Partial Differential Equations
SIAM Monographs on Mathematical Modeling and Computation

About the Series

In 1997, SIAM began a new series on mathematical modeling and computation. Books in the series develop a focused topic from its genesis to the current state of the art; these books

• present modern mathematical developments with direct applications in science and engineering;
• describe mathematical issues arising in modern applications;
• develop mathematical models of topical physical, chemical, or biological systems;
• present new and efficient computational tools and techniques that have direct applications in science and engineering; and
• illustrate the continuing, integrated roles of mathematical, scientific, and computational investigation.

Although sophisticated ideas are presented, the writing style is popular rather than formal. Texts are intended to be read by audiences with little more than a bachelor’s degree in mathematics or engineering. Thus, they are suitable for use in graduate mathematics, science, and engineering courses.

By design, the material is multidisciplinary. As such, we hope to foster cooperation and collaboration between mathematicians, computer scientists, engineers, and scientists. This is a difficult task because different terminology is used for the same concept in different disciplines. Nevertheless, we believe we have been successful and hope that you enjoy the texts in the series.

Mark Holmes

Editor-in-Chief

Mark Holmes
Rensselaer Polytechnic Institute

Editorial Board

Ivo Babuska
University of Texas at Austin
H. Thomas Banks
North Carolina State University
Margaret Cheney
Rensselaer Polytechnic Institute
Paul Davis
Worcester Polytechnic Institute
Stephen H. Davis
Northwestern University
Jack J. Dongarra
University of Tennessee at Knoxville and Oak Ridge National Laboratory
Christopher Hoffmann
Purdue University
George M. Homsy
Stanford University
Joseph B. Keller
Stanford University
J. Tinsley Oden
University of Texas at Austin
James Sethian
University of California at Berkeley
Barna A. Szabo
Washington University

Johnny T. Ottesen, Mette S. Olufsen, and Jesper K. Larsen, Applied Mathematical Models in Human Physiology
Ingemar Kaj, Stochastic Modeling in Broadband Communications Systems
Peter Salamon, Paolo Sibani, and Richard Frost, Facts, Conjectures, and Improvements for Simulated Annealing
Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook, Credit Scoring and Its Applications
Frank Natterer and Frank Wübbeling, Mathematical Methods in Image Reconstruction
Per Christian Hansen, Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion
Michael Griebel, Thomas Dornseifer, and Tilman Neuhausseifer, Numerical Simulation in Fluid Dynamics: A Practical Introduction
Khosrow Chadan, David Colton, Lassi Paivarinta, and William Rundell, An Introduction to Inverse Scattering and Inverse Spectral Problems
Charles K. Chui, Wavelets: A Mathematical Tool for Signal Analysis
Partial Differential Equations
Modeling, Analysis, Computation

R. M. M. Mattheij
S. W. Rienstra
J. H. M. ten Thije Boonkkamp

Technische Universiteit Eindhoven
Eindhoven, The Netherlands

siam
Society for Industrial and Applied Mathematics
Philadelphia
Dedicated to Marie-Anne, Rita, and Liesbeth
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>List of Figures</strong></td>
<td></td>
<td>xvii</td>
</tr>
<tr>
<td><strong>List of Tables</strong></td>
<td></td>
<td>xxv</td>
</tr>
<tr>
<td><strong>Notation</strong></td>
<td></td>
<td>xxvii</td>
</tr>
<tr>
<td><strong>Preface</strong></td>
<td></td>
<td>xxxi</td>
</tr>
<tr>
<td>1</td>
<td>Differential and Difference Equations</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Nomenclature</td>
<td>6</td>
</tr>
<tr>
<td>1.3</td>
<td>Difference Equations</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>Discussion</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Characterisation and Classification</td>
<td>13</td>
</tr>
<tr>
<td>2.1</td>
<td>First Order Scalar PDEs in Two Independent Variables</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>First Order Linear Systems in Two Independent Variables</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>Second Order Scalar PDEs in Two Independent Variables</td>
<td>20</td>
</tr>
<tr>
<td>2.4</td>
<td>Linear Second Order Equations in Several Space Variables</td>
<td>23</td>
</tr>
<tr>
<td>2.5</td>
<td>Reduction to ODEs; Similarity Solutions</td>
<td>24</td>
</tr>
<tr>
<td>2.6</td>
<td>Initial and Boundary Conditions; Well-Posedness</td>
<td>27</td>
</tr>
<tr>
<td>2.7</td>
<td>Discussion</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>Fourier Theory</td>
<td>33</td>
</tr>
<tr>
<td>3.1</td>
<td>Fourier Series</td>
<td>33</td>
</tr>
<tr>
<td>3.2</td>
<td>Fourier Transforms</td>
<td>38</td>
</tr>
<tr>
<td>3.3</td>
<td>Discrete Fourier Transforms</td>
<td>42</td>
</tr>
<tr>
<td>3.4</td>
<td>Fourier Analysis Applied to PDEs</td>
<td>44</td>
</tr>
<tr>
<td>3.5</td>
<td>Discussion</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>47</td>
</tr>
</tbody>
</table>
4 Distributions and Fundamental Solutions ........................................... 51
  4.1 Introduction ............................................................................ 51
  4.2 Distributions in One Variable .............................................. 53
  4.3 Distributions in Several Variables ......................................... 57
  4.4 Strong and Weak Solutions .................................................. 58
  4.5 Fundamental Solutions ........................................................ 61
  4.6 Initial (Boundary) Value Problems; Duhamel Integrals .......... 63
  4.7 Discussion ............................................................................. 67
Exercises .................................................................................. 68

5 Approximation by Finite Differences .............................................. 71
  5.1 Basic Methods ................................................................. 71
    5.1.1 Interpolation, Numerical Differentiation, and Quadrature . . 71
    5.1.2 Finite Difference Equations ......................................... 76
  5.2 Finite Difference Methods for Spatial Variables ..................... 79
    5.2.1 Finite Difference Approximations .................................. 79
    5.2.2 Finite Difference Approximations in Several Dimensions . . 81
    5.2.3 Other Coordinate Systems ............................................ 82
  5.3 Finite Volume Methods ....................................................... 84
    5.3.1 The One-Dimensional Case .......................................... 85
    5.3.2 The Multidimensional Case .......................................... 86
  5.4 Difference Methods for Initial Value Problems ....................... 88
    5.4.1 One-Step and Multistep Difference Schemes ................. 88
    5.4.2 Time Scales and Stiffness ........................................... 92
  5.5 Discretisation, Convergence, and Stability .............................. 96
    5.5.1 The Method of Lines ................................................ 96
    5.5.2 Consistency, Stability, and Convergence of the MOL ....... 97
  5.6 Fourier Mode Analysis .......................................................... 101
  5.7 Discussion ........................................................................... 105
Exercises .................................................................................. 105

6 The Equations of Continuum Mechanics and Electromagnetics ......... 109
  6.1 Introduction ........................................................................... 109
  6.2 Eulerian and Lagrangian Coordinates ..................................... 110
  6.3 The Transport Theorem ....................................................... 111
  6.4 Conservation Equations ....................................................... 112
  6.5 Conservation of Mass .......................................................... 112
  6.6 Conservation of Momentum ............................................... 113
  6.7 Conservation of Energy ....................................................... 115
  6.8 Constitutive Equations and Thermodynamic Relations .......... 116
    6.8.1 Heat Conduction and Mass Diffusion ............................ 117
    6.8.2 Newtonian Viscous Fluid .......................................... 118
    6.8.3 Linear Elastic and Viscoelastic Deformations ............... 120
  6.9 Maxwell Equations ............................................................. 122
    6.9.1 Constitutive Relations .............................................. 123
6.9.2 Energy Conservation and Poynting's Theorem .......... 124
6.9.3 Electromagnetic Waves and Lorentz's Force .......... 124
6.10 Discussion .................................. 125
Exercises ..................................... 126

7 The Art of Modeling ................................ 129
7.1 Introduction .................................. 129
7.2 Models ....................................... 131
  7.2.1 Systematic Models ...................... 131
  7.2.2 Constructing Models ..................... 132
  7.2.3 Canonical Models ....................... 135
7.3 Nondimensionalisation and Scaling ..................... 137
  7.3.1 General Concepts ....................... 137
  7.3.2 Dimensional Analysis .................... 140
  7.3.3 Similarity Solutions ..................... 143
7.4 Scaling and Reduction of the Navier–Stokes Equations .... 148
  7.4.1 Scaling; Nondimensionalisation ............... 148
  7.4.2 Some Dimensionless Groups with Their Common Names .. 149
  7.4.3 Asymptotic Reductions of the Navier–Stokes Equations .. 151
7.5 Discussion .................................. 155
Exercises ..................................... 156

8 The Analysis of Elliptic Equations ....................... 159
8.1 The Laplace Operator .......................... 159
  8.1.1 Problem Types ........................ 160
  8.1.2 Uniqueness .......................... 163
8.2 Eigenvalues and Eigenfunctions ..................... 163
  8.2.1 The One-Dimensional Eigenvalue Problem ......... 163
  8.2.2 Eigenvalue Problems in Several Dimensions ........ 166
8.3 Separation of Variables ........................ 168
8.4 Fundamental Solutions ........................ 171
8.5 Green's Functions; Superposition .................... 173
8.6 The Maximum Principle ........................ 177
8.7 The Stokes Equations .......................... 178
8.8 Discussion .................................. 182
Exercises ..................................... 182

9 Numerical Methods for Elliptic Equations .................. 187
9.1 Discretisation and Boundary Conditions ................. 187
9.2 The Maximum Principle .......................... 189
9.3 Estimates of the Global Error ........................ 193
  9.3.1 Estimates Based on the Maximum Principle .......... 193
  9.3.2 Error Estimates Using the Matrix Method .......... 197
  9.3.3 The Effect of Errors Originating from the Boundary .. 202
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.4</td>
<td>Solution of Linear Systems</td>
<td>205</td>
</tr>
<tr>
<td>9.4.1</td>
<td>One-Dimensional Dirichlet Problems</td>
<td>205</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Basic Iterative Methods</td>
<td>206</td>
</tr>
<tr>
<td>9.4.3</td>
<td>Closure of Iterative Methods; Operator Splitting</td>
<td>211</td>
</tr>
<tr>
<td>9.5</td>
<td>Green's Functions</td>
<td>213</td>
</tr>
<tr>
<td>9.6</td>
<td>Nonlinear Problems</td>
<td>216</td>
</tr>
<tr>
<td>9.6.1</td>
<td>Newton Iteration</td>
<td>216</td>
</tr>
<tr>
<td>9.6.2</td>
<td>Gauss–Jacobi Iteration</td>
<td>219</td>
</tr>
<tr>
<td>9.6.3</td>
<td>Transient Methods</td>
<td>221</td>
</tr>
<tr>
<td>9.7</td>
<td>A Pressure Correction Method for the Stokes Equations</td>
<td>222</td>
</tr>
<tr>
<td>9.8</td>
<td>Discussion</td>
<td>227</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>228</td>
</tr>
<tr>
<td>10</td>
<td>Analysis of Parabolic Equations</td>
<td>233</td>
</tr>
<tr>
<td>10.1</td>
<td>Cauchy Problems</td>
<td>233</td>
</tr>
<tr>
<td>10.1.1</td>
<td>The Heat Equation in One Space Dimension</td>
<td>233</td>
</tr>
<tr>
<td>10.1.2</td>
<td>The Heat Equation in $d$ Space Dimensions</td>
<td>236</td>
</tr>
<tr>
<td>10.1.3</td>
<td>Problems on Half-Spaces</td>
<td>237</td>
</tr>
<tr>
<td>10.2</td>
<td>The Heat Equation with Spatial Symmetries</td>
<td>240</td>
</tr>
<tr>
<td>10.3</td>
<td>Similarity Solutions</td>
<td>241</td>
</tr>
<tr>
<td>10.4</td>
<td>Initial Boundary Value Problems</td>
<td>243</td>
</tr>
<tr>
<td>10.5</td>
<td>Moving Boundaries; Stefan Problems</td>
<td>245</td>
</tr>
<tr>
<td>10.6</td>
<td>Long-Time Behaviour of Solutions</td>
<td>248</td>
</tr>
<tr>
<td>10.6.1</td>
<td>Linear Initial Boundary Value Problem</td>
<td>249</td>
</tr>
<tr>
<td>10.6.2</td>
<td>Equilibrium and Travelling-Wave Solutions for Nonlinear Problems</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>255</td>
</tr>
<tr>
<td>11</td>
<td>Numerical Methods for Parabolic Equations</td>
<td>259</td>
</tr>
<tr>
<td>11.1</td>
<td>The Explicit Euler Scheme</td>
<td>259</td>
</tr>
<tr>
<td>11.2</td>
<td>Semidiscretisation</td>
<td>265</td>
</tr>
<tr>
<td>11.2.1</td>
<td>The Longitudinal Method of Lines</td>
<td>265</td>
</tr>
<tr>
<td>11.2.2</td>
<td>The Transversal MOL</td>
<td>268</td>
</tr>
<tr>
<td>11.3</td>
<td>Implicit Schemes</td>
<td>269</td>
</tr>
<tr>
<td>11.3.1</td>
<td>The Implicit Euler Scheme</td>
<td>269</td>
</tr>
<tr>
<td>11.3.2</td>
<td>The $\theta$ Scheme</td>
<td>271</td>
</tr>
<tr>
<td>11.4</td>
<td>Analysis of the $\theta$ Method by the Matrix Method</td>
<td>276</td>
</tr>
<tr>
<td>11.5</td>
<td>Initial Boundary Value Problems with Discontinuous Data</td>
<td>280</td>
</tr>
<tr>
<td>11.6</td>
<td>Mixed Boundary Conditions</td>
<td>282</td>
</tr>
<tr>
<td>11.7</td>
<td>Problems in Two Space Dimensions</td>
<td>287</td>
</tr>
<tr>
<td>11.8</td>
<td>Splitting Methods</td>
<td>290</td>
</tr>
<tr>
<td>11.8.1</td>
<td>The ADI Method</td>
<td>291</td>
</tr>
<tr>
<td>11.8.2</td>
<td>Mixed Explicit/Implicit Discretisation of the Advection-Diffusion</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>Equation</td>
<td>296</td>
</tr>
<tr>
<td>11.9</td>
<td>Solving Nonlinear Problems</td>
<td>296</td>
</tr>
</tbody>
</table>
## Contents

### 11.10 Stefan Problems ............................. 298
### 11.11 Discussion ................................ 303
### Exercises ..................................... 303

### 12 Analysis of Hyperbolic Equations 307

#### 12.1 First Order Scalar Equations .......................... 307
   12.1.1 Semilinear Equations .......................... 308
   12.1.2 Quasi-linear Equations .......................... 310
   12.1.3 Nonlinear Equations .......................... 314

#### 12.2 Weak Formulation of First Order Scalar Equations ........... 317
   12.2.1 Weak Solutions ............................. 317
   12.2.2 The Riemann Problem .......................... 322

#### 12.3 First Order Systems ........................... 327
   12.3.1 Linear Systems ............................. 327
   12.3.2 Quasi-linear Systems .......................... 330
   12.3.3 Method of Characteristics ....................... 333

#### 12.4 Weak Formulation of First Order Systems ............... 334
   12.4.1 Weak Solutions ............................. 334
   12.4.2 The Riemann Problem .......................... 336

#### 12.5 The Shallow-Water Equations ..................... 342

#### 12.6 The Wave Equation ........................... 349
   12.6.1 One-Dimensional Problems ....................... 349
   12.6.2 Solutions in Several Dimensions ............... 351

#### 12.7 Boundary Conditions .......................... 354

#### 12.8 Discussion ................................ 359
### Exercises ..................................... 359

### 13 Numerical Methods for Scalar Hyperbolic Equations 363

#### 13.1 Explicit One-Step Schemes for the Advection Equation .. 363
   13.1.1 The Upwind Scheme .......................... 364
   13.1.2 The Lax–Wendroff Scheme ....................... 368

#### 13.2 Dissipation and Dispersion of Numerical Schemes ........... 370

#### 13.3 The Advection-Diffusion Equation .................... 373
   13.3.1 An Explicit Scheme for the Advection-Diffusion Equation . 373
   13.3.2 Numerical and Artificial Diffusion ................ 376

#### 13.4 Nondissipative Schemes ........................ 378
   13.4.1 The Box Scheme ............................. 379
   13.4.2 The Leapfrog Scheme .......................... 381
   13.4.3 Propagation of Wave Packets ..................... 384

#### 13.5 The Godunov Scheme for Nonlinear Conservation Laws ....... 385

#### 13.6 High-Resolution Schemes ........................ 391

#### 13.7 A Flux Limiter Scheme for the Advection Equation .......... 396

#### 13.8 Slope Limiter Methods .......................... 402
   13.8.1 A Slope Limiter Method for the Advection Equation .... 403

#### 13.9 Numerical Boundary Conditions .......................... 408
13.10 Discussion ................................ 413
Exercises ..................................... 414

14 Numerical Methods for Hyperbolic Systems 417
14.1 The Upwind Scheme ........................ 417
14.2 The Godunov Scheme ........................ 421
14.2.1 Derivation of the Scheme ............... 421
14.2.2 The Linear Case ....................... 423
14.3 Roe’s Approximate Riemann Solver ........... 425
14.4 Slope Limiter Methods ..................... 430
14.4.1 A Slope Limiter Method for Linear Systems ......... 430
14.4.2 A Slope Limiter Method for Nonlinear Systems ....... 432
14.5 Numerical Solution of the Shallow-Water Equations ........ 434
14.5.1 Numerical Solution by Godunov’s Method ......... 434
14.5.2 Numerical Solution by Roe’s Method ............. 436
14.6 Numerical Solution of the Wave Equation ........... 440
14.6.1 Difference Methods Based on the Scalar Form ........ 440
14.6.2 The Staggered Leapfrog Scheme ............... 444
14.7 Numerical Boundary Conditions .................. 447
14.8 Discussion ................................ 456
Exercises ..................................... 456

15 Perturbation Methods 459
15.1 Introduction ................................ 459
15.2 Asymptotic Approximations and Expansions .......... 461
15.2.1 Asymptotic Approximations ................. 461
15.2.2 Asymptotic Expansions .................... 462
15.2.3 Perturbation Problems ..................... 464
15.2.4 Asymptotic Expansions of Poincaré Type .......... 466
15.3 Regular Perturbation Problems ................ 468
15.3.1 Method of Slow Variation .................. 468
15.3.2 Lindstedt–Poincaré Method ................. 473
15.4 Singular Perturbation Problems ............... 475
15.4.1 Matched Asymptotic Expansions ............. 475
15.4.2 Multiple Scales ........................ 483
15.5 Discussion ................................ 494
Exercises ..................................... 495

16 Modeling, Analysing, and Simulating Problems from Practice 501
16.1 Production of Resin-Containing Panels ............. 501
16.1.1 Modeling the Curing Process ............... 502
16.1.2 Numerical Solution Method ................ 504
16.1.3 Discussion and Related Problems .......... 506
16.1.4 Exercises ........................... 507
16.2 Mechanical Etching of Glass by Powder Blasting ....... 508
16.2.1 Problem Setting ........................ 508
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.2.2</td>
<td>Mathematical Model for Powder Erosion</td>
<td>509</td>
</tr>
<tr>
<td>16.2.3</td>
<td>Characteristic Strip Equations</td>
<td>511</td>
</tr>
<tr>
<td>16.2.4</td>
<td>Solution of the Characteristic Strip Equations</td>
<td>512</td>
</tr>
<tr>
<td>16.2.5</td>
<td>Discussion and Related Problems</td>
<td>516</td>
</tr>
<tr>
<td>16.2.6</td>
<td>Exercises</td>
<td>516</td>
</tr>
<tr>
<td>16.3</td>
<td>Thermal Explosion in a Vessel</td>
<td>517</td>
</tr>
<tr>
<td>16.3.1</td>
<td>Problem Formulation</td>
<td>517</td>
</tr>
<tr>
<td>16.3.2</td>
<td>Mathematical Model</td>
<td>517</td>
</tr>
<tr>
<td>16.3.3</td>
<td>Estimating the Induction Time</td>
<td>518</td>
</tr>
<tr>
<td>16.3.4</td>
<td>Numerical Solution Method</td>
<td>519</td>
</tr>
<tr>
<td>16.3.5</td>
<td>Discussion and Related Problems</td>
<td>521</td>
</tr>
<tr>
<td>16.3.6</td>
<td>Exercises</td>
<td>523</td>
</tr>
<tr>
<td>16.4</td>
<td>Determining Viscoelastic Material Parameters</td>
<td>523</td>
</tr>
<tr>
<td>16.4.1</td>
<td>Problem Formulation</td>
<td>523</td>
</tr>
<tr>
<td>16.4.2</td>
<td>The Model</td>
<td>524</td>
</tr>
<tr>
<td>16.4.3</td>
<td>Analysis and Solution</td>
<td>526</td>
</tr>
<tr>
<td>16.4.4</td>
<td>Example</td>
<td>529</td>
</tr>
<tr>
<td>16.4.5</td>
<td>Discussion and Related Problems</td>
<td>529</td>
</tr>
<tr>
<td>16.4.6</td>
<td>Exercises</td>
<td>530</td>
</tr>
<tr>
<td>16.5</td>
<td>Galloping Transmission Lines</td>
<td>531</td>
</tr>
<tr>
<td>16.5.1</td>
<td>Problem Formulation</td>
<td>531</td>
</tr>
<tr>
<td>16.5.2</td>
<td>The Model</td>
<td>531</td>
</tr>
<tr>
<td>16.5.3</td>
<td>Asymptotic Analysis</td>
<td>534</td>
</tr>
<tr>
<td>16.5.4</td>
<td>Solutions</td>
<td>537</td>
</tr>
<tr>
<td>16.5.5</td>
<td>Examples</td>
<td>538</td>
</tr>
<tr>
<td>16.5.6</td>
<td>Discussion and Related Problems</td>
<td>538</td>
</tr>
<tr>
<td>16.5.7</td>
<td>Exercises</td>
<td>540</td>
</tr>
<tr>
<td>16.6</td>
<td>Groundwater Flow and Rain</td>
<td>540</td>
</tr>
<tr>
<td>16.6.1</td>
<td>Problem Formulation</td>
<td>540</td>
</tr>
<tr>
<td>16.6.2</td>
<td>The Model</td>
<td>540</td>
</tr>
<tr>
<td>16.6.3</td>
<td>Analysis</td>
<td>542</td>
</tr>
<tr>
<td>16.6.4</td>
<td>Discussion and Related Problems</td>
<td>545</td>
</tr>
<tr>
<td>16.6.5</td>
<td>Exercises</td>
<td>546</td>
</tr>
<tr>
<td>16.7</td>
<td>Cooling a Monocrystalline Bar</td>
<td>547</td>
</tr>
<tr>
<td>16.7.1</td>
<td>Problem Formulation</td>
<td>547</td>
</tr>
<tr>
<td>16.7.2</td>
<td>The Model</td>
<td>547</td>
</tr>
<tr>
<td>16.7.3</td>
<td>Asymptotic Analysis</td>
<td>549</td>
</tr>
<tr>
<td>16.7.4</td>
<td>Numerical Solution Method</td>
<td>553</td>
</tr>
<tr>
<td>16.7.5</td>
<td>Discussion and Related Problems</td>
<td>554</td>
</tr>
<tr>
<td>16.7.6</td>
<td>Exercises</td>
<td>554</td>
</tr>
<tr>
<td>16.8</td>
<td>A Catalytic Reaction Problem in Pellets</td>
<td>555</td>
</tr>
<tr>
<td>16.8.1</td>
<td>Problem Formulation</td>
<td>555</td>
</tr>
<tr>
<td>16.8.2</td>
<td>Numerical Solution Method</td>
<td>557</td>
</tr>
<tr>
<td>16.8.3</td>
<td>Asymptotic Analysis</td>
<td>558</td>
</tr>
<tr>
<td>16.8.4</td>
<td>Discussion and Related Problems</td>
<td>561</td>
</tr>
<tr>
<td>16.8.5</td>
<td>Exercises</td>
<td>562</td>
</tr>
</tbody>
</table>
## Contents

### 16.9 Outdoor Noise Enhancement by Atmospheric Conditions
- 16.9.1 Problem Formulation ........................................ 563
- 16.9.2 The Model .................................................. 563
- 16.9.3 Asymptotic Analysis ...................................... 564
- 16.9.4 Solutions ..................................................... 568
- 16.9.5 Discussion and Related Problems ...................... 569
- 16.9.6 Exercises ..................................................... 570

### 16.10 Thin-Layer Flow Along a Curved Surface
- 16.10.1 The Model .................................................. 571
- 16.10.2 Curvilinear Coordinates ................................ 571
- 16.10.3 Perturbation Analysis ................................... 572
- 16.10.4 The Inclined Plane Surface ............................ 576
- 16.10.5 The Inclined Cylinder .................................. 578
- 16.10.6 Discussion and Related Problems ...................... 580
- 16.10.7 Exercises ..................................................... 581

### 16.11 Forming Container Glass
- 16.11.1 Problem Formulation ...................................... 581
- 16.11.2 Governing Equations .................................... 582
- 16.11.3 Slender Geometry Approximation .................... 583
- 16.11.4 Constant Temperature ................................... 584
- 16.11.5 Boundary Conditions .................................... 585
- 16.11.6 Solution ..................................................... 586
- 16.11.7 Example ..................................................... 588
- 16.11.8 Discussion and Related Problems ...................... 589
- 16.11.9 Exercises ..................................................... 589

### 16.12 Laser Percussion Drilling
- 16.12.1 Problem Formulation ...................................... 589
- 16.12.2 The Model ..................................................... 590
- 16.12.3 The Stefan Problem ...................................... 592
- 16.12.4 Finding Suitable Initial Conditions .................. 594
- 16.12.5 The Enthalpy Formulation .............................. 595
- 16.12.6 Discussion and Related Problems ...................... 598
- 16.12.7 Exercises ..................................................... 598

### 16.13 Determining Chemical Composition by Electrophoresis
- 16.13.1 Introduction .................................................. 599
- 16.13.2 Mathematical Model of Capillary Electrophoresis ............................. 600
- 16.13.3 Numerical Solution Method ............................ 601
- 16.13.4 A Practical Example ...................................... 602
- 16.13.5 Numerical Simulations .................................... 603
- 16.13.6 Discussion and Related Problems ...................... 605
- 16.13.7 Exercises ..................................................... 605

### 16.14 Pulse Tube Refrigerators
- 16.14.1 Problem Formulation ...................................... 606
- 16.14.2 The Model ..................................................... 607
- 16.14.3 Nondimensionalisation ................................... 607
- 16.14.4 Derivation of Velocity and Temperature Equations ............... 608
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.14.5</td>
<td>Numerical Solution</td>
<td>610</td>
</tr>
<tr>
<td>16.14.6</td>
<td>Discussion and Related Problems</td>
<td>612</td>
</tr>
<tr>
<td>16.14.7</td>
<td>Exercises</td>
<td>613</td>
</tr>
<tr>
<td>16.15</td>
<td>Flow in a Glass Oven</td>
<td>613</td>
</tr>
<tr>
<td>16.15.1</td>
<td>Problem Formulation</td>
<td>613</td>
</tr>
<tr>
<td>16.15.2</td>
<td>The Model</td>
<td>614</td>
</tr>
<tr>
<td>16.15.3</td>
<td>Defining the Domain</td>
<td>615</td>
</tr>
<tr>
<td>16.15.4</td>
<td>Discretising the Convection-Diffusion Operator</td>
<td>617</td>
</tr>
<tr>
<td>16.15.5</td>
<td>Discretising the Pressure Gradient and the Continuity Equation</td>
<td>619</td>
</tr>
<tr>
<td>16.15.6</td>
<td>Simulations</td>
<td>620</td>
</tr>
<tr>
<td>16.15.7</td>
<td>Discussion and Related Problems</td>
<td>621</td>
</tr>
<tr>
<td>16.15.8</td>
<td>Exercises</td>
<td>621</td>
</tr>
<tr>
<td>16.16</td>
<td>Attenuation of Sound in Aircraft Engine Ducts</td>
<td>622</td>
</tr>
<tr>
<td>16.16.1</td>
<td>Problem Formulation</td>
<td>622</td>
</tr>
<tr>
<td>16.16.2</td>
<td>The Model</td>
<td>622</td>
</tr>
<tr>
<td>16.16.3</td>
<td>Asymptotic Solution</td>
<td>623</td>
</tr>
<tr>
<td>16.16.4</td>
<td>Numerical Evaluation of a Typical Example</td>
<td>626</td>
</tr>
<tr>
<td>16.16.5</td>
<td>Discussion and Related Problems</td>
<td>628</td>
</tr>
<tr>
<td>16.16.6</td>
<td>Exercises</td>
<td>628</td>
</tr>
</tbody>
</table>

### Appendices. Useful Definitions and Properties

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Asymptotic Order Symbols</td>
<td>629</td>
</tr>
<tr>
<td>B</td>
<td>Trigonometric Relations</td>
<td>630</td>
</tr>
<tr>
<td>C</td>
<td>Convergence of Series</td>
<td>630</td>
</tr>
<tr>
<td>D</td>
<td>Multistep Formulas</td>
<td>631</td>
</tr>
<tr>
<td>E</td>
<td>Solution of Recursions</td>
<td>635</td>
</tr>
<tr>
<td>F</td>
<td>Eigenvalues and Eigenvectors of a Tridiagonal Matrix</td>
<td>635</td>
</tr>
<tr>
<td>G</td>
<td>Norms</td>
<td>637</td>
</tr>
<tr>
<td>H</td>
<td>Similarity</td>
<td>639</td>
</tr>
<tr>
<td>I</td>
<td>Estimates of Eigenvalues and Consequences</td>
<td>640</td>
</tr>
<tr>
<td>J</td>
<td>Theorems from Vector Calculus</td>
<td>642</td>
</tr>
<tr>
<td>K</td>
<td>Cartesian, Cylindrical, and Spherical Coordinates</td>
<td>643</td>
</tr>
<tr>
<td>L</td>
<td>Tensors</td>
<td>644</td>
</tr>
<tr>
<td>M</td>
<td>Dimensionless Numbers</td>
<td>646</td>
</tr>
</tbody>
</table>

### Bibliography

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>649</td>
</tr>
</tbody>
</table>

### Index

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>659</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Sketch of dye diffusion. ........................................ 2
1.2 Sketch of traffic flow. .......................................... 3
1.3 Chain of coupled springs. ..................................... 4
1.4 A transmission line model of a telegraph wire. ............. 5
1.5 An array of accommodating individuals. ....................... 6
1.6 Stencil of Example 1.9(i). ..................................... 9
1.7 Stencil of Example 1.9(ii). ..................................... 9
1.8 Stencil of Example 1.10(ii). ................................... 10

2.1 Initial curve $J$ and a characteristic $C$ on the integral surface $S$. .... 14
2.2 Base characteristics of $(\ast)$ of Example 2.2. ............... 17
2.3 Region of influence and region of dependence. In the case of constant coefficients these characteristics are straight lines. .......... 19

3.1 Sawtooth function. ........................................... 36
3.2 Example of Gibbs's phenomenon (50 terms). ................. 37
3.3 Aliasing. .................................................... 44

4.1 A top hat delta sequence. .................................... 56
4.2 An exponential delta sequence ($\lambda = 0.01, 0.04, 0.09, 0.25, 0.49, 1.0$). .. 57

5.1 Stencil of the Laplace operator in compass notation (left) and index notation (right). ........................................... 81
5.2 Stencil of the Laplace operator near a boundary. ............. 82
5.3 Cylindrical (left) and spherical (right) coordinates. ......... 84
5.4 Control volume for a cell-centered (left) and a vertex-centered (right) finite volume method. ........................................... 85
5.5 Two-dimensional control volume for a cell-centered finite volume method. 87
5.6 The solution (5.84) on the intervals (0, 1) (left) and (0, 0.02) (right). .. 93
5.7 Exact and numerical solutions of (5.83a) with initial condition $u(0) = 1$. The numerical solution is computed with the forward Euler method with $\Delta t = 1/45$. ........................................... 94
5.8 Grids for the longitudinal (left) and transversal (right) method of lines. 96
xviii List of Figures

6.1 Contact angle. ........................................ 120

7.1 Sketch of bar element. Side view (i) and cross section (ii). .......... 134

7.2 A temperature distribution. ................................ 139

7.3 A piece of metal heated by an electric field. ......................... 140

7.4 A semi-infinite bar heated from $x = 0$ ................................ 144

7.5 A wedge-shaped conductor heated by an electric field. ................. 146

7.6 Radial temperature distribution in wedge of $v = \frac{1}{2} \pi$ (left) and $v = \frac{3}{2} \pi$ (right) for $\sigma A^2/4k = 1$ and $4k \tau /\rho C = \frac{1}{4}, \frac{3}{4}, \frac{2}{4}, ...$ ............... 147

7.7 A Joukowski airfoil. .................................... 154

8.1 The eigenvalues contained in the first quadrant of the circle $k^2 + l^2 < |\lambda|/\pi$. .................................. 168

8.2 A point $\xi \in \Omega$ and its mirror point $\xi^*$. .......................... 174

8.3 Asymptotic boundary condition for the velocity $u$. ...................... 181

9.1 Five-point stencil for the difference equation (9.2). ..................... 188

9.2 Stencil for the Robin boundary condition (9.5) without (left) and with (right) a virtual grid point $W$. ................................. 189

9.3 An unconnected grid; grid point $x_c$ is not connected to any other grid point. .......................................................... 190

9.4 Stencil at the left boundary. .................................... 201

9.5 The eigenvalues of $A$ compared with the exact eigenvalues for $M = 20$. .......................................................... 207

9.6 Lexicographic ordering of grid points for $M = 4$. ....................... 208


9.8 Five-point stencil for (9.115). ..................................... 225

9.9 A patch of a staggered grid in the interior (left) and near a boundary (right). ......................................................... 225

9.10 The velocity (top) and the pressure field (bottom) of a Stokes flow in a rectangular domain. ............................................ 227

10.1 Solution of the heat shock problem at various time levels. .............. 239

10.2 Similarity coordinate $y$ of the interface as a function of the Stefan problem parameter $\alpha$. ........................................ 247

10.3 Enthalpy function $H(T)$. ...................................... 248

10.4 Phase portrait of Fisher’s travelling-wave problem (10.71) with $c = 2.25$. Arrows indicate the positive $\xi$ direction. Note the trajectory that connects saddle point $(1, 0)^T$ with stable node $(0, 0)^T$. ...................... 254

10.5 The (numerically obtained) travelling-wave solution of Fisher’s problem (10.69), (10.70) with $c = 2.25$ and $U(0) = \frac{1}{2}$. ....................... 255

11.1 Stencil for the explicit scheme (11.4). .................................. 260
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.2</td>
<td>Numerical solution of the heat equation computed with the explicit Euler scheme with time steps $\Delta t = 1.2 \times 10^{-3}$ (left) and $\Delta t = 1.3 \times 10^{-3}$ (right).</td>
<td>262</td>
</tr>
<tr>
<td>11.3</td>
<td>A solution having a sharp moving front of the heat equation with source term at three different time levels.</td>
<td>269</td>
</tr>
<tr>
<td>11.4</td>
<td>Stencil for the implicit scheme (11.41).</td>
<td>270</td>
</tr>
<tr>
<td>11.5</td>
<td>Numerical solution of the heat equation computed with the implicit Euler scheme with time steps $\Delta t = 1.2 \times 10^{-3}$ (left) and $\Delta t = 1.3 \times 10^{-3}$ (right).</td>
<td>272</td>
</tr>
<tr>
<td>11.6</td>
<td>Stencil for the $\theta$ method (11.49).</td>
<td>273</td>
</tr>
<tr>
<td>11.7</td>
<td>Numerical solution of the heat equation with discontinuous data, computed with the $\theta$ scheme.</td>
<td>275</td>
</tr>
<tr>
<td>11.8</td>
<td>Eigenvalues of the matrices $B_{\theta}$ (left) and $C$ (right) for $\theta = 0.5$ (□), 0.75 (○), and 1 (△).</td>
<td>278</td>
</tr>
<tr>
<td>11.9</td>
<td>Amplification factor $e^z$ compared with the rational function $r(z)$ for $\theta = 0.5, 0.75,$ and 1.</td>
<td>282</td>
</tr>
<tr>
<td>11.10</td>
<td>A two-dimensional spatial grid for initial boundary value problem (11.110) with $M = 5$ and $N = 4$.</td>
<td>289</td>
</tr>
<tr>
<td>11.11</td>
<td>Patch of the grid with a moving interface $x = S(t)$.</td>
<td>299</td>
</tr>
<tr>
<td>11.12</td>
<td>Numerical solution of the Stefan problem (11.151).</td>
<td>302</td>
</tr>
<tr>
<td>12.1</td>
<td>Characteristics of the Burgers’ equation that intersect (left) or fan out (right).</td>
<td>312</td>
</tr>
<tr>
<td>12.2</td>
<td>The solution of (12.12) at $t = 0$ and $0.2$ and $t^* = 1/\pi$ and $0.8$ for the initial condition $v(x) = \sin \pi x$.</td>
<td>313</td>
</tr>
<tr>
<td>12.3</td>
<td>Weak solution, discontinuous across $E$.</td>
<td>319</td>
</tr>
<tr>
<td>12.4</td>
<td>Illustration of the equal-area rule: The multivalued “solution” on the left should be replaced by the shock on the right.</td>
<td>321</td>
</tr>
<tr>
<td>12.5</td>
<td>Shock wave and corresponding characteristics.</td>
<td>322</td>
</tr>
<tr>
<td>12.6</td>
<td>Rarefaction wave and corresponding characteristics.</td>
<td>323</td>
</tr>
<tr>
<td>12.7</td>
<td>Characteristics corresponding to an expansion shock of the Burgers’ equation.</td>
<td>324</td>
</tr>
<tr>
<td>12.8</td>
<td>Flux function and solution of the Buckley–Leverett equation.</td>
<td>326</td>
</tr>
<tr>
<td>12.9</td>
<td>Illustration of the method of Massau.</td>
<td>334</td>
</tr>
<tr>
<td>12.10</td>
<td>Similarity solution of the Riemann problem for a 3 \times 3 linear system.</td>
<td>338</td>
</tr>
<tr>
<td></td>
<td>The triple $(\beta_1, \alpha_2, \alpha_3)$ denotes the solution $u = \beta_1 s_1 + \alpha_2 s_2 + \alpha_3 s_3$, etc.</td>
<td></td>
</tr>
<tr>
<td>12.11</td>
<td>Wave pattern of a $k$-simple wave.</td>
<td>339</td>
</tr>
<tr>
<td>12.12</td>
<td>Wave pattern of a contact discontinuity.</td>
<td>340</td>
</tr>
<tr>
<td>12.13</td>
<td>Wave pattern of a $k$-shock.</td>
<td>341</td>
</tr>
<tr>
<td>12.14</td>
<td>Possible wave patterns of the Riemann problem for the shallow-water equations.</td>
<td>343</td>
</tr>
<tr>
<td>12.15</td>
<td>The 1-wave is either a shock (left) or a rarefaction wave (right).</td>
<td>344</td>
</tr>
<tr>
<td>12.16</td>
<td>The function $g(z)$.</td>
<td>345</td>
</tr>
<tr>
<td>12.17</td>
<td>The 2-wave is either a shock (left) or a rarefaction wave (right).</td>
<td>346</td>
</tr>
<tr>
<td>12.18</td>
<td>Characteristics of the hyperbolic system (12.171a) corresponding to $\lambda_k &lt; 0$ (left) and $\lambda_k &gt; 0$ (right).</td>
<td>355</td>
</tr>
</tbody>
</table>
13.1 Stencil for the upwind scheme. The dotted line is the characteristic through the point \((x_j, t^{n+1})\) facing back to the initial line \(t = 0\) ........ 365
13.2 Numerical approximation of a sine wave computed with the upwind scheme. .................................................. 367
13.3 Stencil for the Lax–Wendroff scheme. The dotted line is the characteristic through the point \((x_j, t^{n+1})\) facing back to the initial line \(t = 0\) ........ 368
13.4 Numerical approximation of a sine wave computed with the Lax–Wendroff scheme. .................................................. 370
13.5 Amplitude and phase errors of the upwind scheme (left) and the Lax–Wendroff scheme (right) .................. 372
13.6 Numerical solution computed with the upwind scheme (left) and the Lax–Wendroff scheme (right), after 25, 50, and 75 time steps, respectively .................. 374
13.7 Numerical solution of the advection-diffusion equation computed with the central difference scheme. .................. 376
13.8 Numerical solution of the advection-diffusion equation computed with the upwind scheme. .................. 378
13.9 Stencil for the box scheme. The open circles indicate the intermediate grid points ........ 380
13.10 Phase error of the box scheme ........ 381
13.11 Numerical sine wave (left) and step function (right) computed with the box scheme .................. 382
13.12 Stencil for the leapfrog scheme .................. 382
13.13 Relative phase error of the leapfrog scheme .................. 384
13.14 Numerical group velocity of the box scheme .................. 385
13.15 The effect of the numerical group velocity on a wave packet computed with the box scheme .................. 386
13.16 Piecewise constant numerical solution at time level \(t^n\) .................. 387
13.17 A shock wave solution of the Burgers’ equation computed with Godunov’s scheme (left) and the upwind scheme (right) .................. 391
13.18 The van Leer, superbee, and minmod limiters in the TVD region .................. 400
13.19 Piecewise linear numerical solution at time level \(t^n\) .................. 403
13.20 Numerical solution computed with the slope limiter method, after 25, 50, and 75 time steps, respectively .................. 406
13.21 Rarefaction wave computed with the slope limiter method .................. 408
13.22 Numerical solution of the leapfrog scheme, corrupted by the erroneous boundary condition \(u(1, t) = 0\) .................. 409
13.23 Eigenvalues of the leapfrog scheme with upwind numerical boundary condition .................. 413
14.1 Schematic representation of a similarity solution in the \((x, t)\) plane .................. 427
14.2 Numerical approximation of the dam break problem at time levels 0.5 (top) and 1.0 (bottom) .................. 436
14.3 Schematic representation of a critical 1-rarefaction wave for the shallow-water equations .................. 439
14.4 Stencil for the centred scheme (14.102b) .................. 441
14.5 Patch of the staggered grid and stencil for the leapfrog scheme (14.124) .................. 445
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.6</td>
<td>The $C_2$ characteristic through the boundary grid point $(x_{M+1}, t^{n+1})$.</td>
</tr>
<tr>
<td>14.7</td>
<td>Numerical approximation of the geopotential $\varphi = gh$ at time levels $0.3, 0.4, 0.8$.</td>
</tr>
<tr>
<td>14.8</td>
<td>Two time levels in the staggered grid for the leapfrog scheme. The variables $u$ and $r$ are defined at points marked by $\bullet$ and $\circ$, respectively.</td>
</tr>
<tr>
<td>15.1</td>
<td>A plot of $x + \sin(\varepsilon x) + e^{-x/\varepsilon}$ and its nonuniform asymptotic approximation $x + \varepsilon x$ for $\varepsilon = 0.01$.</td>
</tr>
<tr>
<td>15.2</td>
<td>Analysis of distinguished limits.</td>
</tr>
<tr>
<td>15.3</td>
<td>Slowly varying bar.</td>
</tr>
<tr>
<td>15.4</td>
<td>Plots of the approximate solution and a numerically &quot;exact&quot; solution $y(t; \varepsilon)$ of the air-damped resonator problem for $\varepsilon = 0.1$.</td>
</tr>
<tr>
<td>15.5</td>
<td>Pattern formation in glass by masked erosion (courtesy P.J. Slikkerveer, Philips Research Laboratories, Eindhoven).</td>
</tr>
<tr>
<td>15.6</td>
<td>Cross section of a hole in the substrate.</td>
</tr>
<tr>
<td>15.7</td>
<td>Characteristics and shocks of a two-dimensional trench for $\delta = 0.1$ and $k = 2$.</td>
</tr>
<tr>
<td>15.8</td>
<td>Solution for the surface position (left) and its slope (right) of a one-dimensional trench. Parameter values are $\delta = 0.1$ and $k = 2.33$.</td>
</tr>
<tr>
<td>15.9</td>
<td>Analytical solution (left) and measurements (right) of the surface position for a two-dimensional trench at $t = 0.5, 1.1, 1.6, 2.1$ (experimental results by P.J. Slikkerveer [155]).</td>
</tr>
<tr>
<td>15.10</td>
<td>Temperature and concentration distribution in a reactor vessel.</td>
</tr>
<tr>
<td>15.11</td>
<td>Temperature and concentration at the centre of a reactor vessel.</td>
</tr>
<tr>
<td>15.12</td>
<td>Viscoelastic material suddenly compressed.</td>
</tr>
<tr>
<td>15.13</td>
<td>$u_z$, $t_r$, and $t_z$ at $z = L$, $z = 0.9L$, $z = 0.6L$, and $z = 0$, plotted as a function of time. At the left side are the results in the boundary layer, while at the right side are the results for larger time.</td>
</tr>
<tr>
<td>15.14</td>
<td>Impression of a suspended cable consisting of three interconnected spans.</td>
</tr>
<tr>
<td>15.15</td>
<td>Sketch of cable element stretched under tension.</td>
</tr>
<tr>
<td>15.16</td>
<td>Sketch of cables connected at a suspension string.</td>
</tr>
<tr>
<td>15.17</td>
<td>An oscillating cable section (from top to bottom): the vertical displacement $dY$ in $sL$ for various time steps, the position at the middle ($s = \frac{1}{2}$) of the displacement in time, and the varying part of the tension $dT$ in time. All variables are dimensional.</td>
</tr>
<tr>
<td>15.18</td>
<td>Sketch of groundwater and rain problem.</td>
</tr>
<tr>
<td>15.19</td>
<td>Plot of boundary layer profile $\psi_0(\xi)$ for $K = O(1)$.</td>
</tr>
<tr>
<td>15.20</td>
<td>Sketch of a silicon bar.</td>
</tr>
<tr>
<td>15.21</td>
<td>Temperature variation along the monocrystalline bar.</td>
</tr>
</tbody>
</table>
16.22 Temperature distribution in the monocrystalline bar. ................................. 554
16.23 A spherical catalyst pellet. ............................................................. 556
16.24 $r_1$ as a function of $1/\lambda$. .......................................................... 560
16.25 Leading-order approximation for $\alpha \to 0$ and $\lambda = \mathcal{O}(1)$. ............ 561
16.26 Acoustic refraction in shear flow. ...................................................... 563
16.27 A ray tube. ......................................................................................... 568
16.28 A ray tube starting from a circle of radius $\varepsilon$ in the $YZ$ plane, refracting downward in right-running shear flow. .................................................. 569
16.29 Mass conservation between streamlines .............................................. 575
16.30 Inclined plane surface. ................................................................. 576
16.31 Streamlines along a flat plate. ......................................................... 577
16.32 Inclined cylinder. ............................................................................. 578
16.33 Streamlines along an inclined cylinder issuing from a point source in the wall. Right: analytical solution; left: experiments by one of the authors (SWR). Note the realistic but (strictly speaking) not modeled crossing streamlines! ................................................................. 580
16.34 Sketch of configuration. $R_p = R_p(z - z_p)$ is the plunger surface, $R_m = R_m(z)$ is the mould surface, $z_p(t)$ is the position of the plunger top, and $b(t)$ is the position of the glass surface. ........................................................................... 583
16.35 Sketch of the control surfaces to calculate the axial flux. ....................... 587
16.36 Dimensionless position (left) and minus pressure gradient (right) correspond- ing to a plunger entering a mould fully filled with fluid glass. ......................... 588
16.37 The three events in a laser percussion drilling process where a melting model is needed: (a) melting, (b) splashing, and (c) resolidification. ......................... 590
16.38 The geometry of the laser-induced melting problem. .......................... 592
16.39 The position of the solid-liquid interface for the case of aluminium. The absorption of latent heat is neglected. ......................................................... 594
16.40 Relation between enthalpy and temperature for pure crystalline substances and for glassy substances and alloys. ................................................. 595
16.41 Time history plots for temperature at the interface in the enthalpy problem. ................................................................................................. 597
16.42 Time development of concentration of methyl-hippuric and mandelic acid after (from left to right) 1, 3, and 5 s of migration. ........................................ 603
16.43 Distribution of local pH (lower trace) and local electric field (upper trace) once the steady state is reached. ....................................................... 604
16.44 Migration of mandelic acid from steady state with 2, 5, 10, and 20 s of diffusion (overlapping profiles on the left) and after restoration of the driving current in six 200 ms intervals. ........................................ 604
16.45 A schematic picture of the Stirling-type pulse tube refrigerator: (a) the basic pulse tube and (b) the orifice pulse tube. From right to left the basic system consists of a piston, a regenerator, a cold heat exchanger (CHX), a tube, and a hot heat exchanger (HHX). In (b) the system is extended with an orifice and a reservoir (buffer). ................................................................. 606
16.46 Net heat and enthalpy flows in the tube. .................................................. 607
List of Figures

16.47  Dimensionless pressure variation used in the simulation. ............ 612
16.48  Temperature profiles for the different parts of the pressure cycle. ........ 612
16.49  Sketch of a model furnace. ........................................ 620
16.50  Velocity field as a function of \((x, z)\) for \(y = 2.0\). ................. 621
16.51  Sketch of an aircraft turbofan engine. .................................. 622
16.52  Acoustic pressure field in an aircraft engine inlet duct in the \((x, r)\) plane. 627
i

i

i

pdete
2005/9
page x
i

List of Tables
12.1

Number of boundary conditions for the shallow-water equations.

14.1

Conditions for shocks and rarefaction waves. . . . . . . . . . . . . . . . 434

16.1
16.2
16.3
16.4
16.5

16.7

Parameters involved in the curing process. . . . . . . . . . . . . . . .
Parameters for a thermal explosion. . . . . . . . . . . . . . . . . . . .
Typical parameter values of silicon. . . . . . . . . . . . . . . . . . . .
Typical values of the parison problem parameters. . . . . . . . . . . .
Parameters for the laser drilling process applied to aluminium using an
Nd:YAG-laser. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Component data in the simulation.
The mobilities are in
10−9 m2 /(V · s) and the diffusion coefﬁcients are in 10−9 m2 /s. . . . . .
Physical data for a typical single-inlet pulse tube (values at 300 K). . .

D.1
D.2
D.3

Coefﬁcients for the BDF. . . . . . . . . . . . . . . . . . . . . . . . . . . 633
Coefﬁcients for the Adams–Bashforth formula. . . . . . . . . . . . . . . 634
Coefﬁcients for the Adams–Moulton formula. . . . . . . . . . . . . . . . 634

16.6

. . . . 358

.
.
.
.

503
518
548
584

. 591
. 602
. 611

xxv

i

i
i

i


Notation

Variables and Operators

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>time</td>
</tr>
<tr>
<td>$x$</td>
<td>space coordinate in $\mathbb{R}$</td>
</tr>
<tr>
<td>$x = (x, y)^T, x = (x, y, z)^T$</td>
<td>space coordinate in $\mathbb{R}^d$ ($d = 2, 3$)</td>
</tr>
<tr>
<td>$v(x, t), v(x, t)$</td>
<td>scalar function</td>
</tr>
<tr>
<td>$v(x, t), v(x, t)$</td>
<td>vector function</td>
</tr>
<tr>
<td>$(u, v)$</td>
<td>inner product of the scalar functions $u$ and $v$</td>
</tr>
<tr>
<td>$u \cdot v$</td>
<td>inner product of the vectors $u$ and $v$</td>
</tr>
<tr>
<td>$u \times v$</td>
<td>vector product of the vectors $u$ and $v$</td>
</tr>
<tr>
<td>$L(x, t)[v]$</td>
<td>differential operator $L(x, t)$ applied to $v$</td>
</tr>
<tr>
<td>$L^*(x, t)$</td>
<td>adjoint operator of $L(x, t)$</td>
</tr>
<tr>
<td>$\frac{dv}{dt}, v'$</td>
<td>derivative of $v(x)$</td>
</tr>
<tr>
<td>$\frac{dv}{dx}, v_x$</td>
<td>partial derivative of $v(x, t)$</td>
</tr>
<tr>
<td>$\frac{dv}{dn}, n \cdot \nabla v$</td>
<td>directional derivative of $v$ in the direction of $n$</td>
</tr>
<tr>
<td>$\nabla v$</td>
<td>gradient of $v$</td>
</tr>
<tr>
<td>$\nabla \cdot v$</td>
<td>divergence of $v$</td>
</tr>
<tr>
<td>$\nabla \times v$</td>
<td>curl of $v$</td>
</tr>
<tr>
<td>$\nabla^2 v$</td>
<td>Laplace operator applied to $v$</td>
</tr>
<tr>
<td>$\int_{\Omega} v , dV$</td>
<td>generic integral over a domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$)</td>
</tr>
<tr>
<td>$\int_{\Omega} v(x, t) , dx , dy$</td>
<td>integral over a domain $\Omega \subset \mathbb{R}^2$</td>
</tr>
<tr>
<td>$\int_{\partial \Omega} v \cdot n , dS$</td>
<td>integral over a closed surface $\partial \Omega \subset \mathbb{R}^3$</td>
</tr>
<tr>
<td>$\oint_C v \cdot d\ell$</td>
<td>integral over a closed contour $C \subset \mathbb{R}^d$ ($d = 2, 3$)</td>
</tr>
<tr>
<td>${v}^+$</td>
<td>jump of $v$ across a discontinuity</td>
</tr>
<tr>
<td>$e_x, e_y, e_z$</td>
<td>unit vectors in the Cartesian coordinate system $(x, y, z)$</td>
</tr>
<tr>
<td>$e_r, e_\phi, e_z$</td>
<td>unit vectors in the cylindrical coordinate system $(r, \phi, z)$</td>
</tr>
<tr>
<td>$e_r, e_\theta, e_\phi$</td>
<td>unit vectors in the spherical coordinate system $(r, \theta, \phi)$</td>
</tr>
</tbody>
</table>
### Notation

#### Numerical Parameters and Variables

- $\Delta t$: step size
- $\Delta x$: grid size in $x$ direction
- $h$: generic grid size
- $x_j$: $j$th grid point
- $x_{j+1/2}$: location of a control volume boundary
- $t^n$: time level $n\Delta t$
- $v^n_j$: numerical approximation of $v(x, t)$ at grid point $x_j$ and time level $t^n$
- $F_{j+1/2}^n$: numerical approximation of flux $f(x, t)$ at control volume boundary $x_{j+1/2}$ and time level $t^n$
- $C, E, \text{etc}.$: location names of grid points
- $x_C$: generic grid point
- $v_C$: numerical approximation of $v(x_C)$
- $v^\Delta$: grid function
- $L^\Delta[v^n_j]$: difference operator $L^\Delta$ applied to $v^n_j$
- $L^{-1}_\Delta$: inverse of $L^\Delta$
- $d^n_j = O(\Delta x^2)$: $d^n_j$ is of the order $\Delta x^2$ for $\Delta x \to 0$

#### Vectors and Matrices

- $v = (v_1, v_2, \ldots, v_n)^T$: column vector in $\mathbb{R}^n$
- $v^T = (v_1, v_2, \ldots, v_n)$: row vector in $\mathbb{R}^n$
- $v^k$: $k$th vector in a sequence
- $\|v\|_p$: $p$-norm of $v$
- $A = (a_{ij})$: matrix with $a_{ij}$ in $i$th row and $j$th column
- $A = (a_1, a_2, \ldots, a_n)$: matrix with $a_i$ as the $i$th column
- $A^T$: the transpose of $A$
- $A^{-1}$: the inverse of $A$
- $I$: identity matrix
- $\text{diag}(a_1, a_2, \ldots, a_n)$: diagonal matrix with $a_i$ in $i$th row and column
- $\det A$: determinant of $A$
- $\|A\|_p$: $p$-norm of $A$
- $\rho(A)$: spectral radius of $A$
## Miscellaneous

- \( :=, =: \): is defined as, defines
- \( \doteq \): is equal to when neglecting terms of higher order
- \( \sim \): is asymptotically equal to
- \( \mathcal{O} \): asymptotic order symbol ("big O")
- \( o \): asymptotic order symbol ("small o")
- \( e \): base of the natural logarithm (\( e = 2.71828 \ldots \))
- \( i \): imaginary unit
- \( |z| \): absolute value or modulus of \( z \in \mathbb{C} \)
- \( \bar{z} \): complex conjugate of \( z \in \mathbb{C} \)
- \( [v] \): dimension of variable/constant \( v \) (e.g., in SI units)
- \( \mathcal{C} \): characteristic/curve in \( \mathbb{R}^d \) (\( d = 2, 3 \))
- \( \partial \Omega \): boundary of a domain \( \Omega \subset \mathbb{R}^d \) (\( d = 2, 3 \))
Preface

There exist a number of good textbooks on both the analytical and the numerical aspects of PDEs. So why another book on PDEs, then? Well, most existing texts deal with either analytical or numerical aspects, but not both. There are understandable reasons for this. For one thing, it is the traditional approach. The impressive achievements in understanding a large variety of physical phenomena, long before computers came into use, have made the study of PDEs, often called applied mathematics, a well-established area in its own right, which will no doubt remain so for many years to come. At the same time, although this area has grown impressively, it has become clear that nowadays the power of this branch of mathematics does not so much lie in constructive solutions but rather in giving qualitative answers, thus having a usefulness in its own right. Notwithstanding the achievements it has brought by answering some deep questions such as those regarding existence and uniqueness, and by developing a host of instruments to grasp the solution, its machinery for concretely solving problems is still limited to simplified geometries and the use of expansions in terms of special functions and the like. Yet often the actual solution is the ultimate goal of the scientist or engineer who needs concrete answers. The advent of fast computers and the equally fast development of numerical methods filled this gap to a large extent, so that modern engineers can use a large variety of packages to find numerical approximations to the solutions of their problems. This is reflected in the literature on numerical books. In fact, the latter subject has grown to such an extent that different approaches, like finite elements and finite differences, often appear to be as far apart from each other as they are from a purely analytical text. But despite the importance of standard software, problems are more often than not standard and a thorough knowledge of at least a well-chosen subset of analytical and numerical tools and methodologies is necessary when dealing with real-life problems.

Only when dealing with PDEs in practice does it become clear that numerical treatment and analytical treatment of the subject are both needed. A numerical analyst devoted to computing a Dirichlet problem on a square by yet another method is as esoteric as an applied mathematician who is trying to see the world as built up by spheres and cylinders. This gives the main motivation for this text. We are deeply convinced that we need to treat PDEs by combining analytical and numerical aspects. The two supplement each other but have, conceptually, much in common, too. Nevertheless, concepts and theoretical problems, however sophisticated, are not enough to teach us how to deal with PDEs in practice. In a practical situation one is supposed to make a mathematical model first, obtain insights into the behaviour of the solution, and finally compute the solution (or sometimes be satisfied with a qualitative understanding of the problem). Hence, as a third component, insight into
modeling real-life problems needs to complement and deepen theoretical and numerical knowledge.

This book, therefore, intends to address three aspects: the analysis of PDEs (including some basic knowledge of tools), numerical solution methods, and, last but not least, modeling. In a way, the analytical part follows some of the classical lines. As for numerics, we had to make a choice. Most numerical books focus on either finite difference methods or finite element methods (FEMs). This has a simple explanation: the setting is so large that some choices have to be made. The FEM is quite versatile with respect to the domains on which a problem can be defined. Therefore it is most appropriate for boundary value problems. It requires, however, quite a bit of preparatory work, so that a textbook with these ideals would become too big to handle. We wanted to provide a general introduction to solving PDEs numerically, and for this purpose finite differences (and finite volumes, where they come in handy as well) are sufficiently general. Their introduction is straightforward, as is their use. Moreover, given the insights gained from this text, it should not be hard to use FEMs at proper places instead. Another choice we made deals with the number of nonlinear problems one can treat. Due to lack of space (and also because of didactical constraints), we deal with nonlinearities mostly in a generic way, but certainly do not avoid them. As for the modeling part, we believe that dimensional analysis is an obligatory ingredient. Moreover, modeling can only be learned by doing, so we need a larger number of case examples. We have therefore included a separate chapter, with problems from practice, all from the personal experience of the authors.

We have not attempted to mix analytical and numerical aspects to the extreme to avoid blurring the various approaches. Instead, we have sandwiched chapters on analysis and numerics, thereby frequently cross-referencing. Apart from the aforementioned threefold setup, the book also has a second kind of partitioning. Indeed, the first five chapters are preparatory, both introducing and summarising concepts and properties from a variety of areas, in particular Fourier analysis, distributions, and difference methods. Chapters 8 to 14 deal with analytical and numerical aspects of the major types of PDEs. We have followed the traditional distinction between elliptic, parabolic, and hyperbolic equations, but we do realise that there are, of course, many mixed forms in practice. Although parabolic and hyperbolic equations often have a truly evolutionary character, hyperbolic equations may also arise in a nonevolutionary context. Elliptic equations, which give rise to boundary value problems, on the other hand, may be seen as steady states of parabolic problems. To have fewer academic problems as well, we analyse, e.g., the shallow-water equations and the Stokes equations to some extent. The modeling chapters in this book are Chapters 6, 7, 15, and 16. Chapter 6 introduces the basic equations from continuum mechanics, which are the major reasons why people study a subject like PDEs. Chapter 7 deals with the art of mathematical modeling and expands on the idea that a model should be lean and based on a systematic neglect of what is unimportant. The latter two chapters together provide some essential knowledge for dealing with practical problems. The mainly methodological analytical and numerical Chapters 8 to 14 can be followed without explicit knowledge of Chapters 6 and 7, however. Chapter 15 is a self-contained introduction to perturbation methods and complements the methodologies and tools treated in the preceding PDE chapters. It extends the modeling philosophy of “neglecting the very small” by “including the small” but utilising its smallness. By using the language of asymptotic analysis, we introduce four of the most important methods, two of which are of regular perturbation type (method of slow
variation and method of strained coordinates) and two of which are of singular perturbation type (method of matched asymptotic expansions and method of multiple scales, including rays). It is shown how often remarkably sharp analytic approximations to solutions can be obtained while revealing a lot of the problem’s structure. This plays an important role in actual modeling, as explicit parametric dependence clearly gives more insight than numbers often do. The last chapter, Chapter 16, contains a large number of case studies. Here a true amalgamation of all previous theories and techniques takes place. It can be used alongside other relevant chapters, where references to this chapter enable the user to work out a related practical problem. The various sections in Chapter 16 also give many suggestions for further modeling, which could typically be used in a course with room for larger projects.

The computations for this book were done using MATLAB®. There exists a host of packages and toolboxes to solve PDEs numerically. We have tried to avoid exercises for which elaborate computer usage would be required. A course taught from this book will benefit, however, from assignments dealing with some larger problems, including (less trivial) numerical computations. The exercises at the end of each chapter and in particular the case studies in Chapter 16 should give ideas for this.

The book is intended for use in courses at an advanced undergraduate or a graduate level. It has been designed so that it should be useful both for engineers, who may be more interested in the methods as such, and for more mathematically interested readers. To this end, we have made it self-contained. Teachers may judge for themselves which parts they deem more relevant, given the needs or interests of their audience. By including an extensive index and ample cross-referencing, we hope that this book is also quite suitable for self-study and for reference. The authors have had positive experiences with courses taught from the material of this book. Over the years we have benefited a great deal from the suggestions and remarks of many colleagues, and of course of our students. We owe them a debt, as we will owe to those readers who will share with us their experiences and recommendations. We would like to gratefully acknowledge, nevertheless, the advice of a few people in particular: A.E. Dahoe, J. de Graaf, R. Horvath, I. Lyulina, B. O’Malley, J. Molenaar, V. Nefedov, N.C. Ovenden, M. Patricio, N. Peake, J. Rijenga, W.H.A. Schilders, P.J. Slikkerveer, W.R. Smith, A.A.F. van de Ven, and K. Verhoeven.

We are indebted to the staff of SIAM for their help and patience with this project, which took us much longer to complete than originally anticipated.

R.M.M. Mattheij, S.W. Rienstra, and J.H.M. ten Thije Boonkamp
Eindhoven, December 2004
Chapter 1

Differential and Difference Equations

In this chapter we give a brief introduction to PDEs. In Section 1.1 some simple problems that arise in real-life phenomena are derived. (A more detailed derivation of such problems will follow in later chapters.) We show by a number of examples how they may often be seen as continuous analogues of discrete formulations (i.e., based on difference equations). In Section 1.2 we briefly summarize the terminology used to describe various PDEs. Thus concepts like order and linearity are introduced. In Chapter 2 we shall discuss the classification of the various types of PDEs in more detail. Finally, we introduce difference equations and notions like scheme and stencil, which play a role in numerical approximation, in Section 1.3.

1.1 Introduction

Many phenomena in nature may be described mathematically by functions of a small number of independent variables and parameters. In particular, if such a phenomenon is given by a function of spatial position and time, its description gives rise to a wealth of (mathematical) models, which often result in equations, usually containing a large variety of derivatives with respect to these variables. Apart from the spatial variable(s), which are essential for the problems to be considered, the time variable will play a special role. Indeed, many events exhibit gradual or rapid changes as time proceeds. They are said to have an evolutionary character and an essential part of their modeling is therefore based on causality; i.e., the situation at any time is dependent on the past. As far as (mathematical) modeling leads to PDEs, the latter will be called evolutionary, i.e., involve the time \( t \) as a variable. The other type of problems are often referred to as steady state. We will give some examples to illustrate this background.

A typical PDE arises if one studies the flow of quantities like density, concentration, heat, etc. If there are no restoring forces, they usually have a tendency to spread out. In particular, one may, e.g., think of particles with higher velocities (or rather energy) colliding with particles with lower velocities. The former are initially rather clustered. The energy will gradually spread out, mainly because the high-velocity particles collide with other ones, thereby transferring some of the energy. This is called dissipation. A similar effect can be
observed for a material dissolved in a fluid with concentrations varying in space. Brownian motion will gradually spread out the material over the entire domain. This is called diffusion.

Example 1.1 Consider a long tube of cross section $A$ filled with water and a dye. Initially the dye is concentrated in the middle. Let $u(x, t)$ denote the concentration or density (mass per unit length) of the dye at position $x$ and time $t$; then we see that in a small volume $A \Delta x$, positioned between $x - \frac{1}{2} \Delta x$ and $x + \frac{1}{2} \Delta x$ (Figure 1.1), the total amount of dye equals approximately $u(x, t) \Delta x$. Now consider a similar neighbouring volume $A \Delta x$ between $x + \frac{1}{2} \Delta x$ and $x + \frac{3}{2} \Delta x$, with a corresponding dye concentration $u(x + \Delta x, t)$. The mass that flows per unit time through a cross section is called the mass flux. From the physics of solutions it is known that the dye will move from the volume with higher concentration to one with lower concentration such that the mass flux $f$ between the respective volumes is proportional to the difference in concentration between both volumes and is thus given by

$$f \left( x + \frac{1}{2} \Delta x, t \right) = \alpha \left( u \left( x + \frac{1}{2} \Delta x, t \right) - u(x, t) \right) \frac{u(x + \Delta x, t) - u(x, t)}{\Delta x},$$

where $\alpha$, the diffusion coefficient, usually depends on $u$. This relation is called Fick's law for mass transport by diffusion, which is the analogue of Fourier's law for heat transport by conduction.

As there is a similar flux between the centre volume and its left neighbour, we have a rate of change of total amount of mass in the centre volume equal to the difference between both fluxes given by

$$\frac{\partial}{\partial t} u(x, t) \Delta x = f \left( x + \frac{1}{2} \Delta x, t \right) - f \left( x - \frac{1}{2} \Delta x, t \right).$$

If the diffusion coefficient $\alpha$ is a constant, we have

$$\frac{\partial}{\partial t} u(x, t) = \alpha \frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{\Delta x^2}.\quad (\ast)$$

By taking the limits for small volumes (i.e., $\Delta x \to 0$), we find

$$\frac{\partial}{\partial t} u(x, t) = \alpha \frac{\partial^2}{\partial x^2} u(x, t),$$

which is called the one-dimensional diffusion equation. As heat conduction satisfies the same equation, it is also called the heat equation if $u$ denotes temperature. \hfill \square

\begin{center}
\begin{tikzpicture}
\draw[->] (0,0) -- (6,0);
\draw[->] (2,0) -- (2,2);
\draw[->] (4,0) -- (4,2);
\draw[->] (6,0) -- (6,2);
\node at (0,-0.5) {$x - \frac{1}{2} \Delta x$};
\node at (2,-0.5) {$x - \Delta x$};
\node at (4,-0.5) {$x$};
\node at (6,-0.5) {$x + \Delta x$};
\node at (1,1) {$\Delta x$};
\node at (3,1) {$\Delta x$};
\node at (5,1) {$\Delta x$};
\end{tikzpicture}
\end{center}

\textbf{Figure 1.1.} Sketch of dye diffusion.
Another kind of PDE occurs in the transport of particles. Here a flow typically has a dominant direction; mutual collision of particles (which is felt globally as a kind of internal friction, or viscosity) is neglected.

**Example 1.2** Consider a road with heavy traffic moving in one direction, say the $x$ direction (Figure 1.2). Let the number of cars at time $t$ on a stretch $[x, x + \Delta x]$ be denoted by $\Delta N(x, t)$. Furthermore, let the number of cars passing a point $x$ per time period $\Delta t$ be given by $f(x, t)\Delta t$.

In that period the number of cars $\Delta N(x, t + \Delta t)$ can only be changed by a difference between inflow at $x$ and outflow at $x + \Delta x$; i.e.,

$$\Delta N(x, t + \Delta t) = \Delta N(x, t) - \left(f(x + \Delta x, t) - f(x, t)\right)\Delta t.$$

Rather than the number of cars $\Delta N$ per interval of length $\Delta x$, it is convenient to consider a car density $n(x, t)$, which is defined by

$$\frac{\Delta N(x, t)}{\Delta x} = n(x, t)\Delta x.$$  

Hence we obtain the relation

$$\frac{n(x, t + \Delta t) - n(x, t)}{\Delta t} = -\frac{f(x + \Delta x, t) - f(x, t)}{\Delta x}.$$  

Assuming sufficient smoothness (which implies that we have to allow for fractions of cars . . .), this leads in the limit of $\Delta t, \Delta x \to 0$ to

$$\frac{\partial n}{\partial t} + \frac{\partial f}{\partial x} = 0,$$

which takes the form of a conservation law. We may recognize $f$ again as a flux. If this flux only depends on the local car density, i.e., $f = f(n)$, and $f$ is sufficiently smooth, we obtain

$$\frac{\partial n}{\partial t} + f(n)\frac{\partial n}{\partial x} = 0,$$

also known as the transport equation.

**Figure 1.2.** Sketch of traffic flow.

An important class of problems arises from classical mechanics, i.e., Newtonian systems.

**Example 1.3** Consider a chain consisting of elements, each with mass $m$, and springs, with spring constant $\beta > 0$ and length $\Delta x$; see Figure 1.3. Denote the elements by $V_1, V_2, \ldots$ with position of the masses $x = u_1, u_2, \ldots$. Assuming linear springs, the force necessary to increase the original length $\Delta x$ of the spring of element $V_i$ by an amount $\delta_i = u_i - u_{i-1} - \Delta x$ is equal to $F_i = \beta\delta_i$. Apart from the endpoints, all masses are free to move in the $x$ direction, their inertia being balanced by the reaction forces of the springs. Noting that each element $V_i$ (except for the endpoints) experiences a spring force from the neighbouring $i$th and $(i + 1)$th springs, we have from Newton’s law for the $i$th element that

$$m\frac{d^2 u_i}{dt^2} = F_{i+1} - F_i = \beta(u_{i+1} - u_i - u_{i-1}) = f_i, \quad i = 1, 2, \ldots$$

(*)
Chapter 1. Differential and Difference Equations

If the chain elements increase in number, while the springs and masses decrease in size, it is natural and indeed more convenient not to distinguish the individual elements, but to blend the discrete description of (∗) into a continuous analogue. The small masses are conveniently described by a density \( \rho \) such that \( m = \rho \Delta x \), while the large spring constants are best described by a stiffness \( \sigma = \beta \Delta x \). Then we obtain from (∗) for the position function \( u(x,t) \) the PDE

\[
\frac{\partial^2 u}{\partial t^2} = \frac{\sigma}{\rho} \frac{\partial^2 u}{\partial x^2}.
\]  

As solutions of this equation are typically wave like, it is known as the wave equation, with a wave velocity equal to \( \sqrt{\sigma/\rho} \). In our example it describes longitudinal waves along the suspended chain of masses. In the context of pressure-density perturbations of a compressible fluid like air, the equation describes one-dimensional sound waves, e.g., as they occur in organ pipes. In that case the air stiffness is equal to \( \sigma = \gamma p \), where \( \gamma = 1.4 \) is a gas constant and \( p \) is the atmospheric pressure (see Section 6.8.2).

In the following example we mention the analogue in electrical circuits of the motion of coupled spring-dashpot elements.

**Example 1.4** The time-behaviour of electric currents in a network may be described by the variables potential \( V \), current \( I \), and charge \( Q \). If the network is made of simple wires connecting isolated nodes, resistances, capacities, and coils, and the frequencies are low, it may be modeled (a posteriori confirmed by analysis of the Maxwell equations) one dimensionally by a series of elements with the material properties resistance \( R \), capacitance \( C \), and inductance \( L \). Such a model is called an electrical circuit. If the frequencies are high, such that the wavelength is comparable with the length of the conductors, we have to be more precise. As the signal cannot change instantaneously at all locations, it propagates as a wave of voltage and current along the line. In such a case we cannot neglect the resistance and inductance properties of the wires. By considering the wires as being built up from a series of (infinitesimally) small elements, we can model the system by what is called a transmission line, leading to PDEs in time and space.

In or across each element we have the following relations. The current is defined as the change of charge in time, \( I = \frac{d}{dt} Q \). The capacitance of a pair of conductors is given by \( C = Q/V \), where \( V \) is the potential difference and \( Q \) is the charge difference between the conductors (Coulomb’s law). The resistance between two points is given by \( R = V/I \), where \( V \) is the potential difference between these points and \( I \) is the corresponding current (Ohm’s law). A changing electromagnetic current in a coil with inductance \( L \) induces a counteracting potential, given by \( V = -L \frac{d}{dt} I \) (Faraday’s law). At a junction no charge can accumulate, and we have the condition \( \sum I = 0 \), while around a loop the summed potential vanishes, \( \sum V = 0 \) (Kirchhoff’s laws). With these building blocks we can construct transmission line models.
1.1. Introduction

A famous example is the telegraph equation, where an infinitesimal piece of telegraph wire is modeled (Figure 1.4) as an electrical circuit consisting of a resistance \( R \Delta x \) and an inductance \( L \Delta x \), while it is connected to the ground via a resistance \((G \Delta x)^{-1}\) and a capacitance \( C \Delta x \).

Let \( i(x, t) \) and \( u(x, t) \) denote the current and voltage through the wire at position \( x \) and time \( t \). The change of voltage across the piece of wire is now given by

\[
\begin{align*}
\Delta u(x, t) & = -i R \Delta x - \frac{\partial i}{\partial t} L \Delta x,
\end{align*}
\]

The amount of current that disappears via the ground is

\[
\begin{align*}
\Delta i(x, t) & = -u G \Delta x - \frac{\partial u}{\partial t} C \Delta x,
\end{align*}
\]

By taking the limit \( \Delta x \to 0 \), we get

\[
\begin{align*}
\frac{\partial u}{\partial x} & = -R i - L \frac{\partial i}{\partial t}, \quad \frac{\partial i}{\partial x} = -G u - C \frac{\partial u}{\partial t}.
\end{align*}
\]

By eliminating \( i \), we may combine these equations into the telegraph equation for \( u \), i.e.,

\[
\begin{align*}
\frac{\partial^2 u}{\partial x^2} & = LC \frac{\partial^2 u}{\partial t^2} + (LG + RC) \frac{\partial u}{\partial t} + RG u.
\end{align*}
\]

\( \ast \)

![Figure 1.4. A transmission line model of a telegraph wire.](image)

**Example 1.5** Consider the following crowd of \( N^2 \) very accommodating people (Figure 1.5), for convenience ordered in a square of size \( L \times L \), while each person, labelled by \((i, j)\), is positioned at \( x_i = ih, y_j = jh \), with \( h = L/N \). Each person has an opinion given by the (scalar) number \( p_{ij} \) and can only communicate with his or her immediate neighbours. Assume that each person tries to minimize any conflict with his or her neighbours and is willing to take an opinion that is the average of their opinions. So we have

\[
p_{ij} = \frac{1}{4}(p_{i+1,j} + p_{i-1,j} + p_{i,j+1} + p_{i,j-1}).
\]

\( \ast \)

Only at the borders of the square are the individuals provided with information such that \( p \) is fixed.
If the number of people becomes so large that we may take the limit $N \to \infty$ (i.e., $h \to 0$) and $p$ becomes a continuous function of $(x, y)$, (*) becomes

$$p(x, y) = \frac{1}{4} (p(x + h, y) + p(x - h, y) + p(x, y + h) + p(x, y - h)).$$

This may be recast into

$$\left[ p(x + h, y) - 2p(x, y) + p(x - h, y) \right] + \left[ p(x, y + h) - 2p(x, y) + p(x, y - h) \right] = 0.$$

If this is true for any $h$, we may divide by $h^2$, and the equation becomes in the limit

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} = 0.$$

This equation is called the Laplace equation and describes phenomena where, in some sense, information is exchanged in all directions until equilibrium is achieved. From the above sociological example it is not difficult to appreciate that discontinuities and sharp gradients are smoothed out, while extremes only occur at the boundary. The best-known problem described by this equation is the stationary distribution of the temperature in a heat-conducting medium.

\[ \square \]

1.2 Nomenclature

In the previous section we met a number of equations with derivatives with respect to more than one variable. In general, such equations are called partial differential equations. Let $x$ and $t$ be two independent variables and let $u(x, t)$ denote a quantity depending on $x$ and $t$. Furthermore, let

$$t \in [0, T], \quad 0 \leq T \leq \infty, \quad x \in [a, b] \subset \mathbb{R}. \quad (1.1)$$

For an integer $n$ a general form for a scalar PDE (in two independent variables) reads

$$F \left( \frac{\partial^n u}{\partial t^n}, \frac{\partial^n u}{\partial t^{n-1} \partial x}, \ldots, \frac{\partial^n u}{\partial x^n}, \frac{\partial^{n-1} u}{\partial t^{n-1} \partial x}, \ldots, \frac{\partial^{n-1} u}{\partial x^{n-1}}, \frac{\partial u}{\partial t}, \frac{\partial u}{\partial x}, u, x, t \right) = 0. \quad (1.2)$$

The highest-order derivative is called the order of the PDE; not all partial derivatives (except the highest of at least one variable) need to be present. The form (1.2) is an
1.2. Nomenclature

*implicit* formulation, i.e., the highest-order derivative(s), the *principal part*, do(es) not appear explicitly. If the latter is the case, we call it an *explicit* PDE. The generalization to more than two independent variables is obvious.

**Example 1.6** Some important examples of PDEs are as follows:

\[(i) \quad \frac{\partial u}{\partial t} + c \left(1 + \frac{3}{2}u\right) \frac{\partial u}{\partial x} + \frac{1}{2} c h^2 \frac{\partial^3 u}{\partial x^3} = 0 \quad \text{(Korteweg–de Vries equation).} \]

This is a third order PDE.

\[(ii) \quad \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \quad \text{(nonlinear transport equation).} \]

If \(f\) is differentiable, we see that this is a first order PDE in \(u\).

\[(iii) \quad \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \varepsilon \frac{\partial^2 u}{\partial x^2} \quad \text{(the Burgers’ equation).} \]

If \(\varepsilon = 0\), this may be referred to as the inviscid Burgers’ equation, which is a special case of the transport equation.

\[(iv) \quad \frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} - \frac{1}{3} h^2 \frac{\partial^3 u}{\partial x^3 \partial t} = 0 \quad \text{(linearized Boussinesq equation).} \]

\[(v) \quad EI \frac{\partial^4 u}{\partial x^4} - T \frac{\partial^2 u}{\partial x^2} + m \frac{\partial^2 u}{\partial t^2} = 0 \quad \text{(vibrating beam equation).} \]

\[(vi) \quad \frac{\partial u}{\partial y} \frac{\partial^2 u}{\partial y \partial x} - \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial y^2} + \nu \frac{\partial^3 u}{\partial y^3} = 0 \quad \text{(Prandtl’s boundary layer equation).} \]

In quite a few cases the order can only be deduced after some (trivial) manipulation.

**Example 1.7**

\[\frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left(D(u) \frac{\partial u}{\partial x}\right) = f(x) \quad \text{(nonlinear diffusion equation).} \]

It is clear that this PDE is second order. There is no analytical, numerical, or practical need to rework this and have \(\frac{\partial^2 u}{\partial x^2}\) appear explicitly.

**Example 1.8** A few other examples are as follows:

\[(i) \quad \frac{\partial u}{\partial t} - \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) = 0 \quad \text{(heat equation in three dimensions).} \]
We prefer to write this as \( \frac{\partial}{\partial t} u - \alpha \nabla^2 u = 0 \). \( \nabla^2 \) is referred to as the Laplace operator.

(ii) \( \frac{\partial^2 u}{\partial t^2} - c^2 \nabla^2 u = 0 \) \( \) (wave equation in three dimensions).

(iii) \( \nabla^2 u + k^2 u = 0 \) \( \) (Helmholtz or reduced wave equation).

(iv) \( (1 - M^2) \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0 \) \( \) (equation for small perturbations in steady subsonic \( (M^2 < 1) \) or supersonic \( (M^2 > 1) \) flow).

Sometimes one also denotes a partial derivative of a certain variable by an index:

\[
\begin{align*}
  u_t &:= \frac{\partial u}{\partial t}, & u_{tx} &:= \frac{\partial^2 u}{\partial t \partial x}.
\end{align*}
\]

If we can write (1.2) as a linear combination of \( u \) and its derivatives with respect to \( x \) and \( t \), and with coefficients only depending on \( x \) and \( t \), the PDE is called linear. Moreover, it is called homogeneous if it does not depend explicitly on \( x \) and/or \( t \). If the PDE is a linear combination of derivatives but the coefficients of the highest derivative, say \( n \), depend on \( (n - 1) \)th order derivatives at most, then we call it quasi-linear [29].

For any differential equation we have to prescribe certain initial conditions and boundary conditions for the time and space variable(s), respectively. In evolutionary problems they often both appear as initial boundary conditions. We shall encounter various types and combinations in later chapters.

We finally remark that we may look for solutions that satisfy the PDE in a weak sense. In particular, the derivatives may not exist everywhere on the domain of interest. Again we refer to later chapters for further details.

### 1.3 Difference Equations

Initially, the actual form of the equations we derived in the examples in Section 1.1 was of a difference equation. Like a PDE, we may define a partial difference equation as any relation between values of \( u(x, t) \) where \( (x, t) \in \mathcal{F} \subset [a, b] \times [0, T) \), \( \mathcal{F} \) being a finite set of points of the domain \( [a, b] \times [0, T) \). We shall encounter difference equations when solving a PDE numerically, so they should approximate the PDE in some well-defined way. The simplest way to describe the latter is by defining a scheme, i.e., a discrete analogue of the (continuous) PDE. Since we shall mainly deal with finite difference approximations in this book, we perceive a scheme as the result of replacing the differentials by finite differences. To this end we have to indicate some (generic) points in the domain \( [a, b] \times [0, T) \) at which the function values \( u(x, t) \) are taken. The latter set of points is called a stencil. We shall clarify this with some examples.

**Example 1.9**

(i) Consider Example 1.1 again. If we replace \( \frac{\partial}{\partial t} u(x, t) \) in equation (✱) by a straightforward discretisation, then we obtain the scheme

\[
\begin{align*}
  \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} = \alpha \frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{\Delta x^2},
\end{align*}
\]

and the stencil is the set of bullets (✱) in Figure 1.6.
1.3. Difference Equations

\[ t + \Delta t \]
\[ t \]
\[ x - \Delta x \quad x \quad x + \Delta x \]

**Figure 1.6. Stencil of Example 1.9(i).**

(ii) Consider the wave equation (†) of Example 1.3. A discrete version may be found to be

\[
\frac{u(x, t + \Delta t) - 2u(x, t) + u(x, t - \Delta t)}{\Delta t^2} = \frac{\sigma}{\rho} \frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{\Delta x^2}.
\]

The stencil is given in Figure 1.7.

**Figure 1.7. Stencil of Example 1.9(ii).**

Given the special role of time and the implication it has for the actual computation, which should be based on the causality of the problem, we may distinguish schemes according to the number of time levels involved. If \((k + 1)\) such time levels are involved, we call the scheme a \textit{k-step scheme}. If the scheme involves only spatial differences at earlier time levels, it is called \textit{explicit}; otherwise it is called \textit{implicit}.

**Example 1.10**

(i) The schemes in Example 1.9 are both explicit, the first being a one-step and the second a two-step scheme.

(ii) We could also approximate the \(u_{xx}\) term in the heat equation at time level \(t + \Delta t\) and obtain the scheme

\[
\frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} = \alpha \frac{u(x + \Delta x, t + \Delta t) - 2u(x, t + \Delta t) + u(x - \Delta x, t + \Delta t)}{\Delta x^2}.
\]
This scheme has the stencil given in Figure 1.8. Clearly, it is an implicit one-step scheme.

\[
\begin{array}{c}
t + \Delta t \\
t \\
x - \Delta x & x & x + \Delta x
\end{array}
\]

Figure 1.8. Stencil of Example 1.10(ii).

1.4 Discussion

- The use of the variables \(x\) and \(y\) in an equation does not mean that the PDE cannot have an evolutionary character. There are some cases where they refer to spatial coordinates, yet the corresponding equation may be hyperbolic, a type of equation we will encounter in the next chapter as an instance of evolutionary type.

- If in a system of time-dependent PDEs all spatial derivatives are replaced by suitable difference approximations, we obtain a system of ODEs in time. If one of the PDEs is independent of time, we obtain a differential-algebraic system. A typical example is the condition that an incompressible flow is divergence free (equivalent to conservation of mass), as in the Stokes equations. This problem will be discussed in Section 8.7.

Exercises

1.1. Show that a nonconstant diffusivity \(\alpha(u)\) leads to the equation

\[
\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( \alpha(u) \frac{\partial u}{\partial x} \right).
\]

1.2. Determine the order of the eikonal equation

\[
\left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 + \left( \frac{\partial u}{\partial z} \right)^2 = c^2.
\]

1.3. Determine the order of the PDE

\[
\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2}.
\]

Derive a first order system by writing \(p := \frac{\partial u}{\partial x}, q := \frac{\partial u}{\partial y}\).
1.4. Determine the order of the PDE (where $a$ and $b$ are parameters)

$$\frac{\partial u}{\partial t} = a\nabla^2 u + b\frac{\partial u}{\partial x} + c(u).$$

1.5. Verify that the solution $u = u(x, t)$ of the transport equation (cf. Example 1.2 or 1.6(ii))

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \quad u(x, 0) = v(x),$$

for sufficiently smooth $f$ is implicitly given by

$$u = v(x - f'(u)t).$$
Chapter 2

Characterisation and Classification

The study of PDEs is quite versatile. Therefore it makes sense to first characterise them according to certain properties that will provide us with guidelines for investigating them further. It is useful to start with first order equations in two independent variables. Therefore we start in Section 2.1 by describing scalar first order equations, thereby introducing the important notion of characteristics. This is generalised for first order systems in Section 2.2, leading to the definition of hyperbolicity. A well-known class of PDEs consists of second order scalar equations. In Section 2.3 we reformulate them as a first order system of equations and then discuss their classification as hyperbolic, parabolic, and elliptic equations. Quite naturally, this can be generalised to several (space) dimensions, as is shown in Section 2.4. Sometimes the underlying structure of a problem is simpler than suggested, and after a suitable transformation the PDE may be transformed into an ODE. Examples are given in Section 2.5. PDEs need further conditions to make their solutions meaningful and (we hope) unique. In Section 2.6 we briefly discuss, from a more theoretical point of view, how to properly choose the initial and boundary conditions. For this we have the Hadamard condition, which states the conditions necessary for a problem to be well posed.

2.1 First Order Scalar PDEs in Two Independent Variables

Consider the quasi-linear (explicit) first order PDE

\[
a(x, t, u) \frac{\partial u}{\partial t} + b(x, t, u) \frac{\partial u}{\partial x} = c(x, t, u). \tag{2.1}
\]

Usually, the independent variables \(x\) and \(t\) denote a space coordinate and time, respectively, although, strictly speaking, \(t\) might denote a space coordinate as well. Let \(u = \varphi(x, t)\) be a solution of (2.1). A geometrical interpretation of this solution is as follows. The independent variables \(x\) and \(t\) and the dependent variable \(u\) constitute a two-parameter family of vectors \((x, t, u)\) that is lying on a surface \(S \subset \mathbb{R}^3\). This surface \(S\), given by \(F(x, t, u) := \varphi(x, t) - u = 0\), is called the integral surface of (2.1). A normal \(n\) on \(S\) is
given by
\[ n := \nabla F = \left( \frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial t}, -1 \right)^T. \] (2.2)
Hence, for an infinitesimal displacement \( du := (dx, dt, du)^T \) along the surface, we find
\[ n \cdot du = \frac{\partial \phi}{\partial x} \, dx + \frac{\partial \phi}{\partial t} \, dt - du = 0. \] (2.3)
Comparing (2.1) and (2.3), we conclude that for a solution \( u = \varphi(x, t) \) of (2.1) the following should hold on the integral surface \( S \):
\[ \begin{pmatrix} a & b \\ dt & dx \end{pmatrix} \begin{pmatrix} \frac{\partial \phi}{\partial t} \\ \frac{\partial \phi}{\partial x} \end{pmatrix} = \begin{pmatrix} c \\ du \end{pmatrix}. \] (2.4)
The solution of this system is unique if and only if \( a \, dx - b \, dt \neq 0 \). In the following we will simply write \( u = u(x, t) \) instead of \( u = \varphi(x, t) \) to indicate the solution of (2.1).

This result can be interpreted as follows (see Figure 2.1): suppose we have a smooth, one-parameter curve \( \mathcal{J} = \{ (x(\sigma), t(\sigma), u(\sigma)) \mid \sigma \in I \subset \mathbb{R} \} \) on \( S \), where the condition \( a \, dx - b \, dt \neq 0 \) holds. Then the derivatives \( u_x \) and \( u_t \) are uniquely determined on \( \mathcal{J} \) through (2.4). If, moreover, \( u \) is given along \( \mathcal{J} \), then the solution \( u = u(x, t) \) exists and

![Figure 2.1. Initial curve \( \mathcal{J} \) and a characteristic \( C \) on the integral surface \( S \).](image)
is unique, at least in some neighbourhood of $\mathcal{J}$. The curve $\mathcal{J}$ is referred to as a curve of initial values or, briefly, an initial curve.

The actual construction of the solution proceeds as follows. Suppose $u$ is given along an initial curve $\mathcal{J}$. Consider a curve $\mathcal{C}$ on $\mathcal{S}$ for which \(a \, dx - b \, dt = 0\). Then system (2.4) has either no solution or infinitely many. In the latter case the relations

\[
\frac{dt}{a} = \frac{dx}{b} = \frac{du}{c}
\]

should hold along $\mathcal{C}$. Clearly, the vector $(a, b, c)^T$ is everywhere tangent to $\mathcal{C}$. We can now introduce a parametrization $C = \{(x(s), t(s), u(s))^T \mid s \in I \subset \mathbb{R}\}$ such that $ds = dt/a = dx/b = du/c$ and $s = 0$ on the initial curve $\mathcal{J}$. This way, we obtain the set of ODEs

\[
\frac{dt}{ds} = a, \quad \frac{dx}{ds} = b, \quad \frac{du}{ds} = c
\]

(2.6a) coupled with an initial condition of the form

\[
u(0; \sigma) = v(x(\sigma), t(\sigma)) \quad \text{for} \quad (x(\sigma), t(\sigma)) \in \mathcal{J'},
\]

(2.6b) where $\mathcal{J'}$ is the projection of $\mathcal{J}$ on the $(x, t)$ plane. The ODEs (2.6a) are referred to as the characteristic equations. Consequently, the curve $\mathcal{C}$ is a solution of (2.6). $\mathcal{C}$ is called a characteristic and its projection on the $(x, t)$ plane is called a base characteristic. Note, however, that there is no uniformity in the nomenclature in the literature; usually, no distinction is made between characteristics and base characteristics. In order to construct the integral surface, we compute for each point on the initial curve $\mathcal{J}$ the characteristic passing through that point from (2.6). We formally obtain the solution

\[
\begin{align*}
t &= t(s; \sigma), \quad x &= x(s; \sigma), \quad u &= u(s; \sigma).
\end{align*}
\]

Inverting the first two relations, we find $s = s(t, x), \sigma = \sigma(t, x)$, and substituting these in the expression for $u$ gives $u(x, t) := u(s(t, x); \sigma(t, x))$. This inversion is possible only if the Jacobian $t_x x_s - t_s x_x$ is nonzero. Thus the integral surface $\mathcal{S}$ is generated by a one-parameter family of characteristics $\mathcal{C}$ all passing through an initial curve $\mathcal{J}$. We will demonstrate this by an example.

**Example 2.1** Consider the following initial value problem for the inviscid Burgers’ equation:

\[
\begin{align*}
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} &= 0, \quad x \in \mathbb{R}, \quad t > 0, \\
u(x, 0) &= v(x) := \begin{cases} 
1 & \text{if } x \leq 0, \\
1 - x & \text{if } 0 < x \leq 1, \\
0 & \text{if } x \geq 1.
\end{cases}
\end{align*}
\]

The characteristic equations and the corresponding initial conditions read

\[
\begin{align*}
\frac{dt}{dx} &= 1, \quad \frac{dx}{ds} = u, \quad \frac{du}{ds} = 0, \\
t(0; \sigma) &= 0, \quad x(0; \sigma) = \sigma, \quad u(0; \sigma) = v(\sigma).
\end{align*}
\]
Chapter 2. Characterisation and Classification

The solution of this set of ODEs is given by

\[ t(s; \sigma) = s, \quad x(s; \sigma) = \sigma + v(\sigma)s, \quad u(s; \sigma) = v(\sigma). \]

We can easily invert the first two relations provided that the Jacobian \( t_s x_s - t_x s = -1 - v'(\sigma)s \) is nonzero. This way, we obtain \( s(x, t) = t \) and \( \sigma(x, t) = x - t \) for \( x \leq t \); \( \sigma(x, t) = (x - t)/(1 - t) \) for \( t < x \leq 1 \); and \( \sigma(x, t) = x \) for \( x \geq 1 \). Consequently, the solution is defined for \( 0 < t < 1 \) and is given by

\[
\begin{align*}
    u(x, t) &= \begin{cases} 
        1 & \text{if } x \leq t, \\
        1 - x & \text{if } t < x \leq 1, \\
        0 & \text{if } x \geq 1.
    \end{cases}
\end{align*}
\]

Instead of using \( s \), we can parametrize the characteristics by the variable \( t \) provided that \( a \neq 0 \). We thus obtain the ODEs

\[
\begin{align*}
    \frac{dx}{dt} &= \frac{b}{a}, \\
    \frac{du}{dt} &= \frac{c}{a}.
\end{align*}
\]

(2.7)

We also refer to (2.7) as the characteristic equations. The first equation gives the location of the base characteristics, possibly depending on the solution \( u \), and the second gives \( u \) along the base characteristics. We see that the existence of solutions of (2.1) can be established from studying the ODEs (2.7). It is easy to see that such a solution is composed of solutions of initial value problems defined along base characteristics. If the right-hand side of (2.7) satisfies a Lipschitz condition, then (2.7), together with an initial value for \( u \), determines the solution \( u \) on \( C \).

Example 2.2 Consider the transport equation

\[
\frac{\partial u}{\partial t} + b(u) \frac{\partial u}{\partial x} = 0 \quad (*)
\]

subject to an initial condition of the form \( u(x, 0) = v(x) \). For this equation system (2.7) reduces to

\[
\frac{dx}{dt} = b(u), \quad \frac{du}{dt} = 0,
\]

implying that \( u(x, t) = \text{constant} \) along the base characteristics, which, however, depend on the solution we seek. Since \( u(x, t) = \text{constant} \), \( x - b(u)t = \text{constant} \), and we obtain the following (implicit) representation of the solution:

\[
u(x, t) = v(x - b(u(x, t))t).
\]

In the special case of the linear advection equation, for which \( b(u) = b = \text{constant} \), this representation reduces to \( u(x, t) = v(x - bt) \), i.e., the initial profile is propagated undisturbed with speed \( b \) along the base characteristics; see Figure 2.2.
2.2 First Order Linear Systems in Two Independent Variables

The previous theory for scalar equations can be extended to systems. In the rest of this text we will often encounter systems of the form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = c,$$

where $u : I_1 \times I_2 \to \mathbb{R}^m$, $I_1 \subset \mathbb{R}$, and $I_2 \subset [0, \infty)$. Introducing the Jacobi matrix

$$B := \frac{\partial f}{\partial u} = \left( \frac{\partial f_i(u)}{\partial u_j} \right),$$

we find the quasi-linear first order system of PDEs

$$\frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = c. \quad (2.9)$$

More generally, a quasi-linear first order system of PDEs may read

$$A \frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = c. \quad (2.10)$$

In this section we assume that the coefficient matrices $A$ and $B$ are constant and that $c = c(x, t, u)$. The quasi-linear case where $A = A(x, t, u)$ and $B = B(x, t, u)$ is discussed in Chapter 12.

In order to employ the theory of the previous section, we try to decouple system (2.10) into $n$ scalar equations. For the sake of simplicity we further assume that the matrix $A$ is nonsingular. Then we look for a nonsingular transformation matrix $S$ such that

$$S^{-1}(BA^{-1})S = A, \quad (2.11)$$
where $\mathbf{A}$ is a diagonal matrix. If such an $\mathbf{S}$ exists, we can introduce the characteristic variable $\tilde{\mathbf{u}}$ defined by

$$\tilde{\mathbf{u}}(x, t) := \mathbf{S}^{-1} \mathbf{A} \mathbf{u}(x, t),$$

satisfying the decoupled system

$$\frac{\partial \tilde{\mathbf{u}}}{\partial t} + \mathbf{\Lambda} \frac{\partial \tilde{\mathbf{u}}}{\partial x} = \tilde{\mathbf{c}} := \mathbf{S}^{-1} \mathbf{c}.$$  \hspace{1cm} (2.13a)

If we write this system componentwise, we have the scalar equations

$$\frac{\partial \tilde{u}_k}{\partial t} + \lambda_k \frac{\partial \tilde{u}_k}{\partial x} = \tilde{c}_k, \quad k = 1, 2, \ldots, m. \hspace{1cm} (2.13b)$$

The previous derivation gives rise to the following definition.

**Definition 2.3.** The linear system (2.10) with nonsingular matrix $\mathbf{A}$ is called hyperbolic if $\mathbf{B} \mathbf{A}^{-1}$ has $m$ real eigenvalues and $m$ linearly independent eigenvectors.

Consequently, when system (2.10) is hyperbolic, the matrices $\mathbf{A}$ and $\mathbf{S}$ do exist and are given by

$$\mathbf{A} := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m), \quad \mathbf{S} := (s_1, s_2, \ldots, s_m).$$

(2.14)

where $\lambda_k$ and $s_k$ are the eigenvalues and corresponding eigenvectors of $\mathbf{B} \mathbf{A}^{-1}$, respectively. Thus the $k$th column of $\mathbf{S}$ is the eigenvector $s_k$. Note that (2.10) is always hyperbolic if $\mathbf{B} \mathbf{A}^{-1}$ is symmetric; for a general matrix hyperbolicity is assured if all eigenvalues are real and distinct.

Each equation in (2.13b) induces a characteristic $C_k$ corresponding to the eigenvalue $\lambda_k$ and eigenvector $s_k$. The characteristic equations (2.7) in this case read

$$\frac{dx}{dt} = \lambda_k, \quad \frac{d\tilde{u}_k}{dt} = \tilde{c}_k.$$  \hspace{1cm} (2.15)

The solution of (2.10) can be recovered from solutions of (2.15). This is demonstrated in the following example.

**Example 2.4** Consider the initial value problem

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \frac{\partial \mathbf{u}}{\partial t} + \begin{pmatrix} 1 & 4 \\ 1 & 0 \end{pmatrix} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, \quad x \in \mathbb{R}, \quad t > 0,$$

$$\mathbf{u}(x, 0) = \mathbf{v}(x), \quad x \in \mathbb{R}.$$

We can easily verify that the eigenvalues and corresponding eigenvectors of $\mathbf{B} \mathbf{A}^{-1}$ are given by

$$\lambda_1 = -1, \quad \lambda_2 = 2, \quad s_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$
2.2. First Order Linear Systems in Two Independent Variables

The characteristic variable $\tilde{u}$ defined in (2.12) is now given by

$$\tilde{u}_1 = \frac{1}{3}(u_1 - 4u_2), \quad \tilde{u}_2 = \frac{1}{3}(u_1 + 2u_2).$$

These variables can be computed from (2.15) and we find $\tilde{u}_1(x, t) = \tilde{u}_1(x + t, 0)$ and $\tilde{u}_2(x, t) = \tilde{u}_2(x - 2t, 0)$. Combining this result with the above relations for the characteristic variables, we obtain the final solution

$$u_1(x, t) = \frac{1}{3}(v_1(x + t) - 4v_2(x + t) + 2v_1(x - 2t) + 4v_2(x - 2t)),$$

$$u_2(x, t) = \frac{1}{6}(-v_1(x + t) + 4v_2(x + t) + v_1(x - 2t) + 2v_2(x - 2t)).$$

Clearly, the solution contains waves propagating along base characteristics $x + t = \text{constant}$ and $x - 2t = \text{constant}$, respectively.

In general we conclude that (2.10) should not be subject to an initial condition prescribed on a characteristic. In fact, one should prescribe $u$ on a curve $\mathcal{J}$ that does not intersect any of these characteristics twice.

Finally, we introduce the following notions; see Figure 2.3. The domain of dependence of a point $(x_0, t_0)$ is the region in the $(x, t)$ plane such that $u(x_0, t_0)$ depends on all values $u(x, t)$ with $(x, t)$ in this domain. It is bounded by the two extreme characteristics through $(x_0, t_0)$ facing back to the initial line $t = 0$. On the other hand, the domain of influence of $(x_0, t_0)$ is the region in the $(x, t)$ space where the solution is influenced by $u(x_0, t_0)$.

In the next section we shall consider the special case of systems arising from scalar second order problems.

![Figure 2.3](image_url)

**Figure 2.3.** Region of influence and region of dependence. In the case of constant coefficients these characteristics are straight lines.
2.3 Second Order Scalar PDEs in Two Independent Variables

For PDEs with higher-order derivatives the classification may be reduced to first order systems met before, at least if they are scalar. In particular, we consider the second order linear equation

\[ a \frac{\partial^2 u}{\partial t^2} + b \frac{\partial^2 u}{\partial t \partial x} + c \frac{\partial^2 u}{\partial x^2} + d \frac{\partial u}{\partial t} + e \frac{\partial u}{\partial x} = f, \]  

(2.16)

where the coefficients \( a, b, \ldots, e \) are assumed constant and where the right-hand side \( f \) possibly depends on \( x, t, \) and \( u \). The independent variable \( x \) is a space coordinate, whereas \( t \) is either time or a space coordinate. Introducing the variables

\[ p := \frac{\partial u}{\partial t}, \quad q := \frac{\partial u}{\partial x}, \]  

(2.17)

we obtain the linear system

\[ \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} \frac{\partial}{\partial t} \begin{pmatrix} p \\ q \end{pmatrix} + \begin{pmatrix} b & c \\ -1 & 0 \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} f - d p - e q \\ 0 \end{pmatrix}. \]  

(2.18)

Note that this form is not unique. Clearly, this system is of the form (2.10), with the coefficient matrices given by

\[ A := \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}, \quad B := \begin{pmatrix} b & c \\ -1 & 0 \end{pmatrix}. \]  

(2.19)

We now investigate the eigenvalues and eigenvectors of this system.

In the following we assume that \( a \neq 0 \), so that \( A \) is nonsingular. Like in the previous section, we look for a transformation matrix \( S \) that can diagonalise the matrix

\[ BA^{-1} = \begin{pmatrix} b & c \\ -1/a & 0 \end{pmatrix}. \]  

(2.20)

This will succeed if the characteristic equation

\[ \det(BA^{-1} - \lambda I) = \lambda^2 - \frac{b}{a} \lambda + \frac{c}{a} = 0 \]  

(2.21)

has two different, real roots, since then the corresponding eigenvectors are linearly independent. Consequently, system (2.18) is hyperbolic if \( b^2 - 4ac > 0 \). If \( b^2 - 4ac = 0 \), we have a degeneracy of the eigensystem and so only one “double” characteristic exists; in fact no \( S \) and \( A \), as required in (2.11), can be found. If, finally, \( b^2 < 4ac \), there are no real characteristic values at all. This leads to the following definition.

**Definition 2.5.** The PDE (2.16) is called

(i) **hyperbolic** if \( b^2 - 4ac > 0 \);

(ii) **parabolic** if \( b^2 = 4ac = 0 \);

(iii) **elliptic** if \( b^2 - 4ac < 0 \).
The nomenclature in this definition is adopted from the theory of quadratic forms. In particular, the corresponding quadratic curve \( at^2 + btx + cx^2 + dt + ex = \text{constant} \) is a hyperbola, parabola, or ellipse, depending on the value of \( b^2 - 4ac \).

Next we will derive the normal, or canonical, form of (2.16) in these three different cases, which only depends on the principal part of the equation, i.e., the first three terms containing the second order derivatives.

In the hyperbolic case we find a transformation matrix

\[
S = \begin{pmatrix}
\lambda_1 & \lambda_2 \\
-1 & -1 \\
a & a
\end{pmatrix},
\tag{2.22}
\]

where \( \lambda_1 < \lambda_2 \) are the eigenvalues of \( BA^{-1} \). The characteristic variables \( \tilde{u} \) defined in (2.12) are now given by

\[
\tilde{u} = \frac{a}{\lambda_2 - \lambda_1} \begin{pmatrix}
-p - \lambda_2 q \\ p + \lambda_1 q
\end{pmatrix}.
\tag{2.23}
\]

Inserting the first component \( \tilde{u}_1 \) into (2.13b), we obtain an equation of the form

\[
\left( \frac{\partial}{\partial t} + \lambda_1 \frac{\partial}{\partial x} \right) \left( \frac{\partial}{\partial t} + \lambda_2 \frac{\partial}{\partial x} \right) u = \tilde{f}.
\tag{2.24}
\]

Taking into account the characteristic equations (2.15), we see that the first and second differential operators in (2.24) are just internal derivatives along characteristics of the \( C_1 \) and \( C_2 \) families, respectively. Introducing coordinates \( \xi \) and \( \eta \) along these characteristics, we obtain the canonical form

\[
\frac{\partial^2 u}{\partial \xi \partial \eta} = \tilde{f}.
\tag{2.25}
\]

Note that we would obtain the same equation starting from the equation for the second characteristic variable \( \tilde{u}_2 \).

In the parabolic case there is no transformation matrix \( S \) possible that can diagonalise \( BA^{-1} \). However, we have the Jordan normal form (see Section H in the appendix)

\[
S^{-1} (BA^{-1}) S = J := \begin{pmatrix}
\lambda & 1 \\
0 & \lambda
\end{pmatrix},
\tag{2.26}
\]

with \( \lambda = b/(2a) \) the eigenvalue of \( BA^{-1} \), having algebraic multiplicity two and geometric multiplicity one. A possible matrix \( S \) is given by

\[
S = \begin{pmatrix}
\lambda & 1 \\
-1 & 0 \\
a & a
\end{pmatrix}.
\tag{2.27}
\]

We can now reduce system (2.18) to

\[
\frac{\partial \tilde{u}}{\partial t} + J \frac{\partial \tilde{u}}{\partial x} = \tilde{f}
\tag{2.28}
\]
for the variable $\tilde{u}$, given by

$$\tilde{u} = a \begin{pmatrix} -q \\ p + \lambda q \end{pmatrix}.$$  

(2.29)

From (2.13b) for $\tilde{u}_2$ we easily deduce

$$\left( \frac{\partial}{\partial t} + \lambda \frac{\partial}{\partial x} \right)^2 u = \tilde{f},$$

(2.30)

in which we recognize an internal differentiation along a characteristic. The canonical form for (2.16) is thus given by

$$\frac{d^2 u}{d\xi^2} = \tilde{f},$$

(2.31)

where $\xi$ is the coordinate along the characteristic.

Finally, in the elliptic case, we have complex characteristics, and the transformation matrix $S$ is also complex. Completely analogously to the hyperbolic case, we obtain the form

$$\frac{\partial^2 u}{\partial\xi\partial\eta} = \tilde{f},$$

(2.32)

with $\xi$ and $\eta$ the (complex) coordinates along characteristics of the $C_1$ and $C_2$ families, respectively. One can prove that $\eta = \bar{\xi}$; see, e.g., [29]. Introducing the new coordinates

$$\mu := \frac{1}{2} (\xi + \eta) = \text{Re}(\xi), \quad \nu := \frac{1}{2i} (\xi - \eta) = \text{Im}(\xi),$$

(2.33)

we obtain the canonical form

$$\frac{\partial^2 u}{\partial\mu^2} + \frac{\partial^2 u}{\partial\nu^2} = 4\tilde{f}.$$  

(2.34)

**Example 2.6** The standard examples of hyperbolic, parabolic, and elliptic equations are, respectively,

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \quad \text{(wave equation)},$$

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad \text{(heat equation)},$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \text{(Laplace equation)}. \quad \square$$

The classification as given in Definition 2.5 also holds for linear equations with coefficients depending on $x$ and $t$ and even for quasi-linear equations. The definition should then be applied pointwise, as is demonstrated in the next example.

**Example 2.7** Two well-studied equations in the theory of transonic flow are the Tricomi equation and the isentropic potential flow equation. The Tricomi equation (see, e.g., [4]) is given by

$$y \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0.$$
2.4. Linear Second Order Equations in Several Space Variables

For \( y > 0 \) we apparently have a hyperbolic equation (related to supersonic flow), whereas for \( y < 0 \) the equation is elliptic (related to subsonic flow)! The isentropic potential flow equation for the velocity potential \( \phi \) reads (see, e.g., [63, 64])

\[
\left[ c^2 - \left( \frac{\partial \phi}{\partial x} \right)^2 \right] \frac{\partial^2 \phi}{\partial x^2} + \left[ c^2 - \left( \frac{\partial \phi}{\partial y} \right)^2 \right] \frac{\partial^2 \phi}{\partial y^2} - 2 \frac{\partial \phi}{\partial x} \frac{\partial \phi}{\partial y} \frac{\partial^2 \phi}{\partial x \partial y} = 0,
\]

with the speed of sound \( c \) related to the velocity via Bernoulli’s equation for compressible flow (see (7.12))

\[
\left( \frac{\partial \phi}{\partial x} \right)^2 + \left( \frac{\partial \phi}{\partial y} \right)^2 + \frac{2c^2}{\gamma - 1} = \frac{2c_0^2}{\gamma - 1},
\]

where \( \gamma \) is the specific heat ratio and \( c_0 \) is the sound speed for stagnant flow. Evidently, the equation is hyperbolic for supersonic flow \((\phi^2_x + \phi^2_y > c^2)\) and elliptic for subsonic flow \((\phi^2_x + \phi^2_y < c^2)\).

Hyperbolic, parabolic, and elliptic equations are distinctly different. Solutions of each type of equation show an entirely different behaviour, which is also reflected in the solution methods, either analytically or numerically. In the rest of this book we will extensively address these three types of equations.

2.4 Linear Second Order Equations in Several Space Variables

The general linear second order PDE of constant coefficients is given by

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu = f. \tag{2.35}
\]

Here the variables \( x_1, x_2, \ldots, x_n \) can be time and/or any number of space variables. More precisely, for time-dependent problems we have \( n = d + 1 \) and \( x_n = t \), whereas for stationary problems \( n = d \) and all variables are space coordinates. In Section 2.3 we used characteristics to define new variables and obtained a normal form. Here we shall simply consider transformations of the variables without such a theory and look for simplified forms of (2.35) from a geometrical point of view.

To start with we may associate with (2.35) a symmetric matrix \( A \), where

\[
A := \left( \frac{1}{2} (a_{ij} + a_{ji}) \right); \tag{2.36}
\]

actually, we have distributed the coefficient \( a_{ij} \) symmetrically between the entries \((i, j)\) and \((j, i)\) of the matrix \( A \). Now also define

\[
x := (x_1, \ldots, x_n)^T, \quad b := (b_1, \ldots, b_n)^T. \tag{2.37}
\]

Then (2.35) may be formulated as

\[
\nabla_x \cdot (A \nabla_x u) + b \cdot \nabla_x u + cu = f, \tag{2.38}
\]
where $\nabla_x$ denotes the gradient with respect to $x$. This can be simplified by diagonalising $A$. Since $A$ is symmetric, this is possible through an orthogonal similarity transformation. So let $Q$ be such that

$$Q^T A Q = \Lambda,$$  \hspace{1cm} (2.39)

where $Q^T = Q^{-1}$ and $A$ is the real diagonal matrix of eigenvalues of $A$. Next we introduce a new set of variables $y := (y_1, \ldots, y_n)^T$ by

$$y := Q^T x.$$  \hspace{1cm} (2.40)

Using the relations $\nabla_x = Q \nabla_y$ and $(Qv) \cdot w = v \cdot (Q^T w)$, we obtain the simplified form

$$\nabla_y \cdot (A \nabla_y u) + (Q^T b) \cdot \nabla_y u + cu = f.$$  \hspace{1cm} (2.41)

Viewing $\nabla_y u$ as a vector in $\mathbb{R}^n$, the quadratic form can now be classified as a (generalised)

(i) **ellipse** if all eigenvalues have the same sign,

(ii) **parabola** if at least one of the eigenvalues is zero, or

(iii) **hyperbola** if all eigenvalues are nonzero and have the same sign, except for one.

These geometric descriptions make sense in $\mathbb{R}^2$ at least. The corresponding PDEs are classified similarly: elliptic, parabolic, and hyperbolic. If there are at least two positive and negative eigenvalues (and the others are all nonzero), one sometimes calls the PDE ultrahyperbolic.

If we scale the variables $y_1, \ldots, y_n$ by multiplying them by $\sqrt{|\lambda_1|}, \ldots, \sqrt{|\lambda_n|}$, respectively (unless $\lambda_i = 0$), we will obtain a quadratic form with $\pm 1, 0$ as eigenvalues. Hence it is not restrictive to assume that this has already been done. It then follows that the multidimensional Laplace operator $\nabla^2$ becomes an important symbol to describe second order PDEs in several dimensions. In particular we have the following:

the **(elliptic) Laplace equation**

$$\nabla^2 u = 0,$$  \hspace{1cm} (2.42a)

the **(parabolic) heat equation**

$$\frac{\partial u}{\partial t} = \nabla^2 u,$$  \hspace{1cm} (2.42b)

and the **(hyperbolic) wave equation**

$$\frac{\partial^2 u}{\partial t^2} = \nabla^2 u.$$  \hspace{1cm} (2.42c)

### 2.5 Reduction to ODEs; Similarity Solutions

In the foregoing we have seen that it is sometimes possible to reformulate a (set of) PDE(s) as a set of ODEs that describe the solution along characteristics. In this section we will reiterate this in a slightly different way. To that purpose, consider the initial value problem

$$\frac{\partial u}{\partial t} + b(u) \frac{\partial u}{\partial x} = c(u), \quad x \in \mathbb{R}, \quad t > 0,$$  \hspace{1cm} (2.43a)

$$u(x, 0) = v(x), \quad x \in \mathbb{R}.$$  \hspace{1cm} (2.43b)
2.5. Reduction to ODEs; Similarity Solutions

For the special case \( c(u) \equiv 0 \) we have derived in Example 2.2 from the corresponding characteristic equations the following (implicit) representation of the solution:

\[
 u(x, t) = v(x - b(u(x, t))).
\]

If, furthermore, \( b(u) = b = \text{constant} \), this solution reduces to \( u(x, t) = v(x - bt) \), which is a wave propagating undisturbed with speed \( b \).

Based on these observations, we try a solution of the form

\[
 u(x, t) = \hat{u}(\xi), \quad \xi := x - st. \quad (2.44)
\]

This solution is called a travelling-wave solution with wave speed \( s \). Substituting (2.44) in (2.43a), we obtain the ODE

\[
 (b(\hat{u}) - s) \frac{d\hat{u}}{d\xi} = c(\hat{u}). \quad (2.45)
\]

The travelling wave \( \hat{u}(\xi) \) and its wave speed \( s \) have to be determined from (2.45) and (2.43b). We will illustrate this in the next example.

**Example 2.8** An example from combustion theory reads [88]

\[
 \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = c(u) := u(1 - u)(u - \beta)/\tau, \quad (*)
\]

with parameters \( \tau > 0 \) and \( 0 < \beta < 1 \). The corresponding ODE \( u'(t) = c(u(t)) \) has the stable equilibrium solutions \( u(x, t) \equiv 0, 1 \) and the unstable one \( u(x, t) \equiv \beta \). This means that any initial value \( v(x) \neq \beta \) approaches either one of the stable solutions for \( t \to \infty \). Substituting (2.44) in \((*)\), we obtain the ODE

\[
 (\hat{u} - s) \frac{d\hat{u}}{d\xi} = \frac{1}{\tau} \hat{u}(1 - \hat{u})(\hat{u} - \beta). \quad (**)\]

Suppose that the initial solution \( v(x) \) increases monotonically from zero to one. Then the appropriate boundary conditions for \((**\)) read

\[
 \lim_{\xi \to -\infty} \hat{u}(\xi) = 0, \quad \lim_{\xi \to \infty} \hat{u}(\xi) = 1.
\]

Equation \((**\)) can only satisfy these boundary conditions if \( s = \beta \); otherwise \( \hat{u}(\xi) \) cannot cross the unstable solution \( u(x, t) \equiv \beta \). The resulting solution then reads

\[
 \hat{u}(\xi) = (1 + e^{-\xi/\tau})^{-1}. \quad \Box
\]

Travelling waves occur naturally in transport equations but also in a number of other equations. We return to this in Chapter 10, where we consider travelling waves for a parabolic equation.

Another class of solutions consists of the similarity solutions, which are functions of a (dimensionless) combination of \( x \) and \( t \). We will introduce these solutions for the homogeneous transport equation (2.43a); i.e., \( c(u) \equiv 0 \). A more systematic discussion of similarity solutions, based on dimensional analysis, is presented in Chapter 7. Ignoring initial and boundary conditions, we see that if \( u(x, t) \) is a solution of (2.43a), then \( u_*(x, t) := u(x, t) \).

For the special case \( c(u) \equiv 0 \) we have derived in Example 2.2 from the corresponding characteristic equations the following (implicit) representation of the solution:
Chapter 2. Characterisation and Classification

\( u(\alpha x, \alpha t) \) is a solution as well for any \( \alpha > 0 \). Therefore we may try a similarity solution of the form

\[
    u(x, t) = \hat{u}(\eta), \quad \eta := \frac{x}{t};
\]

(2.46)

i.e., \( u(x, t) = \text{constant along rays } x/t = \text{constant through the origin of the } (x, t) \text{ plane.} \)

In Chapter 12 we will use this formulation to compute solutions of hyperbolic equations. Substituting (2.46) into (2.43a), we have

\[
    \left( b(\hat{u}) - \eta \right) \frac{d\hat{u}}{d\eta} = 0,
\]

(2.47)

implies that either \( \hat{u}(\eta) = \text{constant}, \) resulting in the trivial solution \( u(x, t) \equiv \text{constant}, \) or \( b(\hat{u}) = \eta. \) In the latter case we obtain the solution

\[
    u(x, t) = b^{-1}(x/t),
\]

(2.48)

with \( b^{-1}(u) \) the inverse of \( b(u), \) assuming it exists.

**Example 2.9** Consider the traffic flow problem of Example 1.2 given by the transport equation

\[
    \frac{\partial n}{\partial t} + \frac{\partial f(n)}{\partial x} = 0.
\]

A model for the flux \( f(n) \) proposed in [88] reads

\[
    f(n) := u_m n \left(1 - \frac{n}{n_m}\right),
\]

with \( u_m \) the maximum speed of vehicles and \( n_m \) the maximum density of cars. We may verify that \( \hat{u} \) satisfies the equation

\[
    b(n) = f'(n) = u_m \left(1 - \frac{2n}{n_m}\right) = \eta,
\]

resulting in the similarity solution

\[
    u(x, t) = \frac{1}{2} n_m \left(1 - \frac{x}{u_m t}\right).
\]

See also Example 12.14.

Similarity solutions are also of importance for parabolic equations. In Chapter 10 we give a full account; here we restrict ourselves to an example.

**Example 2.10** Consider the heat equation

\[
    \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.
\]

(\*)

Note that if \( u(x, t) \) is a solution of (\*), then so is \( u_\alpha(x, t) := u(\alpha x, \alpha^2 t) \) for any \( \alpha > 0. \) Therefore an obvious choice for a similarity solution is

\[
    u(x, t) = \hat{u}(\eta), \quad \eta := \frac{x}{\sqrt{t}}.
\]
2.6 Initial and Boundary Conditions; Well-Posedness

Substituting \( \hat{u}(\eta) \) in (\ast), we obtain the ODE
\[
\frac{d^2 \hat{u}}{d\eta^2} + \frac{1}{2^9} \frac{d\hat{u}}{d\eta} = 0.
\]

This equation can be solved, with solution
\[
\hat{u}(\eta) = C_1 \int_0^{\frac{1}{2}} e^{-\tau^2} d\tau + C_2,
\]
where \( C_1, C_2 \) are constants to be determined from the initial and boundary conditions.

2.6 Initial and Boundary Conditions; Well-Posedness

For any differential equation one needs to specify the solution somewhere and somehow. The actual problem is then to find a solution of the PDE subject to certain conditions. If the latter are given at a time point onward from which the evolution takes place, we call them initial values. For the space domain we have (possibly) boundary conditions. For time-dependent problems we usually have both initial and boundary conditions. If we have purely initial conditions, we call the problem a Cauchy problem.

To determine whether the problem is meaningful we use the following definition.

Definition 2.11 (Hadamard’s well-posedness conditions). A problem is well posed if

(i) a solution exists,

(ii) the solution is unique, and

(iii) the solution depends continuously on the data, in particular the initial and boundary values.

Note that these conditions imply (i) that one should not have too many (conflicting) initial and boundary conditions, (ii) that we should not have too few, and (iii) that the effect of small perturbations is also small. The latter may be interpreted either in a local (small strip, interval, etc.) or global (infinite strip, interval, etc.) region.

In an elliptic problem the interaction between the coordinate points, as described by the equation, is in all directions. In time this is of course not possible. Therefore it is physically very unlikely that a time-dependent equation is of elliptic type. This is made precise by the following example.

Example 2.12 Consider the elliptic equation
\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad x \in \mathbb{R}, \quad y > 0,
\]
where \( u = u(x, y; n) \) is subject to
\[
u(x, 0; n) = 0, \quad x \in \mathbb{R},
u_y(x, 0; n) = \frac{1}{n} \sin nx, \quad x \in \mathbb{R}, \quad n \in \mathbb{N}.
\]
Chapter 2. Characterisation and Classification

This initial value problem was used by Hadamard to show that a Cauchy problem setting is not appropriate for elliptic problems. This can be seen as follows. One can easily check that

\[ u(x, y; n) := \frac{1}{n^2} \sin nx \sinh ny \]

is a solution on \( \mathbb{R} \times [0, \infty) \times \mathbb{N} \). For \( y > 0, x = \frac{1}{2} \pi \), and \( n \) odd we note that \( |u(x, y; n)| \to \infty \), however small \( y \) is. On the other hand, the initial conditions \( u(x, 0) = \frac{1}{n^2} u(x, 0) = 0 \) give the solution \( u(x, y) = 0 \). This shows that \( u \) does not depend continuously on the initial data; i.e., it violates criterion (iii).

Elliptic PDEs typically give rise to boundary value problems. We remark that elliptic operators will play a role by themselves as well as being part of parabolic or hyperbolic problems. Now consider a hyperbolic problem. Let \( D \) be a smooth curve and \( n \) denote the normal direction. Then a Cauchy problem has as initial values

\[ u = A(x, t), \quad \frac{\partial u}{\partial n} = B(x, t), \quad (x, t)^T \in D. \quad (2.49) \]

**Theorem 2.13.** Let \( D \) be a curve in \( \mathbb{R}^2 \) such that \( D \) intersects the characteristics only once. Then the Cauchy problem consisting of equation (2.16) and initial conditions (2.49) is well posed.

**Proof.** Let \( D \) have a representation \( \varphi(x, t) = 0 \). Let \( \psi(x, t) \) be such that \( \varphi, \psi \) is a genuine coordinate transformation (i.e., the Jacobian is nonzero). Then we can reformulate (2.16) in terms of \( \varphi, \psi \), giving

\[ \alpha \frac{\partial^2 u}{\partial \varphi^2} + \beta \frac{\partial^2 u}{\partial \varphi \partial \psi} + \gamma \frac{\partial^2 u}{\partial \psi^2} + \delta \frac{\partial u}{\partial \varphi} + \varepsilon \frac{\partial u}{\partial \psi} = \xi, \quad (*) \]

where

\[ \alpha = a \left( \frac{\partial \varphi}{\partial x} \right)^2 + b \frac{\partial \varphi}{\partial x} \frac{\partial \psi}{\partial x} + c \left( \frac{\partial \psi}{\partial x} \right)^2, \quad \text{etc.} \]

For \( u = u(\varphi, \psi) \) we then have the initial values \( u(0, \psi) = \hat{A}(\varphi), u_{\varphi}(0, \psi) = \hat{B}(\psi) \). Hence we can find \( u_{\varphi}(0, \psi), u_{\psi}(0, \psi), \) and \( u_{\varphi\psi}(0, \psi) \) but not \( u_{\varphi\psi}(0, \psi) \) from the initial conditions. If \( \alpha \neq 0 \), we can find \( u_{\varphi\psi}(0, \psi) \) and thus all higher-order derivatives from the transformed PDE (*). A formal solution away from \( D \) can now be found through a Taylor expansion:

\[ u(\hat{\varphi}, \hat{\psi}) = \sum_{i=0}^{\infty} \sum_{j=0}^{i} \frac{1}{j!(i-j)!} \frac{\partial^i u}{\partial \varphi^j \partial \psi^{i-j}} (\varphi - \hat{\varphi})^j (\psi - \hat{\psi})^{i-j}. \]

So at least the solution exists in a neighbourhood of \( D \); i.e., the problem is well posed. Thus we only need to prove \( \alpha \neq 0 \). However, this follows from the fact that if \( \xi(t, x) = 0 \) denotes a characteristic, then \( \det(A\xi + B\xi) = a\xi^2 + b\xi^2 + c\xi^2 = 0 \), while \( D \) was assumed not to be a characteristic.

The preceding theorem also applies to parabolic problems.
Example 2.14 Consider the PDE
\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in I \subset \mathbb{R}, \quad t > 0,
\]
subject to the initial conditions
\[
u(x, 0) = \alpha(x), \quad \frac{\partial u}{\partial t}(x, 0) = \beta(x).
\]
We can trivially see that \( u_{xx}(x, 0) \) should be equal to \( \beta(x) \), which is not true in general. If, on the other hand, we had a Cauchy problem in the \( x \) variable, i.e.,
\[
u(0, t) = \bar{\alpha}(t), \quad \frac{\partial u}{\partial x}(0, t) = \bar{\beta}(t),
\]
then we could still show the solution to exist; cf. Theorem 2.13.

Theorem 2.15. If (2.16) is parabolic, i.e., \( a = b = 0 \), and \( u(x, 0) = v(x) \) is given, then this defines a well-posed Cauchy problem, at least locally.

Proof. As in the proof of Theorem 2.13, one finds that \( u_x(x, 0) \) and \( u_{xx}(x, 0) \) are well defined. Hence \( u_t(x, 0) \) is well defined and then so are their higher derivatives. Hence we conclude the existence of \( u(x, t) \) in a neighbourhood of \( t = 0 \), whence we obtain local existence.

2.7 Discussion

• Classifying equations as elliptic, parabolic, and hyperbolic is more tradition than natural. The most important distinction is between boundary value problems and initial (boundary) value problems, which have an evolutionary character. In the latter the information travels with a “finite speed,” while for the former everything happens with “infinite speed.” As remarked in the discussion of the previous chapter, the use of spatial coordinates does not preclude the problem from having an evolution. The boundary data are then typically needed on only a part of the boundary.

• A possible classification of evolutionary PDEs is whether or not they allow for wave-like solutions; by this we mean solutions of the type \( f(x - st) \). Hyperbolic equations clearly have such solutions. But there also exist dispersive waves, which do not necessarily fit into the definition of hyperbolicity that was given here and will be used later in Chapter 12. For more details see [169].

Exercises

2.1. Consider the PDE
\[
\frac{\partial^2 u}{\partial x^2} + 4 \frac{\partial^2 u}{\partial x \partial y} + 3 \frac{\partial^2 u}{\partial y^2} + 3 \frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} + 2u = 0.
\]
(a) Show that it is hyperbolic.
(b) Find the characteristics and bring the PDE to normal form.
(c) Find a coordinate transformation such that the first order terms vanish in the resulting equation.

2.2. Consider the PDE
\[ \frac{\partial^2 u}{\partial x^2} + 2 \frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} + 5 \frac{\partial u}{\partial x} + 5 \frac{\partial u}{\partial y} + u = 0. \]

(a) Show that it is parabolic.
(b) Find the normal form.
(c) Find a coordinate transformation such that the first order terms vanish in the resulting equation.

2.3. Consider the PDE
\[ \frac{\partial^2 u}{\partial x^2} - 6 \frac{\partial^2 u}{\partial x \partial y} + 12 \frac{\partial^2 u}{\partial y^2} + 4 \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 0. \]

(a) Show that it is elliptic.
(b) Find the normal form.
(c) Find a coordinate transformation such that the first order terms vanish in the resulting equation.

2.4. Classify the PDE
\[ \frac{\partial^2 u}{\partial x^2} + 2 \frac{\partial^2 u}{\partial y^2} + \cos x \frac{\partial u}{\partial z} - e^y u = \cosh z. \]

2.5. Show that in \( d \)-dimensional space any second order elliptic PDE with constant coefficients can be brought to the form
\[ \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2} + cu = f. \]

2.6. Show that in \( n \)-dimensional space any second order hyperbolic PDE with constant coefficients can be brought to the form
\[ \frac{\partial^2 u}{\partial x_n^2} = \sum_{i=1}^{n-1} \frac{\partial^2 u}{\partial x_i^2} + cu + f. \]

2.7. Can you classify the PDE
\[ x^\alpha y^\beta \frac{\partial^2 u}{\partial x^2} + x^\gamma y^\delta \frac{\partial^2 u}{\partial y^2} + \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 0? \]
2.8. Consider the hyperbolic equation
\[
\frac{\partial^2 u}{\partial x \partial y} = 0
\]
on the unit square, where \( u \) is given on the boundary. Show that this problem is not well posed.

2.9. Consider the parabolic equation
\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}
\]
for \((x, t)\) in the positive \((x, t)\) plane. Let \( u(x, 0) \) be given. Show that this Cauchy problem is not well posed.

2.10. Determine the solution of
\[
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = u, \quad x \in \mathbb{R}, \quad t > 0,
\]
\[
u(x, 0) = e^{x^2}, \quad x \in \mathbb{R}.
\]

2.11. Find the characteristics of the equation
\[
\frac{\partial u}{\partial t} + x \frac{\partial u}{\partial x} = 0.
\]

2.12. Show that all travelling-wave solutions of the wave equation
\[
\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}
\]
are of the form \( u(x, t) = \tilde{u}_1(x - ct) \) and \( u(x, t) = \tilde{u}_2(x + ct) \). Consequently, the general solution (see Chapter 12) is given by
\[
u(x, t) = \tilde{u}_1(x - ct) + \tilde{u}_2(x + ct).
\]
Chapter 3

Fourier Theory

Fourier theory plays an important role in applied analysis. In this chapter we give an overview of the most important aspects needed in this book. First, we introduce an inner product and (orthonormal) basis functions in Section 3.1. Here we also define Fourier series, consider their convergence, and give Parseval’s identity. We give both the complex and the trigonometric representations. Finally, the integral analogue of a series is introduced and exemplified. Next, in Section 3.3, the discrete form of the Fourier transform is considered, derived from the continuous version. Again, convergence and Parseval’s identity are studied. Also, important phenomena, such as aliasing that show the restrictions of the discrete Fourier transform are treated. One very important application of this Fourier transform is in analysing linear equations with periodic boundary values. Despite the limitations of this problem class, it turns out that many physically meaningful concepts, such as stability, dissipation, and dispersion, can be studied quite fruitfully for the transformed equation, both in the continuous and in the discrete cases. In Section 3.4, therefore, the use of these transformations is demonstrated, leading to the important concept of the dispersion relation.

3.1 Fourier Series

A powerful tool in analysis is the expansion of a function $f$ in terms of suitably chosen functions that form a basis. There are several ways to find such expansions. The most elegant way to describe the function mathematically is to use projections, for which we need the concept of “orthogonality.” This is provided by using an inner product much like the one met in linear algebra (the “natural inner product”). Since this involves integration, we shall restrict ourselves to a finite interval, which is typically chosen as $(0, L)$. The functions we consider are square integrable, i.e., $L^2$-functions. Moreover, we shall assume that they are periodically extended to the full real axis. We then define the inner product for two such functions as

$$(f, g) := \int_0^L f(x)\overline{g(x)} \, dx,$$  \hspace{1cm} (3.1)$$

where the overbar $\overline{}$ denotes the complex conjugate (i.e., $a + ib = a - ib$). Note that from periodicity it follows that any integral over an interval of length $L$ is equivalent. As can
simply be verified, the functions
\[ p_j(x) := e^{i\alpha_j x}, \quad j = 0, \pm 1, \pm 2, \ldots, \]  
(3.2a)
where
\[ \alpha_j = \frac{2\pi j}{L}, \]  
(3.2b)
are orthogonal. If we change (3.1) into
\[ (f, g)_L := \frac{1}{L} \int_0^L f(x) \overline{g(x)} \, dx, \]  
(3.1')
the functions are even orthonormal; i.e., \((p_j, p_k) = \delta_{jk}\). We call \(\alpha_j\) a wave number if \(x\) denotes a spatial variable and a frequency if the independent parameter is time. In physical texts a wave number is usually denoted by the letter \(k\) and a frequency is denoted by \(\omega\).

Introducing
\[ c_j := \frac{1}{L} \int_0^L f(y) e^{-i\alpha_j y} \, dy, \]  
(3.3)
we can form a Fourier series
\[ \hat{f}(x) := \sum_{j=-\infty}^{\infty} c_j e^{i\alpha_j x}. \]  
(3.4)
The function \(p_j(x) = e^{i\alpha_j x}\) is called a Fourier mode with wave number (frequency) \(\alpha_j\). The important question is whether \(\hat{f}\) can be identified with \(f\) (and of course whether \(\hat{f}\) makes sense at all). We then say that the Fourier series of \(f\) converges to \(f\). In fact, no simple test is known that is both necessary and sufficient to relate a periodic function with its Fourier coefficients [25]. There is, however, a vast amount of partial results.

We have the following theorems.

**Theorem 3.1.** If for all points \(x \in (0, L)\) the left and right derivatives of \(f\), i.e.,
\[ \lim_{h \downarrow 0} \frac{f(x + h) - f(x)}{h} \quad \text{and} \quad \lim_{h \downarrow 0} \frac{f(x) - f(x - h)}{h} \]
exist, then the Fourier series (3.4) of \(f\) at \(x\) converges to \(f(x)\).

**Theorem 3.2.** If \(f\) is continuous in \(x\) and \(f(0) = f(L)\), then the Fourier series (3.4) of \(f\) converges uniformly to \(f\) (Section C in the appendix); i.e.,
\[ \lim_{N \to \infty} \sup_{x \in [0, L]} \left| \sum_{j=-N}^{N} c_j e^{i\alpha_j x} - f(x) \right| = 0. \]

**Definition 3.3.** The function \(f\) is piecewise continuous on \([0, L]\) if there is a finite number of open subintervals \(0 < x < x_1 < x_2 < \ldots < x_N < L\) on which \(f\) is continuous, while the limits \(f(0^+), f(x_1^\pm), \ldots, f(L^-)\) exist. The function \(f\) is piecewise smooth on \([0, L]\) if \(f\) and its derivative \(f'\) are both piecewise continuous.
3.1. Fourier Series 35

For such a function \( f \) we have the following theorem.

**Theorem 3.4 (existence of Fourier series).** If a function \( f \) is piecewise smooth on the interval \([0, L]\), while \( f(x) = \frac{1}{2}[f(x+) + f(x-)]\), then the Fourier series of \( f \) converges for every \( x \) to the \( L \)-periodic continuation of \( f \).

For a given Fourier series \( \sum c_j e^{i\alpha_j x} \) we have the following theorem.

**Theorem 3.5 (continuity of Fourier series).** If a Fourier series is absolutely convergent, i.e., \( \sum |c_j| < \infty \), then it converges absolutely and uniformly to a continuous periodic function \( f \) such that \( c_j \) are just the Fourier coefficients of \( f \).

If \( f \) is a function only in \( L^2 \), then we still have at least the following identity.

**Theorem 3.6 (Parseval’s identity).** Let \( f \in L^2(0, L) \) have Fourier coefficients \( c_j \). Then

\[
(f, f)_L = \sum_{j=-\infty}^{\infty} c_j^2.
\]

**Proof.** For a proof of Theorems 3.1, 3.2, and 3.6, see [25, 112].

**Corollary 3.7.** If \( f \) and \( f' \) are piecewise smooth, then the Fourier coefficients \( c_j \) of \( f \) behave asymptotically for \( j \to \infty \) like \( c_j = O(j^{-1}) \).

**Proof.** Integration by parts yields

\[
c_j = \frac{1}{i\alpha_j} \sum_{x=x_d} \left[ f(x) e^{-i\alpha_j x} \right]_{x_d}^{x_d+} - \frac{1}{i\alpha_j} c_j',
\]

where the summation runs over all points \( x_d \) of discontinuity of \( f \) (possibly including the endpoints) and \( c_j' \) is the \( j \)th Fourier coefficient of \( f' \). As \( c_j' \to 0 \), the result follows.

In a suitable context the inner product \((f, f)\) may be interpreted as an *energy content* and thus \((f, f)_L\) as a *mean energy*. Therefore Theorem 3.6 is sometimes referred to as Parseval’s energy theorem.

Often it is useful to rewrite the Fourier series in terms of trigonometric functions. Using the well-known relation

\[
e^{iz} = \cos z + i \sin z,
\]

we thus find from (3.4) that

\[
\tilde{f}(x) = a_0 + \sum_{j=1}^{\infty} a_j \cos(\alpha_j x) + b_j \sin(\alpha_j x),
\]

where

\[
a_0 = c_0, \quad a_j = c_j + c_{-j}, \quad b_j = i(c_j - c_{-j}).
\]
We can more directly write
\[ a_0 = \frac{1}{L} \int_0^L f(y) \, dy, \quad a_j = \frac{2}{L} \int_0^L f(y) \cos(\alpha_j y) \, dy, \]
\[ b_j = \frac{2}{L} \int_0^L f(y) \sin(\alpha_j y) \, dy, \quad j = 1, 2, \ldots \quad (3.7) \]

Theorems 3.1 and 3.2 carry over to the trigonometric representations and for Parseval’s identity (Theorem 3.6) we have
\[ (f, f)_L = a_0^2 + \frac{1}{2} \sum_{j=1}^\infty (a_j^2 + b_j^2). \quad (3.8) \]

Example 3.8 Consider the function
\[ f(x) = x, \quad x \in (-\pi, \pi]. \]
In order to make it periodic, we extend this function periodically beyond \((-\pi, \pi]\) to obtain the sawtooth function, as in Figure 3.1. It is then straightforward to see that \(f \in L_2(-\pi, \pi).\) Since
\[ a_j \propto \int_{-\pi}^{\pi} y \cos(jy) \, dy = 0 \quad \text{for all } j \]
and
\[ b_j = \frac{1}{\pi} \int_{-\pi}^{\pi} y \sin(jy) \, dy = \frac{1}{\pi} \left[ \frac{-y \cos jy}{j} + \frac{\sin jy}{j^2} \right]_{-\pi}^{\pi} = \frac{2}{j} (-1)^{j+1}, \]
we deduce, on account of Theorem 3.1 (we take \(L = 2\pi\) and the interval \((-\pi, \pi]\) in (3.3)), that the resulting series
\[ \tilde{f}(x) = 2 \left[ \sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \cdots \right] \]
converges to \(f(x)\) for any \(x \in (-\pi, \pi).\)

However, we cannot guarantee uniform convergence (see section C in the appendix) on the whole interval. At \(x = \pi\) the series does not converge to \(f(\pi) = \pi\) but rather to zero, the average of the left and right limits, as is shown graphically by Figure 3.2. We see an interesting phenomenon at the discontinuities of \(f\): there is an overshoot to the left and an undershoot to the right. This is known as Gibbs’s phenomenon. \(\square\)

![Figure 3.1. Sawtooth function.](image-url)
3.1. Fourier Series

The function in Example 3.8 was clearly odd and so were the Fourier terms. This is generally true. If \( f \) is odd (i.e., \( f(-x) = -f(x) \)), we have a Fourier sine series, and if \( f \) is even (i.e., \( f(-x) = f(x) \)), we have a Fourier cosine series.

Example 3.9 The following Fourier sine and cosine series define periodic functions with period one:

\[
\begin{align*}
\sum_{n=1}^{\infty} \frac{\sin(2\pi nx)}{\pi n} &= \left[ \frac{1}{2} - x \right], & \sum_{n=1}^{\infty} \frac{\cos(2\pi nx)}{n} &= -\log |2 \sin \pi x|, \\
\sum_{n=1}^{\infty} \frac{\cos(2\pi nx)}{n^2} &= \left[ x^2 - x + \frac{1}{6} \right], & \sum_{n=1}^{\infty} \frac{\cos(2\pi nx)}{n^2 - \frac{1}{4}} &= 2 - \pi |\sin \pi x|.
\end{align*}
\]

\([\cdot]\) denotes a function originally defined on \([0, 1]\) and continued periodically. Another interesting example is the block-wave function, defined along \([-1, 1]\) by

\[
4 \sum_{n=0}^{\infty} \frac{\sin((2n+1)\pi x)}{(2n+1)\pi} = \text{sign}(x)
\]

and 2-periodically continued otherwise. \(\Box\)

In a straightforward way Fourier series are generalised to several dimensions. Suppose that the square-integrable function \( f : \mathbb{R}^d \to \mathbb{R} \) is periodic in every argument \( x_j \), \( j = 1, \ldots, d \). Then

\[
f(x_1, \ldots, x_d) := \sum_{m_1 = -\infty}^{\infty} \cdots \sum_{m_d = -\infty}^{\infty} C_{m_1, \ldots, m_d} e^{i\alpha_{m_1}^{(j)} x_1 + \cdots + i\alpha_{m_d}^{(j)} x_d}, \quad (3.9)
\]

where

\[
\alpha_{m_j}^{(j)} = \frac{2\pi}{L_j} m_j
\]

and

\[
C_{m_1, \ldots, m_d} = \frac{1}{L_1 \cdots L_d} \int_0^{L_1} \cdots \int_0^{L_d} f(y_1, \ldots, y_d) e^{-i\alpha_{m_1}^{(j)} y_1 - \cdots - i\alpha_{m_d}^{(j)} y_d} \, dy_1 \cdots dy_d. \quad (3.10)
\]
If \( f \) is scaled such that it is \( 2\pi \)-periodic in each independent variable, (3.9) may be written more compactly as
\[
f(x) := \sum_{m=-\infty}^{\infty} C_m e^{im \cdot x}, \quad C_m = \frac{1}{(2\pi)^d} \int_0^{2\pi} f(y) e^{-im \cdot y} dy,
\]
where \( m \in \mathbb{Z}^d \) denotes the index vector \( m = [m_1, \ldots, m_d] \) and the sum and integral signs are to be interpreted \( d \)-fold.

### 3.2 Fourier Transforms

There exists an integral analogue to the Fourier series. Recalling (3.2), we may let \( L \to \infty \); i.e., \( \Delta \alpha \to 0 \). Writing
\[
\Delta \alpha := \frac{1}{L},
\]
we have for an \( L \)-periodic, square-integrable function \( f \) that
\[
c_j = \Delta \alpha \int_0^{(\Delta \alpha)^{-1}} f(y) e^{-2\pi i j \Delta \alpha y} dy,
\]
whence
\[
f(x) = \sum_{j=-\infty}^{\infty} e^{2\pi i j \Delta \alpha x} \Delta \alpha \int_0^{(\Delta \alpha)^{-1}} f(y) e^{-2\pi i j \Delta \alpha y} dy.
\]
Here we recognize a Riemann sum of the function \( g(z) \), where
\[
g(z) := \frac{1}{2\pi} e^{izx} \int_0^{(\Delta \alpha)^{-1}} f(y) e^{-izy} dy.
\]
One should just take piecewise constant approximations of \( g \) at the points \( 2\pi j \Delta \alpha \), \( j = 0, \pm 1, \pm 2, \ldots \), multiplied by the interval width \( 2\pi \Delta \alpha \). Hence by a limit argument we find
\[
f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} f(y) e^{-i\alpha y} dy \right] e^{i\alpha x} d\alpha,
\]
which leads to \( \hat{f} \), the Fourier transform of \( f \), which is, together with its inversion, given by
\[
\hat{f}(\alpha) := \int_{-\infty}^{\infty} f(x) e^{-i\alpha x} dx, \quad (3.17a)
\]
\[
f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\alpha) e^{i\alpha x} d\alpha. \quad (3.17b)
\]
(Note that other, equivalent, definitions frequently occur, causing a lot of confusion.) The Fourier transform (also called the Fourier integral) plays an important role in analysing problems with a continuous spectrum of wave numbers (or frequencies). One can show that \(|f|\) and \(|\hat{f}|\) cannot vanish simultaneously outside a finite domain. Note that it is sufficient for the existence of \( \hat{f} \) that \( f \in L_1 \), but if \( f \in L_2 \), then also \( \hat{f} \in L_2 \) [25] and therefore both (3.17a) and (3.17b) exist.
Example 3.10

(i) The “top hat” function, given by
\[
f(x) = \begin{cases} 
1, & -1 \leq x \leq 1, \\
0, & |x| > 1,
\end{cases}
\]
has a Fourier transform that is closely related to the sinc-function:
\[
\hat{f}(\alpha) = \int_{-1}^{1} e^{-i\alpha x} \, dx = \frac{e^{i\alpha} - e^{-i\alpha}}{i\alpha} = 2 \frac{\sin \alpha}{\alpha} = 2 \text{sinc} \left( \frac{\alpha}{\pi} \right).
\]

(ii) The decaying exponential, vanishing for \( x < 0 \), defined by
\[
f(x) = \begin{cases} 
e^{-px}, & x > 0, \\
0, & x < 0,
\end{cases}
\]
where \( p > 0 \), has a Fourier transform consisting of a single pole in the upper complex \( \alpha \) plane:
\[
\hat{f}(\alpha) = \int_{0}^{\infty} e^{-px} e^{-i\alpha x} \, dx = \left[ -\frac{e^{-p(\alpha+i)x}}{p+i\alpha} \right]_{0}^{\infty} = \frac{1}{p+i\alpha}.
\]

(iii) Another important example is
\[
f(x) = p^{-1/2} e^{-\frac{1}{2}(x/p)^2},
\]
with the similar
\[
\hat{f}(\alpha) = \int_{-\infty}^{\infty} p^{-1/2} e^{-\frac{1}{2}(x/p)^2} e^{-i\alpha x} \, dx = (2\pi p)^{1/2} e^{-\frac{1}{2}(p\alpha)^2}.
\]

Note that where \( f \) depends on \( p \), \( \hat{f} \) depends on its reciprocal.

If \( \alpha^n \hat{f}(\alpha) \) is Fourier transformable, where \( n \in \mathbb{N} \) and \( \hat{f}(\alpha) \) is the Fourier transform of \( f(x) \), then
\[
\frac{d^n}{dx^n} f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\alpha)^n \hat{f}(\alpha) e^{i\alpha x} \, d\alpha.
\]  

Definition 3.11. The convolution product \( f * g \) of \( f, g \in L_2 \) is defined as
\[
(f * g)(x) = \int_{-\infty}^{\infty} f(x - y)g(y) \, dy.
\]

Theorem 3.12 (convolution theorem). The Fourier-transformed convolution product of \( f, g \in L_2 \) is the product of their Fourier transforms:
\[
\hat{f} \ast \hat{g} = \hat{f} \hat{g}.
\]
In other words, the inverse Fourier transform of \( \hat{f} \hat{g} \) is equal to the convolution product:
\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\alpha)\hat{g}(\alpha) e^{i\alpha x} \, d\alpha = \int_{-\infty}^{\infty} f(x - y)g(y) \, dy.
\]
Proof. As \( f \) and \( g \) are square integrable, we may change the order of integration to get
\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\alpha) \hat{g}(\alpha) e^{i\alpha x} \, d\alpha = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\alpha) \int_{-\infty}^{\infty} g(y) e^{-i\alpha y} \, dy \, e^{i\alpha x} \, d\alpha
\]
\[
= \frac{1}{2\pi} \int_{-\infty}^{\infty} g(y) \int_{-\infty}^{\infty} \hat{f}(\alpha) e^{i\alpha(x-y)} \, d\alpha \, dy = \int_{-\infty}^{\infty} f(x-y)g(y) \, dy. \tag{3.19}
\]

If we consider Theorem 3.12 for \( x = 0 \) and take \( g(y) = \overline{f(-y)} \) with \( \hat{g}(\alpha) = \overline{\hat{f}(\alpha)} \), we obtain the analogue of Parseval’s identity (Theorem 3.6) for integrals:
\[
\int_{-\infty}^{\infty} |f(x)|^2 \, dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{f}(\alpha)|^2 \, d\alpha. \tag{3.19}
\]
This is sometimes referred to as the energy theorem for the continuous case.

Example 3.13 Consider Example 3.10(ii) again. We obtain indeed
\[
\int_{-\infty}^{\infty} |f(x)|^2 \, dx = \int_{0}^{\infty} e^{-2px} \, dx = \frac{1}{2p},
\]
\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{f}(\alpha)|^2 \, d\alpha = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\alpha^2/2p^2} \, d\alpha = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\alpha^2/2p^2} \, d\alpha = \frac{1}{2p}. \tag{3.20}
\]

Let \( f \) be a Fourier-transformable function on \( \mathbb{R} \). Noting that the function
\[
\sum_{m=-\infty}^{\infty} f(mL + x)
\]
is periodic in \( x \) with period \( L \), we can write (following [20])
\[
\sum_{m=-\infty}^{\infty} f(mL + x) = \sum_{j=-\infty}^{\infty} \left[ \frac{1}{L} \int_{0}^{L} \sum_{m=-\infty}^{\infty} f(mL + y) e^{-i\alpha_j y} \, dy \right] e^{i\alpha_j x}
\]
\[
= \frac{1}{L} \sum_{j=-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} f(\eta) e^{-i\alpha_j \eta} \, d\eta \right] e^{i\alpha_j x} = \frac{1}{L} \sum_{j=-\infty}^{\infty} \hat{f}(\alpha_j) e^{i\alpha_j x}
\]
(with \( \alpha_j \) as defined in (3.2b)), which leads, for \( x = 0 \), to Poisson’s formula
\[
\sum_{m=-\infty}^{\infty} f(mL) = \frac{1}{L} \sum_{m=-\infty}^{\infty} \hat{f} \left( \frac{2\pi}{L} m \right), \tag{3.21}
\]
where the average of the left and right limit is to be taken at any discontinuities.

Example 3.14 Poisson’s formula is an excellent tool for accelerating slowly converging series. From Example 3.10(ii) we find, for \( p > 0 \) and \( L = 2\pi \), that
\[
\frac{1}{2\pi p} + \frac{p}{\pi} \sum_{m=1}^{\infty} \frac{1}{m^2 + p^2} = \frac{1}{2} + \sum_{m=1}^{\infty} e^{-2\pi pm}.
\]
The left-hand side converges algebraically slowly, in contrast to the fast exponential convergence of the right-hand side. As a bonus, we have in this case an explicit expression if we recognize the geometric series with common ratio \( e^{-2\pi p} \).
3.2. Fourier Transforms

We are interested in knowing when a given \( \hat{f}(\omega) \) is a time-Fourier transform of a physical signal \( f(t) \). First, in order for \( f \) to be real, \( \hat{f} \) has to satisfy the \textit{reality condition}

\[
\hat{f}(\omega) = \hat{f}(-\omega). \tag{3.22}
\]

No physical process can exist for all time. A process \( f(t) \) that starts by some cause at some finite time \( t = t_0 \) and vanishes before \( t_0 \) is called \textit{causal}. The corresponding Fourier transform

\[
\hat{f}(\omega) = \int_{t_0}^{\infty} f(t) \ e^{-i\omega t} \ dt
\]

has the property that \( \hat{f}(\omega) \) is analytic in the lower complex half-plane \( \text{Im}(\omega) < 0 \). So this is a necessary condition on \( \hat{f} \) for \( f \) to be causal. A sufficient condition is the following \textit{causality condition} [112].

\textbf{Theorem 3.15 (causality condition).} If \( \hat{f}(\omega) \) is analytic in \( \text{Im}(\omega) \leq 0 \), \( |\hat{f}(\omega)|^2 \) is integrable along the real axis, and there is a real \( t_0 \) such that \( \hat{f}(\omega) e^{i\omega t_0} \to 0 \) uniformly with regard to \( \text{arg}(\omega) \) for \( |\omega| \to \infty \) in the lower complex half-plane, then \( f(t) \) is causal and vanishes for \( t < t_0 \).

\textbf{Proof.} It is no restriction for the proof to assume that \( t_0 = 0 \). Consider, for \( t < 0 \), the integral

\[
\int \hat{f}(\omega) e^{i\omega t} \ d\omega
\]

along the real contour \([-R, R]\) to be closed via a semicircle of radius \( R \) in the lower complex half-plane. As the integrand is analytic, the integral is zero. Let \( R \to \infty \).

The contribution \( I_R \) of the integral along the semicircle tends to zero because

\[
|I_R| \leq \int_0^\pi |\hat{f}(\omega)| e^{-|R \sin \theta|} R \, d\theta \leq 2R \max_\theta |\hat{f}(\omega)| \int_0^{\pi/2} e^{-|\theta|R^2/\pi} \, d\theta \to 0,
\]

where \( \omega = Re^{i\theta} \). So the contribution from the real axis, being equal to \( 2\pi f(t) \), is also zero.

Note that the \textit{lower} complex half-space becomes the \textit{upper} half-space if the opposite Fourier sign convention is taken.

\textbf{Example 3.16} The Fourier transform \( \hat{f}(\omega) = (p + i\omega)^{-1} \) is causal if \( p > 0 \), as may be confirmed by the inverse transform

\[
f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{i\omega t}}{p + i\omega} \ d\omega = \begin{cases} e^{-pt} & \text{if } t > 0, \\ 0 & \text{if } t < 0. \end{cases}
\]

In the limit of no damping (\( p \downarrow 0 \)) the singularity at \( \omega = i\omega \) moves to \( \omega = 0 \), which is on the real axis. The integral is to be interpreted via a suitable indentation of the contour \textit{under} the pole in order to retain causality.
In a straightforward way Fourier transforms can be generalised to several dimensions. For the square-integrable function \( f: \mathbb{R}^d \to \mathbb{R} \) we have the couple

\[
\hat{f}(\alpha) := \int_{-\infty}^{\infty} f(x) e^{-i \alpha \cdot x} \, dx,
\]

(3.23a)

\[
f(x) = \frac{1}{(2\pi)^d} \int_{-\infty}^{\infty} \hat{f}(\alpha) e^{i \alpha \cdot x} \, d\alpha,
\]

(3.23b)

where \( \alpha \in \mathbb{R}^d \) denotes the Fourier wave number vector and the integral signs are to be interpreted \( d \)-fold.

### 3.3 Discrete Fourier Transforms

Often a function is known at a certain finite number of points only. For instance, \( f \) may be measured at a few values of the argument (“sampled”). If one uses a numerical approximation method based on discretising of the argument, as in the case of finite differences, there is also a natural set of points on which \( f \) is “monitored” only. Given the grid points \( x_k, k = 0, \ldots, N - 1 \), we can give straightforward analogues of the continuous case by considering the discrete inner product

\[
\langle f, g \rangle := \sum_{k=0}^{N-1} f(x_k) g(x_k),
\]

(3.24)

where \( x_k \in [0, L] \). Because of the special properties exhibited by \( e^{i \alpha x} \) when the points \( x_k \) are chosen equispaced, i.e., \( x_{k+1} - x_k \) is constant \( (k = 0, \ldots, N - 1 \) and \( x_N = x_0 + L \)), we shall take them as

\[
x_k := \frac{k}{N} L, \quad k = 0, \ldots, N - 1.
\]

(3.25)

It can be verified that the functions

\[
p_j(x) := e^{i \alpha_j x}
\]

(3.26)

with

\[
\alpha_j = \frac{2\pi j}{L}
\]

are orthogonal. This is stated in the following theorem.

**Theorem 3.17.** The polynomials \( p_j \) defined in (3.26) are orthogonal with respect to (3.24) in the sense that \( \langle p_j, p_l \rangle = 0 \) for all \( j, l \), with \( j - l \) not a multiple of \( N \).

**Proof.** Substituting (3.26) in (3.24), we have

\[
\langle p_j, p_l \rangle = \sum_{k=0}^{N-1} \exp \left[ 2\pi i (j - l) \frac{k}{N} \right] = \sum_{k=0}^{N-1} \sigma^k = \begin{cases} 
\sigma^N - 1 & \text{if } \sigma \neq 1, \\
\sigma - 1 & \text{if } \sigma = 1,
\end{cases}
\]

with \( \sigma := \exp \left( 2\pi i \frac{j - l}{N} \right) \). Apparently, \( \sigma^N = 1 \), while \( \sigma \neq 1 \) if \( (j - l) \mod N \neq 0 \), which proves the result. \( \square \)
Corollary 3.18. \( (p_j, p_j) = N \).

We therefore may as well consider the inner product

\[
(f, g)_N := \frac{1}{N} \sum_{k=0}^{N-1} f(x_k) g(x_k),
\]

which makes \( \{p_j(x)\}_{j=0}^{N-1} \) an orthonormal set of basis functions. We can now give a discrete Fourier transform (DFT) of a function \( f \):

\[
c_j := (f, p_j)_N = \frac{1}{N} \sum_{k=0}^{N-1} f(x_k) e^{-2\pi i jk/N}.
\]

Due to the special choice of the grid, we immediately find that \( \exp(i\alpha x_k) = 1 \) when \( j \) is a multiple of \( N \). Hence for \( j \geq N \) we have

\[
e^{i\alpha jx_k} = e^{2\pi i \hat{j}k/N}, \quad \hat{j} = j \mod N.
\]

This relation tells us that the basis functions \( p_j(x) \), for \( j \geq N \), will not provide additional information to represent (an approximation of) \( f \). The phenomenon in which we cannot distinguish discrete Fourier components of \( p_j \) and \( p_{j+IN} \), \( l \in \mathbb{Z} \), is called aliasing. Hence we will be satisfied to have the finite series

\[
\hat{f}(x) := \sum_{j=0}^{N-1} c_j e^{2\pi i jx}, \quad x \in [0, L].
\]

Of particular interest is \( \hat{f}(x) \) at the points \( x = x_k \), because here the original values of \( f \) are exactly recovered; i.e., \( f(x_k) = \hat{f}(x_k) \). So at these points \( f \) is completely defined by the coefficients \( c_j \) and vice versa. If we define

\[
\hat{c}_k := \frac{1}{\sqrt{N}} f(x_k),
\]

the following reciprocity relation between \( c_j \) and \( \hat{c}_k \) may be shown.

Property 3.19 (reciprocity of DFT).

\[
\hat{c}_k = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} c_j e^{2\pi i jk/N}, \quad c_j = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \hat{c}_k e^{-2\pi i jk/N}.
\]

We remark that the DFT has very important applications, in particular through its efficient implementation the fast Fourier transform (FFT; see [149]).

From an approximation point of view, aliasing implies that we cannot obtain more information about a function \( f \) than the sampling rate (the density of the grid) allows us; see Example 3.20. In particular, we thus conclude that the discrete wave numbers generated by the grid will limit the accuracy of the approximation \( \hat{f}(x) \approx f(x) \).
Example 3.20 In Figure 3.3 we have drawn two sine functions, one with wave number 1, sampled with a rate of \(20/2\pi\), and one with wave number 21. As \(\sin x\) coincides with \(\sin(21x)\) right at the sampling points, \(\sin(21x)\) cannot be represented with this sampling rate.

![Figure 3.3. Aliasing.](image)

3.4 Fourier Analysis Applied to PDEs

Consider now the linear PDE with constant coefficients

\[
\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial u}{\partial x} + cu. \tag{3.31}
\]

We may look for solutions of this equation by using a Fourier series (on a finite domain) or Fourier integrals (on an infinite domain). Of course this does not make much sense in general if we do not specify the initial and boundary conditions. Yet one may hope that an analysis based on Fourier expansions—if successful—might give insight into the general case. If for instance such an analysis leads to prediction of instabilities or other undesirable phenomena, it may imply such a result for a more general case.

Since we have at least two variables, space and time, our ansatz will be

\[
u(x, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{u}(\kappa, \omega) e^{i\kappa x - i\omega t} \, d\kappa \, d\omega, \tag{3.32}
\]

where \(\omega\) denotes frequency and \(\kappa\) is the wave number. For the natural definition of phase and group velocity below, these Fourier variables are defined with opposite signs; we could have used \(e^{i\kappa x - i\omega t}\) equally well in (3.32). This expression is a very general one that is valid for any function \(u(x, t)\). We are not dealing with any \(u\), but with a solution of (3.31), and usually a simpler form is possible, e.g., if we require each Fourier mode to be a solution of the defining equation (3.31). In order to analyse the solution, we consider the single mode

\[
u(x, t) = A e^{i\kappa x - i\omega t}. \tag{3.33}
\]

(Such a planar wave solution is also useful for analyzing nonlinear evolution equations; see [169].) Upon substituting (3.33) into (3.31), we obtain that the mode is a solution if

\[
-i\omega = -a\kappa^2 + ib\kappa + c. \tag{3.34}
\]

Since this relation gives information about the propagation properties of the various modes, it is often referred to as the dispersion relation (dispersion is explained below).
If we follow a modal wave crest, i.e., such that the phase $\kappa x - \omega t = \text{constant}$, we move with the *phase velocity* or *wave speed*

$$v_p := \frac{\omega}{\kappa}. \quad (3.35)$$

As a mode has an infinite extension in $x$, it is hard to tell by which velocity any associated properties, like energy, propagate. Therefore we consider a “localised mode,” or wave packet, that decays slowly to zero for large $|x|$. This is not exactly one mode any more (with a single frequency and wave number), but a superposition of modes near a main frequency and wave number. To be more precise let $f(x)$ be an absolute integrable smooth function and $\varepsilon$ ($\varepsilon$ positive) be small compared to $\kappa_0$ such that $f(\varepsilon x)$ is the slowly varying envelope of the wave at $t = 0$:

$$u(x, 0) = e^{i\kappa_0 x} f(\varepsilon x). \quad (3.36a)$$

The Fourier transform of (3.36a) is given by

$$\int_{-\infty}^{\infty} e^{-i(x-k_0 \lambda)} f(\varepsilon x) dx = \varepsilon^{-1} \hat{f} \left( \frac{\kappa - \kappa_0}{\varepsilon} \right). \quad (3.36b)$$

Let the modes be defined by a dispersion relation $\omega = \omega(\kappa)$. Hence we can write the wave packet as

$$u(x, t) := \frac{1}{2\pi} \int_{-\infty}^{\infty} \varepsilon^{-1} \hat{f} \left( \frac{\kappa - \kappa_0}{\varepsilon} \right) e^{i\kappa x - i\omega(\kappa) t} \, d\kappa$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\eta) e^{i\kappa_0 x + i\varepsilon \eta x - i\omega_0(t + \varepsilon \eta) t} \, d\eta. \quad (3.37)$$

By Taylor expanding $\omega(\kappa_0 + \varepsilon \eta) = \omega_0 + \varepsilon \eta \omega_0' + \mathcal{O}(\varepsilon^2)$ for small $\varepsilon$—where $\omega_0 = \omega(\kappa_0)$ and $\omega_0' = \omega'(\kappa_0)$—and rearranging terms, we find in good approximation that

$$u(x, t) = \frac{1}{2\pi} e^{i\kappa_0 x - i\omega_0 t} \int_{-\infty}^{\infty} \hat{f}(\eta) e^{i\varepsilon \eta x - i\omega'_0 t} \, d\eta$$

$$= e^{i\kappa_0 x - i\omega_0 t} f(\varepsilon(x - \omega'_0 t)). \quad (3.37)$$

We see that the wave crests indeed propagate with the phase velocity $\omega_0/\kappa_0$, whereas the group as a whole propagates with a velocity $\omega'(\kappa_0)$. This velocity is called the *group* or *energy* or *signal velocity* and is thus given by

$$v_g := \frac{d\omega}{d\kappa}. \quad (3.38)$$

In general, all modes propagate with their own speed, and a group superposed of many different modes gets dispersed. The shape of the group remains intact, i.e., shows no dispersion, if all modes propagate with the same speed, i.e., if

$$\frac{d^2 v_g}{d\kappa^2} = \frac{d^2 \omega}{d\kappa^2} = 0. \quad (3.39)$$
Chapter 3. Fourier Theory

Note that a mode like $e^{-i\omega t} e^{i(kx - bt)}$ is clearly not dispersive, so condition (3.39) is not exactly equivalent to a condition of constant phase speed.

We return to (3.31) and consider two special cases. First, let $b$ and $c$ be zero; i.e., we have the standard heat equation. We then find upon substituting $-i\omega = -a\kappa^2$ that

$$u(x, t) = e^{ikx} e^{-a\kappa^2 t}. \quad (3.40)$$

The first factor is just the spatial Fourier component, but the second is an exponentially growing/decaying quantity. Hence we see that we need $a > 0$ to have a stable mode, and we may call solution (3.40) dissipative.

The other special case is when $a = c = 0$. Then we have

$$\omega = -bx. \quad (3.41)$$

This means that each mode is propagating with the same wave speed $-b$. In particular, we find

$$u(x, t) = e^{i(x+bt)x}. \quad (3.42)$$

Hence on a line in the $(x, t)$ plane where $x + bt$ is constant, i.e., on a characteristic (cf. Chapter 2), we note that $u$ is constant. We may therefore call solutions like $u$ in (3.41) conservative.

In the context of numerical methods for PDEs based on finite differences, analysing error propagation leads to typically linearized equations with slowly varying coefficients. Take, e.g., the "true" solution $y$ satisfying

$$y_t = (A(y)y_x)_x + C(y), \quad (3.43)$$

the numerical solution $\tilde{y}$, and the error $u := y - \tilde{y}$. The error is by assumption small. Its typical fluctuations, due to the small grid and time steps, vary over a much shorter time and length scale than the "true" solution. As we are only interested in the error, i.e., the behaviour along the short time and length scales, the linearized equation for $u$ in the neighbourhood of $x = x_0$ and $t = t_0$ may look like

$$\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial u}{\partial x} + cu, \quad (3.44)$$

where $a$, $b$, and $c$ are assumed constant, depending on $y(x_0, t_0)$. (Note that this equation is just an example and has no other purpose than to illustrate.) As before, the behaviour of $u$ in $x$ and $t$ may be analysed by Fourier analysis. Assume for $u$ a single mode of the form (3.33), i.e., with constant amplitude and phase. Substitution in (3.44) yields dispersion relation (3.34).

Fourier analysis in difference equations will turn out to be a powerful tool to determine necessary conditions for a numerical method to be practically useful. We shall defer details to the specific chapters where we assess the numerical methods. Refinements based on the method of multiple scales (Chapter 15, Section 15.4.2) allow us to include the variation of amplitude and phase with the (relatively) slowly varying $x_0$, $t_0$. 
3.5 Discussion

- Fourier theory is an essential tool in many applications that are far beyond the goals of this book. Traditionally, the Fourier series have been used to approximate functions. There is a host of other choices for this, depending on the application; see, e.g., books on special functions, like [107]. Quite another application is the efficient solution of systems through the FFT; see [118].

- Dispersion is a very important concept for the theory of waves. As it will turn out in Chapters 12, 13, and 14 in both the analytical and the numerical study of hyperbolic problems, the actual behaviour of the solution critically depends on properties like dissipation and dispersion. In fact, one of the major problems in numerically solving hyperbolic problems is to capture the physical behaviour, i.e., not to introduce too much “numerical” dissipation or dispersion. Finally, we remark that there are many equations with wave-like solutions that are not of hyperbolic type. These are called dispersive waves [169].

Exercises

3.1. Let \( L \) be some positive number. Show that the functions \( \frac{1}{2} \sqrt{2} s, s \cos(j \pi s^2 x), \) and \( s \sin(j \pi s^2 x), \) with \( s := L^{-\frac{1}{2}}, j = 1, 2, \ldots, \) form an orthonormal basis on \((-L, L)\).

3.2. Show that an even function on \((-L, L)\) is orthogonal to an odd function.

3.3. Prove the reciprocity relation for DFT (Property 3.19).

3.4. (a) Find the Fourier series of \( \cos^2 x \).

(b) Find the Fourier series of \( x^2 \) defined on \([0, 1]\).

(c) Consider the function \( f(x) = x(x - 1) \) defined on \([0, 1]\). Let \( \sum_{j=1}^{\infty} \gamma_j \sin(j \pi x) \) be the Fourier sine series of \( f \). Show that \( \gamma_j = 0 \) for \( j \) even and \( \gamma_j = -8/(j \pi)^3 \) for \( j \) odd.

3.5. Given the Fourier coefficients of \( f(x) \), determine the Fourier coefficients of the first and second derivatives.

3.6. If \( f(x) \) is a periodic \( L^2 \)-function with a jump condition at \( a \) say, then the Fourier series at that point converges to \( \frac{1}{2} (f(a+) + f(a-)) \), where \( a+ \) denotes the right limit and \( a- \) the left limit. Show this for \( f(x) = x \) on \((-\pi, \pi)\).

3.7. Find the Fourier series of the following function, defined on \([0, 1]\):

\[
f(x) = \frac{2}{3} x^3 - x^2 + \frac{1}{3} x.
\]

3.8. Show that the functions given by

\[
f_j(x) = \sin \left( \left( j + \frac{1}{2} \right) \pi x \right), \quad j \in \mathbb{Z},
\]

are orthonormal on \([-1, 1]\).
3.9. Show by utilizing Poisson’s formula that
\[ \sum_{m=1}^{\infty} \frac{\sin m\pi}{m} = \frac{\pi - 1}{2}. \]

3.10. Determine
\[ \sum_{j=0}^{\infty} \frac{(-1)^j}{2j + 1} \]
from the Fourier series of \( \frac{1}{2} - x \) on \([0, 1]\).

3.11. (a) Determine the dispersion relation for the beam equation
\[ \frac{\partial^2 u}{\partial t^2} + c^2 \frac{\partial^4 u}{\partial x^4} = 0. \]
(b) Repeat part (a) for the Korteweg–de Vries equation
\[ \frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} + d \frac{\partial^3 u}{\partial x^3} = 0. \]
(c) Repeat part (a) for the Boussinesq equation
\[ \frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = b^2 \frac{\partial^4 u}{\partial x^2 \partial t^2}. \]

3.12. The dispersion relation for water waves is given by
\[ \omega^2 = \left(1 + \frac{T}{\rho g} \kappa^2 \right) g \kappa \tanh(\kappa h), \]
where \(\omega\) is the frequency, \(\kappa\) is the wave number, \(h\) is the undisturbed water height, \(g = 9.8 \text{ m/s}^2\) is the gravitational acceleration, \(\rho = 1000 \text{ kg/m}^3\) is the water density, and \(T = 0.074 \text{ N/m}\) is the surface tension. Waves controlled by surface tension \((T\kappa^2/\rho g\) is not small) are called ripples. Waves controlled by just gravity are called gravity waves.

(a) Verify that for deep water the phase velocity of gravity waves is twice the group velocity, so the waves are dispersive.

(b) Verify that for long waves \((\kappa \to 0)\) the group and phase velocity become the same (and equal to \(\sqrt{gh}\)), so long waves propagate without dispersion.

(c) A practical parameter to maximize the wave number range of dispersionless waves is the water depth \(h\). A device that uses these waves to physically model sound waves (which are dispersionless) is called a ripple tank [90]. Consider \(V(z, \beta) = v_z/\sqrt{gh}\) as a function of \(z = \kappa h\) and \(\beta = T/\rho gh^2\). Our aim is to select \(\beta\) such that \(V\) remains close to one for a considerable interval \(0 < z < z_0\). Verify that \(V(0, \beta) = 1\) and \(V_z(0, \beta) = V_{zzz}(0, \beta) = 0\). For what value of \(\beta\) is \(V_z(0, \beta) = 0\)? This value produces practically dispersionless waves for \(z\) between 0 and 0.5, i.e., any wavelength larger than \(4\pi h\).
Exercises

By decreasing $\beta$ slightly, the range in $z$ may be increased with an acceptable deviation of $V$. In terms of $h$, suitable values are found for 5 mm to 8 mm at a wave speed of 22 cm/s to 28 cm/s.

3.13. Show that for higher-dimensional waves the group velocity, implied by the dispersion relation $\omega = \omega(\kappa)$, is given by

$$v_g = \frac{\partial \omega}{\partial \kappa}.$$ 

$\frac{\partial}{\partial \kappa}$ denotes the gradient with respect to $\kappa$. 
Chapter 4

Distributions and Fundamental Solutions

This chapter is devoted to rather fundamental concepts. In Section 4.1 we first sketch the idea behind a fundamental solution. For a number of properties and phenomena of PDEs the concept of distribution is needed. Important functions such as Dirac delta functions and Heaviside functions are an instance of this. In Section 4.2 we first consider distributions in one dimension and define what we mean by convergence in a distributional sense. The extension to higher dimensions, which is rather straightforward, is treated in Section 4.3. Distributions play a crucial role in problems that do not possess solutions “in a classical sense,” i.e., that are nonsmooth. This then leads to a notion of solutions in a so-called weak form, which are solutions in distributional sense. They are discussed in Section 4.4. Another use of distributions is describing particular solutions of linear PDEs. These are called fundamental solutions; see Section 4.5. These fundamental solutions are defined on the whole \( \mathbb{R}^d \). A special form of such a solution in the distributional sense is the Green’s function, which moreover satisfies the homogeneous boundary condition. In fact the latter leads to expressions of the solution in terms of the source term of the equation. A more classical approach is to use a Duhamel integral, giving an expression for the solution of a PDE by superposing elementary solutions that represent the source term. These Duhamel integrals are discussed in Section 4.6. In fact, it turns out that there is a natural relationship between these two forms.

4.1 Introduction

Consider the Cauchy problem for the ODE

\[
\frac{du}{dt} = \lambda u + f(t), \quad t > 0, \tag{4.1a}
\]

\[
u(0) = u_0. \tag{4.1b}
\]

If \( f(t) \equiv 0 \), the solution, \( v \) say, is simply given by

\[
v(t) = e^{\lambda t} u_0. \tag{4.2}
\]
In order to find the general solution of (4.1), one can use the *variation of constant* method, i.e., substitute

\[ u(t) = c(t) v(t) \]

and determine \( c(t) \) from (4.1a). The well-known result is

\[ u(t) = e^{\lambda t} u_0 + \int_0^t v(t)(v(\tau))^{-1} f(\tau) \, d\tau. \tag{4.3} \]

Here \( w(t; \tau) \), defined as

\[ w(t; \tau) := v(t)(v(\tau))^{-1}, \quad t \geq \tau, \tag{4.4} \]

is sometimes called a *fundamental solution*. For \( \tau \) fixed it satisfies the homogeneous part of (4.1a). One can view the integral in (4.3) as a superposition of initial value solutions propagating values \( f(\tau) \, d\tau \) until \( t \). One can also describe this as follows. Let

\[ \xi := t - \tau, \tag{4.5a} \]

\[ \overline{w}(\xi) := w(t; \tau). \tag{4.5b} \]

Clearly, \( \overline{w}(\xi) = e^{\lambda \xi} \) for \( \xi \geq 0 \). From (4.3) we conclude that it makes sense to define \( \overline{w}(\xi) = 0 \) for \( \xi < 0 \). But beyond the point \( \xi = 0 \), \( \overline{w}(\xi) \) satisfies the ODE

\[ \frac{d\overline{w}}{d\xi} = \lambda \overline{w}, \quad \xi \neq 0, \tag{4.6} \]

with solution proportional to \( e^{\lambda \xi} \), so at the point \( \xi = 0 \) a jump occurs. This may be described by adding a (for the moment hypothetical) source term \( \delta(\xi) \) to the right-hand side of (4.6) in such a way that

\[ \overline{w}(\xi) = \int_{-\infty}^{\xi} e^{\lambda(s-\tau)} \delta(s) \, ds = e^{\lambda \xi}, \quad \xi > 0. \tag{4.7} \]

This is a nontrivial matter (even more: it is a condition impossible for any normal function to satisfy!), but in loose terms we apparently require \( \delta(\xi) \) to be a spike-like function that integrates to zero along any interval except for an infinitesimally small interval around \( \xi = 0 \) where its integral contribution adds up to one; i.e.,

\[ \int_{-\infty}^{\infty} \delta(\xi) \, d\xi = 1. \tag{4.8} \]

This function \( \delta \) is sometimes called the *(Dirac)* delta function. In the next section we shall give a more rigorous definition. Summarizing, we may view the fundamental solution \( w(t; \tau) \) to satisfy the Cauchy problem

\[ \frac{d}{d\tau} w(t; \tau) = \lambda w(t; \tau) + \delta(t - \tau), \quad t > \tau, \tag{4.9a} \]

\[ w(t; \tau) = 0. \tag{4.9b} \]

This notion can simply be extended to vector-valued ODEs and to boundary value problems as well. We finally remark that for Cauchy problems the value \( \tau = 0 \) needs special consideration; here one just has a solution of the homogeneous problem satisfying \( w(0, 0) = 1 \).
4.2 Distributions in One Variable

In Section 4.1 we saw that we needed a notion of a function of which some properties only made sense after integration. It took mathematicians awhile before they had the proper formulation for the delta function. It shows up as a special instance of a distribution, to be discussed below.

Let $\mathcal{D}$ be some class of functions $\mathbb{R} \to \mathbb{R}$, to be called test functions, and consider mappings, or functionals, from $\mathcal{D}$ to $\mathbb{R}$. A simple but important class of examples is generated by the “inner” product of a test function $\varphi \in \mathcal{D}$ and a given integrable real function $f : \mathbb{R} \to \mathbb{R}$; i.e.,

$$ (f, \varphi) := \int_{-\infty}^{\infty} f(x)\varphi(x) \, dx. \quad (4.10) $$

If the functional can be written in this way, it is identified with the function $f$ and we call it “the functional $f$.” We see that this functional is even linear; i.e.,

$$ (f, \alpha \varphi_1 + \beta \varphi_2) = \alpha (f, \varphi_1) + \beta (f, \varphi_2). \quad (4.11) $$

A suitable test space can be found as follows. First we define the support of a function $\varphi : \mathbb{R} \to \mathbb{R}$ as the closure of the set of all points $x$ such that $\varphi(x) \neq 0$:

$$ \text{supp}(\varphi) = \{ x \in \mathbb{R} | \varphi(x) \neq 0 \}. $$

**Example 4.1** The infinitely many times differentiable real function $\varphi$, defined by

$$ \varphi(x) = \begin{cases} \exp[(x^2 - 1)^{-1}] & \text{for } |x| < 1, \\ 0 & \text{for } |x| \geq 1, \end{cases} $$

clearly has $\text{supp}(\varphi) = [-1, 1]$. \qed

If $\text{supp}(\varphi)$ is a bounded set, then $\varphi$ is said to have a compact support. Now define the test function space

$$ \mathcal{D} = C^\infty_0(\mathbb{R}) := \{ \varphi \in C^\infty(\mathbb{R}) | \varphi \text{ has compact support} \}. \quad (4.12) $$

This space of test functions $\mathcal{D}$ will be used throughout unless indicated otherwise. In order to have a practically meaningful linear functional, we would like it to be continuous. For this we need a convergence concept. In view of the compact support property the following makes sense.

**Definition 4.2.** A sequence of test functions $\{\varphi_i\}_{i \geq 0}$, where $\varphi_i \in C^\infty_0(\mathbb{R})$, is called convergent to zero if

(i) there is a closed and finite interval $I$ such that $\text{supp}(\varphi_i) \subseteq I$;

(ii) $\lim_{i \to \infty} \frac{d^k}{dx^k} \varphi_i(x) = 0$ uniformly on $\mathbb{R}$ for $k = 0, 1, 2, \ldots$.

From now on we shall identify $\mathcal{D}$ with $C^\infty_0(\mathbb{R})$, equipped with the convergence notion as defined in Definition 4.2.
**Definition 4.3.** A linear functional $f$ is continuous if for any sequence $\{\phi_i\}_{i \geq 0} \subset D$ that is convergent to zero, also $(f, \phi_i) \to 0$.

(Because of linearity it suffices to consider continuity at zero.) The space of continuous linear functionals on $D$ is the dual space $D'$. A continuous linear functional is called a *distribution*. Actually, we see that $D'$ is a linear space by writing, for $f, g \in D'$, that

$$
(\alpha f + \beta g, \phi) = \alpha (f, \phi) + \beta (g, \phi) \quad \text{for all} \quad \phi \in D. \quad (4.13)
$$

If $f$ is locally integrable (i.e., the integral exists on any finite interval), then by (4.10) $f$ generates a distribution. Such a distribution is called *regular* and may be identified with $f$.

The delta function encountered in Section 4.1 can now be defined by

$$
(\delta, \phi) = \phi(0) \quad \text{for all} \quad \phi \in D. \quad (4.14)
$$

This is not a regular distribution (see, e.g., [147]). However, we will write this, as tradition does, like (4.10); i.e.,

$$
(\delta, \phi) = \int_{-\infty}^{\infty} \delta(x) \phi(x) \, dx. \quad (4.15)
$$

Take due note that this is just symbolism. The delta function $\delta(x)$ should be interpreted via its definition (4.14). It is not a function in the classical sense.

If $f, g \in D'$ are regular distributions, it follows from (4.10) that if $(f, \phi) = (g, \phi)$ for all $\phi \in D$, then $f(x) = g(x)$ almost everywhere. In general we call two distributions $f$ and $g$ *identical* if

$$
(f, \phi) = (g, \phi) \quad \text{for all} \quad \phi \in D. \quad (4.16)
$$

If we shift the argument in a distribution $f$ by $\xi$, i.e.,

$$
f_{\xi}(x) := f(x - \xi), \quad (4.17)
$$

we find

$$
(f_{\xi}, \phi) = (f, \phi_{-\xi}). \quad (4.18)
$$

For a shifted delta function $\delta(x - \xi)$ we then define $\delta_{\xi}(x) = \delta(x - \xi)$ by

$$
(\delta_{\xi}, \phi) = (\delta, \phi_{-\xi}) = \phi(\xi) \quad \text{for all} \quad \phi \in D. \quad (4.19)
$$

We can also define a multiplication of $f \in D'$ by an integrable function $g$ through the relation

$$
(gf, \phi) := (f, g\phi). \quad (4.20)
$$

Note that $gf$ is again a distribution. It follows that the product of a distribution and a regular distribution is defined. This is not necessarily the case for the product of two nonregular distributions.

**Example 4.4** If we multiply the delta function by $g(x) = x$, we obtain

$$
(g\delta, \phi) = (\delta, g\phi).
$$

Hence the test functions $g\phi$ are all zero at $x = 0$. Consequently, $(g\delta, \phi) = 0$, or

$$
x\delta(x) \equiv 0.
$$
4.2. Distributions in One Variable

Using integration by parts, we find, for an ordinary differentiable function \( f \), that

\[
(f', \varphi) = \left[ f(x)\varphi(x) \right]_{-\infty}^{\infty} - (f, \varphi') = -(f, \varphi'),
\]

(4.21) since \( \varphi \) has compact support. Therefore we define the derivative of a distribution \( f \) in general by

\[
(f', \varphi) := -(f, \varphi') \quad \text{for all } \varphi \in \mathcal{D}.
\]

(4.22)

Clearly, \( f' \) is a distribution again. Note that \( f \) is, in this sense, arbitrarily many times differentiable.

An important application is the following. Let \( H(x) \) be defined by

\[
H(x) = \begin{cases} 
0 & \text{for } x < 0, \\
\frac{1}{2} & \text{for } x = 0, \\
1 & \text{for } x > 0,
\end{cases}
\]

(4.23)

which is the Heaviside function. It generates the Heaviside distribution if we write, for any \( \varphi \in \mathcal{D} \),

\[
(H, \varphi) = \int_{-\infty}^{\infty} H(x)\varphi(x) \, dx = \int_{0}^{\infty} \varphi(x) \, dx.
\]

(4.24a)

From (4.22) we obtain

\[
(H', \varphi) = -(H, \varphi') = -\int_{0}^{\infty} \varphi'(x) \, dx = \varphi(0).
\]

(4.24b)

Hence we may identify the distribution \( H'(x) \) with \( \delta(x) \).

As it will turn out in later chapters, it is useful to have representations of the delta function as a convergent sequence of regular distributions. First, we introduce the following definition.

**Definition 4.5.** Let \( \{f_i\}_{i \geq 0} \) be a sequence of distributions. Then this sequence converges to \( f \), denoted by \( f_i \to f \in \mathcal{D} \), if

\[
\lim_{i \to \infty} (f_i, \varphi) = (f, \varphi) \quad \text{for all } \varphi \in \mathcal{D}.
\]

(4.25a)

Let \( I \) be some interval in \( \mathbb{R} \). If \( \{f_\lambda\}_{\lambda \in I} \) is a family of regular distributions, continuously parametrized by a parameter \( \lambda \), then we say \( f_\lambda \to f \) as \( \lambda \to \lambda_0 \in I \) if

\[
\lim_{\lambda \to \lambda_0} (f_\lambda, \varphi) = (f, \varphi) \quad \text{for all } \varphi \in \mathcal{D}.
\]

(4.25b)

With these definitions we have the following theorem.

**Theorem 4.6.** Let \( f_\lambda \) be a locally integrable function with

(i) \( f_\lambda(x) \geq 0, x \in \mathbb{R}, \lambda > 0, \)
(ii) \( \int_{-\infty}^{\infty} f_\lambda(x) \, dx = 1, \lambda > 0, \)
(iii) \( \lim_{\lambda \to 0} \int_{a}^{b} f_\lambda(x) \, dx = 1 \) for any \( a < 0 < b. \)
Then \( \{ f_\lambda \}_{\lambda \in (0, \infty)} \) is a family of regular distributions with
\[
f_\lambda(x) \to \delta(x) \quad \text{for} \quad \lambda \to 0.
\] (4.26)

**Proof.** Consider for any \( \varphi \in \mathcal{D} \) the difference \( I(\lambda) := (f_\lambda, \varphi) - \varphi(0) \). From the properties of \( f_\lambda \) it follows that
\[
I(\lambda) = \int_{-\infty}^{\infty} f_\lambda(x)(\varphi(x) - \varphi(0)) \, dx.
\]
We split the integration interval into three parts: \((-\infty, -a), (-a, a), \) and \((a, \infty)\) for some \( a > 0 \). Apparently \( \varphi \) is bounded on \( \mathbb{R} \), say, \(|\varphi(x)| < M\) for \( x \in \mathbb{R} \). Hence \(|\varphi(x) - \varphi(0)| \leq 2M\). Furthermore, \( \varphi \) is continuous, so for a given value of \( \varepsilon \), choose \( a \) such that \(|\varphi(x) - \varphi(0)| \leq \frac{1}{2} \varepsilon \), \( x \in (-a, a) \). Finally, let \( \lambda_0 \) be such that \( \int_{-a}^{a} f_\lambda(x) \, dx > 1 - (4M)^{-1} \varepsilon \) for \( 0 < \lambda < \lambda_0 \). Then
\[
|I(\lambda)| \leq |\int_{-\infty}^{-a} \varphi \, dx + \int_{a}^{\infty} \varphi \, dx| + |\int_{-a}^{a} \varphi \, dx| \leq 2M(4M)^{-1} \varepsilon + \frac{1}{2} \varepsilon = \varepsilon.
\]
This shows (4.26).

A sequence \( f_\lambda \) as introduced in Theorem 4.6 is sometimes referred to as a *delta sequence*. Such a delta sequence has both theoretical and practical significance; the latter in a situation where one needs, e.g., numerical approximations.

**Example 4.7**

(i) The top hat sequence (see Figure 4.1)
\[
f_\lambda(x) := \begin{cases} \frac{1}{2} \lambda^{-1} & \text{for} \ x \in (-\lambda, \lambda), \\ 0 & \text{for} \ |x| > \lambda \end{cases}
\]
satisfies all requirements of Theorem 4.6.

![Figure 4.1. A top hat delta sequence.](image)

(ii) Another well-known delta sequence is given by (see Figure 4.2)
\[
f_\lambda(x) := \frac{1}{\sqrt{2\pi} \lambda} \exp \left( -\frac{x^2}{2\lambda} \right), \quad x \in \mathbb{R}, \ \lambda > 0,
\]
which is the probability density function for a normal distribution (in the ordinary statistical sense) with variance \( \lambda \). The requirements from Theorem 4.6 can be verified by introducing the transformation \( y^2 = x^2/2\lambda \) and the error function
\[
\text{erf}(z) := \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2} \, dt
\]
(cf. [1]), which has the property \( \text{erf}(\infty) = 1 \). This sequence has been important in probability theory, but also in parabolic problems; see Chapter 10. Finally, note that \( f_\lambda \in C^\infty(\mathbb{R}) \).
4.3 Distributions in Several Variables

For a multivariate real function mapping \( \mathbb{R}^n \rightarrow \mathbb{R} \), depending on, say, \( x = (x_1, x_2, \ldots, x_n)^T \), we can give straightforward generalisations of the foregoing concepts. The test space \( \mathcal{D} \) now consists of functions in \( C^\infty(\mathbb{R}^n) \) defined by

\[
C^\infty_0(\mathbb{R}^n) := \{ \varphi \in C^\infty(\mathbb{R}^n) \mid \text{\varphi has compact support} \},
\]  

equipped with the following notion of convergence.

**Definition 4.8.** A sequence \( \{ \varphi_i \}_{i \geq 0} \), with \( \varphi_i \in C^\infty_0(\mathbb{R}^n) \), is called convergent to zero if

(i) there is a closed and bounded set \( \Omega \subset \mathbb{R}^n \) such that for all \( i \) \( \text{supp}(\varphi_i) \subset \Omega \),
(ii) \( \lim_{i \to \infty} \varphi_i(x) = 0 \) uniformly in \( \mathbb{R}^n \), and likewise all partial derivatives.

A distribution is a continuous linear functional on \( \mathcal{D} \). If \( f(x) \) is locally integrable, then we can identify \( f \) with its distribution, where

\[
(f, \varphi) := \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x) \varphi(x) \, dx_1 \cdots dx_n. \tag{4.28}
\]

In particular, we can find the delta function \( \delta(x) \) defined by

\[
(\delta, \varphi) = \varphi(0) \quad \text{for all} \quad \varphi \in \mathcal{D}. \tag{4.29}
\]

Differentiation is defined by the relation

\[
\left( \frac{\partial f}{\partial x_i} \right, \varphi) = -\left( f, \frac{\partial \varphi}{\partial x_i} \right) \quad \text{for all} \quad \varphi \in \mathcal{D}. \tag{4.30}
\]

Note that this “derivative” is again a distribution. Theorem 4.6 can be generalised straightforwardly.
Example 4.9 We can give simple generalisations of the top hat and normal distribution representations of the multivariate delta function as follows:

\[(i)\quad f_λ(x) = \begin{cases} 
(2λ)^{-n} & \text{for } \|x\|_∞ \leq λ, \\
0 & \text{for } \|x\|_∞ ≥ λ, 
\end{cases} \quad (4.31)\]

\[(ii)\quad f_λ(x) = \left(\frac{1}{\sqrt{2πλ}}\right)^n \exp\left(-\frac{\|x\|^2_2}{2λ}\right), \quad (4.32)\]

where \(x \in \mathbb{R}^n\), \(\|x\|_∞ = \max_{1 ≤ i ≤ n}(|x_i|)\), and \(\|x\|^2_2 = x_1^2 + \cdots + x_n^2\).

If we have two distributions of a single variable, we can form a direct product to obtain a multivariate distribution. Let \(f_1\) and \(f_2\) be distributions. Then

\[(f_1(x_1) f_2(x_2), ϕ(x_1, x_2)) := (f_1(x_1), (f_2(x_2), ϕ(x_1, x_2))), \quad (4.33)\]

with the inner products defined in an obvious way. We have retained the arguments in \(f_1\), \(f_2\), and \(ϕ\) for clarity.

Example 4.10

\[δ(x) = δ(x_1)δ(x_2) \cdots δ(x_n). \quad \square\]

## 4.4 Strong and Weak Solutions

In order to facilitate the following discussion, we shall denote a general linear second order PDE in \(d\) spatial variables, \(x_1, \ldots, x_d\) say, as

\[L[u] = f, \quad (4.34a)\]

where

\[L[u] := \sum_{i,j=1}^d a_{ij} \frac{∂^2 u}{∂x_i ∂x_j} + \sum_{i=1}^d b_i \frac{∂u}{∂x_i} + cu. \quad (4.34b)\]

Let \(x := (x_1, x_2, \ldots, x_d)^T \in Ω \subseteq \mathbb{R}^d\). If the operator \(L\) is really second order, i.e., \(a_{ij} ≠ 0\) for at least one pair \((i, j)\), then we call \(u\) a strong solution of (4.34) if \(u ∈ C^2(Ω)\). If \(L\) is first order, i.e., all \(a_{ij} = 0\), then \(u\) is a strong solution if \(u ∈ C^1(Ω)\).

Assuming the coefficients in (4.34b) to be \(C^∞\), we may look for solutions in a distributional sense. We call \(u\) a generalised or weak solution if (4.34) is satisfied in a distributional sense, i.e., if

\[(L[u], ϕ) = (f, ϕ) \quad \text{for all } \ ϕ ∈ \mathcal{D}. \quad (4.35)\]

In other words, \(L[u] − f\) is identical to the zero distribution. We may restrict \(\mathcal{D}\) to test functions with support within \(Ω\). Then \(u\) is a weak solution in \(Ω\). A strong solution is also a weak solution, but not the other way around.

We next introduce the concept of an adjoint operator, defined in general by

\[(L[u], v) := \langle u, L^*[v] \rangle. \quad (4.36)\]
4.4. Strong and Weak Solutions

If \( v \) is a test function, then using (4.30) we find in particular that

\[
\mathcal{L}^*[v] := \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j} (a_{ij} v) - \sum_{i=1}^{d} \frac{\partial}{\partial x_i} (b_i v) + cv.
\] (4.37)

For the operator \( \mathcal{L} \) and its adjoint \( \mathcal{L}^* \) the following property holds.

**Property 4.11.** For \( u, v \in C^2(\mathbb{R}^d) \) we have

\[
v \mathcal{L}[u] - u \mathcal{L}^*[v] = \nabla \cdot w,
\] (4.38a)

where the vector function \( w := (w_1, \ldots, w_d)^T \) is given by

\[
w_j = \sum_{j=1}^{d} \left( a_{ij} \frac{\partial u}{\partial x_j} v - u \frac{\partial}{\partial x_j} (a_{ij} v) \right) + b_j uv, \quad i = 1, \ldots, d.
\] (4.38b)

**Proof.** The proof is trivial by direct substitution of (4.38b) into (4.38a). \( \square \)

Next we have the following result.

**Corollary 4.12.** If \( u \) is a distribution and \( \varphi \in \mathcal{D} \), then

\[
\int_{\Omega} (\varphi \mathcal{L}[u] - u \mathcal{L}^*[\varphi]) \, dV = \oint_{\partial \Omega} w \cdot n \, dS,
\] (4.39)

where \( n \) is the unit outward normal of the boundary \( \partial \Omega \) of \( \Omega \).

**Proof.** Apply the divergence theorem (J.13) Section J in the appendix to (4.38a). \( \square \)

**Example 4.13** Consider the Cauchy problem

\[
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \\
u(x, 0) = H(x), \quad x \in \mathbb{R}.
\]

We see that the PDE has characteristics \( x + t = c, c \in \mathbb{R} \). Hence we have a discontinuity along the line \( x + t = 0 \). So the solution appears to be

\[
u(x, t) = H(x + t),
\]

which is clearly not a classical solution. Let \( \varphi \) be a test function and consider (cf. (4.30))

\[
I := -\int_{0}^{\infty} \int_{-\infty}^{\infty} \left( \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} \right) \varphi \, dx \, dt = \int_{0}^{\infty} \int_{-\infty}^{\infty} u \left( \frac{\partial \varphi}{\partial t} - \frac{\partial \varphi}{\partial x} \right) \, dx \, dt.
\]
We split the domain into the regions $R_1 = \{ x + t > 0 \}$, where $u = 1$, and $R_2 = \{ x + t < 0 \}$, where $u = 0$. Then

$$I = \int_{R_1} u \left( \frac{\partial \varphi}{\partial t} - \frac{\partial \varphi}{\partial x} \right) \, dx \, dt = \int_{0}^{\infty} \int_{-\infty}^{\infty} \left( \frac{\partial \varphi}{\partial t} - \frac{\partial \varphi}{\partial x} \right) \, dx \, dt$$

$$= \int_{-\infty}^{0} \int_{-\infty}^{\infty} \frac{\partial \varphi}{\partial t} \, dx \, dt + \int_{0}^{\infty} \int_{-\infty}^{0} \frac{\partial \varphi}{\partial t} \, dx \, dt = \int_{-\infty}^{0} -\varphi(x, -x) \, dx + \int_{0}^{\infty} +\varphi(-x, 0) \, dx + \int_{-\infty}^{\infty} \varphi(-t, t) \, dt = 0$$

because $\varphi$ vanishes along the borders of its domain, i.e. $\varphi(x, 0) = 0$. Hence $I = 0$ and so $u$ is a weak solution.

The example above shows that the weak solution concept nicely captures “shock structures.” We shall encounter this in Chapter 12, where the “jump” condition in $u$ will be met again as a Rankine–Hugoniot condition.

We finally show that a weak solution with sufficient smoothness is always a strong solution.

**Theorem 4.14.** Let $u \in C^2(\Omega)$ be a weak solution of (4.34a) (or at least $C^1(\Omega)$ if $a_{ij} \equiv 0$). Then $u$ is a strong solution on $\Omega$.

**Proof.** From (4.35) and (4.39) we find

$$\int_{\Omega} \mathcal{L}[u] \varphi \, dV = \int_{\Omega} u \mathcal{L}^\ast[\varphi] \, dV = \int_{\Omega} f \varphi \, dV,$$

since $w \equiv 0$ on $\partial \Omega$. Hence

$$\int_{\Omega} (\mathcal{L}[u] - f) \varphi \, dV = 0.$$

Since $\mathcal{L}[u] - f \in C^0(\Omega)$ we conclude that $\mathcal{L}[u] = f$.

In the rest of this book we will often encounter systems of PDEs of the form

$$\frac{\partial \mathbf{u}}{\partial t} = \mathcal{L}[\mathbf{u}], \quad x \in \mathbb{R}, \quad t > 0,$$

(4.40)

where $\mathcal{L}$ is a vector-valued differential operator containing only spatial derivatives that works on the components of the vector function $\mathbf{u} \in \mathbb{R}^m$. Analogously to (4.35), we can define a weak solution of (4.40) by the requirement

$$\int_{0}^{\infty} \int_{-\infty}^{\infty} \left( \frac{\partial \mathbf{u}}{\partial t} - \mathcal{L}[\mathbf{u}] \right) \cdot \varphi \, dx \, dt = 0 \quad \text{for all} \quad \varphi \in \mathcal{D}^m.$$

(4.41)

In this definition $\varphi$ is a vector-valued test function with components in $\mathcal{D}$. We will apply this definition in Chapter 12 to systems of hyperbolic equations.
4.5 Fundamental Solutions

We now have the tools to generalise the findings of Section 4.1. The “solution” $w$ found there will now be recognized as a weak solution.

Consider (4.34a). For a general linear differential operator $L$ and for any $\xi$, we call $w(x; \xi)$ a fundamental solution if

$$L[w](x; \xi) = \delta(x - \xi), \quad x \in \mathbb{R}^d. \quad (4.42)$$

**Example 4.15** A fundamental solution of the Laplace equation in $\mathbb{R}^2$ satisfying

$$\nabla^2 w = \delta(x - \xi)$$

is found to be given by

$$w(x; \xi) := \frac{1}{2\pi} \ln(||x - \xi||).$$

This is not evident because, when considered as an ordinary function, $\nabla^2 w$ is identically zero, except at $x = \xi$, where it does not exist. As a result, any integral $\iint_{\mathbb{R}^2} \psi \nabla^2 w \, dx$ is zero rather than equal to $\psi(\xi)$. However, when we interpret $\nabla^2 w$ as a distribution, the gradient is defined by (4.30), and we have

$$\iint_{\mathbb{R}^2} \psi \nabla^2 w \, dx = -\iint_{\mathbb{R}^2} \nabla \psi \cdot \nabla w \, dx = -\int_0^{2\pi} \int_0^{\infty} \frac{\partial \psi}{\partial r} \frac{\partial w}{\partial r} \, r \, dr \, d\theta = \frac{1}{2\pi} \int_0^{2\pi} \psi(\xi) \, d\theta$$

with $x - \xi$ written in polar coordinates $r$ and $\theta$.

Now consider a general linear inhomogeneous problem with linear boundary value operator $B$, where we would like the solution $u$ on a domain $\Omega$; i.e.,

$$L[u] = f, \quad x \in \Omega, \quad (4.43a)$$

$$B[u] = \beta, \quad x \in \partial \Omega. \quad (4.43b)$$

Then we can construct $u$ as follows. First, we seek a particular solution, $u_p(x)$ say, such that

$$L[u_p] = f, \quad x \in \Omega, \quad (4.44a)$$

$$B[u_p] = 0, \quad x \in \partial \Omega. \quad (4.44b)$$

In order to find $u_p$, we employ special weak solutions derived from a fundamental solution by adding a suitable homogeneous (strong) solution of (4.44) (i.e., with $f = 0$), giving rise to Green’s functions $G(x; \xi)$:

$$L[G](x; \xi) = \delta(x - \xi), \quad x \in \Omega, \quad (4.45a)$$

$$B[G](x; \xi) = 0, \quad x \in \partial \Omega. \quad (4.45b)$$
(Note the difference between a fundamental solution and the Green’s function.) Since we can write \( f(x) \) formally as a superposition of delta functions in the form
\[
f(x) = \int_{\Omega} \delta(x - \xi) f(\xi) \, dV_\xi,
\]
where the integration is with respect to the \( \xi \) coordinates, we can construct \( u_p \) by convolution as
\[
u_p(x) = \int_{\Omega} G(x; \xi) f(\xi) \, d\xi.
\]
(4.47)
The complete solution of (4.43) is now found by adding a (homogeneous) solution \( u_h \) of
\[
L[u_h] = 0, \quad x \in \Omega, \quad (4.48a)
\]
\[
B[u_h] = \beta, \quad x \in \partial \Omega. \quad (4.48b)
\]
Such adding of a suitable homogeneous solution to a particular solution (e.g., in order to make the boundary data fit) is called superposition.

Note that \( u_h \), if it exists, need not be unique. This then gives rise to Fredholm’s alternative: either \( u_h \) is unique on \( \Omega \) or there exist infinitely many solutions for (4.48). The general solution of (4.43) is given by
\[
u(x) = \int_{\Omega} G(x; \xi) f(\xi) \, dV_\xi + u_h(x),
\]
(4.49)
which is an analogue of (4.3).

We remark that the fundamental solution \( w(x; \xi) \) can be found by translation of the argument from \( w(x; 0) \) if \( L \) has constant coefficients; i.e.,
\[
w(x; \xi) = w(x - \xi; 0). \quad (4.50)
\]

We can generalise this concept of the fundamental solution by taking \( t \) as an additional independent variable. So consider
\[
L_1[u] + L_2[u] = f,
\]
(4.51)
where \( L_1 \) denotes a differential operator with respect to \( t \) and \( L_2 \) is a differential operator with respect to \( x \). For each \((\xi, \tau)\) a fundamental solution \( w(x, t; \xi, \tau) \) then satisfies the problem
\[
L_1[w(x, t; \xi, \tau)] + L_2[w](x, t; \xi, \tau) = \delta(x - \xi) \delta(t - \tau), \quad x \in \mathbb{R}^d, \quad t > 0,
\]
(4.52a)
\[
w(x, t; \xi, \tau) = 0, \quad x \in \mathbb{R}^d, \quad t < \tau.
\]
(4.52b)
Note that this formulation takes care of the evolutionary character of the problem (causality principle) with time intervals \((\tau, \infty)\). We shall see later (Chapter 10) that solutions of these problems can be found using such fundamental solutions. An alternative way is to make use of Duhamel integrals; see the next section.
4.6 Initial (Boundary) Value Problems; Duhamel Integrals

Most problems we deal with in this book have both a space and a time dependence. In the linear case the equations are in fact either one of the following type:

\[
\frac{\partial u}{\partial t} = L_1[u] + f(x, t), \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0,
\]
(4.53)

\[
\frac{\partial^2 u}{\partial t^2} = L_2[u] + f(x, t), \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0.
\]
(4.54)

Here $L_1$ and $L_2$ are homogeneous second order differential operators in $x$. In the chapters that will follow we shall investigate the solutions of the various types of PDEs. Generally speaking, we need to specify a condition on the boundary $\partial \Omega$ of the spatial domain $\Omega$, say

\[
B[u] = \beta(t), \quad x \in \partial \Omega, \quad t > 0.
\]
(4.55)

Moreover, we need initial conditions at $t = 0$. For (4.53) they have the form

\[
u(x, 0) = a(x), \quad x \in \Omega.
\]
(4.56)

For (4.54) we have, besides (4.56),

\[
\frac{\partial u}{\partial t}(x, 0) = b(x), \quad x \in \Omega.
\]
(4.57)

Now consider the following family of problems defined on $\Omega \times (\tau, \infty)$, associated with (4.53), (4.55), and (4.56):

\[
\frac{\partial w}{\partial t}(x, t; \tau) = L_1[w](x, t; \tau), \quad x \in \Omega, \quad t > \tau,
\]
(4.58a)

\[
B[w](x, t; \tau) = 0, \quad x \in \partial \Omega, \quad t > \tau,
\]
(4.58b)

\[
w(x, \tau; \tau) = f(x, \tau), \quad x \in \Omega.
\]
(4.58c)

These initial boundary value problems together build the actual solution; see the following theorem.

**Theorem 4.16 (Duhamel integral I).** Assume that the initial boundary value problem (4.53), (4.55) with $\beta(t) \equiv 0$, and (4.56) with $a(x) \equiv 0$ has a unique solution $u$. Then it is given by the Duhamel integral

\[
u(x, t) = \int_0^t w(x, t; \tau) \, d\tau,
\]
(4.59)

where $w(\cdot, \cdot; \tau)$ is a solution of (4.58).
Proof. If we differentiate the right-hand side of (4.59) with respect to $t$, we get
\[
\frac{\partial}{\partial t} \int_0^t w(x, t; \tau) \, d\tau = w(x, t; t) + \int_0^t \frac{\partial w}{\partial t}(x, t; \tau) \, d\tau \\
= f(x, t) + \int_0^t L_1[w](x, t; \tau) \, d\tau = f(x, t) + \left[ \int_0^t w(x, t; \tau) \, d\tau \right].
\]
Hence $\int_0^t w(x, t; \tau) \, d\tau$ satisfies (4.53). Moreover, it is easy to see that (4.55), with $\beta(t) \equiv 0$, and (4.56), with $a(x) \equiv 0$, are satisfied, so that $\int_0^t w(x, t; \tau) \, d\tau$ may be identified with the solution $u(x, t)$.

Example 4.17 Consider the problem
\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x), \quad x \in \mathbb{R}, \quad t > 0,
\]
\[
u(x, 0) = 0, \quad x \in \mathbb{R}.
\]
It can be checked that the solution of this initial value problem is given by (4.59), with
\[
w(x, t; \tau) = \frac{1}{2\sqrt{\pi(t - \tau)}} \int_{-\infty}^{\infty} \exp \left( -\frac{(x - \xi)^2}{4(t - \tau)} \right) f(\xi) \, d\xi.
\]
Similarly, consider a problem associated with (4.54), and (4.55) to (4.57):
\[
\frac{\partial^2 w}{\partial t^2}(x, t; \tau) = L_2[w](x, t; \tau), \quad x \in \Omega, \quad t > \tau, \quad (4.60a)
\]
\[
B[w](x, t; \tau) = 0, \quad x \in \partial\Omega, \quad t > \tau, \quad (4.60b)
\]
\[
w(x, \tau; \tau) = 0, \quad x \in \Omega, \quad (4.60c)
\]
\[
\frac{\partial w}{\partial t}(x, \tau; \tau) = f(x, \tau), \quad x \in \Omega. \quad (4.60d)
\]
Then we find the following result.

Theorem 4.18 (Duhamel integral II). Assume that the initial boundary value problem (4.54), (4.55) with $\beta(t) \equiv 0$, (4.56) with $a(x) \equiv 0$, and (4.57) with $b(x) \equiv 0$ has a unique solution. Then $u$ is given by
\[
u(x, t) = \int_0^t w(x, t; \tau) \, d\tau, \quad (4.61)
\]
where $w(\cdot, \cdot; \tau)$ is a solution of (4.60).

Proof. If $w$ is a solution of (4.60), we find from differentiating the right-hand side of (4.61) once that
\[
\frac{\partial}{\partial t} \int_0^t w(x, t; \tau) \, d\tau = \int_0^t \frac{\partial w}{\partial t}(x, t; \tau) \, d\tau.
\]
4.6. Initial (Boundary) Value Problems; Duhamel Integrals

Differentiating this relation once more then gives

\[ \frac{\partial^2}{\partial t^2} \int_0^t w(x, t; \tau) \, d\tau = \frac{\partial^2 w(x, t; t)}{\partial t^2} + \int_0^t \frac{\partial^2 w(x, t; \tau)}{\partial t^2} \, d\tau \]

which shows that (4.61) is the solution of (4.54) to (4.57) with homogeneous conditions.

If \( a \neq 0 \) in (4.56), we find, instead of (4.61), that

\[ u(x, t) = w(x, t; 0) a(x) + \int_0^t w(x, t; \tau) \, d\tau. \tag{4.62a} \]

Moreover, if \( b \neq 0 \) in (4.57), we obtain

\[ u(x, t) = v(x, t) a(x) + w(x, t; 0) b(x) + \int_0^t w(x, t; \tau) \, d\tau, \tag{4.62b} \]

where \( w \) is a solution of (4.60) and \( v \) is a solution of

\begin{align}
\frac{\partial^2 v}{\partial t^2}(x, t) &= \mathcal{L}[v](x, t), \quad x \in \Omega, \quad t > 0, \\
\mathcal{B}[v](x, t) &= 0, \quad x \in \partial \Omega, \quad t > 0, \\
v(x, 0) &= 1, \quad x \in \Omega, \\
\frac{\partial v}{\partial t}(x, 0) &= 0, \quad x \in \Omega. \tag{4.63a-d}
\end{align}

The Duhamel integral is in fact nothing but the superposition of fundamental solutions (with respect to time), and we have the following theorem.

**Theorem 4.19.** Let \( \mathcal{D}' \) be the set of distributions in the parameter \( t \).

(i) Then \( w \) satisfying (4.58), with \( f(x, t) \equiv 1 \), is a fundamental solution of

\[ \frac{\partial w}{\partial t}(x, t; \tau) = \mathcal{L}_1[w](x, t; \tau) + \delta(\tau). \tag{4.64a} \]

(ii) Also, \( w \) satisfying (4.60), with \( f(x, t) \equiv 1 \), is a fundamental solution of

\[ \frac{\partial^2 w}{\partial t^2}(x, t; \tau) = \mathcal{L}_2[w](x, t; \tau) + \delta(\tau). \tag{4.64b} \]

Here we have defined \( w(x, t) \equiv 0 \) for \( t < \tau \).
Chapter 4. Distributions and Fundamental Solutions

Proof.

(i) From (4.58) we obtain
\[
\int_0^\infty \left( \frac{\partial w}{\partial t} - L_1[w] \right) \varphi \, dt = -\int_0^\infty \left( \frac{\partial \varphi}{\partial t} w + \varphi L_1[w] \right) \, dt
\]
\[
= -\int_\tau^\infty \left( \frac{\partial \varphi}{\partial t} w + \varphi L_1[w] \right) \, dt = \varphi(\tau) + \int_\tau^\infty \left( \frac{\partial w}{\partial t} - L_1[w] \right) \varphi \, dt = \varphi(\tau).
\]

(ii) Likewise, from (4.60) we deduce
\[
\int_0^\infty \left( \frac{\partial w}{\partial t} - L_2[w] \right) \varphi \, dt = \int_0^\infty \left( \frac{\partial^2 \varphi}{\partial t^2} w - \varphi L_2[w] \right) \, dt
\]
\[
= \int_\tau^\infty \left( \frac{\partial^2 \varphi}{\partial t^2} w - \varphi L_2[w] \right) \, dt = \varphi(\tau) + \int_\tau^\infty \left( \frac{\partial^2 w}{\partial t^2} - L_2[w] \right) \varphi \, dt = \varphi(\tau). \quad \Box
\]

Example 4.20 Consider the same problem as in Example 4.17. As one can easily see, a fundamental solution, i.e., a solution of
\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \delta(x-\xi)\delta(t-\tau), \quad x \in \mathbb{R}, \quad t > 0,
\]
\[
u(x,t) = 0, \quad x \in \mathbb{R}, \quad t < \tau,
\]
is given by
\[
w(x,t;\xi,\tau) := \frac{1}{\sqrt{2\pi(t-\tau)}} \exp\left(-\frac{(x-\xi)^2}{4(t-\tau)}\right).
\]
Hence \(u(x,t;\tau) = \int_{-\infty}^{\infty} w(x,t;\xi,\tau) \, d\xi\). \quad \Box

As a final application of the Duhamel integral, consider initial boundary value problems on a one-dimensional semi-infinite domain, i.e., \(\Omega = [0, \infty)\), with homogeneous source term and with homogeneous initial condition(s) (i.e., \(a(x) \equiv 0\) in (4.56) for equation (4.53) and \(a(x) = b(x) \equiv 0\) in (4.56) and (4.57) for equation (4.54)). Let the boundary condition (4.55) be given by
\[
u(0,t) = \beta(t), \quad t > 0.
\]
(4.65)

Assume that the function \(w(x,t;\tau)\) can actually be written as
\[
\overline{w}(x,t - \tau) := w(x,t;\tau).
\]
(4.66)
4.7. Discussion

Let \( \overline{w} \) satisfy the following initial boundary value problem corresponding to either (4.53) or (4.54), i.e.,

\[
\frac{\partial^i \overline{w}}{\partial t^i} = \mathcal{L}_i [\overline{w}], \quad x \in [0, \infty), \quad t > \tau \quad (i = 1, 2),
\]

\[
\overline{w}(0, t) = 1, \quad t > 0,
\]

\[
\overline{w}(x, 0) = 0, \quad x \in [0, \infty) \text{ for (4.53)}.
\]

\[
\frac{\partial \overline{w}}{\partial t}(x, 0) = 0, \quad x \in [0, \infty) \text{ for (4.54)}
\]

Then we have the following representation.

\textbf{Property 4.21.} The solution of (4.53) or (4.54) with \( f(x, t) \equiv 0 \), (4.65), (4.56) with \( a(x) = 0 \), and (for eq. (4.54)) (4.57) with \( b(x) \equiv 0 \), is given by

\[
u(x, t) = \frac{\partial}{\partial t} \int_0^t \overline{w}(x, t - \tau) \beta(\tau) \, d\tau.
\]

\textbf{Proof.} Denote the right-hand side of (4.68) by \( q(x, t) \). Then

\[
q(x, t) = \overline{w}(x, t) \beta(0) + \int_0^t \overline{w}(x, t - \tau) \beta'(\tau) \, d\tau,
\]

and, using all properties of \( \overline{w} \), we find

\[
\frac{\partial q}{\partial t}(x, t) = \frac{\partial}{\partial t} \overline{w}(x, t) \beta(0) + \int_0^t \frac{\partial}{\partial t} \overline{w}(x, t - \tau) \beta'(\tau) \, d\tau,
\]

\[
\frac{\partial^2 q}{\partial t^2}(x, t) = \frac{\partial^2}{\partial t^2} \overline{w}(x, t) \beta(0) + \int_0^t \frac{\partial^2}{\partial t^2} \overline{w}(x, t - \tau) \beta'(\tau) \, d\tau,
\]

\[
\mathcal{L}_i[q](x, t) = \mathcal{L}_i[\overline{w}](x, t) \beta(0) + \int_0^t \mathcal{L}_i[\overline{w}](x, t - \tau) \beta'(\tau) \, d\tau \quad (i = 1, 2).
\]

This shows that \( q \) satisfies the PDE (4.53) or (4.54). Furthermore, we have

\[
q(0, t) = \beta(0) + \int_0^t \beta'(\tau) \, d\tau = \beta(t),
\]

so that it also satisfies the boundary condition.

\[\square\]

\section*{4.7 Discussion}

- Distributions are a fundamental concept and a bit of an outsider in this book as far as rigour is concerned; cf. [119]. From a mathematical point of view they open entire new vistas in which concepts like differentiability can be treated for larger classes of
problems, like in the theory pseudodifferential operators; see, e.g., [154]. They will turn out to be essential, though, when it comes to hyperbolic equations. Here one cannot solve actual problems without taking recourse to weak solutions.

- Fundamental solutions have a value in their own right, as it will turn out, e.g., in Chapters 8 and 10. Another interesting application is their use in the boundary element method, where a boundary value problem is first transformed into an integral equation by using Green’s theorem, after which the latter is discretised. For boundary value problems on infinite domains, as well as for certain other problems, this method can be very attractive; see, e.g., [16, 51].

**Exercises**

4.1. Consider the set of functions

\[ f(x; t) = \frac{t}{\pi(x^2 + t^2)}, \quad x \in \mathbb{R}, \quad t > 0. \]

Show that \( \lim_{t \downarrow 0} f(x; t) = \delta(x) \).

4.2. Let the following set of (locally) integrable functions for \( x \in \mathbb{R}, t > 0 \), be given:

\[ f(x; t) \geq 0, \]

They have the properties

\[ \int_{-\infty}^{\infty} f(x; t) \, dx = 1, \quad t > 0, \]

\[ \lim_{t \downarrow 0} \int_a^b f(x; t) \, dx = 1 \quad \text{for} \quad a < 0 < b. \]

Show that \( \lim_{t \downarrow 0} f(x; t) = \delta(x) \).

4.3. Show that a (fundamental) solution of the problem

\[ \nabla^2 w(x; \xi) = \delta(\xi), \quad x \in \mathbb{R}^3, \]

is given by

\[ w(x; \xi) = -\frac{1}{4\pi \| x - \xi \|_2}. \]

4.4. Determine the Green’s function of the problem

\[ \frac{du}{dx} = 0, \quad x \in (\alpha, \beta), \]

\[ u(\alpha) = u_\alpha, \quad u(\beta) = u_\beta. \]
4.5. Given the form of a fundamental solution as found in Exercise 4.3, determine a Green’s function for the problem

\[ \nabla^2 u = 0, \quad x \in \Omega := B(0; 1), \]
\[ u = g(x), \quad x \in \partial \Omega, \]

where \( B(0; 1) \) denotes the unit ball of radius one centred at the origin. \textit{Hint:} Use an appropriate mirror point.

4.6. Consider the problem

\[ \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in \mathbb{R}, \quad t > 0, \]
\[ u(x, 0) = v(x), \quad x \in \mathbb{R}. \]

Use a Duhamel integral to show that the solution is given by

\[ \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} \exp\left( -\frac{(x - \xi)^2}{4t} \right) v(\xi) \, d\xi. \]

4.7. Under the same assumptions as in Property 4.21, show that the following holds:

\[ u(x, t) = \int_0^t \beta(t) \frac{\partial}{\partial t} \bar{w}(x, t - \tau) \, d\tau. \]
Chapter 5
Approximation by Finite Differences

This chapter gives the basic methods that will be used in this book for numerical computations. In Section 5.1 we start with simple polynomial interpolation, which is then used to derive differentiation and integration formulas. These form the basis of the finite difference approximations treated in Section 5.2 for space variables, both for one- and higher-dimensional problems. Some attention is paid to curved boundaries and coordinate systems other than Cartesian ones. Reformulating a problem in terms of integrals leads to finite volume methods, dealt with in Section 5.3. These are akin to finite difference methods but have a "conservation property" that is useful in certain applications. In Section 5.4 the finite differences are used to define methods for solving initial value problems. We briefly overview both one-step methods and multistep methods. Also in this section we encounter the notion of stiffness, which plays an essential role in evolution equations. The methods described in this chapter all lead to difference equations, which are solved numerically to obtain an approximate solution. The last two sections are devoted to the stability and convergence of such numerical methods. In Section 5.5 we discuss the celebrated Lax equivalence theorem, stating that for consistent discretizations stability is equivalent to convergence. In Section 5.6 a practical criterion based on Fourier analysis is worked out to establish necessary conditions for stability.

5.1 Basic Methods

The method of choice for approximating the differential equations in this book is finite differences. In this chapter we therefore discuss the basic tools underlying this method. These tools are interpolation, numerical differentiation, and numerical quadrature, which are considered first in Section 5.1.1. The remainder of this section is devoted to difference methods proper.

5.1.1 Interpolation, Numerical Differentiation, and Quadrature

If we want to approximate a function $u = u(\xi)$, we can use a variety of methods. The most basic ones (appropriate for our goal) are those based on interpolation. In view of the
available computer arithmetic we opt for polynomials since these can be evaluated directly by the simple algebraic operations (+, −, ×, ÷) that a computer can carry out.

Suppose \( u = u(ξ) \) is defined on an interval \([α, β]\). Let \( ξ_1, ξ_2, \ldots, ξ_m \) be some given points, ordered as \( α ≤ ξ_1 < ξ_2 < \cdots < ξ_m ≤ β \), and let \( p \) be the interpolation polynomial with

\[
p(ξ_i) = u(ξ_i), \quad i = 1, 2, \ldots, m.
\]

The points \( ξ_i \) (\( i = 1, 2, \ldots, m \)) are referred to as interpolation points; \( m \) is sometimes referred to as the interpolation order. Then \( p(ξ) \), the Lagrangian interpolation polynomial, is given by

\[
p(ξ) := \sum_{i=1}^{m} L_i(ξ)u(ξ_i), \quad L_i(ξ) := \prod_{j=1, j \neq i}^{m} \frac{ξ - ξ_j}{ξ_i - ξ_j},
\]

which is a polynomial of degree not more than \( m - 1 \). The terms \( L_i(ξ) \) are called Lagrangian basis polynomials. For any \( ξ ∈ (α, β) \) the interpolation error of \( p(ξ) \) is given by

\[
u(ξ) - p(ξ) = \frac{u^{(m)}(ζ)}{m!} ω(ξ), \quad ζ ∈ (α, β),
\]

where

\[
ω(ξ) := \prod_{j=1}^{m}(ξ - ξ_j).
\]

Note that \( ζ \) generally depends on \( ξ \); i.e., \( ζ = ζ(ξ) \). One can even prove [117] that

\[
\frac{d}{dξ} \left( \frac{u^{(m)}(ζ)}{m!} \right) = \frac{u^{(m+1)}(η)}{(m+1)!}, \quad η ∈ (α, β).
\]

Often the interpolation points \( ξ_i \) are equispaced; i.e., for some \( h \) we have \( ξ_i - ξ_{i-1} =: h \) \((i = 2, 3, \ldots, m)\). Then the factor \( ω(ξ) \) should be scaled by a factor \( h^{-m} \). More generally, it makes sense to use a scaling with

\[
h := \max_{2 ≤ i ≤ m} |ξ_i - ξ_{i-1}|,
\]

resulting in the error term

\[
u(ξ) - p(ξ) = h^m \frac{u^{(m)}(ζ)}{m!} (h^{-m} ω(ξ)).
\]

Let \( ξ \) be any of the interpolation points, say \( ξ = ξ_n \). We can approximate the derivative of \( u \) in \( ξ_n \) by differentiating the interpolation polynomial \( p(ξ) \). This way, we find

\[
\frac{d}{dξ} u(ξ_n) ≡ \frac{d}{dξ} p(ξ_n) = \sum_{i=1}^{m} \frac{d}{dξ} L_i(ξ_n) u(ξ_i).
\]

Applying the proper scaling, we can rewrite (5.5) as (suppressing the obvious dependence of the coefficients on \( ξ_n \))

\[
\frac{d}{dξ} u(ξ_n) ≡ \frac{1}{h} \sum_{i=1}^{m} a_i u(ξ_i) \quad (n = 1, 2, \ldots, m),
\]
where the weights $a_i$ are given by
\begin{equation}
 a_i := \frac{d}{d\xi} L_i(\xi_n) \quad (i = 1, 2, \ldots, m). \tag{5.6b}
\end{equation}

Note that these weights are independent of the scale in which $\xi$ is “measured.” Such “scaled” (or from a physical point of view “dimensionless”) variables have obvious advantages. Actually, in this book we will encounter many such “dimensionless” variables, often with a physical interpretation. A formula like the one on the right-hand side of (5.6a) is called a finite difference.

Differentiating (5.2b) and taking into account that $\omega(\xi_n) = 0$ for any interpolation point $\xi_n$, we find the following expression for the error in (5.6a):
\begin{equation}
 \frac{du}{d\xi}(\xi_n) - \frac{1}{h} \sum_{i=1}^{m} a_i u(\xi_i) = u^{(m)}(\xi) \frac{d\omega}{d\xi}(\xi_n), \quad \xi \in (\alpha, \beta). \tag{5.7}
\end{equation}

In view of (5.2b) we thus conclude that this error can be written as
\begin{equation}
 \frac{du}{d\xi}(\xi_n) - \frac{1}{h} \sum_{i=1}^{m} a_i u(\xi_i) = c_m h^{m-1} \frac{u^{(m)}(\xi)}{m!}, \quad c_m \in \mathbb{R}, \quad \xi \in (\alpha, \beta). \tag{5.8}
\end{equation}

Example 5.1 For $m = 2$ (linear interpolation) we choose $\xi_1 = \alpha$ and $\xi_2 = \beta$, giving the interpolation polynomial
\begin{equation}
p(\xi) = L_1(\xi) u(\xi_1) + L_2(\xi) u(\xi_2), \quad L_1(\xi) = -\frac{1}{h}(\xi - \xi_2), \quad L_2(\xi) = \frac{1}{h}(\xi - \xi_1),
\end{equation}
where $h$ is the interval length; i.e., $h = \xi_2 - \xi_1$. The error for $\xi \in (\xi_1, \xi_2)$ is given by
\begin{equation}
u(\xi) - p(\xi) = \frac{u^{(2)}(\xi)}{2} (\xi - \xi_1)(\xi - \xi_2), \quad \xi \in (\xi_1, \xi_2).
\end{equation}

Differentiation of $p(\xi)$ gives
\begin{equation}
\frac{du}{d\xi}(\xi) = \frac{dp}{d\xi}(\xi) = \frac{1}{h}(u(\xi_2) - u(\xi_1)).
\end{equation}
The error here is given by
\begin{equation}
\frac{du}{d\xi}(\xi_n) - \frac{1}{h} (u(\xi_2) - u(\xi_1)) = c_n h^{n-1} \frac{u^{(2)}(\xi)}{2}, \quad n = 1, 2, \quad \xi \in (\alpha, \beta),
\end{equation}
where $c_1 = -\frac{1}{2}$ and $c_2 = \frac{1}{2}$. \hfill \Box

Example 5.2 If $m = 3$ and $\alpha = \xi_1 < \xi_2 < \xi_3 = \beta$, the interpolation polynomial reads
\begin{equation}
p(\xi) = L_1(\xi) u(\xi_1) + L_2(\xi) u(\xi_2) + L_3(\xi) u(\xi_3), \quad L_1(\xi) = \frac{1}{h_1(h_1 + h_2)} (\xi - \xi_2)(\xi - \xi_3), \quad L_2(\xi) = \frac{1}{h_1 h_2} (\xi - \xi_1)(\xi - \xi_3), \quad L_3(\xi) = \frac{1}{(h_1 + h_2)h_2} (\xi - \xi_1)(\xi - \xi_2).
\end{equation}
where \( h_1 = \xi_2 - \xi_1 \) and \( h_2 = \xi_3 - \xi_2 \). Hence we find for the central point \( \xi_2 \) that
\[
\frac{du}{d\xi}(\xi_2) = \frac{dp}{d\xi}(\xi_2) = \frac{-h_2}{h_1(h_1 + h_2)}u(\xi_1) + \frac{h_2 - h_1}{h_1 h_2}u(\xi_2) + \frac{h_1}{(h_1 + h_2)h_2}u(\xi_3).
\]

The error is given by
\[
\frac{du}{d\xi}(\xi_2) - \frac{dp}{d\xi}(\xi_2) = \frac{-h_1 h_2}{6}u^{(3)}(\zeta), \quad \zeta \in (\alpha, \beta).
\]

If the interpolation points are equispaced, i.e., \( h_1 = h_2 = h \), the approximation becomes
\[
\frac{du}{d\xi}(\xi_2) = \frac{1}{2h}(u(\xi_1) - u(\xi_3)).
\]

For higher-order derivatives we can follow the same idea. In particular, we obtain for the \( k \)th derivative, where \( k \leq m \), that
\[
\frac{d^k u}{d\xi^k}(\xi_n) = \frac{1}{h^k} \sum_{i=1}^{m} a_i(k)u(\xi_i), \quad a_i(k) := h^k \frac{d^k L_i}{d\xi^k}(\xi_n).
\]

Using Leibniz’s rule and repeatedly applying (5.3), we find the error in the difference formula to be
\[
\frac{d^k u}{d\xi^k}(\xi_n) - \frac{1}{h^k} \sum_{i=1}^{m} a_i(k)u(\xi_i) = \sum_{l=1}^{k} \binom{k}{l} \frac{u^{(m+k-l)}(\zeta_{k,l})}{(m+k-l)!} \frac{d^l \omega}{d\xi^l}(\xi_n), \quad \zeta_{k,l} \in (\alpha, \beta).
\]

In the derivation of (5.9b) we have used that \( \omega(\xi_n) = 0 \).

**Example 5.3** Consider the second order interpolation polynomial \( p \) as derived in Example 5.1. Differentiating \( p \) twice, we obtain
\[
\frac{d^2 u}{d\xi^2}(\xi_2) = \frac{d^2 p}{d\xi^2}(\xi_2) = \frac{2}{h_1(h_1 + h_2)}u(\xi_1) - \frac{2}{h_1 h_2}u(\xi_2) + \frac{2}{(h_1 + h_2)h_2}u(\xi_3).
\]

In particular, if \( h_1 = h_2 = h \), then we find at \( \xi_2 \) that
\[
\frac{d^2 u}{d\xi^2}(\xi_2) = \frac{1}{h^2}(u(\xi_1) - 2u(\xi_2) + u(\xi_3)),
\]
with an error
\[
\frac{d^2 u}{d\xi^2}(\xi_2) - \frac{1}{h^2}(u(\xi_1) - 2u(\xi_2) + u(\xi_3)) = \frac{h^2}{12}u^{(4)}(\xi_2). \quad \square
\]

We can also use Lagrangian interpolation to obtain approximations of integrals. Consider
\[
\int_{\alpha}^{\beta} u(\xi) \, d\xi.
\]

Integrating the interpolation polynomial for \( u(\xi) \), we find the approximation
\[
\int_{\alpha}^{\beta} u(\xi) \, d\xi \approx h \sum_{i=1}^{m} b_i u(\xi_i), \quad (5.11a)
\]
5.1. Basic Methods

where the weights \( b_i \) are given by

\[
b_i := \frac{1}{h} \int_{\alpha}^{\beta} L_i(\xi) \, d\xi \quad (i = 1, 2, \ldots, m).
\]

The right-hand side in (5.11a) is called a quadrature formula with weights \( b_i \). Again, \( b_i \) is “dimensionless.” The expression for the error is slightly more complicated to derive; it may, e.g., contain a derivative higher than the interpolation order would suggest. For all relevant methods it turns out to be of the form

\[
\int_{\alpha}^{\beta} u(\xi) \, d\xi - h \sum_{i=1}^{m} b_i u(\xi_i) = c h^{q+1} u^{(q)}(\zeta), \quad c \in \mathbb{R}, \quad q \in \mathbb{N}, \quad \zeta \in (\alpha, \beta).
\] (5.12)

Clearly \( q \geq m \), since the interpolation formula is exact for polynomials of degree not more than \( m - 1 \).

**Example 5.4** Consider the linear interpolation polynomial introduced in Example 5.1. Then integration gives

\[
\int_{\xi_1}^{\xi_2} u(\xi) \, d\xi = \int_{\xi_1}^{\xi_2} p(\xi) \, d\xi = \frac{h}{2} (u(\xi_1) + u(\xi_2)).
\]

which is called the trapezoidal formula. In order to find the error, we use that the method is exact for polynomials of degree zero and one. Employing the method for \( u(\xi) = \xi^2 \), we find that the error is not zero. This then easily shows that

\[
\int_{\xi_1}^{\xi_2} u(\xi) \, d\xi - \frac{h}{2} (u(\xi_1) + u(\xi_2)) = -\frac{1}{12} h^3 u^{(2)}(\zeta), \quad \zeta \in (\alpha, \beta).
\]

**Example 5.5** Let us consider the case \( m = 1 \), but with the point \( \xi_1 \) chosen in a smart way, i.e., the middle of the interval, or \( \xi_1 := \frac{1}{2}(\alpha + \beta) \). We obtain

\[
u(\xi) \equiv p(\xi) = u\left(\frac{1}{2}(\alpha + \beta)\right),
\]

resulting in the approximation

\[
\int_{\alpha}^{\beta} u(\xi) \, d\xi = \int_{\alpha}^{\beta} p(\xi) \, d\xi = (\beta - \alpha) u\left(\frac{1}{2}(\alpha + \beta)\right).
\]

This formula is called the midpoint formula. It is exact for both constant polynomials and linear functions, as one can easily check. It is not exact for \( u(\xi) = \xi^2 \). As in the previous example, one can thus deduce that the error is given by

\[
\int_{\alpha}^{\beta} u(\xi) \, d\xi - (\beta - \alpha) u\left(\frac{1}{2}(\alpha + \beta)\right) = \frac{1}{24} h^3 u^{(3)}(\zeta), \quad \zeta \in (\alpha, \beta).
\]

There are special quadrature methods that can achieve high order by choosing the points \( \xi_n \) in a special way; see section D in the appendix.
5.1.2 Finite Difference Equations

We would now like to use these approximations in the context of differential equations. Typically, we then have to approximate an equation like
\[ \frac{du}{d\xi} = f(u). \] (5.13)

Here \( f \) may still be a complicated function (in particular containing derivatives of other variables). If we replace the derivative \( \frac{du}{d\xi} \) at the points \( \xi_n \) (\( n = 1, 2, \ldots, m \)) by the finite difference formula (5.6), we obtain, for the numerical approximation \( u_i \) of \( u(\xi_i) \) (i.e., \( u_i \approx u(\xi_i) \)), the difference equation
\[ \frac{1}{h} \sum_{i=1}^{m} a_i u_i = f(u_n), \quad n = 1, 2, \ldots, m, \] (5.14)

with the coefficients \( a_i \) defined in (5.6b). If we have a \( k \)th order ODE
\[ \frac{d^ku}{d\xi^k} = f(u), \] (5.15)

the difference equation reads
\[ \frac{1}{h^k} \sum_{i=1}^{m} a_i(k) u_i = f(u_n), \quad n = 1, 2, \ldots, m. \] (5.16)

Naturally, in order to make (5.16) nontrivial, we need to require \( m \geq k + 1 \).

**Example 5.6** Let us summarise what we have derived in the previous examples. For (5.13) we obtain from Example 5.1 the difference equations

\[ \frac{1}{h} (u(\xi_2) - u(\xi_1)) = f(u(\xi_1)) \] (explicit Euler difference equation),
\[ \frac{1}{h} (u(\xi_2) - u(\xi_1)) = f(u(\xi_2)) \] (implicit Euler difference equation),

and from Example 5.2 we obtain
\[ \frac{-h_2}{h_1(h_1 + h_2)} u(\xi_1) + \frac{h_2 - h_1}{h_1h_2} u(\xi_2) + \frac{h_1}{(h_1 + h_2)h_2} u(\xi_3) = f(u(\xi_n)) \quad (n = 1, 2, 3). \]

For (5.15) with \( k = 2 \) we obtain from Example 5.3 the difference scheme
\[ \frac{1}{h^2} (u(\xi_1) - 2u(\xi_2) + u(\xi_3)) = f(u(\xi_2)). \]

There is an alternative way to obtain a difference equation for (5.13). We can integrate (5.13) over an interval \((\xi_i, \xi_n)\) say; i.e.,
\[ u(\xi_n) - u(\xi_i) = \int_{\xi_i}^{\xi_n} f(u(\xi)) \, d\xi. \] (5.17)
5.1. Basic Methods

So we have reformulated the differential equation as an integral equation. The integral in (5.17) is then approximated by a suitable quadrature rule; cf. (5.11). This leads to the relation

\[ u(\xi_n) - u(\xi_i) = \frac{1}{h} \int_{\xi_i}^{\xi_n} L_j(\xi) \, d\xi. \]  

(5.18)

The resulting difference equation can be written as (again with \( u_i = u(\xi_i) \) and \( h \):= \( \max_{2 \leq n \leq m} |\xi_n - \xi_{n-1}| \))

\[ u_n - u_i = h \sum_{j=1}^{m} b_j f(u_j), \quad n = 2, 3, \ldots, m, \quad i = 1, 2, \ldots, n-1. \]  

(5.19)

Difference equations are used to solve a differential equation on a domain. So if the differential equation (5.13) is defined for \( \xi \in (\alpha, \beta) \), we have to choose specific points at which we would like to apply the difference equation. This process is called discretisation and the result is called a grid or mesh, which consists of a set of points \( \{\eta_k\}_{k=1}^{N} \), say. If the points are equispaced, we call the grid uniform; in this case we see that the grid size is \( h = (\beta - \alpha)/(N - 1) \). At each grid point \( \eta_k \) we have to apply a difference equation like (5.14) or (5.19). If the number of interpolation points \( m \) equals one or two, this is trivial. If \( m \geq 2 \), two typical choices are often taken:

(i) The nodes \( \xi_1, \xi_2, \ldots, \xi_m \) are chosen from (neighbouring) grid points \( \eta_n \).

(ii) The nodes \( \xi_2, \ldots, \xi_{m-1} \) are chosen as intermediate points on \( (\xi_1, \xi_m) = (\eta_k, \eta_{k+1}) \).

For the choice (i) we take the grid points \( \eta_{k+1-r-(m-1)}, \ldots, \eta_{k+1-r} \) as interpolation points and apply either (5.14) or (5.19). We would like to write the latter difference equation in a form commensurate with the differential equation; i.e.,

\[ \frac{1}{h} (u_{k+1} - u_k) = \sum_{j=r}^{r+m-1} b_j f(u_{k+1-j}), \quad r = 0, 1, \ldots, m-1. \]  

(5.20)

which is a multistep difference equation (because it involves more than one step, at least for \( m \geq 1 \)). In an initial value context it makes sense to choose \( r = 0 \) or \( r = 1 \). We may consider more general difference equations, of which (5.14) and (5.20) are special cases:

\[ \frac{1}{h} \sum_{j=0}^{m-1} a_j u_{k+1-j} = \sum_{j=r}^{r+m-1} b_j f(u_{k+1-j}), \quad r = 0, 1. \]  

(5.21)

For the choice (ii) the intermediate points need to be eliminated from the equation somehow because they do not correspond to grid points. Let us start with the relation (5.17), with \( \xi_1 = \eta_k \) and \( \xi_n = \eta_{k+1} \), so

\[ u(\eta_{k+1}) - u(\eta_k) = \int_{\eta_k}^{\eta_{k+1}} f(u(\xi)) \, d\xi = h \sum_{j=1}^{m} b_j f(u(\xi_j)). \]  

(5.22)
Chapter 5. Approximation by Finite Differences

The idea now is to simply use (5.17) again in order to find extra relations for each of the function values \( u(\xi_j) \) such that they can be eliminated. Of course we again have to approximate the introduced integrals by some quadrature formula. So we obtain

\[
 u(\xi_j) - u(\xi_1) = u(\xi_j) - u(\eta_k) = \int_\eta_k^{\xi_j} f(u(\xi)) \, d\xi = h \sum_{l=1}^{m} \gamma_{jl} f(u(\xi_l)),
\]

for some \( \gamma_{j1}, \ldots, \gamma_{jm} \). From this we thus find

\[
 \frac{1}{h}(u_{k+1} - u_k) = \sum_{i=1}^{m} b_i k_i, \quad k_i := f \left( u_k + h \sum_{j=1}^{m} \gamma_{ij} k_j \right).
\]

This process is sometimes called collocation. Formula (5.24) is clearly a one-step difference method. The integer \( m \) is called the number of stages of the method.

**Example 5.7** Consider the (first order) interpolation formula defined on the interval \((\eta_k, \eta_{k+1})\):

\[
p(\xi) = f(u(\xi_1)), \quad \xi_1 := \frac{1}{2}(\eta_k + \eta_{k+1}).
\]

This way, we find the approximation

\[
 u(\eta_{k+1}) - u(\eta_k) = \int_{\eta_k}^{\eta_{k+1}} f(\xi) \, d\xi = hf(u(\xi_1)),
\]

with \( h = \eta_{k+1} - \eta_k \) the grid size, from which we have to eliminate \( u(\xi_1) \). To this end we use the relation found from a simple first order quadrature formula

\[
 u(\xi_1) - u(\eta_k) = \frac{1}{2} hf(u(\eta_k)).
\]

Combining of these two relations gives

\[
 u_{k+1} - u_k = hf \left( u_k + \frac{1}{2} hf(u_k) \right).
\]

Since we are replacing the differential equation by a difference equation, we are making an error in general. A useful concept to measure this is the local discretisation error \( \delta(\eta_k) \), say. In order to let the difference equation be commensurate with the differential equation, we shall define the discretisation error for the properly scaled form (cf. (5.21)). We thus have

\[
 \delta(\eta_k) := \frac{1}{h} \sum_{j=0}^{m-1} a_j u(\eta_{k+1-j}) - \sum_{j=r}^{r+m-1} b_j f(u(\eta_{k+1-j})).
\]

Hence \( \delta(\eta_k) \) can be found by substituting the exact solution and determining the residual. Similarly, for (5.24) we define

\[
 \delta(\eta_k) := \frac{1}{h} (u(\eta_{k+1}) - u(\eta_k)) - \sum_{i=1}^{m} b_i k_i^*, \quad k_i^* := f \left( u(\eta_k) + h \sum_{j=1}^{m} \gamma_{ij} k_j^* \right).
\]
5.2 Finite Difference Methods for Spatial Variables

The local discretisation error \( \delta(\eta_k) \) is always an expression of the form

\[
\delta(\eta_k) = c h^p f^{(p+1)}(\vartheta), \quad c \in \mathbb{R}, \quad p \in \mathbb{N}, \quad \vartheta \in (\eta_k, \eta_{k+1}), \quad (5.27)
\]

and we say that \( \delta(\eta_k) \) is of order \( p \). The higher the order, the larger the value of \( m \) required in general. However, there is no simple rule relating these (cf. [21]).

**Example 5.8** Consider the difference equation (cf. Example 5.6)

\[
u_{k+1} - u_k = hf(u_k).
\]

Substituting \( f(u(\eta_k)) = u'(\eta_k) \), we find

\[
\delta(\eta_k) = \frac{1}{h} (u(\eta_{k+1}) - u(\eta_k)) - u'(\eta_k),
\]

where \( h := \eta_{k+1} - \eta_k \). If we use a Taylor series expansion around \( \eta_k \), we obtain

\[
\delta(\eta_k) = \frac{1}{h} \left( u(\eta_k) + hu'(\eta_k) + \frac{1}{2} h^2 u''(\vartheta) - u(\eta_k) \right) - u'(\eta_k)
\]

\[
= \frac{1}{2} hu''(\vartheta), \quad \vartheta \in (\eta_k, \eta_{k+1}). \quad \square
\]

5.2 Finite Difference Methods for Spatial Variables

The difference equations we defined in the previous section can be applied to space and time derivatives separately. In this section we consider spatial derivatives first.

5.2.1 Finite Difference Approximations

By far the oldest method of approximating derivatives is by using finite differences. To start with, consider for a function \( u = u(x) \) approximations of \( \frac{d}{dx}u \) and \( \frac{d^2}{dx^2}u \). Using the basic ideas outlined in Section 5.1, we find the difference formulas

\[
\frac{du}{dx}(x) \approx \frac{1}{h} (u(x + h) - u(x)) \quad \text{(forward difference),} \quad (5.28a)
\]

\[
\frac{du}{dx}(x) \approx \frac{1}{h} (u(x) - u(x - h)) \quad \text{(backward difference),} \quad (5.28b)
\]

\[
\frac{du}{dx}(x) \approx \frac{1}{2h} (u(x + h) - u(x - h)) \quad \text{(central difference).} \quad (5.28c)
\]

Employing Taylor series expansions, we obtain the following expressions for the corresponding local discretisation errors:

\[
\delta(x) = -\frac{1}{2} h \frac{d^2 u}{dx^2}(\xi), \quad \xi \in (x, x + h), \quad (5.29a)
\]

\[
\delta(x) = +\frac{1}{2} h \frac{d^2 u}{dx^2}(\xi), \quad \xi \in (x - h, x), \quad (5.29b)
\]

\[
\delta(x) = -\frac{1}{6} h^2 \frac{d^3 u}{dx^3}(\xi), \quad \xi \in (x - h, x + h). \quad (5.29c)
\]
For the second order derivative \( \frac{d^2 u}{dx^2} \) we find the (second order) central difference
\[
\frac{d^2 u}{dx^2}(x) = \frac{-2}{h_1(h_1 + h_2)} u(x - h_1) - \frac{2}{h_1 h_2} u(x) + \frac{2}{(h_1 + h_2)h_2} u(x + h_2). \tag{5.30}
\]
Expanding in a Taylor series around \( x \), we find for the local discretisation error (defined in an obvious way, similar to the first order case)
\[
\delta(x) = \frac{h_2 - h_1}{3} \frac{d^3 u}{dx^3}(x) - \frac{1}{12 h_1 + h_2} \frac{d^3 u}{dx^3}(\xi), \quad \xi \in (x - h_1, x + h_2). \tag{5.31}
\]
Clearly, if \( h_1 \) is significantly different from \( h_2 \), we have \( \delta(x) = O(h) \), where \( h := \max(h_1, h_2) \).
On the other hand, if \( h_1 = h_2 = h \), we have \( \delta(x) = O(h^2) \).

Often second order derivatives appear in a differential equation as \( \frac{d}{dx}(a(x) \frac{du}{dx}) \). One may approximate these by applying (5.28a) and (5.28b) as follows:
\[
\frac{d}{dx}(a(x) \frac{du}{dx})(x) \doteq \frac{1}{h} \frac{d}{dx}(a(x)(u(x + h) - u(x)))
\]
\[
= \frac{1}{h} \left( a(x)(u(x + h) - u(x)) - a(x - h)(u(x) - u(x - h)) \right)
\]
\[
= \frac{1}{h^2} \left( a(x)u(x + h) - (a(x) + a(x - h))u(x) + a(x - h)u(x - h) \right). \tag{5.32a}
\]

One can do better by making the second derivative more symmetric through the use of central differences. For this we note that (5.28c) can also be seen as an average of (5.28a) and (5.28b). Applying this idea, we find
\[
\frac{d}{dx}(a(x) \frac{du}{dx})(x) \doteq \frac{1}{h} \frac{d}{dx}\left( a \left( x + \frac{1}{2}h \right) \right) \left( u(x + h) - u(x) \right)
\]
\[
= \frac{1}{h^2} \left[ a \left( x + \frac{1}{2}h \right) u(x + h) - a \left( x - \frac{1}{2}h \right) u(x) \right] - \frac{1}{h^2} \left[ a \left( x + \frac{1}{2}h \right) u(x + h) - a \left( x + \frac{1}{2}h \right) u(x) \right]
\]
\[
= \frac{1}{h^2} \left[ a \left( x + \frac{1}{2}h \right) u(x + h) - a \left( x + \frac{1}{2}h \right) u(x) \right]
= \frac{1}{h^2} \left[ a \left( x + \frac{1}{2}h \right) u(x + h) - a \left( x + \frac{1}{2}h \right) u(x) \right]
= \frac{1}{h^2} \left[ a \left( x + \frac{1}{2}h \right) u(x + h) - a \left( x + \frac{1}{2}h \right) u(x) \right] + a \left( x + \frac{1}{2}h \right) u(x - h). \tag{5.32b}
\]

Both difference formulas can be shown to be second order; we leave this as an exercise.

The finite difference approximations mentioned in this section are often referred to as a difference scheme. Such a scheme is typically of a local nature and at a certain point describes the relation of the value of the unknown to that of its neighbours. The points involved in this are called the stencil. Since a numerical scheme will usually only give an approximation of the derivatives, we shall use a special notation for the approximate solution (see Section 5.1). Given the points \( x_k, k = 1, 2, \ldots, M \), on the grid we shall write \( u_k \doteq u(x_k) \).
5.2. Finite Difference Methods for Spatial Variables

5.2.2 Finite Difference Approximations in Several Dimensions

If we have several spatial dimensions, we can apply the difference schemes given in Section 5.2.1 to each of the variables separately. In particular, for the Laplace operator $\nabla^2$ in two dimensions we thus find, with mesh widths $\Delta x$ in the $x$ direction and $\Delta y$ in the $y$ direction, that

$$
\mathcal{L}[u] := \nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{1}{\Delta x^2} \left[ u(x + \Delta x, y) - 2u(x, y) + u(x - \Delta x, y) \right] + \frac{1}{\Delta y^2} \left[ u(x, y + \Delta y) - 2u(x, y) + u(x, y - \Delta y) \right].
$$

(5.33)

In this book we shall often use subscripts $j$ and $k$ for the approximate solution at the mesh points in the $x$ and $y$ directions, respectively; more precisely, we have $u_{j,k} \equiv u(x_j, y_k)$. If we denote the numerical approximation of the Laplace operator by $\mathcal{L}_{\Delta}$, we may write (5.33) as (cf. Figure 5.1)

$$
\mathcal{L}_{\Delta}[u_{j,k}] := \frac{1}{\Delta x^2} \left( u_{j+1,k} - 2u_{j,k} + u_{j-1,k} \right) + \frac{1}{\Delta y^2} \left( u_{j,k+1} - 2u_{j,k} + u_{j,k-1} \right).
$$

(5.34a)

Alternatively, to indicate the positions of points in a stencil one often employs directions on a compass, like in Figure 5.1, with respect to a central point C, say. The approximate solution at C is written as $u_c$. So (5.33) then becomes

$$
\mathcal{L}_{\Delta}[u_c] = \frac{1}{\Delta x^2} (u_e - 2u_c + u_w) + \frac{1}{\Delta y^2} (u_n - 2u_c + u_s).
$$

(5.34b)

Unlike the one-dimensional case, the two-dimensional case poses a problem related to the geometry of the domain, even for nonuniform grids. Indeed, a rectangular grid will rarely fit in with boundary curves. Hence we have to adapt the stencil appropriately for points neighbouring the boundaries. If we have, e.g., a problem like in Figure 5.2, with uniform mesh width $h$, then we can easily employ the difference formula (5.30). Defining $\alpha := \frac{1}{h} (x_e - x_c)$, we simply find

$$
\left( \frac{\partial^2 u}{\partial x^2} \right)_c = \frac{2}{h^2} \left( \frac{1}{\alpha + 1} u_w - \frac{1}{\alpha} u_v + \frac{1}{\alpha(\alpha + 1)} u_e \right).
$$

(5.35)

Figure 5.1. Stencil of the Laplace operator in compass notation (left) and index notation (right).
and likewise for $\frac{\partial^2 u}{\partial y^2}$. Note that this only makes sense if $u_e$ is given, i.e., if we know $u$ at the boundary point $E'$. If we have a derivative in the boundary condition, the situation is more complicated. Consider Figure 5.2, where we have to employ the boundary condition $\frac{\partial u}{\partial n}$ given at $S'$. Let us define

$$
\alpha := \frac{1}{h}(x_e - x_C), \quad \beta := \frac{1}{h}(y_e - y_S'), \quad \gamma := \sqrt{\alpha^2 + \beta^2}.
$$

Then

$$
\left( \frac{\partial u}{\partial n} \right)_{S'} \approx \frac{1}{\gamma h} (u_{S'} - u_T). \tag{5.36}
$$

By linear interpolation we find

$$
u_T \approx \alpha u_W + (1 - \alpha) u_C. \tag{5.37}
$$

Using (5.36), we can therefore express $u_{S'}$ in values at grid points and the given boundary condition $\left( \frac{\partial u}{\partial n} \right)_{S'}$:

$$
u_{S'} \approx \alpha u_W + (1 - \alpha) u_C + \gamma h \left( \frac{\partial u}{\partial n} \right)_{S'}. \tag{5.38}
$$

The value $u_{S'}$ shows up in discretising $\frac{\partial^2 u}{\partial y^2}$ (cf. (5.30)):

$$
\left( \frac{\partial^2 u}{\partial y^2} \right)_{C'} \approx \frac{2}{h^2} \left( \frac{1}{\beta(\beta + 1)} u_{S'} - \frac{1}{\beta} u_C + \frac{1}{\beta + 1} u_N \right). \tag{5.39}
$$

Clearly, upon substituting (5.38) into (5.39), we obtain a scheme for discretising $\left( \frac{\partial^2 u}{\partial y^2} \right)_{C'}$ with unknowns at either grid points or given boundary conditions. The treatment for $\left( \frac{\partial^2 u}{\partial x^2} \right)_{E'}$ goes in a similar way. Note that (5.37) is second order in $h$; cf. Section 5.1. Although we have considered two-dimensional schemes exclusively, the extension to three dimensions is straightforward and is therefore omitted.

### 5.2.3 Other Coordinate Systems

We conclude this section by considering some special coordinate systems, e.g., because these are better suited to the geometry of the problem or because we want to take advantage
5.2. Finite Difference Methods for Spatial Variables 83

of certain symmetries. Let us start with the Poisson problem in polar coordinates for smooth source \( f \) and regular and continuous \( u \):

\[
\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = f(r, \phi). \tag{5.40}
\]

First we define a grid with radial mesh width \( \Delta r \) and angular mesh width \( \Delta \phi \). The points on the circle with radius \( r \) have an angular coordinate \( \frac{k}{\Delta \phi} \) where \( k \) increases in a counterclockwise direction. Let \( u_{j,k} \) be the approximation of \( u(r_{j/\Delta r}, \phi_{k/\Delta \phi}) \). To discretise the various terms in the Laplace operator we can now use any of the difference methods we have seen so far; see, e.g., (5.30) and (5.32). We will not elaborate on this here in general. The main problem is the singularity of the coordinate system at the origin. One way to deal with this is to apply averaging of the values around \( r = 0 \). Let \( \Omega := D(0, \frac{1}{2} \Delta r) \) denote the disc with radius \( \frac{1}{2} \Delta r \) around the origin 0. Then we find

\[
\int_{\Omega} \left( \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} \right) r \, dr \, d\phi = \frac{1}{2} \Delta r \int_{0}^{2\pi} \frac{\partial u}{\partial r} \left( \frac{1}{2} \Delta r, \phi \right) \, d\phi. \tag{5.41}
\]

We now use central differences to approximate \( \frac{1}{r} \frac{\partial}{\partial r} u \left( \frac{1}{2} \Delta r, \phi \right) \) in (5.41). We thus have

\[
\frac{1}{2} \Delta r \int_{0}^{2\pi} \frac{\partial u}{\partial r} \left( \frac{1}{2} \Delta r, \phi \right) \, d\phi = \frac{1}{2} \Delta r \sum_{k=1}^{N} \frac{u_{1,k} - u_{0,k}}{\Delta r} \Delta \phi,
\]

where \( N = 2\pi / \Delta \phi \) is the number of grid points in the \( \phi \) direction. In general the source term \( f \) also needs to be integrated. To do this numerically we approximate it by its value at 0. So we have

\[
\int_{\Omega} f(r, \phi) r \, dr \, d\phi \approx \frac{1}{4} \pi \Delta r^2 f(0).
\]

Combining these results, we thus find an equation involving \( u_{0,0} \):

\[
-u_{0,0} + \frac{1}{N} \sum_{k=1}^{N} u_{1,k} = \frac{1}{4} \Delta r^2 f(0), \tag{5.42}
\]

where we have used continuity so that \( u_{0,k} = u_{0,0} \) for \( k = 0, 1, \ldots, N \).

Another way to handle the singular origin is by applying Taylor series. We will demonstrate this approach in three dimensions as follows. In cylindrical coordinates \( r, \phi, z \) (see Figure 5.3) we have

\[
\nabla^2 u = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2}, \tag{5.43}
\]

and in spherical coordinates \( r, \theta, \phi \) we have

\[
\nabla^2 u = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2}. \tag{5.44}
\]

With cylindrical symmetry the second term on the right-hand side of (5.43) vanishes, while all but the first term on the right-hand side of (5.44) vanish with spherical symmetry.
Suppressing any dependence on other variables, we expand $u(r)$ around $u(0)$ to obtain

$$u(r) = u(0) + \frac{1}{2} r^2 \frac{\partial^2 u}{\partial r^2}(0) + O(r^3), \quad (5.45)$$

where we have $\frac{\partial u}{\partial r}(0) = 0$ because of regularity. By inserting this expansion into (5.43) or (5.44), we find

$$\nabla^2 u = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{\partial^2 u}{\partial z^2} = 2 \frac{\partial^2 u}{\partial r^2}(0) + \frac{\partial^2 u}{\partial z^2}(0) + O(r) \quad (5.46a)$$

in cylindrical coordinates and

$$\nabla^2 u = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) = 3 \frac{\partial^2 u}{\partial r^2}(0) + O(r) \quad (5.46b)$$

in spherical coordinates. Note that the $\theta$ and $\phi$ derivatives vanish at $r = 0$ because $u$ is continuous. Once more employing the Taylor expansion (5.45) at $r = \Delta r$, we can derive the difference approximations

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) \approx \frac{4}{\Delta r^2} (u(\Delta r) - u(0)), \quad (5.47a)$$

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) \approx \frac{6}{\Delta r^2} (u(\Delta r) - u(0)) \quad (5.47b)$$

in cylindrical and spherical coordinates, respectively.

### 5.3 Finite Volume Methods

Rather than directly using finite differences to approximate spatial derivatives, we may first rewrite the PDE in its integral form. In later chapters we shall see that this is actually physically more natural, allowing the occurrence of discontinuities that are excluded in the differential formulation.
5.3. Finite Volume Methods

5.3.1 The One-Dimensional Case

Consider the ODE
\[
d\frac{f(u)}{dx} = s(u),
\]
where \( f(u) \) is a (complicated) function of \( u \) and/or derivatives of \( u \); however, for the sake of brevity we just write \( f(u) \). We shall call \( f(u) \) the flux, anticipating a concept that will be encountered later. In this section we consider the simple form

\[
f(u) := a \frac{du}{dx} + bu.
\]

The separate terms \( a \frac{du}{dx} \) and \( bu \) are sometimes referred to as the conductive flux and the convective flux, respectively. Furthermore, \( s(u) \) is a function of \( u \) only and represents a source term in physical applications.

On an arbitrary interval \((\alpha, \beta)\) we find by integration that

\[
f(u(\beta)) - f(u(\alpha)) = \int_{\alpha}^{\beta} s(u(x)) \, dx.
\]

If \( a = 0 \), we obtain a pure integral equation in (5.50), which we can solve numerically by applying suitable quadrature formulas, as was done in Section 5.1. If \( a \neq 0 \), we need to find a way to approximate the derivative term appropriately, e.g., through a finite difference formula. The finite volume method is based on this integral formulation of (5.48). In fact, we can distinguish between two variants of the finite volume method based on the choice of the interval \((\alpha, \beta)\).

Suppose we have a uniform grid \( x_1, x_2, \ldots, x_M \) of grid size \( \Delta x \) on which we want to compute the solution \( u(x) \) of (5.48). In the cell-centered method we apply the integral relation to intervals \((\alpha, \beta) = (x_{j-1/2}, x_{j+1/2}) := V_j\), where the intermediate point \( x_{j+1/2} \) is located halfway between the grid points \( x_j \) and \( x_{j+1} \), i.e., \( x_{j+1/2} := \frac{1}{2}(x_j + x_{j+1}) \), and likewise for \( x_{j-1/2} \). The interval \( V_j \) is called the control volume; see Figure 5.4. For this choice of the control volume it is obvious to approximate the integral of \( s(u) \) in (5.50) by the midpoint rule; cf. Example 5.5. We find

\[
\int_{x_{j-1/2}}^{x_{j+1/2}} s(u(x)) \, dx \approx s(u(x_j)) \Delta x.
\]

Next we have to approximate the flux \( f(u) \) at the cell boundaries \( x_{j \pm 1/2} \). For the conductive part it is natural to employ central differences, and for the convective part we simply use

\[
\text{Figure 5.4. Control volume for a cell-centered (left) and a vertex-centered (right) finite volume method.}
\]
linear interpolation, resulting in the approximation
\[ f(u(x_{j+1})) \approx \frac{a}{\Delta x} (u(x_{j+1}) - u(x_j)) + \frac{1}{2} b (u(x_j) + u(x_{j+1})). \]  
(5.52)

A similar expression holds for \( f(u(x_{j-1/2})) \). Combining the approximations in (5.51) and (5.52) with the integral formulation (5.50), we obtain the difference scheme
\[ \frac{a}{\Delta x} (u_{j+1} - 2u_j + u_{j-1}) + \frac{b}{2} (u_{j+1} - u_{j-1}) = s(u_j) \Delta x, \]  
(5.53)

where \( u_j \) denotes the numerical approximation of \( u(x_j) \). Of course, in this case the method turns out to be exactly the same as would result from applying central differences (5.28c) and (5.30) directly. In Chapter 13 this finite volume approach will turn out to have a clear advantage over straightforward finite differences when solving conservation laws numerically.

Alternatively, in the vertex-centered method, we take for the \( j \)th control volume \( V_j = (x_j, x_{j+1}) \); i.e., the control volumes coincide with the intervals between grid points. The integral of \( s(u) \) is in this case conveniently approximated by the trapezoidal rule; i.e.,
\[ \int_{x_j}^{x_{j+1}} s(u(x)) \, dx \approx \frac{1}{2} \Delta x \left[ s(u(x_j)) + s(u(x_{j+1})) \right]. \]  
(5.54)

To approximate the flux \( f(u(x_j)) \) at the left edge (vertex) of \( V_j \) we may take the backward difference formula (BDF) for the conductive part, whereas the convective part needs no approximation. This yields
\[ f(u(x_j)) \approx \frac{a}{\Delta x} (u(x_j) - u(x_{j-1})) + bu(x_j), \]  
(5.55)

and likewise for \( f(u(x_{j+1})) \). Hence we obtain the scheme
\[ \frac{a}{\Delta x} (u_{j+1} - 2u_j + u_{j-1}) + b (u_{j+1} - u_{j}) = \frac{1}{2} \Delta x (s(u_j) + s(u_{j+1})). \]  
(5.56)

Note that this is not symmetric with respect to the middle point \( x_j \). Another possibility is to choose \( V_j = (x_{j-1}, x_j) \) and employ the forward difference formula for the conductive part of the flux.

### 5.3.2 The Multidimensional Case

The derivation of the finite volume method for a differential equation in a two- or three-dimensional domain is similar to the one-dimensional case discussed in the previous section. The starting point is the generalisation of (5.48) to
\[ \nabla \cdot f(u) = s(u), \]  
(5.57)

where \( f(u) \) is now the flux vector, generally depending on \( u \) and \( \nabla u \), and \( s(u) \) is the source term, solely depending on \( u \). Equation (5.57) is an example of a conservation law. For the flux vector we adopt the model
\[ f(u) = a \nabla u + bu, \]  
(5.58)

with \( a \nabla u \) and \( bu \) the conductive and convective parts, respectively.
5.3. Finite Volume Methods

Integrating (5.57) over an arbitrary domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) and using Gauss’s theorem (see appendix section (J.13)), we obtain

$$\oint_{\partial \Omega} f(u) \cdot n \, dS = \int_\Omega s(u) \, dV,$$  \hspace{1cm} (5.59)

where $n$ is the unit outward normal on the boundary $\partial \Omega$ of $\Omega$. The finite volume formulation is based on this integral formulation of (5.57). More precisely, in the finite volume method the computational domain is covered with a grid and a set of nonoverlapping control volumes, where a numerical approximation of (5.59) is applied. The arrangement of grid points and control volumes determines the method. In what follows we restrict ourselves to the cell-centred method, where grid points are located in the centre of the control volumes. Furthermore, for the sake of simplicity we restrict ourselves to a two-dimensional domain covered with a rectangular grid.

Consider the control volume $V_c$ as shown in Figure 5.5. In the cell-centred finite volume method we apply relation (5.59) to this control volume and approximate the integrals involved by suitable quadrature rules. Naturally, the midpoint rule is again an obvious choice to approximate the integral of $s(u)$, and we find, by applying this in both variables, that

$$\int_{V_c} s(u) \, dV \approx s(u(x_c)) \Delta x \Delta y.$$ \hspace{1cm} (5.60)

The contour integral of the flux in (5.59) has to be split into four parts corresponding with the four sides of $\Gamma = \partial \Omega$. The orientation of $\Gamma$ is as indicated in Figure 5.5. Let $f = (f_1, f_2)^T$. Then we have

$$\oint_{\partial V_c} f(u) \cdot n \, dS = \int_{\Gamma_e} f_1(u(x_e, \tau)) \, d\tau + \int_{\Gamma_n} f_2(u(\tau, y_n)) \, d\tau$$

$$- \int_{\Gamma_s} f_1(u(x_s, \tau)) \, d\tau - \int_{\Gamma_w} f_2(u(\tau, y_s)) \, d\tau,$$ \hspace{1cm} (5.61)

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{figure5.5.png}
\caption{Two-dimensional control volume for a cell-centered finite volume method.}
\end{figure}
where \( \tau \) is the parameter describing \( \Gamma \). Next we approximate each of these integrals by the midpoint rule to find

\[
\oint_{\partial V_c} f(u) \cdot n \, dS = \left[ f_1(u(x_e)) - f_1(u(x_w)) \right] \Delta y + \left[ f_2(u(x_n)) - f_2(u(x_s)) \right] \Delta x, \tag{5.62}
\]

where \( x_e \) denotes the centre of the eastern cell face \( \Gamma_e \), etc.; see Figure 5.5. Substituting the approximations (5.60) and (5.62) into (5.59), we obtain

\[
(F_{e} - F_{w}) \Delta y + (F_{s} - F_{n}) \Delta x = s(u_c) \Delta x \Delta y, \tag{5.63}
\]

with \( u_c \) and \( F_{i} \) the numerical approximations of \( u(x_c) \) and \( f_i(u(x_i)) \), respectively, etc. Finally, we have to determine the numerical fluxes. Like in (5.52) we have

\[
F_{i} = a \left( \frac{1}{\Delta x} (u_{n} - u_{c}) + \frac{1}{2} b_{1} (u_{c} + u_{e}) \right), \tag{5.64}
\]

and similar expressions hold for the other numerical fluxes. Substituting these expressions in the discrete conservation law (5.63) and rearranging terms, we finally obtain

\[
\frac{a}{\Delta x^2} (u_{e} - 2u_{c} + u_{w}) + \frac{b_{1}}{2 \Delta x} (u_{e} - u_{w}) + \frac{a}{\Delta y^2} (u_{n} - 2u_{c} + u_{s}) + \frac{b_{2}}{2 \Delta y} (u_{n} - u_{s}) = s(u_c). \tag{5.65}
\]

5.4 Difference Methods for Initial Value Problems

When discretising the time derivative, we don’t only encounter problems due to approximation. Indeed, evolution equations, continuous or discrete, involve questions relating to (non)growth. The growth behaviour of the discrete solution may be completely different from that of the continuous one. For obvious reasons we would like this behaviour to be the same; i.e., the numerical solution should mimic the physically relevant properties. Instability, in particular, will become apparent on longer time intervals. This section and the next two will be devoted to these questions.

5.4.1 One-Step and Multistep Difference Schemes

In Section 5.1 we derived a number of methods to approximate the derivative \( \frac{d}{dt} u \). As was stated there, we may use values at either two consecutive (time) points (one-step) or a larger number of points (multistep). So consider the ODE

\[
\frac{du}{dt} = f(u, t), \tag{5.66}
\]

where \( f(u, t) \) may contain derivatives with respect to spatial variables (but not with respect to \( t \)). If we denote the approximation of \( u(t^n) \) at time \( t^n \) by \( u^n \), then we have, for a step size \( \Delta t \), the general one-step scheme

\[
\frac{1}{\Delta t} (u^{n+1} - u^n) = \Phi(u^n, u^{n+1}, t^n; \Delta t). \tag{5.67}
\]
5.4. Difference Methods for Initial Value Problems

In particular, a collocation method with \( m \) stages is called a Runge–Kutta method. We have then (cf. (5.24))

\[
\frac{1}{\Delta t}(u^{n+1} - u^n) = \sum_{j=1}^{m} \beta_j k_j^n, \tag{5.68a}
\]

\[
k_j^n := f \left( t^n + \rho_j \Delta t, u^n + \Delta t \sum_{l=1}^{m} \gamma_{j,l} k_l^n \right). \tag{5.68b}
\]

Now recall that collocation was based on using quadrature formulas on certain intervals. Since any such quadrature formula is at least exact for the constant function (as it is based on polynomial interpolation), it follows that the sum of the weights must be equal to the interval length. Hence we find the following algebraic relations for the coefficients \( \beta_j, \rho_j, \) and \( \gamma_{j,l} \):

\[
\sum_{j=1}^{m} \beta_j = 1, \quad \rho_j = \sum_{l=1}^{m} \gamma_{j,l} \quad (j = 1, 2, \ldots, m). \tag{5.69}
\]

Note that \( 0 \leq \rho_j \leq 1 \) (\( j = 1, 2, \ldots, n \)).

The five most widely used one-step methods are as follows:

\[
\frac{1}{\Delta t}(u^{n+1} - u^n) = f(u^n, t^n) \quad \text{(explicit (or forward) Euler method),} \tag{5.70}
\]

\[
\frac{1}{\Delta t}(u^{n+1} - u^n) = f(u^{n+1}, t^{n+1}) \quad \text{(implicit (or backward) Euler method),} \tag{5.71}
\]

\[
\frac{1}{\Delta t}(u^{n+1} - u^n) = (1 - \vartheta)f(u^n, t^n) + \vartheta f(u^{n+1}, t^{n+1}), \quad \vartheta \in [0, 1] \quad \text{(\vartheta method),} \tag{5.72}
\]

\[
\frac{1}{\Delta t}(u^{n+1} - u^n) = \frac{1}{2} \left( f(u^n, t^n) + f(u^{n+1}, t^{n+1}) \right) \quad \text{(trapezoidal rule),} \tag{5.73}
\]

\[
\frac{1}{\Delta t}(u^{n+1} - u^n) = f \left( \frac{1}{2} (u^n + u^{n+1}), t^n + \frac{1}{2} \Delta t \right) \quad \text{(midpoint rule).} \tag{5.74}
\]

Note that (5.72) is a linear combination of (5.70) and (5.71) and that (5.73) is a special case of (5.72), i.e., with \( \vartheta = \frac{1}{2} \). From what we have found in Section 5.1 we conclude that (5.70) and (5.71) are first order, and thus we expect (5.72) to be first order in general as well. Since (5.73) is second order, this clearly also applies to (5.72) for \( \vartheta = \frac{1}{2} \). Finally, we found in Section 5.1 that (5.74) is a second order method.

Quite another matter is the way the method produces the (approximate) value at the new time point \( t^{n+1} \), given \( u^n \). From the set above only (5.70) directly gives a value for \( u^{n+1} \); it is therefore called an explicit method. All other methods (except for (5.72), with \( \vartheta = 0 \), which is of course just (5.70)) result in an equation that should be solved to find \( u^{n+1} \). Therefore such methods are called implicit. We briefly discuss how to deal with the latter problem at the end of this section.

**Example 5.9** The well known, classical Runge–Kutta formula, going back to Runge (1895) and Kutta (1901), is a fourth order method and reads

\[
\frac{1}{\Delta t}(u^{n+1} - u^n) = \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4),
\]
where
\[\begin{align*}
k_1 & := f(u^n, t^n), \\
k_2 & := f \left( u^n + \frac{1}{2} \Delta t k_1, t^n + \frac{1}{2} \Delta t \right), \\
k_3 & := f \left( u^n + \frac{1}{2} \Delta t k_2, t^n + \frac{1}{2} \Delta t \right), \\
k_4 & := f(u^n + \Delta t k_3, t^n + \Delta t).
\end{align*}\]

Clearly, \(\frac{1}{2} + \frac{1}{3} + \frac{1}{3} + \frac{1}{2} = 1\) (cf. (5.69)). Furthermore, we see that (cf. (5.69)) \(\rho_1 = 0, \rho_2 = \frac{1}{2} = \gamma_2, \rho_3 = \frac{1}{3} = \gamma_3, \) and \(\rho_4 = 1 = \gamma_4.\) \(\square\)

A multistep method has the form (5.21) with \(h = \Delta t, r = 0,\) and \(k\) and \(m\) replaced by \(n\) and \(k + 1,\) respectively:
\[\frac{1}{\Delta t} \sum_{j=0}^{k} a_j u^{n+1-j} = \sum_{j=0}^{k} b_j f(u^{n+1-j}, t^{n+1-j}). \tag{5.75}\]

Since \(k\) is equal to the number of steps, (5.75) is a \(k\)-step method. Two special cases of (5.75) are most widely used. The first one is the family of Adams methods; see section D in the appendix. They are based on direct integration of the interpolation polynomial of \(f\) and can be obtained from (5.20) by choosing \(m = k + 1\) and \(r = 0\) or \(r = 1\) and read
\[\frac{1}{\Delta t} (u^{n+1} - u^n) = \sum_{j=0}^{k} b_j f(u^{n+1-j}, t^{n+1-j}). \tag{5.76}\]

If \(b_0 = 0,\) the methods are clearly explicit (Adams–Bashforth); if \(b_0 \neq 0,\) they are implicit (Adams–Moulton); see appendix section D. The latter are based on finite differences using previous time points and are called backward differences. This results in the BDF family:
\[\frac{1}{\Delta t} \sum_{j=0}^{k} a_j u^{n+1-j} = f(u^{n+1}, t^{n+1}). \tag{5.77}\]

It is important to note that \(k\)-step methods require more than one initial value. This reflects the fact that their solution space is higher dimensional. One may use a one-step method as a start-up procedure. Yet there is still another problem. To see this consider their use for the model problem
\[\frac{du}{dt} = \lambda u. \tag{5.78}\]

Then (5.75) yields the difference equation
\[\sum_{j=0}^{k} (a_j - \Delta t \lambda b_j) u^{n+1-j} = 0. \tag{5.79}\]

This difference equation with constant coefficients has basis solutions of the form
\[u^n = \mu^n(\Delta t \lambda),\]
where \( \mu \) is a root of the characteristic polynomial

\[
\sum_{j=0}^{k} (a_j - \Delta t \lambda b_j) \mu^{k-j} = 0.
\] (5.80)

If \( \mu \) is a double root, then \( u^n = n \mu^n (\Delta t \lambda) \) is also a solution and so are ones with higher powers of \( n \) as coefficient if the root has even higher multiplicity. For the exact solution of (5.78) one has a mode \( u(t) \) satisfying \( u(t^n) = \exp(\Delta t \lambda) u(t^{n-1}) \). Hence one should expect at least one root of (5.80) to be approximately equal to \( \exp(\Delta t \lambda) \) for \( \Delta t \lambda \) small. In particular, this means that for this root we have \( \mu(\Delta t \lambda) \to 1 \) for \( \Delta t \lambda \to 0 \). In fact, this is a necessary condition for the method to be meaningful and is exactly equal to saying that the polynomial

\[
\sum_{j=0}^{k} a_j \mu^{k-j} = 0
\]

has a root equal to one. The corresponding root of (5.80) is called the essential root. All other roots of (5.80) are called parasitic roots. The corresponding solutions of the difference equation are then referred to as the essential solution and the parasitic solutions, respectively. The latter do not have any practical significance and may blur the numerically relevant one if no precautions are taken. This is a stability problem that is absent in (implicit) Adams methods and BDF methods. For a more general discussion one may consult [97].

We demonstrate a typical problem in the following example.

**Example 5.10** Consider the explicit midpoint rule for (5.66), i.e.,

\[
\frac{1}{\Delta t} (u^{n+1} - u^{n-1}) = 2 f(u^n, t^n).
\]

The characteristic polynomial for the model problem reads

\[
\mu^2 - 2 \lambda / \Delta t \mu - 1 = 0.
\]

Hence we find \( \mu = \pm 1 + \lambda / \Delta t + \mathcal{O}(1/\Delta t^2) \). So to leading order we have the general solution

\[
u^n = A(1 + \lambda / \Delta t)^n + B(-1 + \lambda / \Delta t)^n, \quad A, B \in \mathbb{R}.
\]

The constants \( A, B \) follow from the initial conditions. We may approximate \( 1 + \lambda / \Delta t \) by \( \exp(\lambda / \Delta t) \). Therefore we have \( u^n \approx A \exp(\lambda t^n) + B(-1)^n \exp(-\lambda t^n) \). If \( \lambda < 0 \), as typically holds for stable problems, we have an oscillatory (absolutely) growing parasitic mode; hence the problem is unstable.

Both one-step and \( k \)-step methods have advantages and disadvantages. An advantage of one-step methods is their flexibility in choosing the time step \( \Delta t \) (which is preferably not necessarily constant in order to have an efficient time integration). A disadvantage is that each stage requires the evaluation of \( f \) (which is often a boundary value problem in our setting!). The latter problem is not present for a \( k \)-step method, as it can simply use previously computed “\( f \) values.” The disadvantage of a \( k \)-step method is the change of the time step \( \Delta t \). One often does this by using interpolation of \( f \) values (not \( u \) values). This process, together with using an entire family (e.g., Adams–Moulton or BDF), has resulted in variable step, variable order methods. For more details see, e.g., [97].
If we have an implicit method, we are faced with the problem of “retrieving” the value $u^{n+1}$. There are basically two ways to solve such equations: (i) successive substitution and (ii) root finding. We consider them both briefly.

(i) We can symbolically write an implicit method as

$$u^{n+1} = \varphi(u^{n+1}).$$

(5.81)

If we have some reasonable estimate of $u^{n+1}$, say $\nu_0$, from some explicit method, the “predictor,” then we can perform successive substitution to produce a (we hope converging) sequence $\nu_j$, $j = 0, 1, 2, \ldots$, via the “corrector”

$$\nu_{j+1} = \varphi(\nu_j).$$

In order to let this converge, we not only need $\nu_0$ to be close enough to $u^{n+1}$ but also should have $|\varphi'(\nu)| < 1$. The latter condition typically implies that

$$\Delta t \left| \frac{\partial f}{\partial u} \right| < 1$$

in a neighbourhood of $u^{n+1}$; cf. (5.76). Unfortunately, problems are often such that this requirement implies $\Delta t$ be taken unduly small. Another problem caused by the latter behaviour of $f$ is discussed in the next section.

(ii) We can also rewrite the implicit method as

$$\psi(u^{n+1}) = 0.$$  

(5.82)

Again, given some initial guess $\nu_0$ for $u^{n+1}$, e.g., found from extrapolating values at previous time points, we can carry out a Newton algorithm for $j = 0, 1, 2, \ldots$:

$$\nu_{j+1} = \nu_j - \frac{\psi(\nu_j)}{\psi'(\nu_j)}.$$

One can show that this procedure does not suffer from the restriction that we have with (i). For more details one may consult [97].

### 5.4.2 Time Scales and Stiffness

Many time-dependent problems have solutions that are fairly smooth, except for some (smaller) regions where they exhibit rapid motion. This phenomenon is reflected by the PDE, which then has more than one time scale. The relatively smooth part is related to the larger time scales and the highly active part is related to the smaller time scales. These high-activity zones are called layers; they often occur at boundaries of the domain (boundary layers) but sometimes also at the interior of the domain (interior layers). A typical, although academic, example is the initial value problem

$$\frac{du}{dt} = -100\left(u - \cos t + \frac{1}{100}\sin t\right),$$

(5.83a)

$$u(0) = 0.$$  

(5.83b)
5.4. Difference Methods for Initial Value Problems

The solution (5.84) on the intervals (0, 1) (left) and (0, 0.02) (right).

Its solution is easily seen to be

\[ u(t) = \cos t - e^{-100t}. \]  

In Figure 5.6 we give a graph of \( u \) and an enlarged part of the interval (0, 0.02).

Clearly, \( u \) exhibits a boundary layer of about \( 10^{-2} \) thickness; the time scales differ by a factor of 100. It is important to realize that the notion of time scale is a relative one. For if we stretched the time via

\[ \tau := 100t, \]

then the differential equation (5.83a) could be rewritten as

\[ \frac{du}{d\tau} = -u + \cos(0.01\tau) - 0.01\sin(0.01\tau). \]

Its solution,

\[ u(\tau) = \cos(0.01\tau) - e^{-\tau}, \]

does not look like one with a highly active region. However, it is clear that the actual behaviour of the solution for larger values of \( \tau \) should not be measured on the unit scale, but rather on one a factor of 100 larger. In this book we will often encounter such problems after nondimensionalising the equations, as parameters like the factor 100 in (5.83a) show up naturally. In fact, one may even rewrite (5.83a) as

\[ \frac{1}{100} \frac{du}{dt} = -u + \cos t - \frac{1}{100} \sin t \]

to see that \( \frac{1}{100} \frac{du}{dt} \cong 0 \) outside the boundary layer. Taking this idea even further, we can rewrite the equation as

\[ \varepsilon \left( \frac{du}{dt} + \sin t \right) = -u + \cos t, \]

where \( \varepsilon = 0.01 \). If we take \( \varepsilon = 0 \) in (5.87), we obtain the reduced equation

\[ -u(t) + \cos t = 0. \]

Outside the layer the solution is accurate within a bound \( \varepsilon \).
The observations above have important numerical implications. We only need to compute the solution with smaller time scales in the boundary layer. Outside this layer the solution satisfies a simpler problem, so the larger time scales apply. In terms of difference methods this can be translated to smaller values of \( \Delta t \) in the layer and larger values outside the layer, at least from the viewpoint of approximation (i.e., local discretisation) errors. However, let us look at the next example.

**Example 5.11** If we use the forward Euler method to discretise (5.83a), we find

\[
\begin{align*}
  u^{n+1} &= u^n - 100\Delta t \left( u^n - \cos r^n + \frac{1}{100} \sin r^n \right) \\
  &= (1 - 100\Delta t)u^n + \Delta t (100 \cos r^n - \sin r^n).
\end{align*}
\]

Suppose we want to have a local error smaller than \( 10^{-2} \) in absolute value. Since the local error is given by \( \frac{1}{2} \Delta t u''(\tau^n) \) (cf. Example 5.8), we can estimate a value for \( \Delta t \), sufficiently small to resolve the layer, from the inequality

\[
\frac{1}{2} \Delta t 10^4 e^{-100t} \leq 10^{-2}.
\]

This implies a time step restriction

\[
\Delta t \leq 2 \times 10^{-6} e^{100t}.
\]

Hence for \( t \) close to zero this would require steps as small as \( 2 \times 10^{-6} \). Outside the layer we have \( u(t) \approx \cos t \), whence a guaranteed step size \( \Delta t \) there is estimated from

\[
\frac{1}{2} \Delta t \leq 10^{-2} \Rightarrow \Delta t \leq 2 \times 10^{-2}.
\]

In order to see if the latter finding is correct, we consider a variant of problem (5.83), i.e., where we replace (5.83b) by the initial value \( u(0) = 1 \). As one can easily see, there is no boundary layer then. So a step size \( \Delta t = 2 \times 10^{-2} \) seems reasonable to find an approximation of the solution with an absolute accuracy of \( 10^{-2} \). However, applying of the explicit Euler method to this problem might lead to unstable numerical solutions, as shown in Figure 5.7.

![Figure 5.7](image_url)

**Figure 5.7.** Exact and numerical solutions of (5.83a) with initial condition \( u(0) = 1 \). The numerical solution is computed with the forward Euler method with \( \Delta t = 1/45 \).

As the previous example shows, accuracy of the approximation is not enough for the method to be effective. It is easy to see that there are other requirements. To find this...
5.4. Difference Methods for Initial Value Problems

out let us consider two distinct initial values for the forward Euler method discussed in Example 5.11:

\begin{align}
    u^0 &= 1, \quad (5.89a) \\
    \tilde{u}^0 &= 1 + \delta, \quad (5.89b)
\end{align}

with \(0 < \delta \ll 1\). Clearly, (5.89a) and (5.89b) generate sequences \(\{u^n\}\) and \(\{\tilde{u}^n\}\), respectively. We now find

\[\tilde{u}^{n+1} - u^{n+1} = (1 - 100\Delta t)(\tilde{u}^n - u^n).\]

(5.90)

Hence

\[\tilde{u}^n - u^n = (1 - 100\Delta t)^n \delta.\]

If \(\Delta t > 2 \times 10^{-2}\), we see that \(\{\tilde{u}^n - u^n\}\) becomes unbounded. A necessary stability requirement is therefore

\[|1 - 100\Delta t| \leq 1.\]

(5.91)

The phenomenon that accuracy requirements following from smoothness of the solutions allow for larger time steps, whereas stability arguments, such as in the forward Euler method, require smaller time steps, relates to the stiffness of the problem. In general, a problem is called stiff if the actual time scale is (much) larger than small time scales potentially existing in the problem. In this sense the problem is not stiff in the boundary layer, only outside the layer.

Fortunately, not all difference methods have a problem with stiffness. Indeed, consider the implicit Euler method applied to (5.83a):

\[u^{n+1} = u^n - 100\Delta t \left( u^{n+1} - \cos t^{n+1} + \frac{1}{100} \sin t^{n+1} \right).\]

(5.92)

If we use the same initial values as in (5.89), we now find

\[\tilde{u}^{n+1} - u^{n+1} = \tilde{u}^n - u^n - 100\Delta t(\tilde{u}^{n+1} - u^{n+1}),\]

(5.93)

whence

\[\tilde{u}^{n+1} - u^{n+1} = \frac{1}{1 + 100\Delta t}(\tilde{u}^n - u^n).\]

(5.94)

Obviously, \(|\frac{1}{1 + 100\Delta t}| < 1\) for all \(\Delta t > 0\), so no stability constraint applies.

**Example 5.12** If we employ the \(\vartheta\) method (5.72), we find that the stability requirement for (5.83a) is

\[\left|\frac{1 - 100(1 - \vartheta)\Delta t}{1 + 100\vartheta \Delta t}\right| \leq 1.\]

From this inequality we can easily derive the time step restriction \((1 - 2\vartheta)\Delta t \leq 0.02\), implying that the method is unconditionally stable for \(0.5 \leq \vartheta \leq 1\).

To conclude this section we remark that only implicit schemes (like backward Euler) can handle stiffness. Yet not all implicit schemes have this property. There exist special Runge–Kutta schemes, but most widely used are still the BDF (multistep) methods. For more details see [58].
Chapter 5. Approximation by Finite Differences

5.5 Discretisation, Convergence, and Stability

The difference methods outlined so far for space and time variables separately have to be combined to obtain a numerical method for approximating solutions of time-dependent PDEs. This brings up the question of how accurate the actual approximations are. This is not only a matter of choosing a proper mesh, i.e., commensurate with the “activity” of the solution. Indeed, for time-dependent problems there is also the question of ensuring that the method has suitable numerical stability properties.

5.5.1 The Method of Lines

A general time-dependent (scalar) PDE can be written as

\[ A(t)[u] = B(x)[u] + c(x, t, u), \]

where \( A(t) \) is a differential operator in \( t \) and \( B(x) \) is a differential operator in \( x \), whereas \( c \) does not contain any derivatives. Typically, \( A(t) \) is either \( \frac{\partial}{\partial t} \) or \( \frac{\partial^2}{\partial t^2} \). If we denote the relevant space-time domain (where we would like to compute the solution) by \( \Omega \times I \) with \( \Omega \subset \mathbb{R}^d \) and \( I \subset [0, \infty) \), we basically have two options. The first one is to discretise \( B(x) \) by finite differences (directly or through a finite volume method). To do this we have to find a grid on \( \Omega \); this discretisation makes the resulting (semidiscrete) problem finite dimensional. From this we obtain a relation between (numerical) solution values on that grid. If we consider this grid as fixed on \( \Omega \), we thus find from (5.95) an initial value ODE for solution values on the grid, coupled through the finite difference scheme. In a one-dimensional setting we typically have a situation like that in Figure 5.8.

The lines there represent the domains of the thus obtained (time-dependent) variables. This discretisation is therefore called the method of lines (MOL); more accurately this is actually the longitudinal MOL. The second option is to discretise the time derivative first, and this approach is called the transversal MOL; see Figure 5.8. Here we have a boundary value problem (PDE or ODE) at discrete time levels. These are coupled through the finite

Figure 5.8. Grids for the longitudinal (left) and transversal (right) method of lines.
differences of the time derivative(s). In particular, for parabolic problems this transversal MOL leads to a sequence of elliptic problems.

Both types of MOL have advantages and disadvantages. The longitudinal is the simplest and therefore most attractive, as it transforms the PDE to an ODE, which lends itself to easy computation. However, one should realize that a fixed (let alone a uniform) mesh for all \( t \in I \) is not likely to give good results if the high activity of the solution is moving in time. If a good boundary value problem solver is available, then the transversal MOL seems attractive for the latter type of problems. However, changing meshes from one time step to the next is not trivial here either. In practice a hybrid MOL is often used. We shall base our derivations in this book mainly on uniform fixed meshes, used in the longitudinal MOL. When needed we shall use more sophisticated gridding.

### 5.5.2 Consistency, Stability, and Convergence of the MOL

After discretising the spatial derivatives in (5.95), we end up with a vectorial system of equations

\[
A(t)[u] = Bu + f,
\]

where \( u(t) \in \mathbb{R}^M \), with \( M \) the number of grid points, is the vector of unknown values and \( Bu \) is the result of discretising \( B(x)[u] \). We consider the case where \( A(t) \) is either \( \frac{\partial}{\partial t} \) or \( \frac{\partial^2}{\partial t^2} \), so that we actually have the ODE

\[
\frac{du}{dt} = Bu + f
\]

or

\[
\frac{d^2u}{dt^2} = Bu + f.
\]

As is standard procedure for ODEs, one may reduce (5.98) to one of type (5.97) by introducing a new variable \( v \), defined as

\[
v := \begin{pmatrix} u \\ \frac{du}{dt} \end{pmatrix}.
\]

In the remainder of this section we will consider (5.97). Although this looks like an innocuous system of ODEs, one has to bear in mind that for a convergence analysis the dimension \( M \) goes to infinity as the step size \( \Delta x \) approaches zero. A rigorous analysis of the numerical integration of (5.97) will therefore be complicated. Below we shall be satisfied with a finite-dimensional convergence analysis nevertheless, keeping an eye on the infinite-dimensional limit case (for a more functional analytic approach see, e.g., [120]).

First we need a way to measure errors such that they are independent of the dimension \( M \). This can be achieved by using Hölder norms; see the appendix, section G. In particular,
we shall use for \( \mathbf{u} = (u_1, u_2, \ldots, u_M)^T \) one of the norms

\[
\| \mathbf{u} \|_\infty := \max_{1 \leq i \leq M} |u_i| \quad \text{(infinity norm),} \tag{5.100a}
\]
\[
\| \mathbf{u} \|_2 := \left( \sum_{i=1}^{M} |u_i|^2 \right)^{\frac{1}{2}} \quad \text{(2-norm).} \tag{5.100b}
\]

One easily sees that the length of a vector depends on the norm used. Rather than (5.100b), we may use the norm

\[
\| \mathbf{u} \|_{2,m} := \frac{1}{\sqrt{M}} \left( \sum_{i=1}^{M} |u_i|^2 \right)^{\frac{1}{2}} \quad \text{(median 2-norm).} \tag{5.100c}
\]

Note that, e.g., for \( \mathbf{e} := (1, 1, \ldots, 1)^T \) we have \( \| \mathbf{e} \|_2 = \sqrt{M} \) and \( \| \mathbf{e} \|_{2,m} = 1 \).

Let us now assume the PDE to be linear with \( c = c(x, t) \). Then a typical one-step method results in, say,

\[
\frac{1}{\Delta t} (u^{n+1} - u^n) = R^n u^{n+1} + S^n u^n + f^{n,n+1}. \tag{5.101a}
\]

Assuming that \( \mathbf{I} - \Delta t \mathbf{R}^n \) is nonsingular, we have

\[
u^{n+1} = (\mathbf{I} - \Delta t \mathbf{R}^n)^{-1} \left( (\mathbf{I} + \Delta t \mathbf{S}^n) u^n + f^{n,n+1} \right). \tag{5.101b}
\]

Here \( \mathbf{R}^n \) and \( \mathbf{S}^n \) are \( M \times M \) matrices and \( f^{n,n+1} \in \mathbb{R}^M \). Note that, as before, the superscript \( n \) denotes a time index, not a power, whereas the double superscript is a reminder that the source term arose from two time level contributions. If we denote the exact solution of (5.95) restricted to the grid by \( \mathbf{u}(\cdot, t^n) \), i.e., \( \mathbf{u}(\cdot, t^n) := (u(x_1, t^n), u(x_2, t^n), \ldots, u(x_M, t^n))^T \), then we can define the local discretisation error (vector) \( \mathbf{d}(\cdot, t^n) \) by

\[
\mathbf{d}(\cdot, t^n) := \frac{1}{\Delta t} (\mathbf{u}(\cdot, t^{n+1}) - \mathbf{u}(\cdot, t^n)) - R^n \mathbf{u}(\cdot, t^{n+1}) - S^n \mathbf{u}(\cdot, t^n) - f^{n,n+1}. \tag{5.102}
\]

Now let the time step \( \Delta t \) be related to \( \Delta x \) by \( \Delta t = \varphi(\Delta x) \), where we require \( \varphi(\Delta x) \to 0 \) for \( \Delta x \to 0 \). We shall refer to the latter procedure as a refinement path. The analysis below will be valid on a finite time interval \([0, T]\), say. Then we have the following definition.

**Definition 5.13.** The difference scheme (5.101) is consistent for \( \Delta t = \varphi(\Delta x) \) if \( \| \mathbf{d}(\cdot, t^n) \| \to 0 \) for \( \Delta x \to 0 \) and for all time levels \( t^n \leq T \).

One should note that this consistency definition requires the (scalar) local discretisation errors to go to zero if \( \Delta t, \Delta x \to 0 \) (the finite difference methods we have constructed so far have this property automatically); however, it also requires a proper handling of the initial and boundary conditions.

**Example 5.14** If we have to solve a second order problem like in (5.98), then we clearly have to specify two initial conditions at \( t = 0 \). Often this is a given value of the solution and its derivative. The latter should be discretised with a similar accuracy as the time-stepping method to prevent the initial errors from being persistent.
5.5. Discretisation, Convergence, and Stability

The next property we need deals with appropriate error propagation. Recalling the scalar situation (5.90), we therefore consider (cf. (5.101))

$$(I - \Delta t R')^{-1}(I + \Delta t S').$$

Let us now consider system (5.97) defined on some time interval $[0, T]$. We then have the following definition.

**Definition 5.15.** The difference scheme (5.101) is called stable for $\Delta t = \varphi(\Delta x)$ if, for some $\varepsilon$ and $\Delta x < \varepsilon$, there exists a constant $C(\varepsilon)$ such that, for all $k, n$ with $k \leq n$ and $n \Delta t \leq T$,

$$\left\| \prod_{i=k}^{n} (I - \Delta t R')^{-1}(I + \Delta t S') \right\| \leq C(\varepsilon).$$

The latter concept needs some further explanation. We give this through an example.

**Example 5.16** If $R' = R$ and $S' = S$ for all $i$, we may write

$$(I - \Delta t R)^{-1}(I + \Delta t S) = TJ(\Delta t)T^{-1},$$

where $J(\Delta t)$ is the Jordan normal form; see the appendix, section H. Hence

$$\prod_{i=k}^{n} (I - \Delta t R)^{-1}(I + \Delta t S) = T(J(\Delta t))^{n-k+1}T^{-1}.$$ 

So the required bound exists if and only if $\|J(\Delta t))^{n-k+1}\|$ is uniformly bounded; this is the case if and only if $J(\Delta t)$ has eigenvalues smaller than one in modulus and/or unimodular eigenvalues that are geometrically simple; see the appendix, section I. Note that such a property may hold, even if $\|I - \Delta t R\| > 1$, for the 2-norm $\| \cdot \|_2$ or the infinity norm $\| \cdot \|_{\infty}$. 

Finally, we introduce a convergence concept. To this end we define the *global discretisation error* (vector) $e(\cdot, t^n)$ as

$$e(\cdot, t^n) := u(\cdot, t^n) - u^n. \quad (5.103)$$

This has to be understood with $e$ and $u$ being vectors with the appropriate dimension. For a first order time-dependent problem it is trivial to assume that $e(\cdot, t^0) = 0$. For second order problems this is generally not true. We simply assume the starting error to be small enough and disregard it from now on. As a further (but trivial) simplification we let $t^0 = 0$. We now have the following definition.

**Definition 5.17.** The difference scheme (5.101) is convergent for $\Delta t = \varphi(\Delta x)$ if the following holds: for any $c(x, t, u)$ and initial value $u(x, t^*)$ ($0 \leq t^* \leq T$) for which (5.95) has a bounded solution, $\|e(\cdot, t^n)\| \to 0$ for $\Delta x \to 0$ and $n \Delta t \leq T$.

Again we remark that convergence for a PDE is much more complicated than for an ODE, the major problem being the dimension $M$ approaching infinity. The concepts of
consistency and stability are, for many problem classes, equivalent. This is the famous Lax theorem [120], of which we give a proof for the restricted setting sketched above.

**Theorem 5.18 (Lax equivalence theorem).** Let (5.101) be consistent. Then it is convergent if and only if it is stable.

**Proof.** Suppose (5.101) is stable. Then we first have to prove that it is convergent. From (5.101) and (5.102) we obtain

\[(I - \Delta t R^n) e(\cdot, t^{n+1}) = (I + \Delta t S^n) e(\cdot, t^n) + \Delta t d(\cdot, t^n).\]

Since \(e(\cdot, t^n) = 0\), we thus find

\[e(\cdot, t^n) = \Delta t \sum_{k=0}^{n-1} \prod_{i=k+1}^{n} (I - \Delta t R^i)^{-1} (I + \Delta t S^i) (I - \Delta t R^k)^{-1} d(\cdot, t^k),\]

where the matrix product is defined as \(\prod_{i=k+1}^{n} C^i := C^{n-1} \cdots C^{k+2} C^{k+1}\). Hence we obtain

\[\|e(\cdot, t^n)\| \leq \Delta t \|C(\epsilon)\| \sum_{k=0}^{n-1} \|(I - \Delta t R^k)^{-1}\| \|d(\cdot, t^k)\| \leq C(\epsilon) DT \max_k \|d(\cdot, t^k)\| .\]

Here \(C(\epsilon)\) is the stability bound as implied by Definition 5.15 and \(D := \max_k \|(I - \Delta t R^k)^{-1}\|\). In the applications that follow the latter bound is trivial; often even \(D = 1\). The convergence then follows from the consistency.

Conversely, suppose (5.101) is convergent. Then we have to prove that it is stable. We first remark that it is not restrictive to assume \(c(x, t, u) \equiv 0\), and consequently \(f \equiv 0\), and moreover we assume that \(u^k = u(\cdot, t^k)\) for all \(k \leq n \leq T/\Delta t\). The latter assumption implies that we may look at problem (5.95) on the interval \([t^k, T]\) with \(k \Delta t \leq T\). Under these assumptions we have for any “initial” value \(u(\cdot, t^0)\) that

\[\|u(\cdot, t^n) - u^k\| = \left\| \prod_{k=0}^{n-1} (I - \Delta t R^i)^{-1} (I + \Delta t S^i) u(\cdot, t^k) - u(\cdot, t^n) \right\| .\]

Since (5.101) is convergent, the left-hand side is bounded by some value \(\epsilon(\Delta x)\), with \(\epsilon(\Delta x) \to 0\) if \(\Delta x < \delta x\) and \(\delta x \to 0\). Moreover, on a finite interval \([0, T]\), \(\|u(\cdot, t^n)\|\) must be bounded by a constant \(E\), say. Applying the triangle inequality, we thus obtain

\[\left\| \prod_{k=0}^{n-1} (I - \Delta t R^i)^{-1} (I + \Delta t S^i) u(\cdot, t^k) \right\| \leq E + \epsilon\]

for \(\Delta x \leq \delta x\). Hence for \(M\) (dimension of vector space) fixed we find that the set

\[\left\{ \prod_{k=0}^{n-1} (I - \Delta t R^i)^{-1} (I + \Delta t S^i) u(\cdot, t^k) \right\} \]
is bounded for \( k, n \leq T/\Delta t \) and thus
\[
\left\| \prod_{i=k}^{n-1} (I - \Delta t R_i)^{-1} (I + \Delta t S_i) \right\|
\]
is bounded for \( k, n \leq T/\Delta t \); i.e., we have stability. \(\square\)

**Example 5.19** Let \( R_i \equiv R \) and \( S_i \equiv S \), where \( R \) and \( S \) commute and have simple eigenvalues, say equal to \( r_1, r_2, \ldots, r_M \) and \( s_1, s_2, \ldots, s_M \), respectively. Then it is easily verified that
\[
(I - \Delta t R_i)^{-1} (I + \Delta t S_i) = T A(\Delta t) T^{-1},
\]
where \( A(\Delta t) := \text{diag} \left( \frac{1 + \Delta t s_1}{1 - \Delta t r_1}, \frac{1 + \Delta t s_2}{1 - \Delta t r_2}, \ldots, \frac{1 + \Delta t s_M}{1 - \Delta t r_M} \right) \) is a diagonalising matrix \( T \). Thus we have
\[
\prod_{i=k}^{n} (I - \Delta t R_i)^{-1} (I + \Delta t S_i) = T (A(\Delta t))^{n+1} T^{-1},
\]
and, consequently, stability (and thus convergence) follows if
\[
\max_i \left| \frac{1 + \Delta t s_i}{1 - \Delta t r_i} \right| \leq 1.
\]
Clearly, the eigenvalues \( \{s_i\} \) and \( \{r_i\} \) depend on the spatial discretisation. \(\square\)

### 5.6 Fourier Mode Analysis

The method outlined in the previous section is often too heavy a tool for establishing stability results. Indeed, as we saw in Example 5.19, one may need quantities like eigenvalues of a matrix, which are hard to get. One has to remember that the bulk of the matrix elements in \( R_i \) and \( S_i \) are formed by applying a difference scheme locally. Hence it is tempting to look at the latter exclusively in order to find out about stability (and how the refinement path \( \Delta t = \varphi(\Delta x) \) should be chosen, for that matter). This should then give at least necessary conditions. In order to mathematically justify this approach, we may restrict ourselves to one-dimensional problems with periodic boundary conditions. The latter is at least convenient for the operator notation that will follow; however, we shall not aim at a complete derivation, given the restrictive (yet extremely useful!) nature of the setting. The basic idea is to view the difference method as an operator defined on the discrete set of points \( \{x_1, x_2, \ldots, x_M\} \). So, rather than deriving a matrix system like (5.96) (which includes the boundary conditions as well) we investigate the discrete analogue of (5.95). To do this let us introduce the notation for a function defined on the grid \( x_1, x_2, \ldots, x_M \):
\[
u_j^n := \{\nu_j^n\}_{j=1}^M, \quad (5.104)
\]
where \( \nu_j^n \equiv u(x_j, t^n) \). Apparently, \( \nu_j^n : \{1, 2, \ldots, M\} \to \mathbb{R} \), which is the analogue of \( u(\cdot, t^n) : \Omega \to \mathbb{R} \). If we set \( c(x, t, u) \equiv 0 \) in (5.95) and apply a one-step method, we may write the result as
\[
A_{\Delta} [\nu_j^{n+1}] = B_{\Delta} [\nu_j^n], \quad (5.105)
\]
where \( A_{\Delta} \) and \( B_{\Delta} \) denote linear (difference) operators.
Example 5.20 Consider the equation \( \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0 \). Obviously, \( A(t) = \frac{\partial}{\partial t} \), \( B(x) = \frac{\partial}{\partial x} \), and \( c(x, t, u) = 0 \). Applying forward differences yields

\[
\frac{\partial u}{\partial t}(x_j, t^n) \approx \frac{1}{\Delta t}(u_{j+1}^n - u_j^n), \quad \frac{\partial u}{\partial x}(x_j, t^n) \approx \frac{1}{\Delta x}(u_{j+1}^n - u_j^n).
\]

Hence

\[
A\Delta[u_j^{n+1}] := \frac{1}{\Delta t}u_j^{n+1}, \quad B\Delta[u_j^n] := \frac{1}{\Delta t}u_j^n + \frac{1}{\Delta x}(u_{j+1}^n - u_j^n).
\]

The indices \( j \) and \( n \) are generic, just like \( x \) and \( t \). Note that \( A\Delta \) and \( B\Delta \) are determined up to a scaling factor. We could, for example, have multiplied both operators by \( \Delta t \). However, the above representation has the same dimension as the corresponding PDE.

We shall need norms for operators like \( A\Delta \) and \( B\Delta \). Anticipating the use of Fourier series, we first recall the 2-norm for \( u/\Delta \) as

\[
\|u/\Delta\|_2 := \left( \sum_{j=1}^{M} u_j^2 \right)^{\frac{1}{2}}.
\]

We can then define the 2-norm for an operator like \( A\Delta \) as

\[
\|A\Delta\|_2 := \max_{u_\Delta \neq 0} \frac{\|A\Delta u_\Delta\|_2}{\|u_\Delta\|_2} = \max_{u_\Delta \neq 0} \left( \frac{\sum_{j=1}^{M} (A\Delta[u_j])^2}{(\sum_{j=1}^{M} u_j^2)^2} \right)^{\frac{1}{2}}.
\]

Example 5.21 For the choices in Example 5.20 we find

\[
\|A\Delta^{-1}B\Delta\|_2 := \max_{u_\Delta \neq 0} \frac{\|A\Delta^{-1}B\Delta u_\Delta\|_2}{\|u_\Delta\|_2} = \max_{u_\Delta \neq 0} \left( \frac{\sum_{j=1}^{M} \left( u_j + \frac{\Delta t}{\Delta x}(u_{j+1} - u_j) \right)^2}{M \sum_{j=1}^{M} u_j^2} \right)^{\frac{1}{2}}.
\]

Define \( \sigma := \frac{\Delta t}{\Delta x} \) and let \( 0 < \sigma \leq 1 \). Then we have, from the triangle inequality for the 2-norm applied to the numerator, that

\[
\sum_{j=1}^{M} ((1 - \sigma)u_j + \sigma u_{j+1})^2 \leq \left[ \left( \sum_{j=1}^{M} (1 - \sigma)^2 u_j^2 \right)^{\frac{1}{2}} + \left( \sum_{j=1}^{M} \sigma^2 u_{j+1}^2 \right)^{\frac{1}{2}} \right]^2 = \left( 1 - \sigma + \sigma \right) \left( \sum_{j=1}^{M} u_j^2 \right)^{\frac{1}{2}} = \sum_{j=1}^{M} u_j^2.
\]

The equality sign is taken for \( u_j \equiv 1 \). Hence we obtain \( \|A\Delta^{-1}B\Delta\|_2 = 1 \). Note that we have used periodicity.

We now obtain from (5.105) that

\[
u_j^{n+1} = (A\Delta^{-1}B\Delta)[u_j^n].\]
5.6. Fourier Mode Analysis

Hence
\[
\frac{\|u_\Delta^{n+1}\|_2}{\|u_\Delta^n\|_2} \leq \left\| \left( A_\Delta^{-1}B_\Delta \right)^i \right\|_2. \tag{5.108}
\]

As a consequence, we have
\[
\sup_{\epsilon_\Delta \neq 0} \frac{\|u_\Delta^{n+1}\|}{\|u_\Delta^n\|} \leq \left\| A_\Delta^{-1}B_\Delta \right\|_2. \tag{5.109}
\]

Stability of (5.105) is therefore seen to be equivalent to
\[
\left\{ \left\| \left( A_\Delta^{-1}B_\Delta \right)^i \right\|_2 \right\}_{i \geq 1}
\]
being bounded. Like in Section 5.5.2, this has to be interpreted in terms of a suitable refinement path, say \( \epsilon = \psi(\Delta x) \), and for \( \Delta x \) small enough.

A useful tool in estimating growth is the discrete Fourier transform (DFT) (see Chapter 3). Recall that \( e^{2\pi i j/(M-1)} \), where \( j/(M-1) = x_j \), represents a typical mode. We shall apply the DFT to the solution \( u_\Delta^n \), which is defined at the grid points \( x_1, x_2, \ldots, x_M \). Then we can write
\[
u^n_j = \sum_{k=1}^M \hat{u}_k^n e^{2\pi i k j/(M-1)}, \tag{5.110a}
\]
where
\[
\hat{u}_k^n := \sum_{j=1}^M u^n_j e^{-2\pi i k j/(M-1)}. \tag{5.110b}
\]

Clearly, in (5.110a) and (5.110b) the periodicity of \( u_\Delta^n \) is tacitly assumed (which should carry over from \( u(\cdot, t^n) \)). We can use the expansion (5.110a) in (5.105), giving
\[
\sum_{k=1}^M \hat{u}_k^{n+1} e^{2\pi i k j/(M-1)} = \sum_{k=1}^M \hat{u}_k^n \left( A_\Delta^{-1}B_\Delta \right) \left[ e^{2\pi i k j/(M-1)} \right]. \tag{5.111}
\]

If we collect all terms containing the factor \( e^{2\pi i k j/(M-1)} \) in (5.111), we can write this as
\[
\hat{u}_k^{n+1} = \hat{G}_k \hat{u}_k^n. \tag{5.112}
\]

The quantity \( \hat{G}_k \) is called the growth factor (which does not depend on \( n \), apparently).

**Example 5.22** If we apply this idea to the operators \( A_\Delta \) and \( B_\Delta \) from Example 5.20, we obtain
\[
A_\Delta \left[ e^{2\pi i k j/(M-1)} \right] = \frac{1}{M} e^{2\pi i k j/(M-1)},
\]
\[
B_\Delta \left[ e^{2\pi i k j/(M-1)} \right] = \frac{1}{M} e^{2\pi i k j/(M-1)} + \frac{1}{\Delta x} \left( e^{2\pi i (j+1) k/(M-1)} - e^{2\pi i j k/(M-1)} \right).
\]

Consequently, we find for the growth factor
\[
\hat{G}_k = e^{-2\pi i k j/(M-1)} \left( A_\Delta^{-1}B_\Delta \right) \left[ e^{2\pi i k j/(M-1)} \right] = 1 + \frac{\Delta t}{\Delta x} \left( e^{2\pi i j (k+1)/(M-1)} - e^{2\pi i j k/(M-1)} \right). \tag{5.113}
\]
The usefulness of this growth factor is shown by the following. We apparently have

\[ u_{n+1}^{j} = \sum_{k=1}^{M} \hat{G}_{k} e^{2\pi ikj/(M-1)} \]

Hence we deduce from Parseval's identity (cf. Chapter 3) that

\[ \| u_{n+1}^{j} \|^{2} = \sum_{k=1}^{M} (|G_{k}|^{2} \| \hat{u}_{k}^{n} \|^{2}) \leq \max_{k} |G_{k}|^{2} \sum_{k=1}^{M} |\hat{u}_{k}^{n}|^{2} = \max_{k} |G_{k}|^{2} \| u_{n}^{n} \|^{2}. \]

Equality in (5.114) can be obtained if all \( |G_{k}| \) are equal. Hence we conclude that

\[ \sup_{u_{n}^{n}} \left\{ \max_{k} |G_{k}|^{2} \right\} \text{to be bounded. This leads to the following property.} \]

**Property 5.23 (von Neumann).** A necessary condition for (5.104) to be stable on \([0, T]\) is that

\[ |G_{k}| \leq 1 + O(\Delta t) \]  

uniformly in \( k \).

**Proof.** Let \( |G_{k}| \leq 1 + c \Delta t \) for some \( c > 0 \). Then

\[ |G_{k}|^{2} \leq (1 + c \Delta t)^{2} \leq e^{c \Delta t} \leq e^{c T}, \]

which is bounded for every finite \( T \). \( \square \)

Property 5.23 will often turn out to be the most convenient way to find at least necessary conditions for (5.105) to be stable. One should recall, however, that our setting was quite restricted anyway (linear, constant coefficients, periodic boundary conditions). Hence we may use this property as a test to reject certain methods (and refinement paths \( \Delta t = \phi(\Delta x) \)), as a method should at least be meaningful for a member of the above-mentioned restricted class of problems.
5.7 Discussion

- Finite differences are the oldest and simplest methods of finding approximate solutions of PDEs. Therefore there exist a host of books on the subject. It is interesting to see that finite volume methods lead to finite difference methods as well. Finite element methods (FEMs) can often be formulated such that they lead to difference schemes; see [82]. In particular, for complicated domains (in two or three space dimensions), however, the finite difference schemes may result in complicated bookkeeping that FEMs do not encounter. Bearing in mind that there is much similarity, finite differences are the most transparent methods of understanding typical phenomena like stability, etc., of numerical methods, and therefore they are the method of choice.

- Implicit time integrators that address stiffness may look very attractive at first sight. There exist many such methods, in particular implicit Runge–Kutta methods that are not dealt with here; see, e.g., [58, 21]. They come with a price, however. Indeed, implicit methods require the solution of a (possibly nonlinear) system. This may lead to such complexity that despite the relatively small step required by an explicit method, the overall complexity is still lower.

- An important question is how to find an appropriate mesh in general, both for spatial and for temporal discretisation. The simple approach of a uniform mesh requires it to be small enough to cope with the “activity” (in terms of magnitude of derivatives) of the solution. Clearly, for nonuniform behaviour of the solution this is not efficient. There are all sorts of adaptive methods to cope with this problem. For time discretisation this is relatively easy; see, e.g., [97, 57]. For spatial discretisation one often employs equidistribution techniques; see [3, 11].

- The Lax theorem proved in this chapter has a more universal character. In many instances one can show that for a consistent scheme stability is equivalent to convergence; see [148]. The statement and proof in this chapter are for scalar one-step methods only. For multistep methods the amplification factor becomes an amplification matrix. The estimates then typically should hold for eigenvalues of the matrix.

Exercises

5.1. Although numerical derivatives are more accurate the smaller the grid size chosen, this is only true in exact arithmetic. In practice we have to deal with rounding errors as well. This exercise deals with this problem.

(a) Let the function \( f(\xi) \) be perturbed by some \( \delta(\xi) \), where \( |\delta(\xi)| \leq \varepsilon \), for some \( \varepsilon \) and \( \xi \in I \), \( I \) being some relevant interval. Show that a finite difference approximation of \( f'(\xi_1) \), \( \xi_1 \in I \), has a rounding error approximately proportional to \( \varepsilon/h \), where \( h \) is the grid size.

(b) If we use a \( k \)th order approximation for \( f' \), then we have an approximately optimal grid size (i.e., the sum of the absolute approximation error and the rounding error
is minimal) if \( h \approx \varepsilon^{1/(k+1)} \). Note that if, e.g., \( k = 1 \) (as we have for a forward difference), this result means that the last half of the digits in the mantissa of the result may not be significant.

5.2. Derive a Runge–Kutta method by taking the trapezoidal rule as the method of integration and the Euler forward method to approximate the implicit part on the right-hand side.

5.3. A way to determine a finite difference approximation is to impose that it is exact for polynomials of a degree as large as possible. As an example we consider the two-step backward difference

\[
a_1 u(\xi_1) + a_2 u(\xi_2) + a_3 u(\xi_3) \quad (\xi_1 > \xi_2 > \xi_3)
\]
as an approximation of \( \frac{du}{d\xi}(\xi_1) \).

(a) Why can you ensure this to be exact for \( u(\xi) := 1, u(\xi) := \xi, \) and \( u(\xi) := \xi^2 \)?
(b) Find the coefficients \( a_1, a_2, \) and \( a_3 \).
(c) Try the formula for \( u(\xi) := \xi^3 \) and see that it is not differentiated exactly. Assume then that the error is of the form \( cu^{(q)}(\zeta) \) and determine \( c, q, \) and \( \zeta \).

5.4. To see that the error term in integration is of the form (5.12), we consider the case of the trapezoidal rule; see Example 5.4. So consider the interval \( (\xi_1, \xi_2) \). The error is given by

\[
\int_{\xi_1}^{\xi_2} \frac{1}{2} u^{(2)}(\xi) (\xi - \xi_1)(\xi - \xi_2) \, d\xi.
\]

The mean value theorem says that \( \int_{\xi_1}^{\xi_2} f(\xi) g(\xi) \, d\xi = f(\eta) \int_{\xi_1}^{\xi_2} g(\xi) \, d\xi \) for some \( \eta \in (\xi_1, \xi_2) \) if \( g \) has a constant sign. Now prove the error term given in Example 5.4.

5.5. The weights and error term of a quadrature formula can be determined by requiring that it be exact for polynomials up to a certain degree. As an example we consider the integral

\[
\int_{-h}^{h} u(\xi) \, d\xi,
\]

which we would like to approximate by the three-point formula

\[
b_1 u(-h) + b_2 u(0) + b_3 u(h).
\]

We assume that the error term is of the form \( cu^{(q)}(\xi) \).

(a) Show that \( b_1, b_2, \) and \( b_3 \) can be chosen such that the formula is exact for polynomials of degree up to two.
(b) Show that this formula is also exact for any odd function. Thus it is also exact for \( u(\xi) = \xi^3 \).
(c) Determine the error term.

5.6. The approximations of the second derivatives in (5.32a) and (5.32b) are second order. Show this by using Taylor expansions.

5.7. Let \( \Omega \) be the trapezium defined by the corner points \((0,0), (0,2), (2,1), \) and \((2,0)\).

We want to discretise the Poisson equation

\[
\nabla^2 u = f, \quad (x, y) \in \Omega.
\]
Let the grid size be given by \( h = \Delta x = \Delta y = 2/M \) for some \( M \).

(a) Assume that Dirichlet boundary conditions are given. Derive the scheme at a point neighbouring the part of the boundary connecting \((0, 2)\) and \((2, 1)\).

(b) Repeat part (a) for Neumann boundary conditions.

5.8. Consider the difference scheme (5.42).

(a) Give the stencil.

(b) Show that it is second order accurate.

5.9. Show that the finite volume scheme (5.53) is second order accurate.

5.10. The multistep method (5.75) is called root stable if the scheme with \( f(u, t) \equiv 0 \) is a stable recursion, i.e., all roots of the characteristic polynomial

\[
\sum_{j=0}^{k} a_j x^{k-j}
\]

are smaller than one (in absolute value) or equal to one and simple. Now let the multistep method not be root stable. If we apply this method to the ODE

\[
\frac{dx}{dt} = f(t),
\]

then show that the difference equation

\[
\frac{1}{\Delta t} \sum_{j=0}^{k} a_j x^{n+1-j} = \sum_{j=0}^{k} b_j f(t^{n+1-j})
\]

grows unstable.

Hint: Any solution is a linear combination of basis solutions; a typical basis function corresponding to the root \( \lambda \) is given by \( \lambda^n \), or \( n! \lambda^n \), \( j = 1, \ldots, l - 1 \), if \( \lambda \) is an \( l \)-fold root.

5.11. Consider the explicit midpoint rule

\[
\frac{u^{n+1}}{u^{n-1}} + 2\Delta t f(u^n, t^n).
\]

(a) Show that this method is root stable (see Exercise 5.10).

(b) Apply this method to the initial value problem

\[
\frac{du}{dt} = \lambda u, \quad t > 0,
\]

\[
u(0) = 1.
\]

5.12. Consider the ODE

\[
\frac{du}{dt} = i u,
\]

where \( i \) is the imaginary unit.
(a) Show that the Euler forward method is unstable (i.e., the solution grows unbounded) for any step size (and the number of steps large enough).
(b) Show that the solution of the Euler backward method becomes arbitrarily small for any step size (and the number of steps large enough).
(c) Show that the midpoint rule gives a solution that remains constant in absolute value.

5.13. The BDF-2 method reads
\[ \frac{3}{2} u^{n+1} - 2u^n + \frac{1}{2} u^{n-1} = \Delta t f(u^{n+1}, t^{n+1}). \]
Show that this method applied to the initial value problem
\[ \frac{du}{dt} = \lambda u, \quad t > 0, \]
\[ u(0) = 1, \]
is unconditionally stable for \( \lambda \leq 0. \)

Hint: Note that the recursion you obtain has constant coefficients and thus the problem boils down to finding the characteristic roots.

5.14. Define the matrix
\[ J := \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}. \]
Show that \( \rho(J) < 1 \) but that \( \|J\|_{\infty} > 1. \)

5.15. Consider the following PDE
\[ \frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial u}{\partial x} + cu, \quad a > 0, \quad b, c \in \mathbb{R}. \]
Discretise this equation by means of the explicit Euler method for the time derivative and with central differences for the space derivatives.
(a) Give the operators \( A_{\Delta t} \) and \( B_{\Delta t} \) (cf. (5.105)) explicitly.
(b) Determine the growth factor \( G_k. \)

5.16. Consider the PDE
\[ \frac{\partial u}{\partial t} + b \frac{\partial u}{\partial x} = 0, \quad b \in \mathbb{R}. \]
Discretise this equation in space and time with central differences. Show that this leads to an unconditionally unstable difference equation.
Chapter 6

The Equations of Continuum Mechanics and Electromagnetics

Important areas of application of the methods described in this book are found in continuum physics. They are based on an almost axiomatic footing of conservation laws, completed with problem-dependent constitutive relations. Some care is needed when the well-established laws for mass, momentum, and energy conservation are reformulated for a continuum. These laws will therefore be given in detail. The equations for compressible viscous flow and for linear elastic deformations will be written out in full. In view of their importance the equations for electromagnetic fields will also be derived, but without taking full account of possible coupling between electromagnetic forces and stresses in the material.

6.1 Introduction

The major areas of application of the problems and methods considered in this book are found in the physics of continua, which encompasses the theories of fluid dynamics, deformation of elastic media, and electromagnetic phenomena. Therefore we will summarise the respective theories here. In the context of this book this is necessarily very brief and concise. We will, for example, not consider the combined effect of electromagnetic and inertial forces. The reader is advised to consult the extensive existing literature for this and other details and for further background information.

Continuum physics considers the deformation and motion of matter under internal and external forces (inertia, stresses, gravity, or electromagnetic fields) at a macroscopic level, disregarding the molecular structure other than its integrated effects. The prevailing equations are based on the postulates that mass, momentum, and energy are conserved. Therefore these equations are called conservation equations. They are universal and do not contain the properties of the material in question. The number of unknowns is larger than the number of equations. Therefore the equations are not sufficient to determine the problem, and we need in addition constitutive and thermodynamic relations. These relations represent the properties of the material considered, and their choice is part of the modeling process (see Chapter 7).

We will start with the equations for fluid flow and elastic deformation, leaving aside any electromagnetic effects. Then we will present the equations for electromagnetic fields, without dilating upon any mechanical coupling (except the production of heat).
6.2 Eulerian and Lagrangian Coordinates

Consider the deformation and motion of material continuously distributed in some physical domain. We adopt the continuum hypothesis to consider the material as being made up of a coherent collection of “particles,” each consisting of sufficiently many, but not too many, molecules in order to allow us to speak pointwise of velocity, pressure, temperature, etc. Thus we study the matter at a macroscopic level and do not consider explicitly the interaction between the individual molecules. The particles interact with each other via contact forces (stresses) that depend on the material considered.

There are two approaches to describing a deformation or motion. In the Lagrangian description we follow particles that move with the continuum. In the alternative Eulerian description we maintain a fixed position and consider particles that pass this position. Let the position \( x \) of a particle at time \( t \) be written as

\[
x = \varphi(x^*, t),
\]

where \( x^* \) is a reference position for which we take the position of this particle at an initial time \( t_0; \) i.e., \( x^* = \varphi(x^*, t_0) \). Thus a particle can be specified by its Lagrangian coordinates \( (x^*, t) \), when following the motion, or by its Eulerian coordinates \( (x, t) \). A generic variable \( f \) can be expressed in terms of the Lagrangian or the Eulerian coordinates, and we write

\[
f = f(x, t) = \varphi^*(x^*, t) = \varphi(x^*, t, t).
\]

The velocity \( v^*(x^*, t) \) of a particle is the time rate of change of the position of this particle expressed in Lagrangian coordinates:

\[
v^* := \frac{\partial \varphi}{\partial t}(x^*, t).
\]

The time rate of change of a variable \( f^*(x^*, t) \) may be expressed in Eulerian coordinates as the material or convective derivative of \( f \) and is given by

\[
\frac{\partial f^*}{\partial t}(x^*, t) = \frac{df}{dt}\varphi(x^*, t, t) = \frac{\partial f}{\partial t}(x, t) + v(x, t) \cdot \nabla f(x, t).
\]

The first term on the right-hand side of (6.3) is the local time derivative and the second term is the convective derivative, which is in fact the directional derivative of \( f \) in the direction of velocity \( v \).

The displacement vector is defined by

\[
u = x - x^* = \varphi(x^*, t) - x^*.
\]

In the theory of small elastic deformations we need the linear deformation tensor or linear strain tensor

\[
\mathcal{E} := \frac{1}{2} \nabla u + \frac{1}{2} (\nabla u)^T.
\]

Under the assumption of the deformations being small, the gradient \( \nabla u \) may be interpreted with respect to either \( x \) or \( x^* \).

In the theory of Newtonian viscous fluid flow (Section 6.8.2) we need the deformation velocity tensor or the rate of deformation tensor

\[
\mathcal{D} := \frac{1}{2} \nabla v + \frac{1}{2} (\nabla v)^T.
\]
6.3. The Transport Theorem

A rigorous derivation of the conservation laws is based on the transport theorem, which we will derive first. The theorem is most relevant for a moving fluid, although it remains equally valid for a deforming solid.

Let \( q(x, t) \) be a quantity per unit volume of the material (a fluid, say). Consider a control volume \( \Omega(t) \) moving with the flow. \( \Omega \) is called a material volume and its surface \( \partial\Omega \) is called a material surface. Define

\[
F(t) := \int_{\Omega(t)} q(x, t) \, dV. \tag{6.7}
\]

For example, if \( q(x, t) = \rho(x, t) \) is the mass density of the fluid, then \( F(t) \) is the total mass of fluid contained in the control volume. The time rate of change of \( F(t) \) is given by

\[
\frac{dF(t)}{dt} = \int_{\Omega(t)} \left( \frac{\partial q}{\partial t}(x, t) + \nabla \cdot (qv)(x, t) \right) \, dV. \tag{6.8}
\]

Equation (6.8) is called the transport theorem (see also (J.20) in Appendix J).

A concise proof is as follows. Consider \( F(t + h) - F(t) \) asymptotically for small \( h \).

We write symbolically \( \Omega(t + h) = \Omega(t) + \partial\Omega \). Then we have

\[
F(t + h) - F(t) \approx \int_{\Omega(t) + \partial\Omega} q(x, t + h) \frac{\partial q}{\partial t}(x, t) \, dV - \int_{\Omega(t)} q(x, t) \, dV
\]

\[
\approx \int_{\Omega(t)} q(x, t) \, dV + h \int_{\partial\Omega} \frac{\partial q}{\partial t}(x, t) \, dS.
\]

We apply locally near the surface \( \partial\Omega \) of \( \Omega \) an orthogonal coordinate system \( x = \sigma + \lambda n \), where \( \lambda = 0 \) denotes the surface \( \partial\Omega \) and \( n \) is the unit outward normal of \( \partial\Omega \). As the volume \( \Omega \) is a material volume, it moves with the fluid velocity \( v \), and therefore the surface \( \partial\Omega \) has moved a distance \( hv \) between times \( t \) and \( t + h \). So we have

\[
\int_{\partial\Omega} q(x, t) \, dS = \int_{\partial\Omega} q(\sigma + \lambda n, t) \, d\lambda \, dS \approx \int_{\partial\Omega} q(\sigma) h(v \cdot n) \, dS = h \int_{\Omega} \nabla \cdot (qv) \, dV,
\]
where we used Gauss’s divergence theorem for the last step. After dividing by $h$ and taking the limit $h \to 0$, we obtain (6.8).

**Example 6.3** A flow field $\mathbf{v}$ that keeps the content of any convected volume $\Omega(t)$ constant is called incompressible. It satisfies

$$
\frac{d}{dt} \int_{\Omega(t)} 1 \, dV = \int_{\Omega(t)} \nabla \cdot \mathbf{v} \, dV = 0
$$

and is therefore divergence free; i.e., $\nabla \cdot \mathbf{v} = 0$.

### 6.4 Conservation Equations

Consider the deformation and motion of matter in some domain $\Omega(t)$ that is moving with the material (i.e., $\Omega$ is a material volume). From physics we know that the matter is subject to some very strict limitations called conservation laws. These laws postulate that, without a source, certain properties like mass, momentum, and energy remain unchanged. Any such property $P$ can be described by a density $E$ (the amount of $P$ per unit volume) and an associated flux density $F$ (the amount of $P$ that flows per unit time through a unit material surface normal to $F$) such that the change of $P$ of a given material volume $\Omega$ must be exactly equal to the sum of the net influx through the volume’s surface $\partial \Omega$ and any possible production from a source distribution $Q$:

$$
\frac{d}{dt} \int_{\Omega} E \, dV = -\oint_{\partial \Omega} F \cdot n \, dS + \int_{\Omega} Q \, dV,
$$

(6.9)

where $n$ denotes the unit outward normal of $\partial \Omega$. As $\Omega$ moves with the flow with velocity $\mathbf{v}$, this is, according to the transport theorem and Gauss’s divergence theorem, equivalent to

$$
\int_{\Omega} \left( \frac{\partial E}{\partial t} + \nabla \cdot (E \mathbf{v} + F) - Q \right) \, dV = 0.
$$

(6.10)

If this is true for any material volume in our region of interest, the integrand itself must vanish, so we have

$$
\frac{\partial E}{\partial t} + \nabla \cdot (E \mathbf{v} + F) = Q.
$$

(6.11)

It is clear that while $F$ denotes the flux density through a material surface, $E \mathbf{v} + F$ is the flux through a fixed unit surface. Relation (6.11) is the general form of a conservation law for a continuum. In the following we will derive specific versions of this general format. Note that we will often use just “flux” instead of “flux density.”

### 6.5 Conservation of Mass

As there exist in the present context no mass sources, the mass of any material volume $\Omega(t)$ is constant, so

$$
\frac{d}{dt} \int_{\Omega(t)} \rho \, dV = 0.
$$

(6.12)
6.6. Conservation of Momentum

Applying the transport theorem to (6.12) with \( q(x, t) = \rho \) gives

\[
\int_{\Omega(t)} \left( \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) \right) \, dV = 0.
\]  

(6.13)

Since this conservation law holds for any \( \Omega(t) \), the differential form of the mass conservation law

\[
\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0
\]

(6.14)

must be satisfied. Equation (6.14) is called the continuity equation, written in conservative form. By using the material derivative (6.3), we can rewrite (6.14) into convective form:

\[
\frac{d\rho}{dt} + \rho \nabla \cdot \mathbf{v} = 0.
\]

(6.15)

6.6 Conservation of Momentum

The equations of motion of a continuum describe conservation of linear momentum and angular momentum. First, the law of conservation of linear momentum reads

\[
\frac{d}{dt} \int_{\Omega(t)} \rho \mathbf{v} \, dV = \oint_{\partial \Omega(t)} \mathbf{t}(n) \cdot dS + \int_{\Omega(t)} \rho \mathbf{f} \, dV,
\]

(6.16)

with \( \mathbf{t}(n) \) the stress vector, i.e., the internal or contact force field per unit area, acting on the boundary \( \partial \Omega \) of the material volume \( \Omega \), and \( \mathbf{f} \) the specific (i.e., per unit mass) external or volume force field acting on the material contained by \( \Omega \). In (6.16) we have explicitly written \( \mathbf{t}(n) \) to denote the dependence of the stress vector on the outward unit normal \( n \) on \( \partial \Omega \). The conservation law in (6.16) states that the rate of change of momentum of the material contained in \( \Omega \), due to the movement of \( \Omega \) with velocity \( \mathbf{v} \), is equal to the sum of contact forces and volume forces acting on the material.

Second, the law of conservation of angular momentum is expressed by

\[
\frac{d}{dt} \int_{\Omega(t)} \mathbf{x} \times \rho \mathbf{v} \, dV = \oint_{\partial \Omega(t)} \mathbf{x} \times \mathbf{t}(n) \cdot dS + \int_{\Omega(t)} \mathbf{x} \times \rho \mathbf{f} \, dV,
\]

(6.17)

meaning that the rate of change of angular momentum of the material in \( \Omega(t) \), when the control volume moves with the continuum, is equal to the sum of the moment of the contact forces and the moment of the volume forces acting on it.

In the following the conservation law of linear momentum will be used to develop the equations of motion for a continuum. From the conservation law of angular momentum a certain symmetry in the stress vector will be derived.

Consider the \( i \)th \((i = 1, 2, 3)\) component of the conservation law of linear momentum (6.16). Applying the transport theorem (6.8) to its left-hand side with \( q(x, t) = \rho(x, t)v_i(x, t) \) yields

\[
\frac{d}{dt} \int_{\Omega(t)} \rho v_i \, dV = \int_{\Omega(t)} \left( \frac{\partial}{\partial t} (\rho v_i) + \nabla \cdot (\rho \mathbf{v} v_i) \right) \, dV.
\]

(6.18)
The stress vector \( \mathbf{t}(\mathbf{n}) \) acting on a surface with normal \( \mathbf{n} \) is completely determined by the stress vectors \( \mathbf{t}(\mathbf{e}_j) \) for the unit vectors \( \mathbf{e}_j \) \((j = 1, 2, 3)\). The stress tensor \( \mathbf{T} \) is defined by

\[
\mathbf{T} = (T_{ij}) = (\mathbf{t}(\mathbf{e}_1), \mathbf{t}(\mathbf{e}_2), \mathbf{t}(\mathbf{e}_3)) \quad \text{with} \quad T_{ij} := t_i(\mathbf{e}_j); \quad (6.19)
\]

i.e., \( T_{ij} \) is the \( ij \)th component of the stress vector \( \mathbf{t} \) acting on a surface with unit normal \( \mathbf{e}_j \).

(Note that there is no uniformity in the nomenclature. Some authors define the stress tensor as \( \mathbf{T}^T \). Since \( \mathbf{T} \) will in general be symmetric, this is usually of no concern.) Applying the principle of local equilibrium, it follows that \([2, 24, 41]\)

\[
t_i(\mathbf{n}) = \sum_{j=1}^{3} T_{ij} n_j, \quad i = 1, 2, 3, \quad (6.20a)
\]

or

\[
\mathbf{t}(\mathbf{n}) = \mathbf{T} \mathbf{n}. \quad (6.20b)
\]

The surface integral in (6.16) can now be replaced by a volume integral using Gauss’s theorem \((J.13)\), and we find

\[
\oint_{\partial/\Omega(t)} \mathbf{t}(\mathbf{n}) \, d\mathbf{S} = \oint_{\partial/\Omega(t)} \mathbf{T} \mathbf{n} \, d\mathbf{S} = \int_{\Omega(t)} \nabla \cdot \mathbf{T}^T \, dV, \quad (6.21)
\]

where \( \nabla \cdot \mathbf{T}^T \) is defined by

\[
\nabla \cdot \mathbf{T}^T := \begin{pmatrix} \nabla \cdot T_{1*} \\ \nabla \cdot T_{2*} \\ \nabla \cdot T_{3*} \end{pmatrix} \quad (6.22)
\]

(see the appendix \((L.3)\)) and \( T_{is} \) denotes the \( is \)th row of \( \mathbf{T} \). Combining (6.16), (6.18), and (6.21), we get the \( i \)th component of the conservation law of momentum:

\[
\int_{\Omega(t)} \left( \frac{\partial}{\partial t} (\rho v_i) + \nabla \cdot (\rho \mathbf{v} \mathbf{v}_i) \right) \, dV = \int_{\Omega(t)} \nabla \cdot T_{is} \, dV + \int_{\partial/\Omega(t)} \rho f_i \, dV. \quad (6.23)
\]

This becomes, in differential, conservative form,

\[
\frac{\partial}{\partial t} (\rho v_i) + \nabla \cdot (\rho \mathbf{v} \mathbf{v}_i) = \nabla \cdot T_{is} + \rho f_i, \quad i = 1, 2, 3, \quad (6.24a)
\]

or, in vectorial notation,

\[
\frac{\partial}{\partial t} (\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v} \mathbf{v}^T) = \nabla \cdot \mathbf{T}^T + \rho \mathbf{f}. \quad (6.24b)
\]

The divergence of the dyadic product \( \rho \mathbf{v} \mathbf{v}^T \) is defined analogously to (6.22).

From the conservation of angular momentum (6.17) it happens (see below) that \( \mathbf{T} = \mathbf{T}^T \); in other words, the stress tensor \( \mathbf{T} \) is symmetric (for nonpolar media). Therefore in (6.24b) we can safely change \( \nabla \cdot \mathbf{T}^T \) into \( \nabla \cdot \mathbf{T} \). If, in addition, we use the continuity equation (6.14), we finally get the equation in its best-known, convective, form

\[
\rho \frac{d}{dt} \mathbf{v} = \nabla \cdot \mathbf{T} + \rho \mathbf{f}. \quad (6.25)
\]

This is known as Cauchy’s equation of motion \([24]\).
A brief proof of the symmetry of $\mathcal{T}$ is as follows [2]. By employing the transport theorem (6.8) and the continuity equation (6.14), we can derive for the left-hand side of (6.17)

$$\frac{d}{dt} \int_{\Omega(t)} x \times \rho v \, dV = \int_{\Omega(t)} \rho \frac{d}{dt} (x \times v) \, dV = \int_{\Omega(t)} \rho x \times \frac{dv}{dt} \, dV. \quad (6.26)$$

The second integral in (6.17) can be rewritten by means of Gauss’s theorem (J.13) as

$$\oint_{\partial \Omega(t)} (x \times T) \cdot n \, dS = \int_{\Omega(t)} (x \times \nabla \cdot T + t^*) \, dV, \quad (6.27)$$

where $t^*$ denotes a vector, related to the antisymmetric part of $\mathcal{T}$, given by

$$t^* := \begin{pmatrix} T_{32} - T_{23} \\ T_{13} - T_{31} \\ T_{12} - T_{21} \end{pmatrix}. \quad (6.28)$$

By combining (6.17) with (6.26) and (6.27), we get

$$\int_{\Omega(t)} x \times \left( \rho \frac{d}{dt} v - \nabla \cdot T - \rho f \right) \, dV = \int_{\Omega(t)} t^* \, dV. \quad (6.29)$$

Together with (6.14) and (6.24b) this implies that $t^*$ vanishes; i.e., $\mathcal{T}$ is symmetric.

### 6.7 Conservation of Energy

Finally, we have the law of conservation of energy. The rate of change of kinetic energy ($\frac{1}{2} \rho |v|^2$) and internal energy ($\rho e$) should equal the mechanical power of the stresses ($t(n) \cdot v$) and volume forces ($\rho f \cdot v$) acting on the material, plus the heat supplied by internal sources ($\rho r$) and exchanged across the border ($-q \cdot n$). This is given by

$$\frac{d}{dt} \int_{\Omega(t)} \rho E \, dV = \oint_{\partial \Omega(t)} t(n) \cdot v \, dS + \int_{\Omega(t)} \rho f \cdot v \, dV - \oint_{\partial \Omega(t)} q \cdot n \, dS + \int_{\Omega(t)} \rho r \, dV, \quad (6.30)$$

with $q$ the heat flux vector and $r$ the specific heat supply, e.g., by a distribution of radioactive sources or electric (Joule) heating (see (6.64) below). The specific energy $E$ is defined by

$$E := e + \frac{1}{2} |v|^2, \quad (6.31)$$

with $e$ the specific internal energy (a thermodynamic property) of the material and $\frac{1}{2} |v|^2$ the specific kinetic energy of the continuum. The term $-q \cdot n$ is the amount of energy per unit area and per unit time that is transmitted through $\partial \Omega(t)$ to the fluid in material volume $\Omega(t)$.

By using the transport theorem (6.8) and Gauss’s theorem (J.13), we can convert (6.30) into the volume integral

$$\int_{\Omega(t)} \left( \frac{\partial}{\partial t} (\rho E) + \nabla \cdot (\rho v E) \right) \, dV = \int_{\Omega(t)} (\nabla \cdot (T v) + \rho f \cdot v - \nabla \cdot q + \rho r) \, dV. \quad (6.32)$$
which holds for any volume that moves with the material. Therefore we have the energy equation in differential form:

\[
\frac{\partial}{\partial t} (\rho E) + \nabla \cdot (\rho v E) = \nabla \cdot (T v) + \rho f \cdot v - \nabla \cdot q + \rho r. \quad (6.33)
\]

This may be further simplified by taking the inner product between the momentum equations (6.24b) and the velocity \( v \) and subtracting the result from (6.33). This yields, using the symmetry of \( T \), the equation in the conservative form

\[
\frac{\partial}{\partial t} (\rho e) + \nabla \cdot (\rho v e) = T : \nabla v - \nabla \cdot q + \rho r, \quad (6.34)
\]

where the double inner product \( T : \nabla v \) is given by (see (L.5) in Appendix L)

\[
T : \nabla v := \sum_{i=1}^{3} \sum_{j=1}^{3} T_{ij} \frac{\partial v_i}{\partial x_j}. \quad (6.35)
\]

If we use the equation of mass conservation (6.12), we obtain the equation in convective form:

\[
\rho \frac{d}{dt} e = T : \nabla v - \nabla \cdot q + \rho r. \quad (6.36)
\]

### 6.8 Constitutive Equations and Thermodynamic Relations

The conservation laws for deformable media are the continuity equation (6.14), the momentum equations (6.24b), and the energy equation (6.34). These 5 equations have 14 unknowns. Therefore the mathematical description is not complete without some closure relations describing properties of the matter and certain instantaneous (in both time and space) interactions between material parameters. These relations are called the constitutive equations and thermodynamic relations. In particular, the constitutive equations are not as basic as the conservation laws, and their form will depend on the model adopted. In fact, this part of the model, the constitutive equations, may be called the physical model.

The conservation laws have to be completed with models for the stress tensor \( T \) and the heat flux vector \( q \), thermodynamic relations, internal heat sources \( r \), and the volume force field \( f \). Apart from possible apparent forces in an accelerating coordinate system (called Coriolis forces), gravity is usually the only external force field working on the matter, so \( f = g e_g \), with \( g \) the acceleration of gravity and \( e_g \) the unit vector in the direction of gravitation. Except when indicated otherwise, we will further assume no internal heat sources, so \( r = 0 \).

It should be noted that these constitutive relations between \( T, D, \) and \( E \) are based on the physics, and are therefore subject to certain compatibility conditions. The material may be isotropic [24], i.e., have no preferred directions, and the relation must be independent of orientation. The material may be homogeneous, in which case the relation is independent of position. A very important condition is the compatibility of the relations with the principle of objectivity. This means that the relation should be equivalent for any observer, i.e., in...
any frame of reference [24]. This may be formalized as follows. The transformation that connects two frames of reference by the relations

\[ x' = x'_0(t) + Q(t)x, \quad t' = t - t_0, \quad (6.37a) \]

where \( Q \) is a rotation, i.e., an orthogonal tensor with \( \det Q = 1 \), is called an observer transformation. Suppose such an observer transformation identifies a scalar field \( \phi \), vector field \( u \), and tensor field \( T \) with corresponding fields \( \phi' \), \( u' \), and \( T' \) in the following way:

\[ \phi'(x', t') = \phi(x, t), \quad u'(x', t') = Q(t)u(x, t), \quad T'(x', t') = Q(t)T(x, t)Q^T(t). \quad (6.37b) \]

Then these fields are called objective. For a further discussion on this important restriction we refer to the literature.

In the remainder of this section we will consider some examples of physical models, defined by constitutive equations and, where relevant, supplemented by suitable thermodynamic relations.

A simple but important example is the problem of heat conduction in rigid material. This assumption of rigidity implies the absence of any deformation or any response to external forcing. The stress tensor, therefore, plays no role. The resulting heat equation is similar to the equation that describes mass diffusion.

We will further consider two important types of material that are characterized by their stress tensor: viscous fluids and elastic material. A fluid at rest (i.e., in equilibrium) sustains normal stresses by compression, but cannot sustain any shear stress, and \( T \) only depends on the rate of deformation tensor \( D \) (6.6). Elastic material, on the other hand, is characterized by a response to any deformation and for small deformations the stress tensor depends on the deformation tensor \( E \) (6.5). If the material is purely elastic, the stress depends on \( E \) only and vanishes when \( E \) vanishes (the material is in its undeformed reference state).

### 6.8.1 Heat Conduction and Mass Diffusion

If the material is rigid, allowing no motion or deformation, all conservation laws are trivially satisfied except for the heat balance in the energy equation. If the internal energy \( e \) is a function of the temperature \( T \) only, we can define \( C = \frac{d}{dT}e \) as the heat capacity or specific heat of the material. Furthermore, if the heat flux satisfies Fourier’s law of heat conduction

\[ q = -\kappa(T)\nabla T, \quad (6.38) \]

with \( \kappa \) the coefficient of heat conduction, (6.36) reduces to

\[ \rho C \frac{\partial}{\partial t} T = \nabla \cdot (\kappa \nabla T) + \rho r. \quad (6.39) \]

It is worth mentioning here that diffusion of heat and diffusion of molecules of a solute in a liquid are essentially similar processes. If the concentration of the solute is \( c \) and its flux is \( j \), a widely used constitutive law that relates \( j \) to \( c \) in the absence of a temperature gradient is given by [83]

\[ j = -D(c)\nabla c, \quad (6.40) \]
which is called *Fick’s law*, the analogue of Fourier’s law. From the equation of mass conservation for the solute, i.e.,
\[
\frac{\partial c}{\partial t} + \nabla \cdot j = Q,
\]
where \(Q\) is a source of solute, we obtain the diffusion equation
\[
\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c) + Q. \quad (6.41)
\]

### 6.8.2 Newtonian Viscous Fluid

For fluids it is useful to split the stress tensor into a part depending on the thermodynamic pressure \(p\), representing the stationary normal components, and a viscous part \(\tau\), related to the velocity gradients:
\[
\mathcal{T} = -p I + \tau, \quad (6.42)
\]
where \(\tau\) is called the *viscous stress tensor*. Let us recall the just-derived conservation equations:

- **mass**:
  \[
  \frac{\partial}{\partial t} \rho + \nabla \cdot (\rho v) = 0, \quad (6.43a)
  \]
  
- **momentum**:
  \[
  \frac{\partial}{\partial t} (\rho v) + \nabla \cdot (\rho vv^T) = -\nabla p + \nabla \cdot \tau + \rho f, \quad (6.43b)
  \]
  
- **energy**:
  \[
  \frac{\partial}{\partial t} (\rho e) + \nabla \cdot (\rho ve) = -\nabla \cdot q + \nabla \cdot (-pv + \tau v). \quad (6.43c)
  \]

Depending on the application, it is often convenient to introduce the specific *enthalpy* \(h := e + p/\rho\), or *entropy* \(s\) and absolute temperature \(T\), via the fundamental law of thermodynamics for a reversible process:
\[
Td s = de + p d\rho^{-1} = dh - \rho^{-1} dp. \quad (6.44)
\]

With the convective derivative (6.3), and noting that \(\tau : \nabla v = \nabla \cdot (\tau v) - v \cdot (\nabla \cdot \tau)\) since \(\tau\) is symmetric, the above conservation laws may be rewritten in their convective form:

- **mass**:
  \[
  \frac{d}{dt} \rho = -\rho \nabla \cdot v, \quad (6.45a)
  \]
  
- **momentum**:
  \[
  \rho \frac{d}{dt} v = -\nabla p + \nabla \cdot \tau + \rho f, \quad (6.45b)
  \]
  
- **energy**:
  \[
  \rho \frac{d}{dt} e = -\nabla \cdot q - p \nabla \cdot v + \tau : \nabla v, \quad (6.45c)
  \]
  
- **enthalpy**:
  \[
  \rho \frac{d}{dt} h = \frac{d}{dt} p - \nabla \cdot q + \tau : \nabla v, \quad (6.45d)
  \]
  
- **entropy**:
  \[
  \rho T \frac{d}{dt} s = -\nabla \cdot q + \tau : \nabla v. \quad (6.45e)
  \]
6.8. Constitutive Equations and Thermodynamic Relations

Equations (6.45c), (6.45d), and (6.45e) are three equivalent forms of the energy equation. For an ideal fluid, $e$ and $h$ depend on $T$ only and we may introduce $de = C_VdT$ and $dh = C_PdT$ such that we get

**energy 1:** \[ \rho C_V \frac{d}{dt} T = -\nabla \cdot q - p \nabla \cdot \mathbf{v} + \tau : \nabla \mathbf{v}, \quad (6.45f) \]

**energy 2:** \[ \rho C_P \frac{d}{dt} T = \frac{d}{dt} p - \nabla \cdot q + \tau : \nabla \mathbf{v}. \quad (6.45g) \]

For an incompressible fluid we have $Tds = C_VdT$ with

**energy 3:** \[ \rho C_V \frac{d}{dt} T = -\nabla \cdot q + \tau : \nabla \mathbf{v}. \quad (6.45h) \]

$C_V = T \left( \frac{\partial s}{\partial T} \right)_p$ is the **heat capacity** or **specific heat at constant volume**. $C_P = T \left( \frac{\partial s}{\partial T} \right)_p$ is the **heat capacity** or **specific heat at constant pressure** [83, 26]. In general (for an ideal gas) they may depend on temperature. For a **perfect gas** they are constants. (There is no uniformity in nomenclature. Some authors use perfect gas for what we define as ideal gas.) For an incompressible fluid like a liquid, $C_V$ is practically equal to $C_P$. A notable exception is matter during phase transition (fusion, vaporization), where all the added heat (enthalpy) is used for the solid-liquid or liquid-vapor transition, rather than an increase of temperature. The amount of heat involved with the phase transition is usually called **latent heat**.

We will consider here an ideal, heat-conducting, and viscous fluid, which is described by the following relations:

- **Ideal gas relation:** \[ p = \rho R T, \quad (6.46a) \]
- **Fourier’s heat flux model:** \[ q = -\kappa \nabla T, \quad (6.46b) \]
- **Newton’s viscous stress tensor:** \[ \tau = 2\mu \mathbf{D} + \lambda (\nabla \cdot \mathbf{v}) \mathbf{I}, \quad (6.46c) \]

where $\lambda$ and $\mu$ are viscosity coefficients and $\mathbf{D}$ is the deformation velocity tensor, defined in (6.6), while $\mathcal{R}$ is the specific gas constant. For air $\mathcal{R} = 286.73 \text{ J/(kg} \cdot \text{K)}$. For an ideal gas we also have the relationship $\mathcal{R} = C_P - C_V$. Under these assumptions, (6.45a) and (6.45b), usually supplemented by an energy equation (6.45f) or (6.45g), are called the **Navier–Stokes equations**. (There is no uniformity in the nomenclature. Sometimes this name is only given to the momentum equation (6.45b), and sometimes only to the equations for incompressible flow.)

For an isentropic process we have

\[ C_P dT - \rho^{-1} dp = C_V dT - \rho p^{-2} d\rho = 0, \quad \text{so} \quad \frac{dp}{d\rho} = \gamma \frac{p}{\rho}, \quad (6.47) \]

where $\gamma = C_P/C_V$ is the **specific heat ratio** ($= 1.4$ for air). From the definition of the speed of sound $c^2 := (\partial p/\partial \rho)_s$, we have

\[ c = (\gamma p/\rho)^{1/2} \quad \text{or} \quad c = (\gamma RT)^{1/2}. \quad (6.48) \]

It is instructive to introduce (see (L.2) in Appendix L) the deviatoric deformation velocity tensor $\mathbf{D}' := \mathbf{D} - \frac{1}{3} (\nabla \cdot \mathbf{v}) \mathbf{I}$ such that

\[ \mathbf{T} = \left( -p + \left( \lambda + \frac{2}{3} \mu \right) (\nabla \cdot \mathbf{v}) \right) \mathbf{I} + 2\mu \mathbf{D}'. \quad (6.49) \]
The first part of the stress represents the fluid’s resistance against dynamic compression; the second part is its resistance against shear. It shows that the mechanical pressure \( p_m := \frac{1}{3} \text{tr}(T) \) is not equivalent to the thermodynamic pressure \( p \). From the continuity equation (6.14) the difference is found to be proportional to relative changes of density:

\[
p - p_m = \left( \lambda + \frac{2}{3} \mu \right) (\nabla \cdot v) = \left( \lambda + \frac{2}{3} \mu \right) \frac{1}{\rho} \frac{d}{dt} \rho,
\]

where \( \lambda + \frac{2}{3} \mu \) is called the coefficient of bulk viscosity (or expansion viscosity or second viscosity). If, according to Stokes’s hypothesis, the fluid is in local thermodynamic equilibrium and both pressures are the same, this coefficient vanishes (see [2, 15, 41]). The coefficient \( \mu \) is sometimes called the coefficient of dynamic viscosity, in contrast to the ratio \( \nu = \mu/\rho \), which is called the coefficient of kinematic viscosity. It should be noted that the viscosity coefficients in general depend on the temperature.

Important simplifications are obtained (see Chapter 7) if we may neglect viscosity (the fluid is called a gas) or if the fluid is incompressible (the fluid is called a liquid). In the latter case the energy equation is decoupled from the mass and momentum equations and may be solved separately.

Finally, we note that at a free surface \( S \) of a fluid the surface tension produces a pressure jump across \( S \), which is proportional to the sum of the principal curvatures of the surface. The factor of proportionality \( \sigma \) (say) is usually called “surface tension,” but this is really a force per unit length. If \( n \) denotes a unit vector field, (outward) normal to \( S \), then the pressure jump \( [p]_S := p_{\text{inside}} - p_{\text{outside}} \) is given by [46]

\[
[p]_S = \sigma \nabla \cdot n \quad \text{at} \quad S.
\]

It can be proved that any smooth \( n \) yields the same \( \nabla \cdot n \) at \( S \). The contact angle \( \theta \) (Figure 6.1) between the fluid-free surface and the wetted solid surface (e.g., of the container of the fluid) is—in equilibrium—a material property that does not depend on the shape of the fluid.

![Figure 6.1. Contact angle.](image)

### 6.8.3 Linear Elastic and Viscoelastic Deformations

It can be shown that for deformations small enough to allow linearization of homogeneous isotropic elastic material the stress tensor satisfies Hooke’s law, i.e., the linear relation

\[
\mathbf{T} = \lambda \text{tr}(\mathbf{E}) \mathbf{I} + 2\mu \mathbf{E},
\]

or, written out componentwise,

\[
t_{ij} = \lambda \delta_{ij} (e_{11} + e_{22} + e_{33}) + 2\mu e_{ij},
\]
6.8. Constitutive Equations and Thermodynamic Relations

where we followed the tradition to write the Cartesian components of $\mathbf{T}$ as $t_{ij}$ and of $\mathbf{E}$ as $e_{ij}$. The material parameters $\lambda$ and $\mu$ are called Lamé coefficients. The material is called linear elastic.

From the observation that with small deformations density changes of the material are negligible we may consider $\rho$ constant and we can ignore the continuity equation (6.14). Similarly, in the absence of heat sources the energy equation (6.36) may usually be decoupled from the elastic deformation problem. Another important simplification implied by the assumption of small displacements is the fact that the acceleration in the equation of motion (6.25) simplifies to a double time-derivative of displacement $u$. By eliminating $e_{ij} = \frac{1}{2}(\frac{\partial}{\partial x_j} u_i + \frac{\partial}{\partial x_i} u_j)$, we may finally obtain Navier’s equations

$$\rho \frac{\partial^2}{\partial t^2} \mathbf{u} = (\lambda + \mu) \nabla(\nabla \cdot \mathbf{u}) + \mu \nabla^2 \mathbf{u} + \rho \mathbf{f}. \tag{6.52}$$

If a piece of elastic material is uniformly stretched in one direction, we have, say, $t_{11} > 0$, while the other $t_{ij}$ are zero. This does not imply that only $e_{11} > 0$. The material is usually contracted in the transverse directions. From the inverted relation (6.51),

$$E = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu} \frac{t_{11}}{e_{11}}, \quad e_{22} = e_{33} = -\frac{\lambda}{2\mu(3\lambda + 2\mu)} t_{11}, \quad e_{ij} = 0 \ (i \neq j).$$

From the ratios $t_{11}/e_{11}$ and $-e_{22}/e_{11}$ two constants naturally appear:

$$E := \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu} \quad \text{(Young’s modulus)}, \quad \nu := \frac{\lambda}{2(\lambda + \mu)} \quad \text{(Poisson’s ratio)}.$$

$E$ is positive; $\nu$ is less than 0.5 and usually positive. The inverted equation (6.51) is thus

$$E\mathbf{E} = (1 + \nu)\mathbf{T} - \nu \text{tr}(\mathbf{T})\mathbf{I}.$$

Since $\int_{\Omega} \nabla \cdot \mathbf{u} \, dV = \int_{\partial \Omega} \mathbf{u} \cdot \mathbf{n} \, dS$ describes (linearised) the increase of a volume $\Omega$ by a displacement field $\mathbf{u}$, the material is incompressible (no volume changes) if $\nabla \cdot \mathbf{u} = \text{tr}(\mathbf{E}) = 0$. As $\text{tr}(\mathbf{T}) = \frac{3}{2}\text{tr}(\mathbf{E})$ is to be finite, while $E$ is finite, we have necessarily $\nu = 0.5$ for incompressible elastic material.

Viscoelastic material is characterised by a stress-strain relation that depends on the history of the deformation and allows a certain amount of dissipation. The simplest and most common models are as follows:

- **Kelvin–Voigt model:**

$$\mathbf{T} = \lambda(\text{tr}(\mathbf{E}) + \theta_1 \text{tr}(\dot{\mathbf{E}}))\mathbf{I} + 2\mu(\mathbf{E} + \theta_2 \dot{\mathbf{E}}). \tag{6.53a}$$

where the dot denotes a derivative with respect to time. $\lambda$ and $\mu$ are equivalent Lamé coefficients, while $\theta_1$ and $\theta_2$ are time parameters. Since there exists an equilibrium with constant stress, this material is sometimes called a viscoelastic solid. Note that $\text{tr}(\mathbf{T}) = (3\lambda + 2\mu) \text{tr}(\mathbf{E}) + (3\lambda \theta_1 + 2\mu \theta_2) \text{tr}(\dot{\mathbf{E}})$. The coefficient of $\text{tr}(\dot{\mathbf{E}})$, $3\lambda \theta_1 + 2\mu \theta_2$, is usually small enough to be neglected.
• Maxwell model:

\[
E \dot{E} = (1 + \nu) (\dot{T} + \gamma_1 \mathbf{T}) - \nu (\text{tr} (\dot{T}) + \gamma_2 \text{tr} (\mathbf{T})) \mathbf{I},
\]

(6.53b)

where \( E \) and \( \nu \) are equivalents of Young’s modulus and Poisson’s ratio, while \( \gamma_1 \) and \( \gamma_2 \) are time parameters. As a constant stress produces a constant flow rate, this material is sometimes called a viscoelastic fluid.

Both models may be considered as special cases of further generalisations.

**Example 6.4** In the case of pure shear stress in the 1,2-direction (i.e., \( t_{12} \) is constant), we have with Kelvin–Voigt material the equation \( t_{12} = 2 \mu (e_{12} + \theta_2 \dot{e}_{12}) \), yielding an exponential decay to a limiting deformation gradient \( e_{12}(t) = \frac{E}{2 \mu} t_{12} + C e^{-t/\theta_2} \). In the case of pure shear in the 1,2-direction with a fixed velocity gradient (i.e., \( \dot{e}_{12} \) is constant), we have with Maxwell material the equation \( E \cdot e_{12} = (1 + \nu) (\dot{t}_{12} + \gamma_1 t_{12}) \), yielding an exponential decay to a limiting shear stress \( t_{12}(t) = \frac{E}{1 + \nu} e_{12} + C e^{-t/\gamma_1} \).

### 6.9 Maxwell Equations

The above discussion was aimed at a derivation from first principles of the Navier–Stokes equations, describing the motion of fluids, and the equations of linear elastic deformations. Although this relates to the main area of application considered in this book, we cannot leave unmentioned another monument in applied mathematics, the Maxwell equations for electromagnetic fields.

Electric charge is described by a charge density \( Q \) and a current density \( J \), corresponding to changes in motion. These charges and currents produce electromagnetic fields described by (i) the electric field intensity \( E \), which applies a force \( qE \) to a point charge \( q \), and (ii) the magnetic-flux density or magnetic induction \( B \), which applies a torque \( m \times B \) to a magnetic dipole with magnetic moment \( m \). Further, we introduce the derived fields (iii) \( D \), the electric displacement, and (iv) \( H \), the magnetic field intensity.

For these fields we have the following equations.

- **Coulomb’s law.** The net effect of a charge distribution in a fixed volume \( \Omega \) is equivalent to the total flux of electric displacement \( D \) out through \( \Omega \)’s surface \( \partial \Omega \) (with \( n \) the outward unit normal):

\[
\int_{\Omega} Q \, dV = \oint_{\partial \Omega} D \cdot n \, dS.
\]

(6.54)

- **Ampère–Maxwell’s law.** A current or a changing electric displacement causes a magnetic field as follows. The work done by a magnetic field \( H \) along a closed contour \( C \) is equal to the flux of the total current \( J + \frac{\partial}{\partial t} D \) through the enclosed surface \( S \):

\[
\oint_{C} \mathbf{H} \cdot d\mathbf{l} = \int_{S} \left( J + \frac{\partial}{\partial t} D \right) \cdot n \, dS.
\]

(6.55)

- **Faraday–Henry’s law.** A changing magnetic field causes an electric field as follows. The electromotive force induced by \( E \) around a circuit \( C \) is equal but opposite to the
6.9. Maxwell Equations

rate of change of the magnetic flux $B$ through the enclosed surface $S$:

$$\oint_C E \cdot d\ell = -\frac{d}{dt} \int_S B \cdot n \, dS. \quad (6.56)$$

- **Absence of free magnetic poles.** This yields a vanishing total flux of magnetic induction $B$ through the surface of a volume $\Omega$:

$$\oint_{\partial\Omega} B \cdot n \, dS = 0. \quad (6.57)$$

- **Maxwell equations.** As these equations are valid for any volume $\Omega$ or contour $C$, the identities are valid locally and may be reformulated into the differential equations

$$\nabla \cdot D = Q, \quad \nabla \times H = J + \frac{\partial}{\partial t} D, \quad (6.58)$$

$$\nabla \times E + \frac{\partial}{\partial t} B = 0, \quad \nabla \cdot B = 0,$$

known as the Maxwell equations.

By taking the divergence of Ampère–Maxwell’s law (in differential form), we obtain the law of conservation of electric charge:

$$\frac{\partial}{\partial t} Q + \nabla \cdot J = 0. \quad (6.59)$$

In integral form this says that the variation of the total charge of a fixed volume $\Omega$ (in the absence of sources) is equal to the net flux of the current into $\Omega$’s surface $\partial\Omega$:

$$\frac{d}{dt} \int_{\Omega} Q \, dV = -\oint_{\partial\Omega} (J \cdot n) \, dS. \quad (6.60)$$

Taking the divergence of Faraday–Henry’s law yields that $\nabla \cdot B$ is stationary (and therefore equals zero if it starts that way). So the Maxwell equations consist of only 7 independent relations for 16 unknowns. Therefore we need additional relations to describe the field uniquely.

### 6.9.1 Constitutive Relations

The fields $D$, $H$, and $J$ are related to $E$ and $B$ via constitutive equations that depend on the problem considered. For example, in vacuum and isotropic diamagnetic and paramagnetic media we have the simple linear relationship

$$B = \mu H, \quad (6.61)$$

where $\mu$ is the magnetic permeability and is denoted by $\mu_0$ for vacuum. Its numerical value is $\mu_0 = 1.2566371 \times 10^{-6} \text{ H/m}$. In vacuum and isotropic dielectric media we have the linear relation

$$D = \varepsilon E, \quad (6.62)$$
where \( \epsilon \) is the electric permittivity and is denoted by \( \epsilon_0 \) for vacuum. Its numerical value is \( \epsilon_0 = 8.8541853 \times 10^{-12} \text{ F/m} \). Note that \( \mu_0 \epsilon_0 c^2 = 1 \), where \( c = 2.99792458 \times 10^8 \text{ m/s} \) is the speed of light in vacuum and \( \mu_0 = 4\pi \times 10^{-7} \text{ H/m} \).

The relation between the current and the electric field, the generalized Ohm’s law, is, for a wide range of conditions, linear, and is given by

\[
J = \sigma E,
\]

(6.63)

where \( \sigma \) is the conductivity of the medium.

**Example 6.5** A stationary point charge has a charge distribution \( Q(x) = q\delta(x) \). As the field is stationary, we have \( \nabla \times E = 0 \). Hence \( E \) is conservative and may be written as the gradient of a potential \( E = -\nabla \phi \). If the field is in vacuum, we have \( D = \epsilon_0 E \), and so \( \phi \) satisfies Poisson’s equation \( \nabla^2 \phi = -q/\epsilon_0 \delta(x) \). In infinite space this has solution \( \phi = \frac{q}{4 \pi \epsilon_0 r} \), where \( r = |x| \). \( \square \)

### 6.9.2 Energy Conservation and Poynting’s Theorem

Energy is dissipated in conducting media (otherwise \( J = 0 \)) by Joule heating, given (per unit volume) by the power density \( J \cdot E \). With Ohm’s law (6.63) this simplifies to \( J \cdot E = \sigma |E|^2 \).

This is illustrated by the following energy conservation law. If we define the rate of change of energy density \( \frac{\partial}{\partial t} u := E \cdot \frac{\partial}{\partial t} D + B \cdot \frac{\partial}{\partial t} H \) and the energy flux vector \( S := E \times H \) (also called the Poynting vector), the Maxwell equations may be recast into the identity

\[
\frac{\partial u}{\partial t} + \nabla \cdot S = -J \cdot E,
\]

(6.64)

which is known as Poynting’s theorem. It shows that the rate of change of electromagnetic energy within a certain volume plus the energy flowing out through the boundaries per unit time is equal to minus the work done by the field inside the volume.

### 6.9.3 Electromagnetic Waves and Lorentz’s Force

Disturbances of an electromagnetic field propagate like waves. For linear material, where \( B = \mu H \) and \( D = \epsilon E \), the Maxwell equations can be recast into a set of wave equations as follows.

As \( B \) is solenoidal (divergence free), it can be written as the curl of a vector potential \( A \). Since \( \nabla \times \nabla = 0 \), this vector potential is defined up to a scalar potential \( \alpha \), so we write

\[
B = \nabla \times (A + \nabla \alpha).
\]

From \( \nabla \times (E + \frac{\partial}{\partial t} A + \frac{\partial}{\partial t} \nabla \alpha) = 0 \) it follows that there is a scalar potential \( \psi \) such that

\[
E = -\nabla \psi - \frac{\partial}{\partial t} A
\]

because we can absorb \( \frac{\partial}{\partial t} \alpha \) into \( \psi \). This yields

\[
\nabla^2 \psi + \frac{\partial}{\partial t}(\nabla \cdot A) = -\frac{1}{\epsilon} Q.
\]
With the vector identity $\nabla \times (\nabla \times A) = \nabla (\nabla \cdot A) - \nabla^2 A$, we have

$$\nabla^2 A - \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} - \nabla \left( \nabla \cdot A + \frac{1}{c^2} \frac{\partial \psi}{\partial t} \right) = -\mu J,$$

where we introduced the speed of light $c$ in the medium considered, with $\mu \varepsilon = c^{-2}$. If we choose $\psi$ and $A$ such that $\nabla \cdot A + \frac{1}{c^2} \frac{\partial \psi}{\partial t} = 0$, the Lorentz gauge condition, we have finally the set of wave equations

$$\nabla^2 A - \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} = -\mu J,$$  \hspace{1cm} (6.65a)

$$\nabla^2 \psi - \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} = -\frac{1}{\varepsilon} Q.$$  \hspace{1cm} (6.65b)

Note that the freedom in $\alpha$ provides us with a fairly large class of possible potentials $\psi$.

Other gauge conditions are possible, e.g., the Coulomb gauge condition, where $\nabla \cdot A = 0$, leading to a Poisson, rather than a wave, equation for $\psi$:

$$\nabla^2 \psi = -\frac{1}{\varepsilon} Q.$$  \hspace{1cm} (6.65c)

We finally remark that the Lorentz force $q(E + v \times B)$ describes the force acting on a point charge $q$ moving with velocity $v$ in the presence of an electromagnetic field. The path of the particle may be determined by recalling Newton’s equations describing the change of momentum due to an external force.

**Example 6.6** In free space, in the absence of charge $Q$ or current $J$, in a medium satisfying the linear relations (6.61) and (6.62), we have the time-harmonic plane-wave solution (see Chapter 3) given by the real part of

$$E = E_0 e^{i\omega t - ik \cdot x}, \quad H = H_0 e^{i\omega t - ik \cdot x},$$

where $k = k \kappa$ is the wave vector and $E_0 = E_0 \kappa$ and $H_0 = H_0 \hbar$ are the vectorial amplitudes. Unit vectors $\kappa$ and $\epsilon$ and $\hbar$ denote the direction of propagation and the polarization of $E$ and $H$, respectively. They form an orthonormal triple with $\kappa = \epsilon \times \hbar$, $\epsilon = \hbar \times \kappa$, and $\hbar = \kappa \times \epsilon$. The modulus of the wave vector is $|k| = o/c$, where $c = (\varepsilon \mu)^{-1/2}$, while the moduli of the vectorial amplitudes satisfy $E_0 = Z H_0$, where $Z = (\mu/\varepsilon)^{1/2}$ is the impedance of the medium.

\[ \square \]

**6.10 Discussion**

- As myriad applications of mathematics are found in physics, in particular in continuum physics, it is useful to have the pertaining equations concisely summarized. It is, however, very rare that the equations are used in the very comprehensive way presented. Usually, the problem is much more limited, and it is wise to simplify the equations first before attempting to tackle them mathematically. This highly nontrivial step is called “modeling,” and the next chapter will be devoted to it.

- The section on conservation laws is not only of interest purely physically. These laws form the basis of many numerical methods based on the integral formulations of the conservation laws and are known as finite volume methods.
Exercises

6.1. Verify that (6.45), with relations (6.46), for inviscid nonconducting fluids simplify to the Euler equations
\[ \frac{d}{dt} \rho = -\rho \nabla \cdot \mathbf{v}, \quad \rho \frac{d}{dt} \mathbf{v} = -\nabla p + \rho \mathbf{f}, \quad \frac{d}{dt} \mathbf{s} = 0. \]

6.2. Verify that (6.45), with relations (6.46), for incompressible fluids ($\rho = \rho_0$) of constant viscosity simplify to
\[ \nabla \cdot \mathbf{v} = 0, \quad \rho_0 \frac{d}{dt} \mathbf{v} = -\nabla p + \mu \nabla^2 \mathbf{v} + \rho_0 \mathbf{f}, \quad \rho_0 T \frac{d}{dt} \mathbf{s} = -\nabla \cdot \mathbf{q} + \mathbf{\tau} : \nabla \mathbf{v}. \]

6.3. (a) Derive for an ideal gas the relation
\[ ds = C_V \frac{dp}{p} - C_P \frac{d\rho}{\rho}. \]
When is it possible, and when is it beneficial, to integrate this expression?
(b) Show that a homentropic ($s$ is constant) perfect gas is a polytropic gas ($p/\rho^n$ is constant).
(c) Show that for a homentropic perfect gas flow the following relation holds:
\[ \frac{1}{\rho} \nabla p = \frac{C_P}{R} \frac{p}{\rho} = \frac{\gamma}{\gamma - 1} \nabla \frac{p}{\rho}. \]
(d) Prove Kelvin’s theorem of the conservation of circulation: in an inviscid homentropic flow field any circulation
\[ \Gamma := \oint_{\partial A} \mathbf{v} \cdot d\mathbf{l}, \]
where $\partial A$ is a closed material contour (the boundary of a material surface $A$, i.e., moving with the flow), is a constant while it is convected with the flow.

6.4. Show, by using (6.52), that $\nabla^2 (\nabla \cdot \mathbf{u}) = 0$ and $\nabla^4 \mathbf{u} = \mathbf{0}$ if $f = \frac{\partial^2}{\partial t^2} \mathbf{u} = \mathbf{0}$.

6.5. Verify the identity
\[ \frac{\partial^2 e_{ij}}{\partial x_i \partial x_j} + \frac{\partial^2 e_{k\ell}}{\partial x_k \partial x_\ell} = \frac{\partial^2 e_{ik}}{\partial x_j \partial x_k} + \frac{\partial^2 e_{j\ell}}{\partial x_i \partial x_\ell}, \]
and show that it yields six compatibility relations for the components of $\mathbf{E}$.

6.6. Show that the harmonic plane wave $\mathbf{u} = U \mathbf{v} e^{\omega t - i\mathbf{k} \cdot \mathbf{r}}$ with frequency $\omega$, wave vector $\mathbf{k}$, amplitude $U$, and polarization vector $\mathbf{v}$ that satisfies Navier’s equation (6.52) with $f = \mathbf{0}$ has the dispersion relation
\[ \rho \omega^2 = (\lambda + 2\mu)(\mathbf{k} \cdot \mathbf{k}) \]
if the wave is longitudinal ($\mathbf{v} \parallel \mathbf{k}$) and
\[ \rho \omega^2 = \mu (\mathbf{k} \cdot \mathbf{k}) \]
if the wave is transverse ($\mathbf{v} \perp \mathbf{k}$).
Exercises 127

6.7. Consider two electromagnetic media satisfying relations (6.61) and (6.62) with constants \( \epsilon_1, \mu_1 \) and \( \epsilon_2, \mu_2 \). The media are separated from each other by a plane interface with unit normal vector \( \kappa \) directed from medium 1 to medium 2. The interface is given by \( \kappa \cdot x = 0 \). In medium 1 a plane wave propagates in direction \( \kappa \) (see Example 6.6) and reflects and transmits at the interface. In medium 1 the wave is given by

\[
E = E_0 (e^{-i k_1 \cdot x} + R e^{i k_1 \cdot x}) e^{i \omega t}, \quad H = Z_1^{-1} E_0 (e^{-i k_1 \cdot x} - R e^{i k_1 \cdot x}) e^{i \omega t},
\]

while in medium 2 it is given by

\[
E = E_0 T e^{i \omega t - i k_2 \cdot x}, \quad H = Z_2^{-1} E_0 T e^{i \omega t - i k_2 \cdot x}.
\]

Given that at an interface the normal components of \( D \) and \( B \), and the tangential components of \( E \) and \( H \), are continuous, determine the reflection and transmission coefficients \( R \) and \( T \).
Chapter 7
The Art of Modeling

This chapter describes the ideas and principles of modeling a real-life problem. It is necessarily a bit contemplative in nature. In Section 7.1 we discuss how modeling may be defined. There are various ways to model real-life situations. We then discuss the kinds of models that can be distinguished in Section 7.2. They all relate to the (thrifty) way a real-world problem description is translated to a mathematical formulation and the usefulness of the conclusions drawn from the latter for the actual problem at hand. In order to decide which part of the resulting equations really matters, it is necessary to make them dimensionless. To do this one has to scale the parameters and variables first, which is treated in Section 7.3. We show that by Buckingham’s theorem there is always a subset of relevant (maybe redefined) variables in which the problem can be formulated. As a result, one can often indicate large or small coefficients for some of the contributions in the equation. If they are small enough, we can neglect them, at least in most of the domain. This kind of treating the equations of mass, momentum, and energy for fluid flow is discussed to some extent in Section 7.4. It is shown which simpler equations (models) may result from them.

7.1 Introduction

Mathematics has, historically, its major sources of inspiration in applications [33, 47, 91]. It is just the unexpected question from practice that forces one to go off the beaten track. Also it is usually easier to portray properties of a mathematical abstraction with a concrete example at hand. Therefore it is safe to say that most mathematics is applied, applicable, or emerges from applications.

Before mathematics can be applied to a real problem, the problem must be described mathematically. We need a mathematical representation of its primitive elements and their relations, and the problem must be formulated in equations and formulas to render it amenable to formal manipulation and to clarify the inherent structure. This is called mathematical modeling. An informal definition could be as follows:

Mathematically modeling is describing a real-world problem in a mathematical way by what is called a model, such that it becomes possible to deploy mathematical tools for its solution. The model should be based on first principles.
and elementary relations and it should be accurate enough to have reasonable claims to predict both quantitative and qualitative aspects of the original problem. The accuracy of the description should be limited in order that the model not be unnecessarily complex.

This is evidently a very loose definition. Apart from the question of what is meant by a problem being described in a mathematical way, there is the confusing paradox that we only know the precision of our model if we can compare it with a better model, but this better model is exactly what we try to avoid as it is usually unnecessarily complex! In general we do not know a problem and its accompanying model well enough to be absolutely sure that the sought description is consistent, complete, and sufficiently accurate for the purpose, and not too formidable for any treatment. A model is, therefore, to a certain extent a vague concept. Nevertheless, modeling plays a key role in applied mathematics since mathematics cannot be applied to any real-world problem without the intermediate steps of modeling. Therefore a more structured approach is necessary, which is the aim of this chapter.

Some people define modeling as the process of translating a real-world problem into mathematical terms. We will not do so, as this definition is too wide to include the subtle aspects of “limited precision” (to be discussed in a minute). Therefore we will introduce the word mathematising, defined as the process of translating a real-world problem into mathematical terms. It is a translation in the sense that we translate from the inaccurate, verbose “everyday” language to the language of mathematics. For example, the geometrical presence and evolution of objects in space and time may be described parametrically in a suitable coordinate system. Any properties or fields that are expected to play a role may be formulated by functions in time and space, explicitly or implicitly, e.g., as a differential equation.

Mathematising is an elementary but not trivial step. In fact, it forms probably the single most important step in the progress of science. It requires the distinction, naming, and exact specification of the essential relevant elementary objects and their interrelations, where mathematics acts as a language in which the problem is described. If theory is available for the mathematical problem obtained in this way, the problem considered may be subjected to the strict logic of mathematics, and reasoning in this language will transcend the limited and inaccurate ordinary language. Mathematising is, therefore, as well as being the link between the mathematical world and the real world, also important for science in general.

A very important point to note is the fact that such a mathematised formulation is always at some level simplified. Earth can be modeled by a point or a sphere in astronomical applications, or by an infinite half-space or not at all in problems of human scale. Based on the level of simplification, sophistication, or accuracy, we can associate an inherent hierarchy with the set of possible descriptions. A model may be too crude, but also it may be too refined. It is too crude if it just doesn’t describe the problem considered, or if the numbers it produces are not accurate enough to be acceptable. It is too refined if it includes irrelevant effects that make the problem untreatable, or make the model so complicated that important relations or trends remain hidden. According to Barenblatt [14], every mathematical model is based on “intermediate asymptotics.”

The ultimate goal of mathematising a problem is a deeper understanding and a more profound analysis and solution of the problem. Usually, a more refined problem translation
is more accurate but also more complicated and more difficult—if not impossible!—
to analyse and solve than a simpler one. Therefore not every mathematical translation is a
good one. We will call a good mathematical translation a model or mathematical model if it is
lean or thrifty in the sense that it describes our problem quantitatively or qualitatively to
a suitable or required accuracy with a minimal number of essentially different parameters
and variables. (We say “essentially different” in view of a reduction that is always possible
by writing the problem in dimensionless form. See Buckingham’s Theorem 7.11 below.)
Again, this definition is rather subjective, as it greatly depends on the context of the problem
considered and our knowledge and resources. So there will rarely be one “best” model. At
the same time, it shows that modeling, even if relying significantly on intuition, is part of
the mathematical analysis.

7.2 Models

We will distinguish the following three classes of models.

7.2.1 Systematic Models

Other possible names are asymptotic models or reducing models. The starting point here is
to use available complete models, which are adequate, but overcomplete, so that effects are
included that are irrelevant, uninteresting, or negligibly small, thereby making the mathe-
tatical problem unnecessarily complex. By using available additional information (order
of magnitude of the parameters), we can make assumptions that minimize in a systematic
way the overcomplete model into a good model by taking a parameter that is already large
or small to its asymptotic limit: small parameters are taken zero; large parameters become
infinite; an almost symmetry becomes a full symmetry.

Examples of systematic models are found in particular in the well-established fields
of continuum physics, considered in Chapter 6. An ordinary flow is usually described by a
model that is reduced from the full, i.e., compressible and viscous, Navier–Stokes equations.
This will be elaborated on in detail in Section 7.4.

Example 7.1 (Convection-diffusion) Consider the following simple convection-diffusion
problem. For a temperature field $T$ and given velocity field $v$ in space $x$ and time $t$ and thermal
diffusion coefficient $\alpha$ we have the assumed “overcomplete” model

$$\frac{\partial T}{\partial t} + v \cdot \nabla T = \alpha \nabla^2 T,$$

which is, however, difficult to solve. If we have reasons to assume that the diffusion term may
be ignored, we obtain the reduced problem

$$\frac{\partial T}{\partial t} + v \cdot \nabla T = 0, \quad (\ast)$$

which is far more attractive than the full problem, as it may be solved exactly. Along the
streamlines $x = \xi(t)$ given by

$$\frac{d\xi}{dt} = v,$$
equation (\ast) simplifies to
\[ \frac{d}{dt} T(\xi(t), t) = 0 \]
with solution \( T = T(\xi(t), t) = \text{constant} \).

Example 7.2 (the pendulum) The dynamics of an idealised frictionless undriven pendulum consisting of a point mass suspended by a weightless cord of length \( L \) may be described by the difficult nonlinear pendulum equation
\[ \frac{d^2\theta}{dt^2} = -\frac{g}{L} \sin \theta, \quad \theta(0) = \theta_0, \quad \frac{d}{dt}\theta(0) = 0, \]
where \( g \) is the gravitational acceleration and \( \theta = \theta(t) \) is the angle at time \( t \) of the cord with the vertical. Since \( \frac{d}{dt} \theta(0) = 0 \) and the pendulum is undriven, \( \theta(t) \) will never be larger than \( \theta_0 \). So if \( |\theta_0| \ll 1 \), we may assume that \( \theta \) is always much smaller than one, and we may approximate, at least for some time, the nonlinear term \( \sin \theta \) by \( \theta \). This yields the much simpler model
\[ \frac{d^2\theta}{dt^2} = -\frac{g}{L} \theta, \quad \theta(0) = \theta_0, \quad \frac{d}{dt}\theta(0) = 0, \]
which has solution \( \theta(t) = \theta_0 \cos(\omega t) \) with \( \omega = \sqrt{g/L} \).

### 7.2.2 Constructing Models

Other possible names are building block models or lumped-parameter models. Here we build our problem description step by step from low to high, from simple to more complex, by adding effects and elements lumped together in building blocks, until the required accuracy or adequacy is obtained.

Example 7.3 (the air pump) The following problem of air release by a simple air pump may be an example of a building block model. Consider a pump of cross section \( S \) and length \( a(t) \) that depends on the piston position. Initially, \( a(0) = L \). Under pressure, the enclosed volume of air \( Sa(t) \) leaves the pump through a small hole, forming a jet of cross section \( S_j \) and (mean) velocity \( v_j \). From time \( t = 0 \) a spring pushes against the piston with a force \( F = \lambda a(t) \). Assuming any inertia effects of the piston to be much smaller than the inertia of the flow, the piston force is balanced by a pressure increase from atmospheric pressure \( p_\infty \) outside to the value \( p_0 \) inside the pump. So \( F = S(p_0 - p_\infty) \). Assuming incompressible air of density \( \rho_0 \)

and a practically vanishing velocity inside the pump, conservation of mass (see (6.12)) tells us that \{the variation inside\} = \{-what goes out\}. Written as a formula this is
\[ \frac{d}{dt}(\rho_0 Sa(t)) = -\rho_0 v_j S_j, \]
leading to
\[ v_j = -\frac{S}{S_j} \frac{da}{dt}. \]

From Bernoulli’s law (to be introduced below; see (7.10)), relating pressure \( p \) and velocity \( v \) by \( p + \frac{1}{2} \rho_0 v^2 = \text{constant} \), and noting that the pressure inside the jet is equal to the atmospheric pressure \( p_\infty \) (the jet cannot support a pressure difference), we can deduce that
\[ p_0 = p_j + \frac{1}{2} \rho_0 v_j^2 = p_\infty + \frac{1}{2} \rho_0 v_j^2, \]
resulting in
\[ \frac{1}{2} \rho_0 v_j^2 = \frac{\lambda a}{S}. \]

Putting these together, we have the model
\[ \frac{da}{dt} = -K \sqrt{a} \quad \text{with} \quad K = \frac{S_j}{S} \sqrt{\frac{2\kappa}{\rho_0 S}}, \]
which is easily solved by
\[ a(t) = L \left( 1 - \frac{1}{2} \frac{Kt}{L} \right)^2 \quad \text{along} \quad 0 \leq t \leq \frac{2\sqrt{L}}{K}. \]

**Example 7.4 (the flexible bar)** A brilliant example of a constructive model is the Bernoulli–Euler model of elastic deformation of slender bars, in which case the bar is described by a flexible line of vanishing cross section. The essentials of the theory were developed long before the general results described by (6.51) and (6.52) were available. In principle, the equations for the line should be implied by the general three-dimensional theory by utilizing the slenderness of the bar in the limit to zero. This, however, is not straightforward, therefore the classical derivation is still important [48, 151].

For simplicity we will restrict the analysis to the case of deformation and motion in the vertical plane. Torsion and friction with any surroundings are neglected. The two-dimensional equations of motion are written as a differential equation for the position vector.

The line is described by the position vector \( x(s, t) \) as a function of curve length \( s \) and time \( t \), with natural local coordinate \( s \) such that \( |x'| = 1 \), where \( \{ \} = \frac{\partial}{\partial s} \{ \} \) and \( \{ \} = \frac{\partial}{\partial t} \{ \} \) (see, e.g., [92]). Introduce the right-hand orthogonal basis \([t, n, b]\) consisting of the tangential unit vector \( t = x' \), the principal normal unit vector \( n \), and the binormal unit vector \( b \) such that \( b = t \times n \), \( n = b \times t \), and \( t = n \times b \). The curvature vector is \( k = k' = x'' \), with curvature \( |k| = |k| \) defined such that \( k = k n \). The torsion or second curvature vector is \( b' = -\tau n \), with torsion \( \tau \). Note that \( n' = -\kappa t + \tau b \).

Introduce a bar element of length \( ds \), loaded by an external line load \( q \) and internal forces \( F \) and moments \( M \) at both ends. The basic equations are derived from the equilibrium of the dynamic forces, the equilibrium of the moments, and from the constitutive equations as follows [84].

For a beam there is a moment around \( b \) (bending) and around \( t \) (torsion), so \( M = M_b b + M_t t \) is the moment at both ends. Torsion will be assumed to be zero, and \( M_b \) is given by the following Bernoulli hypothesis. See Figure 7.1.

Consider a small bar of length \( \ell \) and cross section \( A \), bent over an angle \( \psi \). From \( R\psi = \ell \) and \( (R + \xi)\psi = \ell + d\ell \) it follows that \( d\ell/\ell = \xi/R = \xi \kappa \). The residual force \( df \) at a cross-sectional slice \( dA \) that causes the bar to bend is by Hooke’s law given by \( df = Ex\xi dA \). The
moment applied by $df$ is then

$$M_B = \int_A \xi \, df = E \int_A \xi^2 \, dA = EI \kappa,$$

where the bending stiffness $EI$ is the product of Young’s modulus $E$ and the second moment of cross-sectional area $I$.

Since the force $F$ is the only cause of the deformation, $F$ lies in the plane of the tangent and the principal normal, so $F = T \xi + S \eta$, where $T$ is called the normal force and $S$ is called the shearing force. The dynamic force equilibrium $dF + q \, ds = m_0 \ddot{x} \, ds$ (where $m_0$ is the mass per unit length) and the moment equilibrium $dM + dx \times F = 0$ yield

$$F' + q = m_0 \ddot{x}, \quad M' + t \times F = 0.$$

From the vector identity $t \times (M' + t \times F) = t \times M' + T \xi - F = 0$ we obtain $t \times M' = (t \times M) + q = m_0 \ddot{x}$. With (3) and $\tau = M_T = 0$, we have $t \times M' = -\kappa n \times (M_B \eta) = -(M_B n) - M_T \kappa t = -E1 \kappa^2 - E1 \kappa^2 t$, which yields the following differential equation for the position vector:

$$[-E1(x'' + \kappa^2 x') + T \xi'] + q = m_0 \ddot{x}.$$

With $x(s, t) = (x(s, t), y(s, t), 0)$, we denote the angle between horizon and tangent $t(s, t)$ as $\phi(s, t)$, so we have

$$x(s, t) = x(0, t) + \int_0^t \cos \phi(\zeta, t) \, d\zeta, \quad y(s, t) = y(0, t) + \int_0^t \sin \phi(\zeta, t) \, d\zeta.$$

Note that $\kappa^2 = |x'|^2 = (\phi')^2$. If the line is loaded by its own weight only, we have $q = (0, -Q, 0)$, where $Q = m_0 g$ is the weight per unit length and $g$ is the acceleration of gravity.
In $x$ and $y$ coordinates, we have finally

$$\frac{\partial}{\partial s} \left( EI \frac{\partial^2 \phi}{\partial s^2} \sin \phi + T \cos \phi \right) = m_0 \frac{\partial^2 x}{\partial t^2},$$

$$\frac{\partial}{\partial s} \left( -EI \frac{\partial^2 \phi}{\partial s^2} \cos \phi + T \sin \phi \right) - Q = m_0 \frac{\partial^2 y}{\partial t^2}.$$

In the stationary state we can integrate the equations (with integration constants $H_0$ and $V_0$, say), and by eliminating $T$ we get the single equation

$$EI \frac{d^2 \phi}{ds^2} = H_0 \sin \phi - (Q_s + V_0) \cos \phi. \quad (†)$$

Note that the present theory allows large deflections of the bar, although the elastic compression and extension at each element $ds$ are small enough to apply linear elastic theory. If the deflections are small, we may write $x = s$, $y' = \phi$, and $T = \text{constant}$ and derive the linear beam equation

$$EI \frac{d^4 y}{dx^4} - T \frac{d^2 y}{dx^2} + Q + m_0 \frac{d^2 y}{dt^2} = 0.$$

### 7.2.3 Canonical Models

Another possible name is characteristic models or quintessential models. Here an existing model is further reduced to describe only a certain aspect of the problem considered. These models are particularly important if the mathematical analysis of a model from one of the other categories lacks available theory. The development of such a theory is usually hindered by too many irrelevant details. These models are useful for understanding but are usually far away from the original full problem setting and therefore not suitable for direct industrial application.

**Example 7.5 (Burgers equation)** The Navier–Stokes equations for incompressible viscous flow (see Chapter 7), given in terms of a velocity and pressure field, are

$$\frac{\partial}{\partial t} \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 \mathbf{v}, \quad \nabla \cdot \mathbf{v} = 0.$$

These equations are in general very complex. In particular, the coupling between the nonlinear and viscous terms, yielding instabilities and turbulence, is complicated and difficult to analyse. Therefore Burgers proposed to consider the following very simplified version of the equations, where the pressure gradient has been neglected and only behaviour in one dimension is taken into account. This equation,

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2},$$

is called the Burgers equation [169]. A great deal of insight is obtained by the remarkable transformation

$$u = -2\nu \varphi^{-1} \frac{\partial \varphi}{\partial x}$$

found independently by Cole (1951) and Hopf (1950), by which the nonlinear equation is reduced to a linear equation

$$\frac{\partial \varphi}{\partial t} - \nu \frac{\partial^2 \varphi}{\partial x^2} = C(t) \varphi,$$

related to the heat equation, where $C(t)$ is an arbitrary function of $t$. This equation is well understood and allows many exact solutions.
Example 7.6 (Sommerfeld's diffraction problem) An important model problem for the understanding of acoustic or electromagnetic wave scattering at sharp edges is Sommerfeld's flat plate diffraction problem [145]. Consider a flat plate of vanishing thickness at \( y = 0, \ x < 0 \). Time-harmonic acoustic plane waves, with circular frequency \( \omega \), propagate in the direction \( e_x \cos \theta + e_y \sin \theta \), and scatter at the solid plate. The velocity potential is independent of \( z \) and may be written in the usual complex notation \( \phi^*(x, y, t) = \text{Re}(\phi(x, y) e^{i\omega t}) \). As \( \phi^* \) satisfies the wave equation, we have for \( \phi \) the Helmholtz equation with boundary conditions

\[
\nabla^2 \phi + \kappa^2 \phi = 0, \quad \text{where} \quad \kappa = \frac{\omega}{c},
\]

\[
\frac{\partial \phi}{\partial y} = 0 \quad \text{at} \quad y = 0, \ x < 0,
\]

while the incident wave is given by \( \phi_i(x, y) = e^{-i\kappa r \cos(\theta - \theta_i)} \), where \( x = r \cos \theta, \ y = r \sin \theta \). It is important to note that the problem as stated does not have a unique solution. This is caused by three modeling simplifications that we made. We assumed that

1. the field is time harmonic, i.e., exists for all time;
2. the medium is nondissipative;
3. the plate edge is infinitely sharp.

From assumptions 1 and 2 we lost information about the propagation direction of the field, which is (except from the incident part) supposed to radiate away from the scattering edge. This is to be compared with the (acceptable) outward radiating solution \( f(t - r/c)/r \) and (unacceptable) inward radiating solution \( g(t + r/c)/r \) of the three-dimensional wave equation; see Example 12.31 in Section 12.6.2. We have to add radiation conditions in the form of a causality condition (assume the field to be switched on at some time long ago) or allow a small amount of dissipation, usually by giving \( c \) a vanishingly small imaginary part. If the source region is finite, a third option may be to prescribe the direction of the energy flux vector in the far field. In the present problem it comes down to the condition that for large \( \kappa r \) the radiated part of the solution should behave like \( \exp(i\omega t - i\kappa r) \). Therefore it is of utmost importance to note the sign convention \( +i\omega t \) in the exponential. Assumption 3 is of another kind. To represent the solid wall we prescribed the normal component of the gradient to vanish at the plate. At any discontinuous change of the wall's normal vector (i.e., at any sharp edge) the boundary condition is not defined any more, and a line source (delta function and its derivatives) may "hide" itself, producing a false additional field. In the present problem we would have the following false solutions (decaying to zero sufficiently fast for \( r \to \infty \)) [18]:

\[
\psi(x, y) = H^{(2)}_{\nu}(\kappa r) \sin(\nu \theta), \quad \nu = n + \frac{1}{2}, \quad n = 1, 2, \ldots,
\]

where \( H^{(2)}_{\nu} \) is the Hankel function of the second kind [1] (chosen in compliance with the radiation condition) of order \( \nu \). To exclude these solutions we have to add the edge condition of an integrable energy \( \sim |\nabla \phi|^2 \) in any neighbourhood of the edge. This tells us that, e.g., \( \nu = \frac{1}{2} \).
is allowed, but any higher order is not. When we take these extra conditions into account, the solution is found to be

$$\phi(x, y) = e^{-i\kappa r} \left( F(\gamma^+) + F(\gamma^-) \right), \quad \gamma^\pm = (2\kappa r)^{1/2} \sin \frac{1}{2} (\theta \mp \theta_i),$$

where $F$ denotes a version of Fresnel's integral

$$F(z) = \pi^{-1/2} e^{-1/4 \pi i} \int_{z}^{\infty} e^{-t^2} dt.$$

Note that an asymptotic model may start as a building block model, which is only found at a later stage to be too comprehensive. Similarly, a canonical model may reduce from an asymptotic model if the latter appears to contain a particular, not yet understood, effect, which should be investigated in isolation before any progress with the original model can be made.

### 7.3 Nondimensionalisation and Scaling

Modeling means that one has to decide which effects are relevant and should be included and which are irrelevant and can be ignored. More in general, we may expect a hierarchy in relevance from most dominant, via less relevant and locally irrelevant, to absolutely unimportant effects or contributions. Relevant and irrelevant are rather vague qualifications. To make this operational we will relate them to small and large terms in our mathematical description (equations, etc.).

#### 7.3.1 General Concepts

Small and large have no absolute meaning as long as we have not defined our “measuring stick.” To illustrate this we may imagine the following science fiction scenario. Suppose we are lost in outer space, with all planets, stars, and galaxies so far away that they are only seen as sizeless spots on our retina. Then a rock drifts slowly into our field of vision. As long as we are not close enough for a stereoscopic view with both our eyes, we are not able to compare its size or distance with anything we know. There is no way to estimate if it is big and far away, or small and nearby. Only the rock itself is our scale of reference. A similar experience is found when we look into a microscope of unknown amplification. An object, visible but not recognizable, may be as big as an amoeba, or as small as a virus or a molecule. Reinterpreting the famous saying of Protagoras, “Man is the measure of all things,” nothing we observe is small or large, fast or slow, in any absolute sense. It is only by comparison that these qualifications have a meaning.

The next question is, what do we use for comparing? We can use an absolute or universal measuring stick, like a meter or a kilogram, to archive the observations and be able to reproduce them exactly. However, we use a natural scale, like typical sizes in the problem itself, if we want to classify the type of phenomenon.

The following concepts are important in this respect.

When we model, we need to understand the problem in advance to a certain degree, such that we are able to formulate the relevant physical laws and relations. Therefore, in modeling, the natural scaling is the appropriate one to use. We introduce for all our dependent and independent variables typical values taken from the problem in question, e.g., a length $L$ for the independent spatial variable $x$, and a velocity $V$ for the dependent...
variable \( v \), and thus an intrinsic time \( L/V \) for time coordinate \( t \). We refer to this as inherent scaling.

When more than one problem parameter in the same units is available, e.g., a length \( L \) and a width \( D \), or a time \( L/V \) and an inverse frequency \( \omega^{-1} \), it is inevitable that if one is selected for the scaling, the combination with the others gives us new parameters, like \( D/L \) or \( \omega L/V \). These are now independent of the units (meters, seconds) and are therefore called dimensionless parameters. Incidentally, this meaning of the word “dimension” has nothing to do with the mathematical meaning of the number of independent basis vectors in a vector space. Dimensionless parameters are very important for a systematic classification of types of problems. They measure the relative importance of certain effects in an absolute way.

Consider a model depending on \( n \) physical quantities \( q_1, q_2, \ldots, q_n \). Each quantity \( q \) has a dimension (unit of scale, dimensional unit) denoted by \([q]\), such that \( q \) can be written as

\[ q = u[q]. \]

The dimension is derived from a set of \( r \) independent base units \( d_i \), e.g., the SI base units \([m, kg, s, A, K, mol, cd]\) [153]. If the model is a proper one, reflecting the intrinsic relations between the variables, it should not depend on the arbitrary use of meters or inches, etc. Let the model be formally given by the relation

\[ f(q_1, q_2, \ldots, q_n) = 0. \]

This relation should be equivalent for all choices of sets of independent base units. In other words, it should be dimensionally homogeneous. We refer to this as the principle of dimensional homogeneity. In order to achieve this, the dimension function has to satisfy the following conditions.

- Terms that are added, like \( q_1 + q_2 \), should have the same dimensions; i.e., \([q_1] = [q_2]\).
- The dimension of a product should be the product of the dimensions; i.e., if \( q_0 = q_1 q_2 \), then \([q_0] = [q_1][q_2]\).
- Terms that occur as the argument of a dimensionless function like \( \sin \) or \( \exp \) should have dimension one, i.e., be dimensionless. So if \( \sin(q_1 q_2) \) occurs, then \([q_1][q_2] = 1\).

It can be shown (see [91, 13, 14]) that this is only possible if the dimension function is written as a monomial of powers of \( d_i \), so

\[ [q_j] = d_1^{\mu_{1j}} d_2^{\mu_{2j}} \cdots d_r^{\mu_{rj}} = \prod_{i=1}^{r} d_i^{\mu_{ij}}. \]

**Example 7.7 (a simple scaling problem)** Consider the following model of a quantity \( x \) satisfying the equation

\[ ax^2 + bx + c = 0. \]

Assume that \( x \) denotes a length, with units in meters, denoted by \([x] = m\), and \( c \) is a velocity with units in meters per second, or \([c] = m/s\). If the equation is dimensionally homogeneous with \([ax^2] = [bx] = [c]\), the units of the other parameters \( a \) and \( b \) cannot be anything but \([a] = [c]/[x^2] = 1/(m \cdot s)\) and \([b] = [c]/[x] = 1/s\). Therefore we can scale time and length in several combinations to obtain a reduced problem as follows.
7.3. Nondimensionalisation and Scaling

If $b, c \neq 0$: 
\[ x := \frac{c}{b} X, \quad a := \frac{b^2}{c} \alpha, \quad \alpha X^2 + X + 1 = 0. \]

If $ac > 0$: 
\[ x := \frac{c}{\sqrt{a}} X, \quad b := \sqrt{ac} \beta, \quad X^2 + \beta X + 1 = 0. \]

If $a \neq 0$: 
\[ x := \frac{b}{a} X, \quad c := \frac{b^2}{a} \gamma, \quad X^2 + X + \gamma = 0. \]

The constants $\alpha$, $\beta$, and $\gamma$ are dimensionless, parametrising the respective reduced problem. It should be noted that all of these scalings are equivalent (no information is lost), but they are not equally useful. The preferred reduction is the one in which $x$ is scaled on a value typically occurring in the situation considered, and $X$ is henceforth of order unity. So a careful inspection of the range of numerical values of $x$ and the parameters $a, b, c$ is essential. Only then can the dimensionless parameter $\alpha$, $\beta$, or $\gamma$ tell us more about the behaviour of $X$.

**Example 7.8 (a cooking problem)** Consider an object \(\Omega\) of typical size $L$ that initially has a temperature distribution $T(x, 0) = T_0(x)$. The temperature $T$ satisfies the following heat diffusion equation with thermal diffusion constant $\alpha$:

\[
\frac{\partial T}{\partial t} = \alpha \nabla^2 T, \quad x \in \Omega, \quad t > 0,
\]

\[
T(x, t) = 0, \quad x \in \partial\Omega, \quad t > 0,
\]

\[
T(x, 0) = T_0(x), \quad x \in \Omega.
\]

The edges of the object are kept at a constant temperature $T(\partial\Omega, t) = 0$ (Figure 7.2). Note that the steady state solution is $T(x, t) \equiv 0$. So any gradient of $T$ is always coupled to a variation in time. We scale $x$ on $L$, the only length scale in the problem. As the problem is linear, it is not really necessary to scale $T$, but we could use the mean or maximum value of $T_0$. There is no explicit time scale, $t_0$ say, in the problem, e.g., from an external source. If we leave it unspecified for the moment and write $x = L \xi$ and $t = t_0 \tau$, then we obtain

\[
\frac{1}{t_0} \frac{\partial T}{\partial \tau} = \frac{\alpha}{L^2} \nabla^2 \xi T.
\]

As is also clear from the equation, the only parameter with the dimension of time is the number $L^2/\alpha$. Therefore, as long as no steady state is achieved, the balance between decay and diffusion implies that the typical decay time (the half-life, say) is given, in order of magnitude, by this number. It is thus the natural time to scale on and so we have $t_0 = L^2/\alpha$.

**Figure 7.2.** A temperature distribution.
Example 7.9 (electrically heated metal) A piece of metal $\Omega$ of size $L$ is heated from an initial state $T(x, t) \equiv 0$ to a temperature distribution $T$ by applying an electric field with potential $\psi$ and typical voltage $V$ (Figure 7.3). This heat source, amounting to the energy dissipation of the electric field (see Section 6.9.2), is given by the inhomogeneous term $\sigma|\nabla \psi|^2$ in the inhomogeneous heat equation

$$\rho C \frac{\partial T}{\partial t} = \kappa \nabla^2 T + \sigma|\nabla \psi|^2, \quad x \in \Omega, \quad t > 0,$$

$$T(x, t) = 0, \quad x \in \partial\Omega, \quad t > 0,$$

$$T(x, 0) = 0, \quad x \in \Omega.$$

The edges are kept at $T = 0$. If we introduce the formal scaling $T = T_0 u, t = t_0 \tau, x = L \xi$, and $\psi = V \Psi$, then we get

$$\frac{\rho C T_0}{t_0} \frac{\partial u}{\partial \tau} = \frac{\kappa}{L^2} \nabla^2 \xi u + \frac{\sigma V^2}{L^2} |\nabla \xi \Psi|^2.$$

Assuming a balance between the storage (first) and dissipation (second) terms during the initial phase of the process (although details may vary with the applied field $\psi$), it follows that the generated heat is dissipated through the metal with a typical decay time of $O(\rho CL^2/\kappa)$, which is therefore a natural choice for the scaling time $t_0$.

Assuming a balance between the dissipation and source (third) terms in the stationary state, it follows that the final temperature of the stationary state is typically $O(\sigma V^2/\kappa)$, which is therefore a suitable choice for $T_0$, the scaling temperature.

Note that the boundary conditions are rather important. If the edges were thermally isolated, we would, at least initially, have no temperature gradients scaling on $L$. Only the storage term would balance the source term, and there would be no other temperature to scale on than $\sigma V^2 t_0/\rho CL^2$. In other words, the temperature would rise approximately linearly in time.

See Example 7.19 for an extensive description.

\[\begin{array}{c}
\includegraphics[width=0.2\textwidth]{figure}\n\end{array}\]

**Figure 7.3. A piece of metal heated by an electric field.**

7.3.2 Dimensional Analysis

In any description of reality the variables and parameters have physical dimensions and are therefore dimensionally related. However, any physical law is only universally valid if the expression or expressions are dimensionally homogeneous and independent of the physical dimensions used. We will see below that this fact alone implies that the set of variables and parameters can always be reduced to a smaller set of essentially independent, dimensionless, variables and parameters.
7.3. Nondimensionalisation and Scaling

Definition 7.10. Dimensionless groups consisting of a combination of problem parameters only are called dimensionless numbers. Groups consisting of a combination of parameters and several variables are called similarity variables.

As the number of dimensionless groups tells us something about the complexity of the model involved (and for a good model this should not be too high; see Sections 7.1 and 7.2), we are interested in determining how many dimensionless groups are possible at most.

A fundamental result quantifying this is known as Buckingham's $\Pi$ theorem ($\Pi$ stands for the products).

Theorem 7.11 (Buckingham's $\Pi$ theorem). If a physical problem is described by $n$ variables and parameters in $r$ dimensions, the number of essentially different problem parameters (dimensionless groups) is at most $n - r$.

Proof. Evidently, $n > r$. Suppose we have $n$ physical variables and parameters $q_1, \ldots, q_n$ in $r$ independent physical dimensions $d_1, \ldots, d_r$. All $d_i$ occur at least once. So we postulate a relation

$$f(q_1, q_2, \ldots, q_n) = 0,$$

while the dimension $[q_j]$ of each $q_j$ may be written as

$$[q_j] = d_1^{\mu_{1j}} d_2^{\mu_{2j}} \cdots d_r^{\mu_{rj}} = \prod_{i=1}^r d_i^{\mu_{ij}}.$$

From the principle of dimensional homogeneity (this relation should be equivalent for all base units) and the fact that any dimension can be written as a power law monomial, it can be shown [91, 13, 14] that $f$ can be written as

$$\Phi(R_1, \ldots, R_m) = 0,$$

where $\Phi$ depends on $m$ dimensionless groups of $q_1, \ldots, q_n$ of the form $R_k = q_1^{\alpha_{1k}} \cdots q_n^{\alpha_{nk}}$. The smallest possible $m$ is the number to be determined.

Since each group is dimensionless, we have the dimension of $R_k$ given by

$$[R_k] = [q_1^{\alpha_{1k}} q_2^{\alpha_{2k}} \cdots q_n^{\alpha_{nk}}] = [q_1^{\alpha_{11}} || q_2^{\alpha_{21}} || \cdots || q_n^{\alpha_{n1}}]$$

$$= \prod_{j=1}^n [q_j^{\alpha_{jk}}] = \prod_{j=1}^n \prod_{i=1}^r d_i^{\mu_{ij} \alpha_{jk}} = \prod_{i=1}^r d_i^{\sum_{j=1}^n \mu_{ij} \alpha_{jk}} = 1$$

for $k = 1, \ldots, m$. This is only possible if any of the exponents of $d_i$ is zero. In other words, $m$, the number of possible groups, is the number of independent (nontrivial) solutions $\xi = (\xi_1, \ldots, \xi_n)^T$ of

$$\sum_{j=1}^n \mu_{ij} \xi_j = 0 \quad \text{for} \quad i = 1, 2, \ldots, r,$$
or, in matrix notation,
\[
\begin{pmatrix}
\mu_{11} & \mu_{12} & \ldots & \mu_{1n} \\
\mu_{21} & \ddots & & \\
\vdots & & \ddots & \\
\mu_{r1} & \cdots & \ldots & \mu_{rn}
\end{pmatrix}
\begin{pmatrix}
\xi_1 \\
\xi_2 \\
\vdots \\
\xi_n
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
\vdots \\
0
\end{pmatrix}.
\]

Hence we have at least \(n - r\) nontrivial solutions (i.e., nonzero and apart from a multiplicative factor) because the number of independent solutions is equal to the dimension \(n\) of the solution vector minus the rank of the matrix, which is at most \(r\), the number of equations. However, the base units \(d_1, \ldots, d_r\) are by assumption independent and occur at least once, so all \(r\) equations are independent and the rank of the matrix is exactly \(r\), and we have \(n - r\) nontrivial solutions.

On the other hand, as long as we have not solved the problem in detail, we may not be certain that all \(R_1, \ldots, R_{n-r}\) are indeed necessary to describe the problem. Therefore the number of dimensionless groups is at most \(n - r\).

**Corollary 7.12.** If a physical quantity \(q_0\) is expressed by \(n\) quantities \(q_1, \ldots, q_n\) in \(r\) dimensional base units, it depends on (at most) \(n - r\) dimensionless parameters.

**Proof.** From Theorem 7.11 it follows directly that if \(q_0 = f(q_1, \ldots, q_n)\), it can be written as \(q_0 = q_1^{\gamma_1} \cdots q_n^{\gamma_n} F(R_1, \ldots, R_m)\), where \(m = n - r\).

**Example 7.13** (viscous drag) Consider the drag \(D\)—the reaction force due to the surrounding flow—of a sphere of radius \(a\) moving with velocity \(V\) in a viscous fluid with viscosity \(\mu\) and density \(\rho\). We assume, as our model, that the drag \(D\) depends only on \(\rho, V, \mu,\) and \(a\). (This is true for relatively low velocities, an infinite medium, and a relatively large sphere.)

Now we verify the dimensions of the parameters \([D] = \text{kg} \cdot \text{m/s}^2\), \([\rho] = \text{kg}/\text{m}^3\), \([V] = \text{m/s}\), \([\mu] = \text{kg}/(\text{m} \cdot \text{s})\), and \([a] = \text{m}\). Presented in the form of a table, with at each entry the corresponding exponent of the base units kg, m, and s, this is as follows:

<table>
<thead>
<tr>
<th>(\rho)</th>
<th>(V)</th>
<th>(\mu)</th>
<th>(a)</th>
<th>(D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kg</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>m</td>
<td>-3</td>
<td>1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>s</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>-2</td>
</tr>
</tbody>
</table>

We conclude that we can express \(D\) in a functional relationship with at most \(4 - 3 = 1\) dimensionless constant. The best-known form is the Reynolds number \(Re = \rho VL/\mu\). The drag is usually scaled either on the pressure difference \(\frac{1}{2} \rho V^2 a^2\) between front and back of the sphere or on the viscous friction \(aV \mu\) due to wall shear stress. This leads to the functional relations

\[D = \frac{1}{2} \rho V^2 a^2 F(Re) = aV \mu G(Re),\]

where \(Re = \frac{\rho VL}{\mu}\).

The first expression is the proper scaling for nearly inviscid flow (\(Re\) large) and the second is for very viscous flow (\(Re\) small).

**Example 7.14** (an intense explosion) A famous example, originally due to G. I. Taylor [13, 12, 14], is the analysis of the shock front propagation of a very intense (e.g., nuclear bomb)
7.3. Nondimensionalisation and Scaling

explosion. From physical considerations the radius of the shock wave front $R$ depends, during the early stages of the explosion when the pressure inside the shock wave is much higher than outside, only on the time interval $t$ since the explosion, the initial energy $E$, and the initial air density $\rho_0$. Since $[R] = m$, $[t] = s$, $E = kg \cdot m^2/s^2$, and $\rho_0 = kg/m^3$, we have $3 - 3 = 0$ dimensionless groups. In other words, we can express $R$ as

$$R = \text{constant} \left( \frac{E}{\rho_0} \right)^{1/5} t^{2/5}.$$ 

The full solution to the appropriate gas dynamical problem showed that the constant has a value close to unity.

Example 7.15 (membrane resonance) The resonance frequency $\omega$ of a freely suspended membrane (no resonance cavity) is determined by the air density $\rho_a$ and sound speed $c_a$, and the membrane tension $T$, density $\sigma$, and diameter $a$. As $[\rho_a] = kg/m^3$, $[c_a] = m/s$, $[T] = kg/s^2$, $[\sigma] = kg/m^2$, and $[a] = m$, we have $5 - 3 = 2$ dimensionless numbers determining $\omega$. A possible choice is

$$\omega = \frac{c_M}{a} F \left( \frac{c_M}{c_a}, \frac{\rho_a a}{\sigma} \right),$$

where for convenience we introduced $c_M = (T/\sigma)^{1/2}$, the propagation speed of transversal waves in the membrane in the absence of air loading.

Example 7.16 (a sessile drop with surface tension) The height $h$ of a drop of liquid at rest on a horizontal surface with the effect of gravity balanced by surface tension is a function of liquid density $\rho$, volume $L^3$, acceleration of gravity $g$, surface tension $\gamma$, and contact angle $\theta$. As $[h] = m$, $[\rho] = kg/m^3$, $[L] = m$, $[g] = m/s^2$, $[\gamma] = kg/s^2$, and $[\theta] = 1$, we have $5 - 3 = 2$ dimensionless numbers. Possible choices are ($\theta$ is already dimensionless)

$$h = L F(Bo, \theta) = \left( \frac{\gamma}{\rho g} \right)^{1/2} G(Bo, \theta),$$

where $Bo = \rho g L^2 / \gamma$. $Bo$ is known as the Bond number. The first form is suitable when $Bo$ is small (high relative surface tension). The drop becomes spherical and $h$ is comparable with $L$. The second form is the proper scaling when $Bo$ is large (low relative surface tension). The drop is flat as a pancake and $h$ is comparable with $\sqrt{\gamma / \rho g}$ [124].

7.3.3 Similarity Solutions

If the problem contains no other length scale than the spatial variable $x$ itself and no other time scale than the time variable $t$ itself, dimensionless groups can only occur by combinations of $x$ and $t$. As a result, the spatial distribution of the solution develops in time but remains geometrically self-similar [14]. We call this a similarity solution. Dimensional analysis of the independent variables naturally suggests the possibility of such solutions, although not always the form. Following Zeldovich and Barenblatt [13], we call self-similar solutions that can be constructed using dimensional analysis alone self-similar solutions of the first kind. There is complete similarity in all the parameters and variables, independent as well as dependent. Self-similar solutions of the second kind are self-similar solutions with incomplete similarity in the dependent variables. They are connected to an eigenvalue problem. We refer to [13] for an extensive discussion.

We will illustrate the concept of similarity solutions by the following examples.
Example 7.17 (the heated bar) Consider the following heat conduction problem. A very long thermally isolated bar, initially at uniform temperature zero, is heated at one end by a constant flux. There is no source at the other end. The bar is modeled as semi-infinite (Figure 7.4), with a crosswise constant temperature distribution, while the temperature $T$ is described by the following one-dimensional equation for heat conduction, with constant heat diffusion coefficient $\alpha$:

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad x \in [0, \infty), \quad t > 0,$$

with initial and boundary conditions

$$\frac{\partial T}{\partial x}(0, t) = -Q_0, \quad t > 0,$$

$$T(x, 0) = 0, \quad x \in [0, \infty),$$

$$0 \leq T(x, t) < \infty, \quad x \to \infty, \quad t > 0.$$

If we try to scale, we find that we have modeled any explicit length, time, or temperature scale out of our problem. So we can only make the problem dimensionless on the available implicit scales:

- As there is no length scale in $x$ or $t$, the intrinsic length scale can only be $\sqrt{\alpha t}$.
- The only temperature in the problem is $Q_0 x$ or $Q_0 \sqrt{\alpha t}$.

Therefore we assume

$$T(x, t) = Q_0 x g(\eta),$$

where the similarity variable $\eta$ is given by

$$\eta = \frac{x}{\sqrt{4\alpha t}}.$$

It follows that $g$ satisfies the reduced ODE

$$\frac{1}{2} \eta \gamma'' + (1 + \eta^2) \gamma' = 0$$

with boundary conditions

$$\lim_{\eta \to \infty} g(\eta) = 0, \quad \lim_{\eta \to 0} \{g(\eta) + \eta g'(\eta)\} = -1.$$

This has the solution

$$g(\eta) = \frac{1}{\eta \sqrt{\pi}} \exp(-\eta^2) - \text{erfc}(\eta),$$

where $\text{erfc}(x) := \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-\xi^2} \, d\xi$ is the complementary error function. Hence we have

$$T(x, t) = Q_0 \left[ \frac{4\alpha t}{\pi} \exp(-\eta^2) - x \text{erfc}(\eta) \right].$$

Figure 7.4. A semi-infinite bar heated from $x = 0$. 
7.3. Nondimensionalisation and Scaling 145

(Note that there exists no stationary solution!) The found solution is completely similar, both in the independent and in the dependent variables. Therefore it is a similarity solution of the first kind.

Example 7.18 (convection) In the convection problem
\[
\frac{\partial u}{\partial t} + U_0 \frac{\partial u}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0,
\]
\[u(x, 0) = H(x), \quad x \in \mathbb{R},\]
there is the length scale given by \(U_0 t\) and a length scale, say \(L\), hidden in the initial profile \(H(x)\), as \(x\) cannot occur on its own. The dimensions of \(u\) and \(H\) are the same, say \(H_0\), and we write \(H(x) = H_0 h(x/L)\). We scale \(x = L \xi, t = L U_0 \tau, u = H_0 \nu,\) and \(v(\xi, 0) = h(\xi)\) to get
\[
\frac{\partial \nu}{\partial \tau} + \frac{\partial \nu}{\partial \xi} = 0,
\]
with solution \(\nu(\xi, \tau) = h(\xi - \tau)\).

Example 7.19 (Ohmic heating at a corner) Consider the following canonical model, originating from the problem of Joule heating in miniature soldering [126], for the edge singularity of the time-dependent temperature field generated in a homogeneous and isotropic conductor by an electric field (see (6.58)). The electric current density \(J\) and the electric field \(E\) satisfy Ohm’s law (6.63) \(J = \sigma E\), where \(\sigma\) is the electric conductivity, i.e., the inverse of the specific electric resistance. For an effectively stationary current flow the conservation of electric charge (6.59) leads to a vanishing divergence of the electric current density, \(\nabla \cdot J = 0\). The electric field \(E\) satisfies \(\nabla \times E = 0\) and therefore has a potential \(\psi\), with \(E = -\nabla \psi\), satisfying \(\nabla \cdot (\sigma \nabla \psi) = 0\). The electric conductivity \(\sigma\) is a material parameter that depends quite strongly on temperature. Nevertheless, to make progress we will assume a constant \(\sigma\), independent of \(T\). This, then, leads to the Laplace equation for \(\psi\):
\[
\nabla^2 \psi = 0.\tag{\ast}
\]
The heat dissipated as a result of the work done by the field per unit time and volume (Ohmic heating) is given by Joule’s law \(J \cdot E\) (see Section 6.9.2) and leads to the heat source distribution
\[
\sigma |\nabla \psi|^2.
\]
Since energy is conserved, the net rate of heat conduction and the rate of increase of internal energy are balanced by the heat source (6.39) with \(\rho r\) the above heat source), which yields the equation for temperature \(T\):
\[
\rho C \frac{\partial T}{\partial t} = \kappa \nabla^2 T + \sigma |\nabla \psi|^2.
\]
The thermal conductivity \(\kappa\), the density \(\rho\), and the specific heat of the material \(C\) depend mildly on temperature, but we assume these parameters constant.

Since we are interested in the role of the edge only, the conductor is modeled, in cylindrical \((r, \phi)\) coordinates, as an electrically and thermally isolated infinite wedge-shaped two-dimensional region (without any geometrical length scale; Figure 7.5) \(0 \leq \phi \leq \nu\) with an electric field with potential
\[
\psi(x, y) = (v/\pi) A r^{\phi/\nu} \cos(\phi \pi/v)
\]
Figure 7.5. A wedge-shaped conductor heated by an electric field.

(a similarity solution of the second kind of (\ref{eq:heat-source})), while the temperature distribution $T$ due to the heat generated by this source is given by

$$
\rho C \frac{\partial T}{\partial t} = \kappa \nabla^2 T + \sigma A^2 r^{2\pi/\nu-2}, \tag{**}
$$

with boundary and initial conditions

$$
\frac{\partial T}{\partial \phi} = 0 \quad \text{at} \quad \phi = 0, \, \phi = \nu, \quad T(x, y, 0) \equiv 0.
$$

Since there are no other (point) sources at $r = 0$, we have the additional condition of a finite field at the origin: $0 \leq T(0, 0, t) < \infty$. Boundary conditions and the symmetric source imply that $T$ is a function of $t$ and $r$ only, so that (**) reduces to

$$
\rho C \frac{\partial T}{\partial t} = \kappa \left( \frac{\partial^2 T}{\partial r^2} + \frac{1}{r} \frac{\partial T}{\partial r} \right) + \sigma A^2 r^{2\pi/\nu-2}. \tag{†}
$$

Owing to the homogeneous initial and boundary conditions, the infinite geometry, and the fact that the source is a monomial in $r$, homogeneous of order $2\pi/\nu - 2$, there is no length scale in the problem other than $\sqrt{\kappa \rho C / \kappa}$, while the temperature $T$ can only scale on $\sigma A^2 r^{2\pi/\nu} / \nu$. This indicates that a similarity solution (of the first kind) is possible of the form

$$
T(r, t) = \frac{\sigma}{4\kappa} A^2 r^{2\pi/\nu} h(X), \quad X = \frac{\rho C r^2}{4\kappa t},
$$

where $X$ is a similarity variable, reducing (†) to

$$
X^2 h'' + X \left( 1 + \frac{2\pi}{\nu} + X \right) h' + \frac{\pi^2}{\nu^2} h = -1,
$$

with boundary conditions, corresponding to the behaviour near $r = 0$ and $t = 0$, of

$$
0 \leq X^{\nu/\nu} h(X) < \infty \quad \text{if} \quad X \downarrow 0, \quad h(X) \to 0 \quad \text{if} \quad X \to \infty.
$$

This equation may be recognized as related to the confluent hypergeometric equation in $-X$. It has solution, with the required behaviour in $X = 0$, given by

$$
h(X) = \text{constant} \times X^{-\nu/\nu} M \left( -\frac{\pi}{\nu}, 1; -X \right) - \frac{\nu^2}{\pi^2},
$$

where $M(a; b; z)$ is Kummer’s function or the regular confluent hypergeometric function \cite[Ch. 13]{1}. From the asymptotic expansion of $M(-\nu/\nu; 1; -X)$ and the condition for $X \to \infty$
7.3. Nondimensionalisation and Scaling

the unknown constant is found to be \((\nu/\pi)\Gamma(\pi/\nu)\). Putting everything together, we have the solution

\[
T(r, t) = \frac{\sigma \nu^2 A^2}{4\pi^2 \kappa} r^{-2\pi/\nu} \left[ \Gamma\left(1 + \frac{\pi}{\nu}\right) \left(\frac{\rho C r^2}{4\kappa t}\right)^{-\pi/\nu} M\left(-\frac{\pi}{\nu}; 1; -\frac{\rho C r^2}{4\kappa t}\right) - 1 \right].
\]

At the edge we have \(T(0, t) \sim t^{\pi/\nu}\). This shows, together with the radial temperature distribution given in Figure 7.6, a marked difference in behaviour between outward \((\nu < \pi)\) and inward \((\nu > \pi)\) pointed wedges. For the first category the temperature at the corner rises smoothly and so slowly that it always remains behind the temperatures for larger \(r\). For the other category it is the other way around. The corner temperature rises abruptly and so quickly that the values for larger \(r\) are always lower.

![Figure 7.6. Radial temperature distribution in wedge of \(\nu = \frac{1}{2} \pi\) (left) and \(\nu = \frac{3}{2} \pi\) (right) for \(\sigma A^2/4\kappa = 1\) and \(4\kappa t/\rho C = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \ldots\).](image)

**Example 7.20** (decibels) Since the range of our human audible sensitivity is incredibly large \((10^{14}\) in energy), the loudest and quietest levels are practically infinitely far away. Therefore we have no reference or scaling level to compare with other than the sound itself we are hearing. As a result, variations in loudness \(dL\) are perceived proportional to relative variations of the physical sound intensity \(dI/I\) and thus \(L\) varies logarithmically in \(I\). As the intensity (the time-averaged energy flux) \(I\) is, for a single tone, proportional to the mean squared acoustic pressure \(p_{rms}^2\), we have for suitable constants \(K\) and \(L_0\) the relation \(L = K \log(p_{rms}) + L_0\). When

\[
L = 2 \log_{10}(p_{rms}/p_0)
\]

for a reference value \(p_0 = 2 \times 10^{-5}\) Pa, we call \(L\) the sound pressure level in bels. The usual unit is one tenth of this, the decibel.

**Example 7.21** (duct modes) When the geometrical restrictions of a problem are invariant in one direction (say \(z\)), the absence of any length scale in \(z\) leads naturally to a trial solution independent of \(z\). Usually, however, there are infinitely many solutions that are self-similar in \(z\). These solutions are called modes. They are self-similar of the second kind and indeed related to an eigenvalue problem in \((x, y)\). The missing length scale in the \(z\) direction is inherited from the available length scale in the crosswise \((x, y)\) plane by a dispersion relation.

Consider, as an example, the acoustic wave problem in the hard-walled duct given by

\[
\mathcal{D} = \{x = (x, y, z) | (x, y) \in A\}.
\]
where \( A \) is a simply connected two-dimensional area in the \((x, y)\) plane. If the time-harmonic potential field is given by \( \text{Re}[\phi(x) e^{i\omega t}] \), then \( \phi(x) \) satisfies the Helmholtz equation with boundary condition

\[
\nabla^2 \phi + \kappa^2 \phi = 0 \quad \text{for} \quad x \in D, \quad n \cdot \nabla \phi = 0 \quad \text{for} \quad x \in \partial D,
\]

where \( \kappa = \omega/c \) and \( n \) is the normal of the duct wall. In view of the invariance in \( z \) of the geometry and boundary conditions, we try solutions of the form

\[
\phi(x) = \psi(x, y) F(z)
\]

to find that

\[
-\nabla^2 \psi = \alpha^2 \psi \quad \text{for} \quad (x, y) \in A, \quad n \cdot \nabla \psi = 0 \quad \text{for} \quad (x, y) \in \partial A.
\]

For the rectangular duct \( A = [0, a] \times [0, b] \) we have solutions \( \psi = \cos(p \pi x / a) \cos(q \pi y / b) \), where \( \alpha^2 = (p \pi / a)^2 + (q \pi / b)^2 \) and \( p \) and \( q \) are nonnegative integers. For the circular duct \( A = \{ r < R \} \) in polar coordinates \((r, \phi)\) we have the solutions \( \psi = J_m(\alpha r) e^{\pm i m \phi} \), where \( J_m \) is the \( m \)th order ordinary Bessel function of the first kind [1], \( m \) is a nonnegative integer, and \( \alpha R = J'_m(\alpha R) \geq 0 \) is a nontrivial zero of \( J'_m \). In general, it is true that for this boundary condition the eigenvalues \( \alpha_n^2 \) are real and positive, except for the first one, which is \( \alpha_1 = 0 \).

Note that the eigenvalue problem is independent of \( \kappa \). For a finite number of eigenvalues, i.e., with \( 0 \leq \alpha_n < \kappa, \gamma_n \) is real and the mode propagates in \( z \). For the infinitely many others, i.e., with \( \alpha_n > \kappa, \gamma_n \) is imaginary and the mode is evanescent, i.e., exponentially decaying in \( z \). If \( \alpha_n = \kappa \), then \( \gamma_n = 0 \) and the mode is in resonance, i.e., independent of \( z \). The set of modes forms an \( L_2 \)-complete set of orthogonal basis functions to represent any solution of the problem.

### 7.4 Scaling and Reduction of the Navier–Stokes Equations

One of the most fruitful uses of scaling is the hierarchy it provides in comprehensive and rich models. In most applications such overcomplete models are not truly adapted to the problem, and therefore they are not true “lean” models, as we discussed above. From a suitable inherent scaling the order of magnitude of the various contributions or effects can be estimated, leading to a hierarchy on which we can base an asymptotic modeling.

We will give here one of the most important examples, the scaling and reduction of the compressible Navier–Stokes equations. Note that a typical, but not universal, approach will be worked out, since sometimes problem-dependent choices have to be made.

We recall the mass, momentum, and energy conservation equations (6.45) with constitutive relations (6.46) as derived in Chapter 6.

#### 7.4.1 Scaling; Nondimensionalisation

Assume that we have the following typical scales for velocity, pressure, density, and temperature:

\[
v \text{ with } v_0, \Delta v; \quad p \text{ with } p_0, \Delta p; \quad \rho \text{ with } \rho_0, \Delta \rho; \quad T \text{ with } \Delta T; \quad x \text{ with } L; \quad t \text{ with } f^{-1},
\]
where a subscript 0 refers to the primary variable and \( \Delta \) denotes a typical difference. The typical time is written as the inverse of a frequency \( f \). Note that sometimes we have more candidates for a scaling parameter. For example, the typical frequency may be enforced by an external source or instability, or it may be inherited from the intrinsic hydrodynamic frequency \( v_0/L \) or the inverse diffusion time \( \kappa/\rho_0 C_p L^2 \).

Assuming that the relations (6.46) hold, the mass, momentum, and energy equations
\[
\frac{d}{dt} \rho = -\rho \nabla \cdot v, \tag{6.45a}
\]
\[
\rho \frac{d}{dt} v = -\nabla p + \nabla \cdot \tau, \tag{6.45b}
\]
\[
\rho C_p \frac{d}{dt} T = \frac{d}{dt} \rho - \nabla \cdot q + \tau : \nabla v \tag{6.45g}
\]
become, symbolically, after scaling,
\[
f \frac{\rho_0 \Delta p}{L} = \frac{\rho_0 \Delta v}{L},
\]
\[
f \frac{\rho_0 v_0 \Delta v}{L} = \frac{\Delta p}{L}, \frac{\mu \Delta v}{L^2},
\]
\[
f \frac{\rho_0 C_p \Delta T}{L} = \frac{\rho_0 C_p v_0 \Delta T}{L} = \frac{f \Delta p}{L}, \frac{v_0 \Delta p}{L}, \frac{\kappa \Delta T}{L^2}, \frac{\mu (\Delta v)^2}{L^2}.
\]
From the gas laws we know that the typical sound speed squared \( c_0^2 \) scales with both \( P_0/\rho_0 \) and \( \Delta p/\rho \) (not exactly but in order of magnitude), so we can take \( P_0 = \rho_0 c_0^2 \) and \( \Delta \rho = \Delta p/c_0^2 \). Furthermore, we will take here the common situation where \( \Delta v = \mathcal{O}(v_0) \) and \( \kappa \) and \( \mu \) are constants. (This may not always be the case!) With a rescaling such that all coefficients become dimensionless, we have
\[
f \frac{L \Delta p}{v_0 \rho_0 c_0^2} = 1,
\]
\[
f \frac{L}{v_0} = \frac{\Delta p}{\rho_0 v_0^2}, \frac{\mu}{\rho_0 v_0 L},
\]
\[
f \frac{L}{v_0} = \frac{f L \Delta p}{\rho_0 C_p v_0 \Delta T}, \frac{\Delta p}{\rho_0 C_p \Delta T}, \frac{\kappa}{\rho_0 C_p v_0 L}, \frac{v_0 \mu}{\rho_0 C_p L \Delta T}.
\]
From these tables we can collect by inspection all the potentially relevant dimensionless numbers that may occur in problems described by the Navier–Stokes equations.

### 7.4.2 Some Dimensionless Groups with Their Common Names

In a natural way a large number of dimensionless numbers or groups, characterizing a flow, have appeared. Each represents a certain balance between two or more effects. If the typical values used for the scaling are well chosen, these numbers already provide an enormous amount of information about the physical problem considered. A more complete list, derived also from other fields, is given in the appendix, Section M.
Chapter 7. The Art of Modeling

Reynolds: \( Re = \frac{\rho_0 v_0 L}{\mu} \),
Prandtl: \( Pr = \frac{C_P \mu}{\kappa} \),
Mach: \( Ma = \frac{v_0}{c_0} \).

Strouhal: \( Sr = \frac{f L}{v_0} \),
Fourier: \( Fo = \frac{\kappa f^{-1}}{\rho_0 C_p L^2} \),
Euler: \( Eu = \frac{\Delta p}{\rho_0 v_0^2} \).

Helmholtz: \( He = \frac{f L}{c_0} \),
Eckert: \( Ec = \frac{v_0^2}{C_p \Delta T} \),
Péclet: \( Pe = \frac{\rho_0 C_p v_0 L}{\kappa} \).

These dimensionless numbers have the following physical interpretations.

• The Reynolds number \( Re \) describes how important viscous forces are compared to inertial forces and tells us if either viscosity or inertia may be neglected.

• The Prandtl number \( Pr \) describing the relative importance of viscous against heat diffusion, depends only on the material, and is for most gases and fluids of order one.

• The Mach number \( Ma \) compares the occurring velocities with the speed of sound and tells us whether the stationary velocity is so high that compressibility effects should be taken into account.

• The Strouhal number \( Sr \) compares an externally enforced frequency \( f \) with the hydrodynamically induced frequency \( v_0/L \).

• The Fourier number \( Fo \) compares a time scale \( f^{-1} \) with the typical time necessary for the diffusion of heat along a distance \( L \).

• The Euler number \( Eu \) compares the available pressure difference with the typical pressure difference that can be expected from hydrodynamical effects alone.

• The Helmholtz number \( He \) compares the typical wavelength of sound with the size of a scattering object or a source, which tells us a lot about the effectiveness of the scatterer or source.

• The Eckert number \( Ec \) compares the kinetic energy of the flow with available differences in enthalpy.

• The Pécel number \( Pe \) compares forced convection of heat with heat conduction.

Note that a dimensionless number is not always best interpreted by a balance between two effects. In that case the problem at hand may be better described by another selection of numbers. This is not difficult, as many dimensionless numbers are related. For example, \( Pe = Pr Re, He = Sr Ma, \) and \( Sr Fo Pr Re = 1 \).

The scaled equations are now, in dimensionless variables,

\[ Eu Ma^2 \left( Sr \frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \nabla \rho \right) = -\rho \nabla \cdot \mathbf{v}, \quad (7.1a) \]

\[ Sr \rho \frac{\partial \mathbf{v}}{\partial t} + \rho \mathbf{v} \cdot \nabla \mathbf{v} = -Eu \nabla p + \frac{1}{Re} \nabla \cdot \mathbf{\tau}, \quad (7.1b) \]

\[ Sr \rho \frac{\partial T}{\partial t} + \rho \mathbf{v} \cdot \nabla T = Eu Ec \left( Sr \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p \right) + \frac{1}{Pe} \nabla^2 T + \frac{Ec}{Re} \mathbf{\tau} : \nabla \mathbf{v}. \quad (7.1c) \]
7.4. Scaling and Reduction of the Navier–Stokes Equations

7.4.3 Asymptotic Reductions of the Navier–Stokes Equations

When $Ma$, $Re$, $Sr$, $Ec$, etc., become small or large, we may derive from (7.1) various reduced models by assuming the respective terms to vanish or dominate completely. Other terms are sometimes best taken equal to one, in particular if one of the involved scales is inherent (i.e., a result) rather than externally enforced.

A Fourier number $Fo = 1$ corresponds to a time scale $\rho_0 C_P L^2 / k$ determined by heat diffusion. A Strouhal number $Sr = 1$ corresponds to a frequency $f = v_0 / L$, which is connected to hydrodynamical convection processes, e.g., the shedding of vortices from a blunt trailing edge. For a cylinder of diameter $D$, experiments indeed show that the shedding frequency is practically equal to $f_S = 0.2 v_0 / D$. The Euler number $Eu = 1$ corresponds to a present pressure gradient comparable to hydrodynamic pressure variations, $\rho_0 v_0^2$, which is the usual situation in incompressible inviscid (separated) flow around nonstreamlined bodies. We have to select $Eu Re = \Delta p / \rho_0 c_0^2 = 1$ if the pressure gradient is mainly balanced by viscous forces, e.g., in very slow or very viscous flow. The combination $Eu Ma^2 = \Delta p / \rho_0 c_0^2 = 1$ corresponds to a pressure scaled on pressure variations due to compressible effects, which is common for acoustic problems (pressure-density coupling while $Ma$ is small) or high-speed problems ($v_0$ comparable with $c_0$). We select $Eu Ec = \Delta p / C_P \Delta T = 1$ if enthalpy changes are mainly coupled to pressure variations and the flow is at least for the larger part isentropic (cf. (6.44)).

In the following we will derive some important examples of asymptotic models, systematically deduced from the Navier–Stokes equations.

- **Incompressible flow.** When the Mach number $Ma$ tends to zero, while we take $Eu = Sr = 1$, we have incompressible flow described by

  \[
  \nabla \cdot v = 0, \quad \rho \left( \frac{\partial v}{\partial t} + v \cdot \nabla v \right) = -\nabla p + \frac{1}{Re} \nabla \cdot \tau. \tag{7.2}
  \]

  Very often an incompressible flow will have a uniform constant density, but this is not necessary. Note that the energy equation does not disappear but is decoupled from the other equations if viscosity and density may be taken independent of the temperature. The pressure does not play a role thermodynamically, as only its gradient occurs as a reaction force. If the fluid is Newtonian and $\rho$ and $\mu$ are constant (they may be taken equal to one), we obtain for (7.2) the form

  \[
  \nabla \cdot v = 0, \quad \frac{\partial v}{\partial t} + v \cdot \nabla v = -\nabla p + \frac{1}{Re} \nabla^2 v. \tag{7.3}
  \]

  A fully developed laminar flow in a circular pipe, called a Poiseuille flow, satisfies

  \[
  v = U(r) e_z, \quad \nabla p = -Ke_z, \quad \frac{1}{r} \frac{d}{dr} \left( r \frac{d}{dr} U \right) = -Re K.
  \]

  If the duct radius is unity, the solution is described by the parabolic profile

  \[
  U(r) = \frac{1}{4} (1 - r^2) Re K.
  \]

  The equation for mass conservation of (7.3) can be solved by introducing a stream function $\Psi$ defined by $v = \nabla \times \Psi$. This is particularly useful in two-dimensional flow when $\Psi =$
(0, 0, ψ)T and \(v = (ψ_y, -ψ_x, 0)^T\). By taking the curl of the momentum equation, we remove the dependence on pressure and obtain (using (J.9)–(J.12)) the following equation for the transport of vorticity \(\omega := \nabla \times v\):

\[
\frac{\partial \omega}{\partial t} + v \cdot \nabla \omega = \omega \cdot \nabla v + \frac{1}{Re} \nabla^2 \omega.
\]

(7.4)

Note that in two dimensions \(\omega \cdot \nabla v = 0\), so (7.4) becomes a convection-diffusion equation in \(\omega\).

- **Inviscid compressible flow.** If the Reynolds number tends to infinity, usually the Péclet number does too, because \(Pe = Pr Re\) and the Prandtl number is for most fluids and gases of order one. If we further take \(Eu = Sr = Ma = Ec = 1\), we obtain a compressible inviscid flow described by

\[
\begin{align*}
\frac{\partial \rho}{\partial t} + v \cdot \nabla \rho &= -\rho \nabla \cdot v, \\
\rho \left( \frac{\partial v}{\partial t} + v \cdot \nabla v \right) &= -\nabla p, \\
\rho \left( \frac{\partial T}{\partial t} + v \cdot \nabla T \right) &= \frac{\partial p}{\partial t} + v \cdot \nabla p.
\end{align*}
\]

(7.5a–7.5c)

In terms of entropy, the last equation is equivalent to \(\frac{1}{\rho} \frac{\partial}{\partial t} s = 0\), so the flow is isentropic everywhere the assumptions hold.

- **Stokes flow.** If the velocities of a viscous flow are so low that \(Re\) tends to zero, while the time scales are all determined by the flow itself (no external forcing) such that \(Sr\) remains finite, we need to scale the pressure gradients on the inverse Reynolds number in order to have flow at all; i.e., \(Eu\) tends to infinity such that \(Eu Re = 1\). If, in addition, the velocities remain so small compared to the sound speed, i.e., \(Ma \ll 1\), that \(Eu Ma^2\) tends to zero, we obtain the very viscous incompressible Stokes flow given by the Stokes equations

\[
\nabla \cdot v = 0, \quad -\nabla p + \nabla \cdot \tau = 0.
\]

(7.6)

Again, it should be noted that the energy equation is not negligible but only decoupled from the other equations (provided that the viscosity is not temperature dependent).

- **Sound waves.** Consider small pressure-density perturbations in an atmosphere of uniform mean pressure. Assume that the frequencies are relatively high and the typical velocities are small but large enough to neglect viscosity, such that \(Ma \to 0\) while we choose \(Eu = Sr = Ma^{-1}\) and \(Ec = Ma\), and \(Re \) and \(Pe\) remain finite or large. We then retain

\[
\begin{align*}
\frac{\partial \rho}{\partial t} &= -\rho \nabla \cdot v, \\
\frac{\partial v}{\partial t} &= -\nabla p, \\
\rho \frac{\partial T}{\partial t} &= \frac{\partial p}{\partial t}.
\end{align*}
\]

Written in terms of entropy \(s\), the last equation is equivalent to \(\frac{1}{\rho} \frac{\partial}{\partial t} s = 0\). This means that pressure and density perturbations are coupled isentropically by \(\frac{1}{\rho} \frac{\partial}{\partial t} p = c^2 \frac{\partial}{\partial t} \rho\). Noting that
7.4. Scaling and Reduction of the Navier–Stokes Equations

$\rho c^2$ is proportional to the mean pressure, which is constant, we can now eliminate $\rho$ and $v$ to obtain finally the wave equation

$$\frac{\partial^2 p}{\partial t^2} = \nabla \cdot (c^2 \nabla p). \quad (7.7)$$

If $\rho$ and $c$ are constant, we have the usual wave equation with constant coefficients.

- **Convection-diffusion.** In a given flow field with $Sr = 1$, $Pe$ finite, and $EuEc$ and $Ec/Re$ negligible, we get for the temperature the convection-diffusion problem

$$\rho \left( \frac{\partial T}{\partial t} + v \cdot \nabla T \right) = \frac{1}{Pe} \nabla^2 T. \quad (7.8)$$

We end with two important reductions, not immediately obtainable from small-parameter considerations.

- **Potential flow.** If the flow is irrotational, i.e., the vorticity vector is $\omega = \nabla \times v = 0$, a scalar velocity potential $\phi$ may be introduced with

$$v = \nabla \phi.$$ 

For example, in inviscid homentropic flow, any vorticity is convected with the flow (see Kelvin’s theorem, Exercise 6.3d), and if the flow starts out irrotational, it stays that way. In incompressible flow this potential is independent of pressure (except indirectly via boundary conditions) and satisfies Laplace’s equation

$$\nabla^2 \phi = 0. \quad (7.9)$$

In two dimensions an important class of solutions may be generated by using the property of analytic functions $F(z)$ in the complex variable $z = x + iy$ that both their real and imaginary parts satisfy Laplace’s equation. If we introduce the complex potential $F(z) = \phi(x, y) + i\psi(x, y)$, then the velocity $v = (u, v)$ is given by $u - i v = F'(z)$. Note that solutions may be constructed by superposition of elementary solutions (the problem is only nonlinear in pressure). For example, a uniform flow $Uz$ and a dipole source flow $R^2Uz^{-1}$ yield together the flow past a cylinder of radius $R$. As the flow is inviscid, this solution is not unique and any multiple of a line vortex flow $-i \Gamma \ln(z)/2\pi$ may be added to get

$$F(z) = Uz + \frac{R^2U}{z} - i \frac{\Gamma}{2\pi} \ln(z).$$

By itself this solution is not very practical because no high-Reynolds-number flow will pass a cylinder without separating and creating a turbulent wake. It may, however, be a starting point for a larger family of solutions $F(\zeta(z))$ to be obtained by conformal mappings $\zeta \mapsto z$. For example, the Joukowski transformation

$$z = \left( \zeta + \frac{\lambda^2}{\zeta} \right) e^{-i\alpha}$$
maps the circle $|\zeta - \zeta_c| = R$ in $\zeta$ domain to an airfoil in $z$ domain if $\lambda = \xi_c + \sqrt{R^2 - \eta_c^2}$, where \( \zeta_c = \xi_c + i \eta_c \) (see Figure 7.7). (Take e.g., \( \xi_c = -0.03, \eta_c = 0.03, R = 1, \) and $\alpha = 0.05$.) The corresponding flow around the airfoil is given by

$$F(z) = U\zeta e^{-ia} + \frac{R^2 U e^{ia}}{z - \zeta_c} - i \frac{\Gamma}{2\pi} \ln(z - \zeta_c).$$

The undetermined circulation $\Gamma$ is found by requiring the flow to be nonsingular at the trailing edge $\zeta = \lambda$ or $z = 2\lambda e^{-ia}$ (the so-called Kutta condition) and we obtain $\Gamma = -4\pi RU \sin(\alpha + \beta)$, where $\beta = \arcsin(\eta_c/R)$. This condition is a remainder of the effect of viscosity near the trailing edge. Note that when we drop viscosity in our model the no-slip boundary condition cannot be maintained as no solution would exist. However, dropping the no-slip condition altogether is too much and will produce a nonunique solution. It can be shown that for small angles of incidence the inviscid limit yields a condition between slip and no-slip: the no-slip condition can be dropped almost everywhere, except near the trailing edge, where it degenerates to the Kutta condition of nonsingular velocity.

Figure 7.7. A Joukowski airfoil.

- **Bernoulli’s law.** In stationary, incompressible and inviscid flow of constant density $\rho_0$ the momentum equation

$$\rho_0 \left( \frac{\partial}{\partial t} v + v \cdot \nabla v \right) = -\nabla p$$

can be integrated along a streamline, leading to an equation known as Bernoulli’s law that describes conservation of mechanical energy density. By using the vector identity

$$v \cdot \nabla v = \frac{1}{2} |\nabla v|^2 + \omega \times v$$

and noting that $(\omega \times v) \cdot d\ell = 0$, we can integrate along a streamline

$$\int \left[ \frac{1}{2} \rho_0 |v|^2 + \rho_0 \omega \times v + \nabla p \right] \cdot d\ell = \int \nabla \left[ \frac{1}{2} \rho_0 |v|^2 + p \right] \cdot d\ell = 0$$

to get the famous and useful identity

$$\frac{1}{2} \rho_0 |v|^2 + p = \text{constant}. \quad (7.10)$$

Bernoulli’s equation in irrotational flow is valid everywhere rather than only along a streamline. By introducing a potential, we can generalise to unsteady flow with gravity in the $z$ direction as follows:

$$\frac{\partial \psi}{\partial t} + \frac{1}{2} |\nabla \psi|^2 + \frac{p}{\rho_0} + gz = C(t), \quad (7.11)$$

where $C$ is an arbitrary function of time. A typical example is water. Bernoulli’s equation may be generalised to include compressibility if the flow is *barotropic*, i.e., the pressure is
7.5. Discussion

a function of density alone. For example, if we have an irrotational inviscid homentropic perfect gas flow, we obtain, with the result of Exercise 6.3c, that

\[ \frac{\partial \phi}{\partial t} + \frac{1}{2} | \nabla \phi |^2 + \frac{\gamma}{\gamma - 1} \frac{p}{\rho} = C(t), \tag{7.12} \]

where \( \gamma \) is the specific heat ratio and \( C \) is an arbitrary function of time.

7.5 Discussion

• The mathematical solution of a real-world problem starts with the modeling phase, where the problem is described in a mathematical representation of its primitive elements and their relations.

• As the solution is not served by unnecessary complexity, we are interested in an adequate mathematical description with the lowest number of essentially independent parameters and variables. We will call this a model.

• A model may be constructed ad hoc by collecting as many pieces as possible of information and combining them into what we call a building block model. Usually, this is the first type of model if a new area of research is being explored.

Once the foundations are laid, it is not necessary any more to start from scratch, but we can use the complete and comprehensive descriptions that have become available. By removing any unnecessary details, we can simplify these usually overcomplete models to the model required. This is what we call a systematic or asymptotic model.

• A very important aspect in modeling is the identification of a hierarchy of importance to distinguish between the important, the less important, and the unimportant effects. From this hierarchy we may decide which aspects can be included and which can be neglected in the model. An important tool in this respect is the notion of scaling and dimensionless numbers. If this is done properly, it is possible to characterize a problem without any calculation.

• A natural consequence of dimensional analysis is the possibility of similarity solutions. Self-similar solutions that can be constructed using dimensional analysis alone are called self-similar of the first kind. Self-similar solutions of the second kind have incomplete similarity in the dependent variables. They are connected to an eigenvalue problem.

• A sometimes underrated aspect is the question of existence and uniqueness. While making modeling assumptions for a reduced or asymptotic model, we have to make sure that the boundary and related conditions are always consistent with the other equations and accurately reflect the physics of the problem. Examples have been given in the above. When we neglect the effects of viscosity, the no-slip boundary condition of no-slip has to be dropped, but sometimes this is too drastic a simplification, and amendments (like the Kutta condition or the causality condition) have to be made that were not necessary in the higher-level model. The same is true for waves scattered
at edges that are simplified to sharp edges. In problems of sound with mean flow the
edge condition and the Kutta condition may occur in combination [121, 122]. There
is no doubt that many more examples exist from other fields of application.

- Although modeling is the basis of any application of mathematics, there are not really
many books that deal with this aspect from a more general point of view. A few
representative texts are [152, 47, 33, 91].

**Exercises**

7.1. Simplify, by suitable scaling of the variables, the Korteweg–de Vries equation

\[
\frac{\partial u}{\partial t} + c \left(1 + \frac{3}{2}u\right) \frac{\partial u}{\partial x} + \frac{1}{6}c^2h^2 \frac{\partial^3 u}{\partial x^3} = 0
\]

and the linearized Boussinesq equation

\[
\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} - \frac{1}{3}h^2 \frac{\partial^4 u}{\partial x^2 \partial t^2} = 0
\]

such that the coefficients become equal to one.

7.2. Nondimensionalise the telegraph equation (⋆) in Example 1.4,

\[
\frac{\partial^2 u}{\partial t^2} + (a + b) \frac{\partial u}{\partial t} + abu - c^2 \frac{\partial^2 u}{\partial x^2} = 0,
\]

such that the resulting problem depends on one dimensionless number only.

7.3. The drag \(D\) of a moving ship, due to viscous effects and wave generation, depends
on its length \(L\) and velocity \(V\), water viscosity \(\mu\), gravity acceleration \(g\) and water
density \(\rho\). The dimensional units are \([D]\) = kg \cdot m/s², \([L]\) = m, \([V]\) = m/s, \([\mu]\) =
kg/(m \cdot s), \([g]\) = m/s², and \([\rho]\) = kg/m³. By how many dimensionless groups is
the problem completely described? Give an example of such a set of dimensionless
groups (these are not uniquely defined).

7.4. Is the self-similar solution of Example 7.18 of the first or of the second kind?

7.5. Analyse Example 7.19 in terms of Buckingham’s theorem. Verify the dimensional
groups, including their number. Note that \([\sigma]\) = \(A^2 \cdot s^3/(m^3 \cdot kg)\), \([\rho C]\) = kg/(m \cdot
s² \cdot K), \([\psi]\) = kg \cdot m²/(s³ \cdot A), and \([\kappa]\) = kg \cdot m/(s³ \cdot K).

7.6. Reconsider Example 7.1. Make the problem complete by adding boundary and initial
conditions. Make the problem dimensionless by scaling on the inherent time and
length scales. Determine conditions, in terms of a dimensionless number, for which
the diffusion term can be neglected. Note that the order of the differential equation
is then reduced from two to one. What are the consequences for the boundary and/or
initial conditions?

7.7. Consider (†) of Example 7.4 to describe a stationary suspended flexible bar of length
\(L\). Make the problem complete by adding suitable boundary conditions at \(x = 0\) and
Exercises 157

\[ x = D, \] where \( 1 - D/L \) is positive and not small. Note the two integration constants \( H_0 \) and \( V_0 \), so we need four conditions. Make the problem dimensionless by scaling lengths on \( L \) and forces on \(QL\). Under what condition, in terms of a dimensionless parameter, can we neglect bending stiffness (i.e., the term multiplying \( \phi_{ss} \))? The result describes a cable with vanishing bending stiffness, or a catenary. What are the consequences for the boundary conditions? Solve this equation.

7.8. Material of concentration \( c \) is diffused from a container located at \( |x| \leq a \) through a membrane at \( |x| = a \) into the outer medium \( |x| > a \). In the container and in the medium the diffusion is described by

\[ \frac{\partial c}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial c}{\partial x} \right). \]

The diffusion coefficient in the container is \( D = D_i \); in the outer medium it is \( D = D_o \). Initially, \( c = 0 \) for \( |x| > a \) and \( c = c_0 \) in \( |x| \leq a \). At the interfaces \( x = \pm a \), \( c \) and \( Dc \) are continuous.

(a) Describe the problem in dimensionless variables.

(b) Approximate the solution for \( t \) large (which is the same as for \( a \to 0 \)). Hint: Use the fact that eventually the majority of the material is diffused to the outer medium, while conservation of mass requires that \( \int_{-\infty}^{\infty} c(x,t) \, dx = \) constant. Then derive a similarity solution. Note the symmetry in \( x \).

(c) Do the same for the analogous problems in two and three dimensions by utilising Chapter 10, Section 10.3.

7.9. We are interested in knowing at what distance \( D \) a boat of height \( H \) is still visible above the horizon. Criticize the following (incorrect) model. The height of a person is negligible compared to Earth’s radius \( R \). So our field of visibility is just in the tangent plane of Earth’s at the position of the observer. A boat is visible in this plane if \( \cos \left( \frac{D}{R} \right) \geq \frac{R}{R+a} \). Since \( D/R \) and \( H/R \) are small, this is equivalent to \( D^2 \leq 2RH \).

7.10. A simple model for the temperature \( T \) in the ground, at time \( t \) and depth \( z \), is

\[ \rho C \frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial z^2}, \quad -\infty < z \leq 0, \]

where \( T(0, t) = T_0(t) \) is the given air temperature at the surface \( z = 0 \). \( \rho \) denotes the density, \( C \) the specific heat, and \( k \) the thermal conductivity of the soil. For dry sand \( \rho = 1600 \text{ kg/m}^3 \), \( C = 800 \text{ J/(kg} \cdot \text{K}) \), and \( k = 0.3 \text{ W/(m} \cdot \text{K}) \), and for saturated sand \( \rho = 2000 \text{ kg/m}^3 \), \( C = 1480 \text{ J/(kg} \cdot \text{K}) \), \( k = 2.2 \text{ W/(m} \cdot \text{K}) \). Estimate, by scaling, the typical penetration depth of the yearly temperature variations (steady state). Compare this with an exact solution if \( T_0 \) varies harmonically like \( T_0(t) = A + B \cos(\omega t) \), where \( \omega = 2\pi (1 \text{ year})^{-1} \), and \( T \) is in steady state. If the temperature for \( z \to -\infty \) is below 0°C, we call this soil permafrost.

7.11. A reasonably complete model of a kettle drum consists of a cavity in free space, with acoustic pressure perturbations \( p(x, t) = \text{Re}(\hat{p}(x) e^{i\omega t}) \) inside and outside the cavity given by

\[ \nabla^2 \hat{p} + \kappa^2 \hat{p} = 0, \quad i\omega\rho_0 \hat{v} + \nabla \hat{p} = 0, \]
where $\kappa = \omega/c_0$, $\rho_0$ denotes the mean air density, and $\hat{v}$ represents the corresponding velocity perturbations. The cavity is hard walled on all sides ($\hat{v} \cdot n = 0$) except one, which is closed by an elastic membrane (tension $T$, mass density $\sigma$). The membrane displacement $\hat{\eta}$ is driven by, and drives, the pressure difference across the membrane:

$$T \nabla^2 \hat{\eta} + \omega^2 \sigma \hat{\eta} = \hat{p}_{\text{upper}} - \hat{p}_{\text{lower}}.$$  

As the air follows the membrane, the normal velocity $\hat{v} \cdot n$ at both sides of the membrane is equal to $i \omega \hat{\eta}$. The basic question of interest is to determine the spectrum of (complex) resonance frequencies.

Consider the following canonical model of the kettle drum problem. A semi-infinite pipe ($0 \leq x < \infty$) of radius $a$, closed at $x = 0$, contains a piston-like element at $x = L$ (modeling the membrane), which is driven by the pressure difference across $x = L$ and kept in position by a spring:

$$\frac{\partial^2 \hat{p}}{\partial x^2} + \kappa^2 \hat{p} = 0 \quad \text{for} \quad x \in (0, L) \cup (L, \infty)$$

$$-\frac{8}{a^2} T \hat{\eta} + \omega^2 \sigma \hat{\eta} = \hat{p}(L^+) - \hat{p}(L^-) \quad \text{at} \quad x = L$$

$$\frac{\partial \hat{p}}{\partial x} = 0 \quad \text{at} \quad x = 0, \quad \frac{\partial \hat{p}}{\partial x} = \omega^2 \rho_0 \hat{\eta} \quad \text{at} \quad x = L, \quad \text{outgoing waves for} \quad x \to \infty.$$  

Determine the equation for $\omega$, solve this for some simple cases, and try to indicate the general solution graphically in the complex $\omega$ plane for dimensionless groups of parameters. Are there solutions with $\text{Im}(\omega) = 0$? How are these to be interpreted physically?
Chapter 8

The Analysis of Elliptic Equations

In this chapter we discuss analytical methods for elliptic equations. We define in Section 8.1 several boundary value problems for elliptic equations and investigate the uniqueness of the solution. The concepts of eigenvalues and eigenfunctions of an elliptic boundary value problem are introduced in Section 8.2. An important analytical solution method for linear elliptic equations is separation of variables, which is discussed in Section 8.3. In Section 8.4 we introduce the fundamental solution of the Poisson equation, which is a solution of the Poisson equation with a Dirac delta function as source term. If, furthermore, the fundamental solution satisfies certain homogeneous boundary conditions, it is called a Green’s function. Next, in Section 8.5, we derive integral representations for the solution of elliptic boundary value problems using these Green’s functions. A qualitative description of the solution of the Poisson equation is based on the maximum principle and is presented in Section 8.6. As an application of elliptic equations, we study in Section 8.7 the Stokes equations, which describe creeping flow. In particular, we compute the (slow) flow around a sphere.

8.1 The Laplace Operator

Elliptic differential equations typically occur in problems that describe stationary situations; i.e., time has no explicit role. The simplest and most well known elliptic equation is the Laplace equation, defined on a domain \( \Omega \subset \mathbb{R}^d \) (\( d = 1, 2, 3 \)), say:

\[
\mathcal{L}[u] := \nabla^2 u = 0, \quad x \in \Omega. \tag{8.1}
\]

In the inhomogeneous case we have the Poisson equation

\[
\mathcal{L}[u] := \nabla^2 u = f(x), \quad x \in \Omega. \tag{8.2}
\]

A further type that is often encountered is the Helmholtz equation, which is actually related to the eigenvalue problem of (8.1):

\[
\mathcal{L}[u] := \nabla^2 u - \lambda u = 0, \quad x \in \Omega, \quad \lambda \in \mathbb{R}. \tag{8.3}
\]
The Laplace operator $\nabla^2$ also occurs quite often in time-dependent problems like the heat equation or the wave equation. This provides for additional interest in investigating problems like (8.1), (8.2), and (8.3).

### 8.1.1 Problem Types

The Laplace operator typically occurs in situations where the flux $f$ of a variable is proportional to its gradient. As in Example 1.1 we may consider a concentration $c$ that causes a flow to areas with lower concentration; i.e.,

$$f = -D\nabla c,$$  \hspace{1cm} \text{(8.4)}

with $D > 0$ the diffusion coefficient. If we apply Gauss’s divergence theorem (see (J.13) in the appendix) to an arbitrary volume $W \subset \Omega$, we find

$$-\int_W \nabla \cdot (D\nabla c) \, dV = -\oint_{\partial W} (D\nabla c) \cdot n \, dS = \oint_{\partial W} f \cdot n \, dS. \hspace{1cm} \text{(8.5)}$$

Without sources or sinks, the net flux through $\partial W$ should be 0. If, furthermore, the diffusion coefficient $D$ is constant, we obtain (8.1). Any solution in $C^2(\Omega)$ of (8.1) is called a **harmonic function**.

In order to define a solution more precisely, we have to specify a **boundary condition**. Three common cases are distinguished for $x \in \partial \Omega$:

$$u(x) = a(x) \hspace{1cm} \text{(Dirichlet),} \hspace{1cm} \text{(8.6a)}$$

$$\frac{\partial u}{\partial n}(x) = b(x) \hspace{1cm} \text{(Neumann),} \hspace{1cm} \text{(8.6b)}$$

$$\alpha u(x) + \beta \frac{\partial u}{\partial n}(x) = c(x), \hspace{0.5cm} \alpha, \beta \neq 0 \hspace{1cm} \text{(Robin),} \hspace{1cm} \text{(8.6c)}$$

with $n$ the outward unit normal on $\partial \Omega$. $\frac{\partial}{\partial n}$ denotes the normal derivative; i.e., $\frac{\partial}{\partial n} u := n \cdot \nabla u$. Condition (8.6a) is called a **Dirichlet boundary condition**, (8.6b) is called a **Neumann boundary condition**, and (8.6c) is called a **Robin boundary condition**. We can easily establish the uniqueness of a solution of (8.1) and either one of the two boundary conditions (8.6a) or (8.6c). It is instructive to illustrate these problems for the one-dimensional case, where the Poisson equation reduces to an ODE.

**Example 8.1** Consider the two-point Dirichlet boundary value problem

$$L[u] := \frac{d^2 u}{dx^2} = f(x), \quad x \in (0, \pi),$$

$$u(0) = u(\pi) = 0,$$

where $f$ is piecewise smooth on $(0, \pi)$. We now use Fourier sine series to find the solution in a formal way. We take the ansatz

$$u(x) = \sum_{k=1}^{\infty} a_k \sin(kx),$$
8.1. The Laplace Operator

which has the advantage that \( u \) satisfies the boundary conditions identically if the series converges uniformly and the found solution is continuous at the endpoints (appendix, Section C). This is to be verified afterward. We also expand \( f(x) \) in a Fourier sine series

\[
f(x) = \sum_{k=1}^{\infty} f_k \sin(kx),
\]

where

\[
f_k = \frac{2}{\pi} \int_{0}^{\pi} f(x) \sin(kx) \, dx, \quad k = 1, 2, \ldots,
\]

and \( f_k \to 0 \) for \( k \to \infty \). We then find

\[
\sum_{k=1}^{\infty} (k^2a_k + f_k) \sin(kx) = 0.
\]

From the uniqueness of the Fourier coefficients of the null function it follows that \( k^2a_k = -f_k \), yielding the solution

\[
u(x) = -\sum_{k=1}^{\infty} \frac{f_k}{k^2} \sin(kx), \quad x \in (0, \pi),
\]

which is indeed uniformly convergent.

**Example 8.2** Consider the two-point Neumann boundary value problem

\[
\mathcal{L}[u] := \frac{d^2u}{dx^2} = f(x), \quad x \in (0, \pi),
\]

\[
\frac{du}{dx}(0) = \frac{du}{dx}(1) = 0,
\]

where \( f \) and \( f' \) are both piecewise smooth on \( (0, \pi) \). Because of the form of the boundary conditions, it seems advantageous to expand the solution \( u(x) \) in a Fourier cosine series as

\[
u(x) = b_0 + \sum_{k=1}^{\infty} b_k \cos(kx).
\]

If \( u' \) converges uniformly (to be verified afterward), the boundary conditions are automatically satisfied. Likewise, we have

\[
f(x) = f_0 + \sum_{k=1}^{\infty} f_k \cos(kx),
\]

with coefficients \( f_k \) given by

\[
f_0 = \frac{1}{\pi} \int_{0}^{\pi} f(x) \, dx, \quad f_k = \frac{2}{\pi} \int_{0}^{\pi} f(x) \cos(kx) \, dx, \quad k = 1, 2, \ldots.
\]

Note that \( f_k = O(k^{-1}) \) for \( k \to \infty \) (see Corollary 3.7). Substituting these expansions into the differential equation, we find the relation

\[
\sum_{k=0}^{\infty} (k^2b_k + f_k) \cos(kx) = 0,
\]
so that $k^2 b_k + f_k = 0$. In particular, $f_0 = 0$, implying that $f(x)$ should satisfy the consistency condition

$$\int_0^\pi f(x) \, dx = 0$$

stating that the average value of $f(x)$ over $(0, \pi)$ vanishes. As a consequence, $b_0$ is undetermined and the solution reads

$$u(x) = b_0 - \sum_{k=1}^{\infty} \frac{f_k}{k^2} \cos(kx),$$

with $b_0$ arbitrary. Thus $u(x)$ is determined up to an additive constant. Since $f_k/k = O(k^{-2})$, it is readily verified that $u'$ converges uniformly and satisfies the boundary conditions.

Example 8.3 Consider the two-point Helmholtz boundary value problem

$$\mathcal{L}[u] := \frac{d^2 u}{dx^2} - \lambda u = 0, \quad x \in (0, \pi),$$

$$u(0) = u(\pi) = 0.$$  

To solve this linear ODE in a formal way, we look for solutions of the form $u(x) = e^{\mu x}$. Substituting this solution into the differential equation, we find that $\mu$ should satisfy the characteristic equation

$$\mu^2 - \lambda = 0.$$  

So as a general solution we obtain

$$u(x) = A e^{\sqrt{\lambda} x} + B e^{-\sqrt{\lambda} x}, \quad A, B \in \mathbb{R}.$$  

Using the boundary conditions, we find

$$A + B = 0, \quad A e^{\sqrt{\lambda} \pi} + B e^{-\sqrt{\lambda} \pi} = 0.$$  

For arbitrary $\lambda$ this system has only the trivial solution $A = B = 0$. Nontrivial solutions (the eigenvalue problem) exist if the determinant of the system is zero:

$$\begin{vmatrix} 1 & 1 \\ e^{\sqrt{\lambda} \pi} & e^{-\sqrt{\lambda} \pi} \end{vmatrix} = e^{-\sqrt{\lambda} \pi} - e^{\sqrt{\lambda} \pi} = 0.$$  

This relation implies that $e^{\sqrt{\lambda} \pi} = 1$, which has the solutions $2\sqrt{\lambda} \pi = k 2\pi i$ ($k = 0, 1, 2, \ldots$). Apparently, the only possible values of $\lambda$ that allow solutions are given by

$$\lambda = \lambda_k = -k^2, \quad k = 0, 1, 2, \ldots.$$  

Choosing $A = -B = -\frac{1}{2} i$, we find the corresponding solutions

$$u(x) = u_k(x) = \sin(kx), \quad k = 0, 1, 2, \ldots.$$  

So we have either no solutions or infinitely many solutions. Note that $k = 0$ corresponds to the trivial solution $u_0(x) \equiv 0$ and should therefore be discarded.
8.1.2 Uniqueness

One can simply investigate the uniqueness of the Laplace equation (8.1) for a solution satisfying either one of the boundary conditions (8.6). This is done in the next theorem.

**Theorem 8.4.** A harmonic function satisfying the Dirichlet boundary condition (8.6a) is unique. A harmonic function satisfying the Neumann boundary condition (8.6b) is unique up to an additive constant. If \( \text{sign}(\alpha) = \text{sign}(\beta) \), then a harmonic function satisfying the Robin boundary condition (8.6c) is unique.

**Proof.** First, consider the boundary condition (8.6a) with \( a(x) \equiv 0 \). Using the first identity of Green (J.16), we obtain

\[
\int_{\Omega} (u \nabla^2 u + \nabla u \cdot \nabla u) \, dV = \int_{\partial \Omega} u \frac{\partial u}{\partial n} \, dS,
\]

so that

\[
\int_{\Omega} |\nabla u|^2 \, dV = 0, \tag{*}
\]

whence \( u \) is constant in \( \Omega \). Because of continuity we conclude that \( u(x) \equiv 0 \). If we now have two harmonic solutions, \( u_1(x) \) and \( u_2(x) \) say, both satisfying (8.6a), then obviously \( u_1(x) - u_2(x) \equiv 0 \) in \( \Omega \), which implies uniqueness. For the boundary condition (8.6b) with \( b(x) \equiv 0 \) we derive (*) for a harmonic function \( u(x) \) in a similar manner. This trivially implies \( u(x) \) to be constant in \( \Omega \). If \( u_1(x) \) and \( u_2(x) \) are two harmonic functions satisfying the boundary condition (8.6b), then \( u_1(x) - u_2(x) \) can be identified with \( u(x) \), as before. Hence a solution is unique, apart from an additive constant. Finally, for a harmonic function \( u(x) \) satisfying the boundary condition (8.6c) we obtain

\[
\int_{\Omega} |\nabla u|^2 \, dV = -\int_{\partial \Omega} \frac{\alpha}{\beta} u^2 \, dS \leq 0.
\]

Obviously this can only be true if \( u(x) \equiv 0 \). Uniqueness then follows in the same fashion as for the first case. \( \square \)

8.2 Eigenvalues and Eigenfunctions

Eigenvalue problems play an important role, either directly, like the Helmholtz equation, or indirectly, when we characterise a PDE. In this section we will mainly consider the latter aspect. We first consider the one-dimensional case.

8.2.1 The One-Dimensional Eigenvalue Problem

Probably the simplest eigenvalue problem is given by

\[
\frac{d^2u}{dx^2} = \lambda u, \quad x \in (0, a), \tag{8.7a}
\]

\[
u(0) = u(a) = 0. \tag{8.7b}
\]
Note that (8.7a) is just the one-dimensional Helmholtz equation. Parameter $\lambda$ is not prescribed but is part of the solution. It is called an eigenvalue and its value is determined by the condition that (8.7) has a nontrivial solution. The general solution of (8.7a) can be written like

$$u(x) = A \, e^{\sqrt{\lambda} x} + B \, e^{-\sqrt{\lambda} x},$$

(8.8)

where $A$ and $B$ follow from the boundary conditions (8.7b). We find the equations

$$A + B = 0, \quad A \, e^{\sqrt{\lambda} a} + B \, e^{-\sqrt{\lambda} a} = 0.$$

(8.9)

This homogeneous system only has a nontrivial solution if the following determinant is zero:

$$\begin{vmatrix} 1 & 1 \\ e^{\sqrt{\lambda} a} & e^{-\sqrt{\lambda} a} \end{vmatrix} = e^{-\sqrt{\lambda} a} - e^{\sqrt{\lambda} a} = 0.$$  

(8.10)

We conclude that $e^{\sqrt{\lambda} a} = 1 = e^{2k\pi i}$, and the eigenvalues are found to be

$$\lambda = \lambda_k = -\left( \frac{k\pi}{a} \right)^2, \quad k = 0, 1, 2, \ldots.$$  

(8.11)

Since $B = -A$, we find for the corresponding eigenfunctions

$$u(x) = u_k(x) = \sin \left( \frac{k\pi x}{a} \right), \quad k = 0, 1, 2, \ldots.$$  

(8.12)

Any multiple of $u_k(x)$ is an eigenfunction as well. Note that $u_0(x) \equiv 0$ and therefore $\lambda_0$ and $u_0(x)$ should be discarded as trivial solutions.

The inner product $(f, g)$, defined in (3.1), reduces for real-valued functions to

$$(f, g) := \int_0^a f(x) g(x) \, dx.$$  

(8.13)

It is readily seen that $(u_k, u_l) = 0$ for $k \neq l$. In other words, the eigenfunctions $u_k$ are orthogonal with respect to the inner product (8.13). We can put this in a more formal setting by considering the operator

$$\mathcal{L}[u] := \frac{d^2}{dx^2} u.$$  

(8.14)

If we apply integration by parts twice, we find for a suitable “test function” $v = v(x)$ with $v(0) = v(a) = 0$ the relation

$$(\mathcal{L}[u], v) = \int_0^a \frac{d^2 u}{dx^2}(x)v(x) \, dx = \int_0^a u(x)\frac{d^2 v}{dx^2}(x) \, dx = (u, \mathcal{L}[v]).$$  

(8.15)

This is a special property of this operator. More specifically, we define the adjoint operator $A^*$ of an operator $A$ by (cf. Chapter 4)

$$(u, A[v]) = (A^*[u], v).$$  

(8.16)
8.2. Eigenvalues and Eigenfunctions

Then we see from (8.15) that \( \mathcal{L} \) defined in (8.14) has the symmetry property
\[
(\mathcal{L}[u], v) = (u, \mathcal{L}[v]) = (\mathcal{L}^*[u], v).
\]
(8.17)

We therefore call this \( \mathcal{L} \) self-adjoint. More generally, if we have the operator \( \mathcal{L} \) defined by
\[
\mathcal{L}[u] := \frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u(x),
\]
(8.18)
where \( p \) and \( q \) are real, then \( \mathcal{L} \) is self-adjoint. We leave it as an exercise to see that an operator with an explicit first derivative term is not self-adjoint. There is a remarkable similarity with symmetric matrices. Indeed, symmetric matrices have an orthogonal system of eigenvectors that correspond to real eigenvalues. This similarity is often exploited in numerical approaches that try to preserve the self-adjointness of (8.18) in a discrete form; see also Chapter 9.

We can also derive the following (cf. (8.15)). Let \( \lambda \) be an eigenvalue of (8.18) with homogeneous boundary conditions and \( u \) be the corresponding eigenfunction. Then
\[
\lambda (u, u) = (\mathcal{L}[u], u) = \int_0^a \left( -p(x) \left| \frac{du}{dx} \right|^2 + q(x)|u|^2 \right) dx.
\]
(8.19)
Hence \( \lambda \) is real, which is in agreement with the eigenvalues of (8.7) and is in line with what we know for symmetric matrices. Moreover, we see that \( \lambda < 0 \) for \( p > 0 \) and \( q \leq 0 \). With these requirements on \( p \) and \( q \) we have found the analogue of what we call negative definite symmetric matrices. (A matrix \( C \) is negative definite if \( z^T C z < 0 \) for any \( z > 0 \).)

Remark. Sometimes one rather prefers to use the operator \( -\mathcal{L} \) in order to have strictly positive eigenvalues.

Example 8.5 Consider the operator \( \mathcal{L} \), with
\[
\mathcal{L}[u] := \frac{d^2 u}{dx^2} + qu, \quad q \in \mathbb{R},
\]
and let \( u(0) = u(a) = 0 \). If we try to find the eigenvalues as we did for (8.7), we obtain
\[
\lambda_k = -\left( \frac{k\pi}{a} \right)^2 + q, \quad k = 1, 2, \ldots.
\]
Clearly, all eigenvalues are negative if \( q < (\pi/a)^2 \), which is slightly more relaxed than the requirement \( q \leq 0 \) that we used before.

Example 8.6 If we have purely Neumann boundary conditions, we get eigensolutions different from (8.11) and (8.12). Thus consider the boundary value problem
\[
\frac{d^2 u}{dx^2} = \lambda u, \quad x \in (0, a),
\]
\[
\frac{du}{dx}(0) = \frac{du}{dx}(a) = 0.
\]
Analogously to the derivation of the eigensolution (8.11) and (8.12), we find a general solution of the form (8.8), where the coefficients \( A \) and \( B \) have to satisfy the equations
\[
A - B = 0, \quad A e^{\sqrt{\lambda} a} - B e^{-\sqrt{\lambda} a} = 0.
\]
We again find the eigenvalues 
\[ \lambda_k = -\left(\frac{k\pi}{a}\right)^2, \quad k = 0, 1, 2, \ldots \]
but now they correspond to the eigenfunctions 
\[ u_k(x) = \cos\left(\frac{k\pi x}{a}\right), \quad k = 0, 1, 2, \ldots. \]

Note that \( u_0(x) \equiv 1 \) corresponds to any constant that can be added to the solution of the Neumann problem for the Poisson equation, illustrating that this solution is not unique.

There is a close relationship between eigenfunctions and Fourier series (cf. Chapter 3). If we expand a function \( f \) with \( f(0) = f(a) = 0 \) on \([0, a]\), then we obtain the Fourier sine series
\[ f(x) = \sum_{k=1}^{\infty} f_k \sin\left(\frac{k\pi x}{a}\right), \quad (8.20) \]
so \( f(x) \) is actually expanded in terms of eigenfunctions of (8.7). This property holds more generally for eigenfunctions and has a host of consequences. An appropriate setting for this is variational calculus, which is outside the scope of this text.

### 8.2.2 Eigenvalue Problems in Several Dimensions

The analysis of eigenvalue problems can be extended to higher dimensions. The starting point is a self-adjoint generalization of the Helmholtz equation defined on a domain \( \Omega \) with homogeneous Dirichlet boundary condition; i.e.,
\[ \mathcal{L}[u] := \nabla^2 u + qu = \lambda u, \quad x \in \Omega; \quad (8.21a) \]
\[ u(x) = 0, \quad x \in \partial \Omega. \quad (8.21b) \]

We first introduce an inner product on \( \Omega \). Let \( u \) and \( v \) be defined on \( \Omega \) and satisfy the homogeneous Dirichlet boundary condition as in (8.21b). Then
\[ (u, v) := \int_{\Omega} uv \, dV. \quad (8.22) \]

Our first goal is to show that \( \mathcal{L} \) is self-adjoint. We have, using the second identity of Green, that
\[ (\mathcal{L}[u], v) = \int_{\Omega} (\nabla^2 u + qu) v \, dV = \int_{\Omega} u (\nabla^2 v + q v) \, dV = (u, \mathcal{L}[v]) = (\mathcal{L}^*[u], v). \quad (8.23) \]

Note that we have used that \( v = u = 0 \) on \( \partial \Omega \). If we now let \( u \) be an eigensolution corresponding to \( \lambda \) (presupposing its existence), then we find by Green’s first identity that
\[ \lambda (u, u) = (\mathcal{L}[u], u) = \int_{\Omega} (\nabla^2 u + qu) u \, dV = -\int_{\Omega} (|\nabla u|^2 - qu^2) \, dV. \quad (8.24) \]

Hence we see that \( \lambda \in \mathbb{R} \). Moreover, if \( q < 0 \), then \( \lambda < 0 \).
8.2. Eigenvalues and Eigenfunctions 167

Now let \( u_k \) and \( u_l \) be eigenfunctions corresponding to the eigenvalues \( \lambda_k \) and \( \lambda_l \), respectively, with \( \lambda_k \neq \lambda_l \). Then

\[
u_l \mathcal{L}[u_k] - u_k \mathcal{L}[u_l] = u_l \nabla^2 u_k - u_k \nabla^2 u_l = (\lambda_k - \lambda_l) u_k u_l. \tag{8.25}\]

After integration we obtain from (8.25) that

\[
\int_{\Omega} (u_l \nabla^2 u_k - u_k \nabla^2 u_l) \, dV = (\lambda_k - \lambda_l) \int_{\Omega} u_k u_l \, dV. \tag{8.26}\]

For the left-hand side in (8.26), we can apply Green's second identity, which then results in

\[
\oint_{\partial \Omega} (u_l \frac{\partial u_k}{\partial n} - u_k \frac{\partial u_l}{\partial n}) \, dS = (\lambda_k - \lambda_l) \int_{\Omega} u_k u_l \, dV. \tag{8.27}\]

Clearly, the integral on the left in (8.27) is zero, from which we derive

\[
(u_k, u_l) = \int_{\Omega} u_k u_l \, dV = 0; \tag{8.28}\]

i.e., the eigensolutions form an orthogonal set.

For a general domain \( \Omega \) one has to use numerical methods to find (an approximation of) the eigenvalues. If we have a rectangular domain, say \( \Omega = \{(x, y) \in \mathbb{R}^2 | 0 \leq x \leq a, 0 \leq y \leq b\} \), we can simply obtain the eigensolutions from what we found for the one-dimensional case when \( q(x) = 0 \). It is left as an exercise to show that

\[
\lambda_{k,l} = -\pi^2 \left( \frac{k^2}{a^2} + \frac{l^2}{b^2} \right), \tag{8.29a}\]

\[
u_{k,l}(x, y) = \sin \left( \frac{k \pi x}{a} \right) \sin \left( \frac{l \pi y}{b} \right) \quad (k, l = 1, 2, \ldots) \tag{8.29b}\]

are eigenvalues and eigenfunctions, respectively. We conclude this section with a general property, which we give without proof (see, e.g., [36]).

**Property 8.7.** Let the area of a two-dimensional domain be \( A \). Then there exists an infinite sequence of eigenvalues of the Laplace equation. They can be ordered such that the \( m \)th eigenvalue \( \lambda_m \) satisfies, for large \( m \),

\[
\lambda_m \sim -\frac{4 \pi m}{A}. \tag{8.30}\]

**Example 8.8** Consider a Dirichlet problem on the unit square \( \Omega := \{(x, y) \in \mathbb{R}^2 | 0 < x < 1, 0 < y < 1\} \). From (8.29a) we see that \( \lambda_{k,l} = -\pi^2 (k^2 + l^2) \). Let \( m(\lambda) \) be the number of eigenvalues, still larger than a fixed negative \( \lambda \), or, equivalently, the number of index pairs \( (k, l) \) satisfying \( k^2 + l^2 < |\lambda| / \pi^2 \). We can relabel the eigenvalues \( \{\lambda_{k,l}\} \) as \( \{\mu_m\} \) by letting \( \lambda_{k,l} \) gradually decrease, starting from \( \lambda_{1,1} \). In Figure 8.1 we have indicated the set of index pairs \( \{(k, l)\} \) that lie in the first quadrant of a circle with radius \( 1 / \pi \sqrt{2} \). The area of this quarter circle is equal to \( |\lambda| / 4 \pi \). Hence

\[
m(\lambda) \sim \frac{|\lambda|}{4 \pi}, \quad \mu_m \sim -4 \pi m \text{, in agreement with (8.30).} \]

Chapter 8. The Analysis of Elliptic Equations

8.3 Separation of Variables

The method of separation of variables is useful for linear problems with constant coefficients and homogeneous boundary conditions. In this section we will apply the method to the two-dimensional Laplace equation (8.1). The method is based on the assumption that the solution \( u = u(x, y) \) can be written as a product of a function \( v \), say, depending solely on \( x \), and a function \( w \), say, depending solely on \( y \); i.e.,

\[
u(x, y) = v(x)w(y). \tag{8.31}\]

If we substitute this in (8.1), we obtain

\[
\frac{d^2 v}{dx^2}(x)w(y) = -v(x)\frac{d^2 w}{dy^2}(y).
\]

Dividing by \( v(x)w(y) \), we can write this as

\[
\frac{1}{v(x)}\frac{d^2 v}{dx^2}(x) = -\frac{1}{w(y)}\frac{d^2 w}{dy^2}(y). \tag{8.32}
\]

Since the left-hand side in (8.32) is a function of \( x \) only and the right-hand side is a function of \( y \) only, they must be independent of both, i.e., constant. Let us denote this constant by \( \lambda \). Then we apparently have the two eigenvalue problems

\[
\frac{d^2 v}{dx^2}(x) = \lambda v(x), \tag{8.33a}
\]

\[
\frac{d^2 w}{dy^2}(y) = -\lambda w(y). \tag{8.33b}
\]

The constant \( \lambda \) is called the separation constant. Let us now assume that \( \Omega \) is the unit square, i.e., \( 0 < x, y < 1 \), and that, e.g., the following boundary conditions are given:

\[
u(0, y) = u(1, y) = 0, \quad u(x, 0) = u(x, 1) = x(1 - x). \tag{8.34}\]
8.3. Separation of Variables

From the homogeneous boundary conditions we conclude that \( v(0) = v(1) = 0 \) and so (8.33a) is a genuine eigenvalue problem. We then find that the eigenvalues \( \lambda_k \) are given by

\[
\lambda_k = -k^2 \pi^2, \quad k = 1, 2, \ldots,
\]

and the corresponding eigenfunctions are given by

\[
v_k(x) = \sin(k \pi x), \quad k = 1, 2, \ldots; \tag{8.35b}
\]

cf. Section 8.2.1. The functions \( w(y) = w_k(y) \) corresponding to \( \lambda_k \) can be determined from the ODE (8.33b) and we find

\[
w_k(y) = \alpha_k e^{k\pi y} + \beta_k e^{-k\pi y}, \quad k = 1, 2, \ldots, \tag{8.36}
\]

for some \( \alpha_k, \beta_k \). In order to determine the desired solution \( u(x, y) \), we apply the superposition principle; i.e., we assume that

\[
u(x, t) = \sum_{k=1}^{\infty} v_k(x) w_k(y) = \sum_{k=1}^{\infty} \left( \alpha_k e^{k\pi y} + \beta_k e^{-k\pi y} \right) \sin(k \pi x). \tag{8.37}
\]

This superposition is possible since the PDE is linear. Note that (8.37) is in fact a Fourier-sine series in the \( x \) variable, so that the homogeneous boundary conditions are automatically satisfied if the series converges uniformly. This is to be verified afterward. The coefficients \( \alpha_k \) and \( \beta_k \) follow from the remaining boundary conditions. First we have to find a Fourier-sine series for the function \( x(x - 1) \). We obtain

\[
x(x - 1) = \sum_{k=1}^{\infty} \gamma_k \sin(k \pi x), \tag{8.38a}
\]

where

\[
\gamma_k = \begin{cases} 
0 & \text{for } k \text{ even,} \\
-\frac{8}{(k \pi)^2} & \text{for } k \text{ odd.} 
\end{cases} \tag{8.38b}
\]

The coefficients now follow from comparing (8.38a) and (8.37) for \( x = 0 \) and \( x = 1 \). This gives the set of equations

\[
\begin{align*}
\alpha_k + \beta_k &= \gamma_k, \tag{8.39a} \\
\alpha_k e^{k\pi} + \beta e^{-k\pi} &= \gamma_k. \tag{8.39b}
\end{align*}
\]

With some straightforward arithmetic one finds from (8.39) the coefficients resulting in the final (indeed uniformly converging) series solution

\[
u(x, y) = \sum_{k=1}^{\infty} \gamma_k \sin(k \pi x) \frac{\cosh(k \pi (y - \frac{1}{2}))}{\cosh(\frac{1}{2} k \pi)}. \tag{8.40}
\]

The method of separation of variables apparently requires knowledge of the eigenvalues and eigenfunctions of the separate systems. This means that the boundary conditions
should allow the separation to result in homogeneous boundary conditions for either one of the subproblems. The method is not restricted to Cartesian coordinates. Indeed, consider the PDE

\[ \nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0 \quad (8.41) \]

defined on a finite wedge \{ (r, \phi) | 0 < r < R, 0 < \phi < \Phi \}, where \( u \) satisfies the boundary conditions

\[ u(r, 0) = u(r, \Phi) = 0, \quad u(R, \phi) = g(\phi) \quad (8.42) \]

for some \( g(\phi) \). Writing

\[ u(r, \phi) = v(r)w(\phi), \]

we obtain from (8.41) that

\[ \frac{r^2}{v} \frac{d^2 w}{d\phi^2} + \frac{r}{v} \frac{dw}{dr} = -\frac{\lambda}{w} \]

We now have the two eigenvalue problems

\[ \frac{d^2 w}{d\phi^2} = \lambda w, \]
\[ r^2 \frac{d^2 v}{dr^2} + \frac{r}{r} \frac{dv}{dr} = -\lambda v. \]

From the homogeneous boundary conditions in (8.42) we see that \( w(0) = w(\Phi) = 0 \) and consequently \( w \) is the solution of a genuine eigenproblem with the typical eigenvalues

\[ \lambda_k = -\omega_k^2, \quad \omega_k := \frac{k\pi}{\Phi} \quad (k = 1, 2, \ldots) \quad (8.45) \]

and corresponding eigenfunctions

\[ w_k(\phi) = \sin(\omega_k \phi) \quad (k = 1, 2, \ldots). \quad (8.46) \]

For each \( \lambda_k \) the equation for \( v \) can now be solved by simple substitution of this \( \lambda_k \) in (8.44b). Substituting a solution of the form \( v(r) = r^\mu \), we can easily see that \( v_k(r) \) has the form

\[ v(r) = \alpha_k r^{\omega_k} \]

By requiring \( v_k \) to be bounded, we see that \( \beta_k = 0 \). The resulting solution of the boundary value problem (8.41), (8.42) is then given by

\[ u(r, \phi) = \sum_{k=1}^{\infty} \alpha_k r^{\omega_k} \sin(\omega_k \phi). \quad (8.48) \]

Like in the previous case, if we have a Fourier-sine expansion of the function \( g(\phi) \), we can formally determine the coefficients \( \alpha_k \).
Example 8.9 Denote by \( \gamma_k \) the coefficients of the Fourier-sine expansion of \( g(\phi) \). Then we have
\[
\gamma_k = \frac{2}{\Phi} \int_0^\Phi g(\phi) \sin(\omega_k \phi) \, d\phi.
\]
Hence from the boundary condition at \( r = R \) we obtain
\[
\alpha_k = \gamma_k R^{-\omega_k},
\]
so the final solution reads
\[
u(r, \phi) = \sum_{k=1}^\infty \gamma_k \left( \frac{r}{R} \right)^{\omega_k} \sin(\omega_k \phi).
\]
This result is interesting as it shows the smoothness of the solution on a wedge in the neighbourhood of the corner point. Indeed, if we consider, e.g., \( \frac{\partial \nu}{\partial r} \), we obtain
\[
\frac{\partial \nu}{\partial r} = \frac{1}{R} \sum_{k=1}^\infty \gamma_k \omega_k \left( \frac{r}{R} \right)^{\omega_k-1} \sin(\omega_k \phi).
\]
We see already that for \( \Phi > \pi \) the first term \( (k = 1) \) is not bounded. The conclusion more generally, therefore, is that corners in a domain imply less smoothness. In particular, for reentrant corners (i.e., those with angles larger than \( \pi \)) this already holds true for the first derivative. This corner problem has, of course, consequences when solving a problem numerically.

8.4 Fundamental Solutions

Before looking at general Poisson problems, as we shall do in Section 8.5, it is meaningful to investigate the fundamental solution \( w \) of the Poisson equation, i.e., the solution of the equation
\[
\nabla^2 w(x; \xi) = \delta(x - \xi).
\]
We look for symmetry solutions depending only on the distance \( r = \|x - \xi\| \), i.e., circular and spherically symmetric solutions in the two-dimensional and three-dimensional spaces, respectively. This simplifies the problem to an ODE. Indeed, let us denote this solution by \( \bar{w}(r) \). Then it satisfies the differential equation
\[
\nabla^2 \bar{w} = \frac{d^2 \bar{w}}{dr^2} + \frac{d-1}{r} \frac{d\bar{w}}{dr} = \delta(r) \quad (d = 2, 3).
\]
Clearly, (8.50) has a general solution, for \( r > 0 \), of
\[
\bar{w}(r) = \begin{cases} A \ln r + B & \text{if } d = 2, \\ A + B & \text{if } d = 3, \end{cases}
\]
with \( A, B \in \mathbb{R} \). For \( d = 3 \) we may choose \( \bar{w}(r) \to 0 \) for \( r \to \infty \), implying that \( B = 0 \). We shall also choose \( B = 0 \) for \( d = 2 \). Therefore we are left to find \( A \) such that (8.50) is satisfied. To this end we define a ball \( B(\theta; \rho) \) around \( \theta \) with radius \( \rho \) and denote by \( \partial B(\theta; \rho) \) its sphere. From Gauss’s theorem we then formally obtain
\[
\int_{B(\theta; \rho)} \nabla^2 \bar{w}(r) \, dV = \oint_{\partial B(\theta; \rho)} \frac{\partial \bar{w}}{\partial n}(r) \, dS = \oint_{\partial B(\theta; \rho)} \frac{d\bar{w}}{dr}(r) \, dS.
\]
Since \( \int_{B(0, \rho)} \nabla^2 \bar{w}(r) \, dV = 1 \), we can determine \( A \). For \( d = 2 \) we find
\[
\int_{\partial B(0, \rho)} \frac{d\bar{w}}{dr}(r) \, dS = \int_{\partial B(0, \rho)} \frac{A}{\rho} \, dS = 2\pi A,
\]
from which we conclude that \( A = 1/(2\pi) \). Likewise, for \( d = 3 \) we obtain
\[
\int_{\partial B(0, \rho)} \frac{d\bar{w}}{dr}(r) \, dS = \int_{\partial B(0, \rho)} -\frac{A}{\rho^2} \, dS = -4\pi A,
\]
so that in this case \( A = -1/(4\pi) \). Using this in (8.51), we obtain for the fundamental solution
\[
w(x, \xi) = \begin{cases} 
\frac{1}{2\pi} \ln \|x - \xi\|_2 & \text{if } d = 2, \\
-\frac{1}{4\pi \|x - \xi\|_2} & \text{if } d = 3.
\end{cases}
\] (8.53)

We leave it as an exercise to show that \( w(x, \xi) \) is the (weak) solution of (8.49).

**Example 8.10** Consider the scalar problem
\[
\mathcal{L}[u] := \frac{d^2u}{dx^2} = 0.
\]
We would like to find a fundamental solution \( w(x; \xi) \) satisfying
\[
\frac{d^2w}{dx^2}(x; \xi) = \delta(x - \xi).
\]
Since the general solution of the ODE for \( x \neq \xi \) can be written as \( A + Bx \), we may take as ansatz for \( w \)
\[
w(x; \xi) = \begin{cases} 
A_1 + B_1x & \text{if } x < \xi, \\
A_2 + B_2x & \text{if } x > \xi.
\end{cases}
\]
Using integration, we obtain for sufficiently small \( \rho \) that
\[
\frac{dw}{dx}(\xi + \rho; \xi) - \frac{dw}{dx}(\xi - \rho; \xi) = \int_{\xi-\rho}^{\xi+\rho} \delta(x - \xi) \, dx = 1.
\]
Applying this condition to the solution above, we find
\[
B_2 - B_1 = 1.
\]
Since \( w \) is apparently continuous at \( x = \xi \), we also find
\[
A_1 + B_1\xi = A_2 + B_2\xi.
\]
Hence we obtain the fundamental solution
\[
w(x; \xi) = \begin{cases} 
A_1 + B_1x & \text{if } x \leq \xi, \\
A_1 - \xi + (B_1 + 1)x & \text{if } x > \xi,
\end{cases}
\]
where \( A_1 \) and \( B_1 \) are arbitrary constants. \( \square \)
8.5. Green's Functions; Superposition

Of course the fundamental solution relates to the (homogeneous) operator only. In the next section we shall construct solutions of (8.49) that satisfy the homogeneous form of the boundary condition as well. It is constructive to illustrate this here for the one-dimensional case.

Example 8.11 Consider the fundamental solution derived in Example 8.5. We now want to solve the boundary value problem

\[ L[w](x; \xi) := \frac{d^2w}{dx^2}(x; \xi) = \delta(x - \xi), \quad x \in (0, 1), \]
\[ w(0; \xi) = w(1; \xi) = 0. \]

Applying the boundary conditions to the general form in Example 8.5, we find

\[ A_1 = 0, \quad A_1 - \xi + B_1 + 1 = 0. \]

Therefore the fundamental solution is given by

\[ w(x, \xi) = \begin{cases} x(\xi - 1) & \text{if } x \leq \xi, \\ (x - 1)\xi & \text{if } x > \xi. \end{cases} \]

8.5 Green's Functions; Superposition

Consider the Poisson problem (8.2). We can formally write the right-hand side \( f(x) \) as

\[ f(x) = \int_{\Omega} \delta(x - \xi) f(\xi) \, dV_\xi, \quad x \in \Omega, \]

where the subscript \( \xi \) denotes integration with respect to \( \xi \). Using this representation of \( f(x) \), we can derive a particular solution \( u_p(x) \) in the following way. We multiply (8.49) by \( f(\xi) \) and integrate with respect to \( \xi \) over the domain \( \Omega \). This way, we find

\[ u_p(x) = \int_{\Omega} w(x; \xi) f(\xi) \, dV_\xi. \]

Inserting the fundamental solution (8.53), we obtain

\[ u_p(x) = \begin{cases} \frac{1}{4\pi} \int_{\Omega} \ln \|x - \xi\|_2 f(\xi) \, dV_\xi & \text{if } d = 2, \\ -\frac{1}{4\pi} \int_{\Omega} \frac{1}{\|x - \xi\|_2} f(\xi) \, dV_\xi & \text{if } d = 3. \end{cases} \]

In general \( u_p(x) \) does not satisfy the prescribed boundary condition (see (8.6)). However, it is clear that \( u(x) - u_p(x) \) satisfies the Laplace equation (8.1). Hence there exists a harmonic function \( u_h(x) \) such that

\[ u(x) = u_p(x) + u_h(x). \]
This is called superposition. The harmonic function \( u_h(x) \) has to be determined such that \( u(x) \) satisfies the prescribed boundary condition. Suppose we have a Dirichlet problem. Then it is obvious that \( u_h(x) \) is the solution of the boundary value problem

\[
\nabla^2 u_h = 0, \quad x \in \Omega, \quad (8.58a)
\]
\[
u_h(x) = u(x) - u_p(x), \quad x \in \partial \Omega. \quad (8.58b)
\]

In a similar way, we can derive a solution for a Neumann or a Robin boundary value problem.

**Example 8.12** Consider a two-dimensional Dirichlet problem on the half-space \( \Omega := \{(x,y) \in \mathbb{R}^2 \mid y > 0\} \) (Figure 8.2). The fundamental solution (8.53), and consequently also the particular solution (8.56), does not satisfy the homogeneous boundary condition on the line \( y = 0 \). In order to overcome this problem, we have to modify the fundamental solution. This can be done in the following way. Let \( \xi = (\xi, \eta) \) be an arbitrary point in \( \Omega \). Then we take a mirror point \( \xi^* := (\xi, -\eta) \) with respect to the line \( y = 0 \) and modify the fundamental solution as follows:

\[
w(x; \xi) = \frac{1}{2\pi} \ln \|x - \xi\|_2 - \frac{1}{2\pi} \ln \|x - \xi^*\|_2 = \frac{1}{2\pi} \ln \left( \frac{\|x - \xi\|_2}{\|x - \xi^*\|_2} \right).
\]

It is obvious that this fundamental solution satisfies the homogeneous boundary condition \( w(x; \xi) = 0 \) on the line \( y = 0 \). Moreover, since \( \xi^* \notin \Omega \), \( w(x; \xi) \) is also a solution of (8.49).

The solution of the Dirichlet problem is given by

\[
u(x) = \frac{1}{2\pi} \int_{\Omega} \ln \left( \frac{\|x - \xi\|_2}{\|x - \xi^*\|_2} \right) f(\xi) \, dV_{\xi}. \quad \square
\]

**Figure 8.2.** A point \( \xi \in \Omega \) and its mirror point \( \xi^* \).

Another way to use this superposition principle is to find a particular solution \( u_p(x) \) satisfying the inhomogeneous equation with homogeneous boundary conditions and a harmonic function \( u_h(x) \) satisfying the Laplace equation with the appropriate boundary conditions. This leads to the introduction of the Green’s function \( G(x; \xi) \). The Green’s function \( G(x; \xi) \) for the Dirichlet problem (8.2) and (8.6a) is by definition the solution of the boundary value problem

\[
\nabla^2 G(x; \xi) = \delta(x - \xi), \quad x \in \Omega, \quad (8.59a)
\]
\[
G(x; \xi) = 0, \quad x \in \partial \Omega. \quad (8.59b)
\]
Using superposition, we immediately see that the Green’s function $G(x; \xi)$ equals the fundamental solution $w(x; \xi)$ (see Section 8.4), apart from a harmonic function. Using Theorem 8.4, we therefore conclude that the solution is unique. From the second identity of Green

$$\int_{\Omega} (u \nabla^2 v - v \nabla^2 u) \, dV = \oint_{\partial \Omega} \left( u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) \, dS,$$  

(8.60)

we can find a representation of $u(x)$ as announced. Indeed, if we take $v = G(x; \xi)$, then we obtain

$$u(\xi) = \int_{\Omega} G(x; \xi) f(x) \, dV + \oint_{\partial \Omega} a(x) \frac{\partial G}{\partial n}(x; \xi) \, dS.$$  

(8.61)

Also, for Neumann problems, we can construct a solution by superposition. First, we remark that the divergence theorem provides for a constraint relating the source term $f(x)$ and the boundary function $b(x)$ (see (8.6b)). Indeed, we have

$$\int_{\Omega} f(x) \, dV = \int_{\Omega} \nabla^2 u \, dV = \oint_{\partial \Omega} \frac{\partial u}{\partial n} \, dS = \oint_{\partial \Omega} b(x) \, dS.$$  

(8.62)

As a candidate for the Green’s function $G(x; \xi)$ we might take the function satisfying (8.49) and the homogeneous Neumann boundary condition. However, this choice of $G(x; \xi)$ obviously does not satisfy the constraint (8.62). Another possibility is to define the Green’s function as the solution of the boundary value problem

$$\nabla^2 G(x; \xi) = \delta(x - \xi) - \frac{1}{\kappa}, \quad \kappa := \int_{\Omega} \, dV, \quad x \in \Omega,$$  

(8.63a)

$$\frac{\partial G}{\partial n}(x; \xi) = 0, \quad x \in \partial \Omega,$$  

(8.63b)

where the right-hand side of (8.63a) has been modified to enforce (8.62). We then obtain from (8.60) with $v = G(x; \xi)$ that

$$u(\xi) = \int_{\Omega} G(x; \xi) f(x) \, dV + \frac{1}{\kappa} \int_{\Omega} u(x) \, dV - \oint_{\partial \Omega} G(x; \xi) b(\xi) \, dS.$$  

(8.64)

First, note that for this Green’s function the constraint (8.62) is satisfied. Indeed, we have

$$\oint_{\partial \Omega} \frac{\partial G}{\partial n}(x; \xi) \, dS = \int_{\Omega} \nabla^2 G(x; \xi) \, dV = 1 - \frac{1}{\kappa} \int_{\Omega} \, dV = 0.$$  

Second, we remark that the first integral in (8.64) represents the particular solution and the last one represents the harmonic function. The term $\frac{1}{\kappa} \int_{\Omega} u(x) \, dV$ is just an additive constant.

Finally, for the boundary value problem with Robin boundary condition (8.6c), we define the Green’s function $G(x; \xi)$ as solution of

$$\nabla^2 G(x; \xi) = \delta(x - \xi), \quad x \in \Omega,$$  

(8.65a)

$$\alpha G(x; \xi) + \beta \frac{\partial G}{\partial n}(x; \xi) = 0, \quad x \in \partial \Omega.$$  

(8.65b)
Analogously to the previous derivations, we obtain the following integral representation of $u(\xi)$:

$$u(\xi) = \int_{\Omega} G(x; \xi) f(x) \, dV - \oint_{\partial \Omega} \left( \frac{\alpha}{\beta} u(x) + \frac{\partial u}{\partial n}(x) \right) G(x; \xi) \, dS$$

$$= \int_{\Omega} G(x; \xi) f(x) \, dV - \frac{1}{\beta} \oint_{\partial \Omega} G(x; \xi) c(x) \, dS,$$

(8.66)

provided that $\beta \neq 0$.

A particular solution of (8.63a) can be found by subtracting the following solution $v_p(x)$ (which is independent of $\xi$) of the fundamental solution $w(x; \xi)$:

$$v_p(x) = \begin{cases} \frac{1}{4\kappa} \|x\|_2^2 & \text{if } d = 2, \\ \frac{1}{6\kappa} \|x\|_2^2 & \text{if } d = 3. \end{cases}$$

(8.67a) \hspace{1cm} (8.67b)

The Green's function $G(x; \xi)$ then differs from $w(x; \xi) - v_p(x)$ by a harmonic function, which can be found formally by solving a (homogeneous) Neumann problem with known boundary conditions.

Example 8.13 Consider the following Dirichlet boundary value problem on the half-space:

$$\nabla^2 u = f(x), \quad x \in \Omega := \{(x, y) \in \mathbb{R}^2 \mid y > 0\},$$

$$u(x, 0) = a(x), \quad u(x, y) \to 0 \text{ for } y \to \infty,$$

where the boundary function $a(x) \to 0$ as $|x| \to \infty$. From Example 8.12 we derive for the Green’s function with $\xi = (\xi, \eta)$ that

$$G(x, y; \xi, \eta) = \frac{1}{2\pi} \ln \left( \frac{(x - \xi)^2 + (y - \eta)^2}{(x - \xi)^2 + (y + \eta)^2} \right)^{\frac{1}{2}}.$$

Hence on $\partial \Omega$ we find

$$\frac{\partial G}{\partial n}(x, 0; \xi, \eta) = \frac{\eta}{\pi (x - \xi)^2 + \eta^2}.$$

So, formally, the solution of the boundary value problem above is given by

$$u(\xi, \eta) = \frac{1}{2\pi} \int_0^\infty \int_{-\infty}^\infty \ln \left( \frac{(x - \xi)^2 + (y - \eta)^2}{(x - \xi)^2 + (y + \eta)^2} \right)^{\frac{1}{2}} f(x, y) \, dx \, dy$$

$$+ \frac{\eta}{\pi} \int_{-\infty}^\infty \frac{a(x)}{(x - \xi)^2 + \eta^2} \, dx. \quad \Box$$

The last example shows the power and weakness of an analytical approach. On the one hand, by mere construction, one can show that a solution exists, and one can give estimates for it. On the other hand, the expressions one obtains are often complicated and usually not directly solvable in closed form.

We would like to point out another important fact. The expressions we have derived in this section for the solutions can be seen as inverting the operator form for the original
problem; i.e., both the equation and the boundary condition. In particular, the Green’s function can be interpreted as the (“constrained”) inverse of the Laplace operator. In the next chapter we will investigate numerical methods for such problems, where this “inversion” will be met again, now in terms of matrices.

8.6 The Maximum Principle

The Laplace operator has a nice property that gives rise to a number of useful results. If the second derivative of a scalar function \( u(x) \) is zero, it is linear. On a finite interval the absolute maximum is attained at either end of the interval. For a Poisson problem this is described in the following property.

**Property 8.14 (maximum principle).** Let \( \nabla^2 u(x) = 0 \) for all \( x \in \Omega \). Then \( u(x) \) satisfies the inequalities

\[
\begin{align*}
m &:= \min_{\xi \in \partial \Omega} u(\xi) \leq u(x) \leq \max_{\xi \in \partial \Omega} u(\xi) := M.
\end{align*}
\]

**Proof.** Define the function \( v_\varepsilon(x) := u(x) + \varepsilon \|x\|_2^2 \) for \( \varepsilon > 0 \). Clearly, we have

\[
\nabla^2 v_\varepsilon(x) = \nabla^2 u(x) + 2d\varepsilon > 0, \quad x \in \Omega \quad (d = 2, 3).
\]

Suppose \( v_\varepsilon(x) \) has a maximum in the interior of \( \Omega \), say at \( x_0 \). Then

\[
\nabla^2 v_\varepsilon(x_0) \leq 0.
\]

Since this contradicts the inequality in \((*)\), we conclude that \( v_\varepsilon \) can only attain its maximum at the boundary \( \partial \Omega \). If we denote the latter by \( M_\varepsilon \), we derive the required upper bound by letting \( \varepsilon \downarrow 0 \). The lower bound follows from a similar argument, now using \(-u(x)\) and \( v_\varepsilon(x) := -u(x) + \varepsilon \|x\|_2^2 \) instead.

The theorem above is sometimes also referred to as the minimum-maximum principle. In particular, the maximum principle can be extended to Poisson problems.

**Property 8.15.** Let \( \nabla^2 u(x) \geq 0 \) for all \( x \in \Omega \). Then \( u(x) \) attains its maximum at the boundary; i.e.,

\[
\begin{align*}
u(x) \leq \max_{x \in \partial \Omega} u(\xi).
\end{align*}
\]

**Proof.** It is easy to see that the arguments in the proof of Property 8.14 still apply for this case.

The latter property is quite powerful in that it gives the possibility of comparing solutions of two Poisson problems. The following property states this more precisely.

**Property 8.16 (comparison theorem).** Consider the two Poisson equations \( \nabla^2 u_1(x) = f_1(x) \) and \( \nabla^2 u_2(x) = f_2(x) \) with \( f_1(x) \geq f_2(x) \) for all \( x \in \Omega \). Then

\[
\begin{align*}
u_1(x) \leq u_2(x) + \max_{x \in \partial \Omega} (u_1(x) - u_2(x)).
\end{align*}
\]

**Proof.** Since \( \nabla^2 (u_1(x) - u_2(x)) \geq 0 \), the result follows directly from Property 8.15.
Corollary 8.17. A Dirichlet problem has a unique solution that depends continuously on the boundary data.

Proof. If there are two solutions, then the difference is a harmonic function satisfying the homogeneous boundary condition. We can apply Property 8.14 to conclude that this difference must be zero. The continuous dependence is a consequence of the comparison theorem.

8.7 The Stokes Equations

The problems met in practice are often more complex than the ones discussed thus far. A typical example of these are the Navier–Stokes equations in fluid dynamics and their creeping flow simplification, the Stokes equations. Here we shall not dwell on either their derivation or their application. This is done in Chapters 7 and 16. The first complication we meet is that a flow in one dimension is not of interest and thus a nontrivial formulation requires a Laplacian operating on a vector. Let \( u \) and \( p \) be a velocity and a pressure field. Then the Stokes equations for a domain \( \Omega \subset \mathbb{R}^d \) (\( d = 2, 3 \)) read

\[
\begin{align*}
\nabla^2 u - \nabla p &= 0, \\
\nabla \cdot u &= 0. 
\end{align*}
\]

Note that \( \nabla^2 u \) is to be taken componentwise. From (8.71a) it is clear that \( \nabla p \), rather than \( p \), is a (dependent) variable, so we always need to specify \( p \) somewhere in the domain. We shall moreover prescribe the Dirichlet boundary condition

\[
\mathbf{u} \cdot \mathbf{n} = b(x), \quad x \in \partial \Omega. 
\]

First, consider the case where \( b(x) \equiv 0 \). We now want to show that the solution of boundary value problem (8.71) and (8.72) is unique. We shall need a bit of vector calculus to do this. First, we use the relation (see appendix J)

\[
\nabla \cdot (p \mathbf{u}) = p \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla p. 
\]

Applying the divergence theorem to this relation and taking into account (8.71b) and (8.72), we find

\[
\int_{\Omega} \mathbf{u} \cdot \nabla p \, dV = \int_{\Omega} \nabla \cdot (p \mathbf{u}) \, dV = \oint_{\partial \Omega} (p \mathbf{u}) \cdot \mathbf{n} \, dS = \oint_{\partial \Omega} pb(x) \, dS. 
\]

Since \( b(x) \equiv 0 \), we obtain

\[
\int_{\Omega} \mathbf{u} \cdot \nabla p \, dV = 0, 
\]

which states that the velocity field and the pressure gradient are orthogonal. Furthermore, we have the following relation for the velocity field \( \mathbf{u} := (u, v, w)^T \):

\[
\nabla \cdot (\mathbf{u} \cdot \nabla \mathbf{u}) = |\nabla \mathbf{u}|^2 + \mathbf{u} \cdot \nabla^2 \mathbf{u}, 
\]
8.7. The Stokes Equations

where \( \mathbf{u} \cdot \nabla \mathbf{u} := (\mathbf{u} \cdot \mathbf{u}_x, \mathbf{u} \cdot \mathbf{u}_y, \mathbf{u} \cdot \mathbf{u}_z)^T \) and \( |\nabla \mathbf{u}|^2 := |\nabla u|^2 + |\nabla v|^2 + |\nabla w|^2 \). Integrating this relation yields

\[
\int_{\Omega} \nabla \cdot (\mathbf{u} \cdot \nabla \mathbf{u}) \, dV = \int_{\Omega} |\nabla \mathbf{u}|^2 \, dV + \int_{\Omega} \mathbf{u} \cdot \nabla^2 \mathbf{u} \, dV. \tag{8.77}
\]

The divergence theorem now implies for the left-hand side of (8.77) that

\[
\int_{\Omega} \nabla \cdot (\mathbf{u} \cdot \nabla \mathbf{u}) \, dV = \oint_{\partial \Omega} (\mathbf{u} \cdot \nabla \mathbf{u}) \cdot \mathbf{n} \, dS = 0. \tag{8.78}
\]

The zero value follows from (8.72). Let us now take the inner product of \( \nabla^2 \mathbf{u} - \nabla p \) and \( \mathbf{u} \); this trivially gives

\[
\int_{\Omega} \mathbf{u} \cdot \nabla^2 \mathbf{u} \, dV - \int_{\Omega} \mathbf{u} \cdot \nabla p \, dV = 0. \tag{8.79}
\]

The second term in (8.79) is zero on account of (8.75). So the first integral in (8.79) is zero as well. Combining (8.77), (8.78), and (8.79), we therefore conclude that

\[
\int_{\Omega} |\nabla \mathbf{u}|^2 \, dV = 0. \tag{8.80}
\]

Clearly \( |\nabla \mathbf{u}|^2 = 0 \) only if \( \nabla \mathbf{u} = 0 \), so \( \mathbf{u} \) is constant. On account of the boundary condition (8.72), with \( b(x) \equiv 0 \), we conclude that \( \mathbf{u}(x) \equiv 0 \). It is trivial now to see that this implies uniqueness of the general problem (8.71), (8.72).

The linearity of the problem allows us to construct fundamental solutions \( \mathbf{u}_i, p_i \) called the Stokeslets; cf. Section 8.4. They are named after G.G. Stokes but were really first discovered by H.A. Lorentz in 1896; see [93]. In two dimensions they are the solution of

\[
\nabla^2 \mathbf{u}_i - \nabla p_i = \delta(x - \xi) e_i \quad (i = 1, 2), \tag{8.81a}
\]

\[
\nabla \cdot \mathbf{u}_i = 0. \tag{8.81b}
\]

Here \( e_1 = (1, 0)^T \) and \( e_2 = (0, 1)^T \). We have for \( \mathbf{u}_i = (u^1_i, u^2_i)^T \) that

\[
u^j_i(x; \xi) = \frac{1}{4\pi} \left( \delta_{ij} \ln \|x - \xi\| - \frac{(x_j - \xi_j)(x_i - \xi_i)}{\|x - \xi\|^3} \right) \quad (i, j = 1, 2), \tag{8.82a}
\]

\[
p_i(x; \xi) = -\frac{\xi_i}{2\pi \|x - \xi\|^2}. \tag{8.82b}
\]

These Stokeslets can be used to derive an integral formulation for the Stokes equations.

We conclude this section with an example of one of the few Stokes problems that can be solved analytically. It regards the flow past a sphere of radius one. We assume that the velocity \( \mathbf{u} \) is equal to zero at the sphere and equal to one in the direction of the flow (say the \( z \) direction) at infinity. The pressure \( p \) approaches a limit value there. This example will give us an opportunity to show some more vector calculus. To start with we shall use the spherical coordinates \( (r, \theta, \phi) \), and we obtain, with \( \mathbf{u} = (u_r, u_\theta, u_\phi)^T \), that

\[
\nabla \cdot \mathbf{u} = \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 u_r) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\sin \theta u_\theta) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} u_\phi = 0. \tag{8.83}
\]
Chapter 8. The Analysis of Elliptic Equations

We assume axial symmetry, implying that \( \frac{\partial}{\partial \phi} = 0 \) and \( u_\phi = 0 \). The construction of the solution now employs the notions of a stream function, \( \psi = \psi(r, \theta) \) say. In particular, we require

\[
\begin{align*}
  u_r &= \frac{1}{r^2 \sin \theta} \frac{\partial \psi}{\partial \theta}, \\
u_\theta &= -\frac{1}{r \sin \theta} \frac{\partial \psi}{\partial r}.
\end{align*}
\] (8.84)

Substituting these relations into (8.83) (with \( u_\phi = 0 \)) gives

\[
\frac{1}{r^2 \sin \theta} \frac{\partial^2 \psi}{\partial r \partial \theta} - \frac{1}{r \sin \theta} \frac{\partial^2 \psi}{\partial \theta \partial r} = 0.
\]

So we have \( \nabla \cdot u = 0 \), indeed. The relations in (8.84) imply that the velocity can be written as

\[
u = \nabla \times \tilde{\psi}, \quad \tilde{\psi} := \frac{\psi}{r \sin \theta} e_\phi.
\] (8.85)

Next it is convenient to introduce the vorticity vector \( \omega = \nabla \times u \). Obviously, \( \tilde{\psi} \) and \( \omega \) are related by

\[
\nabla \times (\nabla \times \tilde{\psi}) = \omega.
\] (8.86)

Applying the curl operator twice to the vector \( \tilde{\psi} \) defined in (8.85) results in a vector in the same direction. More precisely, we have

\[
\omega = -\frac{1}{r \sin \theta} D[\psi] e_\phi,
\] (8.87)

where the differential operator \( D \) is defined by

\[
D[\psi] := \frac{\partial^2 \psi}{\partial r^2} + \frac{1}{r^2 \sin \theta} \frac{\partial^2 \psi}{\partial \theta^2} - \cot \theta \frac{\partial \psi}{\partial r}.
\] (8.88)

Using the relation \( \nabla^2 u = \nabla (\nabla \cdot u) - \nabla \omega \) (see the appendix, Section J), we can rewrite (8.71a) as

\[
\nabla \times \omega = -\nabla p.
\] (8.89)

In order to eliminate the pressure gradient, we apply the curl operator to (8.89) and find

\[
\nabla \times (\nabla \times \omega) = 0.
\] (8.90)

Analogously to the derivation of (8.87), we also find

\[
\nabla \times (\nabla \times \omega) = \frac{1}{r \sin \theta} D^2[\psi] e_\phi.
\] (8.91)

Hence we conclude from (8.90) that

\[
D^2[\psi] = 0.
\] (8.92)

At the sphere we have \( u = 0 \), so we obtain as boundary conditions for \( \psi \) that

\[
\frac{\partial \psi}{\partial r} = \frac{\partial \psi}{\partial \theta} = 0 \quad \text{at} \quad r = 1,
\] (8.93)
so apparently $\psi(r, \theta) = \text{constant at } r = 1$ and therefore we may as well take $\psi(1, \theta) = 0$. As for the asymptotic boundary condition, we see from Figure 8.3 that $u_r \sim \cos \theta$ and $u_\theta \sim -\sin \theta$. Using this in (8.84) gives $\frac{\partial \psi}{\partial r} \sim r^2 \sin \theta \cos \theta$, so

$$\psi(r, \theta) \sim \frac{1}{2} r^2 \sin^2 \theta \quad \text{for } r \to \infty. \quad (8.94)$$

Note that an integration constant is immaterial.

Returning now to (8.92), we apply separation of variables, i.e.,

$$\psi(r, \theta) = f(r) g(\theta). \quad (8.95)$$

Actually, with some further (tedious) analysis it turns out that $g(\theta)$ may be taken equal to $\sin^2 \theta$. Straightforward computation then reveals

$$D[\psi] = (\frac{d^2 f}{dr^2} - \frac{2 f}{r^2}) \sin^2 \theta. \quad (8.96)$$

Once more applying the rule (8.96) we find

$$D^2[\psi] = \left(\frac{d^4 f}{dr^4} - \frac{4 d^2 f}{r^2 dr^2} + \frac{8 f}{r^3 dr} - \frac{8 f}{r^4}\right) \sin^2 \theta = 0. \quad (8.97)$$

We have finally arrived at an ODE whose general solution is

$$f(r) = \alpha r^4 + \beta r^2 + \gamma r + \frac{\delta}{r}. \quad (8.98)$$

Using the condition at infinity (8.94), we find $\alpha = 0$, $\beta = \frac{1}{2}$, and using the boundary condition at the sphere (8.93), we then find $\gamma = -\frac{3}{4}$, $\delta = \frac{1}{4}$. We thus find for the stream function

$$\psi(r, \theta) = \left(\frac{1}{2} r^2 - \frac{3}{4} r + \frac{1}{4r}\right) \sin^2 \theta. \quad (8.99)$$
The actual sought flow field $u$ is then given by

$$u_r = \left(1 - \frac{3}{2r} + \frac{1}{2r^3}\right) \cos \theta, \quad u_\theta = \left(-1 + \frac{3}{4r} + \frac{1}{4r^3}\right) \sin \theta,$$

(8.100)

while for the pressure $p$ we obtain

$$p = -\frac{3 \cos \theta}{2r^2} + p_0,$$

(8.101)

with $p_0$ the constant pressure at infinity.

### 8.8 Discussion

- Problems of elliptic type are probably the most common. At the least, elliptic operators appear in a large number of problems in mathematical physics. As we saw in Chapter 2, we encounter the Laplace operator in parabolic problems as well as hyperbolic problems. A hyperbolic problem like the wave equation reduces to an elliptic problem when we consider only time-harmonic solutions. When the equation is of parabolic type, elliptic problems appear as steady state problems, letting the time go to infinity. Thus we may consider situations like the temperature distribution in a room with heat sources and sinks or the concentration in a vessel that is draining with constant replenishing [30, 31]. The name “potential problem” comes from electrostatics. It refers to the potential from electrical charges, following from the Maxwell equations. In fact, the Maxwell equations are a good example of a hyperbolic problem [67].

- In fluid mechanics the velocity potential and stream function of an incompressible irrotational flow satisfy an elliptic equation [15]. Typical mechanical problems involve the deflection of a thin membrane in two dimensions. For deflection of beams one encounters the biharmonic equation. This is still elliptic but contains the $\nabla^4$ operator. There are many similarities with harmonic analysis [44].

- In Chapter 16 we shall encounter a number of elliptic problems arising in a practical context. To start with, in Section 16.6 the groundwater table is modeled as a nonlinear boundary value problem in one dimension. This nonlinearity arises in the second derivative term, which contains a small coefficient. It is shown when and how an asymptotic expansion can “solve” this problem. Section 16.8 deals with chemical reactions in small pellets. The problem is described as the steady state of a reaction-diffusion equation of a spherically symmetric problem. Here too, solutions can be obtained in terms of an asymptotic expansion. An example of a viscous flow problem leading to the Stokes equation is the forming of glass products; see Section 16.11. The problem discussed here can be solved analytically using thin-layer approximations.

### Exercises

8.1. In this exercise we determine the spherically symmetric fundamental solution $w(x)$ of the $d$-dimensional Poisson equation, i.e., $w(x) = \tilde{w}(r)$ with $r = ||x||_2$. 
(a) Show that $\tilde{w}(r)$ is a solution of the ODE
\[
\frac{d^2 \tilde{w}}{dr^2} + \frac{d - 1}{r} \frac{d \tilde{w}}{dr} = \delta(r).
\]

(b) For $d > 2$ and the requirement that $\lim_{r \to \infty} \tilde{w}(r) = 0$, show that
\[
\tilde{w}(r) = \frac{A_d}{r^{d-2}},
\]
where the integration constants $A_d$ satisfy
\[
A_d \int_{S(0; \rho)} \frac{d}{dr} \left( \frac{r^{2-d}}{d} \right) dS = 1,
\]
with $S(0; \rho)$ the sphere with centre $0$ and radius $\rho$. One can prove that the constants $A_d$ can be found from
\[
A_d = \frac{1}{(2-d) B_d},
\]
where $B_d$ is the surface area of the $d$-dimensional unit sphere $S(0; 1)$. Note that $B_d = 2\pi^{d/2} \Gamma(d/2)$.

8.2. Let $\nabla^2 u(x) = 0$ for all $x \in \Omega$ and $u(x) > 0$ for all $x \in \partial \Omega$. Prove that $u(x) > 0$ for all $x \in \Omega$.

8.3. Consider the Neumann problem
\[
\nabla^2 u = c, \quad x \in \Omega := \{(x, y) \in R^2 \mid x^2 + y^2 < 1\},
\]
\[
\frac{\partial u}{\partial n} = 1, \quad x \in \partial \Omega,
\]
where $c$ is a constant. Show that this boundary value problem only has solutions for $c = 2$. Determine all of these solutions.

8.4. Let $\nabla^2 u_i(x) = 0$ ($i = 1, 2, \ldots, N$) for all $x \in \Omega$. Let $u_{i-1}(x) \leq u_i(x) \leq u_{i+1}(x)$ ($i = 2, 3, \ldots, N - 1$) for all $x \in \partial \Omega$. Show that $u_{i-1}(x) \leq u_i(x) \leq u_{i+1}(x)$ ($i = 2, 3, \ldots, N - 1$) for all $x \in \Omega$.

8.5. Let $\nabla^2 u(x) = 0$ for all $x \in \Omega$. Split the boundary $\partial \Omega$ into two simply connected parts, $\partial \Omega_1$ and $\partial \Omega_2$, such that $\partial \Omega_1 \cup \partial \Omega_2 = \partial \Omega$. Let $u(x)$ satisfy the boundary conditions
\[
u(x) = \alpha(x), \quad x \in \partial \Omega_1,
\]
\[
\frac{\partial u}{\partial n}(x) = \beta(x), \quad x \in \partial \Omega_2.
\]
Show that the solution $u(x)$ is unique.

8.6. Consider the domain $\Omega := \{(x, y) \in R^2 \mid x > 0, \ y > 0\}$. Construct a Green’s function for the Dirichlet and Neumann problem using mirror points.
8.7. Let $a, b \in \mathbb{R}$ with $a < b$. Define the domain $\Omega := \{(x, y) \in \mathbb{R}^2 \mid a < y < b\}$. Construct a Green’s function for the Dirichlet problem on $\Omega$ using mirror points. Show that the Green’s function is symmetric; i.e.,

$$G(x; \xi) = G(\xi; x).$$

8.8. Consider on the unit disc $\Omega := \{(r, \phi) \in \mathbb{R}^2 \mid 0 \leq r < 1, -\pi \leq \phi < \pi\}$ the Neumann problem

$$\nabla^2 u = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0$$

satisfying the boundary condition

$$\frac{\partial u}{\partial n}(1, \phi) = g(\phi),$$

where the function $g(\phi)$ satisfies the constraint

$$\int_{-\pi}^{\pi} g(\phi) \, d\phi = 0.$$

Find a formal solution.

8.9. Prove the mean value theorem: If $u(x)$ is harmonic in $\Omega \subset \mathbb{R}^3$ and $S(x; \rho)$ denotes the sphere in $\Omega$ with centre $x$ and radius $\rho$, then

$$u(x) = \frac{1}{4\pi \rho^2} \int_{S(x; \rho)} u(y) \, dS_y,$$

where the subscript $y$ denotes that the integration is carried out in the $y$ variables.

8.10. Consider on the strip $\Omega := \{(x, y) \in \mathbb{R}^2 \mid 0 < x < 1, 0 < y\}$ the Dirichlet problem

$$\nabla^2 u = 0, \quad x \in \Omega,$nabla^2 u = 0, \quad x \in \Omega,$$

$$u(0, y) = u(1, y) = 0, \quad 0 \leq y < \infty,$$

$$u(x, 0) = x(1-x), \quad 0 \leq x \leq 1,$$

$$u \text{ is bounded in } \Omega.$$

Show by separation of variables that the solution is given by

$$u(x, y) = \frac{8}{\pi} \sum_{k=0}^{\infty} \frac{1}{(2k+1)^2} e^{-(2k+1)^2 \pi y} \sin((2k+1)\pi x).$$

8.11. The Helmholtz equation may not allow for the maximum principle. Prove this for the one-dimensional case. Consider the equation

$$\frac{d^2 u}{dx^2} + \lambda u = 0, \quad x \in \Omega := (0, \pi),$$

where $\lambda > 1$. 


8.12. The maximum principle may also apply to nonlinear problems. For this consider the boundary value problem

\[ \nabla^2 u = \lambda e^u, \quad x \in \Omega, \\
   u = 0, \quad x \in \partial \Omega. \]

(a) Prove that \( u(x) > 0 \) for all \( x \in \Omega \) if \( \lambda > 0 \).
(b) Construct a comparison function to estimate \( u(x) \).

8.13. If we apply an affine coordinate transformation (corresponding to rigid body motions), the Laplace operator remains invariant. Prove this for \( x \in \mathbb{R}^2 \).

8.14. Consider the Helmholtz equation on a unit square, i.e.,

\[ \nabla^2 u + ku = 0, \quad x \in \Omega := \{(x, y) \in \mathbb{R}^2 \mid 0 < x, y < 1\}, \\
   u = 0, \quad x \in \partial \Omega, \]

where \( k < 0 \). Compute the eigenvalues and eigenfunctions.

8.15. Show that the operator

\[ \mathcal{L}[u] := p(x) \frac{d^2 u}{dx^2} + q(x) \frac{du}{dx} \]

is not self-adjoint unless \( q(x) = \frac{d}{dx} p(x) \).

8.16. Consider the eigenvalue problem

\[ \mathcal{L}[u] := \frac{d^2 u}{dx^2} = \lambda u, \quad x \in (0, 1), \\
   u(0) = 0, \quad \frac{du}{dx}(1) = 0. \]

Determine the eigenvalues and eigenfunctions.

8.17. Consider the eigenvalue problem

\[ \mathcal{L}[u] := \frac{d^2 u}{dx^2} = \lambda u, \quad x \in (0, 1), \\
   u(0) - \frac{du}{dx}(0) = 0, \quad u(1) = 0. \]

Show that the eigenvalues are negative.

8.18. Consider the equation

\[ \mathcal{L}[u] := \nabla^2 u + qu = \lambda u, \quad x \in \Omega, \]

where \( q < 0 \). Show that the eigenvalues \( \lambda_k \) are nonpositive if we impose homogeneous Dirichlet or Neumann boundary conditions. Also show that \( \lambda_k \leq 0 \) if we impose the homogeneous Robin boundary condition (8.6c) with \( \text{sign}(\alpha) = \text{sign}(\beta) \).
8.19. Show that for the operator

\[ L[u] := \nabla^2 u + a(x) \frac{\partial u}{\partial x} + b(x) \frac{\partial u}{\partial y} \]

the adjoint operator is given by

\[ L^*[v] = \nabla^2 v - \frac{\partial}{\partial x} (a(x)v) - \frac{\partial}{\partial y} (b(x)v). \]
Chapter 9

Numerical Methods for Elliptic Equations

This chapter is devoted to the numerical solution of elliptic equations. In Section 9.1 we recall how the difference methods that were derived in Chapter 5 are used. Moreover, the discretisation of derivative boundary conditions is described. In Section 9.2 the maximum principle for discrete problems is discussed. As it turns out, there are many analogies with the continuous case. The discrete maximum principle is then applied in Section 9.3 to estimate the global discretisation error. As an alternative, the global discretisation error is estimated using the matrix method. Both methods enable one to assess the global effect of discretisation errors that are exclusively made at the boundary. In Section 9.4 some basic solution methods for linear systems resulting from discretising an elliptic equation are discussed. In particular, Gauss-type iterative methods are analyzed. These iterative methods are evolutionary in character and can therefore be interpreted in terms of a parabolic equation. This latter equation is called the closure of the iterative method. The discrete analogue of Green’s functions is presented in Section 9.5. This provides a method of explicitly computing the inverse of the discretisation matrix. Nonlinear problems are dealt with in Section 9.6 and the Newton and Gauss–Jacobi iteration and transient methods of solution are discussed. Finally, in Section 9.7 we consider the numerical solution of the Stokes equations, in particular a special iteration method called pressure correction.

9.1 Discretisation and Boundary Conditions

In this chapter we shall restrict ourselves to problems with at most two independent variables. First, let us consider the linear equation

\[
\frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial y^2} + c \frac{\partial u}{\partial x} + d \frac{\partial u}{\partial y} + eu = f, \quad (x, y) \in \Omega \subseteq \mathbb{R}^2, \quad a, b > 0, \quad e \leq 0,
\]

where the coefficients \(a, b, c, d, \) and \(e\) are constant. In Chapter 5 we introduced second order central difference approximations for both first and second order derivatives. Applying these central differences to (9.1) on a uniform grid with grid sizes \(\Delta x\) and \(\Delta y\) in the \(x\) and
Chapter 9. Numerical Methods for Elliptic Equations

Figure 9.1. Five-point stencil for the difference equation (9.2).

y directions respectively, we obtain the difference scheme (cf. (5.33))

\[
\frac{1}{\Delta x^2} \left( a + \frac{1}{2} c \Delta x \right) u_E + \frac{1}{\Delta x^2} \left( a - \frac{1}{2} c \Delta x \right) u_W + \frac{1}{\Delta y^2} \left( b + \frac{1}{2} d \Delta y \right) u_N + \frac{1}{\Delta y^2} \left( b - \frac{1}{2} d \Delta y \right) u_S + \left( e - \frac{2a}{\Delta x^2} - \frac{2b}{\Delta y^2} \right) u_C = f(x_C),
\]

(9.2)

where \( u_C \) denotes the numerical approximation of \( u(x_C) \), etc. The stencil of (9.2) is depicted in Figure 9.1. The difference equation (9.2) can also formally be written as

\[
\alpha_0 u_C + \alpha_1 u_N + \alpha_2 u_E + \alpha_3 u_S + \alpha_4 u_W = f(x_C).
\]

(9.3)

From our assumptions we immediately see that the following property holds.

**Property 9.1.** The coefficients in the difference scheme (9.3) satisfy the inequalities

\[
\alpha_0 < 0,
\]

(9.4a)

\[
\alpha_0 + \sum_{i=1}^{4} \alpha_i \leq 0, \quad (i = 1, 2, 3, 4)
\]

(9.4b)

\[
\alpha_i > 0 \quad (i = 1, 2, 3, 4) \quad \text{if } \Delta x \text{ and } \Delta y \text{ are small enough.}
\]

(9.4c)

The grid points considered thus far are interior grid points. If we have a Dirichlet boundary condition, we can simply substitute the given data into the scheme to eliminate any one of the points \( x_E, x_S, x_W \) or \( x_N \) or (at corners) two of them. In Section 5.2.2 we considered how to deal with curved boundaries. Below we shall deal with straight lines being the edges of a rectangular domain. In particular, let us consider the treatment of a Robin boundary condition, which automatically includes the case of Neumann boundary conditions. Consider a situation like in Figure 9.2. Suppose that along a left vertical boundary we have

\[
\frac{\partial u}{\partial x} = q u + g(y).
\]

(9.5)

A straightforward, though only first order, discrete approximation of (9.5) would be the forward difference

\[
\frac{1}{\Delta x} (u_E - u_N) = qu + g(y_C).
\]

(9.6)
9.2. The Maximum Principle

This then gives

\[ u_c = \frac{u_e - \Delta x g(y_c)}{1 + q \Delta x}, \]  

(9.7)

Using this value for a boundary point \( x_c \) differs from using a Dirichlet boundary value in two ways. First, it changes the scheme, as the coefficient of \( u_e \) appearing in the difference equation at \( x_e \) is clearly updated with a value \( \alpha/(1 + q \Delta x) \), where \( \alpha \) is the coefficient of \( u_c \). Second, it also affects the local discretisation error of such a point. Indeed, since (9.6) leaves an error of \( O(\Delta x) \), using \( u_c \) computed from (9.7) induces an additional error of \( O(\Delta x^2) \) in any scheme where it is substituted. In (9.2) this therefore introduces local errors of \( O(1) \). A better way is to employ virtual grid points. For example, if we use an extra grid point \( x_w \) like in Figure 9.2 and apply central differences to (9.5), we obtain

\[ \frac{1}{2 \Delta x} (u_e - u_w) = qu_c + g(y_c). \]  

(9.8)

This leads to

\[ u_w = u_e - 2 \Delta x (qu_c + g(y_c)). \]  

(9.9)

Substitution into (9.2) then gives a stencil containing grid points within the domain only. Since (9.8) is second order, this then induces an extra local error of \( O(\Delta x) \). This order is still less than desired. However, we shall show in Example 9.14 that this does not affect the error globally.

**Example 9.2** An alternative method defines a grid such that the boundary is at the middle of the first interval. We can then employ a box-scheme approximation of (9.5); i.e.,

\[ \frac{1}{\Delta x} (u_e - u_c) = \frac{1}{2} q (u_c + u_e) + g(y_c). \]

From this relation we can eliminate \( u_c \) and find

\[ u_c = \frac{(1 - \frac{1}{2} q \Delta x) u_e - \Delta x g(y_c)}{1 + \frac{1}{2} q \Delta x}. \]


9.2 The Maximum Principle

In the previous section we constructed the central difference scheme, which may be written as a discrete operator \( \mathcal{L}_\Delta \) acting on a grid function \( u_\Delta \) at a generic point of a grid \( \Omega_\Delta \). Figure 9.2. Stencil for the Robin boundary condition (9.5) without (left) and with (right) a virtual grid point \( W \).
Chapter 9. Numerical Methods for Elliptic Equations

Figure 9.3. An unconnected grid; grid point $x_c$ is not connected to any other grid point.

covering the domain. So for an interior grid point $x_c$ we have

$$
\mathcal{L}_\Delta [u_c] = f_c,
$$

(9.10)

where $u_c = u_\Delta(x_c)$ is the numerical approximation of $u(x_c)$ and $f_c$ is the source term, including contributions from the boundary conditions when applicable. If $x_c$ is a genuine interior grid point, i.e., it has no neighbouring grid points on the boundary, then $f_c = f(x_c)$. For the theory that follows it is important that such a point be a generic interior point. We have illustrated this in Figure 9.3. Apparently, the grid point $x_c$ is not connected to any of its neighbours, so the total set of difference equations is uncoupled. We now require that any two interior grid points can be connected by a curve that follows grid lines in the interior of the domain. Such a grid will be called connected. One should note that a grid can always be assumed connected for $\Delta x, \Delta y$ small enough. We remark that an interior grid point corresponds to an unknown value. Also, a grid point on that part of the boundary where either a Neumann or a Robin condition is specified contains an unknown value. For such points we proceed as follows. Consider, e.g., the discrete Robin condition (9.6), which we rewrite as

$$
\frac{1}{\Delta x} \left( 0 u_w - (1 + q \Delta x) u_c + u_w \right) = g(y_c).
$$

(9.11)

Then we can view $x_c$ as an interior grid point. Note that $x_w$ is a virtual point that does not play a role as such; neither does the value $u_w$ in the difference equation. By doing this for all such points, we see that we can interpret our discrete problem as one with Dirichlet boundary conditions exclusively; the actual solution value at such points may be taken as zero so that the points correspond to (virtual) homogeneous Dirichlet boundary conditions. In more complex situations the difference equation, arising after having eliminated the virtual point(s) from a discrete Robin or Neumann boundary condition, can be treated likewise: the values at virtual points are incorporated in the difference equation with a zero coefficient. In what follows we need at least one real Dirichlet or Robin boundary point. This corresponds of course to the well-posedness of the continuous problem, which requires Robin or Dirichlet data, at least on part of the boundary. So for a two-dimensional problem with central differences and with $x_c$ as any interior point in the general sense as described
9.2. The Maximum Principle

above, we write the difference equation as

\[ \mathcal{L}_\Delta[u_c] := \alpha_0 u_c + \sum_{i=1}^{4} \alpha_i u_{n_i}. \] (9.12)

Here we have written \( u_{n_1}, \ldots, u_{n_4} \) for \( u_n, u_e, u_s, \) and \( u_w \) in a more generic way. Note that the coefficients in (9.12) coincide with those in (9.3) if \( x_c \) is a genuine interior point. We first deal with the case where we have a strict equality in (9.4b), i.e., \( \alpha_0 + \sum_{i=1}^{4} \alpha_i = 0 \) (which is, e.g., the case for a Dirichlet problem on a rectangle with a uniform grid).

**Theorem 9.3.** Let the grid \( \Omega_\Delta \) be connected and let \( \mathcal{L}_\Delta[u_c] = 0 \) for all interior grid points \( x_c \), where \( \mathcal{L}_\Delta \) is defined in (9.12). Moreover, the coefficients \( \alpha_0, \ldots, \alpha_4 \) satisfy the conditions

\[ \alpha_0 + \sum_{i=1}^{4} \alpha_i = 0, \quad \alpha_i > 0 \quad (i = 1, 2, 3, 4). \]

Then for any grid point \( x_c \) we have

\[ m := \min_{x_{n_1} \in \partial \Omega_\Delta} u_{n_1} \leq u_c \leq \max_{x_{n_1} \in \partial \Omega_\Delta} u_{n_1} =: M, \]

where \( \partial \Omega_\Delta \) denotes the set of grid points on the boundary.

**Proof.** We prove this by contradiction. Suppose that \( u_c > M \) at some interior point \( x_c \). Since \( \alpha_0 \neq 0 \) (because \( \alpha_0 < 0 \)), we find from (9.3) that

\[ u_c = -\sum_{i=1}^{4} \frac{\alpha_i}{\alpha_0} u_{n_i}, \]

i.e., \( u_c \) is a weighted average of its neighbouring values \( u_{n_i} \). This equality only makes sense if \( u_{n_i} = u_c \) for all \( i \). Repeating this argument, we conclude that all interior points must have a value equal to \( u_c \). At least one is connected to a boundary point where the value \( M \) is taken. This leads to a contradiction. A similar argument can be applied to see that \( u_c \geq m \) at all interior grid points. The case that \( x_c \) is a boundary point is trivial. \( \square \)

**Property 9.4 (maximum principle).** Let the grid \( \Omega_\Delta \) be connected and let \( \mathcal{L}_\Delta[u_c] \geq 0 \) for all interior grid points. Then for any interior grid point \( x_c \) we have

\[ u_c \leq M := \max_{x_{n_1} \in \partial \Omega_\Delta} u_{n_1}. \]

**Proof.** Suppose that \( u_c > M \) for some interior grid point \( x_c \). From \( \mathcal{L}_\Delta[u_c] \geq 0 \) and \( \alpha_0 < 0 \) we deduce that

\[ u_c \leq -\sum_{i=1}^{4} \frac{\alpha_i}{\alpha_0} u_{n_i} \leq u_c. \]

This leads to the same conclusion; i.e., \( u_{n_i} = u_c \), etc. \( \square \)
Example 9.5 Consider the two-point boundary value problem

\[ \mathcal{L}[u] := \frac{d^2 u}{dx^2} = 6x, \quad x \in (0, 1), \]

\[ u(0) = u(1) = 0, \]

which has the solution \( u(x) = x^3 - x \). If we use central differences, we obtain the difference equation

\[ \mathcal{L}_\Delta[u_c] := \frac{1}{\Delta x^2} (u_e + u_w - 2u_c) = 6x_c. \]

Since central differences are exact for polynomials up to degree three, we see that \( u_c(x_c) < 0 \). This is in agreement with Property 9.4, from which we conclude that \( u_c < \max(u(0), u(1)) = 0 \), indeed.

Like in the continuous case there exists a comparison theorem as well.

**Theorem 9.6 (comparison theorem).** Consider the two operator equations

\[ \mathcal{L}_\Delta[u_c] = f_c, \quad \mathcal{L}_\Delta[v_c] = g_c, \]

with \( f \geq g \) pointwise (i.e., at every grid point). Then

\[ u_c \leq v_c + \max_{x_P \in \partial \Omega_\Delta} (u_P - v_P). \]

**Proof.** Since \( \mathcal{L}_\Delta[u_c - v_c] \geq 0 \), we can deduce this directly from Property 9.4.

Example 9.7 If we have the two-point boundary value problem

\[ \mathcal{L}[u] := \frac{d^2 u}{dx^2} = \epsilon^*, \quad x \in (0, 1), \]

\[ u(0) = u(1) = 0, \]

we can use the comparison theorem and estimate the solution of

\[ \mathcal{L}_\Delta[u_c] = \exp(x_c) \]

in terms of the solution given in Example 9.5. Indeed, we have \( \epsilon^* \geq (e/6) 6x \) for \( x \in (0, 1) \).

If we use the auxiliary function \( v(x) := \frac{\epsilon}{6} (x^3 - x) \), which is the solution of

\[ \mathcal{L}_\Delta[v_c] = \epsilon, \quad v(0) = v(1) = 0, \]

as the comparison function, we obtain from the comparison theorem that

\[ u_c \leq v_c + \max(u(0) - v(0), u(1) - v(1)) = \frac{\epsilon}{6} (x_c^3 - x_c). \]

Again, like in the continuous case, we can easily show uniqueness of a solution on a connected grid.

**Theorem 9.8.** On a connected grid the solution of (9.3) with suitable boundary data is unique.
9.3. Estimates of the Global Error

Proof. Let $u_\Delta$ and $v_\Delta$ be two such solutions. Then $u_\Delta - v_\Delta$ satisfies the homogeneous problem with homogeneous data. From Theorem 9.4 it then follows that $u_\gamma - v_\gamma \leq 0$. By interchanging $u_\Delta$ and $v_\Delta$, we similarly find $v_\gamma - u_\gamma \leq 0$. This shows that $u_\gamma = v_\gamma$ for any interior point $x_\gamma$. □

9.3 Estimates of the Global Error

The actual global discretisation error for an elliptic problem is typically found by solving a system with the local error as the known right-hand side. In this section we shall consider two approaches to estimating this error. In Section 9.3.1 our results will use the maximum principle to obtain suitable bounds. Although this method is not totally unrelated to the matrix method (actually solving the aforementioned system), we consider the latter separately in Section 9.3.2. Finally, in Section 9.3.3, we investigate the effect of errors originating from the boundary.

9.3.1 Estimates Based on the Maximum Principle

Let us denote by $A$ the matrix with coefficients as appearing in the difference equations, i.e., from collecting all difference equations (9.3) for all interior grid points, including possible boundary points where $u$ is not given. In the following we will denote this set of grid points by $\Omega_\Delta$. We can order the unknowns in many ways. Although this is a nontrivial matter, we leave this question of ordering for the time being. It is extremely important, however, when computing the numerical solution. Writing the vector of unknowns as $u$ and the known forcing and boundary data as $f$, we have

$$ Au = f. \quad (9.13) $$

So we have two different representations for the numerical solution: viz. the grid function $u_\Delta$ satisfying the operator equation (9.10) and the vector of unknowns $u$ satisfying the linear system (9.13). We will use both formulations for the numerical solution and also for the discretisation errors to be defined next.

The local discretisation error is actually the residual or defect when the exact solution of (9.1) is substituted in (9.13). Let us denote the vector of values at the (similarly ordered!) grid points of the exact solution in a similar fashion by $u^\ast$. Then the local discretisation error $d$ is defined by

$$ Au^\ast =: f + d. \quad (9.14) $$

Since the matrix $A$ is constant, the global discretisation error

$$ e := u^\ast - u \quad (9.15) $$

simply follows from (9.13) and (9.14); i.e.,

$$ Ae = d. \quad (9.16) $$

We see that we have to solve a system to find the global error. Employing the maximum principle for this is conceptually simple. We first look for a solution of the problem with
a right-hand side bounded below by one. We then use an appropriate scaling and use the resulting solution as the comparison function. This is borne out in the following theorems. Inequalities for vectors are meant to hold elementwise. To start with, we restrict ourselves to Dirichlet boundary data on a suitable domain so that there are no approximation errors coming from the boundary (one may think of a domain where the boundary coincides with grid lines). It is important to realise that a difference equation for the global error holds at any interior point (where the solution is unknown); here the coefficient corresponding to the boundary point(s) is taken as zero. As remarked before, we may as well take the boundary value for the error to be zero. Thus we have a homogeneous Dirichlet problem for the error.

**Theorem 9.9.** Assume that there exists a vector \( \mathbf{v} \) with nonnegative entries such that
\[
A \mathbf{v} \geq \mathbf{e} := (1, 1, \ldots, 1)^T.
\]
Then the global discretisation error is bounded by
\[
\| \mathbf{e} \|_\infty \leq \max_{x_C \in \partial \Omega_\Delta} (\mathbf{v}_C) \| \mathbf{d} \|_\infty.
\]

**Proof.** Since \( L_\Delta [v_C] \geq 1 \) for all grid points \( x_C \in \Omega_\Delta \), we clearly have
\[
L_\Delta [\varphi v_C - e_C] \geq \varphi - d_C \geq 0,
\]
where \( \varphi := \| \mathbf{d} \|_\infty \). Applying the maximum principle (see Property 9.4), we obtain
\[
-e_C \leq \varphi v_C - e_C \leq \max_{x_C \in \partial \Omega_\Delta} (\varphi v_C - e_C) = \varphi \max_{x_C \in \partial \Omega_\Delta} v_C.
\]
In the derivation of (*) we have used that \( e_C = 0 \) at the boundary grid points. Likewise, from
\[
L_\Delta [\varphi v_C + e_C] \geq \varphi + d_C \geq 0,
\]
we can derive the inequalities
\[
e_C \leq \varphi v_C + e_C \leq \max_{x_C \in \partial \Omega_\Delta} (\varphi v_C + e_C) = \varphi \max_{x_C \in \partial \Omega_\Delta} v_C.
\]
Combining the inequalities in (*) and (**), we find
\[
|e_C| \leq \varphi \max_{x_C \in \partial \Omega_\Delta} v_C
\]
for all \( x_C \in \Omega_\Delta \).

We remark that a comparison function \( v_\Delta \), like in the theorem above, is often chosen as the solution of a suitable continuous problem restricted to the grid \( \Omega_\Delta \). Since the right-hand side is not so important (although its maximum at \( \partial \Omega \) should not be too large of course), one can comfortably restrict oneself to polynomial solutions of sufficiently low degree.

**Example 9.10** Consider the Poisson problem
\[
L[u] := \nabla^2 u = f(x, y), \quad (x, y) \in \Omega := (0, 1) \times (0, 1),
\]
\[
u(x, y) = 0, \quad (x, y) \in \partial \Omega.
\]
9.3. Estimates of the Global Error

For a constant grid size \( h = \Delta x = \Delta y \) we obtain the difference scheme

\[
\mathcal{L}_h[u_c] := \frac{1}{h^2} \left( u_n + u_e + u_s + u_w - 4u_c \right) = f(x_c).
\]

As comparison function \( v_\Delta \) we use the restriction of

\[
v(x, y) := \frac{1}{4} (x^2 + y^2).
\]

Clearly we have \( \mathcal{L}_h[v_c] = 1 \) and \( \max_{x,y \in \Omega_1} (v_{x c}) = \frac{1}{2} \). Applying Theorem 9.9, we conclude that

\[
|u_c| \leq \frac{1}{2} \max_{x,y \in \Sigma} |f(x_c)|.
\]

Applying this result to (9.16), we can determine an upper bound for the global discretisation error.

Next we can prove convergence of the scheme in the following sense.

**Theorem 9.11.** If the local discretisation error of (9.10) is \( O(h^p) \), then for a well-posed problem we also have an \( O(h^p) \) global discretisation error (pointwise).

**Proof.** For a well-posed problem we have at least one point where a Dirichlet/Robin condition is given, so that a maximum principle holds. The result then follows from Theorem 9.9, applied to (9.16), noting that we should have zero data for the error at the boundary.

**Example 9.12** Consider the boundary value problem

\[
\mathcal{L}[u] := \nabla^2 u = -2 \sin x \sin y, \quad (x, y) \in \Omega := (0, 1) \times (0, 1),
\]

\[
u(x, y) = \begin{cases} 
0 & \text{if } x = 0, y \in (0, 1), \\
0 & \text{if } y = 0, x \in (0, 1), \\
sin 1 \sin y & \text{if } x = 1, y \in (0, 1), \\
sin x \sin 1 & \text{if } y = 1, x \in (0, 1).
\end{cases}
\]

Clearly, this problem has the solution \( u(x, y) = \sin x \sin y \). Discretising with central differences and uniform grid size \( h \) gives a local discretisation error equal to (cf. (5.30))

\[
d(x_j, y_k) = -\frac{1}{12} h^2 \left( \frac{\partial^4 u}{\partial x^4}(\xi, y_k) + \frac{\partial^4 u}{\partial y^4}(x_j, \eta) \right),
\]

with \( \xi \in (x_j - h, x_j + h) \) and \( \eta \in (y_k - h, y_k + h) \). This can be bounded by

\[
|d(x_j, y_k)| \leq \frac{1}{6} h^2.
\]

Using the comparison function \( v_\Delta \) as defined in Example 9.10, we obtain

\[
\max_{x,y \in \Sigma} |v_c| \leq \frac{1}{2} \frac{1}{6} h^2 = \frac{1}{12} h^2.
\]

\( \square \)
We can sometimes improve our error estimates by distinguishing between two subgrids and considering estimates there separately. In particular, this may be useful when some of the boundary data are corrupted with errors. The latter may be due to the fact that the grid does not coincide with a boundary (cf. (5.35)), or it may be due to discretisation errors arising from Robin or Neumann boundary conditions. The next theorem deals with this question.

**Theorem 9.13.** Let $\Omega_\Delta^*$ be a subgrid of $\Omega_\Delta$ containing a part of the boundary. Denote by $x_r^*$ a grid point in $\Omega_\Delta^*$ and by $x_r$ an interior grid point in $\Omega_\Delta \setminus \Omega_\Delta^*$. Assume there exists a nonnegative grid function $v_\Delta$ such that

$$L_\Delta[v_r] \geq 1, \quad x_r \in \Omega_\Delta \setminus \Omega_\Delta^*,
$$

$$L_\Delta[v_r^*] \geq \beta > 0, \quad x_r^* \in \Omega_\Delta^*.$$

Then the grid function $e_\Delta$ corresponding to the global discretisation error $e$ (with homogeneous boundary conditions) is estimated by

$$|e_c| \leq \max_{x_r \in \Omega_\Delta} (v_r) \frac{\max(\psi, \psi^*/\beta)}{\max_{x_r^* \in \Omega_\Delta^*} |d_r^*|},$$

where

$$\psi := \max_{x_r \in \Omega_\Delta} |d_r|, \quad \psi^* := \max_{x_r^* \in \Omega_\Delta^*} |d_r^*|.$$ 

**Proof.** For $x_r \in \Omega_\Delta \setminus \Omega_\Delta^*$ and $x_r^* \in \Omega_\Delta^*$ we have, respectively,

$$L_\Delta[\psi^* v_r \pm e_r] = \psi L_\Delta[v_r] \pm L_\Delta[e_r] \geq \psi \pm \psi^* \geq 0,$$

$$L_\Delta[\frac{1}{\beta} \psi^* v_r^* \pm e_r^*] = \frac{1}{\beta} \psi^* L_\Delta[v_r^*] \pm L_\Delta[e_r^*] \geq \psi^* \pm \psi^* \geq 0.$$

Combining both formulas, we conclude that

$$L_\Delta[\tilde{\psi} v_r \pm e_r] \geq 0$$

for all $x_r \in \Omega_\Delta$, where $\tilde{\psi} := \max(\psi, \psi^*/\beta)$. The proof then follows the same formal reasoning as that of Theorem 9.9. \hfill \Box

To show how Theorem 9.13 may be used we have the following example.

**Example 9.14** Consider the two-point boundary value problem

$$L[u] := \frac{d^2u}{dx^2} = e', \quad x \in (0, 1),$$

$$\frac{du}{dx}(0) = 1, \quad u(1) = e.$$

If we discretise the problem on the grid $\Omega_\Delta = \{x_0, x_1, \ldots, x_M\}$, where $x_j = j \Delta x$, $\Delta x = 1/(M + 1)$, we obtain the scheme

$$L_\Delta[u_j] := \frac{1}{\Delta x^2} (u_{j+1} - 2u_j + u_{j-1}) = \exp(x_j), \quad j = 0, 1, 2, \ldots, M.$$
9.3. Estimates of the Global Error

The right boundary condition is straightforward. In order to deal with the left boundary condition, we introduce a virtual point \( x_{-1} = -\Delta x \) and apply central differences at \( x_0 \), giving

\[ u_1 - u_{-1} = 2\Delta x. \]

Eliminating \( u_{-1} \) from this relation and substituting into the difference equation for \( j = 0 \), we obtain

\[ \mathcal{L}_\Delta[u_0] := \frac{2}{\Delta x^2} (u_1 - u_0) = 1 + \frac{2}{\Delta x}. \tag{*} \]

We may view (\( * \)) as a difference equation at the points \( x_{-1}, x_0, x_1 \) with a (Dirichlet) boundary condition; i.e.,

\[ \mathcal{L}_\Delta[u_0] = \frac{1}{\Delta x^2} (2u_1 - 2u_0 + 0u_{-1}). \]

As a (discrete) comparison function we now take

\[ v_j := \frac{1}{2} x_j (x_j - p) \]

for some \( p \). We can choose \( p \) such that the comparison function satisfies our requirements and even such that it minimises the global error. We have

\[ \mathcal{L}_\Delta[v_j] = 1, \quad j = 1, 2, \ldots, M, \]

and

\[ \mathcal{L}_\Delta[v_0] = 1 - \frac{p}{\Delta x}. \]

Obviously, \( \Omega_\Delta^* = \{x_0\} \). So a possible choice is \( p = -1 \), implying that \( \beta = 1 - 1/\Delta x \). Since the problem is smooth we see that \( \varphi = \mathcal{O}(\Delta x^2) \) and \( \varphi^* = \mathcal{O}(\Delta x) \). Moreover, \( \max_{x \in \Omega_\Delta}(v_0) = 1 \).

Application of Theorem 9.13 then yields \( \|e\|_\infty = \mathcal{O}(\Delta x^2) \), which is sharper than \( \mathcal{O}(\Delta x) \), which one would have obtained by direct application of the comparison theorem (and might have expected because the local error arising from the boundary condition is first order). Optimising for \( p \) will not give a higher order (as second order is the best one can get, of course). One may, however, try to optimise \( p \) such that the coefficient in the error estimate is as small as possible. We omit this.

\[ \square \]

9.3.2 Error Estimates Using the Matrix Method

Although the maximum principle is a very powerful tool, there are situations where employing the matrix representation may be at least as useful, as will be illustrated here and in Chapter 11. At this point it is important to remark that some useful properties only hold if the grid points are ordered in a special way. For our purposes (and two-dimensional problems) it is sufficient to assume that they are ordered lexicographically (in reading order). Another useful ordering is red-black, i.e., first the red squares of a chess board and then the black ones (again in reading order).

To start we introduce a number of concepts for matrices arising from discretised PDEs with suitable orderings. The first one is positivity of vectors and matrices.

**Definition 9.15.** A vector \( \mathbf{v} = (v_1, v_2, \ldots, v_n)^T \) is called positive and denoted \( \mathbf{v} > 0 \) if \( v_i > 0 \) for \( i = 1, 2, \ldots, n \). A matrix \( \mathbf{A} = (a_{ij}) \) is called positive and denoted \( \mathbf{A} > 0 \) if \( a_{ij} > 0 \) for \( i, j = 1, 2, \ldots, n \). It is called semipositive and denoted \( \mathbf{A} \geq 0 \) if \( a_{ij} \geq 0 \) for \( i, j = 1, 2, \ldots, n \).
**Theorem 9.17.** Let \( \hat{I} + I \geq A \geq 0 \) for \( i = 1, 2, \ldots, n \) and \( a_{i,j} \leq 0 \) for \( j = 1, 2, \ldots, n, \ j \neq i \). An \( n \times n \) matrix \( A = (a_{i,j}) \) is called an \( M \)-matrix if \( a_{i,j} \leq 0 \) for \( i, j = 1, 2, \ldots, n, \ j \neq i \), \( A \) is nonsingular, and its inverse \( A^{-1} \) is semipositive.

Matrix and vector inequalities like \( A \leq B \) follow in an obvious way from Definition 9.15. It is not difficult to see that an \( M \)-matrix is an \( L \)-matrix. The following theorem gives a sufficient condition for the converse.

**Theorem 9.17.** Let \( A \) be an \( L \)-matrix. Let \( D \) denote the diagonal of \( A \), \( C := D - A \), and \( \hat{B} := D^{-1} C \). Then \( A \) is an \( M \)-matrix if and only if \( \rho(\hat{B}) < 1 \).

**Proof.** Suppose \( \rho(\hat{B}) < 1 \). Then the matrix \( I - \hat{B} \) is nonsingular and the series \( I + \hat{B} + \hat{B}^2 + \cdots \) converges to \((I - \hat{B})^{-1}\). Since \( D, C \geq O \), we conclude that \( B \) and \((I - \hat{B})^{-1}\) are semipositive. Moreover, since \( D \) and \( I - \hat{B} \) are nonsingular, it follows that \( A = D(I - \hat{B}) \) is also nonsingular and \( A^{-1} = (I - \hat{B})^{-1} D^{-1} \) is semipositive. Hence \( A \) is an \( M \)-matrix. Conversely, let \( A \) be an \( M \)-matrix. Define

\[ \hat{A} := D^{-1/2} A D^{-1/2}. \]

Then \( \hat{A} = I - \hat{B}^{1/2} \hat{B} \hat{B}^{-1/2} \). Clearly, \( \hat{A} \) is an \( M \)-matrix. If we finally define

\[ \hat{B} := D^{1/2} B D^{-1/2}, \]

which is clearly similar to \( B \), we can write

\[ \hat{A}^{-1} = (I - \hat{B})^{-1} = (I + \hat{B} + \hat{B}^2 + \cdots + \hat{B}^m) + (I - \hat{B})^{-1} \hat{B}^{m+1}. \]

Since \((I - \hat{B})^{-1}\) and \( \hat{B} \) are semipositive as well, we conclude that \( I + \hat{B} + \hat{B}^2 + \cdots + \hat{B}^m \) is bounded above elementwise by the corresponding elements of \((I - \hat{B})^{-1}\), which implies that the series \( I + \hat{B} + \hat{B}^2 + \cdots \) converges; i.e., \( \rho(\hat{B}) = \rho(B) < 1 \).

One can give a useful property that gives a pointwise estimate of a matrix and thus can be seen as a matrix analogue of the comparison theorem.

**Property 9.18.** Let \( B \) be an \( M \)-matrix. Let \( A := B + F \), where we have the diagonal matrix \( F \geq O \). Then \( A^{-1} \leq B^{-1} \).

**Proof.** Write \( B := D(I - J) \) and \( A := E(I - K) \), with \( D \) and \( E \) the diagonal parts of \( B \) and \( A \), respectively. Since \( \rho(J) < 1 \) (see Theorem 9.17), \((I - J)^{-1}\) exists. By assumption we moreover have \( K \leq J \), which implies that \( K^i \to 0 \) for \( i \to \infty \); i.e., \((I - K)^{-1}\) exists. Actually, we can even deduce from \( K^i \leq J^i \) that \((I - K)^{-1} \leq (I - J)^{-1} \). Combining results, we see that \((I - K)^{-1} E^{-1} \leq (I - J)^{-1} D^{-1} \); i.e., \( A^{-1} \leq B^{-1} \).

Typically, the matrices we have encountered will give rise to the negative of \( M \)-matrices. However, rather than considering the problem

\[ \nabla^2 u = f, \]
9.3. Estimates of the Global Error

one may write
\[-\nabla^2 u = \tilde{f} := -f,\]
as we shall do henceforth in this chapter. The usual central differences then give rise to an M-matrix, as is borne out in the next example.

Example 9.19 If we use central differences to discretise the boundary value problem
\[-\frac{d^2 u}{dx^2} = -bu, \quad x \in (0, 1),\]
\[u(0) = u(1) = 0,\]
where \(b > 0\), we obtain the matrix
\[A = B + bI, \quad B := \frac{1}{\Delta x^2} \begin{pmatrix} 2 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \cdots & 0 \\ 0 & 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{pmatrix},\]
where \(\Delta x = 1/(M + 1)\) is the grid size. Clearly, \(B\) is an L-matrix. The matrix \(D^{-1}C\), where \(D\) is the diagonal part of \(B\) and \(C = D - B\), is symmetric and has Gershgorin circles with radius one, centred around zero. The only possible eigenvalues \(\lambda\) with \(|\lambda| = 1\) might be \(+1, -1\). It is trivial to see that neither value is an eigenvalue, which implies that \(\rho(D^{-1}C) < 1\). So, according to Theorem 9.17, \(B\) is an M-matrix. In fact, we can give the explicit inverse of \(B^{-1}\) as
\[B^{-1} := \frac{1}{(M + 1)^2} \begin{pmatrix} M & M - 1 & M - 2 & \cdots & 1 \\ M - 1 & (M - 1)2 & (M - 2)2 & \cdots & 2 \\ M - 2 & (M - 2)2 & (M - 3)3 & \cdots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \cdots & M \end{pmatrix},\]
see also Example 9.28. We find \(\|B^{-1}\|_\infty = \frac{1}{2}\). Following the same reasoning as above, we can prove that \(A\) is also an M-matrix, and, moreover, \(A^{-1} \leq B^{-1}\). From this inequality we see that \(\|A^{-1}\|_\infty \leq \|B^{-1}\|_\infty\). \(\square\)

We will now apply these properties of M-matrices to estimate the global discretisation error. Consider the linear system (9.13). The corresponding global discretisation error \(e\) satisfies the system (9.16). If the matrix \(A\) is an M-matrix, we find
\[|e| \leq A^{-1} |d|, \quad (9.17)\]
where the absolute value of the vectors has to be taken elementwise; i.e., \(|e| := (|e_1|, |e_2|, \ldots, |e_N|)^T\), with \(e_j\) the global discretisation error in the \(j\)th grid point and \(N\) the number of points in \(\Omega_\Delta\), and \(|d|\) is defined likewise. This immediately gives, for the problem in Example 9.19, that
\[\|e\|_\infty \leq \|A^{-1}\|_\infty \max_j |d_j| \leq \frac{C}{8} \Delta x^2, \quad (9.18)\]
where $C$ is a bound on the fourth derivatives of the exact solution $u(x)$. Of course it is crucial to have an estimate of $\|A^{-1}\|_\infty$. For more general problems it is often not so easy to obtain an estimate for $\|e\|_\infty$ and we proceed as follows. Introducing the vector

$$e := (1, 1, \ldots, 1)^T,$$

we can write, instead of (9.17),

$$|e| \leq A^{-1} e \max_j |d_j| = v \max_j |d_j|,$$

where the vector $v$ is the solution of the system

$$Av = e.$$  (9.19)

In many situations the (positive) vector $v$ can be estimated rather sharply from above, thus making it possible to estimate the global discretisation error through (9.19). We demonstrate this by an example. One should note the similarity with the use of a comparison function in Section 9.3.1.

**Example 9.20** For the problem in Example 9.19 with $b = 0$ we define the vector $v$ to be such that its $j$th coordinate is $v_j = \frac{1}{2} x_j (1 - x_j)$. This is based on the fact that the difference equation is exact for polynomial solutions of degree up to two. It is easy to see that (9.20) holds. Clearly, $\|v\|_\infty = \frac{1}{8}$, confirming the estimate in (9.18).

Usually $v$ can only be estimated from above by some $\tilde{v}$, say, such that

$$e = Av \leq A\tilde{v}.$$  (9.20)

The next example demonstrates this for a problem that has a Robin boundary condition at one side. In the next subsection we will consider the global error due to discretising this. Here we will be concerned with estimating the effect of the local discretisation errors arising from the interior.

**Example 9.21** Consider the boundary value problem

$$-\frac{d^2 u}{dx^2} = f(x), \quad x \in (0, 1),$$

$$\frac{du}{dx}(0) = u(0), \quad u(1) = 0.$$  

To discretise this problem we employ central differences on the grid $\Omega_{1/\Delta x} = \{x_0, x_1, \ldots, x_M\}$ with $x_j = j/\Delta x, \Delta x = 1/(M + 1)$. We introduce the virtual point $x_{-1}$ and apply a central difference to discretise the Robin boundary condition at the left boundary as

$$\frac{u_1 - u_{-1}}{2\Delta x} = u_0.$$  

After eliminating $u_{-1}$ using the difference scheme at the points $x_{-1}, x_0, x_1$, we obtain the global system

$$A := \frac{1}{\Delta x^2} \begin{pmatrix} 2(1 + \Delta x) & -2 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}.$$
9.3. Estimates of the Global Error  

As one can easily see, the vector \( \tilde{v} = (\tilde{v}_0, \tilde{v}_1, \ldots, \tilde{v}_M)^T \) with \( \tilde{v}_i = \frac{1}{4}(1 - x_i^2) \) satisfies \( A\tilde{v} \geq e \). Since \( \|\tilde{v}\|_\infty \leq \frac{1}{2} \), we find the estimate \( \|e\|_\infty \leq \frac{1}{2} C \Delta x^2 \), with \( C \) a bound on the third and fourth derivatives of \( u(x) \). □

We conclude this subsection by demonstrating the analysis for a problem in two dimensions.

**Example 9.22**  Consider the boundary value problem

\[
\begin{align*}
-\nabla^2 u &= f(x), & x &\in \Omega := (0, 1) \times (0, 1), \\
\frac{\partial u}{\partial y}(x, 0) &= u(x, 0), & x &\in (0, 1), \\
u(x, 1) &= g_2(x), & x &\in (0, 1), \\
\frac{\partial u}{\partial x}(0, y) &= u(0, y), & y &\in (0, 1), \\
u(1, y) &= g_4(y), & x &\in (0, 1).
\end{align*}
\]

To discretise this problem we use central differences on a uniform grid with grid spacing \( h = \Delta x = \Delta y = 1/(M + 1) \) for some \( M \). In order to discretise the boundary conditions at \( y = 0 \) and \( x = 0 \), we have to introduce virtual grid points at the lines \( y = -h \) and \( x = -h \), respectively. The virtual points arising in the difference equation are then eliminated via the discrete boundary condition, like in Example 9.14. Typically, a stencil at the left boundary then looks like Figure 9.4. So we have (away from the corners)

\[
\frac{1}{h^2}((4 + 2h)u_c - u_N - u_S - 2u_E) = f(x_c).
\]

One thus obtains \( M + 1 \) grid lines in the \( y \) direction (at \( x_j, j = 0, 1, \ldots, M \)) and \( M + 1 \) grid lines in the \( x \) direction (at \( y_k, k = 0, 1, \ldots, M \)). In analogy with the previous example, we choose

\[
\tilde{v}_P = \frac{1}{4}(1 - x_j^2) + \frac{1}{4}(1 - y_k^2),
\]

with \( \tilde{v}_P \) the component of \( \tilde{v} \) corresponding to the grid point \( x_P = (x_j, y_k) = (jh, kh) \). We can again easily check that \( \tilde{v} \) satisfies \( A\tilde{v} \geq e \). Moreover, \( \|\tilde{v}\|_\infty = \frac{1}{2} \), implying an error estimate of the form \( \|e\|_\infty \leq \frac{1}{2} C h^2 \) for some constant \( C \). □

![Figure 9.4. Stencil at the left boundary.](image)

\[ \begin{array}{c}
\text{N} \\
\text{C} \\
\text{E} \\
\text{S}
\end{array} \]
9.3.3 The Effect of Errors Originating from the Boundary

The boundary data need special attention. For the sake of simplicity, we restrict ourselves to a uniform grid of size \( h \) in either one or two space dimensions. Recalling the vector \( e \) introduced in Section 9.3.2, we define a vector \( e_{\text{bnd}} \) to be found from \( e \) by making all components zero if the corresponding grid point is a genuine interior point, i.e., has no neighbours on the boundary. The genuine interior points induce a vector \( e_{\text{int}} \) defined by

\[
e_{\text{int}} := e - e_{\text{bnd}}.
\]

From (9.17) we thus find

\[
|e| \leq A^{-1} \left( e_{\text{int}} \max_{\text{int}} |d_{\text{int}}| + e_{\text{bnd}} \max_{\text{bnd}} |d_{\text{bnd}}| \right) .
\]

(9.22)

Here \( d_{\text{int}} \) and \( d_{\text{bnd}} \) denote the local discretisation errors at the internal and boundary points, respectively. The crucial point is now to realise that \( A \) is usually weakly diagonally dominant; i.e., \( |a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \) for all \( i \), with an inequality for at least one \( i \), corresponding to a grid point that is not a genuine interior point. Hence, for some constant \( C_{\text{bnd}} > 0 \), which depends on the problem at hand, we have \( A e \geq C_{\text{bnd}} e_{\text{bnd}} / h^2 \), giving

\[
A^{-1} e_{\text{bnd}} \leq \frac{h^2}{C_{\text{bnd}}} e.
\]

(9.23)

Since \( A^{-1} e_{\text{int}} \leq A^{-1} e \), we estimate the global effect of the interior local errors, say \( \|A^{-1} e\|_{\infty} \leq C_{\text{int}} \) for some \( C_{\text{int}} \). Combining this with (9.22), we then obtain

\[
\|e\|_{\infty} \leq C_{\text{int}} \max_{\text{int}} |d_{\text{int}}| + \frac{h^2}{C_{\text{bnd}}} \max_{\text{bnd}} |d_{\text{bnd}}| .
\]

(9.24)

The result in (9.24) implies that local errors at the boundary may be two orders less than the local errors made in the interior without affecting the order of the global error. This phenomenon explains why the often thoughtless treatment of derivative boundary conditions may not affect the overall result dramatically.

**Example 9.23** Consider the two-point boundary value problem

\[
-\frac{d^2 u}{dx^2} = f(x), \quad x \in (0, 1),
\]

\[
u(0) = 0, \quad u(1) = R,
\]

where \( R \) stands for a representation that can only be approximated; say we use \( \bar{R} \) instead of \( R \), with \( \bar{R} = R + \delta \). After discretising the problem with central differences with grid size \( h \), we obtain the matrix

\[
A := \frac{1}{h^2} \begin{pmatrix}
2 & -1 & & \\
-1 & 2 & -1 & \\
& \ddots & \ddots & \ddots \\
& & -1 & 2 & -1 \\
& & & -1 & 2
\end{pmatrix}.
\]

From this we see that we can take \( C_{\text{bnd}} = 1 \) in (9.24). The effect of the approximation error in the boundary condition is only globally felt as \( h^2 \delta \). \( \square \)
9.3. Estimates of the Global Error

We can generalise the preceding analysis to problems where boundary condition errors occur at several places on the boundary. To start with let us split the boundary $\Gamma := \partial \Omega$ into a number of subboundaries, $\Gamma_k, k = 1, 2, \ldots, K$, for some $K$, say, and introduce the vectors (cf. (9.22))

$$e = e_{\text{int}} + \sum_{k=1}^{K} e_{b,k},$$

(9.25)

where the vector $e_{b,k}$ contains zero entries if the corresponding grid point has no neighbours on $\Gamma_k$. Still assuming that $A^{-1}$ is positive, we then obtain, in an obvious notation (cf. (9.22)), that

$$|e| \leq \max_{\text{int}} |d_{\text{int}}| A^{-1} e_{\text{int}} + \sum_{k=1}^{K} \max_{\Gamma_k} |d_{b,k}| A^{-1} e_{b,k}.$$  

(9.26)

For problems with derivative boundary conditions we now try to find positive vectors $w_{b,k}$ and positive constants $C_{b,k}$ such that

$$A w_{b,k} \geq \frac{C_{b,k}}{h^2} e_{b,k},$$

(9.27)

so that we have (cf. (9.23))

$$A^{-1} e_{b,k} \leq \frac{h^2}{C_{b,k}} w_{b,k}, \quad k = 1, 2, \ldots, K.$$  

(9.28)

The eventual estimate then reads

$$\|e\|_\infty \leq C_{\text{int}} \max_{\text{int}} |d_{\text{int}}| + h^2 \sum_{k=1}^{K} \max_{\Gamma_k} |d_{b,k}| \frac{\|w_{b,k}\|_\infty}{C_{b,k}}.$$  

(9.29)

If some part of the boundary, say with index $k$, has a Dirichlet boundary condition, the corresponding estimates above are trivial. Indeed, one can take $w_{b,k} = e$ and $C_{b,k} = 1$, from which it follows that the contribution to the global error is just $h^2 \max_{\Gamma_k} |d_{b,k}|$. The next two examples show how these estimates may be used.

**Example 9.24** Let us replace the boundary condition at $x = 1$ in Example 9.23 by the Robin condition

$$\frac{du}{dx}(1) + u(1) = g.$$  

We can approximate this boundary condition by the box scheme:

$$\frac{1}{h} \left( u \left( 1 + \frac{1}{2} h \right) - u \left( 1 - \frac{1}{2} h \right) \right) + \frac{1}{2} \left( u \left( 1 - \frac{1}{2} h \right) + u \left( 1 + \frac{1}{2} h \right) \right) = g.$$  

As one can easily see this scheme has a discretisation error $O(h^2)$. To discretise this problem we employ central differences on the interior grid $\{x_1, x_2, \ldots, x_M\}$, with $x_j = (j - 1) \frac{h}{2}, h = 1/M$. This way, we obtain a standard problem as in the previous example, but for the last equation. Introducing the virtual grid point $x_{M+1} = 1 + \frac{1}{2} h$, we can apply the box scheme and obtain
a relation between \( u_M \) and \( u_{M+1} \). We can use this latter relation to eliminate \( u_{M+1} \) from the difference equation at \( x_M \). This gives

\[
\frac{1}{h^2} \left( 2 - \frac{1 - \frac{1}{2}h}{1 + \frac{1}{2}h} \right) u_M - u_{M-1} = f_M + \frac{1}{h} \frac{g}{1 + \frac{1}{2}h}.
\]

As a result, the local discretisation error due to discretising the boundary condition is \( O(h) \). It is clear that we should split the boundary into a left part \( \Gamma_{\Delta,1} \), say, and a right part \( \Gamma_{\Delta,2} \), the latter being the only troublesome one. If we take \( w_{b,2} = (1, 2, \ldots, M)^T \), we find that \( A w_{b,2} \) has zeros everywhere except for the last element, which is given by

\[
\frac{1}{h^2} \left( 2 - \frac{1 - \frac{1}{2}h}{1 + \frac{1}{2}h} \right) M - M + 1 = \frac{1}{h^2} \left( 2 + \frac{1}{2}h \right) \geq \frac{1}{h^2}.
\]

Thus the constant \( C_{b,2} \) can be chosen equal to one and we find, for the contribution of the discretisation error of the Robin boundary condition,

\[
h^2 \| w_{b,2} \|_{\infty} \leq \frac{1}{2} M h^2 O(h) = O(h^2).
\]

This result is similar to what we found from applying the maximum principle when using a virtual point and the second order central difference scheme in Example 9.14.

The next example deals with a two-dimensional problem already encountered in Example 9.22.

**Example 9.25** Consider the boundary value problem of Example 9.22. The grid points on the edges \( y = 0, y = 1, x = 0, \) and \( x = 1 \) will be denoted by \( \Gamma_{\Delta,1}, \Gamma_{\Delta,2}, \Gamma_{\Delta,3}, \) and \( \Gamma_{\Delta,4} \), respectively. As in Example 9.22 we use central differences to discretise the Poisson equation and the boundary conditions. An appropriate vector \( w_{b,3} \) then has entries

\[
(w_{b,3})_j = M + 1 - j,
\]

corresponding to grid point \( x_j = (x_j, y_k) \). It is easy to see that

\[
A w_{b,3} \geq \frac{2}{h^2} e_{b,3}.
\]

Thus we find

\[
A^{-1} e_{b,3} \leq \frac{h^2}{2} w_{b,3} \leq \frac{h}{2} e.
\]

We can treat the lower boundary in a similar way to obtain

\[
A^{-1} e_{b,1} \leq \frac{h}{2} e.
\]

Since the local errors from discretising the Neumann boundary conditions are first order (though the central difference is second order), we can conclude again that they have a global effect of second order only. The errors arising form discretising internal points were discussed in Example 9.22.
9.4 Solution of Linear Systems

After discretising a linear elliptic equation, we obtain a system of equations as in (9.13). In Section 9.4.1 we first consider the one-dimensional case, for which this system can most efficiently be solved by standard LU-decomposition. For problems in higher dimensions this becomes inefficient and therefore iterative methods are often employed. This topic is briefly introduced in Section 9.4.2. To conclude we discuss in Section 9.4.3 the closure of iterative methods.

9.4.1 One-Dimensional Dirichlet Problems

Before we deal with problems in higher dimensions, it is instructive to start with a one-dimensional problem. So consider the boundary value problem

\[ a \frac{d^2 u}{dx^2} + c \frac{du}{dx} + e u = f, \quad x \in (0, 1), \quad (9.30a) \]

\[ u(0) = A, \quad u(1) = B. \quad (9.30b) \]

If the coefficient functions \(a, c,\) and \(e\) are constant, we obtain the central difference scheme (cf. (9.2))

\[ \frac{1}{\Delta x^2} \left( a + \frac{1}{2} c \Delta x \right) u_E + \left( a - \frac{1}{2} c \Delta x \right) u_W - (2a - \Delta x^2 e) u_C = f(x_C). \quad (9.31) \]

If we label the grid points \(x_j,\) with the index \(j\) running from 0 to \(M + 1,\) so

\[ x_j = j \Delta x, \quad j = 0, 1, \ldots, M + 1, \quad \Delta x := 1/(M + 1), \]

and denote the numerical approximation of \(u(x_j)\) by \(u_j,\) then (9.31) can be written as a second order difference equation of the form

\[ \alpha_j u_{j-1} + \beta_j u_j + \gamma_j u_{j+1} = f_j, \quad j = 1, 2, \ldots, M. \quad (9.32a) \]

Moreover, from the boundary conditions (9.30b) we conclude that

\[ u_0 = A, \quad u_{M+1} = B. \quad (9.32b) \]

The equations (9.32a) and (9.32b) result in the linear system

\[ Au = f, \quad (9.33) \]

where

\[ A := \begin{pmatrix} \beta_1 & \gamma_1 & & & \\ \alpha_2 & \beta_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha_{M-1} & \beta_{M-1} & \gamma_{M-1} \\ & & & \alpha_M & \beta_M \end{pmatrix}, \quad u := \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{M-1} \\ u_M \end{pmatrix}, \quad b := \begin{pmatrix} f_1 - \alpha_1 A \\ f_2 \\ \vdots \\ f_{M-1} - \alpha_{M-1} A \\ f_M - \gamma_M B \end{pmatrix}. \quad (9.34) \]
Clearly, \( A \) is sparse; i.e., has a large number of systematic zero entries. In this (still rather simple) case we may use a direct solver based on LU-decomposition [52]. It is instructive to work this out. Let us anticipate the existence of this LU-decomposition without pivoting, i.e., without interchanging rows. Say

\[
A = LU,
\]

(9.35)

where the matrices \( L \) and \( U \) are given by

\[
L := \begin{pmatrix}
1 & & & & \\
I_2 & 1 & & & \\
& \ddots & \ddots & & \\
& & I_{M-1} & 1 & \\
& & & I_M & 1
\end{pmatrix}, \quad U := \begin{pmatrix}
v_1 & u_1 & & & \\
v_2 & u_2 & & & \\
& \ddots & \ddots & & \\
& & \ddots & \ddots & \\
v_{M-1} & u_{M-1} & & & \\
v_M & & & & 
\end{pmatrix}.
\]

(9.36)

It is easy to see that the matrices \( L \) and \( U \) maintain the same sparsity structure as \( A \). Simply comparing the coefficients in \( LU \) and \( A \) reveals

\[
w_j = \gamma_j, \quad j = 1, 2, \ldots, M - 1, \\
v_1 = \beta_1, \\
l_j = \alpha_j/v_{j-1}, \quad j = 2, 3, \ldots, M, \\
v_j = \beta_j - l_j \gamma_{j-1}, \quad j = 2, 3, \ldots, M.
\]

(9.37)

In the simplest case where \( a(x) \equiv 1, c(x) \equiv 0, \) and \( e(x) \equiv 0, \) we can easily show that \( |v_j| > 1, \) so that the algorithm (9.37), without partial pivoting, does not break down.

The matrix \( A \) in the latter case is symmetric. One can show that its eigenvalues \( \lambda_j \) are given by

\[
\lambda_j = -\frac{4}{\Delta x^2} \sin^2 \left( \frac{1}{2} \pi x_j \right), \quad j = 1, 2, \ldots, M,
\]

(9.38)

which should be approximations of the eigenvalues \( -(j \pi)^2 \) (\( j = 1, 2, \ldots \)) of the exact operator, i.e., the differential operator \( \frac{\partial^2}{\partial x^2} \) combined with (homogeneous) Dirichlet boundary conditions; cf. (8.11). As an example, we have plotted in Figure 9.5 both the eigenvalues of \( A \) for \( M = 20 \) and the exact eigenvalues. In particular, we see that the smallest eigenvalue in absolute value, \( \lambda_1, \) equals

\[
\lambda_1 = -\pi^2 + O(\Delta x^2).
\]

(9.39)

Hence \( \lambda_1 \) is a sharp approximation of the exact eigenvalue \( -\pi^2. \) However, the largest eigenvalue, in absolute value, \( \lambda_M, \) is approximately equal to \(-4/\Delta x^2, \) which is not a good approximation of the \( M \)th eigenvalue \( -(M\pi)^2 \) of the exact operator.

### 9.4.2 Basic Iterative Methods

For two-dimensional problems the situation is similar, but more complicated. Below we shall only discuss the simple case of a Poisson equation on the unit square with homogeneous
9.4. Solution of Linear Systems

Dirichlet boundary conditions, i.e.,

\[ \nabla^2 u = f, \quad (x, y) \in \Omega := (0, 1) \times (0, 1), \quad (9.40a) \]
\[ u = 0, \quad (x, y) \in \partial \Omega, \quad (9.40b) \]
discretised with central differences on the uniform grid

\[ (x_j, y_k) := (jh, kh) \quad j, k = 0, 1, \ldots, M + 1, \quad h := 1/(M + 1). \quad (9.41) \]

More general PDEs and discretisations can be dealt with in a straightforward way. Note that there is no natural ordering of the (internal) grid points. A common way to label these is in lexicographical order, i.e., we number the grid points along horizontal grid lines, as indicated in Figure 9.6, starting in the lower left corner. If we use the notation \( u_{j,k} \) for the numerical approximation of \( u(x_j, y_k) \), the central difference scheme for (9.40a) reads

\[ \frac{1}{h^2} \left( u_{j+1,k} + u_{j-1,k} + u_{j,k+1} + u_{j,k-1} - 4u_{j,k} \right) = f(x_j, y_k) \quad j, k = 1, 2, \ldots, M; \quad (9.42a) \]

cf. (9.2). These equations have to be completed with the numerical boundary conditions

\[ u_{j,k} = 0 \quad \text{if} \quad j, k = 0, M + 1. \quad (9.42b) \]

Introducing the vector of unknowns \( u \)

\[ u := \left( u_{1,1}, u_{2,1}, \ldots, u_{M,1} | u_{1,2}, u_{2,2}, \ldots, u_{M,2} | \ldots | u_{1,M}, u_{2,M}, \ldots, u_{M,M} \right)^T, \quad (9.43) \]

we can write (9.42a) and (9.42b) as the linear system

\[ Au = f, \quad (9.44) \]
where the matrix $A$ can be written as

$$ A = \frac{1}{h^2} \begin{pmatrix} B & I & 0 & \cdots & 0 \\ I & B & I & \cdots & 0 \\ 0 & I & B & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & B \end{pmatrix}, \quad B := \begin{pmatrix} -4 & 1 & 0 & \cdots & 0 \\ 1 & -4 & 1 & \cdots & 0 \\ 0 & 1 & -4 & 1 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \cdots & -4 \end{pmatrix}. \quad (9.45) $$

The partitioning of $u$ and $A$ is based on the numbering along horizontal grid lines.

The matrix $A$ is called block tridiagonal. Since all elements below the $M$th lower codiagonal as well as above the $M$th upper codiagonal are zero, it is also referred to as a band matrix, with bandwidth $2M + 1$. Note that we only need to store five elements per row (or rather five linear arrays). LU-decomposition will give a fill-in within this band, however, as one can easily check. This is typical for systems arising from elliptic (higher-dimensional) PDEs. A host of alternatives, in particular iterative methods, have been designed. They are all meant to reduce memory usage and to be at least as fast as LU-decomposition. Most of them are faster because one does not need full accuracy of the solution, which is perturbed by discretisation errors anyway. This important branch of numerical linear algebra is, however, too vast to even be superficially covered by this book; one may consult, e.g., [52, 135]. Nevertheless, we shall briefly discuss two classical iterative methods, i.e., Gauss–Jacobi and Gauss–Seidel, which have a merit of their own and lend themselves to a simple description.

The basic idea is to obtain an iteration of the form

$$ u^0 \text{ given,} $$

$$ u^{l+1} = M u^l + c^l \quad (l = 0, 1, 2, \ldots). \quad (9.46) $$

The matrix $M$ is called the iteration matrix. Gauss-type methods try to update the vector elementwise in sequential order (i.e., components with increasing index). To see how this

**Figure 9.6.** Lexicographic ordering of grid points for $M = 4$. 
9.4. Solution of Linear Systems

is done let us split the matrix \( A = (a_{i,j}) \) into three parts as follows:

\[
A := L + D + U,
\]

(9.47)

where \( L \) is the lower triangular part \(((a_{i,j})_{i>j})\), \( D \) is the diagonal diag\((a_{i,i})\), and \( U \) is the upper triangular part \(((a_{i,j})_{i<j})\) of \( A \). The first method we discuss is Gauss–Jacobi. Since \( D \) is nonsingular, we can define

\[
\frac{4}{h^2} u_{j,k}^{l+1} = \frac{1}{h^2} (u_{j+1,k}^l + u_{j-1,k}^l + u_{j,k+1}^l + u_{j,k-1}^l) - f(x_j, y_k).
\]

(9.48)

This corresponds to

\[
-Du_{j,k}^{l+1} = (L + U)u_{j,k}^l - f.
\]

(9.49)

We see that we can identify the iteration matrix \( M = M_J \), say, and the vector \( c \) above with

\[
M_J := -D^{-1}(L + U), \quad c^l = D^{-1}f.
\]

(9.50)

We will analyse this method now. To start with we note that the eigenvalues \( \lambda_{j,k} \) of \( A \) are given by (see the appendix, Section F)

\[
\lambda_{j,k} = -4 \sin^2 \left( \frac{1}{2} \pi x_j \right) - \sin^2 \left( \frac{1}{2} \pi y_k \right).
\]

(9.51)

Hence \( A \) is symmetric and negative definite. We have convergence if \( \rho(M_J) < 1 \); see Appendix I. Since

\[
M_J = \frac{h^2}{4} A + I,
\]

(9.52)

we find that the eigenvalues of \( M_J \) \( \mu_{j,k} \) say, are given by

\[
\mu_{j,k} = 1 - \sin^2 \left( \frac{1}{2} \pi x_j \right) - \sin^2 \left( \frac{1}{2} \pi y_k \right).
\]

(9.53)

Apparently, \( \rho(M_J) \approx 1 - (\pi h)^2/2 \). In order to assess this, we remark that we should be content if, after \( m \) iterations, the error has shrunk to \( \mathcal{O}(h^2) \), being the global discretisation error of this problem. Since \( (1 - \frac{1}{2} (\pi h)^2)^m \approx e^{-\frac{1}{2}m(\pi h)^2} \), we thus find that

\[
m \approx \frac{4}{\pi^2} M^2 \ln M.
\]

An improvement is to employ all updated values per row, instead of just the diagonal element. This gives the Gauss–Seidel method, which reads

\[
-\frac{1}{h^2} (u_{j-1,k}^{l+1} + u_{j+1,k}^{l+1}) + \frac{4}{h^2} u_{j,k}^{l+1} = \frac{1}{h^2} (u_{j+1,k}^l + u_{j,k+1}^l) - f(x_j, y_k).
\]

(9.54)

or, in matrix-vector formulation,

\[
-(L + D)u_{j,k}^{l+1} = Uu_{j,k}^l - f.
\]

(9.55)
Note that the system (9.55) can be “solved” directly (back substitution). The iteration matrix, $M_S$ say, and the vector $c^l$ are now given by

$$ M_S := -(L+D)^{-1}U, \quad c^l := (L+D)^{-1}f. \quad (9.56) $$

In order to analyse the spectral radius of the system, we need the following technical property.

Property 9.26. For the block-tridiagonal matrix $A$ in (9.45) the following holds: $\lambda_1$ is an eigenvalue of $M_J$ in (9.50) if and only if $\lambda_S = \lambda_J^2$ is an eigenvalue of $M_S$ in (9.56).

Proof. Consider the diagonal matrix $E$

$$ E := \text{diag} \left( \alpha, \alpha^2, \ldots, \alpha^M \right) $$

and the block diagonal matrix

$$ \hat{E} := \text{diag} \left( \alpha E, \alpha^2 E, \ldots, \alpha^M E \right). $$

Then

$$ \hat{E}(L + U)\hat{E}^{-1} = \alpha L + \frac{1}{\alpha} U. $$

So $\alpha L + \frac{1}{\alpha} U$ has the same eigenvalues as $L + U$. Now we first note that $\det(I + D^{-1}L) = 1$, whence

$$ \det(\lambda I - M_S) = \det \left( (I + D^{-1}L)(\lambda I - M_J) \right) $$

$$ = \det(\lambda I + \lambda D^{-1}L + D^{-1}U) $$

$$ = \det \left( \lambda I - \frac{1}{4} h^2 (\lambda J L + U) \right). $$

If $\lambda = \lambda_S = \lambda_J^2$, we find

$$ \det(\lambda_S I - M_S) = \lambda_J^{M^2} \det \left( \lambda_J I - \frac{1}{4} h^2 \left( \lambda_J L + \frac{1}{\lambda_J} U \right) \right) $$

$$ = \lambda_J^{M^2} \det \left( \lambda_J I - \frac{1}{4} h^2 (L + U) \right) $$

$$ = \lambda_J^{M^2} \det (\lambda_J I - M_J) = 0. $$

For the second equality we used $\alpha = \lambda_J$ in the assertion above. The derivation can also be carried out starting with $\lambda_J$ as an eigenvalue of $M_J$. \qed

From this property we immediately see that

$$ \rho(M_S) \approx \left( 1 - \frac{1}{2} (\pi h)^2 \right)^2 \approx 1 - (\pi h)^2. $$
9.4. Solution of Linear Systems

The number of iterations \( m \) to obtain an iteration error of \( O(h^2) \) is therefore approximately equal to

\[
m \approx \frac{2}{\pi^2} M^2 \ln M,
\]

which is about a factor of two faster than Gauss–Jacobi, at least asymptotically. There are many improvements to the Gauss-type methods. For example, one may apply relaxation; i.e., the contribution of old “values” is damped by a factor, say \( \omega \). We shall not discuss this any further. Another class of methods is related to solving a variational form of the equation. An example is the conjugate gradient method. Together with preconditioning techniques this gives a very powerful method. For further details we refer to the literature [135].

**Example 9.27** In this example we compute a numerical solution of the boundary value problem (9.40) with \( f(x, y) = 2(x(x - 1) + y(y - 1)) \). The exact solution is in this case given by \( u(x, y) = x(1 - x)y(1 - y) \). We choose \( M = 20 \) and apply the Gauss–Jacobi and Gauss–Seidel methods to system (9.44). A convergence history of both methods is given in Figure 9.7; i.e., \( \|r_l\|_2/\|r_0\|_2 \) is given as a function of the iteration index \( l \), where \( r_l := Au - f \) is the residual after \( l \) iterations. From this figure we see, e.g., that approximately 800 Gauss–Jacobi iterations are necessary to reduce the 2-norm of the initial residue \( r_0^2 \) by a factor of \( 10^4 \), whereas approximately 400 Gauss–Seidel iterations suffice. This nicely confirms that Gauss–Seidel is twice as fast as Gauss–Jacobi.

![Figure 9.7](image)

**Figure 9.7.** Convergence history of the Gauss–Jacobi and Gauss–Seidel iterative methods for the Dirichlet problem (9.40) with \( M = 20 \).

9.4.3 Closure of Iterative Methods; Operator Splitting

Iterative methods change for increasing \( l \) and are therefore evolutionary in character. It is sometimes possible to write an iterative scheme as a time integration method for a set of
ODEs. This set of ODEs is then called the \textit{time closure} of the method. To start with we show this for the Gauss–Jacobi method. Recall from (9.49), with $D = -\frac{4}{h^2}I$, that
\begin{equation}
\mathbf{u}^{l+1} = \frac{h^2}{4} (L + U) \mathbf{u}^l - \frac{h^2}{4} \mathbf{f}.
\end{equation}
(9.57)
Hence we find the increment
\begin{equation}
\mathbf{u}^{l+1} - \mathbf{u}^l = \frac{h^2}{4} (L + U + D) \mathbf{u}^l - \frac{h^2}{4} \mathbf{f} = \frac{h^2}{4} (A\mathbf{u}^l - \mathbf{f}).
\end{equation}
(9.58)
Introducing the artificial time step $\tau := h^2/4$, we see that (9.58) actually amounts to
\begin{equation}
\frac{1}{\tau} (\mathbf{u}^{l+1} - \mathbf{u}^l) = A\mathbf{u}^l - \mathbf{f}.
\end{equation}
(9.59)
This is nothing but an explicit Euler discretisation with step size $\tau$ of the ODE
\begin{equation}
\frac{d\mathbf{u}}{dt} = A\mathbf{u} - \mathbf{f}.
\end{equation}
(9.60)
Hence (9.60) may be called the time closure of (9.49), i.e., the continuous version of the iterative method. In fact, the \textit{closure} of the method as such is the parabolic PDE
\begin{equation}
\frac{\partial \mathbf{u}}{\partial t} = \nabla^2 \mathbf{u} - \mathbf{f}.
\end{equation}

For the Gauss–Seidel method (9.55) we can do something similar; note that the method is implicit since we employ components of the “solution” at the new iteration level; see (9.54). Let us recall the $\vartheta$ method (see (5.72)), which is a weighted average of the forward and the backward Euler schemes. We now generalize this to a form where $\vartheta$ might be different for each component. So we seek to identify Gauss–Seidel with a scheme of the form
\begin{equation}
\frac{1}{\tau} (\mathbf{u}^{l+1} - \mathbf{u}^l) = (\mathbf{I} - B)A\mathbf{u}^l + B\mathbf{u}^{l+1} - \mathbf{f}
\end{equation}
(9.61) for some matrix $B$. Collecting terms we actually find
\begin{equation}
(\mathbf{I} - \tau BA)\mathbf{u}^{l+1} = (\mathbf{I} + \tau (\mathbf{I} - B)A)\mathbf{u}^l - \tau \mathbf{f}.
\end{equation}
(9.62)
We now need to identify $(\mathbf{I} - \tau BA)^{-1}(\mathbf{I} + \tau (\mathbf{I} - B)A)$ with $-(L + D)^{-1}U$. It makes sense to scale the factors in the latter product by $D = -\frac{4}{h^2}I$. Hence we try
\begin{align}
\mathbf{I} - \tau BA &= -\frac{1}{4}h^2L + \mathbf{I},
\end{align}
(9.63a)
\begin{align}
\mathbf{I} + \tau (\mathbf{I} - B)A &= \frac{1}{4}h^2U.
\end{align}
(9.63b)
From this we obtain the expression
\begin{equation}
B = LA^{-1} = \mathbf{I} - (D + U)A^{-1}.
\end{equation}
(9.64)
9.5. Green’s Functions

We can then rewrite (9.62) as an (implicit) time step scheme for solving (9.60) as

\[
\frac{1}{\tau} (u^{i+1} - u^i) = (D + U)u^i + Lu^{i+1} - f, \tag{9.65}
\]

from which we find the time closure of the Gauss–Seidel method

\[
\frac{du}{dt} = (D + U)u + Lu - f = Au - f.
\]

Apparently, we have split the operator into two terms. This type of scheme, of which (9.65) is just an example, is also referred to as operator splitting.

It is possible to split the operator in many other ways. We shall consider here one of these, i.e., alternating direction implicit (ADI). The idea is to split the matrix \( A \) into a matrix derived from discretising \( \frac{\partial^2}{\partial x^2} u \) and \( \frac{\partial^2}{\partial y^2} u \). In an obvious notation we have

\[
A = A_x + A_y. \tag{9.66}
\]

The scheme then reads

\[
u^{i+1} = u^i + \tau A_x u^i + \tau A_y u^{i+1} - \tau f, \tag{9.67a}
\]

\[
u^{i+2} = u^{i+1} + \tau A_x u^{i+2} + \tau A_y u^{i+1} - \tau f. \tag{9.67b}
\]

The “implicit” direction is used for the discretisation in the \( y \) step first (and only!) and then for the discretisation in the \( x \) step next. We need to perform two steps per ADI iteration. We clearly have two discretisations of (9.60). A more detailed convergence analysis is therefore deferred to Section 11.8.1. The two-step iteration is

\[
(I - \tau A_y)u^{i+1} = (I + \tau A_x)u^i - \tau f, \tag{9.68a}
\]

\[
(I - \tau A_x)u^{i+2} = (I + \tau A_y)u^{i+1} - \tau f. \tag{9.68b}
\]

In contrast to Gauss–Jacobi and Gauss–Seidel, scheme (9.68) contains a matrix on the left that does not allow for a direct solve. However, both \( I - \tau A_y \) and \( I - \tau A_x \) are tridiagonal, so that each system allows for a simple (and cheap) LU-decomposition.

9.5 Green’s Functions

The linear systems in (9.33) and (9.44) are analogues of the continuous (partial) differential equation and the boundary conditions combined. It is therefore no surprise that one can find a discrete analogue to the Green’s function. Indeed, consider the system of equations

\[
A \mathbf{v}^j = \mathbf{e}^j, \quad \mathbf{e}^j := (0, \ldots, 0, \frac{1}{\Delta}, 0, \ldots, 0)^T. \tag{9.69}
\]

The vector \( \mathbf{e}^j \) can be seen as a discrete “Dirac function.” By construction the corresponding grid function \( v^j \) satisfies the difference equation (9.2) (or (9.31) in the one-dimensional
case) with homogeneous source term, except for one point, where this source term equals one; furthermore, it satisfies homogeneous boundary conditions. In operator form we have

\[ L_\Lambda[v^j_\Lambda](x_k) = \delta_{j,k}, \quad x_k \in \Omega_\Lambda, \] 
\[ v^j_\Lambda(x_k) = 0, \quad x_k \in \partial\Omega_\Lambda, \]  

(9.70a)  
(9.70b)

where \( x_k \) is the \( k \)th grid point in \( \Omega_\Lambda \), assuming some ordering of the grid points. Therefore \( v^j_\Lambda \) and its vector representation \( v^j \) are called a discrete Green’s function.

Apparently, the Green’s functions \( v^j \) are the columns of the inverse of \( A \); i.e.,

\[ A^{-1} = \begin{pmatrix} v^1 & v^2 & \cdots & v^N \end{pmatrix}, \]  

(9.71)

with \( N \) the number of grid points. It is interesting to see that these discrete Green’s functions can be used to construct a particular solution of the following inhomogeneous problem with homogeneous boundary conditions:

\[ L_\Lambda[v](x_k) = f(x_k), \quad x_k \in \Omega_\Lambda, \] 
\[ v(x_k) = 0, \quad x_k \in \partial\Omega_\Lambda. \]  

(9.72a)  
(9.72b)

Comparing (9.70) and (9.72), we conclude that

\[ v(x_k) = \sum_{j=1}^{N} f(x_j)v^j_\Lambda(x_k), \quad x_k \in \Omega_\Lambda, \]  

(9.73)

or, in matrix-vector notation,

\[ v = A^{-1}f. \]  

(9.74)

There is a close analogy with the continuous case, dealt with in Section 8.5. Note that Green’s functions and fundamental solutions differ only by a (“discrete”) harmonic function. The complete solution \( u_\Lambda \) of the inhomogeneous problem

\[ L_\Lambda[u_\Lambda](x_k) = f(x_k), \quad x_k \in \Omega_\Lambda, \] 
\[ u_\Lambda(x_k) = g(x_k), \quad x_k \in \partial\Omega_\Lambda, \]  

(9.75a)  
(9.75b)

is given by

\[ u_\Lambda(x_k) = v_\Lambda(x_k) + w_\Lambda(x_k), \quad x_k \in \Omega_\Lambda, \]  

(9.76)

with \( v_\Lambda \) the solution of (9.72) and \( w_\Lambda \) the solution of the following homogeneous problem with inhomogeneous boundary conditions:

\[ L_\Lambda[w_\Lambda](x_k) = 0, \quad x_k \in \Omega_\Lambda, \] 
\[ w_\Lambda(x_k) = g(x_k), \quad x_k \in \partial\Omega_\Lambda. \]  

(9.77a)  
(9.77b)

In matrix-vector formulation we have

\[ Aw = g. \]  

(9.78)
where the vector $\mathbf{g}$ has zero components except for points where the scheme includes boundary points; at such a point the component is equal to minus the coefficient in the difference equation times the value at that boundary (or, when there are more, a sum of such contributions). As before, (9.76) is an instance of the superposition principle; cf. (8.60). We thus have

$$u = v + w, \quad A u = f + g.$$  \hfill (9.79)

**Example 9.28** Consider the boundary value problem

$$L[u] := \frac{d^2 u}{dx^2} = f(x), \quad x \in (0, 1),$$

$$x(0) = A, \quad x(1) = B.$$  \hfill (9.76)

The fundamental solution of this differential equation is given by $u(x) = A + Bx$. Since this is linear it will be integrated exactly by the central difference scheme and hence it is also the solution of the discrete problem

$$L_{\Delta}[u_{\Delta}](x_i) = f(x_i), \quad i = 1, 2, \ldots, M$$

$$u_0 = A, \quad u_{M+1} = B,$$

where $L_{\Delta}$ is the central difference operator defined on the grid

$$x_j := j \Delta x, \quad j = 0, 1, \ldots, M + 1, \quad \Delta x = 1/(M + 1).$$

The resulting coefficient matrix of the linear system $A u = f$ reads

$$A := \frac{1}{\Delta x^2} \left( \begin{array}{ccccc} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{array} \right).$$

The solution of $A \mathbf{v} = \mathbf{e}^i$ can now be constructed as follows: For $i = 0, 1, \ldots, j$ it must have the form

$$v'_j = B'_1 x_j$$

for some $B'_1$. For $i = j, j + 1, \ldots, M + 1$ it must have the form

$$v'_i = B'_2 (1 - x_j)$$

for some $B'_2$. Note that this function $v'$ satisfies the homogeneous boundary conditions $v'(x_0) = v'(x_{M+1}) = 0$. At $i = j$ we have

$$v'_{j-1} - 2v'_j + v'_{j+1} = \Delta x^2, \quad B'_1 x_j = B'_2 (1 - x_j).$$

From these relations we can determine $B'_1$ and $B'_2$. We obtain

$$B'_1 \left( x_{j-1} - 2x_j + \frac{x_j}{1 - x_j} (1 - x_{j+1}) \right) = \Delta x^2.$$  \hfill (9.77)

Hence

$$B'_1 = -\Delta x (1 - x_j), \quad B'_2 = -\Delta x x_j.$$
So we find for the discrete Green’s function
\[ v_j^i = \begin{cases} 
\Delta x (x_j - 1) x_i & \text{for } i = 0, 1, \ldots, j, \\
\Delta x x_j (x_i - 1) & \text{for } i = j + 1, j + 2, \ldots, M + 1.
\end{cases} \]

We immediately find, for the inverse of the matrix A,
\[ A^{-1} = -\frac{1}{(M+1)^3} \begin{pmatrix} 
M & M-1 & M-2 & \ldots & 1 \\
M-1 & (M-1)2 & (M-2)2 & \ldots & 2 \\
M-2 & (M-2)2 & (M-2)3 & \ldots & 3 \\
M-3 & (M-3)2 & (M-3)3 & \ldots & 4 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 2 & 3 & \ldots & M
\end{pmatrix}. \]

### 9.6 Nonlinear Problems

For nonlinear problems we have to adapt our solution procedure. We shall only consider PDEs with a linear differential operator. So we consider
\[ \mathcal{L}[u] = f(u), \quad x \in \Omega, \quad (9.80a) \]
\[ u = g(x), \quad x \in \partial \Omega, \quad (9.80b) \]
where \( f(u) \) is nonlinear in \( u \). After discretisation, we obtain a nonlinear system of equations for the unknowns at the grid points, say \( u \in \mathbb{R}^N \):
\[ N(u) = 0. \quad (9.81) \]

Since nonlinear problems occur quite frequently, we shall discuss three often used approaches: Newton iteration in Section 9.6.1, Gauss–Jacobi iteration in Section 9.6.2, and transient methods in Section 9.6.3.

#### 9.6.1 Newton Iteration

The rationale behind Newton’s method is the linearisation of \( N(u) \) at an approximation of the solution, \( \tilde{u} \) say:
\[ N(v) \approx N(\tilde{u}) + J(\tilde{u})(v - \tilde{u}). \quad (9.82) \]

Here \( J(u) \) is the Jacobi matrix of \( N(u) \), defined by
\[ J(u) := \frac{\partial N(u)}{\partial u} = \left( \frac{\partial N_i(u)}{\partial u_j} \right). \quad (9.83) \]

Anticipating \( v \) to be a better approximation of the exact solution \( u \) than \( \tilde{u} \), we take \( N(v) = 0 \), and so we obtain, assuming \( J(\tilde{u}) \) is nonsingular,
\[ v - \tilde{u} = -J^{-1}(\tilde{u}) N(\tilde{u}). \quad (9.84) \]
Let $u^0$ be some known approximate value. Then we have Newton’s method:

given $u^0$,

$$\text{solve } J(u^0)s' = -N(u'),$$

$$u^{l+1} = u^l + s'. \quad \text{(9.85)}$$

The quantity $s'$ is called the Newton update. Note that $s'$ satisfies a linear equation, which can therefore be solved by a method described in Section 9.4. The convergence of the Newton method (9.85) is assessed in the following famous theorem, which we give without proof (see [68]).

**Theorem 9.29 (Newton–Kantorovich).** Let $N(u)$ be continuously differentiable in a neighbourhood $B(u^0; \rho)$ of $u^0$ (i.e., a ball with centre $u^0$ and radius $\rho$) and let the constants $C_1, C_2,$ and $C_3$ be such that

$$\|J^{-1}(u^0)N(u^0)\| \leq C_1,$$

$$\|J^{-1}(u^0)\| \leq C_2,$$

$$\|J(u_1) - J(u_2)\| \leq C_3\|u_1 - u_2\|, \quad u_1, u_2 \in B(u^0; \rho).$$

Define $\sigma := C_1C_2C_3$. If $\sigma \leq \frac{1}{2}$ and $\rho \geq \rho_\ast := (1 - \sqrt{1 - 2\sigma})/(C_2C_3)$, then the following hold:

(i) There is a unique root $u^\ast \in B(u^0; \rho_\ast)$.

(ii) The sequence of Newton iterates $\{u^l\}$ converges to $u^\ast$.

If $\sigma < \frac{1}{2}$, then, moreover,

(iii) $u^\ast$ is the unique root in $B(u^0; \min(\rho, \rho_\ast))$, where $\rho_\ast := (1 + \sqrt{1 - 2\sigma})/(C_2C_3)$.

(iv) $\|u^l - u^\ast\| \leq (2\sigma)^{2m} \frac{1}{C_2C_3}.$

Unfortunately, this theorem does not reveal that the method is actually asymptotically quadratically convergent. The main problem, however, is to find a suitable initial guess $u^0$. The Newton–Kantorovich theorem at least shows that this will be less problematic the smoother $f$ is.

The Jacobi matrix $J(u)$ has a special form for our problem (9.80). Let $A$ be the matrix that would have been found from discretising $\mathcal{L}[u] = 0$. Then we have

$$J(u) = A - \text{diag} \left( f'(u_i) \right). \quad \text{(9.86)}$$

If we assume that $f'(u) > 0$, then the important requirement (9.4a) in Property 9.1 is certainly satisfied (even in a strict inequality sense).

**Example 9.30** Consider the one-dimensional problem

$$\mathcal{L}[u] := \frac{d^2u}{dx^2} = -\alpha e^u, \quad x \in (0, 1),$$

$$\frac{du}{dx}(0) = 0, \quad u(1) = 0,$$
where \( \alpha > 0 \) is some parameter. This boundary value problem only has a solution for \( 0 < \alpha < \alpha^* := 0.878457764 \) [3]. If we use central differences on the grid

\[
x_j := j \Delta x, \quad j = 0, 1, \ldots, M + 1, \quad \Delta x = 1/(M + 1),
\]

we obtain the following nonlinear system for \( \mathbf{u} := (u_1, u_2, \ldots, u_M)^T \)

\[
\mathbf{N}(\mathbf{u}) := \mathbf{A} \mathbf{u} + \alpha \text{diag}(e^{u_1}, e^{u_2}, \ldots, e^{u_M}) = 0,
\]

with

\[
\mathbf{A} = \frac{1}{\Delta x^2} \begin{pmatrix}
-2 & 2 & 0 & \cdots & 0 \\
1 & -2 & 1 & & \\
& \ddots & \ddots & \ddots & \\
0 & \cdots & 1 & -2 & 1 \\
0 & \cdots & 0 & 1 & -2
\end{pmatrix}.
\]

The Jacobi matrix for this problem is obviously

\[
\mathbf{J}(\mathbf{u}) = \mathbf{A} + \alpha \text{diag}(e^{u_1}, e^{u_2}, \ldots, e^{u_M}).
\]

We have solved this boundary value problem for several values of \( \alpha \) between 0.5 and \( \alpha^* \) on a grid with \( \Delta x = 0.01 \). The nonlinear system (\( \ast \)) is solved with Newton’s method, and it is continued until \( \|\mathbf{N}(\mathbf{u}_l)\|_2 < \text{tol} \) and \( \|\mathbf{s}_l\|_2 < \text{tol} \) for some tolerance tol. Convergence is very fast; typically 5 to 10 iterations are needed for \( \text{tol} = 10^{-4} \). This value is of the same order of magnitude as the discretisation error; a smaller value is not needed since the numerical solution contains a discretisation error anyway.

In practice the Newton process does not converge without additional effort. To understand how to improve the method it is useful to define the object function

\[
q(\mathbf{u}) := \|\mathbf{N}(\mathbf{u})\|_2^2 = \mathbf{N}(\mathbf{u})^T \mathbf{N}(\mathbf{u}).
\]

The Newton update \( \mathbf{s} \) is a descent direction; i.e., it tends to smaller values of \( q(\mathbf{u}) \). Since \( \nabla q(\mathbf{u}) = 2\mathbf{J}(\mathbf{u})^T \mathbf{N}(\mathbf{u}) \), we readily see that

\[
\mathbf{s}^T \nabla q(\mathbf{u}) = 2 \mathbf{s}^T \mathbf{J}(\mathbf{u})^T \mathbf{N} = -2q(\mathbf{u}) < 0.
\]

If a certain Newton step does not give a smaller value of the object function, only a fraction may do a better job. Indeed, from a Taylor expansion we see

\[
q(\mathbf{u} + \lambda \mathbf{s}) = q(\mathbf{u}) + \lambda \mathbf{s}^T \nabla q(\mathbf{u}) + \mathcal{O}(\lambda^2 \|\mathbf{s}\|_2^2).
\]

If \( \lambda \|\mathbf{s}\|_2 \) is small enough, we see that (\( \lambda > 0! \))

\[
q(\mathbf{u} + \lambda \mathbf{s}) < q(\mathbf{u}).
\]

Such a \( \lambda \) is called a damping factor and the variant is called damped Newton: given \( \mathbf{u}^0 \), solve

\[
\mathbf{J}(\mathbf{u})^T \mathbf{s}^l = -\mathbf{N}(\mathbf{u}^l),
\]

\[
\mathbf{u}^{l+1} = \mathbf{u}^l + \lambda \mathbf{s}^l.
\]
Although one may try to find “optimal” values for $\lambda_l$ from the model above, an often used method is to halve the damping factor if “convergence” fails (and repeat this until a smaller value of the object function is reached); if such a decrease is found, one then only cautiously increases the damping parameter; for more details see, e.g., [3].

Another obvious modification is to keep the Jacobi matrix fixed during a few iterations. Only at the very end, when quadratic convergence is becoming visible, might a new Jacobi evaluation at every step be meaningful.

### 9.6.2 Gauss–Jacobi Iteration

Rather than using the complete operator in Newton’s method, we can use a Gauss–Jacobi or Gauss–Seidel splitting first. So let us rewrite the discrete nonlinear equation (9.81) as

$$Au = f(u),$$

with $A$ the matrix resulting from discretising the homogeneous problem $L[u] = 0$. In the rest of this section we take for $L$ the two-dimensional Laplacian. If we split the matrix $A$ as in (9.47), we obtain the nonlinear variant of the Gauss–Jacobi iteration

$$Du^{l+1} - f(u^{l+1}) = -(L + U)u^l.$$  \hspace{1cm} (9.92)

An alternative to Newton’s method is employing some form of operator splitting; see Section 9.4.3. Since $f(u)$ is nonlinear, it is convenient to leave this out of the implicit part. So, alternatively, we may use the following form, which is linear in $u^{l+1}$:

$$Du^{l+1} = -(L + U)u^l + f'(u^l).$$  \hspace{1cm} (9.93)

Likewise, a nonlinear Gauss–Seidel variant is

$$(L + D)u^{l+1} = -Uu^l + f'(u^l).$$  \hspace{1cm} (9.94)

In order to prove convergence of (9.93), we need the following lemma. Note that symmetric matrices have real eigenvalues.

**Lemma 9.31.** Let $B$ and $C$ be symmetric matrices. Denote by $\lambda_1, \ldots, \lambda_M$, $\mu_1, \ldots, \mu_M$, and $v_1, \ldots, v_M$ the eigenvalues of $B$, $B - C$, and $C$, respectively. Then

$$\max_i \lambda_i \geq \max_j \mu_j + \min_k v_k,$$

$$\min_i \lambda_i \leq \min_j \mu_j + \max_k v_k.$$  \hspace{1cm} (9.95)

**Proof.** The maximum eigenvalue of $B$ satisfies

$$\max_i \lambda_i = \max_{\|u\|_2 = 1} u^T Bu$$

$$= \max_{\|u\|_2 = 1} (u^T (B - C)u + u^T Cu)$$

$$\geq v^T (B - C)v + v^T Cv$$

$$\geq v^T (B - C)v + \min_k v_k,$$
where $v$ is the eigenvector, with $\|v\|_2 = 1$, belonging to the largest eigenvalue of $B - C$, so

$$v^T (B - C)v = \max_j \mu_j.$$  

If we consider the minimum eigenvalue of $B$, we obtain

$$\min_i \lambda_i = \min_{\|u\|_2 = 1} u^T Bu$$

$$\leq w^T (B - C)w + w^T Cw$$

$$\leq w^T (B - C)w + \max_k v_k,$$

where $w$ is the eigenvector, with $\|w\|_2 = 1$, belonging to the smallest eigenvalue of $B - C$, so

$$v^T (B - C)v = \min_j \mu_j.$$  

From this the result immediately follows.

We can now show the following property.

**Property 9.32.** Assume we use central differences to discretise boundary value problem (9.80). Let $C$ be a constant such that $f'(u) \geq C > 0$ for all $u$. Then the Gauss–Jacobi method (9.93) converges and the error $e_l := u_l - u$ satisfies

$$\|e^{l+1}\|_2 \leq \left(1 - h^2 \left(\frac{1}{2} \pi^2 + \frac{1}{4} C\right)\right) \|e^l\|_2.$$  

**Proof.** For the error vector $e^l$ in the Gauss–Jacobi iteration we obtain

$$De^{l+1} = -(L + U)e^l + f(u') - f(u) = -(L + U - \text{diag}(f'(v_l^i)))e^l,$$

where $v_l^i$ is some intermediate value between $u_i$ and $u_l$. We write this simply as

$$-De^{l+1} = (L + U - E^l)e^l, \quad E^l := \text{diag}(f'(v_l^i)).$$

Note that $E^l$ is a diagonal matrix with positive entries since $f'(u) > 0$. Hence

$$e^{l+1} = \left(-D^{-1}(L + U) + D^{-1}E^l\right)e^l.$$  

Taking the 2-norm, we obtain the upper bound

$$\|e^{l+1}\|_2 \leq \|D^{-1}(L + U) - D^{-1}E^l\|_2 \|e^l\|_2.$$  

We can now invoke Lemma 9.31 to estimate this matrix norm. Let us identify $D^{-1}(L + U)$ with $B$ and $D^{-1}E^l$ with $C$. Moreover, we remark that the eigenvalues of $D^{-1}(L + U)$ can be found from (9.53); in particular, we obtain in good approximation $\lambda_1 := \min_j \lambda_j = \lambda_1 := \max_j \lambda_j = \lambda_1 := 1 - \frac{(\pi h)^2}{2}$. We thus obtain in good approximation

$$-1 + \frac{(\pi h)^2}{2} - \max_k v_k \leq \mu_j \leq 1 - \frac{(\pi h)^2}{2} - \min_k v_k.$$  

9.6. Nonlinear Problems

Apparently the matrix $D^{-1}(L + U) - D^{-1}E$ is symmetric and positive definite. Since $v_k \geq \frac{\nu}{4} C$, we see that

$$\|D^{-1}(L + U) - D^{-1}E\|_2 \leq 1 - h^2 \left( \frac{1}{2} \pi^2 + \frac{1}{4} C \right).$$

Note that the convergence is slightly enhanced by a larger value of $C$ in the property above. Actually, the result is not essentially different from that of a linear problem where the PDE has a term proportional to $u$. For Gauss–Seidel a similar result holds.

9.6.3 Transient Methods

Quite another approach is to invoke the time closure of (9.91) directly, i.e.,

$$\frac{du}{dt} = Au - f(u). \quad (9.96)$$

The time closure automatically gives the proper sign, i.e., not $-Au + f(u)$, which could have been chosen as well from a purely theoretical point of view. The basic idea is to compute the steady state solution of (9.96) by applying a suitable time integration technique; cf. Section 5.4. If we have an inaccurate initial guess $u^0$, we expect to need many time steps to arrive at the steady state where $\frac{du}{dt} = 0$. This kind of approach is often referred to as the (false) transient method. Note that (9.96) is the semidiscretisation of the parabolic equation

$$\frac{\partial u}{\partial t} = \nabla = L[u] - f(u). \quad (9.97)$$

Solution methods for this equation are deferred to Chapter 11.

Since we are only interested in the steady state solution of (9.96), the stability of a time integrator is of more importance than high accuracy. Therefore the implicit Euler method is a good candidate, resulting in the following nonlinear system for the unknown $v = u^{n+1}$:

$$F(v) := \left( A - \frac{1}{\Delta t} I \right) v - f(v) + \frac{1}{\Delta t} u^n = 0, \quad (9.98)$$

with $\Delta t$ the time step and $u^n$ the approximation of $u$ at time level $n \Delta t$. To solve (9.96) we can suitably use Newton iteration. Note that the Jacobi matrix $J_F$ for (9.98) is given by

$$J_F = A - \frac{1}{\Delta t} I - \text{diag} \left( f'(v_i) \right). \quad (9.99)$$

It contains the extra term $(1/\Delta t)I$ compared to the Jacobi matrix in (9.86), thus enhancing condition (9.4a) in Property 9.1. As a result, $\text{cond}(J_F)$ is often smaller than $\text{cond}(J)$ and Newton iteration for (9.98) usually converges well. We have to solve (9.98) repeatedly until $\|u^{n+1} - u^n\| < \Delta t \text{tol}$ for some tolerance tol.

We conclude this section by noting that Newton’s method also has a time closure. Consider the damped Newton iteration

$$u^{i+1} - u^i = -\lambda_i J^{-1}(u^i) N(u^i). \quad (9.100)$$
Then, letting $\lambda_l \to 0$, we arrive at
\[ \frac{du}{dt} = -J^{-1}(u)N(u), \] (9.101)
the *Davidenko equation*. Here, too, one can use a transient method as a practical way to determine the damping factor. Note that apparently $\lambda_l$ should not be larger than one. This is intuitively more restrictive than a stability requirement for the Euler forward method would give.

### 9.7 A Pressure Correction Method for the Stokes Equations

In Section 8.7 we encountered the Stokes equations
\[
\begin{align*}
\nabla^2 u - \nabla p &= 0, \quad (9.102a) \\
\nabla \cdot u &= 0, \quad (9.102b)
\end{align*}
\]
which describe creeping, steady flow of an incompressible fluid in a domain $\Omega \subset \mathbb{R}^d \ (d = 2, 3)$. In this section we restrict ourselves to two-dimensional problems; the extension to three dimensions is straightforward. In system (9.102) $u = (u(x, y), v(x, y))^T$ and $p = p(x, y)$ are the velocity and the pressure field, respectively. We assume that we have Dirichlet boundary conditions for $u$; i.e.,
\[ u = b, \quad x \in \partial\Omega. \] (9.103)

These boundary conditions have to satisfy the constraint (cf. (9.102b))
\[ \int_{\partial\Omega} b \cdot n \, ds = 0, \] (9.104)
with $n$ the unit outward normal on $\partial\Omega$.

If we apply the divergence operator to the momentum equations (9.102a) and invoke the incompressibility condition (9.102b), we obtain the *pressure Poisson equation*
\[ \nabla^2 p = 0; \] (9.105)
i.e., the pressure is a harmonic function. Conversely, from (9.102a) and (9.105) we can simply deduce that $\nabla^2 d = 0$ with $d := \nabla \cdot u$, and this equation has only the trivial solution $d(x) \equiv 0$ if $d = \nabla \cdot u = 0$ on the boundary $\partial\Omega$. This is then the proper boundary condition for (9.105). It can be shown that $\nabla \cdot u = 0$ on the boundary implies the Neumann condition
\[ \frac{\partial p}{\partial n} = n \cdot \nabla^2 u, \quad x \in \partial\Omega, \] (9.106)
which is just the normal component of (9.102a) applied at the boundary; cf. [55]. So we have two alternative formulations for the Stokes problem: either (9.102) with boundary condition (9.103) or (9.102a) and (9.105) together with the boundary conditions (9.103)
and (9.106). In both formulations \( p \) is determined up to an additive constant. In the rest of this section we will derive a numerical scheme for (9.102).

In order to compute a numerical solution for (9.102), we cover the domain \( \Omega \) with a uniform grid with grid sizes \( \Delta x \) and \( \Delta y \) in the \( x \) and \( y \) directions, respectively, and apply a finite difference scheme. First, we consider central differences. Let \( N \) denote the number of (unknown) internal grid points. Introducing the vectors

\[
\mathbf{u} := (u_1, v_1, u_2, v_2, ..., u_N, v_N)^T, \\
\mathbf{p} := (p_1, p_2, ..., p_N)^T
\]

(9.107)

containing the velocity components and pressure at the grid points, ordered lexicographically, we obtain an algebraic system of the form (cf. (9.44))

\[
\mathbf{A} \mathbf{u} - \mathbf{B} \mathbf{p} = \mathbf{f}_1, \\
\mathbf{B}^T \mathbf{u} = \mathbf{f}_2.
\]

(9.108a, 9.108b)

In (9.108), \( \mathbf{A}, \mathbf{B}, \) and \( \mathbf{B}^T \) are the central difference approximations of the Laplace, gradient, and divergence operators, respectively. Nonzero boundary conditions are taken into account through the vectors \( \mathbf{f}_1 \) and \( \mathbf{f}_2 \). The direct numerical solution of (9.108) is tedious because the pressure is not included in (9.108b), which precludes the straightforward application of standard iterative methods.

For the solution of (9.108) we consider the pressure correction method. The basic idea is to decouple the computation of \( \mathbf{u} \) and \( \mathbf{p} \) in a predictor-corrector fashion in combination with a (false) transient method. The transient method is obvious here, since (9.102a) is the time limit solution (i.e., for \( t \to \infty \)) of

\[
\frac{\partial \mathbf{u}}{\partial t} = \nabla^2 \mathbf{u} - \nabla p.
\]

(9.109)

Applying the method of lines to (9.109), we obtain the semidiscrete system

\[
\frac{d\mathbf{u}}{dt} = \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{p} - \mathbf{f}_1,
\]

(9.110)

which should replace (9.108a). Since we are interested only in the time limit solution of (9.110) and (9.108b), we look for an unconditionally stable time integration method. As noted in Section 9.6.3, we are not interested in the accuracy of this integration as such. Therefore the implicit Euler method is an obvious choice. We find

\[
\mathbf{u}^{n+1} = \mathbf{u}^n + \Delta t(\mathbf{A}\mathbf{u}^{n+1} - \mathbf{B}\mathbf{p}^{n+1} - \mathbf{f}_1), \\
\mathbf{B}^T \mathbf{u}^{n+1} = \mathbf{f}_2.
\]

(9.111a, 9.111b)

A predictor-corrector method for (9.111) then runs as follows. First, compute a predictor \( \mathbf{u}^* \) for the velocity by simply replacing \( \mathbf{p}^{n+1} \) in (9.111a) by \( \mathbf{p}^n \); i.e.,

\[
\mathbf{u}^* = \mathbf{u}^n + \Delta t(\mathbf{A}\mathbf{u}^* - \mathbf{B}\mathbf{p}^n - \mathbf{f}_1).
\]

(9.112)
Chapter 9. Numerical Methods for Elliptic Equations

Note that this predictor does not satisfy (9.108b), a condition we will impose on the corrector. Indeed, the corrector is evaluated via (9.111a) with \( u^{n+1} \) on the right-hand side replaced by \( u^* \). Thus the corrected value, also denoted by \( u^{n+1} \), is given by

\[
   u^{n+1} = u^n + \Delta t (Au^* - Bp^{n+1} - f_i) .
\]

(9.113)

Subtracting (9.112) from (9.113), we obtain

\[
   u^{n+1} - u^* = -\Delta t Bq^n , \quad q^n := p^{n+1} - p^n .
\]

(9.114)

Since we require the corrector \( u^{n+1} \) to satisfy the discrete incompressibility condition (9.111b), we obtain, upon substituting (9.114) in (9.111b),

\[
   L q^n = \frac{1}{\Delta t} (B^T u^* - f_2) , \quad L := B^T B.
\]

(9.115)

Equation (9.115) is a “discrete” Poisson equation, from which we can compute the pressure update \( q^n \). The velocity update simply follows from (9.114). This time integration procedure should be repeated until \( \| u^{n+1} - u^n \| < \text{tol} \Delta t \) for some tolerance \( \text{tol} \).

Collecting the various steps, we arrive at the pressure correction algorithm

\[
   (I - \Delta t A) u^* = u^n - \Delta t (Bp^n + f_i) \quad \text{(predictor)},
\]

(9.116a)

\[
   L q^n = \frac{1}{\Delta t} (B^T u^* - f_2) , \quad p^{n+1} = p^n + q^n \quad \text{(pressure update)},
\]

(9.116b)

\[
   u^{n+1} = u^* - \Delta t Bq^n \quad \text{(velocity update)}.
\]

(9.116c)

There are a few difficulties associated with the discrete Poisson equation (9.115). First, the matrix \( L \) is singular, since \( Le = 0 \), with \( e := (1, 1, \ldots, 1)^T \). This is in agreement with the fact that the pressure is determined up to an additive constant. In order to determine \( p \), we have to specify the pressure at one single point of the domain, e.g., somewhere near the boundary. The drawback is that the incompressibility condition (9.111b) is not satisfied at this point. Alternatively, we might solve the modified equation

\[
   (L - \alpha I) q^n = \frac{1}{\Delta t} (B^T u^* - f_2) - \alpha q^{n-1} ,
\]

(9.117)

where \( 0 < \alpha \ll 1 \). In this way we achieve that the coefficient matrix \( L - \alpha I \) has an acceptable condition number. Condition (9.111b) then only holds in the limit \( n \to \infty \).

Second, combining the central difference approximations of \( \nabla p \) and \( \nabla \cdot u \) at an interior grid point, we see that the matrix \( L \) is the standard five-point discretisation of the Laplace operator, except on a grid with grid sizes \( 2\Delta x \) and \( 2\Delta y \), as indicated in Figure 9.8. Clearly, there is no direct coupling between neighbouring values of \( q \), and consequently also not between neighbouring values of the pressure. This means that the pressure field is decoupled in the following way. Ignoring boundary conditions, let \( p_{j,k} \) denote the numerical approximation of the pressure at grid point \( x_c = (x_j, y_k) \). Then \( p_{j,k} + (-1)^j C, \, C \in \mathbb{R} \), is also a solution of (9.116). Apparently, the grid function \((-1)^j \) is a spurious mode of the numerical pressure. Likewise, \((-1)^k \) and \((-1)^{j+k} \) are spurious modes.
A way to avoid spurious modes is to use a staggered grid. In a staggered grid the nodes for the velocity components \( u \) and \( v \) and for the pressure \( p \) are displaced as indicated in Figure 9.9. The momentum equations (9.102a) are discretised at the \( u, v \) points and the continuity equation (9.102b) is discretised at the \( p \) points. In a staggered grid arrangement all first derivatives can be conveniently discretised by central differences on a grid with grid sizes \( \Delta x/2 \) and \( \Delta y/2 \). Indeed, we have for \( \nabla p \)

\[
\frac{\partial p}{\partial x}(x_{j+\frac{1}{2}}, y_k) \doteq \frac{1}{\Delta x} \left( p(x_{j+1}, y_k) - p(x_j, y_k) \right),
\]

(9.118a)

and for \( \nabla \cdot \mathbf{u} \)

\[
\nabla \cdot \mathbf{u}(x_j, y_k) \doteq \frac{1}{\Delta x} \left( u(x_{j+\frac{1}{2}}, y_k) - u(x_{j-\frac{1}{2}}, y_k) \right) + \frac{1}{\Delta y} \left( v(x_j, y_{k+\frac{1}{2}}) - v(x_j, y_{k-\frac{1}{2}}) \right).
\]

(9.118c)
Combining the discretisations in (9.118), we see that

\[ \mathcal{L}_\Delta[q](x_j, y_k) = \frac{1}{\Delta x^2} (q(x_{j+1}, y_k) - 2q(x_j, y_k) + q(x_{j-1}, y_k)) + \frac{1}{\Delta y^2} (q(x_j, y_{k+1}) - 2q(x_j, y_k) + q(x_j, y_{k-1})) , \]

(9.119)

where \( \mathcal{L}_\Delta \) is the difference operator corresponding to \( \mathbf{L} \). Thus \( \mathbf{L} \) is the standard five-point discretisation of the Laplace operator, so that coupling between neighbouring \( p \) values is guaranteed.

Finally, we do not need explicit boundary conditions for \( q \) since there are no nodes for \( q \) on the boundary. To illustrate this consider the pressure node \((x_1, y_k)\) adjacent to a vertical boundary; see Figure 9.9. For this node (9.111b) can be written as

\[ \frac{1}{\Delta x} u^{n+1}_{1,k} + \frac{1}{\Delta y} (v^{n+1}_{1,k+1} - v^{n+1}_{1,k-1}) = \frac{1}{\Delta x} u^1_{1,k} , \]

(9.120)

where \( u_{1/2,k} \) is the normal component of the velocity at the boundary. Using (9.114) in (9.120), we obtain the equation

\[ \frac{1}{\Delta x} (q^{n+1}_{2,k} - q^n_{1,k}) + \frac{1}{\Delta y} (q^n_{1,k+1} - 2q^n_{1,k} + q^n_{1,k-1}) = \frac{1}{\Delta t} \left( \frac{1}{\Delta x} (u^n_{1,k} - u^1_{1,k}) + \frac{1}{\Delta y} (v^n_{1,k+1} - v^1_{1,k-1}) \right) , \]

(9.121)

which only contains \( q \) values in the interior domain. Other boundaries can be treated in a similar way.

**Example 9.33** In this example we compute a numerical solution of the Stokes equations (9.102) on a rectangular domain \( \Omega := (0, a) \times (0, b) \) subject to the following boundary conditions:

- **inflow:** \( u(x) = 16U_{\max} \frac{y}{b} \left( \frac{1}{2} - \frac{y}{b} \right) \), \( v(x) = 0 \), \( x \in \partial \Omega_{in} := \left\{ x \mid x = 0, 0 \leq y \leq \frac{1}{2}b \right\} \).

- **outflow:** \( u(x) = 0 \), \( v(x) = 16V_{\max} \left( \frac{x}{a} - \frac{1}{2} \right) \left( 1 - \frac{x}{a} \right) \), \( x \in \partial \Omega_{out} := \left\{ x \mid \frac{1}{2}a \leq x \leq a, y = b \right\} \).

- **solid wall:** \( u(x) = 0 \), \( x \in \partial \Omega \setminus (\partial \Omega_{in} \cup \partial \Omega_{out}) \).

The constraint (9.104) leads to \( aV_{\max} = bU_{\max} \). More specifically, we take \( a = 2, b = 1, \) and \( U_{\max} = 1 \). Furthermore, we choose the following numerical parameters: \( \Delta x = \Delta y = 0.05, \Delta t = 0.1, \) and \( \text{tol} = 10^{-6} \). We compute the pressure (update) from the modified equation (9.117) with \( \alpha = 10^{-6} \). The steady solution we obtained after 886 time steps satisfies \[ \| \mathbf{B}^T \mathbf{u} - \mathbf{f} \|_2 \leq 6.25 \times 10^{-4} \] and is given in Figure 9.10.
9.8 Discussion

Elliptic problems can be numerically treated in a variety of ways. Although we considered finite difference methods here, this appears to be the area where finite element methods turn out to be very powerful [9]. The derivation and analysis of these methods is, however, so much different from the finite differences or finite volumes we use that it would require too much space for this book. These methods are based on the weak form of a differential equation, which, after applying Gauss’s formula, results in a bilinear form, which is discretised in turn. The domain is then divided into small triangles (or quadrilaterals), which is the basis for the large flexibility as well as the possibility of working fairly easily for complex geometries. Another approach is to use Green’s identity and employ knowledge of fundamental solutions to find an
Chapter 9. Numerical Methods for Elliptic Equations

An integral equation on the boundary; this is the basis of the boundary element method [16].

• A crucial point in the discretisation of any problem is the concern for “sampling” the solution properly. So one needs a denser grid in subdomains where the solution exhibits higher activity than outside. The problem of finding adaptive grids is very hard for finite difference methods. In principle one could use numerical derivatives as a measure and adapt the grid in an iterative way. For a Poisson problem the source term is a natural criterion for equidistribution of the error [11]. The difficulty for finite difference methods is that one has to take recourse to interpolation to find the difference scheme for points in “transition” areas. Such schemes also lead to unstructured matrices (where no simple numbering of the grid points leads to a nice form).

• As appears from the theory in this chapter, numerical linear algebra is an important issue in solving elliptic problems numerically. Again, this is an issue by itself. There exist a host of good books that treat linear systems, in particular methods for sparse systems. The majority of these methods are iterative in nature. Yet a method like multigrid may require a number of arithmetical operations of the same order as the number of unknowns, clearly the best one can get [167].

• In Chapter 16 the following practical problems are related to the material of this chapter. In Section 16.6 the nonlinear groundwater table problem is solved. Catalyzer pellets, discussed in Section 16.8, are modeled as a steady state reaction diffusion problem with spherical symmetry, leading to a boundary value problem with the radial distance \( r \) as the only variable. Finally, in Section 16.15 we model the flow in a glass oven. Here the full steady state Navier–Stokes equations are solved and various aspects of the formulation, including the (handling of the) boundary conditions, are discussed.

Exercises

9.1. One can use higher-order difference approximations for the Laplace operator. One possibility is to use five points per direction. Show the stencil of this scheme and determine the coefficients. Finally, show that this has fourth order accuracy.

9.2. Suppose we have the differential operator

\[
\mathcal{L}[u] := \nabla^2 u + \frac{\partial^2 u}{\partial x \partial y}.
\]

Use central differences on a grid with grid sizes \( \Delta x = \Delta y = h \) to obtain a nine-point scheme on a square stencil whose coefficients are given by

\[
\frac{1}{h^2} \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \frac{1}{4h^2} \begin{bmatrix} -1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & -1 \end{bmatrix}.
\]
9.3. Consider the Poisson equation on the unit square, i.e.,

\[ \mathcal{L}[u] := \nabla^2 u = f, \quad x \in \Omega := (0, 1) \times (0, 1). \]

Let the boundary condition be given by \( u(x) = 0 \) at the lower, left, and right boundaries, while \( u(x) = \sin(\pi x) \) at the upper boundary. Assume that we use central differences to approximate the solution.

(a) Show that the numerical solution at a grid point \( x = (x, y) \) is bounded by \( x(1-x)y + \sin \pi \).

(b) Show that the global error vector \( e \) is bounded by

\[ \|e\|_{\infty} \leq \left( \frac{1}{8} + \sin \pi \right) \|d\|_{\infty}, \]

where \( d \) is the local error vector.

9.4. Consider the problem

\[ \mathcal{L}[u] := \nabla^2 u - \frac{\partial u}{\partial x} = 0 \]

on the unit square. We use central differences to discretise the problem.

(a) Determine a maximum value for \( \Delta x \) such that the scheme satisfies the requirements of Property 9.1.

(b) Use a comparison function to give an estimate for the norm of the global error in terms of the norm of the local error.

Hint: You may use, e.g., \( v(x, y) = x^2 + y^2 \).

9.5. Consider the mixed problem

\[ \mathcal{L}[u] := \nabla^2 u = f \]

defined on the unit square with homogeneous Dirichlet boundary conditions except at the right boundary, where \( \frac{\partial u}{\partial x} = 0 \).

(a) Use a one-sided difference to discretise the Neumann boundary conditions. Show how the stencil looks at a point with a right boundary point as neighbour. Show that this gives errors like \( O(\Delta x)^{-1} \).

(b) Show that the global error is at least first order accurate in \( \Delta x \).

9.6. In this exercise we want to determine the coefficients in the LU-decomposition of the matrix resulting from the central difference discretisation of

\[ \mathcal{L}[u] := \frac{d^2 u}{dx^2} = f, \quad x \in (0, 1), \]

with Dirichlet and with mixed boundary conditions.

(a) Do this for the case where \( u \) is given at \( x = 0 \) and \( x = 1 \).

(b) Next consider the case where \( u \) is given at \( x = 0 \) and \( \frac{du}{dx} \) is given at \( x = 1 \). Determine the LU-decomposition when \( \frac{du}{dx}(1) \) is approximated by a backward difference.
(c) Repeat part (b) with Neumann and Dirichlet conditions interchanged. Which difference approximation would you use for $\frac{d^2 u}{dx^2}$?

9.7. Show that LU-decomposition of the matrix $A$ in (9.45) gives $L$ and $U$ matrices whose $M$ diagonals immediately below and above the main diagonal (the codiagonals) will fill in (at least eventually).

9.8. Consider the matrix

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$ Determine the spectral radii $\rho(M_J)$ and $\rho(M_S)$.

9.9. Assume that a matrix $A = (a_{i,j})$ is strongly diagonally dominant; i.e.,

$$\sum_{j \neq i} a_{i,j} < a_{i,i} \quad \text{for all } i.$$ Show that both the Gauss–Jacobi and the Gauss–Seidel methods converge.

9.10. Assume that the matrices $A_x$ and $A_y$ in (9.66) commute. Let us denote their eigenvalues by $\lambda_j$ and $\mu_k$, respectively.

(a) Show that the convergence of ADI is determined by the quotient

$$\max_{\lambda_j, \mu_k} \left| \frac{1 + \tau \lambda_j}{1 - \tau \lambda_j} \frac{1 + \tau \mu_k}{1 - \tau \mu_k} \right|.$$ (b) Show that convergence is assured if $A_x$ and $A_y$ have negative eigenvalues.

(c) Check the requirement under part (b) for a simple case.

9.11. Consider the problem

$$L[u] := \frac{d^2 u}{dx^2} = e^x, \quad x \in (0, 1),$$

with $u$ given at the boundary points $x = 0$ and $x = 1$. We discretise this boundary value problem with central differences. Show that the Newton algorithm for the resulting nonlinear system converges, at least locally.

9.12. Let $\Omega$ be the trapezium defined by the corner points $(0, 0), (0, 2), (2, 1), (2, 0)$. We want to discretise the problem

$$\nabla^2 u = f, \quad x \in \Omega.$$ Let the grid size be given by $h = \Delta x = \Delta y = 2/M$ for some $M$.

(a) Assume that Dirichlet boundary conditions are given. Use central differences to discretise this problem. Pay special attention to points with a neighbour at the right boundary; see also Exercise 7 in Chapter 5.

(b) Show that the global discretisation error is second order.

9.13. Consider the Laplace equation in polar coordinates

$$\nabla^2 u := \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0,$$
defined on the unit circle, with boundary value

\[ u(1, \phi) = 1. \]

Let \( \Delta r \) and \( \Delta \phi \) be the grid size in the \( r \) and \( \phi \) directions, respectively, and use central differences.

(a) Find the system of equations for the unknowns at the grid points, ordered by taking them counterclockwise and from the innermost circle to the outermost circle of grid points.

(b) Find the equation at \((0, 0)\); cf. Section 5.2.3.

9.14. Consider the Laplace equation (cf. (8.41)).

\[ \nabla^2 u := \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0 \]

defined on a finite wedge \( \{(r, \phi) \mid 0 < r < 1, 0 < \phi < \frac{1}{4} \pi\} \), where \( u \) satisfies the boundary conditions

\[ u(r, 0) = u(r, \pi/4) = 0, \quad u(1, \phi) = \sin 4 \phi. \]

(a) Use central differences to discretise the problem. Give the explicit form of the system for the unknowns at the grid points.

(b) Solve the system for various values of the grid sizes \( \Delta r \), \( \Delta \phi \) and make a table of the error at a number of points (on the intersection of the grids) close to \( 0 \).

Note: You can find the exact solution from (8.41).

9.15. In (9.108) we have a linear system that we hope has a unique solution. To see this we do the following: Instead of (9.108b), we take

\[ -B^T u = f_2. \]

Show that (9.108a) together with this equation leads to a positive definite symmetric system. Note: A matrix \( C \) is positive definite if \( z^T C z > 0 \) for \( z \neq 0 \). Complete the proof.
Chapter 10

Analysis of Parabolic Equations

The analytical theory of parabolic equations in this chapter starts (Section 10.1) with deriving solutions for Cauchy problems, i.e., pure initial value problems. The dependence of (fundamental) solutions on the time and space variables turns out to respect certain symmetries (Section 10.2). We study similarity solutions in Section 10.3, where we also derive formulas for the diffusion operator in other than Cartesian coordinates. Next we investigate the role of boundary conditions in Section 10.4. First we analyse problems on finite (spatial) domains. A special class of problems is formed by PDEs with moving boundaries called Stefan problems (Section 10.5). The last section, Section 10.6, is devoted to steady state solutions and travelling-wave solutions. A point of interest here is that stationary solutions constitute proper solutions of corresponding elliptic boundary value problems (which are the subject of Chapter 8).

10.1 Cauchy Problems

Parabolic equations arise in a variety of applications, mainly associated with diffusive processes. In Section 1.1 we saw an example of (chemical) diffusion. Heat flow also has a diffusive character, often called conduction. This may be combined with convection if the medium is a fluid in motion. In mechanics of fluids, internal friction, called viscosity, produces diffusion of momentum. In this section we shall consider the simplest form of a diffusive problem: the linear heat equation on infinite domains.

10.1.1 The Heat Equation in One Space Dimension

In order to investigate solutions of parabolic equations, we consider the following initial value problem for the simplest (nondimensional) form of the heat equation:

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in \mathbb{R}, \quad t > 0, \quad (10.1a)
\]

\[
u(x, 0) = v(x), \quad x \in \mathbb{R}, \quad (10.1b)
\]
where \(|v|\) and \(|u|\) are integrable over \(\mathbb{R}\). (Note that limiting conditions for \(x \to \pm\infty\) are to be expected if we portray the present initial value problem as the limit of an initial boundary value problem. See Example 10.2.) We seek a solution of the form
\[ u(x, t) = p(x)q(t); \quad (10.2) \]
i.e., we separate the independent variables so that the functions \(p\) and \(q\) in the product depend on either one of the variables \(x\) and \(t\) exclusively. Upon substitution we find
\[ \frac{dq}{dt} = \frac{d^2 p}{dx^2} q, \]
or, assuming \(p(x), q(t) \neq 0\),
\[ \frac{1}{q} \frac{dq}{dt} = \frac{1}{p} \frac{d^2 p}{dx^2}. \quad (10.3) \]
The left-hand side of (10.3) depends solely on \(t\) and the right-hand side depends solely on \(x\). This is only possible if both sides are equal to some constant \(\lambda\), the separation constant. (As will appear below, it is sufficient to assume that \(\lambda\) is real.) So we obtain the two eigenvalue problems
\[ \frac{d^2 p}{dx^2} = \lambda p, \quad (10.4a) \]
\[ \frac{dq}{dt} = \lambda q. \quad (10.4b) \]

From (10.4b) we conclude that only nonpositive values of \(\lambda\) are allowed for a stable solution \(q(t)\). Indeed, if \(\lambda > 0\), then (10.4a) will have solutions \(p(x)\) that are exponentially increasing for either \(x \to -\infty\) or \(x \to \infty\). Also, for \(\lambda = 0\), (10.4a) has an unbounded solution given by \(p(x) = C_1 x + C_2\). Therefore, in order for the solution \(u(x, t)\) of (10.1a) to be bounded, we assume (with \(\kappa\) real) that
\[ \lambda = -\kappa^2 < 0. \quad (10.5) \]
A general solution of (10.4) is then given by
\[ p(x)q(t) = Ce^{i\kappa x - \kappa^2 t}, \quad (10.6) \]
where \(C\) is a constant. Here we recognize a planar wave from the dispersion relation for the heat equation (10.1a). Apparently, we can view this planar wave as a Fourier mode. By Fourier analysis (Chapter 3), a general bounded solution of (10.1) can now be found by superposition over all possible \(\kappa\). The suggested solution is now given by
\[ u(x, t) = \int_{-\infty}^{\infty} \hat{v}(\kappa) e^{i\kappa x - \kappa^2 t} \, dx, \quad (10.7) \]
where, using the initial value \(u(x, 0) = v(x)\),
\[ \hat{v}(\kappa) := \frac{1}{2\pi} \int_{-\infty}^{\infty} v(\xi) e^{-i\kappa \xi} \, d\xi. \quad (10.8) \]
By substituting (10.8) in (10.7) and interchanging the order of integration, we obtain

\[ u(x,t) = \int_{-\infty}^{\infty} v(\xi) \psi(x - \xi, t) \, d\xi, \quad (10.9a) \]

where

\[ \psi(\zeta, t) := \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\kappa\zeta} e^{-\kappa^2 t} \, d\kappa = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(\kappa/2i)^2 - \zeta^2/4t} \, d\kappa = \frac{e^{-\zeta^2/4t}}{2\sqrt{\pi t}}. \quad (10.9b) \]

We shifted the \( \kappa \) contour by an amount \( i\zeta/2t \) into the complex plane. Altogether we obtain the solution

\[ u(x,t) = \frac{1}{\sqrt{\pi t}} \int_{-\infty}^{\infty} v(\xi) \exp \left( -\frac{(x-\xi)^2}{4t} \right) \, d\xi. \quad (10.10a) \]

We remark that the derivation closely follows that of the Duhamel integral in Section 4.6. One can easily verify that this solution satisfies (10.1a). By changing the coordinate of integration to \( \xi = x + 2\sqrt{t} \eta \), we find

\[ u(x,t) = \frac{1}{\sqrt{\pi t}} \int_{-\infty}^{\infty} v(x + 2\sqrt{t} \eta) e^{\eta^2} \, d\eta. \quad (10.10b) \]

The initial condition is now readily verified to be

\[ u(x,0) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} v(x) e^{-\eta^2} \, d\eta = v(x) \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\eta^2} \, d\eta = v(x). \]

In Section 10.1.2 we shall show that the solution is also unique.

Next we show that \( \psi(\zeta, t) \) is equivalent to a fundamental solution; see (4.5). By taking \( v(x) = \delta(x) \) in (10.10a), we immediately see that \( \psi \) satisfies the initial value problem

\[ \frac{\partial \psi}{\partial t} - \frac{\partial^2 \psi}{\partial \zeta^2} = 0, \quad \zeta \in \mathbb{R}, \quad t > 0, \quad (10.11a) \]

\[ \psi(\zeta, 0) = \delta(\zeta), \quad \zeta \in \mathbb{R}. \quad (10.11b) \]

A fundamental solution \( w(x, t; \xi, \tau) \) of (10.1) satisfies

\[ \frac{\partial w}{\partial t} - \frac{\partial^2 w}{\partial x^2} = \delta(x-\xi)\delta(t-\tau), \quad x \in \mathbb{R}, \quad t > \tau, \quad (10.12a) \]

\[ w(x, \xi, t, \tau) = 0, \quad t < \tau. \quad (10.12b) \]

**Property 10.1.** The function \( w(x, t; \xi, \tau) = \psi(x - \xi, t - \tau)H(t - \tau) \), where \( H \) is the Heaviside step function, is a fundamental solution of (10.1).
Proof. $H(t - \tau) \equiv 0$ for $t < \tau$. Furthermore, the difference between

$$w_t = \psi_t H + \psi \delta = \psi_t H + \delta(x - \xi) \delta(t - \tau)$$

and $w_{xx} = \psi_{\xi \xi} H$ is just $\delta(x - \xi) \delta(t - \tau)$.

It follows that for $t > \tau$ we can identify $w(x, t; \xi, \tau)$ with $\psi(x - \xi, t - \tau)$. We remark that the $\psi$ found in (10.9b) apparently satisfies

$$\int_{-\infty}^{\infty} \psi(\zeta, t) \, d\zeta = \frac{1}{\sqrt{\pi t}} \int_{-\infty}^{\infty} e^{-\zeta^2/4t} \, d\zeta = 1. \quad (10.13a)$$

Hence $\psi$ can be seen as a probability distribution. In fact, it is the famous Gaussian or normal distribution. From the rightmost side of (10.13a) we see that

$$\text{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-\mu^2} \, d\mu, \quad (10.13b)$$

the error function, which plays such an important role in statistics, is related to the problem here. The same is true for the complementary error function defined by

$$\text{erfc}(x) := \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-\mu^2} \, d\mu = 1 - \text{erf}(x). \quad (10.13c)$$

We will frequently encounter this function further on.

Example 10.2 If we do not restrict ourselves, in problems on infinite domains, to bounded solutions, we may easily encounter solutions that grow rather than decay in time. This growth may be associated with possible sources at infinity. Consider the examples

$$u(x, 0) = e^t \quad \text{with} \quad u(x, t) = e^{t^2};$$

$$u(x, 0) = \frac{1}{2} x^2 \quad \text{with} \quad u(x, t) = \frac{1}{2} x^2 - t. \quad \Box$$

Example 10.3 Rather than (10.1a), one may have

$$\frac{\partial u}{\partial t} = k \frac{\partial^2 u}{\partial x^2},$$

where $k$ is a positive constant called the (thermal) diffusivity. We may apply a change of variables, $s := tk$ or $z := x/\sqrt{k}$, to obtain either $u_s = u_{xx}$ or $u_z = u_{zz}$. \Box

10.1.2 The Heat Equation in $d$ Space Dimensions

Our formulation in Section 10.1.1 was chosen such that it lends itself easily to heat equations in more than one space variable. So consider

$$\frac{\partial u}{\partial t} = \nabla^2 u, \quad x \in \mathbb{R}^d, \quad t > 0, \quad (10.14a)$$

$$u(x, 0) = v(x), \quad x \in \mathbb{R}^d, \quad (10.14b)$$
10.1. Cauchy Problems

where $|v|$ and $|u|$ are integrable over $\mathbb{R}^d$. A fundamental solution $w$ (cf. Section 4.5) then satisfies (cf. (10.12))

$$
\frac{\partial w}{\partial t} - \nabla^2 w = \delta(x - \xi)\delta(t - \tau), \quad x \in \mathbb{R}^d, \quad t > 0 \tag{10.15a}
$$

$$
w(x, t; \xi, \tau) = 0, \quad t < \tau. \tag{10.15b}
$$

By analogy with Property 10.1, we have

$$
w(x, t; \xi, \tau) = \left(\frac{1}{2\sqrt{\pi(t-\tau)}}\right)^{d/2} \exp\left(-\frac{\|x - \xi\|^2}{4(t-\tau)}\right) H(t-\tau). \tag{10.16}
$$

Superposition then gives (cf. Chapter 4)

$$
u(x, t) = \left(\frac{1}{2\sqrt{\pi t}}\right)^d \int_{\mathbb{R}^d} v(\xi) \exp\left(-\frac{\|x - \xi\|^2}{4t}\right) d\xi. \tag{10.17}
$$

This solution exists and is unique.

**Theorem 10.4.** Let $v$ be continuous and integrable on $\mathbb{R}^d$. Then the function given by (10.17) satisfies initial value problem (10.14) uniquely.

**Proof.** By virtue of the exponential and the fact that $v$ is bounded, the defining integral of $u$ is uniformly integrable for all $x$ and $t > 0$. So $u$ and likewise $\int w v(\xi) d\xi$ and $\int \nabla^2 w v(\xi) d\xi$ exist, and we can interchange differentiation and integration. Since $w = \nabla^2 w$, (10.14a) readily follows. To show that $u$ also satisfies the initial condition we change variables $\xi = x + 2\eta \sqrt{t}$ to obtain

$$
u(x, t) = \left(\frac{1}{\sqrt{\pi}}\right)^d \int_{\mathbb{R}^d} v(x + 2\eta \sqrt{t}) e^{-\|\eta\|^2} d\eta.
$$

With identity (10.13a), this is clearly equal to $v(x)$ for $t = 0$.

To prove that the solution is unique we assume that $u_1$ and $u_2$ are two solutions. Then the difference $w(x, t) := u_1(x, t) - u_2(x, t)$ satisfies the initial value problem with $v(x, t) \equiv 0$. From Green’s first identity we have

$$
\frac{d}{dt} \int_{\mathbb{R}^d} \frac{1}{2} w^2 \, dV = \int_{\mathbb{R}^d} \frac{\partial w}{\partial t} \nabla^2 w \, dV = -\int_{\mathbb{R}^d} \nabla^2 w \, dV = -\int_{\mathbb{R}^d} |\nabla w|^2 \, dV < 0.
$$

Hence

$$
0 \leq \int_{\mathbb{R}^d} w^2(x, t) \, dV \leq \int_{\mathbb{R}^d} w^2(x, 0) \, dV = 0,
$$

so $w \equiv 0$. \qed

### 10.1.3 Problems on Half-Spaces

It is of interest to study the solution of problems defined on a semi-infinite domain with values prescribed at the boundary. Although such a setting can also be handled in higher
dimensions, we restrict ourselves to the one-dimensional case, with the domain \([0, \infty)\). We are looking for the solution of
\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in [0, \infty), \quad t > 0,
\]  
(10.18)
satisfying either
\[
\begin{align*}
u(x, 0) &= v(x), \quad x \in [0, \infty), \\
u(0, t) &= 0, \quad t > 0,
\end{align*}
\]
(10.19a)
for some given initial value \(v(x)\) or
\[
\begin{align*}
u(x, 0) &= 0, \quad x \in [0, \infty), \\
u(0, t) &= \beta(t), \quad t > 0,
\end{align*}
\]
(10.20a)
for some boundary value \(\beta(t)\). Clearly, the conditions (10.19) are different from but related to the one we encountered in (10.1b).

First consider initial boundary value problem (10.18) and (10.19). Formally, we may try to use a fundamental solution approach by employing a continuation of the initial condition on the left interval \((-\infty, 0]\). Let \(\tilde{v}(x)\) be equal to \(v(x)\) for \(x \geq 0\) and some continuous extension for \(x < 0\) but such that the representation
\[
u(x, t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} \tilde{v}(\xi) \exp \left( -\frac{(x - \xi)^2}{4t} \right) \, d\xi
\]
(10.21)
still satisfies boundary condition (10.19b). Simple substitution then reveals
\[
\frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} \tilde{v}(\xi) \exp \left( -\frac{\xi^2}{4t} \right) \, d\xi = \frac{1}{2\sqrt{\pi t}} \int_{0}^{\infty} \left( \tilde{v}(\xi) + \tilde{v}(-\xi) \right) \exp \left( -\frac{\xi^2}{4t} \right) \, d\xi = 0,
\]
(10.22)
which is true for all \(t\) if
\[
\tilde{v}(\xi) = -v(-\xi), \quad \xi < 0.
\]
(10.23)
As a consequence we find the representation
\[
u(x, t) = \frac{1}{2\sqrt{\pi t}} \int_{0}^{\infty} v(\xi) \left[ \exp \left( -\frac{(x - \xi)^2}{4t} \right) - \exp \left( -\frac{(x + \xi)^2}{4t} \right) \right] \, d\xi.
\]
(10.24)

**Example 10.5** An important application is the heat shock. Suppose an object with a certain given temperature is suddenly exposed to a heat source (or sink). If this has a temperature different from the object, this means that the temperature profile initially exhibits a discontinuity. The simplest model is given by assuming that the initial temperature is constant; i.e.,
\[
v(x) = v_0 > 0.
\]
(See Example 7.17 for the related problem of a given heat flux.) The solution of initial boundary value problem (10.18) and (10.19) is then given by
\[
u(x, t) = \frac{v_0}{2\sqrt{\pi t}} \left[ \int_{0}^{\infty} \exp \left( -\frac{(x - \xi)^2}{4t} \right) \, d\xi - \int_{0}^{\infty} \exp \left( -\frac{(x + \xi)^2}{4t} \right) \, d\xi \right].
\]
10.1. Cauchy Problems

Substituting $z := -(x - \xi)/2\sqrt{t}$ in the first integral and $z := (x + \xi)/2\sqrt{t}$ in the second one gives

\[
u(x, t) = \frac{v_0}{\sqrt{\pi}} \left[ \int_{-\infty}^{\infty} e^{-z^2} \, dz - \int_{\sqrt{2}/2\sqrt{t}}^{\infty} e^{-z^2} \, dz \right] = v_0 \text{erf} \left( \frac{x}{2\sqrt{t}} \right).
\]

In Figure 10.1 we have displayed various stages of the solution for $v_0 = 1$.

**Figure 10.1. Solution of the heat shock problem at various time levels.**

**Example 10.6** We can also apply the analysis to a two-sided heat shock. Indeed, consider an object that is suddenly heated in a finite region, $[-1, 1]$ say. A very simple model is then given by the piecewise constant initial condition

\[
u(x) = \begin{cases} 
0 & \text{if } x \in (-\infty, -1), \\
v_0 & \text{if } x \in [-1, 1], \\
0 & \text{if } x \in (1, \infty), 
\end{cases}
\]

where $v_0 \neq 0$, yielding the solution

\[
u(x, t) = \frac{v_0}{2\sqrt{\pi t}} \int_{-1}^{1} \exp \left( -\frac{(x - \xi)^2}{4t} \right) \, d\xi,
\]

or, upon substituting $z := (x - \xi)/2\sqrt{t}$,

\[
u(x, t) = \frac{v_0}{\sqrt{\pi}} \int_{(x-1)/2\sqrt{t}}^{(x+1)/2\sqrt{t}} e^{-z^2} \, dz = \frac{v_0}{2} \left( \text{erf} \left( \frac{x+1}{2\sqrt{t}} \right) - \text{erf} \left( \frac{x-1}{2\sqrt{t}} \right) \right),
\]

which can also be found by employing the solution of the previous example.

The other situation, i.e., initial boundary value problem (10.18) and (10.20), can be solved using the solution of the initial boundary value problem (10.18) and (10.19). First, consider the problem where $\beta(t) \equiv 1$. It is simple to see that

\[
u(x, t) = 1 - \text{erf} \left( \frac{x}{2\sqrt{t}} \right) = \text{erfc} \left( \frac{x}{2\sqrt{t}} \right)
\]

(10.25)
Chapter 10. Analysis of Parabolic Equations

satisfies (10.18) and (10.20). We can now invoke the Duhamel integral (see Theorem 4.16) to obtain the solution. Note that we can define solutions \( w(x, t - \tau) \) by

\[
\begin{align*}
  w(x, t - \tau) & := \operatorname{erfc} \left( \frac{x}{2\sqrt{t - \tau}} \right).
\end{align*}
\]

(10.26)

Hence we obtain as the formal solution of (10.18) and (10.20) from (4.68) that

\[
\begin{align*}
  u(x, t) & = \frac{\partial}{\partial t} \int_0^t \beta(\tau) \operatorname{erfc} \left( \frac{x}{2\sqrt{t - \tau}} \right) \, d\tau.
\end{align*}
\]

(10.27)

We can work this out to get

\[
\begin{align*}
  u(x, t) & = \frac{x}{2\sqrt{\pi}} \int_0^t \beta(\tau) \left( t - \tau \right)^{3/2} \exp \left( - \frac{x^2}{4(t - \tau)} \right) \, d\tau.
\end{align*}
\]

(10.28)

One may check that this solution satisfies the initial and boundary conditions.

Other than for the simplest forms of \( \beta(\tau) \) this formula does not provide an explicit answer. Nevertheless, it may be very useful for order of magnitude estimates or the analysis of trends or asymptotic behaviour. On the other hand, if we are interested in actual numbers for the general case, we have to evaluate the integral numerically. Although this is not a major problem (the apparent square root singularity in \( \tau = t \) is completely cancelled by the exponential), we have to compare the effort with other, more direct, numerical methods for solving the problem.

10.2 The Heat Equation with Spatial Symmetries

In Section 10.1.2 we formulated the solution of the \( d \)-dimensional heat equation

\[
\frac{\partial u}{\partial t} = \nabla^2 u, \quad x \in \mathbb{R}^d, \quad t > 0,
\]

(10.29)

for a given initial value for \( u \). If boundary and initial conditions are cylindrically or spherically symmetric, so is the solution, and it is beneficial to rewrite the Laplacian \( \nabla^2 \) in the corresponding coordinates (see the appendix, Section K) and suppress any of the nonradial derivatives. With circular (in two dimensions) or cylindrical (in three dimensions) symmetry, with \( r^2 = x^2 + y^2 \), we have

\[
\frac{\partial u}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) \quad \text{or} \quad \frac{\partial u}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{\partial^2}{\partial z^2} u.
\]

(10.30a)

With spherical symmetry, with \( r^2 = x^2 + y^2 + z^2 \), we have

\[
\frac{\partial u}{\partial t} = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right).
\]

(10.30b)

In general we have thus for a radially symmetric field in \( d = 1, 2, 3 \)-dimensional space that

\[
\frac{\partial u}{\partial t} = \frac{1}{r^{d-1}} \frac{\partial}{\partial r} \left( r^{d-1} \frac{\partial u}{\partial r} \right) \quad \text{for} \quad r = |x|, \quad x \in \mathbb{R}^d, \quad t > 0.
\]

(10.30c)
10.3. Similarity Solutions

Evidently, these PDEs are defined on the half-space \( r \geq 0 \). If \( u \) is smooth at \( r = 0 \), the radial symmetry induces the boundary condition

\[
\frac{\partial u}{\partial r} = 0.
\] (10.31)

Finally, we remark that if for a circular or cylindrically symmetric problem the \( r \) domain consists of values \( r \geq R_0 > 0 \), for some \( R_0 \), we can use the transformation

\[
v(\rho, t) := u(r, t), \quad \rho := \ln r,
\] (10.32a)

resulting, for (10.30a), in

\[
\frac{\partial v}{\partial t} = e^{-2\rho} \frac{\partial^2 v}{\partial \rho^2} \quad \text{or} \quad \frac{\partial v}{\partial t} = e^{-2\rho} \frac{\partial^2 v}{\partial \rho^2} + \frac{\partial^2 v}{\partial z^2}.
\] (10.32b)

A three-dimensional spherically symmetric problem can be transformed, for \( r \geq R_0 \), for some \( R_0 > 0 \), by

\[
v(r, t) := u(r, t)r
\] (10.33a)

to a form equivalent to the one-dimensional standard heat equation

\[
\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial r^2}.
\] (10.33b)

This shows that any of the foregoing one-dimensional solutions corresponds immediately to a similar three-dimensional solution.

10.3 Similarity Solutions

In Section 7.3 we found by dimensional arguments that, in the absence of explicit length and time scales other than \( x \) and \( t \) themselves, dimensionless groups can only occur by mutual combinations of \( x \) and \( t \). In the present problems related to the heat equation the prevailing combination appears to be \( x/\sqrt{t} \). In the same way the dependent variable (the temperature, say) will scale on a given value if this is available, but will have to depend on a combination of \( x \) and \( t \), and the flux if no value but only a flux is given. These observations give rise to similarity solutions (see also Section 2.5).

As these symmetry properties should be present in the problem independently of the physics, it seems useful to look for similarity solutions as soon as the spatial and time coordinates have an infinite extension. Assume that the problem is radially symmetric, so \( u = u(r, t) \) where \( d = 1, 2, \) or 3, and \( r \) denotes the Euclidean distance to the origin. Consider, therefore,

\[
\frac{\partial u}{\partial t} = \frac{1}{r^{d-1}} \frac{\partial}{\partial r} \left( r^{d-1} \frac{\partial u}{\partial r} \right) \quad \text{for} \quad r = |x|, \quad x \in \mathbb{R}^d, \quad t > 0.
\] (10.30c)

We will look for solutions of the form

\[
u(r, t) = r^m f(\eta), \quad \eta = \frac{r}{r^m}.
\] (10.34)
As \( r = t^n \eta \), there is no need to include any factor of the type \( t^p \). Differentiation with respect to \( t \) and \( r \) yields

\[
\frac{\partial u}{\partial t} = -n r^{m-2} \eta^3 t^{2n-1} f'(\eta), \quad (10.35a)
\]

\[
\frac{\partial u}{\partial r} = r^{m-1} (mf(\eta) + \eta f'(\eta)), \quad (10.35b)
\]

\[
\frac{\partial^2 u}{\partial r^2} = r^{m-2} \left( m(m-1) f(\eta) + 2m\eta f'(\eta) + \eta^2 f''(\eta) \right) \quad (10.35c)
\]

(the prime denotes differentiation with respect to \( \eta \)). After substitution in (10.30c), we find

\[
\eta^2 f'' + \eta \left( \frac{1}{2} \eta^2 + 2m + d - 1 \right) f' + m(m + d - 2) f + n\eta^3 t^{2n-1} f' = 0, \quad (10.36)
\]

which only has proper solutions if \( n = \frac{1}{2} \). So we finally have the equation

\[
\eta^2 f'' + \eta \left( \frac{1}{2} \eta^2 + 2m + d - 1 \right) f' + m(m + d - 2) f = 0. \quad (10.37)
\]

A solution of this equation for arbitrary \( m \) and \( d \) can be given in terms of hypergeometric or similar functions (see [1]), but this is too general to be of interest. It is more expedient to look for specific solutions once \( d \) and \( m \) are known. The value of \( m \) has to follow from the available boundary conditions and is sometimes immediately clear from dimensional arguments. A typical case may be found in Example 7.17.

**Example 10.7** An important example of a similarity solution is the field of a steady point source. This is most efficiently found from the fundamental solution given in (10.16). This gives us instantly the field of a stroke of heat from the point source \( \delta(x) \delta(t) \), i.e.,

\[
u(x, t) = \left( \frac{1}{2\sqrt{\pi t}} \right)^d \exp \left( \frac{-x^2}{4t} \right).
\]

By integrating this solution with respect to time, we find the field of a steady point source \( \delta(x)H(t) \) at the origin, switched on at \( t = 0 \), thus satisfying

\[
\frac{\partial U}{\partial t} - \nabla^2 U = \delta(x)H(t).
\]

In one dimension this is

\[
U_1(x, t) = \int_0^t e^{-r^2/4t} \sqrt{4\pi \tau} d\tau = \sqrt{\frac{t}{\pi}} \exp \left( -\frac{x^2}{4t} \right) - \frac{1}{2} |x| \operatorname{erfc} \left( \frac{|x|}{\sqrt{4t}} \right),
\]

which we already encountered in Example 7.17. In two dimensions we have

\[
U_2(r, t) = \int_0^t \frac{e^{-r^2/4t}}{4\pi \tau} d\tau = \frac{1}{4\pi} E_1 \left( \frac{r^2}{4\tau} \right), \quad \text{where} \quad E_1(z) = \int_1^\infty \frac{e^{-z}}{t} \, dt.
\]

\( E_1(z) \) is known as the exponential integral [1]. In three dimensions we obtain

\[
U_3(r, t) = \int_0^t \frac{e^{-r^2/4t}}{(4\pi \tau)^{3/2}} d\tau = \frac{1}{4\pi r} \operatorname{erfc} \left( \frac{r}{\sqrt{4t}} \right).
\]
10.4. Initial Boundary Value Problems

It may be verified that indeed the respective fluxes out of the source are given by
\[
\lim_{h \to 0} \left[ -\frac{\partial U_1}{\partial x} \right]_{x=h} = \lim_{h \to 0} \left[ \frac{1}{2} \text{sign}(x) \operatorname{erfc}\left(\frac{|x|}{\sqrt{4t}}\right) \right]_{x=-h} = 1,
\]
\[
\lim_{h \to 0} \left[ -2\pi r \frac{\partial U_2}{\partial r} \right]_{r=h} = \lim_{h \to 0} \left[ \exp\left(-\frac{r^2}{4t}\right) \right]_{r=h} = 1,
\]
\[
\lim_{h \to 0} \left[ -4\pi r^2 \frac{\partial U_3}{\partial r} \right]_{r=h} = \lim_{h \to 0} \left[ \frac{r}{\sqrt{\pi t}} \exp\left(-\frac{r^2}{4t}\right) + \operatorname{erfc}\left(\frac{r}{\sqrt{4t}}\right) \right]_{r=h} = 1.
\]

It is remarkable that the field of a point source of constant output sometimes (if \(d = 3\)) converges to a stationary state, but not always (if \(d = 1, 2\)). This depends on \(d\), the number of spatial dimensions of the problem. For large \(t\) we find asymptotically \([1]\)
\[
U_1(x, t) = \sqrt{\frac{T}{\pi}} - \frac{1}{2} |x| + \mathcal{O}(t^{-1/2}),
\]
\[
U_2(r, t) = \frac{1}{4\pi} \left[ \ln(t) - \ln\left(\frac{1}{4} r^2\right) - \gamma \right] + \mathcal{O}(t^{-1}),
\]
\[
U_3(r, t) = \frac{1}{4\pi r} + \mathcal{O}(t^{-1/2}),
\]
where \(\gamma = 0.5772\ldots\). Apparently, only the capacity of three-dimensional space is big enough to absorb all the heat of a stationary source! □

10.4 Initial Boundary Value Problems

When the problem is defined on a domain with boundaries, we need to provide boundary conditions. Sometimes these boundaries are part of the evolution process; we shall address the latter in the next section. In order to fix our thoughts, we think of \(u\) as a temperature here. Consider, as before, a PDE with given initial value:
\[
\frac{\partial u}{\partial t} = \nabla \cdot (k \nabla u) = k \nabla^2 u, \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0,
\]
\[
u(x, 0) = v(x),
\]
where \(k\) denotes the thermal diffusion coefficient and \(-k \nabla u\) is the heat flux vector. The heat flux across a surface with unit normal vector \(\mathbf{n}\) is thus given by \(-k \frac{\partial u}{\partial n}\). The first form of (10.38) is more general and is valid if \(k\) depends on \(x\) or \(u\) (or both). We assume that \(\Omega\) is a compact simply connected domain with boundary \(\partial \Omega\). Apart from a prescribed initial value, the following boundary conditions are usually considered:

(i) If the temperature at the boundary is prescribed, we call this a boundary condition of Dirichlet type:
\[
u(x) = g(x, t), \quad x \in \partial \Omega, \quad t > 0.
\]

(ii) If the heat flux across the boundary is prescribed, we have a boundary condition of Neumann type:
\[
-k \frac{\partial u}{\partial n} (x, t) = g(x, t), \quad x \in \partial \Omega, \quad t > 0.
\]

Here \(\frac{\partial u}{\partial n} := \mathbf{n} \cdot \nabla u\), where \(\mathbf{n}\) is the unit outward normal vector on \(\partial \Omega\).
A linear combination of Dirichlet and Neumann type boundary conditions is called a \textit{Robin} or a \textit{Newton} boundary condition:

\[-k \frac{\partial u}{\partial n}(x, t) = s(u(x, t) - g(x, t)), \quad x \in \partial \Omega, \quad t > 0, \quad (10.41)\]

with the heat transfer coefficient \(s > 0\). Put in physical terms, the heat flux is proportional to the difference between the temperature of the medium and the ambient.

If the loss of heat is caused by radiation, we have a \textit{radiative} boundary condition:

\[-k \frac{\partial u}{\partial n}(x, t) = \sigma (u(x, t) - u_\infty^4), \quad x \in \partial \Omega, \quad t > 0, \quad (10.42)\]

where \(u_\infty\) is the specified ambient temperature of the surrounding medium and the parameter \(\sigma\) quantifies the emissivity properties of the surface.

We remark that boundary conditions without external forcing, i.e., where \(g(x, t) \equiv 0\), are called homogeneous.

\textbf{Example 10.8} Consider the initial boundary value problem

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0, \\
u(x, 0) &= v(x), \quad x \in (0, 1), \\
u(0, t) &= u(1, t) = 0, \quad t > 0.
\end{align*}
\]

Clearly the boundary conditions are of Dirichlet type. Assume that \(v\) and \(v'\) are piecewise smooth such that the coefficients of the Fourier-sine expansion

\[v(x) = \sum_{k=1}^{\infty} v_k \sin(k\pi x)\]

converge according to \(v_k = O(1/k)\) as \(k \to \infty\) (Corollary 3.7). Consider the formal series

\[u(x, t) = \sum_{k=1}^{\infty} v_k e^{-k^2 \pi^2 t} \sin(k\pi x)\]

Due to the exponential, the coefficients behave as \(v_k e^{-k^2 \pi^2 t} = O(k^{-n})\) as \(k \to \infty\) for any \(n\) if \(t > 0\), so the series and all of its derivatives converge uniformly for \(t > 0\) (appendix, Section C). As a result, \(u\) is continuous, while differentiation and summation may be exchanged. It is then easily verified that \(u\) satisfies the differential equation, initial profile \(v\), and boundary values \(u(0, t) = u(1, t) = 0\). Note in particular that in the heat equation no discontinuity in any derivative can be sustained: (practically) any initial profile \(v\) yields immediately, for any \(t > 0\), an infinitely differentiable temperature distribution \(u\).

\textbf{Example 10.9} Consider the initial boundary value problem

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0, \\
u(x, 0) &= v(x), \quad x \in (0, 1), \\
u_x(0, t) &= \nu_x(1, t) = 0, \quad t > 0.
\end{align*}
\]
This is a boundary condition of Neumann type. Analogously to the previous example, we assume that \( v \) and \( v' \) are piecewise smooth and expand \( v \) into the Fourier-cosine series
\[
v(x) = \sum_{k=0}^{\infty} v_k \cos(k\pi x),
\]
whose coefficients converge like \( v_k = O(1/k) \) as \( k \to \infty \). The formal solution
\[
u(x, t) = \sum_{k=0}^{\infty} v_k e^{-k^2 \pi^2 t} \cos(k\pi x)
\]
and any of its derivatives converge uniformly for \( t > 0 \) to a continuous function. So this solution satisfies the equation and initial condition. In particular, it satisfies the boundary conditions.

Uniqueness of solutions can be shown by integral arguments.

**Theorem 10.10.** The solution of the linear initial boundary value problem (10.38), i.e., \( k \) is independent of \( u \), with any linear boundary condition (10.39), (10.40), or (10.41), is unique.

**Proof.** Let \( u_1 \) and \( u_2 \) be two solutions. Then \( w(x, t) := u_1(x, t) - u_2(x, t) \) satisfies either initial boundary value problem with \( g(x, t) \equiv 0 \) and \( v(x, t) \equiv 0 \). From a variation of Green’s first identity we have (for a fixed \( \Omega \))
\[
\frac{d}{dt} \int_{\Omega} \frac{1}{2} w^2 \, dV = \int_{\Omega} w \frac{\partial w}{\partial t} \, dV = \int_{\Omega} w \nabla \cdot (k \nabla w) \, dV = \int_{\Omega} (\nabla \cdot (k w \nabla w) - k |\nabla w|^2) \, dV
\]
\[
= -\int_{\Omega} k |\nabla w|^2 \, dV + \oint_{\partial \Omega} k w \frac{\partial w}{\partial n} \, dS.
\]

\[
= \begin{cases} 
-\int_{\Omega} k |\nabla w|^2 \, dV & \text{for (10.39) or (10.40)} \\
-\int_{\Omega} k |\nabla w|^2 \, dV - \oint_{\partial \Omega} s w^2 \, dS & \text{for (10.41)}
\end{cases} < 0.
\]

The first option corresponds to a Dirichlet or Neumann boundary condition, the other to a Robin boundary condition. Hence
\[
0 \leq \int_{\Omega} w^2(x, t) \, dV \leq \int_{\Omega} w^2(x, 0) \, dV = 0,
\]
so \( w \equiv 0 \).

**10.5 Moving Boundaries; Stefan Problems**

Diffusion processes often involve a moving boundary. Typical examples are contact problems between two materials like two fluids or a solid material and a dissolvent, between two
phases of the same material like water and ice, and between wet and dry materials. In these problems the boundary is determined by some additional kinematic constraint. In Chapter 11 we shall consider one such problem in more detail. Here we will restrict ourselves to a class of problems referred to as Stefan problems [172]. One may typically think of ice of 0°C that is melting at a surface, the front, denoted by $S$. Apparently $S = S(t)$. In one dimension, its simplest form, the (dimensionless) temperature $u(x, t)$ of the water satisfies the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < S(t), \quad t > 0. \quad (10.43)$$

Before $t = 0$, the temperature is everywhere $u(x, t) = 0$, but at $t = 0$ the temperature at one end, $x = 0$, is suddenly increased to unity:

$$u(0, t) = 1, \quad t \geq 0. \quad (10.44a)$$

This enforces the input of energy because the temperature elsewhere is lower. At $x = S(t)$, the interface with the ice, we have a continuous temperature, so

$$u(S, t) = 0, \quad t > 0. \quad (10.44b)$$

As the position of the interface is unknown, we need an additional condition. This extra condition is found from the physics of the phase transition. The amount of specific latent heat released during the water-to-ice phase transition is again to be added when the ice melts. As the front travels with a certain speed, the produced heat travels with the same speed and is therefore equal to the heat flux. As we assumed that the ice is of constant temperature, i.e., $u_{ic} \equiv 0$, the flux into the ice vanishes, and we get (nondimensionally)

$$-\frac{\partial u}{\partial x}(S(t), t) = \alpha \frac{dS}{dt}, \quad S(0) = 0. \quad (10.45)$$

The constant $\alpha$ in (10.45) is called the Stefan constant. Let us try to solve problem (10.43–10.45) by introducing, following (10.34) to (10.37), the similarity solution

$$u(x, t) = x^m f(\eta), \quad \eta := \frac{x}{2 \sqrt{t}}. \quad (10.46)$$

From (10.44a) it follows that $m = 0$, while from (10.44b) we can write for some constant $\gamma$ that

$$S = 2 \gamma \sqrt{t}. \quad (10.47)$$

Equation (10.43) and boundary conditions turn into

$$f'' + 2\eta f' = 0, \quad f(0) = 1, \quad f(\gamma) = 0, \quad f'(\gamma) = -2\alpha \gamma, \quad (10.48)$$

along $0 \leq \eta \leq \gamma$, with solution

$$f(\eta) = 1 - \alpha \gamma^2 \sqrt{\pi} \operatorname{erf}(\eta), \quad (10.49)$$

while the unknown $\gamma$ is given by the real positive root of

$$\alpha \gamma^2 \sqrt{\pi} \operatorname{erf}(\gamma) = 1. \quad (10.50)$$
10.5. Moving Boundaries; Stefan Problems 247

Figure 10.2. Similarity coordinate $\gamma$ of the interface as a function of the Stefan problem parameter $\alpha$.

Altogether we thus have

$$u(x, t) = 1 - \frac{\text{erf}(x/2\sqrt{t})}{\text{erf}(\gamma)}.$$  (10.49)

We may solve (10.48) numerically to obtain $\gamma$ as a function of the Stefan parameter $\alpha$. See, e.g., Figure 10.2. For small $\gamma$ we can approximate $\gamma \approx (1 - \frac{1}{2} \ln \alpha)^{1/2}$. For large $\gamma$ we have a dominating exponential that produces for small $\alpha$ $\gamma \approx (1 - \frac{1}{2} \ln \alpha)^{1/2}$.

Example 10.11 This single-phase problem, where the temperature varies only in the water, is easily generalised to a double-phase problem, where the temperature varies in both the ice and the water. In this case we need the more general Stefan condition

$$\rho L \frac{d}{dt} S = \left[ -k_i \frac{\partial}{\partial x} T_i + k_w \frac{\partial}{\partial x} T_w \right]_{x=S(t)}$$  (10.50)

(now dimensional), where $L$ denotes the amount of specific latent heat released during the water-to-ice phase transition.

Consider a Stefan problem of melting ice in the semi-infinite domain $x \in [0, \infty)$. Let the phase change be at $x = S$. Let the temperature of the ice at $t = 0$ be given by $T_0 < 0$. It follows that the temperature for $x \to \infty$ also equals $T_0$. At time $t = 0$ the temperature at $x = 0$ is suddenly increased to $T_1 > 0$. We have

$$\alpha \frac{d}{dt} S = \left[ -k_w \frac{\partial}{\partial x} T_w + k_i \frac{\partial}{\partial x} T_i \right]_{x=S(t)}, \quad \left[ T_w(x, t) = T_{wa}(x, t) = 0 \right]_{x=S(t)}.$$  (10.51)

Inspired by the previous solution, we find solutions of the form

$$S = 2\gamma \sqrt{t}, \quad T_{wa} = T_1 \left(1 - \frac{\text{erf}(x/2\sqrt{k_w t})}{\text{erf}(\gamma_{wa})}\right), \quad T_i = T_0 \left(1 - \frac{\text{erfc}(x/2\sqrt{k_i t})}{\text{erfc}(\gamma_i)}\right).$$
Chapter 10. Analysis of Parabolic Equations

Figure 10.3. Enthalpy function $H(T)$.

where $\gamma_{wa} = \gamma / \sqrt{k_{wa}}$ and $\gamma_{ic} = \gamma / \sqrt{k_{ic}}$. The constant $\gamma$ is to be determined from the equation

$$\alpha \sqrt{\pi} = T_1 \exp(-\gamma_{wa}^2) + T_0 \exp(-\gamma_{ic}^2) \gamma_{wa} \text{erf}(\gamma_{wa}) \gamma_{ic} \text{erfc}(\gamma_{ic}).$$

From a physical point of view it is more sensible to use the enthalpy rather than the temperature as the dependent variable for the melting ice problem, as we have both “sensible” temperature and latent heat. The enthalpy, being the sum of both, is given as a function of the temperature. This then closes the equations. Let us denote this enthalpy by $H$. Then we have for given $H = H(T)$

$$\frac{\partial H}{\partial t} = \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < S(t), \quad t > 0,$$

$$(10.51a)$$

$$T(x, 0) = 0,$$

$$(10.51b)$$

$$T(0, t) - \beta \frac{\partial T}{\partial x}(0, t) = 1.$$

$$(10.51c)$$

To start with we know the temperature at $t = 0$, say, and so we know the initial enthalpy as well. Two typical graphs of $H(T)$ are given in Figure 10.3. One shows a simple discontinuous $H$, with a discontinuity at $T = 0$ between $H_{-}$ and $H_{+}$, whereas the other one shows an $H$ of a material with a “mushy region,” i.e., where the phase change is more gradually taking place (as happens in melting of alloys).

10.6 Long-Time Behaviour of Solutions

The solutions of parabolic problems are typically diffusive and smoothing steep gradients. As a result, any initial value tends to be “forgotten” and is therefore sometimes not as important as the long-time behaviour. We will consider some occasions where this is the case.
10.6. Long-Time Behaviour of Solutions

10.6.1 Linear Initial Boundary Value Problem

Parabolic equations are of special interest in studying phenomena that tend to a steady state situation or equilibrium. In fact, when considering the PDE

\[ \frac{\partial u}{\partial t} = \mathcal{L}[u], \]  

(10.52)

where \( \mathcal{L} \) is a second order differential operator, it is natural to ask for solutions for which

\[ \frac{\partial u}{\partial t} \to 0; \]

i.e., \( u \) will ultimately satisfy the stationary elliptic equation

\[ \mathcal{L}[u] = 0. \]

This shows that there is an intimate relation between some parabolic and elliptic PDEs. Indeed, one often even tries to solve a difficult elliptic boundary value problem numerically by embedding it in a corresponding parabolic problem like (10.52) and solving the latter by implicit time stepping. This false transient approach is the basis of quite a few numerical methods. As we saw in Example 10.7, this will not always work and the problems should be analyzed carefully. At least the elliptic problem should have a solution, and the solution of the associated parabolic problem should converge for large time to this solution.

We shall illustrate the long-time behaviour of solutions of parabolic problems by a simple linear one-dimensional case. Consider for \( U(x) \) the steady elliptic problem

\[ \mathcal{L}[U] := a \frac{\partial^2 U}{\partial x^2} + b \frac{\partial U}{\partial x} + c U = 0, \quad x \in (0, 1), \]  

(10.53a)

\[ U(0) = p, \quad U(1) = q, \]  

(10.53b)

where \( a > 0, b, c, p, \) and \( q \) are constants. Suppose that we try to approach this solution by the auxiliary parabolic problem

\[ \frac{\partial y}{\partial t} = a \frac{\partial^2 y}{\partial x^2} + b \frac{\partial y}{\partial x} + cy, \quad x \in (0, 1), \quad t > 0, \]  

(10.54a)

\[ y(x, 0) = v(x) + U(x), \quad x \in (0, 1), \]  

(10.54b)

\[ y(0, t) = p, \quad y(1, t) = q, \quad t > 0, \]  

(10.54c)

where \( v \) is an arbitrary, in general unknown, but reasonably smooth, initial error to the finally sought solution. We are interested in knowing under which conditions solution \( y \) asymptotically tends to \( U \). In other words, when does the difference \( u = y - U \), satisfying

\[ \frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial u}{\partial x} + cu, \quad x \in (0, 1), \quad t > 0, \]  

(10.55a)

\[ u(x, 0) = v(x), \quad x \in (0, 1), \]  

(10.55b)

\[ u(0, t) = u(1, t) = 0, \quad t > 0, \]  

(10.55c)

tend to zero for \( t \to \infty \)?
Chapter 10. Analysis of Parabolic Equations

After trying a transformation of the type $u(x, t) = e^{\alpha t - \beta x} \tilde{u}(x, \gamma t)$, we find that the transformation $u(x, t) = \exp(c - \frac{b^2}{4a})t - \frac{b}{2a}x] \tilde{u}(x, at)$ reduces the problem to the one in Example 10.8. So we deduce a Fourier series solution of (10.55) of the form

$$u(x, t) = e^{(c - \frac{b^2}{4a})t - \frac{b}{2a}x} \sum_{k=1}^{\infty} A_k e^{-\pi^2 k^2 at} \sin(k\pi x), \quad (10.56)$$

where the coefficients $A_k$ are found from the Fourier-sine series expansion

$$v(x) e^{\frac{b}{2a}x} = \sum_{k=1}^{\infty} A_k \sin(k\pi x) \quad (10.57)$$

in order to satisfy condition (10.55b). Although in general (10.57) is not uniformly convergent, the coefficients will at least decay like $A_k = O(k^{-1})$ as $k \to \infty$. So, with the help of the exponential, series (10.56) will converge uniformly for any $t > 0$, and indeed $u(0, t) = u(1, t) = 0$ (appendix, Section C).

We conclude that $u(x, t) \to 0$ for $t \to \infty$ if the least attenuated mode tends to zero, i.e., if $c < \frac{b^2}{4a} + \pi^2 a$. In this case solution $\gamma(x, t)$ will approach the stationary solution $U(x)$.

The corresponding problem with boundary conditions of Neumann type is more involved. Consider the initial boundary value problem

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial u}{\partial x} + cu, \quad x \in (0, 1), \quad t > 0, \quad (10.58a)$$

$$u(x, 0) = v(x), \quad x \in (0, 1), \quad (10.58b)$$

$$\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = 0, \quad t > 0. \quad (10.58c)$$

After a similar transformation to before and separation of variables, we can construct a Fourier-type series expansion

$$u(x, t) = A_0 e^{\alpha t} + e^{(c - \frac{b^2}{4a})t - \frac{b}{2a}x} \sum_{k=1}^{\infty} A_k e^{-\pi^2 k^2 at} f_k(x), \quad (10.59)$$

where $\{f_k(x)\}$ form the orthonormal set

$$f_k(x) = \left( \frac{2}{4a^2 \pi^2 k^2 + b^2} \right)^{\frac{1}{2}} (2a\pi k \cos k\pi x + b \sin k\pi x). \quad (10.60)$$

Except for $A_0$ (note that the solution is not unique), the coefficients $A_k$ are determined by

$$A_k = \int_0^1 (v(x) - A_0) e^{\frac{b}{2a}x} f_k(x) \, dx. \quad (10.61)$$

We conclude that the restrictions on $c$ for a decaying $u$ are slightly more stringent than in the previous case. Unless we are able to make sure that $A_0 = 0$, we must have $c < 0$ for $u$ to tend to zero for large $t$. 
10.6. Long-Time Behaviour of Solutions

Example 10.12 Let \( \Omega := [0, 1] \times [0, 1] \). Consider the initial boundary value problem

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \nabla^2 u + 1, \quad x \in \Omega, \quad t > 0 \\
u(x, y, 0) &= 0, \quad x \in \Omega, \\
u(x, y, t) &= 0, \quad x \in \partial \Omega, \quad t > 0.
\end{align*}
\]

We shall use the Duhamel integral (see Theorem 4.16). To this end we have to find a solution of

\[
\begin{align*}
\frac{\partial w}{\partial t}(x, t; \tau) &= \nabla^2 w(x, t; \tau), \quad x \in \Omega, \quad t > \tau, \\
w(x, \tau; \tau) &= 1, \quad x \in \Omega, \\
w(x, t; \tau) &= 0, \quad x \in \partial \Omega, \quad t > \tau.
\end{align*}
\]

Because of the boundary conditions it is convenient to have Fourier basis modes of the form \( \sin(j \pi x) \sin(k \pi y) \) only. We find

\[
1 = 16 \sum_{j, k \geq 0} \frac{\sin \alpha_j x}{\alpha_j} \frac{\sin \alpha_k y}{\alpha_k}, \quad \text{where} \quad \alpha_j := (2j + 1)\pi.
\]

Using, e.g., the dispersion relation, we find the (uniformly converging) expansion

\[
w(x, y, t; \tau) = 16 \sum_{j, k \geq 0} \frac{\sin \alpha_j x}{\alpha_j} \frac{\sin \alpha_k y}{\alpha_k} e^{-\left(\alpha_j^2 + \alpha_k^2\right)(t - \tau)}.
\]

Hence

\[
u(x, y, t) = 16 \sum_{j, k \geq 0} \frac{\sin(\alpha_j x) \sin(\alpha_k y)}{\alpha_j \alpha_k \alpha_j^2 + \alpha_k^2} \int_{\tau}^{t} e^{-\left(\alpha_j^2 + \alpha_k^2\right)(t - \tau)} d\tau.
\]

If \( t \to \infty \), we find

\[
u(x, y, t) \to 16 \sum_{j, k \geq 0} \frac{\sin(\alpha_j x) \sin(\alpha_k y)}{\alpha_j \alpha_k \alpha_j^2 + \alpha_k^2} =: v(x, y).
\]

By direct substitution and noting the uniform convergence, we may verify immediately that this is the solution of the Poisson problem

\[
\begin{align*}
\nabla^2 v &= 1, \quad x \in \Omega, \\
v &= 0, \quad x \in \partial \Omega.
\end{align*}
\]

10.6.2 Equilibrium and Travelling-Wave Solutions for Nonlinear Problems

Quite often diffusion equations arise in the analysis of chemical reactions or population dynamics or in the modeling of epidemics. Typically, the equation includes a nonlinear source or reaction term and takes on the form

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \nabla \cdot (D \nabla u) + r(u) \\
\end{align*}
\] (10.62)
The reaction term \( r \) arises from the application. For example, \( u \) may be a temperature and \( r \) the energy release from the chemical reaction; or \( u \) may be a concentration of some species and \( r \) a source or sink. A common model for \( r \) is the polynomial

\[
r(u) = \gamma (u - \alpha)(\beta - u),
\]

(10.63) where \( \alpha < \beta \) and \( \gamma > 0 \). If we further assume that \( D \) is a positive constant, we may rescale this equation by the transformation

\[
\tilde{u} := \frac{u - \alpha}{\beta - \alpha}, \quad \tilde{x} := \left( \frac{\gamma}{D(\beta - \alpha)} \right)^{1/2} x, \quad \tilde{t} := \gamma (\beta - \alpha) t
\]

(10.64)

(where we will omit the tilde henceforth) into the standard form

\[
\frac{\partial u}{\partial t} = \nabla^2 u + u(1 - u).
\]

(10.65)

This is known as Fisher’s equation. It was introduced originally to model the spread of a gene in a population [173].

To demonstrate the basic ideas we restrict our discussion of (10.62) to the one-dimensional case

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + r(u).
\]

(10.66)

Clearly, \( r(u) \) represents a source when it is positive and a sink when it is negative. Any zero of \( r \), say \( u = u_0 \), is evidently also a solution of the equation if the boundaries are compatible with this solution, e.g., if the walls are isolated such that no heat or matter is lost. These stationary solutions are called equilibrium solutions. It is not clear in advance if any such solution will be attained for \( t \to \infty \) if we start in its neighbourhood. This will be seen to be critically dependent on the sign of \( r'(u_0) \).

Let us therefore study the initial boundary value problem—with isolated boundaries and arbitrary initial condition—for a perturbation \( \delta e(x, t) \) of \( u = u_0 \) given by

\[
u(x, t) := u_0 + \delta e(x, t),
\]

(10.67)

where \( \delta \) is small and \( r'(u_0) \neq 0 \). After linearisation of \( r(u) = \delta r_0 e + \cdots \) (where we neglect any term smaller than \( O(\delta) \)), we have the following problem for \( e(x, t) \):

\[
\begin{align*}
\frac{\partial e}{\partial t} &= \frac{\partial^2 e}{\partial x^2} + r_0' e, & x \in (0, 1), & t > 0, \quad (10.68a) \\
e(x, 0) &= v(x), & x \in (0, 1), & (10.68b) \\
\frac{\partial e}{\partial x}(0, t) &= \frac{\partial e}{\partial x}(1, t) = 0, & t > 0. \quad (10.68c)
\end{align*}
\]

This has the same form as the problem we studied in (10.58). We can immediately infer that \( u = u_0 \) is an asymptotically stable equilibrium, i.e., \( e \to 0 \) as \( t \to \infty \), if \( r_0' < 0 \). Similarly,
it is asymptotically unstable if \( r'_0 > 0 \). In particular, for the Fisher equation (10.65) we obtain that \( u(x, t) \equiv 1 \) is stable, while \( u(x, t) \equiv 0 \) is unstable.

It should be noted that other equilibrium solutions than the zeros of \( r(u) \) are in principle also possible. They are, however, in general more difficult to analyse.

**Example 10.13** The stability of solutions is an important issue in many reaction-diffusion equations. A typical question is whether the solution becomes unbounded in finite time, i.e., "blow up" occurs. We shall give a simple example of such a problem. Consider the following PDE with initial and boundary conditions:

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u^3, \quad x \in (0, 1), \quad t > 0,
\]

\[
u(0, 0) = 1, \quad x \in (0, 1),
\]

\[
\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = 0, \quad t > 0.
\]

One easily verifies that a solution exists that is independent of \( x \) and hence is given by

\[
u(x, t) = \frac{1}{\sqrt{1 - 2t}}.
\]

It has the remarkable property that \( \nu \to \infty \) for \( t \to \frac{1}{2} \).

Reaction-diffusion equations allow for further analysis by looking for travelling-wave solutions (see Section 2.5). These may be considered as similarity solutions of a particular type (Section 7.3.3, Example 7.18). As an example we consider Fisher’s equation

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(1 - u), \quad x \in \mathbb{R}, \quad t > 0.
\]  
\[ (10.69) \]

We now try to find similarity solutions of the form

\[ U(\xi) := U(x/c - t) = u(x, t). \]

By substituting this in (10.69), we obtain the autonomous ODE

\[ U'' + c^2 U' + c^2 U(1 - U) = 0, \quad \xi \in \mathbb{R}, \]  
\[ (10.70) \]

where the prime denotes differentiation with respect to \( \xi \). The physically interesting solutions are those that remain finite for \( \xi \to \pm \infty \). Note that \( U \) depends on \( c^2 \), so the behaviour for left- and right-running waves is the same.

To facilitate the analysis it is useful to rewrite (10.70) as a first order system, i.e., with \( y_1 := U \) and \( y_2 := U' \):

\[
y'_1 = y_2,
\]

\[
y'_2 = -c^2 y_2 - c^2 y_1 (1 - y_1).
\]  
\[ (10.71) \]

Insight into the behaviour of possible solutions is obtained by using the phase plane (see, e.g., [97]), i.e., graphs (trajectories) of \( y_2 \) as a function of \( y_1 \); see Figure 10.4. Of particular importance are the stationary points \((0, 0)^T\) and \((1, 0)^T\). The Jacobi matrix of the system...
(10.71), linearised around \((0, 0)^T\), is given by
\[
\begin{pmatrix}
0 & 1 \\
-c^2 & -c^2
\end{pmatrix}
\]
The corresponding eigenvalues
\[
\lambda_{1,2} = -\frac{1}{2} c^2 \pm \frac{1}{2} \sqrt{c^4 - 4c^2}
\]
are negative if \(|c| \geq 2\) and complex with a negative real part otherwise. Hence the point \((0, 0)^T\) is a stable spiral point for \(|c| < 2\) and a stable node for \(|c| \geq 2\). At the other stationary point \((1, 0)^T\) one finds the Jacobi matrix
\[
\begin{pmatrix}
0 & 1 \\
c^2 & -c^2
\end{pmatrix}
\]
with the real eigenvalues
\[
\lambda_{1,2} = -\frac{1}{2} c^2 \pm \frac{1}{2} \sqrt{c^4 + 4c^2},
\]
one of which is always negative and the other always positive. So this point is an unstable saddle point (see Figure 10.4). Note that conclusions for both \((0, 0)^T\) and \((1, 0)^T\) are entirely in agreement with the results of (10.68).

Most of the trajectories are physically not interesting because they tend to infinity for \(\xi \to \infty\) or \(-\infty\). For each \(c\) there is one trajectory, however, that starts at \((1, 0)^T\) and ends at \((0, 0)^T\). This trajectory (see Figure 10.4) corresponds to the solution we are looking for. This solution may be understood as a transition of the system from the unstable stationary state \(u(x, t) \equiv 0\) to the stable stationary state \(u(x, t) \equiv 1\).

For models where \(u\) cannot be negative we have to assume additionally that \((0, 0)^T\) is not a spiral point; i.e., \(|c| \geq 2\). In Figure 10.5 we have depicted the right-running wave for
the value \( c = 2.25 \), normalized by \( U(0) = \frac{1}{2} \). The wave’s activity is exponentially localized around \( x - ct = 0 \).

### 10.7 Discussion

- Parabolic equations describe phenomena like conduction and diffusion. They are diffusive in nature; i.e., an initial profile is spread out rather than propagated in a certain direction. Conduction typically occurs in heat transfer, whereas diffusion refers to the “spreading” of a material, often in a chemistry context.

- Since solutions of elliptic equations can be seen as a steady state of solutions of parabolic equations, there are a number of correspondences between the two classes. An interesting one is that there is also a maximum principle for parabolic equations. The theory and statements follow similar lines as those outlined in Section 8.7. We may refer to [115].

- There exists an extensive literature on nonlinear problems. They occur in a multitude of problems, like population dynamics, chemical reactions, etc. In particular, the study of solitary waves has brought many interesting phenomena to light. For further reading we recommend [35].

- In Chapter 16 we will encounter parabolic problems in a number of practical situations. An example is that of thermal explosion in vessels in Section 16.3, which deals with chemical species that may spontaneously ignite at a certain temperature. An example of a Stefan problem is given in Section 16.12, where the melting of a solid by a laser beam is studied.
Exercises

10.1. Consider the initial boundary value problem
\[ \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - cu + \sin \pi x + \frac{1}{8} \sin 3\pi x, \quad x \in (0, 1), \quad t > 0, \]
\[ u(x, 0) = 0, \quad x \in (0, 1), \]
\[ u(0, t) = u(1, t) = 0, \quad t > 0. \]
Find its solution \( u(x, t) \).

10.2. Consider the initial boundary value problem
\[ \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0, \]
\[ u(x, 0) = v(x), \quad x \in (0, 1), \]
\[ u(0, t) = u(1, t) = 0, \quad t > 0. \]
(a) Find the Green’s function \( w(x, t; \tau) \) such that
\[ u(x, t) = \int_0^t w(x, t; \tau)v(\tau) \, d\tau. \]
(b) Determine \( u(x, t) \) for \( v(x) = \sin^2(\pi x) \).

10.3. Consider the problem of Exercise 10.2, now with boundary conditions
\[ \frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = 0, \quad t > 0. \]
(a) Compute the solution \( u(x, t) \).
(b) Determine \( \lim_{t \to \infty} u(x, t) \) both from the expression under (a) and by direct computation.

10.4. Solve the initial boundary value problem
\[ \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - cu + \frac{1}{2}(1 - \cos \pi x), \quad x \in (0, 1), \quad t > 0, \]
\[ u(x, 0) = 2(1 - \cos \pi x), \quad x \in (0, 1), \]
\[ \frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = 0, \quad t > 0. \]

10.5. Solve the initial boundary value problem
\[ \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (-1, 1), \quad t > 0, \]
\[ u(x, 0) = 1, \quad x \in (-1, 1), \]
\[ u(-1, t) = u(1, t) = 0, \quad t > 0. \]
10.6. A sphere with radius \( R \) has a constant initial temperature \( u_0 \). At \( t = 0 \) the sphere is cooled at \( u = 0 \).

(a) Show that the temperature \( u(r, t) \) satisfies the initial boundary value problem

\[
\frac{\partial u}{\partial t} = k \left( \frac{\partial^2 u}{\partial r^2} + \frac{2 \partial u}{r \partial r} \right), \quad r \in (0, R), \quad t > 0,
\]

\[
u(r, 0) = u_0, \quad r \in (0, R),
\]

\[
\frac{\partial u}{\partial r} (0, t) = u(R, t) = 0, \quad t > 0.
\]

(b) Determine \( u(r, t) \).

10.7. Use Green’s functions to solve the following problem, defined for \( 0 < x < L \) (for some \( L \)) and \( t > 0 \):

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + g(x, t), \quad x \in (0, L), \quad t > 0,
\]

\[
u(x, 0) = v(x), \quad x \in (0, L),
\]

\[
u(0, t) = 0, \quad \frac{\partial u}{\partial x} (L, t) + u(L, t) = 0, \quad t > 0.
\]

10.8. Consider the following initial boundary value problem for \( x > 0 \), \( t > 0 \) for some parameter \( \alpha > 0 \):

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, \infty), \quad t > 0,
\]

\[
u(x, 0) = 0, \quad x \in (0, \infty),
\]

\[
u(0, t) = t^\alpha, \quad t > 0.
\]

Find a similarity solution \( \hat{u}(\eta) \) of the form

\[
u(x, t) = t^\alpha \hat{u}(\eta) \]

with \( \eta = x/\sqrt{t} \).

10.9. Find the solution of the following initial boundary value problem with periodic boundary conditions:

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (-1, 1), \quad t > 0,
\]

\[
u(x, 0) = v(x), \quad x \in (-1, 1),
\]

\[
u(-1, t) = u(1, t), \quad \frac{\partial u}{\partial x} (-1, t) = \frac{\partial u}{\partial x} (1, t), \quad t > 0.
\]

10.10. Find a similarity solution, for \( x > 0 \), \( t > 0 \), of

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + 1, \quad x \in (0, \infty), \quad t > 0,
\]

\[
u(x, 0) = 0, \quad x \in (0, \infty),
\]

\[
u(0, t) = 0, \quad t > 0.
\]
10.11. Let $0 < a < \frac{1}{2}$.

(a) Find an ODE for the travelling-wave solution $\hat{u}(\xi) := \hat{u}(x - bt)$ of

$$\frac{\partial \hat{u}}{\partial t} = \frac{\partial^2 \hat{u}}{\partial \xi^2} - \hat{u}(1 - \hat{u})(a - \hat{u}).$$

(b) Draw a phase plane for $\hat{u}$ and investigate when a solution exists subject to the conditions

$$\hat{u}(\xi) \to 1 \text{ for } \xi \to \infty, \quad \hat{u}(\xi) \to a \text{ for } \xi \to -\infty.$$

10.12. Use separation of variables to compute the solution of

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0,$$

$$u(x, 0) = \sin(\pi x) \cos(\pi x), \quad x \in (0, 1),$$

$$u(0, t) = u(1, t) = 0, \quad t > 0.$$

10.13. Use separation of variables to compute the solution of

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0,$$

$$u(x, 0) = \sin(\pi x) \cos(\pi x), \quad x \in (0, 1),$$

$$u(0, t) = 0, \quad \frac{\partial u}{\partial x}(1, t) = 0, \quad t > 0.$$

10.14. Consider the initial boundary value problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \alpha u, \quad x \in (0, 1), \quad t > 0,$$

$$u(x, 0) = v(x), \quad x \in (0, 1),$$

$$u(0, t) = u(1, t) = 0, \quad t > 0.$$

(a) Determine the formal solution.

(b) Show that this solution is stable for $\alpha < \frac{\pi}{2}$.

10.15. Consider the problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \alpha u, \quad x \in (0, 1), \quad t > 0,$$

$$u(x, 0) = 0, \quad x \in (0, 1),$$

$$u(0, t) = u(1, t) = 0, \quad t > 0.$$

(a) Determine the stationary solution, i.e., the solution for $t \to \infty$.

(b) Use the result in part (a) to compute the solution.
Chapter 11

Numerical Methods for Parabolic Equations

This chapter gives an overview of a few numerical techniques for solving parabolic problems, as well as their analysis for accuracy and stability. It starts with an introduction of the simple explicit Euler scheme in Section 11.1. A matrix-vector formulation appears to enhance the analysis of this scheme a great deal. This analysis is carried out in Section 11.2, where semidiscretisations, both of the spatial derivatives and of the time derivative, are considered. The former results in a vector ODE. Due to the inherent stiffness, it makes sense to consider implicit schemes, which is done in Section 11.3. Both the implicit Euler scheme and a weighted average of the Euler schemes (the $\theta$ method) are analysed. In Section 11.4 the matrix method, based on the semidiscretisation with respect to the spatial derivatives, is used to analyse the stability and accuracy of the one-step schemes of the earlier sections. A case where the $\theta$ scheme may fail is for an initial value problem with discontinuous data. This is investigated in detail in Section 11.5. Mixed boundary conditions require a special treatment as they may have an impact on stability; they are studied in Section 11.6. Problems in two space dimensions are briefly introduced in Section 11.7. It is shown that the explicit Euler scheme has a similar time step restriction as in the one-dimensional case. Implicit methods, however, involve the solution of linear systems that may lose their sparseness in direct methods (LU-decomposition). Hence we derive in Section 11.8 the alternating direction implicit (ADI) method, which has both favorable stability properties and good efficiency. We also present some mixed explicit/implicit methods for the advection-diffusion equation in this section. In Section 11.9 we turn to nonlinear parabolic equations. We discuss a number of options for a variety of nonlinear terms. Finally, in Section 11.10, we treat some numerical questions related to the Stefan problem, which is a parabolic problem with moving boundaries.

11.1 The Explicit Euler Scheme

In Chapter 10 we considered various ways to find an analytic solution of a parabolic PDE. Sometimes this is the method of choice, but most often the analytical expressions are less useful for actually finding numerical values. As an example, a series solution might converge very slowly, or an integral expression might still require a nontrivial numerical quadrature
procedure. The major advantage of analytical methods is to obtain qualitative insight. If more precise (numerical) information is needed, one has to resort to numerical methods. To start with we consider the heat equation in one space variable. In fact, most of the ideas and methods can be applied to more complex problems, yet this fairly simple case demonstrates most of the algorithmic and analysis aspects of parabolic problems in general. Higher-dimensional as well as nonlinear cases are specifically treated in later sections. So let us study the problem

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad x \in (0, 1), \quad t > 0, \quad (11.1a)
\]

\[
u(x, 0) = \psi(x), \quad x \in (0, 1), \quad (11.1b)
\]

\[
u(0, t) = g_l(t), \quad t > 0, \quad (11.1c)
\]

\[
u(1, t) = g_r(t), \quad t > 0. \quad (11.1d)
\]

For the source term \(f(x, t) \equiv 0\), (11.1a) is often called the heat equation.

Recalling what we did in Chapter 5, we now introduce a grid for the domain \(\Omega := \{(x, t) \in \mathbb{R}^2 \mid 0 < x < 1, \ t > 0\}\); see Figure 11.1. If \(\Delta x\) denotes the spatial grid size and \(\Delta t\) the time step, the grid points \((x_j, t^n)\) can be chosen as

\[
x_j := j\Delta x \quad (j = 0, 1, 2, \ldots, M + 1), \quad \Delta x = 1/(M + 1), \quad (11.2a)
\]

\[
t^n := n\Delta t \quad (n = 0, 1, 2, \ldots). \quad (11.2b)
\]

We discretise (11.1a) by replacing the time derivative \(\frac{\partial u}{\partial t}(x_j, t^n)\) by the forward difference approximation defined in (5.28c) and the space derivative \(\frac{\partial^2 u}{\partial x^2}(x_j, t^n)\) by the central difference approximation given in (5.30). Writing \(u^n_j\) for the numerical approximation of \(u(x_j, t^n)\) and defining \(f^n_j := f(x_j, t^n)\), we obtain the difference scheme

\[
\frac{1}{\Delta t}(u^n_{j+1} - u^n_j) = \frac{1}{\Delta x^2}(u^n_{j+1} - 2u^n_j + u^n_{j-1}) + f^n_j, \quad (11.3)
\]

\[\text{Figure 11.1. Stencil for the explicit scheme (11.4).}\]
11.1. The Explicit Euler Scheme

which holds at all internal grid points \((x_j, t^n)\). The scheme (11.3) can simply be rewritten as

\[
\begin{align*}
u_{j}^{n+1} &= \nu_{j}^{n} + d \left( \nu_{j+1}^{n} - 2\nu_{j}^{n} + \nu_{j-1}^{n} \right) + \Delta t f_{j}^{n},
\end{align*}
\]  

(11.4)

where \(d\) is defined by

\[
d := \frac{\Delta t}{\Delta x^2}.
\]  

(11.5)

The coefficient \(d\) is sometimes called the diffusion number. The dependence of the numerical value \(v_{j+1}^{n+1}\) at the new time level \(t^{n+1}\) on the numerical values \(v_{j}^{n} (k = j \pm 1, j)\) at the old time level \(t^{n}\) is depicted in the stencil in Figure 11.1 (black dots). Clearly, each value at time level \(t^{n+1}\) can be computed independently from the other values at the same time level; i.e., the difference scheme (11.4) is explicit; it is often referred to as the explicit Euler scheme.

The initial condition (11.1b) and the boundary conditions (11.1c), (11.1d) also have to be prescribed for the numerical solution. For the initial boundary value problem (11.1) this is trivial and we find

\[
u_{j}^{0} = \nu(x_j) \quad (j = 1, 2, \ldots, M)
\]  

(11.6)

for the initial condition and

\[
u_{n}^{0} = g_{\ell}(t^{n}), \quad \nu_{M+1}^{n} = g_{r}(t^{n}) \quad (n = 0, 1, 2, \ldots)
\]  

(11.7)

for the boundary conditions.

**Example 11.1** In this example we compute the numerical solution of the initial boundary value problem (11.1) for the initial function

\[
u(x) = 4x(1 - x), \quad 0 < x < 1,
\]

and homogeneous boundary conditions (i.e., \(g_{\ell}(t), g_{r}(t) \equiv 0\)). Furthermore, we take \(f(x, t) \equiv 0\).

We have employed the difference scheme (11.4) with grid size \(\Delta x = 5 \times 10^{-2}\) and time steps \(\Delta t = 1.2 \times 10^{-3}\) and \(\Delta t = 1.3 \times 10^{-3}\). The resulting numerical solutions are shown in Figure 11.2. In the left column the solution is computed with \(\Delta t = 1.2 \times 10^{-3}\), which is a good approximation of the exact solution. However, the column on the right shows the numerical solution for \(\Delta t = 1.3 \times 10^{-3}\), which is obviously incorrect. When the number of time steps increases further, the numerical scheme eventually results in floating-point overflow. This example shows that the difference scheme (11.4) has to be applied with care.

We can investigate the stability of this scheme in various ways. The simplest way is often to resort to Fourier analysis; see Chapter 5. As we saw there, such an analysis can detect when a scheme may be unstable (it rather gives necessary conditions for stability). The prerequisite is that the boundary conditions be periodic (like in the example), which we therefore assume. Consider the homogeneous scheme (11.4); i.e., we take \(f_{j}^{n} \equiv 0\). We look for a numerical solution of the form

\[
u_{j}^{n} = e^{\iota x \Delta t},
\]  

(11.8)

where \(\kappa\) is the wave number and \(\omega\) is the frequency; see Section 3.4. Suppressing the dependencies on wave number and frequency, we use two other variables, i.e., the amplification factor \(\lambda\) and phase angle \(\phi\), which are defined as

\[
\lambda := e^{-\iota \omega \Delta t}, \quad \phi := \kappa \Delta x.
\]  

(11.9)
This results in the representation

\[ u^n_j = \lambda^n e^{ix_j} . \]  

(11.10)

Substituting (11.10) in the homogeneous scheme and using the relations

\[ u^{n+1}_j = \lambda u^n_j, \quad u^n_{j \pm 1} = e \pm i\varphi u^n_j, \]  

(11.11)

we find the amplification factor

\[ \lambda(\varphi) = 1 + d(e^{i\varphi} - 2 + e^{-i\varphi}) = 1 - 4d \sin^2 \left( \frac{1}{2} \varphi \right) . \]  

(11.12)
11.1. The Explicit Euler Scheme

Thus at each time step the Fourier mode $e^{i(x,t)}$ is amplified by the factor $\lambda(\varphi)$ in (11.12). The undesirable oscillations in the numerical solution in Figure 11.2 can be prevented if we impose the condition

$$|\lambda(\varphi)| \leq 1 \quad \text{for} \quad 0 \leq \varphi \leq \pi. \quad (11.13)$$

It is clear that this condition is satisfied if

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}. \quad (11.14)$$

The inequality in (11.14) is called the stability condition of the numerical scheme (11.4). For the grid size $\Delta x = 5 \times 10^{-2}$ in Example 11.1 the stability condition (11.14) gives the time step restriction $\Delta t \leq 1.25 \times 10^{-3}$. Hence the numerical solution in the left column in Figure 11.2 is stable and the numerical solution in the right column is not.

The next question is the accuracy of the method, which is related to the choice of both $\Delta x$ and $\Delta t$. In (5.25) the local discretisation error was defined as the residual, $d^n_j$, found from substituting the exact solution $u(x,t)$ of (11.1a) in the difference scheme (11.3); the latter should be scaled such that it has the same dimensions as (11.1a). We thus find

$$d^n_j := \frac{1}{\Delta t} (u(x_j, t^n + 1) - u(x_j, t^n)) - \frac{1}{\Delta x^2} (u(x_{j+1}, t^n) - 2u(x_j, t^n) + u(x_{j-1}, t^n)) - f^n_j. \quad (11.15)$$

Inserting Taylor series expansions of $u(x,t)$ about $(x_j, t^n)$ and assuming that $u(x,t)$ is sufficiently smooth, we obtain

$$d^n_j = \frac{\partial u}{\partial t}(x_j, t^n) - \frac{\partial^2 u}{\partial x^2}(x_j, t^n) - f^n_j$$

$$+ \frac{1}{2} \Delta t \frac{\partial^4 u}{\partial t^2}(x_j, t^n) - \frac{1}{12} \Delta x^2 \frac{\partial^4 u}{\partial x^4}(x_j, t^n) + O(\Delta t^2) + O(\Delta x^4)$$

$$= \frac{1}{2} \Delta t \frac{\partial^2 u}{\partial t^2}(x_j, t^n) - \frac{1}{12} \Delta x^2 \frac{\partial^4 u}{\partial x^4}(x_j, t^n) + O(\Delta t^2) + O(\Delta x^4), \quad (11.16)$$

which is in agreement with the discretisation errors in (5.29) and (5.31). From (11.16) we conclude that the local discretisation error $d^n_j$ approaches zero for $\Delta t, \Delta x \to 0$ with $n \Delta t, j \Delta x$ fixed; i.e., the scheme is consistent; cf. Definition 5.13.

For this scheme there is a simple way to estimate the global discretisation error $e^n_j$ (cf. (5.103)), defined by

$$e^n_j := u(x_j, t^n) - u^n_j. \quad (11.17)$$

Subtracting (11.3) from (11.15) and multiplying by $\Delta t$, we find the following recurrence relation for $e^n_j$:

$$e^{n+1}_j = e^n_j + d(e^n_{j+1} - 2e^n_j + e^n_{j-1}) + \Delta td^n_j. \quad (11.18)$$

Furthermore, from the boundary conditions (11.7) we see that

$$e^{n+1}_0 = e^{n+1}_{M+1} = 0. \quad (11.19)$$

Note that the homogeneous part of the recurrence relation (11.18) is exactly equal to the homogeneous part of the numerical scheme (11.4). Consequently, the time step restriction
and the expression (11.16) for the local discretisation error \( d^n_j \) imply that both the numerical solution \( u^n_j \) and the global discretisation error \( e^n_j \) remain bounded for increasing \( n \) and fixed \( \Delta t \).

To estimate this global discretisation error \( e^{n+1}_j \), we consider its maximum value over all indices \( j \) and apply the triangle inequality in (11.18) to find

\[
\max_{1 \leq j \leq M} |e^{n+1}_j| \leq |(1 - 2d)e^n_j + d(e^n_{j+1} + e^n_{j-1})| + \Delta t |d^n_j| = (1 - 2d)\max_{1 \leq j \leq M} |e^n_j| + d \max_{1 \leq j \leq M} \left( |e^n_{j+1}| + |e^n_{j-1}| \right) + \Delta t \max_{1 \leq j \leq M} |d^n_j|.
\]  

(11.20)

The latter inequality can also be written as

\[
\|e^{n+1}\|_\infty \leq \|e^n\|_\infty + \Delta t \|d^n\|_\infty,
\]  

(11.22)

where the vectors \( d \) and \( e \) are defined by

\[
d^n := \begin{pmatrix} d^n_1 \\ d^n_2 \\ \vdots \\ d^n_M \end{pmatrix}, \quad e^n := \begin{pmatrix} e^n_1 \\ e^n_2 \\ \vdots \\ e^n_M \end{pmatrix}.
\]  

(11.23)

For the definition of the vector norm \( \| \cdot \|_\infty \) the reader is referred to the appendix, Section G.

Let us assume that the initial condition (11.6) is exact, so that \( \|e^0\|_\infty = 0 \).

(11.24)

From the inequality (11.22) and the initial error in (11.24), we can deduce (and show by induction) the following estimate for \( e^n \):

\[
\|e^n\|_\infty \leq \Delta t \sum_{l=0}^{n-1} \|d^l\|_\infty.
\]  

(11.25)

Hence for the special case that \( f(x, t) \equiv 0 \) and the derivative \( \frac{\partial^4 u}{\partial x^4} \) in the local error is bounded, say

\[
\frac{\partial^4 u}{\partial x^4}(x, t) \leq C \quad \text{for all} \quad (x, t) \in (0, 1) \times (0, \infty), \quad C > 0,
\]  

(11.26)

we find from (11.25) and (11.16) the following upper bound for \( \|e^n\|_\infty \):

\[
\|e^n\|_\infty \leq \frac{1}{2} C \left| d - \frac{1}{6} \right| t^n \Delta x^2 + O(\Delta t^2) + O(\Delta x^4).
\]  

(11.27)

From the above inequality it is clear that \( \|e^n\|_\infty \to 0 \) if \( \Delta t, \Delta x \to 0 \), while \( t^n = n\Delta t \leq T, T < \infty \) fixed. In that case we deduce that the numerical scheme (11.4) is convergent; see Definition 5.17.
11.2 Semidiscretisation

In the previous section we showed how to discretise the heat equation by the simple explicit scheme (11.4). It will turn out to be helpful to use a matrix-vector formulation, both for a more direct and simple description of methods and also for analysis purposes. As was pointed out in Section 5.5.1, we can discretise either the spatial variable or the time first. In this section we consider both types of semidiscretisation.

11.2.1 The Longitudinal Method of Lines

In this subsection we consider the case where the spatial variable is discretised first. This gives rise to the (longitudinal) method of lines (MOL). For this we use a finite number of grid points, fixed for all \( t \), where the second order derivative is discretised. So on the grid lines \( x = x_j (j = 1, 2, \ldots, M) \) in the interior domain we define vector functions

\[
\mathbf{u}(t) := \begin{pmatrix}
u(x_1, t) \\
u(x_2, t) \\ \vdots \\
u(x_{M-1}, t) \\
u(x_M, t)
\end{pmatrix}.
\]  

(11.28)

If we use a central difference approximation for \( \frac{\partial^2}{\partial x^2} u \) in (11.1a), we obtain a set of ODEs (cf. (11.3))

\[
\frac{d}{dt} u_j = \frac{1}{\Delta x^2} (u_{j+1} - 2u_j + u_{j-1}) + f_j(t) \quad (j = 1, 2, \ldots, M),
\]  

(11.29)

where \( f_j(t) := f(x_j, t) \). The boundary conditions in (11.1c) and (11.1d) plus the source term give a discrete source term \( f \), say. Hence we find the ODE system

\[
\frac{d\mathbf{u}}{dt} = A\mathbf{u} + f,
\]  

(11.30a)

where the discretisation matrix \( A \) is defined by

\[
A := \frac{1}{\Delta x^2} \begin{pmatrix}
-2 & 1 \\
1 & -2 & 1 \\
& \ddots & \ddots \\
& & 1 & -2 & 1 \\
& & & 1 & -2
\end{pmatrix}
\]  

(11.30b)
and where

\[
\mathbf{f}(t) := \begin{pmatrix}
    f(x_1, t) + \frac{1}{\Delta x^2} g_1(t) \\
    f(x_2, t) \\
    \vdots \\
    f(x_{M-1}, t) \\
    f(x_M, t) + \frac{1}{\Delta x^2} g_M(t)
\end{pmatrix}.
\] (11.30c)

Clearly, the initial condition for (11.30) is given by

\[
\mathbf{u}(0) = \begin{pmatrix}
    v(x_1) \\
    v(x_2) \\
    \vdots \\
    v(x_{M-1}) \\
    v(x_M)
\end{pmatrix}.
\] (11.30d)

In order to analyse the behaviour of the solutions of (11.30a), we can try to find the various possible modes. In this relatively simple case this can be done indeed. This approach is meant to analyse the problem, and not as an option to be carried out when solving problems generally. The point is that the tridiagonal matrix \( A \) can simply be seen to have the eigenvalues \( \lambda_k \), where

\[
\lambda_k := -\frac{4}{\Delta x^2} \sin^2 \left( \frac{1}{2} \pi x_k \right) \quad (k = 1, 2, \ldots, M). \] (11.31)

The corresponding eigenvectors are given by

\[
\mathbf{w}_k := \sqrt{\frac{2}{M}} \begin{pmatrix}
    \sin(\pi x_k) \\
    \sin(2\pi x_k) \\
    \vdots \\
    \sin((M-1)\pi x_k) \\
    \sin(M\pi x_k)
\end{pmatrix} \quad (k = 1, 2, \ldots, M). \] (11.32)

The vectors \( \mathbf{w}_k \) form a complete basis; i.e., the matrix

\[
\mathbf{W} := (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M)
\] (11.33)

is nonsingular; in fact, it is even an orthogonal matrix, implying that

\[
\mathbf{W}^{-1} \mathbf{A} \mathbf{W} = \mathbf{A}, \] (11.34)
11.2. Semidiscretisation

where

\[ \Lambda := \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{M-1} \\ \lambda_M \end{pmatrix}. \]  

(11.35)

Using this matrix to transform the variable \( u(t) \) as

\[ \tilde{u}(t) := W^{-1}u(t), \]

and likewise \( \tilde{f}(t) := W^{-1}f(t) \), we can derive the following decoupled system for (11.30a):

\[ \frac{d\tilde{u}}{dt} = \Lambda \tilde{u} + \tilde{f}. \]  

(11.36)

Hence we can conclude that the stability properties of the system (11.30a) are determined by the scalar equations

\[ \frac{d\tilde{u}_k}{dt} = \lambda_k \tilde{u}_k + \tilde{f}_k, \quad (k = 1, 2, \ldots, M). \]  

(11.37)

For \( k = 1 \) we find the eigenvalue

\[ \lambda_1 = -\frac{4}{\Delta x^2} \sin^2 \left( \frac{1}{2} \pi x_1 \right) \approx -\pi^2, \]

whereas for \( k = M \) we have

\[ \lambda_M = -\frac{4}{\Delta x^2} \sin^2 \left( \frac{1}{2} \pi x_M \right) = -4(M + 1)^2 \sin^2 \left( \frac{\pi M}{2(M + 1)} \right) \approx -4(M + 1)^2. \]

We thus conclude that (11.37) is stiff for larger values of \( M \); see Section 5.4.2. In fact, there is an entire sequence of small time scales, from very small (i.e., \( (4(M + 1)^2)^{-1} \)), gradually increasing to approximately one.

If we want to approximate the solution only moderately accurately by the explicit Euler method, we are faced with time step constraints, i.e., once we have passed the various boundary layers, or, in absence of these, even right from the start. Indeed, explicit Euler would give

\[ u^{n+1} = (I + \Delta t A)u^n + \Delta t f^n, \]  

(11.38)

where \( u^n \) is the numerical approximation of \( u(t^n) \) and \( f^n = f(t^n) \). Stability is guaranteed if

\[ \| I + \Delta t A \| \leq 1. \]

If we take the 2-norm (\( \| \cdot \|_2 \)), then the latter requirement is satisfied if

\[ \| I + \Delta t A \|_2 \leq 1. \]
To see this one should use the fact that \( \|Q_1BQ_2\|_2 = \|B\|_2 \) for any matrix \( B \) and orthogonal matrices \( Q_1 \) and \( Q_2 \). Hence (11.38) is stable if
\[
|1 + \Delta t \lambda_k| \leq 1 \quad (k = 1, 2, \ldots, M).
\]
This leads to the same result as obtained from the Fourier mode analysis in the previous section, i.e., \( d \leq \frac{1}{2} \).

It should be noted that the generalisation of this analysis to other problems is obvious. In general the resulting matrix is no longer constant or symmetric. This complicates the error analysis, but not in a fundamental way (e.g., one may need a more subtle use of norms). Of course, for nonlinear problems we obtain a nonlinear vector equation for which we cannot directly apply the matrix method as outlined above.

### 11.2.2 The Transversal MOL

If we discretise the time derivative first, we see that we have a method that differs from the previous type of semidiscretisation only if this discretisation is implicit. In this section we restrict ourselves to the implicit Euler method. So let us denote the numerical approximation of \( u(x,t_n) \) by \( u^n(x) \), which is defined on the grid lines \( t = t^n (n = 0, 1, 2, \ldots) \). Then we obtain from (11.1a) the following equation defined along these transversal lines:
\[
u^{n+1} - \Delta t \frac{d^2}{dx^2}u^{n+1} = u^n + \Delta t f(x,t^{n+1}).
\]
(11.39a)

Equation (11.39a) is subject to the boundary conditions (cf. (11.1c) and (11.1d))
\[
u^{n+1}(0) = g_l(t^{n+1}), \quad u^{n+1}(1) = g_r(t^{n+1}).
\]
(11.39b)

Since \( u(x,0) \) is given we can find the solution \( u \) of the PDE by successively solving a series of boundary value problems (11.39). The real advantage to discretising the spatial derivative first is the greater flexibility with respect to the choice of the spatial variables at each step. Once we have found \( u^n \) on some grid, we can use interpolation to obtain \( u^n \) at any other desired point in \((0,1)\). In particular, for problems that require a finer grid in an area that is moving in time, this can be of great advantage. Indeed, a fixed spatial grid would require a dense grid on a larger subdomain and thus would contain many redundant points at each time level (i.e., that could be omitted without affecting the global accuracy).

**Example 11.2** An example of a problem where we have a sharp front moving in the domain \((-1,1)\) back and forth if time evolves is given by the (slightly contrived but typical) problem
\[
u = \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \alpha(2au - \cos t)(1 - u^2),
\]
which has the solution
\[
u(x,t) = \tanh(\alpha(x - \sin t))
\]
for suitable initial and boundary conditions. In Figure 11.3 we have drawn \( u(\cdot, t) \) for \( \alpha = 50 \) at three different time levels. It is clear that it has a sharp front, requiring a fine spatial grid, but at different positions for different time points.
11.3 Implicit Schemes

The scheme (11.4) turned out to be first order accurate in the time step $\Delta t$ and second order accurate in the spatial grid size $\Delta x$, as expected. The severe time step restriction (11.14), due to the inherent stiffness (as found in the previous section), is a major drawback indeed. For instance, if we compute the numerical solution of the initial boundary value problem (11.1) on a grid with grid size $\Delta x = 10^{-2}$ at time level say $t^n = 1$, then the number of required time steps will be at least 20,000! In general, the number of time steps is prohibitively large for this scheme. From what we have seen in Chapter 5 we are therefore better to take recourse to implicit schemes.

11.3.1 The Implicit Euler Scheme

Rather than using forward differences for the time discretisation, let us now use the backward difference approximation defined in (5.71) to approximate the derivative $\frac{\partial}{\partial t} u(x_j, t^n)$. If we again use the central difference approximation (5.30) for the space derivative $\frac{\partial^2}{\partial x^2} u(x_j, t^n)$, we get the difference equation

$$\frac{1}{\Delta t} (u_{j+1}^{n+1} - u_j^n) = \frac{1}{\Delta x^2} (u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) + f_{j+1}^{n+1}$$

(11.40)

for the numerical solution $u_j^n$. This equation can be rewritten as

$$u_j^{n+1} - d(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) = u_j^n + \Delta t f_j^{n+1}.$$  

(11.41)
The stencil of the numerical scheme (11.41), which is called the implicit Euler scheme, is given in Figure 11.4. From this stencil it is obvious that the numerical value \( u_{n+1}^j \) cannot be computed directly from the numerical values \( u_n^j \) at the previous time level \( t^n \). Indeed, writing the difference equations as a matrix-vector recursion like in (11.38), we can formulate the scheme as

\[
(I - \Delta t A)u_{n+1}^j = u_n^j + \Delta t f_{n+1}^j.
\]

(11.42)

So for each new time level a linear system of equations has to be solved. As before, the boundary conditions are represented in the source term, i.e., \( f_{n+1} \), whereas for the initial value we have \( u^0 = u(0) \), cf. (11.30d). The coefficient matrix \( I - \Delta t A \) is a symmetric positive definite tridiagonal matrix and consequently the system (11.42) can be solved very efficiently using LU-decomposition without partial pivoting (in fact, Choleski decomposition; see [52]). In principle the linear system (11.42) has to be solved at each time level. Note that in this (constant coefficient) case the LU-decomposition of the matrix A has to be done only once; hence only the forward and backward substitutions need to be performed at each time step, making the time stepping very efficient. For nonconstant matrices this advantage is lost, however.

Next we apply Fourier mode analysis to investigate the stability of scheme (11.41). If we substitute the mode (11.10) in the homogeneous scheme (11.41), assuming periodic boundary conditions, and apply the relations (11.11), we find

\[
\lambda - d\lambda (e^{i\varphi} - 2 + e^{-i\varphi}) = 1,
\]

from which we deduce that

\[
\lambda(\varphi) = \frac{1}{1 + 4d \sin^2 \left( \frac{\varphi}{2} \right)}.
\]

(11.43)

It is obvious from (11.43) that the stability condition (11.13) is always satisfied. This means that we do not have to restrict the time step \( \Delta t \) in order to enforce the stability condition (11.13). Therefore the numerical scheme (11.41) is unconditionally stable, confirming our expectations.

Due to its similarity with the explicit Euler scheme we expect the accuracy of the implicit scheme (11.40) to be first order as well. We obtain for the local discretisation...
error $d^n_j$

$$d^n_j := \frac{1}{\Delta t} \left( u(x_j, t^{n+1}) - u(x_j, t^n) \right)$$

$$- \frac{1}{\Delta t} \left( u(x_{j+1}, t^{n+1}) - 2u(x_j, t^{n+1}) + u(x_{j-1}, t^{n+1}) \right) - f_j^{n+1}; \quad (11.44)$$

cf. (11.15). Assuming $u(x, t)$ to be sufficiently smooth, straightforward substitution of
taylor series expansions about $(x_j, t^{n+1})$ into (11.44) gives

$$d^n_j = -\frac{1}{2} \Delta t \frac{\partial^2 u}{\partial t^2}(x_j, t^{n+1}) - \frac{1}{12} \Delta x^2 \frac{\partial^4 u}{\partial x^4}(x_j, t^{n+1}) + O(\Delta t^2) + O(\Delta x^4). \quad (11.45)$$

Hence the implicit scheme (11.40) is consistent.

From this we can find an upper bound for the global discretisation error $e_j^{n+1}$. From
(11.40) and (11.44) we find the following recurrence relation for $e_j^n$:

$$e_j^{n+1} - d(e_j^{n+1} - 2e_j^{n+1} + e_{j-1}^{n+1}) = e_j^n + \Delta td^n_j. \quad (11.46)$$

Using the triangle inequality, we derive from (11.46)

$$(1 + 2d) \max_{1 \leq j \leq M} |e_j^{n+1}| =: (1 + 2d)|e_j^{n+1}|$$

$$\leq d \left( |e_{j+1}^{n+1}| + |e_{j-1}^{n+1}| \right) + |e_j^n| + \Delta t |d_j^n|$$

$$\leq 2d \max_{1 \leq j \leq M} |e_j^{n+1}| + \max_{1 \leq j \leq M} |e_j^n| + \Delta t \max_{1 \leq j \leq M} |d_j^n| \quad (11.47)$$

for some (possibly not unique) index $j^*$ ($1 \leq j^* \leq M$). It is now obvious that (11.47)
again leads to the inequality (11.22); however, this time no restriction on $\Delta t$ is imposed.
Following the same line of reasoning as in Section 11.1, we again would obtain the error
estimate (11.27) and, consequently, the implicit scheme (11.41) is convergent.

**Example 11.3** Continuing Example 11.1, we employ the implicit Euler scheme (11.41) to solve
the same initial boundary value problem. We take a grid size $\Delta x = 5 \times 10^{-3}$ and time steps
$\Delta t = 1.2 \times 10^{-3}$ and $\Delta t = 1.3 \times 10^{-3}$, respectively. The numerical solutions are shown in
Figure 11.5. In this case both solutions are a good approximation of the exact solution. Thus
the implicit scheme does not display the erroneous behaviour of the explicit scheme, which
was caused by the instability of the scheme. \hfill \Box

### 11.3.2 The $\theta$ Scheme

Given the form of the semidiscrete problem (11.30), which is an ODE, it is natural to ask
how we can fare with more general time-stepping methods. We shall restrict ourselves here
to the class of $\theta$ methods. These methods are linear combinations of the explicit and implicit
Euler schemes, i.e., (11.3) and (11.40). They read

$$\frac{1}{\Delta t} \left( u_j^{n+1} - u_j^n \right) = \frac{1 - \theta}{\Delta x^2} \left( u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1} \right)$$

$$+ \frac{\theta}{\Delta x^2} \left( u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1} \right) + (1 - \theta) f_j^n + \theta f_j^{n+1} \quad (11.48)$$

for some parameter $\theta$ with $0 \leq \theta \leq 1$. Obviously, the difference quotient on the left-
hand side of (11.48) is an approximation of the weighted average time derivative $(1 - \vartheta) \frac{\partial}{\partial t} u(x_j, t^n) + \vartheta \frac{\partial}{\partial t} u(x_j, t^{n+1})$. Note that for $\vartheta = 0$ we recover the explicit scheme (11.3) and for $\vartheta = 1$ we have the implicit scheme (11.40). For $\vartheta = \frac{1}{2}$ the scheme (11.48) is called the Crank–Nicolson scheme. The parameter $\vartheta$ can be used to optimize the accuracy and/or stability of the scheme. Rearranging terms, with all numerical values $u_k^{n+1}$ ($k = j, j \pm 1$) on the left-hand side and all numerical values $u_k^n$ ($k = j, j \pm 1$) on the right-hand side, we find

$$u_j^{n+1} - \vartheta d(u_j^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) = u_j^n + (1 - \vartheta)d(u_j^{n+1} - 2u_j^n + u_{j-1}^n) + \Delta t((1 - \vartheta)f_j^n + \vartheta f_j^{n+1}),$$

(11.49)

Figure 11.5. Numerical solution of the heat equation computed with the implicit Euler scheme with time steps $\Delta t = 1.2 \times 10^{-3}$ (left) and $\Delta t = 1.3 \times 10^{-3}$ (right).
11.3. Implicit Schemes

with the diffusion number \( d \) defined in (11.5). The numerical scheme (11.49) is called the \( \vartheta \) method. The stencil of (11.49) is shown in Figure 11.6. It is obvious that for \( \vartheta \neq 0 \) the scheme is implicit. Consequently, to compute the numerical solution requires a linear system to be solved at each time step. Finally, the numerical scheme has to be completed with the initial conditions (11.6) and boundary conditions (11.7).

The difference equations (11.49) together with the boundary conditions (11.7) lead to the linear system

\[
(\mathbf{I} - \vartheta \Delta t \mathbf{A}) \mathbf{u}^{n+1} = (\mathbf{I} + (1 - \vartheta) \Delta t \mathbf{A}) \mathbf{u}^n + \Delta t ((1 - \vartheta) \mathbf{f}^n + \vartheta \mathbf{f}^{n+1}),
\]

(11.50)

where the vector \( \mathbf{u}^{n+1} \) is the numerical approximation of \( \mathbf{u}(t^{n+1}) \) and the discretisation matrix \( \mathbf{A} \) is defined in (11.30b). The coefficient matrix \( \mathbf{B}_\vartheta := \mathbf{I} - \vartheta \Delta t \mathbf{A} \) is again a symmetric positive definite tridiagonal matrix, which makes the system (11.50) easy to solve using LU-decomposition. The computational costs per time step for the \( \vartheta \) method are only slightly higher than for the implicit Euler scheme.

We investigate the stability of the \( \vartheta \) scheme (11.49) by the Fourier mode analysis. Thus, substituting the mode (11.10), we find for the amplification factor \( \lambda \)

\[
\lambda - \vartheta d \lambda (e^{i\varphi} - 2 + e^{-i\varphi}) = 1 + (1 - \vartheta) d (e^{i\varphi} - 2 + e^{-i\varphi}),
\]

from which we conclude that

\[
\lambda(\varphi) = \frac{1 - 4(1 - \vartheta) d \sin^2(\frac{\varphi}{2})}{1 + 4 \vartheta d \sin^2(\frac{\varphi}{2})}.
\]

(11.51)

Applying the inequality (11.13) to (11.51) leads to the stability conditions

\[
\begin{cases}
0 \leq \vartheta < \frac{1}{2} \\
\frac{1}{2} \leq \vartheta \leq 1
\end{cases}
\]

(11.52)

For \( \frac{1}{2} \leq \vartheta \leq 1 \) the scheme is unconditionally stable, while for \( \vartheta = 0 \) the stability condition (11.14) is recovered.
Chapter 11. Numerical Methods for Parabolic Equations

As in the previous sections, we take the maximum value of the following upper bound for the inequalities (11.22) and thus (11.25). Combining this with the estimate (11.54), we find

$$ d^n_j := \frac{1}{\Delta t} (u(x_j, t^{n+1}) - u(x_j, t^n)) - \frac{1 - \vartheta}{\Delta x^2} (u(x_{j+1}, t^n) - 2u(x_j, t^n) + u(x_{j-1}, t^n)) $$

$$ - \vartheta \frac{\partial^2 u}{\partial x^2} (u(x_{j+1}, t^{n+1}) - 2u(x_j, t^{n+1}) + u(x_{j-1}, t^{n+1})) - (1 - \vartheta) f^n_j - \vartheta f^{n+1}_j. \quad (11.53) $$

Expanding all function values in (11.53) about \( (x_j, t^n) \) and assuming that \( u(x, t) \) is sufficiently smooth, we find

$$ d^n_j = - \left( \vartheta - \frac{1}{2} \right) \Delta t \frac{\partial^2 u}{\partial t^2} (x_j, t^n) - \frac{1}{12} \Delta x^2 \frac{\partial^4 u}{\partial x^4} (x_j, t^n) + O(\Delta t^2) + O(\Delta x^4). \quad (11.54) $$

From this expression we see that the \( \vartheta \)-method is first order accurate in the time step \( \Delta t \) and second order accurate in the grid size \( \Delta x \), except for \( \vartheta = \frac{1}{2} \), in which case the scheme is also second order accurate in \( \Delta t \). Apparently, \( \vartheta = \frac{1}{2} \) is the best choice, since it gives the smallest discretisation error and unconditional stability; cf. (11.52). This optimality result is well known for the trapezoidal method when solving ODEs (see [97]). However, the Crank–Nicolson scheme does not always produce reliable results, as is demonstrated in the next example.

Example 11.4. In this example we employ the Crank–Nicolson scheme to compute a numerical solution of the initial boundary value problem (11.1) with initial data \( v(x) = 0 \) and boundary conditions (11.1c) and (11.1d) replaced by \( u(0, t) = 0 \) and \( u(1, t) = 1 \). Note that \( v(1) \neq u(1, 0) \), and we say that the data of the initial boundary value problem (11.1) are discontinuous. We take \( \Delta x = \Delta t = 5 \times 10^{-2} \). The numerical solution is shown at time levels \( t^n = n \Delta t \) \( (n = 1, 2, \ldots, 6) \) in Figure 11.7. Obviously, the numerical solution tends to the steady solution \( u_\infty(x) = x \); however, the oscillatory behaviour in the grid point adjacent to the boundary \( x = 1 \) is rather peculiar. This erroneous behaviour of the Crank–Nicolson scheme is due to the discontinuity in the data of the initial boundary value problem to be solved, as will be explained in Section 11.5.

The analysis of the global discretisation error \( e^{n+1}_j \) is quite similar to what we have done for the explicit and implicit Euler schemes, and therefore we only give a brief outline. From (11.48) and the expression for the local discretisation error \( d^n_j \) in (11.53) we get the following recurrence relation for \( e_j^{n+1} \):

$$ e_j^{n+1} - \vartheta d(e_{j+1}^{n+1} - 2e_j^{n+1} + e_{j-1}^{n+1}) = e^n_j + (1 - \vartheta) d(e_{j+1}^{n} - 2e_j^{n} + e_{j-1}^{n}) + \Delta t d_j^n. \quad (11.55) $$

If \( 2(1 - \vartheta) d \leq 1 \), the coefficient of \( e_j^n \) in the recurrence relation (11.55) is positive. Proceeding as in the previous sections, we take the maximum value of \( |e_j^{n+1}| \) and obtain precisely the inequalities (11.22) and thus (11.25). Combining this with the estimate (11.54), we find the following upper bound for \( ||e^n||_\infty \):

$$ ||e^n||_\infty \leq C \left( \frac{1}{2} - \vartheta \right) \Delta t + \frac{1}{12} \Delta x^2 + O(\Delta t^2) + O(\Delta x^4). \quad (11.56) $$
11.3. Implicit Schemes

Figure 11.7. Numerical solution of the heat equation with discontinuous data, computed with the $\vartheta$-scheme.

with $C$ defined in (11.26). From this inequality we can conclude that the $\vartheta$ method is convergent. More precisely, we say that the method is first order convergent in $\Delta t$ and second order convergent in $\Delta x$ in the $\infty$-norm, except when $\vartheta = \frac{1}{2}$, in which case the $\vartheta$ method is second order convergent in both $\Delta t$ and $\Delta x$. From a practical point of view a choice $\vartheta = \frac{1}{2} + O(\Delta t)$ is also advantageous. In order to have unconditional stability, $\vartheta$ should be chosen larger than $\frac{1}{2}$. 
11.4 Analysis of the $\vartheta$ Method by the Matrix Method

Since Fourier mode analysis is rather restrictive (requiring periodic boundary conditions, e.g.), we shall analyse the stability of the $\vartheta$ method by the matrix method as well. The latter will thus be the method of choice to analyse stability as well as to obtain error estimates for the global discretisation error in more general situations.

Consider the matrix-vector recursion (11.50) for the $\vartheta$ method. The global error satisfies a similar relation. Indeed, substituting $e^m_0 = e^m_{M+1} = 0$ $(m = n, n + 1)$ in (11.55), we find the relation

$$B_{\vartheta} e^{n+1} = B_{\vartheta - 1} e^n + \Delta t d^n,$$

where the matrices $B_{\vartheta}$ and $B_{\vartheta - 1}$ are defined by

$$B_{\vartheta} := I - \vartheta \Delta t A \quad (\vartheta = \vartheta, \vartheta - 1).$$

Since we know that $A$ has negative eigenvalues only, it follows that $B_{\vartheta}$ is nonsingular. Because of the linearity of the problem we can make the recursion explicit. So

$$e^{n+1} = C e^n + \Delta t B_{\vartheta}^{-1} d^n,$$

with

$$C := B_{\vartheta}^{-1} B_{\vartheta - 1}.$$  

As initial condition for (11.57a) we trivially have $e^0 = 0$. Hence we find the following expression for the global discretisation error vector $e^n$:

$$e^n = \Delta t \sum_{l=0}^{n-1} C^{n-1-l} B_{\vartheta}^{-1} d^l.$$  

Thus the global discretisation error vector $e^n$ contains the local discretisation error vectors $d^l$ at all previous time levels. We will now investigate the stability and accuracy of the $\vartheta$ method based on the error expression (11.59). At this point we have to specify which norms to use. The 2-norm is a convenient choice for the following reason. The matrices $C$ and $B_{\vartheta}^{-1}$ can be jointly transformed to diagonal form, for which the 2-norm is simply the spectral radius; see (G.5c). Indeed, using the definitions of $B_{\vartheta}$ and $B_{\vartheta - 1}$ in (11.57b), we find

$$W^{-1} B_{\vartheta}^{-1} W = (I - \vartheta \Delta t A)^{-1},$$

$$W^{-1} C W = (I - \vartheta \Delta t A)^{-1} (I + (1 - \vartheta) \Delta t A),$$

with $W$ the orthogonal eigenvector matrix and $A$ the diagonal matrix defined in (11.33) and (11.35), respectively. The 2-norm of these matrices is given by

$$\|W^{-1} B_{\vartheta}^{-1} W\|_2 = \rho((I - \vartheta \Delta t A)^{-1}) = \rho(B_{\vartheta}^{-1}),$$

$$\|W^{-1} C W\|_2 = \rho((I - \vartheta \Delta t A)^{-1} (I + (1 - \vartheta) \Delta t A)) = \rho(C),$$

where $\rho(A)$ denotes the spectral radius of the matrix $A$. Therefore, the 2-norm of the global discretisation error vector $e^n$ is

$$\|e^n\|_2 = \Delta t \sum_{l=0}^{n-1} \rho(C^{n-1-l} B_{\vartheta}^{-1} d^l).$$

This implies that the stability and accuracy of the $\vartheta$ method are closely related to the spectral radius of $B_{\vartheta}$ and $B_{\vartheta - 1}$.
11.4. Analysis of the \( \theta \) Method by the Matrix Method

where \( \rho(\cdot) \) denotes the spectral radius of the corresponding matrix. Finally, since the 2-norm is invariant under orthogonal transformations, we have

\[
\| C^{n-1-l} B_\theta^{-1} \|_2 \leq \| C \|_2^{n-1-l} \| B_\theta^{-1} \|_2 = \| W^{-1} C W \|_2^{n-1-l} \| W^{-1} B_\theta^{-1} W \|_2 = \rho(C)^{n-1-l} \rho(B_\theta^{-1}).
\]

(11.62)

From (11.59) we obtain the following upper bound for \( \| e \|_2 \):

\[
\| e \|_2 \leq \frac{1}{\Delta t} \rho(B_\theta^{-1}) \sum_{l=0}^{n-1} \rho(C)^{n-1-l} \| d_l \|_2.
\]

(11.63)

If the matrices \( C^{n-1-l} \) are bounded uniformly in \( l \) and if \( B_\theta \) is nonsingular, we can decide for the stability of the \( \theta \) scheme; see Definition 5.15. This is so in particular if

\[
\rho(C) \leq 1, \quad \rho(B_\theta^{-1}) < \infty.
\]

(11.64)

In order to compute the spectral radii in (11.64), we employ the eigenvalues \( \lambda_k \) of \( A \); see (11.31). Let the eigenvalues of \( B_\theta^{-1} \) and \( C \) be denoted by \( \mu_k \) and \( \eta_k \), respectively. Since \( B_\theta \) and \( C \) have the same set of eigenvectors as \( A \), it easily follows that

\[
\mu_k = (1 - \theta \Delta t \lambda_k)^{-1}, \quad \eta_k = \frac{1 + (1 - \theta) \Delta t \lambda_k}{1 - \theta \Delta t \lambda_k} \quad (k = 1, 2, \ldots, M).
\]

(11.65)

From (11.31) and (11.65) it is obvious that \( 0 < \mu_k < 1 \) for \( k = 1, 2, \ldots, M \), and thus \( \rho(B_\theta^{-1}) < 1 \). The first inequality in (11.64) is really the interesting one to prove and actually reduces to

\[
-1 \leq \frac{1 + (1 - \theta) \Delta t \lambda_k}{1 - \theta \Delta t \lambda_k} \leq 1 \quad (k = 1, 2, \ldots, M).
\]

(11.66)

The right inequalities in (11.66) are always satisfied and thus we only have to investigate the left inequalities. Rearranging terms, we find the conditions

\[
(2\theta - 1) \Delta t \lambda_k \leq 2 \quad (k = 1, 2, \ldots, M).
\]

(11.67)

Note that for \( \frac{1}{2} \leq \theta \leq 1 \) (11.67) is simply satisfied. For \( 0 \leq \theta < \frac{1}{2} \) we have to substitute expression (11.31) for the eigenvalues \( \lambda_k \). This results in the stability conditions

\[
\begin{cases} 
0 \leq \theta < \frac{1}{2} \quad \text{and} \quad d \leq \frac{1}{2(1 - 2\theta)}, \\
\frac{1}{2} \leq \theta \leq 1.
\end{cases}
\]

(11.68)

Note that these stability conditions are exactly the same as the ones given in (11.52), which are obtained by Fourier mode analysis.
Example 11.5 We have plotted in Figure 11.8 the eigenvalues of $B^{-1}_\vartheta$ and $C$ for $\Delta x = \Delta t = 5 \times 10^{-3}$ and $\vartheta = 0.5, 0.75,$ and 1. We see that for $\vartheta = 0.5$, we have $\eta_k \approx -1$ when $k \to M$. This means that high-frequency eigenvector components in $e^n$, corresponding to (in absolute value) large eigenvalues $\lambda_k$, are badly damped by the Crank–Nicolson scheme. On the other hand, for $\vartheta = 1$, we have $\eta_k \approx 0$ when $k \to M$, implying that high-frequency components are very well damped for the implicit Euler scheme.

With the above stability result we can give an upper bound for the global discretisation error. From (11.54) we find the following upper bound for $\|d^n\|_2$:

$$
\|d^n\|_2 \leq \sqrt{M} \|d^n\|_\infty \\
\leq \frac{1}{\sqrt{\Delta x}} \left( \left( \frac{1}{2} - \vartheta \right) \Delta t + \frac{1}{12} \Delta x^2 \right) C + O(\Delta t^2) + O(\Delta x^4).
$$

where we have assumed that $f(x, t) \equiv 0$ and $u(x, t)$ satisfies (11.26). Now we can determine an upper bound for $\|e^n\|_2$. Assuming the stability conditions (11.68) to be satisfied, we can deduce from (11.63) that

$$
\|e^n\|_2 \leq \Delta t \sum_{l=0}^{n-1} \|d^l\|_2 \\
\leq \frac{\rho^n}{\sqrt{\Delta x}} \left( \left( \frac{1}{2} - \vartheta \right) \Delta t + \frac{1}{12} \Delta x^2 \right) C + O(\Delta t^2) + O(\Delta x^4).
$$

This error bound is a bit disappointing in view of the second order accuracy found in (11.56). One way to interpret (11.70) is by considering the median of the norm, i.e., $\|e^n\|_2/\sqrt{M}$. A more technical proof of a sharper $\infty$-norm for the Crank–Nicolson scheme is given in the following property.

Property 11.6. If $\Delta t = O(\Delta x)$, then the global error in the $\infty$-norm for the Crank–Nicolson scheme is second order; i.e., $\|e^n\|_\infty = O(\Delta x^2)$.
11.4. Analysis of the $\vartheta$ Method by the Matrix Method

**Proof.** For the proof of this estimate we have to consider the expression

$$C^{n-1-i}B^{-1}_{\vartheta}$$

more closely. We note that this matrix is similar to $\Lambda$. Using the transformation matrix $W$ (see (11.33)), we obtain

$$W^{-1}C^{n-1-i}B^{-1}_{\vartheta}W = \text{diag} \left( \eta_k^{n-1-i} \mu_k \right) =: P_{n-1-i},$$

where $\eta_k$ and $\mu_k$ are given in (11.65). Taking the absolute value of all matrix entries and estimating elementwise, we find

$$\left| C^{n-1-i}B^{-1}_{\vartheta} \right| \leq \frac{2}{M} \sum_{k=1}^{M} |\eta_k^{n-1-i} \mu_k|.$$

After working out the product of the last two matrices, we conclude that

$$\left\| C^{n-1-i}B^{-1}_{\vartheta} \right\|_{\infty} \leq \frac{2}{M} M \sum_{k=1}^{M} |\eta_k^{n-1-i} \mu_k|.$$

Consider the expression for $\eta_k$ in (11.65). Note that the function $\psi(\xi) := (1 + \frac{1}{2} \xi)/(1 - \frac{1}{2} \xi)$ is monotonically increasing. Furthermore, we have $\frac{1}{2} \xi \leq \sin \zeta$ for $0 \leq \zeta \leq \frac{1}{2} \pi$, so $\lambda_k \leq -4k^2$. We now distinguish between two possibilities: $\eta_k \geq 0$ (case (i)) and $\eta_k \leq 0$ (case (ii)). First, consider case (i). Note that we need the bound for the $\lambda_k$ here. We obtain

$$\sum_{l=1}^{n-1} |\eta_k|^{n-1-i} \mu_k \leq \left( 1 - \frac{1 + \frac{1}{2} \Delta t \lambda_k}{1 - \frac{1}{2} \Delta t \lambda_k} \right)^{-1} \frac{1}{1 - \frac{1}{2} \Delta t \lambda_k} \leq \frac{1}{4k^2 \Delta t}.$$

In case (ii) we find

$$\sum_{l=1}^{n-1} |\eta_k|^{n-1-i} \mu_k \leq \left( 1 + \frac{1}{2} \Delta t \lambda_k + 1 \right)^{-1} \frac{1}{1 - \frac{1}{2} \Delta t \lambda_k} \leq \frac{1}{2}.$$

The first case clearly gives an upper bound $O(\Delta t^{-1})$, while the second gives an upper bound $O(\Delta x^{-1})$; for the latter one should realize that $\Delta x = O(M^{-1})$. Combining the results in case (i) and case (ii), we thus conclude that

$$\sum_{l=1}^{n-1} \left\| C^{n-1-i}B^{-1}_{\vartheta} \right\|_{\infty} \leq \sum_{k=1}^{M} \sum_{l=1}^{n-1} |\eta_k|^{n-1-i} \mu_k = O \left( \frac{1}{\Delta t} \right) + O \left( \frac{1}{\Delta x} \right).$$

Using this estimate in the expression for the local error in (11.59) yields the required result. \( \square \)

We remark that one often prefers to have a slightly more dissipative scheme than Crank–Nicolson. In order not to lose the actual second order error, one can take $\vartheta = \frac{1}{2} + O(\Delta t)$, where of course $\vartheta > \frac{1}{2}$.
11.5 Initial Boundary Value Problems with Discontinuous Data

To show another use of the matrix method we analyse the phenomenon noticed in Example 11.4. Consider the following initial boundary value problem for $u(x, t)$:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1) \quad t > 0,$$  

(11.71a)

$$u(x, 0) = 0, \quad x \in (0, 1),$$  

(11.71b)

$$u(0, t) = 0, \quad u(1, t) = 1, \quad t > 0.$$  

(11.71c)

Note that at the right boundary $x = 1$ and at time $t = 0$ the solution $u$ abruptly jumps from zero to one; i.e., the initial boundary value problem (11.71) has discontinuous data. If we employ the central difference approximation (5.30) for the spatial derivative and apply the boundary conditions in (11.71c), i.e.,

$$u_0 = 0, \quad u_{M+1} = 1,$$  

(11.72)

we obtain the initial value problem (cf. (11.30))

$$\frac{du}{dt} = Au + f,$$  

(11.73a)

$$u(0) = 0,$$  

(11.73b)

where the vector $f$ is given by

$$f := \frac{1}{\Delta x^2} (0, 0, \ldots, 0, 1)^T.$$  

(11.74)

The exact solution of the initial value problem (11.73) reads

$$u(t) = \int_0^t \exp(A(t-s))f \, ds.$$  

(11.75)

We can formally compute the matrix exponential $\exp(A(t-s))$ in (11.75) using the spectral decomposition $A = W \Lambda W^{-1}$; cf. (11.34). We find

$$\exp(A(t-s)) = W \exp(\Lambda(t-s)) W^{-1},$$  

(11.76a)

$$\exp(A(t-s)) = \text{diag}(e^{\lambda_1(t-s)}, e^{\lambda_2(t-s)}, \ldots, e^{\lambda_M(t-s)}).$$  

(11.76b)

Substitution of (11.76) in (11.75) and subsequent integration gives

$$u(t) = WA^{-1}(\exp(At) - I)W^{-1}f$$  

$$= A^{-1}(\exp(At) - I)f.$$  

(11.77)

As can be seen from Figure 11.7, we are mainly interested in the transient part of the solution in (11.77), $v$, say, i.e.,

$$v(t) = A^{-1}\exp(At)f.$$  

(11.78)
11.5. Initial Boundary Value Problems with Discontinuous Data

From (11.78) we deduce that the increment of \( v(t) \) on a time interval \( \Delta t \) is given by

\[
v(t + \Delta t) = \exp(A \Delta t) v(t),
\]

(11.79)

where the increment matrix \( \exp(A \Delta t) \) is given by

\[
\exp(A \Delta t) = W \text{diag}(e^{\lambda_1 \Delta t}, e^{\lambda_2 \Delta t}, \ldots, e^{\lambda_M \Delta t}) W^{-1}.
\]

(11.80)

Now suppose we use the \( \vartheta \) method for the system of differential equations in (11.73). Then in matrix terms we obtain

\[
\frac{1}{\Delta t} (u^{n+1} - u^n) = (1 - \vartheta) (Au^n + f) + \vartheta (Au^{n+1} + f).
\]

(11.81)

Multiplying by \( \Delta t \) and rearranging terms, we find the recurrence relation

\[
u^{n+1} = Cu^n + \Delta t B_\vartheta^{-1} f,
\]

(11.82)

where the matrices \( B_\vartheta \) and \( C \) are defined in (11.57b) and (11.58b), respectively. Since one is not an eigenvalue of \( C \), the matrix \( C - I \) is invertible. Using the initial condition \( u^0 = 0 \), we can see that the solution of (11.82) is given by

\[
u^n = \Delta t \left( \sum_{l=0}^{n-1} C^l \right) B_\vartheta^{-1} f = \Delta t (C - I)^{-1} (C^n - I) B_\vartheta^{-1} f,
\]

(11.83)

where we used the expression

\[
(C - I)(I + C + C^2 + \cdots + C^{n-1}) = C^n - I.
\]

By further noting that

\[
\Delta t (C - I)^{-1} B_\vartheta^{-1} = A^{-1}
\]

and that the matrices \( (C - I)^{-1} \) and \( C^n \) commute, we see that the solution of (11.83) is given by

\[
u^n = A^{-1} (C^n - I) f.
\]

(11.84)

This is a nice (time) discrete analogue of (11.78). In particular, the discrete transient part is given by

\[
v^n := A^{-1} C^n f.
\]

(11.85)

Clearly, the effect of the initial condition is damped out by the decay of the matrix \( C^n \). Therefore it makes sense to investigate the increment matrix \( C \) as given in (11.82); i.e., we compare this to the increment matrices \( \exp(A \Delta t) \) (the exact decay matrix per step \( \Delta t \)). Using the spectral decomposition of \( A \), we can rewrite the matrix \( C \) as

\[
C = (I - \vartheta \Delta t A)^{-1} (I + (1 - \vartheta) \Delta t A)
\]

\[
= W (I - \vartheta \Delta t A)^{-1} (I + (1 - \vartheta) \Delta t A) W^{-1}
\]

\[
= W \text{diag}(r(\lambda_1 \Delta t), r(\lambda_2 \Delta t), \ldots, r(\lambda_M \Delta t)) W^{-1}.
\]

(11.86)
Figure 11.9. Amplification factor $e^{\lambda}$ compared with the rational function $r(z)$ for $\theta = 0.5, 0.75,$ and $1.$

with the rational function $r(z)$ defined by

$$r(z) := \frac{1 + (1 - \theta)z}{1 - \theta z}. \quad (11.87)$$

Comparing (11.80) and (11.86), we see that the incremental factors $e^{\lambda \Delta t}$ in $\exp(A \Delta t)$ are approximated by the fractions $r(\lambda_k \Delta t)$, which might be close to one in modulus and even become negative for larger negative eigenvalues $\lambda_k$; see Figure 11.9. This then introduces oscillatory components in the numerical solution, which explains the curious behaviour in Example 11.4.

11.6 Mixed Boundary Conditions

So far we have only considered Dirichlet boundary conditions. However, in practice the flux of a variable is often prescribed at the boundary of a domain, leading to mixed boundary conditions; see Section 10.4. Treating mixed boundary conditions, in particular the fluxes, requires special attention. In this section we investigate how to discretise these and look for sufficient conditions to ensure the stability of the method.

Consider the following initial boundary value problem for $u(x, t)$ (cf. (10.41)):

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0, \quad (11.88a)$$

$$u(x, 0) = v(x), \quad x \in (0, 1), \quad (11.88b)$$

$$\frac{\partial u}{\partial x}(0, t) - \alpha (u(0, t) - g_\ell(t)) = 0, \quad t > 0, \quad (11.88c)$$

$$\frac{\partial u}{\partial x}(1, t) + \beta (u(1, t) - g_r(t)) = 0, \quad t > 0, \quad (11.88d)$$
with parameters $\alpha, \beta \geq 0$ and $g_\ell(t), g_r(t)$ some given functions. When $\alpha = \beta = 0$, we obtain the Neumann boundary conditions

$$\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = 0, \quad t > 0,$$

and when $\alpha, \beta \to \infty$ the boundary conditions in (11.88) reduce to the Dirichlet boundary conditions $u(0, t) = g_\ell(t)$ and $u(1, t) = g_r(t)$.

For the numerical solution of the initial boundary value problem (11.88) we employ central differences for space discretisation and the $\vartheta$ method for time integration. This gives the following numerical scheme, which holds at all internal grid points $x_j$ ($j = 1, 2, \ldots, M$):

$$u_{j}^{n+1} - \vartheta d\left(u_{j+1}^{n+1} - 2u_{j}^{n+1} + u_{j-1}^{n+1}\right) = u_{j}^{n} + (1 - \vartheta)d\left(u_{j+1}^{n} - 2u_{j}^{n} + u_{j-1}^{n}\right).$$

(11.90)

To facilitate the notation we take in the following that $g_\ell(t) = g_r(t) \equiv 0$. For the initial boundary value problem (11.88) the boundary values $u(0, t)$ and $u(1, t)$ are not explicitly known, since they are related to the (unknown) derivatives $\frac{\partial}{\partial x}u(0, t)$ and $\frac{\partial}{\partial x}u(1, t)$, respectively. Therefore the numerical variables $u_m^n, u_{m+1}^n$ ($m = n, n + 1$) are unknown and have to be determined somehow. We can achieve this by requiring the PDE to hold on a slightly larger domain, introducing virtual grid points $x_{-1} := -\Delta x$ and $x_{M+2} := 1 + \Delta x$, and using the numerical scheme at the boundary points $x_0$ and $x_{M+1}$. First consider the left boundary, for which $j = 0$. Substituting $j = 0$ into (11.90), we obtain

$$u_{0}^{n+1} - \vartheta d\left(u_{0}^{n+1} - 2u_{0}^{n+1} + u_{-1}^{n+1}\right) = u_{0}^{n} + (1 - \vartheta)d\left(u_{1}^{n} - 2u_{0}^{n} + u_{-1}^{n}\right).$$

(11.91)

This equation contains the numerical values $u_m^n$ ($m = n, n + 1$), which can be eliminated from (11.91) by applying a finite difference approximation of the boundary condition at $x = 0$. To discretise the derivative $\frac{\partial}{\partial x}u(0, t)$, we choose the central difference approximation (5.28c), because it is second order accurate like the central difference approximation of $\frac{\partial^2}{\partial x^2}u(x_j, t)$ at all grid points. However, we remark that there is a snag here, which will be addressed at the end of this section. We now obtain

$$\frac{1}{2\Delta x}(u_{1}^{n} - u_{-1}^{n}) - \alpha u_m^n = 0 \quad (m = n, n + 1),$$

(11.92a)

or, equivalently,

$$u_{-1}^{n} = u_{1}^{n} - 2\alpha \Delta x u_m^n \quad (m = n, n + 1).$$

(11.92b)

Substituting the discrete boundary condition (11.92b) into (11.91), we get

$$u_{0}^{n+1} - \vartheta d\left(2u_{1}^{n+1} - 2(1 + \alpha \Delta x)u_{0}^{n+1}\right) = u_{0}^{n} + (1 - \vartheta)d\left(2u_{1}^{n} - 2(1 + \alpha \Delta x)u_{0}^{n}\right).$$

(11.93)

Likewise, at the right boundary point $x_{M+1}$, we have the equation

$$u_{M+1}^{n+1} - \vartheta d\left(u_{M+2}^{n+1} - 2u_{M+1}^{n+1} + u_{M}^{n+1}\right) = u_{M+1}^{n} + (1 - \vartheta)d\left(u_{M+2}^{n} - 2u_{M+1}^{n} + u_{M}^{n}\right).$$

(11.94)

containing the numerical values $u_m^n$ ($m = n, n + 1$) at the virtual grid point $x_{M+2} = 1 + \Delta x$. These values can be eliminated if we combine (11.94) with the central difference approximation of the boundary condition at $x = 1$, i.e.,

$$\frac{1}{2\Delta x}(u_{M+2}^{n} - u_{M}^{n}) + \beta u_{M+1}^{n} = 0 \quad (m = n, n + 1).$$

(11.95a)
From this equation we immediately conclude that
\[ u_{M+2}^m = u_M^m - 2\beta \Delta x u_{M+1}^m \quad (m = n, n + 1). \]  \hfill (11.95b)

Substituting the latter equation into (11.94), we obtain
\[ u_{M+1}^{n+1} - \vartheta d (2u_M^{n+1} - 2(1 + \beta \Delta x)u_{M+1}^{n+1}) = u_{M+1}^n + (1 - \vartheta) d (2u_M^n - 2(1 + \beta \Delta x)u_{M+1}^n). \]  \hfill (11.96)

The numerical scheme for the initial boundary value problem (11.88) consists of the equations (11.90), (11.93), and (11.96) for the unknowns \( u_n := (u_0^n, u_1^n, \ldots, u_{M+1}^n)^T \).

The above numerical scheme can be written as the linear system
\[
(I - \vartheta \Delta t A) u_{n+1} = (I + (1 - \vartheta) \Delta t A) u_n,
\]  \hfill (11.97)

with \( I \) the \((M+2) \times (M+2)\) identity matrix and \( A \) the discretisation matrix given by
\[
A := \frac{1}{\Delta x^2} \begin{pmatrix}
-2(1 + \alpha \Delta x) & 2 & & \\
1 & -2 & 1 & \\
& \ddots & \ddots & \ddots \\
& & 1 & -2 & 1 \\
& & & 2 & -2(1 + \beta \Delta x)
\end{pmatrix}.
\]  \hfill (11.98)

This matrix \( A \) is the discretisation matrix of the differential operator \( \frac{\partial^2}{\partial x^2} \) and the boundary conditions together. The first and last rows of \( A \) represent the boundary conditions at \( x = 0 \) and \( x = 1 \), respectively. Note that for \( \alpha = \beta = 0 \), i.e., for Neumann boundary conditions, the matrix \( A \) is singular. In particular, the vector \( e := (1, 1, \ldots, 1)^T \) is then a basis for the null space of \( A \), which in turn implies that the numerical solution computed from (11.97) is determined up to \( \gamma e \), with \( \gamma \) an arbitrary constant.

Next we want to apply the matrix method to investigate the stability of (11.97). Completely analogously to the derivation in Section 11.4, we find that the global discretisation error vector satisfies the recurrence relation (11.57a), but with the discretisation matrix \( A \) now as defined in (11.98). The solution of (11.57a) is given in (11.59), with matrices \( B_\vartheta \), \( B_{\vartheta-1} \), and \( C \) defined like in (11.57b) and (11.58b). The stability conditions are again given in (11.64).

The problem is now that the matrix \( A \) is not symmetric any more as a result of the (discrete) boundary conditions. Consequently, the matrices \( B_\vartheta \), \( B_{\vartheta-1} \), and \( C \) in (11.58b) are also not symmetric and therefore the 2-norm does not seem to be appropriate. However, as we can see from (11.98), the lack of symmetry of \( A \) only occurs at two places and can easily be overcome. To that end we introduce the scaling matrix
\[
S := \text{diag}(\sqrt{2}, 1, 1, \ldots, 1, \sqrt{2}).
\]  \hfill (11.99)

We define a matrix \( \hat{A} \) similar to \( A \) (which implies that \( \hat{A} \) has the same eigenvalues as \( A \)) as
\[
\hat{A} := S^{-1} AS.
\]  \hfill (11.100)
11.6. Mixed Boundary Conditions

Then we find

\[
\hat{A} = \frac{1}{\Delta x^2} \begin{pmatrix}
-2(1 + \alpha \Delta x) \sqrt{2} & \sqrt{2} & -2 & 1 \\
\sqrt{2} & -2 & 1 & \ddots & \ddots & \ddots \\
-2 & 1 & -2 & 1 & \ddots & \ddots & \ddots \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
1 & -2 & 1 & \sqrt{2} & \ddots & \ddots & \ddots \\
1 & -2 & 1 & -2 & \sqrt{2} & -2(1 + \beta \Delta x)
\end{pmatrix}
\]  

(11.101)

Clearly, \(\hat{A}\) is symmetric. For ease of notation let us introduce the matrices (cf. (11.57b) and (11.58b))

\[
\hat{B}_\varphi := S^{-1}B_\varphi S, \\
\hat{B}_{\varphi-1} := S^{-1}B_{\varphi-1}S, \\
\hat{C} := S^{-1}CS.
\]

(11.102a)  

(11.102b)  

(11.102c)

Trivially, these matrices are symmetric, because \(\hat{A}\) is symmetric and the matrices \(\hat{B}_\varphi\) and \(\hat{B}_{\varphi-1}\) commute.

Recalling the expression (11.59) for the global discretisation error vector \(e^n\), we find

\[
e^n = \Delta t \sum_{l=0}^{n-1} C^{n-l-1} B^{-1}_\varphi d' = \Delta t \sum_{l=0}^{n-1} S \hat{C}^{n-l-1} \hat{B}^{-1}_\varphi S^{-1} d'.
\]

(11.103)

Since all matrices involved are symmetric and \(\|S\|_2 = \rho(S) = \sqrt{2}\) and \(\|S^{-1}\|_2 = \rho(S^{-1}) = 1\), we get the following upper bound for \(\|e^n\|_2\):

\[
\|e^n\|_2 \leq \Delta t \sqrt{2} \|\hat{B}^{-1}_\varphi\|_2 \sum_{l=0}^{n-1} \|\hat{C}\|_2^{n-l} \|d'\|_2
\]

\[
= \Delta t \sqrt{2} \rho(B^{-1}_\varphi) \sum_{l=0}^{n-1} \rho(C)^{n-l} \|d'\|_2.
\]

(11.104)

where we used that the matrices \(\hat{B}^{-1}_\varphi\) and \(\hat{C}\) are similar to \(B^{-1}_\varphi\) and \(C\), respectively, and consequently have the same eigenvalues.

Next we want to derive sufficient conditions such that the numerical scheme (11.97) is stable. Let the eigenvalues of the matrix \(A\) be denoted by \(\lambda_k\). Unlike in the derivation in the appendix, Section F, we will not explicitly compute the eigenvalues \(\lambda_k\). Instead, we apply Gershgorin’s theorem to estimate these eigenvalues; see the appendix, Section I. First, note that the matrix \(A\) has real eigenvalues since it is similar to the symmetric matrix \(\hat{A}\). Then by Gershgorin’s theorem we conclude that

\[
\Delta x^2 \lambda_k \in [-4, 0] \cup [-4 - 2\alpha \Delta x, -2\alpha \Delta x] \cup [-4 - 2\beta \Delta x, -2\beta \Delta x].
\]

(11.105)
Consequently, we obtain the following inequalities for the eigenvalues $\lambda_k$:

$$-\frac{4}{\Delta x^2} - \frac{2}{\Delta x} \max(\alpha, \beta) \leq \lambda_k \leq 0.$$  (11.106)

Obviously, the condition $\rho(B^{-1}) < \infty$ is again trivially satisfied. Like we saw in Section 11.4, the requirement $\rho(C) < 1$ leads to the inequalities in (11.66). Combining (11.66) and (11.106), we finally obtain that the method is stable if

$$\begin{cases}
0 \leq \vartheta < \frac{1}{2} \quad \text{and} \quad d \leq \frac{1}{\frac{1}{2} - 2\vartheta} \frac{1}{2 + \Delta x \max(\alpha, \beta)}, \\
\frac{1}{2} \leq \vartheta \leq 1.
\end{cases}$$  (11.107)

We see that for $\alpha, \beta > 0$ the stability condition (11.107) is more restrictive than the stability condition given in (11.66).

We would finally like to address the accuracy problem. If we use central differences, we obtain a discretisation error (i.e., the residual) in (11.92a) of order $O(\Delta x^2)$. As a result, $u_{n+1}^m$ has an $O(\Delta x^3)$ error. After substituting this in (11.93), we get a local error of order $O(\Delta t)$. So we may expect that this will spoil the second order behaviour of the scheme as a whole. Fortunately, this is not the case. We recall Section 9.3.3, where we saw that the (local) discretisation error in the boundary conditions may be two orders lower than the discretisation error in internal grid points and still not spoil the overall accuracy.

In order to apply the theory developed in Section 9.3.3, we need a property.

**Property 11.7.** Let $d \leq 1/(2(1 - \vartheta))$. Then the matrix $C := (I - \vartheta \Delta t A)^{-1}(I + (1 - \vartheta) \Delta t A)$ is positive and satisfies

$$O < \sum_{k=0}^{n-1} C^{n-k} (I - \vartheta \Delta t A)^{-1} \leq -\frac{1}{\Delta t} FA^{-1},$$  (11.108)

where $\|F\|_{\infty} \leq 2$.

**Proof.** Note that the condition for $d$ implies that $I + (1 - \vartheta) \Delta t A > O$, since $I - \vartheta \Delta t A$ clearly is an M-matrix, its inverse is positive as well. Also note that one is not an eigenvalue of $C$, implying that $I - C$ is nonsingular. Moreover, we have

$$(I - C)^{-1} = -\frac{1}{\Delta t} A^{-1}(I - \vartheta \Delta t A).$$

Hence we deduce from (11.83) that

$$\sum_{k=0}^{n-1} C^k (I - \vartheta \Delta t A)^{-1} = (I - C^n)(I - C)^{-1}(I - \vartheta \Delta t A)^{-1} = -(I - C^n)\frac{1}{\Delta t} A^{-1}.$$

Let us now define $F := I - C^n$. It is not complicated to see that

$$\|(I - \vartheta \Delta t A)x\|_{\infty} \geq \|x\|_{\infty}.$$
Hence \( \|(I - \vartheta \Delta t A)^{-1}\|_\infty \leq 1 \). It can also be directly seen that
\[
\|I + (1 - \vartheta)\Delta t A\|_\infty \leq 1.
\]
So we have \( \|C\|_\infty \leq 1 \), whence \( \|C^e\|_\infty \leq 1 \); i.e., \( \|F\|_\infty \leq 2 \).

Like we did in Section 9.3.3, we will split up the vector of local discretisation errors \( d^k \) into a part arising from discretising the boundary and a part from discretising the interior. Thus let \( d_{\text{bnd}} \) and \( d_{\text{int}} \) be estimates such that
\[
\|d^k\|_\infty \leq d_{\text{bnd}} e_{\text{bnd}} + d_{\text{int}} e_{\text{int}}.
\]
(11.109)

In order to find the global error, we will again estimate \( |A^{-1}| e_{\text{bnd}} \) and \( |A^{-1}| e_{\text{int}} \) separately using Property 11.7. For ease of writing we assume \( \alpha = \beta \). We find
\[
-A e_{\text{bnd}} \geq 2 x^2 / (2 \alpha),
\]
whence
\[
\|A^{-1}| e_{\text{bnd}} \|_\infty \leq x / (2 \alpha).
\]
Moreover, the vector \( v := (v_0, v_1, \ldots, v_{M+1})^T \) with \( v_j := - \frac{1}{2} (x_j^2 - x_{j-1}) \) yields
\[
-A v \leq e.
\]
Note that this vector can be found as from the continuous problem, for which the function \( u(x) = - \frac{1}{2} x^2 + \frac{1}{2} x + 1 / (2 \alpha) \) satisfies the boundary conditions and is numerically differentiated exactly. We therefore obtain
\[
|A^{-1}| e_{\text{bnd}} \leq \left( \frac{1}{8} + \frac{1}{2} \alpha \right) e.
\]
Finally, we thus obtain the following estimates for the global discretisation error (cf. (11.59)):
\[
\|e^e\|_\infty \leq \frac{2}{\Delta t} \Delta x / \alpha + \frac{1}{\Delta t} \left( \frac{1}{4} + \alpha \right) d_{\text{int}}.
\]
Since \( d_{\text{bnd}} = O(\Delta t \Delta x) \) and \( d_{\text{int}} = O(\Delta t \Delta x^2) \), we have \( \|e^e\|_\infty = O(\Delta x^2) \) only if \( \Delta t \leq \Delta x^2 / (2(1 - \vartheta)) \).

11.7 Problems in Two Space Dimensions

The methods we have encountered so far have obvious counterparts in several space dimensions. In this section we shall restrict ourselves to two dimensions. In order to keep the derivations in this section simple, we assume that the domain \( \Omega \) is a rectangle in \( \mathbb{R}^2 \), say \( \Omega := \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x \leq a, 0 \leq y \leq b\} \) for some \( a, b > 0 \). Then we look for the solution \( u(x, y, t) \) of the initial boundary value problem
\[
\begin{align*}
\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f(x, y, t), \quad (x, y) \in \Omega, \quad t > 0, \quad (11.110a) \\
u(x, y, 0) &= v(x, y), \quad (x, y) \in \Omega, \quad (11.110b) \\
u(x, y, t) &= g(x, y, t), \quad (x, y) \in \partial \Omega, \quad t > 0, \quad (11.110c)
\end{align*}
\]
where \( f, v, \) and \( g \) are given functions.
In order to compute a numerical solution, we cover $\Omega$ with a rectangular grid with grid sizes $\Delta x$ in the $x$ direction and $\Delta y$ in the $y$ direction, with $M + 2$ and $N + 2$ lines, respectively, i.e., we have $(M + 2)^2 (N + 2)^2$ grid points, including the boundary points, which is the same as $M^2 N^2$ interior points; see Figure 11.10.

Let us indicate the approximation of $u(j\Delta x, k\Delta y, t) =: u(x_j, y_k, t)$ by $u_{jk}(t)$. As remarked in Chapter 9, there is a host of possibilities to order the points in the spatial grid. As before, we shall assume that they are ordered lexicographically; see Figure 11.10. We then define the solution vector as

$$u(t) := (u_{11}(t), \ldots, u_{M1}(t) | u_{12}(t), \ldots, u_{M2}(t) | \cdots | u_{1N}(t), \ldots, u_{MN}(t))^T.$$  

This leads via the MOL to the ODE system

$$\frac{du}{dt} = Au + f(t),$$  

where the matrix $A$ is defined by

$$A := \begin{pmatrix} \alpha & \beta & \gamma \\ \beta & \alpha & \beta \\ \cdot & \cdot & \gamma \\ \cdot & \cdot & \cdot \\ \beta & \alpha & \beta \\ \cdot & \cdot & \gamma \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \gamma & \cdot \\ \cdot & \gamma & \cdot \\ \gamma & \cdot & \cdot \\ \gamma & \cdot & \cdot \\ \cdot & \gamma & \cdot \\ \cdot & \cdot & \gamma \\ \cdot & \gamma & \cdot \\ \cdot & \cdot & \cdot \\ \gamma & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \gamma & \cdot \\ \cdot & \cdot & \cdot \\ \gamma & \cdot & \cdot \\ \cdot & \gamma & \cdot \\ \cdot & \cdot & \cdot \\ \gamma & \cdot & \cdot \\ \cdot & \gamma & \cdot \\ \cdot & \cdot & \cdot \\ \gamma & \cdot & \cdot \end{pmatrix}$$  

(11.111b)
11.7. Problems in Two Space Dimensions

and \( f(t) \) contains both the source term and contributions from the boundary points. In (11.111b) we have introduced the shorthand notation

\[
\alpha := -\frac{2}{\Delta x^2}, \quad \beta := \frac{1}{\Delta y^2}, \quad \gamma := \frac{1}{\Delta y^2}. \tag{11.111c}
\]

As indicated in (11.111b), \( A \) is an \( N \times N \) block matrix, each block being an \( M \times M \) matrix. Let us have a look at the explicit Euler method first. If \( u^n_{jk} \) denotes the numerical approximation of \( u(j/\Delta x,k/\Delta y,n/\Delta t) =: u(x_j,y_k,t^n) \) and if we introduce \( f^n_{jk} := f(x_j,y_k,t^n) \), we obtain

\[
 u^{n+1}_{jk} = u^n_{jk} + \frac{\Delta t}{\Delta x^2} \left( 2u^n_{j+1,k} - 2u^n_{j,k} + u^n_{j-1,k} \right) + \frac{\Delta t}{\Delta y^2} \left( 2u^n_{j,k+1} - 2u^n_{j,k} + u^n_{j,k-1} \right) + \Delta t f^n_{jk}, \tag{11.112}
\]

where the diffusion numbers \( d_x \) and \( d_y \) are defined by

\[
d_x := \frac{\Delta t}{\Delta x^2}, \quad d_y := \frac{\Delta t}{\Delta y^2}. \tag{11.113}
\]

Defining the numerical solution vector as

\[
 u^n := \left( u^n_{1,1}, \ldots, u^n_{1,2}, \ldots, u^n_{M,1}, \ldots, u^n_{M,N} \right)^T,
\]

we obtain the following matrix formulation for (11.112):

\[
 u^{n+1} = (I + \Delta t A)u^n + \Delta t f^n, \tag{11.114}
\]
with $f^n := f(t^n)$. It is simple to see that
\[ \| I + \Delta t A \|_\infty = |1 + \alpha \Delta t| + |\alpha \Delta t|. \]
We therefore conclude that the method is stable if $|\alpha \Delta t| \leq \frac{1}{2}$, i.e.,
\[ \Delta t \left( \frac{1}{\Delta x^2} + \frac{1}{\Delta y^2} \right) \leq \frac{1}{2}. \]
This result is a sheer analogue of the one-dimensional time step restriction (11.14). This restriction of the time step is again a result of the inherent stiffness of the problem and is, like in the one-dimensional case, undesirable for time domains where the solution $u$ is relatively smooth. Therefore we should use implicit methods. Consider as an example the implicit Euler method given by
\[ u_{j,k}^{n+1} = u_{j,k}^{n} + \Delta t f_{n}^{j,k}. \]
This system is less trivial to solve than the one-dimensional case. The sparsity structure, i.e., five nonzero (co)diagonals, will be lost if an LU-decomposition is carried out; by this we mean that the lower triangular and the upper triangular matrices will have fill-in, i.e., have potential nonzeros on the first $M$ codiagonals. Moreover, the complexity, i.e., the number of elementary operations, is significantly larger due to this fill-in. A further discussion of this linear algebra topic is outside the scope of this book. We refer to [52]. In the next section we shall consider an alternative that has low complexity and good stability properties.

### 11.8 Splitting Methods

In some situations it will turn out to be useful to make a distinction between implicit and explicit discretisations on the operator level; i.e., if one part of the operator is discretised on the new time level and another part on the old time level, some efficiency gains can be achieved. While such a “splitting” may lead to lower complexity of the computations, it may also interfere with the stability of the overall method. In Section 11.8.1 we treat methods that split the elliptic operator as such and in Section 11.8.2 we discuss methods that deal with advection and diffusion operators separately.
11.8. Splitting Methods

11.8.1 The ADI Method

In order to lower the complexity needed for solving (11.116), we may try to take only the \( x \) derivative implicitly in (11.115). This would give

\[
\begin{align*}
\frac{u_{j,k}^{n+1} - d_t (u_{j,k+1}^{n+1} - u_{j,k}^{n+1})}{\Delta t} &= u_{j,k}^{n+1} + d_t (u_{j,k+1}^{n} - 2u_{j,k}^{n} + u_{j,k-1}^{n}) + \Delta t f_{jk}^*.
\end{align*}
\]

(11.117)

The source term \( f_{jk}^* \) may be evaluated at some combination of the time points \( t^n \) and \( t^{n+1} \).

Introducing the splitting \( A = A_x + A_y \), where the matrices \( A_x \) and \( A_y \) are defined by

\[
A_x := \frac{1}{\Delta x^2}
\begin{pmatrix}
-2 & 1 & & & & \\
1 & -2 & 1 & & & \\
& \ddots & \ddots & \ddots & & \\
& 1 & -2 & 1 & & \\
& & 1 & -2 & & \\
& & & \ddots & \ddots & \ddots \\
& & & 1 & -2 & 1 \\
& & & & 1 & -2 \\
& & & & & \ddots & \ddots \\
& & & & & 1 & -2 \\
& & & & & & \ddots & \ddots \\
& & & & & & 1 & -2 \\
& & & & & & & \ddots & \ddots \\
& & & & & & & 1 & -2 \\
& & & & & & & & 1 & -2 \\
& & & & & & & & & \ddots & \ddots \\
& & & & & & & & & 1 & -2 \\
& & & & & & & & & & \ddots & \ddots \\
& & & & & & & & & & & \ddots & \ddots \\
& & & & & & & & & & & 1 & -2 \\
& & & & & & & & & & & & 1 & -2 \\
\end{pmatrix}
\]

(11.118a)
we can write (11.117) as
\[
(I - \Delta t A_x)u^{n+1} = (I + \Delta t A_y)u^n + \Delta t f^*_x.
\] (11.119)

Note that the coefficient matrix $I - \Delta t A_x$ is tridiagonal again and hence solving the system (11.119) is inexpensive (cf. the one-dimensional case).

Let us now investigate the stability of this system, i.e., $\| (I - \Delta t A_x)^{-1}(I + \Delta t A_y) \|$. We choose the 2-norm again. Recall the transformation matrix $W$ in (11.33) that brings the one-dimensional discretisation matrix in (11.30b) into diagonal form (11.35). We now build a larger matrix, composed of such transformation matrices, $\tilde{W}$ say, defined by
\[
\tilde{W} := \begin{pmatrix}
W \\
W \\
\vdots \\
W
\end{pmatrix}
\] (11.120)
11.8. Splitting Methods

and thus transform \( I - \Delta t A_x \) as follows:

\[
\tilde{W}^{-1}(I - \Delta t A_x)\tilde{W} = \begin{pmatrix}
A_x & & & \\
& A_x & & \\
& & \ddots & \\
& & & A_x
\end{pmatrix},
\]

(11.121)

where \( A_x \) is defined by

\[
A_x := I - \Delta t A;
\]

cf. (11.35). We now obtain

\[
\tilde{W}^{-1}(I - \Delta t A_x)^{-1}\tilde{W}^{\top}(I + \Delta t A_y)\tilde{W} = \begin{pmatrix}
A_x^{-1} & & & \\
& A_x^{-1} & & \\
& & \ddots & \\
& & & A_x^{-1}
\end{pmatrix}(I + \Delta t A_y),
\]

(11.123)

which follows from the simple structure of the matrix \( A_y \). Next we can bring the split matrix \( A_y \) into tridiagonal form, like \( A_x \), if we renumber the unknowns along vertical grid lines. This renumbering corresponds to a permutation of the rows and columns of \( A_y \), and we can formally define a tridiagonal matrix \( B_y \) by

\[
B_y := \tilde{P}^{-1} A_y \tilde{P},
\]

where \( \tilde{P} \) is the corresponding permutation matrix. Applying the same permutations to the matrix in (11.123), we obtain

\[
\tilde{P}^{-1}\tilde{W}^{-1}(I - \Delta t A_x)^{-1}\tilde{W}^{\top}\tilde{P}^{-1}\tilde{W}^{\top}(I + \Delta t A_y)\tilde{W}^{\top} = \tilde{P}^{-1}\begin{pmatrix}
A_x^{-1} & & & \\
& A_x^{-1} & & \\
& & \ddots & \\
& & & A_x^{-1}
\end{pmatrix}\tilde{P}(I + \Delta t B_y).
\]

(11.124)

It is simple to see then that \( \tilde{P} \) transforms the matrix with the \( A_x^{-1} \) blocks into another diagonal matrix:

\[
\begin{pmatrix}
(1 - \Delta t \lambda_1)I & & & \\
& (1 - \Delta t \lambda_2)I & & \\
& & \ddots & \\
& & & (1 - \Delta t \lambda_M)I
\end{pmatrix}.
\]

Since the 2-norm of a matrix does not change when pre- or postmultiplied by orthogonal (permutation) matrices, it is easy to see that

\[
\| (I - \Delta t A_x)^{-1}(I + \Delta t A_y) \|_2 = \max_{1 \leq i \leq M} \frac{1 - 4d_i \sin^2\left(\frac{\pi}{2} y_k\right)}{1 + 4d_i \sin^2\left(\frac{\pi}{2} x_j\right)}.
\]

(11.125)
We conclude that we end up with a similar stability restriction as before and that at this stage not much is gained. However, we now do a second step, thereby alternating the implicit direction, which gave the method its name ADI. In matrix terms the method goes as follows:

\[
(I - \Delta t A_x)u^{n+1} = (I + \Delta t A_x)u^n + \Delta t f_x^n, \quad \text{(11.126a)}
\]

\[
(I - \Delta t A_y)u^{n+2} = (I + \Delta t A_y)u^{n+1} + \Delta t f_y^{n+1}, \quad \text{(11.126b)}
\]

with \(f_x^n\) and \(f_y^n\) the vectors of the source term and boundary values. For stability we now investigate the matrix \(K\) and its similar form \(L\), defined by

\[
K := (I - \Delta t A_x)^{-1}(I + \Delta t A_x)(I - \Delta t A_y)^{-1}(I + \Delta t A_y), \quad \text{(11.127a)}
\]

\[
L := (I - \Delta t A_y)K(I - \Delta t A_x)^{-1}. \quad \text{(11.127b)}
\]

Actually, we need to transform \(L\) first, like we did above in (11.124), and work with the matrix \(L\), say. Then we easily obtain

\[
\|L\|_2 = \|L\|_2 = \max_{1 \leq j \leq M} \left| \frac{1 - 4d_s \sin^2 \left( \frac{x_j}{2} \right)}{1 + 4d_s \sin^2 \left( \frac{x_j}{2} \right)} \right|, \quad \text{(11.128)}
\]

which is obviously bounded by one, independent of the grid sizes. Although \(L\) may be a skew transformation of \(K\), this does not affect the stability as such (one can give a sharper estimate of the increment matrices, which we shall omit here).

There is another way to look at this process. Consider the matrix ODE (11.111a). As discussed for the one-dimensional case, the homogeneous solution of this problem has an increment matrix

\[
\exp(2\Delta t \mathbf{A})
\]

relating the solution at time levels \(t^n\) and \(t^{n+2}\). Since \(\mathbf{A} = \mathbf{A}_x + \mathbf{A}_y\) and the matrices \(\mathbf{A}_x\) and \(\mathbf{A}_y\) commute, i.e., \(\mathbf{A}_x \mathbf{A}_y = \mathbf{A}_y \mathbf{A}_x\), we can split the increment matrix as follows:

\[
\exp(2\Delta t \mathbf{A}) = \exp(2\Delta t \mathbf{A}_x + 2\Delta t \mathbf{A}_y) = \exp(\Delta t \mathbf{A}_x + \Delta t \mathbf{A}_y + \Delta t \mathbf{A}_x) = \exp(\Delta t \mathbf{A}_x) \exp(\Delta t \mathbf{A}_y) \exp(\Delta t \mathbf{A}_x) \exp(\Delta t \mathbf{A}_y). \quad \text{(11.129)}
\]

ADI can thus be interpreted as taking an implicit Euler approximation for the first and third factors in this product and an explicit Euler approximation for the second and fourth factors. There are many other splitting methods, such as local one-dimensional methods; see, e.g., [96].

In (9.67) we have already encountered ADI. For the solution of the linear system \(\mathbf{A} \mathbf{u} + \mathbf{f} = \mathbf{0}\) it reads

\[
(I - \tau \mathbf{A}_x)u^{n+1} = (I + \tau \mathbf{A}_x)u^n + \tau \mathbf{f},
\]

\[
(I - \tau \mathbf{A}_y)u^{n+2} = (I + \tau \mathbf{A}_y)u^{n+1} + \tau \mathbf{f}.
\]
where $\tau$ is a time-step-like parameter that still needs to be determined. These recursions have the same form as (11.126) if we take $\Delta t = \tau$. In order to estimate the convergence rate of ADI as an iterative method for solving Poisson problems, we use the bound for the increment matrix (11.128) and work this out a bit. First we note that if

$$\alpha \leq d\lambda_k \leq \beta \quad (k = 1, 2, \ldots, M),$$

an upper bound for (11.128) (with $M = N$ and $d_x = d_y = d$) is given by

$$\max_{\alpha, \beta} \left( \left| \frac{1 - \tau \alpha}{1 + \tau \alpha} \right|, \left| \frac{1 - \tau \beta}{1 + \tau \beta} \right| \right).$$

Clearly this is minimal if both factors are equal. Hence $\tau$ should satisfy

$$\tau^2 - \frac{1}{\alpha \beta} = 0.$$ 

So we obtain the estimate

$$\tau = \frac{1}{\sqrt{\alpha \beta}} \approx \frac{1}{2} \Delta x.$$

(11.130)

Using this value for $d = \tau / \Delta x^2$ in (11.128), we find that the convergence factor is approximately equal to $(1 - \sin(\pi \Delta x))/(1 + \sin(\pi \Delta x)) \approx 1 - 2 \pi \Delta x$.

### 11.8.2 Mixed Explicit/Implicit Discretisation of the Advection-Diffusion Equation

The stability requirement, forcing the time step to be taken very small or else forcing the use of an implicit method, is one of the most cumbersome problems in solving evolutionary equations. There are various ways to tackle this problem for equations including an advection term, i.e., first order spatial derivatives. So consider first the linear advection-diffusion equation

$$\frac{\partial u}{\partial t} + b \frac{\partial u}{\partial x} = D \frac{\partial^2 u}{\partial x^2},$$

(11.131)

where $b$ and $D > 0$ are the advection velocity and diffusion coefficient, respectively. Applying the MOL to this equation, we obtain an ODE system of the form

$$\frac{du}{dt} = A_d u - A_a u + f,$$

(11.132)

where $A_d u$ originates from discretising the diffusion term $D \frac{\partial^2 u}{\partial x^2}$ and $A_a u$ comes from discretising the advection term $b \frac{\partial u}{\partial x}$. For the diffusion term we usually employ central differences, whereas for the advection term we can choose either central or one-sided differences. In Chapter 13 we will discuss this matter in detail.

Since the discrete diffusion operator $A_d u$ causes the problem to be very stiff on a fine spatial grid, we decide to use the implicit Euler method for time integration. This way, we obtain the linear system

$$(I + \Delta t (A_a - A_d)) u^{n+1} = u^n + \Delta t f^{n+1}.$$  

(11.133)
296 Chapter 11. Numerical Methods for Parabolic Equations

The coefficient matrix $I + \Delta t (A_d - A_a)$ is tridiagonal so that (11.133) can be very efficiently solved at each time step. However, the symmetry is lost due to the term $A_a$. This is not a problem for equations in one space dimension. On the other hand, for two- or three-dimensional problems we have to solve (11.133) iteratively, which is often considerably more expensive than solving a symmetric system by an iterative procedure. Based on these considerations, we might take the advection term explicitly, leading to the system

$$
(I - \Delta t A_d)u^{n+1} = (I - \Delta t A_a)u^n + \Delta t f^{n+1}.
$$

(11.134)

Concerning stability, we will show in Chapter 13 that the explicit advection term leads to a favorable time step restriction of the form $\Delta t / \Delta x \leq C$ for some (small) constant $C > 0$. Other combinations of implicit discretisations of the diffusion term and explicit discretisations of the advection term are of course possible. We leave this as an exercise.

The need to take the advection term explicitly becomes more urgent if the advection term is nonlinear. We will demonstrate this for the nonlinear advection-diffusion equation

$$
\frac{\partial u}{\partial t} + b(u) \frac{\partial u}{\partial x} = D \frac{\partial^2 u}{\partial x^2},
$$

(11.135)

where the advection velocity $b(u)$ is now a (nonlinear) function of $u$. After space discretisation, we obtain the ODE system

$$
\frac{du}{dt} = A_d u - A_a(u)u + f,
$$

(11.136)

where the discretisation matrix $A_a(u)$ for the advection term depends on $u$. Once more applying the implicit Euler method, we find the nonlinear system

$$
(I - \Delta t A_d)u^{n+1} + \Delta t A_a(u^{n+1})u^{n+1} = u^n + \Delta t f^{n+1},
$$

(11.137)

which we have to solve using a Newton-type method. We can circumvent this nonlinear system by using $A_a(u^n)$ instead of $A_a(u^{n+1})$. This way, we obtain the linear system

$$
(I - \Delta t (A_a(u^n) - A_d))u^{n+1} = u^n + \Delta t f^{n+1}.
$$

(11.138)

This method has comparable stability properties to (11.137), whereas it only requires the solution of a linear system at each time step.

11.9 Solving Nonlinear Problems

Nonlinear parabolic equations are in principle not much different from linear ones. In order to solve them, one can often employ the same discretisations as in the linear case. The main difficulty is the solution of the nonlinear algebraic system resulting from an implicit time-stepping method. Yet, or rather because of this, one may take recourse to special methods that alleviate the burden of nonlinearity.

To start with let us investigate stability for a problem with nonlinear diffusion. So consider the equation

$$
\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( D(u) \frac{\partial u}{\partial x} \right) + f(x, t), \quad x \in (0, 1), \quad t > 0.
$$

(11.139)
where the diffusion coefficient $D(u) > 0$ is a function of $u$. We can find a difference scheme that has a symmetric form for the spatial part of the equation; see also (5.32b), where this was derived for the linear case. If we define

$$D^n_j := D(u(x_j, t^n)), \quad D^n_{j+1/2} := \frac{1}{2}(D^n_j + D^n_{j+1}),$$

we have, e.g., with the explicit Euler method, that

$$u^{n+1}_j = u^n_j + d(D^n_{j+1/2}(u^{n+1}_{j+1} - u^n_j) - D^n_{j-1/2}(u^n_j - u^n_{j-1})) + \Delta t f^n_j.$$ \hfill (11.140)

By “freezing” the coefficients, we can investigate stability for a particular value of $D$. Clearly, we have

$$\tilde{d} = \frac{D}{\Delta x^2} \leq \frac{1}{2}. \quad \text{By requiring this to hold for all values of } D \text{ that are relevant for the problem, we can find an intuitively reasonable bound for the time step.}$$

**Example 11.8** If the value of $D$ in (11.139) is given by $D(u) := 1 + e^{-u}$, then $1 < D < 2$ for $u > 0$. In this case stability requires that $\Delta t \leq \frac{1}{(4\Delta x^2)}$. \hfill □

From the foregoing it again follows that one may need a fairly small time step. However, if we want to overcome this by using an implicit method, there is a big problem since the resulting system is nonlinear. For moderate dependencies of $D$ on $u$ we can simply use the values of $u$ found at previous time levels. This gives the following implicit alternative for (11.140):

$$u^{n+1}_j = u^n_j + d(D^n_{j+1/2}(u^{n+1}_{j+1} - u^n_j) - D^n_{j-1/2}(u^n_j - u^n_{j-1})) + \Delta t f^n_j.$$ \hfill (11.141)

Finally, let us consider problems with a nonlinear source term, i.e.,

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(u) \quad x \in (0, 1), \quad t > 0.$$ \hfill (11.142)

In Section 10.6.2 we encountered Fisher's equation as an example of this type. The treatment of this term follows the same principles as before. If $f$ is not strongly dependent on $u$ (i.e., $|f'(u)|$ is not large), then we can again simply take the values at the previous time step, so

$$u^{n+1}_j = u^n_j + \frac{\Delta t}{\Delta x^2}(u^n_{j+1} - 2u^n_j + u^n_{j-1}) + \Delta t f^n_j.$$ \hfill (11.143)

However, if $|f'(u)|$ is large, this may require very small time steps again, as demonstrated in the next example.

**Example 11.9** Consider the problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \exp\left(-\frac{1}{\varepsilon} u\right), \quad x \in (0, 1), \quad t > 0,$$

where $\varepsilon$ is a small positive number. Such equations typically occur in chemical kinetics (Arrhenius law), where $\varepsilon$ is something like the reaction temperature. Clearly, $f'(u) = -f(u)/\varepsilon$. For (11.143) to be stable an intuitively reasonable guess is

$$\Delta t \left(\frac{1}{\Delta x^2} + \frac{1}{\varepsilon} \max_u \exp\left(-\frac{1}{\varepsilon} u\right)\right) \leq 1,$$

where the maximum should be taken over the actually attained values. If $\varepsilon$ is much smaller than $\Delta x^2$, this is obviously a strong time step restriction indeed. \hfill □
For very stiff problems, as in Example 11.9, we have to use an implicit method for the source term. In fact, if we use an explicit time-stepping method for the diffusion term, we have to solve the following nonlinear system for y := u^{n+1}:

\[ F(y) := y - \Delta t f(y) - u^n - \Delta t (A u^n + b^n) = 0, \]  

(11.144)

where \( b^n \) is a term coming from the boundary conditions and \( A \) is the matrix in (11.30b). In order to find the solution of (11.144), we can use Newton’s method, which produces a sequence \( y^{(1)}, y^{(2)}, \ldots \), with \( y^{(l)} \to u^{n+1} \), satisfying

\[ y^{(l+1)} = y^{(l)} - J^{-1}(y^{(l)}) F(y^{(l)}). \]  

(11.145a)

Here \( J(y) \) is the Jacobi matrix given by

\[ J(y) = I - \Delta t \text{diag}(f'(y_j)). \]  

(11.145b)

A suitable initial guess for (11.145) is \( y^{(0)} = u^n \), provided that the time step \( \Delta t \) is small enough.

### 11.10 Stefan Problems

In Section 10.5 we encountered parabolic problems with moving boundaries or interfaces. In this section we will discuss some numerical issues, though it should be stated that we cannot cover this vast area in any detail in this book. Let us start with the following simple one-phase problem (cf. (10.43) to (10.45)):

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, & x \in (0, S(t)), & t > 0, \\
u(0, t) - \beta \frac{\partial u}{\partial x}(0, t) &= 1, & t > 0, \\
u(S(t), t) &= 0, & t > 0, \\
\frac{\partial u}{\partial x}(S(t), t) &= -\alpha \frac{dS}{dt}, & S(0) = 0, & t > 0.
\end{align*}
\]  

(11.146a-d)

In the following we will employ a fixed spatial grid. The points to the right of the interface \( x = S(t) \) are redundant and may be discarded. During the computation new grid points are added if the boundary passes by one or a few grid points. Typically, we have a grid as sketched in Figure 11.11. Here the boundary point \( x = S(t^n) \) lies between the grid points \( x_{i(n)} \) and \( x_{i(n)+1} \). Let \( S^n \) be the numerical approximation of \( S(t^n) \). The distance between \( S^n \) and \( x_{i(n)} \) is written as

\[ S^n - x_{i(n)} =: r^n \Delta x, \quad 0 < r^n < 1. \]

We want to apply the explicit Euler time integration method to (11.146). Applying the explicit Euler central difference scheme (11.4) to (11.146a) in any grid point not neighbouring the boundary \( S^n \) is trivial. However, for the grid point \((x_{i(n)}, t^n)\) the scheme
The next step is to use condition (11.146d). In order to find an approximation of \( u(\mathbf{S}^n, t^n) \), we employ the derivative of the second order Lagrangian interpolation polynomial through the grid points \( (x_{i(n)} - 1, t^n), (x_{i(n)}, t^n) \), and \( (\mathbf{S}^n, t^n) \); see Example 5.2. Given the values of \( u \) in these points, we find
\[
\frac{\partial^2 u}{\partial x^2}(\mathbf{S}^n, t^n) \approx \frac{2}{\Delta x^2} \left( \frac{1}{1 + r^n} u(x_{i(n)} - 1, t^n) - \frac{1}{r^n} u(x_{i(n)}, t^n) \right). \tag{11.147}
\]
where we have used the condition \( u(\mathbf{S}^n, t^n) = 0 \) because of (11.146c). The explicit Euler scheme in \( (x_{i(n)}, t^n) \) then reads
\[
u_{i(n)}^{n+1} = u_{i(n)}^n + 2d \left( \frac{1}{1 + r^n} u_{i(n)}^n - \frac{1}{r^n} u_{i(n)}^n \right). \tag{11.148}
\]
The next step is to use condition (11.146d). In order to find an approximation of \( \frac{\partial u}{\partial x}(\mathbf{S}(t^n), t^n) \), we employ the derivative of the second order Lagrangian interpolation polynomial through the grid points \( (x_{i(n)} - 1, t^n), (x_{i(n)}, t^n) \), and \( (\mathbf{S}^n, t^n) \); see Example 5.2. Given the values of \( u \) in these points, we find
\[
\frac{\partial u}{\partial x}(\mathbf{S}(t^n), t^n) \approx \frac{1}{\Delta x} \left( \frac{r^n}{1 + r^n} u(x_{i(n)} - 1, t^n) - \frac{1 + r^n}{r^n} u(x_{i(n)}, t^n) \right). \tag{11.149}
\]
Hence the new interface location is given by
\[
\mathbf{S}^n+1 = \mathbf{S}^n - \frac{\Delta t}{\alpha \Delta x} \left( \frac{r^n}{1 + r^n} u_{i(n)}^n - \frac{1 + r^n}{r^n} u_{i(n)}^n \right). \tag{11.150}
\]
Finally, note that (11.146) has mixed boundary conditions. However, this problem has already been addressed in Section 11.6.

We have to use implicit methods to overcome the inherent stiffness resulting from discretising the term \( \frac{\partial^2 u}{\partial x^2} \). We now demonstrate how to do this for a two-phase problem,
Chapter 11. Numerical Methods for Parabolic Equations

300

describing the moving interface between, say, water (wa) and ice (ic). Thus consider the following problem:

\[
\begin{align*}
\frac{\partial u_{wa}}{\partial t} &= \frac{\partial^2 u_{wa}}{\partial x^2}, & x \in (0, S(t)), & t > 0, \\
\frac{\partial u_{ic}}{\partial t} &= \frac{\partial^2 u_{ic}}{\partial x^2}, & x \in (S(t), 1), & t > 0, \\
u_{wa}(x, 0) &= u_\ell, & x \in (0, S(0)), \\
u_{ic}(x, 0) &= u_r, & x \in (S(0), 1), \\
u_{wa}(0, t) &= u_\ell, & \quad t > 0, \\
u_{ic}(1, t) &= u_r, & \quad t > 0, \\
u_{wa}(S(t), t) &= u_{ic}(S(t), t) = 0, & \quad t > 0, \\
\frac{\partial u_{ic}}{\partial x}(S(t), t) - \frac{\partial u_{wa}}{\partial x}(S(t), t) &= \frac{\alpha dS}{dt}, & \quad t > 0,
\end{align*}
\]

with \( u_\ell > 0 > u_r \). Since we have to deal with a moving interface, the transversal MOL is a suitable approach; i.e., we discretise all time derivatives first. Suppose we employ the implicit Euler method. This way, for (11.151a) and (11.151b), we have

\[
\begin{align*}
u_{wa}^{n+1} - \Delta t \frac{\partial^2 u_{wa}}{\partial x^2} &= u_{wa}^n, & x \in (0, S^{n+1}), \\
u_{ic}^{n+1} - \Delta t \frac{\partial^2 u_{ic}}{\partial x^2} &= u_{ic}^n, & x \in (S^{n+1}, 1),
\end{align*}
\]

and, for the Stefan condition (11.151g), we have

\[
\frac{d}{dx} u_{ic}^{n+1}(S^{n+1}) - \frac{d}{dx} u_{wa}^{n+1}(S^{n+1}) = \frac{\alpha dS}{dt} (S^{n+1} - S^n).
\]

Note that (11.152) and (11.153) are coupled, since the numerical interface position \( S^{n+1} \) is not known.

Next we have to choose the proper space discretisation. For the second derivatives in (11.152) we take the central difference approximation. For grid points other than \((x_{(n+1)1}, \rho^{n+1})\) and \((x_{(n+1)+1}, \rho^{n+1})\) this is trivial. However, for these two grid points, neighbouring the interface at \( x = S^{n+1} \), we have to apply the central difference approximation on a nonuniform grid. As an example, consider the grid point \((x_{(n+1)+1}, \rho^{n+1})\) just left of the interface. In this point we have the approximation

\[
\frac{d^2}{dx^2} u_{ic}(x_{(n+1)+1}) = \frac{2}{\Delta x^2} \left( \frac{-1}{1 - \rho^{n+1}} u_{ic}(x_{(n+1)1}) + \frac{1}{2 - \rho^{n+1}} u_{ic}(x_{(n+1)+2}) \right),
\]

which we subsequently have to combine with the implicit Euler time integration method. In the derivation of (11.154) we have assumed that \( u_{ic}(S^{n+1}) = 0 \); cf. (11.151f). Proceeding this way, we obtain the linear systems

\[
\begin{align*}
(I - \Delta t A_{wa}) u_{wa}^{n+1} &= u_{wa}^n + \Delta t f_{wa}^{n+1}, \\
(I - \Delta t A_{ic}) u_{ic}^{n+1} &= u_{ic}^n + \Delta t f_{ic}^{n+1}.
\end{align*}
\]
where the discretisation matrices $A_{wa}$ and $A_{ic}$ are given by

\[
A_{wa} := \frac{1}{\Delta x^2} \begin{pmatrix}
-2 & 1 \\
1 & -2 & 1 \\
& \ddots & \ddots & \ddots \\
& & 1 & -2 & 1 \\
& & & 2 & -2 \\
& & & & \frac{1}{1 + r^n+1}
\end{pmatrix},
\]

(11.156)

\[
A_{ic} := \frac{1}{\Delta x^2} \begin{pmatrix}
-2 & 1 \\
1 & -2 & 1 \\
& \ddots & \ddots & \ddots \\
& & 1 & -2 & 1 \\
& & & 1 & -2
\end{pmatrix},
\]

Note that $A_{wa}$ is an $(i(n+1) - 1) \times (i(n+1) - 1)$ matrix, so its dimension is still unknown. Likewise, $A_{ic}$ is an $(M - 1 - i(n+1)) \times (M - 1 - i(n+1))$ matrix with $M$ the total number of spatial grid points. For the two derivatives in (11.153) we use one-sided differences based on the second order Lagrangian interpolation polynomial through $(S^{n+1}, t^{n+1})$ and the two closest grid points on either side of the interface. For the first derivative we have, e.g.,

\[
\frac{d}{dx} u_{ic}(S^{n+1}) = \frac{1}{\Delta x} \left( \frac{2 - r^{n+1}}{1 - r^{n+1}} u(x_{i(n+1)+1}, t^{n+1}) - \frac{1}{2} \frac{r^{n+1}}{1 - r^{n+1}} u(x_{i(n+1)+2}, t^{n+1}) \right),
\]

(11.157)

where we have once more used the condition $u_{ic}(S^{n+1}) = 0$. We can obtain a similar expression for $\frac{d}{dx} u_{wa}(S^{n+1})$. The new location of the interface then follows from (11.153) and is given by

\[
S^{n+1} = S^n - \frac{\Delta t}{\alpha \Delta x} \left( - \frac{2 - r^{n+1}}{1 - r^{n+1}} u_{ic,i(n+1)+1}^{n+1} + \frac{1}{2} \frac{r^{n+1}}{1 - r^{n+1}} u_{ic,i(n+1)+2}^{n+1} \\
+ \frac{r^{n+1}}{1 + r^{n+1}} u_{wa,i(n+1)-1}^{n+1} - \frac{1}{2} \frac{r^{n+1}}{1 + r^{n+1}} u_{wa,i(n+1)}^{n+1} \right).
\]

(11.158)

Summarizing, we have the linear systems (11.155) together with (11.158) for the unknowns $p := u_{wa}^{n+1}$, $q := u_{ic}^{n+1}$, and $\sigma := S^{n+1}$. Furthermore, to facilitate the discussion, we introduce the shorthand notation $I := i(n+1)$ and $R := r^{n+1}$. An iterative solution procedure for (11.155) and (11.158) is the following:

- **initialisation, $l = 0$:**
  - Choose the initial guess $\sigma^{(l)} := S^n$ and set $I^{(l)} = i(n)$ and $R^{(l)} = r^n$. Compute the coefficient matrices in (11.155).
• iteration, \( l = 1, 2, \ldots \):
  
  – Compute \( p^{(l)} \) from (11.155a).
  
  – Compute \( q^{(l)} \) from (11.155b).
  
  – Compute \( \sigma^{(l)} \) from (11.158), determine \( I^{(l)} \) and \( R^{(l)} \), and update the coefficient matrices in (11.155).

Note that when \( I^{(l)} \) changes, then the dimensions of the linear systems in (11.155) also change. We decide to stop the iteration when \(| \sigma^{(l+1)} - \sigma^{(l)} | < \text{tol} \) for some tolerance tol. Typically, one or two iteration steps are sufficient to get a convergent solution, depending on the time step.

**Example 11.10** We have actually computed a numerical solution of (11.151) with \( u_l = 4 \), \( u_r = -1 \), \( \alpha = 1 \), and \( S(0) = 1/3 \). For the grid size and time step we have chosen \( \Delta x = 10^{-2} \) and \( \Delta t = 2 \times 10^{-3} \). The numerical solution after 1, 5, 10, and 50 time steps is shown in Figure 11.12. Clearly, the solution tends to a steady state where derivatives \( \frac{\partial}{\partial x} u^c \) and \( \frac{\partial}{\partial x} u^w \) balance at the interface.

![Figure 11.12. Numerical solution of the Stefan problem (11.151).](image)

Another possible approach for solving (11.151) is to use a coordinate transformation. In particular, if we introduce the normalized coordinates \( \xi \) and \( \eta \) defined by

\[
\xi(x, t) := x/S(t), \quad \eta(x, t) := (1 - x)/(1 - S(t)),
\]

(11.159)
we obtain the following equations instead of (11.151a) and (11.151b).

\[
\frac{\partial u_{wa}}{\partial t} = \frac{1}{S^2} \frac{\partial^2 u_{wa}}{\partial \xi^2} + \frac{\xi}{S} \frac{dS}{dt} \frac{\partial u_{wa}}{\partial \xi}, \quad \xi \in (0, 1), t > 0 \quad (11.160a)
\]

\[
\frac{\partial u_{ic}}{\partial t} = \frac{1}{(S-1)^2} \frac{\partial^2 u_{ic}}{\partial \eta^2} + \frac{\eta}{S-1} \frac{dS}{dt} \frac{\partial u_{ic}}{\partial \eta}, \quad \xi \in (0, 1), t > 0 \quad (11.160b)
\]

Note that both equations are defined on the interval \((0, 1)\). Furthermore, for the interface condition we have

\[
\frac{1}{S-1} \frac{\partial u_{ic}}{\partial \eta}(1, t) - \frac{1}{S} \frac{\partial u_{wa}}{\partial \xi}(1, t) = \frac{dS}{dt}. \quad (11.161)
\]

Analogously to the above, we can apply the implicit Euler central difference scheme to the system (11.160) and (11.161). We will not go into further detail.

11.11 Discussion

• As appears from the text, the stability of a scheme is of paramount importance when approximating the solution of a parabolic equation. Generally speaking, explicit methods require a very small time step for stability, making them prohibitively expensive. Therefore we seem to have to use implicit schemes, which do not suffer from a time step restriction. However, the effort in solving the resulting equations is often so large that explicit methods may still be competitive. This is even more the case when there is high temporal activity.

• Problems often exhibit convection as well as diffusion. Technically, problems with a large convection term are still parabolic. However, these convection-dominated problems “behave” more like hyperbolic problems, which is why we defer their discussion to Chapter 13; see Section 13.3.

• In Chapter 16 we discuss a few problems of a parabolic nature. In Section 16.1 we encounter a problem that deals with the effect of heat in the pressing of (resin) plates. A thermal explosion of materials can be found in Section 16.3, the melting of material (a free boundary problem) in laser drilling in Section 16.12, and a pulse-tube cooler in Section 16.14.

Exercises

11.1. Consider the heat equation (11.1a) defined on the domain \(\Omega := [0, 1] \times [0, \infty)\). Suppose that we discretise both \(\frac{\partial}{\partial t} u\) and \(\frac{\partial^2}{\partial x^2} u\) by central differences, so that we have the scheme

\[
\frac{1}{2\Delta t} (u_{j}^{n+1} - u_{j}^{n-1}) = \frac{1}{\Delta x^2} (u_{j+1}^{n} + u_{j-1}^{n} - 2u_{j}^{n+1}).
\]

Show that this scheme is unconditionally unstable.
11.2. Suppose we apply the second order backward difference formula (BDF2) to the ODE system (11.30a).
(a) Give the resulting scheme.
(b) Compute the local discretisation error and show that it is second order in $\Delta t$ and $\Delta x$.
(c) Investigate the stability of the scheme.

11.3. Consider the problem of Section 11.6 with mixed boundary conditions.
(a) Show that on a finite time interval the requirement $\rho(C) \leq 1 + O(\Delta t)$ is sufficient for stability.
(b) How is the stability restriction (11.107) changed by this?

11.4. Rather than using virtual grid points to solve a problem with mixed boundary conditions, we can employ one-sided differences.
(a) Find the scheme that would result from discretising (11.88c) with the forward difference approximation and (11.88d) with the backward difference approximation.
(b) Show that the method found in part (a) is stable.
(c) If we used the backward difference approximation for (11.88c) and the forward difference approximation for (11.88d), we would find an unstable method. Prove this.

11.5. Consider the initial boundary value problem
\[
\frac{\partial u}{\partial t} + b \frac{\partial u}{\partial x} = D \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0,
\]
\[u(x, 0) = 0, \quad x \in (0, 1),
\]
\[u(0, t) = u(1, t) = \sin t, \quad t > 0.
\]
We want to solve this problem with the explicit Euler central difference scheme. Give sufficient conditions for $\Delta t, \Delta x$ to have a stable scheme. Hint: Use Gershgorin’s theorem; see the appendix, Section I.

11.6. Consider problem (11.88) with $\alpha = 0$ in (11.88c); i.e., we have a Neumann boundary condition at $x = 0$.
(a) Show that the discretisation matrix $A$ in (11.98) is nonsingular.
(b) Investigate the stability and accuracy of the $\vartheta$ method.

11.7. Consider the viscous Burgers’ equation
\[
\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left( \frac{1}{2} u^2 \right) = \frac{\partial^2 u}{\partial x^2}, \quad x \in (-1, 1), \quad t > 0.
\]
(a) Show that a particular solution is given by
\[u(x, t) = 1 - 2\alpha \tanh (\alpha (x - t)), \quad \alpha \in \mathbb{R}.
\]
(b) Plot $u(x, t)$, taking $\alpha$ rather large.
Exercises 305

(c) Compute a numerical solution, choosing the initial and boundary conditions corresponding to the given solution. Give a proper motivation for why you chose a particular method.

11.8. Consider scheme (11.141). Compute the local discretisation error and show that it is second order in $\Delta x$ and first order in $\Delta t$.

11.9. The DuFort–Frankel scheme for the heat equation reads

$$u_{j}^{n+1} = u_{j}^{n-1} + 2d(u_{j+1}^{n} - (u_{j+1}^{n+1} + u_{j-1}^{n})�

(a) Obviously this is a two-step method (three time levels involved). What is the stencil of this method?
(b) Show that this method has a local discretisation error estimated by $O(\Delta t^2) + O(\Delta x^2) + O((\Delta x/\Delta t)^2)$.
(c) Assume we have homogeneous Dirichlet boundary conditions. Show that the method in matrix-vector form can be written as

$$(1 + 2d)u_{n+1} = (1 - 2d)u_{n-1} + 2dBu^n,$$

where the matrix $B$ is given by

$$B := \begin{pmatrix}
0 & 1 \\
1 & 0 & 1 \\
& \ddots & \ddots & \ddots \\
& & 1 & 0 & 1 \\
& & & 1 & 0
\end{pmatrix}.$$n

(d) Show that the matrix $\frac{1}{\Delta x^2}B$ has eigenvalues $\lambda_k + \frac{2}{\Delta x^2}$ ($k = 1, 2, \ldots, M$), where $\lambda_k$ are the eigenvalues of the discretisation matrix $A$ in (11.30b).
(e) Use the matrix $W$ in (11.33) to transform the matrix recursion to diagonal form and show that it is unconditionally stable.

11.10. The leapfrog scheme is based on using central differences for both the space and the time derivatives. Applied to the heat equation it reads

$$u_{j}^{n+1} = u_{j}^{n-1} + 2d(u_{j+1}^{n} - 2u_{j}^{n} + u_{j-1}^{n}).$$

(a) Give the stencil of this scheme.
(b) Find a matrix-vector formulation of this method, like in Exercise 11.9.
(c) Bring the recursion in part (b) to diagonal form and show that it is unstable.

11.11. Consider the following spherically symmetric problem:

$$\frac{\partial u}{\partial t} = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right), \quad r \in (0, 1), \quad t > 0,$$

$$u(r, 0) = 1, \quad r \in (0, 1),$$

$$\frac{\partial u}{\partial r} (0, t) = 0, \quad u(1, t) = 0, \quad t > 0.$$n

For space discretisation we use the finite volume method.
11.12. Consider the problem
\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad (x, y) \in \Omega, \quad t > 0,
\]
with domain \( \Omega := \{(x, y) \in \mathbb{R}^2 \mid x^2/4 + y^2 \leq 1 \} \). This domain is covered by a uniform grid with grid size \( h \). This means that we have to use different approximations for grid points near the boundary of \( \Omega \), which are first order in \( h \). Show that the resulting (large) local discretisation error does not influence the global error.

11.13. Consider the advection-diffusion equation (11.131). For space discretisation we use central differences. To integrate the resulting ODE system (11.132) we use the second order Adams–Bashforth scheme for the advection term and the Crank–Nicolson scheme for the diffusion term; i.e.,
\[
\frac{1}{\Delta t} (u^{n+1} - u^n) = \frac{1}{2} (A_d u^n + A_d u^{n+1}) - \frac{1}{2} (3A_a u^n - A_a u^{n-1}) + f.
\]
(a) Give the scheme.
(b) Compute the local discretisation error and show that it is second order in \( \Delta t \) and \( \Delta x \).
(c) Investigate the stability of the scheme.

11.14. Consider Fisher’s equation, i.e.,
\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(1 - u).
\]
(a) Give the implicit Euler scheme for this equation.
(b) Give the nonlinear system (11.144) for \( y := u^{n+1} \).
Chapter 12

Analysis of Hyperbolic Equations

This chapter is devoted to hyperbolic problems. We mainly restrict ourselves to equations with one space variable only. In Section 12.1 we start with first order scalar equations and describe solutions in terms of characteristics. Difficulties are encountered for nonlinear equations, for which characteristics may intersect or fan out. We introduce the corresponding solutions, i.e., shock waves and rarefaction waves. It turns out that a classical approach, with smooth solutions, is not suited to treating these phenomena. For that reason, we introduce weak solutions in Section 12.2. As an important example, we investigate the solution of the Riemann problem, which is an initial value problem with piecewise constant initial condition. We extend the above to systems in the next two sections. First, in Section 12.3, we introduce the definition of hyperbolicity. A system is hyperbolic if its Jacobi matrix has real eigenvalues and linearly independent eigenvectors. This then means that the system can be diagonalised. Subsequently, we can apply the scalar theory to each of the resulting equations. Furthermore, for a linear system of two unknowns we give a concise description of the method of characteristics. Next, in Section 12.4, we consider weak solutions of the Riemann problem. In particular, we introduce the elementary solutions, i.e., shock waves, rarefaction waves, and contact discontinuities. In Section 12.5 we apply this theory to the shallow-water equations. In Section 12.6 the wave equation is treated. We derive the solution in one space dimension first and then study higher-dimensional problems. The last section deals with choosing the proper boundary conditions. Since the characteristics determine the propagation directions of the solution, these determine what boundary conditions can be prescribed.

12.1 First Order Scalar Equations

In Chapter 2 we introduced first order scalar equations. In this section we shall have a closer look at these equations, starting with the linear case. After that, we will deal with the nonlinear case in more detail. A crucial element in the construction of solutions will turn out to be the behaviour of the characteristics.
12.1.1 Semilinear Equations

Consider the first order scalar equation (cf. (2.1))

\[
a \frac{\partial u}{\partial t} + b \frac{\partial u}{\partial x} = c \quad (a, b \neq 0).
\]  \hspace{1cm} (12.1)

Equation (12.1) is a scalar hyperbolic equation. In this subsection we allow the coefficients \(a, b\), and \(c\) to depend on \(x\) and \(t\) but not on \(u\), so that we have a linear equation in \(u\). We showed in Chapter 2 that the PDE (12.1) leads to the set of ODEs

\[
\frac{dx}{dt} = \frac{b}{a}, \quad (12.2a)
\]

\[
\frac{du}{dt} = \frac{c}{a}. \quad (12.2b)
\]

Equation (12.2a) defines the location of the (base) characteristics \(C\) in the \((x, t)\) plane and (12.2b) defines the solution \(u\), as a function of \(t\), along these base characteristics. In the following we will omit the adjective “base,” and use the term “characteristics” to denote solution curves of (12.2a) in the \((x, t)\) plane. In order to find solutions of (12.1), it will be sufficient to study the ODEs (12.2). Let us first consider the homogeneous problem, i.e., \(c(x, t) = 0\). If \(u\) is given at a single point on some characteristic \(C\), it is completely determined along \(C\) by virtue of (12.2b). Consequently, we may not prescribe a solution \(u(x, t)\) on a characteristic unless it satisfies (12.2b). The solution would then exist on this single characteristic only. Thus consider a curve \(J\) that is not a characteristic and that intersects each characteristic at most once. So for some interval \(I \subset \mathbb{R}\) let

\[
J := \{ (x(\sigma), t(\sigma)) \mid \sigma \in I \}. \quad (12.3)
\]

Then we may prescribe \(u\) on \(J\), say

\[
u(x(\sigma), t(\sigma)) = v(\sigma), \quad \sigma \in I, \quad (12.4)
\]

where \(v\) is some given function. We can now compute \(u\) from the differential equations (12.2) and the condition (12.4).

Next we assume that \(a\) and \(b\) are constant. Special cases are when \(J\) coincides with either the \(x\) axis or the \(t\) axis. In the first case we have

\[
J = \{ (x(\sigma), 0)^T \mid \sigma \in \mathbb{R} \},
\]

and we may identify \(\sigma\) with \(x\). The characteristic intersecting \(J\) at a point \((x_0, 0)^T\), say, is given by

\[
x - \frac{b}{a} t = x_0.
\]

The solution along this characteristic reads

\[
u(x, t) = v(x_0) = v \left( x - \frac{b}{a} t \right). \quad (12.5)
\]
12.1. First Order Scalar Equations

In a similar vein, if we have
\[ \mathcal{J} = \{(0, t(\sigma))^T \mid \sigma \geq 0\}, \]
we may identify \( \sigma \) with \( t \). At the characteristic intersecting \( \mathcal{J} \) at a point \((0, t_0)^T\) we then have
\[ t - \frac{a}{b} x = t_0. \]
Consequently, for the solution along this characteristic we find
\[ u(x, t) = v(t_0) = v \left( t - \frac{a}{b} x \right). \quad (12.6) \]
Note that \( b/a \) can be interpreted as a velocity, implying that the solution in (12.5) is just the initial solution \( v(x) \) propagated over a distance \( bt/a \). A similar interpretation holds for (12.6).

Example 12.1 Consider the PDE
\[ \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0. \]
The characteristics of this equation are given by \( x - t = C \) (\( C \in \mathbb{R} \)). If we prescribe \( u(x, t) = v(t) \) on \( \mathcal{C} = \{(x, t) \mid x = t\} \), i.e., the characteristic through the origin, we apparently need \( v(t) \) to be constant. The solution then only exists on \( \mathcal{C} \). Prescribing \( u(x, t) \) along any line \( \mathcal{J} \) that is not parallel to this curve \( \mathcal{C} \) will then be meaningful. Suppose we take for \( \mathcal{J} \) the \( x \)-axis and choose
\[ u(x, 0) = \sin x. \]
Clearly, we then have
\[ u(x, t) = \sin(x - t), \]
implying that \( u(x, t) \) is indeed constant along the characteristics.

If \( a \) and \( b \) are not constant and \( c \neq 0 \), we can still find an integral representation of the solution along characteristics from (12.2b). Suppose a characteristic \( \mathcal{C} \) intersects the “initial” line \( \mathcal{J} \) at the point \((x_0, t_0)^T = (x(s_0), t(s_0))^T\). We can express the value \( u(x, t) \), for \((x, t)^T \in \mathcal{C} \), in terms of the initial value \( u(x(s_0), t(s_0)) \) as follows:
\[ u(x, t) = u(x(s_0), t(s_0)) + \int_{t_0}^{t} \frac{c(x(\tau), \tau)}{a(x(\tau), \tau)} \, d\tau. \quad (12.7) \]
In particular, if \( u \) is given along the \( x \)-axis, i.e.,
\[ u(x, 0) = v(x), \]
and if the coefficients \( a \) and \( b \) are constant, we find from (12.7) that
\[ u(x, t) = v(x_0) + \frac{1}{a} \int_{0}^{t} c(x(\tau), \tau) \, d\tau \]
\[ = v \left( x - \frac{b}{a} t \right) + \frac{1}{a} \int_{0}^{t} c \left( x - \frac{b}{a} (t - \tau), \tau \right) \, d\tau, \quad (12.8) \]
where we have used that \( x - \frac{b}{a} t = x_0 = x(\tau) - \frac{b}{a} \tau \) along characteristics. Also, for more general \( c \), i.e., \( c = c(x, t, u) \), in which case the PDE is sometimes called semilinear, it is fairly simple to find a solution.
Example 12.2 Consider the initial value problem
\[
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = -u, \quad x \in \mathbb{R}, \quad t > 0,
\]
\[u(x, 0) = \sin x, \quad x \in \mathbb{R}.
\]
From (12.2) we conclude that \(u(x(t), t) = Ce^{-t} (C \in \mathbb{R})\) along the characteristic with \(x(t) - t = x_0\). Applying the initial condition, we obtain \(u(x_0, 0) = \sin x_0 = C\). So the solution is given by
\[u(x, t) = e^{-t} \sin (x - t).
\]

If \(a\) and \(b\) are not constant, the characteristics are not straight lines in general. Yet the preceding outline for the construction of a solution is still valid, as is shown by the following example.

Example 12.3 Consider the initial value problem
\[
\frac{\partial u}{\partial t} + b(x, t, u) \frac{\partial u}{\partial x} = c(x, t, u),
\]
\[(12.9)\]
Moreover, if we assume that \(b = b(u)\), i.e., \(b\) does not explicitly depend on \(x\) or \(t\), we can define a flux function \(f(u)\) as
\[f(u) := \int_{u_0}^u b(v) \, dv, \quad (12.10)\]
where \(u_0\) is some reference value for \(u\). We can then rewrite equation (12.9) in conservation form as
\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = c(x, t, u).
\]

12.1.2 Quasi-linear Equations

If the coefficients \(a, b,\) and \(c\) depend on \(u\) as well, we call the PDE quasi-linear; cf. Example 1.2. Since we have assumed that \(a \neq 0\), it is not restrictive to take \(a(x, t, u) \equiv 1\). So consider the equation
\[
\frac{\partial u}{\partial t} + b(x, t, u) \frac{\partial u}{\partial x} = c(x, t, u).
\]
\[(12.9)\]

Moreover, if we assume that \(b = b(u)\), i.e., \(b\) does not explicitly depend on \(x\) or \(t\), we can define a flux function \(f(u)\) as
\[f(u) := \int_{u_0}^u b(v) \, dv, \quad (12.10)\]
where \(u_0\) is some reference value for \(u\). We can then rewrite equation (12.9) in conservation form as
\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = c(x, t, u).
\]

(12.11)
12.1. First Order Scalar Equations

Equation (12.11) has the form of a conservation equation. It is also known as a transport equation.

Below we consider the case \( c(x, t, u) \equiv 0 \). In order to determine a solution of (12.11), or equivalently (12.9), we proceed like before. So the governing equations for the solution along a characteristic \( C \) read

\[
\frac{dx}{dt} = b(u) = f'(u), \quad (12.12a)
\]

\[
\frac{du}{dt} = 0. \quad (12.12b)
\]

From (12.12a) we see that the location of characteristics depends on \( u \), in contrast to the linear case where we could compute the characteristics independently from the solution \( u \). We can, in principle, compute \( u \) from these equations if \( u \) is given on some initial curve \( J \) intersecting the characteristics at most once. Let \( J \) be the \( x \) axis. Then

\[
u(x, 0) = v(x) \quad (12.13)
\]

for some given function \( v \). From (12.12b) we conclude that \( u \), and therefore also \( b(u) \), is constant along a characteristic. Integration of the differential equations in (12.12) is then trivial, and we find the solution

\[
x - x_0 = b(v(x_0))t, \quad (12.14a)
\]

\[
u(x, t) = v(x_0) = v(x - b(v(x_0)))t, \quad (12.14b)
\]

which holds on the characteristic through the point \( (x_0, 0)^T \); see also Example 2.2. Equation (12.14a) implicitly defines \( x_0 \) as a function of \( x \) and \( t \); i.e., \( x_0 = x(x, t) \). By substituting the latter relation into (12.14b), we can find the solution \( u(x, t) \).

A well-known example of (12.9) is the (inviscid) Burgers’ equation. (In Chapter 12 and 13 we simply write “Burgers’ equation” to denote the inviscid Burgers’ equation.)

**Example 12.4** Consider the Burgers’ equation

\[
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0.
\]

Clearly \( b(u) = u \) and we have \( u(x, t) = v(x - v(x_0))t \) along the characteristics given by \( x - x_0 = v(x_0)t \). \( \square \)

As before, we note that the initial condition is propagated along characteristics, now with a speed that depends on the location. This induces a problem, certainly from a mathematical point of view.

**Example 12.5** Consider the Burgers’ equation again. Let the initial condition \( v(x) \) be piecewise constant, say

\[
v(x) := \begin{cases} 
\alpha & \text{if } x < 0, \\
\beta & \text{if } x \geq 0.
\end{cases}
\]

The characteristic through a point \( (x_0, 0)^T \) is given by

\[
x = x_0 + v(x_0)t, \quad x_0 \in \mathbb{R},
\]
and has a slope $\frac{dt}{dx} = 1/v(x_0)$. Therefore we can encounter two typical situations: $\alpha > \beta$ and $\alpha \leq \beta$. In the first case the slope of characteristics emanating from the negative $x$-axis is smaller than the slope of characteristics emanating from the positive $x$ axis; see Figure 12.1. This would imply a multivalued solution where characteristics intersect. In the second case the characteristics on the left have a larger slope than the characteristics on the right, leading to a wedge-shaped region in the $(x, t)$ plane where the solution is not defined.

![Figure 12.1. Characteristics of the Burgers’ equation that intersect (left) or fan out (right).](image)

Apparently, the nonlinearity is causing the problems met in Example 12.5. We emphasize that this is not necessarily a consequence of discontinuous initial data. To see this we shall analyse the influence of the initial condition in more detail. For simplicity we restrict ourselves to the Burgers’ equation. Consider as an example the initial condition $v(x) = \sin \pi x$ ($0 \leq x \leq 1$) shown in Figure 12.2; $v(x)$ is monotonically increasing on $(0, \frac{1}{2})$ and monotonically decreasing on $(\frac{1}{2}, 1)$. From (12.14) we conclude that the initial condition $v(x)$ is propagated along characteristics with velocity $b(v(x)) = v(x)$. This means that the characteristics emanating from $(0, \frac{1}{2})$ fan out and consequently the initial solution on this interval expands. On the other hand, characteristics originating from $(\frac{1}{2}, 1)$ are approaching each other, leading to a compression of the initial solution on this interval. This means that the left part of the solution overtakes the right part, leading to an increasingly steep profile, as shown in Figure 12.2. The solution will eventually break down when $\frac{\partial u}{\partial x}$ tends to infinity at some point $(x^*, t^*)^T$, say, where a discontinuity starts. We can compute $t^*$ as follows. Consider the characteristic through $(x_0, 0)^T$ with $x_0 \in (\frac{1}{2}, 1)$ where the initial solution is monotonically decreasing. The location of this characteristic is given by the relation (12.14a), from which we can deduce

$$
(1 + v'(x_0)t) \frac{\partial x_0}{\partial x} = 1. \tag{12.15}
$$

Furthermore, the solution $u$ along this characteristic is implicitly given by (12.14b). From this relation and (12.15) we can compute $\frac{\partial u}{\partial x}$ to find

$$
\frac{\partial u}{\partial x} = v'(x_0) \frac{\partial x_0}{\partial x} = v'(x_0)(1 + v'(x_0)t)^{-1}. \tag{12.16}
$$
12.1. First Order Scalar Equations

Figure 12.2. The solution of (12.12) at \( t = 0 \) and 0.2 and \( t^* = 1/\pi \) and 0.8 for the initial condition \( v(x) = \sin \pi x \).

Obviously \( \frac{\partial u}{\partial x} \to -\infty \) when \( 1 + v'(x_0) t^* = 0 \) for some \( t^* > 0 \). Note that this condition also implies that \( x_0 \) cannot be determined any more from relation (12.14a). The time \( t^* \) when a discontinuity first emerges is thus given by

\[
t^* = -\frac{1}{\min_{0 \leq x \leq 1} v'(x)}
\]

and only holds when \( v'(x) < 0 \) somewhere.

If we computed the solution from (12.14) beyond the critical time \( t^* \), we would obtain the physically incorrect triple-valued function shown in Figure 12.2. To determine the correct physical behaviour one should realise that the true physical model is an equation of the form

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \varepsilon \frac{\partial^2 u}{\partial x^2},
\]

including the viscous term \( \varepsilon \frac{\partial^2 u}{\partial x^2} \) with \( 0 < \varepsilon \ll 1 \), instead of (12.11) with \( c(x, t, u) = 0 \). Equation (12.11) is an appropriate model only if \( \varepsilon \) is small and the solution is smooth. In this case the viscous term \( \varepsilon \frac{\partial^2 u}{\partial x^2} \) is negligible. However, when a discontinuity starts to develop, (12.11) loses its validity and we must return to (12.18). In the vicinity of the emerging discontinuity the term \( \varepsilon \frac{\partial^2 u}{\partial x^2} \) becomes gradually larger, thus balancing the left-hand side in (12.18) and preventing breakdown of the solution. For decreasing \( \varepsilon \) the solution
becomes gradually steeper. In fact, one can prove that the vanishing viscosity solution, for which $\varepsilon \to 0$, is the discontinuous solution discussed above [87, 70].

### 12.1.3 Nonlinear Equations

Surprisingly, it appears that it is possible to generalize the above results to the general, fully nonlinear case. Under relatively mild conditions of smoothness (derivatives should exist), any first order scalar PDE may be rewritten in characteristic form, i.e., as a system of ODEs. This can be solved when two characteristics do not pass through the same point and boundary conditions are consistent (no more than one boundary condition prescribed on the same characteristic).

**Theorem 12.6 (characteristic form of first order PDEs).** The $n$-dimensional first order nonlinear scalar PDE in $u = u(x)$ given by

$$F(x, u, q) = 0, \quad q := \nabla u,$$

where $F$ is smooth, may be recast into the following system of ODEs

$$\frac{d\chi}{ds} = \frac{\partial F}{\partial q}, \quad \frac{du}{ds} = q \cdot \frac{\partial F}{\partial q}, \quad \frac{dq}{ds} = -q \frac{\partial F}{\partial u} - \frac{\partial F}{\partial x},$$

(12.19)

where $\frac{\partial F}{\partial q}$ and $\frac{\partial F}{\partial x}$ denote the gradients with respect to $q$ and $x$, respectively. The curve $x = \chi(s)$, with parameter $s$, is called a characteristic. Since $s$ is only an auxiliary variable, other equivalent forms exist. They may be constructed by varying the defining equation $F(x, u, q) = 0$.

**Proof.** We have the problem $F(x_1, \ldots, x_n, u, q_1, \ldots, q_n) = 0$, where $q_i = \frac{\partial u}{\partial x_i}$ ($i = 1, 2, \ldots, n$). We look for characteristics, given by $x_i = \chi_i(s)$, along which the PDE can be written as an ODE. If we compare the expressions

$$\frac{dF}{dx_j} = \sum_{j=1}^{n} \frac{\partial q_j}{\partial x_j} \frac{\partial F}{\partial q_j} + \frac{\partial u}{\partial x_i} \frac{\partial F}{\partial u} + \frac{\partial F}{\partial x_i} = 0,$$

$$\frac{dq_i}{ds} = \sum_{j=1}^{n} \frac{\partial q_j}{\partial x_j} \frac{d\chi_j}{ds} = \sum_{j=1}^{n} \frac{\partial^2 u}{\partial x_j \partial x_i} \frac{d\chi_j}{ds} = \sum_{j=1}^{n} \frac{\partial q_j}{\partial x_i} \frac{d\chi_j}{ds},$$

we observe that if we take

$$\frac{d\chi_j}{ds} = \frac{\partial F}{\partial q_j},$$

we obtain the relations

$$\frac{dq_i}{ds} = \sum_{j=1}^{n} \frac{\partial q_j}{\partial x_j} \frac{\partial F}{\partial q_j} = -q_i \frac{\partial F}{\partial u} - \frac{\partial F}{\partial x_i},$$

$$\frac{du}{ds} = \sum_{j=1}^{n} \frac{\partial u}{\partial x_j} \frac{d\chi_j}{ds} = \sum_{j=1}^{n} q_j \frac{\partial F}{\partial q_j},$$

and the problem is indeed rewritten in characteristic form. \qed
12.1. First Order Scalar Equations

Example 12.7 A famous example is the eikonal equation

\[ F(x, u, q) = q \cdot q - \frac{1}{c^2(x)} = 0, \quad q := \nabla u, \]

which becomes, along the characteristics \( x = \chi(s) \),

\[ \frac{d\chi}{ds} = 2q, \quad \frac{du}{ds} = 2q \cdot q = \frac{2}{c^2(x)}, \quad \frac{dq}{ds} = -\frac{2}{c^2(x)} \nabla c. \]

A much neater form, with the same characteristics but a different parametrization, is found if we take

\[ F(x, u, q) = \frac{1}{2}\left(c^2(x) q \cdot q - 1\right) = 0, \quad q := \nabla u. \]

Then we have

\[ \frac{d\chi}{ds} = c^2(x) q, \quad \frac{du}{ds} = c^2(x) q \cdot q = 1, \quad \frac{dq}{ds} = -\frac{1}{c} \nabla c. \]

Obviously, \( u = s + u(0) \), implying that the parametrization corresponds now to the level surfaces of \( u \).

A geometrical interpretation is possible for the special case of two independent variables, say \( x \) and \( y \); see also Section 2.1. Therefore consider the equation

\[ F(x, y, u, p, q) = 0, \quad p := \frac{\partial u}{\partial x}, \quad q := \frac{\partial u}{\partial y}, \quad (12.20) \]

for \( u = u(x, y) \). The solution(s) of (12.20) can be represented as vectors \((x, y, u)^T\) lying on an integral surface \( S \) in \( \mathbb{R}^3 \). All possible integral surfaces of (12.20) at a given point \( P_0(x_0, y_0, u_0) \) have normal vectors \( n = (p, q, -1)^T \), where \( p \) and \( q \) satisfy the nonlinear equation \( F(x_0, y_0, u_0, p, q) = 0 \). This means that all possible normal vectors at \( P_0 \) sweep out a curved surface with apex at \( P_0 \), referred to as the cone of normals. This nomenclature is based on the eikonal equation, for which the cone of normals is actually a right circular cone. Consequently, all possible tangent vectors at \( P_0 \), each of which is perpendicular to the cone of normals, generate a curved surface with apex at \( P_0 \) called the Monge cone. Hence all possible integral surfaces at \( P_0 \) are tangent to the Monge cone.

If (12.20) is quasi-linear, the relation \( F(x_0, y_0, u_0, p, q) = 0 \) is linear in \( p \) and \( q \), defining only one direction vector for the normal at \( P_0 \). In that case the Monge cone reduces to a single characteristic direction.

Suppose that the solution is given along some initial value curve \( J = \{(x(\sigma), y(\sigma), u(\sigma)) \mid \sigma \in I \subset \mathbb{R}\} \). We want to construct the integral surface through \( J \). In the quasi-linear case this is relatively straightforward—compute all characteristics emanating from \( J \). In the fully nonlinear case this approach does not give a unique solution since there is more than one integral surface that is tangent to all Monge cones on \( J \). To determine the generator of the cone that is tangent to the integral surface we proceed as follows. Suppose we parametrize the solution of \( F(x_0, y_0, u_0, p, q) = 0 \); i.e., \( p = p(\lambda) \) and \( q = q(\lambda) \) for some \( \lambda \). Then we have

\[ F(x_0, y_0, u_0, p(\lambda), q(\lambda)) = 0, \quad (12.21) \]

Differentiating with respect to \( \lambda \), we obtain

\[ \frac{\partial F}{\partial p} p'(\lambda) + \frac{\partial F}{\partial q} q'(\lambda) = 0, \quad (12.22) \]
where the derivatives of $F$ should be taken at $(x_0, y_0, u_0, p(\lambda), q(\lambda))$. A tangent plane at $P_0$ on the integral surface is given by

$$u - u_0 = (x - x_0)p(\lambda) + (y - y_0)q(\lambda).$$  (12.23)

The Monge cone is the envelope of all these tangent planes and can be obtained if we differentiate (12.23) with respect to $\lambda$; i.e.,

$$0 = (x - x_0)p'(\lambda) + (y - y_0)q'(\lambda).$$  (12.24)

This relation should hold for the generator of the Monge cone. Comparing (12.22) and (12.24), we conclude that

$$\frac{dx}{F_p} = \frac{dy}{F_q} = ds,$$  (12.25a)

with $s$ the parameter along the generator. Obviously, these relations can be rewritten as

$$\frac{dx}{ds} = \frac{\partial F}{\partial p}, \quad \frac{dy}{ds} = \frac{\partial F}{\partial q}.$$  (12.25b)

Along the generator the solution $u$ satisfies

$$\frac{du}{ds} = \frac{\partial u}{\partial x} \frac{dx}{ds} + \frac{\partial u}{\partial y} \frac{dy}{ds} = \frac{\partial F}{\partial p} + q \frac{\partial F}{\partial q}.$$  (12.26)

Since (12.25b) and (12.26) contain $p$ and $q$, we need differential equations for these variables as well. Differentiating (12.20) with respect to $x$ and $y$, respectively, we obtain

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial F}{\partial p} \frac{\partial p}{\partial x} + \frac{\partial F}{\partial q} \frac{\partial q}{\partial x} = 0,$$  (12.27a)

$$\frac{\partial F}{\partial y} + \frac{\partial F}{\partial u} \frac{\partial u}{\partial y} + \frac{\partial F}{\partial p} \frac{\partial p}{\partial y} + \frac{\partial F}{\partial q} \frac{\partial q}{\partial y} = 0.$$  (12.27b)

Using (12.25b), (12.27), and the consistency relation $p_y = q_x$, we obtain the following ODEs for $p$ and $q$:

$$\frac{dp}{ds} = \frac{\partial p}{\partial x} \frac{dx}{ds} + \frac{\partial p}{\partial y} \frac{dy}{ds} = \frac{\partial F}{\partial p} \frac{\partial p}{\partial x} + \frac{\partial F}{\partial q} \frac{\partial q}{\partial x} = \left( \frac{\partial F}{\partial x} + \frac{\partial F}{\partial u} p \right).$$  (12.28a)

$$\frac{dq}{ds} = \frac{\partial q}{\partial x} \frac{dx}{ds} + \frac{\partial q}{\partial y} \frac{dy}{ds} = \frac{\partial F}{\partial p} \frac{\partial p}{\partial y} + \frac{\partial F}{\partial q} \frac{\partial q}{\partial y} = \left( \frac{\partial F}{\partial y} + \frac{\partial F}{\partial u} q \right).$$  (12.28b)

To summarize, we have the ODE system (12.25b), (12.26), and (12.28). These equations are called the characteristic strip equations or Charpit’s equations. They are a special case of the ODE system in Theorem 12.6.

Initial conditions for $x, y, u$ are determined by the initial curve $\mathcal{J}$; i.e., we have

$$x(0, \sigma) = x_0(\sigma), \quad y(0, \sigma) = y_0(\sigma), \quad u(0, \sigma) = u_0(\sigma).$$  (12.29a)
12.2. Weak Formulation of First Order Scalar Equations

The remaining initial conditions for \( p \) and \( q \) follow from the relations

\[
F \left( x_0(\sigma), y_0(\sigma), u_0(\sigma), p_0(\sigma), q_0(\sigma) \right) = 0, \tag{12.29b}
\]

\[
u_0'(\sigma) = p_0(\sigma)x_0'(\sigma) + q_0(\sigma)y_0'(\sigma), \tag{12.29c}
\]

where the prime \( (\prime) \) denotes differentiation with respect to \( \sigma \). The last relation allows us to move characteristics along the initial curve \( J \) to generate the integral surface \( S \). It follows from the condition that \( J \subset S \).

Using (12.25b), (12.26), and (12.27), we can show that

\[
\frac{d}{ds} F(x(s), y(s), u(s), p(s), q(s)) = 0.
\]

This means that if \( F(x_0, y_0, u_0, p_0, q_0) = 0 \) holds, then the curve determined by the characteristic strip equations and the initial conditions (12.29) satisfies the differential equation \( F(x, y, u, p, q) = 0 \).

The formal solution procedure is as follows. We solve the characteristic strip equations subject to the initial conditions (12.29) to find \( x = x(s, \sigma), y = y(s, \sigma), u = u(s, \sigma), p = p(s, \sigma), \) and \( q = q(s, \sigma) \). Inverting the first two relations, which is only possible if \( x_s y_\sigma - x_\sigma y_s \neq 0 \), we obtain \( s = s(x, y) \) and \( \sigma = \sigma(x, y) \), which we then substitute in the expression for \( u \). In this way we find the representation \( u = u(s(x, y), \sigma(x, u)) =: u^*(x, y) \).

12.2 Weak Formulation of First Order Scalar Equations

As we saw in the previous section, a solution of (12.11) that is computed from the ODE system (12.12) is not necessarily continuous, let alone differentiable. This means that we have to reconsider our concept of a solution of (12.11). For this we need distribution theory; cf. Chapter 4.

12.2.1 Weak Solutions

Let us start by observing that hyperbolic conservation equations are often derived in integral form rather than as a differential equation. As an example, think of gas flowing in a tube of constant cross section. Let \( x \) denote the coordinate along the tube and \( \rho(x, t) \) and \( v(x, t) \) denote the mass density and flow velocity, respectively, at position \( x \) and time \( t \). Then conservation of mass in an arbitrary segment \( (x_1, x_2) \) is given by the relation

\[
\frac{d}{dt} \int_{x_1}^{x_2} \rho(x, t) \, dx = (\rho v)(x_1, t) - (\rho v)(x_2, t), \tag{12.30}
\]

stating that the increase of mass in \( (x_1, x_2) \) is balanced by the net influx of mass. If we replace \( \rho \) by a variable \( u \) and the mass flux \( \rho v \) by a generic flux \( f(u) \), (12.30) generalises to

\[
\frac{d}{dt} \int_{x_1}^{x_2} u(x, t) \, dx = f(u(x_1, t)) - f(u(x_2, t)). \tag{12.31}
\]
By integrating this equation over an arbitrary time interval \([t_1, t_2]\), we find
\[
\int_{t_1}^{t_2} (u(x, t_2) - u(x, t_1)) \, dx = \int_{t_1}^{t_2} \left( f(u(x_1, t)) - f(u(x_2, t)) \right) \, dt.
\] (12.32)

If \(u\) and \(f(u)\) are continuously differentiable, this equation is equivalent to
\[
\int_{t_1}^{t_2} \int_{x_1}^{x_2} \left( \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} \right) \, dx \, dt = 0.
\] (12.33)

Since this equation should hold for arbitrary \(x_1, x_2, t_1,\) and \(t_2\), the integrand has to be zero
necessarily; i.e.,
\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0.
\] (12.34)

A function \(u\) is called a weak solution of (12.34) if it satisfies (12.32) for arbitrary \(x_1, x_2, t_1,\) and \(t_2\). Note that a solution of (12.34) is always a solution of (12.32); the converse need not
be true.

Since the verification of (12.32) for arbitrary \(x_1, x_2, t_1,\) and \(t_2\) is rather cumbersome,
we prefer another definition of weak solution, which is based on distribution theory; see
Chapter 4. Here we define the space of test functions \(\mathcal{D}\) as follows:
\[
\mathcal{D} = C^1_0(\mathbb{R} \times [0, \infty)) := \{ \varphi \in C^1(\mathbb{R} \times [0, \infty)) \mid \varphi \text{ has compact support for any } t \}. \tag{12.35}
\]
The basic idea is then to multiply (12.34) by such a test function \(\varphi(x, t)\), integrate over
\(\mathbb{R} \times [0, \infty)\), and subsequently apply integration by parts. Using the fact that \(\varphi(x, t)\) vanishes
for \(|x| + t \to \infty\), we obtain
\[
\int_0^\infty \int_{-\infty}^\infty \left( u \frac{\partial \varphi}{\partial t} + f(u) \frac{\partial \varphi}{\partial x} \right) \, dx \, dt = - \int_{-\infty}^\infty u(x, 0) \varphi(x, 0) \, dx. \tag{12.36}
\]
This relation then gives rise to the following definition.

**Definition 12.8.** A function \(u(x, t)\) is called a weak solution of conservation law (12.34) if
(12.36) holds for all test functions \(\varphi \in C^1_0(\mathbb{R} \times [0, \infty))\).

Obviously, when \(u(x, t)\) satisfies (12.34), it is a weak solution. The converse is only
true when \(u(x, t)\) is continuously differentiable. In the following, when we speak of a
solution of (12.34), we mean a weak solution in the sense of this definition.

Note that relation (12.36) allows for discontinuous solutions. However, not every
discontinuous function can be a solution of (12.34), as is shown in the following theorem.

**Theorem 12.9.** Let \(u\) be a piecewise smooth solution of (12.34) that has a discontinuity
across a curve \(E : x = x(t)\). Then \(u\) satisfies the condition
\[
[f(u)]^+ = s[u]^+,
\] (12.37)
where \([v]^+ := v(x(t)^+, t) - v(x(t)^-, t) (v = u, f(u))\) is the jump of \(v\) across \(E\) and \(s\) is the speed of \(E\).
12.2. Weak Formulation of First Order Scalar Equations

**Proof.** Assume that $E$ separates a domain $\Omega \supset \text{supp}(\phi)$ into a left part $\Omega_1$ and a right part $\Omega_1^r$; see Figure 12.3. The solution is smooth both in $\Omega_1$ and in $\Omega_1^r$. Since (12.34) holds in $\Omega_1$, we have

$$\int \int_{\Omega_1} \left( \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} \right) \phi \, dx \, dt = 0$$

for every test function $\phi \in \mathcal{D}$. We can rewrite this equation as

$$\int \int_{\Omega_1} \left( \frac{\partial (u \phi)}{\partial t} + \frac{\partial (f(u) \phi)}{\partial x} \right) \, dx \, dt = \int \int_{\Omega_1} \left( u \frac{\partial \phi}{\partial t} + f(u) \frac{\partial \phi}{\partial x} \right) \, dx \, dt.$$

If we apply the two-dimensional Gauss’s theorem to the integral on the left-hand side, we find

$$- \oint_{\partial \Omega_1} \phi(u \, dx - f(u) \, dr) = \int \int_{\Omega_1} \left( u \frac{\partial \phi}{\partial t} + f(u) \frac{\partial \phi}{\partial x} \right) \, dx \, dr,$$

with $\partial \Omega_1$ the boundary of $\Omega_1$. Next, using that $\varphi(x, t) = 0$ for $(x, t) \in \partial \Omega \cap \{ t > 0 \}$, with $\partial \Omega$ the boundary of $\Omega$, we obtain

$$- \int_a^b \varphi(x, 0) \, dx + \int_E \varphi(u_r \, dx - f(u_r) \, dr) = \int \int_{\Omega_1^r} \left( u \frac{\partial \phi}{\partial t} + f(u) \frac{\partial \phi}{\partial x} \right) \, dx \, dr,$$

where $x_s$ is the intersection of $E$ with the $x$ axis and $u_r := u(x(t) + t)$ is the limit value of $u$ just left of the discontinuity. Carrying out a similar procedure for $\Omega_1^r$, we find

$$- \int_a^b \varphi(x, 0) \, dx + \int_E \varphi(u_r \, dx - f(u_r) \, dr) = \int \int_{\Omega_1^r} \left( u \frac{\partial \phi}{\partial t} + f(u) \frac{\partial \phi}{\partial x} \right) \, dx \, dr,$$

with $u_r := u(x(t) - t)$ the limit value of $u$ from the right of $E$. The integral over $E$ in $(**)\text{ is evaluated in the same direction as in (**) (see Figure 12.3) and therefore has a + sign in

![Figure 12.3. Weak solution, discontinuous across $E$.](image-url)
front. Adding \((\ast)\) and \((\ast\ast)\), we obtain
\[
- \int_a^b u(x, 0) \varphi(x, 0) \, dx + \int_E \varphi([u]^- \, dx - [f(u)]^+ \, dt) = \int\int_{\Omega} \left( \frac{\partial \varphi}{\partial t} + f(u) \frac{\partial \varphi}{\partial x} \right) \, dx \, dt.
\]
Combining this relation with (12.36) and taking into account that \(\varphi(x, t)\) vanishes outside \(\Omega\), we find
\[
\int_E \varphi([u]^- \, dx - [f(u)]^+ \, dt = 0.
\]
This relation holds for arbitrary test functions \(\varphi \in D\), so that \([u]^- \, dx - [f(u)]^+ \, dt = 0\). Finally, since \(s = \frac{dx}{dt}\), this implies relation (12.37).

The propagation speed \(s\) of a discontinuity is thus given by
\[
f(u_r) - f(u_\ell) = s(u_r - u_\ell), \tag{12.38}
\]
with \(u_r := u(x(t)+, t)\) and \(u_\ell := u(x(t)-, t)\) the limit values of \(u(x, t)\) just right and left of the discontinuity, respectively. Relation (12.38) is called the Rankine–Hugoniot jump condition. Inserting (12.10) into (12.38), we find the following alternative expression for \(s\):
\[
s = \frac{1}{u_r - u_\ell} \int_{u_\ell}^{u_r} b(v) \, dv; \tag{12.39}
\]
i.e. \(s\) is the average advection velocity \(b(v)\) over the interval \(\text{int} (u_\ell, u_r)\).

**Example 12.10** Consider the Burgers’ equation, written in conservation form,
\[
\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left( \frac{1}{2} u^2 \right) = 0,
\]
subject to the piecewise constant initial condition
\[
u(x, 0) = \begin{cases} 
\alpha & \text{if } x < 0, \\
\beta & \text{if } x > 0.
\end{cases}
\]
Let \(\alpha > \beta\), so that we have a discontinuity. We apparently have, for the speed \(s\) of the discontinuity,
\[
s = \frac{dx}{dt} = \frac{\frac{1}{2} (\beta^2 - \alpha^2)}{\beta - \alpha} = \frac{1}{2} (\beta + \alpha).
\]
The discontinuity is thus a straight line with a directional coefficient that is the average of those of the characteristics to the left and the right, respectively.

It is important to note that the weak solution depends on the formulation of the conservation equation, as the next example clearly shows.

**Example 12.11** Consider again the Burgers’ equation from the previous example, subject to the same initial condition. If we multiply this equation by \(u\), we can easily derive the following conservation equation for \(w := u^2\):
\[
\frac{\partial w}{\partial t} + \frac{\partial}{\partial x} \left( \frac{2}{3} w^{\gamma/2} \right) = 0. \tag{\ast}
\]
If we apply the Rankine–Hugoniot jump condition (12.38) to (∗), we find, for the propagation speed \( s \) of a discontinuity,

\[
s = \frac{2}{3} \frac{w_3^2 - w_1^2}{w_1 - w_3} = \frac{2}{3} \frac{\beta^2 - \alpha^2}{\beta^2 - \alpha^2} = \frac{2}{3} \frac{\beta^2 + \beta \alpha + \alpha^2}{\beta + \alpha}.
\]

This is generally not equal to the propagation speed \( \frac{1}{2} (\beta + \alpha) \) found in Example 12.10. So, although the Burgers’ equation and (∗) are equivalent for smooth solutions, they have different weak solutions.

What we learn from this example is that we cannot manipulate the strong formulation of a problem when dealing with discontinuous solutions. In fact, the Rankine–Hugoniot jump condition is the reformulation of the correct physical conservation law across a discontinuity. Stated otherwise, the Rankine–Hugoniot jump condition is an extra condition that should be imposed for discontinuous solutions along with the corresponding PDE.

It is instructive to look back now to the problem depicted in Figure 12.2. The breaking at \( t = t^* \) starts a shock, which then propagates at speed \( s \) given by (12.38). From the shock speed we can determine the location of the shock. Now consider the initial profile in Figure 12.2, which is zero outside \((0, 1)\). Then if \((a, b)\) is a sufficiently large interval, we have from the integral form of the conservation equation (12.34) that

\[
\frac{d}{dt} \int_a^b u(x, t) \, dx = f(u(a, t)) - f(u(b, t)) = 0. \tag{12.40}
\]

Hence \( \int_a^b u(x, t) \, dx \) is constant. At a time point \( t > t^* \) a classical solution would have a positive and a negative integral part, as shown in Figure 12.4. Of course, the shape as depicted does not make sense practically as we would have a multivalued solution, yet we may formally do the integration. The total effect would be the same as when we integrated up to the point \( x_s \) (starting from \( a \)); see Figure 12.4. Conservation means that our weak solution should also be conserved, and hence we may identify the point \( x_s \) with the propagated breaking point on the shock line. Actually, one may reverse the argument and determine \( x_s \) graphically: choose the point \( x_s \) such that the shaded regions on either side of the line \( x = x_s \) have equal areas.

![Figure 12.4](image)

**Figure 12.4. Illustration of the equal-area rule: The multivalued “solution” on the left should be replaced by the shock on the right.**
12.2.2 The Riemann Problem

The initial value problem for a conservation equation with piecewise constant initial condition and a single discontinuity is known as the Riemann problem. The canonical form of the Riemann problem for (12.34) reads

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \quad (12.41a)
\]

\[
u(x, 0) = \begin{cases} 
u_\ell & \text{if } x < 0, \\ \nu_r & \text{if } x > 0, \end{cases} \quad (12.41b)
\]

Note that if \( u(x, t) \) is a solution of (12.41), then so is \( u(\alpha x, \alpha t) \) for arbitrary \( \alpha > 0 \), implying that the solution of (12.41) is a similarity solution of the form \( u(x, t) = \hat{u}(x/t) \); cf. Section 2.5. In Section 12.1 we showed that solutions of an initial value problem like (12.41) are propagated along characteristics with velocity \( b(u) = f'(u) \). Based on this observation we can distinguish two cases, \( b(\nu_\ell) > b(\nu_r) \) and \( b(\nu_\ell) < b(\nu_r) \), to be discussed below separately.

**Case 1.** \( b(\nu_\ell) > b(\nu_r) \).

The characteristics emanating from the negative \( x \) axis have a slope smaller than the slope of the characteristics coming from the positive \( x \) axis. As a consequence, characteristics intersect, which would lead to multivalued solutions. Instead, we have a discontinuous solution. We can easily verify by substitution into (12.36) that the solution of the Riemann problem (12.41) is indeed given by

\[
u(x, t) = \begin{cases} \nu_\ell & \text{if } x/t < s, \\ \nu_r & \text{if } x/t > s, \end{cases} \quad (12.42)
\]

where \( s \) is defined in (12.38). The solution in (12.42) represents a discontinuity traveling with speed \( s \) and is called a shock wave; \( s \) is called the shock speed. A typical shock wave and the corresponding characteristics are shown in Figure 12.5. Note that the characteristics move into the shock for increasing \( t \).

![Figure 12.5. Shock wave and corresponding characteristics.](image_url)
12.2. Weak Formulation of First Order Scalar Equations

Case 2. $b(u_\ell) < b(u_r)$.

In this case the characteristics emanating from the negative $x$ axis have a larger slope than those emanating from the positive $x$ axis. So we have separating characteristics and we do not expect a discontinuous solution. The solution of (12.41) is now given by

$$u(x,t) = \begin{cases} 
  u_\ell & \text{if } x/t < b(u_\ell), \\
  w(x/t) & \text{if } b(u_\ell) < x/t < b(u_r), \\
  u_r & \text{if } x/t > b(u_r), 
\end{cases}$$

where $w(\eta)$ is the solution of the relation $b(w(\eta)) = \eta$. This solution is called a rarefaction wave. It is continuous, despite the fact that the initial condition is discontinuous, and consists of the constant states $u(x,t) = u_\ell$ and $u(x,t) = u_r$, connected by the intermediate solution $u(x,t) = w(x/t)$. The latter solution is a similarity solution of (12.34), as we can easily verify by direct substitution. The constant states are an immediate consequence of (12.12).

A typical rarefaction wave and the corresponding characteristics are depicted in Figure 12.6.

**Example 12.12** Consider once more the Burgers' equation; cf. Example 12.10. Since the convection velocity $b(u) = u$, we have for the solution either a shock wave when $u_\ell > u_r$ or a rarefaction wave when $u_\ell < u_r$. The shock wave solution reads

$$u(x,t) = \begin{cases} 
  u_\ell & \text{if } x/t < s, \\
  u_r & \text{if } x/t > s, 
\end{cases}$$

with shock speed $s = \frac{1}{2}(u_\ell + u_r)$. The rarefaction wave is given by

$$u(x,t) = \begin{cases} 
  u_\ell & \text{if } x/t < u_\ell, \\
  x/t & \text{if } u_\ell < x/t < u_r, \\
  u_r & \text{if } x/t > u_r. 
\end{cases}$$

We can show that (12.42) is a solution of the Riemann problem (12.41) by substitution into (12.36), regardless of whether $b(u_\ell) > b(u_r)$ or $b(u_\ell) < b(u_r)$. In the latter case this solution is called an expansion shock. In Figure 12.7 we have sketched the corresponding characteristics. Note that the characteristics move out of the shock for increasing $t$. This
poses a problem, since for \( b(u_\ell) < b(u_r) \) we have at least two solutions of the Riemann problem (12.41). The expansion shock is physically not correct and should be discarded. One of the reasons is that this solution is not stable under small perturbations in the initial data, as is demonstrated in the next example.

**Example 12.13** Consider the Burgers’ equation subject to the piecewise linear initial condition

\[
    u(x, 0) = \begin{cases} 
    0 & \text{if } x < 0, \\
    x/\delta & \text{if } 0 < x < \delta, \\
    1 & \text{if } x > \delta,
    \end{cases}
\]

with \( 0 < \delta \ll 1 \). Note that for \( \delta \to 0 \) we obtain the standard Riemann problem. The characteristics in this initial value problem fan out and we can compute its solution from (12.14). This way, we find

\[
    u(x, t) = \begin{cases} 
    0 & \text{if } x < 0, \\
    x/(t + \delta) & \text{if } 0 < x < t + \delta, \\
    1 & \text{if } x > t + \delta.
    \end{cases}
\]

Clearly, for \( \delta \to 0 \) this solution changes into the rarefaction wave (12.43), which is completely different from an expansion shock.

The next example concerns the modeling of traffic flow; cf. Example 1.2.

**Example 12.14** A simple model for traffic flow on a highway is given by the following conservation equation:

\[
    \frac{\partial n}{\partial t} + \frac{\partial f(n)}{\partial x} = 0, \quad f(n) := u_m n(1 - n/n_m), \tag{*}
\]

where \( x \) is the coordinate along the highway, \( n(x, t) \) is the density of cars, and \( n_m \) is the maximum density and \( u_m \) is the maximum speed of vehicles. Obviously, \( n_m, u_m > 0 \). Consider the corresponding Riemann problem with initial condition

\[
    n(x, 0) = \begin{cases} 
    n_\ell & \text{if } x < 0, \\
    n_r & \text{if } x > 0.
    \end{cases}
\]
where $0 \leq n_\ell, n_r \leq n_m$. We can easily verify that the advection velocity $b(n)$ is given by
\[
b(n) = u_m \left( 1 - \frac{2n}{n_m} \right),
\]
which is a monotonically decreasing function of $n$. This implies that a shock wave occurs when $n_\ell < n_r$. The shock wave solution is given by
\[
n(x, t) = \begin{cases} 
  n_\ell & \text{if } x/t < s, \\
  n_r & \text{if } x/t > s,
\end{cases}
\]
with shock speed $s = u_m \left( 1 - (n_\ell + n_r)/n_m \right)$. Note that the shock speed can be either positive or negative, depending on the values of $n_\ell$ and $n_r$. This corresponds, e.g., to the situation when cars approach a red traffic light. On the other hand, when $n_\ell > n_r$, we have a rarefaction wave given by
\[
n(x, t) = \begin{cases} 
  n_\ell & \text{if } x/t < b(n_\ell), \\
  \frac{1}{2}n_m \left( 1 - \frac{x}{u_m t} \right) & \text{if } b(n_\ell) < x/t < b(n_r), \\
  n_r & \text{if } x/t > b(n_r).
\end{cases}
\]
This solution describes, e.g., the situation when cars speed up after the traffic light has turned green.

We have seen that not every discontinuous solution of (12.41) is physically correct. Therefore we would like to have a simple criterion to determine whether a discontinuous solution is admissible. In fact, the physically relevant solution is the solution of (12.18) for $\varepsilon \to 0$. One can show that this vanishing viscosity solution for the Burgers' equation reduces to a shock wave when $u_\ell > u_r$ and to a rarefaction wave when $u_\ell < u_r$; see, e.g., [70]. However, the computation of the solution of (12.18) is often very tedious and not very practical to work with. A simple criterion is suggested by the requirement that characteristics move into a shock for increasing $t$, as shown in Figure 12.5. This then gives rise to the following definition.

**Definition 12.15 (Lax entropy condition).** A discontinuous solution of (12.34) that has a convex flux function satisfies the entropy condition if
\[
b(u_\ell) > s > b(u_r),
\]
with $s$ the propagation speed of the discontinuity, given by (12.38).

One can show that this entropy solution is the unique physically correct solution; cf. [142]. When this condition is generalized to the Euler equations for compressible gas flow, one can prove that the entropy of the flow increases across the discontinuity, in agreement with the second law of thermodynamics, and therefore (12.44) is referred to as the entropy condition.

**Example 12.16** The flux function for the Burgers’ equation is convex, and therefore the entropy condition (12.44) simply reduces to $u_\ell > u_r$. On the other hand, for the traffic flow problem we have a concave flux function, leading to the criterion $n_\ell < n_r$. 

\[\square\]
Integrating the ODE system (12.12) along the characteristic that goes into the shock from the left gives a relation between \( u_\ell \) and the initial data. Likewise, the characteristic going into the shock from the right provides a relation for \( u_r \). Together with the jump condition (12.38), these relations suffice to compute the three unknowns \( u_\ell, u_r, \) and \( s \).

A more general definition, which is also applicable when \( f(u) \) is neither convex nor concave, is the following; cf. [105].

**Definition 12.17 (Oleinik’s entropy condition).** A weak solution of (12.34) is the entropy solution if all discontinuities, which propagate at speed \( s \) given by (12.38), satisfy

\[
\frac{f(u) - f(u_\ell)}{u - u_\ell} \geq s \geq \frac{f(u) - f(u_r)}{u - u_r}
\]

for all \( u \) between \( u_\ell \) and \( u_r \).

We have seen that for convex or concave flux functions the Riemann problem (12.41) has either a shock or a rarefaction wave as solution. For general flux functions the entropy solution might involve both, as demonstrated by the next example.

**Example 12.18** A model equation for two-phase flow is the *Buckley–Leverett equation*, given by (12.34) with flux function

\[
f(u) := \frac{u^2}{u^2 + a(1-u)^2};
\]

see Figure 12.8. Here \( u \) typically represents the saturation of water, so that \( 0 \leq u \leq 1 \) and \( a \) is a constant. It is applied in oil reservoir simulation; for more details see, e.g., [61, 87]. Consider the Riemann problem for (\( * \)) with initial data

\[
u(x, 0) = \begin{cases} 
1 & \text{if } x < 0, \\
0 & \text{if } x > 0;
\end{cases}
\]

The characteristic velocity \( b(u) \) is given by

\[
b(u) = \frac{2au(1-u)}{(u^2 + a(1-u)^2)^2}.
\]

![Figure 12.8. Flux function and solution of the Buckley–Leverett equation.](image)
12.3. First Order Systems

Note that \( b(1) = b(0) = 0 \), so that all characteristics emanate vertically from the \( x \) axis. For all intermediate values \( u (0 < u < 1) \) we have \( b(u) > 0 \). As a consequence, the two constant states \( u_r = 0 \) and \( u_\ell = 1 \) have to be connected by a shock followed by a rarefaction wave. The solution is thus given by

\[
    u(x, t) = \begin{cases} 
    1 & \text{if } x/t < 0, \\
    w(x/t) & \text{if } 0 < x/t < s, \\
    0 & \text{if } x/t > s,
    \end{cases}
\]

where \( w = w(\eta) \) is a similarity solution satisfying \( b(w(\eta)) = \eta \) and where \( s \) is the shock speed; see Figure 12.8. Suppose the shock connects the right state \( u_r = 0 \) with an intermediate state \( u_s (0 < u_s < 1) \). Then according to (12.38), we have

\[
    s = f(u_s)/u_s.
\]

To determine the value of \( u_s \) we have to invoke the entropy condition (12.45), which in this case reduces to

\[
    \frac{f(u) - f(u_s)}{u - u_s} \geq s \geq \frac{f(u)}{u}. \quad (**) 
\]

It will turn out that \( u_s \) satisfies the relation

\[
    f(u_s)/u_s = f'(u_s),
\]

which means that the straight line through \((0,0)\) and \((u_s, f(u_s))\) is tangent to \( y = f(u) \) in \( u_s \). Moreover, the shock moving with speed \( s = f'(u_s) \) is then parallel to the characteristic just left of it. Suppose the shock were connected to a state \( u^* < u_s \). Then the shock speed \( f(u^*)/u^* \) would be smaller than \( f'(u_s) \), leading to a triple-valued function. On the other hand, if the shock were connected to a state \( u^* > u_s \), then the entropy condition (**) would be violated. \( \square \)

12.3 First Order Systems

In this section we shall deal with first order systems of hyperbolic equations, starting with linear systems in Section 12.3.1, followed by nonlinear ones in Section 12.3.2. Finally, in Section 12.3.3, we briefly describe the method of characteristics for a system of two equations, which gives a nice illustration of the role of characteristics for hyperbolic equations.

12.3.1 Linear Systems

Consider the first order system of equations

\[
    A \frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = e, \quad (12.46)
\]

with \( A \) and \( B \) constant \( m \times m \) matrices and \( e \) an \( m \) vector, possibly depending on \( x, t \), and \( u \). In Section 2.2 we considered the case where \( A \) was nonsingular. We can easily generalise this procedure for singular matrices. For this we have to consider the generalised eigenvalue problem: Find left eigenvectors \( t^T \) and corresponding eigenvalues \( \lambda \) such that

\[
    \lambda t^T A = t^T B. \quad (12.47)
\]
Obviously, (12.47) only holds for nonzero \( t^T \) if
\[
\det(\lambda A - B) = 0. \tag{12.48}
\]
The linear combination
\[
\lambda A - B, \quad \lambda \in \mathbb{C}, \tag{12.49}
\]
is called a matrix pencil \([49]\). This matrix pencil is called singular if \( \det(\lambda A - B) = 0 \) for all \( \lambda \in \mathbb{C} \); otherwise it is called regular. We now assume the matrix pencil (12.49) to be regular. In this case there exist \( m \) eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_m \) such that (12.48) holds. Furthermore, we assume that these eigenvalues are real and that the matrix is not defect; i.e., there exists a complete set of (left) eigenvectors \( t_1^T, t_2^T, \ldots, t_m^T \). We can take together all (generalised) eigenvalue problems, giving the relation
\[
TB = \Lambda TA, \tag{12.50}
\]
where the matrices \( \Lambda \) and \( T \) are defined by
\[
\Lambda := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m), \quad T := \begin{pmatrix}
t_1^T \\
t_2^T \\
\vdots \\
t_m^T
\end{pmatrix}. \tag{12.51}
\]
Note that \( T \) is nonsingular since its rows \( t_k^T \) are linearly independent.

Left multiplication of (12.46) by \( T \) results in
\[
TA \frac{\partial u}{\partial t} + \Lambda TA \frac{\partial u}{\partial x} = Tc. \tag{12.52}
\]
In analogy of what we did in Section 2.2, we define the characteristic variable \( \tilde{u} \) by
\[
\tilde{u} := TA u, \tag{12.53}
\]
and since \( T \) is a constant matrix, we obtain the decoupled system (cf. (2.13a))
\[
\frac{\partial \tilde{u}}{\partial t} + \Lambda \frac{\partial \tilde{u}}{\partial x} = \tilde{c} := Tc. \tag{12.54a}
\]
Written componentwise, this is
\[
\frac{\partial \tilde{u}_k}{\partial t} + \lambda_k \frac{\partial \tilde{u}_k}{\partial x} = \tilde{c}_k \quad (k = 1, 2, \ldots, m). \tag{12.54b}
\]
When this decoupling is possible, we call the system (12.46) hyperbolic. We thus have the following formal definition.

**Definition 12.19.** The system (12.46) is called hyperbolic if there exists a real diagonal matrix \( \Lambda \) and a nonsingular matrix \( T \) such that (12.50) holds.
Example 12.20 Consider tidal waves travelling along a straight canal of uniform depth $h$. Let $x$ denote the coordinate along the canal. For small-amplitude waves the water elevation $\eta(x,t)$ above the still-water level satisfies the standard wave equation [116]

$$\frac{\partial^2 \eta}{\partial t^2} = a^2 \frac{\partial^2 \eta}{\partial x^2},$$

with $a := \sqrt{gh}$ and $g$ the gravitational acceleration. Introducing the auxiliary variables

$$u_1 := \frac{1}{a} \frac{\partial \eta}{\partial t}, \quad u_2 := \frac{\partial \eta}{\partial x}, \quad u := \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

we can, e.g., reformulate the wave equation for $\eta$ as the $2 \times 2$ linear system

$$\frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = 0, \tag{*}$$

with the coefficient matrix $B$ given by

$$B := -\begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}.$$

The eigenvalues $\lambda_1, \lambda_2$ and corresponding eigenvectors $t^*_1, t^*_2$ are given by

$$\lambda_1 = -a < 0, \quad \lambda_2 = a > 0, \quad t^*_1 = (1 \ 1), \quad t^*_2 = (1 \ -1),$$

and, consequently, the linear system (*) is hyperbolic. \qed

Since there is no bias with respect to $A$ or $B$ we may consider pairs $(\lambda, \mu)$ of solutions of

$$\det(\lambda A - \mu B) = 0. \tag{12.55}$$

Note that for each pair $(\lambda, \mu)$ satisfying (12.55) any nonzero multiple of it also satisfies this equation. Analogously to the above, the matrix pencil

$$\lambda A - \mu B, \quad \lambda, \mu \in \mathbb{C}, \tag{12.56}$$

is called singular if (12.55) holds for all $\lambda, \mu \in \mathbb{C}$; otherwise it is called regular. We shall again assume that the matrix pencil is regular. In this case there exist $m$ eigenvalues pairs $(\lambda_1, \mu_1), (\lambda_2, \mu_2), \ldots, (\lambda_m, \mu_m)$, apart from a multiplicative constant. Note that for each nontrivial eigenvalue pair $(\lambda_k, \mu_k)$, at least one eigenvalue is nonzero. The appropriate generalisation of (12.50) then reads as follows: There exist nonsingular matrices $T, S$ and real diagonal matrices $\Lambda_A, \Lambda_B$ such that

$$T A S^{-1} = \Lambda_A, \quad T B S^{-1} = \Lambda_B \tag{12.57}$$

i.e., the matrices $A$ and $B$ can be diagonalised simultaneously. In this case (12.46) is called hyperbolic; note that both $A$ and $B$ may be singular.

Premultiplying (12.46) by $T$ and substituting

$$\tilde{u} := Su \tag{12.58}$$
yields the decoupled system
\[ \Lambda A \frac{\partial \tilde{u}}{\partial t} + \Lambda B \frac{\partial \tilde{u}}{\partial x} = \tilde{c} := Tc \]  
(12.59)

If a diagonal element of \( \Lambda A \), say \( \lambda_{A,k} \), is zero, then for the corresponding characteristic \( C_k \) we have
\[ \frac{dt}{dx} = 0; \]
i.e., \( C_k \) is parallel to the \( x \) axis. This implies that the information is propagating with infinite velocity along this characteristic. We shall explicitly exclude such cases here and in the rest of this book. As a consequence, we may assume \( A \) to be nonsingular.

### 12.3.2 Quasi-linear Systems

We now consider the more general case where the matrices \( A \) and \( B \) may depend on \( x, t, \) and \( u \):
\[ A(x, t, u) \frac{\partial u}{\partial t} + B(x, t, u) \frac{\partial u}{\partial x} = c(x, t, u). \]  
(12.60)

Assuming now that \( A \) is nonsingular (cf. what we said in the previous section), we may as well take it equal to \( I \); i.e., we consider the system
\[ \frac{\partial u}{\partial t} + B(x, t, u) \frac{\partial u}{\partial x} = c(x, t, u). \]  
(12.61)

As a straightforward generalisation of Definition 12.19, we can define hyperbolicity of the system (12.61).

**Definition 12.21.** The system (12.61) is called hyperbolic in \( (x, t, u) \) if there exists a real diagonal matrix \( \Lambda (x, t, u) \) and a nonsingular matrix \( T(x, t, u) \) such that
\[ T(x, t, u)B(x, t, u) = \Lambda(x, t, u)T(x, t, u). \]  
(12.62)

In this definition \( A(x, t, u) = \text{diag}(\lambda_1(x, t, u), \lambda_2(x, t, u), \ldots, \lambda_m(x, t, u)) \) and \( T(x, t, u) \) is the matrix of corresponding left eigenvectors; cf. (12.51). Note that the hyperbolicity of (12.61) depends on \( x, t, \) and \( u \). In the following we will suppress this dependency and simply write \( A \) instead of \( A(x, t, u) \), etc.

Left multiplication of (12.61) by the matrix \( T \) gives
\[ T \frac{\partial u}{\partial t} + \Lambda T \frac{\partial u}{\partial x} = Tc := \tilde{c}. \]  
(12.63)

Next we introduce the characteristic variables \( \tilde{u} \) by the relation
\[ \tilde{d} \tilde{u} := Td u, \]  
(12.64)
where \( d \) is an arbitrary differential. Equation (12.64) is a Pfaffian differential equation and expresses the differentials of \( \tilde{u} \) as a linear combination of the differentials of \( u \). Conditions for the solvability of (12.64) can be found in, e.g., [173]. Note that for constant \( T \) relation...
12.3. First Order Systems 331

(12.64) is equivalent to (12.53) in the case \( A = I \). Furthermore, for many nonlinear systems of conservation equations the eigenvectors \( t_k^T(u) \) can be scaled such that (12.64) is integrable. Assuming (12.64) has a solution, system (12.63) reduces to

\[
\frac{\partial \tilde{u}}{\partial t} + A \frac{\partial \tilde{u}}{\partial x} = \tilde{c}.
\]  (12.65)

Thus, like in the linear case, system (12.61) can be diagonalised. Note, however, that system (12.65) is still coupled through the eigenvalue matrix \( \Lambda_1 \), which in general depends on \( u \).

Alternatively, we can premultiply (12.61) by an arbitrary left eigenvector \( t_k^T \) of \( B \), giving

\[
t_k^T \frac{\partial u}{\partial t} + t_k^T B \frac{\partial u}{\partial x} = t_k^T \left( \frac{\partial u}{\partial t} + \lambda_k \frac{\partial u}{\partial x} \right) = t_k^T \tilde{c} =: \tilde{c}_k.
\]  (12.66)

This is in fact a linear combination of the equations of (12.61). We would like (12.66) to be equivalent to an ODE of the form

\[
\frac{\partial \tilde{u}}{\partial s} = \tilde{c}_k.
\]  (12.67)

which should hold on some curve \( K := \{(x(s), t(s)) \mid s \in I \subset \mathbb{R} \} \). Since we have

\[
\frac{\partial u}{\partial s} = \frac{\partial u}{\partial t} \frac{\partial t}{\partial s} + \frac{\partial u}{\partial x} \frac{\partial x}{\partial s},
\]  (12.68)

we thus find by comparing (12.66) and (12.67) and using relation (12.64) that

\[
\frac{\partial t}{\partial s} = 1, \quad \frac{\partial x}{\partial s} = \lambda_k, \quad \frac{\partial \tilde{c}_k}{\partial t} = \tilde{c}_k.
\]  (12.69)

The curve \( K \) is apparently the characteristic \( C_k \) corresponding to the \( k \)th eigenvalue \( \lambda_k \). If system (12.61) is hyperbolic, there exist \( m \) such characteristics. Equation (12.67) is said to be in normal form or characteristic form. The variables \( \tilde{u} \) are also called Riemann variables. If \( c = 0 \), the variables \( \tilde{a}_k \) are constant along the corresponding characteristic \( C_k \) and are therefore often referred to as Riemann invariants.

In many practical applications the coefficient matrix \( B \) and the right-hand side vector \( c \) depend only on \( u \). In the rest of this book, we will restrict ourselves to this case. Then system (12.61) can be rewritten as

\[
\frac{\partial u}{\partial t} + B(u) \frac{\partial u}{\partial x} = c(u),
\]  (12.70)

where the flux vector \( f(u) \) is related to \( B(u) \) through

\[
B(u) = \frac{\partial f(u)}{\partial u} = \left( \frac{\partial f_i(u)}{\partial u_j} \right).
\]  (12.71)

So \( B(u) \) is the Jacobi matrix of \( f(u) \). The formulation in (12.70) is again called the conservation form. According to Definition 12.21, the system (12.70) is hyperbolic if the Jacobi matrix \( B(u) \) is diagonalisable through its left eigenvectors. Alternatively, \( B(u) \) can be brought into diagonal form by its right eigenvectors. Indeed, introducing the matrix

\[
S = \left( s_1, s_2, \ldots, s_m \right) := T^{-1},
\]  (12.72)
we readily see from (12.62) that

$$BS = SA;$$

(12.73)
i.e., the $k$th column $s_k$ of $S$ is the right eigenvector of $B(u)$ corresponding to the eigenvalue $\lambda_k$. Note that the right eigenvectors are linearly independent since $S$ is nonsingular. Changing to the characteristic variables $\tilde{u}$, which are now defined by $d\tilde{u} = S^{-1}du$, we obtain in a similar way as before the diagonalised system (12.65).

**Example 12.22** Referring to Chapter 6, we note that the Euler equations for isentropic gas flow can be written in the standard form (12.70), with $u$, $f(u)$, and $c(u)$ given by

$$u = \begin{pmatrix} \rho \\ \rho u \end{pmatrix}, \quad f(u) = \begin{pmatrix} \rho u \\ \rho u^2 + p(\rho) \end{pmatrix}, \quad c(u) = 0,$$

where $\rho$, $u$, and $p$ are the density, velocity, and pressure, respectively, of the gas flow. For isentropic flow, the pressure is given by the relation

$$p(\rho) = p_0 \rho^\gamma,$$

with $\gamma = C_p/C_v$ the specific heat ratio and $p_0$ a reference pressure. The Jacobi matrix is given by

$$B(u) = \begin{pmatrix} 0 & 1 \\ -u^2 + p'(\rho) & 2u \end{pmatrix},$$

and its eigenvalues $\lambda_k(u)$ and (left) eigenvectors $t_k^T(u)$ are given by

$$\lambda_1(u) = u - c, \quad \lambda_2(u) = u + c, \quad c := \sqrt{p'(\rho)} = \sqrt{\gamma p/\rho},$$

$$t_1^T(u) = (-u - c, 1)^T, \quad t_2^T(u) = (-u + c, 1).$$

The variable $c$ is the speed of sound. Note that the eigenvectors are determined up to a multiplicative constant. Clearly, the isentropic Euler equations are hyperbolic. To decouple these equations we introduce the characteristic variables $\tilde{u}$ through the relation (12.64). This way, we obtain the system

$$d\tilde{u}_1 = -(u + c)d\rho + d(\rho u) = \rho du - cd\rho,$$

$$d\tilde{u}_2 = -(u + c)d\rho + d(\rho u) = \rho du + cd\rho.$$  

Unfortunately, these equations are not integrable. However, scaling the eigenvectors by a factor $1/\rho$, we find the relations

$$d\tilde{u}_1 = du - \frac{c}{\rho}d\rho, \quad d\tilde{u}_2 = du + \frac{c}{\rho}d\rho.$$

Using relation (*), these equations are easy to integrate, and we find

$$\tilde{u}_1 = u - \frac{2c}{\gamma - 1}, \quad \tilde{u}_2 = u + \frac{2c}{\gamma - 1}.$$

Finally, we obtain the decoupled system

$$\frac{\partial}{\partial t} \left( u - \frac{2c}{\gamma - 1} \right) + (u - c) \frac{\partial}{\partial x} \left( u - \frac{2c}{\gamma - 1} \right) = 0,$$

$$\frac{\partial}{\partial t} \left( u + \frac{2c}{\gamma - 1} \right) + (u + c) \frac{\partial}{\partial x} \left( u + \frac{2c}{\gamma - 1} \right) = 0,$$

implying that the Riemann variables $u - \frac{2c}{\gamma - 1}$ and $u + \frac{2c}{\gamma - 1}$ are constant along the $C_1$ and $C_2$ characteristics, respectively.
12.3. First Order Systems

12.3.3 Method of Characteristics

For systems of two equations the normal form (12.67) together with the equations for the characteristics (12.69) lend themselves to a (theoretically) simple solution method. We shall work this out below for the general linear case (12.46), which can be written as

\[
\begin{align*}
\frac{\partial u_1}{\partial t} + \frac{\partial u_2}{\partial t} + b_{11} \frac{\partial u_1}{\partial x} + b_{12} \frac{\partial u_2}{\partial x} &= c_1, \\
\frac{\partial u_1}{\partial t} + \frac{\partial u_2}{\partial t} + b_{21} \frac{\partial u_1}{\partial x} + b_{22} \frac{\partial u_2}{\partial x} &= c_2.
\end{align*}
\]  

(12.74)

Now let \(v^T := (v_1, v_2)\) be a left eigenvector corresponding to the eigenvalue \(\lambda_v\) of (12.47); i.e.,

\[v^T B = \lambda_v v^T A.\]  

(12.75)

If we write

\[v^T A =: (\tilde{v}_1, \tilde{v}_2), \quad v^T c =: \tilde{c}_1\]

and let \(\xi\) be the independent variable along the characteristic defined by

\[\frac{dt}{d\xi} = 1, \quad \frac{dx}{d\xi} = \lambda_v,\]

then we have

\[\tilde{v}_1 \frac{\partial u_1}{\partial \xi} + \tilde{v}_2 \frac{\partial u_2}{\partial \xi} = \tilde{c}_1.
\]  

(12.78)

There also exists a second left eigenvector \(w\) corresponding to the other eigenvalue \(\lambda_w\), say. Define

\[w^T A =: (\tilde{w}_1, \tilde{w}_2), \quad w^T c =: \tilde{c}_2,\]

and, finally, let \(\eta\) be the independent variable along the corresponding characteristic, with

\[\frac{dt}{d\eta} = 1, \quad \frac{dx}{d\eta} = \lambda_w.\]

(12.80)

Then

\[\tilde{w}_1 \frac{\partial u_1}{\partial \eta} + \tilde{w}_2 \frac{\partial u_2}{\partial \eta} = \tilde{c}_2.
\]  

(12.81)

The equations (12.77), (12.78), (12.80), and (12.81) form a complete system that determines both the characteristics and the solutions along them. This system lends itself to a numerical method in an obvious way. Using, e.g., forward differences leads to the method of Massau. In Figure 12.9 we have sketched the idea. We denote by \(t_P\) the value of the variable \(t\) at the point \(P\), etc.

First we discretise (12.77) and (12.80), relating them to step sizes \(\Delta \xi\) and \(\Delta \eta\), respectively, which can be chosen to be constant during the process, i.e.,

\[
\begin{align*}
t_Q - t_P &\equiv \Delta \xi, \quad x_Q - x_P \equiv \lambda_v \Delta \xi, \\
t_Q - t_R &\equiv \Delta \eta, \quad x_Q - x_R \equiv \lambda_w \Delta \eta.
\end{align*}
\]  

(12.82a)
respectively. Suppose we are dealing with a Cauchy problem, so the data are given at $t = 0$. Then typically (12.82) determines the point $Q$, while from (12.83) we can find $u_{1,Q}$ and $u_{2,Q}$. From the previous approximation method it immediately follows that we will face complications if $m > 2$.

### 12.4 Weak Formulation of First Order Systems

In this section we consider the nonlinear system (12.70). An example of such a system is the shallow-water equations, which we will discuss in detail in the next section. In many applications nonlinear systems have discontinuous solutions, which is why we need to take recourse to weak solutions. In Section 12.4.1 we generalise the concept of weak solutions for hyperbolic systems and in Section 12.4.2 we investigate (weak) solutions of the Riemann problem.

#### 12.4.1 Weak Solutions

In this section we give the mathematical definition of a weak solution, which is a straightforward generalization of Definition 12.8. The discussion is rather concise, since it is very similar to the scalar case.

Let $\varphi \in \mathcal{D}^m$ be the space of test functions, with $\mathcal{D}$ defined in (12.35). If we take the inner product of (12.70) with $\varphi$, integrate over $\mathbb{R} \times [0, \infty)$, and subsequently apply Green’s theorem, we get the relation

$$
\int_{-\infty}^{\infty} \int_{0}^{\infty} \left( u \cdot \frac{\partial \varphi}{\partial t} + f(u) \cdot \frac{\partial \varphi}{\partial x} \right) \, dx \, dt = - \int_{-\infty}^{\infty} u(x, 0) \cdot \varphi(x, 0) \, dx - \int_{0}^{\infty} \int_{-\infty}^{\infty} c \cdot \varphi \, dxdt. 
$$

(12.84)
In the derivation of (12.84) we have used that \( \phi(x, t) \) vanishes for \( |x| + t \to \infty \). We then have the following definition.

**Definition 12.23.** A function \( u(x, t) \) is called a weak solution of system (12.70) if relation (12.84) holds for all test functions \( \phi \in (C_0^1(\mathbb{R} \times [0, \infty)))^m \).

A weak solution that often occurs is a piecewise smooth solution, where the smooth parts are connected by discontinuities. These discontinuities cannot be of arbitrary size, as is apparent from the following theorem.

**Theorem 12.24.** Let \( u \) be a piecewise smooth solution of (12.70) that has a discontinuity across a curve \( E : x = x(t) \). Then \( u \) satisfies the condition

\[
[f(u)]^+ = s[u]^+,
\]

where \( [v]^+ := v(x(t) + t) - v(x(t) - t) \) (\( v = u, f(u) \)) is the jump of \( v \) across \( E \) and \( s \) is the speed of \( E \).

**Proof.** The curve \( E \) separates a domain \( \Omega \supset \supp(\phi) \) into a left part \( \Omega_\ell \) and a right part \( \Omega_r \). The solution \( u \) is smooth in both subdomains. Since (12.70) holds in \( \Omega_\ell \), we have

\[
\int\int_{\Omega_\ell} \left( \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} \right) \cdot \phi \, dx \, dt = \int\int_{\Omega_\ell} c \cdot \phi \, dx \, dt
\]

for every test function \( \phi \in D^m \). Using the product rule of differentiation, we can rewrite this relation as

\[
\int\int_{\Omega_\ell} \left( \frac{\partial}{\partial t} (u \cdot \phi) + \frac{\partial}{\partial x} (f(u) \cdot \phi) \right) \, dx \, dt = \int\int_{\Omega_\ell} \left( u \cdot \frac{\partial \phi}{\partial t} + f(u) \cdot \frac{\partial \phi}{\partial x} \right) \, dx \, dt + \int\int_{\Omega_\ell} c \cdot \phi \, dx \, dt.
\]

Next we apply the two-dimensional Gauss theorem to the integral on the left-hand side to find

\[
\int\int_{\Omega_\ell} \left( \frac{\partial}{\partial t} (u \cdot \phi) + \frac{\partial}{\partial x} (f(u) \cdot \phi) \right) \, dx \, dt = -\int_{\partial \Omega_\ell} \phi \cdot (u \, dx - f(u) \, dt)
\]

\[
= -\int_a^t u(x, 0) \cdot \phi(x, 0) \, dx - \int_E \phi \cdot (u_\ell - f(u_\ell)) \, dt,
\]

with \( \partial \Omega_\ell \) the boundary of \( \Omega_\ell \) and \( u_\ell := u(x(t) - t) \) the limit value of \( u \) just left of \( E \); see Figure 12.3. In the derivation of this relation we have used that \( \phi(x, t) = 0 \) for \( (x, t) \in \partial \Omega \cap \{ t > 0 \} \), with \( \partial \Omega \) the boundary of \( \Omega \). Combining these relations, we obtain

\[
-\int_a^t u(x, 0) \cdot \phi(x, 0) \, dx - \int_E \phi \cdot (u_\ell - f(u_\ell)) \, dt
\]

\[
= \int\int_{\Omega_\ell} \left( u \cdot \frac{\partial \phi}{\partial t} + f(u) \cdot \frac{\partial \phi}{\partial x} \right) \, dx \, dt + \int\int_{\Omega_\ell} c \cdot \phi \, dx \, dt. \quad (*)
\]
In the same fashion we find for the right subdomain \( \Omega_1 \) that
\[
- \int_{a}^{b} u(x,0) \cdot \varphi(x,0) \, dx + \int_{E} \varphi \cdot (u_t - f(u)) \, dx \, dt = \int_{\Omega} \left( u \cdot \frac{\partial \varphi}{\partial t} + f(u) \cdot \frac{\partial \varphi}{\partial x} \right) \, dx \, dt + \int_{\Omega} c \cdot \varphi \, dx \, dt, \tag{**}
\]
with \( u_t := u(x(t) +, t) \) the limit value of \( u \) just right of \( E \). The integral over \( E \) in (**i) is evaluated in the same direction as in (*) (see Figure 12.3) and therefore has a plus sign in front of it. Adding the relations (*) and (**), we find
\[
- \int_{a}^{b} u(x,0) \cdot \varphi(x,0) \, dx + \int_{E} \varphi \cdot ([u]^+_{t} - [f(u)]^+) \, dx \, dt + \int_{\Omega} (u \cdot \frac{\partial \varphi}{\partial t} + f(u) \cdot \frac{\partial \varphi}{\partial x}) \, dx \, dt + \int_{\Omega} c \cdot \varphi \, dx \, dt = 0.
\]
Combining this last relation with (12.84) and taking into account that \( \varphi(x,t) \) vanishes outside \( \Omega \), we have
\[
\int_{E} \varphi \cdot ([u]^+_{t} - [f(u)]^+) \, dx \, dt = 0.
\]
Since this relation holds for arbitrary test functions \( \varphi \in \mathcal{D}^{m} \), we conclude that
\[
[u]^+_{t} \, dx - [f(u)]^+ \, dt = 0,
\]
and since \( s = \frac{dx}{dt} \) it is obvious that (12.85) holds.

Written in full, the jump relation (12.85) reads
\[
f(u_r) - f(u_l) = s(u_r - u_l), \tag{12.86}
\]
with \( u_r := u(x(t) +, t) \) and \( u_l := u(x(t) -, t) \), and is also called the Rankine–Hugoniot jump condition. Relation (12.86) provides \( m \) equations for the \( 2m + 1 \) variables \( u_l, u_r, \) and \( s \). This result is needed to construct the shock wave solutions; see next section.

### 12.4.2 The Riemann Problem

In this section we will derive the elementary solutions of the Riemann problem for (12.70). We assume the system to be homogeneous; i.e., \( c = 0 \). The canonical form of the Riemann problem for (12.70) reads

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{12.87a}
\]

\[
u(x,0) = \begin{cases} u_l & \text{if } x < 0, \\ u_r & \text{if } x > 0. \end{cases} \tag{12.87b}
\]
Like in the scalar case, any solution of (12.87) is a similarity solution of the form \( u(x, t) = \tilde{u}(x/t) \).

We first consider the linear case; i.e., \( f(u) = Bu \) with \( B \) a constant matrix. Since the eigenvectors \( s_1, s_2, \ldots, s_m \) of the coefficient matrix \( B \) are linearly independent, we can decompose the initial state vectors \( u_\ell \) and \( u_r \) as

\[
\begin{align*}
  u_\ell &= \sum_{k=1}^{m} \alpha_k s_k, \\
  u_r &= \sum_{k=1}^{m} \beta_k s_k. 
\end{align*}
\] (12.88)

Alternatively, the initial condition can be written as

\[
\begin{align*}
  u(x, 0) &= S\tilde{u}(x, 0) = \sum_{k=1}^{m} \tilde{u}_k(x, 0)s_k. 
\end{align*}
\] (12.89)

Comparing (12.87b) and (12.89) and using that the eigenvectors \( s_k \) are linearly independent, we see that

\[
\tilde{u}_k(x, 0) = \begin{cases} 
  \alpha_k & \text{if } x < 0, \\
  \beta_k & \text{if } x > 0.
\end{cases}
\] (12.90)

Since the eigenvalues \( \lambda_k \) are constant, the variables \( \tilde{u}_k \) can be readily computed from the initial value problem (12.54b) and (12.90), and we find

\[
\tilde{u}_k(x, t) = \tilde{u}_k(x - \lambda_k t, 0) = \begin{cases} 
  \alpha_k & \text{if } x/t < \lambda_k, \\
  \beta_k & \text{if } x/t > \lambda_k.
\end{cases}
\] (12.91)

Inserting this in the relation \( u = S\tilde{u} \), we obtain the solution

\[
\begin{align*}
  u(x, t) &= \sum_{k=1}^{m} \tilde{u}_k(x, t)s_k = \sum_{x/t < \lambda_k} \alpha_k s_k + \sum_{x/t > \lambda_k} \beta_k s_k. 
\end{align*}
\] (12.92)

The solution \( u \) is thus piecewise constant because the initial discontinuity at \( x = 0 \) propagates along all characteristics. The patches of constant \( u \) in the \((x, t)\) plane are separated by the characteristics. As an example, we show in Figure 12.10 the solution for a 3 \times 3 system with \( \lambda_1 < 0 \) and \( \lambda_2, \lambda_3 > 0 \).

Next we consider the quasi-linear case. Note that since \( B \) depends on \( u \), all eigenvalues and eigenvectors depend on \( u \) as well. The general solution of a Riemann problem is hard to obtain. Instead, we will derive specific elementary wave solutions corresponding to an eigenvalue, i.e., a simple wave, a contact discontinuity, and a shock. In the next section we will solve the Riemann problem for the shallow-water equations in full detail. First, we introduce the following definitions.

**Definition 12.25.** An eigenvector \( s_k \) is called genuinely nonlinear if

\[
\left( \frac{\partial \lambda_k}{\partial u} (u), s_k (u) \right) \neq 0
\] (12.93a)
Chapter 12. Analysis of Hyperbolic Equations

Figure 12.10. Similarity solution of the Riemann problem for a $3 \times 3$ linear system. The triple $(\beta_1, \alpha_2, \alpha_3)$ denotes the solution $u = \beta_1 s_1 + \alpha_2 s_2 + \alpha_3 s_3$, etc.

for all $u$, where $\frac{\partial}{\partial u} := (\partial / \partial u_1, \partial / \partial u_2, \ldots, \partial / \partial u_m)$. Likewise, an eigenvector $s_k$ is called linearly degenerate if

$$\left( \frac{\partial \lambda_k}{\partial u}(u), s_k(u) \right) = 0 \quad (12.93b)$$

for all $u$.

Case 1. Simple wave solution.

Assume that $\lambda_k(u_\ell) < \lambda_k(u_r)$ and that $s_k$ is genuinely nonlinear. In this case we can normalize $s_k$ such that

$$\left( \frac{\partial \lambda_k}{\partial u}(u), s_k(u) \right) = 1 \quad (12.94)$$

for all $u$. For an arbitrary left state vector $u_\ell$ we consider the initial value problem

$$\frac{d}{d\eta} (\hat{u}(\eta)) = s_k(\hat{u}(\eta)), \quad \eta > 0, \quad (12.95a)$$

$$\hat{u}(0) = u_\ell, \quad (12.95b)$$

which defines an integral curve in phase space that is everywhere tangent to $s_k$. Let $u_\ell = \hat{u}(\eta_\ell)$ for some $\eta_\ell > 0$. Since

$$\frac{d}{d\eta} \lambda_k(\hat{u}(\eta)) = \left( \frac{\partial \lambda_k}{\partial u}(\hat{u}(\eta)), s_k(\hat{u}(\eta)) \right) = 1,$$

we have

$$\lambda_k(\hat{u}(\eta)) = \eta + \lambda_k(u_\ell), \quad \lambda_k(u_\ell) = \eta_\ell + \lambda_k(\hat{u}(\eta)) > \lambda_k(u_\ell).$$
Next we will show that the solution of the Riemann problem (12.87) is given by

\[
\begin{cases}
  u_\ell & \text{if } x/t < \lambda_k(u_\ell), \\
  \hat{u}(x/t - \lambda_k(u_\ell)) & \text{if } \lambda_k(u_\ell) < x/t < \lambda_k(u_r), \\
  u_r & \text{if } \lambda_k(u_r) < x/t.
\end{cases}
\]

This solution is called a \textit{k-simple wave} or a \textit{k-rarefaction wave}. We restrict ourselves to the nontrivial case \(\lambda_k(u_\ell) < x/t < \lambda_k(u_r)\). We readily see that

\[
\lambda_k(u(x,t)) = \lambda_k(\hat{u}(x/t - \lambda_k(u_\ell))) = x/t - \lambda_k(\hat{u}(\eta)) + \lambda_k(u_\ell) = x/t.
\]

Therefore, using (12.95a) and the previous equation, we have

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \frac{\partial u}{\partial t} + B(u) \frac{\partial u}{\partial x} = -\frac{x}{t^2} s_k(u_\ell) + \frac{1}{t} B(\hat{u}(\eta)) s_k(\hat{u}(\eta))
\]

\[
= \frac{1}{t} \left( -\frac{x}{t} + \lambda_k(\hat{u}(\eta)) \right) s_k(\hat{u}(\eta)) = 0.
\]

So \(u(x,t)\) defined in (12.96) is indeed the solution of the Riemann problem (12.87). An illustration of this solution is given in Figure 12.11.

For the computation of \(k\)-simple waves from the initial value problem (12.95) the \(k\)-\textit{Riemann invariants} are useful. They are defined as follows.

**Definition 12.26.** \textit{A} \(k\)-\textit{Riemann invariant} of (12.87a) is a continuously differentiable function \(w_k: \mathbb{R}^m \to \mathbb{R}\) such that

\[
\left( \frac{\partial w_k}{\partial u}(u), s_k(u) \right) = 0
\]

for all \(u\).

Relation (12.97) is a first order equation that can usually be integrated exactly. Let \(\hat{u}(\eta)\) be a solution of (12.95). Then we have

\[
\frac{d}{d\eta} w_k(\hat{u}(\eta)) = \left( \frac{\partial w_k}{\partial u}(\hat{u}(\eta)), s_k(\hat{u}(\eta)) \right) = 0;
\]

Figure 12.11. Wave pattern of a \(k\)-simple wave.
i.e., $w_k(\hat{u}(\eta))$ is constant along the integral curve of (12.95). One can prove that there exist $m - 1$ such $k$-Riemann invariants $w_k^{(1)}, w_k^{(2)}, \ldots, w_k^{(m-1)}$ with linearly independent gradients [50]. Then it is clear that the integral curve of (12.95) is part of the curve $\mathcal{K}$ given by

$$\mathcal{K} := \{ u \in \mathbb{R}^m \mid w_k^{(j)}(u) = w_k^{(j)}(u_c), \ j = 1, 2, \ldots, m - 1 \}. \quad (12.99)$$

We will use this result in the next section to compute $k$-simple waves for the shallow-water equations.

**Case 2.** Contact discontinuity.

Assume that $s_k$ is linearly degenerate. Let $\hat{u}(\eta)$ be the solution of (12.95) with $\hat{u}(\eta_r) = u_c$. Then we readily see that

$$\frac{d}{d\eta} (\lambda_k(\hat{u}(\eta))) = \left( \frac{\partial \lambda_k}{\partial u}(\hat{u}(\eta)), s_k(\hat{u}(\eta)) \right) = 0,$$

implying that $\lambda_k(\hat{u}(\eta)) = \lambda_k(u_c) = \lambda_k(u_r)$ for all $\eta \in [0, \eta_r]$. Thus $\lambda_k(\hat{u}(\eta))$ is constant along integral curves in the phase space that are tangent to $s_k$. We will show that the solution of the Riemann problem is given by

$$u(x, t) = \begin{cases} u_l & \text{if } x/t < \lambda_k(u_l), \\ u_r & \text{if } \lambda_k(u_l) < x/t. \end{cases} \quad (12.100)$$

This solution is called a contact discontinuity. To this purpose, we have to show that (12.100) satisfies the Rankine–Hugoniot condition (12.86). Since $\lambda_k(\hat{u}(\eta))$ is constant, we have

$$\frac{d}{d\eta} (f(\hat{u}(\eta)) - \lambda_k(\hat{u}(\eta))\hat{u}(\eta)) = B(\hat{u}(\eta)) \frac{d\hat{u}}{d\eta} - \lambda_k(\hat{u}(\eta)) \frac{d\hat{u}}{d\eta}$$

$$= (B(\hat{u}(\eta)) - \lambda_k(\hat{u}(\eta)) I) s_k(\hat{u}(\eta)) = 0,$$

and, consequently, $f(\hat{u}(\eta)) - \lambda_k(\hat{u}(\eta))\hat{u}(\eta) = C (C \in \mathbb{R}^m)$. In particular, we have $f(u_r) - \lambda_k(u_r)u_r = f(u_l) - \lambda_k(u_l)u_l$, so that the Rankine–Hugoniot condition is satisfied with $s = \lambda_k(u_r) = \lambda_k(u_l)$. An illustration is given in Figure 12.12.

![Figure 12.12. Wave pattern of a contact discontinuity.](image)
12.4. Weak Formulation of First Order Systems

Figure 12.13. Wave pattern of a $k$-shock.

**Case 3.** Shock wave.

Assume that $\lambda_k(u_\ell) > \lambda_k(u_r)$ and that $s_k$ is genuinely nonlinear. In this case the solution of the Riemann problem (12.87) is given by the $k$-shock wave

$$u(x, t) = \begin{cases} u_\ell & \text{if } x/t < s, \\ u_r & \text{if } s < x/t, \end{cases} \quad (12.101)$$

where the shock speed $s$ has to satisfy the Rankine–Hugoniot jump condition (12.86). An illustration of a $k$-shock wave is given in Figure 12.13.

Like in the scalar case, we need a simple criterion to determine whether a shock wave is physically correct. This is given by the following definition.

**Definition 12.27 (Lax entropy condition).** The $k$-shock wave (12.101) satisfies the entropy condition if the inequalities

$$\lambda_{k-1}(u_\ell) < s < \lambda_k(u_\ell), \quad (12.102a)$$

$$\lambda_k(u_r) < s < \lambda_{k+1}(u_r) \quad (12.102b)$$

hold with $s$ the shock speed given by (12.86).

These inequalities imply that $m - k + 1$ characteristics move into the shock from the left and $k$ characteristics move into the shock from the right. Integrating the ODE system (12.69) along these characteristics, we find $m - k + 1$ relations between $u_\ell$ and the initial condition left of the shock and $k$ relations between $u_r$ and the initial condition right of the shock. Together with the Rankine–Hugoniot jump condition these constitute $m - k + 1 + k + 1 = 2m + 1$ equations for the same number of unknowns, i.e., $u_\ell$, $u_r$, and $s$.

The general solution of a Riemann problem involves the elementary solutions introduced above, as described in the following theorem; for a proof see, e.g., [142].

**Theorem 12.28.** Suppose that the system (12.87a) is hyperbolic and that each eigenvector of the Jacobi matrix of $f(u)$ is either genuinely nonlinear or linearly degenerate. Then for any $u_\ell \in \mathbb{R}^m$ there exists a neighbourhood $N$ of $u_\ell$ such that the Riemann problem (12.87) has a unique solution if $u_r \in N$. This solution consists of at most $m + 1$ constant states separated by shocks, simple waves, or contact discontinuities.
12.5 The Shallow-Water Equations

In this section we apply the theory of the previous section to the shallow-water equations. The one-dimensional shallow-water equations describe flow in a straight canal and read [159]

\[
\frac{\partial \varphi}{\partial t} + \frac{\partial}{\partial x}(\varphi u) = 0, \tag{12.103a}
\]

\[
\frac{\partial}{\partial t}(\varphi u) + \frac{\partial}{\partial x}\left(\varphi u^2 + \frac{1}{2} \varphi^2\right) = 0, \tag{12.103b}
\]

where \(x\) is the coordinate along the canal, \(u\) is the flow velocity, and \(\varphi := gh\) is the geopotential, with \(h > 0\) the depth of the canal and \(g\) the gravitational acceleration. The first equation in (12.103) describes conservation of mass and the second describes conservation of momentum. An alternative formulation is presented in Chapter 15. These equations can be written in the standard form (12.70) with \(c(u) = 0\) and \(u\) and \(f(u)\) defined by

\[
u := \begin{pmatrix} \varphi \\ \varphi u \end{pmatrix}, \quad f(u) := \begin{pmatrix} \varphi u \\ \varphi u^2 + \frac{1}{2} \varphi^2 \end{pmatrix}. \tag{12.104}
\]

The Jacobi matrix \(B(u)\) of the flux is

\[
B(u) = \begin{pmatrix} 0 & 1 \\ \varphi - u^2 & 2u \end{pmatrix}, \tag{12.105}
\]

and its eigenvalues \(\lambda_k(u)\) and eigenvectors \(s_k(u)\) \((k = 1, 2)\) are given by

\[
\lambda_1(u) = u - c, \quad \lambda_2(u) = u + c, \quad c := \sqrt{\varphi}, \tag{12.106a}
\]

\[
s_1(u) = \begin{pmatrix} 1 \\ u - c \end{pmatrix}, \quad s_2(u) = \begin{pmatrix} 1 \\ u + c \end{pmatrix}. \tag{12.106b}
\]

The shallow-water equations are thus a \(2 \times 2\) hyperbolic system of equations. Moreover, by direct substitution in (12.97), we see that \(w_1(u) = u + 2c\) and \(w_2(u) = u - 2c\) are the 1- and 2-Riemann invariants of the shallow-water equations, respectively.

Consider the Riemann problem for these equations, i.e.,

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{12.107a}
\]

\[
u(x, 0) = \begin{cases} u_\ell & \text{if } x < 0, \\ u_r & \text{if } x > 0, \end{cases} \tag{12.107b}
\]

with \(u\) and \(f(u)\) as defined in (12.104). Since the eigenvectors \(s_k(u)\) are genuinely nonlinear, the solution of (12.107) consists of (at most) three constant states, separated by shocks and/or rarefaction waves. The possible wave patterns for the Riemann problem (12.107) are shown in Figure 12.14.
Figure 12.14. Possible wave patterns of the Riemann problem for the shallow-water equations.

Suppose that the constant states \( u_\ell \) and \( u_* \) are separated by a shock, referred to as the 1-shock. We want to establish a relation between \( u_\ell \) and \( u_* \) using the entropy condition and the Rankine–Hugoniot jump condition. Let \( s_1 \) denote the speed of the shock. For ease of presentation we introduce the variables

\[
v_1 := u - s_1, \quad m_1 := \varphi v_1;
\]

i.e., \( v_1 \) is the flow velocity relative to the 1-shock. The entropy condition for the 1-shock reads

\[
s_1 < \lambda_1(u_\ell),
\]

\[
\lambda_1(u_\ell) < s_1 < \lambda_2(u_*),
\]

which means that three characteristics go into the shock, as shown in Figure 12.15. From (12.108) and the entropy condition (12.109) we derive the inequalities

\[
v_{1,\ell} > c_\ell, \quad |v_{1,*}| < c_\ast.
\]

The 1-shock also satisfies the jump conditions

\[
s_1[\varphi]^- = [\varphi u]^+, \quad s_1[\varphi u]^+ = \left[ \varphi u^2 + \frac{1}{2} \varphi^2 \right]^+.
\]
where \([a]_+^\ell := a_\ell - a_\ell^*\) for a generic variable \(a\). If we introduce the variables \(v_1\) and \(m_1\) into (12.111), these relations simplify to

\[
\begin{align*}
[m_1]_+^- &= 0, \quad (12.112a) \\
\left[ m_1 v_1 + \frac{1}{2} \psi^2 \right]^-_+ &= 0. \quad (12.112b)
\end{align*}
\]

Combining both jump conditions in (12.112) and using that \(m_1 = \varphi_\ell v_1, \varphi_\ell > 0\), we find, for \(m_1\),

\[
m_1 = \sqrt{\frac{1}{2}} (\varphi_\ell + \varphi_\ell^*). \quad (12.113)
\]

Furthermore, from the second jump condition (12.112b) we obtain the relation

\[
u^* - u_\ell = \frac{1}{2} \frac{\psi_\ell^2 - \psi_\ell^*}{m_1}. \quad (12.114)
\]

Let the variable \(z_1\) be defined by

\[
z_1 := \frac{\varphi_\ell^*}{\varphi_\ell}. \quad (12.115)
\]

From the first jump condition (12.112a) and the inequalities in (12.110) we can conclude that \(z_1 > 1\); i.e., the geopotential increases when crossing the 1-shock from left to right. Substituting (12.113) into (12.114), we get the relation

\[
\frac{u^* - u_\ell}{c_\ell} = \frac{1}{2} \sqrt{2(1 - z_1)} \sqrt{1 + \frac{1}{z_1}}, \quad (12.116)
\]

from which we conclude that the flow velocity decreases when passing the 1-shock from left to right.

Figure 12.15. The 1-wave is either a shock (left) or a rarefaction wave (right).
12.5. The Shallow-Water Equations

Alternatively, let the constant states \( u_\ell \) and \( u_* \) be separated by a rarefaction wave, which we will call the 1-rarefaction wave. In this case we have

\[
\lambda_1(u_*) > \lambda_1(u_\ell),
\]

and the rarefaction wave consists of rays \( x/t = \lambda_1(u) \) and is bounded on both sides by the 1-characteristics, as shown in Figure 12.15. As shown in Section 12.4.2, the Riemann invariant \( w_1 \) is constant across the 1-rarefaction wave, implying that

\[
u + 2c = u_\ell + 2c_\ell = u_* + 2c_*.
\]

From (12.118) we easily find

\[
\frac{u_* - u_\ell}{c_\ell} = 2(1 - \sqrt{z_1}),
\]

with \( z_1 \) defined in (12.115). From (12.117) and (12.118) we see that \( 0 < z_1 < 1 \); i.e., the geopotential decreases and the flow velocity increases when passing the 1-rarefaction wave from left to right.

Summarizing, we have the following relation for the 1-wave connecting the states \( u_\ell \) and \( u_* \):

\[
\frac{u_* - u_\ell}{c_\ell} = g(z_1),
\]

with \( z_1 \) defined in (12.115) and the function \( g(z) \) defined by

\[
g(z) := \begin{cases} 
2(1 - \sqrt{z}) & \text{if } 0 < z \leq 1, \\
\frac{1}{2} \sqrt{2(1 - z)} \sqrt{1 + \frac{1}{z}} & \text{if } z > 1;
\end{cases}
\]

see Figure 12.16. The case \( 0 < z_1 \leq 1 \) corresponds to a 1-rarefaction wave, and for \( z_1 > 1 \) we have a 1-shock.

Now we consider the 2-wave connecting the constant states \( u_* \) and \( u_r \). First, consider the case of a 2-shock. Let \( s_2 \) denote the speed of the shock. Analogously to (12.108), we introduce the variables

\[
v_2 := u - s_2, \quad m_2 := \varphi v_2.
\]
The entropy condition for the 2-shock reads
\[
\lambda_2(u_r) < s_2, \tag{12.123a}
\]
\[
\lambda_1(u_s) < s_2 < \lambda_2(u_s), \tag{12.123b}
\]
and in this case three characteristics go into the shock, as shown in Figure 12.17. From (12.122) and the entropy condition (12.123) we find the inequalities
\[
v_{2,r} < -c_r, \quad |v_{2,s}| < c_s. \tag{12.124}
\]
The jump conditions for the 2-shock are identical to (12.111), with \(s_1\) replaced by \(s_2\) and where now \([a]^+_- := a_r - a_s\) for a generic variable \(a\). In a similar way as for the 1-shock we then obtain the jump conditions
\[
[m_2]^+_- = 0, \tag{12.125a}
\]
\[
\left[ m_2 v_2 + \frac{1}{2} \varphi^2 \right]^+_- = 0. \tag{12.125b}
\]
Combining both jump conditions in (12.125) and using that \(m_2 = \varphi_r v_{2,r} < 0\), we find
\[
m_2 = -\frac{1}{2} \frac{\varphi_r + \varphi_s}{\varphi_r \varphi_s}. \tag{12.126}
\]
Also, from the second jump condition (12.125b) we can derive the relation
\[
u_r - u_s = -\frac{1}{2} \frac{\varphi_r^2 - \varphi_s^2}{m_2}. \tag{12.127}
\]
Substituting (12.126) in (12.127), we get
\[
\frac{u_s - u_r}{c_r} = \frac{1}{2} \sqrt{2(\varphi_2 - 1)} \sqrt{1 + \frac{1}{\varphi_2}}, \tag{12.128}
\]
with the variable $z_2$ defined by
\[ z_2 := \frac{\varphi_r}{\varphi_l}. \] (12.129)
From the first jump condition in (12.125) and the inequalities in (12.124) we see that $z_2 > 1$, and consequently both the geopotential and the flow velocity decrease when passing the 2-shock from left to right.

Alternatively, when the constant states $u_*$ and $u_r$ are connected by a 2-rarefaction wave, we have
\[ \lambda_2(u_*) > \lambda_2(u_r), \] (12.130)
and rays $x/t = \lambda_2(u)$ are bounded by the 2-characteristics, as shown in Figure 12.17. In this case the Riemann invariant $w_2$ is constant across the rarefaction wave, leading to
\[ u - 2c = u_* - 2c_* = u_r - 2c_r. \] (12.131)
Rearranging terms, we obtain
\[ \frac{u_* - u_r}{c_r} = 2(\sqrt{z_2} - 1). \] (12.132)
We conclude from (12.130) and (12.131) that $0 < z_2 < 1$ and consequently both the geopotential and the flow velocity increase when passing the 2-rarefaction wave from left to right.

Summarizing, for the 2-wave connecting the states $u_*$ and $u_r$, the following relation holds:
\[ \frac{u_* - u_r}{c_r} = -g(z_2), \] (12.133)
with the variable $z_2$ and the function $g(z)$ defined in (12.129) and (12.121), respectively. For $0 < z_2 < 1$ the 2-wave is a rarefaction wave and for $z_2 > 1$ it is a shock.

By eliminating of the intermediate values $\varphi_*$ and $u_*$ from the definitions in (12.115) and (12.129) and the equations (12.120) and (12.133), we obtain the equations
\[ z_1 \varphi_\ell = z_2 \varphi_r, \] (12.134a)
\[ u_\ell + c_\ell g(z_1) = u_r - c_r g(z_2). \] (12.134b)
Introducing the auxiliary variables
\[ A := \frac{\varphi_r}{\varphi_\ell}, \quad B := \frac{u_r - u_\ell}{c_\ell} \] (12.135)
and eliminating $z_2 = z_1/A$, we find the following nonlinear equation for $z_1$:
\[ G(z_1) := g(z_1) + \sqrt{A} g(z_1/A) - B = 0. \] (12.136)
We can verify that the function $G(z_1)$, defined in (12.136), has the properties $G(0) = 2(1 + \sqrt{A}) - B$, $G'(z_1) < 0$, and $G''(z_1) > 0$. These conditions imply that the nonlinear equation (12.136) has a unique solution provided that $G(0) > 0$. In terms of the variables $u_\ell$ and $u_r$, this latter inequality boils down to
\[ u_r - u_\ell < 2(c_\ell + c_r). \] (12.137)
The Riemann problem (12.107) has a unique solution if the inequality (12.137) holds.
To summarize, the Riemann problem (12.107) can be solved as follows:

1. Compute $A$ and $B$ from (12.135).
2. Solve (12.136) for $z_1$.
3. Compute $z_2 = z_1/A$.
4. Determine $\psi_*$ and $u_*$ for the intermediate state, from, e.g., (12.115) and (12.120).

In the last step we have computed the intermediate state $u_*$ from the left state $u_\ell$. We could have equally well computed $u_*$ from $u_r$ using (12.129) and (12.133). One should note that step 2 above can conveniently be done numerically. Indeed, since $G(z_1) > 0$ and $G''(z_1) > 0$ on the interval $[0, z_1]$, Newton iteration for the numerical solution of (12.136) is bound to converge for an initial guess in $[0, z_1]$; cf. [149]. We still have to determine which wave pattern from Figure 12.14 is the solution of the Riemann problem. We have seen that the 1-wave is a shock if $z_1 > 1$. This condition is equivalent to $G(1) > 0$, or, stated in terms of the variables $A$ and $B$, $B < \sqrt{Ag(1/A)}$. Otherwise, if $B \geq \sqrt{Ag(1/A)}$, the 1-wave is a rarefaction wave. Likewise, the 2-wave is a shock if $z_2 > 1$, or, equivalently, $G(A) > 0$. This latter inequality gives the condition $B < g(A)$. On the other hand, if $B \geq g(A)$, the 2-wave is a rarefaction wave.

Putting everything together, we have the following similarity solution $u(x,t) = \hat{u}(x/t; u_\ell, u_r)$ of the Riemann problem (12.107):

1-shock if $B < \sqrt{Ag(1/A)}$:

$$u(x,t) = \begin{cases} u_\ell & \text{if } x/t < s_1, \\ u_* & \text{if } x/t > s_1, \end{cases}$$

with shock speed $s_1$ given by

$$s_1 = u_\ell - c_\ell \sqrt{\frac{1}{2}(1 + z_1)z_1}$$

1-rarefaction wave if $B \geq \sqrt{Ag(1/A)}$:

$$u(x,t) = \begin{cases} u_\ell & \text{if } x/t < u_\ell - c_\ell \\ u + 2c = u_\ell + 2c_\ell & \text{if } u_\ell - c_\ell < x/t < u_* - c_* \\ u - c = x/t & \text{if } x/t > u_* - c_* \end{cases}$$

2-shock if $B < g(A)$:

$$u(x,t) = \begin{cases} u_* & \text{if } x/t < s_2, \\ u_r & \text{if } x/t > s_2, \end{cases}$$

with shock speed $s_2$ given by

$$s_2 = u_r + c_r \sqrt{\frac{1}{2}(1 + z_2)z_2}$$

12.138a

12.138b

12.139

12.140a

12.140b
2-rarefaction wave if $B \geq g(A)$:
\[
\begin{align*}
 u(x, t) &= u_* & \text{if } x/t < u_* + c_* , \\
 u - 2c &= u_* - 2c_t & \text{if } u_* + c_* < x/t < u_t + c_t , \\
 u + c &= x/t & \text{if } x/t > u_t + c_t .
\end{align*}
\]
(12.141)

Expression (12.138b) for the shock speed $s_1$ follows readily from (12.108) and (12.113), and, likewise, expression (12.140b) for $s_2$ can be derived from (12.122) and (12.126).

12.6 The Wave Equation

A special kind of hyperbolic equation is given by second order problems, as discussed in Section 2.3. In particular, the wave equation
\[
\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} 
\]
(given here in a one-dimensional medium) occurs in the modeling of many phenomena. We shall first consider solutions of this equation in one space dimension in Section 12.6.1. In Section 12.6.2 we discuss solutions in more space dimensions.

12.6.1 One-Dimensional Problems

We can easily derive a first order system associated with (12.142). Introducing the variables
\[
p := \frac{\partial u}{\partial t}, \quad q := \frac{\partial u}{\partial x},
\]
we obtain
\[
\begin{align*}
\frac{\partial}{\partial t} \begin{pmatrix} p \\ q \end{pmatrix} - \begin{pmatrix} 0 & a^2 \\ 1 & 0 \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} p \\ q \end{pmatrix} &= 0.
\end{align*}
\]
(12.143)

Hence we immediately see that the system matrix has two eigenvalues $-a$ and $a$ corresponding to the two characteristics $C_1$ and $C_2$ given by
\[
C_1 := \{(x, t) \mid x + at = \text{constant}\}, \quad C_2 := \{(x, t) \mid x - at = \text{constant}\}.
\]
(12.144a, 12.144b)

These characteristics imply that we should describe two initial boundary conditions. In the simple case of a Cauchy problem, i.e., data given on the line $t = 0$, we can, e.g., prescribe $u$ and $\frac{\partial u}{\partial t}$. So let
\[
\begin{align*}
u(x, 0) &= v(x), \quad \frac{\partial u}{\partial t}(x, 0) = w(x).
\end{align*}
\]
(12.145a, 12.145b)
for some given functions \(v(x)\) and \(w(x)\). There is a nice way to construct a solution, named after d’Alembert. Since (12.143) has constant coefficients, we first note that we can find the normal form (cf. (12.69))

\[
\frac{\text{d}\tilde{u}_k}{\text{d}s} = 0 \quad \text{on} \quad C_k \quad (k = 1, 2).
\]

(12.146)

Because of linearity we have

\[
u(x, t) = \tilde{u}_1(x + at) + \tilde{u}_2(x - at).
\]

(12.147)

Note that \(\tilde{u}_1\) and \(\tilde{u}_2\) are unique but for a multiplicative constant. From (12.145a) we derive

\[
\tilde{u}_1(x) + \tilde{u}_2(x) = v(x),
\]

(12.148)

and from (12.145b) we derive

\[
a(\tilde{u}_1'(x) - \tilde{u}_2'(x)) = w(x),
\]

where the prime \(\prime\) denotes differentiation with respect to \(x\). Integration of the latter equation results in

\[
\tilde{u}_1(x) - \tilde{u}_2(x) = \frac{1}{a} \int_0^x w(\xi) \, \text{d}\xi + C, \quad C \in \mathbb{R}.
\]

(12.149)

From (12.148) and (12.149) we then deduce

\[
u(x, t) = \frac{1}{2} \left(v(x + at) + v(x - at)\right) + \frac{1}{2a} \int_{x-at}^{x+at} w(\xi) \, \text{d}\xi.
\]

(12.150)

Formula (12.150) is the d’Alembert solution of (12.142) and (12.145) and holds on the whole real line.

Wave phenomena are often defined on a semi-infinite or finite interval. One may, e.g., think of a (simplified) model for a string attached at one end or two ends. A natural way to solve such problems is by employing reflections. Consider first (12.142), subject to the boundary and initial conditions

\[
u(0, t) = 0, \quad t \geq 0
\]

(12.151a)

\[
u(x, 0) = v(x), \quad \frac{\partial u}{\partial t}(x, 0) = w(x), \quad x > 0.
\]

(12.151b)

We then define a problem on \((-\infty, \infty)\) by continuing \(v\) and \(w\) as odd functions for negative values of the argument; i.e., the points in the graph are reflected with respect to the origin. So we have functions \(\tilde{v}\) and \(\tilde{w}\) with

\[
\tilde{v}(x) = -v(-x), \quad \tilde{w}(x) = -w(-x), \quad x < 0.
\]

(12.152)

We shall omit the bar below again and consider the condition (12.151b) as defined on \((-\infty, \infty)\) now. The solution is then formally given by (12.150). At \(x = 0\) we thus find

\[
u(0, t) = \frac{1}{2} \left(v(at) + v(-at)\right) + \frac{1}{2a} \int_{-at}^{at} w(\xi) \, \text{d}\xi = 0,
\]

(12.153)
12.6. The Wave Equation

as is required by (12.151a). The complete solution of (12.142) and (12.151) is given by

\[
\begin{align*}
    u(x, t) &= \begin{cases} 
        \frac{1}{2} \left( v(x + at) + v(x - at) \right) + \frac{1}{2a} \int_{x-at}^{x+at} w(\xi) \, d\xi & \text{if } x > at, \\
        \frac{1}{2} \left( v(at + x) - v(at - x) \right) + \frac{1}{2a} \int_{at-x}^{at+x} w(\xi) \, d\xi & \text{if } x < at.
    \end{cases}
\end{align*}
\]

(12.154)

We assume \( w(x) = 0 \). Then the initial profile \( v(x) \) is split into two parts, one travelling to the right and one to the left. The left-travelling part is reflected and “inverted” at \( x = 0 \). This part of the solution can be interpreted as the inverted profile originating from \(-x\). For a finite interval the procedure is similar, taking a reflection on the right boundary as well.

12.6.2 Solutions in Several Dimensions

We now turn to the wave equation in three space variables. So consider

\[
\frac{\partial^2 u}{\partial t^2} = a^2 \nabla^2 u
\]

subject to the initial conditions

\[
u(x, 0) = v(x), \quad \frac{\partial u}{\partial t}(x, 0) = w(x).
\]

(12.156)

One can find a d’Alembert-type solution to this problem using averaging. Define for \( \xi \in \mathbb{R}^3 \) the average of \( u \) on a sphere \( S(\xi; r) \) with centre \( \xi \) and radius \( r \); i.e.,

\[
\bar{u}(r, t; \xi) := \frac{1}{4\pi r^2} \int_{S(\xi; r)} u(x, t) \, dS, \quad r \neq 0.
\]

(12.157)

Introducing the variable \( \tilde{x} := (x - \xi)/r = n \), with \( n \) the outward unit normal on \( S(\xi; r) \), we may as well take the average over the unit sphere; i.e.,

\[
\bar{u}(r, t; \xi) := \frac{1}{4\pi} \int_{S(0, 1)} u(\xi + rn, t) \, d\tilde{S},
\]

(12.158)

with \( d\tilde{S} = dS/r^2 \). One easily verifies that \( u \) satisfies

\[
u(\xi, t) = \bar{u}(0, t; \xi).
\]

For \( \bar{u} \) we have the following property.

**Property 12.29.** The variable \( \bar{u} \), with \( \bar{u} \) defined in (12.158), satisfies the one-dimensional equation

\[
\frac{\partial^2}{\partial t^2}(\bar{u}) = a^2 \frac{\partial^2}{\partial r^2}(\bar{u}).
\]

(12.159)
**Proof.** From (12.158) we find
\[
\frac{\partial^2 \bar{u}}{\partial t^2} = \frac{1}{4\pi} \int_{S(\xi)} \frac{\partial^2 u}{\partial t^2}(\xi + r n, t) d\tilde{S} = \frac{a^2}{4\pi} \int_{S(\Omega)} \nabla^2 u(\xi + r n, t) d\tilde{S} \quad (\star)
\]
On the other hand, we find from Gauss’s theorem (cf. the appendix, Section J) that
\[
\frac{r^2}{r^2} \frac{\partial^2 \bar{u}}{\partial r^2} = \frac{r^2}{4\pi} \int_{S(\xi)} \nabla u(\xi + r n, t) \cdot n d\tilde{S} = \frac{1}{4\pi} \int_{S(\xi, r)} \nabla^2 u(x, t) \cdot n d\tilde{S} = \frac{1}{4\pi} \int_{B(\xi, r)} \nabla^2 u(x, t) dV.
\]
where $B(\xi; r)$ is the ball of the sphere $S(\xi; r)$. The latter integral can be rewritten as
\[
\frac{1}{4\pi} \int_{S(\xi, r)} \nabla^2 u(x, t) d\tilde{S}.
\]
Hence we have
\[
\frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \bar{u}}{\partial r} \right) = \frac{1}{4\pi r^2} \int_{S(\xi, r)} \nabla^2 u(x, t) dS = \frac{1}{4\pi} \int_{S(\Omega)} \nabla^2 u(\xi + r n, t) d\tilde{S} \quad (**)
\]
From $(\star)$ and $(**)$ we conclude that
\[
\frac{\partial^2 \bar{u}}{\partial t^2} = \frac{a^2}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \bar{u}}{\partial r} \right).
\]
By straightforward manipulation this can then be rewritten as (12.159). □

From (12.159) we can obtain an expression for the solution $r \bar{u}$ as a function of $r$ and $t$; i.e.,
\[
\bar{u}(r, t; \xi) = \bar{v}_1(r + at) + \bar{v}_2(r - at)
\]
for some $\bar{v}_1, \bar{v}_2$. By substituting $r = 0$, we immediately see that $\bar{v}_1(at) = -\bar{v}_2(-at)$, so that (12.160) turns into
\[
\bar{u}(r, t; \xi) = \bar{v}_1(r + at) - \bar{v}_1(at - r).
\]
We can obtain a simpler relation now for $u(\xi, t) = \bar{u}(0, t; \xi)$ by differentiating both sides in (12.161) with respect to $r$ and setting $r = 0$ (the prime $(\prime)$ denotes differentiation with respect to the argument of $\bar{v}_1$); i.e.,
\[
u(\xi, t) = 2\bar{v}_1'(at).
\]
We now determine $\bar{v}_1$. First we differentiate (12.161) with respect to $r$ and set $t = 0$, resulting in
\[
\frac{\partial}{\partial r} (r \bar{u})(r, 0; \xi) = \bar{v}_1'(r) + \bar{v}_1'(-r).
\]
Next we differentiate (12.161) with respect to \( t \) and set \( t = 0 \) to find
\[
\frac{\partial}{\partial t} (\bar{r} u)(r, 0; \xi) = a (\bar{v}_1'(r) - \bar{v}_1'(-r)).
\] (12.164)

In (12.163) and (12.164) we can eliminate \( \bar{v}_1'(r) \) to find an expression for \( 2\bar{v}_1(r) \) and use this to determine \( u(\xi, t) \). We obtain
\[
2\bar{v}_1'(r) = \frac{\partial}{\partial r} (r\bar{u})(r, 0; \xi) + r \frac{\partial}{\partial r} \left( \frac{\partial}{\partial t} (r\bar{u})(r, 0; \xi) \right) = \frac{\partial}{\partial r} \left( \frac{r}{4\pi} \int_{S(0,1)} u(\xi + r\nu; 0) \, d\tilde{S} \right) + \frac{r}{4\pi a} \frac{\partial}{\partial t} \left( \frac{r}{4\pi} \int_{S(0,1)} u(\xi + r\nu; 0) \, d\tilde{S} \right)
\] (12.165)

The anticipated expression for \( u(\xi, t) \) is now obtained from (12.165) by substituting \( r = at \):
\[
u(\xi, t) = \frac{\partial}{\partial t} \left( \frac{1}{4\pi} \int_{S(\xi,at)} v(\xi + at\nu; a(t - \tau)) \, d\tilde{S} \right) + \frac{1}{4\pi a^2} \int_{S(\xi,at)} w(\xi + at\nu; a(t - \tau)) \, d\tilde{S} + \frac{1}{4\pi a^2} \int_{S(\xi,at)} w(x) \, dS
\] (12.166)

where \( \bar{v} \) and \( \bar{w} \) are defined analogously to (12.157). The expression in (12.166) lends itself to obtaining the solution in \( \mathbb{R}^2 \) by employing the fact that \( v \) and \( w \) depend on two variables only. We leave this as an exercise.

We can also find a solution of an inhomogeneous problem employing the Duhamel principle. Consider the problem
\[
\frac{\partial^2 u}{\partial t^2} = a^2 \nabla^2 u + s(x, t),
\] (12.167)
subject to homogeneous initial conditions (12.156), i.e., \( v(x) = w(x) = 0 \). We recall from Section 4.6 that we can solve this initial value problem if we can find a solution \( \bar{u}(x, t; \tau) \) such that
\[
\frac{\partial^2 \bar{u}}{\partial t^2} = a^2 \nabla^2 \bar{u}, \quad x \in \mathbb{R}^3, \quad t > \tau,
\] (12.168a)
\[
\bar{u}(x, \tau; \tau) = 0, \quad x \in \mathbb{R}^3,
\] (12.168b)
\[
\frac{\partial \bar{u}}{\partial t}(x, \tau; \tau) = s(x, \tau), \quad x \in \mathbb{R}^3.
\] (12.168c)

Clearly, \( \bar{u} \) follows from (12.166) and is given by
\[
\bar{u}(\xi, t; \tau) = \frac{1}{4\pi a^2(t - \tau)} \int_{S(\xi,at)} s(x, \tau) \, d\tilde{S}.
\] (12.169)
Hence we find the following representation for the solution of (12.167):

\[ u(\xi, t) = \frac{1}{4\pi a^2} \int_0^t \frac{d\tau}{t - \tau} \int_{S(\xi, \tau)} s(x, \tau) \, dS. \]  

(12.170)

**Example 12.30** Consider the one-dimensional case. We can directly apply Duhamel's principle. We then find the representation

\[ u(\xi, t) = \frac{1}{2a} \int_0^t d\tau \int_{\xi + a(t - \tau)}^{\xi - a(t - \tau)} s(x, \tau) \, dx. \]  

\[ \square \]

**Example 12.31** An important application of solution (12.160) is the following initial value problem in \( \mathbb{R}^3 \) of the field generated by a spherical source at \( r = R \), where \( r = |x| \):

\[ \frac{\partial^2 u}{\partial t^2} + a^2 \nabla^2 u, \quad r > R, \quad t > 0 \]

\[ u = U(t), \quad r = R, \quad t > 0 \]

\[ u \equiv 0, \quad r > R, \quad t < 0. \]

The general solution is thus

\[ u = \frac{1}{r} F\left(t - \frac{r}{a}\right) + \frac{1}{r} G\left(t + \frac{r}{a}\right), \]

where \( F \) corresponds apparently to outgoing waves and \( G \) to incoming waves. From the causality condition \( u \equiv 0 \) for \( t < 0 \) it follows that \( G = 0 \). From the boundary condition we obtain \( U(t) = F\left(t - R/a\right)/R \), and so we have

\[ u = \frac{R}{r} U\left(t - \frac{r - R}{a}\right) \quad \text{for} \quad r \geq R, \quad t > 0. \]

\[ \square \]

### 12.7 Boundary Conditions

Proper initial and boundary conditions are crucial for having unique solutions of PDEs. As we have seen, for hyperbolic equations these conditions are propagated along characteristics. In order to find out whether the problem is well posed, it is therefore important to know where the characteristics emanate. As we shall see, the number of conditions we can impose at a boundary, often referred to as physical boundary conditions, depends on (the sign of) the eigenvalues of the system.

We first discuss the linear case. Consider the following \( m \)-dimensional linear initial boundary value problem for \( u(x, t) \):

\[ \frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = 0, \quad x \in (0, 1), \quad t > 0, \]  

(12.171a)

\[ u(x, 0) = v(x), \quad x \in (0, 1), \]  

(12.171b)

\[ C_l u(0, t) = g_l(t), \quad C_r u(1, t) = g_r(t), \quad t > 0. \]  

(12.171c)

Since the system in (12.171) is hyperbolic, the coefficient matrix \( B \) has \( m \) real eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_m \), of which \( p \), say, are positive and \( m - p \) are negative. Without restriction we may assume that they are ordered as

\[ \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{m-p} < 0 < \lambda_{m-p+1} \leq \cdots \leq \lambda_m. \]  

(12.172)
12.7. Boundary Conditions

Furthermore, let $C_ℓ$ be an $m_ℓ \times m$ matrix and $g_ℓ$ an $m_ℓ$ vector, which means that we impose $m_ℓ$ boundary conditions at the left boundary $x = 0$. Likewise, let $C_r$ be an $m_r \times m$ matrix and $g_r$ be an $m_r$ vector, implying that we have $m_r$ boundary conditions at $x = 1$. Obviously, $0 \leq m_ℓ, m_r \leq m$.

To investigate the boundary conditions in (12.171) we need to decouple the system and write it in terms of the characteristic variables $\tilde{u}$. In Section 12.3.2 we showed that the characteristic variable $\tilde{u}_k$ corresponding to the $k$th eigenvalue and its characteristic satisfy the differential equations

$$\frac{d\tilde{u}_k}{dt} = 0, \quad \frac{dx}{dt} = \lambda_k,$$

(12.173)

implying that $\tilde{u}_k(x, t) = C_k (C_k \in \mathbb{R})$ along the characteristic $C_k : x - \lambda_k t = C (C \in \mathbb{R})$. Consider the case $\lambda_k < 0$ first. From Figure 12.18 it is clear that through every point on the left boundary $x = 0$ a characteristic passes emanating from a point on either the initial line $t = 0$ or the right boundary $x = 1$. This means that we have to prescribe $\tilde{u}_k$ at $t = 0$ and $x = 1$; then $\tilde{u}_k$ is completely determined at the boundary $x = 0$, where the characteristics leave the domain. Consequently, we may not even impose any boundary condition for $\tilde{u}_k$ at $x = 0$. The case $\lambda_k > 0$ is similar. Characteristics $C_k$ enter the domain at $t = 0$ and $x = 0$, and $\tilde{u}_k$ has to be specified there. On the other hand, characteristics $C_k$ leave the domain at $x = 1$, so that no boundary conditions for $\tilde{u}_k$ may be given there. Summarizing, the characteristic variable $\tilde{u}_k$ must be given at the boundary from which characteristics $C_k$ emanate.

We introduce the following partitioning of the characteristic variables $\tilde{u}$:

$$\tilde{u} = \begin{pmatrix} \tilde{u}^- \\ \tilde{u}^+ \end{pmatrix} \quad \text{with} \quad \tilde{u}^- := \begin{pmatrix} \tilde{u}_1 \\ \vdots \\ \tilde{u}_{m-p} \end{pmatrix}, \quad \tilde{u}^+ := \begin{pmatrix} \tilde{u}_{m-p+1} \\ \vdots \\ \tilde{u}_m \end{pmatrix};$$

(12.174)

Figure 12.18. Characteristics of the hyperbolic system (12.171a) corresponding to $\lambda_k < 0$ (left) and $\lambda_k > 0$ (right).
i.e., \( \tilde{u}^- \) and \( \tilde{u}^+ \) contain the characteristic variables corresponding to negative and positive eigenvalues, respectively. Likewise, we split the right eigenvector matrix \( S := T^{-1} \) as follows:

\[
S = (S^- | S^+) \quad \text{with} \quad S^- := (s_1, \ldots, s_{m-p}), \quad S^+ := (s_{m-p+1}, \ldots, s_m). \tag{12.175}
\]

The matrices \( S^- \) and \( S^+ \) contain the eigenvectors corresponding to negative and positive eigenvalues, respectively. Now consider the boundary condition at \( x = 0 \) for the variable \( u \). In terms of the characteristic variable \( \tilde{u} \), it can be written as

\[
C_j S \tilde{u}(0, t) = g_j(t). \tag{12.176}
\]

Substituting the partitionings of (12.174) and (12.175) in (12.176), we find

\[
C_j S^- \tilde{u}^-(0, t) + C_j S^+ \tilde{u}^+(0, t) = g_j(t). \tag{12.177}
\]

From the preceding discussion we conclude that \( \tilde{u}^- (0, t) \) cannot be prescribed. On the other hand, \( \tilde{u}^+ (0, t) \) has to be specified. This means that the \( m_\ell \times p \) matrix \( C_j S^+ \) in (12.177) has to be invertible. Then at least \( m_\ell = p \) should hold; i.e., the number of boundary conditions at \( x = 0 \) is equal to the number of positive eigenvalues. Analogously to (12.177), we obtain, for the boundary condition at \( x = 1 \), that

\[
C_r S^- \tilde{u}^-(1, t) + C_r S^+ \tilde{u}^+(1, t) = g_r(t). \tag{12.178}
\]

In this case \( \tilde{u}^- (1, t) \) has to be given, which means that the \( m_r \times (m - p) \) matrix \( C_r S^- \) in (12.178) has to be invertible. This in turn implies that we must have \( m_r = m - p \), so the number of boundary conditions at \( x = 1 \) is equal to the number of negative eigenvalues. Finally, the boundary conditions in (12.171) have to satisfy the conditions

\[
x = 0: \quad m_\ell = p, \quad C_j S^+ \text{ invertible}, \tag{12.179a}\]

\[
x = 1: \quad m_r = m - p, \quad C_r S^- \text{ invertible}. \tag{12.179b}
\]

Example 12.32 Consider the linear system from Example 12.20 describing tidal waves in a canal. We can easily verify that

\[
S = (s_1, s_2) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad S^- = s_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad S^+ = s_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix},
\]

corresponding to the eigenvalues \( \lambda_1 = -a < 0 \) and \( \lambda_2 = a > 0 \), respectively. We thus have to impose one boundary condition at each boundary \( x = 0 \) and \( x = 1 \). We can write the boundary condition at \( x = 0 \) in the form

\[
\text{or, equivalently,}
\]

\[
C_{\ell,1} \frac{\partial \eta}{\partial t} (0, t) + C_{\ell,2} \frac{\partial \eta}{\partial x} (0, t) = g_\ell(t) \quad \text{with} \quad C_{\ell,1} - C_{\ell,2} \neq 0.
\]

Likewise, we have at \( x = 1 \) the boundary condition

\[
C_{r,1} \frac{\partial \eta}{\partial t} (1, t) + C_{r,2} \frac{\partial \eta}{\partial x} (1, t) = g_r(t) \quad \text{with} \quad C_{r,1} + C_{r,2} \neq 0.
\]
Next consider the following nonlinear initial boundary value problem for \( u(x, t) \):

\[
\begin{align*}
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} &= 0, & x \in (0, 1), & t > 0, \\
 u(x, 0) &= v(x), & x \in (0, 1), \\
 C_t(u)(0, t) &= g_t(t), & \quad t > 0.
\end{align*}
\]

(12.180a) (12.180b) (12.180c)

The nonlinear system (12.180a) is hyperbolic, which means that the Jacobi matrix \( B(u) \) has \( m \) real eigenvalues \( \lambda_k(u) \). The ordering is the same as in (12.172). Since the eigenvalues depend on the solution \( u \), so does the number of positive eigenvalues \( p \). In particular, \( p \) can be different on the boundaries \( x = 0 \) and \( x = 1 \). Therefore, we will use the notation \( p(\xi, t) \) (\( \xi = 0, 1 \)) to denote the number of positive eigenvalues at \( x = \xi \). In (12.180c) \( C_t(u) \) is a vector function mapping the \( m \)-dimensional space onto the \( m \times \) -dimensional space and \( g_t \) is an \( m_t \) vector. We thus impose \( m_t \) boundary conditions at \( x = 0 \). At \( x = 1 \) we have \( m_t \) boundary conditions, where \( C_t(u) \) is a vector function mapping the \( m \)-dimensional space onto the \( m_t \)-dimensional space and \( g_t \) is an \( m_t \) vector.

To investigate the boundary conditions for (12.180) we introduce the characteristic variables \( \tilde{u} \) by the Pfaffian differential equation

\[
T \tilde{u} = d\tilde{u}.
\]

(12.181)

Assuming that (12.181) is integrable, we can determine \( \tilde{u} \) as a function of \( u \). Formally inverting this function, we may write

\[
u = C(\tilde{u}).
\]

(12.182)

In Section 12.3.2, we showed that the characteristic variables \( \tilde{u}_k \) satisfy the differential equations (12.69) and, consequently, \( \tilde{u}_k \) is constant along the characteristic \( C_k \). Repeating the previous discussion on boundary conditions for \( \tilde{u}_k \), we conclude that \( \tilde{u}_k \) must be specified at the boundary at which the corresponding characteristic \( C_k \) enters the domain. This means that \( \tilde{u}_k(0, t) \) and \( \tilde{u}_k(1, t) \) must be specified if \( \lambda_k(0, t) > 0 \) and \( \lambda_k(1, t) < 0 \), respectively. Alternatively, in terms of the partitioning (12.174), this means that \( \tilde{u}^- (0, t) \) and \( \tilde{u}^+ (1, t) \) must be given.

Now consider the boundary conditions for \( u \). Substituting the relation (12.182), we can write the boundary condition at \( x = 0 \) as

\[
(C_t \circ C)(\tilde{u}^-, \tilde{u}^+)(0, t) - g_t(t) = 0,
\]

(12.183)

where \( C_t \circ C \) is the composition of the mappings \( C_t \) and \( C \), i.e., \( (C_t \circ C)(\tilde{u}) := C_t(C(\tilde{u})) \).

In (12.183) we explicitly distinguish the variables \( \tilde{u}^- \) and \( \tilde{u}^+ \). From this relation \( \tilde{u}^+ (0, t) \) has to be determined, and according to the implicit function theorem, this can be done if the \( m_t \times p(0, t) \) matrix \( \frac{\partial}{\partial u}(C_t \circ C)(0, t) \) is invertible. Therefore we must at least have that \( m_t = p(0, t) \); i.e., the number of boundary conditions at \( x = 0 \) equals the number of positive eigenvalues at \( x = 0 \). Analogously to (12.183), the boundary condition at \( x = 1 \) can be written as

\[
(C_t \circ C)(\tilde{u}^-, \tilde{u}^+)(1, t) - g_t(t) = 0.
\]

(12.184)
from which \( \tilde{u}^{-}(1, t) \) has to be determined. This leads to the requirement that the \( m_{\ell} \times (m - p(1, t)) \) matrix \( \frac{\partial}{\partial \tilde{u}}(C_{\ell} \circ C)(1, t) \) be invertible. This in turn implies that at least \( m_{\ell} = m - p(1, t) \); i.e., the number of boundary conditions at \( x = 1 \) is equal to the number of negative eigenvalues at \( x = 1 \). To conclude, we have the following requirements on the boundary conditions in (12.180):

\[
\begin{align*}
  &x = 0: \quad m_{\ell} = p, \quad \frac{\partial}{\partial \tilde{u}}(C_{\ell} \circ C) \text{ invertible,} \quad (12.185a) \\
  &x = 1: \quad m_{\ell} = n - p, \quad \frac{\partial}{\partial \tilde{u}}(C_{\ell} \circ C) \text{ invertible.} \quad (12.185b)
\end{align*}
\]

**Example 12.33** Recall from Section 12.5 that the eigenvalues \( \lambda_{k}(u) \) and eigenvectors \( s_{k}(u) \) of the shallow-water equations are given by

\[
\begin{align*}
  &\lambda_{1}(u) = u - c, \quad \lambda_{2}(u) = u + c, \quad c := \sqrt{\phi}, \\
  &s_{1}(u) = \begin{pmatrix} 1 \\
  u - c \end{pmatrix}, \quad s_{2}(u) = \begin{pmatrix} 1 \\
  u + c \end{pmatrix}.
\end{align*}
\]

Next we have to scale the eigenvectors such that the Pfaffian differential equation \( d\tilde{u} = Tdu \) with \( T = S^{-1} \) is integrable. It turns out that we have to choose

\[
T = \frac{1}{c^{2}} \begin{pmatrix} -(u + c) & 1 \\
-u + c & 1 \end{pmatrix}.
\]

Inserting this matrix into (12.181), we obtain the relations

\[
\begin{align*}
  &d\tilde{u}_{1} = \left( du - \frac{d\phi}{\sqrt{\phi}} \right), \quad d\tilde{u}_{2} = \left( du + \frac{d\phi}{\sqrt{\phi}} \right). \quad (*)
\end{align*}
\]

These equations can be trivially solved, and we find

\[
\begin{align*}
  &\tilde{u}_{1} = u - 2c, \quad \tilde{u}_{2} = u + 2c.
\end{align*}
\]

Subsequently, we can give the explicit inverse

\[
\begin{align*}
  &u = C(\tilde{u}) := \frac{1}{16}(\tilde{u}_{2} - \tilde{u}_{1})^{2} \begin{pmatrix}
  1 \\
  \frac{1}{2}(\tilde{u}_{1} + \tilde{u}_{2})
\end{pmatrix}. \quad (***)
\end{align*}
\]

The number of boundary conditions for this problem is summarized in Table 12.1. We can distinguish three cases at each boundary. We only discuss the boundary conditions at \( x = 0 \); the situation at \( x = 1 \) is similar. First, when \( u \geq c \), i.e., \( \lambda_{1}(u) \geq 0 \), we have supercritical inflow and both characteristic variables \( \tilde{u}_{1} \) and \( \tilde{u}_{2} \) have to be specified; in other words, \( \tilde{u}^{-} = \tilde{u} \). From

|       | \( u \geq c \) | \( |u| < c \) | \( u \leq -c \) |
|-------|----------------|-------------|------------|
| \( x = 0 \) | 2              | 1, \( \tilde{u}_{2} \) | 0          |
| \( x = 1 \) | 0              | 1, \( \tilde{u}_{1} \) | 2          |

**Table 12.1.** Number of boundary conditions for the shallow-water equations.
Exercises (359)

(12.185) we deduce that the matrix \( \frac{∂}{∂u} (C_ℓ \circ C) = \frac{∂}{∂u} C_1 \frac{∂}{∂u} C \) has to be invertible. Using the mappings \((*)\) and \((**)*\), we can see that this is equivalent to the requirement

\[
\det \left( \frac{∂}{∂u} (C_ℓ \circ C) \right) = -\frac{1}{2} c^3 \det \left( \frac{∂C_1}{∂u} \right) \neq 0.
\]

Second, when \(|u| < c\) or, equivalently \(λ_1(u) < 0 < λ_2(u)\), the inflow or outflow is \textit{subcritical}. The characteristics \(C_2\) enter the domain and consequently \(\tilde{u}_2\) has to be imposed; i.e., \(\tilde{u}^+ = \tilde{u}_2\).

From (12.185) and \((**)*\) we can easily see that \(c_ℓ(ϕ, u)\) should satisfy the condition

\[
\frac{∂c_ℓ}{∂ϕ} + (u + c) \frac{∂c_ℓ}{∂(ϕu)} \neq 0.
\]

Third, when \(u \leq -c\), i.e., \(λ_2(u) \leq 0\), we have supercritical outflow and no boundary conditions are required. \(\square\)

12.8 Discussion

- Hyperbolic equations describe propagation phenomena such as the evolution of surfaces or the propagation of waves. Examples of the latter are water waves [90, 32, 150, 169], sound waves [114, 101], and electromagnetic waves [67, 94].
- Hyperbolic problems in more than one spatial dimension are more complicated than those in one dimension. For one thing the characteristics generalise to Monge cones and are already much more complex from a geometrical point of view. A large problem area is that of fluid dynamics. We refer the interested reader to [95].
- In Chapter 16 a number of problems deal with hyperbolic equations. In Section 16.2 the etching effect of high-speed powder blasting on glass is investigated. In Sections 16.9 and 16.10 the effect of atmospheric conditions on noise and the flow along curved surfaces (as in bowls or gutters) for thin layers, respectively, are investigated.

Exercises

12.1. Determine the solution of the problem

\[
x \frac{∂u}{∂t} - t \frac{∂u}{∂x} = 0, \quad x \in (0, ∞), \quad t > 0,
\]

\[
u(x, 0) = x^2, \quad x \in (0, ∞).
\]

12.2. Consider the equation

\[
\frac{∂u}{∂t} + \frac{∂u}{∂x} = u^2
\]

subject to the condition

\[
u(x, t) = t \quad \text{for} \quad (x, t) \in \mathcal{J} := \{(x, t) ∈ \mathbb{R}^2 | x + t = 0\}.
\]

Determine the solution on the relevant domain.
12.3. Determine the solution of the Cauchy problem

\[ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 1, \quad x \in \mathbb{R}, \quad t > 0, \]

\[ u(x, 0) = x, \quad x \in \mathbb{R}. \]

12.4. Find the weak solution of the Burgers’ equation if the initial condition is given by

\[ u(0, x) = \begin{cases} 
1 & \text{if } x \in (-\infty, -1), \\
0 & \text{if } x \in [-1, 1], \\
1 & \text{if } x \in (1, \infty). 
\end{cases} \]

12.5. Verify that the shock wave (12.42) is a weak solution of (12.34).

12.6. Verify that the rarefaction wave (12.43) is a weak solution of (12.34).

12.7. Consider the system of equations

\[ \frac{\partial u_1}{\partial t} + \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial x} = 0, \]

\[ \frac{\partial u_2}{\partial t} + \frac{\partial u_1}{\partial x} + 2\frac{\partial u_2}{\partial x} + \frac{\partial u_3}{\partial x} = 0, \]

\[ \frac{\partial u_3}{\partial t} - \frac{\partial u_1}{\partial x} + 2\frac{\partial u_2}{\partial x} = 0. \]

(a) Show that this system is hyperbolic.

(b) Determine the Riemann invariants.

(c) Discuss possible boundary conditions on an interval \((0, L)\).

12.8. A one-dimensional model problem from linear acoustics reads

\[ \frac{\partial p}{\partial t} + K_0 \frac{\partial u}{\partial x} = 0, \]

\[ \rho_0 \frac{\partial u}{\partial t} + \frac{\partial p}{\partial x} = 0, \]

where \(p\) and \(u\) are small perturbations to the pressure and velocity of the ambient fluid. \(K_0\) and \(\rho_0\) are constants. Repeat the questions from Exercise 12.7.

12.9. A model equation for one-dimensional electromagnetic waves propagating in the \(x\) direction reads

\[ \frac{\partial E}{\partial t} + \frac{1}{\epsilon_0 \mu_0} \frac{\partial B}{\partial x} = 0, \]

\[ \frac{\partial B}{\partial t} + \frac{\partial E}{\partial x} = 0, \]

where \(E\) and \(B\) are the electric field and magnetic inductance, respectively. Repeat the questions from Exercise 12.7.
12.10. A model problem from gas dynamics is the $p$ system [50], given by

\[
\begin{align*}
\frac{\partial v}{\partial t} - \frac{\partial u}{\partial x} &= 0, \\
\frac{\partial u}{\partial t} + \frac{\partial p(v)}{\partial x} &= 0,
\end{align*}
\]

where $v$, $u$, and $p(v)$ are the specific volume, velocity, and pressure of the gas, respectively.

(a) Show that this system is hyperbolic if $p'(v) < 0$ for all $v > 0$.

(b) Investigate the Riemann problem for this system.

(c) Discuss possible boundary conditions on an interval $(0, L)$.

12.11. Show that the Euler equations for a perfect gas are hyperbolic. Compute the Riemann invariants and give the decoupled system.

12.12. Consider the wave equation with a source term, i.e.,

\[
\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + 2 - 6x, \quad x \in (0, 1), \quad t > 0,
\]

subject to the boundary and initial conditions

\[
\begin{align*}
u(0, t) &= 0, & u(1, t) &= 1, & t > 0, \\
u(x, 0) &= x^3, & \frac{\partial u}{\partial t}(x, 0) &= 2x, & x \in (0, 1).
\end{align*}
\]

Determine the solution.

12.13. The expression (12.166) lends itself to obtaining the solution in $\mathbb{R}^2$ by employing the fact that $v$ and $w$ depend on $x$ and $y$ only.

(a) Show, using spherical coordinates, that the average $\bar{v}(r, t; \xi)$ with $\xi = (\xi, \eta, 0)$ reduces to

\[
\bar{v}(r, t; \xi) = \frac{1}{2\pi a t} \int_{D(\xi; at)} \frac{v(x)}{\sqrt{a^2 t^2 - \|x - \xi\|^2}} \, dx \, dy,
\]

with $D(\xi; at)$ the disc with centre $\xi$ and radius $at$, i.e., $D(\xi; at) := \{ x \in \mathbb{R}^2 \mid \|x - \xi\| \leq at \}$.

(b) Show then that the solution of the two-dimensional wave equation, subject to the initial conditions (12.156), is given by

\[
u(\xi, t) = \frac{\partial}{\partial t} \left( \frac{1}{2\pi a} \int_{D(\xi; at)} \frac{v(x)}{\sqrt{a^2 t^2 - \|x - \xi\|^2}} \, dx \, dy \right) + \frac{1}{2\pi a} \int_{D(\xi; at)} \frac{w(x)}{\sqrt{a^2 t^2 - \|x - \xi\|^2}} \, dx \, dy.
\]
12.14. Consider the damped wave equation
\[
\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}.
\]

(a) Write this equation as a linear first order system.
(b) Show that this system is hyperbolic.
(c) Give the decoupled system.
Chapter 13

Numerical Methods for Scalar Hyperbolic Equations

In this chapter we discuss a few numerical methods for scalar hyperbolic equations in one space dimension. We pay special attention to the advection equation as a model problem to introduce and analyse numerical methods. These numerical methods can then be generalized to nonlinear equations and even systems of equations. Therefore we start in Section 13.1 with the upwind and Lax–Wendroff schemes for the advection equation. These two schemes are the building blocks for many other numerical methods that we consider. Next, in Section 13.2, we investigate dissipation and dispersion of a numerical scheme. An alternative description of many numerical methods is based on the advection-diffusion equation. This will be discussed in Section 13.3. A special class of schemes are the nondissipative schemes, which are considered in Section 13.4. The Godunov method is a generalization of the upwind scheme to nonlinear equations and is presented in Section 13.5. From the analysis in Section 13.2 we can conclude that the upwind scheme is too dissipative and that the Lax–Wendroff scheme causes spurious oscillations in the vicinity of discontinuities. These shortcomings motivate us to introduce high-resolution schemes in Section 13.6. Two special examples of high-resolution schemes are the flux limiter method and slope limiter method. These are discussed in Sections 13.7 and 13.8, respectively. Finally, in Section 13.9 we give a concise description of numerical boundary conditions needed to carry out a computation on a finite domain.

13.1 Explicit One-Step Schemes for the Advection Equation

In Section 12.1 we introduced the advection equation, which will be our model equation here, with \( a = 1 \) and \( c = 0 \). So we shall study the equation

\[
\frac{\partial u}{\partial t} + b \frac{\partial u}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0,
\]

(13.1a)

where the advection velocity \( b \) is assumed constant. The methods we will discuss will, in principle, be applied to a Cauchy problem: i.e., for \( u \) the initial condition

\[
u(x, 0) = v(x), \quad x \in \mathbb{R},
\]

(13.1b)
is given, where \( v \) may be a discontinuous function. We recall that the exact solution of the initial value problem (13.1) is given by

\[
u(x, t) = v(x - bt).
\]

Clearly, we can use this problem as a benchmark to derive and analyse numerical schemes; these schemes are easily generalised for more complex problems (albeit these are often not as easy to analyse in such situations). In this section we will consider two typical explicit one-step methods: the upwind and Lax–Wendroff schemes. There exist a few more that are used in practice, but their derivation is similar to the two types discussed below.

### 13.1.1 The Upwind Scheme

In order to compute a numerical solution of the initial value problem (13.1), we use a grid with points \((x_j, t^n)\) defined by

\[
x_j := j/Delta_1x \quad (j = 0, \pm 1, \pm 2, \ldots), \quad t^n := n/Delta_1t \quad (n = 0, 1, 2, \ldots),
\]

for some spatial grid size \( Delta_1x \) and time step \( Delta_1t \). As we saw in Chapter 12, the characteristics play an important role and so we make a distinction between the cases \( b > 0 \) and \( b < 0 \). First, we take \( b > 0 \); i.e., (13.2) is a "wave" propagating in the positive \( x \) direction.

For the time derivative \( \partial u/\partial t (x_j, t^n) \) we use the forward difference approximation (5.70). In order to make the proper choice for the spatial derivative \( \partial u/\partial x (x_j, t^n) \), we note that \( u(x, t) \) is constant along characteristics \( x - bt = C \) \((C \in \mathbb{R})\); see (12.5). In particular, we have \( u(x_j, t^{n+1}) = u(x_j - bDt, t^n) \) and, since \( b > 0 \), this equation suggests the backward difference approximation (5.28b) for the spatial derivative \( \partial u/\partial x \). Denoting by \( u^n_j \) the numerical approximation of \( u(x_j, t^n) \), we thus obtain the difference scheme

\[
\frac{1}{Delta_1t} (u^{n+1}_j - u^n_j) + \frac{b}{Delta_1x} (u^n_j - u^n_{j-1}) = 0.
\]

This scheme is called the upwind scheme because it uses the numerical solution at the grid point \((x_{j-1}, t^n)\), which lies upwind of \((x_j, t^{n+1})\). The dependence of the numerical value \( u^{n+1}_j \) at the new time level \( t^{n+1} \) on the numerical values \( u^n_k \) \((k = j, j - 1)\) at the old time level \( t^n \) is displayed in the stencil in Figure 13.1. The scheme (13.4a) can be rewritten as

\[
u^{n+1}_j = u^n_j - c(u^n_j - u^n_{j-1}),\]

where \( c \) is the Courant number defined as

\[
c := bDelta_1t/Delta_1x.
\]

An alternative derivation of the upwind scheme is the following. Assume that \( bDelta_1t \leq Delta_1x \), so that \( x_j - bDelta_1t \in [x_{j-1}, x_j) \). Then \( u(x_j - bDelta_1t, t^n) \), and thus also \( u(x_j, t^{n+1}) \), can easily be approximated by linear interpolation, resulting in the scheme (13.4b); see Figure 13.1.

To analyse the stability of the upwind scheme we look for a planar wave solution of the form (cf. (3.33))

\[
u^n_j = e^{i(kx_j - i\omega t^n)}.
\]

where \( k \) is the wave number and \( \omega \) is the frequency. This planar wave is a solution of (13.4b), disregarding any boundary conditions, if the following discrete dispersion relation

\[
\frac{\omega}{Delta_1t} (k - c) = \frac{\omega}{Delta_1x},
\]

is satisfied. This gives the Courant–Friedrichs–Lewy (CFL) condition

\[
c = b/Delta_1t < 1.
\]

It is not difficult to show that the upwind scheme satisfies this condition for \( b > 0 \).
13.1. Explicit One-Step Schemes for the Advection Equation

Figure 13.1. Stencil for the upwind scheme. The dotted line is the characteristic through the point \((x_j, t^{n+1})\) facing back to the initial line \(t = 0\).

holds (see Section 3.4):

\[ e^{-i\omega \Delta t} = 1 - c \left( 1 - e^{-i\kappa \Delta x} \right). \]  

This relation is rather difficult to handle and therefore we introduce, dropping the dependencies on frequency and wave number, two new variables, the amplification factor \(\lambda\) and phase angle \(\phi\), i.e.,

\[ \lambda := e^{-i\omega \Delta t}, \quad \phi := \kappa \Delta x. \]  

(13.8)

Note that we thus have

\[ u^n_j = \lambda^n e^{i\phi}. \]  

(13.9)

Combining (13.7) and (13.8) results in

\[ \lambda(\phi) = 1 - c \left( 1 - e^{-i\phi} \right) = 1 - 2c \sin^2 \left( \frac{\phi}{2} \right) - ic \sin \phi. \]  

(13.10)

The Fourier modes (13.9) that can be represented on a fixed spatial grid have a phase \(\phi\) ranging from zero, corresponding to the constant mode \(u^n_j = \lambda^n\), to \(\pi\), corresponding to the high-frequency mode \(u^n_j = \lambda^n(-1)^l\) of wavelength \(2\pi/\kappa = 2\Delta x\). Therefore we have to investigate \(\lambda(\phi)\) for \(0 < \phi < \pi\). Taking the modulus of (13.10), we find

\[ |\lambda(\phi)|^2 = \left( 1 - 2c \sin^2 \left( \frac{\phi}{2} \right) \right)^2 + c^2 \sin^2 \phi \]  

(13.11)

Stability requires that \(|\lambda(\phi)| \leq 1\) and, obviously, this requirement is fulfilled if

\[ c = \frac{b \Delta t}{\Delta x} \leq 1. \]  

(13.12)

Therefore (13.12) yields the stability condition of the upwind scheme. Since it restricts the choice of the time step, the upwind scheme is only conditionally stable.
An interpretation of the stability condition (13.12) is the following. The inequality 
\[ \frac{b}{\Delta t} \frac{1}{\Delta x} \leq 1 \] 
means that the distance the solution of the advection equation propagates during one time step is smaller than the spatial grid size \( \Delta x \). Referring to Figure 13.1, we can see that this means that the characteristic through grid point \( P \) and facing back to the initial line \( t = 0 \), which is the (physical) domain of dependence of grid point \( P \), is contained in the triangle \( \triangle PQR \). From (13.4b) it is clear that the numerical solution in \( P \) depends on the numerical solution in all grid points in this triangle, and for this reason we call this set of grid points the **numerical domain of dependence** of grid point \( P \). The stability condition (13.12) thus means that the (physical) domain of dependence of \( P \) is enclosed by the numerical domain of dependence. Stated otherwise, the numerical scheme should include all physical information needed to compute the solution in grid point \( P \). This is called the **CFL condition** (named after Courant, Friedrichs, and Lewy).

Next we would like to determine the accuracy of the scheme. To that end we compute the **local discretisation error** \( d^n_j \) defined as
\[
 d^n_j := \frac{1}{\Delta t} \left( u(x_j, t^{n+1}) - u(x_j, t^n) \right) + \frac{b}{\Delta x} \left( u(x_j, t^n) - u(x_{j-1}, t^n) \right). 
\]
(13.13)
The local discretisation error \( d^n_j \) is thus the residual of the difference scheme (13.4a) after substituting the exact solution of (13.1a). Assuming that \( u(x, t) \) is sufficiently smooth and using Taylor series expansions of \( u(x, t) \) about \( (x_j, t^n) \), we find
\[
 d^n_j = \frac{1}{2} \frac{\Delta t}{\Delta x} \frac{\partial^2 u}{\partial t^2} (x_j, t^n) - \frac{1}{2} \frac{b \Delta x}{\Delta t} \frac{\partial^2 u}{\partial x^2} (x_j, t^n) + O(\Delta t^2) + O(\Delta x^2) 
\]
(13.14)
The second line in (13.14) follows from the relation \( \frac{\partial^2 u}{\partial t^2} = b^2 \frac{\partial^2 u}{\partial x^2} \), which readily follows from (13.1a). Consequently, the upwind scheme is **first order consistent** in the time step \( \Delta t \) and the spatial grid size \( \Delta x \), which was to be expected, since we only used one-sided differences for both \( \frac{\partial u}{\partial t} \) and \( \frac{\partial u}{\partial x} \).

For the other situation, where \( b < 0 \), the derivation is similar. We then take the forward difference approximation (5.28a) for the spatial derivative \( \frac{\partial u}{\partial x} (x_j, t^n) \) to obtain
\[
 \frac{1}{\Delta t} \left( u_{j+1}^{n+1} - u_{j}^{n+1} \right) + \frac{b}{\Delta x} \left( u_{j+1}^{n} - u_{j}^{n} \right) = 0. 
\]
(13.15)
The Courant number is negative and the stability condition now reads
\[
 \frac{b \Delta t}{\Delta x} \geq -1. 
\]
(13.16)
This scheme is again first order consistent in \( \Delta t \) and \( \Delta x \), since we used one-sided differences for the derivatives \( \frac{\partial u}{\partial t} \) and \( \frac{\partial u}{\partial x} \).

We can combine the difference schemes (13.4a) and (13.15) for the general case, i.e., when \( \text{sign}(b) \) is not known beforehand. To do this we introduce a splitting of the velocity \( b \) as follows:
\[
 b = b^+ + b^-, \quad b^+ := \max(b, 0) \geq 0, \quad b^- := \min(b, 0) \leq 0. 
\]
(13.17)
13.1. Explicit One-Step Schemes for the Advection Equation 367

A straightforward generalisation of both upwind schemes then reads

\[
\frac{1}{\Delta t}(u_j^{n+1} - u_j^n) + \frac{b^+}{\Delta x}(u_j^n - u_{j-1}^n) + \frac{b^-}{\Delta x}(u_{j+1}^n - u_j^n) = 0. \tag{13.18}
\]

The $b^+$ term in (13.18) is the backward difference approximation of $b^+\frac{\partial u}{\partial x}(x_j, t^n)$ and corresponds to a wave propagating in the positive $x$ direction. Likewise, the $b^-$ term in (13.18) is the forward difference approximation of $b^-\frac{\partial u}{\partial x}(x_j, t^n)$ and corresponds to a wave propagating in the negative $x$ direction. Thus the sign of $b^\pm$ determines the upwind direction, i.e., the direction of the one-sided approximation of $\frac{\partial u}{\partial x}(x_j, t^n)$. In Section 14.7.1 we will generalize this approach for hyperbolic systems. It is obvious that this version of the upwind scheme is first order consistent in $\Delta t$ and $\Delta x$.

Finally, we want to assess the stability of the scheme (13.18). Substituting the Fourier mode (13.9) in (13.18), we find (in a similar fashion as done for (13.11)) the following expression for the amplification factor:

\[
|\lambda(\phi)|^2 = 1 - 4|c|(1 - |c|) \sin^2\left(\frac{1}{2}\phi\right). \tag{13.19}
\]

From (13.19) we deduce that the upwind scheme (13.18) is stable under the condition

\[
\frac{|b|\Delta t}{\Delta x} \leq 1. \tag{13.20}
\]

This condition can again be interpreted as a CFL condition; i.e., the domain of dependence of a grid point is entirely enclosed by its numerical domain of dependence.

**Example 13.1** In this example we computed the upwind numerical solution of (13.1a) subject to the initial condition $v(x) = \sin 8\pi x$. We have chosen the (numerical) parameters $b = 1$, $\Delta x = 2.5 \times 10^{-2}$, and $\Delta t = 2 \times 10^{-2}$, resulting in a stable scheme for which $c = 0.8$. Comparing the numerical solution after 50 time steps with the exact solution (dotted line), we conclude that the upwind scheme suffers from severe damping (see Figure 13.2).

![Figure 13.2. Numerical approximation of a sine wave computed with the upwind scheme.](image)
13.1.2 The Lax–Wendroff Scheme

The upwind scheme in the previous section is only first order accurate. For practical problems, however, first order accuracy is often not enough. There are various schemes that give higher-order results, the best known being the Lax–Wendroff scheme; it is second order accurate in both $\Delta t$ and $\Delta x$. The derivation is somewhat unconventional. In Section 13.3 we address the idea behind this. For now let us consider the Taylor series expansion

$$u(x_j, t^{n+1}) = u(x_j, t^n) + \Delta t \frac{\partial u}{\partial t}(x_j, t^n) + \frac{1}{2} \Delta t^2 \frac{\partial^2 u}{\partial t^2}(x_j, t^n) + \mathcal{O}(\Delta t^3),$$

which includes terms up to and including $\frac{1}{2} \Delta t^2 \frac{\partial^2 u}{\partial t^2}$ in order to achieve second order accuracy in $\Delta t$. Next we replace the time derivatives in (13.21) by space derivatives. We use the relation $\frac{\partial^2 u}{\partial t^2} = b^2 \frac{\partial^2 u}{\partial x^2}$ (cf. (13.1a)) and substitute this in (13.21) to obtain the expansion

$$u(x_j, t^{n+1}) = u(x_j, t^n) - b \Delta t \frac{\partial u}{\partial x}(x_j, t^n) + \frac{1}{2} b^2 \Delta t^2 \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + \mathcal{O}(\Delta t^3).$$

In order to get a scheme that is also second order accurate in $\Delta x$, we approximate the space derivatives in (13.22) by central differences to find the Lax–Wendroff scheme

$$u_j^{n+1} = u_j^n - \frac{1}{2} c (u_{j+1}^n - u_{j-1}^n) + \frac{1}{2} c^2 (u_{j+1}^n - 2u_j^n + u_{j-1}^n).$$

The stencil of this scheme is shown in Figure 13.3. An alternative derivation of (13.23) is to approximate $u(x_j, t^{n+1}) = u(x_j - b \Delta t, t^n)$ by quadratic interpolation from the values $u_k^n (k = j, j \pm 1)$ at time level $t^n$.

![Figure 13.3. Stencil for the Lax–Wendroff scheme. The dotted line is the characteristic through the point $(x_j, t^{n+1})$ facing back to the initial line $t = 0.$]
13.1. Explicit One-Step Schemes for the Advection Equation 369

As before, we investigate the stability of (13.23) by substituting the Fourier mode (13.9). We find the following expression for the amplification factor:

$$\lambda(\varphi) = 1 - 2c^2 \sin^2\left(\frac{1}{2} \varphi\right) - i c \sin \varphi. \quad (13.24)$$

Taking the absolute value of $\lambda(\varphi)$, we find

$$|\lambda(\varphi)|^2 = 1 - 4c^2(1 - c^2) \sin^4\left(\frac{1}{2} \varphi\right), \quad (13.25)$$

from which we see that the Lax–Wendroff scheme is stable if

$$|c| = \frac{|b|}{\Delta x} \leq 1. \quad (13.26)$$

In order to find the local discretisation error, we have to scale (13.23) so that it is compatible with the advection equation (13.1a), i.e.,

$$\frac{1}{\Delta t} (u_{j+1}^{n+1} - u_j^n) + \frac{b}{2\Delta x} (u_{j+1}^{n} - u_{j-1}^{n}) - \frac{b^2\Delta t}{2\Delta x^2} (u_{j+1}^{n} - 2u_j^n + u_{j-1}^{n}) = 0. \quad (13.27)$$

The local discretisation error $d_j^n$ of the Lax–Wendroff scheme is then found as the residual of (13.27) after substituting the exact solution. So we have

$$d_j^n := \frac{1}{\Delta t} (u(x_j, t^{n+1}) - u(x_j, t^n)) + \frac{b}{2\Delta x} (u(x_{j+1}, t^n) - u(x_{j-1}, t^n)) - \frac{b^2\Delta t}{2\Delta x^2} (u(x_{j+1}, t^n) - 2u(x_j, t^n) + u(x_{j-1}, t^n)). \quad (13.28)$$

Assuming $u(x, t)$ is sufficiently smooth, we can compute $d_j^n$ by substituting Taylor expansions of $u(x, t)$ around $(x_j, t^n)$ in (13.28) to find

$$d_j^n = \frac{1}{6} \Delta t^2 \frac{\partial^3 u}{\partial t^3}(x_j, t^n) + \frac{1}{6} b \Delta x^2 \frac{\partial^3 u}{\partial x^3}(x_j, t^n) + O(\Delta t^3) + O(\Delta x^3) \quad (13.29)$$

The second line in (13.29) follows from the relation $\frac{\partial^3 u}{\partial t^3} = -b^3 \frac{\partial^3 u}{\partial x^3}$; see (13.1a). Thus, as was to be expected, the Lax–Wendroff scheme is second order consistent in $\Delta t$ and $\Delta x$.

Example 13.2 We repeat the computation of Example 13.1 using the Lax–Wendroff scheme. This time the damping of the numerical solution is less severe; however, now the numerical solution lags the exact solution (see Figure 13.4).
13.2 Dissipation and Dispersion of Numerical Schemes

In this section we will give a Fourier analysis for both the upwind scheme and the Lax–Wendroff scheme. As it will turn out, the numerical behaviour of the solution may differ from that of the exact solution not only in growth character (cf. stability criteria for the former) but also in what is called the phase. We will investigate this for the advection equation.

Consider the equation (13.1a). Like in Section 3.4, we look for a planar wave solution of the form

\[ u(x, t) = e^{i(\kappa x - \omega t)}, \]  

(13.30)

where \( \kappa \) is the wave number and \( \omega \) is the frequency. Substituting (13.30) in (13.1a), we find the dispersion relation

\[ \omega(\kappa) = b\kappa. \]  

(13.31)

This means that every mode of the form (13.30) propagates with phase velocity \( b = \omega/\kappa \). We thus could also write

\[ u(x, t) = e^{i\kappa(x - bt)}. \]  

(13.32)

From (13.30) and (13.31) we conclude that the exact planar wave amplifies by a factor \( e^{-i b \kappa/\Delta t} = e^{-i c \phi/\Delta t} \) per time step \( \Delta t \), so we have

\[ u(x_j, t^{n+1}) = e^{-i b \kappa \Delta t} u(x_j, t^n) = e^{-i c \phi} u(x_j, t^n). \]  

(13.33)

We compare this amplification to that of the numerical solution. Using the numerical dispersion relation, which we preferred to formulate in terms of the amplification factor and the phase angle, we obtain a relation \( \lambda = \lambda(\phi) \). Obviously, the amplification factor \( \lambda(\phi) \) of the numerical solution is an approximation of the amplification factor \( e^{-i c \phi} \) of the exact planar wave. So it makes sense to write \( \lambda \) in polar form:

\[ \lambda = |\lambda| e^{-i \phi}. \]  

(13.34)
where $|\lambda|$ is called the *amplitude* and $\psi$ is the *phase* of the amplification factor. The argument $\psi$ is apparently an approximation of $c\varphi = b\kappa \Delta t$. Hence we define a *numerical phase velocity* $\beta$ as

$$
\beta := \frac{\psi}{\kappa \Delta t}. 
$$

(13.35)

Inserting (13.34) and (13.35), we obtain the following representation of the numerical mode (13.9):

$$
u^n_j = |\lambda|^n e^{i(\kappa(x_j - \beta t^n))}.
$$

(13.36)

Thus the mode propagates with speed $\beta$ indeed.

In order to compare (13.36) with (13.32), it turns out to be convenient to define an *amplitude error* $\epsilon_a$ and a *relative phase error* $\epsilon_f$:

$$
\epsilon_a := 1 - |\lambda|, \quad \epsilon_f := \frac{c\varphi - \psi}{c\varphi} = 1 - \frac{\beta}{b}.
$$

(13.37)

In order to understand the definition of the amplitude error $\epsilon_a$, one should realise that the exact solution has an amplitude $|e^{-i\varphi}| = 1$. Clearly, only if $0 \leq \epsilon_a < 1$ is the scheme stable. Furthermore, we conclude that for $\epsilon_f > 0$ the numerical solution lags behind the exact solution, while for $\epsilon_f < 0$ it propagates faster. Combining the definitions of $\epsilon_a$ and $\epsilon_f$ with the representation (13.36), we find the following relation between the exact and numerical modes:

$$
u^n_j = ((1 - \epsilon_a) \exp(i\epsilon_f c\varphi))^n u(x_j, t^n),
$$

(13.38)

which confirms our previous conclusions.

Let us now apply this to the schemes we have seen so far, to begin with to the upwind scheme (13.4b). Combining the polar representation (13.34) with the expression in (13.10), we find the following relations for the amplitude $|\lambda|$ and phase $\psi$:

$$
|\lambda| = \sqrt{1 - 4c(1-c) \sin^2\left(\frac{1}{2}\varphi\right)}, \quad \tan \psi = \frac{c \sin \varphi}{1 - 2c \sin^2\left(\frac{1}{2}\varphi\right)}.
$$

(13.39)

From the first relation we can trivially determine the amplitude error. In order to derive an expression for the phase error, we have to invert the second relation in (13.39), and since $0 \leq \psi \leq \pi$, we have

$$
\psi = \tan^{-1}\left(\frac{c \sin \varphi}{1 - 2c \sin^2\left(\frac{1}{2}\varphi\right)}\right),
$$

(13.40)

where $\tan^{-1} : \mathbb{R} \rightarrow [0, \pi)$ is given by

$$
\tan^{-1}(q) := \begin{cases} 
\arctan q & \text{if } q \geq 0, \\
\arctan q + \pi & \text{if } q < 0.
\end{cases}
$$

(13.41)

In Figure 13.5 we have drawn $\epsilon_a$ and $\epsilon_f$ as functions of $\varphi$. We see from these pictures that low-frequency components, characterized by a small wave number $\kappa$ and a small phase angle $\varphi$, indeed have amplitude and phase errors close to zero. This means that these components are hardly damped and propagate with the correct velocity. On the other hand,
high-frequency components, for which \( \kappa \approx \pi / \Delta x \) and \( \varphi \approx \pi \), have a significant amplitude and phase errors. These components are badly damped and propagate at the wrong speed.

Next consider the Lax–Wendroff scheme (13.23). The amplification factor \( \lambda(\varphi) \) is given in (13.24). For the amplitude \( |\lambda| \) and the phase \( \psi \) we find

\[
|\lambda| = \sqrt{1 - 4c^2(1 - c^2) \sin^4 \left( \frac{1}{2} \varphi \right)}, \quad \tan \psi = \frac{c \sin \varphi}{1 - 2c^2 \sin^2 \left( \frac{1}{2} \varphi \right)}.
\]

Combining (13.37) and (13.42), we trivially obtain the amplitude and phase errors of the Lax–Wendroff scheme, which are shown in Figure 13.5. Essentially the same conclusions hold as for the upwind scheme. Moreover, comparing the upwind scheme with the Lax–Wendroff scheme, we see that the amplitude error of the latter is much closer to zero for low-frequency modes. This means that these Fourier modes are not so badly damped by the Lax–Wendroff scheme as by the upwind scheme.

The growth/decay of the numerical mode (13.36) is determined by the amplitude \( |\lambda(\varphi)| \). Ideally, we would like to have \( |\lambda(\varphi)| = 1 \), as is true for the exact solution. In practice we have to be content with \( |\lambda(\varphi)| \leq 1 \); i.e., the scheme has to be stable. Based on this observation we introduce the following definition.

Figure 13.5. Amplitude and phase errors of the upwind scheme (left) and the Lax–Wendroff scheme (right).
13.3. The Advection-Diffusion Equation

Definition 13.3. A numerical scheme for the advection equation (13.1a) is called dissipative if \(|\lambda(\varphi)| \leq 1\) for all \(\varphi (0 \leq \varphi \leq \pi)\) and the inequality holds for at least one \(\varphi\). On the other hand, a numerical scheme is called nondissipative if \(|\lambda(\varphi)| = 1\) for all \(\varphi (0 \leq \varphi \leq \pi)\).

A nondissipative scheme does not damp the numerical solution and thus corresponds to an amplitude error \(\epsilon_a = 0\). In Section 13.4 we will introduce two nondissipative schemes.

It is more difficult to recover the correct phase velocity. In fact, the numerical phase velocity \(\beta\) of any scheme depends on the wave number \(\kappa\). This leads to the following definition.

Definition 13.4. A numerical scheme for the advection equation (13.1a) is called dispersive if the numerical phase velocity \(\beta\) depends on the wave number \(\kappa\), i.e., \(d\beta/d\kappa \neq 0\).

In a dispersive scheme each mode in a solution will propagate with its own numerical phase velocity \(\beta\) depending on the wave number \(\kappa\), so that the solution will disperse.

Example 13.5 Let us compute numerical solutions of the initial value problem (13.1) using the upwind and the Lax–Wendroff schemes for the initial function

\[
v(x) = \begin{cases} 
1 & \text{if } 0 \leq x \leq 0.2, \\
0 & \text{if } x < 0, x > 0.2.
\end{cases}
\]

We choose \(\Delta x = 10^{-2}\), \(\Delta t = 0.8 \times 10^{-2}\), and \(b = 1\), and consequently \(c = 0.8\). The results are shown in Figure 13.6. From this figure we see that the upwind solution is damped too much, whereas the Lax–Wendroff solution contains oscillations. These oscillations are due to the large phase errors for high-frequency components: each component in the Fourier series of the initial function \(v(x)\) has a different numerical propagation velocity \(\beta = (1 - \epsilon_f)b\). In Section 13.7 we will combine the upwind and Lax–Wendroff schemes into a high-resolution scheme, which has an amplitude error comparable to the Lax–Wendroff scheme but does not produce spurious oscillations.

13.3 The Advection-Diffusion Equation

In practice we often encounter an advection-diffusion equation which has both an advective term and a (possibly small) diffusion term. Although this equation is of parabolic type, we have to use schemes from this chapter to discretise the advective term. Furthermore, it turns out that many numerical schemes for the advection equation (13.1a) can be described in terms of an advection-diffusion equation. Therefore we will introduce in Section 13.3.1 an explicit scheme for the advection-diffusion equation and investigate its stability. Next, in Section 13.3.2, we will give an alternative formulation of the upwind and Lax–Wendroff schemes.

13.3.1 An Explicit Scheme for the Advection-Diffusion Equation

In this section we consider the advection-diffusion equation

\[
\frac{\partial u}{\partial t} + b \frac{\partial u}{\partial x} = D \frac{\partial^2 u}{\partial x^2},
\]

(13.43)
Figure 13.6. Numerical solution computed with the upwind scheme (left) and the Lax–Wendroff scheme (right), after 25, 50, and 75 time steps, respectively.

where $D > 0$ is the diffusion coefficient. If we discretise the spatial derivatives by central differences and use the explicit Euler method for time integration, we obtain the scheme

$$u_j^{n+1} = u_j^n - \frac{1}{2} c (u_{j+1}^n - u_{j-1}^n) + d (u_{j+1}^n - 2u_j^n + u_{j-1}^n),$$

(13.44)

where $d := D\Delta t/\Delta x^2$ is the diffusion number; cf. (11.5). Obviously, this scheme is first order accurate in $\Delta t$ and second order in $\Delta x$. 
13.3. The Advection-Diffusion Equation

It is instructive to derive the dispersion relation for (13.43). Substituting the planar wave (13.30), we find

$$\omega(\kappa) = b\kappa - iD\kappa^2,$$

which then results in the damped mode

$$u(x, t) = e^{-D\kappa^2 t} e^{i\kappa(x-bt)}.$$

Clearly, the addition of a diffusion term leads to damping of the planar wave mode. We will see that a similar conclusion holds for numerical schemes.

To analyse the stability of (13.44) we substitute the mode (13.9) to find the amplification factor

$$\lambda(\varphi) = 1 - 4g(\varphi) \sin^2 \left(\frac{1}{2} \varphi\right) - i c \sin \varphi.$$

The modulus of $\lambda(\varphi)$ is given by

$$|\lambda(\varphi)|^2 = 1 - 4g(\varphi) \sin^2 \left(\frac{1}{2} \varphi\right),$$

where the function $g(\varphi)$ is defined as

$$g(\varphi) := 2d - 4d^2 \sin^2 \left(\frac{1}{2} \varphi\right) - c^2 \cos^2 \left(\frac{1}{2} \varphi\right).$$

For stability we need $|\lambda(\varphi)|^2 \leq 1$ and this requirement is met if $g(\varphi) \geq 0$ for $0 \leq \varphi \leq \pi$. We can easily verify that $g'(\varphi) = (-2d^2 + \frac{1}{4}c^2) \sin \varphi$, implying that $g(\varphi)$ is a monotone function, either increasing or decreasing, depending on the sign of $-2d^2 + \frac{1}{4}c^2$. In either case $g(\varphi) \geq 0$ only if $g(0) \geq 0$ and $g(\pi) \geq 0$. The inequality $g(0) \geq 0$ leads to the time step restriction

$$\Delta t \leq \frac{2D}{b^2}.$$

This condition is rather peculiar since it does not depend on $\Delta x$. It is sometimes referred to as the advection-diffusion barrier. The inequality $g(\pi) \geq 0$ results in

$$\Delta t \leq \frac{\Delta x^2}{(2D)},$$

which is the usual time step restriction for the explicit Euler scheme applied to the heat equation. Combining both results, we have

$$\Delta t \leq \min \left(\frac{2D}{b^2}, \frac{\Delta x^2}{(2D)}\right).$$

From (13.49c) we conclude that the scheme (13.44) becomes unconditionally unstable if $D = 0$, or, stated otherwise, a diffusion term is needed to keep the scheme stable.

There is another issue that needs further analysis, as is apparent from the following example.

Example 13.6 We compute a numerical solution of the advection-diffusion equation (13.43) at $t = 1$ with $b = 1$ and $D = 2.5 \times 10^{-3}$. This means that the advective term is highly dominant, and, consequently, the exact solution, indicated by the dotted line in Figure 13.7, has a very thin boundary layer near $x = 1$. We have applied scheme (13.44) with $\Delta x = 2.5 \times 10^{-2}$ and $\Delta t = 4 \times 10^{-3}$. We can easily check that the stability condition (13.49c) is satisfied. However, the condition $c/d \leq 2$ is clearly violated, leading to the numerical oscillations in the vicinity of $x = 1$. □
From the example above it is clear that, although the scheme (13.44) is stable, it can produce unrealistic numerical solutions. To investigate the oscillatory behaviour of the numerical solution we write (13.44) in the alternative form
\[ u^n_{j+1} = u^n_j - d \left( 1 + \frac{1}{2} Pe \right) (u^n_j - u^n_{j-1}) + d \left( 1 - \frac{1}{2} Pe \right) (u^n_{j+1} - u^n_j), \] (13.50)
where \( Pe := c/d = b \Delta x / D \) is the Péclet number. Replacing index \( j \) by \( j + 1 \) and subtracting (13.50) from the resulting equation, we find the relation
\[ u^{n+1}_{j+1} - u^{n+1}_j = (1 - 2d)(u^n_{j+1} - u^n_j) + d \left( 1 + \frac{1}{2} Pe \right) (u^n_j - u^n_{j-1}) + \left( 1 - \frac{1}{2} Pe \right) (u^n_{j+2} - u^n_{j+1}). \] (13.51)
If all coefficients on the right-hand side of (13.51) are positive, i.e., if \( d \leq \frac{1}{2} \) and \( |Pe| \leq 2 \), we see that
\[ \sum_j |u^{n+1}_{j+1} - u^{n+1}_j| \leq \sum_j |u^n_{j+1} - u^n_j|, \] (13.52)
where the summation is over all possible values of \( j \). This inequality suggests that an initially monotone solution cannot be corrupted by oscillations as in the example above. Indeed, we can prove that unphysical oscillations cannot occur if (13.52) holds; for a more rigorous analysis the reader is referred to Section 13.6. To conclude, scheme (13.44) should satisfy the stability condition (13.49c) and in addition the inequality \( |Pe| \leq 2 \) in order to prevent unphysical oscillations.

### 13.3.2 Numerical and Artificial Diffusion

In this section we consider again the advection equation (13.1a). As was noted in Section 13.1.2, the derivation of the Lax–Wendroff scheme was somewhat unconventional. Instead,
13.3. The Advection-Diffusion Equation

let us (naively) try to use the central difference approximation for the space derivative and the forward Euler method for time integration, leading to the scheme

\[ u^n_j = u^n_j - \frac{1}{2} c (u^n_{j+1} - u^n_{j-1}) \]  
(13.53)

To investigate stability we substitute the Fourier mode (13.9), resulting in the amplification factor

\[ \lambda(\varphi) = 1 - i c \sin \varphi. \]  
(13.54)

Hence we find

\[ |\lambda(\varphi)|^2 = 1 + c^2 \sin^2 \varphi \geq 1. \]  
(13.55)

We therefore conclude that this scheme is unconditionally unstable, which is in agreement with the analysis from the previous section. With the derivation of the Lax–Wendroff scheme in mind, we try to add a diffusion term to the scheme such that \( \text{Re}(\lambda) < 1 \), resulting in a (conditionally) stable scheme. Comparing (13.23) and (13.53), we see that the Lax–Wendroff scheme is precisely the central difference scheme (13.53) with a diffusion term added. Indeed, introducing the coefficient \( D := \frac{1}{2} b^2 \Delta t \), we can rewrite the Lax–Wendroff scheme in the form

\[ u^{n+1}_j = u^n_j - \frac{1}{2} c (u^n_{j+1} - u^n_{j-1}) + D \frac{\Delta t}{\Delta x^2} (u^n_{j+1} - 2u^n_j + u^n_{j-1}). \]  
(13.56)

Viewing \( D \) as a constant, (13.56) is the central difference approximation in space and the forward Euler approximation in time of the advection-diffusion equation (13.43). Consequently, the Lax–Wendroff scheme introduces the discretisation of the term \( D \frac{\partial^2 u}{\partial x^2} \), which can be interpreted as a diffusion. Therefore the central difference approximation of \( D \frac{\partial^2 u}{\partial x^2} \) is often called the numerical diffusion. We can now easily derive a stability condition from the theory of the previous section. For the particular choice \( D = \frac{1}{2} b^2 \Delta t \), the inequality (13.49a) is trivially satisfied, whereas (13.49b) is equivalent to \( |c| \leq 1 \); i.e., the stability condition (13.26) of the Lax–Wendroff scheme is recovered. The main conclusion of the analysis above is that the Lax–Wendroff scheme stabilizes the central difference scheme by adding numerical diffusion.

For the upwind scheme a similar situation applies, though it is less obvious. To see how we rewrite the upwind scheme (13.4b) for \( b > 0 \) as the central difference scheme (13.53) with a correction term added:

\[ u^{n+1}_j = u^n_j - \frac{1}{2} c (u^n_{j+1} - u^n_{j-1}) - c (u^n_j - u^n_{j-1}) + \frac{1}{2} c (u^n_{j+1} - u^n_{j-1}) \]  
(13.57)

Defining \( D := \frac{1}{2} b \Delta x \) and taking it constant, we see that (13.57) is the forward Euler approximation in time and the central difference approximation in space of (13.43). Since \( D > 0 \), the upwind scheme can thus be interpreted as the central difference scheme stabilized by adding numerical diffusion. Note that for \( 0 < c < 1 \) the upwind scheme has more numerical diffusion than the Lax–Wendroff scheme. This then explains the damping character of the upwind scheme. Of course, similar reasoning holds for \( b < 0 \).

Since it appears to be the diffusion (or in some physical applications the viscosity) that makes a scheme stable, one can also try to add some artificial diffusion before discretisation. Hence, instead of a typical hyperbolic problem, say

\[
\frac{\partial u}{\partial t} + \mathcal{L}[u] = 0, \quad (13.58a)
\]

where \( \mathcal{L} \) is some quasi-linear first order operator, one then solves

\[
\frac{\partial u}{\partial t} + \mathcal{L}[u] = D \frac{\partial^2 u}{\partial x^2}. \quad (13.58b)
\]

The parameter \( D \) is then chosen to produce a stable solution.

It may be clear that stability obtained by either numerical or artificial diffusion will have an effect on the accuracy of the numerical solution. This is demonstrated by the next example.

**Example 13.7** We once more compute a numerical solution of the advection-diffusion equation (13.43), this time using the upwind discretisation for the advective term \( b \frac{\partial u}{\partial x} \). All parameters are the same as in Example 13.6. The result is shown in Figure 13.8. Clearly, the numerical oscillations are gone; however, the boundary layer at \( x = 1 \) is only inaccurately resolved.

![Figure 13.8. Numerical solution of the advection-diffusion equation computed with the upwind scheme.](image)

### 13.4 Nondissipative Schemes

The numerical schemes we have encountered so far are dissipative, implying that the numerical solution is damped. In this section we introduce two nondissipative schemes: the box scheme and the leapfrog scheme. These schemes are very useful for computing smooth solutions of (13.1a). On the other hand, for nonsmooth solutions these methods are not suitable due to the phase error and the lack of damping.
13.4. Nondissipative Schemes

13.4.1 The Box Scheme

Apart from the explicit schemes studied so far, implicit schemes are also used. Consider, e.g., the scheme we obtain by employing the trapezoidal rule for time integration; cf. (5.73). Besides being second order accurate, this method is interesting for another reason. To see this note that the time-dependent behaviour of the planar wave (13.30) is characterized by the ODE

\[ \frac{du}{dt} = -i \omega u, \quad \omega \in \mathbb{R}. \]  

(13.59)

A typical solution of (13.59) is given by

\[ u(x,t) = Ce^{-i \omega t}, \text{ which has the property that} \]

\[ |u(t)| = |C| \text{ for all } t. \]

Applying the trapezoidal rule to (13.59) gives

\[ u^{n+1} = \frac{1 - \frac{1}{2}i \omega \Delta t}{1 + \frac{1}{2}i \omega \Delta t} u^n, \]  

(13.60)

with \( u^n \) the numerical approximation of \( u(t^n) \). Clearly, \( |u^{n+1}| = |u^n| \) which is a “non-growth/nondecay” property it shares with the exact solution. Therefore we call the time integration scheme (13.60) nondissipative.

Next we return to the advection equation (13.1a). In the following we assume that (13.1a) is defined on a finite interval, let’s say \( x \in (0,1) \). Thus, apart from an initial condition, we need a boundary condition as well. If \( b > 0 \), we have to prescribe \( u \) at the left boundary \( x = 0 \), and if \( b < 0 \), we have to prescribe it at \( x = 1 \); see Chapter 12 for a discussion of boundary conditions. In order to compute a numerical solution, we cover the domain \( \Omega = [0,1] \times [0,\infty) \) with a grid containing points \((x_j,t^n)\) defined as

\[ x_j := j \Delta x \quad (j = 0, 1, 2, \ldots, M), \quad t^n := n \Delta t \quad (n = 0, 1, 2, \ldots), \]  

(13.61)

with \( \Delta x = 1/M \) the spatial grid size and \( \Delta t > 0 \) the time step. For time integration we use the trapezoidal rule because we expect it to be nondissipative. In order to match the second order behaviour of the trapezoidal rule, we use the central difference approximation for the space derivative. However, in order to keep the stencil compact, we apply the central difference scheme at intermediate points \( x_{j+1/2} \) given by

\[ x_{j+1/2} := \frac{1}{2}(x_j + x_{j+1}), \quad (j = 0, 1, \ldots, M - 1). \]

This way, we obtain the scheme

\[ \frac{1}{\Delta t} (u^{n+1}_{j+1/2} - u^n_{j+1/2}) + \frac{1}{2} \frac{b}{\Delta x} \left( (u^n_{j+1} - u^n_j) + (u^{n+1}_{j+1} - u^{n+1}_j) \right) = 0. \]  

(13.62)

Finally, we approximate the intermediate values \( u(x_{j+1/2}, t^n) \) \((m = n, n + 1)\) by straight-forward linear interpolation, which is second order accurate as well, to find the box scheme

\[ u^{n+1}_j + u^{n+1}_{j+1} - (u^n_j + u^n_{j+1}) + \frac{c}{2} (u^n_{j+1} - u^n_j + u^{n+1}_{j+1} - u^{n+1}_j) = 0. \]  

(13.63)

The stencil of this scheme is displayed in Figure 13.9.

This scheme is implicit with two unknown variables at the new time level \( t^{n+1} \), thus resulting in a bidiagonal linear system. Such a system can be very efficiently solved without
Figure 13.9. Stencil for the box scheme. The open circles indicate the intermediate grid points.

Although the box scheme is implicit, the computational costs per time step are comparable to those of an explicit scheme!

Let us now investigate the numerical dispersion relation. Substituting the mode (13.9) in (13.64a), we obtain the relation
\[
\lambda e^{i\varphi} = 1 + \frac{1 - c}{1 + c} (e^{i\varphi} - \lambda),
\]
where \(\varphi = \kappa \Delta x\). Hence we find
\[
\lambda(\varphi) = \frac{(1 + c) e^{-i\varphi/2} + (1 - c) e^{i\varphi/2}}{(1 + c) e^{i\varphi/2} + (1 - c) e^{-i\varphi/2}} = \frac{\cos \frac{1}{2} \varphi - i c \sin \frac{1}{2} \varphi}{\cos \frac{1}{2} \varphi + i c \sin \frac{1}{2} \varphi}.
\]
\[\text{(13.66)}\]

We conclude that \(|\lambda(\varphi)| = 1\), as we hoped for. Although this scheme looks attractive, there is a snag since it suffers from dispersion. Comparing (13.66) with the polar representation (13.34) and using elementary trigonometric formulas, we find
\[
\cos \frac{1}{2} \psi \cos \varphi + c \sin \frac{1}{2} \psi \sin \varphi = \cos \frac{1}{2} \psi,
\]
\[\text{(13.67a)}\]
\[-c \sin \frac{1}{2} \psi \cos \varphi + \cos \frac{1}{2} \psi \sin \varphi = c \sin \frac{1}{2} \psi.
\]
\[\text{(13.67b)}\]
From this system we obtain for the phase angle $\psi$ that
\[
\tan \frac{1}{2} \psi = c \tan \frac{1}{2} \varphi.
\] (13.68)

The phase error of the box scheme is then given by
\[
\epsilon_f = 1 - \frac{2}{c \varphi} \tan^{-1} \left( c \tan \frac{1}{2} \varphi \right),
\] (13.69)
with $\tan^{-1}$ defined in (13.41); see Figure 13.10.

**Example 13.8** Let us compute two numerical solutions of the advection equation (13.1a) in the interval $(0, 1)$ using the box scheme (13.64a). The initial condition is either $v(x) = \sin 8\pi x$ or the step function
\[
v(x) = \begin{cases} 
1 & \text{if } x \leq 0.2, \\
0 & \text{if } x > 0.2.
\end{cases}
\]
Furthermore, we choose the parameter values $b = 0.2$, $\Delta x = 2.5 \times 10^{-2}$, and $\Delta t = 5 \times 10^{-2}$. The numerical solution after 40 time steps is displayed in Figure 13.11. The sine wave is well resolved and only a small phase shift is visible. On the other hand, the step function is completely corrupted by oscillations caused by the large phase error of the high-frequency components in the solution.

### 13.4.2 The Leapfrog Scheme

It is also possible to use explicit methods and still have nondissipation. The method we shall discuss in this section is, however, a two-step method; i.e., it uses the numerical solution at two previous time levels.

Like in the previous section, we consider the advection equation on the interval $(0, 1)$ and for its numerical solution we employ the grid defined in (13.61). The numerical scheme
Figure 13.11. Numerical sine wave (left) and step function (right) computed with the box scheme.

is the midpoint rule for time integration in combination with the central difference scheme for the space derivative. Both are second order accurate. The resulting scheme reads

\[
\frac{1}{2\Delta t}(u_{j+1}^{n+1} - u_{j-1}^{n+1}) + \frac{b}{2\Delta x}(u_{j+1}^{n} - u_{j-1}^{n}) = 0. \quad (13.70a)
\]

Rearranging terms, we find the leapfrog scheme

\[
u_{j+1}^{n+1} = u_{j-1}^{n} - c(u_{j+1}^{n} - u_{j-1}^{n}); \quad (13.70b)
\]

see Figure 13.12 for its stencil. Clearly, we have to find a second “initial” condition at \( t_1 = \frac{\Delta t}{2} \) in order to be able to start the computation. A possibility is to use a one-step method like the box scheme. For the second order wave equation, as will be treated in Chapter 14, a special gridding takes care of this problem more directly.

As before, we can obtain a numerical dispersion relation:

\[
\lambda = \frac{1}{\lambda} - c(e^{i\omega} - e^{-i\omega}). \quad (13.71)
\]

Figure 13.12. Stencil for the leapfrog scheme.
13.4. Nondissipative Schemes

Hence the amplification factor satisfies the quadratic equation

$$\lambda^2 + 2i\epsilon \lambda \sin \varphi - 1 = 0,$$

(13.72)

which has the two solutions

$$\lambda(\varphi) = \pm \sqrt{1 - c^2 \sin^2 \varphi - i c \sin \varphi}.$$  

(13.73)

Clearly, if $|c| \leq 1$, then $|\lambda(\varphi)| = 1$. Furthermore, the product of the roots is $-1$; hence they are truly complex, i.e., have a nonzero imaginary part. As a consequence, both roots are unimodular, so that the scheme is nondissipative indeed. On the other hand, if $|c| > 1$, then $|\lambda(\varphi)| > 1$ for some $\varphi$, resulting in an unstable scheme.

A particular feature of the leapfrog scheme, and in fact of any multistep method, is the fact that it introduces additional solutions, as is evidenced by the fact that more than one initial condition is needed to compute the numerical solution. In order to investigate this, we apply the midpoint rule to the ODE model problem (13.59). The result is

$$u^{n+1} = u^{n-1} - 2i\omega \delta t u^n,$$

with $u^n$ the approximation of $u(t^n)$. A typical solution $u^n = \mu^n$ exists if $\mu$ satisfies the characteristic equation

$$\mu^2 + 2i\omega \delta t \mu - 1 = 0.$$

For $\delta t \to 0$ we obtain the two expansions for the two roots

$$\mu_1 = 1 - i\omega \delta t - \frac{1}{2} o^2 \delta t^2, \quad \mu_2 = -1 - i\omega \delta t + \frac{1}{2} o^2 \delta t^2$$

as approximations of the exact amplification factor $e^{-i\omega \delta t}$. Obviously, only the first essential root makes sense, giving rise to an essential solution, i.e., a numerical solution that is related to the actual solution of (13.59). The other root is often referred to as spurious; see, e.g., [97]. This root generates a spurious solution that does not have a physical meaning at all. The spurious solution is often generated by initial or boundary conditions, and since it is not damped, it will manifest itself by oscillatory components in the solution.

For the essential root we can deduce from (13.73) and its polar representation (13.34) the relation

$$\sin \psi = c \sin \varphi, \quad \cos \psi \geq 0$$

(13.74)

for the phase $\psi$. Inverting this relation, we obtain for the phase error

$$\epsilon_\ell = 1 - \frac{1}{c \varphi} \arcsin(c \sin \varphi);$$

(13.75)

see Figure 13.13. From this figure we conclude that the phase error of the leapfrog scheme is significant, and, consequently, the scheme is highly dispersive. The performance of the leapfrog scheme is comparable to that of the box scheme: smooth solutions are well resolved on a fine enough grid and discontinuous solutions are completely destroyed by spurious oscillations.
13.4.3 Propagation of Wave Packets

In Section 3.4 we encountered wave packets, i.e., superpositions of planar waves all having nearly the same wave number and phase velocity. It is interesting to see what this implies for a numerical scheme. Recall the wave packet at time $t = 0$, as was defined in (3.36a); i.e., for a particular wave number $\kappa_0$ and a parameter $\epsilon$ we have

$$u(x, 0) = e^{i\kappa_0 x} f(\epsilon x).$$

(13.76)

Suppose that the constituent modes of the wave packet satisfy a dispersion relation $\omega = \omega(\kappa)$. In Section 3.4 we showed that the wave packet corresponding to (13.76) is then given by

$$u(x, t) = e^{i\kappa_0 x - i\omega_0 t} f(\epsilon (x - \omega_0 t)).$$

(13.77)

with $\omega_0 = \omega(\kappa_0)$ and $\omega_0' = \omega'(\kappa_0)$. This means that all wave crests propagate with phase velocity $\omega_0 / \kappa_0$, whereas the envelope of the packet propagates with the group velocity $\omega_0'$. For the special case of the advection equation we have $\omega(\kappa) = \beta \kappa$, and, consequently, $\omega_0 / \kappa_0 = \omega_0' = \beta$.

Modes propagate at different phase velocities with numerical approximations of wave packets, since any numerical scheme is dispersive. In order to investigate dispersion of a nondissipative scheme, it is convenient to formulate the discrete dispersion relation in terms of the (real) numerical frequency $\Omega$ as

$$\Omega(\kappa) := \beta \kappa = \psi / \Delta t,$$

(13.78)

with $\beta$ the numerical phase velocity; cf. (13.35). In analogy with the continuous case, we define the numerical group velocity $v_{ng}$ by

$$v_{ng} := \frac{d\Omega}{d\kappa} = \frac{1}{\Delta t} \frac{d\psi}{d\kappa}$$

(13.79)
As an illustration we compute the numerical group velocity of the box scheme (13.63). Combining relation (13.68) with the definition of $v_{ng}$, we obtain the expression

$$
\frac{v_{ng}}{b} = \frac{\cos^2 \frac{\psi}{2}}{\cos^2 \psi} = \left( \cos^2 \left( \frac{1}{2} \psi \right) + c^2 \sin^2 \left( \frac{1}{2} \psi \right) \right)^{-1}
$$

for the numerical group velocity. In Figure 13.14 we have drawn graphs of this function for several values of $c$.

**Example 13.9** Let us illustrate the effect of the numerical group velocity on a wave packet. To that purpose, we compute a numerical solution of the advection equation (13.1a) on the interval $(0, 4)$, subject to the initial condition

$$
v(x) = e^{-\xi(x-1)^2} \sin(\alpha \pi x).
$$

We choose the parameter values $b = 1$, $\xi = 10$, and $\alpha = 16$. We apply the box scheme with $\Delta x = 1.25 \times 10^{-2}$ and $\Delta t = 2.5 \times 10^{-2}$, so that $c = 2$. The situation at $t = 2$ is given in Figure 13.15; the solid line denotes the numerical wave packet and the dotted line denotes the envelope of the exact solution. In this case the numerical solution has a phase $\psi = \alpha \pi \Delta x = \pi/5$ and, consequently, $v_{ng} \approx 0.78$; see Figure 13.14. The numerical and the exact wave packets have travelled distances 1.55 and 2, respectively, at $t = 2$, as can be seen from Figure 13.15 when comparing the maximums of both solutions. □

### 13.5 The Godunov Scheme for Nonlinear Conservation Laws

In the preceding sections we analysed schemes based on finite difference methods. Moreover, we only dealt with the advection equation. In this section we will consider finite volumes instead (see Section 5.3). At the same time we will view them in the more general
context of nonlinear conservation laws. So consider the initial value problem

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{13.80a}
\]

\[
u(x, 0) = v(x), \quad x \in \mathbb{R}, \tag{13.80b}
\]

for \(u(x, t)\), where the flux \(f(u)\) is generally a nonlinear function of \(u\). Recall that for the advection equation we have \(f(u) = bu\). For the most part we will assume that \(f(u)\) is a convex function; i.e., \(f''(u) > 0\) for all \(u\). Nonlinear problems are essentially more difficult than linear ones. In particular, nonlinear equations allow for discontinuous solutions even if the initial condition is smooth; see Chapter 12. As we will see, the correct numerical computation of a discontinuous solution to a nonlinear equation puts extra requirements on a scheme.

The finite volume method we will discuss can be seen as a generalisation of the upwind scheme in the sense that both schemes use the exact local solution of the conservation law. Let us begin by introducing control volumes or cells \(V_j\) as follows:

\[
V_j := [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}], \quad x_{j+\frac{1}{2}} := \frac{1}{2}(x_j + x_{j+1}), \quad j = 0, \pm 1, \pm 2, \ldots, \tag{13.81}
\]

with \(x_j\) defined in (13.3); see Figure 13.16. The numerical solution \(u^n_j\) has to be interpreted as an approximation of the average value of \(u(x, t)\) over control volume \(V_j\) at time level \(t^n\). Associated with \(u^n_j\) is the function \(\bar{u}(x, t)\), defined as the solution of the following initial value problem with piecewise constant initial condition at \(t = t^n\):

\[
\frac{\partial \bar{u}}{\partial t} + \frac{\partial f(\bar{u})}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > t^n, \tag{13.82a}
\]

\[
\bar{u}(x, t^n) = u^n_j, \quad x \in V_j \quad (j = 0, \pm 1, \pm 2, \ldots). \tag{13.82b}
\]

As shown in Section 5.3 for a stationary conservation law, the finite volume method is based on the integral formulation of the corresponding equation. We modify this derivation for the

---

**Figure 13.15.** The effect of the numerical group velocity on a wave packet computed with the box scheme.
Figure 13.16. Piecewise constant numerical solution at time level $t^n$.

initial value problem (13.82). Integration of the conservation law (13.82a) over $V_j$ results in

$$\frac{d}{dt} \int_{V_j} \bar{u}(x, t^n) \, dx + f(\bar{u}(x_{j+\frac{1}{2}}, t^n)) - f(\bar{u}(x_{j-\frac{1}{2}}, t^n)) = 0. \quad (13.83)$$

Subsequent integration over the time interval $[t^n, t^{n+1}]$ gives

$$\int_{V_j} \bar{u}(x, t^{n+1}) \, dx - \int_{V_j} \bar{u}(x, t^n) \, dx + \int_{t^n}^{t^{n+1}} f(\bar{u}(x_{j+\frac{1}{2}}, t)) \, dt - \int_{t^n}^{t^{n+1}} f(\bar{u}(x_{j-\frac{1}{2}}, t)) \, dt = 0. \quad (13.84)$$

Introducing the numerical flux $F(u^n_j, u^n_{j+1})$ as the time average flux at the control volume edge $x_{j+\frac{1}{2}}$, i.e.,

$$F(u^n_j, u^n_{j+1}) := \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(\bar{u}(x_{j+\frac{1}{2}}, t)) \, dt, \quad (13.85)$$

we can write the “integral balance” (13.84) in the form

$$\int_{V_j} \bar{u}(x, t^{n+1}) \, dx - \int_{V_j} \bar{u}(x, t^n) \, dx = -\Delta t (F(u^n_j, u^n_{j+1}) - F(u^n_{j-1}, u^n_j)). \quad (13.86)$$

In (13.85) we have explicitly written the numerical flux at the cell boundary $x_{j+\frac{1}{2}}$ as a function of the numerical values $u^n_j$ and $u^n_{j+1}$ in the adjacent cells. The next step is to compute the numerical fluxes.

To compute the numerical flux $F(u^n_j, u^n_{j+1})$ we solve a local Riemann problem (see Section 12.2.2) at the cell edge $x_{j+\frac{1}{2}}$ consisting of the conservation law (13.82a) and the
piecewise constant initial condition

\[
\tilde{u}(x, t^n) = \begin{cases} 
  u^n_j & \text{if } x < x_{j+\frac{1}{2}}, \\
  u^n_{j+1} & \text{if } x > x_{j+\frac{1}{2}}.
\end{cases}
\]

(13.87)

The solution of this Riemann problem is a similarity solution of the form

\[
\tilde{u}(x, t) = u_R(\eta; u^n_j, u^n_{j+1}), \quad \eta := \frac{x - x_{j+\frac{1}{2}}}{t - t^n},
\]

(13.88)

where the function \( u_R \) depends on a single variable \( \eta \) and two parameters \( u^n_j \) and \( u^n_{j+1} \).

Since \( \eta = 0 \) for \( x = x_{j+\frac{1}{2}} \), the computation of the numerical flux from (13.85) is particularly simple, and we find

\[
F(u^n_j, u^n_{j+1}) = f(u_R(0; u^n_j, u^n_{j+1})).
\]

(13.89)

In the examples below we will actually compute the numerical flux (13.89) for two specific cases: the advection equation and the Burgers’ equation. In general the procedure may be rather involved; see, e.g., [87, 158].

As a final step, the numerical value \( u^n_{j+1} \) is then found as the average value of \( \tilde{u}(x, t^{n+1}) \) over the control volume \( V_j \):

\[
u^n_{j+1} := \frac{1}{\Delta x} \int_{V_j} \tilde{u}(x, t^{n+1}) \, dx.
\]

(13.90)

Substituting (13.90) in (13.86) leads to the Godunov scheme

\[
u^n_{j+1} = u^n_j - \frac{\Delta t}{\Delta x} (F(u^n_j, u^n_{j+1}) - F(u^n_{j-1}, u^n_j)),
\]

(13.91)

with the numerical flux \( F(u^n_j, u^n_{j+1}) \) as defined in (13.89).

**Example 13.10** Consider the advection equation (13.1a). The Riemann problem associated with this equation reads

\[
\frac{\partial \tilde{u}}{\partial t} + b \frac{\partial \tilde{u}}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > t^n,
\]

\[
\tilde{u}(x, t^n) = \begin{cases} 
  u^n_j & \text{if } x < x_{j+\frac{1}{2}}, \\
  u^n_{j+1} & \text{if } x > x_{j+\frac{1}{2}}.
\end{cases}
\]

We can easily verify that the solution of this problem is given by \( \tilde{u}(x, t) = \tilde{u}(x - b(t - t^n), t^n) \).

Consequently, the similarity solution \( u_R(\eta; u^n_j, u^n_{j+1}) \) is given by

\[
u_R(\eta; u^n_j, u^n_{j+1}) = \begin{cases} 
  u^n_j & \text{if } \eta < b, \\
  u^n_{j+1} & \text{if } \eta > b.
\end{cases}
\]

For the numerical flux \( F(u^n_j, u^n_{j+1}) = b u_R(0; u^n_j, u^n_{j+1}) \) we then find

\[
F(u^n_j, u^n_{j+1}) = \begin{cases} 
  bu^n_j & \text{if } b > 0, \\
  bu^n_{j+1} & \text{if } b < 0.
\end{cases}
\]

Substituting this flux into the Godunov scheme (13.91), we recover the upwind scheme (13.18). \( \square \)
Example 13.11 Consider the Burgers’ equation, for which $f(u) = \frac{1}{2}u^2$. The Riemann problem is then given by

$$\frac{\partial \tilde{u}}{\partial t} + \frac{\partial}{\partial x} \left( \frac{1}{2} \tilde{u}^2 \right) = 0, \quad x \in \mathbb{R}, \quad t > t^n,$$

$$\tilde{u}(x, t^n) = \begin{cases} u^n_j & \text{if} \quad x < x_{j+\frac{1}{2}}, \\ u^n_{j+1} & \text{if} \quad x > x_{j+\frac{1}{2}}. \end{cases}$$

For the similarity solution of this problem we distinguish two cases: a shock wave and a rarefaction wave; see Example 12.13. If $u^n_j > u^n_{j+1}$, the solution of the Riemann problem is a shock wave propagating with speed $s^n_j := \frac{1}{2}(u^n_j + u^n_{j+1})$. The similarity solution $u_R(\eta; u^n_j, u^n_{j+1})$ is then given by

$$u_R(\eta; u^n_j, u^n_{j+1}) = \begin{cases} u^n_j & \text{if} \quad \eta < s^n_j, \\ u^n_{j+1} & \text{if} \quad \eta > s^n_j. \end{cases}$$

Alternatively, if $u^n_j < u^n_{j+1}$, the solution of the Riemann problem is the rarefaction wave

$$u_R(\eta; u^n_j, u^n_{j+1}) = \begin{cases} u^n_j & \text{if} \quad \eta < u^n_j, \\ \eta & \text{if} \quad u^n_j < \eta < u^n_{j+1}, \\ u^n_{j+1} & \text{if} \quad \eta > u^n_{j+1}. \end{cases}$$

For the numerical flux $F(u^n_j, u^n_{j+1}) = \frac{1}{2} u^n_R(0; u^n_j, u^n_{j+1})$ we obtain the following result:

- If $u^n_j > u^n_{j+1}$, then

$$F(u^n_j, u^n_{j+1}) = \begin{cases} \frac{1}{2} (u^n_j)^2 & \text{if} \quad s^n_j > 0, \\ \frac{1}{2} (u^n_{j+1})^2 & \text{if} \quad s^n_j < 0, \end{cases}$$

with $s^n_j := \frac{1}{2}(u^n_j + u^n_{j+1})$.

- If $u^n_j < u^n_{j+1}$, then

$$F(u^n_j, u^n_{j+1}) = \begin{cases} \frac{1}{2} (u^n_j)^2 & \text{if} \quad u^n_j > 0, \\ 0 & \text{if} \quad u^n_j < 0 < u^n_{j+1}, \\ \frac{1}{2} (u^n_{j+1})^2 & \text{if} \quad u^n_{j+1} < 0. \end{cases}$$

The Godunov scheme for the Burgers’ equation is thus given by (13.91) with numerical flux $F(u^n_j, u^n_{j+1})$ given above.

Next we will briefly comment on the stability and accuracy of the Godunov scheme. In the derivation of the Godunov scheme we have tacitly assumed that $\tilde{u}(x, t)$ at the cell edges is not influenced by adjacent Riemann problems, so that $\tilde{u}(x_{j+1/2}, t) = u_R(0; u^n_j, u^n_{j+1}) = C (C \in \mathbb{R})$ for $t^n \leq t \leq t^{n+1}$. This condition is satisfied if the time step restriction

$$\frac{\Delta t}{\Delta x} |f'(u^n_j)| \leq 1 \quad (j = 0, \pm 1, \pm 2, \ldots), \quad (13.92)$$
holds, which means that the fastest of all waves emanating from the cell boundaries traverses at most one control volume during a time step; for more details see, e.g., [158, 168, 88]. The time step restriction (13.92) is a nonlinear variant of the CFL condition (13.20). Furthermore, it can be proved that the Godunov scheme is only first order accurate; see exercises. This is to be expected since it reduces to the upwind scheme for the linear advection equation.

The Godunov scheme is an example of a \textit{conservative scheme}, which has the general form

\[ u_{j}^{n+1} = u_{j}^{n} - \frac{\Delta t}{\Delta x} \left( F_{j+\frac{1}{2}}^{n} - F_{j-\frac{1}{2}}^{n} \right), \]

with $F_{j+\frac{1}{2}}^{n}$ the numerical flux at the cell interface $x_{j+\frac{1}{2}}$. Conservative schemes have the following nice property. Summing the relation (13.93) over any number of adjacent control volumes, ranging from $j = J_{1}$ to $j = J_{2}$, and rearranging terms, we find

\[ \Delta x \sum_{j=J_{1}}^{J_{2}} (u_{j}^{n+1} - u_{j}^{n}) = -\Delta t \left( F_{J_{2}+\frac{1}{2}}^{n} - F_{J_{1}-\frac{1}{2}}^{n} \right), \]

since at each interface numerical fluxes cancel, except at the utmost left and right ones at $x = x_{J_{1}-\frac{1}{2}}$ and $x_{J_{2}+\frac{1}{2}}$, respectively. This relation can be seen as a discrete version of the integral balance over the interval $(a, b) := (x_{J_{1}-\frac{1}{2}}, x_{J_{2}+\frac{1}{2}})$, i.e.,

\[ \int_{a}^{b} (u(x, t^{n+1}) - u(x, t^{n})) \, dx = -\int_{t^{n}}^{t^{n+1}} \left( f(u(b, t)) - f(u(a, t)) \right) \, dt, \]

which is the basic conservation law; see Section 12.2. Consequently, (discrete) conservation is guaranteed for any union of adjacent control volumes. For this reason a scheme of the form (13.93) is called conservative. The conservative form of a scheme is sometimes advantageous, as is evident from the following example.

**Example 13.12** We would like to compute the shock wave solution of the Burgers’ equation corresponding to the initial condition

\[ u(x, 0) = \begin{cases} 2 & \text{if } x < 0, \\ 1 & \text{if } x > 0. \end{cases} \]

According to (12.38), this shock wave is propagating at speed $s = 1.5$. In this case the Godunov scheme reduces to

\[ u_{j}^{n+1} = u_{j}^{n} - \frac{\Delta t}{\Delta x} \left( \frac{1}{2} \left( u_{j}^{n} \right)^{2} - \frac{1}{2} \left( u_{j-1}^{n} \right)^{2} \right). \]

Alternatively, we can write the Burgers’ equation in the equivalent form

\[ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0 \]

and apply the upwind scheme to this equation. Since $u > 0$, we then find the nonconservative formulation

\[ u_{j}^{n+1} = u_{j}^{n} - \frac{\Delta t}{\Delta x} u_{j}^{n} (u_{j}^{n} - u_{j-1}^{n}). \]

For both schemes we choose $\Delta x = 2.5 \times 10^{-2}$ and $\Delta t = 1.25 \times 10^{-2}$ and compute the solution at $t = 1$. The result is shown in Figure 13.17. Clearly, the Godunov scheme gives the correct location of the shock, although it is not very sharp. On the other hand, the upwind numerical solution propagates too slowly, resulting in the wrong position of the shock. □
From this example we conclude that nonconservative schemes are not suitable for computing discontinuous solutions of a nonlinear conservation law, since the exact location of the discontinuity is not correctly captured. The reason for this is that the discrete conservation property (13.94) is violated. On the other hand, conservative schemes do produce numerical solutions with the discontinuity at the correct location. Moreover, it can be shown that under certain (mild) conditions, the solution of a conservative scheme converges to the correct, entropy-satisfying, weak solution of the conservation law (13.80a); for more details see, e.g., [88].

13.6 High-Resolution Schemes

In Section 13.2 we found that the upwind scheme is very dissipative; i.e., the numerical solution is damped very much. On the other hand, the Lax–Wendroff scheme is less dissipative, so that numerical dispersion may become annoying. The latter is visible as spurious oscillations in the vicinity of discontinuities. In this and the following sections we will construct high-resolution schemes, which satisfy the following two requirements:

1. They are (at least) second order accurate for smooth solutions.

2. They give an accurate, nonoscillatory resolution of discontinuities.

Many numerical schemes for the scalar conservation law (13.80a) can be written in the conservative form (13.93), where the numerical flux \( F^n_{j+\frac{1}{2}} \) is an approximation of the exact flux \( f(u(x_{j+\frac{1}{2}}, t^n)) \). The numerical flux \( F^n_{j+\frac{1}{2}} \) depends on the neighbouring values \( u^n_j \) and \( u^n_{j+1} \), and we write

\[
F^n_{j+\frac{1}{2}} = F(u^n_j, u^n_{j+1}),
\]

where \( F = F(u, v) \) is called the numerical flux function. The dependence of the numerical flux function \( F \) on two values of the numerical solution leads, of course, to a three-point scheme.
Example 13.13 Comparing the upwind scheme (13.18) with the generic form in (13.93), we easily see that the upwind numerical flux is given by

\[ F(u^n_j, u^n_{j+1}) = b^+ u^n_j + b^- u^n_{j+1}. \]

Likewise, from (13.89) we find for the Godunov numerical flux

\[ F(u^n_j, u^n_{j+1}) = f(u^n_R(0; u^n_j, u^n_{j+1})). \]

It is sometimes convenient to write a three-point scheme in incremental form, i.e.,

\[ u^{n+1}_j = u^n_j - C^n_{j-\frac{1}{2}} (u^n_j - u^n_{j-1}) + D^n_{j+\frac{1}{2}} (u^n_{j+1} - u^n_j), \] (13.97)

with coefficients \( C^n_{j-\frac{1}{2}} \) and \( D^n_{j+\frac{1}{2}} \) depending on the numerical values \( u^n_k \) (\( k = j, j \pm 1 \)).

We first address the second requirement for high-resolution schemes mentioned above; i.e., the scheme should not produce spurious oscillations in the vicinity of discontinuities. This means that we require a scheme to be monotonicity preserving, according to the following definition.

**Definition 13.14.** A numerical scheme for conservation law (13.80a) is called monotonicity preserving if the following holds: If the set \( \{u^n_j\} \) is monotonically increasing, then \( \{u^{n+1}_j\} \) is also monotonically increasing, or, if \( \{u^n_j\} \) is monotonically decreasing, so is \( \{u^{n+1}_j\} \).

In order to quantify oscillations in a numerical solution, we define the total variation TV of a numerical solution \( u^n = (u^n_j) \) as

\[ TV(u^n) := \sum_{j=-\infty}^{\infty} |u^n_{j+1} - u^n_j|. \] (13.98)

Hence we also have the following definition.

**Definition 13.15.** A numerical scheme for conservation law (13.80a) is called total variation diminishing (TVD) if its solution satisfies \( TV(u^{n+1}) \leq TV(u^n) \) for any set of data \( u^n \) and for all \( n \geq 0 \).

It is clear that the total variation of the Lax–Wendroff numerical solution in Figure 13.6 is larger than the total variation of the initial solution due to the presence of oscillations. It is intuitively clear that a numerical scheme is monotonicity preserving when it is TVD, as is confirmed by the following theorem.

**Theorem 13.16.** A TVD scheme for conservation law (13.80a) is monotonicity preserving.

For a proof one may consult [59, 87]. Consequently, we have to look for schemes that are TVD. Conditions under which a numerical scheme in incremental form (13.97) is TVD are given in the following theorem.
Theorem 13.17 (Harten). A numerical scheme for conservation law (13.80a), written in incremental form (13.97), is TVD if the following inequalities hold:

\[ C^n_{j+\frac{1}{2}} \geq 0, \quad D^n_{j+\frac{1}{2}} \geq 0, \quad 0 \leq C^n_{j+\frac{1}{2}} + D^n_{j+\frac{1}{2}} \leq 1. \]  

(13.99)

Proof. Replacing index \( j \) by \( j+1 \) in difference scheme (13.97) yields

\[ u^n_{j+1} = u^n_j - C^n_{j+\frac{1}{2}}(u^n_{j+1} - u^n_j) + D^n_{j+\frac{1}{2}}(u^n_{j+2} - u^n_{j+1}). \]

Subtracting (13.97) from the latter equation, we get

\[ u^n_{j+1} - u^n_j = (1 - C^n_{j+\frac{1}{2}} - D^n_{j+\frac{1}{2}})(u^n_{j+1} - u^n_j) + D^n_{j+\frac{1}{2}}(u^n_{j+2} - u^n_{j+1}) + C^n_{j+\frac{1}{2}}(u^n_{j+1} - u^n_j). \]

Taking the absolute value of \( u^n_{j+1} - u^n_j \), summing over all indices \( j \), and using the triangle inequality results in

\[
\begin{align*}
\sum_{j=-\infty}^{\infty} |u^n_{j+1} - u^n_j| &\leq \sum_{j=-\infty}^{\infty} (1 - C^n_{j+\frac{1}{2}} - D^n_{j+\frac{1}{2}})|u^n_{j+1} - u^n_j| \\
&+ \sum_{j=-\infty}^{\infty} D^n_{j+\frac{1}{2}}|u^n_{j+2} - u^n_{j+1}| + \sum_{j=-\infty}^{\infty} C^n_{j+\frac{1}{2}}|u^n_{j+1} - u^n_{j-1}|.
\end{align*}
\]

The coefficients in front of the difference terms are all positive by virtue of the inequalities in (13.99). If we apply a shift in index in the second \((j' = j+1)\) and in the third \((j' = j-1)\) sums on the right-hand side of the inequality above, it is clear that these terms cancel against the corresponding terms in the first sum on the right-hand side. Consequently, this inequality reduces to

\[ \sum_{j=-\infty}^{\infty} |u^n_{j+1} - u^n_j| \leq \sum_{j=-\infty}^{\infty} |u^n_{j+1} - u^n_j|, \]

or, in other words, \( TV(u^{n+1}) \leq TV(u^n) \), proving the statement. \( \square \)

Example 13.18 If we compare the upwind scheme (13.18) to the incremental form (13.97), we see that

\[ C^n_{j+\frac{1}{2}} = c^+ := b^+ \Delta t/\Delta x, \quad -D^n_{j+\frac{1}{2}} = c^- := b^- \Delta t/\Delta x. \]

Hence the inequalities \( C^n_{j+\frac{1}{2}} \geq 0 \) and \( D^n_{j+\frac{1}{2}} \geq 0 \) are trivially satisfied. The third inequality in Theorem 13.17, \( C^n_{j+\frac{1}{2}} + D^n_{j+\frac{1}{2}} \leq 1 \), is satisfied if the stability condition (13.20) holds. Thus the upwind scheme is TVD if the CFL condition holds. \( \square \)

Example 13.19 The Lax–Wendroff scheme (13.23) can be written in incremental form as follows:

\[ u^n_{j+1} = u^n_j - \frac{1}{2}c((u^n_{j+1} - u^n_j) + (u^n_j - u^n_{j-1})) + \frac{1}{2}c^2((u^n_{j+1} - u^n_j) - (u^n_j - u^n_{j-1})) \]

\[ = u^n_j - \frac{1}{2}c(c+1)(u^n_j - u^n_{j-1}) + \frac{1}{2}c(c-1)(u^n_{j+1} - u^n_j). \]
Using (13.97), we can conclude that
\[ C_{j+\frac{1}{2}} = \frac{1}{2}c(c + 1) \] and
\[ D_{j+\frac{1}{2}} = \frac{1}{2}c(c - 1). \] If \( 0 < c \leq 1 \), we have the inequalities
\[ C_{j+\frac{1}{2}} \geq 0, \quad D_{j+\frac{1}{2}} \leq 0, \] and, alternatively, if \( -1 \leq c < 0 \), then
\[ C_{j+\frac{1}{2}} \leq 0, \quad D_{j+\frac{1}{2}} \geq 0, \] and, consequently, the Lax–Wendroff scheme can never be TVD.

Now we have to address the first requirement for high-resolution schemes, i.e., the second order accuracy. So let us determine the local discretisation error \( d_j^n \) of the numerical scheme (13.93) with the numerical flux given in (13.96). The local discretisation error for this scheme is defined by
\[
d_j^n := \frac{1}{\Delta t} (u(x_j, t^{n+1}) - u(x_j, t^n)) + \frac{1}{\Delta x} (F(u(x_j, t^n), u(x_{j+1}, t^n)) - F(u(x_{j-1}, t^n), u(x_j, t^n))). \tag{13.100}
\]
with \( u(x, t) \) the solution of the conservation law (13.80a). A general expression for the local discretisation error is given in the following property.

**Property 13.20.** If the numerical flux function \( F(u, v) \) satisfies \( F(u, u) = f(u) \) and if it is twice continuously differentiable, then the local discretisation error is given by
\[
d_j^n = \frac{1}{\Delta t} \frac{\partial^2 u}{\partial t^2}(x_j, t^n) + \frac{\partial^2 F}{\partial v \partial u}(x_j, t^n). \tag{13.101}
\]
where all derivatives of \( F \) have to be evaluated at \((u(x_j, t^n), u(x_j, t^n))\).

**Proof.** Straightforward Taylor series expansion of \( u(x_j, t^{n+1}) \) gives, for the first term on the right-hand side of (13.100),
\[
\frac{1}{\Delta t} (u(x_j, t^{n+1}) - u(x_j, t^n)) = \frac{\partial u}{\partial t}(x_j, t^n) + \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_j, t^n) + O(\Delta t^2). \tag{*}
\]
Let the auxiliary variables \( A \) and \( B \) be defined by
\[
A := u(x_{j+1}, t^n) - u(x_j, t^n), \quad B := u(x_{j-1}, t^n) - u(x_j, t^n).
\]
Again, straightforward Taylor series expansion of \( u(x_{j\pm 1}, t^n) \) gives
\[
A = \Delta x \frac{\partial u}{\partial x}(x_j, t^n) + \frac{1}{2} \Delta x^2 \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta x^3),
B = -\Delta x \frac{\partial u}{\partial x}(x_j, t^n) + \frac{1}{2} \Delta x^2 \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta x^3).
\]
Expanding the flux terms on the right-hand side of (13.100), we get
\[
F(u(x_j, t^n), u(x_{j+1}, t^n)) = F + A \frac{\partial F}{\partial v} + \frac{1}{2} A^2 \frac{\partial^2 F}{\partial v^2} + O(A^3),
F(u(x_{j-1}, t^n), u(x_j, t^n)) = F + B \frac{\partial F}{\partial u} + \frac{1}{2} B^2 \frac{\partial^2 F}{\partial u^2} + O(B^3),
\]
where $F$ and all its derivatives on the right-hand sides have to be evaluated at $(u(x_j, t^n), u(x_{j+1}, t^n))$. Substituting the expressions for $A$ and $B$ into the fluxes above, we find

$$F(u(x_j, t^n), u(x_{j+1}, t^n)) - F(u(x_j-1, t^n), u(x_j, t^n))$$

$$= \Delta x \left( \frac{\partial F}{\partial u} + \frac{\partial F}{\partial v} \right) \frac{\partial u}{\partial x}(x_j, t^n) + \frac{1}{2} \Delta x^2 \left( \frac{\partial F}{\partial v} - \frac{\partial F}{\partial u} \right) \frac{\partial^2 u}{\partial x^2}(x_j, t^n)$$

$$+ \frac{1}{2} \Delta x^2 \left( \frac{\partial^2 F}{\partial u^2} - \frac{\partial^2 F}{\partial v^2} \right) \left( \frac{\partial u}{\partial x} \right)^2(x_j, t^n) + O(\Delta x^3), \quad (**)$$

The first term on the right-hand side of this equation equals

$$\Delta x \frac{\partial}{\partial x} F(u, u) = \Delta x \frac{\partial}{\partial x} f(u),$$

since $F(u, u) = f(u)$. Substitution of $(*)$ and $(**)$ into (13.100) finally gives the expression in (13.101).

Consider, as an example, the advection equation (13.1a) with $b > 0$. In this case $F(u, v) = bu$ and, substituting this numerical flux function in (13.101), we precisely find (13.14).

In the remainder of this section we restrict ourselves to the advection equation (13.1a) with $b > 0$. A high-resolution scheme for this equation should be TVD and second order accurate. These two requirements are not easily realized, as shown by the following theorem.

**Theorem 13.21.** A linear, three-point scheme for the advection equation (13.1a) that is TVD is at most first order accurate.

**Proof.** Since we are dealing with a linear equation, the numerical flux function $F(u, v)$ can generally be written as

$$F(u, v) = bu + \gamma b(v - u)$$

for some constant $\gamma$. Note that this flux function satisfies $F(u, u) = bu$ and, consequently, the resulting scheme is consistent. Substituting this flux function in (13.93) gives

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( bu_j^n + \gamma b(u_{j+1}^n - u_j^n) - bu_{j-1}^n - \gamma b(u_j^n - u_{j-1}^n) \right)$$

$$= u_j^n - c((1 - \gamma)(u_j^n - u_{j-1}^n) + \gamma(u_{j+1}^n - u_j^n)). \quad (*)$$

Using (13.101), we obtain for the local discretisation error

$$d_j^n = \frac{1}{2} \Delta t \frac{\partial^2 u}{\partial t^2}(x_j, t^n) + \frac{1}{2} \Delta x(2\gamma - 1) b \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta t^2) + O(\Delta x^2).$$

Since $\frac{\partial^2 u}{\partial x^2} = b^2 \frac{\partial^2 u}{\partial x^2}$, we thus find

$$d_j^n = \frac{1}{2} b \Delta x(c + 2\gamma - 1) \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta t^2) + O(\Delta x^2).$$
Let the scheme be second order accurate. From the expression for $d^n_j$ we see that this second order accuracy can only be obtained if $\gamma = (1 - c)/2$. Consider now the following numerical solution at time level $t^n$:

$$u^n_j = \begin{cases} 0 & \text{if } j < 0, \\ 1 & \text{if } j \geq 0. \end{cases}$$

It is obvious that $TV(u^n) = 1$. We want to compute $TV(u^{n+1})$. Replacing index $j$ by $j + 1$ in scheme $(\ast)$, we find

$$u^{n+1}_{j+1} = u^n_{j+1} - c((1 - \gamma)(u^n_{j+1} - u^n_j) + \gamma(u^n_{j+2} - u^n_{j+1})).$$

(\ast\ast)

Subtracting $(\ast)$ from $(\ast\ast)$, we obtain

$$u^{n+1}_{j+1} - u^n_{j+1} = (1 + c(2\gamma - 1))(u^n_{j+1} - u^n_j) - c\gamma(u^n_{j+2} - u^n_{j+1}) + c(1 - \gamma)(u^n_j - u^n_{j-1}).$$

This then makes it possible to compute $TV(u^{n+1})$. For the specific choice of $u^n_j$ only three terms contribute to $TV(u^{n+1})$, i.e., the terms corresponding to $j = -2, -1, 0$ in $(\ast\ast)$. Thus we find

$$TV(u^{n+1}) = c(|\gamma| + |1 - \gamma|) + 1 + c(2\gamma - 1),$$

For the choice $\gamma = (1 - c)/2$, which is necessary to have second order accuracy, this expression reduces to

$$TV(u^{n+1}) = \frac{1}{2}c(|1 - c| + |1 + c|) + 1 - c^2.$$

We only have to consider the case $0 < c < 1$, because for $c > 1$ the scheme is unstable. For $0 < c < 1$ we have $TV(u^{n+1}) = 1 + c(1 - c) > 1 = TV(u^n)$, implying that the scheme is not TVD. This is a contradiction, and therefore our assumption in the beginning that the scheme was second order accurate is false.

### 13.7 A Flux Limiter Scheme for the Advection Equation

In the previous section we introduced the notion of high-resolution schemes for the scalar conservation law (13.80a). For the special case of the advection equation (13.1a) we proved that the two requirements for high-resolution schemes, second order accuracy and nonoscillatory representation of discontinuities, cannot be combined into a linear scheme. Consequently, we have to look for nonlinear high-resolution schemes. In this section we will construct a high-resolution scheme for the advection equation (13.1a) by combining the upwind and Lax–Wendroff schemes in a suitable manner. More specifically, we will combine the numerical fluxes of both schemes and apply a flux limiter in order to make the resulting scheme second order accurate and TVD.

Recall the general form of a three-point conservative scheme for (13.1a):

$$u^{n+1}_j = u^n_j - \frac{\Delta t}{\Delta x} (F(u^n_j, u^n_{j+1}) - F(u^n_{j-1}, u^n_j)).$$

(13.102)
We now take for the numerical flux $F(u^n_j, u^n_{j+1})$ a combination of the upwind flux $F_{uw}(u^n_j, u^n_{j+1})$ and the Lax–Wendroff flux $F_{LW}(u^n_j, u^n_{j+1})$ as follows:

\[
F(u^n_j, u^n_{j+1}) := F_{uw}(u^n_j, u^n_{j+1}) + \Phi^n_{j+\frac{1}{2}}(F_{LW}(u^n_j, u^n_{j+1}) - F_{uw}(u^n_j, u^n_{j+1})) ,
\]

where $\Phi^n_{j+\frac{1}{2}}$ is the flux limiter, which depends on the numerical solution at time level $t^n$. In regions where the solution is smooth we would like to have $\Phi^n_{j+\frac{1}{2}} = 0$, resulting in the upwind scheme. The limiter $\Phi^n_{j+\frac{1}{2}}$ is given by

\[
\Phi^n_{j+\frac{1}{2}} := \Phi(r^n_{j+\frac{1}{2}}),
\]

with $\Phi(r) \geq 0$ the limiter function and the variable $r^n_{j+\frac{1}{2}}$ a measure of the smoothness of the numerical solution near the cell boundary $x_{j+\frac{1}{2}}$. There are many possible choices for $r^n_{j+\frac{1}{2}}$. Usually, it is defined by

\[
r^n_{j+\frac{1}{2}} := \begin{cases} 
\frac{u^n_j - u^n_{j-1}}{u^n_{j+1} - u^n_j} & \text{if } b > 0, \\
\frac{u^n_{j+2} - u^n_{j-1}}{u^n_{j+1} - u^n_j} & \text{if } b < 0;
\end{cases}
\]

i.e., $r^n_{j+\frac{1}{2}}$ is the ratio of the upwind difference and the central difference of the numerical solution at the cell boundary $x_{j+\frac{1}{2}}$. Several limiter functions are used in practice; we will consider a few later on.

Let us first elaborate the numerical fluxes of the upwind and Lax–Wendroff schemes. For the upwind scheme (13.18) we find

\[
F_{uw}(u^n_j, u^n_{j+1}) = b^+u^n_j + b^-u^n_{j+1},
\]

with $b^+$ and $b^-$ as defined in (13.17). The determination of the Lax–Wendroff numerical flux is somewhat more involved. We actually need the flux difference $F_{LW}(u^n_j, u^n_{j+1}) - F_{uw}(u^n_j, u^n_{j+1})$, and for that reason we write the Lax–Wendroff scheme (13.23) as the upwind scheme plus a correction term. If $b > 0$, we have

\[
u^n_{j+1} = u^n_j - c(u^n_j - u^n_{j-1}) - \frac{1}{2}c(u^n_{j+1} - 2u^n_j + u^n_{j-1}) + \frac{1}{2}c^2(u^n_{j+1} - 2u^n_j + u^n_{j-1}) \\
= u^n_j - c(u^n_j - u^n_{j-1}) - \frac{1}{2}c(1-c)(u^n_{j+1} - 2u^n_j + u^n_{j-1}).
\]

Since stability requires $0 < c < 1$, we can interpret the Lax–Wendroff scheme as the upwind scheme with the central difference approximation of the antidiffusion term $-\frac{1}{2}b(1-c)\Delta x \frac{\partial}{\partial x}(x_j, t^n)$ added. Using the relation $u^n_{j+1} - 2u^n_j + u^n_{j-1} = (u^n_{j+1} - u^n_j) - (u^n_j - u^n_{j-1})$, we can rewrite (13.107) in the generic form (13.102), with numerical flux

\[
F_{LW}(u^n_j, u^n_{j+1}) = bu^n_j + \frac{1}{2}b(1-c)(u^n_{j+1} - u^n_j).
\]
Likewise, if \( b < 0 \), the Lax–Wendroff scheme can be written as

\[
\begin{align*}
       u_j^{n+1} &= u_j^n - c(u_{j+1}^n - u_{j-1}^n) + \frac{1}{2}c(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \\
       &= u_j^n - c(u_{j+1}^n - u_{j-1}^n) + \frac{1}{2}c(1 + c)(u_{j+1}^n - 2u_j^n + u_{j-1}^n).
\end{align*}
\]

(13.109)

Since we now need \(-1 < c < 0\) to have stability, we thus find that

\[
F_{\text{LW}}(u_j^n, u_j^{n+1}) = bu_j^{n+1} - \frac{1}{2}b(1 + c)(u_j^{n+1} + u_j^n).
\]

(13.110)

Combining (13.108) and (13.110) for the Lax–Wendroff flux into a single expression yields

\[
F_{\text{LW}}(u_j^n, u_j^{n+1}) = F_{\text{uw}}(u_j^n, u_j^{n+1}) + \frac{1}{2}|b|(1 - |c|)(u_{j+1}^n - u_j^n).
\]

(13.111)

The correction term in (13.111) is an antidiffusion flux that has the effect of steepening the upwind approximation of a discontinuity. Substituting (13.106) and (13.111) into (13.103), we obtain for the high-resolution flux

\[
F(u_j^n, u_{j+1}^n) = b^+u_{j+1}^n + b^-u_j^n + \frac{1}{2}|b|(1 - |c|)\Phi_{j+\frac{1}{2}}(u_{j+1}^n - u_j^n).
\]

(13.112)

In what follows we will derive constraints on the limiter function \( \Phi(r) \) such that the scheme (13.102) with numerical flux (13.112) is both TVD and second order accurate.

We now investigate when the scheme (13.102) is TVD. Substituting (13.112) into (13.102), we obtain the high-resolution scheme

\[
\begin{align*}
u_j^{n+1} &= u_j^n - c^+(u_j^+ - u_{j-1}^-) - c^-(u_{j+1}^- - u_j^+) \\
       &\quad - \frac{1}{2}|c|(1 - |c|)\Phi_{j+\frac{1}{2}}^+(u_{j+1}^+ - u_j^-) - \Phi_{j-\frac{1}{2}}^-(u_j^+ - u_{j-1}^-),
\end{align*}
\]

(13.113)

where \( c := b\Delta t/\Delta x, c^+ := b^+\Delta t/\Delta x \), and \( c^- := b^-\Delta t/\Delta x \). In order to determine under what conditions (13.113) is TVD, we have to rewrite it in the incremental form (13.97). An obvious choice for the coefficients \( C_{j-\frac{1}{2}}^n \) and \( D_{j+\frac{1}{2}}^n \) is

\[
C_{j-\frac{1}{2}}^n = c^+ - \frac{1}{2}|c|(1 - |c|)\Phi_{j+\frac{1}{2}}^+, \quad D_{j+\frac{1}{2}}^n = -c^- - \frac{1}{2}|c|(1 - |c|)\Phi_{j-\frac{1}{2}}^-.
\]

Unfortunately, \( D_{j+\frac{1}{2}}^n < 0 \) if \( b > 0 \) and \( C_{j-\frac{1}{2}}^n < 0 \) if \( b < 0 \), so that we cannot apply Theorem 13.17 to prove that (13.113) is TVD. However, other choices for the coefficients \( C_{j-\frac{1}{2}}^n \) and \( D_{j+\frac{1}{2}}^n \) are possible. Let \( b > 0 \). Using the definition of \( r_{j+\frac{1}{2}}^n \), we can rewrite (13.113) as a modification of the upwind scheme in the following way:

\[
u_j^{n+1} = u_j^n - c\left(1 + \frac{1}{2}(1 - c)\left(\Phi_{j+\frac{1}{2}}^+ - \Phi_{j-\frac{1}{2}}^-\right)\right)(u_j^n - u_{j-1}^n).
\]

(13.114)
13.7. A Flux Limiter Scheme for the Advection Equation

This then corresponds to the choice

\[ C^n_{j-\frac{1}{2}} = c \left( 1 + \frac{1}{2} (1 - c) \left( \frac{\Phi^n_{j+\frac{1}{2}}}{r^n_{j+\frac{1}{2}}} - \Phi^n_{j-\frac{1}{2}} \right) \right), \quad D^n_{j+\frac{1}{2}} = 0. \] (13.115)

From (13.99) we see that (13.114) is TVD if \( 0 \leq C^n_{j-\frac{1}{2}} \leq 1 \). For \( b < 0 \) we can rewrite (13.113) as

\[ u^n_{j+1} = u^n_j - c \left( 1 - \frac{1}{2} (1 + c) \left( \frac{\Phi^n_{j+\frac{1}{2}}}{r^n_{j+\frac{1}{2}}} - \frac{\Phi^n_{j-\frac{1}{2}}}{r^n_{j-\frac{1}{2}}} \right) \right) (u^n_{j+1} - u^n_j). \] (13.116)

Obviously, the coefficients \( C^n_{j-\frac{1}{2}} \) and \( D^n_{j+\frac{1}{2}} \) are now given by

\[ C^n_{j-\frac{1}{2}} = 0, \quad D^n_{j+\frac{1}{2}} = -c \left( 1 - \frac{1}{2} (1 + c) \left( \frac{\Phi^n_{j+\frac{1}{2}}}{r^n_{j+\frac{1}{2}}} - \frac{\Phi^n_{j-\frac{1}{2}}}{r^n_{j-\frac{1}{2}}} \right) \right). \] (13.117)

From (13.99) we conclude that scheme (13.116) is TVD if \( 0 \leq D^n_{j+\frac{1}{2}} \leq 1 \).

We have to elaborate the inequalities for \( C^n_{j-\frac{1}{2}} \) and \( D^n_{j+\frac{1}{2}} \) mentioned above. For \( b > 0 \) and \( 0 < c \leq 1 \) the inequalities \( 0 \leq C^n_{j-\frac{1}{2}} \leq 1 \) are equivalent to

\[ \frac{-2}{1 - c} \leq \frac{\Phi^n_{j+\frac{1}{2}}}{r^n_{j+\frac{1}{2}}} - \Phi^n_{j-\frac{1}{2}} \leq \frac{2}{c}. \] (13.118)

Likewise, for \( b < 0 \) and \(-1 \leq c < 0 \) the inequalities \( 0 \leq D^n_{j+\frac{1}{2}} \leq 1 \) can be reformulated as

\[ \frac{2}{c} \leq \frac{\Phi^n_{j+\frac{1}{2}}}{r^n_{j+\frac{1}{2}}} - \Phi^n_{j-\frac{1}{2}} \leq \frac{2}{1 + c}. \] (13.119)

Maximising the lower bounds and minimising the upper bounds, we can derive sufficient conditions for the inequalities in (13.118) and (13.119) to hold. More precisely, the upper bound \( 2/c \) in (13.118) attains its minimum value for \( c = 1 \) and the lower bound \( 2/c \) in (13.119) has its maximum value for \( c = -1 \). Likewise, the choice \( c = 0 \) gives the maximum value for the lower bound \(-2/(1-c)\) in (13.118) and the minimum value for the upper bound \(2/(1+c)\) in (13.119). This way, we find the sufficient condition

\[ \left| \frac{\Phi(r_1)}{r_1} - \Phi(r_2) \right| \leq 2, \quad \text{for all } r_1, r_2 \] (13.120)

for the inequalities in (13.118) and (13.119). For \( r_1, r_2 > 0 \) the inequality in (13.120) holds if \( \Phi(r) \) satisfies

\[ 0 \leq \Phi(r) \leq \min(2, 2r) \quad \text{for all } r > 0. \] (13.121)

The region in the \((r, \Phi)\) plane satisfying (13.121) is referred to as the TVD region and is shown in Figure 13.18. Recall that near extrema we apply the upwind scheme, which
means that \( \Phi(r) = 0 \) for \( r \leq 0 \). Thus near extrema the generation of spurious oscillations is prevented at the cost of a lower accuracy.

It remains to be investigated whether (13.113) is second order accurate. From the foregoing it is clear that (13.113) can only be second order accurate away from extrema. This is stated in the following theorem.

**Theorem 13.22.** If the limiter function \( \Phi(r) \) satisfies \( \Phi(1) = 1 \) and is Lipschitz continuous near \( r = 1 \), then the high-resolution scheme (13.113) with \( \Phi^n_{j+\frac{1}{2}} \) given by (13.104) and (13.105) is second order consistent on smooth regions where \( \frac{\partial u}{\partial x} \) is bounded away from zero.

**Proof.** We have to prove that the local discretisation error is second order in \( \Delta t \) and \( \Delta x \).

The high-resolution scheme (13.113) can be written in the form

\[
\frac{1}{\Delta t} (u^+_{j+1} - u^+_j) + \frac{b^+}{\Delta x} (u^n_j - u^n_{j-1}) + \frac{b^-}{\Delta x} (u^n_{j+1} - u^n_j) \\
+ \frac{1}{2} |b|(1 - |c|) \left[ \Phi^n_{j+\frac{1}{2}} \Phi^n_{j+\frac{1}{2}} \frac{1}{\Delta x} (u^n_{j+1} - u^n_j) - \Phi^n_{j-\frac{1}{2}} \frac{1}{\Delta x} (u^n_j - u^n_{j-1}) \right] = 0.
\]

Substituting the exact solution of the advection equation (13.1a) into this scheme gives the following expression for the local discretisation error \( d^n_j \):

\[
d^n_j = \frac{1}{\Delta t} (u(x_j, t^{n+1}) - u(x_j, t^n)) + \frac{b^+}{\Delta x} (u(x_j, t^n) - u(x_{j-1}, t^n)) \\
+ \frac{b^-}{\Delta x} (u(x_{j+1}, t^n) - u(x_j, t^n)) + \frac{1}{2} |b|(1 - |c|) \left[ \Phi^n_{j+\frac{1}{2}} \Phi^n_{j+\frac{1}{2}} \frac{1}{\Delta x} (u(x_{j+1}, t^n) - u(x_j, t^n)) \\
- \Phi^n_{j-\frac{1}{2}} \frac{1}{\Delta x} (u(x_j, t^n) - u(x_{j-1}, t^n)) \right].
\]
with \( \bar{\phi}_{j+\frac{1}{2}}^n = \Phi(\widetilde{\phi}_{j+\frac{1}{2}}^n) \) and \( \bar{\phi}_{j+\frac{1}{2}}^n \) given by

\[
\bar{\phi}_{j+\frac{1}{2}}^n = \begin{cases} 
\frac{u(x_j, t^n) - u(x_{j-1}, t^n)}{u(x_{j+1}, t^n) - u(x_j, t^n)} & \text{if } b > 0, \\
\frac{u(x_{j+1}, t^n) - u(x_{j+\frac{1}{2}}, t^n)}{u(x_{j+\frac{1}{2}}, t^n) - u(x_j, t^n)} & \text{if } b < 0.
\end{cases}
\]

Introducing the auxiliary variables

\[
A := u(x_{j+1}, t^n) - u(x_j, t^n), \quad B := u(x_j, t^n) - u(x_{j-1}, t^n),
\]

we find the following condensed expression for \( d^n_j \):

\[
d^n_j = \frac{1}{\Delta t} \left( u(x_j, t^{n+1}) - u(x_j, t^n) \right) + \frac{1}{\Delta x} \left( b^+ + \frac{1}{2} |b| (1 - |c|) \bar{\phi}_{j+\frac{1}{2}}^n \right) A \\
+ \frac{1}{\Delta x} \left( b^+ - \frac{1}{2} |b| (1 - |c|) \bar{\phi}_{j-\frac{1}{2}}^n \right) B. \tag{*}
\]

Straightforward Taylor series expansion of \( u(x_j, t^{n+1}) \) gives, for the first term on the right-hand side of (*),

\[
\frac{1}{\Delta t} \left( u(x_j, t^{n+1}) - u(x_j, t^n) \right) = \frac{\partial u}{\partial t}(x_j, t^n) + \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_j, t^n) + O(\Delta t^2).
\]

Likewise, Taylor series expansion of \( u(x_{j\pm 1}, t^n) \) gives

\[
A = \Delta x \frac{\partial u}{\partial x}(x_j, t^n) + \frac{1}{2} \Delta x^2 \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta x^2),
\]

\[
B = \Delta x \frac{\partial u}{\partial x}(x_j, t^n) - \frac{1}{2} \Delta x^2 \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta x^3).
\]

Substituting these Taylor series expansions into expression (*) for the local discretisation error, we obtain

\[
d^n_j = \frac{\partial u}{\partial t}(x_j, t^n) + \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_j, t^n) + O(\Delta t^2) + \frac{1}{2} \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta x^2)
\]

\[
+ \left( b^+ + \frac{1}{2} |b| (1 - |c|) \bar{\phi}_{j+\frac{1}{2}}^n \right) \left( \frac{\partial u}{\partial x}(x_j, t^n) + \frac{1}{2} \Delta x \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta x^2) \right)
\]

\[
+ \left( b^+ - \frac{1}{2} |b| (1 - |c|) \bar{\phi}_{j-\frac{1}{2}}^n \right) \left( \frac{\partial u}{\partial x}(x_j, t^n) - \frac{1}{2} \Delta x \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta x^2) \right).
\]

Replacing in this expression the time derivatives by the appropriate space derivatives, i.e.,

\[
\frac{\partial u}{\partial t} = -b \frac{\partial u}{\partial x} \quad \text{and} \quad \frac{\partial^2 u}{\partial t^2} = b^2 \frac{\partial^2 u}{\partial x^2},
\]

and rearranging terms, we find

\[
d^n_j = \frac{1}{2} |b| (1 - |c|) \left( \bar{\phi}_{j+\frac{1}{2}}^n - \bar{\phi}_{j-\frac{1}{2}}^n \right) \frac{\partial u}{\partial x}(x_j, t^n)
\]

\[
+ \frac{1}{4} \Delta x |b| (1 - |c|) \left( \bar{\phi}_{j+\frac{1}{2}}^n + \bar{\phi}_{j-\frac{1}{2}}^n - 2 \right) \frac{\partial^2 u}{\partial x^2}(x_j, t^n) + O(\Delta t^2) + O(\Delta x^2).
\]
Applying the triangle inequality and assuming that \( u(x, t) \) is sufficiently smooth, we find the following upper bound for \( d^n_j \):

\[
|d^n_j| \leq \frac{1}{2} |b|(1 - |c|) \left| \Phi^n_{j+\frac{1}{2}} - \Phi^n_{j-\frac{1}{2}} \right| C_1 + \frac{1}{4} \Delta x |b|(1 - |c|) \left| \Phi^n_{j+\frac{1}{2}} + \Phi^n_{j-\frac{1}{2}} - 2 \right| C_2 + O(\Delta t^2) + O(\Delta x^2) \quad (**) \]

for some constants \( C_1, C_2 > 0 \). Next we have to estimate the terms involving the flux limiters \( \Phi^n_{j+\frac{1}{2}} \). Since the limiter function \( \Phi(r) \) is Lipschitz continuous near \( r = 1 \), with Lipschitz constant \( L \), say, we have the following estimates:

\[
\left| \Phi^n_{j+\frac{1}{2}} + \Phi^n_{j-\frac{1}{2}} - 2 \right| = \left| \Phi(\tilde{r}_{n+\frac{1}{2}}) - \Phi(1) + \Phi(\tilde{r}_{n-\frac{1}{2}}) - \Phi(1) \right| \leq L \left| \tilde{r}_{n+\frac{1}{2}} - 1 \right| + \left| \tilde{r}_{n-\frac{1}{2}} - 1 \right| \quad (***)
\]

In the derivation of the second inequality in (**) we have used that \( \Phi(1) = 1 \). Finally, we have to determine the ratios \( \tilde{r}_{n+\frac{1}{2}} \). To that purpose, we assume that \( b > 0 \). The case \( b < 0 \) is completely similar. Expanding \( u(x_{j+\frac{1}{2}}, t^n) \) in a Taylor series, we find

\[
\tilde{r}_{n+\frac{1}{2}} = 1 - \Delta x \frac{\partial^2 u}{\partial x^2}(x_j, t^n) / \frac{\partial u}{\partial x}(x_j, t^n) + O(\Delta x^2), \quad (****)
\]

provided that \( \frac{\partial u}{\partial x}(x_j, t^n) \neq 0 \). A similar expression holds for \( \tilde{r}_{n-\frac{1}{2}} \). Combining (**), (***) and (****), we find that \( d^n_j = O(\Delta t^2) + O(\Delta x^2) \) in regions where \( \frac{\partial u}{\partial x} \) is bounded away from zero.

We conclude this section by giving two limiter functions. A smooth limiter function is due to van Leer [85] and is given for \( r > 0 \) by

\[
\Phi(r) = \frac{2r}{1 + r}. \quad (13.122a)
\]

Another limiter function, which is especially suitable for representing discontinuities, is the superbee limiter function of Roe [133]. It is defined for \( r > 0 \) by

\[
\Phi(r) = \max(\min(1, 2r), \min(r, 2)). \quad (13.122b)
\]

Both limiter functions are shown in Figure 13.18.

### 13.8 Slope Limiter Methods

In Section 13.7 we constructed a high-resolution scheme for the advection equation (13.1a) by combining the upwind and Lax–Wendroff numerical fluxes and applying a flux limiter.
13.8. Slope Limiter Methods

to make the scheme TVD. In this section we will present an alternative formulation of high-resolution schemes that is easier to generalise to nonlinear equations. This formulation is based on a piecewise linear representation of the numerical solution. Applying this to the advection equation gives the Lax–Wendroff scheme, which is not TVD. In order to make the scheme TVD, we need slope limiters. As we will show, the resulting slope limiter method is identical to the flux limiter method derived in the previous section. Finally, we will briefly describe the extension to nonlinear equations.

13.8.1 A Slope Limiter Method for the Advection Equation

As we did for the derivation of the Godunov scheme discussed in Section 13.5, we introduce a function \( \bar{u}(x, t) \) that satisfies the initial value problem

\[
\frac{\partial \bar{u}}{\partial t} + b \frac{\partial \bar{u}}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > t^n, \quad (13.123a)
\]

\[
\bar{u}(x, t^n) = u^n_j + s^n_j (x - x_j), \quad x \in V_j \quad (j = 0, \pm 1, \pm 2, \ldots), \quad (13.123b)
\]

where \( u^n_j \) and \( s^n_j \) are the average value and slope, respectively, of \( \bar{u}(x, t^n) \) in the control volume \( V_j \); see Figure 13.19. The initial value \( \bar{u}(x, t^n) \) is thus piecewise linear in each control volume \( V_j \). The slope \( s^n_j \) is still unknown and has to be determined such that the resulting scheme is TVD. Note that for \( s^n_j = 0 \) we recover the upwind scheme. The slope limiter method is based on the integral formulation of the initial value problem (13.123).

Integrating the advection equation (13.123a) over control volume \( V_j \) gives

\[
\frac{d}{dt} \int_{V_j} \bar{u}(x, t) \, dx + b \left( \bar{u}(x_{j+\frac{1}{2}}, t) - \bar{u}(x_{j-\frac{1}{2}}, t) \right) = 0, \quad (13.124)
\]

![Figure 13.19. Piecewise linear numerical solution at time level \( t^n \).](image)
and subsequent integration over the time interval \([t^n, t^{n+1}]\) results in

\[
\int_{V_j} \tilde{u}(x, t^{n+1}) \, dx - \int_{V_j} \tilde{u}(x, t^n) \, dx + b \left[ \int_{t^n}^{t^{n+1}} \tilde{u}(x_{j+\frac{1}{2}}, t) \, dt - \int_{t^n}^{t^{n+1}} \tilde{u}(x_{j-\frac{1}{2}}, t) \, dt \right] = 0.
\] (13.125)

We can write the integral formulation (13.125) in the condensed form

\[
\int_{V_j} \bar{u}(x, t^{n+1}) \, dx - \int_{V_j} \bar{u}(x, t^n) \, dx = -\Delta t \left( F(u_j^n, u_{j+1}^n) - F(u_{j-1}^n, u_j^n) \right),
\] (13.126)

where the numerical flux \(F(u_j^n, u_{j+1}^n)\) is now given by

\[
F(u_j^n, u_{j+1}^n) = \frac{b}{\Delta t} \int_{t^n}^{t^{n+1}} \bar{u}(x_{j+\frac{1}{2}}, t) \, dt;
\] (13.127)

cf. (13.85). In order to complete the derivation of this version of the Godunov scheme, we assume that \(\bar{u}(x, t^{n+1})\) also has a piecewise linear distribution analogous to (13.123b). The integrals in (13.126) are then trivial and we find

\[
u_{j+1}^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( F(u_j^n, u_{j+1}^n) - F(u_{j-1}^n, u_j^n) \right).
\] (13.128)

The next step in the derivation of the slope limiter scheme is to compute the numerical fluxes from (13.127).

In order to compute the numerical flux \(F(u_j^n, u_{j+1}^n)\), we have to compute \(\bar{u}(x_{j+\frac{1}{2}}, t)\). Locally, i.e., in the vicinity of the cell boundary \(x_{j+\frac{1}{2}}\), the function \(\bar{u}(x, t)\) is the solution of the generalized Riemann problem consisting of (13.123a) and the piecewise linear initial condition

\[
\bar{u}(x, t^n) = \begin{cases} 
  u_j^n + s_j^n (x - x_j) & \text{if } x < x_{j+\frac{1}{2}}, \\
  u_{j+1}^n + s_{j+1}^n (x - x_{j+1}) & \text{if } x > x_{j+\frac{1}{2}}.
\end{cases}
\] (13.129)

As we have seen before, the general solution of (13.123a) is given by \(\tilde{u}(x, t) = \tilde{u}(x - b(t - t^n), t^n)\) \((t > t^n)\); i.e., the initial solution \(\tilde{u}(x, t^n)\) is propagated over a distance \(b(t - t^n)\). Applying this result to the initial solution in (13.129), we find, for \(b > 0\), that

\[
\tilde{u}(x_{j+\frac{1}{2}}, t) = \tilde{u}(x_{j+\frac{1}{2}} - b(t - t^n), t^n) = u_j^n + s_j^n(x_{j+\frac{1}{2}} - b(t - t^n) - x_j)
\] (13.130)

Likewise, we obtain for \(b < 0\) the solution

\[
\tilde{u}(x_{j+\frac{1}{2}}, t) = u_{j+1}^n - s_{j+1}^n \left( \frac{1}{2} \Delta x - b(t - t^n) \right).
\] (13.131)
We can now compute the numerical flux by substituting (13.130) and (13.131) into (13.127).

First, we consider the case $b > 0$. Then we find for the numerical flux

\[
F(u^n_j, u^n_{j+1}) = \frac{b}{\Delta t} \int_{t^n}^{t^{n+1}} \left( u^n_j + s^n_j \left( \frac{1}{2} \Delta x - b(t - t^n) \right) \right) \, dt
\]

\[
= \frac{b}{\Delta t} \left( u^n_j \Delta t + s^n_j \left( \frac{1}{2} \Delta x \Delta t - b \frac{1}{2} \Delta t^2 \right) \right)
\]

\[
= bu^n_j + \frac{1}{2} b(1 - c) \Delta x s^n_j,
\]

(13.132)

The computation of the numerical flux $F(u^n_j, u^n_{j+1})$ for $b < 0$ is completely similar to what we found in (13.132), and we obtain

\[
F(u^n_j, u^n_{j+1}) = bu^n_{j+1} - \frac{1}{2} b(1 + c) \Delta x s^n_{j+1}.
\]

(13.133)

For either sign of $b$ we can combine the expressions (13.132) and (13.133) to find for the numerical flux

\[
F(u^n_j, u^n_{j+1}) = F_{uw}(u^n_j, u^n_{j+1}) + \frac{1}{2} b(1 - |c|) \Delta x s^n_{j_u},
\]

(13.134)

where the index $j_u$ is defined by

\[
j_u := \begin{cases} 
  j & \text{if } b > 0, \\
  j + 1 & \text{if } b < 0.
\end{cases}
\]

(13.135)

The numerical flux in (13.134) can thus be interpreted as the upwind flux plus a correction term depending on the slope in the cell upwind of the boundary $x_{j+\frac{1}{2}}$.

Finally, we have to specify the slope $s^n_{j_u}$. Note that the numerical flux (13.134) is identical to the Lax–Wendroff flux (13.111) for the choice $s_{j_u} = (u^n_{j+1} - u^n_j)/\Delta x$, i.e., when the slope $s^n_{j_u}$ in the upwind cell is given by the central difference approximation of $u^n/\Delta x(x_{j+\frac{1}{2}}, t^n)$. As seen in the previous section, the Lax–Wendroff scheme is not TVD, and therefore this choice is not allowed. However, if we take

\[
s_{j_u} = \Phi^n_{j+\frac{1}{2}} \frac{1}{\Delta x} (u^n_{j+1} - u^n_j),
\]

(13.136)

where $\Phi^n_{j+\frac{1}{2}}$ is a limiter on the slope $(u^n_{j+1} - u^n_j)/\Delta x$, the numerical flux (13.134) is precisely the high-resolution flux (13.112). $\Phi^n_{j+\frac{1}{2}}$ is now referred to as the slope limiter and is defined by (13.104) and (13.105). If we choose a limiter function $\Phi(r)$ in the TVD region shown in Figure 13.18, the resulting scheme (13.128) is TVD. The numerical scheme (13.128) with numerical flux defined by (13.134) to (13.136) is also referred to as the slope limiter scheme.

Example 13.23 We repeat the computations of Example 13.5, now using the slope limiter method (13.128) with numerical flux given by (13.134) to (13.136). We have used the van
Leer limiter function in (13.122a). The numerical parameters are once more $\Delta x = 10^{-2}$, $\Delta t = 0.8 \times 10^{-2}$, and $b = 1$. The results are shown in Figure 13.20. From this figure we can conclude that the slope limiter method gives a sharp and nonoscillatory approximation of the discontinuities in the solution. Apparently, this method is much better suited than either the upwind or Lax–Wendroff scheme for the numerical simulation of problems involving discontinuities.

![Figure 13.20. Numerical solution computed with the slope limiter method, after 25, 50, and 75 time steps, respectively.](image)

13.8.2 A Slope Limiter Method for Nonlinear Conservation Laws

For nonlinear conservation laws one can also use a slope limiter method, which we outline below. As with the advection equation, the slope limiter method is based on the Godunov scheme in combination with a piecewise linear numerical solution. In particular, we need an expression for the numerical flux $F(u^n_j, u^n_{j+1})$. For the sake of simplicity we restrict ourselves to the case $b(u) := f'(u) > 0$. Other situations can be dealt with in a similar way; cf. [53, 156].

The generic form of a slope limiter method for the conservation law (13.80a) is given by (13.102). In order to compute the numerical flux $F(u^n_j, u^n_{j+1})$, we have to solve the generalized Riemann problem consisting of the conservation law (13.80a) and the piecewise linear initial condition (13.123b). In general this is a difficult problem. Since we are approximating the solution anyway it makes sense to overcome this problem by replacing...
13.8. Slope Limiter Methods

the flux \( f(u) \) by its linear interpolant \( \tilde{f}(u) \), i.e.,

\[
\tilde{f}(u) := f(u^n_j) + \tilde{b}^{n}_{j+\frac{1}{2}} (u - u^n_j), \quad u \in \text{int}(u^n_j, u^{n+1}_j). \tag{13.137}
\]

Here \( \tilde{b}^{n}_{j+\frac{1}{2}} \) is the finite difference approximation of \( b(u(x_{j+\frac{1}{2}}, t^n)) \) given by

\[
\tilde{b}^{n}_{j+\frac{1}{2}} := \frac{f(u_{j+1}^n) - f(u^n_j)}{u_{j+1}^n - u^n_j} \quad \text{provided that } u^n_j \neq u^{n+1}_j. \tag{13.138}
\]

Since the interpolated flux \( \tilde{f}(u) \) is piecewise linear, we obtain linear Riemann problems similar to (13.123), from which we can compute the numerical flux. Hence the generalised Riemann problem reduces to the linear initial value problem

\[
\frac{\partial \tilde{u}}{\partial t} + \tilde{b}^{n}_{j+\frac{1}{2}} \frac{\partial \tilde{u}}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > t^n; \tag{13.139a}
\]

\[
\tilde{u}(x, t^n) = u^n_j + s^n_j (x - x_j), \quad x \in V_j \quad (j = 0, \pm 1, \pm 2, \ldots), \tag{13.139b}
\]

where \( \tilde{b}^{n}_{j+\frac{1}{2}} \) is defined in (13.138) and where \( u^n_j \) and \( s^n_j \) are the average value and slope, respectively, of \( \tilde{u}(x, t^n) \) in the control volume \( V_j \); cf. (13.123). Like in (13.85) the numerical flux \( F(u^n_j, u^{n+1}_j) \) is the time average of \( \tilde{f}(\tilde{u}(x_{j+\frac{1}{2}}, t)) \) over \([t^n, t^{n+1}])\); i.e.,

\[
F(u^n_j, u^{n+1}_j) := \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \tilde{f}(\tilde{u}(x_{j+\frac{1}{2}}, t)) \, dt. \tag{13.140}
\]

The computation of \( \tilde{u}(x_{j+\frac{1}{2}}, t) \) from (13.139) is straightforward, and we obtain

\[
\tilde{u}(x_{j+\frac{1}{2}}, t) = u^n_j + s^n_j \left( \frac{1}{2} \Delta x - \tilde{b}^{n}_{j+\frac{1}{2}} (t - t^n) \right), \tag{13.141}
\]

and, subsequently, substituting this expression into (13.140), we find for the numerical flux

\[
F(u^n_j, u^{n+1}_j) = f(u^n_j) + \frac{1}{2} \tilde{b}^{n}_{j+\frac{1}{2}} \left( 1 - \frac{\Delta t}{\Delta x} \tilde{b}^{n}_{j+\frac{1}{2}} \right) \Delta x s^n_j. \tag{13.142}
\]

Note that for \( f(u) = bu \) \((b > 0)\) we recover (13.132).

As a final step we have to determine the slopes \( s^n_j \) such that the resulting scheme is TVD. A possible choice for \( s^n_j \) is the minmod slope \([88]\)

\[
s^n_j = \frac{1}{\Delta x} \minmod(u^n_{j+1} - u^n_j, u^n_j - u^n_{j-1}), \tag{13.143}
\]

where the minmod function is defined as

\[
\minmod(y, z) := \begin{cases} 
  y & \text{if } |y| < |z| \text{ and } yz > 0, \\
  z & \text{if } |z| < |y| \text{ and } yz > 0, \\
  0 & \text{if } yz \leq 0;
\end{cases} \tag{13.144}
\]
see also Figure 13.18. We can rewrite the minmod slope (13.143) in the standard form (13.136) if we set \( \Phi(r) = \max(0, \min(1, r)) \). The minmod limiter thus chooses from the upwind slope \((u^n_j - u^n_{j-1})/\Delta x\) and the downwind slope \((u^n_{j+1} - u^n_j)/\Delta x\) the one that is smallest in absolute value if at least both have the same sign. If the two slopes have opposite signs, then \( s^n_j \) is set to zero.

**Example 13.24** We apply the slope limiter scheme to compute the rarefaction wave of the Burgers’ equation with \( u_L = 1 \) and \( u_R = 2 \); cf. (12.43). We choose \( \Delta x = 2.5 \times 10^{-2} \) and \( \Delta t = 10^{-2} \), resulting in a stable scheme. The result after 40 time steps is shown in Figure 13.21. The slope limiter scheme gives an accurate numerical approximation, with the largest errors occurring near the head and tail of the rarefaction wave.

![Figure 13.21. Rarefaction wave computed with the slope limiter method.](image)

### 13.9 Numerical Boundary Conditions

For most schemes one needs to provide additional boundary conditions in order to carry out the computations. In general this may be a complicated and nontrivial matter. We will not discuss the issue in full detail, so we restrict ourselves to the advection equation (13.1a) defined on the domain \([0, 1] \times [0, \infty)\). Furthermore, we assume that \( b > 0 \), and consequently we can prescribe the solution at the left boundary \( x = 0 \). However, some numerical boundary condition is required at the right boundary \( x = 1 \). We will elaborate this for two schemes. In particular, we will investigate the effect of the numerical boundary condition on stability.

We must prescribe numerical boundary conditions very carefully, as is demonstrated in the next example.

**Example 13.25** We apply the leapfrog scheme to compute the numerical solution of the advection equation subject to the initial and boundary conditions \( u(x, 0) = \sin(8\pi x) \) and \( u(0, t) = -\sin(8\pi bt) \), respectively. Furthermore, we apply the erroneous boundary condition \( u(1, t) = 0 \). We choose the (numerical) parameters \( b = 1 \), \( t_{\text{max}} = 0.5 \), \( \Delta x = 1.25 \times 10^{-2} \), and \( \Delta t = 6.25 \times 10^{-3} \), and consequently \( c = 0.5 \). Hence the scheme is stable. The result is shown in Figure 13.22. From this figure we see that the error propagates undamped to the left over a distance \( bt_{\text{max}} = 0.5 \) and thereby completely corrupts the solution. The left part of the solution only suffers from slight dispersion. \( \square \)
13.9. Numerical Boundary Conditions

First we would like to apply the Lax–Wendroff scheme (13.23) to the advection equation. Assume that the domain is covered with the grid defined in (13.61). The numerical value \( u_{M+1}^{n+1} \) at the right boundary has to be approximated somehow. For this we cannot use the Lax–Wendroff scheme, since it has a three-point coupling in the \( x \) direction, resulting in a virtual grid point outside the domain. Note, however, that the upwind scheme can be applied without difficulty. Thus we have for all interior grid points \( x_j \) \( (j = 1, 2, \ldots, M - 1) \) the scheme

\[
 u_j^{n+1} = u_j^n - \frac{1}{2} c (u_{j+1}^n - u_{j-1}^n) + \frac{1}{2} c^2 (u_{j+1}^n - 2u_j^n + u_{j-1}^n),
\]

whereas at the right boundary we have the numerical boundary condition

\[
 u_M^{n+1} = u_M^n - c (u_M^n - u_{M-1}^n).
\]

Clearly (13.145) gives a method of computing the numerical solution at all grid points \( x_j \) \( (j = 1, 2, \ldots, M) \). We will now prove that this method is stable as long as the CFL condition \( (c \leq 1) \) is satisfied. Since boundary conditions are involved, we have to employ the matrix method for stability; cf. Chapter 11. Introducing the numerical solution vector at time \( t^n \), i.e., \( u^n := (u_1^n, u_2^n, \ldots, u_M^n)^T \), we can write (13.145) in the form

\[
 u^{n+1} = (I + A) u^n + \Delta t f^n,
\]

where the matrix \( A \) is defined as

\[
 A := \begin{pmatrix}
 -c^2 & \frac{1}{2} c (c - 1) \\
 \frac{1}{2} c (c + 1) & -c^2 & \frac{1}{2} c (c - 1) \\
 & \ddots & \ddots & \ddots \\
 & \frac{1}{2} c (c + 1) & -c^2 & \frac{1}{2} c (c - 1) & c \\
 & & \frac{1}{2} c (c + 1) & -c^2 & \frac{1}{2} c (c - 1) & \cdots & \cdots & \cdots \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
 & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \\
 & & & & & & \ddots & \ddots & \ddots & \ddots \\
 & & & & & & & \ddots & \ddots & \ddots \\
 & & & & & & & & \ddots & \ddots \\
 & & & & & & & & & \ddots \\
 & & & & & & & & & & \ddots \\
\end{pmatrix}.
\]

Figure 13.22. Numerical solution of the leapfrog scheme, corrupted by the erroneous boundary condition \( u(1, t) = 0 \).
The right-hand side vector \( \mathbf{f}^n \) contains the boundary data at \( x = 0 \). We will investigate the stability of (13.146) by estimating the spectral radius \( \rho(\mathbf{I} + A) \).

**Property 13.26.** If \( 0 < c \leq 1 \), then scheme (13.145), i.e., the Lax–Wendroff scheme with upwind numerical boundary condition, is stable.

**Proof.** Let us consider the case \( c = 1 \) first. We then find that the diagonal and the upper codiagonal of the matrix \( \mathbf{I} + A \) contain only zeros. The lower codiagonal contains only ones. Clearly, the norm of this matrix (a 1-, 2-, or \( \infty \)-norm) is bounded by one. So let us now assume \( 0 < c < 1 \). We first transform the matrix \( A \) in such a way that it becomes skew symmetric, i.e., the codiagonals have opposite signs. In particular, we introduce the diagonal transformation matrix \( \mathbf{D} \), defined by

\[
\mathbf{D} := \text{diag}(d_1, d_2, \ldots, d_M),
\]

\[
d_l := \frac{1 + c}{1 - c} \left( \frac{l - 1}{2} \right)^{(l-1)/2}, \quad (l = 1, 2, \ldots, M - 1), \quad d_M := \sqrt{\frac{1}{1 - c}} d_{M-1}.
\]

This indeed transforms the matrix \( A \) into the required form:

\[
\mathbf{\tilde{A}} := \mathbf{D}^{-1} \mathbf{AD} = \begin{pmatrix}
-\gamma & \cdots & \gamma \\
\gamma & -\gamma & \cdots \\
\vdots & \ddots & \ddots \\
\gamma & \cdots & -\gamma & \gamma \\
\delta & \cdots & \delta & -\delta
\end{pmatrix},
\]

where \( \gamma := \frac{1}{2} c \sqrt{1 - c^2} \) and \( \delta := \frac{c \sqrt{(1 - c)/2}}{|c|} \). We proceed as follows: First, we introduce a complex inner product, as we have to deal with complex eigenvalues. For vectors with elements consisting of complex numbers the inner product reads

\[
(x, y) := \bar{y}^T x,
\]

where the overbar denotes the complex conjugate. For any eigenvalue \( \lambda \) of \( \mathbf{I} + \mathbf{A} \) and corresponding eigenvector \( x \) we have

\[
((\mathbf{I} + \mathbf{A}) x, x) = \lambda (x, x).
\]

Since \( \mathbf{A} \) and \( \mathbf{\tilde{A}} \) have the same set of eigenvectors, we find for the inner product on the left-hand side

\[
((\mathbf{I} + \mathbf{A}) x, x) = ((\mathbf{I} + \mathbf{\tilde{A}}) x, x) = \gamma \sum_{k=1}^{M-2} (x_k \bar{x}_{k+1} - x_{k+1} \bar{x}_k) + \delta (x_{M-1} \bar{x}_M - x_M \bar{x}_{M-1}) + (1 - c^2) \sum_{k=1}^{M-1} x_k \bar{x}_k + (1 - c) x_M \bar{x}_M.
\]

On the other hand, we find, splitting \( \lambda \) into its real and imaginary parts, i.e., \( \lambda =: \alpha + i \beta \), that \( \lambda (x, x) = (\alpha + i \beta) \sum_{k=1}^{M-1} x_k \bar{x}_k \). We can take \( (x, x) = 1 \) without restriction. Writing \( \rho := |x_M|^2 = x_M \bar{x}_M \), we thus find, from comparing the real parts, that

\[
\alpha = (1 - c^2)(1 - \rho) + (1 - c) \rho = 1 - c^2 + \rho (c^2 - c),
\]

where \( \rho \) is the spectral radius of \( \mathbf{I} + \mathbf{A} \). Therefore, if \( 0 < c \leq 1 \), then we have

\[
\lambda = (1 - c^2) + \rho (c^2 - c) < 1
\]

and hence stability.
and, from the imaginary parts, we find

$$\beta = 2\gamma \text{Im} \left( \sum_{k=1}^{M-2} x_k \bar{x}_{k+1} \right) + 2\delta \text{Im}(x_{M-1}\bar{x}_M).$$

Note that $|\text{Im}(\sum_{k=1}^{M-2} x_k \bar{x}_{k+1})| \leq \sum_{k=1}^{M-2} |x_k \bar{x}_{k+1}|$. Also note that $2\gamma = c\sqrt{1-c^2} \leq c\sqrt{2-2c} = 2\delta$. We thus obtain

$$|\beta| \leq c\sqrt{2-2c} \sum_{k=1}^{M-1} |x_k \bar{x}_{k+1}|.$$ 

We can estimate the latter by applying Cauchy’s inequality (cf. (G.3)), giving

$$|\beta| \leq c\sqrt{2-2c} \sum_{k=1}^{M-1} |x_k \bar{x}_{k+1}| \leq c\sqrt{2-2c} \sum_{k=1}^{M-1} |x_k \bar{x}_{k+1}| \leq c\sqrt{2-2c}.$$ 

Here we have used that $\sum_{k=1}^{M-1} |x_k \bar{x}_{k+1}|$ and $\sum_{k=1}^{M-1} |x_k \bar{x}_{k+1}|$ are not greater than 1. Hence we find

$$\alpha^2 + \beta^2 < (1-c^2)^2 + 2c^2 - 2c^3 = 1 - 2c^3 + c^4 \leq 1.$$

This indeed proves that $|\lambda| = \sqrt{\alpha^2 + \beta^2} \leq 1$. \hfill \Box

Another scheme that needs a numerical boundary condition at $x = 1$ is the leapfrog scheme. Like for the Lax–Wendroff scheme, we apply the upwind scheme at grid point $x_M$, leading to the vector recursion

$$u^{n+1} = Au^n + Du^{n-1} + \Delta t f^n,$$ \hfill (13.147a)

where

$$A := \begin{pmatrix} 0 & -c \\ c & 0 & -c \\ \vdots & \ddots & \ddots \\ c & 0 & -c \\ c & 1 & -c \end{pmatrix}, \quad D := \text{diag}(1, 1, \ldots, 1, 0). \hfill (13.147b)$$

**Property 13.27.** If $0 < c \leq 1$, then scheme (13.147), i.e., the leapfrog scheme with upwind numerical boundary condition, is stable.

**Proof.** We leave it as an exercise to show that the case $c = 1$ is correct and hence can assume that $0 < c < 1$. Let us rewrite (13.147a) as a first order recursion whose homogeneous part reads

$$v^{n+1} = \hat{A} v^n,$$ 

where

$$\hat{A} := \begin{pmatrix} A & D \\ I & 0 \end{pmatrix}, \quad v^n := \begin{pmatrix} u^n \\ u^{n-1} \end{pmatrix}.$$
For stability we investigate the spectral radius $\rho(\hat{A})$. Note that zero is always an eigenvalue of $\hat{A}$. In the following we only consider eigenvalues $\lambda \neq 0$. Let $x$ be the corresponding eigenvector, which we partition into two parts $y$ and $z$, analogous to the partitioning of $\mathbf{v}$, i.e., $\mathbf{x}^T = (y^T, z^T)$. Writing out the relation $\hat{A}\mathbf{x} = \lambda \mathbf{x}$ gives

$$Ay + Dz = \lambda y, \quad y = \lambda z.$$  

This then implies the relation

$$(Az, z) = \lambda (z, z) - \frac{1}{\lambda} (Dz, z). \quad (*)$$

Writing $\lambda =: \alpha + i\beta$, we note that $1/\lambda = (\alpha - i\beta)/(\alpha^2 + \beta^2)$. Collecting the real parts of (*) gives

$$(1 - c)z_M\hat{z}_M = \alpha \sum_{k=1}^{M} z_k\hat{z}_k - \frac{\alpha}{\alpha^2 + \beta^2} \sum_{k=1}^{M-1} z_k\hat{z}_k,$$

while for the imaginary parts we have

$$2\alpha \text{Im} \left( \sum_{k=1}^{M-1} z_k\hat{z}_{k+1} \right) = \beta \sum_{k=1}^{M} z_k\hat{z}_k + \frac{\beta}{\alpha^2 + \beta^2} \sum_{k=1}^{M-1} z_k\hat{z}_k.$$  

As before it is not restrictive to take $\sum_{k=1}^{M} z_k\hat{z}_k = 1$. Defining $\rho := z_M\hat{z}_M$, we thus have the relations

$$(1 - c)\rho = \alpha \left( 1 - \frac{1 - \rho}{\alpha^2 + \beta^2} \right), \quad 2\alpha \sqrt{1 - \rho} \geq \left| \beta \left( 1 + \frac{1 - \rho}{\alpha^2 + \beta^2} \right) \right|, \quad (**)$$

where we have used the Cauchy–Schwarz inequality. Now suppose that we have no stability; i.e., $\alpha^2 + \beta^2 > 1$. Then we find $|\alpha| < (1 - c)\rho/(1 - (1 - \rho)) = 1 - c$. As a consequence, we must have that $\beta^2 > 1 - (1 - c)^2 = 2c - c^2$. This would then give a lower bound for $|\beta|$. Comparing this lower bound with the upper bound in (**) results in

$$\sqrt{2c - c^2} < \frac{2c|\lambda|^2}{|\lambda|^2 + 1 - \rho}.$$  

After some rearrangement, this leads to

$$\left( 2\alpha \sqrt{\frac{1 - \rho}{2c - c^2} - 1} \right) |\lambda|^2 > 1 - \rho.$$

Note that the factor in brackets must be positive in order to let this inequality make sense. Moreover, due to our assumption, we have $|\lambda|^2 = \alpha^2 + \beta^2 > 1$, hence the inequality is still correct if we require it to hold without the factor $|\lambda|^2$, i.e., $2\alpha \sqrt{(1 - \rho)/(2c - c^2)} > 2 - \rho$. After squaring and some simple rewriting, we thus find that we should have

$$g(\rho) := (2 - c)\rho^2 + 8(c - 1)\rho + 8(1 - c) < 0.$$  

It is easy to see that $g(\rho)$ is positive definite. So this inequality cannot be fulfilled for any choice of $\rho$. We conclude that assuming $|\lambda| > 1$ leads to a false statement. Thus we have shown that $|\lambda| \leq 1$. \qed
Example 13.28 For the leapfrog scheme with upwind numerical boundary condition we have drawn in Figure 13.23 the (complex) eigenvalues of the matrix $A$ for $\Delta x = 5 \times 10^{-2}$ and $c = 0.99$. Clearly, all eigenvalues are smaller than one in modulus, and some are even significantly smaller than one. This then leads to the conclusion that the nondissipation property of the leapfrog scheme is annihilated.

![Figure 13.23. Eigenvalues of the leapfrog scheme with upwind numerical boundary condition.](image)

13.10 Discussion

- Hyperbolic problems very rarely appear in practice without some small diffusive (or viscous) term. In this sense they are an idealisation. However, as can be seen from the analysis in this chapter, methods for advection-dominated problems are best studied by analysing methods for advection problems proper. If higher-order terms are present, one may be forced to resort to implicit methods for the latter. As outlined in Chapter 11, the advection term can then often be discretised by an explicit method using operator splitting.

- Most of the methods discussed in this chapter have higher-dimensional analogues. However, as was already indicated in Chapter 12, hyperbolic equations in several spatial dimensions may have solutions with quite complicated behaviour. This has its impact on the numerical treatment of such problems, for instance shock capturing. These complications are reflected in the literature, which is much less extensive for the latter class of problems. For some further reading see, e.g., [42, 88, 158].

- A wealth of methods and applications can be found in fluid dynamics. Indeed, one of the most advanced areas of numerical simulation of problems arising in practice is computational fluid dynamics (CFD). Here all sorts of problems in areas such as aerospace or management of waterways occur. CFD is often also the main ingredient...
in complex physical problems like combustion (flow of gas mixtures and chemistry) and plasma physics (flow of gas and electromagnetism). Some books on CFD are, e.g., [63, 64, 45, 168].

- In Chapter 16 some typical problems modeled by the advection-diffusion equation, with dominant advection, are the following: the determination of chemical species by separation through electrophoresis in Section 16.13 and the cooling device employing a pulse tube in Section 16.14.

### Exercises

13.1. Show that the upwind numerical solution is the exact solution of (13.1a) when \( c = 1 \).

13.2. A well-known scheme for the advection equation (13.1a) is the Beam–Warming scheme given by

\[
u^{n+1}_j = u^n_j - \frac{1}{2} c (3u^n_j - 4u^n_{j-1} + u^n_{j-2}) + \frac{1}{2} c^2 (u^n_j - 2u^n_{j-1} + u^n_{j-2}).\]

(a) Show that this scheme can be found in a similar way as the Lax–Wendroff scheme, but using backward difference approximations for the spatial derivatives in (13.22). Draw the stencil.

(b) Show that the scheme is second order.

(c) Find a sufficient condition for stability.

13.3. Compute the amplitude error and the relative phase error of the Beam–Warming scheme in Exercise 13.2. Investigate the dissipation and dispersion of the scheme.

13.4. Sometimes it is useful to use higher-order differences for the space discretisation. To this end consider the scheme applied to the advection equation (13.1a):

\[
u^{n+1}_j = u^n_j - \frac{1}{12} c (-u^n_{j+2} + 8u^n_{j+1} - 8u^n_{j-1} + u^n_{j-2}).\]

(a) Give the stencil of this scheme.

(b) Show that this method is third order in space and first order in time.

(c) Investigate the stability of this method.

13.5. Consider the following scheme for the advection equation (13.1a):

\[
u^{n+1}_j = \frac{1}{2} (u^n_{j+1} + u^n_{j-1}) - \frac{1}{2} c (u^n_{j+1} - u^n_{j-1}).\]

(a) Draw the stencil of this Lax–Friedrichs scheme.

(b) Show that this scheme is stable if \(|c| \leq 1\).

(c) Show that the local discretisation error of the method is \(O(\Delta t + \frac{\Delta x^2}{\Delta t} + \Delta x^2)\).

(d) Comment on the result in part (c).
Exercises 415

13.6. Consider the advection equation (13.1a).
(a) Show what scheme you get if you use the standard central difference approximation to discretise the spatial derivative and the implicit Euler method for time integration.
(b) Is this scheme stable?

13.7. As is well known, backward difference formula (BDF) methods are dissipative; see, e.g., Section 5.4.2 and also [97]. In fact, the Euler backward method is the lowest-order member of this family (BDF-1). If we use BDF-2 for time integration and the standard central difference approximation to discretise the spatial derivative, we obtain the following scheme for (13.1a):

$$3u_j^{n+1} = 4u_j^n - u_j^{n-1} - \frac{1}{2}c(u_{j+1}^{n+1} - u_{j-1}^{n+1}).$$

(a) Give the stencil.
(b) Show that the scheme is second order in $\Delta x$ and $\Delta t$.
(c) Show that the scheme is unconditionally stable.

13.8. Compute the local discretisation error of scheme (13.44) and show that it is second order in $\Delta x$ and first order in $\Delta t$.

13.9. Prove that scheme (13.44) for the advection-diffusion equation is TVD provided that $d \leq \frac{1}{4}$ and $|Pe| \leq 2$.

13.10. Instead of (13.44), we could use the upwind approximation for the advective term in (13.43).
(a) Give the resulting scheme.
(b) Compute the local discretisation error and show that it is first order in $\Delta x$ and $\Delta t$.
(c) Investigate the stability of the scheme.
(d) Prove that the scheme is TVD.

13.11. Compute the local discretisation error of the box scheme (13.63) and show that it is second order in $\Delta x$ and $\Delta t$.

13.12. Compute the local discretisation error of the leapfrog scheme (13.70b) and show that it is second order in $\Delta x$ and $\Delta t$.

13.13. Compute the numerical group velocity of the leapfrog scheme.


13.15. Compute the local discretisation error of the Godunov scheme applied to the Burgers’ equation and show that it is first order in $\Delta x$ and $\Delta t$.

13.16. Consider the nonlinear conservation law (13.80a) with convex flux function $f(u)$. Derive a general expression for the Godunov numerical flux in terms of $f'(u)$.

13.17. Apply Property 13.20 to derive a general expression for the local discretisation error of the Godunov scheme.

13.18. Give the slope limiter method for the Burgers’ equation when $u > 0$. 

13.19. Consider the advection equation (13.1a) with $b < 0$ defined on the domain $[0, 1] \times [0, \infty)$. Show that the Lax–Wendroff scheme with upwind numerical boundary condition at $x = 0$ is stable if $-1 \leq c < 0$.

Chapter 14

Numerical Methods for Hyperbolic Systems

In this chapter we extend the numerical methods in the previous chapter to systems. We start in Section 14.1 with a description of the upwind scheme for linear systems. In Section 14.2 we present the Godunov scheme. For linear systems the method reduces to the upwind scheme, like in the scalar case. For nonlinear systems the computational costs of the Godunov scheme are high since it requires the solution of local Riemann problems. Approximate Riemann solvers are often a cheaper alternative. The Roe scheme is a well-known example of these and is presented in Section 14.3. Next, in Section 14.4, we construct high-resolution schemes based on slope limiters. As an example, we present in Section 14.5 a detailed elaboration of both the Godunov and Roe schemes applied to the shallow-water equations. The wave equation can be reformulated as a $2 \times 2$ hyperbolic system and for that reason it is covered in this chapter. In Section 14.6 we present a few numerical methods for this equation, based on either the scalar form or the equivalent system formulation. Finally, in Section 14.7, we address the important topic of boundary conditions.

14.1 The Upwind Scheme

The simplest scheme we encountered in Chapter 13 was the upwind scheme, introduced in Section 13.1.1. To generalise this to linear hyperbolic systems consider the following initial value problem for $u(x, t) \in \mathbb{R}^m$:

\[
\frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{14.1a}
\]

\[
u(x, 0) = v(x), \quad x \in \mathbb{R}, \tag{14.1b}
\]

with $B \in \mathbb{R}^{m \times m}$ a constant matrix. The linear system (14.1a) is a model problem for advective processes; e.g., the linearised equations of gas dynamics can be written in this form [158]. Moreover, the one-dimensional wave equation can be reformulated as in (14.1a); see Example 12.20. We assume that the linear system (14.1a) is hyperbolic. Then, according to Definition 12.19, the matrix $B$ has $m$ real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ and $m$ linearly independent left eigenvectors $t_1, t_2, \ldots, t_m$ that make up the matrix $T$; see (12.51). Alter-
natively, $B$ has a complete set of (right) eigenvectors $s_1, s_2, \ldots, s_m$ such that

$$B = SAS^{-1},$$

(14.2)

where the matrices $A$ and $S$ are defined as

$$A := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m), \quad S := (s_1, s_2, \ldots, s_m) = T^{-1}.$$

(14.3)

In Section 12.3 we showed that the characteristic variable $\tilde{u}$, defined by

$$\tilde{u} := Tu = S^{-1}u,$$

(14.4)

satisfies the decoupled system

$$\frac{\partial \tilde{u}}{\partial t} + \Lambda \frac{\partial \tilde{u}}{\partial x} = 0.$$

(14.5)

Written componentwise, this is

$$\frac{\partial \tilde{u}_k}{\partial t} + \lambda_k \frac{\partial \tilde{u}_k}{\partial x} = 0 \quad (k = 1, 2, \ldots, m).$$

(14.6a)

From (14.1b) we can derive the following initial condition for $\tilde{u}$:

$$\tilde{u}(x, 0) = S^{-1}v(x) =: \left(\tilde{v}_1(x), \tilde{v}_2(x), \ldots, \tilde{v}_m(x)\right)^T.$$

(14.6b)

Each of the equations in (14.6a) is an advection equation, and we can verify that the solutions of these equations, together with the initial conditions in (14.6b), are given by

$$\tilde{u}_k(x, t) = \tilde{v}_k(x - \lambda_k t).$$

(14.7)

The solution of (14.1) is then given by

$$u(x, t) = S\tilde{u}(x, t) = \sum_{k=1}^m \tilde{v}_k(x - \lambda_k t)s_k.$$

(14.8)

Thus the solution $u(x, t)$ consists of $m$ components $\tilde{v}_k(x - \lambda_k t)s_k$, each propagating with advection velocity $\lambda_k$ along the $x$ axis.

In order to compute a numerical solution of (14.1), we cover the domain $\Omega := \mathbb{R} \times [0, \infty)$ with grid points $(x_j, t^n)$ defined by

$$x_j := j \Delta x \quad (j = 0, \pm 1, \pm 2, \ldots), \quad t^n := n \Delta t \quad (n = 0, 1, 2, \ldots),$$

(14.9)

with $\Delta x$ the spatial grid size and $\Delta t$ the time step. We first introduce the upwind scheme for the linear advection equations in (14.6a) and subsequently give the scheme for $u$. By virtue of (14.8), each component $\tilde{v}_k(x - \lambda_k t)s_k$ of the solution is constant along characteristics $C_k : x - \lambda_k t = C$ ($C \in \mathbb{R}$) and therefore has an upwind direction determined by the sign of the corresponding eigenvalue $\lambda_k$. Introducing the splitting of the eigenvalues $\lambda_k$ as

$$\lambda_k = \lambda^+_k + \lambda^-_k, \quad \lambda^+_k := \max(\lambda_k, 0), \quad \lambda^-_k := \min(\lambda_k, 0),$$

(14.10)
14.1. The Upwind Scheme

We can write the upwind schemes for the advection equations in (14.6a) as

\[
\frac{1}{\Delta t} (\tilde{u}_{k,j}^{n+1} - \tilde{u}_{k,j}^n) + \frac{\lambda_k^+}{\Delta x} (\tilde{u}_{k,j}^n - \tilde{u}_{k,j-1}^n) + \frac{\lambda_k^-}{\Delta x} (\tilde{u}_{k,j+1}^n - \tilde{u}_{k,j}^n) = 0
\]

\( (k = 1, 2, \ldots, m) \), \hspace{1cm} (14.11)

where \( \tilde{u}_{k,j}^n \) denotes the numerical approximation of \( \tilde{u}_k(x_j, t^n) \), etc.; see Section 13.1. The upwind schemes for all characteristic variables \( \tilde{u}_k \) can be combined in vector form as

\[
\frac{1}{\Delta t} (\tilde{u}_j^{n+1} - \tilde{u}_j^n) + \frac{1}{\Delta x} A^+ (\tilde{u}_j^n - \tilde{u}_{j-1}^n) + \frac{1}{\Delta x} A^- (\tilde{u}_{j+1}^n - \tilde{u}_j^n) = 0,
\]

\( \hspace{1cm} (14.12) \)

with

\[
A^+ := \text{diag}(\lambda_1^+, \lambda_2^+, \ldots, \lambda_m^+), \hspace{0.5cm} A^- := \text{diag}(\lambda_1^-, \lambda_2^-, \ldots, \lambda_m^-).
\]

(14.13)

We obtain the upwind scheme in terms of the original variable \( u \) by substituting (14.4) in (14.12) and multiplying the resulting scheme by the matrix \( S \), giving

\[
\frac{1}{\Delta t} (u_j^{n+1} - u_j^n) + \frac{1}{\Delta x} B^+ (u_j^n - u_{j-1}^n) + \frac{1}{\Delta x} B^- (u_{j+1}^n - u_j^n) = 0,
\]

\( \hspace{1cm} (14.14) \)

where \( u_j^n \) is obviously the numerical approximation of \( u(x_j, t^n) \) and the matrices \( B^+ \) and \( B^- \) are defined by

\[
B^+ := SA^+ S^{-1}, \hspace{0.5cm} B^- := SA^- S^{-1}.
\]

(14.15)

We see that the coefficient matrix \( B \) is split as \( B = B^+ + B^- \), with \( B^+ \) having the non-negative eigenvalues \( \lambda_k^+ \) and \( B^- \) the nonpositive eigenvalues \( \lambda_k^- \). Thus the \( B^+ \) term in (14.14) is the backward difference approximation of \( \frac{\partial u}{\partial x} \) and corresponds to components propagating in the positive \( x \) direction. Likewise, the \( B^- \) term in (14.14) is the forward difference approximation of \( \frac{\partial u}{\partial x} \) and corresponds to components propagating in the negative \( x \) direction.

Stability of the upwind scheme (14.14) is ensured if the corresponding scalar upwind schemes for all variables \( \tilde{u}_k \) are stable. This latter requirement gives the Courant, Friedrichs, and Lewy (CFL) stability condition

\[
\max_k |\lambda_k| \frac{\Delta t}{\Delta x} \leq 1;
\]

(14.16)

cf. (13.20). This requirement means that all characteristics through the point \((x_j, t^n+1)\) intersect the grid line \( t = t^n \) at points in the interval \([x_{j-1}, x_{j+1}]\).

The upwind scheme (14.12) is first order accurate in \( \Delta t \) and \( \Delta x \). Consequently, this also holds for scheme (14.14). We leave it as an exercise to show this.

**Example 14.1** Consider the hyperbolic system in Example 12.20, which describes small-amplitude waves in a canal of uniform depth \( h \):

\[
\frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = 0,
\]

\( (\ast) \)
where $u$ and $B$ are given by

$$u = \begin{pmatrix} \frac{\partial}{\partial t} \eta \\ \frac{\partial}{\partial x} \eta \end{pmatrix}, \quad B = -\begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}, \quad a = \sqrt{gh},$$

with $\eta(x, t)$ the water elevation. The eigenvalues $\lambda_1$, $\lambda_2$ and corresponding eigenvectors $s_1$, $s_2$ are given by

$$\lambda_1 = -a < 0, \quad \lambda_2 = a > 0, \quad s_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

The splitting of the eigenvalues according to (14.10) is particularly simple, and we find

$$\lambda^-_1 = -a, \quad \lambda^+_1 = 0, \quad \lambda^-_2 = 0, \quad \lambda^+_2 = a.$$

From (14.13) and (14.15) we can compute the matrices $B^+$ and $B^-$ to obtain

$$B^+ = \frac{1}{2} \begin{pmatrix} a & -a \\ -a & a \end{pmatrix}, \quad B^- = -\frac{1}{2} \begin{pmatrix} a & a \\ a & a \end{pmatrix}.$$

We can verify that the matrix $B^+$ has the eigenvalues zero and $a$, whereas $B^-$ has the eigenvalues $-a$ and zero. Substituting these matrices in (14.14), we obtain the upwind scheme for system $(\ast)$:

$$u^{n+1}_j = u^n_j - \frac{\Delta t}{2\Delta x} \left( B^+ (u^n_j - u^n_{j-1}) + B^- (u^n_{j+1} - u^n_j) \right),$$

with $c := a \Delta t / \Delta x$ the Courant number. Finally, the water elevation $\eta$ has to be computed numerically from either $\frac{\partial}{\partial x} \eta = u_2$ or $\frac{\partial}{\partial t} \eta = au_1$.

As we have seen in Section 13.3.2, the scalar upwind scheme can be interpreted as the central difference scheme stabilized by adding some numerical diffusion. The upwind scheme in the example above also contains terms reminiscent of numerical diffusion. Indeed, we can rewrite the upwind scheme (14.14) as the central difference scheme with some numerical diffusion added, as follows:

$$u^{n+1}_j = u^n_j - \frac{\Delta t}{\Delta x} \left( B^+ (u^n_j - u^n_{j-1}) + B^- (u^n_{j+1} - u^n_j) \right)$$

$$= u^n_j - \frac{\Delta t}{2\Delta x} B (u^n_{j+1} - u^n_{j-1})$$

$$+ \frac{\Delta t}{2\Delta x} \left( (B - 2B^+) (u^n_j - u^n_{j-1}) + (B - 2B^-) (u^n_{j+1} - u^n_j) \right)$$

$$= u^n_j - \frac{\Delta t}{2\Delta x} B^+ (u^n_{j+1} - u^n_{j-1}) + \frac{\Delta t}{2\Delta x} B^* (u^n_{j+1} - 2u^n_{j} + u^n_{j-1}).$$

(14.17)

where the matrix $B^*$ is defined as

$$B^* := B^+ - B^-.$$  (14.18)
14.2 The Godunov Scheme

14.2.1 Derivation of the Scheme

Consider the following initial value problem for \( u(x,t) \in \mathbb{R}^m \):

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{14.21a}
\]

\[
u(x,0) = v(x), \quad x \in \mathbb{R}, \tag{14.21b}
\]

where the flux \( f(u) \) is generally a nonlinear function of \( u \). Note that \( f(u) = Bu \) for the linear system (14.1a). An example of such a system is the shallow-water equations, introduced in Section 12.5. Let us assume that the system (14.21a) is hyperbolic; i.e., the Jacobi matrix \( B(u) := \frac{\partial f(u)}{\partial u} \) has \( m \) real eigenvalues, assumed to be ordered as \( \lambda_1(u) \leq \lambda_2(u) \leq \cdots \leq \lambda_m(u) \), and \( m \) corresponding linearly independent right eigenvectors \( s_1(u), s_2(u), \ldots, s_m(u) \); see Section 12.3.2.
The Godunov scheme for the hyperbolic system (14.21a) that we derive below is a natural generalisation of the scalar method and therefore our presentation can be brief. Like in Section 13.5, we introduce control volumes or cells $V_j$ as follows:

$$V_j := [x_{j-rac{1}{2}}, x_{j+rac{1}{2}}], \quad x_{j+rac{1}{2}} = \frac{1}{2}(x_j + x_{j+1}) \quad (j = 0, \pm 1, \pm 2, \ldots), \quad (14.22)$$

with $x_j$ defined in (14.9). The numerical solution $u^n_j$ has to be interpreted as an approximation of the average value vector of $u(x, t)$ over the control volume $V_j$ at time level $t^n$.

Associated with $u^n_j$ is defined the vector function $\bar{u}(x, t)$:

$$\frac{\partial \bar{u}}{\partial t} + \frac{\partial f(\bar{u})}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > t^n, \quad (14.23a)$$

$$\bar{u}(x, t^n) = u^n_j, \quad x \in V_j \quad (j = 0, \pm 1, \pm 2, \ldots). \quad (14.23b)$$

The Godunov scheme is based on the integral formulation of the initial value problem (14.23).

Integrating the conservation law (14.23a) over $V_j$, we find

$$\int_{V_j} \bar{u}(x, t) \, dx + f(\bar{u}(x_{j+rac{1}{2}}, t)) - f(\bar{u}(x_{j-rac{1}{2}}, t)) = 0. \quad (14.24)$$

Subsequently integrating (14.24) over the time interval $[t^n, t^{n+1}]$, we obtain

$$\int_{V_j} \bar{u}(x, t^{n+1}) \, dx - \int_{V_j} \bar{u}(x, t^n) \, dx + \int_{t^n}^{t^{n+1}} f(\bar{u}(x_{j+rac{1}{2}}, t)) \, dt - \int_{t^n}^{t^{n+1}} f(\bar{u}(x_{j-rac{1}{2}}, t)) \, dt = 0. \quad (14.25)$$

Introducing the numerical flux $F(u^n_j, u^n_{j+1})$ as the time average flux at the control volume edge $x_{j+rac{1}{2}}$, i.e.,

$$F(u^n_j, u^n_{j+1}) := \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(\bar{u}(x_{j+rac{1}{2}}, t)) \, dt, \quad (14.26)$$

we can write the integral balance (14.25) in the form

$$\int_{V_j} \bar{u}(x, t^{n+1}) \, dx - \int_{V_j} \bar{u}(x, t^n) \, dx = -\Delta t \left( F(u^n_j, u^n_{j+1}) - F(u^n_{j-1}, u^n_{j}) \right). \quad (14.27)$$

In (14.26) we have explicitly written the numerical flux at the cell boundary $x_{j+rac{1}{2}}$ as a function of the numerical vectors $u^n_j$ and $u^n_{j+1}$ in the adjacent cells. The next step is to compute the numerical fluxes.

To compute the numerical flux $F(u^n_j, u^n_{j+1})$ we solve the local Riemann problem at the cell edge $x_{j+rac{1}{2}}$ consisting of the conservation law (14.23a) and the piecewise constant initial condition

$$\bar{u}(x, t^n) = \begin{cases} u^n_j & \text{if } x < x_{j+rac{1}{2}}, \\ u^n_{j+1} & \text{if } x > x_{j+rac{1}{2}}. \end{cases} \quad (14.28)$$
14.2. The Godunov Scheme

The solution of this Riemann problem is a similarity solution of the form
\[ \bar{u}(x,t) = u_R(\eta; u^n_j, u^n_{j+1}), \quad \eta := \frac{x - x_j + \frac{1}{2}}{t - t^n}, \]  
(14.29)
where the function \( u_R \) depends on a single variable \( \eta \) and the parameter vectors \( u^n_j \) and \( u^n_{j+1} \). Since \( \eta = 0 \) for \( x = x_j + \frac{1}{2} \), the numerical flux from (14.26) is given by
\[ F(u^n_j, u^n_{j+1}) = f(u_R(0; u^n_j, u^n_{j+1})). \]  
(14.30)
Thus we need the solution of local Riemann problems. For general nonlinear problems, i.e., for general flux vectors \( f(u) \), this can be quite complicated.

The numerical vector \( u^{n+1}_j \) is defined as the average value vector of \( \bar{u}(x, t^{n+1}) \) over the control volume \( V_j \); i.e.,
\[ u^{n+1}_j := \frac{1}{\Delta x} \int_{V_j} \bar{u}(x, t^{n+1}) \, dx. \]  
(14.31)
If we finally substitute (14.31) in (14.27), we obtain the Godunov scheme, i.e.,
\[ u^{n+1}_j = u^n_j - \frac{\Delta t}{\Delta x} (F(u^n_j, u^{n+1}_j) - F(u^n_{j-1}, u^n_j)), \]  
(14.32)
with the numerical flux \( F(u^n_j, u^{n+1}_j) \) defined in (14.30).

As for stability, we note that in the derivation of the Godunov scheme it is tacitly assumed that \( \bar{u}(x, t) \) at the cell boundaries is not influenced by neighbouring Riemann problems, implying that \( \bar{u}(x_j + \frac{1}{2}, t) = u_R(0; u^n_j, u^n_{j+1}) = C \ (C \in \mathbb{R}) \) for \( t^n \leq t \leq t^{n+1} \). This condition is satisfied if the time step restriction
\[ \frac{\Delta t}{\Delta x} \max_{\eta} S_\eta \leq 1 \]  
(14.33)
holds, with \( \max_{\eta} S_\eta \) the maximum wave speed (in absolute value) occurring in the Riemann problems involved at time level \( t^n \); see [158] for more details.

14.2.2 The Linear Case

If the flux is linear, we can derive an explicit form of the Godunov flux and hence for the method as such. Consider the hyperbolic system (14.1a) with flux \( f(u) := Bu \). In order to compute the Godunov numerical flux \( F(u^n_j, u^n_{j+1}) = f(u_R(0; u^n_j, u^n_{j+1})) \) for generic parameter vectors \( u_\ell \) and \( u_r \), we consider the Riemann problem
\[ \frac{\partial u}{\partial t} + B \frac{\partial u}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \]  
(14.34a)
\[ u(x, 0) = \begin{cases} 
    u_\ell & \text{if } x < 0, \\
    u_r & \text{if } x > 0. 
\end{cases} \]  
(14.34b)
Since the right eigenvectors $s_k$ of $B$ are linearly independent, we can decompose $u_\ell$ and $u_r$ as
\[
u_\ell = \sum_{k=1}^{m} \alpha_k s_k, \quad u_r = \sum_{k=1}^{m} \beta_k s_k \tag{14.35}\]
for some $\alpha_k$ and $\beta_k$. Note that the coefficients $\alpha_k$ and $\beta_k$ depend on the parameter vectors $u_\ell$ and $u_r$; i.e., $\alpha_k = \alpha_k(u_\ell, u_r)$ and $\beta_k = \beta_k(u_\ell, u_r)$. In the following we often suppress this dependency. In Section 12.4.2 we showed that the similarity solution of the Riemann problem (14.34) is given by
\[
u_R(x/t; u_\ell, u_r) = \sum_{\lambda_k < x/t} \beta_k s_k + \sum_{\lambda_k > x/t} \alpha_k s_k. \tag{14.36}\]
The solution $\nu_R$ is thus piecewise constant because the initial discontinuity at $x = 0$ propagates along all characteristics. The patches of constant-value vectors in the $(x,t)$ plane are separated by the characteristics. Along the centre line $x/t = 0$ the similarity solution is given by
\[
u_R(0; u_\ell, u_r) = \sum_{\lambda_k < 0} \beta_k s_k + \sum_{\lambda_k > 0} \alpha_k s_k; \tag{14.37}\]
i.e., $\nu_R(0; u_\ell, u_r)$ is a superposition of left-going (first term) and right-going (second term) waves. Rearranging terms in (14.37), we can derive the alternative expressions
\[
u_R(0; u_\ell, u_r) = u_\ell + \sum_{\lambda_k < 0} (\beta_k - \alpha_k) s_k \tag{14.38}\]
\[
u_R(0; u_\ell, u_r) = u_r - \sum_{\lambda_k > 0} (\beta_k - \alpha_k) s_k. \tag{14.38}\]
Using both expressions for $\nu_R(0; u_\ell, u_r)$ in (14.38), we find for the Godunov numerical flux $F(u_\ell, u_r) = B\nu_R(0; u_\ell, u_r)$ the two representations
\[
F(u_\ell, u_r) = Bu_\ell + \sum_{\lambda_k < 0} (\beta_k - \alpha_k) \lambda_k s_k \\
F(u_\ell, u_r) = Bu_r - \sum_{\lambda_k > 0} (\beta_k - \alpha_k) \lambda_k s_k. \tag{14.39}\]

We can find still another expression for the numerical flux as follows. With $\lambda_k^-$ as defined in (14.10), we can simplify the sum in the first expression in (14.39) as
\[
\sum_{\lambda_k < 0} (\beta_k - \alpha_k) \lambda_k s_k = \sum_{k=1}^{m} (\beta_k - \alpha_k) \lambda_k^- s_k = B^- (u_r - u_\ell) \tag{14.40}\]
to find for the numerical flux
\[
F(u_\ell, u_r) = Bu_\ell + B^- (u_r - u_\ell) = B^+ u_\ell + B^- u_r. \tag{14.41}\]
The expression (14.41) can also be found using the second representation for the numerical flux in (14.39).
14.3. Roe’s Approximate Riemann Solver

All three alternative expressions for the numerical flux can be combined with the Godunov scheme. If we insert, for example, the expression (14.41) in the Godunov scheme (14.32), we obtain

\[ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( (B^+ u_j^n + B^- u_{j+1}^n) - (B^+ u_j^{n-1} + B^- u_j^n) \right) \]

(14.42)

\[ = u_j^n - \frac{\Delta t}{\Delta x} \left( B^+ (u_j^n - u_{j-1}^n) + B^- (u_{j+1}^n - u_j^n) \right). \]

Like in the scalar case, the Godunov scheme reduces to the upwind scheme for linear hyperbolic systems.

We can derive an alternative formulation of the Godunov scheme if we use the first expression in (14.39) for the numerical flux \( F(u_j^n, u_{j+1}^n) \) at the right cell interface \( x_{j+1/2} \) and the second expression in (14.39) for the flux \( F(u_{j-1}^n, u_j^n) \) at the left interface \( x_{j-1/2} \).

Then the common term \( Bu_j^n \) cancels and we find the scheme

\[ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( \sum_{\lambda_k < 0} \gamma_{k,j+1/2}^n \lambda_k s_k + \sum_{\lambda_k > 0} \gamma_{k,j-1/2}^n \lambda_k s_k \right), \]

(14.43)

with \( \gamma_{k,j+1/2}^n := \beta_k(u_j^n, u_{j+1}^n) - \alpha_k(u_j^n, u_{j+1}^n) \), etc. Thus the cell average in cell \( V_j \) is only affected by left-going waves from \( x_{j+1/2} \) and right-going waves from \( x_{j-1/2} \). This form of Godunov’s scheme is sometimes referred to as the wave propagation form.

**Example 14.3** Consider once more the linear system from Example 14.1. Taking into account that \( \lambda_1 = -a < 0 \) and \( \lambda_2 = a > 0 \), we obtain the following wave propagation form of the Godunov scheme:

\[ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( \gamma_{1,j+1/2}^n \lambda_1 s_1 + \gamma_{2,j-1/2}^n \lambda_2 s_2 \right) \]

\[ = u_j^n + c(\gamma_{1,j+1/2}^n s_1 - \gamma_{2,j-1/2}^n s_2). \]

We have to compute the coefficients \( \gamma_{k,j+1/2}^n \) \((k = 1, 2)\) from the relation

\[ u_{j+1}^n - u_j^n = \gamma_{1,j+1/2}^n s_1 + \gamma_{2,j+1/2}^n s_2. \]

The solution is given by

\[ \gamma_{1,j+1/2}^n = \frac{1}{2}(u_{1,j+1}^n - u_{1,j}^n + u_{2,j+1}^n - u_{2,j}^n), \]

\[ \gamma_{2,j+1/2}^n = \frac{1}{2}(u_{1,j+1}^n - u_{1,j}^n + u_{2,j+1}^n + u_{2,j}^n). \]

Substituting this result in the wave propagation form above, we recover the upwind scheme from Example 14.1.

**14.3 Roe’s Approximate Riemann Solver**

The Godunov scheme requires the solution of local Riemann problems at control volume interfaces for the computation of the numerical flux; cf. (14.30). For nonlinear systems
this is a difficult and expensive computation and, moreover, it is not necessary since we only use the average value of the state variable over the control volumes. Therefore one often employs schemes based on approximate Riemann solvers, where the numerical flux is computed from an approximate and easier to solve Riemann problem. A straightforward choice is to replace the nonlinear Riemann problem by a linear one. In this section we consider a method named after Roe [134].

Like we did for the Godunov scheme (14.32), we will write the Roe scheme as

\[ u^n_{j+1} = u^n_j - \frac{\Delta t}{\Delta x} \left( F_R(u^n_j, u^n_{j+1}) - F_R(u^n_{j-1}, u^n_j) \right), \] (14.44)

where the Roe numerical flux \( F_R(u^n_j, u^n_{j+1}) \) is an approximation of the Godunov flux (14.30).

The problem thus is to find \( F_R(u_\ell, u_r) \) for generic state vectors \( u_\ell \) and \( u_r \). The derivation goes briefly as follows. First, we derive integral expressions for the Godunov numerical flux of the original Riemann problem. Subsequently, we replace this Riemann problem by a linearisation and again determine integral expressions for the corresponding Godunov numerical flux. Finally, combining these Godunov fluxes, we can determine the Roe numerical flux.

Consider the general Riemann problem

\[ \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0, \] (14.45a)

\[ u(x, 0) = \begin{cases} u_\ell & \text{if } x < 0, \\ u_r & \text{if } x > 0. \end{cases} \] (14.45b)

A schematic picture of the similarity solution \( u_R(x/t; u_\ell, u_r) \) in the \((x, t)\) plane is given in Figure 14.1. Here \( s_\ell \) and \( s_r \) are the minimum and maximum propagation speeds, respectively, involved in the Riemann problem. Thus the similarity solution holds in the wedge-shaped region in the centre and is bounded by the constant states \( u_\ell \) and \( u_r \). Integrating (14.45a) over the left control volume \([x_\ell, 0] \times [0, \Delta t]\) for some time step \( \Delta t > 0 \) and taking into account that the solution is constant along the \( t \) axis and outside the wedge-shaped region, we find

\[ F(u_\ell, u_r) = f(u_\ell) - \frac{1}{\Delta t} \left( \int_{x_\ell}^{0} u(x, \Delta t) \, dx + x_\ell u_\ell \right), \] (14.46)

with \( F(u_\ell, u_r) \) the Godunov flux corresponding to (14.45a), i.e.,

\[ F(u_\ell, u_r) = f(u_R(0; u_\ell, u_r)). \] (14.47)

Likewise, integration of (14.45a) over the right control volume \([0, x_r] \times [0, \Delta t]\) gives

\[ F(u_\ell, u_r) = f(u_r) + \frac{1}{\Delta t} \left( \int_{0}^{x_r} u(x, \Delta t) \, dx - x_r u_r \right). \] (14.48)

Equality of the right-hand sides of (14.46) and (14.48) can be established by integration of (14.45a) over \([x_\ell, x_r] \times [0, \Delta t]\).
Figure 14.1. Schematic representation of a similarity solution in the $(x,t)$ plane.

The computation of the Roe numerical flux is based on the approximate Riemann problem

\[
\frac{\partial \tilde{u}}{\partial t} + \tilde{B}(u_\ell, u_r) \frac{\partial \tilde{u}}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t > 0,
\]

(14.49a)

\[
\tilde{u}(x,0) = \begin{cases} 
    u_\ell & \text{if } x < 0, \\
    u_r & \text{if } x > 0,
\end{cases}
\]

(14.49b)

with $\tilde{B}(u_\ell, u_r)$ a constant matrix depending only on the initial states $u_\ell$ and $u_r$. Associated with (14.49) is the linear flux $\tilde{f}(u) := \tilde{B}(u_\ell, u_r) u$. The flux $\tilde{f}(u)$ and the Roe matrix $\tilde{B}(u_\ell, u_r)$ should be an approximation of the flux $f(u)$ in (14.45a) and the corresponding Jacobi matrix $B(u) := \frac{\partial}{\partial u} f(u)$, respectively. The following three requirements are imposed on the Roe matrix $\tilde{B}(u_\ell, u_r)$:

1. $f(u_\ell) - f(u_r) = \tilde{B}(u_\ell, u_r)(u_r - u_\ell)$.
2. $\tilde{B}(u, u) = B(u)$.
3. $\tilde{B}(u_\ell, u_r)$ has $m$ real eigenvalues $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_m$ and $m$ corresponding linearly independent eigenvectors $\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_m$.

The first property implies that an isolated shock or contact discontinuity connecting the constant states $u_\ell$ and $u_r$ in the true Riemann solution is approximated by a single discontinuity by virtue of the Rankine–Hugoniot jump relations; cf. (12.86). The second property means
that the Roe matrix $\tilde{B}(u_\ell, u_r)$ is consistent with the exact Jacobi matrix $B(u)$. Finally, the third property means that the linear system (14.49a) is hyperbolic, like the original system. Note that the eigenvalues $\tilde{\lambda}_k$ and eigenvectors $\tilde{s}_k$ depend on $u_\ell$ and $u_r$, i.e., $\tilde{\lambda}_k = \tilde{\lambda}_k(u_\ell, u_r)$ and $\tilde{s}_k = \tilde{s}_k(u_\ell, u_r)$; however, we will mostly use the shorthand notation $\tilde{\lambda}_k$ and $\tilde{s}_k$.

Repeating the previous analysis, we can derive the following expressions for the Godunov numerical flux $\tilde{F}(u_\ell, u_r) := \tilde{B}(u_\ell, u_r) u_R(0; u_\ell, u_r)$ associated with the linear Riemann problem (14.49):

$$
\tilde{F}(u_\ell, u_r) = \tilde{f}(u_\ell) - \frac{1}{\Delta t} \left( \int_{x_\ell}^{0} \tilde{u}(x, \Delta t) \, dx + x_\ell u_\ell \right) = \tilde{f}(u_\ell) + \frac{1}{\Delta t} \left( \int_{0}^{x_r} \tilde{u}(x, \Delta t) \, dx - x_r u_r \right). 
$$

(14.50)

In the derivation of (14.50) we assumed that the similarity solution $\tilde{u}_R(x/t; u_\ell, u_r)$ is completely contained in $[x_\ell, x_r] \times [0, \Delta t]$. Using the first expression on the right-hand side in (14.50) and the expression in (14.46), we find a relation between the numerical fluxes $F(u_\ell, u_r)$ and $\tilde{F}(u_\ell, u_r)$:

$$
F(u_\ell, u_r) - \tilde{F}(u_\ell, u_r) = f(u_\ell) - \tilde{f}(u_\ell) - \frac{1}{\Delta t} \int_{x_\ell}^{0} \left( u(x, \Delta t) - \tilde{u}(x, \Delta t) \right) \, dx. 
$$

(14.51)

Omitting the last, presumably small, term in (14.51), we obtain the following expression for the Roe numerical flux $F_R(u_\ell, u_r)$:

$$
F_R(u_\ell, u_r) = f(u_\ell) + \tilde{F}(u_\ell, u_r) - \tilde{f}(u_\ell). 
$$

(14.52)

Likewise, using (14.48) and the second expression in (14.50) and omitting the resulting integral, we find the alternative expression

$$
F_R(u_\ell, u_r) = f(u_r) + \tilde{F}(u_\ell, u_r) - \tilde{f}(u_r). 
$$

(14.53)

The equality of the right-hand sides of (14.52) and (14.53) is an immediate consequence of the first property of the Roe matrix.

Hence we have to find suitable expressions for the numerical flux $\tilde{F}(u_\ell, u_r)$. To do this let us introduce the decomposition

$$
u_\ell - u_\ell = \sum_{k=1}^{m} \gamma_k \tilde{s}_k, 
$$

(14.54)

where the coefficients $\gamma_k = \gamma_k(u_\ell, u_r)$ are referred to as the wave strengths. Then the exact solution of the linear Riemann problem (14.49) at $x/t = 0$ can be written as (see (14.38)),

$$
\tilde{u}_R(0; u_\ell, u_r) = u_\ell + \sum_{\lambda_k < 0} \gamma_k \tilde{s}_k = u_\ell - \sum_{\lambda_k > 0} \gamma_k \tilde{s}_k. 
$$

(14.55)
14.3. Roe's Approximate Riemann Solver

We can now compute \( \tilde{F}(u_\ell, u_r) \) to find

\[
\tilde{F}(u_\ell, u_r) = \tilde{f}(u_\ell) + \sum_{\lambda_k > 0} \gamma_k \tilde{\lambda}_k \tilde{s}_k,
\]

\[
= \tilde{f}(u_r) - \sum_{\lambda_k < 0} \gamma_k \tilde{\lambda}_k \tilde{s}_k. \tag{14.56}
\]

Combining (14.52), (14.53), and (14.56), we thus obtain for the Roe numerical flux

\[
F_R(u_\ell, u_r) = f(u_\ell) + \sum_{\lambda_k < 0} \gamma_k \tilde{\lambda}_k \tilde{s}_k
\]

\[
= f(u_r) - \sum_{\lambda_k > 0} \gamma_k \tilde{\lambda}_k \tilde{s}_k. \tag{14.57}
\]

From the first expression in (14.57) we conclude that \( F_R(u_\ell, u_r) = f(u_\ell) \) if all eigenvalues \( \tilde{\lambda}_k \) are positive; i.e., all waves in the Riemann problem (14.49) propagate in the positive \( x \) direction. On the other hand, \( \tilde{F}(u_\ell, u_r) = f(u_r) \) if all \( \tilde{\lambda}_k \) are negative. Taking the average value of both expressions in (14.57), we finally obtain the alternative expression

\[
F_R(u_\ell, u_r) = \frac{1}{2} \left( f(u_\ell) + f(u_r) \right) - \frac{1}{2} \sum_{k=1}^m \gamma_k |\tilde{\lambda}_k| \tilde{s}_k. \tag{14.58}
\]

The computation of the Roe numerical flux \( F_R(u_\ell, u_r) \) thus requires the following steps:

- the Roe matrix \( \tilde{B}(u_\ell, u_r) \) from \( f(u_r) - f(u_\ell) = \tilde{B}(u_\ell, u_r)(u_r - u_\ell) \),
- its eigenvalues \( \tilde{\lambda}_k \) and eigenvectors \( \tilde{s}_k \),
- the wave strengths \( \gamma_k \) from (14.54), and
- the flux from either expression in (14.57).

The crucial step is the computation of the Roe matrix \( \tilde{B}(u_\ell, u_r) \). This is generally a difficult and expensive task, and we will not discuss it here. However, for many relevant conservation laws there do exist efficient methods to compute the Roe matrix. As an example, we will compute in Section 14.5.2 the Roe matrix for the shallow-water equations.

A drawback of the Roe scheme is that it cannot represent rarefaction waves because the underlying linear Riemann problem (14.49) does not allow such solutions; cf. Chapter 12. A way to overcome this disadvantage is the Harten–Hyman entropy fix. We will discuss this as well in Section 14.5.2.

An alternative derivation of the Roe scheme is based on the wave propagation form (14.43) of the Godunov scheme. Consider the Godunov scheme (14.32) rewritten in the form

\[
u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( \left( f(u_R(0; u_j^n, u_{j+1}^{n+1})) - f(u_j^n) \right) \\
+ \left( f(u_j^n) - f(u_R(0; u_j^n, u_{j-1}^{n+1})) \right) \right), \tag{14.59}
\]
The flux differences \( f(u_R(0; u^0; u''^0)) - f(u^0) \) and \( f(u_R^0) - f(u_R(0; u''^0, u^0)) \) include the waves emanating from the interfaces \( x = x_{j+1/2} \) and \( x = x_{j-1/2} \), respectively, and for this reason (14.59) is referred to as the wave propagation form. Using (14.39), we see that for the special case \( f(u) = Bu \) the scheme (14.59) simply reduces to the linear wave propagation form (14.43). The basic idea of the Roe scheme is to approximate the flux differences in (14.59) from local linear Riemann problems of the form (14.49). Obviously, we take the Roe matrices \( \hat{B}(u_j^0, u_{j+1}^0) \) and \( \hat{B}(u_{j-1}^0, u_j^0) \) to approximate \( f(u_R(0; u^0; u''^0)) - f(u^0) \) and \( f(u_R^0) - f(u_R(0; u''^0, u^0)) \), respectively. Using the appropriate expressions for the Roe flux in (14.57), we obtain the wave propagation form

\[
\begin{align*}
    u_{j+1}^n &= u_j^n - \frac{\Delta t}{\Delta x} \left( \sum_{i,j+1/2 < 0} \gamma_{k,j+1/2}^n \tilde{\lambda}_{k,j+1/2}^n \tilde{u}_{k,j+1/2} \right. \\
    &\quad \left. + \frac{\Delta t}{\Delta x} \sum_{i,j-1/2 > 0} \gamma_{k,j-1/2}^n \tilde{\lambda}_{k,j-1/2}^n \tilde{u}_{k,j-1/2} \right),
\end{align*}
\]

(14.60)

where \( \tilde{\lambda}_{k,j+1/2}^n = \tilde{\lambda}_k(u_j^0, u_{j+1}^0), \tilde{\lambda}_{k,j-1/2}^n = \tilde{\lambda}_k(u_{j-1}^0, u_j^0), \) etc., and where the wave strengths \( \gamma_{k,j+1/2}^n \) follow from the decomposition (14.54) with \( u_\ell = u_j^0 \) and \( u_i = u_{j+1}^0 \).

### 14.4 Slope Limiter Methods

For the linear system (14.1a) we have derived the upwind scheme (14.14), which is only first order accurate. This is not very satisfactory, and therefore we look for a high-resolution scheme, i.e., a scheme that is second order accurate in regions where the solution is smooth and suppresses spurious oscillations in the vicinity of discontinuities. There are two possible approaches to achieve this: the flux limiter method and the slope limiter method. We only treat the slope limiter method, because this approach, which is considered in Section 14.4.1 for the linear case, can easily be generalised to nonlinear systems, as shown in Section 14.4.2.

#### 14.4.1 A Slope Limiter Method for Linear Systems

In the derivation of the slope limiter method for (14.1a), we first decouple the system and then apply the basic slope limiter method to the advection equation for each characteristic variable \( u_k \). Thus consider the decoupled system of equations (14.6a). The basic slope limiter method for these equations is based on the Godunov scheme in combination with a piecewise linear distribution for the numerical solution in each control volume. Applying the basic slope limiter method from Section 13.8 to (14.6a), we obtain the following scheme for the characteristic variables \( \tilde{u}_k \):

\[
\begin{align*}
    \tilde{u}_{k,j+1}^n &= \tilde{u}_{k,j}^n - \frac{\Delta t}{\Delta x} \left( \tilde{\lambda}_k^c \left( \tilde{u}_{k,j}^n - \tilde{u}_{k,j-1}^n \right) + \tilde{\lambda}_k^c \left( \tilde{u}_{k,j+1}^n - \tilde{u}_{k,j}^n \right) \right) \\
    &\quad + \frac{1}{2} |\tilde{\lambda}_k^c| \left( 1 - \frac{\Delta t}{\Delta x} |\tilde{\lambda}_k^c| \right) \Delta x \left( \tilde{\sigma}_{k,j(k)}^n - \tilde{\sigma}_{k,j(k)-1}^n \right),
\end{align*}
\]

(14.61)
14.4. Slope Limiter Methods

with \( \tilde{\sigma}_{k,j}(k) \) the slope of the numerical solution in the control volume labeled \( j_u(k) \), where the index \( j_u(k) \) is defined by

\[
  j_u(k) := \begin{cases} 
  j & \text{if } \lambda_k > 0, \\
  j + 1 & \text{if } \lambda_k < 0;
  \end{cases}
\]

(14.62)
cf. (13.134) and (13.135). The slopes \( \tilde{\sigma}_{k,j}(k) \) are chosen according to

\[
  \tilde{\sigma}_{k,j}^n = \Phi_{k,j+1/2}^n \frac{1}{\Delta x} (\tilde{u}_{k,j}^n - \tilde{u}_{k,j}^n),
\]

(14.63)
with \( \Phi_{k,j+1/2}^n \) the slope limiter given by

\[
  \Phi_{k,j+1/2}^n = \Phi(r_{k,j+1/2}^n), \quad r_{k,j+1/2}^n := \begin{cases} 
  \tilde{u}_{k,j}^n - \tilde{u}_{k,j}^n & \text{if } \lambda_k > 0, \\
  \tilde{u}_{k,j+1}^n - \tilde{u}_{k,j}^n & \text{if } \lambda_k < 0;
  \end{cases}
\]

(14.64)
cf. (13.104) and (13.105). The limiter function \( \Phi(r) \) should lie in the total variation diminishing (TVD) region of Figure 13.18 in order that scheme (14.61) is TVD.

It is most convenient to formulate the slope limiter method for (14.1a) in terms of the variable \( u \). Therefore we multiply (14.61) by the eigenvector \( s_k \) and sum over all indices \( k \). Taking into account that \( u = S\tilde{u} = \sum_{k=1}^{m} \tilde{u}_k s_k \), we obtain

\[
  u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( \sum_{k=1}^{m} (\tilde{u}_{k,j}^n - \tilde{u}_{k,j,-1}^n) \lambda_k^+ s_k + \sum_{k=1}^{m} (\tilde{u}_{k,j+1}^n - \tilde{u}_{k,j}^n) \lambda_k^- s_k + \right.
\]

\[
  \left. + \frac{1}{2} \sum_{k=1}^{m} |\lambda_k| \left( 1 - \frac{\Delta t}{\Delta x} |\lambda_k| \right) \Delta x (\tilde{\sigma}_{k,j}(k) - \tilde{\sigma}_{k,j}(k-1)) s_k \right).
\]

(14.65)

Since \( \lambda_k^+ \) and \( \lambda_k^- \) are the eigenvalues of the split matrices \( B^+ \) and \( B^- \), respectively, the first and second sums on the right-hand side of (14.65) reduce to

\[
  \sum_{k=1}^{m} (\tilde{u}_{k,j}^n - \tilde{u}_{k,j,-1}^n) \lambda_k^+ s_k = B^+(u_j^n - u_{j-1}^n),
\]

(14.66a)

\[
  \sum_{k=1}^{m} (\tilde{u}_{k,j+1}^n - \tilde{u}_{k,j}^n) \lambda_k^- s_k = B^-(u_{j+1}^n - u_j^n).
\]

(14.66b)

Substitution of (14.66) in (14.65) gives the slope limiter scheme

\[
  u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( B^+(u_j^n - u_{j-1}^n) + B^-(u_{j+1}^n - u_j^n) \right. \right.
\]

\[
  \left. + \frac{1}{2} \sum_{k=1}^{m} |\lambda_k| \left( 1 - \frac{\Delta t}{\Delta x} |\lambda_k| \right) \Delta x (\tilde{\sigma}_{k,j}(k) - \tilde{\sigma}_{k,j}(k-1)) s_k \right).
\]

(14.67)
The numerical flux $F(u^n_j, u^n_{j+1})$ for (14.67) is obviously given by

$$F(u^n_j, u^n_{j+1}) = B^+ u^n_j + B^- u^n_{j+1} + \frac{1}{2} \sum_{k=1}^{m} |\lambda_k| \left(1 - \frac{\Delta t}{\Delta x} |\lambda_k| \right) \Delta x \tilde{\sigma}^n_{k,j}(k) s_k$$  (14.68)

and is a straightforward generalisation of the slope limiter flux in (13.134) for the linear advection equation. The numerical flux $F(u^n_j, u^n_{j+1})$ is thus equivalent to the upwind flux plus a correction term containing the slopes $\tilde{\sigma}^n_{k,j}(k)$ of the numerical approximations of the variables $\tilde{u}_k$, taken in the control volume that lies upwind (with respect to $\tilde{u}_k$) of the cell boundary $x_{j+1/2}$.

The coefficients $\gamma_{k,j+1/2} := \tilde{u}^n_{k,j+1} - \tilde{u}^n_{k,j}$ in (14.63) for the slopes $\tilde{\sigma}^n_{k,j}(k)$ can be computed from the decomposition

$$u^n_{j+1} - u^n_j = \sum_{k=1}^{m} \gamma_{k,j+1/2} s_k.$$  (14.69)

Combining (14.62), (14.63), and (14.64), we find the following expressions for the slope $\tilde{\sigma}^n_{k,j}$:

$$\tilde{\sigma}^n_{k,j} = \begin{cases} \Phi \left( \frac{\gamma_{k,j+1/2}}{\gamma_{k,j-1/2}} \right) \frac{1}{\Delta x} \gamma_{k,j+1/2} & \text{if } \lambda_k > 0, \\ \Phi \left( \frac{\gamma_{k,j+1/2}}{\gamma_{k,j-1/2}} \right) \frac{1}{\Delta x} \gamma_{k,j-1/2} & \text{if } \lambda_k < 0. \end{cases}$$  (14.70)

**Example 14.4** Consider the linear system from Example 14.1. For this system expression (14.68) for the numerical flux $F(u^n_j, u^n_{j+1})$ reduces to

$$F(u^n_j, u^n_{j+1}) = B^+ u^n_j + B^- u^n_{j+1} + \frac{1}{2} a(1-c) \Delta x (\tilde{\sigma}^n_{1,j+1/2} s_1 + \tilde{\sigma}^n_{2,j} s_2),$$

with $c = a \Delta t / \Delta x$. The matrices $B^+$ and the eigenvectors $s_k$ ($k = 1, 2$) are given in Example 14.1. Since $\lambda_1 < 0$ and $\lambda_2 > 0$, we have to select the second expression in (14.70) for $\tilde{\sigma}^n_{1,j}$ and the first one for $\tilde{\sigma}^n_{2,j}$. The coefficients $\gamma_{k,j+1/2}$ can be determined from (14.69), and we find

$$\gamma_{1,j+1/2} = \frac{1}{2} \left( u^n_{1,j+1} - u^n_{1,j} + u^n_{2,j+1} - u^n_{2,j} \right),$$

$$\gamma_{2,j+1/2} = \frac{1}{2} \left( u^n_{1,j+1} - u^n_{1,j} - u^n_{2,j+1} + u^n_{2,j} \right).$$

**14.4.2 A Slope Limiter Method for Nonlinear Systems**

In this section we briefly discuss the slope limiter method for the nonlinear system (14.21a). In particular, we give a generalization of the numerical flux $F(u^n_j, u^n_{j+1})$ in (14.68). The basic idea is to replace the nonlinear system (14.21a) by a suitable linearization in the neighbourhood of each cell interface $x_{j+1/2}$ and apply the slope limiter method of the previous section.
14.4. Slope Limiter Methods

We compute the numerical flux \( F(u_j^n, u_{j+1}^n) \) from the local linearization

\[
\frac{\partial \tilde{u}}{\partial t} + \tilde{B}_j^n \frac{\partial \tilde{u}}{\partial x} = 0,
\]

which holds in the vicinity of the cell interface \( x_{j+1/2} \). We take \( \tilde{B}_j^n = \tilde{B}(u_j^n, u_{j+1}^n) \), where \( \tilde{B}(u_j, u_r) \) is the Roe matrix satisfying the three conditions in Section 14.3. Let the eigenvalues and eigenvectors of \( \tilde{B}_j^n \) be denoted by \( \tilde{\lambda}_{k,j}^n \) and \( \tilde{s}_{k,j}^n \), respectively. Then the Roe numerical flux \( F_R(u_j^n, u_{j+1}^n) \) is given by

\[
F_R(u_j^n, u_{j+1}^n) = f(u_j^n) + \sum_{\tilde{k}, j \neq n < 0} \tilde{\lambda}_{k,j}^n \tilde{s}_{k,j}^n \tilde{B}_j^n \tilde{u}_{j+1/2}^n
\]

\[
= f(u_{j+1}^n) - \sum_{\tilde{k}, j \neq n > 0} \tilde{\lambda}_{k,j}^n \tilde{s}_{k,j}^n \tilde{B}_j^n \tilde{u}_{j+1/2}^n
\]

(14.72)

where the coefficients \( \tilde{\lambda}_{k,j}^n \) have to be computed from the decomposition

\[
u_j^{n+1} - u_j^n = \sum_{k=1}^{m} \gamma_{k,j+1/2}^n \tilde{s}_{k,j+1/2}^n.
\]

Recall that the decomposition (14.73) is always possible, since the eigenvectors \( \tilde{s}_{k,j+1/2}^n \) are linearly independent. For the linear system (14.1a) we have \( f(u) = Bu \), and the expressions for the Roe numerical flux in (14.72) reduce to

\[
F_R(u_j^n, u_{j+1}^n) = B^+ u_j^n + B^- u_{j+1}^n.
\]

(14.74)

which is the first part of the numerical flux in (14.68). In the derivation of (14.74) we have used that \( \tilde{\lambda}_{k}^n \) are the eigenvalues of \( B^\pm \); see Section 14.1.

A natural generalization of the numerical flux (14.68) to the nonlinear system (14.21a) is

\[
F(u_j^n, u_{j+1}^n) = F_R(u_j^n, u_{j+1}^n) + \frac{1}{2} \sum_{k=1}^{m} |\gamma_{k,j+1/2}^n| \left( 1 - \frac{\Delta t}{\Delta x} |\tilde{\lambda}_{k,j+1/2}^n| \right) \Delta x \tilde{s}_{k,j+1/2}^n \tilde{s}_{k,j+1/2}^n
\]

(14.75)

where \( F_R(u_j^n, u_{j+1}^n) \) is the Roe numerical flux given in (14.72) and \( \tilde{s}_{k,j+1/2}^n \) is the limited slope in control volume \( V_{j+1/2} \) of the characteristic variable \( \tilde{u}_{k,j+1/2}^n \) associated with the linear system (14.71). The index \( j_k(k) \) defined in (14.62), with \( \lambda_k \) replaced by \( \tilde{\lambda}_{k,j+1/2}^n \), denotes the control volume that lies upwind (with respect to \( \tilde{u}_{k,j+1/2}^n \)) of the interface \( x_{j+1/2} \). Analogously to (14.70), the slopes \( \tilde{s}_{k,j}^n \) are given by

\[
\tilde{s}_{k,j}^n = \begin{cases} \Phi \frac{\gamma_{k,j+1/2}^n}{\gamma_{k,j+1/2}^n} \frac{1}{\Delta x} \gamma_{k,j+1/2}^n & \text{if } \tilde{\lambda}_{k,j+1/2}^n > 0, \\ \Phi \frac{\gamma_{k,j+1/2}^n}{\gamma_{k,j-1/2}^n} \frac{1}{\Delta x} \gamma_{k,j-1/2}^n & \text{if } \tilde{\lambda}_{k,j-1/2}^n < 0, \end{cases}
\]

(14.76)

where the limiter function \( \Phi \) has to be chosen such that the resulting numerical method is TVD; see Section 13.7.
14.5 Numerical Solution of the Shallow-Water Equations

The shallow-water equations were introduced in Section 12.5 as an example of a nonlinear hyperbolic system. In this section we will give a detailed elaboration of the Godunov and Roe schemes for this system.

14.5.1 Numerical Solution by Godunov’s Method

In this section we will determine the Godunov numerical flux \( F(u_\ell, u_r) \), which is defined in (14.30), for the shallow-water equations. Obviously, this flux depends on the solution of the corresponding Riemann problem, which is presented in Section 12.5. The four possible wave patterns of the Riemann problem are displayed in Figure 12.14. In all four cases we have the constant states \( u_\ell, u_* \), and \( u_r \) connected by shocks and/or rarefaction waves. We compute the intermediate state \( u_* \) from the following algorithm:

1. Compute \( A = \frac{\varphi_r}{\varphi_\ell} \) and \( B = \frac{(u_r - u_\ell)}{c_\ell} \).
2. Solve (12.136) for \( z_1 \).
3. Compute \( z_2 = \frac{z_1}{A} \).
4. Compute \( \varphi_* = \frac{z_1 \varphi_\ell}{A} \) and \( u_* = u_\ell + g(z_1)c_\ell \).

Recall that \( c = \sqrt{\varphi} \). The function \( g = g(z) \) in this algorithm is defined in (12.121). The conditions for which shocks and/or rarefaction waves occur are given in Table 14.1. The speeds \( s_k \) \( (k = 1, 2) \) of possible shocks are given by

\[
\begin{align*}
  s_1 &= u_\ell - c_\ell \sqrt{\frac{1}{2}(1 + z_1)z_1}, \\
  s_2 &= u_r + c_\ell \sqrt{\frac{1}{2}(1 + z_2)z_2}.
\end{align*}
\]  

(14.77)

The four elementary waves of the Riemann problem are detailed in (12.138) to (12.141). Combining these, we can determine the four possible wave patterns in Figure 12.14. However, for the Godunov flux we only need the similarity solution along the \( t \) axis, i.e., \( u_R(0; u_\ell, u_r) \), for which we can distinguish the following four cases:

Case 1. 1- and 2-shocks.

\[
 u_R(0; u_\ell, u_r) = \begin{cases} 
 u_\ell & \text{if } s_1 > 0, \\
 u_* & \text{if } s_1 < 0 < s_2, \\
 u_r & \text{if } s_2 < 0.
\end{cases}
\]  

(14.78)

<table>
<thead>
<tr>
<th>( k )-wave</th>
<th>shock</th>
<th>rarefaction wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( B &lt; \sqrt{A}g(1/A) )</td>
<td>( B \geq \sqrt{A}g(1/A) )</td>
</tr>
<tr>
<td>2</td>
<td>( B &lt; g(A) )</td>
<td>( B \geq g(A) )</td>
</tr>
</tbody>
</table>

Table 14.1. Conditions for shocks and rarefaction waves.
14.5. Numerical Solution of the Shallow-Water Equations

Case 2. 1-rarefaction wave and 2-shock.

\[
\begin{align*}
\mathbf{u}_R(0; \mathbf{u}_\ell, \mathbf{u}_r) &= \begin{cases} 
\mathbf{u}_\ell & \text{if } \mathbf{u}_\ell - c_\ell > 0, \\
\varphi_1 \left( \frac{1}{u_1} \right) & \text{if } \mathbf{u}_\ell - c_\ell < 0 < \mathbf{u}_s - c_s, \\
\mathbf{u}_s & \text{if } \mathbf{u}_s - c_s < 0 < s_2, \\
\mathbf{u}_r & \text{if } s_2 < 0,
\end{cases} \\
\text{with } & \varphi_1 := \frac{1}{9} (\mathbf{u}_\ell + 2c_\ell)^2, \quad u_1 := \frac{1}{3} (\mathbf{u}_\ell + 2c_\ell).
\end{align*}
\] (14.79)

Case 3. 1-shock and 2-rarefaction wave.

\[
\begin{align*}
\mathbf{u}_R(0; \mathbf{u}_\ell, \mathbf{u}_r) &= \begin{cases} 
\mathbf{u}_\ell & \text{if } s_1 > 0, \\
\phi_2 \left( \frac{1}{u_2} \right) & \text{if } s_1 < 0 < \mathbf{u}_r + c_r, \\
\mathbf{u}_s & \text{if } s_1 < 0 < \mathbf{u}_r + c_r, \\
\mathbf{u}_r & \text{if } \mathbf{u}_r + c_r < 0,
\end{cases} \\
\text{with } & \phi_2 := \frac{1}{9} (-\mathbf{u}_r + 2c_r)^2, \quad u_2 := \frac{1}{3} (\mathbf{u}_r - 2c_r).
\end{align*}
\] (14.81)

Case 4. 1- and 2-rarefaction waves.

\[
\begin{align*}
\mathbf{u}_R(0; \mathbf{u}_\ell, \mathbf{u}_r) &= \begin{cases} 
\mathbf{u}_\ell & \text{if } \mathbf{u}_\ell - c_\ell > 0, \\
\varphi_1 \left( \frac{1}{u_1} \right) & \text{if } \mathbf{u}_\ell - c_\ell < 0 < \mathbf{u}_s - c_s, \\
\mathbf{u}_s & \text{if } \mathbf{u}_s - c_s < 0 < \mathbf{u}_r + c_r, \\
\varphi_2 \left( \frac{1}{u_2} \right) & \text{if } \mathbf{u}_s + c_s < 0 < \mathbf{u}_r + c_r, \\
\mathbf{u}_r & \text{if } \mathbf{u}_r + c_r < 0.
\end{cases} \\
\text{with } & \varphi_1 := \frac{1}{9} (\mathbf{u}_\ell + 2c_\ell)^2, \quad \varphi_2 := \frac{1}{9} (-\mathbf{u}_r + 2c_r)^2.
\end{align*}
\] (14.83)

We can now simply compute the Godunov flux by substituting the various expressions for \( \mathbf{u}_R(0; \mathbf{u}_\ell, \mathbf{u}_r) \) given in (14.78) to (14.83) in the flux vector \( f(\mathbf{u}) \) defined in (12.104).

**Example 14.5** We apply the Godunov scheme to compute the numerical solution of the *dam break problem*. This is a special case of the Riemann problem for the shallow-water equations with \( u_\ell = u_r = 0 \) and \( h_\ell \neq h_r \), and it describes the situation where a dam situated at \( x = 0 \) separates two water levels. At \( t = 0 \) the dam is removed and waves start to propagate from \( x = 0 \). In this example we choose \( \phi_\ell = 100g \) and \( \phi_\ell = 50g \) with \( g \) the gravitational acceleration, so that \( A = 0.5 \) and \( B = 0. \) We can easily verify the inequalities \( \sqrt{Ag(1/A)} \leq B < g(A) \), implying that the solution consists of a 1-rarefaction wave and a 2-shock. This is indeed confirmed by the numerical results presented in Figure 14.2, for which we take \( \Delta x = 0.8 \) and \( \Delta t = 10^{-2} \). We see that the 2-shock is slightly smeared due to the first order accuracy of the Godunov scheme.

Figure 14.2. Numerical approximation of the dam break problem at time levels 0.5 (top) and 1.0 (bottom).

14.5.2 Numerical Solution by Roe’s Method

We can also apply the Roe scheme (14.44), with numerical flux given in (14.57), to the shallow-water equations. In particular, we focus on the computation of the Roe matrix \( \widetilde{B}(u_\ell, u_r) \). We will also discuss the entropy fix, as mentioned above.

Consider the shallow-water equations (12.103). In order to compute the numerical flux \( F_R(u_\ell, u_r) \), we first have to compute the Roe matrix \( \widetilde{B}(u_\ell, u_r) \) from the relation

\[
\begin{align*}
 f(u_r) - f(u_\ell) &= \widetilde{B}(u_\ell, u_r)(u_r - u_\ell). \tag{14.84}
\end{align*}
\]

The basic idea is to introduce a parameter vector \( z \) such that the differences \( u_r - u_\ell \) and \( f(u_r) - f(u_\ell) \) in (14.84) can be expressed in terms of the difference \( z_r - z_\ell \). In particular, we try to find relations of the form

\[
\begin{align*}
 u_r - u_\ell &= \tilde{P}(u_\ell, u_r)(z_r - z_\ell), \tag{14.85a} \\
 f(u_r) - f(u_\ell) &= \tilde{Q}(u_\ell, u_r)(z_r - z_\ell). \tag{14.85b}
\end{align*}
\]

where \( \tilde{P}(u_\ell, u_r) \) and \( \tilde{Q}(u_\ell, u_r) \) are 2 \times 2 matrices depending on the state vectors \( u_\ell \) and \( u_r \). Combining (14.84) and (14.85) and assuming that \( \tilde{P}(u_\ell, u_r) \) is nonsingular, we readily see that the Roe matrix is given by

\[
\widetilde{B}(u_\ell, u_r) = \tilde{Q}(u_\ell, u_r)\tilde{P}^{-1}(u_\ell, u_r). \tag{14.86}
\]

In the following we will suppress the arguments \( u_\ell \) and \( u_r \).
A suitable choice for the parameter vector \( z \) appears to be

\[
z := \frac{1}{\sqrt{\phi}} u \quad \text{or} \quad \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \sqrt{\phi} \\ \sqrt{\phi} u \end{pmatrix}. \tag{14.87}
\]

We can now express the state vector \( u \) and the flux vector \( f(u) \) from (12.104) in terms of \( z \) as

\[
u = \begin{pmatrix} z_1^2 \\ z_1 z_2 \end{pmatrix}, \quad f(u) = \begin{pmatrix} z_1 z_2 \\ z_1^2 + \frac{1}{2} z_1^4 \end{pmatrix}. \tag{14.88}
\]

Substituting (14.88) into (14.85), we can determine the matrices \( \tilde{P} \) and \( \tilde{Q} \). We can verify that

\[
\phi_t - \phi_\ell = 2 \tilde{z}_1 (z_{1,\ell} - z_{1,\ell}), \tag{14.89a}
\]

\[
\phi_t u_t - \phi_\ell u_\ell = \tilde{z}_2 (z_{1,\ell} - z_{1,\ell}) + \tilde{z}_1 (z_{2,\ell} - z_{2,\ell}), \tag{14.89b}
\]

\[
\phi_t u_t^2 - \phi_\ell u_\ell^2 = 2 \tilde{z}_2 (z_{2,\ell} - z_{2,\ell}), \tag{14.89c}
\]

\[
\frac{1}{2} (\phi_t^2 - \phi_\ell^2) = 2 \tilde{\phi} \tilde{z}_1 (z_{1,\ell} - z_{1,\ell}), \tag{14.89d}
\]

with \( \tilde{v} := \frac{1}{2} (v_t + v_\ell) \) the average value of the variable \( v \), and from these relations we obtain

\[
\tilde{P} = \begin{pmatrix} 2 \tilde{z}_1 & 0 \\ \tilde{z}_2 & \tilde{z}_1 \end{pmatrix}, \quad \tilde{Q} = \begin{pmatrix} \tilde{z}_2 & \tilde{z}_1 \\ 2 \tilde{\phi} \tilde{z}_1 & 2 \tilde{z}_2 \end{pmatrix}. \tag{14.90}
\]

The matrix \( \tilde{P} \) is obviously nonsingular since \( \det(\tilde{P}) = 2 \tilde{z}_1^2 \neq 0 \). From (14.86) and (14.90) we can determine the Roe matrix in terms of the average \( \tilde{z} \) of the parameter vector:

\[
\tilde{B} = \begin{pmatrix} 0 & 1 \\ \tilde{\phi} - \frac{\tilde{z}_2}{\tilde{z}_1} & 2 \tilde{z}_2 - \frac{1}{\tilde{z}_1} \end{pmatrix}. \tag{14.91a}
\]

In terms of the state vector \( u \), the Roe matrix reads

\[
\tilde{B} = \begin{pmatrix} 0 & 1 \\ \tilde{\phi} - \tilde{u}^2 & 2 \tilde{u} \end{pmatrix}, \tag{14.91b}
\]

where we have introduced the weighted average velocity \( \tilde{u} \) defined by

\[
\tilde{u} := \frac{(u, \sqrt{\phi})_t + (u, \sqrt{\phi})_\ell}{\sqrt{\phi_t} + \sqrt{\phi_\ell}} = \frac{\tilde{z}_2}{\tilde{z}_1}. \tag{14.92}
\]

Note that the Roe matrix \( \tilde{B} \) in (14.92) is precisely the Jacobi matrix \( B \) in (12.105), with the geopotential \( \phi \) replaced by its average value \( \tilde{\phi} \) and the flow velocity \( u \) by the weighted average \( \tilde{u} \) defined in (14.92). Consequently, its eigenvalues \( \tilde{\lambda}_k \) and eigenvectors \( \tilde{s}_k \) \( (k = 1, 2) \)
are given by
\[
\tilde{\lambda}_1(u_\ell, u_r) = \bar{u} - \bar{c}, \quad \tilde{\lambda}_2(u_\ell, u_r) = \bar{u} + \bar{c}, \quad \bar{c} := \sqrt{\varphi}, \quad (14.93a)
\]
\[
\tilde{s}_1(u_\ell, u_r) = \left( \frac{1}{\bar{u} - \bar{c}} \right), \quad \tilde{s}_2(u_\ell, u_r) = \left( \frac{1}{\bar{u} + \bar{c}} \right), \quad (14.93b)
\]
and thus the third property of the Roe matrix is automatically satisfied. Moreover, if \( u_\ell = u_r = u \), then \( \varphi = \varphi \) and \( \bar{u} = u \), and, consequently, the second property of the Roe matrix is also trivially fulfilled.

We now compute the Roe numerical flux \( F_R(u_\ell, u_r) \) as follows. From (14.54) and (14.93) we can compute the coefficients \( \gamma_k \) (\( k = 1, 2 \)):
\[
\gamma_1 = \frac{1}{2\bar{c}} \left( (\bar{u} + \bar{c})(\varphi_r - \varphi_\ell) - (\varphi_r u_\ell - \varphi_\ell u_r) \right), \quad (14.94a)
\]
\[
\gamma_2 = \frac{1}{2\bar{c}} \left( -(\bar{u} - \bar{c})(\varphi_r - \varphi_\ell) + (\varphi_r u_\ell - \varphi_\ell u_r) \right). \quad (14.94b)
\]
In the evaluation of the expressions for the numerical flux in (14.57) we have to distinguish three cases: \( \tilde{\lambda}_1 > 0 \), \( \tilde{\lambda}_1 \leq 0 \leq \tilde{\lambda}_2 \), and \( \tilde{\lambda}_2 < 0 \). In the case \( \tilde{\lambda}_1 > 0 \) the first expression in (14.57) is most appropriate, since the sum in it is void. We then have \( F_R(u_\ell, u_r) = f(u_\ell) \). Similarly, when \( \tilde{\lambda}_2 < 0 \), we have to choose the second expression in (14.57), and we obtain \( F_R(u_\ell, u_r) = f(u_\ell) \). Most interesting is the second case, for which we can use either expression in (14.57); we choose the first one. Summarizing, we obtain for the Roe numerical flux
\[
F_R(u_\ell, u_r) = \begin{cases} 
  f(u_\ell) & \text{if } \bar{u} > \bar{c}, \\
  f(u_\ell) + \gamma_1 \tilde{s}_1 \tilde{\lambda}_1 & \text{if } |\bar{u}| \leq \bar{c}, \\
  f(u_r) & \text{if } \bar{u} < -\bar{c},
\end{cases} \quad (14.95)
\]
with \( \gamma_1, \tilde{\lambda}_1, \text{and} \tilde{s}_1 \) defined in (14.94) and (14.93).

We conclude this section with an outline of the Harten–Hyman entropy fix for (12.103); for more details see, e.g., [60, 87, 158]. Consider the particular solution of the Riemann problem (14.45) depicted in Figure 14.3. In this case the solution consists of three constant states \( u_\ell, u_\ell, \) and \( u_t \), connected by a 1-rarefaction wave and a 2-shock. Moreover, the 1-rarefaction wave is critical, which means that the corresponding eigenvalue \( \lambda_1 = u - c \) changes sign across the wave. In particular, we have
\[
\tilde{\lambda}_{1,\ell} < 0 < \tilde{\lambda}_{1,s}, \quad (14.96)
\]
where \( \tilde{\lambda}_{1,\ell} \text{ and } \tilde{\lambda}_{1,s} \) denote the value of \( \tilde{\lambda}_1 \) at the left and right ends of the 1-rarefaction wave, respectively. Thus \( \tilde{\lambda}_{1,\ell} \text{ is the speed of the head of the rarefaction wave and } \tilde{\lambda}_{1,s} \text{ is the speed of the tail.}

The computation of the Roe numerical flux \( F_R(u_\ell, u_r) \) is based on the linearized Riemann problem (14.49). Its solution consists of the constant states \( u_\ell, u_\ell, \) and \( u_t \), now connected by two shocks. The 1-rarefaction wave is thus approximated by a 1-shock propagating with speed \( \tilde{\lambda}_1 \), which is obviously not correct. Suppose we want to compute the Roe numerical flux using the first expression in (14.57). The sum in this expression
accounts for the effect of jumps in the solution of (14.49) propagating with negative speeds \( \lambda_{1,\ell} \). Straightforward application of this formula would lead to an entropy-violating shock as an approximation of the 1-rarefaction wave [158]. In order to prevent this, we apply the Harten–Hyman entropy fix. The idea is to split the jump \( u_* - u_\ell \) travelling at speed \( \lambda_1 \) into two jumps \( u_c - u_\ell \) and \( u_* - u_c \) travelling at speeds \( \lambda_{1,\ell} < 0 \) and \( \lambda_{1,*} > 0 \), respectively. The intermediate state \( u_c \) is referred to as the critical state and is as yet unknown. Note that the single jump \( u_* - u_\ell \) can be written as

\[
u_* - u_\ell = \gamma_1 \tilde{s}_1. \tag{14.97}
\]

Integrating the linear conservation law (14.49a) over a sufficiently large interval, we can derive the following relation between the three jumps:

\[
\tilde{\lambda}_1 (u_* - u_\ell) = \lambda_{1,\ell} (u_c - u_\ell) + \lambda_{1,*} (u_* - u_c). \tag{14.98}
\]

From (14.97) and (14.98) we obtain

\[
u_c - u_\ell = \frac{\lambda_{1,*} - \tilde{\lambda}_1}{\lambda_{1,*} - \lambda_{1,\ell}} \gamma_1 \tilde{s}_1. \tag{14.99}
\]

In the first expression for the Roe numerical flux in (14.57) we only take into account the wave due to the jump \( u_c - u_\ell \) propagating with negative speed \( \lambda_{1,\ell} \). Consequently, the Roe flux reduces to

\[
F_R(u_\ell, u_\ell) = f(u_\ell) + \frac{\lambda_{1,*} - \tilde{\lambda}_1}{\lambda_{1,*} - \lambda_{1,\ell}} \gamma_1 \tilde{s}_1 \tag{14.100a}
\]

or, in a more condensed form,

\[
F_R(u_\ell, u_\ell) = f(u_\ell) + \gamma_1 \tilde{s}_1 \tilde{\lambda}_1, \quad \tilde{\lambda}_1 := \frac{\lambda_{1,*} - \tilde{\lambda}_1}{\lambda_{1,*} - \lambda_{1,\ell}}. \tag{14.100b}
\]

To conclude we need an estimate for the intermediate state \( u_* \) in order to compute \( \lambda_{1,*} \). One possibility is to compute \( u_* \) from (14.97). Other possibilities are to use approximate Riemann solvers; see, e.g., [159] for more details. The case of a sonic 2-rarefaction wave can be handled in a similar manner; we will not discuss this here.
14.6 Numerical Solution of the Wave Equation

In this section we discuss numerical solution methods for the following initial boundary value problem for the wave equation:

\[
\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0, \quad (14.101a)
\]

\[
u(x, 0) = v(x), \quad x \in (0, 1), \quad (14.101b)
\]

\[
\frac{\partial u}{\partial t}(x, 0) = w(x), \quad x \in (0, 1), \quad (14.101c)
\]

\[
u(0, t) = g_\ell(t), \quad t \geq 0, \quad (14.101d)
\]

\[
u(1, t) = g_r(t), \quad t \geq 0. \quad (14.101e)
\]

As we have seen in Section 12.6, we can reduce (14.101a) to a first order system. As such it can be dealt with by the methods we have encountered so far in this chapter. However, we will consider two alternative approaches here. The first one is based on direct discretisation of the wave equation by central differences and is discussed in Section 14.6.1. The second alternative employs the leapfrog scheme for the equivalent first order system, but on a staggered grid, and is the topic of Section 14.6.2.

14.6.1 Difference Methods Based on the Scalar Form

The first method we discuss is based on using central differences for the second order time and space derivatives. Hence let us assume that we have an equispaced grid covering \( \Omega := [0, 1] \times [0, \infty) \) as follows:

\[
x_j := j \Delta x \quad (j = 0, 1, 2, \ldots, M + 1), \quad t^n := n \Delta t \quad (n = 0, 1, 2, \ldots),
\]

with \( \Delta x = 1/(M + 1) \) the (spatial) grid size and \( \Delta t > 0 \) the time step. Let \( u^n_j \) denote the numerical approximation of \( u(x_j, t^n) \). An obvious way to combine the central difference approximations of both derivatives is to centre them around the grid point \( (x_j, t^n) \), leading to the scheme

\[
\frac{1}{\Delta t^2} (u_{j+1}^{n+1} - 2u_j^n + u_{j-1}^n) = \frac{a^2}{\Delta x^2} (u_{j+1}^n - 2u_j^n + u_{j-1}^n) \quad (14.102a)
\]

or, equivalently,

\[
u_{j+1}^{n+1} = 2(1 - c^2)u_j^n + c^2(u_{j+1}^n + u_{j-1}^n) - u_{j}^{n-1}, \quad (14.102b)
\]

where the parameter \( c \) is defined by

\[
c := \frac{a \Delta t}{\Delta x} > 0. \quad (14.103)
\]
14.6. Numerical Solution of the Wave Equation

The stencil of scheme (14.102b) is shown in Figure 14.4. In order to compute the numerical values $u^n_j$ ($n = 2, 3, \ldots$), we need a second “initial” value $u^0_j$, besides $u^0_0$, for which we use (14.101c). To that end we expand $u(x_j, t^1)$ in a Taylor series to obtain

$$u(x_j, t^1) = u(x_j, 0) + \Delta t \frac{\partial u}{\partial t}(x_j, 0) + \frac{1}{2} \Delta t^2 \frac{\partial^2 u}{\partial t^2}(x_j, 0) + \mathcal{O}(\Delta t^3). \quad (14.104)$$

Substituting $\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$, approximating the latter derivative by central differences, and applying the initial conditions, we obtain

$$u(x_j, t^1) = v(x_j) + \Delta t w(x_j) + \frac{1}{2} c^2 (v(x_{j+1}) - 2v(x_j) + v(x_{j-1})) + \mathcal{O}(\Delta t^3) + \mathcal{O}((\Delta t \Delta x)^2). \quad (14.105)$$

From (14.105) we easily deduce the scheme

$$u^1_j = (1 - c^2)v_j + \frac{1}{2} c^2 (v_{j+1} + v_{j-1}) + \Delta t w_j \quad (14.106)$$

for $u^1_j$, where we have introduced the obvious notation $v_j = v(x_j)$, etc.

Writing the unknown values $u^n_j$ ($j = 1, 2, \ldots, M$) as a vector $\mathbf{u}^n$, we obtain from (14.102b) the \textit{(three-term)} recursion

$$\mathbf{u}^{n+1} - (2\mathbf{I} + (a\Delta t)^2 \mathbf{A})\mathbf{u}^n + \mathbf{u}^{n-1} = (a\Delta t)^2 \mathbf{f}^n, \quad (14.107)$$

\textbf{Figure 14.4. Stencil for the centred scheme (14.102b).}
where the discretisation matrix $A$ is defined by

$$A := \frac{1}{\Delta x^2} \begin{pmatrix} -2 & 1 \\ 1 & -2 & 1 \\ & \ddots & \ddots \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}$$

and the vector $f^n$ represents the contributions of the boundary data. We encountered the same discretisation matrix in Section 11.2.1. Recall that $A$ is similar to a diagonal matrix; i.e.,

$$W^{-1}AW = \Lambda,$$

where $\Lambda$ and $W$ are defined by

$$\Lambda := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M), \quad W := (w_1, w_2, \ldots, w_M),$$

with $\lambda_k$ and $w_k$ the eigenvalues and corresponding eigenvectors of $A$. The eigenvalues $\lambda_k$ are given by

$$\lambda_k = -\frac{4}{\Delta x^2} \sin^2\left(\frac{\pi}{2} x_k \right) \quad (k = 1, 2, \ldots, M)$$

and will be needed in the following stability analysis.

To investigate stability we perform the variable transformation

$$\tilde{u}^n := W^{-1}u^n.$$

Inserting (14.110) into (14.107), we obtain the decoupled system

$$\tilde{u}^{n+1} = (2I + (a/\Delta t)^2 \Lambda)\tilde{u}^n + \tilde{u}^{n-1} = (a/\Delta t)^2 \tilde{f}^n,$$

with $\tilde{f}^n := W^{-1}f^n$. The recursion in (14.111) actually consists of $M$ decoupled recursions; i.e.,

$$\tilde{u}_k^{n+1} - (2 + (a/\Delta t)^2 \lambda_k)\tilde{u}_k^n + \tilde{u}_k^{n-1} = (a/\Delta t)^2 \tilde{f}_k^n \quad (k = 1, 2, \ldots, M).$$

The homogeneous recursion determines the stability of the scheme (14.102a) and has a solution of the form

$$\tilde{u}_k^n = A_k \mu_{k,1}^n + B_k \mu_{k,2}^n, \quad A_k, B_k \in \mathbb{R},$$

where $\mu_{k,1}$ and $\mu_{k,2}$ are the roots of the characteristic polynomial

$$p_k(\mu) := \mu^2 - (2 + (a/\Delta t)^2 \lambda_k)\mu + 1.$$

Clearly the method is stable if $|\mu_{k,1}|, |\mu_{k,2}| \leq 1$. The stability condition is given in the next property.

**Property 14.6.** The centred scheme (14.102a) is stable if $c \leq 1$. 
14.6. Numerical Solution of the Wave Equation 443

**Proof.** We note that the product $\mu_{k,1}\mu_{k,2}$ must be equal to one. For ease of notation we introduce the variable $q_k := 2 + (a/Delta_1t)^2\lambda_k = 2 - 4c^2 \sin^2(\pi x_k)$. Obviously, the solutions of (14.113b) are given by

$$\mu_{k,l} = \frac{1}{2} q_k \pm \sqrt{\left(\frac{1}{2} q_k\right)^2 - 1} \quad (l = 1, 2).$$

Hence if $|q_k| \leq 2$, we have two complex conjugate roots with a product of one. This implies that each root has modulus one. If, on the other hand, $|q_k| > 2$, we always have a real root of modulus larger than one. It is simple to check that $c \leq 1$ implies that $|q_k| \leq 2$.

Note that the result of Property 14.6 also follows from the CFL condition.

As usual, the local discretisation error is the residual after substituting the exact solution of (14.101a) in (14.102a), i.e.,

$$d^n_i := \frac{1}{\Delta t^2} \left( a(x_j, t^n+1) - 2u(x_j, t^n) + u(x_j, t^{n-1}) \right)$$

$$- \frac{a^2}{\Delta x^2} \left( u(x_{j+1}, t^n) - 2u(x_j, t^n) + u(x_{j-1}, t^n) \right). \tag{14.114}$$

Expanding all function values in a Taylor series around $(x_j, t^n)$ and assuming that $u(x,t)$ is sufficiently smooth, we find

$$d^n_i = \frac{1}{12} \left( \Delta t^2 \frac{\partial^4 u}{\partial t^4}(x_j, t^n) - a^2 \Delta x^2 \frac{\partial^4 u}{\partial x^4}(x_j, t^n) \right) + O(\Delta t^4) + O(\Delta x^4) \tag{14.115}$$

where we have used that $\frac{\partial^4}{\partial t^4} u = a^4 \frac{\partial^4}{\partial x^4} u$. Clearly, scheme (14.102a) is second order accurate in $\Delta t$ and $\Delta x$.

Assuming that the boundary conditions are exact, we can derive the following recursion for the global discretisation error vector $\mathbf{e}^n$:

$$\mathbf{e}^{n+1} - \left( 2I + (a/Delta_1t)^2 \mathbf{A} \right) \mathbf{e}^n + \mathbf{e}^{n-1} = \Delta t^2 \mathbf{d}^n. \tag{14.116}$$

Analogously to the derivation of (14.112), we find the componentwise error recursion (in an obvious notation)

$$\tilde{e}^{n+1}_k - (2 + (a/Delta_1t)^2 \lambda_k) \tilde{e}^n_k + \tilde{e}^{n-1}_k = \Delta t^2 \tilde{d}^n_k \quad (k = 1, 2, \ldots, M). \tag{14.117}$$

Using (E.7), we find

$$|\tilde{e}^n_k| \leq \Delta t^2 \frac{\sum_{l=1}^{n-1} |\mu_{k,2}^{n-l} - \mu_{k,1}^{n-l}|}{|\mu_{k,2} - \mu_{k,1}|} \max_{1 \leq l \leq n-1} |\tilde{d}^l_k|. \tag{14.118}$$

Let us now assume that $0 \leq t \leq T$ for some $T > 0$. Since $|\mu_{k,1}| = |\mu_{k,2}| = 1$, we find for $n \Delta t \leq T$ that

$$\sum_{l=1}^{n-1} |\mu_{k,2}^{n-l} - \mu_{k,1}^{n-l}| \leq 2(n-1) \leq 2T/\Delta t.$$
For the denominator in (14.118) we have
\[ |\mu_{k,2} - \mu_{k,1}| = 4c \sin \left( \frac{\pi}{2} x_k \right) \sqrt{1 - c^2 \sin^2 \left( \frac{\pi}{2} x_k \right)}. \]

It is simple to check that its minimum value is \( O(\Delta t) \). Hence we obtain
\[ |\tilde{e}_k^n| \leq C \max_{1 \leq \ell \leq n-1} |\tilde{d}_{k,\ell}| \]
for some \( C > 0 \). From this inequality and the relation \( \Delta t = O(\Delta x) \) we eventually obtain the infinity norm estimate
\[ \|e\|_\infty = O(\Delta x^{3/2}). \tag{14.119} \]

Of course, one can use other methods to compute solutions of the wave equation. In particular, in view of the linearity, implicit methods may be attractive to avoid time step constraints. As an example, consider the central difference spatial discretisation at three time levels (a kind of generalisation of Crank–Nicolson). This leads to the scheme
\[
\frac{1}{\Delta t^2} \left( u_j^{n+1} - 2u_j^n + u_j^{n-1} \right) = a^2 \left( \frac{1}{\Delta x^2} (u_{j+1}^{n-1} - 2u_j^{n-1} + u_{j-1}^{n-1}) \right) + \frac{2}{\Delta x^2} (u_{j+1}^{n-1} - 2u_j^{n-1} + u_{j-1}^{n-1}) + \frac{1}{\Delta x^2} (u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}). \tag{14.120}
\]

This results in the matrix recursion
\[
\left( I - \frac{1}{4}(a \Delta t)^2 A \right) u_j^{n+1} - 2 \left( I + \frac{1}{4}(a \Delta t)^2 A \right) u_j^n + \left( I - \frac{1}{4}(a \Delta t)^2 A \right) u_j^{n-1} = (a \Delta t)^2 f^*, \tag{14.121}
\]
for \( u_j \), where \( f^* \) is the vector containing boundary data at all three time levels involved. After transforming this to diagonal form, as we did in (14.111), we find out that in this case the basis solutions of the thus found decoupled recursions are determined by the roots of the characteristic polynomial
\[
p_k(\mu) := \mu^2 - 2 + \frac{1}{4}(a \Delta t)^2 \lambda_k \frac{1}{1 - \frac{1}{2}(a \Delta t)^2 \lambda_k} + 1. \tag{14.122}
\]

Clearly, we have two complex conjugate roots \( \mu_{k,1} \) and \( \mu_{k,2} \) with a product of one. Hence \( |\mu_{k,1}| = |\mu_{k,2}| = 1 \). Since \( \mu_{k,1} \neq \mu_{k,2} \), i.e., the eigenvalues are simple, stability is guaranteed.

### 14.6.2 The Staggered Leapfrog Scheme

We can rewrite equation (14.101a) as a system using the new variables
\[
p := \frac{1}{a} \frac{\partial u}{\partial t}, \quad q := \frac{\partial u}{\partial x};
\]
14.6. Numerical Solution of the Wave Equation

Figure 14.5. Patch of the staggered grid and stencil for the leapfrog scheme (14.124).

cf. Section 12.6. This way, we obtain the first order system

\[
\frac{\partial p}{\partial t} - a \frac{\partial q}{\partial x} = 0, \quad (14.123a)
\]
\[
\frac{\partial q}{\partial t} - a \frac{\partial p}{\partial x} = 0. \quad (14.123b)
\]

As we see from (14.123), we have two coupled equations of identical form, where the time derivative of one variable is coupled to the space derivative of the other. This makes the leapfrog scheme in combination with a staggered grid an attractive choice. Note that we introduced this type of grid in Section 9.7. A patch of such a grid is shown in Figure 14.5. Typically, we take the \( p \) values at the grid points \((x_j, t^n)\), whereas we take the \( q \) values at the shifted points \((x_{j+1/2}, t^{n+1/2})\). The resulting leapfrog scheme then reads

\[
p_j^{n+1} = p_j^n + c(q_j^{n+1/2} - q_j^{n-1/2}), \quad (14.124a)
\]
\[
q_j^{n+3/2} = q_j^{n+1/2} + c(p_j^{n+1} - p_j^n). \quad (14.124b)
\]

see (13.70b). We can interpret this scheme as a one-step method for the vector \((p_j^n, q_j^{n+1/2})^T\).

Next we will investigate the stability of (14.124). Let us now assume that we have periodic boundary conditions. This enables us to carry out a Fourier mode analysis. We find, upon substituting the Fourier modes

\[
p_j^n = \hat{p} \lambda^n e^{ijv}, \quad q_j^{n+1/2} = \hat{q} \lambda^{n+1/2} e^{i(j+1/2)v},
\]
the dispersion relation

\[
\begin{pmatrix}
\lambda - 1 & -2i c \sin \left( \frac{1}{2} \phi \right) \\
-2i c \sin \left( \frac{1}{2} \phi \right) & \lambda - 1
\end{pmatrix}
\begin{pmatrix}
\hat{p} \\
\sqrt{\lambda} \hat{q}
\end{pmatrix} = 0.
\] (14.125)

Note that this dispersion relation is more complicated than we have seen before. It is implicit but also vectorial. Since the vector \((\hat{p}, \sqrt{\lambda} \hat{q})^T\) is nonzero, the coefficient matrix in (14.125) has to be singular. Taking the determinant, we find

\[
(\lambda - 1)^2 + 4c^2 \sin^2 \left( \frac{1}{2} \phi \right) \lambda = 0.
\] (14.126)

Apparently, the product of the roots of this equation is one. If the discriminant of (14.126) is positive, we have at least one root being larger than one in modulus. On the other hand, if the discriminant is negative, we have two complex, unimodular roots. It is easy to see that this is the case if \(c \leq 1\); cf. the proof of Property 14.6. For a stability analysis in the case of nonperiodic boundaries we refer to the next section.

Once we have determined \(p = \frac{1}{2} \frac{\partial u}{\partial t}\) and \(q = \frac{\partial u}{\partial x}\) numerically, we can formally integrate either one to obtain \(u\). There is, however, a more direct way to find \(u\). Indeed, let \(r\) be an auxiliary variable such that

\[
\begin{align*}
\frac{\partial u}{\partial t} - a \frac{\partial r}{\partial x} &= 0, \quad (14.127a) \\
\frac{\partial r}{\partial t} - a \frac{\partial u}{\partial x} &= 0. \quad (14.127b)
\end{align*}
\]

One can simply check that \(u\) satisfies the wave equation (14.101a). In other words, the leapfrog method can also be applied to system (14.127) directly. As for the initial conditions, we directly find from (14.101b) the value of \(u_j^0\). In order to find the initial value \(r_j^{1/2}\), we have to use system (14.127) and the initial conditions (14.101b) and (14.101c). First, integrating (14.127a) and applying initial condition (14.101c), we find

\[
\begin{align*}
&\frac{\partial r}{\partial t} (x_j^{1/2}, 0) = a v' (x_j^{1/2}) = a \Delta x (v(x_{j+1}) - v(x_j)).
\end{align*}
\] (14.129)

Finally, combining the Taylor expansion

\[
\begin{align*}
r \left( x_j^{1/2}, \frac{1}{2} \Delta t \right) &= r(x_j^{1/2}, 0) + \frac{1}{2} \Delta t \frac{\partial r}{\partial t} (x_j^{1/2}, 0) + \mathcal{O}(\Delta t^2) \quad (14.130)
\end{align*}
\]
14.7. Numerical Boundary Conditions

with (14.128) and (14.129), we find the following approximation of $r_{1/2}^{j+1/2}$:

$$r_{1/2}^{j+1/2} = \frac{1}{a} Q(w)(x_{j+1/2}) + \frac{c}{2} \left( v(x_{j+1}) - v(x_{j}) \right),$$  \hspace{1cm} (14.131)

where $Q(w)(x_{j+1/2})$ is some quadrature rule for the integral in (14.128).

There is an interesting observation to be made from the alternative formulation (14.127) of the wave equation. If we apply the leapfrog scheme (14.124) to (14.127) and take a one-sided backward numerical time derivative of (14.124a), we obtain

$$u_{j}^{n+1} - 2u_{j}^{n} + u_{j}^{n-1} = c \left( r_{j+1/2}^{n+1/2} - r_{j-1/2}^{n+1/2} - r_{j+1/2}^{n-1/2} + r_{j-1/2}^{n-1/2} \right).$$  \hspace{1cm} (14.132)

Likewise, if we take a one-sided backward numerical space derivative of (14.124b) and replace $n$ by $n - 1$, we find

$$r_{j+1/2}^{n+1/2} - r_{j+1/2}^{n-1/2} - r_{j-1/2}^{n+1/2} + r_{j-1/2}^{n-1/2} = c \left( u_{j+1}^{n} - 2u_{j}^{n} + u_{j-1}^{n} \right).$$  \hspace{1cm} (14.133)

Combining (14.132) and (14.133), we retrieve the central difference scheme (14.102b)!

### 14.7 Numerical Boundary Conditions

An important question is how to deal with boundary conditions. We will not cover this subject in full detail; instead we will discuss a few selected topics. First, we introduce nonreflecting and numerical boundary conditions for two model problems: the system formulation of the wave equation and the shallow-water equations. Next we investigate the impact of boundary conditions on the behaviour of a numerical solution. In particular, we discuss parasitic solutions of a multistep scheme and we conclude with a stability analysis of the leapfrog scheme.

Hyperbolic systems are often defined on a very large or even infinite spatial domain. In order to make a numerical computation feasible, we have to cut off this domain at some point. At this point we want to impose an artificial boundary condition such that the numerical solution on the smaller domain agrees well with the numerical solution on the larger domain, assuming that there are no incoming waves at this point. Moreover, we almost never have enough physical boundary conditions to carry out a numerical computation. Therefore we have to derive additional numerical boundary conditions. We will illustrate these two issues for an initial boundary value problem of the form

$$\frac{\partial u}{\partial t} + B(u) \frac{\partial u}{\partial x} = 0, \quad x \in (0, \infty), \quad t > 0,$$  \hspace{1cm} (14.134a)

$$u(x, 0) = 0, \quad x \in (0, \infty),$$  \hspace{1cm} (14.134b)

$$C_{\ell} u(0, t) = g_{\ell}(t), \quad t > 0.$$  \hspace{1cm} (14.134c)

Note that the initial condition $u(x, 0) = 0$ implies the radiation condition of no incoming waves from infinity. In the following we take for the hyperbolic system (14.134a) either the linear system from Example 14.1 or the shallow-water equations. In both cases the boundary conditions depend on the characteristic variable $\tilde{u}$. The solution of (14.134) is a right-travelling wave generated by the boundary condition (14.134c) at $x = 0$. 
First, consider the linear system from Example 14.1. The characteristic variable \( \tilde{u} := S^{-1} u \), with \( S \) the (right) eigenvector matrix, is given by

\[
\tilde{u}_1 = \frac{1}{2} (u_1 + u_2), \quad \tilde{u}_2 = \frac{1}{2} (u_1 - u_2).
\]

These variables satisfy the scalar equations

\[
\frac{\partial \tilde{u}_1}{\partial t} - a \frac{\partial \tilde{u}_1}{\partial x} = 0, \quad \frac{\partial \tilde{u}_2}{\partial t} + a \frac{\partial \tilde{u}_2}{\partial x} = 0,
\]

implying that \( \tilde{u}_1 \) is a left-going wave and \( \tilde{u}_2 \) is a right-going wave. Suppose we cut off the spatial domain at \( x = L \) for some \( L > 0 \). At \( x = L \) we impose the condition \( \tilde{u}_1(L,t) = 0 \); i.e., there is no incoming wave at \( x = L \). This gives the absorbing or nonreflecting boundary condition

\[
u_1(L,t) + u_2(L,t) = 0, \quad t > 0.
\]

According to the theory in Section 12.7, this is an admissible boundary condition. The numerical approximation of (14.137a) at time level \( t^{n+1} \) is straightforward and is given by

\[
u_1^{n+1}_{M+1} + u_2^{n+1}_{M+1} = 0,
\]

assuming \([0, L]\) is covered with a uniform (spatial) grid as follows:

\[
x_j = j \Delta x \quad (j = 0, 1, 2, \ldots, M + 1), \quad \Delta x := L/(M + 1).
\]

Most schemes for (14.134a) require two conditions at the boundary point \( x_{M+1} = L \). Therefore we have to complete the boundary condition (14.137b) with a numerical boundary condition. One way to determine such a numerical boundary condition is to use that \( \tilde{u}_2 \) is constant along the \( C_2 \) characteristic through the grid point \((x_{M+1}, t^{n+1})\); see Figure 14.6. If the stability condition \( c := a \Delta t / \Delta x \leq 1 \) is satisfied, this characteristic intersects the grid

\[
C_2 : x - at = \text{Constant}
\]

Figure 14.6. The \( C_2 \) characteristic through the boundary grid point \((x_{M+1}, t^{n+1})\).
14.7. Numerical Boundary Conditions

Line \( t = t^n \) between the grid points \((x_M, t^n)\) and \((x_{M+1}, t^n)\). Applying linear interpolation, we find

\[
\tilde{u}^{n+1}_{2,M+1} = c\tilde{u}^n_{2,M} + (1-c)\tilde{u}^n_{2,M+1},
\]

which is just the upwind scheme applied to the advection equation for the characteristic variable \( \tilde{u}_2 \) in (14.136) at grid point \((x_{M+1}, t^n)\). In terms of the original variables we find the numerical boundary condition

\[
u_{1,M+1}^n - u_{2,M+1}^{n+1} = c(u_{1,M+1}^n - u_{2,M}^n) + (1-c)(u_{1,M+1}^n - u_{2,M+1}^n), \tag{14.138}
\]

and together with (14.137b) this is sufficient to determine both \( u_1 \) and \( u_2 \) at the boundary \( x = L \).

Next suppose system (14.134a) represents the shallow-water equations. In this case the characteristic variables are defined by the Pfaffian differential equation \( d\tilde{u} = S^{-1}d\mathbf{u} \). In Example 12.33 we saw that

\[
\tilde{u}_1 = u - 2c, \quad \tilde{u}_2 = u + 2c, \quad c := \sqrt{\phi}, \tag{14.139}
\]

and these variables satisfy the diagonalised system

\[
\frac{\partial \tilde{u}_1}{\partial t} + (u - c)\frac{\partial \tilde{u}_1}{\partial x} = 0, \quad \frac{\partial \tilde{u}_2}{\partial t} + (u + c)\frac{\partial \tilde{u}_2}{\partial x} = 0. \tag{14.140}
\]

Assume the in/outflow at the artificial boundary \( x = L \) is subcritical, i.e., \(|u| < c\). Then we have \( \lambda_1(u) < 0 < \lambda_2(u) \), implying that we have to prescribe one physical boundary condition and derive one numerical boundary condition at \( x = L \); see Section 12.7. For the physical boundary condition we take again the nonreflecting boundary condition. In this case it reads: the amplitude of the incoming wave \( \tilde{u}_1 \) is constant in time [146]. This means that the incoming wave \( \tilde{u}_1 \) vanishes, since it is generated by changes in amplitude. Thus we have

\[
\frac{\partial \tilde{u}_1}{\partial t}(L,t) + 2\frac{\partial c}{\partial t}(L,t) = 0, \quad t > 0. \tag{14.141a}
\]

The numerical approximation is straightforward, and we find

\[
u_{1,M+1}^n - 2c_{M+1}^n = u_{M+1}^n - 2c_{M+1}^n, \tag{14.141b}
\]

The derivation of the numerical boundary condition is again based on the observation that \( \tilde{u}_2 \) is constant along the \( C_2 \) characteristic through \((x_{M+1}, t^n+1)\). One approach is then to apply linear interpolation in order to express \( \tilde{u}_2^{n+1} \) in terms of the values \( \tilde{u}_2^n \) and \( \tilde{u}_2^n \). Alternatively, we can apply the upwind scheme to the scalar equation for \( \tilde{u}_2 \) in (14.140). This way we obtain

\[
u_{M+1}^{n+1} + 2c_{M+1}^{n+1} = u_{M+1}^n + 2c_{M+1}^n - \frac{\Delta t}{\Delta x} (u_{M+1}^n + c_{M+1}^n)(u_{M+1}^n + 2c_{M+1}^n - u_M^n - 2c_M^n). \tag{14.142}
\]

From (14.141b) and (14.142) we can determine \( u_{M+1}^{n+1} \) and \( c_{M+1}^{n+1} \) and subsequently we can compute \( \phi_{M+1}^{n+1} \).
Example 14.7 We compute tidal waves in a canal, which we model by the shallow-water equations. We assume that at $x = 0$ the mouth of the canal enters the sea, where we prescribe the periodic boundary condition

$$\varphi(0, t) = g(h_s + H \sin \omega t),$$

with $H = 0.1h_s = 0.2$ and $\omega = 8$. At $x = 1$ we put an artificial boundary where we prescribe the nonreflecting and numerical boundary conditions (14.141b) and (14.142), respectively. We have applied the Roe scheme to compute a numerical solution. Numerical parameters are $\Delta x = 2.5 \times 10^{-2}$ and $\Delta t = 4 \times 10^{-3}$. The results are shown in Figure 14.7. Clearly, the wave generated by the boundary condition at $x = 0$ leaves the domain at the artificial boundary $x = 1$.

![Figure 14.7](image-url)

Figure 14.7. Numerical approximation of the geopotential $\varphi = gh$ at time levels 0.3, 0.4, ... , 0.8.
14.7. Numerical Boundary Conditions

undisturbed and does not produce spurious reflections. Moreover, we notice that the amplitude of the solution is slightly damped due to the dissipative character of the Roe scheme.

A related question is the occurrence of parasitic solutions of a multistep scheme for (14.134a). These parasitic solutions should be properly controlled by appropriate boundary conditions. We will work this out for the leapfrog scheme, which is a two-step difference scheme; cf. Section 13.4.2. Applying the leapfrog scheme, we obtain

\[ u_j^{n+1} = -\frac{\Delta t}{\Delta x} B(u_{j+1}^n - u_{j-1}^n) + u_j^{n-1}. \]  

(14.143)

Inserting the planar wave solution

\[ u_j^n = s e^{i(x_j - \omega t)}, \quad \lambda := e^{-i\omega \Delta t}, \quad \mu := e^{i\kappa \Delta x}, \]  

(14.144)

for some vector \( s \neq 0 \), we find the relation

\[ \left( \frac{\lambda - 1}{\lambda} \right) s = -\frac{\Delta t}{\Delta x} \left( \frac{\mu - 1}{\mu} \right) Bs. \]

This relation is actually an eigenvalue problem. Indeed, assuming \( \mu^2 \neq 1 \), we have

\[ Bs = -\frac{\Delta x}{\Delta t} \frac{\mu(\lambda^2 - 1)}{\lambda(\mu^2 - 1)} s. \]  

(14.145)

Using the eigenvalues of \( B \), we can compute \( \lambda \) as a function of \( \mu \) or vice versa.

To make the subsequent analysis more transparent, we take for \( B \) the matrix

\[ B = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix} \]

occurring in Example 14.1 on small-amplitude waves. Clearly its eigenvalues\(^1\) \( \nu_k \) and eigenvectors \( s_k \) are given by

\[ \nu_1 = -a < 0, \quad \nu_2 = a > 0, \quad s_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \]

In this case the solution of (14.134a) consists of left- and right-travelling waves and is of the form

\[ u(x,t) = \hat{w}_1(x + at)s_1 + \hat{w}_2(x - at)s_2 \]

for arbitrary continuously differentiable functions \( \hat{w}_1(\xi) \) and \( \hat{w}_2(\xi) \). The first component typically has planar wave modes with dispersion relation \( \omega(\kappa) = -a\kappa \), whereas for the second component we have \( \omega(\kappa) = a\kappa \).

From the eigenvalue problem (14.145) we conclude that

\[ \mu(\lambda^2 - 1) = \pm c\lambda(\mu^2 - 1), \]  

(14.146)

\(^1\)Here we use \( \nu_k \) to denote the eigenvalues of \( B \) instead of \( \lambda_k \) in order to avoid confusion with the amplification factor \( \lambda = e^{-i\omega \Delta t} \).
with $c := a \Delta t / \Delta x$. For given wave number $\kappa$, or equivalently for given $\mu$, we can determine the amplification factor $\lambda$. In particular, for the plus sign in (14.146), corresponding to $v_1 = -a$, we obtain the quadratic equation

$$\lambda^2 - \alpha \lambda - 1 = 0, \quad \alpha := \frac{c \mu^2 - 1}{\mu},$$  \hspace{1cm} (14.147a)

assuming $\mu \neq 0$. This equation has the solutions

$$\lambda_{+,1} = \frac{1}{2} \alpha + \sqrt{1 + \frac{1}{4} \alpha^2}, \quad \lambda_{+,2} = \frac{1}{2} \alpha - \sqrt{1 + \frac{1}{4} \alpha^2}. \hspace{1cm} (14.147b)$$

For small $\omega \Delta t$ and $\kappa \Delta x$ we find that $\alpha \approx 2i a \kappa \Delta t$. Clearly, $\lambda_{+,1}$ is the essential root and we have

$$\lambda_{+,1} = e^{-i\omega \Delta t} \pm i a \kappa \Delta t \approx e^{i\omega \Delta t},$$  \hspace{1cm} (14.148)

and the corresponding dispersion relation reads $\omega(\kappa) = -a \kappa$ for a left-going wave. The other root $\lambda_{+,2}$ does not belong to a physical mode; it is parasitic (see Section 5.4). Indeed, we find

$$e^{-i\omega \Delta t} \pm i a \kappa \Delta t \approx -e^{-i\omega \Delta t};$$  \hspace{1cm} (14.149)

i.e., $\omega(\kappa) = a \kappa - \pi / \Delta t$. This would correspond to a mode that changes sign each time step and travels in the wrong direction. In the same way, when we take the minus sign in (14.146), corresponding to $v_2 = a$, we obtain the relation

$$\lambda^2 + \alpha \lambda - 1 = 0.$$  \hspace{1cm} (14.150a)

It roots are

$$\lambda_{-,1} = -\frac{1}{2} \alpha + \sqrt{1 + \frac{1}{4} \alpha^2}, \quad \lambda_{-,2} = -\frac{1}{2} \alpha - \sqrt{1 + \frac{1}{4} \alpha^2}. \hspace{1cm} (14.150b)$$

In this case $\lambda_{-,1}$ is the essential root corresponding to a right-going mode, while $\lambda_{-,2}$ corresponds to a parasitic mode. Observe that $\lambda_{+,1} \lambda_{-,1} = 1$. Thus we have four linearly independent solutions of the form (14.144). The general solution of (14.134a) is a linear combination of these and reads

$$u^n_j = \left( C_1 \lambda_{+,1}^n s_1 + C_2 \lambda_{+,2}^n s_1 + C_3 \lambda_{-,1}^n s_2 + C_4 \lambda_{-,2}^n s_2 \right) \mu^j \hspace{1cm} (14.151)$$

for arbitrary constants $C_l$ ($l = 1, 2, 3, 4$). The second and fourth terms in (14.151) are parasitic modes.

Because of symmetry with respect to $\mu$ and $\lambda$, we may as well express $\mu$ in terms of $\lambda$. Completely analogously to the previous derivation, we obtain for $\mu$ the quadratic equations

$$\mu^2 \mp \beta \mu - 1 = 0, \quad \beta := \frac{\lambda^2 - 1}{c \lambda},$$  \hspace{1cm} (14.152)

the roots of which are given by

$$\mu_{+,1} = \frac{1}{2} \beta + \sqrt{1 + \frac{1}{4} \beta^2}, \quad \mu_{+,2} = \frac{1}{2} \beta - \sqrt{1 + \frac{1}{4} \beta^2};$$  \hspace{1cm} (14.153a)

$$\mu_{-,1} = -\frac{1}{2} \beta + \sqrt{1 + \frac{1}{4} \beta^2}, \quad \mu_{-,2} = -\frac{1}{2} \beta - \sqrt{1 + \frac{1}{4} \beta^2}. \hspace{1cm} (14.153b)$$
The essential roots are $\mu_{+,1}$ and $\mu_{-,1}$, and, furthermore, $\mu_{+,1}\mu_{-,1} = 1$. The general solution of (14.134a) can thus be rewritten as

$$u_j^a = \lambda^a \left(D_1 \mu_{+,1}^j s_1 + D_2 \mu_{+,2}^j s_1 + D_3 \mu_{-,1}^j s_2 + D_4 \mu_{-,2}^j s_2 \right) \quad (14.154)$$

for arbitrary coefficients $D_l$ ($l = 1, 2, 3, 4$). Also, in this representation, the second and fourth terms are parasitic. Suppose we prescribe a physical boundary condition at $x = L$ of the form

$$C_l u(L, t) = 0, \quad t > 0, \quad (14.155)$$

with $C_l$ a $1 \times 2$ matrix; cf. Section 12.7. Inserting (14.154) into (14.155), we obtain the relation

$$D_1 \mu_{+,1}^{M+1} C_1 s_1 + D_2 \mu_{+,2}^{M+1} C_1 s_1 + D_3 \mu_{-,1}^{M+1} C_1 s_2 + D_4 \mu_{-,2}^{M+1} C_1 s_2 = 0. \quad (14.156)$$

In either case we see that this boundary condition may annihilate either the left- or right-going parasitic wave; i.e., we can choose $C_l$ such that either $C_l s_1 = 0$ or $C_l s_2 = 0$. Thus one parasitic mode is persistent; i.e., it cannot be removed in this way. However, there is a variety of ways to remove these parasitic waves. One is to choose boundary conditions that include derivatives. Below we give an example.

**Example 14.8** Consider the general solution (14.154) and assume that the left-going parasitic mode is absent; i.e., $D_2 = 0$. In this case we have

$$u_j^a = \left(\begin{array}{c} u_{1,j}^a \\ u_{2,j}^a \end{array}\right) = \lambda^a \left[ D_1 \mu_{+,1}^j \begin{pmatrix} 1 \\ 1 \end{pmatrix} + D_2 \mu_{+,2}^j \begin{pmatrix} 1 \\ 1 \end{pmatrix} + D_3 \mu_{-,1}^j \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right].$$

We want to prescribe boundary conditions at $x_{M+1} = L$ such that the remaining parasitic mode is eliminated; i.e., the boundary conditions should imply $D_2 = 0$. A successful choice for this is

$$u_{1,M+1}^a - u_{1,M}^a = 0, \quad u_{2,M+1}^a + u_{2,M}^a = 0.$$

The latter condition can be interpreted as a "smoothing" approximation of $u_2(L, t) = 0$. Applying these boundary conditions, we obtain the linear system

$$(\mu_{+,1} - 1) \dot{D}_1 + (\mu_{+,2} - 1) \dot{D}_2 = 1 - \mu_{-,1},$$

$$(\mu_{+,1} + 1) \dot{D}_1 + (\mu_{+,2} + 1) \dot{D}_2 = 1 + \mu_{-,1}$$

for $\dot{D}_1 := (\mu_{+,1}/\mu_{-,1})^M D_1/D_2$ and $\dot{D}_2 := (\mu_{+,2}/\mu_{-,1})^M D_2/D_3$. Since $\mu_{+,1}/\mu_{-,1} = 1$ and $\mu_{+,2}/\mu_{-,1} = -1$, this system can be simplified to

$$\left(\frac{1}{\eta} - 1\right) \dot{D}_1 - (\eta + 1) \dot{D}_2 = 1 - \eta,$$

$$\left(\frac{1}{\eta} + 1\right) \dot{D}_1 + (-\eta + 1) \dot{D}_2 = 1 + \eta,$$

with $\eta := \mu_{-,1}$. It is simple to see that the solution is given by $\dot{D}_1 = \eta$ and $\dot{D}_2 = 0$, and so $D_2 = 0$. The latter result is the desired one, showing that no parasitic solutions are being generated.
Finally, we want to investigate the impact of boundary conditions on the stability of a scheme. We consider in particular the staggered leapfrog scheme (14.124) applied to system (14.127). We assume that $u$ satisfies the boundary conditions (14.101d) and (14.101e). Since we know the value of $u$ at the boundaries $x = 0$ and $x = 1$, it is tempting to choose the grid in Figure 14.8, where grid points for $u$ are located on the boundaries. Note that there is one more unknown value of $r$ than of $u$ at every time level, because the $u$ values at the boundary are known. Let the number of unknown $u$ values at each line $t = t^n$ be $M$. Introducing the vectors

$$
\mathbf{u}^n := \begin{pmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_M^n \end{pmatrix}, \quad \mathbf{r}^{n+1/2} := \begin{pmatrix} r_1^{n+1/2} \\ r_2^{n+1/2} \\ \vdots \\ r_M^{n+1/2} \end{pmatrix}
$$

and assuming homogeneous boundary conditions, we see that the scheme (14.124) applied to (14.127) can be written as the matrix-vector recursion

$$
\begin{pmatrix} I & 0 \\ C^T & I \end{pmatrix} \begin{pmatrix} \mathbf{u}^{n+1} \\ \mathbf{r}^{n+3/2} \end{pmatrix} = \begin{pmatrix} I & 0 \\ C & I \end{pmatrix} \begin{pmatrix} \mathbf{u}^n \\ \mathbf{r}^{n+1/2} \end{pmatrix},
$$

(14.157)

where the $M \times (M + 1)$ matrix $C$ is defined by

$$
C := c \begin{pmatrix} -1 & 1 \\ & -1 & 1 \\ & & \ddots & \ddots \\ & & & -1 & 1 \\ & & & & -1 & 1 \end{pmatrix}.
$$

(14.158)

In order to analyse the stability of (14.157), we investigate the eigenvalues of the
14.7. Numerical Boundary Conditions

generalised eigenvalue problem

\[
\begin{pmatrix}
I & 0 \\
C^T & I
\end{pmatrix}
\begin{pmatrix}
u \\
r
\end{pmatrix}
= \lambda
\begin{pmatrix}
I & C \\
0 & I
\end{pmatrix}
\begin{pmatrix}
u \\
r
\end{pmatrix}.
\] (14.159)

We leave it as an exercise to show that \( \lambda \neq 0 \). We find from (14.159) that

\[
(1 - \lambda)u = \lambda Cr.
\] (14.160a)

\[
C^T u = (\lambda - 1)r.
\] (14.160b)

Multiplying (14.160a) by \( C^T \) and substituting \( C^T u \) from (14.160b) yields the following eigenvalue problem for \( r \):

\[
C^T Cr = -\frac{(\lambda - 1)^2}{\lambda} r.
\] (14.161)

Next we readily verify that \( C^T C \) is given by

\[
C^T C = c^2 \begin{pmatrix}
1 & -1 & \\ -1 & 2 & -1 \\ & & \ddots \ddots \ddots \\ & & -1 & 2 & -1 \\ & & & -1 & 1
\end{pmatrix}.
\] (14.162)

Let us write \( \mu := -\lambda^{-1}(\lambda - 1)^2 \). Applying Gershgorin’s theorem (cf. Theorem I.7) to \( C^T C \) (see (14.161)), we conclude that \( 0 \leq \mu \leq 4c^2 \leq 4 \). The last inequality is an immediate consequence of the stability condition \( c \leq 1 \). Moreover, \( C^T C \) has a simple eigenvalue \( \mu = 0 \), with eigenvector \( e := (1, 1, \ldots, 1)^T \). Clearly, \( \lambda \) satisfies

\[
\lambda^2 + (\mu - 2)\lambda + 1 = 0,
\]

the roots of which are complex conjugates and have modulus one. However, the eigenvalue \( \mu = 0 \) implies a double root \( \lambda = 1 \), which turns out to have only one eigenvector \((0^T, e^T)^T\). Hence (14.159), and consequently (14.157), cannot be stable; cf. Definition I.5. We therefore conclude that one cannot use the \( u \) values at both boundaries directly.

Alternatively, if we assume that \( r \) is given at the right boundary \( x = 0 \), we obtain a similar system as (14.157), but now with an \( M \times M \) matrix \( C \) given by

\[
C := c \begin{pmatrix}
1 & \\ -1 & 1 \\ & \ddots \ddots \ddots \\ & & -1 & 1 \\ & & & -1 & 1
\end{pmatrix}.
\] (14.163)

If we carry out a similar analysis, now with this matrix, it will be clear that \( C^T C \) is nonsingular and hence all eigenvalues \( \lambda \) in (14.159) are complex conjugate pairs with nonzero imaginary parts. We leave it as an exercise to show that there are no double eigenvalues. Thus the scheme is stable.
14.8 Discussion

- We have only discussed problems in one space dimension. All methods in this chapter can be extended to several space dimensions by dimensional splitting, i.e., the one-dimensional scheme is applied in each direction; see, e.g., [159, 88].

- Hyperbolic systems often include a source term, describing, e.g., chemical reactions. One approach to solving these systems is to use an operator-splitting method. In such a method the system is split into a “convection part” and a “chemistry part,” which are alternately solved at each time step. For a detailed discussion see, e.g., [88].

- Boundary conditions provide a real source of concern. We have only dealt with this problem briefly. In particular, the treatment of reflective waves has been, and still is, a subject of much research. A good survey paper dealing with radiation boundary conditions (Robin type) is [56]. Another good source for the study of numerical boundary conditions is [72].

Exercises

   (a) Verify the stability condition (14.16). Explain that it is just the CFL condition applied to (14.14).
   (b) Show that the scheme is first order accurate in $\Delta t$ and $\Delta x$.

14.2. Consider the model problem from Exercise 12.9, describing one-dimensional electromagnetic waves.
   (a) Give the upwind scheme (14.14).
   (b) Give the wave propagation form (14.43).
   (c) Elaborate the slope limiter method (14.67).

14.3. Compute the Godunov numerical flux (14.30) for the $p$ system in Exercise 12.10. Then elaborate the Godunov scheme (14.32) and the stability condition (14.33).

14.4. Show that an alternative expression for the Roe numerical flux is given by

$$F_R(u_\ell, u_r) = \frac{1}{2} (f(u_\ell) + f(u_r)) - \frac{1}{2} \tilde{B}^*(u_\ell, u_r)(u_r - u_\ell),$$

where $\tilde{B}^*(u_\ell, u_r)$ is defined analogously to (14.18); i.e., the Roe numerical flux can be interpreted as the central difference flux $\frac{1}{2}(f(u_\ell) + f(u_r))$ stabilized by the diffusive flux $-\frac{1}{2} \tilde{B}^*(u_\ell, u_r)(u_r - u_\ell)$.

14.5. Consider the $p$ system from Exercise 12.10.
   (a) Show that the Roe matrix $\tilde{B}(u_\ell, u_r)$ is given by

$$\tilde{B}(u_\ell, u_r) = \begin{pmatrix} 0 & -1 \\ p(v_\ell) - p(v_r) & v_r - v_\ell \end{pmatrix}.$$
(b) Compute the Roe numerical flux from either expression in (14.57). Then elaborate the Roe scheme (14.44).

c) Give the slope limiter flux (14.75).

14.6. Discuss the Harten–Hyman entropy fix for a sonic 2-rarefaction wave of the shallow-water equations.

14.7. Consider the damped wave equation

\[ \frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0, \quad a > 0. \]

(a) Use central differences to discretise this equation.
(b) Investigate for which values of \( c = a \Delta t / \Delta x \) the scheme is stable.

14.8. For the wave equation (14.101a) we can use a variant of the implicit Euler method by discretising \( \frac{\partial^2 u}{\partial t^2} \) at the time point \( t^{n+1} \).

(a) Write down the resulting scheme.
(b) Show that this scheme is unconditionally stable.
(c) Find the dispersion relation and show that this method is dissipative.

14.9. Consider the nonlinear initial value problem

\[ u^2 \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \quad x \in \mathbb{R}, \quad t > 0, \]

\[ u(x, 0) = 1 + x^2, \quad x \in \mathbb{R}, \]

\[ \frac{\partial u}{\partial t}(x, 0) = 0, \quad x \in \mathbb{R}. \]

Show how to construct an explicit scheme.

14.10. Show that we can associate with the wave equation \( \frac{\partial u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \) the system

\[ \frac{\partial u}{\partial t} \pm a \frac{\partial v}{\partial x} = 0, \]

\[ \frac{\partial v}{\partial t} \pm a \frac{\partial u}{\partial x} = 0, \]

where either the plus or the minus sign is assumed in both equations.

14.11. Show the stability of the recursion (14.107) by writing it as a one-step recursion and investigating the eigenvalues of the matrix involved in the latter.

14.12. Prove that the matrix \( C^T C \) with \( C \) defined in (14.163) has eigenvalues on the unit circle but no double ones.

14.13. Consider the wave equation for \( x, t > 0 \) in the form

\[ \frac{\partial u}{\partial t} + a \frac{\partial v}{\partial x} = 0, \]

\[ \frac{\partial v}{\partial t} + a \frac{\partial u}{\partial x} = 0. \]
(a) Derive the Lax–Wendroff scheme for this system.

(b) Assume we use a special boundary condition similar to the one in Example 14.8, i.e.,

\[ \frac{\partial u}{\partial x}(0, t) = 0, \quad v(0, t) = 0, \quad t > 0. \]

Show that there is no parasitic solution arising from this boundary.
Chapter 15

Perturbation Methods

In this chapter we give a concise introduction to a number of important analytical techniques that use available small parameters in a problem. These methods are called perturbation or asymptotic methods, and follow naturally from the ideas of systematic modeling set forth in Chapter 7. After an introduction, explaining why small parameters occur frequently in practice, we start in Section 15.2 by stating some principles of asymptotic approximations and expansions. Then we continue in Section 15.3 with two methods for solving regular perturbation problems: the method of slow variation and the Lindstedt-Poincaré method. Two of the most important methods for singular perturbation problems are considered in Section 15.4: the method of matched asymptotic expansions (MAEs) and some versions of the method of multiple scales.

15.1 Introduction

We saw in Chapter 7 that a real-world problem can be described by a hierarchy of models such that a higher-level model is more comprehensive and more accurate than one from a lower level. Now suppose that we have a fairly good model describing the dominating phenomena in good order of magnitude. And suppose that we are interested in improving this model by adding some previously ignored aspects or effects. In general, this implies a very abrupt change in our model. The equations are more complex and more difficult to solve. As an illustration, consider the simple model \( x^2 = a^2 \) and the more complete model \( x^2 + \varepsilon x^3 = a^2 \). The first one can easily be solved analytically, the second one, with much more effort, only numerically. So it seems that the relation between solution and model is not continuous in the problem parameters. However small \( \varepsilon \) we take, from a transparent and exact solution of the simple model at \( \varepsilon = 0 \), we abruptly face a far more complicated solution of a model that is just a little bit better. This is a pity, because certain types of useful information (parametric dependencies, trends) become increasingly difficult to dig out of the more complicated solution of the complex model. This discontinuity of models in the parameter \( \varepsilon \) may therefore be an argument for retaining the simpler model.

The (complexity of the) model is, however, only discontinuous if we are merely interested in exact or numerically “exact” solutions (e.g., for reasons of benchmarking or validation of solution methods). This is not always the case. As far as our modeling
objectives are concerned, we have to keep in mind that the improved model is only the next step in the modeling hierarchy and is not exact in any absolute sense. So there is no reason to require the solution to be more exact than the corresponding model, as an exact solution of an approximate model is no better than an approximate solution of an exact model. Moreover, the type of information that analytical solutions may provide (functional relationships, etc.) is sometimes so important that it may be worth sacrificing numerical accuracy.

Let us go back to our “fairly good” improved model. The effects we added are relatively small. Otherwise, the previous lower-level model was not fairly good, as we assumed, but just completely wrong. Usually, this smallness is quantified by small dimensionless parameters occurring in the equations and (or) boundary conditions. This is the generic situation. The transition from a lower-level to a higher-level theory is characterized by the appearance of one or more modeling parameters, which are (when made dimensionless) small or large, and yield in the limit a simpler description. Examples are infinitely large or small geometries with circular or spherical symmetry that reduce the number of spatial dimensions, small amplitudes allowing linearization, low velocities and long time scales in flow problems allowing incompressible description, and small relative viscosity allowing inviscid models. In fact, in any practical problem it is really the rule rather than the exception that dimensionless numbers are either small or large (cf. [76]).

If we accept approximate solutions, where the approximation is based on the inherently small or large modeling parameters, it is possible to gradually increase the complexity of a model and study small but significant effects in the most efficient way. The methods utilizing this approach systematically are called “perturbation methods.” The approximation constructed is almost always asymptotic, i.e., where the error decreases with the small or large parameter.

Usually, a distinction is made between regular and singular perturbations. A (loose definition of a) regular perturbation problem is where the approximate problem is everywhere close to the unperturbed problem. This, however, depends of course on the domain of interest and, as we will see, on the choice of coordinates. If a problem is regular without any need for other than trivial reformulations, the construction of an asymptotic solution is straightforward. In fact, this forms the usual strategy in modeling when terms are linearised or effects are neglected. The more interesting perturbation problems are those where this straightforward approach fails.

We will consider here four methods relevant in the presented modeling problems. The first two are examples of regular perturbation methods, but only after a suitable coordinate transformation. The first is called the method of slow variation, where the typical axial length scale is much greater than the transverse length scale. The second is the Lindstedt–Poincaré method or the method of strained coordinates for periodic processes. Here the intrinsic time scale ( ~ the period of the solution) is unknown and has to be found. The other two methods are of singular perturbation type, because there is no coordinate transformation possible that renders the problem into one of regular type. The third one is the method of Matched Asymptotic Expansions. To render the problem into one of regular type, different scalings are necessary in spatially distinct regions (boundary layers). The fourth singular perturbation method considered here is the method of multiple scales and may be considered as a combination of the method of slow variation and the method of strained coordinates, as now several (long, short, shorter) length scales occur in parallel. This cannot be repaired by
15.2 Asymptotic Approximations and Expansions

Before we can introduce the methods, we have to define our terminology for asymptotic approximations and asymptotic expansions.

15.2.1 Asymptotic Approximations

In order to give a qualitative description of the behaviour of a function \( f(\varepsilon) \) near a point of interest, say \( \varepsilon = 0 \) (equivalent to any other value by a simple translation), we have the order symbols \( O, o, \) and \( O_s \); see the appendix, Section A. Often \( \varepsilon = 0 \) is the lower limit of a parameter range, and we have the tacit assumption that \( \varepsilon \downarrow 0 \).

**Definition 15.1.** \( \phi(\varepsilon) \) is an asymptotic approximation to \( f(\varepsilon) \) as \( \varepsilon \to 0 \) if

\[
 f(\varepsilon) = \phi(\varepsilon) + o(\phi(\varepsilon)) \quad \text{as} \quad \varepsilon \to 0.
\]

This is sometimes more compactly denoted by \( f \sim \phi \).

If \( f \) and \( \phi \) depend on \( x \), this definition remains valid pointwise, i.e., for \( x \) fixed. It is, however, useful to extend the definition to uniformly valid approximations.

**Definition 15.2.** Let \( f(x; \varepsilon) \) and \( \phi(x; \varepsilon) \) be continuous functions for \( x \in D \) and \( 0 < \varepsilon < a \). We call \( \phi(x; \varepsilon) \) a uniform asymptotic approximation to \( f(x; \varepsilon) \) for \( x \in D \) as \( \varepsilon \to 0 \) if for any positive number \( \delta \) there is an \( \varepsilon_1 \) (independent of \( x \) and \( \varepsilon \)) such that

\[
 |f(x; \varepsilon) - \phi(x; \varepsilon)| \leq \delta|\phi(x; \varepsilon)| \quad \text{for} \quad x \in D \quad \text{and} \quad 0 < \varepsilon < \varepsilon_1.
\]

We write \( f(x; \varepsilon) = \phi(x; \varepsilon) + o(\phi(x; \varepsilon)) \) uniformly in \( x \in D \) as \( \varepsilon \to 0 \). Note that \( D \) may depend on \( \varepsilon \).

**Example 15.3** Let \( D = [0, 1] \) and \( 0 < \varepsilon < 1 \). Then we have \( \cos(\varepsilon x) = 1 + O(1) \) uniformly in \( D \) as \( \varepsilon \to 0 \), since for any given \( \delta \) we can choose \( \varepsilon_1 = \sqrt{\delta} \) such that \( |\cos(\varepsilon x) - 1| \leq \varepsilon^2 x^2 \leq \varepsilon_1^2 = \delta \).

**Example 15.4** Although \( \cos(\varepsilon x) = O(1) \) uniformly in \( x \in [0, 1] \) for \( \varepsilon \to 0 \), there is no constant \( K \) such that \( \cos(x/\varepsilon) = K + o(1) \).

**Example 15.5** \( x + \sin(\varepsilon x) + e^{-x^3/\varepsilon} = x + \varepsilon x + O(\varepsilon) \) as \( \varepsilon \to 0 \) for all \( x \neq 0 \), but this is not the case uniformly in \( x \in [0, 1] \). More precisely, for \( x \in [\delta(\varepsilon), 1], x + \sin(\varepsilon x) + e^{-x^3/\varepsilon} = x + \varepsilon x + o(\varepsilon) \) uniformly if \( \varepsilon = \theta(\delta) \), but it is not uniformly equal for any \( \delta = O(\varepsilon) \). If \( x = O(\varepsilon) \), the otherwise exponentially small term is not small any more. This is illustrated in Figure 15.1. The difference between the original function and its nonuniform asymptotic approximation is
typically large in a neighbourhood of \(x = 0\), while the size of this neighbourhood is \(x = O(\varepsilon)\). This neighbourhood is an example of a boundary layer. The occurrence and behaviour of boundary layers will be discussed in more detail in Section 15.4.1.

15.2.2 Asymptotic Expansions

Asymptotic approximations are usually structured in the form of a series expansion that helps us to construct an approximation systematically.

**Definition 15.6.** The sequence \(\{\mu_n(\varepsilon)\}_{n=0}^{\infty}\) is called an asymptotic sequence if

\[
\mu_n(\varepsilon) + \mu_{n+1}(\varepsilon) = o(\mu_n(\varepsilon)) \quad \text{as} \quad \varepsilon \to 0
\]

for each \(n = 0, 1, 2, \ldots\)

**Example 15.7** Examples of asymptotic sequences (as \(\varepsilon \to 0\)) are

\[
\begin{align*}
\mu_n(\varepsilon) &= \varepsilon^n; \quad \mu_n(\varepsilon) = \varepsilon^{\frac{n}{2}}; \\
\mu_n(\varepsilon) &= \tan^n(\varepsilon); \quad \mu_n(\varepsilon) = \ln(\varepsilon)^{-n}; \\
\mu_n(\varepsilon) &= \varepsilon^p \ln(\varepsilon)^q, \quad \text{where} \quad p = 0, 1, 2, \ldots, \\
q &= 0, \ldots, p, \quad \text{and} \quad n = \frac{1}{2} p(p + 3) - q.
\end{align*}
\]

**Definition 15.8.** If \(\{\mu_n(\varepsilon)\}_{n=0}^{\infty}\) is an asymptotic sequence, then \(f(\varepsilon)\) has an asymptotic expansion of \(N\) terms with respect to this sequence denoted by

\[
f(\varepsilon) \sim \sum_{n=0}^{N-1} a_n \mu_n(\varepsilon),
\]

where the coefficients \(a_n\) are independent of \(\varepsilon\), if

\[
f(\varepsilon) - \sum_{n=0}^{M} a_n \mu_n(\varepsilon) = o(\mu_M(\varepsilon)) \quad \text{as} \quad \varepsilon \to 0
\]
15.2. Asymptotic Approximations and Expansions 463

for each \( M = 0, \ldots, N - 1 \). \( \mu_n(\varepsilon) \) is called a gauge function. If \( \mu_n(\varepsilon) = \varepsilon^n \), we call the expansion an asymptotic power series.

**Definition 15.9.** Two functions \( f \) and \( g \) are asymptotically equal up to \( N \) terms with respect to the asymptotic sequence \( \{\mu_n\} \) if \( f - g = o(\mu_N) \). If the remaining error is clear from the context, this is sometimes denoted as \( f \sim g \).

Asymptotic expansions based on the same gauge functions may be added. They may be multiplied if the products of the gauge functions can be asymptotically ordered. In contrast to ordinary series expansions, defined for an infinite number of terms, in asymptotic expansions only a finite (\( N \)) number of terms are considered. For \( N \to \infty \) the series may either converge or diverge, but this is irrelevant for the asymptotic behaviour.

For given \( \{\mu_n(\varepsilon)\}_{n=0}^{\infty} \) the coefficients \( a_n \) can be determined uniquely by the following recursive procedure (provided \( \mu_n \) are nonzero for \( \varepsilon \) near zero and each of the limits below exists):

\[
a_0 = \lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{\mu_0(\varepsilon)}, \quad a_1 = \lim_{\varepsilon \to 0} \frac{f(\varepsilon) - a_0\mu_0(\varepsilon)}{\mu_1(\varepsilon)}, \quad \ldots, \quad a_{N-1} = \lim_{\varepsilon \to 0} \frac{f(\varepsilon) - \sum_{n=0}^{N-2} a_n\mu_n(\varepsilon)}{\mu_{N-1}(\varepsilon)}.\]

**Example 15.10** A function may have different asymptotic expansions:

\[
\tan(\varepsilon) = \varepsilon + \frac{1}{3}\varepsilon^3 + \frac{2}{15}\varepsilon^5 + O(\varepsilon^7)
\]

\[
= \sin \varepsilon + \frac{1}{2}(\sin \varepsilon)^3 + \frac{3}{8}(\sin \varepsilon)^5 + O((\sin \varepsilon)^7)
\]

\[
= \varepsilon \cos \varepsilon + \frac{5}{6}((\varepsilon \cos \varepsilon)^3) + \frac{161}{120}(\varepsilon \cos \varepsilon)^5 + O((\varepsilon \cos \varepsilon)^7). \quad \square
\]

**Example 15.11** The asymptotic expansion

\[
e^{-1} e^{-1/\varepsilon} Ei(1/\varepsilon) = \sum_{n=0}^{\infty} a_n e^n + o(e^n), \quad \text{where} \quad Ei(x) = \int_{-\infty}^{x} \frac{e^t}{t} \, dt,
\]

which is related to the exponential integral \( Ei \), diverges as \( N \to \infty \) if \( \varepsilon \neq 0 \). The accuracy increases with \( N \), but on a smaller interval [1]. \( \square \)

**Example 15.12** Different functions may have the same asymptotic expansion:

\[
\cos(\varepsilon) = 1 - \frac{1}{2}\varepsilon^2 + \frac{1}{24}\varepsilon^4 + O(\varepsilon^6),
\]

\[
\cos(\varepsilon) + e^{-1/\varepsilon} = 1 - \frac{1}{2}\varepsilon^2 + \frac{1}{24}\varepsilon^4 + O(\varepsilon^6).
\]

Note that both asymptotic expansions, considered as regular power series in \( \varepsilon \), converge to \( \cos(\varepsilon) \) rather than \( \cos(\varepsilon) + e^{-1/\varepsilon} \). \( \square \)
Theorem 15.13. An asymptotic expansion vanishes only if the coefficients vanish; i.e.,
\[
\{a_0 \mu_0(\varepsilon) + a_1 \mu_1(\varepsilon) + a_2 \mu_2(\varepsilon) + \cdots = 0 \ (\varepsilon \to 0)\} \iff \{a_0 = a_1 = a_2 = \cdots = 0\}.
\]

Proof. The sequence \(\{\mu_n\}\) is asymptotically ordered, so both \(\mu_0 a_0 = -\mu_1 a_1 = \cdots = \mathcal{O}(\mu_1)\) and \(\mu_1 = \mathcal{O}(\mu_0)\). So there is a positive constant \(K\) such that for any positive \(\delta\) there is an \(\varepsilon\) interval where \(|a_0 \mu_0| < \delta K |\mu_0|\), which is only possible if \(a_0 = 0\). This may now be repeated for \(a_1\), etc. This proves \(\Rightarrow\). The proof of \(\Leftarrow\) is trivial. \(\square\)

15.2.3 Perturbation Problems

The assumed existence of an asymptotic expansion yields a class of methods to solve problems that depend on a parameter that is typically small in the range of interest. Such methods are called perturbation methods. The importance of these methods is twofold. They provide analytic solutions to otherwise intractable problems, and the asymptotic structure of the solution provides instant insight into the dominating qualities.

If \(a(\varepsilon)\) is implicitly given as the solution of an algebraic equation
\[
F(a; \varepsilon) = 0
\]
and both \(a(\varepsilon)\) and \(F(a; \varepsilon)\) have an asymptotic series expansion with the same gauge functions, \(a(\varepsilon)\) may be determined asymptotically by the following perturbation method. We expand \(a\), substitute this expansion in \(F\), and expand \(F\) to obtain
\[
a(\varepsilon) = a_0 \mu_0(\varepsilon) + a_1 \mu_1(\varepsilon) + \cdots, \tag{15.2a}
F(a; \varepsilon) = F_0(a_0 \mu_0(\varepsilon)) + F_1(a_1, a_0) \mu_1(\varepsilon) + F_2(a_2, a_1, a_0) \mu_2(\varepsilon) + \cdots = 0. \tag{15.2b}
\]

From Theorem 15.13 it follows that each term \(F_n\) vanishes, and the sequence of coefficients \((a_n)\) can be determined by induction:
\[
F_0(a_0) = 0, \quad F_1(a_1, a_0) = 0, \quad F_2(a_2, a_1, a_0) = 0, \quad \text{etc.} \tag{15.3}
\]

It should be noted that finding the sequence of gauge functions \((\mu_n)\) is of particular importance. This is done iteratively. First, the order of magnitude of \(a\) should be determined by seeking the asymptotic scaling \(a(\varepsilon) = \gamma(\varepsilon) A(\varepsilon)\) that yields a meaningful \(A = \mathcal{O}(1)\) in the limit \(\varepsilon \to 0\). This is called a distinguished limit, while the reduced equation for \(A(0)\), i.e., \(F_0(A) = 0\), is called a significant degeneration (there may be more than one). The first gauge function that occurs is now \(\mu_0(\varepsilon) = \gamma(\varepsilon)\), while \(a_0 = A(0)\). The procedure may be repeated for the new unknown \(a(\varepsilon) - \mu_0(\varepsilon) a_0\), and so on. Often the rest of the sequence \((\mu_n)\) can be guessed from the structure of the defining equation \(F = 0\).

We illustrate this procedure by the following example.

Example 15.14 Consider the roots for \(\varepsilon \to 0\) of the equation
\[
x^3 - \varepsilon x^2 + 2x^3 x + 2x^6 = 0.
\]
From the structure of the equation it seems reasonable to assume that the solutions \(x^{(1)}, x^{(2)}, x^{(3)}\) have an asymptotic expansion in powers of \(\varepsilon\). However, the order of magnitude of the leading-order term is not immediately clear:
\[
x(\varepsilon) = x^0(X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + \mathcal{O}(\varepsilon^3)).
\]
Therefore we have to determine the exponent $n$ first. This is done by balancing terms. We scale

$$x = \varepsilon^n X(\varepsilon), \quad X = \mathcal{O}(1),$$

and we seek an $n$ that produces a nontrivial limit as $\varepsilon \to 0$. We compare asymptotically the coefficients in the equation that remain after scaling:

$$\varepsilon^3 x^3 - \varepsilon^{1+2n} x^2 + 2\varepsilon^{3+n} x + 2\varepsilon^6 = 0.$$ 

In order to have a meaningful (or “significant”) degenerate solution $X(0) = \mathcal{O}(1)$, at least two terms of the equation should be asymptotically equivalent and at the same time of leading order when $\varepsilon \to 0$. So this leaves us with the task of comparing the exponents $3n, 1 + 2n, 3 + n,$ and $6$ as a function of $n$. Consider Figure 15.2. The solid lines denote the exponents of the powers of $\varepsilon$ that occur in the coefficients of the equation considered. At the intersections of these lines, denoted by the open and solid circles, we find the candidates of distinguished limits, i.e., the points where at least two coefficients are asymptotically equivalent. Finally, only the solid circles are the distinguished limits, because these are located along the lower envelope (thick solid line) and therefore correspond to leading-order terms when $\varepsilon \to 0$. We now have three cases.

**Case 1.** $n = 1$

$$\varepsilon^3 x^3 - \varepsilon^3 x^2 + 2\varepsilon^3 x + 2\varepsilon^3 = 0 \quad \text{or} \quad x^3 - x^2 + 2x + 2 = 0.$$ 

If we assume the expansion $X = X_0 + \varepsilon X_1 + \cdots$, we finally have

$$X_0^3 - X_0^2 = 0, \quad 3X_0^2X_1 - 2X_0X_1 + 2X_0 = 0, \quad \text{etc.,}$$

and so $X_0 = 1$ and $X_1 = -2$, etc., leading to $x(\varepsilon) = \varepsilon - 2\varepsilon^2 + \cdots$. Note that the solution $X_0 = 0$ is excluded because that would change the order of the scaling!
Case 2. $n = 2$

$$\varepsilon^6 X^3 - \varepsilon^5 X^2 + 2\varepsilon^5 X + 2\varepsilon = 0 \quad \text{or} \quad \varepsilon X^3 - X^2 + 2X + 2\varepsilon = 0.$$  

If we assume the expansion $X = X_0 + \varepsilon X_1 + \cdots$, we finally have

$$-X_0^2 + 2X_0 = 0, \quad \text{etc.,}$$

and so $X_0 = 2$, etc., leading to $x(\varepsilon) = 2\varepsilon^2 + \cdots$.

Case 3. $n = 3$

$$\varepsilon^3 X^3 - \varepsilon^7 X^2 + 2\varepsilon^6 X + 2\varepsilon = 0 \quad \text{or} \quad \varepsilon^3 X^3 - \varepsilon X^2 + 2X + 2 = 0.$$  

If we assume the expansion $X = X_0 + \varepsilon X_1 + \cdots$, we finally have

$$2X_0 + 2 = 0, \quad \text{etc.,}$$

and so $X_0 = -1$, etc., leading to $x(\varepsilon) = -\varepsilon + \cdots$. \qed

It is not always so easy to guess the general form of the gauge functions. Then all terms have to be estimated iteratively by a similar process of balancing as for the leading-order term. See Exercise 15.4.

### 15.2.4 Asymptotic Expansions of Poincaré Type

An asymptotic expansion of a function of variable $x$ and small parameter $\varepsilon$ of the form

$$f(x; \varepsilon) = \sum_{n=0}^{N-1} a_n(x; \varepsilon) \mu_n(\varepsilon) + O(\mu_N), \quad a_n = O(1),$$

appears to be too general to be of practical use. The restriction that the coefficients $a_n$ depend on $x$ only appears to be fruitful. This is called a Poincaré expansion, defined more precisely here.

**Definition 15.15.** If $\{\mu_n(\varepsilon)\}_{n=0}^{\infty}$ is an asymptotic sequence and $f(x; \varepsilon)$ has an asymptotic expansion of $N$ terms with respect to this sequence given by

$$f(x; \varepsilon) \sim \sum_{n=0}^{N-1} a_n(x) \mu_n(\varepsilon),$$

where the shape functions $a_n(x)$ are independent of $\varepsilon$, this expansion is called a Poincaré expansion. The domain of $x$ may depend on $\varepsilon$.

**Example 15.16** For $x > 0$, but not for $x \in (-\varepsilon, 0)$, we have the Poincaré expansion

$$\ln(x + \varepsilon) = \ln x + \frac{\varepsilon}{x} - \frac{\varepsilon^2}{2x^2} + \frac{\varepsilon^3}{3x^3} + O(\varepsilon^4). \quad \Box$$

**Example 15.17** $\sin\left(\frac{\varepsilon}{x}\right)$ has no Poincaré expansion for $x \neq 0$. \Box
15.2. Asymptotic Approximations and Expansions

Example 15.18  The Poincaré expansion of $e^{-x^4}$ with respect to $[x^a]$ is equal to zero for $x \in [A, \infty)$, is equal to one for $x = 0$, and does not exist for $x \in (-\infty, -A]$. ($A$ is fixed and positive.) \hfill $\Box$

Definition 15.19.  If a Poincaré expansion is uniform in $x$ on a given domain $D$ (Definition 15.2), this expansion is called a regular expansion. Otherwise, the expansion is called a singular expansion.

(Not that there is no uniformity in the literature on the definition of regular and singular expansions.) Regular expansions may be differentiated with respect to the independent variable $x$.

Example 15.20  The Poincaré expansion

$$\cos(x + \epsilon) = \cos(x) - \epsilon \sin(x) - \frac{1}{2} \epsilon^2 \cos(x) + \frac{1}{6} \epsilon^3 \sin(x) + O(\epsilon^4),$$

with respect to the gauge functions $\mu_x(e) = e^a$ and with domain $D = \mathbb{R}$, is uniform since $\cos(x)$ and $\sin(x)$ are bounded for all $x \in \mathbb{R}$. It follows that it is a regular expansion. \hfill $\Box$

Example 15.21  The expansion

$$\cos(x + \epsilon) e^{-x^4} + \sin(x + \epsilon) = \sin(x + \epsilon) \cos x + O(\epsilon^2)$$

is a uniform, and therefore regular, expansion on any interval $[A, \infty)$, where $A > 0$. However, it is a nonuniform, and therefore singular, expansion on $[0, \infty)$. In fact, on any interval $[A\epsilon^a, \infty)$ it is regular if $\alpha < 1$ and singular if $\alpha \geq 1$. \hfill $\Box$

It is important to appreciate the central role of the choice of the independent variable $x$ in a Poincaré expansion. By suitable linear coordinate transformations of the type $x = \lambda(\epsilon) + \delta(\epsilon) \xi$ we can change and optimize the domain of uniformity. This filters out specific behaviour that belongs to one asymptotic length scale.

Example 15.22

- $\sin(x + \epsilon x + \epsilon^2 x) = \sin(x) + \epsilon x \cos(x) + \epsilon^2 x \cos x - \frac{1}{4} \epsilon^2 \sin x + O(\epsilon^3)$, is only uniform on an interval $[0, A]$ but if we introduce $\xi = (1 + \epsilon) x$, we have $\sin(\xi + \epsilon^2 x) = \sin(\xi) + O(\epsilon)$ uniform in $x \in [0, A\epsilon^{-1}]$ for any positive constant $A$.

- $\sin(\epsilon x + \delta) = \epsilon x + \delta + O(\epsilon^3)$ is only uniform on a finite interval and, moreover, does not show any of the inherent periodicity. If we introduce $X = \epsilon x$, we get the much better $\sin(X + \epsilon) = \sin(X) + O(\epsilon)$, which is even uniform in $\mathbb{R}$.

- $e^{-x^4} = 0 + o(x^a)$ is a singular expansion on $x > 0$, but if we introduce $\xi = x/\epsilon$, it becomes the regular expansion $e^{-\xi^4} = O(1)$ on $\xi > 0$.

- On $x > 0$ we have $\frac{\pi}{2} \arctan \left( \frac{x}{1 + x^2} \right) + \sin(\epsilon x)/(1 + x^2)$ equal to $1 + \epsilon (x/(1 + x^2) - 2/\pi x) + O(\epsilon^3)$ in $x$, $1 + \epsilon^2 (\sin(X)/X^2 - 2/\pi X) + O(\epsilon^3)$ in $X = \epsilon x$, and $\frac{\pi}{2} \arctan(\epsilon x) = \epsilon x + \epsilon^2 x + O(\epsilon^3)$ in $\xi = x/\epsilon$.

- $e^{-x^4} \sin(x \sqrt{1 + \epsilon}) = \sin x + \epsilon x \left( \frac{1}{2} \cos x - \sin x \right) + O(\epsilon^2)$ is only uniform on an interval $[0, A]$. This cannot be improved by a single other choice of independent variable. However, if we introduce two variables, $x_1 = \epsilon x$ and $x_2 = (1 + \epsilon x)$, we get the much better $e^{-x^4} \sin x_2 - \frac{1}{2} \epsilon^2 x_2 e^{-x^4} \cos x_2 + O(\epsilon^3)$. \hfill $\Box$
15.3 Regular Perturbation Problems

If a function $\Phi(x; \varepsilon)$ is implicitly given by an equation (usually a differential equation with boundary conditions), say

$$L[\Phi](x; \varepsilon) = 0$$

on a domain $D$, (15.4)

and both $\Phi$ and $L[\Phi](x; \varepsilon)$ have a regular asymptotic expansion on $D$ with the same gauge functions, (15.4) is called a regular perturbation problem [65]. The shape functions $\Phi_n$ are determined as follows. We expand $L[\Phi]$ as

$$L[\Phi](x; \varepsilon) = \mu_0(\varepsilon) L_0[\Phi_0](x) + \mu_1(\varepsilon) L_1[\Phi_1, \Phi_0](x) + \mu_2(\varepsilon) L_2[\Phi_2, \Phi_1, \Phi_0](x) + \cdots = 0.$$ (15.5)

According to Theorem 15.13, each term vanishes, and the sequence $(\Phi_n)$ can be determined by induction:

$$L_0[\Phi_0](x) = 0, \quad L_1[\Phi_1, \Phi_0](x) = 0, \quad L_2[\Phi_2, \Phi_1, \Phi_0](x) = 0, \quad \ldots$$ (15.6)

It should be noted that in many interesting cases the problem is only regular after a suitable coordinate transformation. The major task when solving the problem is then to find this scaled or shifted coordinate. Practically important solution methods of this type are the method of slow variation, for geometrically stretched or slowly varying configurations, and the Lindstedt–Poincaré method, for solutions that are periodic in time with an unknown $\varepsilon$-dependent period.

If (15.4) is not a regular perturbation problem, we call it a singular perturbation problem. Practically important solution methods for singular perturbation problems are the MAE method, where regular expansions exist locally but not in the whole region considered, and the method of multiple scales, where two or more distinct long and short length scales are intertwined.

15.3.1 Method of Slow Variation

Suppose we have a function $\psi(x; \varepsilon)$ of spatial coordinates $x$ and a small parameter $\varepsilon$ such that the typical variation in one direction, say $x$, is of the order of the length scale $\varepsilon^{-1}$. We can express this behaviour most conveniently by writing $\psi(x, y, z; \varepsilon) = \Phi(\varepsilon x, y, z; \varepsilon)$. Now if we were to expand $\Phi$ for small $\varepsilon$, we might, for example, get something like

$$\Phi(\varepsilon x, y, z; \varepsilon) = \Phi(0, y, z; 0) + \varepsilon(\Phi_1(0, y, z; 0) + \Phi_2(0, y, z; 0)) + \cdots,$$

which is only uniform in $x$ on an interval $[0, L]$ if $L = O(1)$, and the inherent slow variation on the longer scale of $x = O(\varepsilon^{-1})$ would be masked. It is clearly much better to introduce the scaled variable $X = \varepsilon x$, and an (assumed) regular expansion of $\Phi(X, y, z; \varepsilon)$

$$\Phi(X, y, z; \varepsilon) = \mu_0(\varepsilon) \phi_0(X, y, z) + \cdots$$ (15.7)

now retains the slow variation in $X$ in the shape functions of the expansion.
This situation frequently happens when the geometry involved is slender [163]. The theories of one-dimensional gas dynamics, lubrication flow, and sound propagation in horns (Webster’s horn equation) are important examples, although they are usually not derived systematically without explicit reference to the slender geometry. We will illustrate the method for heat flow in a varying bar, quasi one-dimensional gas flow, and the shallow-water problem.

**Example 15.23 (heat flow in a bar)** Consider the stationary problem of the temperature distribution \( T \) in a long heat-conducting bar with outward surface normal \( n \) and slowly varying cross section \( \mathcal{A} \). The bar is kept at a temperature difference such that a given heat flux is maintained, but it is otherwise isolated. As there is no leakage of heat, the flux is constant. With spatial coordinates made dimensionless on a typical bar cross section, we have the following equation and boundary conditions:

\[
\nabla^2 T = 0, \quad \nabla T \cdot n = 0, \quad \int_{\mathcal{A}} \frac{\partial T}{\partial x} dS = Q.
\]

After integrating \( \nabla^2 T \) over a slice \( x_1 \leq x \leq x_2 \) and applying Gauss’s theorem, we find that the axial flux \( Q \) is indeed independent of \( x \). The typical length scale of diameter variation is assumed to be much larger than a diameter. We introduce the ratio between a typical diameter and this length scale as the small parameter \( \varepsilon \) and write for the bar surface \( S(X, r, \theta) = 0 \), \( X = \varepsilon x \), where \((x, r, \theta)\) form a cylindrical coordinate system (see Figure 15.3). By writing \( R \) as a continuous function of the slow variable \( X \), rather than \( x \), we have made our formal assumption of slow variation explicit in a convenient and simple way, since \( \varepsilon R_x = \mathcal{O}(\varepsilon) \).

The crucial step will now be the assumption that the temperature is only affected by the geometric variation induced by \( R \). Any initial or entrance effects are ignored or have disappeared. As a result, the temperature field \( T \) is a function of \( X \), rather than \( x \), and its axial gradient scales on \( \varepsilon \) as \( T_x = \mathcal{O}(\varepsilon) \).

We introduce the gradient \( \nabla S \) and the transverse gradient \( \nabla_{\perp} S \):

\[
\nabla S = -\varepsilon R_x e_x + e_r - r^{-1} R_{\theta} e_{\theta}, \quad \nabla_{\perp} S := S_r e_r + r^{-1} S_{\theta} e_{\theta} = e_r - r^{-1} R_{\theta} e_{\theta}.
\]

At the bar surface \( S = 0 \) the gradient \( \nabla S \) is a vector normal to the surface, while the transverse gradient \( \nabla_{\perp} S \), directed in the plane of a cross section \( X = \text{constant} \), is normal to the circumference \( S(X = c, r, \theta) = 0 \). Inside the bar we have the rescaled heat equation

\[
\varepsilon^2 T_{XX} + \nabla_{\perp}^2 T = 0. \quad (\star)
\]

![Figure 15.3. Slowly varying bar.](image-url)
At the wall the boundary condition of vanishing heat flux is
\[ \nabla T \cdot \nabla S = \epsilon^2 T X S_X + \nabla_l T \cdot \nabla_l S = 0 \quad \text{at} \quad S = 0. \tag{1} \]

The flux condition, for lucidity rewritten with \( Q = \epsilon q \), is given by
\[ \int_A \frac{\partial T}{\partial X} dS = q. \]

This problem is too difficult in general, so we try to utilize the small parameter \( \epsilon \) in a systematic manner. Since the perturbation terms are \( O(\epsilon^2) \), we assume the asymptotic expansion
\[ T(X, r, \theta; \epsilon) = T_0(X, r, \theta) + \epsilon^2 T_1(X, r, \theta) + \mathcal{O}(\epsilon^4). \]

After substitution in (\( \ast \)) and boundary condition (\( \dagger \)), further expansion in powers of \( \epsilon^2 \), and equating like powers of \( \epsilon \), we obtain to leading order a Laplace equation in \((r, \theta)\):
\[ \nabla^2_\perp T_0 = 0 \quad \text{with} \quad \nabla_\perp T_0 \cdot \nabla_\perp S = 0 \quad \text{at} \quad S = 0. \]

An obvious solution is \( T_0(X, r, \theta) \equiv 0 \). Since solutions of Laplace’s equation with vanishing normal derivatives at the boundary are unique up to a constant (here a function of \( X \)), we have
\[ T_0 = T_0(X). \]

We could substitute this directly in the flux condition to find that \( A T_{0X} = q \), where \( A(X) \) is the area of cross section \( A(X) \). For the present exposition, however, it is of interest to show that this result also emerges from the equations as follows. To obtain an equation for \( T_0 \) in \( X \) we continue with the \( O(\epsilon^2) \) equation and corresponding boundary condition
\[ \nabla^2_\perp T_1 + T_{0,XX} = 0 \quad \nabla_\perp T_1 \cdot \nabla_\perp S = -T_{0X} S_X. \tag{\( \ddagger \)} \]

The boundary condition can be rewritten as
\[ \nabla_\perp T_1 \cdot \n_\perp = \frac{T_{0X} R_X}{|\nabla_\perp S|} = \frac{T_{0X} R_X}{\sqrt{R^2 + R_0^2}}, \]
where \( n_\perp = \nabla_\perp S/|\nabla_\perp S| \) is the transverse unit normal vector. By integrating (\( \ddagger \)) over a cross section \( A \) of area \( A(X) \), using Gauss’s theorem, and noting that \( A = \frac{2\pi}{\epsilon} R^2 d\theta \), and that a circumferential line element is given by \( d\ell = (R^2 + R_0^2)^{1/2} d\theta \), we obtain
\[ \int_A (\nabla^2_\perp T_1 + T_{0,XX}) dS = \oint_{\partial A} \nabla_\perp T_1 \cdot n_\perp d\ell + A T_{0,XX} \]
\[ = T_{0X} \int_0^{2\pi} R X d\theta + A T_{0,XX} = A_X T_{0X} + A T_{0,XX} = \frac{d}{dX} \left( A \frac{d}{dX} T_0 \right) = 0. \]

The finally obtained equation can be solved easily. Note that we have recovered the conservation law of heat flux \( A T_{0X} = q \). Finally, we have
\[ T_0(X) = \int^X q/A(\zeta) d\zeta + T_{ref}. \]

It should be noted that we did not include in our analysis any boundary conditions at the ends of the bar. It is true that the present method fails here. The found solution is uniformly valid on \( \mathbb{R} \) (since \( R(X) \) is assumed continuous and independent of \( \epsilon \), but only as long as we stay away from any end. Near the ends the boundary conditions induce transverse gradients of \( \mathcal{O}(1) \), which make the prevailing length scale \( x \) again, rather than \( X \). This region is asymptotically of boundary layer type and should be treated differently (see below). \( \Box \)
Example 15.24 (quasi one-dimensional gas dynamics) Consider a slowly varying duct with irrotational inviscid isentropic flow, described (in dimensionless form) by the velocity potential \( \varphi \) and density \( \rho \) satisfying the equation for mass conservation and the compressible form of Bernoulli's equation (see Chapter 7); i.e.,

\[
\nabla \cdot (\rho \nabla \varphi) = 0, \quad \frac{1}{2} |\nabla \varphi|^2 + \frac{\rho \gamma - 1}{\gamma - 1} = E.
\]

The parameter \( \gamma \) is a gas constant (1.4 for air) and \( E \) is a constant of the problem. With the same notation as in the previous example, the duct wall is given by \( S(\varepsilon x, r, \theta) = 0 \), while at the impermeable wall we have \( \nabla \varphi \cdot \nabla S = 0 \). The mass flux, the same at any cross section \( A \), is given by

\[
\int_{A} \rho \varphi_x \, dS = F.
\]

We introduce the slow variable \( X = \varepsilon x \) and assume \( \varphi \) and \( \rho \) to depend essentially on \( X \) rather than \( x \). The dimensionless axial flow velocity \( \varphi_x \), the density \( \rho \), the cross-sectional area \( A \), the flux \( F \), and the thermodynamical constant \( E \) are \( O(1) \). So we have to rescale \( \varphi \) and write

\[
\varphi(x, y, z; \varepsilon) = \varepsilon^{-1} \Phi(X, y, z; \varepsilon).
\]

The equations for \( \Phi \) and \( \rho \) become

\[
e^2 \frac{\partial}{\partial X} \left( \rho \frac{\partial}{\partial X} \Phi \right) + \nabla \cdot (\rho \nabla \Phi) = 0, \quad \frac{1}{2} \Phi_x^2 + \frac{1}{2} \left( \varepsilon^2 |\nabla \Phi|^2 \right) + \frac{\rho \gamma - 1}{\gamma - 1} = E,
\]

with boundary condition

\[
\nabla \Phi \cdot \nabla S = e^2 \Phi_X S_X + \nabla \Phi \cdot \nabla S = 0 \quad \text{at} \quad S = 0.
\]

We assume the expansions

\[
\Phi(X, y, z; \varepsilon) = \Phi_0(X, y, z) + O(\varepsilon), \quad \rho(X, y, z; \varepsilon) = \rho_0(X, y, z) + O(\varepsilon).
\]

From Bernoulli's equation it follows that \( |\nabla \Phi_0|^2 = 0 \), so \( \Phi_0 = \Phi_0(X) \), and hence \( \rho_0 = \rho_0(X) \). From the mass flux equation we get \( \Phi_{0X} = F/\rho_0 A \), while the algebraic equation

\[
\frac{F^2}{2 \rho_0^2 A^2} + \frac{\rho_0^{\gamma - 1}}{\gamma - 1} = E
\]

finally determines \( \rho_0 \) (in general to be solved numerically).

Example 15.25 (shallow-water equations) The irrotational motion under gravity of a horizontal layer of inviscid, incompressible water is described by the equation for mass conservation and the Bernoulli equation (see (7.11)):

\[
\nabla^2 \varphi = 0, \quad \frac{\partial \varphi}{\partial t} + \frac{1}{2} |\nabla \varphi|^2 + \frac{p}{\rho_0} + g z = C(t),
\]

where \( \varphi \) is the velocity potential with velocity \( \mathbf{v} = \nabla \varphi \), \( \rho_0 \) is the density, \( p \) is the pressure, \( g \) is the gravitational acceleration, and \( C \) is an unimportant function of time. The boundary conditions are provided by the impermeability of the bottom at \( z = 0 \), the assumption that the free surface \( z = h(x, y, t) \) consists of streamlines (particles remain there), and the fact that the pressure is uniformly constant along the free surface (the big difference between the densities of water and air makes the water insensitive to any air motion). As any constant in pressure
may be absorbed by $C$, we may assume that the surface pressure is zero. These conditions result in

\[ z = 0: \quad \frac{\partial \psi}{\partial z} = 0, \]
\[ z = h: \quad \frac{\partial \psi}{\partial z} = \frac{\partial h}{\partial t} + \frac{\partial \psi}{\partial x} \frac{\partial h}{\partial x} + \frac{\partial \psi}{\partial y} \frac{\partial h}{\partial y} \quad \text{and} \quad p = 0. \]

We assume that the typical horizontal velocities $U$ are so large and the frequencies $f$ are so low that the corresponding typical length scale $L = U/f$ is large compared to the typical water depth $D$ (e.g., tidal motion). To quantify this slenderness we introduce the small parameter $\varepsilon = D/L$. We are interested in the situation where pressure is coupled to both the inertia of the flow and the effects of gravity. This corresponds to the assumption that $p$ scales on $\rho_0 U^2$ and the inverse-squared Froude number $\gamma = gD/U^2 = O(1)$. Suppose we scale and nondimensionalise as follows:

\[ x = LX, \quad y = LY, \quad z = DZ, \quad t = LU^{-1} \tau, \quad \psi = U \psi, \quad p = \rho_0 U^2 P, \quad h = DH. \]

We also introduce the operator $\nabla_\perp := e_x \frac{\partial}{\partial X} + e_y \frac{\partial}{\partial Y}$, which is $\nabla$ restricted to $X$ and $Y$. Then we obtain

\[ \frac{\partial^2 \psi}{\partial Z^2} + \varepsilon^2 \nabla_\perp^2 \psi = 0, \quad \frac{\partial \psi}{\partial \tau} + \frac{1}{2} |\nabla_\perp \psi|^2 + \frac{1}{2} \varepsilon^2 \left( \frac{\partial \psi}{\partial Z} \right)^2 + P + \gamma Z = \tilde{C}(t), \]

with boundary conditions

\[ Z = 0: \quad \frac{\partial \psi}{\partial Z} = 0, \]
\[ Z = H: \quad \frac{\partial \psi}{\partial Z} = \varepsilon^2 \left( \frac{\partial H}{\partial \tau} + \nabla_\perp \psi \cdot \nabla_\perp H \right) \quad \text{and} \quad P = 0. \]

Assuming no interfering $O(\varepsilon)$ effects (e.g., from initial or boundary conditions), we expand in powers of $\varepsilon^2$ the only small parameter that occurs:

\[ \psi = \Psi_0 + \varepsilon^2 \Psi_1 + O(\varepsilon^4), \quad H = H_0 + \varepsilon^2 H_1 + O(\varepsilon^4), \quad P = P_0 + \varepsilon^2 P_1 + O(\varepsilon^4). \]

From this we obtain to leading order $\Psi_{0,ZZ} = 0$, which integrates to $\Psi_{0,Z} = B_0(X, Y, \tau) = 0$ because of the boundary conditions. So we finally have

\[ \Psi_0 = A_0(X, Y, \tau). \]

To first order we have $\Psi_{1,ZZ} = -\nabla_\perp^2 \Psi_0$, which integrates to

\[ \frac{\partial}{\partial Z} \Psi_1 = -Z \nabla_\perp^2 \Psi_0 \]

when we take into account the boundary condition at $Z = 0$. Next we expand and substitute these results into Bernoulli’s equation (note that $\varepsilon^{-2} (\psi_Z)^2 = O(\varepsilon^2)$) and get

\[ \frac{\partial}{\partial \tau} \Psi_0 + \frac{1}{2} |\nabla_\perp \Psi_0|^2 + P_0 + \gamma Z = \tilde{C}(t). \]

By the pressure boundary condition at $Z = H_0$ this yields

\[ \frac{\partial}{\partial \tau} \Psi_0 + \frac{1}{2} |\nabla_\perp \Psi_0|^2 + \gamma H_0 = \tilde{C}(t). \quad (*) \]
Finally, the streamline condition produces
\[ \frac{\partial}{\partial Z} \Psi_1 = -H_0 \nabla^2 \Psi_0 = \frac{\partial}{\partial \tau} H_0 + \nabla \cdot (H_0 \nabla \Psi_0) \]
or
\[ \frac{\partial}{\partial \tau} H_0 + \nabla \cdot (H_0 \nabla \Psi_0) = 0. \] (†)

Equations (†) are known as a form of the shallow-water equations (see also Chapter 12). They are not generally solvable, and their behaviour requires extensive analysis.

A family of simple wave solutions (i.e., along a characteristic) may be found as follows (cf. [169]). We look for a plane wave in, say, the \( \xi = X \cos \theta + Y \sin \theta \) direction, so we have, after rewriting the equations in terms of \( \xi \), velocity \( V_0 = \Psi_0, \) and height \( H_0, \)

\[ \frac{\partial}{\partial \tau} H_0 + \frac{\partial}{\partial \xi} (V_0 H_0) = 0, \quad \frac{\partial}{\partial \tau} V_0 + \frac{\partial}{\partial \xi} \left( \frac{1}{2} V_0^2 + \gamma H_0 \right) = 0. \]

Evidently, both equations may be written in the same characteristic form if

\[ \frac{d}{dH_0} (V_0 H_0) = \frac{d}{dV_0} \left( \frac{1}{2} V_0^2 + \gamma H_0 \right). \]

This is satisfied if \( (\frac{d}{dH_0} V_0)^2 H_0 = \gamma, \) or

\[ V_0 = \pm 2 \sqrt{\gamma H_0} + c_0, \]

where \( c_0 \) is a constant. We can now assume that we have \( H_0 = H_0(\eta) \) and the corresponding \( V_0 = V_0(\eta), \) where \( \eta = \eta(x, t) \) satisfies

\[ \frac{\partial}{\partial \tau} \eta + (V_0 \pm \sqrt{\gamma H_0}) \frac{\partial}{\partial \xi} \eta = 0. \]

Using the result of Exercise 1.5, we obtain solutions implicitly described by

\[ \eta = F(\xi - f(\eta) \tau), \quad \text{where } f(\eta) = V_0(\eta) \pm \sqrt{\gamma H_0(\eta)} \]

and \( F \) is any suitable differentiable function.

### 15.3.2 Lindstedt–Poincaré Method

When we have a function \( f \) depending on a small parameter \( \varepsilon \) and periodic in \( t \) with fundamental frequency \( \omega(\varepsilon) \), we can write \( f \) as a Fourier series

\[ f(t; \varepsilon) = \sum_{n=-\infty}^{\infty} A_n(\varepsilon) e^{i\omega_n(\varepsilon)t}. \] (15.8)

If amplitudes and frequency have an asymptotic expansion for small \( \varepsilon \), say

\[ A_n(\varepsilon) = A_{n,0} + \varepsilon A_{n,1} + \cdots, \quad \omega(\varepsilon) = \omega_0 + \varepsilon \omega_1 + \cdots, \] (15.9)

we have a natural asymptotic series expansion for \( f \) of the form

\[ f(t; \varepsilon) = \sum_{n=-\infty}^{\infty} A_{n,0} e^{i\omega_n(\varepsilon)t} + \varepsilon \sum_{n=-\infty}^{\infty} \left( A_{n,1} + i\varepsilon \omega_1 A_{n,0} \right) e^{i\omega_n(\varepsilon)t} + \cdots. \] (15.10)
This expansion, however, is only uniform in $t$ on an interval $[0, T]$, where $T = o(\varepsilon^{-1})$. On a larger interval, e.g., $[0, \varepsilon^{-1}]$, the asymptotic hierarchy in the expansion becomes invalid because $\varepsilon t = O(1)$. This happens because of the occurrence of algebraically growing oscillatory terms, called “secular terms.” Secular means “occurring once in a century,” and saeculum means “generation,” referring to its astronomical origins.

**Definition 15.26.** The terms proportional to $t^n \sin(n\omega_0 t)$, $t^n \cos(n\omega_0 t)$ are called secular terms. More generally, the name refers to any algebraically growing terms that limit the region of validity of an asymptotic expansion.

It is therefore far better to apply a coordinate transformation $\tau = \omega(\varepsilon)t$, introduce $F(\tau; \varepsilon) = f(t; \varepsilon)$, and expand $F$, rather than $f$, asymptotically. We get the uniformly valid approximation

$$F(\tau; \varepsilon) = \sum_{n=-\infty}^{\infty} A_n(\varepsilon) e^{i\tau} = \sum_{n=-\infty}^{\infty} A_{n,0} e^{i\tau} + \varepsilon \sum_{n=-\infty}^{\infty} A_{n,1} e^{i\tau} + \cdots. \quad (15.11)$$

The method is called the Lindstedt–Poincaré method or the method of strained coordinates. In practical situations the frequency $\omega$ is of course unknown and has to be found. Therefore, when constructing the solution, we have to allow for an unknown coordinate transformation. In order to construct the unknown $\omega(\varepsilon)$, we expand $\tau$, e.g., like

$$\tau = (\omega_0 + \varepsilon \omega_1 + \varepsilon^2 \omega_2 + \cdots) t, \quad (15.12)$$

but this depends of course on the problem. Note that the purpose of the scaling is to render the asymptotic expansion of $F$ regular, so it is no restriction to assume $\omega_0 = 1$. The other coefficients are determined from the additional condition that the asymptotic hierarchy should be respected as long as possible. In other words, secular terms should not occur. We will illustrate this with the following example.

**Example 15.27 (the pendulum)** Consider the motion of the pendulum described by the initial value problem (see Example 7.2)

$$\ddot{\theta} + K^2 \sin(\theta) = 0, \quad \text{with} \quad \theta(0) = \varepsilon, \quad \theta'(0) = 0,$$

where $0 < \varepsilon \ll 1$. After the transformation $\tau = \omega t$ and noting that $\theta = O(\varepsilon)$, we have

$$\omega^2 \theta'' + K^2 \left( \theta - \frac{1}{6} \theta^3 + \cdots \right) = 0.$$

We expand

$$\omega = 1 + \varepsilon^2 \omega_1 + \cdots, \quad \theta = \varepsilon \theta_0 + \varepsilon^3 \theta_1 + \cdots$$

and find, after substitution, the equations for the first two orders:

$$\theta_0'' + K^2 \theta_0 = 0, \quad \theta_0(0) = 1, \quad \theta_0'(0) = 0,$$

$$\theta_1'' + K^2 \theta_1 = -2 \omega_0 \theta_0'' + \frac{1}{6} K \theta_0^3, \quad \theta_1(0) = 0, \quad \theta_1'(0) = 0.$$

The solution $\theta_0$ is obviously given by

$$\theta_0 = \cos(K \tau),$$
leading to (see the appendix, Section B) the following equation for $\theta_1$:

$$\theta_1'' + K^2 \theta_1 = 2K \left( \omega_1 K + \frac{1}{16} \right) \cos(K\tau) + \frac{1}{24} K \cos(3K\tau).$$

At this point it is essential to observe that the right-hand side consists of two forcing terms: one with frequency $3K$ and one with $K$, the resonance frequency of the left-hand side. This resonance leads to secular terms, as the solutions will behave like $\tau \sin(K\tau)$ and $\tau \cos(K\tau)$. Therefore, in order to suppress the occurrence of secular terms, the amplitude of the resonant forcing term should vanish, which yields the next-order terms $\omega_1$ and $\theta_1$:

$$\omega_1 = -\frac{1}{16K}, \quad \text{leading to} \quad \theta_1 = \frac{1}{192K} (\cos(K\tau) - \cos(3K\tau)).$$

15.4 Singular Perturbation Problems

If the solution of the problem considered does not allow a regular expansion, the problem is singular and the solution has no uniform Poincaré expansion in the same variable. We will consider two classes of problems. In the first one the singular behaviour is of boundary layer type and the solution can be built up from locally regular expansions. The solution method is called the method of Matched Asymptotic Expansions. In the other one more time or length scales occur together and a solution is constructed by considering these length scales as if they were independent. The solution method is called the method of multiple scales.

15.4.1 Matched Asymptotic Expansions

Very often it happens that a simplifying limit applied to a more comprehensive model gives a correct approximation for the main part of the domain but not everywhere: the limit is nonuniform. This nonuniformity may be in space, in time, or in any other variable. For the moment we think of nonuniformity in space, let’s say a small region near $x = 0$. If this region of nonuniformity is crucial for the problem, e.g., because it contains a boundary condition, or a source, the primary reduced problem (which does not include the region of nonuniformity) is not sufficient. This, however, does not mean that no use can be made of the inherent small parameter. The local nature of the nonuniformity itself often makes another reduction possible. In such a case we have apparently two physically connected but different problems, called inner and outer problems, as far as the dominating mechanism is concerned. Depending on the problem, we now have two simpler problems, serving as boundary conditions to each other via continuity or matching conditions.

Nonuniform Asymptotic Approximations Suppose that a given sufficiently smooth function $\Phi(x; \varepsilon)$, with $0 \leq x \leq 1$, and $0 < \varepsilon \leq \varepsilon_1$, does not have a uniform limit $\varepsilon \to 0$, $x \to 0$. Typically, such a function will depend, apart from $x$, on combinations like $x/\delta(\varepsilon)$, where $\delta = o(1)$.

Assume that this function does not have a regular asymptotic expansion on the whole interval $[0, 1]$ but only on partial intervals $x \in [\eta(\varepsilon), 1]$, where $\eta = o(1)$ and $\delta = o(\eta)$:

$$\Phi(x; \varepsilon) \approx \sum_{k=0}^{n} \mu_k(\varepsilon) \psi_k(x) + o(\mu_n) \quad \varepsilon \to 0, \quad x = O(1). \quad (15.13)$$
We call this expansion the *outer expansion*, principally valid in the “$x = O(1)$” outer region, but extendible to $[O(\eta), 1]$. Now consider the *stretched coordinate*

$$\xi = \frac{x}{\delta(\varepsilon)}. \tag{15.14}$$

Assume that the transformed $\Psi(\xi; \varepsilon) = \Phi(x; \varepsilon)$ has a nontrivial regular asymptotic expansion on partial intervals $\xi \in [0, \zeta(\varepsilon)/\delta(\varepsilon)]$, where $\eta(\varepsilon) < \zeta(\varepsilon)$:

$$\Psi(\xi; \varepsilon) = \sum_{k=0}^{m} \lambda_k(\varepsilon) \psi_k(\xi) + o(\lambda_m) \quad \varepsilon \to 0, \quad \xi = O(1). \tag{15.15}$$

We call this expansion the *inner expansion*, principally valid in the “$\xi = O(1)$” inner region, but extendible to $[0, O(\zeta)]$.

**Example 15.28**

$$\Phi(x; \varepsilon) = \arctan \left( \frac{x}{\varepsilon} \right) + \sin(x + \varepsilon)$$

$$= \frac{\pi}{2} + \sin x + \varepsilon \cos x - \frac{\varepsilon}{x} + O(\varepsilon^{3/2}) \quad \text{on } \frac{1}{2} \leq x \leq 1;$$

$$\Psi(\xi; \varepsilon) = \arctan(\xi) + \sin(\varepsilon \xi + \varepsilon)$$

$$= \arctan(\xi) + \varepsilon(\xi + 1) + O(\varepsilon^{3/2}) \quad \text{on } 0 \leq \xi \leq \frac{2}{\varepsilon^{1/2}}. \quad \Box$$

The adjective “nontrivial” is essential: the expansion must be *significant*, i.e., different from the outer expansion in $\Phi_n$ rewritten in the inner variable $\xi$. This determines the choice of the inner variable $\xi = x/\delta(\varepsilon)$. The scaling $\delta(\varepsilon)$ is the asymptotically largest gauge function with this property. We call the expansion for $\Psi$ the inner expansion or *boundary layer expansion*, where the region $\xi = O(1)$ or $x = O(\delta)$ is the boundary layer with thickness $\delta$ and $\xi$ is the boundary layer variable. Boundary layers may be nested and may occur at internal points of the domain of $\Phi$. Then they are called *internal layers*. The assumption $\eta < \xi$, i.e., that the inner and outer expansions may be extended to regions that overlap, is called the *overlap hypothesis*.

Suppose $\Phi(x; \varepsilon)$ has an outer expansion

$$\Phi(x; \varepsilon) = \sum_{k=0}^{n} \mu_k(\varepsilon) \psi_k(x) + o(\mu_n) \tag{15.16}$$

and a boundary layer $x = O(\delta)$ with inner expansion

$$\Psi(\xi; \varepsilon) = \sum_{k=0}^{m} \lambda_k(\varepsilon) \psi_k(\xi) + o(\lambda_m). \tag{15.17}$$

Suppose that both expansions are complementary; i.e., there is no other boundary layer in between $x = O(1)$ and $x = O(\delta)$. Then the overlap hypothesis says that the expansions represent the same function in an intermediate region of overlap. This overlap region may be described by a stretched variable $x = \eta(\varepsilon) \sigma$, asymptotically in between $O(1)$ and $O(\delta)$, so
\[ \delta \ll \eta \ll 1. \] In the overlap region both expansions match, which means that the expansions are asymptotically equivalent and reduce to the same expressions.

A widely used procedure, which is simpler to use than matching by means of intermediate variables but not as universal when the expansions involve logarithms, is Van Dyke’s matchings rule \([162, 38]\): the outer expansion, rewritten in the inner variable, has a regular series expansion that is \( \text{equal} \) to the regular asymptotic expansion of the inner expansion rewritten in the outer variable. Suppose that

\[
\sum_{k=0}^{n} \mu_k(\epsilon) \varphi_k(\delta \xi) = \sum_{k=0}^{m} \lambda_k(\epsilon) \eta_k(\xi) + o(\lambda_m), \quad (15.18a)
\]

\[
\sum_{k=0}^{n} \lambda_k(\epsilon) \psi_k(x/\delta) = \sum_{k=0}^{n} \mu_k(\epsilon) \theta_k(x) + o(\mu_n). \quad (15.18b)
\]

Then the expansion of \( \eta_k \) back to \( x \),

\[
\sum_{k=0}^{n} \lambda_k(\epsilon) \eta_k(x/\delta) = \sum_{k=0}^{n} \mu_k(\epsilon) \zeta_k(x) + o(\mu_n), \quad (15.19)
\]

is such that \( \zeta_k = \theta_k \) for \( k = 0, \ldots, n \).

The idea of matching is very important because it allows one to move smoothly from one regime to the other. The method of constructing local, but matching, expansions is therefore called Matched Asymptotic Expansions (MAE) \([81]\).

**Constructing Asymptotic Solutions** The most important application of this concept of inner and outer expansions is that approximate solutions of certain differential equations can be constructed for which the limit under a small parameter is apparently nonuniform.

The main lines of argument for constructing an MAE solution to a differential equation satisfying some boundary conditions are as follows. Suppose \( \Phi \) is given by the equation

\[
D[\Phi', \Phi](x; \epsilon) = 0 + \text{boundary conditions}, \quad (15.20)
\]

where \( \Phi' = \frac{d}{dx} \Phi \). Then we try to construct an outer solution by looking for \( \text{nontrivial degenerations} \) of \( D \) under \( \epsilon \rightarrow 0 \); i.e., find \( \mu_0(\epsilon) \) and \( v_0(\epsilon) \) such that

\[
\lim_{\epsilon \to 0} v_0^{-1}(\epsilon) D[\mu_0 \varphi_0', \mu_0 \varphi_0](x; \epsilon) = D_0[\psi_0', \varphi_0](x) = 0 \quad (15.21)
\]

has a nontrivial solution \( \varphi_0 \). A series \( \varphi = \mu_0 \varphi_0 + \mu_1 \varphi_1 + \cdots \) is constructed by repeating the process for \( D - v_0 D_0 \), etc.

Suppose the approximation is nonuniform. For example, not all boundary conditions can be satisfied. Then we start looking for an inner expansion if we have reason to believe that the nonuniformity is of boundary layer type. The presence, location, and size of the boundary layer(s) are now found by the correspondence principle, i.e., the (heuristic) idea that if \( \Phi \) behaves somehow differently in the boundary layer, then the defining equation must also be essentially different. Therefore we search for \( \text{significant degenerations} \) or \( \text{distinguished limits} \) of \( D \). These are degenerations of \( D \) under \( \epsilon \rightarrow 0 \), with scaled \( x \) and \( \Phi \), that contain a maximum of information in the sense that they result from at least two effects (i.e., terms) that balance and dominate.
Example 15.29 Under the limit $\varepsilon \to 0$, the equation $\varepsilon y' + y = \sin x$, $y(0) = 1$, reduces to $y = \sin x$ with $y(0) \neq 1$. After the scaling $x = \varepsilon \xi$, the equation reduces to the essentially different $y_\xi + y = 0$.

The next step is then to select from these distinguished limits the one(s) allowing a solution that matches with the outer solution and satisfies any applicable boundary conditions. Symbolically, we have the following algorithm:

- find $x_0, \delta(\varepsilon), \lambda(\varepsilon), \kappa(\varepsilon)$
- with $x = x_0 + \delta \xi$, $\Phi(x; \varepsilon) = \lambda(\varepsilon)\Psi(\xi; \varepsilon)$
- such that $B_0(\psi', \psi_0, \xi) = \lim_{\varepsilon \to 0} \kappa^{-1} \mathcal{D}[\delta^{-1} \lambda, \lambda, \lambda](x_0 + \delta \xi; \varepsilon)$ is a significant degeneration,
- while $B_0(\psi', \psi_0, \xi) = 0$ has a solution satisfying the boundary and matching conditions.

Again, an asymptotic expansion may be constructed inductively by repeating the argument. It is of practical importance to note that the order estimate $\lambda$ of $\Phi$ in the boundary layer is often determined a posteriori by boundary or matching conditions. We can illustrate some of the main arguments by considering

\[ \mathcal{D}[\psi', \psi](x; \varepsilon) = \varepsilon \frac{d^2 \psi}{dx^2} + \frac{d \psi}{dx} - 2x = 0, \quad \psi(0) = \psi(1) = 2. \]  

(15.22)

The leading-order outer equation is evidently (with $\mu_0 = \nu_0 = 1$)

\[ \mathcal{D}_0 = \frac{d \phi_0}{dx} - 2x = 0, \]  

(15.23)

with solution

\[ \phi_0 = x^2 + A. \]  

(15.24)

The integration constant $A$ can be determined by the boundary condition $\phi_0(0) = 2$ at $x = 0$ or $\phi_0(1) = 2$ at $x = 1$, but not both, so we expect a boundary layer at either end. By trial and error we find that no solution can be constructed if we assume a boundary layer at $x = 1$, so, inferring a boundary layer at $x = 0$, we have to use the boundary condition at $x = 1$ and we find that

\[ \phi_0 = x^2 + 1. \]  

(15.25)

The structure of the equation suggests a correction of $O(\varepsilon)$, so we try the expansion

\[ \psi = \phi_0 + \varepsilon \phi_1 + \varepsilon^2 \phi_2 + \cdots. \]  

(15.26)

For $\phi_1$ this results in the equation

\[ \frac{d \phi_1}{dx} + \frac{d^2 \phi_0}{dx^2} = 0, \]  

(15.27)

with $\phi_1(1) = 0$ (the $O(\varepsilon)$ term of the boundary condition), which has the solution

\[ \phi_1 = 2 - 2x. \]  

(15.28)
15.4. Singular Perturbation Problems

Higher orders are straightforward; i.e.,

\[ \frac{d\varphi_n}{dx} = 0, \quad \text{with } \varphi_n(1) = 0, \] (15.29)

leading to solutions \( \varphi_n \equiv 0 \). We find for the outer expansion

\[ \varphi = x^2 + 1 + 2\varepsilon(1-x) + \mathcal{O}(\varepsilon^N). \] (15.30)

We continue with the inner expansion, and find, with \( x_0 = 0, \varphi = \lambda \psi, \) and \( x = \delta \xi \), that

\[ \varepsilon \lambda \frac{d^2 \psi}{d\xi^2} + \frac{\lambda}{\delta} \frac{d\psi}{d\xi} - 2\delta \xi = 0. \] (15.31)

Both from the matching condition (\( \varphi_{\text{outer}} \to 1 \) for \( x \downarrow 0 \)) and from the boundary condition (\( \varphi(0) = 2 \)) we have to conclude that \( \varphi_{\text{inner}} = \mathcal{O}(1) \) and so \( \lambda = 1 \). Furthermore, the boundary layer only has a reason for existence if it comprises new effects not described by the outer solution. From the correspondence principle we expect that new effects are only included if \( (d^2 \psi/d\xi^2) \) is included. So \( \varepsilon \delta^{-2} \) must be at least as large as \( \delta^{-1} \), the largest of \( \delta^{-1} \) and \( \delta \).

From the principle that we should look for the equation with the richest structure, it must be exactly as large, implying a boundary layer thickness \( \delta = \varepsilon \). Thus we have \( \kappa = \varepsilon^{-1} \), and the inner equation is

\[ \frac{d^2 \psi}{d\xi^2} + \frac{d\psi}{d\xi} - 2\varepsilon^2 \xi = 0. \] (15.32)

From this equation it would \textit{seem} that we have a series expansion without the \( \mathcal{O}(\varepsilon) \) term, since the equation for this order would be the same as for the leading order. However, from matching with the outer solution, i.e.,

\[ \psi_{\text{outer}} \to 1 + 2\varepsilon + \varepsilon^2(\xi^2 - 2\xi) + \cdots \quad (x = \varepsilon \xi, \quad \xi = \mathcal{O}(1)), \] (15.33)

we see that an additional \( \mathcal{O}(\varepsilon) \) term is to be included. So we substitute the series expansion

\[ \psi = \psi_0 + \varepsilon \psi_1 + \varepsilon^2 \psi_2 + \cdots. \] (15.34)

It is a simple matter to find

\[ \frac{d^2 \psi_0}{d\xi^2} + \frac{d\psi_0}{d\xi} = 0, \quad \psi_0(0) = 2 \implies \psi_0 = 2 + A_0(e^{-\xi} - 1), \] (15.35a)
\[ \frac{d^2 \psi_1}{d\xi^2} + \frac{d\psi_1}{d\xi} = 0, \quad \psi_1(0) = 0 \implies \psi_1 = A_1(e^{-\xi} - 1), \] (15.35b)
\[ \frac{d^2 \psi_2}{d\xi^2} + \frac{d\psi_2}{d\xi} = 2\xi, \quad \psi_2(0) = 0 \implies \psi_2 = \xi^2 - 2\xi + A_2(e^{-\xi} - 1), \] (15.35c)

where the constants \( A_0, A_1, A_2, \ldots \) are to be determined from the matching condition that the inner and outer solutions should be asymptotically equivalent in the region of overlap.

We can follow the method of intermediate variables and rewrite outer expansion (15.30) and
inner expansion (15.34) in terms of an intermediate variable $x = \eta(\varepsilon)\sigma$, where $\varepsilon \ll \eta \ll 1$, and reexpand as follows:

\begin{align*}
2 + A_0(e^{-\xi} - 1) + \varepsilon A_1(e^{-\xi} - 1) + \varepsilon^2(\xi^2 - 2\xi + A_2(e^{-\xi} - 1)) + \mathcal{O}(\varepsilon^3) \\
\quad \approx 2 - A_0 - \varepsilon A_1 + \eta^2\sigma^2 - 2\varepsilon \eta \sigma - \varepsilon^2 A_2 + \mathcal{O}(\varepsilon^3).
\end{align*}

(15.36a)

Alternatively, we can follow Van Dyke’s matching rule and rewrite outer expansion (15.30) in terms of inner variable $\xi$ and inner expansion (15.34) in terms of outer variable $x$, reexpand, and rewrite the result in $x$, giving

\begin{align*}
2 + A_0(e^{-\xi} - 1) + \varepsilon A_1(e^{-\xi} - 1) + \varepsilon^2(\xi^2 - 2\xi + A_2(e^{-\xi} - 1)) + \mathcal{O}(\varepsilon^3) \\
\quad \approx 2 - A_0 - \varepsilon A_1 + x^2 - 2\varepsilon x - \varepsilon^2 A_2 + \mathcal{O}(\varepsilon^3).
\end{align*}

(15.36b)

In either case the resulting reduced expressions, (15.36a) and (15.36b) (respectively (15.36c) and (15.36d)), must be functionally equivalent. A full matching is thus obtained if we choose $A_0 = 1$, $A_1 = -2$, and $A_2 = 0$.

**Example 15.30** (integration across a boundary layer) The same ideas of overlap and intermediate variables are exploited for an integral across a boundary layer. Take, e.g., the above function $\phi(x; \varepsilon)$ on $x \in [0, 1]$. We break up the integration interval at a point in the region of overlap, say at $x = \eta(\varepsilon)$. With matching outer and inner expansions denoted as before we obtain

\[
\int_0^1 \phi(x; \varepsilon) \, dx = \int_0^{\eta(\varepsilon)} \phi(x; \varepsilon) \, dx + \int_{\eta(\varepsilon)}^1 \phi(x; \varepsilon) \, dx = \varepsilon \int_0^{\eta(\varepsilon)} \psi(\xi; \varepsilon) \, d\xi + \int_0^1 \phi(x; \varepsilon) \, dx
\]

The result will contain terms depending on the auxiliary function $\eta$, but these will disappear after reexpanding the result up to $\mathcal{O}(\varepsilon^2)$ while taking $\eta$ asymptotically about halfway $\mathcal{O}(\varepsilon)$ and $\mathcal{O}(1)$, say $\eta = \sqrt{\varepsilon}$.

**Example 15.31** (Prandtl’s boundary layer analysis) The start of modern boundary layer theory is Prandtl’s analysis of uniform incompressible low-viscosity flow along a flat plate. Consider the stationary two-dimensional version of (7.3)

\[
u_x + v_y = 0, \quad u\Delta x + u\Delta y = -p_x + \varepsilon(u_{xx} + u_{yy}), \quad u\Delta x + u\Delta y = -p_y + \varepsilon(v_{xx} + v_{yy}),
\]

with $\varepsilon = Re^{-1}$ small, subject to boundary conditions $u = v = 0$ at $y = 0, 0 < x < 1$, and outer solution for $y = \mathcal{O}(1)$ to leading order given by $(u, v, p) = (1, 0, 0)$. When we scale $x = X$, $y = \varepsilon^0 Y$, $u = U$, $v = \varepsilon^m V$, and $p = \varepsilon^m P$, we find

\[
U_X + \varepsilon^{m-n} V_Y = 0, \quad U U_X + \varepsilon^{m-n} V U_Y = -\varepsilon^2 P_X + \varepsilon U_{XX} + \varepsilon^{1-2n} U_{YY}, \quad \varepsilon^m U X_X + \varepsilon^{2m-n} V Y_Y = -\varepsilon^{1-m} P_Y + \varepsilon^{1+m} V_{XX} + \varepsilon^{1+2m-n} V_{YY},
\]
This yields the distinguished limits \( m = n = \frac{1}{2} \) and \( k = 1 \), with the significant degeneration
\[
U_X + V_Y = 0, \quad U_U + V_Y = U_Y,
\]
known as Prandtl's boundary layer equations. The same equations, but with other boundary conditions, are valid in Goldstein's viscous wake \( x > 1, y = \mathcal{O}(x^{1/2}) \). The trailing-edge region around \( x = 1, y = 0 \) is far more complicated. Here the boundary layer structure consists of three layers
\[
y = \mathcal{O}(\varepsilon^{5/8}), \mathcal{O}(\varepsilon^{7/8}), \mathcal{O}(\varepsilon^{3/8}) \text{ within } x - 1 = \mathcal{O}(\varepsilon^{3/8}).
\]
This is known as Stewartson's triple deck.

Logarithmic switchback It is not always evident from just the structure of the equation what the necessary expansion will look like. Sometimes it is well concealed and we are only made aware of an invalid initial choice by a matching failure. In fact, it is also the matching process itself that reveals the required sequence of scaling functions. An example of such a back reaction is known as logarithmic switchback.

Consider the following problem for \( y = y(x; \varepsilon) \) on the unit interval:
\[
\varepsilon y'' + x(y' - y) = 0, \quad 0 < x < 1, \quad y(0; \varepsilon) = 0, \quad y(1; \varepsilon) = e. \quad (15.37)
\]
The outer solution appears to have the expansion
\[
y(x; \varepsilon) = y_0(x) + \varepsilon y_1(x) + \varepsilon^2 y_2(x) + \mathcal{O}(\varepsilon^3). \quad (15.38)
\]
By trial and error, we find that the boundary layer appears to be located near \( x = 0 \), so the governing equations and boundary conditions are then
\[
y_0' - y_0 = 0, \quad y_0(0) = 0, \quad y_0(1) = e, \quad (15.39a)
\]
\[
y_n' - y_n = -x^{-1} y_{n-1}'' , \quad y_n(0) = 0, \quad y_n(1) = 0, \quad (15.39b)
\]
with general solution
\[
y_n(x) = A_n e^x + \int_0^x z^{-1} e^{x-z} y_{n-1}''(z) \, dz \quad (15.40a)
\]
such that
\[
y_0(x) = e^x, \quad (15.40b)
\]
\[
y_1(x) = -e^x \ln(x), \quad (15.40c)
\]
\[
y_2(x) = e^x \left( \frac{1}{2} \ln(x)^2 + \frac{3}{2} - 2x^{-1} + \frac{1}{2} x^{-2} \right), \quad (15.40d)
\]
etc. The boundary layer thickness is found from the assumed scaling \( x = \varepsilon^m t \) and from noting that \( y = \mathcal{O}(1) \) because of the matching with the outer solution. This leads to the significant degeneration of \( m = \frac{1}{2} \), or \( x = \varepsilon^{\frac{1}{2}} t \). The boundary layer equation for \( y(x; \varepsilon) = Y(t; \varepsilon) \) is thus
\[
Y'' + t Y' - \varepsilon^{\frac{1}{2}} t Y = 0, \quad Y(0; \varepsilon) = 0. \quad (15.41)
\]
Chapter 15. Perturbation Methods

The obvious choice of expansion of \( Y \) in powers of \( \varepsilon^\frac{1}{2} \) is not correct, as the found solution does not match the outer solution. Therefore we consider the outer solution in more detail for small \( x \). When \( x = \varepsilon^\frac{1}{2} t \), we have for the outer solution

\[
y(\varepsilon^\frac{1}{2} t; \varepsilon) = 1 + \varepsilon^\frac{1}{2} t + \varepsilon \left( -\frac{1}{2} \ln \varepsilon + \frac{1}{2} t^2 - \ln t + \frac{1}{2} t^{-2} + \cdots \right) + \mathcal{O}(\varepsilon^2 \ln \varepsilon). \tag{15.42}
\]

(The dots indicate powers of \( t^{-2} \) that appear with higher-order \( y_n \).) So we apparently need at least

\[
Y(t; \varepsilon) = Y_0(t) + \varepsilon^\frac{1}{2} Y_1(t) + \varepsilon \ln(\varepsilon) Y_2(t) + \varepsilon Y_3(t) + o(\varepsilon), \tag{15.43}
\]

with equations and boundary conditions

\[
\begin{align*}
Y_0'' + t Y_0' &= 0, & Y_0(0) &= 0, \tag{15.44a} \\
Y_1'' + t Y_1' &= t Y_0, & Y_1(0) &= 0, \tag{15.44b} \\
Y_2'' + t Y_2' &= 0, & Y_2(0) &= 0, \tag{15.44c} \\
Y_3'' + t Y_3' &= t Y_1, & Y_3(0) &= 0. \tag{15.44d}
\end{align*}
\]

e等。因此内层展开式给出

\[
Y_0(t) = A_0 \text{erf} \left( \frac{t}{\sqrt{2}} \right), \tag{15.45a}
\]

\[
Y_1(t) = A_1 \text{erf} \left( \frac{t}{\sqrt{2}} \right) + A_0 \left[ t \text{erf} \left( \frac{t}{\sqrt{2}} \right) + 2 \left( \frac{2}{\pi} \right)^{\frac{1}{2}} (e^{-\frac{1}{2} t^2} - 1) \right], \tag{15.45b}
\]

\[
Y_2(t) = A_2 \text{erf} \left( \frac{t}{\sqrt{2}} \right), \tag{15.45c}
\]

\[
Y_3(t) = A_3 \text{erf} \left( \frac{t}{\sqrt{2}} \right) + \int_0^t e^{-\frac{1}{2} z^2} \int_0^z e^{\frac{1}{2} \xi^2} Y_1(\xi) \, d\xi \, dz. \tag{15.45d}
\]

Unfortunately, \( Y_3 \) cannot be expressed in closed form. However, for demonstration purposes it is sufficient to derive the behaviour of \( Y_3 \) for large \( t \). As \( \text{erf}(z) \to 1 \) exponentially fast for \( z \to \infty \), we obtain

\[
Y_1(t) = A_0 t + A_1 - 2 \left( \frac{2}{\pi} \right)^{\frac{1}{2}} A_0 + \text{exponentially small terms}.
\]

If \( Y_3 \) behaves algebraically for large \( t \), then \( t Y_3' \gg Y_3'' \), so \( Y_3 = Y_1 - t^{-1} Y_3' \simeq A_0 t \). By successive substitution it follows that

\[
Y_3(t) = \frac{1}{2} A_0 t^2 + \left( A_1 - 2 \left( \frac{2}{\pi} \right)^{\frac{1}{2}} A_0 \right) t - A_0 \ln(t) + \cdots
\]

To match the inner solution we introduce the intermediate variable \( \eta = \varepsilon^{\frac{1}{2} - \alpha} x = \varepsilon^{\frac{1}{2} - \alpha} t \),
where $0 < \alpha < \frac{1}{2}$, and compare it with expression (15.42). We have

$$A_0 + \varepsilon^{\frac{1}{2} + a} \left( A_1 - 2 \left( \frac{2}{\pi} \right)^{\frac{1}{2}} A_0 \right) + \varepsilon^{\alpha} A_0 \eta + \varepsilon \ln(\varepsilon) A_2 + \frac{1}{2} \varepsilon^{2\alpha} A_0 \eta^2$$

$$+ \varepsilon^{\frac{1}{2} + a} \left( A_1 - 2 \left( \frac{2}{\pi} \right)^{\frac{1}{2}} A_0 \right) \eta - \varepsilon A_0 \ln \eta + \varepsilon \left( \frac{1}{2} - \alpha \right) A_0 \ln \varepsilon$$

$$\equiv 1 + \varepsilon^{\alpha} \eta + \frac{1}{2} \varepsilon^{2\alpha} \eta^2 - \varepsilon \ln \eta - \alpha \varepsilon \ln(\varepsilon) + \frac{1}{2} \varepsilon^{2 - 2\alpha} \eta^{-2}. \quad (15.46)$$

Noting that $2 - 2\alpha > 1$, we find a full matching with

$$A_0 = 1, \quad A_1 = 2 \left( \frac{2}{\pi} \right)^{\frac{1}{2}}, \quad A_2 = -\frac{1}{2}. \quad (15.47)$$

This problem is an example where intermediate matching is preferable.

**The Role of Matching** It is important to note whether a matching is possible at all! Only part of the terms can be matched by selection of the undetermined constants. Other terms are already equal, without free constants, and there is no way to repair a possibly incomplete matching here. This is an important consistency check on the found solution, at least as long as no real proof is available. If no matching appears to be possible, one of the assumptions made with the construction of the solution almost certainly has to be reconsidered. Particularly notorious are logarithmic singularities of the outer solution, as we saw above. See, for other examples, [86].

In summary, matching of inner and outer expansions plays an important role in the following ways:

(i) It provides information about the sequence of order (gauge) functions $\{\mu_k\}$ and $\{\lambda_k\}$ of the expansions.

(ii) It allows us to determine unknown constants of integration.

(iii) It provides a check on the consistency of the solution, giving us confidence in the correctness.

**15.4.2 Multiple Scales**

Suppose a function $\varphi(x; \varepsilon)$ depends on more than one length scale acting together, e.g., $x$, $\varepsilon x$, and $\varepsilon^2 x$. Then the function does not have a regular expansion on the full domain of interest, $x \leq O(\varepsilon^{-2})$ say. It is not possible to bring these different length scales together by a simple coordinate transformation, like in the method of slow variation or the Lindstedt–Poincaré method, or to split up our domain in subdomains, like in the MAE method. Therefore we have to find another way to construct asymptotic expansions that is valid in the full domain of interest. The approach that is followed in the method of multiple scales is at first sight rather radical: the various length scales are temporarily considered as independent variables—$x_1 = x$, $x_2 = \varepsilon x$, $x_3 = \varepsilon^2 x$—and the original function $\varphi$ is identified with a more general function $\psi(x_1, x_2, x_3; \varepsilon)$ depending on a higher-dimensional independent variable.
Example 15.32
\[ \varphi(x; \varepsilon) = A(e) e^{-\varepsilon x} \cos(x + \theta(\varepsilon)) \]
becomes
\[ \psi(x_1, x_2; \varepsilon) = A(e) e^{-x_2} \cos(x_1 + \theta(\varepsilon)). \]

Since this identification is not unique, we may add constraints such that this auxiliary function \( \psi \) does have a Poincaré expansion on the full domain of interest. After having constructed this expansion, we may associate it with the original function along the line \( x_1 = x, x_2 = \varepsilon x, x_3 = \varepsilon^2 x \).

This technique, utilizing this difference between small-scale and large-scale behaviour is the method of multiple scales. As with most approximation methods, this method has grown from practice and works well for certain types of problems. Typically, the multiple scale method is applicable to problems with, on the one hand, a certain global quantity (energy, power), which is conserved or almost conserved, controlling the amplitude, and, on the other hand, two rapidly interacting quantities (kinetic and potential energy) controlling the phase. Usually, this describes slowly varying waves affected by small effects during a long time. Intuitively, it is clear that over a short distance (a few wavelengths) the wave only sees constant conditions and will propagate approximately as in the constant case, but over larger distances it will somehow have to change its shape in accordance with its new environment.

We will illustrate the method by considering a damped harmonic oscillator
\[
\frac{d^2 y}{dt^2} + 2\varepsilon \frac{dy}{dt} + y = 0 \quad (t \geq 0), \quad y(0) = 0, \quad \frac{dy(0)}{dt} = 1, \quad (15.48)
\]
with \( 0 < \varepsilon \ll 1 \). The exact solution is readily found to be
\[
y(t) = e^{-\varepsilon t} \sin \left( \sqrt{1 - \varepsilon^2} t \right), \quad (15.49)
\]
A naive approximation of this \( y(t) \) for small \( \varepsilon \) and fixed \( t \) is
\[
y(t) = \sin t - \varepsilon t \sin t + \mathcal{O}(\varepsilon^2), \quad (15.50)
\]
which appears to be useful for \( t = \mathcal{O}(1) \) only. For large \( t \) the approximation becomes incorrect:
1. If \( t \geq \mathcal{O}(\varepsilon^{-1}) \), the second term is of equal importance to, or larger than, the first term and nothing is left over of the slow exponential decay.
2. If \( t \geq \mathcal{O}(\varepsilon^{-2}) \), the phase has an error of \( \mathcal{O}(1) \) or larger, giving an approximation of which even the sign may be in error.

We would obtain a far better approximation if we adopted two different time variables, \( T = \varepsilon t \) and \( \tau = \sqrt{1 - \varepsilon^2} t \), and changed to \( y(t; \varepsilon) = Y(\tau, T; \varepsilon) \), where
\[
Y(\tau, T; \varepsilon) = e^{-T} \frac{\sin(\tau)}{\sqrt{1 - \varepsilon^2}}.
\]

It is easily verified that a Taylor series of \( Y \) in \( \varepsilon \) yields a regular expansion for all \( t \).

If we construct a straightforward approximate solution directly from (15.48), we will get the same approximation as in (15.50), which is too limited for most applications.
however, knowing the character of the error, we may try to avoid it and look for the auxiliary function \( Y \) instead of \( y \). As we, in general, do not know the occurring time scales, their determination becomes part of the problem.

Suppose we can expand
\[
y(t; \varepsilon) = y_0(t) + \varepsilon y_1(t) + \varepsilon^2 y_2(t) + \cdots.
\]
(15.51)

Substituting in (15.48) and collecting equal powers of \( \varepsilon \) gives

\[
O(\varepsilon^0): \quad \frac{d^2 y_0}{dt^2} + y_0 = 0 \quad \text{with} \quad y_0(0) = 0, \quad \frac{dy_0(0)}{dt} = 1,
\]

\[
O(\varepsilon^1): \quad \frac{d^2 y_1}{dt^2} + y_1 = -2 \frac{dy_0}{dt} \quad \text{with} \quad y_1(0) = 0, \quad \frac{dy_1(0)}{dt} = 0.
\]

We then find

\[
y_0(t) = \sin t, \quad y_1(t) = -t \sin t, \quad \text{etc.}
\]

which indeed reproduces expansion (15.50). The straightforward Poincaré-type expansion (15.51) breaks down for large \( t \) when \( \varepsilon t \geq O(1) \). It is important to note that this is caused by the fact that any \( y_n \) is excited in its eigenfrequency (by the “source”-terms \(-2dy_{n-1}/dt\)), resulting in resonance. We recognise the algebraically generated growing terms of the type \( t^n \sin t \) and \( t^n \cos t \) called secular terms (Definition 15.26). Apart from being of limited validity, the expansion reveals nothing of the real structure of the solution, and we change our strategy to look for an auxiliary function dependent on different time scales. We start with the hypothesis that, besides a fast time scale \( t \), we have the slow time scale

\[
T := \varepsilon t.
\]
(15.52)

Then we identify the solution \( y \) with another suitably chosen function \( Y \) that depends on both variables \( t \) and \( T \):

\[
Y(t, T; \varepsilon) := y(t; \varepsilon).
\]
(15.53)

There exist infinitely many functions \( Y(t, T; \varepsilon) \) that are equal to \( y(t, \varepsilon) \) along the line \( T = \varepsilon t \) in \((t, T)\) space. So we now have some freedom to prescribe additional conditions. It is natural to think of conditions to be chosen such that no unwelcome secular terms occur when we construct an approximation.

Since the time derivatives of \( y \) turn into partial derivatives of \( Y \), i.e.,

\[
\frac{dy}{dt} = \frac{\partial Y}{\partial t} + \varepsilon \frac{\partial Y}{\partial T},
\]
(15.54)
equation (15.48) becomes for \( Y \)

\[
\frac{\partial^2 Y}{\partial t^2} + Y + 2\varepsilon \left( \frac{\partial Y}{\partial t} + \frac{\partial^2 Y}{\partial t \partial T} \right) + \varepsilon^2 \left( \frac{\partial^2 Y}{\partial T^2} + 2 \frac{\partial Y}{\partial T} \right) = 0.
\]
(15.55)

We assume the expansion

\[
Y(t, T; \varepsilon) = Y_0(t, T) + \varepsilon Y_1(t, T) + \varepsilon^2 Y_2(t, T) + \cdots
\]
(15.56)
and substitute this into (15.55) to obtain to leading order
\[
\frac{\partial^2 Y_0}{\partial t^2} + Y_0 = 0,
\]
\[
\frac{\partial^2 Y_1}{\partial t^2} + Y_1 = -2 \frac{\partial Y_0}{\partial t} - 2 \frac{\partial^2 Y_0}{\partial t \partial T},
\]
with initial conditions
\[
Y_0(0, 0) = 0, \quad \frac{\partial}{\partial t} Y_0(0, 0) = 1,
\]
\[
Y_1(0, 0) = 0, \quad \frac{\partial}{\partial t} Y_1(0, 0) = -\frac{\partial}{\partial T} Y_0(0, 0).
\]
The solution for $Y_0$ is easily found to be
\[
Y_0(t, T) = A_0(T) \sin(t + \theta_0(T))
\]
with $A_0(0) = 1, \theta_0(0) = 0$. (15.57)

This gives a right-hand side for the $Y_1$ equation of
\[
-2 \left( A_0 + \frac{\partial A_0}{\partial T} \right) \cos(t + \theta_0) + 2 A_0 \frac{\partial \theta_0}{\partial T} \sin(t + \theta_0).
\]

No secular terms occur (no resonance between $Y_1$ and $Y_0$) if these terms vanish:
\[
A_0 + \frac{\partial A_0}{\partial T} = 0 \text{ yielding } A_0 = e^{-T}, \quad \frac{\partial \theta_0}{\partial T} = 0 \text{ yielding } \theta_0 = 0. \quad (15.58)
\]

Putting this together, we have indeed constructed an approximation of (15.49) that is valid for $t \leq O(\varepsilon^{-1})$:
\[
y(t; \varepsilon) = e^{-\varepsilon t} \sin t + O(\varepsilon).
\]
(15.59)

Note (this is typical of this approach) that we determined $Y_0$ only on the level of $Y_1$ without having to solve $Y_1$ itself.

The present approach is by and large the multiple scale technique in its simplest form. Variations on this theme are sometimes necessary. For example, we have not completely gotten rid of secular terms. On a longer time scale ($t = O(\varepsilon^{-2})$) we again have resonance in $Y_2$ because of the “source” $e^{-T} \sin t$, yielding terms $O(\varepsilon^2 t)$. We see that a second time scale $T_2 = \varepsilon^2 t$ is necessary. From the exact solution we may infer that these longer time scales are not really independent and it may be worth trying a fast time of strained coordinate type: $\tau = \omega(\varepsilon)t = (1 + \varepsilon^2 \omega_1 + \varepsilon^4 \omega_2 + \cdots) t$. In the present example we would recover $\omega(\varepsilon) = \sqrt{1 - \varepsilon^2}$.

The method fails when the slow variation is due to external effects, like a slowly varying problem parameter, as is demonstrated by the next example.

**Example 15.33** Consider the problem
\[
\ddot{x} + \kappa(\varepsilon t)^2 \dot{x} = 0, \quad x(0; \varepsilon) = 1, \quad \dot{x}(0; \varepsilon) = 0,
\]
where $\kappa = O(1)$. It seems plausible to assume two time scales: a fast one $O(\varepsilon^{-1}) = O(1)$ and a slow one $O(\varepsilon^{-1})$. So we introduce besides to $t$ the slow scale $T = \varepsilon t$ and rewrite $x(t; \varepsilon) = X(t, T; \varepsilon)$. We expand $X = X_0 + \varepsilon X_1 + \cdots$, and obtain $X_0 = A_0(T) \cos(\kappa(T)t - \theta_0(T))$. Suppressing secular terms in the equation for $X_1$ requires $A_0' = \kappa' t - \theta_0' = 0$, which is impossible. □
Here the fast time scale is slowly varying itself and the fast variable is to be strained locally by a suitable strain function as follows:

\[ \tau = \int_{t'}^T \omega(\varepsilon t'; \varepsilon) \, dt' = \frac{1}{\varepsilon} \int_{\tau}^{T} \omega(z; \varepsilon) \, dz, \quad \text{where } T = \varepsilon t, \quad (15.60) \]

while for \( x(t; \varepsilon) = X(\tau, T; \varepsilon) \) we have

\[ \dot{x} = \omega X_T + \varepsilon X_T \quad \text{and} \quad \ddot{x} = \omega^2 X_{TT} + \varepsilon \omega T X_T + 2\varepsilon \omega X_{TT} + \varepsilon^2 X_{TT}. \quad (15.61) \]

**Example 15.34** Reconsider 15.33. At the solution of \( X = X_0 + \varepsilon X_1 + \cdots \) and \( \omega = \omega_0 + \varepsilon \omega_1 + \cdots \), we obtain

\[ \omega_0^2 X_{0r} + \kappa^2 X_0 = 0, \]

\[ \omega_0^2 X_{1r} + \kappa^2 X_1 - 2\omega_0 \omega_1 X_{0r} - \omega_0 T X_{0t} - 2\omega_0 X_{0TT}. \quad (\ast) \]

The leading-order solution is \( X_0 = A_0(T) \cos(\lambda(T) \tau - \theta_0(T)) \), where \( \lambda = \kappa/\omega_0 \). The right-hand side of \((\ast)\) is then

\[ 2\omega_0^2 A_0(A_0 \lambda + \lambda_T \tau - \theta_0) \cos(\lambda \tau - \theta_0) + (A_0 \lambda)^{-1} (\omega_0 A_0^2 \lambda^2) \sin(\lambda \tau - \theta_0). \]

Suppression of secular terms requires \( \lambda_T = 0 \). Without loss of generality we can take \( \lambda = 1 \) or \( \omega_0 = \kappa \). Then we need \( \omega_0 = \theta_0 \), which just yields that \( \lambda, \tau - \theta_0 = \tau - \theta_0 = \varepsilon^{-1} \int_{0}^{T} \omega(z) \, dz \) and \( \int_{0}^{T} \omega_0(z) \, dz = \varepsilon^{-1} \int_{0}^{T} \omega_0(z) \, dz + \mathcal{O}(\varepsilon) \). In other words, we may just as well take \( \omega_0 = 0 \) and \( \theta_0 = \text{constant} \). Finally, we have \( \omega_0 A_0^2 \lambda^2 = \kappa A_0^2 = \text{constant} \). \( \square \)

For linear wave-type problems we may anticipate the structure of the solution and assume the WKB hypothesis or ray approximation

\[ y(t; \varepsilon) = A(T; \varepsilon) e^{i \varepsilon^{-1} \int_{0}^{T} \omega(\tau; \varepsilon) \, d\tau}. \quad (15.62) \]

The method is again illustrated by the example of the damped oscillator (15.48), but now in complex form, so we consider the real part of (15.62). After substitution and suppressing the exponential factor, we get

\[ (1 - \omega^2) A + i \varepsilon \left( 2\omega \frac{\partial A}{\partial T} + \frac{\partial \omega}{\partial T} A + 2\omega A \right) + \varepsilon^2 \left( \frac{\partial^2 A}{\partial T^2} + 2 \frac{\partial A}{\partial T} \right) = 0, \quad (15.63) \]

\[ \text{Re}(A) = 0, \quad \text{Re}(i \omega A + \varepsilon A') = 1 \quad \text{at } T = 0. \]

Unlike in the multiple scales method the secular terms will not be explicitly suppressed, at least not to leading order. The underlying additional condition here is that the solution of the present type exists and that each higher-order correction is no more secular than its predecessor. The solution is expanded as

\[ A(T; \epsilon) = A_0(T) + \epsilon A_1(T) + \epsilon^2 A_2(T) + \cdots, \]
\[ \omega(T; \epsilon) = \omega_0(T) + \epsilon^2 \omega_2(T) + \cdots. \quad (15.64) \]

Note that \( \omega_1 \) may be set to zero since the factor \( \exp(i \int_{0}^{T} \omega_1(\tau) \, d\tau) \) may be incorporated in \( A \). By a similar argument, i.e., by reexpanding the exponential for small \( \epsilon \), all other terms \( \omega_2, \omega_3, \ldots \) could be absorbed by \( A \) (this is often done). This is perfectly acceptable for
the time scale \( T = \mathcal{O}(1) \), but for larger times we will not be able to suppress higher-order secular terms. So we will find it more convenient to include these terms and use them whenever convenient.

We substitute the expansions and collect equal powers of \( \varepsilon \) to obtain, to \( \mathcal{O}(\varepsilon^0) \), that
\[
(1 - \omega_0^2)A_0 = 0, \quad (15.65a)
\]
with solution \( \omega_0 = 1 \) (or \(-1\), but that is equivalent for the result). To \( \mathcal{O}(\varepsilon^1) \) we then have
\[
A'_0 + A_0 = 0 \quad \text{with} \quad \text{Re}(A_0) = 0, \quad \text{Im}(\omega_0 A_0) = -1 \quad \text{at} \quad T = 0, \quad (15.65b)
\]
with solution \( A_0 = -i e^{-T} \). To order \( \mathcal{O}(\varepsilon^2) \) the equation reduces to
\[
A'_1 + A_1 = -i \left( \frac{1}{2} + \omega_2 \right) e^{-T}, \quad \text{with} \quad \text{Re}(A_1) = 0, \quad \text{Im}(\omega_0 A_1) = \text{Re}(A'_0) \quad \text{at} \quad T = 0, \quad (15.65c)
\]
with solution
\[
\omega_2 = -\frac{1}{2}, \quad A_1 = 0. \quad (15.65d)
\]
Note that if we had chosen \( \omega_2 = 0 \), the solution would have been \( A_1 = -\frac{1}{2} T e^{-T} \). Although it is by itself correct for \( T = \mathcal{O}(1) \), it renders the asymptotic hierarchy invalid for \( T \geq \mathcal{O}(1/\varepsilon) \) and is therefore better avoided. The solution that emerges is indeed consistent with the exact solution.

**Example 15.35 (the air-damped resonator)** In dimensionless form this is given by
\[
\frac{d^2 y}{dt^2} + \varepsilon \frac{dy}{dt} \left| \frac{dy}{dt} \right| + y = 0, \quad \text{with} \quad y(0) = 1, \quad \frac{dy}{dt}(0) = 0. \quad (\ast)
\]

By rewriting the equation in the form
\[
\frac{d}{dt} \left[ \frac{1}{2}(y')^2 + \frac{1}{2} y^2 \right] = -\varepsilon (y')^2 |y'|
\]
and assuming that \( y \) and \( y' = \mathcal{O}(1) \), we may infer that the damping acts on a time scale of \( \mathcal{O}(\varepsilon^{-1}) \). So we conjecture the presence of the slow time variable \( T = \varepsilon t \) and introduce a new dependent variable \( Y \) that depends on both \( t \) and \( T \). We have
\[
T = \varepsilon t, \quad y(t; \varepsilon) = Y(t, T; \varepsilon), \quad \frac{dy}{dt} = \frac{\partial Y}{\partial t} + \varepsilon \frac{\partial Y}{\partial T},
\]
and we obtain for \((\ast)\)
\[
\frac{\partial^2 Y}{\partial t^2} + Y + \varepsilon \left( \frac{2 \partial^2 Y}{\partial T \partial t} + \frac{\partial Y}{\partial t} \frac{\partial Y}{\partial t} \right) + \mathcal{O}(\varepsilon^2) = 0,
\]
\[
Y(0, 0; \varepsilon) = 1, \quad \left( \frac{\partial}{\partial t} + \varepsilon \frac{\partial}{\partial T} \right) Y(0, 0; \varepsilon) = 0.
\]
The error of \( \mathcal{O}(\varepsilon^2) \) results from the approximation \( \frac{\partial}{\partial t} Y + \varepsilon \frac{\partial}{\partial T} Y = \frac{\partial}{\partial t} Y + \mathcal{O}^{\ast} \) and is of course only valid outside a small neighbourhood of the points where \( \frac{\partial}{\partial t} Y = 0 \). We expand
\[
Y(t, T; \varepsilon) = Y_0(t, T) + \varepsilon Y_1(t, T) + \mathcal{O}(\varepsilon^2)
\]
15.4. Singular Perturbation Problems

489

to find for the leading order

\[ \frac{\partial^2 Y_0}{\partial t^2} + Y_0 = 0, \quad \text{with} \quad Y_0(0, 0) = 1, \quad \left. \frac{\partial}{\partial t} Y_0 \right|_{(0, 0)} = 0. \]

The solution is given by

\[ Y_0 = A_0(T) \cos(t - \Theta_0(T)), \quad \text{where} \quad A_0(0) = 1, \quad \Theta_0(0) = 0. \]

For the first order we have the equation

\[ \frac{\partial^2 Y_1}{\partial t^2} + Y_1 = -2 \frac{\partial^2 Y_0}{\partial T \partial t} - \left. \frac{\partial Y_0}{\partial t} \right| \left. \frac{\partial Y_0}{\partial t} \right| = -2 \frac{d}{dT} Y_0 \cos(t - \Theta_0) + 2A_0 \frac{d\Theta_0}{dT} \cos(t - \Theta_0) + A_0^2 \sin(t - \Theta_0) \sin(t - \Theta_0), \]

with corresponding initial conditions. The secular terms are suppressed if the first harmonics of the right-hand side cancel. For this we use the Fourier series expansion

\[ \sin(t) \sin(t) = -\frac{8}{\pi} \sum_{n=0}^{\infty} \frac{\sin(2n+1)t}{(2n-1)(2n+1)(2n+3)}. \]

We obtain the equations

\[ 2 \frac{dA_0}{dT} + \frac{8}{3\pi} A_0^2 = 0 \quad \text{and} \quad \frac{d\Theta_0}{dT} = 0, \]

with solution \( \Theta_0(T) = 0 \) and

\[ A_0(T) = \frac{1}{1 + \frac{4}{3\pi} T}. \]

Altogether we have the approximate solution

\[ y(t; \epsilon) = \frac{\cos(t)}{1 + \frac{4}{3\pi} \epsilon t} + O(\epsilon). \]

This approximation appears to be remarkably accurate. See Figure 15.4, where plots, made for a parameter value of \( \epsilon = 0.1 \), of the approximate and a numerically “exact” solution are hardly distinguishable. A maximum difference of 0.03 is found.

\[ \text{Figure 15.4. Plots of the approximate solution and a numerically “exact” solution } y(t; \epsilon) \text{ of the air-damped resonator problem for } \epsilon = 0.1. \]
Example 15.36 (sound propagation in a slowly varying duct) Consider a hard-walled circular cylindrical duct with a slowly varying diameter described in polar coordinates \((x, r, \theta)\) as \(r = R(\epsilon x)\) with \(\epsilon\) a dimensionless small parameter. In this duct we have an acoustic medium with constant mean pressure and sound speed \(c\). Sound waves of circular frequency \(\omega\) are most conveniently written in complex form, so the physical pressure perturbation is given by \(\text{Re}(p(x) e^{i\omega t})\), where \(p\) is the complex pressure amplitude described by Helmholtz’s equation (see Example 1.8iii):

\[
\nabla^2 p + \kappa^2 p = 0.
\]

\(\kappa = \omega/c\) is the free-field wave number, and the boundary condition of a vanishing normal gradient at the wall yields

\[
\frac{\partial p}{\partial r} - \epsilon R'(\epsilon x) \frac{\partial p}{\partial x} = 0 \quad \text{at} \quad r = R(\epsilon x).
\]

For constant \(R\) and constant \(\kappa\) the general solution can be built up from a sum of right- and left-running modes (see Example 7.21) of the type

\[
p = \sum_{m=-\infty}^{\infty} e^{-i m \theta} \sum_{\mu=1}^{\infty} J_m(\alpha_m \mu r) \left( A_{m \mu} e^{-i \kappa_m \mu x} + B_{m \mu} e^{i \kappa_m \mu x} \right),
\]

\(\alpha_m = j'_{m \mu}/R\), \(\kappa_m^2 = \kappa^2 - \alpha_m^2\), \(\text{Re}(\kappa_m) \geq 0\), \(\text{Im}(\kappa_m) \leq 0\).

\(J_m\) denotes the \(m\)th order Bessel function and \(J'_m(j'_{m \mu}) = 0\). For the present problem we consider only a single mode and we assume, following the previous section, that the solution for the straight duct is locally close to the one for the varying duct. We introduce the slow variable \(X = \epsilon x\) so that \(R = R(X)\), and we seek a solution of slowly varying modal type:

\[
p = A(X, r; \epsilon) e^{-i m \theta} e^{-i \gamma(\epsilon \xi)}.
\]

For simplicity we assume \(\gamma\) independent of \(\epsilon\). Since

\[
\frac{\partial^2 p}{\partial x^2} = \left( -\gamma^2 A - 2i \epsilon \gamma' \frac{\partial A}{\partial X} - i \epsilon \gamma' A + \epsilon^2 \frac{\partial^2 A}{\partial X^2} \right) \exp(\cdots)
\]

we have for (\(\ast\))

\[
\left[ -\gamma^2 A - 2i \epsilon \gamma' \frac{\partial A}{\partial X} - i \epsilon \gamma' A + \gamma^2 \frac{\partial^2 A}{\partial X^2} + \frac{1}{\epsilon} \frac{\partial A}{\partial r} - \frac{m^2}{r^2} A + \kappa^2 A \right] \exp(\cdots) = 0.
\]

After we suppress the exponential factor, this expression and the accompanying boundary condition are up to order \(O(\epsilon)\),

\[
\mathcal{L}[A] = \frac{i \epsilon}{A} \frac{\partial}{\partial X} (\gamma A^2), \quad \frac{\partial A}{\partial r} + i \epsilon R' \gamma A = 0 \quad \text{at} \quad r = R(X). \quad (\dagger)
\]

Here we introduced for short the Bessel-type operator

\[
\mathcal{L}[A] = \frac{\partial^2 A}{\partial r^2} + \frac{1}{r} \frac{\partial A}{\partial r} + \left( \kappa^2 - \gamma^2 - \frac{m^2}{r^2} \right) A
\]

and rewrote the right-hand side in a form that will turn out to be convenient later. We expand

\[
A(X, r; \epsilon) = A_0(X, r) + \epsilon A_1(X, r) + O(\epsilon^2),
\]
15.4. Singular Perturbation Problems

substitute in (†), and collect like powers of $\varepsilon$ to obtain

$$O(1): \mathcal{L}[A_0] = 0, \quad \frac{\partial A_0}{\partial r} = 0 \quad \text{at } r = R(X),$$

$$O(\varepsilon): \mathcal{L}[A_1] = \frac{i}{A_0} \frac{\partial}{\partial X} \gamma A_0^2, \quad \frac{\partial A_1}{\partial r} = -i \gamma A_0 \quad \text{at } r = R(X).$$

(2)

Since the variable $X$ plays no other role in (‡) than that of a parameter, we have for $A_0$ the “almost mode”

$$A_0(X, r) = P_0(X) J_m(\alpha(X)r), \quad \alpha(X) = j_m j' m \mu / R(X), \quad \gamma^2(X) = \kappa^2 - \alpha^2(X), \quad \text{Re}(\gamma) \geq 0, \quad \text{Im}(\gamma) \leq 0.$$

The amplitude $P_0$ is still undetermined and follows from $A_1$. We do not, however, have to solve $A_1$ if we can find a sufficient solvability condition for $A_1$. For this we multiply the left- and right-hands side of (♯) by $r A_0$ and integrate with respect to $r$ from 0 to $R(X)$. For the left-hand side we utilize the self-adjointness of $\mathcal{L}$:

$$\int_0^R r A_0 \mathcal{L}[A_1] dr = \int_0^R r A_0 \mathcal{L}[A_1] - r A_1 \mathcal{L}[A_0] dr$$

$$= \left[ r A_0 \frac{\partial A_1}{\partial r} - r A_1 \frac{\partial A_0}{\partial r} \right]_0^R = -i \gamma R R' A_0^2.$$

For the right-hand side we apply Leibniz's rule, i.e.,

$$\int_0^R \frac{\partial}{\partial X} (i \gamma A_0^2) dr = \frac{d}{dX} \int_0^R i \gamma A_0^2 dr - i \gamma R R' A_0^2.$$

Hence

$$\frac{d}{dX} \int_0^R r \gamma A_0^2 dr = 0,$$

and so, using properties of the Bessel function [1], we have

$$\int_0^R r \gamma A_0^2 dr = \left[ \frac{1}{2} \gamma P_0^2 \left( r^2 - \frac{m^2}{\alpha^2} \right) J_m(\alpha r) \right]_0^R = \frac{1}{2} \gamma P_0^2 R^2 \left( 1 - \frac{m^2}{j_m^2} \right) J_m(j_m^2) = C$$

(where $C$ denotes any constant) or

$$P_0(X) = \frac{C}{R(X) \sqrt{\gamma(X)}}.$$

The present problem may be generalised for annular ducts with acoustically lined walls and mean flow or modeling sound propagation in the inlet or exhaust duct of a turbofan aircraft engine [127, 128, 132]. See Chapter 16, Section 16.16.

In more dimensions the assumed form of (15.62), where an integral occurs in the argument of the exponential, is not practical. In this case it is more convenient to write

$$\psi(x, t; \varepsilon) := A(X, T; \varepsilon) e^{i \varepsilon^{-1} \Omega(X,T)},$$

(15.66)
while for clarity of notation we leave $\Omega$ independent of $\epsilon$ and introduce the slowly varying frequency and wave vector

$$
\omega := \frac{\partial \Omega}{\partial T}, \quad \kappa := -\nabla \Omega,
$$

where $\nabla := \frac{\partial}{\partial T} e_x + \frac{\partial}{\partial Y} e_y + \frac{\partial}{\partial Z} e_z$. This relation implies the continuity equation of waves

$$
\frac{\partial \kappa}{\partial T} + \nabla \omega = 0.
$$

(15.67)

Consider the following example of a one-dimensional wave equation with slowly varying coefficients:

$$
\frac{\partial}{\partial t} \left( m(X, T) \frac{\partial}{\partial t} \phi \right) = \frac{\partial}{\partial x} \left( C(X, T) \frac{\partial}{\partial x} \phi \right) + B(X, T) \phi,
$$

(15.68)

where $X = \epsilon x$ and $T = \epsilon t$ are slow variables. We assume that the solution $\phi$ takes the form given by (15.66). This yields the equation

$$
-\omega^2 mA + i\epsilon \frac{\partial}{\partial T} \left( \omega m A^2 \right) = -\kappa^2 CA - i\epsilon \frac{\partial}{\partial X} \left( \kappa CA^2 \right) + BA + O(\epsilon^2).
$$

(15.69)

As before, we expand

$$
A = A_0 + \epsilon A_1 + O(\epsilon^2).
$$

After substitution and collecting equal powers of $\epsilon$, we get from leading order the slowly varying dispersion relation for $\omega$ and $\kappa$, or the eikonal-type equation for $\Omega$:

$$
\omega^2 m = \kappa^2 C - B.
$$

(15.70)

Equation (15.67) turns into

$$
\frac{\partial \kappa}{\partial T} + V(\kappa) \frac{\partial \kappa}{\partial X} = 0,
$$

showing that both $\kappa$ and $\omega$ propagate with the group velocity (see (3.38))

$$
V = \frac{d\omega}{dx} = \frac{\kappa C}{\omega m}.
$$

(15.71)

The next order yields a conservation-type equation for $A_0$:

$$
\frac{\partial}{\partial T} \left( \omega m A_0^2 \right) + \frac{\partial}{\partial X} \left( \kappa CA_0^2 \right) = 0.
$$

(15.72)

(It should be noted that this result reflects the underlying physics and therefore depends on the original equation. In general the resulting equation is not of conserved type.) The pair $\omega m A_0^2$ and $\kappa CA_0^2$ are called adiabatic invariants because they correspond to the density and flux of a quantity that is conserved on the slow time and length scales. This is seen
as follows. When we integrate (15.72) between the moving boundaries \( X = X_1(T) \) and \( X = X_2(T) \), we obtain

\[
\int_{X_1}^{X_2} \frac{\partial}{\partial T} (\omega m A_0^2) + \frac{\partial}{\partial X} (\kappa C A_0^2) \, dX = \frac{d}{dT} \int_{X_1}^{X_2} \omega m A_0^2 \, dX \\
- V_2 [\omega m A_0^2]_{X_2} + V_1 [\omega m A_0^2]_{X_1} + [\kappa C A_0^2]_{X_2} - [\kappa C A_0^2]_{X_1} = 0,
\]

where \( V_1 = \frac{d}{dT} X_1 \) and \( V_2 = \frac{d}{dT} X_2 \). This reduces to

\[
\frac{d}{dT} \int_{X_1}^{X_2} \omega m A_0^2 \, dX = 0
\]

if the velocity of either endpoint is equal to the group velocity (15.71). In other words, the integral of \( \omega m A_0^2 \) is conserved between two points moving with the local group velocity.

Suppose, for definiteness, that \( \phi \) denotes position and \( m \) denotes mass. Then if \( \omega \) and \( \kappa \) are real, we can derive from (15.72) with (15.70) the conservation laws for wave action

\[
\frac{\partial}{\partial T} \left( \frac{1}{2} \omega m |A_0|^2 \right) + \frac{\partial}{\partial X} \left( \frac{1}{2} \kappa C |A_0|^2 \right) = 0 \tag{15.73a}
\]

(\( \frac{1}{2} \omega m |A_0|^2 \) is the wave action density [8]) and energy

\[
\frac{\partial}{\partial T} \left( \frac{1}{2} \omega^2 m |A_0|^2 \right) + \frac{\partial}{\partial X} \left( \frac{1}{2} \omega C |A_0|^2 \right) = 0 \tag{15.73b}
\]

(\( \frac{1}{2} \omega^2 m |A_0|^2 \) is the wave energy density) by substituting \( A_0 = |A_0| \exp(i \arg A_0) \), dividing out the complex exponent, and taking the real part of what remains.

**Example 15.37** (ray acoustics in a temperature gradient) A very important application of waves in a slowly varying medium is the theory of rays. When a sound wave propagates in free space through a medium that varies on a much larger scale than the typical wavelength (typically temperature gradients or wind with shear), the same ideas of multiple scales may be applied. In contrast to the duct, where the wave is confined by the duct walls, the waves may now freely refract and follow curved paths. These paths are called rays.

Consider an infinite three-dimensional medium with varying temperature (typical length scale \( L \)) but otherwise with a constant mean pressure, so that we have for the acoustic pressure perturbations the equation (7.7), i.e.,

\[
\frac{\partial^2}{\partial t^2} p = \nabla \cdot (\epsilon(X) \nabla p), \tag{15.74}
\]

where \( X = \epsilon x \) for small \( \epsilon \) and the sound speed \( c \) is slowly varying when compared to the acoustic length scale. This means that \( \omega \), the frequency considered, is high enough for the corresponding wavelength \( \lambda \sim 2\pi c/\omega \) to be small compared to \( L \), so \( \epsilon \sim \lambda/L \). Assuming the field to be locally plane, we try an approximate solution (a ray approximation)

\[
p(x, t) = A \, e^{i\omega t - i\theta(x/t)} \tag{15.75}
\]

with the form of a plane wave but with slowly varying amplitude \( A = A(X; \epsilon) \) and phase \( \epsilon^{-1}(\omega t - \theta(X)) \), where \( T = \epsilon t \). The surfaces \( \theta - \omega T = \text{constant} \) describe the propagating
wave front, with normal vectors $\nabla \theta$. We define the operator $\nabla = \left( \frac{\partial}{\partial X}, \frac{\partial}{\partial Y}, \frac{\partial}{\partial Z} \right)$ so that $\nabla = \varepsilon \nabla$.

Then we define the local wave vector $\nabla \theta = \kappa$. Since

$\nabla p = \left(-i\kappa A + \varepsilon \nabla A\right)e^{i\omega t - \theta/\varepsilon}$,

$\nabla^2 p = \left(-|\kappa|^2 A - 2i\varepsilon \kappa \cdot \nabla A - i\varepsilon (\nabla \cdot \kappa) A + \varepsilon^2 \nabla^2 A\right)e^{i\omega t - \theta/\varepsilon}$,

we obtain by substituting (†) in (∗) the equation

$$(\omega^2 - c^2 |\kappa|^2)A = \frac{i\varepsilon}{A} \nabla \cdot (c^2 \kappa A^2) + O(\varepsilon^2).$$

We expand

$$A = A_0 + \varepsilon A_1 + O(\varepsilon^2)$$

and collect like powers in (‡). We find to leading orders for $\kappa$ and $A_0$ that

$$c^2 |\kappa|^2 = \omega^2,$$

$$\nabla \cdot (c^2 \kappa A_0^2) = 0.$$

Written in terms of $\theta$, (‡) is the eikonal equation, which determines the wave fronts and the ray paths. Equation (‡‡) is called the transport equation. It implies, like in (15.73), the equations for conservation of wave action and conservation of energy, since $\kappa$ is real. It relates the amplitude variation to diverging or converging rays (see the problem considered in Section 16.9 of Chapter 16). The eikonal equation is a nonlinear first order PDE of hyperbolic type, which can always be reduced to a system of ODEs along characteristics (see Chapter 2).

15.5 Discussion

• Inherent in any modeling is the hierarchy in importance of the various effects that constitute the model. Therefore certain effects in any modeling will be small. Sometimes they are small but not small enough to be ignored, and sometimes they are small but in a nonuniform way such that they are important locally.

• For an efficient solution, and to obtain qualitative insight, it makes sense to utilize this “smallness.” Perturbation methods systematically exploit such intrinsic smallness.

• The basis of a systematic approach is formed by the concept of asymptotic approximations and expansions of Poincaré type. If the approximation is uniform on the domain of interest, it is called regular; otherwise it is called singular. The role of the chosen independent variable in this respect cannot be overemphasized.

• Perturbation methods have a long history. Before the time of numerical methods and the computer, perturbation methods were the only way to increase the applicability of available exact solutions to difficult, otherwise intractable problems. Nowadays, perturbation methods have their use as a natural step in the process of systematic modeling, for the insight they provide in the nature of singularities occurring in the problem and typical parameter dependencies, and sometimes in the speed of practical calculations.
Exercises

15.1. Determine asymptotic expansions for $\varepsilon \to 0$ of
   (a) $\varepsilon/\tan \varepsilon$,
   (b) $\varepsilon/(1 - \varepsilon^x)$,
   (c) $\log(\sin \varepsilon)$,
   (d) $(1 - \varepsilon^2 \ln \varepsilon)/(1 - \varepsilon \ln \varepsilon - \varepsilon^2 \ln \varepsilon)$.

15.2. Determine, if possible, uniform asymptotic expansions for $\varepsilon \to 0$ and $x \in [0, 1]$ of
   (a) $\sin(\varepsilon x)$,
   (b) $1/(\varepsilon + x)$,
   (c) $x \log(\varepsilon x)$,
   (d) $e^{-\sin(\varepsilon x)}$,
   (e) $e^{-\sin(\varepsilon x)/\varepsilon}$.

15.3. Derive asymptotic solutions (for $\varepsilon \to 0$) of the equation
   $$\varepsilon x^3 - x + 2 = 0.$$

15.4. Derive, step by step, by iteratively scaling $x(\varepsilon) = \mu_0(\varepsilon)x_0 + \mu_1(\varepsilon)x_1 + \cdots$ and balancing, that a third order asymptotic solution (for $\varepsilon \to 0$) of the equation
   $$\ln(\varepsilon x) + x = a$$
   is given by
   $$x(\varepsilon) = \ln \varepsilon^{-1} - \ln(\ln \varepsilon^{-1}) + a + o(1).$$
   Find a more efficient expansion based on an alternative asymptotic sequence of gauge functions by combining $e^{-\alpha \varepsilon}$.

15.5. Analyse asymptotically for $\varepsilon \to 0$ the zeros of
   $$e^{-x^2/\varepsilon^2} + x - \varepsilon.$$
15.6. Derive the *Webster's horn equation* for sound of long wavelength propagating in slowly varying horns, by the method of slender approximation [86]. The reduced wave equation for pressure perturbations $p$ and wave number $k$ is given by

$$\nabla^2 p + k^2 p = 0$$

within a duct given in cylindrical polar coordinates by

$$S = r - R(X, \theta) = 0, \quad X = \varepsilon x, \quad \varepsilon \text{ is small.}$$

The wave number is $O(\varepsilon)$, so we scale $k = \varepsilon k$. The duct wall is hard, so we have the boundary condition

$$\nabla p \cdot \nabla S = 0 \quad \text{at} \quad S = 0.$$

15.7. Consider the incompressible Navier–Stokes equations to describe *lubrication flow* in a two-dimensional narrow and slowly varying channel with prescribed volume flux. (In actual practice this flux is created by a pressure difference.)

(a) Make the channel height and volume flux dimensionless and scale the pressure gradient such that viscous forces are balanced by the pressure gradient, so the Reynolds number $Re \leq O(1)$. Verify that we obtain in dimensionless form that

$$Re(u \cdot \nabla u) + \nabla p = \nabla^2 u, \quad \nabla \cdot u = 0,$$

for the velocity $u = (u, v)^T$ and pressure $p$ in the channel given by $-\infty < x < \infty$ and $g(\varepsilon x) \leq y \leq h(\varepsilon x)$, where $\varepsilon$ is a small parameter. (End conditions in $x$ are not important.) Boundary conditions are as follows: no slip at the walls, i.e., $u = v = 0$ at $y = g(\varepsilon x)$ and $y = h(\varepsilon x)$, and a flux

$$\int_{g(\varepsilon x)}^{h(\varepsilon x)} u(x, y) \, dy = 1.$$

(b) We rewrite $X = \varepsilon x$ and assume that the field varies slowly in $X$ (any end effects are local and irrelevant for the $x$'s considered). Rescale $u$, $v$, and $p$. The order of magnitude of the pressure can be found from the observation that a pressure gradient is necessary to have a flow. The crosswise velocity $v$ will be much smaller than the axial velocity $u$.

(c) Assume for rescaled $u$, $v$, and $p$ an obvious asymptotic expansion in $\varepsilon$ and solve up to leading order.

15.8. Consider the function

$$f(x; \varepsilon) = e^{-x/\varepsilon}(1 + x) + \pi \cos(\pi x + \varepsilon) \quad \text{for} \quad 0 \leq x \leq 1.$$

(a) Construct an outer and an inner expansion of $f$ with error $O(\varepsilon^3)$.

(b) Integrate $f$ from $x = 0$ to $x = 1$ exactly and expand the result up to $O(\varepsilon^3)$.

(c) Compare this with the integral that is obtained by integrating the inner and outer expansions following the method described in Example 15.30.
15.9. Construct an asymptotic approximation for small $\varepsilon$ of
\[ \int_0^\infty \frac{1}{x^2 + x^6 + \varepsilon^2} \, dx. \]

15.10. Reconsider (†) of Example 7.4 to describe a stationary suspended flexible bar of length $L$.
(a) First we consider a cable with clamped ends at equal height. This is described by adding boundary conditions $x = y = \phi = 0$ at $s = 0$ and $x = D$ and $y = \phi = 0$ at $s = L$, where $1 - D/L$ is positive and not small. Note that for given $D$ the necessary horizontal force $H$ is unknown and to be determined. Make the problem dimensionless by scaling lengths on $L$ and forces on $QL$. Introduce the parameters $\varepsilon = \sqrt{(EI/QL^3)}$, $h = h(\varepsilon) = H_0/QL$, and $v = v(\varepsilon) = V_0/QL$. If $\varepsilon$ is small, the equation describes a suspended cable. Find the solution asymptotically to leading order for small $\varepsilon$.
(b) Do the same for a cable with hinged ends, i.e., with $\phi'(0) = \phi(L)' = 0$.
(c) The same differential equation represents a model for laying submarine gas and oil pipelines from a laybarge. The pipe is freely suspended over an unknown length $L$, with prescribed curvature $R$ at the liftoff point at height $y = W$ and a prescribed horizontal tension $H$ in order to avoid buckling of the pipe. Both the angle $\phi$ and the curvature $\phi'$ vanish at the touchdown point $s = 0$. We have thus $\phi(0) = \phi'(0) = 0$, $\phi'(L) = -R^{-1}$, $y(0) = 0$, and $y(L) = W$, while $L$ is unknown. Make dimensionless and solve the resulting problem asymptotically for small $\varepsilon$.

15.11. Determine the asymptotic approximation of the solution $y(x; \varepsilon)$ (first or first and second order terms for positive small parameter $\varepsilon \to 0$) of the following singularly perturbed problems. Let $\alpha$ and $\beta$ be nonzero constants that are independent of $\varepsilon$. Provide arguments for the determined boundary layer thickness and location, and show how free constants are determined by the matching procedure.
(a) $\varepsilon y'' - y' = 2x$, \hspace{1cm} $y(0; \varepsilon) = \alpha, \quad y(1; \varepsilon) = \beta$. 
(b) $\varepsilon y' + y^2 = \cos x$, \hspace{1cm} $y(0; \varepsilon) = 0, \quad 0 \leq x \leq 1$. 
(c) $\varepsilon y'' + (2x + 1)y' + y^2 = 0$, \hspace{1cm} $y(0; \varepsilon) = \alpha, \quad y(1; \varepsilon) = \beta$.

15.12. Determine at least two terms of an asymptotic approximation for small $\varepsilon > 0$ of the solution $\psi(x)$ of the equation
\[ \varepsilon(\psi'' + \psi \psi') = -2 \quad \text{along} \quad 0 \leq x \leq 1, \]
Chapter 15. Perturbation Methods

with boundary conditions

\[ \psi(0) = 1, \quad \psi(1) = 1. \]

Note that there are boundary layers at both ends. Hint: Airy functions play a role, i.e., solutions of \( y'' - xy = 0 \). Consult [1] for their asymptotic behaviour.

15.13. When we stir a cup of tea, the surface of the fluid deforms until equilibrium is attained between gravity, centrifugal force, and surface tension. This last force is only important near the wall. (Note: Surface tension is \( \sigma = 75 \) mN/m for water.) Consider for this problem the following model.

A cylinder (radius \( a \), axis vertical) with fluid (density \( \rho \), surface tension \( \sigma \)) rotates around its axis \( e_z \) (angular velocity \( \Omega \)) in a gravity field \(-g e_z\). By the gravity and the centrifugal force the surface deforms to something that looks like a paraboloid. Within a small neighbourhood of the cylinder wall the contact angle \( \theta = \frac{\pi}{2} - \alpha \) is felt by means of the surface tension (see (6.50)). Because of symmetry we can describe the surface by a radial tangent angle \( \psi \) with the horizon, parametrized by arc length \( s \), such that \( s = 0 \) corresponds to the axis and \( s = L \) corresponds to the wall of the cylinder. \( L \) is unknown. Select the origin on the axis at the surface such that the vertical and radial coordinates are given by \( Z(s) = \int_0^s \sin \psi(s') \, ds' \) and \( R(s) = \int_0^s \cos \psi(s') \, ds' \), respectively.

The balance between hydrostatic pressure and surface tension yields the equation

\[ p_0 - \rho g Z + \frac{1}{2} \rho \Omega^2 R^2 = -\sigma \left( \frac{d\psi}{ds} + \frac{\sin \psi}{R} \right) \]

with unknown \( p_0 \). Boundary conditions are \( \psi(0) = 0, \psi(L) = \alpha \), and \( R(L) = a \).

(a) Scale lengths on \( a \), i.e., \( s = at \), \( R = ar \), \( Z = az \), and \( L = a\lambda \), and introduce \( \beta = p_0/pga \) and the dimensionless parameters \( \varepsilon^2 = \sigma/\rho ga^2 \) and \( \mu = \Omega^2 a/g \).

Can you identify the common names of these dimensionless numbers?

(b) Solve the resulting problem asymptotically for \( \varepsilon \to 0 \), while \( \mu = O(1) \).

15.14. Consider the van der Pol equation for variable \( y = y(t; \varepsilon) \) in \( t \) and small parameter \( \varepsilon \):

\[ y'' + y - \varepsilon(1 - y^2)y' = 0. \]

Construct by using the Lindstedt–Poincaré method an \( O(\varepsilon^2) \) approximation of a periodic solution.

15.15. Consider the system governed by the equation of motion

\[ y'' + y + ay^3 = \varepsilon^2 \cos \omega t. \]
Exercises

499

Find, asymptotically for $\varepsilon \to 0$ and $\alpha = O(1)$, to leading order the steady solution
(a) for $\omega^2 = 1 + O(1)$, i.e., away from resonance;
(b) for $\omega^2 = 1 + \varepsilon \mu$, i.e., near resonance.

Hint: For part (b), introduce the variable $\tau = \omega t$.

15.16. Following Example 15.36, derive a multiple scales solution of sound waves in a slowly varying duct while the sound speed is also a slowly varying function of $x$.

The pertinent equation is therefore (15) of Example 15.37.

15.17. (a) Rewrite the eikonal equation (15) of Example 15.37 in characteristic form by using Theorem 12.6.

(b) Prove that in a medium with a linearly varying sound speed the paths of rays are circles.

15.18. Analyse the error $u = y - \tilde{y}$ of (3.43), described by (3.44), but now including the effects of the slowly varying coefficients $a$, $b$, and $c$. Formulate your result in the form of adiabatic invariants.

15.19. Consider the Fisher travelling-wave problem (10.70) with large $c$

$$U'' + c^2 U' + c^2 U(1 - U) = 0.$$    

(a) Derive an approximate solution on $(-\infty, \infty)$ with $U(-\infty) = 1, U(\infty) = 0$. It is no restriction to assume that $U(0) = \frac{1}{2}$.

(b) Repeat part (a) on $[0, \infty)$ with $U(0) = 0$, while the previous solution is the outer solution.

15.20. A field $\phi$ (satisfying the Helmholtz or reduced wave equation; see Example 1.8iii), radiated by a source $q(x)$ that is distributed within a finite region $\Omega$ of typical diameter $L$, is given by

$$\phi(x; \kappa) = -\frac{1}{4\pi} \int_{\Omega} q(y) \frac{e^{-ixr}}{r} dV, \quad \text{where} \quad r = \|x - y\|.$$    

With the origin inside $\Omega$, evaluate $\phi$ asymptotically for small $\kappa L$ for different regions in $x$. Distinguish in particular the near or static zone, $\kappa L \ll \|x\| \ll 1$; the intermediate or induction zone, $\kappa L \ll \kappa \|x\| = O(1)$; and the far or radiation zone, $\kappa L \ll 1 \ll \kappa \|x\|$.

15.21. The equation

$$y'' + \frac{1}{r} y' + \left(\alpha^2 - \frac{m^2}{r^2}\right) y = 0$$

has solutions in the form of Bessel functions of order $m$ and argument $\alpha r$. Find asymptotic solutions of WKB type for $\alpha \to \infty$ and $r = O(1)$ with $r > m/\alpha$. Consider (i) $m^2 = O(1)$, (ii) $m^2/\alpha = O(1)$, and (iii) $m^2/\alpha^2 = O(1)$.
Chapter 16

Modeling, Analysing, and Simulating Problems from Practice

16.1 Production of Resin-Containing Panels

There exists a variety of synthetic panels for both interior and exterior use. Typical applications are office desktops and facade cladding. One type of panel is made of polymerised resin reinforced with wood fibres or sulphate paper.

In manufacturing these panels we can distinguish two stages. In the first stage sheets of resin-impregnated paper enclosed by two layers of padded paper are pressed together. During this pressing stage the components are compressed to within 10 percent of their final thickness. A schematic three-layer model of a panel is given in Figure 16.1. A panel thus has two bolsters of padded paper and a core consisting of impregnated paper. In the second stage high temperatures are applied at the boundaries of the panel, which causes heating of the panel. This induces a polymerisation reaction in the core. Heat is released during the polymerisation, which again leads to an increase of temperature in the panel. This process is called resin curing and continues until the chemicals in the core are depleted. When the polymerisation reaction is finished, the temperature finally tends to a steady state. In this section we introduce a model for the resin-curing process.

![Figure 16.1. Schematic representation of a three-layer panel.](image)
16.1.1 Modeling the Curing Process

The curing process can be described by the temperature $T$ in the panel and the concentration $C$ of chemicals involved in the polymerisation reaction in the core. To keep the model feasible we make the following assumptions:

1. The materials are incompressible.
2. Resin curing is a first order reaction.
3. Diffusion of resin is negligible.
4. All variables depend only on the transverse space coordinate $x$.

The governing equations are the following. In both bolsters the heat equation holds, which in our particular case is given by [104]

$$ \rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( \lambda \frac{\partial T}{\partial x} \right), \quad (16.1) $$

with $\rho$, $c_p$, and $\lambda$ the density, specific heat, and thermal conductivity, respectively, of the bolster material. These parameters are assumed to be constant. In the core, heat conduction is coupled with the polymerisation reaction, and the governing equations read

$$ \rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( \lambda \frac{\partial T}{\partial x} \right) + \Delta H k C, \quad (16.2a) $$

$$ \frac{\partial C}{\partial t} = -k C, \quad (16.2b) $$

with $\Delta H$ and $k$ the enthalpy of polymerisation and the reaction rate, respectively. We assume also that the parameters $\rho$, $c_p$, $\lambda$, and $\Delta H$ in (16.2a) and (16.2b) are constant. The reaction rate is given by the Arrhenius expression

$$ k = Ae^{-E_a/RT}, \quad (16.3) $$

with $A$, $E_a$, and $R$ the preexponential factor, the activation energy, and the universal gas constant, respectively. The reaction rate $k$ as a function of the temperature $T$ is shown in Figure 16.2. The values of all physical parameters/constants are given in Table 16.1.

Since the parameters $\rho$, $c_p$, $\lambda$, and $\Delta H$ are constant, the heat equations (16.1) and (16.2a) can be simplified. Introducing the thermal conductivity $a$ and the coefficient $b$ of temperature increase per unit concentration decrease, in short the temperature coefficient, by

$$ a := \frac{\lambda}{\rho c_p}, \quad b := \frac{\Delta H}{\rho c_p}, \quad (16.4) $$

we find

$$ \frac{\partial T}{\partial t} = a \frac{\partial^2 T}{\partial x^2} \quad (16.5) $$

for the heat equation in both bolsters and

$$ \frac{\partial T}{\partial t} = a \frac{\partial^2 T}{\partial x^2} + bk C \quad (16.6) $$
16.1. Production of Resin-Containing Panels

![Figure 16.2. The reaction rate \( k \) as a function of the temperature \( T \).](image)

Table 16.1. Parameters involved in the curing process.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>((ρc_p)_{1,3})</td>
<td>1.830 \times 10^6</td>
<td>J/(m^3 \cdot K)</td>
<td>(E_a)</td>
<td>9.4 \times 10^4</td>
<td>J/mol</td>
</tr>
<tr>
<td>((ρc_p)_{2})</td>
<td>2.123 \times 10^6</td>
<td>J/(m^3 \cdot K)</td>
<td>(R)</td>
<td>8.314</td>
<td>J/(mol \cdot K)</td>
</tr>
<tr>
<td>(λ_{1,3})</td>
<td>1.9032 \times 10^{-1}</td>
<td>J/(m \cdot s \cdot K)</td>
<td>(x_{\text{int},1})</td>
<td>2.0 \times 10^{-3}</td>
<td>m</td>
</tr>
<tr>
<td>(λ_{2})</td>
<td>3.16327 \times 10^{-1}</td>
<td>J/(m \cdot s \cdot K)</td>
<td>(x_{\text{int},2})</td>
<td>1.6 \times 10^{-2}</td>
<td>m</td>
</tr>
<tr>
<td>(ΔH)</td>
<td>8.0 \times 10^4</td>
<td>J/(mol \cdot K)</td>
<td>(L)</td>
<td>1.8 \times 10^{-2}</td>
<td>m</td>
</tr>
<tr>
<td>(A)</td>
<td>3 \times 10^9</td>
<td>s^{-1}</td>
<td>(x_{\text{int},1})</td>
<td>2 \times 10^{-3}</td>
<td>m</td>
</tr>
</tbody>
</table>

in the core. Equation (16.6) is coupled with (16.2b) for the concentration.

The model (16.5), (16.6), and (16.2b) has to be completed with initial and boundary conditions and conditions at the interfaces between core and bolsters. As initial conditions, we choose a constant temperature and concentration:

\[
T(x, 0) = T_0, \quad 0 < x < L, \quad (16.7)
\]

\[
C(x, 0) = C_0, \quad x_{\text{int},1} < x < x_{\text{int},2}. \quad (16.8)
\]

At the boundaries we impose a constant temperature:

\[
T(0, t) = T_{\text{top}}, \quad T(L, t) = T_{\text{bot}}, \quad t > 0. \quad (16.9)
\]

Finally, at the interfaces we require that both the temperature \( T \) and the heat flux \( \lambda \frac{∂T}{∂x} \) be continuous. This results in the conditions

\[
T(x -, t) = T(x +, t), \quad \left( \lambda \frac{∂T}{∂x} \right)(x -, t) = \left( \lambda \frac{∂T}{∂x} \right)(x +, t), \quad x = x_{\text{int},1}, x_{\text{int},2}. \quad (16.10)
\]
These conditions mean that there is no accumulation of heat at the interfaces. Note that we do not have interface conditions for the concentration \( C \) and that we do not need them either.

Before we compute a numerical solution, we estimate the typical time scales involved. The time scale \( \tau_C \) due to heat conduction from the walls follows from the balance between the terms \( \frac{\partial}{\partial t} T \) and \( \frac{\partial^2}{\partial x^2} T \) and is given by

\[
\tau_C = L^2/a.
\]

The time scale \( \tau_P \) of the polymerisation reaction is

\[
\tau_P(T) = k(T)^{-1}
\]

and depends strongly on the temperature. Heat conduction and the polymerisation reaction are essentially coupled when both time scales are of the same order of magnitude. This is the case when the order of magnitude of the temperature is as follows:

\[
T \sim T^* := \frac{T_a}{\ln(A L^2/a)}.
\]

where the temperature \( T^* \) follows from \( \tau_P(T^*) = \tau_C \) and where \( T_a := E_a/R \) is the activation temperature. In this problem we typically have \( \tau_C = 2000 \text{ s} \) and \( T^* = 383 \text{ K} \).

### 16.1.2 Numerical Solution Method

In order to compute a numerical solution of the initial boundary value problem (16.5), (16.6), (16.2b), and (16.7) to (16.10), we introduce the spatial grid

- bolster (1): \( x_j = (j - 1) \Delta x_1 \), \( j = 1, 2, \ldots, N_1 \),
- core (2): \( x_{j+N_1} = x_{\text{int},1} + (j - 1) \Delta x_2 \), \( j = 1, 2, \ldots, N_2 \),
- bolster (3): \( x_{j+N_1+N_2} = x_{\text{int},2} + (j - 1) \Delta x_3 \), \( j = 1, 2, \ldots, N_3 \),

with \( \Delta x_r (r = 1, 2, 3) \) the grid size in region \( r \); see Figure 16.1. We have chosen two grid points at each interface, i.e., \( x_{N_1} = x_{N_1+1} = x_{\text{int},1} \) and \( x_{N_1+N_2} = x_{N_1+N_2+1} = x_{\text{int},2} \). Moreover, we choose the discrete time levels

\[
t^n = n \Delta t \quad (n = 0, 1, 2, \ldots, M), \quad \Delta t = \frac{t_{\text{fin}}}{M},
\]

with \( t_{\text{fin}} \) the final simulation time. To discretise the space derivatives in the governing equations we employ central differences and for time integration we use the \( \vartheta \) method. On the other hand, to discretise the interface conditions we use one-sided differences. In the following \( T^n_j \) and \( C^n_j \) denote the numerical approximations of \( T(x_j, t^n) \) and \( C(x_j, t^n) \), respectively. A subscript \( r \) \( (r = 1, 2, 3) \) on a parameter refers to the region it relates to.

For the heat equation (16.5) in both bolsters we find the difference equations

\[
T_{j+1}^{n+1} - \vartheta d_t (T_{j+1}^{n+1} - 2T_j^{n+1} + T_{j-1}^{n+1}) = T_j^n + (1-\vartheta)d_t (T_{j+1}^{n} - 2T_j^{n} + T_{j-1}^{n}) \quad (r = 1, 3).
\]
16.1. Production of Resin-Containing Panels

The coefficient $d_r$ in (16.16) is defined by

$$d_r := \frac{d_r \Delta t}{\Delta x_r^2}$$

(16.17)

and might be called the conduction number. Likewise, for the equations (16.6) and (16.2b) in the core we have

$$T_{n+1}^j - \theta \left( d_2 (T_{n+1}^{j+1} - 2T_n^{j+1} + T_{n-1}^{j+1}) + \Delta t b_2 k(T_n^{j+1})C_n^{j+1} \right)$$

$$= T_n^j + (1 - \theta) \left( d_2 (T_{n+1}^j - 2T_n^j + T_{n-1}^j) + \Delta t b_2 k(T_n^j)C_n^j \right),$$

(16.18a)

$$C_{n+1}^j + \theta \Delta t k(T_{n+1}^j)C_n^{j+1} = C_n^j - (1 - \theta) \Delta t k(T_n^j)C_n^j.$$  

(16.18b)

The coefficient $d_2$ in (16.18a) is defined in (16.17). Finally, at the interfaces the following discrete equations hold:

$$T_{N_{1}+1}^m = T_{N_{1}+1}^m, \quad (m = n, n + 1),$$

(16.19a)

$$\frac{\lambda_1}{\Delta x_1} (T_{N_{1}}^m - T_{N_{1}-1}^m) = \frac{\lambda_2}{\Delta x_2} (T_{N_{1}+2}^m - T_{N_{1}+1}^m),$$

(16.19b)

$$T_{N_{1}+N_{2}}^m = T_{N_{1}+N_{2}+1}^m,$$

(16.19c)

$$\frac{\lambda_2}{\Delta x_2} (T_{N_{1}+N_{2}}^m - T_{N_{1}+N_{2}-1}^m) = \frac{\lambda_3}{\Delta x_3} (T_{N_{1}+N_{2}+2}^m - T_{N_{1}+N_{2}+1}^m).$$

(16.19d)

The first two equations apply to the interface $x = x_{int,1}$ and the second two apply to the interface $x = x_{int,2}$. The complete set of equations is (16.16), (16.18), and (16.19) for the unknowns $T_{n+1}^m$ and $C_{n+1}^j$. In fact, this is a nonlinear system due to the $k(T_{n+1}^j)$ terms. In order to avoid solving a nonlinear system, we replace the nonlinear term $k(T_{n+1}^j)$ in (16.18a) and (16.18b) by $k(T_n^j)$. This corresponds to one Picard iteration step.

As an example, we have computed a numerical solution for a symmetrically heated panel. The initial temperature and concentration are $T_0 = 293 \text{ K}$ and $C_0 = 1250 \text{ mol/m}^3$, respectively. Then, at $t = 0$, the temperatures at top and bottom of the panel are set to $T_{top} = T_{bot} = 433 \text{ K}$. The numerical parameters in our numerical simulation are $\theta = 1$, $N_1 = N_3 = 11$, $N_2 = 41$, $t_{fin} = 1500$, and $M = 25$. The results are presented in Figures 16.3 and 16.4. In Figure 16.3 we show the temperature and concentration distributions at the discrete time levels $t^n$ ($n = 0, 1, 2, \ldots, M$). The time histories of the temperature and concentration at the centre of the panel are shown in Figure 16.4. From these two figures we can see that the temperature in the core initially increases through conduction. The higher temperature initiates a chemical reaction, resulting in heat production and consumption of chemicals. This process continues until the chemicals in the core are depleted. Then the temperature tends to the constant temperature imposed at the boundaries.
16.1.3 Discussion and Related Problems

- The proposed model is rather elementary. In the real resin-curing process more complex chemistry, involving more chemical species and reactions, should be taken into account. Also the diffusion of all species should be modeled. Finally, the one-dimensional approximation is most likely not accurate near the ends of the panel. A three-dimensional model should be used instead.

- A more detailed description of resin-containing panels, including deformation through moisture transport, is given in [54, 23].

- The resin-curing process is an example of a reaction-diffusion problem. These problems occur in many fields of science and engineering, e.g., in models for flame propagation, biochemical reactions, and population dynamics. Examples in this text are the thermal explosion in a vessel (Section 16.3) and the catalyst pellet problem (Section 16.8). Moreover, we discussed Fisher's equation, which is a model problem from population dynamics, in Section 10.6.2.
16.1.4 Exercises

16.1. Derive expression (16.13) for $T^*$.

16.2. To make the equations (16.2) dimensionless we can employ the scaling

\[ \tilde{x} := x/L, \quad \tilde{t} := t/\tau, \quad \tilde{u} := (T^* - T_0)/\Delta T, \quad \tilde{C} := C/C_0, \]

with $\Delta T := T^* - T_0$. Show that the resulting dimensionless equations (omitting the tilde \( ^\sim \)) read

\[
\frac{\partial \tilde{u}}{\partial \tilde{t}} = \frac{\partial^2 \tilde{u}}{\partial \tilde{x}^2} + \gamma \exp\left(\frac{\beta \tilde{u}}{\tilde{u} + \alpha}\right) \tilde{C},
\]

\[
\frac{\partial \tilde{C}}{\partial \tilde{t}} = -\exp\left(\frac{\beta \tilde{u}}{\tilde{u} + \alpha}\right) \tilde{C},
\]

where the coefficients $\alpha$, $\beta$, and $\gamma$ are defined by

\[ \alpha := T^*/\Delta T, \quad \beta = T_a/T^*, \quad \gamma := bc_0/\Delta T. \]

Thus $\beta$ is a dimensionless activation energy and $\gamma$ is a dimensionless heat release parameter.

16.3. Formulate the method of Newton for the solution of the nonlinear algebraic system (16.16), (16.18), and (16.19).

16.4. A more refined model for resin curing would include diffusion of the chemical species.
So consider the model in Section 16.1.1, with (16.2b) replaced by

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial C}{\partial x} \right) - kC,$$

where $D$ is the diffusion coefficient.

(a) Propose boundary conditions for $C$ at the interfaces $x_{\text{int},1}$ and $x_{\text{int},2}$.

(b) Derive the discretised equations for this problem using central differences for the space derivatives and the $\vartheta$ method for time integration.

(c) Formulate Newton's method for the resulting nonlinear algebraic system.

---

### 16.2 Mechanical Etching of Glass by Powder Blasting

#### 16.2.1 Problem Setting

Some modern television displays have a vacuum enclosure that is internally supported by a glass plate. This plate must not hinder the display function. For that reason it has to be accurately patterned with small trenches or holes so that electrons can move from the cathode to the screen. One method of manufacturing such a glass plate is to cover it with an erosion-resistant mask and blast it with an abrasive powder. In this section we present a mathematical model for this powder-blasting or powder-erosion process and solve it using the method of characteristics.

In powder erosion of brittle materials sharp and abrasive particles in the powder cause deformations and microcracks in the target material. Chipping of target material above the cracks is the dominant erosion mechanism. The rate at which substrate material is removed is determined by the erosion rate $E$, which is defined as the ratio of the mass loss of the substrate and the mass of the erodent used. For the erosion rate we take the model

$$E = C (|v| \cos \theta)^k,$$

where $v$ is the particle impact velocity, $\theta$, the angle between $v$ and the inward normal $n$ to the surface of the substrate, and $C$ and $k$ are constants. Experimentally, it is observed that $2 \leq k \leq 4$ [43]. This is in agreement with theoretical models, which predict the value $k = 7/3$; see, e.g., [139].

In Figure 16.5 we show the formation of a pattern in a glass plate by masked erosion. The mask considered has a circular hole and is exposed to a constant particle mass flux; i.e., the particle flux and the impact velocity are constant. Time-lapse photography has exposed a picture at fixed time intervals. Some characteristics of the erosion process can be seen from these photographs. First, the hole is shallow close to the edge of the mask and has a sharp tip in the middle. The reason for this is that, due to the finite particle size, not all particles contribute to the erosion process close to the mask. On the other hand, for a wide, shallow hole we would get a flat bottom in the centre. Second, with increasing depth, the rate of growth of the hole decreases. This can be explained by the dependence of the erosion rate on the angle of impact $\theta$. Finally, with increasing depth, the hole takes a nipple shape.
16.2. Mechanical Etching of Glass by Powder Blasting

This is probably caused by rebounding particles from the steep slopes at the sides of the pattern [140]. These particles are focused on the centre of the pattern, where they generate additional erosion. We will not consider this so-called second-strike effect.

16.2.2 Mathematical Model for Powder Erosion

Consider an initially flat substrate of brittle material, covered with a mask. In Figure 16.6 we introduce an \((x, y, z)\) coordinate system, where the \((x, y)\) plane coincides with the initial substrate and where the positive \(z\) axis is directed into the material. A flux of particles with velocity \(v\) in the positive \(z\) direction hits the substrate and removes material. The position of the surface at time \(t\) is described by the function \(z = \zeta(x, y, t)\). We now derive a kinematic condition for the surface. Let \(c\) denote the velocity of the surface in the direction of the unit normal \(n\) on the surface, which is directed into the material and is given by

\[
n = \frac{1}{\sqrt{1 + \zeta_x^2 + \zeta_y^2}} \begin{pmatrix} -\zeta_x \\ -\zeta_y \\ 1 \end{pmatrix}.
\]

(16.21)

Consider a point \(x(t) = (x(t), y(t), z(t))^T\) on the surface at time \(t\). The displacement of this point satisfies the equation

\[
\frac{dx}{dt} = cn.
\]

(16.22)

Figure 16.5. Pattern formation in glass by masked erosion (courtesy P.J. Slikkerveer, Philips Research Laboratories, Eindhoven).

Figure 16.6. Cross section of a hole in the substrate.
On the other hand, from the relation \( z(t) = \zeta(x(t), y(t), t) \) we readily obtain

\[
\frac{dz}{dt} = \zeta_x \frac{dx}{dt} + \zeta_y \frac{dy}{dt} + \zeta_t.
\]  (16.23)

Inserting (16.22) in (16.23), we find

\[
cn_3 = c(\zeta x_{n_1} + \zeta y_{n_2}) + \zeta_t,
\]

where \( n_1, n_2, \) and \( n_3 \) denote the components of \( n \), and, combining this relation with (16.21), we find the following equation for \( \zeta \):

\[
\zeta_t - c\sqrt{1 + \zeta^2_x + \zeta^2_y} = 0.
\]  (16.24)

We will specify initial and boundary conditions for (16.24) later.

For \( c \) we assume the relation [140]

\[
c = \frac{1}{\rho_s} E \Phi \cdot n = \frac{1}{\rho_s} E \Phi \cos \theta,
\]  (16.25)

where \( \rho_s \) is the mass density of the substrate, \( E \) is the erosion rate, \( \Phi = \Phi e_z \) is the particle mass flux in the direction of \( v \), and \( \theta \) is the angle between \( v \) and \( n \). This means that only the particle flux perpendicular to the surface contributes to the velocity \( c \) of the surface. Likewise, the erosion rate \( E \) depends only on the normal component \( v \cdot n \) of the particle velocity \( v = ve_z \) and is given by (16.20). Substitution of (16.25) and (16.20) in (16.24) gives

\[
\zeta_t - \frac{C}{\rho_s} v^4 \Phi (1 + \zeta^2_x + \zeta^2_y)^{-k/2} = 0
\]

(16.26)

where we have used that \( \cos \theta = 1/\sqrt{1 + \zeta^2_x + \zeta^2_y} \).

Next we make (16.26) dimensionless. Introducing a characteristic length scale \( L \), a characteristic particle mass flux \( \tilde{\Phi} \), and the characteristic time \( T = L \rho_s / (C \tilde{\Phi} v^4) \), we obtain in a straightforward way

\[
\zeta_t - \Phi(x, y) (1 + \zeta^2_x + \zeta^2_y)^{-k/2} = 0.
\]  (16.27)

The characteristic length \( L \) can be the width of a two-dimensional trench or the radius of a rotationally symmetric hole. Note that \( T \) is the time needed to propagate the surface over a distance \( L \) when the particles hit the surface perpendicularly with mass flux \( \Phi \). In the derivation of (16.27) we have assumed that \( v \) is constant, which implies that the dimensionless particle mass flux \( \Phi(x, y) \) is equal to one. However, later we will see that this choice for the particle mass flux does not give the correct solution of (16.27). Therefore we will propose a better model for \( \Phi(x, y) \).

In the following we consider the special case of a two-dimensional trench. Let \( x \) denote the transverse coordinate in the trench, as indicated in Figure 16.6. Then (16.27) reduces to

\[
\zeta_t + \Phi(x) f(\zeta) = 0, \quad 0 < x < 1, \quad t > 0,
\]  (16.28a)

where the function \( f = f(p) \) is defined by

\[
f(p) := -(1 + p^2)^{-k/2}.
\]  (16.28b)
16.2. Mechanical Etching of Glass by Powder Blasting

Equation (16.28a) is supplemented with the initial and boundary conditions

\[
\begin{align*}
\zeta(x, 0) &= 0, & 0 < x < 1, \\
\zeta(0, t) &= \zeta(1, t) = 0, & t > 0.
\end{align*}
\] (16.28c, 16.28d)

The boundary conditions in (16.28d) mean that the trench cannot grow at the ends \(x = 0\) and \(x = 1\). We will see that, for a suitable choice of the particle mass flux \(\Phi(x)\), boundary conditions for \(\zeta\) follow directly from the initial condition. For the special case \(\zeta(x, 0) = 0\) we obtain (16.28d); in this situation these boundary conditions are in fact redundant.

### 16.2.3 Characteristic Strip Equations

Introducing the variables

\[
p := \zeta_x, \quad q := \zeta_t,
\] (16.29)

we can write the PDE (16.28a) in the canonical form

\[
F(x, t, \zeta, p, q) := q - \frac{k p}{\Phi_1(x) (1 + p^2)^{k/2}} = 0.
\] (16.30)

According to Theorem 12.6, the solution of (16.30) can be constructed from the following initial value problem for the characteristic strip equations:

\[
\begin{align*}
\frac{dx}{ds} &= F_p = \Phi(x) \frac{kp}{(1 + p^2)^{k/2+1}}, & x(0; \sigma) = \sigma, \\
\frac{dt}{ds} &= F_q = 1, & t(0; \sigma) = 0, \\
\frac{d\zeta}{ds} &= p F_p + q F_q = \Phi(x) \frac{1 + (k + 1)p^2}{(1 + p^2)^{k/2+1}}, & \zeta(0; \sigma) = 0, \\
\frac{dp}{ds} &= -(F_x + p F_\zeta) = \Phi'(x) \frac{1}{(1 + p^2)^{k/2}}, & p(0; \sigma) = 0, \\
\frac{dq}{ds} &= -(F_t + q F_\zeta) = 0, & q(0; \sigma) = \Phi(\sigma),
\end{align*}
\] (16.31a-e)

where \(s\) and \(\sigma\) are the parameters along the characteristics and the initial curve, respectively.

The initial condition for \(q\) follows from (16.30) and the initial conditions for the other variables. Note that the solution of (16.31b) and (16.31e) is trivial, and we find

\[
t(s; \sigma) = s, \quad q(s; \sigma) = \Phi(\sigma).
\] (16.32)

The formal solution procedure for the other equations is as follows. We solve (16.31a), (16.31c), and (16.31d), to find \(x = x(t; \sigma), \zeta = \zeta(t; \sigma),\) and \(p = p(t; \sigma).\) Inverting the first function, we have \(\sigma = \sigma(x, t),\) and substitution of the latter expression gives the final solution \(\zeta(x, t) := \zeta(t; \sigma(x, t))\) and \(p(x, t) := p(t; \sigma(x, t)).\) In the following we use the notation \(v(t; \sigma)\) or \(v(x; \sigma)\) for a generic variable \(v\) to indicate that an expression holds only along a characteristic parametrized by \(t\) or \(x.\) The parameter \(\sigma\) means that the characteristic passes through the point \((\sigma, 0).\) On the other hand, we use the notation \(v(x, t)\) if an expression holds in a part of the \((x, t)\) plane.
A first obvious choice for the dimensionless particle mass flux is \( \Phi(x) = 1 \). The solution of the initial value problem (16.31) in this case is trivial, and we find \( \zeta(x, t) = t \) and \( p(x, t) = 0 \), corresponding to a flat-bottomed hole. This particular solution is in contrast to experimental results, which show slanted sides near the edges of the mask; see Figure 16.5. As observed earlier, this is due to the finite particle size of the eroding powder which causes some particles not to be effective in the erosion process close to the mask. In order to model this phenomenon, we introduce transition regions of thickness \( \delta \), where we assume that \( \Phi(x) \) increases continuously and monotonically from zero at the boundaries to one at \( x = \delta, 1 - \delta \). The parameter \( \delta \) is characteristic of the (dimensionless) particle size and a typical value is \( \delta = 0.1 \). We adopt the simplest possible choice for \( \Phi(x) \); i.e.,

\[
\Phi(x) = \begin{cases} 
   x/\delta & \text{if } 0 \leq x < \delta, \\
   1 & \text{if } \delta \leq x \leq 1 - \delta, \\
   (1-x)/\delta & \text{if } 1 - \delta < x \leq 1.
\end{cases}
\]

(16.33)

As a result of (16.33), the growth rate of the surface position close to the mask is smaller than in the middle of the hole. Since \( \Phi(0) = \Phi(1) = 0 \), we obtain from (16.31) the solutions \( x(t; 0) = \zeta(t; 0) = 0 \) and \( x(t; 1) = 1, \zeta(t; 1) = 0 \), implying that the boundary conditions (16.28d) for \( \zeta \) are automatically satisfied. Moreover, since \( \Phi'(0) = 1/\delta \), we can directly compute \( p(t; 0) \) from (16.31d); likewise we can compute \( p(t; 1) \). Thus we are not allowed to specify boundary conditions for \( p \). However, the most important consequence of (16.33) is that characteristics originating from the transition regions are directed into the interior domain for increasing \( t \). This implies that the reduced particle mass flux in the transition regions will also influence the solution in the interior domain, which is in agreement with experimental results. We want to emphasize that (16.33) is merely a sensible guess for the particle mass flux that still allows computation of the analytical solution of (16.31). Other choices are possible.

By introducing transition regions, we create intersecting characteristics. Therefore the solution of (16.30) can only be a weak solution and it is anticipated that shocks will emerge from the edges \( x = \delta \) and \( x = 1 - \delta \). Let \( x = \xi_{i,1}(t) \) and \( x = \xi_{i,2}(t) \) denote the location of the shocks at time \( t \) originating at \( x = \delta \) and \( x = 1 - \delta \), respectively. Each point \((\xi_{i,1}(t), t)\) and \((\xi_{i,2}(t), t)\) on these shocks is connected to two different characteristics that exist on both sides of the shocks. The speed of these shocks is given by the jump condition

\[
\frac{d\xi_{i,j}}{dt} [p] = -[\Phi(x)(1 + p^2)^{-1/2}]^+ \quad (i = 1, 2).
\]

(16.34)

Thus we can distinguish the following five regions in the \((x, t)\) plane: the left transition region \( 0 \leq x \leq \delta \) (region 1), the right transition region \( 1 - \delta \leq x \leq 1 \) (region 2), the interior domain left of the first shock (region 3), the interior domain right of the second shock (region 4), and the region between the two shocks (region 5); see Figure 16.7. Note that the location of the shocks depends on the solution through (16.34) and has to be computed as well.

### 16.2.4 Solution of the Characteristic Strip Equations

First, consider region 1. If we choose \( x \) as the parameter along the characteristics instead of \( t \), we can derive from (16.31) the following differential equations for the slope \( p \) and the
16.2. Mechanical Etching of Glass by Powder Blasting

Figure 16.7. Characteristics and shocks of a two-dimensional trench for \( \delta = 0.1 \) and \( k = 2 \).

The surface position \( \xi \):

\[
\frac{dp}{dx} = \frac{1 + p^2}{k \sigma}, \quad \frac{d\xi}{dx} = \frac{1 + (k + 1)p^2}{k \sigma}.
\]  (16.35)

Using the initial conditions \( x(0; \sigma) = \sigma \) and \( p(0; \sigma) = 0 \), we can solve the first equation to find

\[
p(x; \sigma) = \sqrt{\left(\frac{x}{\sigma}\right)^{2/k} - 1}.
\]  (16.36)

Combining both differential equations in (16.35), we obtain the following equation for \( \xi \):

\[
\frac{d\xi}{dx} = p + \frac{1 + p^2}{k \sigma} = p + x \frac{dp}{dx}.
\]  (16.37)

Together with the initial conditions \( \xi(0; \sigma) = p(0; \sigma) = 0 \), (16.37) has the solution

\[
\xi(x; \sigma) = xp(x; \sigma).
\]  (16.38)

The solutions in (16.36) and (16.38) give \( p \) and \( \xi \) along the characteristic through the point \((\sigma, 0)\) on the initial curve. The location of this characteristic follows from the differential equation

\[
\frac{dx}{dt} = \frac{k \sigma}{\delta} \sqrt{(x/\sigma)^{2/k} - 1},
\]  (16.39)
which follows readily after substitution of (16.36) in (16.31a). Integrating (16.39) subject to the initial condition \( x(0; \sigma) = \sigma \) and choosing \( p \) as the independent variable, we arrive at the relation
\[
\frac{t}{\delta} = T(p) := \int_0^p \left(1 + r^2\right)^{k/2} \, dr.
\] (16.40)

Alternatively, formula (16.40) can be obtained by direct integration of (16.31d). Points \((x, t)\) on the characteristic through \((\sigma, 0)\) are thus determined by (16.36) and (16.40). Note that the slope \( p \) in region 1 is a function of \( t/\delta \) only and is independent of the spatial coordinate \( x \). When computing the solution in region 1 at a given time level \( t \), we first solve (16.40) for \( p \) and subsequently compute \( \zeta \) from (16.38).

From (16.36), (16.38), and (16.40) we see that the characteristic through \((\sigma, 0)\) \((0 < \sigma < \delta)\) reaches the edge \( x = \delta \) at time \( t_1(\sigma) \), with slope \( p_1(\sigma) \) and surface position \( \zeta_1(\sigma) \) given by
\[
p_1(\sigma) := \sqrt{\left(\frac{\delta}{\sigma}\right)^{2/k} - 1}, \quad \zeta_1(\sigma) := \delta p_1(\sigma), \quad t_1(\sigma) := \delta T(p_1(\sigma)).
\] (16.41)

These are the ‘initial’ conditions for the solution of initial value problem (16.31) in region 3.

Now consider region 3. Since \( \Phi'(x) = 0 \), the slope \( p \) is constant along characteristics and, consequently, the differential equations in (16.31) can easily be solved. We find
\[
p(x; \sigma) = p_1(\sigma), \quad \zeta(x; \sigma) = \zeta_1(\sigma) + \frac{1 + (k + 1)p_1^2(\sigma)}{kp_1(\sigma)}(x - \delta), \quad t(x; \sigma) = t_1(\sigma) + \frac{(1 + p_1^2(\sigma))^{k/2+1}}{kp_1(\sigma)}(x - \delta).
\] (16.42a, b, c)

The characteristics in this region are straight lines and are displayed in Figure 16.7. To compute the solution at a given point \((x, t)\) we first have to solve (16.42c) for \( p_1(\sigma) \) and subsequently compute \( p(x, t) \) and \( \zeta(x, t) \) from the other two relations in (16.42).

The solution of (16.31) in region 5 is trivial and is given by
\[
x(t; \sigma) = \sigma, \quad p(x, t) = 0, \quad \zeta(x, t) = t;
\] (16.43)

see Figure 16.7. The characteristics are now vertical lines through \((\sigma, 0)\), corresponding to a flat surface.

Since the particle mass flux (16.33) is symmetric around \( x = 0.5 \), it is also clear that the solution of the initial value problem (16.31) is symmetric around \( x = 0.5 \). Consequently, the solution in regions 2 and 4 can easily be obtained from the corresponding solutions in regions 1 and 3, respectively. However, for the sake of brevity, we omit the explicit formulas.

Finally, we have to determine the location of the shocks. We only consider the first shock, emanating from the edge \( x = \delta \). The computation of the second shock is completely analogous. The evolution of the first shock is determined by the jump condition (16.34), which in this case leads to the initial value problem
\[
\frac{d\xi_{s,1}}{dr} = \frac{1}{p_{s,1}} \left(1 - (1 + p_{s,1}^2)^{-k/2}\right), \quad \xi_{s,1}(0) = \delta,
\] (16.44)
with \( p_{s,1} := p_1(\sigma) \) being the value of the slope just left of the shock on the characteristic through \((\delta, t_1(\sigma))\). Thus the following relation holds between \( p_{s,1} \) and \( \xi_{s,1} \):

\[
\frac{\xi_{s,1} - \delta}{t - t_1(\sigma)} = \frac{k p_{s,1}}{(1 + p_{s,1}^2)^{1/2 + \Gamma}}.
\] (16.45)

Combining relation (16.45) with the formulas for \( t_1(\sigma) \) given in (16.41), differentiating the resulting equation with respect to \( t \), and substituting (16.44), we find the following initial value problem for \( p_{s,1} \):

\[
\frac{dp_{s,1}}{dt} = \frac{p_{s,1}^2}{(1 + p_{s,1}^2)^{1/2}} \left( k - \frac{(1 + p_{s,1}^2)^{1/2 - 1}}{p_{s,1}} \right) \left( 1 - p_{s,1}^2 \right), \quad p_{s,1}(0) = 0.
\] (16.46)

The initial value for \( p_{s,1} \) follows, e.g., from (16.40). The propagation of the shock is determined by the differential equations (16.44) and (16.46), which we have to solve numerically. The result is shown in Figure 16.7.

We have collected the results of this section in Figure 16.8, which gives the solution for \( \zeta \) and \( p \) at time levels \( t = 0.0, 0.1, \ldots, 1.0 \) for \( \delta = 0.1 \) and \( k = 2.33 \). This figure nicely displays the features of the solution: a slanted surface in the transition regions, a flat bottom in the interior domain, and a curved surface in between. Also, the inwardly propagating shocks are clearly visible.

To validate our model we compare in Figure 16.9 the analytical solution with experimental results at four different time levels for the erosion of a trench. All surfaces are dimensionless according to the scaling in Section 16.2.2. We have computed the surfaces with \( k = 2.33 \) and \( \delta = 0.1 \). We see a nice qualitative agreement between analytical and experimental results. Both show the shallow, flat-bottomed solution in the middle of the hole at \( t = 0.5 \). However, the analytically computed hole at \( t = 0.5 \) is somewhat deeper than the experimental one. The reason for this is probably the shadow effect of the mask, which is not included in the analytical model. For \( t \geq 1.6 \) the experimental surfaces show
a nipple shape in the middle of the hole, which is probably caused by rebounding particles from the slopes at the sides of the surface. This second-strike effect is also not included in our model. Further differences are the round top of the experimental holes due to the finite particle size and the widening of the experimental holes due to mask wear.

16.2.5 Discussion and Related Problems

- Etching is a technique for removing material from a layer by means of either chemical or mechanical processes. It is used to accurately pattern layers and is often applied in the manufacture of semiconductor devices. A mathematical description of etching processes is given in, e.g., [75].

- All etching techniques involve a moving surface. The kinematic condition describing the evolution of such a surface is a first order nonlinear PDE like (16.28a). In this section we solved this equation analytically using the method of characteristics. Alternatively, we could solve it numerically. Another approach is to apply level set methods [137, 108].

- Moving interfaces occur in many problems. Examples are the growth of crystals or the propagation of flames [137].

16.2.6 Exercises

16.5. Compute the solution of the initial value problem (16.31) in regions 2 and 4.

16.6. Derive the initial value problem for the second shock, emanating from the edge \( x = 1 - \delta \).

16.7. Show that \( p := \zeta_x \) satisfies the quasi-linear equation

\[
p_t + (\Phi(x)f(p))_x = 0.
\]

Use this equation to derive the jump condition (16.34).
16.3 Thermal Explosion in a Vessel

16.3.1 Problem Formulation

In a vessel containing a mixture of reacting species the temperature can increase tremendously, causing great damage to the plant where it is situated. To be more precise, if an exothermic reaction takes place, heat is released, which causes an increase of temperature. Often, the reaction rate increases with increasing temperature, leading to more production of heat, etc. In this way we get a self-accelerating chemical reaction, which continues until the reactant is depleted. Some mixtures can be characterized by the existence of a critical temperature \(T_{\text{crit}}\), below which there is virtually no reaction, whereas above this temperature the reaction proceeds extremely fast. Thus at the critical temperature the reaction rate changes very suddenly, practically discontinuously. This process is called a thermal explosion. In this section we study a simple model for this type of process.

16.3.2 Mathematical Model

As an example, we consider a spherical vessel containing a reacting mixture. Assuming spherical symmetry and a one-step irreversible reaction and neglecting the flow of the mixture, we have the initial boundary value problem [171]

\[
\begin{align*}
\frac{\partial T}{\partial t} &= a \frac{\partial}{\partial r} \left( r^2 \frac{\partial T}{\partial r} \right) + q w(C, T), \quad 0 < r < L, \quad t > 0, \\
\frac{\partial C}{\partial t} &= b \frac{\partial}{\partial r} \left( r^2 \frac{\partial C}{\partial r} \right) - w(C, T), \\
T(r, 0) &= T_0, \quad C(r, 0) = C_0, \quad 0 < r < L, \\
\frac{\partial T}{\partial r} (0, t) &= 0, \quad \frac{\partial C}{\partial r} (0, t) = 0, \quad t > 0, \\
T(L, t) &= T_{\text{amb}}, \quad \frac{\partial C}{\partial r} (L, t) = 0, \quad t > 0,
\end{align*}
\]  

(16.47a) \quad (16.47b) \quad (16.47c) \quad (16.47d) \quad (16.47e)

where \(T\) is the temperature of the mixture and \(C\) is the concentration of the reactants. The constants \(a, b, \) and \(q\) are the thermal diffusivity, the diffusion coefficient, and the heat release
Table 16.2. Parameters for a thermal explosion.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>1.0</td>
<td>m²/s</td>
<td>( A )</td>
<td>( 3.15 \times 10^{10} )</td>
<td>s⁻¹</td>
</tr>
<tr>
<td>( b )</td>
<td>2.0</td>
<td>m²/s</td>
<td>( L )</td>
<td>1</td>
<td>m</td>
</tr>
<tr>
<td>( q )</td>
<td>( 8.1 \times 10^3 )</td>
<td>m³-K/mol</td>
<td>( T_0 )</td>
<td>293</td>
<td>K</td>
</tr>
<tr>
<td>( E_a )</td>
<td>( 7.2 \times 10^4 )</td>
<td>J/mol</td>
<td>( C_0 )</td>
<td>0.3</td>
<td>mol/m³</td>
</tr>
<tr>
<td>( R )</td>
<td>8.314</td>
<td>J/(mol·K)</td>
<td>( T_{amb} )</td>
<td>308</td>
<td>K</td>
</tr>
</tbody>
</table>

parameter, respectively. The variable \( w \) is the reaction rate and is given by

\[
 w(C, T) = AC \exp(-E_a/RT),
\]

with \( A, E_a, \) and \( R \) being the preexponential factor, the activation energy, and the universal gas constant, respectively.

The values of all physical parameters/constants are given in Table 16.2. The initial boundary value problem (16.47) is a characteristic model for thermal explosions; see Chapter 7. Its solution has the characteristic features of a thermal explosion; however, details of, e.g., the chemistry or flow are not included. These require a much more complicated model.

16.3.3 Estimating the Induction Time

The solution of the initial boundary value problem (16.47) is characterized by an initial period of virtually no reaction called the induction period. Then, suddenly, a very intense chemical reaction takes place in which the temperature increases tremendously. In this section we will estimate the induction time \( t_{ind} \) at which this temperature increase occurs. For this we assume that the vessel is a perfectly stirred reactor; i.e., the temperature \( T \) and the concentration \( C \) are independent of the spatial coordinate \( r \). Then (16.47a) and (16.47b) reduce to

\[
\frac{dT}{dt} = qACe^{-T_a/T}, \quad \frac{dC}{dt} = -ACe^{-T_a/T},
\]

where \( T_a := E_a/R \) is the activation temperature. Taking a suitable linear combination of both equations in (16.49), we obtain the integral

\[
T(t) + qC(t) = T_0 + qC_0 = T_{max},
\]

where \( T_{max} \) is the maximum temperature at which the reactant is depleted, i.e., when \( C = 0 \). Substituting this relation in the ODE for \( T \) and integrating the resulting equation, we obtain

\[
\int_{T_0}^{T} \frac{1}{T_b - u} e^{T_a/u} du = At.
\]

Note that the integral in (16.51) diverges for \( T \to T_b \), implying that the mixture reaches the depleted state \( T = T_{max}, C = 0 \) for \( t \to \infty \). Substituting \( v = T_a/u \) and splitting the
integrand into partial fractions, we obtain the relation
\[
E_i \left( \frac{T_a}{T} \right) - E_i \left( \frac{T_a}{T_0} \right) - e^{T_a/T_b} \left[ E_i \left( \frac{1}{T} - \frac{1}{T_b} \right) - E_i \left( \frac{1}{T_0} - \frac{1}{T_b} \right) \right] = At,
\]
(16.52)
where \( E_i(x) \) is the exponential integral defined by
\[
E_i(x) := \int_{-\infty}^{x} \frac{e^t}{t} \, dt.
\]
(16.53)
The induction time \( t_{\text{ind}} \) is now defined as the time needed to reach the temperature \( T = T^* := T_{\text{max}}(1 - T_{\text{max}}/T_a) \) is where the reaction rate \( w^*(T) \), defined by
\[
w^*(T) := w\left((T_{\text{max}} - T)/q, T\right) = A(T_{\text{max}} - T) \exp(-T_a/T),
\]
is maximal. Substituting this expression for \( T^* \) in (16.52), using the asymptotic approximation (see, e.g., [1] or Example 15.11)
\[
E_i(x) = e^x/x (1 + O(x^{-1})) \quad \text{for} \quad x \to \infty
\]
(16.54)
and taking into account that \( T_a/T^*, T_a/T_0, T_a(1/T^* - 1/T_{\text{max}}) \), and \( T_a(1/T_0 - 1/T_{\text{max}}) \) are large, we obtain asymptotically that
\[
\left( T^* \right) \left( \frac{T_{\text{max}}}{T_a} \right) e^{T_a/T^*} - \left( T_0 \right) \left( \frac{T_{\text{max}}}{T_a} \right) e^{T_a/T_0} \simeq At_{\text{ind}}.
\]
(16.55)
Finally, taking into account that \( e^x(1/T^* - 1/T_{\text{max}}) \) is exponentially small, we find the approximation
\[
t_{\text{ind}} \simeq \frac{1}{A} \frac{T_0^2}{T_{\text{max}} - T_0} e^{T_a/T_0},
\]
(16.56)
with \( T_{\text{max}} \) given in (16.50). For the parameter values in Table 16.2 we find \( t_{\text{ind}} = 0.8883 \) s.

### 16.3.4 Numerical Solution Method

For the numerical solution of initial boundary value problem (16.47) we employ the finite volume method for space discretisation in combination with the \( \theta \) method for time integration. Thus we cover the interval \([0, L]\) with control volumes \( V_j \) as follows:
\[
V_j := (r_{j-1/2}, r_{j+1/2}) = ((j - 1)\Delta r, j\Delta r), \quad j = 1, 2, \ldots, N_r,
\]
(16.57)
with \( \Delta r = L/N_r \), the spatial grid size. Each control volume \( V_j \) is a spherical shell with \( r_{j-1/2} \leq r \leq r_{j+1/2} \) centred around \( r_j = (j - 1/2)\Delta r \). We introduce the discrete time levels
\[
t^n := n\Delta t \quad (n = 0, 1, 2, \ldots, M), \quad \Delta t = \frac{t_{\text{fin}}}{M},
\]
(16.58)
with \( t_{\text{fin}} \) the final simulation time. As usual, \( T_j(t) \) and \( T_j^n \) denote the semi- and fully discrete approximations of \( T(r_j, t) \) and \( T(r_j, t^n) \), respectively. \( C_j(t) \) and \( C_j^n \) are defined likewise.
Integrating (16.47a) and (16.47b) over a control volume $V_j$ and applying Gauss’s theorem (see the appendix, Section J), we find

\[
\frac{d}{dt} \int_{V_j} T \, dV = a \oint_{\partial V_j} \frac{\partial T}{\partial n} \, dS + q \int_{V_j} w(C, T) \, dV, \tag{16.59a}
\]

\[
\frac{d}{dt} \int_{V_j} C \, dV = b \oint_{\partial V_j} \frac{\partial C}{\partial n} \, dS - \int_{V_j} w(C, T) \, dV. \tag{16.59b}
\]

Next we have to approximate the integrals in (16.59). For the volume integrals we use the midpoint rule. So we have for the first integral in (16.59a)

\[
\int_{V_j} T \, dV \approx T(r_j, t) |V_j| = \frac{4}{3} \pi \left( r_{j+1/2}^3 - r_{j-1/2}^3 \right) T(r_j, t), \tag{16.60}
\]

with $|V_j|$ the volume of $V_j$. The same goes for the other volume integrals. Computing the surface integrals is slightly more complicated. Consider, e.g., the surface integral in \((16.59a)\). Each control volume is bounded by the spheres $r = r_{j+1/2}$ and $r = r_{j-1/2}$ with outward unit normals $e_r$ and $-e_r$, respectively. Applying the midpoint rule once more, we obtain

\[
\oint_{\partial V_j} \frac{\partial T}{\partial n} \, dS \approx 4 \pi \left( r_{j+1/2}^2 \frac{\partial T}{\partial r}(r_{j+1/2}, t) - r_{j-1/2}^2 \frac{\partial T}{\partial r}(r_{j-1/2}, t) \right). \tag{16.61}
\]

For the spatial derivatives at the interface of $V_j$ we use standard central differences. This way, we obtain

\[
\oint_{\partial V_j} \frac{\partial T}{\partial n} \, dS \approx \frac{4 \pi}{\Delta r} \left( r_{j+1/2}^2 (T(r_{j+1}, t) - T(r_j, t)) - r_{j-1/2}^2 (T(r_j, t) - T(r_{j-1}, t)) \right). \tag{16.62}
\]

A similar expression holds for the surface integral in (16.59b). Inserting the above quadrature rules in (16.59), we obtain the semidiscrete system

\[
\frac{dT_j}{dt} = a \left( \alpha_{j+1/2}(T_{j+1} - T_j) - \beta_{j-1/2}(T_j - T_{j-1}) \right) + q w(C_j, T_j), \tag{16.63a}
\]

\[
\frac{dC_j}{dt} = b \left( \alpha_{j+1/2}(C_{j+1} - C_j) - \beta_{j-1/2}(C_j - C_{j-1}) \right) - w(C_j, T_j), \tag{16.63b}
\]

where the geometrical factors $\alpha_{j+1/2}$ and $\beta_{j-1/2}$ are defined by

\[
\alpha_{j+1/2} := \frac{3}{\Delta r^3} \frac{r_{j+1/2}^3}{r_{j+1/2}^3 + r_{j+1/2} r_{j-1/2} + r_{j-1/2}^3}, \tag{16.64a}
\]

\[
\beta_{j-1/2} := \frac{3}{\Delta r^3} \frac{r_{j-1/2}^3}{r_{j+1/2}^3 + r_{j+1/2} r_{j-1/2} + r_{j-1/2}^3}. \tag{16.64b}
\]
Finally, we apply the \( \vartheta \) method to (16.63) to obtain

\[
T_j^{n+1} - \vartheta \Delta t \left( a \left( \alpha_{j+1/2} \left( T_{j+1}^{n+1} - T_j^n \right) - \beta_{j-1/2} \left( T_j^{n+1} - T_{j-1} \right) \right) + q \left( C_j^{n+1} - T_j^{n+1} \right) \right) = T_j^n + \left( 1 - \vartheta \right) \Delta t \left( a \left( \alpha_{j+1/2} \left( T_{j+1}^{n} - T_j^n \right) - \beta_{j-1/2} \left( T_j^n - T_{j-1} \right) \right) + q \left( C_j^n - T_j^n \right) \right),
\]

(16.65a)

\[
C_j^{n+1} - \vartheta \Delta t \left( b \left( \alpha_{j+1/2} \left( \frac{C_j^{n+1} - C_{j+1}^{n+1}}{\vartheta} \right) - \beta_{j-1/2} \left( \frac{C_j^{n+1} - C_{j-1}^{n+1}}{\vartheta} \right) \right) - w \left( C_j^{n+1} - T_j^{n+1} \right) \right) = C_j^n + \left( 1 - \vartheta \right) \Delta t \left( b \left( \alpha_{j+1/2} \left( \frac{C_j^n - C_{j+1}^n}{\vartheta} \right) - \beta_{j-1/2} \left( \frac{C_j^n - C_{j-1}^n}{\vartheta} \right) \right) - w \left( C_j^n - T_j^n \right) \right).
\]

(16.65b)

The equations in (16.65) constitute a nonlinear system coupled through the reaction term \( w(C_j^{n+1}, T_j^{n+1}) \). Introducing the vector of unknowns \( u := \left( T_1^{n+1}, T_2^{n+1}, C_1^{n+1}, \ldots, T_N^{n+1}, C_N^{n+1} \right)^T \), we can write (16.65) in the form

\[
A(T^{n+1})u = b,
\]

(16.66a)

where the coefficient matrix \( A \) depends only on the temperature vector \( T^{n+1} := \left( T_1^{n+1}, T_2^{n+1}, \ldots, T_N^{n+1} \right)^T \). This system could be solved using (damped) Newton iteration with the solution at time level \( t^n \) as initial guess. However, since the reaction term \( w(C, T) \) is extremely sensitive to the temperature \( T \), it is expected that Newton iteration would not converge unless the initial guess was very close to the exact solution, i.e., the time step \( \Delta t \) was extremely small. Alternatively, we solve the linear system

\[
A(T^n)u = b,
\]

(16.66b)

which we obtain by simply replacing \( T^{n+1} \) by \( T^n \) in the coefficient matrix. This corresponds to one Picard iteration for (16.66a). Moreover, to improve the robustness of the method, we bound the numerical solution according to \( T_0 \leq T \leq T_{\text{max}} \) and \( 0 \leq C \leq C_0 \).

As an example, we have computed a solution for \( T_0 = 293 \, \text{K}, \, C_0 = 0.3 \, \text{mol/m}^3 \), and \( T_{\text{amb}} = 308 \, \text{K} \). The numerical parameters are \( \vartheta = 1, \, N_r = 100, \, t_{\text{fin}} = 2, \, \) and \( M = 1000 \). The results are presented in Figures 16.10 and 16.11. The temperature profiles in Figure 16.10 are plotted every 10 time steps. The solution displays the following behaviour. Initially, the temperature increases close to the boundary \( r = 1 \) through conduction. Gradually, the temperature also slightly increases in the interior domain. However, the concentration is still virtually 0.3 everywhere. The increasing temperature induces a chemical reaction, which in turn induces heat production. This process is very slow, until \( t \approx 0.80 \), which is approximately 10 percent off the induction time computed in Section 16.3.3; see Figure 16.11. Then, all of a sudden, the temperature jumps to its maximum value and the concentration rapidly decreases to zero. Next the temperature decreases slowly through conduction. Finally, the temperature and the concentration tend to their steady values \( T = T_{\text{amb}} \) and \( C = 0 \), respectively.

### 16.3.5 Discussion and Related Problems

- Although our model for a thermal explosion is a simple one, taking into account only conduction/diffusion and a one-step chemical reaction, it contains the essential characteristics. On the other hand, a more realistic model for thermal explosions should
contain a description of the (turbulent) flow; more complex chemistry, involving
many species and reactions; and possibly more complicated boundary conditions.

- The thermal explosion is an example of a reaction-diffusion process. Many of these
  processes occur in, e.g., chemistry, combustion theory, and population dynamics. In
  this sense it is related to the resin-curing process (Section 16.1) and the reaction in a
catalyzer pellet (Section 16.8).

- The activation energy $E_a$ is usually very large, making the problem even more non-
  linear. On the other hand, this makes the problem suitable for large activation energy
  asymptotics [170].

- The system (16.65) is nonlinear in the temperature, putting very high requirements on
  a nonlinear solver. Newton iteration is most likely not to converge unless the initial
guess is extremely close to the exact solution. Therefore Newton iteration has to be
combined with continuation methods to provide a good initial guess (see, e.g., [3])
and embedding techniques to improve the convergence behaviour.

---

**Figure 16.10.** Temperature and concentration distribution in a reactor vessel.
Exercises 523

Figure 16.11. Temperature and concentration at the centre of a reactor vessel.

• The initial boundary value problem (16.47) can suitably be solved using moving grid techniques, since its solution is rapidly varying in a small and moving part of the domain.

16.3.6 Exercises

16.10. The semidiscrete equations (16.63) hold only for control volumes in the interior of \([0, L]\), i.e., for control volumes not adjacent to the boundary. Give the derivation of the semidiscrete equations for the control volumes \(V_1\) and \(V_N\), neighbouring \(r = 0\) and \(r = L\), respectively.

16.11. Formulate the method of Newton to compute \(T^{n+1}\) and \(C^{n+1}\) from (16.66a). An obvious initial guess would be \(T^n\) and \(C^n\). Could you come up with a better initial guess?

16.4 Determining Viscoelastic Material Parameters

16.4.1 Problem Formulation

In order to measure the stress parameters of a viscoelastic material of Kelvin–Voigt type (see Section 6.8.3), the following experiment is done. A cylinder, filled with the material in question, is closed at one end, \(z = 0\), by a solid wall, while at the other end, \(z = L\), it is closed by a freely movable piston; see Figure 16.12. Before time \(t = 0\), the medium is at rest, undeformed, and free of stress. As of \(t = 0\) the medium is suddenly compressed.
by a constant force applied to the piston. The force is large enough to measure but small enough to keep the deformations linear. As the deformation in the radial direction remains constrained by the cylinder walls, the experiment is known as the “constrained compression test.” The sought parameter values are eventually obtained from the time history and the final displacement of the piston and the final pressure at the wall. In order to collect these values accurately, we need a model that relates the parameters to the measured data.

The usual approach is to apply a linear model for the deformations, neglecting any acceleration effects. This is called quasi-static model. Apart from the question of when inertia may be neglected (it depends on the parameters that we are going to measure), this approach is not consistent with an initial value problem because the elastic reaction in the material exists only with deformation. So right at the beginning, when there is no reaction yet, the material deforms with a large acceleration, while at the same time the deformation and stresses predicted by the quasi-static model are completely wrong.

For this reason we are interested in analysing a model that includes the inertia and investigating its role in the initial phase and beyond.

16.4.2 The Model

We introduce $\bm{T}$ denoting the stress tensor (see (6.19)), $\bm{E}$ the deformation tensor, $\bm{s}$ the deviatoric stress tensor (see the appendix, Section L.2), $\bm{e}$ the deviatoric deformation tensor, $\bm{u}$ the displacement vector, $\kappa$ the bulk modulus, $\eta$ the viscosity, $\rho$ the density, and $\gamma$ the relaxation rate. The parameters to be determined are thus $\kappa$, $\eta$, and $\gamma$. The constitutive equations of a Kelvin–Voigt medium are given by (6.53a), with

$$3\lambda \theta_1 + 2\mu \theta_2 = 0, \quad \kappa = \lambda + \frac{2}{3} \mu, \quad \eta = \mu \theta_2, \quad \gamma = \theta_2^{-1}. \quad (16.67)$$

We obtain the equations

$$\text{tr}(\bm{T}) = 3\kappa \text{tr}(\bm{E}) \quad \text{or} \quad \sum_{k=1}^{3} t_{kk} = 3\kappa \sum_{k=1}^{3} e_{kk}, \quad (16.67)$$

$$\frac{s}{2\eta} = \dot{\epsilon} + \gamma \epsilon \quad \text{or} \quad \frac{s_{ij}}{2\eta} = \dot{\epsilon}_{ij} + \gamma \epsilon_{ij} \quad (i, j = 1, 2, 3). \quad (16.68)$$

The linear relations involve $\bm{E} = \frac{1}{2} \nabla \bm{u} + \frac{1}{2} (\nabla \bm{u})^T$. Initially, we have $\epsilon = \mathbf{0}$. As we will investigate the role of inertia, we use Cauchy’s equation of motion (6.25):

$$\nabla \cdot \bm{T} = \rho \ddot{\bm{u}} \quad \text{or} \quad \sum_{j=1}^{3} t_{i,j,j} = \rho \ddot{u}_i \quad (i = 1, 2, 3). \quad (16.69)$$
16.4. Determining Viscoelastic Material Parameters 525

where we have used the compact notation that an index before the comma denotes a vectorial component and an index after the comma denotes a partial derivative. So \( u_{i,j} = \frac{\partial}{\partial x_j} u_i \). Furthermore, the dot denotes a time derivative, so \( \dot{u}_i = \frac{\partial}{\partial t} u_i \). In cylindrical coordinates \((r, \phi, z)\) the axial, radial, and circumferential displacements are given by \((u_z, u_r, u_\phi)\). From the symmetry of the problem there is no displacement or variation in \(\phi\). Furthermore, we assume no displacement or variation in the radial direction due to the rigid container wall. This assumption may be not justified if the fluid pulls at the wall, i.e., if the normal stress in the radial direction is positive. In this case the fluid may separate from the wall.

The radial component of (16.69) is then

\[
(\nabla \cdot \mathbf{T})_r = t_{rr} - t_{\phi\phi} = 0,
\]

and so

\[
t_{\phi\phi} = t_{rr}.
\]

The wall is assumed perfectly smooth, resulting in no shear stresses. From

\[
\sigma_{zz} = \frac{2}{3} \sigma_{zz} - \frac{2}{3} \sigma_{tr} = \frac{4}{3} \eta (u_{zz,rt} + \gamma u_{zz}), \quad t_{zz} = 2t_{tr} = 3\kappa u_{zz},
\]

we obtain

\[
t_{zz} = \frac{4}{3} \eta u_{zz} + \left(\kappa + \frac{4}{3} \eta \gamma\right) u_{zz},
\]

and so

\[
m_{zz} = \rho \ddot{u}_z = \frac{4}{3} \eta u_{zz} + \left(\kappa + \frac{4}{3} \eta \gamma\right) u_{zz},
\]

while the normal stress in the radial direction \(t_{rr}\) (which is minus the pressure at the wall) is given by

\[
t_{rr} = \left(\kappa - \frac{2}{3} \eta \gamma\right) u_{zz} - 2 \frac{\eta}{3} u_{zz}.
\]

At \(z = L\) we apply the external pressure \(p\), so we have the boundary condition

\[
t_{zz} = \frac{4}{3} \eta u_{zz} + \left(\kappa + \frac{4}{3} \eta \gamma\right) u_{zz} = -p \text{ at } z = L.
\]

We make expression (16.72) dimensionless as follows:

\[
z = L z^*, \quad t = \frac{4}{\kappa + \frac{3}{2} \eta \gamma} t^*, \quad u_z = \frac{pL}{\kappa + \frac{3}{2} \eta \gamma} w
\]

(asterisks * will be suppressed in the following) and introduce the parameter

\[
\varepsilon = \left(\frac{3L}{4\eta}\right)^2 \left(\kappa + \frac{4}{3} \eta \gamma\right) \rho.
\]

Neglect of inertia is equivalent to \(\varepsilon = 0\). So we are interested in the solution for \(\varepsilon\) small.

We obtain in dimensionless form the initial boundary value problem

\[
\frac{\partial^3 w}{\partial z^3 \partial t} + \frac{\partial^2 w}{\partial z^2 \partial t} = \varepsilon \frac{\partial^2 w}{\partial t^2} \quad \text{for} \quad t > 0, \quad 0 < z < 1,
\]

\[
w(z, 0) = \frac{\partial w}{\partial t}(z, 0) = 0 \quad \text{for} \quad 0 < z < 1,
\]

\[
w(0, t) = 0 \quad \text{for} \quad t > 0,
\]

\[
\frac{\partial w}{\partial z}(1, t) + \frac{\partial^2 w}{\partial z \partial t}(1, t) = -1 \quad \text{for} \quad t > 0,
\]

\[
\frac{\partial w}{\partial z}(0, t) + \frac{\partial^2 w}{\partial z \partial t}(0, t) = 0 \quad \text{for} \quad t > 0.
\]
where \(0 < \epsilon \ll 1\). For completeness the dimensional normal stresses in the \(z\) and \(r\) directions are given by

\[
t_{zz} = p \left( \frac{\partial^2 w}{\partial z \partial t} + \frac{\partial w}{\partial z} \right),
\]

(16.78a)

\[
t_{rr} = p \left( \frac{\partial w}{\partial z} - \frac{1}{2} \frac{\partial^2 w}{\partial z \partial t} \right), \quad \text{where } \alpha = \frac{\kappa - \frac{2}{3} \eta \gamma}{\kappa + \frac{4}{3} \eta \gamma}.
\]

(16.78b)

### 16.4.3 Analysis and Solution

An explicit solution in series or integral form may be constructed (see the exercises). This is, unfortunately, by its form, of limited value if insight and understanding are the goal. Therefore we will further reduce the complexity of the problem by utilizing the small parameter \(\epsilon\) and constructing an approximate solution by perturbation methods. The problem will be seen to be of singular perturbation type with boundary layer behaviour in the time coordinate near \(t = 0\), and the solution will be a matched asymptotic expansion (MAE) (see Section 15.4.1).

First, we will assume the existence of an asymptotic expansion

\[
w(z, t; \epsilon) = w_0(z, t) + \epsilon w_1(z, t) + \mathcal{O}(\epsilon^2),
\]

(16.79)

which starts with an \(\mathcal{O}(1)\) term in view of the boundary condition at \(z = 1\). The fact that the correction term is of \(\mathcal{O}(\epsilon)\) is an assumption based on the observation that each correction in the equation is of \(\mathcal{O}(\epsilon)\). After substituting the expansion in (16.77) and collecting like powers of \(\epsilon\), we obtain for \(w_0\) the equation with boundary conditions

\[
\frac{\partial^3 w_0}{\partial z^2 \partial t} + \frac{\partial^2 w_0}{\partial z^2} = 0,
\]

(16.80)

\[
w_0(0, t) = 0, \quad \frac{\partial}{\partial z} w_0(1, t) + \frac{\partial^2}{\partial z \partial t} w_0(1, t) = -1.
\]

Since we are alarmed by the fact that the small parameter \(\epsilon\) multiplies the largest derivative in time, we have skipped for the moment the initial conditions. By neglecting this second time derivative under the limit of \(\epsilon \to 0\), we will not see the freedom or capacity any more to satisfy the initial conditions. The present approximate solution (the outer solution) is not uniformly valid, and the equation is of singularly perturbed type. A boundary layer is necessary to describe the local behaviour near \(t = 0\) and the effects of the initial condition.

After integrating (16.80) and applying the boundary conditions, we obtain for \(w_0\)

\[
w_0(z, t) = R_0(z) e^{-t} - z, \quad \text{where } R_0(0) = 0.
\]

(16.81)

In a similar way the \(\mathcal{O}(\epsilon)\) term is found as

\[
\frac{\partial^3 w_1}{\partial z^2 \partial t} + \frac{\partial^2 w_1}{\partial z^2} = \frac{\partial^2 w_0}{\partial t^2},
\]

(16.82)

\[
w_1(0, t) = 0, \quad \frac{\partial}{\partial z} w_1(1, t) + \frac{\partial^2}{\partial z \partial t} w_1(1, t) = 0.
\]
By integrating the equation and applying the boundary conditions, we find

$$w_1(z, t) = (S_0(z)t + R_1(z)) e^{-t},$$

where $S_0'' = R_0$, $S_0(0) = 0$, $S_0'(1) = 0$, $R_1(0) = 0$.

Note that $R_0(z)$ and $R_1(z)$ are still to be determined. In the quasi-static approach $R_0$ would be determined by the initial conditions, while all other $O(\epsilon)$ terms would be neglected. Although it is asymptotically inconsistent, we will see that this result is correct, at least for $R_0$. For the complete asymptotic problem applying the initial conditions is not possible. We could satisfy one condition, but not both. So we do need a boundary layer near $t = 0$ and we leave both conditions for the moment unresolved.

The boundary layer corresponds to a scaling $t = \delta(\epsilon) \tau$ and $w = \beta(\epsilon)/\Omega_1$ (where $\tau, \Omega_1 = O(1)$) such that the resulting equation and boundary conditions degenerate under $\epsilon \to 0$ to another significant limit than the one for the above outer expansion.

The second order time derivative has to play a role in applying two initial conditions. So the first order time derivative, which is $\sim \delta^{-1}$, has to balance $\epsilon$ times the second order time derivative, which is $\sim \epsilon \delta^{-2}$. Therefore it is clear that $\delta(\epsilon) = \epsilon$. Both the boundary condition at $z = 1$, containing the only inhomogeneous term $-1$, and the matchings condition for $\tau \to \infty, t \to 0$, requiring that $w_i = O(1)$, yield the choice of $\beta(\epsilon) = \epsilon$. So altogether we have the inner problem

$$t = \epsilon \tau, \quad w = \epsilon \Omega,$$

$$\frac{\partial^3 \Omega}{\partial z^2 \partial \tau} + \epsilon \frac{\partial^2 \Omega}{\partial z^2} = \frac{\partial^2 \Omega}{\partial \tau^2},$$

$$\Omega(z, 0) = 0, \quad \frac{\partial}{\partial \tau} \Omega(z, 0) = 0, \quad \Omega(0, \tau) = 0, \quad \epsilon \frac{\partial}{\partial z} \Omega(1, \tau) + \frac{\partial^2}{\partial z \partial \tau} \Omega(1, \tau) = -1.$$

As before, we assume the existence of an asymptotic expansion

$$\Omega(z, \tau; \epsilon) = \Omega_0(z, \tau) + \epsilon \Omega_1(z, \tau) + O(\epsilon^2),$$

which gives, after substituting and collecting like powers of $\epsilon$,

$$\frac{\partial^3 \Omega_0}{\partial z^2 \partial \tau} = \frac{\partial^2 \Omega_0}{\partial \tau^2} \quad \text{for } 0 < z < 1, \quad \tau > 0,$$

$$\Omega_0(z, 0) = 0, \quad \frac{\partial}{\partial \tau} \Omega_0(z, 0) = 0 \quad \text{for } \tau = 0,$$

$$\Omega_0(0, \tau) = 0, \quad \epsilon \frac{\partial}{\partial z} \Omega_0(1, \tau) + \frac{\partial^2}{\partial z \partial \tau} \Omega_0(1, \tau) = -1.$$

Integrating the equation gives

$$\frac{\partial^2 \Omega_0}{\partial z^2} = \frac{\partial}{\partial \tau} C_0(z).$$
The function $C_0$ vanishes identically, as can be deduced from the initial conditions. Since $\Omega_0 = 0$ at $\tau = 0$, we also have $\Omega_{0,zz} = 0$. From the defining equation it thus follows that $\Omega_{0,\tau} = \Omega_{0,zz} + C_0(z) = C_0(z) = 0$. In the same way the condition $\Omega_{0,\tau} = \Omega_{0,zz} = 0$ along $\tau = 0$ can be skipped because it is implied by $\Omega_0 = 0$.

The condition $\Omega_{0,zz} = -1$ along $z = 1$ can be integrated to $\Omega_{0,zz} = -\tau + a_0$ with an unknown constant $a_0$. In order to appreciate the meaning of this constant, we consider $\Omega_{0,zz}$ in $(z, \tau) = (1, 0)$. In the limit $\tau \downarrow 0$ along $z = 1$ we find $\Omega_{0,zz} = a_0$. On the other hand, in the limit $z \rightarrow 1$ along $\tau = 0$ we obtain $\Omega_{0,zz} = 0$ since $\Omega_0$ is identically zero. So a nonzero $a_0$ corresponds to $\frac{\partial}{\partial z} \Omega_0$ being discontinuous in $(z, \tau) = (1, 0)$. This is physically unacceptable because it would yield singular stresses (16.78), so apparently this constant is $a_0 = 0$. It is, however, not necessary to add conditions of regularity to our problem. This is already included in the problem and will become available to the inner solution via matching with the outer solution, as we will see below. So we will continue with an unknown $a_0$ and eventually show via matching that $a_0 = 0$.

We introduce the auxiliary variable $\chi(z, \tau)$ by

$$\Omega_0(z, \tau) = z \tau - \frac{1}{6} z^3 + \left( \frac{1}{2} + a_0 \right) z + \chi(z, \tau).$$

(16.88)

So the problem (16.86) reduces to

$$\frac{\partial^2}{\partial z^2} \chi = \frac{\partial}{\partial \tau} \chi$$

for $0 < z < 1$, $\tau > 0$,

(16.89)

$$\chi(z, 0) = \frac{1}{6} z^3 - \left( \frac{1}{2} + a_0 \right) z$$

for $\tau = 0$,

$$\chi(0, \tau) = 0, \quad \frac{\partial}{\partial z} \chi(1, \tau) = 0$$

for $\tau > 0$.

This has a general solution, without considering the initial conditions yet, of

$$\chi(z, \tau) = \sum_{n=0}^{\infty} k_n e^{-\lambda_n^2 \tau} \sin(\lambda_n z), \quad \lambda_n = \frac{1}{2} \pi + n \pi.$$

(16.90)

From Fourier series expansions for $z$ and $z^3$ we find at $\tau = 0$ that

$$\sum_{n=0}^{\infty} k_n \sin \lambda_n z = \frac{1}{6} z^3 - \left( \frac{1}{2} + a_0 \right) z = -2 \sum_{n=0}^{\infty} (-1)^n \left( \frac{1}{\lambda_n^2} + \frac{a_0}{\lambda_n^4} \right) \sin \lambda_n z,$$

(16.91)

which yields $k_n$ by inspection.

Having found the inner solution, we may determine the unknown constants and unknown functions of the outer solution. After expanding the outer solution, written in the inner variable $\tau = t/\varepsilon$, we have

$$w(z, \varepsilon \tau; \varepsilon) \simeq R_0(z)(1 - \varepsilon \tau) - z + \varepsilon R_1(z) + O(\varepsilon^2).$$

(16.92)

After expanding the inner solution, written in the outer variable $t = \varepsilon \tau$, asymptotically for small $\varepsilon$, we have

$$\varepsilon \Omega(z, t/\varepsilon; \varepsilon) \simeq -\varepsilon t - \frac{1}{6} \varepsilon z^3 + \left( \frac{1}{2} + a_0 \right) \varepsilon z + O(\varepsilon^2).$$

(16.93)
Equality of both expressions requires that

\[ R_0(z) = z, \quad R_1(z) = -S_0(z) = -\frac{1}{6}z^3 + \left(\frac{1}{2} + a_0\right)z. \] (16.94)

The constant \( a_0 \) vanishes because of the boundary condition \( S_0'(1) = a_0 = 0 \) of (16.83).

A composite expansion may be assembled as follows:

\[ w(z, t; \varepsilon) = (e^{-t} - 1)z + \varepsilon \left(\frac{1}{2}z - \frac{1}{6}z^3\right)(1 - t) + 2\varepsilon \sum_{n=0}^{\infty} \frac{(-1)^n}{\lambda_n^4} e^{-\lambda_n^2 t} \sin(\lambda_n z) + O(\varepsilon^2). \] (16.95)

The corresponding normal stresses are then

\[ t_z = p \left(-1 + 2 \sum_{n=0}^{\infty} \frac{(-1)^n}{\lambda_n} e^{-\lambda_n^2 t} \cos(\lambda_n z)\right) + O(\varepsilon), \] (16.96)

\[ t_r = p \left(-\alpha + \left(\alpha + \frac{1}{2}\right) e^{-t} - \sum_{n=0}^{\infty} \frac{(-1)^n}{\lambda_n} e^{-\lambda_n^2 t} \cos(\lambda_n z)\right) + O(\varepsilon). \] (16.97)

### 16.4.4 Example

An example of results is plotted in Figure 16.13 for \( \kappa = 1 \times 10^9 \) Pa, \( \eta = 1 \times 10^5 \) Pa s, \( \gamma = 1 \times 10^2 \) s\(^{-1} \), \( \rho = 1 \times 10^3 \) kg/m\(^3\), \( p = 2 \times 10^5 \) Pa, and \( L = 0.04 \) m. This yields \( \varepsilon = 0.0912 \) and \( \alpha = 0.9803 \).

The axial displacement \( u_z \) and radial and axial normal stresses \( t_r \) and \( t_z \) for various \( z \) positions are plotted against time. Along the right-hand side the time scale is used that corresponds to the viscoelastic phenomena. This is much longer than the initial time scale in the boundary layer, which is determined by inertia. Therefore the initial part of the plots, rescaled to this boundary layer time, is given along the left-hand side. Note that the boundary layer behaviour is primarily present in the derivative of \( u_z \) and is therefore difficult to see in \( u_z \) itself.

### 16.4.5 Discussion and Related Problems

- An interesting (and rather unexpected) result of Figure 16.13 is the sign of the normal radial stress \( t_r \), i.e., minus the pressure at the wall. For small time this is positive, which means that the material pulls at the wall. This may be possible at the lee side of a passing wave in reaction to the compression, but at the wave front it is unlikely. It is probably a failure of the viscoelastic model.

- If the material detaches from the wall, the assumption of a vanishing radial deformation is apparently not justified. A model that would include these deformations is, unfortunately, not one-dimensional any more and probably too complicated for analytical treatment. So we would sacrifice the relatively clear results that are obtainable now.
16.4.6 Exercises

16.12. Formulate a strategy to determine the problem parameters from the measurements. Is it possible to use only the final status (deformation and stresses), i.e., for \( t \to \infty \)?

16.13. Derive the Fourier series for \( z \) and \( z^3 \) used in (16.91).
16.5. Galloping Transmission Lines

16.14. Until we match the inner and outer solutions, we are not able to determine the unknown constant \( a_0 \) without invoking physical arguments. Physically interpret the solution of (16.86) with arbitrary \( a_0 \). Note the appearance of the delta function and its derivative along \( \tau = 0 \) by the representation

\[
\sum_{n=0}^{\infty} (-1)^n \sin \lambda_n z = \sum_{n=-\infty}^{\infty} \left\{ \delta(z + 1 + 4n) - \delta(z - 1 + 4n) \right\}.
\]

16.15. Why is the error in the asymptotic results (16.96) and (16.97) for the normal stresses only \( O(\varepsilon) \)?

16.16. Derive the found asymptotic solution directly from the exact solution due to A.A.F. van de Ven (found by Laplace transformation), given by the expression

\[
w(z, t; \varepsilon) = -z - 4\varepsilon \sum_{n=0}^{\infty} (-1)^n \sin(\lambda_n z) \left[ \frac{e^{-\lambda_n^2(1 - \sigma_n)t/2\varepsilon}}{\lambda_n^2(1 - \sigma_n)} + \frac{e^{-\lambda_n^2(1 + \sigma_n)t/2\varepsilon}}{\lambda_n^2(1 + \sigma_n)} - 4\varepsilon \right],
\]

where \( \sigma_n = (1 - 4\varepsilon \lambda_n^2)^{1/2} \). Note the different regions \( t = O(1) \) and \( t = O(\varepsilon) \).

16.5 Galloping Transmission Lines

16.5.1 Problem Formulation

Overhead transmission lines for transport of high-voltage electricity are cables of aluminum alloy suspended between high towers in the countryside [123]. One cable of several kilometers is carried by several towers in a row, and a section between two towers is called a span. As the cable is connected to the towers by a freely movable suspension string or isolator, the dynamical motions of neighbouring spans are coupled. In wintertime, the cables are vulnerable to large-scale vertical vibrations, induced by crosswind, when the cable is covered by snow or ice. This is an aeroelastic instability known as galloping. For high enough amplitudes neighbouring conductors may get close enough for the air-insulation to break down, causing a short circuit and structural damage to the cables.

It is known from observations that only a small wind force is sufficient to maintain a galloping vibration. The motion of the cable is therefore very close to a free vibration, which is what we will consider here.

16.5.2 The Model

Differential Equations and Boundary Conditions Consider a cable that is fixed at the outer ends and divided into \( N \) equal spans by \( N - 1 \) equal supports (Figure 16.14). \( N \) is not large (in the asymptotic sense to be explained below). The supports are inextensible suspension strings of length \( a \) and negligible weight, suspended from fixed pivots separated
by a distance $S$, the span size. The cable is linearly elastic, with negligible bending stiffness, of uniform undeformed cross section $A$, mass per unit length $m$, and Young's modulus $E$. It has a length per span $L$ when the cable is free of tension.

We parametrize the position along the cable by the variable $\ell \in [0, NL]$ such that this is just the arc length when the cable is unstretched. The (dimensional) time variable is $t$. We will here only consider cable motion in a vertical plane, which is provided with a Cartesian coordinate system oriented such that the gravity vector $-g e_y$ points in the negative $y$ direction (in reality the cable moves along a slightly tilted ellipse).

The cable position is given by the position vector $X(\ell, t) = (X(\ell, t), Y(\ell, t))$ with a corresponding tension vector $T(\ell, t) = T(\ell, t) (\cos \psi, \sin \psi)$, where $\psi$ is the angle between the cable tangent and the horizontal. The tension vector is tangent to the cable because of the assumed negligible bending stiffness.

Now consider a small cable element $d\ell$. Due to gravity, cable tension, and inertial forces, this element is stretched (but the mass remains the same). According to Hooke's law, a cable element is elongated in proportion to the tension (Figure 16.15), so

$$\sqrt{(dX^2 + dY^2)} = \left( 1 + \frac{T}{EA} \right) d\ell.$$ \hspace{1cm} (16.98)

According to Newton's law, the internal (tension) and external (gravity) forces are in equilibrium with the inertial forces, so

$$dT = (g e_y + \ddot{X}) m d\ell,$$ \hspace{1cm} (16.99)

where $\ddot{}$ denotes a second derivative with respect to time. When we introduce

$$\frac{\partial X}{\partial \ell} = \left( 1 + \frac{T}{EA} \right) \cos \psi, \quad \frac{\partial Y}{\partial \ell} = \left( 1 + \frac{T}{EA} \right) \sin \psi,$$ \hspace{1cm} (16.100)
16.5. Galloping Transmission Lines

Figure 16.16. Sketch of cables connected at a suspension string.

The equations that finally result in the limit \(d\ell \to 0\) are given by

\[
\frac{\partial}{\partial \ell} \left( \frac{T}{1+T/EA} \frac{\partial X}{\partial \ell} \right) = m \frac{\partial^2 X}{\partial t^2}, \quad (16.101a)
\]

\[
\frac{\partial}{\partial \ell} \left( \frac{T}{1+T/EA} \frac{\partial Y}{\partial \ell} \right) = m \frac{\partial^2 Y}{\partial t^2} + mg, \quad (16.101b)
\]

\[
\left( \frac{\partial X}{\partial \ell} \right)^2 + \left( \frac{\partial Y}{\partial \ell} \right)^2 = \left( 1 + \frac{T}{EA} \right)^2. \quad (16.101c)
\]

The boundary and coupling conditions are (see Figure 16.16) as follows:

(i) The supports are fixed at \(\ell = 0\) and \(\ell = NL\):

\[
X = 0, \quad Y = 0 \quad (\ell = 0), \quad (16.102a)
\]

\[
X = NS, \quad Y = 0 \quad (\ell = NL). \quad (16.102b)
\]

(ii) At the suspension strings \(\ell = nL\), where \(n = 1, \ldots, N - 1\), position vector \(X\) is continuous, the motion is restricted to a circle, while the force component normal to the string is continuous:

\[
\left[ X \right]_{\ell = nL^+} = 0, \quad (16.103a)
\]

\[
(X - nS)^2 + (Y - a)^2 = a^2, \quad (16.103b)
\]

\[
\left[ T\cos(\phi - \psi) \right]_{\ell = nL^+} = 0. \quad (16.103c)
\]

\(\phi\) denotes the angle of the suspension string with the vertical and \(\ell = nL\pm\) denotes the limits from either side of the support.

Small Parameters and the Mode of Vibration

The type of motion of interest allows further reduction of the model and is specified as follows:

- The ratio of sag \(D\) (the maximum stationary vertical displacement) and cable length \(L\) is small (typically on the order of \(1/30\)), so the slenderness

\[
\varepsilon := \frac{D}{L} \to 0 \quad (16.104)
\]

is a small parameter in the problem. This parameter will be used later to further reduce the problem.
• The total vertical nonstationary displacement is of the order of the sag, so
\[ \frac{Y}{L} = \mathcal{O}(\varepsilon). \] (16.105)

• The transversal wavelength \( \lambda_T \) is of the order of \( L \), so
\[ \frac{\lambda_T}{L} = \mathcal{O}(1). \] (16.106)
This is exactly what may be expected from the boundary conditions at the span ends: \( \frac{1}{2}, 1, 2, \) etc., waves per span.

• Apart from the slenderness, there is still another small parameter in the problem, which is rather more invisible, as it is related to the relative elasticity.

In general, it takes some time before a change of tension has been distributed as a longitudinal tensional wave through the cable. This time, however, is very short, since the propagation speed of these waves (the sound speed) is high and \( N \) is not large. In other words, the longitudinal wavelength \( \lambda_L \) is large compared to \( L \). A second small parameter, however, is usually not very convenient in an asymptotic analysis, so we use the estimate, which is consistent with practice, that
\[ \frac{L}{\lambda_L} = \mathcal{O}(\varepsilon). \] (16.107)

• The string length \( a \) is of the order of the sag, so
\[ \frac{a}{L} = \mathcal{O}(\varepsilon). \] (16.108)

• The relative amplitude \( \delta \) of the time-dependent perturbation will be taken small but bigger than the orders of \( \varepsilon \) neglected. We will assume a nearly harmonic vibration with a single dominating frequency \( \omega \) (which is to be found!). Since we are interested in the intrinsic nonlinear interaction between the harmonics, we will analyse the generated higher harmonics by a Lindstedt–Poincaré series in \( \delta \); see Section 15.3.2. To keep the results clear higher harmonics not generated by the first one will be excluded.

16.5.3 Asymptotic Analysis

Reduced Problem The basic small parameter \( \varepsilon \) will be used to reduce the above general problem to an asymptotic model. This model, assumed independent of \( \varepsilon \), will subsequently be analysed asymptotically for small relative amplitude \( \delta \). (So \( \varepsilon \) is not the small parameter that we will use later in the dynamical analysis.)

Since the longitudinal wave speed is \( c_L = (EA/m)^{1/2} \), while \( L/\lambda_L = \omega L/c_L = \mathcal{O}(\varepsilon) \), we can introduce the reference frequency \( \omega_{\text{ref}} = \varepsilon (EA/m)^{1/2}/L \). The dimensionless frequency and time variables are then given by
\[ \omega = \omega_{\text{ref}} \omega^*, \quad t = t^* / \omega_{\text{ref}}. \] (16.109)
Since \( \lambda_T/L = \mathcal{O}(1) \), the spatial coordinate \( \ell \) obviously scales on \( L \). (Note that \( L \) is an unknown of the problem!) Since \( Y/L = \mathcal{O}(\varepsilon) \), \( Y \) scales on \( \varepsilon L \). The transversal wave velocity is \( c_T = (T/m)^{1/2} \), so \( \lambda_T/L = c_T / \omega L = \mathcal{O}(1) \), yielding that the tension scales on \( T_{\text{ref}} = \varepsilon^2 EA \). Altogether we have for \( \ell \in \{ \text{the nth section} \} \) that
\[ \ell = (n - 1 + s)L, \quad Y(\ell, t) = \varepsilon LY^*(s, t^*; n), \quad T(\ell, t) = T_{\text{ref}} T^*(s, t^*; n). \] (16.110)
The suitable scaling for \( x \) is more subtle. Of course, the slender geometry suggests immediately that \( X \approx \ell \). However, we are only interested in the \( x \) displacement, which is very small. By substituting the above estimates in (16.101c), we have \( \frac{\partial}{\partial s} X = 1 + O(\varepsilon^2) \), and so for \( X \) from the \( n \)th section we have

\[
X(\ell, t) = (n - 1)S + Ls + \varepsilon^2 LX^*(s, t^*; n). \tag{16.111}
\]

Finally, we investigate the role of gravity. If we substitute the present estimates into (16.101b), we find the term \( mgL/E \varepsilon^3 \) next to terms of \( O(1) \). So this term has to be \( O(1) \) or smaller. Suppose it is smaller. Then the stationary solution would be to leading order \( Y \equiv 0 \), so \( D = 0 \), which is contradictory to our scaling assumptions. So the term is \( O(1) \) and we introduce

\[
\mu = \frac{mgL}{8EA\varepsilon^3} = O(1) \tag{16.112}
\]

(where the factor 8 is included for convenience later). While we omit the superscript asterisks \(*\) from here on, we obtain, under the approximation for small \( \varepsilon \), the reduced version of problem (16.101) as follows:

\[
\frac{\partial T}{\partial s} = 0, \quad \frac{\partial T}{\partial s} \left( T \frac{\partial Y}{\partial s} \right) = 8\mu + \frac{\partial^2 Y}{\partial t^2}, \quad \frac{\partial X}{\partial s} + \frac{1}{2} \left( \frac{\partial Y}{\partial s} \right)^2 = T. \tag{16.113a,b,c}
\]

In practice, the sag \( D \) and span size \( S \) are known, while the cable length \( L \) is to be determined. So \( \varepsilon \) and \( \mu \) are to be determined from the stationary solution. Moreover, all spans are assumed geometrically the same and so is the stationary solution. It is therefore useful to split up the solution into a stationary and a nonstationary part

\[
X(s, t; n) = X_0(s) + x(s, t; n), \quad Y(s, t; n) = Y_0(s) + y(s, t; n),
\]

\[
T(s, t; n) = T_0(s) + \tau(s, t; n), \tag{16.114}
\]

with boundary conditions for the stationary part of \( Y_0(0) = Y_0(1) = 0 \) and \( X_0(0) = 0 \). As we scaled \( Y \) on \( D \) and, from symmetry, anticipate the location of maximum deflection halfway, we have of course \( Y_0(\frac{L}{2}) = -1 \). The condition for \( X_0 \) at \( s = 1 \) is not really a boundary condition to the scaled problem but implicitly determines the unknown length \( L \) by the relation \( S = L(1 + \varepsilon^2 S_0) \) and (up to the \( \varepsilon \) approximation pursued so far) \( X_0(1) = S_0 \). The result is easily found (a parabola) and is given by

\[
T_0(s) = \mu, \quad Y_0(s) = -4(s - s^2), \quad X_0(s) = \mu s - \frac{4}{3} (1 + (2s - 1)^2), \quad S_0 = \mu - \frac{8}{3}. \tag{16.115a,b,c,d}
\]

If we substitute (16.115) into (16.113), we obtain our fundamental nonstationary problem

\[
\frac{\partial \tau}{\partial s} = 0, \quad (\mu + \tau) \frac{\partial^2 y}{\partial s^2} + 8\tau = \frac{\partial^2 y}{\partial t^2}, \quad \frac{\partial x}{\partial s} + 4(2s - 1) \frac{\partial y}{\partial s} + \frac{1}{2} \left( \frac{\partial y}{\partial s} \right)^2 = \tau. \tag{16.116a,b,c}
\]

The boundary conditions at the rigid supports follow readily:

\[
x(0, t; 1) = x(1, t; N) = y(0, t; 1) = y(1, t; N) = 0. \tag{16.117}
\]
For the conditions at the suspension string \((s = 0, n = 2, 3, \ldots, N, s = 1, n = 1, 2, \ldots, N - 1)\) we observe that \(a \phi / L = O(\varepsilon^2)\) as it is of the order of the \(x\) variation, so \(\phi = O(\varepsilon)\), and we obtain for small \(\varepsilon\) the conditions
\[
y(0, t; n) = y(1, t; n) = 0, \quad x(1, t; n) = x(0, t; n + 1), \quad \tau(1, t; n) = \tau(0, t; n + 1),
\]
with the important consequences that
\[
\sum_{n=1}^{N} \left[ x(1, t; n) - x(0, t; n) \right] = 0, \quad \tau = \text{constant in } s \text{ and independent of } n. \quad (16.119)
\]

**Weakly Harmonic Motion** The Lindstedt–Poincaré method (Section 15.3.2) involves the assumption of a periodic solution with fundamental (as yet unknown!) frequency \(\omega\). The amplitude \(\delta\) will be introduced as a small parameter, and based on this a solution will be constructed as a power series in \(\delta\), where \(\omega\) also depends on \(\delta\).

In order to apply the condition of periodicity, it is necessary to introduce the inherent time scale
\[
t' := \omega t. \quad (16.120)
\]
Now we assume \(y = O(\delta)\) and expand
\[
y = \delta y_1 + \delta^2 y_2 + \cdots. \quad (16.121)
\]
We can do likewise for \(x\) and \(\tau\). We consider \(y_1, y_2, \ldots\) normalized, e.g., as \(y_1(\frac{1}{2}) = y_2(\frac{1}{2}) = \cdots = 1\). We expand \(\omega\) as
\[
\omega = \omega_0 + \delta^2 \omega_2 + \cdots, \quad (16.122)
\]
where we used the result that \(\omega_1 = 0\). This could have been guessed from the intuitive argument that \(\omega\) should not depend on the sign of the vibrational amplitude.

We introduce the notation \(\mu' = \frac{\mu}{\mu'}(\cdot)\) and \(\gamma' = \frac{\gamma}{\gamma'}(\cdot)\). We substitute the \(\delta\) series into (16.116b,c), and collect like powers of \(\delta\):
\[
\begin{align*}
\mu y_1'' + 8\tau_1 &= \omega_0^2 y_1, & x_1' + 4(2s - 1)y_1' &= \tau_1, \quad (16.123) \\
\mu y_2'' + 8\tau_2 + \tau_1 y_1' &= \omega_0^2 y_2, & x_2' + 4(2s - 1)y_2' + \frac{1}{2}(y_1')^2 &= \tau_2. \quad (16.124)
\end{align*}
\]
with \(\tau_1, \tau_2, \ldots\) constant in \(s\). The interesting solutions here are based on the harmonic solution of the linearized problem (i.e., for \(y_1, x_1, \tau_1\)). Since any phase shift in time is unimportant, we may assume \(y_1, x_1, \tau_1\) to vary in time like \(\sin t'\). The next question is now what time dependence this induces in the second order. In (16.124) the terms \(\tau_1 y_1''\) and \((y_1')^2\) act as an inhomogeneous (source) term. Since they vary in time like \(\sin(t')^2 = \frac{1}{2} - \frac{1}{2} \cos(2t')\), we expect therefore a constant and a \(\cos(2t')\) contribution. So we put
\[
y_1(s, t'; n) = y_{11}(s; n) \sin t', \quad y_2(s, t'; n) = y_{20}(s; n) + y_{22}(s; n) \cos(2t'), \quad (16.125)
\]
etc. We have similar results for \(x\) and \(\tau\).
16.5. Galloping Transmission Lines

By substituting (16.125) into (16.123) and (16.124) and collecting the harmonics, we get

\[ \mu y''_{11} + 8\tau_{11} + \omega_0^2 y_{11} = 0, \]  
\[ x'_{11} + 4(2s - 1)y'_{11} = \tau_{11}, \]  
\[ \mu y''_{20} + 8\tau_{20} + \frac{1}{2}\tau_{11}y_{11}'' = 0, \]  
\[ x'_{20} + 4(2s - 1)y'_{20} + \frac{1}{4}(y'_{11})^2 = \tau_{20}, \]  
\[ \mu y''_{22} + 8\tau_{22} - \frac{1}{2}\tau_{11}y_{11}'' + 4\omega_0^2 y_{22} = 0, \]  
\[ x'_{22} + 4(2s - 1)y'_{22} - \frac{1}{4}(y'_{11})^2 = \tau_{22}. \]  

The boundary conditions are the same as for \( x, y, \) and \( \tau \) ((16.117), (16.118), and (16.119)), but now taken for each harmonic independently.

16.5.4 Solutions

In spite of the many assumptions and reductions made, the possible solutions are still numerous. This is exactly in agreement with reality, where the cables are known to vibrate in many possible modes. This indeterminacy is also a very important problem to be dealt with in any numerical simulation.

For illustration we will pick here a possible family of solutions. Others may be obtained in analogous ways.

A very important class of solutions is one in which the tension’s first harmonic \( \tau_{11} \) vanishes. This is rather unexpected (the visible parts \( x, y \) of the cable do vibrate in their first harmonic) and is typical of coupled spans (for a single span it only occurs for higher-order modes). We have then for the \( n \)th span \( (1 \leq n \leq N) \) the general solution for the \( j \)th mode

\[ \tau_{11} = 0, \]  
\[ y_{11} = A_n \sin ks, \quad \text{where} \quad k := \omega_0/\mu^{1/2} \quad \text{and} \quad k_j := (2j + 1)\pi, \]  
\[ x_{11} = B_n - 4(2s - 1)A_n \sin ks - 8A_n \cos ks/k, \]  
\[ \tau_{20} = 3 \frac{3^{3/2}\tau_0^2}{8\omega_0^2 A^2/\mu}(16 + 3\mu), \]  
\[ y_{20} = 4\tau_{20}(s - s^2)/\mu, \]  
\[ x_{20} = C_n + \tau_{20}s + \frac{8}{3}\tau_{20}(1 + (2s - 1)^3)/\mu - \frac{1}{8}k^2A_n^2(s + \sin(2ks)/2k), \]  
\[ \text{where} \quad C_n = (n - 1)\tau_{20}(1 + 16/3\mu) - \frac{1}{8}k^2(A_1^2 + \cdots + A_{n-1}^2) \]  
\[ \text{and} \quad A^2 = (A_1^2 + A_2^2 + \cdots + A_N^2)/N; \]
\[ \tau_{22} = \frac{1}{8} k^2 \omega_0^2 \frac{A_1}{(16 - \omega_0^2)}, \]
\[ y_{22} = 2 \tau_{22} (\cos(2ks) - 1)/\omega_0^2, \]
\[ x_{22} = D_n + \tau_{22} s - 8 \tau_{22} (1 + (2s - 1) \cos(2ks) - \sin(2ks)/k)/\omega_0^2 \]
\[ + \frac{1}{8} k^2 A_0^2 (s + \sin(2ks)/2k), \tag{16.131} \]
\[ \text{where } D_n = -(n - 1) \tau_{22} (16/\omega_0^2 - 1) + \frac{1}{8} k^2 (A_1^2 + \cdots + A_{n-1}^2). \]

From (16.119) it follows that the amplitudes are further restricted by the condition that
\[ A_1 + A_2 + \cdots + A_N = 0. \tag{16.132} \]

For a two-span configuration \((N = 2)\) with equal vibrational energy in each section this becomes \(A_1 = -A_2 = 1\), while \(A_2^2 = 1\).

The frequency is, to leading order, equal to \(\omega_0 = (2j + 1)\sqrt{\mu} \pi^{1/2}\). The higher-order shift \(\omega_2\) is not yet determined in the above results.

16.5.5 Examples

Two practically relevant examples are given in Figure 16.17. In both cases we have \(m = 1\) kg/m, \(g = 9.8\) m/s\(^2\), \(EA = 2.156 \times 10^7\) N, and \(S = 320\) m, while \(A = \overline{A}^2 = 1\) and the lowest mode \((j = 0)\) is selected. Shown are the dimensional varying parts of vertical displacement \(dY := \varepsilon Ly\) and tension \(dT := \varepsilon^2 EA \tau\). In the left column we have a stationary tension \(T = 12,600\) N and sag \(D = 10\) m with \(\delta = 0.5\). In this case \(L = 320.7\) m, \(\mu = 0.6011, \varepsilon = 0.0312,\) and the frequency is \(\omega = 2.436\) or \(0.175\) Hz. We observe a mild contribution of the higher harmonics in \(dY\). These are more present in the case displayed in the right column, where we have \(T = 16,500\) N and \(D = 7.61\) m with \(\delta = 0.3\). Now \(L = 320.3\) m, \(\mu = 1.354, \varepsilon = 0.0238,\) and the frequency is \(3.655\) or \(0.201\) Hz.

16.5.6 Discussion and Related Problems

- In this section we considered an example of asymptotic modeling. The modeling process starts with a global model derived from first principles. Then two small parameters are identified, geometrical slenderness and relative elasticity, that allow us to remove irrelevant elements. In the final model all parts contribute significantly. The solution is based on the assumption that the cable vibrates harmonically with weakly nonlinear perturbations.

- The above solution is a gravity mode: \(\tau_{11} = 0\). The elasticity does not have much of a role; an inelastic catenary would allow practically the same motion. This is not true for elastogravity modes, where \(\tau_{11} \neq 0\) and the elastic tension and the vertical motion have the same frequency. See [123].
16.5. Galloping Transmission Lines

- See the expression for $\tau_{22}$ in (16.131). An interesting feature in this solution is the singularity for $\omega_0 = 4$, apparently indicating a breakdown of the present approximation. There is a resonance between the first and second harmonics, where the second order is not small any more compared to the first order; see Figure 16.17.

- A more complete modeling would include the driving force of the wind and the effect of air friction. This is a small effect, but near resonance the resulting amplitudes may be very large.

**Figure 16.17.** An oscillating cable section (from top to bottom): the vertical displacement $dY$ in $sL$ for various time steps, the position at the middle ($s = \frac{1}{2}$) of the displacement in time, and the varying part of the tension $dT$ in time. All variables are dimensional.
16.5.7 Exercises

16.17. Find the exact solution to the full stationary problem given by (16.101).
16.18. Redo the above analysis for a single-span elastogravity mode of lowest order.
16.19. Explain why a solution with \( \tau_{11} \neq 0 \) is unlikely in a multispans geometry.
16.20. Propose various external driving forces that improve the model.
16.21. Formulate the related problem for motion of the cable in three dimensions.

16.6 Groundwater Flow and Rain

16.6.1 Problem Formulation

Consider a long strip of land between two parallel canals of different water levels. Above a certain level, the ground is porous like sand, allowing the water to flow very slowly from the canal with the higher level to the other one. Below the porous top layer, the ground is semipermeable like clay, and the water disappears through it at a rate proportional to the local hydrostatic pressure. From above there is a vertical influx of precipitation, i.e., rain.

We are interested in the groundwater level in a situation of steady state for moderate to small amounts of rain when one of the canal water levels is below the semipermeable layer.

16.6.2 The Model

We assume that the width of the strip is \( L \) and select a coordinate system such that the \( Z \) axis is parallel to the strip, the \( Y \) axis is along the vertical, and the \( X \) axis is orthogonal to both. \( X = 0 \) corresponds to the border of canal 0 and \( X = L \) to the border of canal 1; see Figure 16.18. We assume that the groundwater level in the \( Z \) direction is constant. The semipermeable layer is situated at \( Y = 0 \), the groundwater level is at \( Y = h(X) \), and the water levels of canals 0 and 1 are at \( Y = h_0 < 0 \) and \( Y = h_1 \geq 0 \), respectively. Above \( Y = h(x) \), the ground is not saturated with water. Here, any available water slowly creeps downward under the effect of gravity only until it merges with the saturated region \( 0 \leq Y \leq h(x) \). Water flows through the semipermeable layer at a rate proportional to the local hydrostatic pressure, which is in turn proportional to \( h \), the water height. So the vertical mass flux density of water at \( Y = 0 \) may be given by

\[
\Phi_0 = \alpha h, \quad (16.133)
\]

where \( \alpha \) is a constant. From the top surface of the strip water comes in due to rainfall. Although, of course, it does not rain continuously, the time scales of the rainfall and the water flow in the unsaturated region \( Y > h \) are so much apart that we consider the average vertical mass flux density from the top of

\[
\Phi_1 = N, \quad (16.134)
\]
to be a constant in the problem. Thus we neglect variations in \( N \) due to different top surface coverage (green, buildings, roads).

Consider a slice of ground between \( X \) and \( X + dX \). The flux into the slice from the top is equal to \( \Phi_1 dX \) and the flux of mass leaving the slice below is \( \Phi_0 dX \). Between two neighbouring positions \( X \) and \( X + dX \) we have a small difference in groundwater level and therefore a small pressure gradient. According to Darcy's law, this produces a slow flow in the \( X \) direction with velocity \( v \) proportional (and opposite) to the pressure gradient and dependent on the porosity of the ground. In the present model we assume that the porosity is uniformly the same and that the water and the ground are incompressible. Since the pressure difference is the same for all \( Y \in [0, h(X)] \), the flow velocity is uniform in \( Y \) and we have

\[
  h(X + dX) - h(X) \sim p(X + dX) - p(X) \sim -v(X) dX. \tag{16.135}
\]

So the horizontal mass flux density \( \Phi_h \), which is proportional to \( v \), is

\[
  \Phi_h = -D \frac{dh}{dX}, \tag{16.136}
\]

where \( D \) is a positive material constant. The flux balance across all sides of a slice \( dX \) is then given by

\[
  \left[ h \Phi_h \right]_{X}^{X+dX} = (\Phi_1 - \Phi_0) dX.
\]

After taking the limit \( dX \to 0 \) in combination with relations (16.133), (16.134), and (16.136), we obtain

\[
  \frac{d}{dX} \left( D \frac{dh}{dX} \right) = \alpha h - N. \tag{16.137}
\]
The boundary condition at \( X = L \) is evidently \( h = h_1 \). From the presence of the semipermeable layer the groundwater level cannot become negative, so the boundary condition at \( X = 0 \) is just \( h = 0 \).

### 16.6.3 Analysis

As \( D \) is constant, this parameter can be taken out of the derivative in (16.137). We make the problem dimensionless with \( L, h_1, \) and \( \alpha \)—\( X = Lx, h(X) = h_1 \phi(x), N = \alpha h_1 K \)—and introduce the positive dimensionless parameter

\[
\varepsilon := \frac{D h_1}{2 \alpha L^2} \tag{16.138}
\]

The equation and boundary conditions are

\[
\varepsilon \frac{d^2}{dx^2} \phi - \phi + K = 0, \quad 0 \leq x \leq 1, \tag{16.139a}
\]

\[
\phi(0) = 0, \quad \phi(1) = 1. \tag{16.139b}
\]

When the strip width \( L \) is not very small compared to \( h_1 \) and the porosity of the semipermeable layer is much smaller than the porosity of the ground such that the vertical flux \( \sim Dh_1^2 / L \) is much bigger than the horizontal flux \( \sim \alpha h_1 L \), then \( \varepsilon \) is small. We will exploit this smallness and analyse the problem asymptotically under the limit \( \varepsilon \to 0 \). As the small parameter multiplies the highest derivative, the problem will appear to be a singular perturbation problem of boundary layer type (Section 15.4.1).

We will distinguish two cases: (i) moderate amounts of rain such that \( K = O(1) \) and (ii) small amounts of rain such that \( K = O(\varepsilon) \).

#### Moderate Amounts of Rain

Assume that \( \phi(x; \varepsilon) = \phi_0(x) + \mu_1(\varepsilon) \phi_1(x) + \cdots \). Then

\[
\phi_0(x) = K, \quad \phi_1(x) \equiv 0 \quad \text{for every } \mu_1(\varepsilon). \tag{16.140}
\]

As far as we can write \( \phi(x; \varepsilon) - K \) as a regular asymptotic approximation, it is equivalent to zero. Since \( \phi_0 \) does not satisfy the boundary conditions at \( x = 0 \) and \( x = 1 \), we expect boundary layers there, while \( \phi_0 \) is an outer solution. We start with \( x = 0 \). Suppose \( x = \delta \xi \) and \( \phi(x; \varepsilon) = \psi(\xi; \varepsilon) \). Since \( \phi_0 = O(1) \), we expect \( \psi = O(1) \). We find a significant degeneration \( \delta = \sqrt{\varepsilon} \).

We substitute \( \psi(\xi; \varepsilon) = \psi_0(\xi) + \lambda_1(\varepsilon) \psi_1(\xi) + \cdots \) and find to leading order that

\[
\frac{d^2}{d\xi^2} \psi_0^2 - \psi_0 + K = 0. \tag{16.141}
\]

We multiply by \( \frac{d}{d\xi} (\psi_0^2) \) and integrate to

\[
\frac{1}{2} \left( \frac{d}{d\xi} (\psi_0^2) \right)^2 - \frac{2}{3} \psi_0^3 + K \psi_0^2 = C_0. \tag{16.142}
\]
16.6. Groundwater Flow and Rain

Figure 16.19. Plot of boundary layer profile $\psi_0(\xi)$ for $K = O(1)$.

$C_0$ follows from the matching condition $\psi_0 \rightarrow K$ such that $(\psi_0')' = 0$, and hence

$$C_0 = \frac{1}{3} K^3,$$

$$\frac{d}{d\xi} \psi_0^2 = \left( \frac{2}{3} K^3 - 2K \psi_0^2 + \frac{4}{3} \psi_0^3 \right)^{1/2} = \frac{2}{\sqrt{3}} (K - \psi_0) \left( \frac{1}{2} K + \psi_0 \right)^{1/2}. \tag{16.143}$$

A positive sign prevails because $\psi_0$ increases from 0 to $K$. Note that with this choice of $C_0$ the following integral has a nonintegrable singularity at $\eta = 1$, and so $\xi \rightarrow \infty$ when $\psi_0/K \rightarrow 1$. We integrate

$$\int_0^{\psi_0/K} \eta \frac{d\eta}{(1-\eta)(\frac{1}{2} + \eta)^{1/2}} = \xi \tag{16.144}$$

to get the implicit solution (see Figure 16.19)

$$\left[ -2\sqrt{\frac{1}{2} + \eta} + \sqrt{\frac{1}{2} + \eta} \ln \frac{\sqrt{\frac{1}{2} + \eta} + \sqrt{\frac{1}{2} - \eta}}{\sqrt{\frac{1}{2} - \eta}} \right]_{\psi_0/K}^0 = \frac{\xi}{\sqrt{3K}}. \tag{16.145}$$

The boundary layer near $x = 1$ is analogous. Assume $x = 1 + \sqrt{\varepsilon} \zeta$, $\phi(x; \varepsilon) = \chi(\zeta; \varepsilon)$, and $\chi(\zeta; \varepsilon) = \chi_0(\zeta) + \cdots$. Then

$$\frac{d}{d\zeta} \chi_0^2 = \pm \frac{2}{\sqrt{3}} (K - \chi_0) \left( \frac{1}{2} K + \chi_0 \right)^{1/2}. \tag{16.146}$$

with a sign given by the sign of $1 - K$, the difference between the boundary value at $x = 1$ and the outer solution.

**Small Amounts of Rain** We write explicitly $K = \varepsilon k$, where $k$ is independent of $\varepsilon$. Suppose $\phi(x; \varepsilon) = \varepsilon \phi_0(x) + \mu_1(\varepsilon) \phi_1(x) + \cdots$. Then

$$\phi_0(x) = k, \quad \phi_1(x) = 0 \quad \text{for every} \quad \mu_1(\varepsilon) = o(\varepsilon). \tag{16.147}$$
We search for significant degenerations by writing $x = a + \delta(\varepsilon)\xi$, $\phi(x; \varepsilon) = \lambda(\varepsilon)\psi(\xi; \varepsilon)$ (where $a$ is the as yet unknown location of the boundary layer). Then we have

$$\varepsilon \lambda^2 \delta^{-2} \frac{d^2}{d\xi^2} \psi^2 - \lambda \psi + \varepsilon k = 0$$

(16.148)

such that from the outer solution ($\varepsilon k$) and the boundary condition ($\phi(1) = 1$) we can expect two types of boundary layers:

(a) $\lambda = \varepsilon$, $\delta = \varepsilon$

(b) $\lambda = 1$, $\delta = \sqrt{\varepsilon}$

Near $x = 0$ there are no problems. Suppose $x = \varepsilon \xi$ and $\phi(x; \varepsilon) = \varepsilon \psi_0(x) + \cdots$. Then

$$\frac{d^2}{d\xi^2} \psi_0^2 - \psi_0 + k = 0, \quad \psi_0(0) = 0, \quad \psi_0 \to k \quad (\xi \to \infty),$$

(16.149)

with solution given by

$$\frac{1}{2} \sqrt{3} \ln \left( \frac{\sqrt{3} - \sqrt{1 + 2\psi_0/k} \sqrt{3} + 1}{\sqrt{3} + \sqrt{1 + 2\psi_0/k} \sqrt{3} - 1} \right) + \sqrt{1 + 2\psi_0/k} - 1 = -\frac{\xi}{\sqrt{6k}}.$$ 

(16.150)

The boundary layer near $x = 1$ is less obvious. In view of the boundary condition, we start with the assumption that $\phi = O(1)$:

$$x = 1 + \sqrt{\varepsilon} \xi, \quad \phi(x; \varepsilon) = \chi_0(\xi) + \cdots \quad (-\infty < \xi \leq 0),$$

(16.151)

with

$$\frac{d^2}{d\xi^2} \chi_0^2 - \chi_0 = 0, \quad \chi_0(0) = 1.$$ 

(16.152)

The solution, up to a constant $C$ given by

$$\int_{\chi_0}^{1} \frac{\eta \, d\eta}{\sqrt{\frac{1}{2} \eta^3 + C}} = -\xi,$$

(16.153)

does not satisfy the matching condition $\chi_0 \to 0$ for $\xi \to -\infty$. Define

$$\xi_0(C) = -\int_{0}^{1} \frac{\eta \, d\eta}{\sqrt{\frac{1}{2} \eta^3 + C}}.$$ 

(16.154)

Then $\xi_0$ is finite for every $C$, and $\chi_0$ is apparently only defined (with these boundary conditions) along the interval $\xi_0 \leq \xi \leq 0$. So there is something wrong.

We consider several possibilities: there is no solution and therefore no approximation (this is, however, not the case); the $O(\sqrt{\varepsilon})$ boundary layer is inside another, wider, boundary layer (this is not the case; the only other possibility is $O(\varepsilon)$, hence smaller); our assumption...
of an outer solution of $O(\varepsilon)$ is not right (this is also not the case; the other possibility of $O(\varepsilon^{-1})$ leads to nothing); we hit a rare example of inner and outer solutions without region of overlap (very hypothetical but not totally impossible).

The solution, however, appears to be an inner boundary layer of $O(\varepsilon)$ around $\zeta = \zeta_0$ where $\phi = O(\varepsilon)$. This is rather unusual: the outer and inner solutions are connected via a small “interface” interior layer.

The behaviour of $\chi_0$ near $\zeta = \zeta_0$ is given by

$$ C \neq 0: \quad \chi_0 \approx (2\sqrt{C}(\zeta - \zeta_0))^{1/2}, \quad (16.155a) $$

$$ C = 0: \quad \chi_0 = (1 - \zeta/\zeta_0)^2 \quad \text{with} \quad \zeta_0 = -2\sqrt{3} \quad (16.155b) $$

(for $C = 0$, $\chi_0$ is exact). The value of $C$ will be determined from matching. Assume $x = 1 + \sqrt{\varepsilon}\zeta_0 + \varepsilon z$ and $\phi(x; \varepsilon) = f(z; \varepsilon)$. Expand

$$ f(z; \varepsilon) = \varepsilon f_0(z) + \cdots. \quad (16.156) $$

Then

$$ \frac{d^2}{dz^2} f_0^2 - f_0 + k = 0 \quad (16.157) $$

with matching conditions in both directions:

$$ f_0 \approx \phi_0 = k \quad (z \to -\infty) \quad (16.158a) $$

$$ \varepsilon f_0 \approx \chi_0 \quad (z \to \infty). \quad (16.158b) $$

The solution that approaches the constant $k$ in an exponential way for $z \to -\infty$ is

$$ \frac{1}{3} \sqrt{3} \ln \left( \frac{\sqrt{1 + \frac{2f_0}{k} - \sqrt{3}}}{\sqrt{1 + \frac{2f_0}{k} + \sqrt{3}}} \right) + \sqrt{1 + \frac{2f_0}{k} - A} = \frac{z}{\sqrt{6k}}. \quad (16.159) $$

If $z \to \infty$, then $f_0 \approx \frac{1}{2}k(A + z/\sqrt{6k})^2$, which is to be compared with (16.155). A successful matching is apparently obtained with $C = 0$ and $A = 0$.

We summarize the behaviour for small amounts of rain: via a boundary layer of $O(\varepsilon)$ near $x = 0$ the solution becomes a constant outer solution that continues via an interior boundary layer of $O(\sqrt{\varepsilon})$ into the parabolic behaviour of the boundary layer of $O(\sqrt{\varepsilon})$ near $x = 1$. This boundary layer has the particular property that it is defined on a finite interval.

### 16.6.4 Discussion and Related Problems

- In order to allow an asymptotic analysis, the proposed model is kept elementary. It is, however, not too difficult to introduce some straightforward generalisations. For example, $D$ and $N$ may be functions of $X$ without any serious change of the analysis.

- Furthermore, the current one-dimensional problem in $X$ may be reformulated to two dimensions in $(X, Z)$ by taking the mass fluxes across small rods $[dX \times dZ \times h(X, Z)]$ rather than slices. Wherever the boundaries are smooth, the boundary layers retain their one-dimensional character, and again the analysis will not be essentially different.
• The present model is rather ad hoc (an example of a constructing model, Section 7.2). For physically more refined models, books on groundwater and porous media flow should be consulted (e.g., [7]). Note that the present model is essentially one of slow variation in $X$ and might well be derived systematically from the fundamental equations of porous media flow by the method of slow variation (Section 15.3.1).

• One of the interesting features that appeared in the analysis is the inner layer for small $K$, separating the asymptotically constant outer solution from a boundary layer where diffusion from the canal dominates. A very similar transition layer may be found in the catalyst pellet problem of Section 16.8.

• An interesting detail in the matching problem of (16.143) can be generalised as follows. Consider a boundary layer equation in $Y(\xi), 0 \leq \xi < \infty$, of the form

\[
\frac{\partial^2}{\partial \xi^2} Y + F'(Y) = 0,
\]

which may be integrated to

\[
\frac{1}{2} \left( \frac{\partial}{\partial \xi} Y \right)^2 + F(Y) = E.
\]

If $Y$ should be matched for $\xi \to \infty$ to an outer solution $y(x)$ of $O(1)$, then the integration constant $E$ may be found by observing that $\frac{\partial}{\partial \xi} Y$ should vanish and $E = F(y(0))$. An important condition for consistency is that the final integral

\[
\int_{Y(0)}^{Y} \frac{d\eta}{\sqrt{E - F(\eta)}} = \pm \sqrt{2} \xi
\]

diverge at $\eta = y(0)$ in order to have $\xi \to \infty$. This did not occur, for example, in (16.152) and (16.153). The unsuccessful matching was indeed the warning sign for the more complex structure with an internal transition layer.

16.6.5 Exercises


16.24. Reformulate the asymptotic analysis for $D = D(X)$ and $N = N(X)$.

16.25. Reformulate the problem generalised to two surface dimensions $X$ and $Z$. Assume that $D = D(X, Z)$ and $N = N(X, Z)$. Try to solve the problem.

16.26. The boundary value problem (16.139) can also be solved numerically.

(a) Give the central difference scheme for (16.139a). Taking into account the boundary condition (16.139b), write the difference equations as a nonlinear algebraic system $f(\phi) = 0$.

(b) Formulate Newton’s method for the system in part (a).

(c) Compute a numerical solution of boundary value problem (16.139) for $\epsilon = 10^{-1}$ and $K = 1$, for $\epsilon = 10^{-3}$ and $K = 1$, and for $\epsilon = 10^{-3}$ and $K = 10^{-3}$. Compare with the asymptotic approximation.
(d) An alternative approach is to reformulate the problem in terms of the variable \( u = \sqrt{\varphi} \). Do the same as above. What are the (numerical) advantages and (possible) disadvantages of this approach?

### 16.7 Cooling a Monocrystalline Bar

#### 16.7.1 Problem Formulation

A common way to grow large silicon monocrystals is to pull the crystal very slowly from a melt [73]. The temperatures at which these processes take place are usually rather high. Silicon, for example, solidifies at 1693 K. This is why the crystal will lose most of its heat by radiation. At the same time, the radiation from the surface is small compared to the conduction in the material. In order to investigate the typical temperature variation along the crystalline bar, we consider the following model, originally proposed by [74].

#### 16.7.2 The Model

Consider a long cylindrical bar \( z \geq 0, \ 0 \leq r \leq a, \ 0 \leq \phi < 2\pi \), where \( (r, \phi, z) \) are cylindrical coordinates. The bar is assumed to be long enough for a semi-infinite approximation. How long is sufficient is not clear yet, but this will follow from the analysis below.

The bar is heated from one end at \( z = 0 \), where the melting front is situated, while it is cooled by radiation from the other side \( r = a \); see Figure 16.20. At these high temperatures cooling by natural convection of the surrounding air is very small compared to radiation. The melting front will be considered planar, although in reality it is slightly curved. The monocrystal is slowly pulled away with constant velocity \( V \) in the positive \( z \) direction. The heat flow in and out of the bar is in equilibrium, so we have a steady state. Inside the bar \( z > 0, \ r < a \), we have the stationary convective heat equation for temperature \( T \)

\[
\rho C_p V \frac{\partial T}{\partial z} = \nabla \cdot (\kappa(T)\nabla T), \tag{16.160}
\]

where \( \rho \) denotes the crystal density and \( C_p \) is the specific heat, which are nearly constant for the temperatures considered. \( \kappa \) is, in general, the temperature-dependent thermal conductivity (typically inversely proportional to the temperature [73]). To make progress we

![Figure 16.20. Sketch of a silicon bar.](image-url)
assume $\kappa$ constant [74] and equal to the value at or near the melting temperature. To complete the problem formulation we have the following boundary conditions to quantify the heating and radiation and the symmetry condition at the axis:

$$T = T_m \quad \text{at} \quad z = 0, \quad (16.161a)$$

$$-\kappa \frac{\partial T}{\partial r} = \varepsilon \sigma (T^4 - T_0^4) \quad \text{at} \quad r = a, \quad (16.161b)$$

$$T \to T_0 \quad \text{for} \quad z \to \infty, \quad (16.161c)$$

$$\frac{\partial T}{\partial r} = 0 \quad \text{at} \quad r = 0. \quad (16.161d)$$

The temperature $T_0$ of the environment represents the heating by radiation from the surroundings back into the crystal. This is obviously small in our region of interest as $T_0$ is much lower than the melting temperature $T_m$. So we will ignore $T_0$, although it will imply that for large $z$ the model will become inaccurate. The universal constant $\sigma$ in the law of radiation is the Stefan–Boltzmann constant. The material constant $\varepsilon$ is called the emission coefficient or emissivity. It is equal to unity for black bodies and between zero and one otherwise. By symmetry, the problem is independent of $\phi$.

We make the problem dimensionless using

$$r = ar^*, \quad z = az^*, \quad T = T_m T^*, \quad (16.162)$$

and we introduce the dimensionless parameters

$$\varepsilon = \frac{\varepsilon \sigma T_m^3 a}{\kappa}, \quad Pe = \frac{\rho C_p V a}{\kappa}. \quad (16.163)$$

The relative importance between heat transfer by radiation and that by conduction is quantified by $\varepsilon$ and may be identified with a Biot number. The relative importance of convection and conduction is given by the Péclet number $Pe$. Using the typical parameter values given in Table 16.3, we find

$$\varepsilon = 9.67 \times 10^{-3}, \quad Pe = 1.91 \times 10^{-2}.$$ 

We see that both are small, but it is not yet clear if they are small enough to be neglected. It does not seem to be a good idea to neglect $\varepsilon$, because without it we would have no

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>$2.33 \times 10^3$</td>
<td>kg/m$^3$</td>
<td>$T_0$</td>
<td>$3.0 \times 10^2$</td>
<td>K</td>
</tr>
<tr>
<td>$C_p$</td>
<td>$9.22 \times 10^2$</td>
<td>J/(kg·K)</td>
<td>$a$</td>
<td>$1.15 \times 10^{-2}$</td>
<td>m</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>$2.16 \times 10^4$</td>
<td>J/(m·s·K)</td>
<td>$V$</td>
<td>$1.67 \times 10^{-5}$</td>
<td>m/s</td>
</tr>
<tr>
<td>$T_m$</td>
<td>$1.693 \times 10^3$</td>
<td>K</td>
<td>$\sigma$</td>
<td>$5.6704 \times 10^{-8}$</td>
<td>J/(s·m²·K^4)</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
cooling at all. At the same time, $Pe$ is of the same order of magnitude as $\varepsilon$, so it would seem at least consistent to include $Pe$ if we include $\varepsilon$. So we will consider the problem asymptotically for small $\varepsilon$ and $Pe$, but not as two independent parameters. It is almost always inadvisable to retain several independent small parameters. This creates undesirable and useless complexities in the asymptotic analysis. Therefore we introduce

$$Pe = \varepsilon c,$$  \hspace{1cm} (16.164)

which is based on the observation that $Pe \simeq \varepsilon$ for the chosen characteristic parameter values. The velocity $V$ we adopted corresponds to 1 mm/min. In [73] higher values (up to 5 mm/min) are considered, leading to $Pe = O(\varepsilon^{1/2})$. We will not consider this case because it leads to a more difficult problem. Ignoring the asterisks $\ast$, we have in dimensionless form the problem

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T}{\partial r} \right) + \frac{\partial^2 T}{\partial z^2} = \varepsilon c \frac{\partial T}{\partial z}, \hspace{1cm} \text{for } 0 < r < 1, \quad 0 < z < \infty$$  \hspace{1cm} (16.165a)

$$\frac{\partial T}{\partial r} = -\varepsilon T^4 \hspace{1cm} \text{at } r = 1,$$  \hspace{1cm} (16.165b)

$$T = 1 \hspace{1cm} \text{at } z = 0,$$  \hspace{1cm} (16.165c)

$$T \rightarrow 0 \hspace{1cm} \text{for } z \rightarrow \infty,$$  \hspace{1cm} (16.165d)

$$\frac{\partial T}{\partial r} = 0 \hspace{1cm} \text{at } r = 0.$$  \hspace{1cm} (16.165e)

In the following this problem will be analysed asymptotically for $\varepsilon \rightarrow 0$.

### 16.7.3 Asymptotic Analysis

At the outset we can anticipate a region near $z = 0$, where the boundary condition at $z = 0$ dominates the solution. Away from this initial zone the temperature profile is evened out radially and sideways radiation is balanced by axial conduction. In this region the temperature variation scales on another, longer, length scale, say $\delta(\varepsilon)^{-1}$, that is as yet unknown. To write the problem as a singular perturbation problem it is convenient to determine this length scale and rescale the axial variable. Hence we introduce the variables

$$Z = \delta(\varepsilon) z, \quad T(r, z) = \mathcal{T}(r, Z).$$  \hspace{1cm} (16.166)

In $(r, Z)$ coordinates we obtained the following singular perturbation problem (see Chapter 15), with $\delta$ to be determined:

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial \mathcal{T}}{\partial r} \right) + \delta^2 \frac{\partial^2 \mathcal{T}}{\partial Z^2} = \varepsilon \delta c \frac{\partial \mathcal{T}}{\partial Z}, \hspace{1cm} \text{for } 0 < Z < \infty,$$  \hspace{1cm} (16.167a)

$$\frac{\partial \mathcal{T}}{\partial r} = -\varepsilon \mathcal{T}^4 \hspace{1cm} \text{at } r = 1,$$  \hspace{1cm} (16.167b)

$$\mathcal{T} = 1 \hspace{1cm} \text{at } Z = 0,$$  \hspace{1cm} (16.167c)

$$\mathcal{T} \rightarrow 0 \hspace{1cm} \text{for } Z \rightarrow \infty,$$  \hspace{1cm} (16.167d)

$$\frac{\partial \mathcal{T}}{\partial r} = 0 \hspace{1cm} \text{at } r = 0.$$  \hspace{1cm} (16.167e)
We start by assuming an asymptotic expansion of $T$. In terms of variable $Z$ this is an outer expansion. Apart from the first one (which has to be $O(1)$ because the bar’s temperature decays only slowly from one at the end), we do not know the order functions in the expansion yet. However, from the equation and boundary conditions it is clear that $\delta^2$ and $\varepsilon$ will appear somewhere. So we start with the crude estimate

$$T(r, Z; \varepsilon) = T_0(r, Z) + O(\delta^2, \varepsilon\delta).$$  \hfill (16.168)

This yields, after substitution in (16.167a), to leading order,

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} T_0 \right) = 0,$$  \hfill (16.169)

with solution $T_0(r, Z) = Q_0(Z)$. From the radiation boundary condition (16.167b) we find

$$\frac{\partial}{\partial r} \left[ Q_0 + O(\delta^2, \varepsilon\delta) \right]_{r=1} = O(\delta^2, \varepsilon\delta) = O(\varepsilon),$$  \hfill (16.170)

which can only be true if

$$\delta = \varepsilon^{1/2}.$$  \hfill (16.171)

This means that the convection term in (16.167a) is asymptotically smaller than $\delta^2$. We return to the asymptotic expansion and write

$$T(r, Z; \varepsilon) = T_0(r, Z) + \varepsilon T_1(r, Z) + \cdots.$$  \hfill (16.172)

We have found $T_0$ already. For $T_1$ we have

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} T_1 \right) = -Q_0''(Z),$$  \hfill (16.173)

with solution

$$T_1(r, Z) = Q_1(Z) - \frac{1}{4} r^2 Q_0''(Z).$$  \hfill (16.174)

Since this $r$-dependent term is the first to contribute to the heat flux at $r = 1$ due to radiation, we arrive at the following equation for $Q_0$:

$$\frac{\partial}{\partial r} \left[ Q_0 + \varepsilon Q_1 - \frac{1}{4} \varepsilon r^2 Q_0'' \right]_{r=1} = -\frac{1}{2} \varepsilon Q_0'' = -\varepsilon Q_0^4.$$  \hfill (16.175)

This equation can be integrated once to

$$\left( Q_0' \right)^2 = k_1 + \frac{4}{5} Q_0^5.$$  \hfill (16.176)

From the far-field condition (16.167d) it follows that $k_1 = 0$. As $T$ has to decay from one to zero, we know that $Q_0' < 0$, and we can integrate another time to obtain the solution

$$T_0(r, Z) = Q_0(Z) = \left( \frac{\left( \frac{3}{4} \sqrt{3} \right)^{2/3}}{Z + k_0} \right)^{2/3},$$  \hfill (16.176)
where \( k_0 \) is still to be determined. Near \( z = 0 \), \( T \) scales on both \( r \) and \( z \), so we assume an asymptotic expansion of \( T(r, z; \varepsilon) \). In terms of the “compressed” variable \( Z \), \( z \) is a stretched variable, and this expansion is an inner expansion. Had we used a more general entrance temperature profile than (16.167c), we would have immediately recognized the need for a boundary layer from the inability of the outer solution to satisfy this condition (see the exercises below).

From the radiation condition it is clear that, in the sequence of order functions, \( \varepsilon \) will show up somewhere. This appears, however, not to be enough. We will also have to consider the condition of matching with the outer solution \( T_0 \). If \( Z = \mathcal{O}(\delta) \), we have

\[
T_0 \sim Q_0(0) + \delta z Q_0'(0) + \cdots
\]  

(16.177)

so before the \( \mathcal{O}(\varepsilon) \) term there is an \( \mathcal{O}(\varepsilon^{1/2}) \) term. We write, therefore,

\[
T(r, z; \varepsilon) = \vartheta_0(r, z) + \varepsilon^{1/2} \vartheta_1(r, z) + \varepsilon \vartheta_2(r, z) + \cdots
\]  

(16.178)

Upon substitution of this expansion in the equation and boundary conditions, it follows that for \( n = 0 \) and \( n = 1 \) we have

\[
-\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \vartheta_n \right) + \frac{\partial^2}{\partial z^2} \vartheta_n = 0, \quad \frac{\partial \vartheta_n}{\partial r} = 0 \quad \text{at} \quad r = 1, \quad \frac{\partial \vartheta_n}{\partial r} = 0 \quad \text{at} \quad r = 0,
\]  

(16.179)

where \( \vartheta_0 = 1 \) and \( \vartheta_1 = 0 \) at \( z = 0 \). This problem has a general solution given by the exponential Bessel function series

\[
\vartheta_n(r, z) = C_n z + D_n + \sum_{k=1}^{\infty} J_0(\alpha_k r) \left( A_{nk} e^{-\alpha_k z} + B_{nk} e^{\alpha_k z} \right),
\]  

(16.180)

where \( J_0 \) is the ordinary Bessel function of the first kind of order zero; \( \alpha_k \) is the \( k \)th positive zero of \( J'_0 = -J_1 \); and \( A_{nk}, B_{nk}, C_n, \) and \( D_n \) are constants to be determined from the boundary conditions.

From the matching condition for \( z \rightarrow \infty \) to the outer expansion of \( T \) it follows that \( B_k^{(\text{outer})} = o(e^{-z}) \) for \( k \rightarrow \infty \) and \( z \rightarrow \infty \). This is only possible if \( B_{nk} = 0 \). From the orthogonality of the set \( \{J_0(\alpha_k r)\} \) on \([0, 1]\), applied to the boundary condition at \( z = 0 \), it is readily found that we have

\[
D_0 = 1, \quad A_{0k} = \frac{\int_0^1 r J_0(\alpha_k r) \, dr}{\frac{1}{2} J_0(\alpha_k)^2} = \frac{\alpha_k^{-1} J_1(\alpha_k)}{\frac{1}{2} J_0(\alpha_k)^2} = 0, \quad (16.181a)
\]

\[
D_1 = 0, \quad A_{1k} = 0. \quad (16.181b)
\]

Furthermore, for \( z = \mathcal{O}(\delta^{-1}) \) the linear leading-order term \( C_0 z \) is too big to match with \( T_0 \), which is \( \mathcal{O}(1) \), so \( C_0 = 0 \). Altogether, our first terms of the inner expansion are greatly simplified to

\[
\vartheta_0(r, z) = 1, \quad \vartheta_1(r, z) = C_1 z.
\]  

(16.182)

Under matching with outer solution \( T_0 \), this leads to \( \lim_{z \to 0} Q_0(Z) = 1 \), which requires

\[
k_0 = \frac{1}{3} \sqrt{5}.
\]  

(16.183)
To determine $C_1$ of $\vartheta$ we note that from matching with the outer solution we have

$$Q_0(0) + \delta z Q_0'(0) + \cdots \sim 1 + \varepsilon^{1/2} C_1 z + \cdots.$$  \hfill (16.184)

The correspondence in $z$ requires

$$C_1 = Q_0'(0) = -\frac{2}{3} k_0^{-1} = -\frac{2}{5} \sqrt{5}.$$  \hfill (16.185)

Altogether we have the following description of the sought solution:

$$T(r, z) = \begin{cases} 
1 + 3 \frac{5}{3} \sqrt{5} Z^{-2/3} + \cdots & \text{if } Z = O(1), \\
1 - \frac{2}{5} (5\varepsilon)^{1/2} z + \cdots & \text{if } Z = O(\varepsilon^{1/2}).
\end{cases}$$  \hfill (16.186)

To leading order there is effectively no boundary layer. Apparently, the boundary condition at $z = 0$ is simple enough to be satisfied by the outer solution up to $O(\varepsilon)$ and we may as well write

$$T(r, z) = \frac{1}{(1 + \frac{3}{5} \sqrt{5} Z)^{2/3}} + O(\varepsilon) \quad \text{for } 0 \leq z < \infty.$$  \hfill (16.187)

Finally, as a consistency test, it is of interest to check whether the heat flux into the bar at $z = 0$ is indeed equal to the total flux out of the bar along $r = 1$. This is indeed the case:

$$2\pi \int_0^1 \frac{\partial T}{\partial z}(r, 0) r \, dr \approx -2\pi \int_0^1 \frac{2}{5} (5\varepsilon)^{1/2} r \, dr = -\frac{2}{5} \pi (5\varepsilon)^{1/2},$$  \hfill (16.188a)

$$2\pi \int_0^\infty -\varepsilon T^4(r, z) \, dz \approx -2\pi \varepsilon^{1/2} \int_0^\infty \left(1 + \frac{3}{5} \sqrt{5} Z\right)^{-8/3} \, dZ = -\frac{2}{5} \pi (5\varepsilon)^{1/2}.$$  \hfill (16.188b)

An example is plotted in Figure 16.21 for the values listed in Table 16.3, where in particular $\varepsilon = 9.6688 \times 10^{-2}$. We see that for $z = 0.34 \, m$ the temperature drops below the environment temperature of (say) 300 K as a result of our neglecting $T_0$. Physically it is impossible for an object to release heat by radiation to a temperature lower than its environment.

Figure 16.21. Temperature variation along the monocrystalline bar.
16.7.4 Numerical Solution Method

An interesting approach to solving the boundary value problem (16.165) numerically is to use alternating direction implicit (ADI); see Section 9.4. Take \( L \) large enough and divide the interval \((0, L)\) into \( M + 1 \) intervals of length \( \Delta z := L/(M + 1) \). Define in addition \( \Delta r := 1/(N + 1) \), \( r_j := j \Delta r \) \((j = 0, 1, \ldots, N + 1)\), and \( z_k := k \Delta z \) \((k = 0, 1, \ldots, M + 1)\). For the second derivative with respect to \( r \) we use the difference approximation

\[
\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T}{\partial r} \right) (r_j, z_k) = \frac{1}{\Delta r^2} (\beta_j T_{j+1,k} - 2T_{j,k} + \alpha_j T_{j-1,k}). \tag{16.189a}
\]

where the geometrical factors \( \alpha_j \) and \( \beta_j \) are defined by

\[
\alpha_j := 1 - \frac{\Delta r}{2r_j}, \quad \beta_j := 1 + \frac{\Delta r}{2r_j}, \quad j = 2, 3, \ldots, N. \tag{16.189b}
\]

Note that \( \frac{\partial}{\partial r} T(0, z) = 0 \) and therefore we have, for \( j = 1 \), that

\[
\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T}{\partial r} \right) (r_1, z_k) \equiv \frac{4}{\Delta r^2} (T_{1,k} - T_{0,k}); \tag{16.190}
\]

see (5.47a). For the \( z \) derivatives we take the standard central difference approximations

\[
\frac{\partial^2 T}{\partial z^2} (r_j, z_k) \equiv \frac{1}{\Delta z^2} (T_{j,k+1} - 2T_{j,k} + T_{j,k-1}), \tag{16.191a}
\]

\[
\frac{\partial T}{\partial z} (r_j, z_k) \equiv \frac{1}{2\Delta z} (T_{j,k+1} - T_{j,k-1}). \tag{16.191b}
\]

Using the boundary conditions in (16.165), we thus obtain a system of equations of the form

\[
Au = f(u), \tag{16.192}
\]

where \( u \) is the vector of unknown values \( T_{j,k} \), \( A \) is the matrix containing the coefficients defined by the discretisations in (16.189) to (16.191), and \( f(u) \) contains the boundary value terms. Note that \( f(u) \) is nonlinear due to the boundary condition at \( r = 1 \). We can split \( A \) into a part containing \( r \) derivative approximations and a part containing \( z \) derivative approximations, say (cf. (9.66))

\[
A = A_r + A_z. \tag{16.193}
\]

The ADI scheme for (16.192) then reads

\[
u^{i+1} = u^i + \tau (A_r u^i + A_z u^{i+1} - f(u^i)), \tag{16.194a}
\]

\[
u^{i+2} = u^{i+1} + \tau (A_r u^{i+2} + A_z u^{i+1} - f(u^{i+2})). \tag{16.194b}
\]

As we found out in Section 11.8.1, \( \tau \) should be chosen \( O(\Delta r) \) and \( O(\Delta z) \). We can directly compute \( u^{i+1} \) from (16.194a). However, (16.194b) is nonlinear. Fortunately, the nonlinearity is quite mild and we can solve this system with only a few Newton iterations. In Figure 16.22 we show the result of applying ADI as outlined above, computed with \( M = N = 30 \). As initial guess \( u^0 \) we take the first order term in (16.187). The numerical solution is practically independent of \( r \), which is indeed in agreement with the asymptotic analysis of the previous section.
16.7.5 Discussion and Related Problems

- An important result is that the role of $V$ (via $Pe$) is negligible for the adopted velocities of 1 mm/min when $Pe = O(\varepsilon)$, but it is certainly to be included for larger values when $Pe \geq O(\varepsilon^{1/2})$.

- The present model is well able to provide insight into the behaviour of the heat flow entering the bar from the reservoir of fluid crystal at $z = 0$ and leaving the bar by radiation. Some simplifications, however, should be removed for more accurate results. For example, one should consider a temperature-dependent $\kappa$, a finite environment temperature $T_0$, the effect of slightly larger velocities $V$, and a finite length of the bar if the bar is relatively short compared to the diameter $a$ or the slow length scale $a/\varepsilon^{1/2}$. Heat release by natural convection of the surrounding air may be considered for the part of the bar with a temperature range near or just above $T_0$.

16.7.6 Exercises

16.27. Formulate the inner solution for an arbitrary initial $(r, \phi)$-dependent temperature profile and verify the consistency check (16.188).

16.28. The present model does not include a varying $\kappa$. Formulate the same problem for $\kappa(T) = \kappa_m T_m / T$ and solve the outer solution. Note that condition (16.167d) is inadequate to determine one integration constant. It is always satisfied for any two integration constants. Since the inner solution is only able to determine one condition, we need to invoke an extra condition. In [73] this is solved by taking the bar of finite length and assuming a zero heat flux out of the far end. What is the real cause of this strange behaviour?

16.29. Estimate the order of magnitude of cooling by natural convection (the relevant dimensionless number is a Biot number). Compare this with cooling due to radiation.

(a) Show that ADI as outlined in Section 16.7.4 converges for \( \tau \) small enough.

(b) Determine the optimal relaxation value of the parameter \( \tau \).

### 16.8 A Catalytic Reaction Problem in Pellets

#### 16.8.1 Problem Formulation

A catalytic reaction is a chemical reaction between reactants, of which one—the catalyst—returns to its original state after the reaction. Its role is entirely to enable the reaction to happen. An example of a catalyst is platinum. The primary reactant is usually a liquid or a gas. As the catalyst and the reactant are immiscible, the reaction occurs at the catalyst surface, which is therefore made as large as possible. A way to achieve this is to apply the catalyst to the pores of porous pellets. The reactant diffuses from the surface to the inside of the pellet. Meanwhile, being in contact with the catalyst, the reactant is converted to the final product.

Assume that reactant \( A \) reacts in an equilibrium reaction with catalyst \( S \) at the pellet pore surface to the intermediate product \( AS \) in a way described by

\[
A + S \rightleftharpoons AS
\]

with concentrations \( [A] \) mol/m\(^3\), \( [S] \) mol/kg, and \( [AS] \) mol/kg satisfying

\[
[AS] = K[A][S], \quad \text{where} \quad [S] = [S]_0 - [AS],
\]

for equilibrium constant \( K \) m\(^3\)/mol and initially available catalyst \( [S]_0 \). It follows that

\[
[AS] = \frac{K[A][S]_0}{1 + K[A]}.
\]

If \( AS \) reacts in a first order reaction to the final product \( B \) with reaction rate constant \( \tilde{k} \), we have the reaction equation

\[
\frac{\partial}{\partial t}[AS] = -\tilde{k}[AS] = -\tilde{k}[S]_0\frac{K[A]}{1 + K[A]} = -k\frac{K[A]}{1 + K[A]}.
\]

The term \( K[A]/(1 + K[A]) \) is called a Langmuir isotherm. This reaction acts as a source term for species \( B \) or, equivalently, as a sink term for \( A \). We define, for ease of notation, the concentration of \( A \) inside a pellet as \( C := [A] \). Under the additional assumption of a well-stirred fluid in order to maintain a constant concentration \( C = C_R \) and \( [B] = 0 \) at the outer surface of spherical pellets (see Figure 16.23), we obtain the initial boundary value problem

\[
\frac{\partial C}{\partial t} - \nabla \cdot (D \nabla C) = -k\frac{KC}{1 + KC}, \quad 0 < \tilde{r} < R, \quad t > 0, \quad (16.195a)
\]

\[
C(r, 0) = 0, \quad 0 < \tilde{r} < R, \quad (16.195b)
\]

\[
C(R, t) = C_R, \quad \frac{\partial}{\partial \tilde{r}} C(0, t) = 0, \quad t > 0, \quad (16.195c)
\]
Chapter 16. Modeling, Analysing, and Simulating Problems from Practice

where $D$ is the diffusion coefficient of $C$ inside the pellet. After a time $t \gg \max(R^2/D, (1 + KC_R)/kK)$, the concentration $C$ is in stationary equilibrium. Assuming spherical symmetry and a constant diffusion coefficient $D$, $C$ satisfies the boundary value problem

$$
D \frac{1}{\tilde{r}^2} \frac{d}{d\tilde{r}} \left( \tilde{r}^2 \frac{dC}{d\tilde{r}} \right) = k \frac{KC}{1 + KC}, \quad 0 < \tilde{r} < R, \quad (16.196a)
$$

$$
C(R) = C_R, \quad \frac{d}{d\tilde{r}}C(0) = 0. \quad (16.196b)
$$

A point of concern mathematically is the singularity at $C = -1/K$. Physically, of course, this is just an irrelevant anomaly of the model because $C$ is only meant to be positive, but it is not clear beforehand if the model per se precludes negative $C$. We therefore assume whenever necessary that the right-hand side of (16.196a) is zero for $C < 0$.

We make the problem dimensionless as follows:

$$
c := \frac{C}{C_R}, \quad r := \frac{\tilde{r}}{R}, \quad \lambda = \frac{kR^2}{DC_R}, \quad \alpha = \frac{1}{KC_R}, \quad (16.197)
$$

such that

$$
1 \frac{d}{r^2} \frac{d}{dr} \left( r^2 \frac{dc}{dr} \right) = \lambda \frac{cH(c)}{\alpha + c}, \quad 0 < r < 1, \quad (16.198a)
$$

$$
c(1) = 1, \quad c'(0) = 0, \quad (16.198b)
$$

where the prime ($'$) denotes differentiation with respect to $r$ and $H$ denotes the Heaviside function $(4.23)$. The net mass flux into the pellet, an important final result, is given by $-4\pi R^2 D \frac{dc}{dr} (R)$ (Fick’s law; see (6.40)). In dimensionless quantities this is thus equal to $-4\pi R DC_R c'(1)$.

We will consider in detail the behaviour of the solution for the particular limit of $\lambda = O(1), \alpha \to 0$. This is an interesting limit because, as we will see, a critical value of $\lambda$ exists above which a (steady) front exists.

In addition, the approximate analytical results will be compared with probably more accurate numerically generated data. It should be remembered, however, that the analytical
results take the form of a functional relation among \( c, r, \lambda, \) and \( \alpha. \) This relationship is therefore different from just the numerical value at a given point. The information it provides includes trends, behaviour at infinity or near singular points, etc., and has therefore to be interpreted in a much wider sense than just at the level of “numerical outcome.”

16.8.2 Numerical Solution Method

To discretise (16.198a) we employ the finite volume method; see Chapter 5. In this section we restrict ourselves to a short outline. Integrating (16.198a) over a control volume \( V_j := (r_{j-\frac{1}{2}}, r_{j+\frac{1}{2}}) = ((j - \frac{1}{2})\Delta r, (j + \frac{1}{2})\Delta r), \) centred around grid point \( r_j = j\Delta r \) with \( \Delta r \) the grid size, and applying Gauss’s theorem, we obtain the integral balance

\[
\oint_{\partial V_j} \frac{\partial c}{\partial n} \, dS = \int_{V_j} f(c) \, dV, \quad f(c) := \frac{eH(c)}{\alpha + e},
\]

(16.199)

where \( n \) is the outward unit normal on the boundary \( \partial V_j \) of the control volume. Next we have to approximate both integrals in (16.199). For the surface integral we have

\[
\oint_{\partial V_j} \frac{\partial c}{\partial n} \, dS = \int_{r=r_{j-\frac{1}{2}}}^{r=r_{j+\frac{1}{2}}} \frac{\partial c}{\partial r} \, dS - \int_{r=r_{j-\frac{1}{2}}}^{r=r_{j-\frac{1}{2}+1}} \frac{\partial c}{\partial r} \, dS
\]

\[
\approx \frac{\partial c}{\partial r} \left( r_{j+\frac{1}{2}} \right) 4\pi r_{j+\frac{1}{2}}^2 - \frac{\partial c}{\partial r} \left( r_{j-\frac{1}{2}} \right) 4\pi r_{j-\frac{1}{2}}^2
\]

\[
= \frac{4\pi}{\Delta r} \left( r_{j+\frac{1}{2}}^3 \alpha_{j+\frac{1}{2}} (c_{j+1} - c_j) - r_{j-\frac{1}{2}}^3 \alpha_{j-\frac{1}{2}} (c_j - c_{j-1}) \right),
\]

(16.200)

where \( c_j \) denotes the numerical approximation of \( c(r_j). \) To derive (16.200) we have used that \( n = e, \) at the sphere \( r = r_{j+\frac{1}{2}} \) and \( n = -e, \) at the sphere \( r = r_{j-\frac{1}{2}}. \) Furthermore, we have approximated the integrals and derivatives involved by the midpoint rule and the central difference scheme, respectively. To approximate the volume integral we use the midpoint rule to obtain

\[
\int_{V_j} f(c) \, dV \approx f(c_j)|V_j| = \frac{4}{3}\pi f(c_j) \left( r_{j+\frac{1}{2}}^3 - r_{j-\frac{1}{2}}^3 \right).
\]

(16.201)

Inserting (16.200) and (16.201) into (16.199) and taking into account the relation \( x^3 - y^3 = (x - y)(x^2 + xy + y^2), \) we obtain the discrete conservation law

\[
\frac{1}{\Delta r^2} \left( \beta_{j+\frac{1}{2}} (c_{j+1} - c_j) - \alpha_{j-\frac{1}{2}} (c_j - c_{j-1}) \right) = f(c_j),
\]

(16.202)

where the coefficients \( \beta_{j+\frac{1}{2}} \) and \( \alpha_{j-\frac{1}{2}} \) are defined by

\[
\beta_{j+\frac{1}{2}} := \frac{3r_{j+1}^2}{r_{j+\frac{1}{2}}^2 + r_{j+\frac{1}{2}}r_{j-\frac{1}{2}} + r_{j-\frac{1}{2}}^2}, \quad \alpha_{j-\frac{1}{2}} := \frac{3r_{j-1}^2}{r_{j+\frac{1}{2}}^2 + r_{j+\frac{1}{2}}r_{j-\frac{1}{2}} + r_{j-\frac{1}{2}}^2}.
\]

(16.203)

Note that the diffusion term, the left-hand side of (16.198a), is not singular in \( r = 0 \) because \( c(r) = a + br^2 + O(r^4). \) Then we have \( r^{-2}(r^2c)' = c'' + 2c'/r = 6b + O(r^2), \) and we have to adjust the numerical scheme (16.202) accordingly. Finally, we use Newton’s method to solve the nonlinear system (16.202); see the exercises.
16.8.3 Asymptotic Analysis

The limit we will consider is one of a strong nonlinear reaction and corresponds to \( \lambda = O(1) \), \( \alpha \to 0 \). We note the following:

(i) As long as \( c \neq 0 \), the small parameter \( \alpha \) is not multiplied by the highest derivative. From the physics, we do know, however, that \( c > 0 \), although \( c \) may be small. So we have to be aware of singular behaviour if \( c \leq O(\alpha) \) somewhere.

(ii) The reduced problem (\( \alpha = 0 \)) is corrected by terms of \( O(\alpha) \).

(iii) From the boundary condition \( c(1) = 1 \) we can, at least in the neighbourhood of \( r = 1 \), estimate the order of magnitude of \( c \) as \( c = O(1) \).

We start with the assumption (see Chapter 15) that we can expand \( c \) as follows:

\[
c(r; \alpha) = c_0(r) + \alpha c_1(r) + O(\alpha^2).
\] (16.204)

Since \( c(1) = 1 \), this will probably be valid in the neighbourhood of \( r = 1 \). Equation (16.198a) becomes

\[
\frac{1}{r^2} (r^2 c_0')' + \frac{1}{r^2} (r^2 c_1')' + O(\alpha^2) = \frac{\lambda}{c + \alpha} - \frac{\alpha}{c_0} + O(\alpha^2).
\] (16.205)

Note, in addition to point (i) above, that the geometrical series expansion used is not valid any more if \( c \leq O(\alpha) \).

To leading order we get

\[
\frac{1}{r^2} (r^2 c_0')' = \lambda
\] (16.206)

with a first integral

\[
c_0' = \frac{1}{3} \lambda r + \frac{A_0}{r^2}.
\] (16.207)

If \( c_0 \) is a valid approximation everywhere, we can apply the boundary condition in \( r = 0 \) such that \( A_0 = 0 \). Integrating (16.207) and applying the boundary condition at \( r = 1 \), we find

\[
c_0(r) = 1 - \frac{1}{6} \lambda (1 - r^2).
\] (16.208)

We see now that there is an important distinction to be made between a solution for \( \lambda < 6 \) and one for \( \lambda > 6 \):

\( \lambda < 6 \): \( c_0 \neq 0 \) everywhere, so our assumption is indeed satisfied.

\( \lambda > 6 \): \( c_0 \) has a zero in \( r_0 = \sqrt{1 - 6/\lambda} \), and the solution is certainly not valid near and below \( r_0 \).

As the solution found is (to leading order) complete for \( \lambda < 6 \), we continue with the assumption \( \lambda > 6 \). The question that immediately arises is whether this solution, invalid near \( r_0 \), is at least valid for \( r \) well above \( r_0 \). The stunning answer is that this solution is not valid anywhere!

As soon as the interval is cut into two pieces by the zero \( r_0 \), we cannot apply to \( c_0 \) the boundary condition at \( r = 0 \). One might say that \( c_0 \) has no contact any more with the other end. So the constant of integration \( A_0 \) (16.207) remains undetermined yet, and so is any
16.8. A Catalytic Reaction Problem in Pellets

zero \( r = r_1 \) of this \( c_0 \). So we have for \( r > r_1 \) that

\[
c_0(r) = 1 - \frac{1}{6} \lambda (1 - r^2) + A_0 \left( 1 - \frac{1}{r} \right).
\]

(16.209)

For the interval \( 0 \leq r < r_1 \) we have to consider the fact that physically \( c \) represents a positive quantity (a concentration) and therefore cannot be negative. (Strictly speaking, this is not a relevant argument mathematically since we are dealing with a mathematical model that may just be invalid for \( \lambda > 6 \). A mathematical proof of the solution being positive is better; see the exercises.) So we assume that \( 0 \leq c \leq \mathcal{O}(\alpha) \). If we scale \( c \) on a gauge function \( \gamma(\alpha) \) to be determined, so \( c = \alpha \gamma(\alpha) \tilde{c} \), then we have

\[
\alpha \frac{1}{r^2} \frac{d}{dr} \left( r^2 \frac{d \tilde{c}}{dr} \right) = \frac{\lambda}{1 + \gamma \tilde{c}}.
\]

(16.210)

and we see that for any choice of \( \gamma \) every asymptotic expansion \( \tilde{c}(r; \alpha) = \tilde{c}_0(r) + \cdots \) leads to \( c(r) \equiv 0 \). Apparently, the solution here is asymptotically equivalent to zero.

What remain to be determined are the position and neighbourhood of the transition point \( r_1 \), where both parts of the solution come together. As noted before, this point is still unknown since we do not know the constant \( A_0 \) yet.

The only way that the solution, asymptotically equivalent to zero, can be set free is by a quick transition in a boundary layer (Section 15.4.1), because only then are the derivatives of \( c \) big enough to outweigh the small factor \( \alpha \). Therefore we will search for a boundary layer

\[
c = \alpha^n \psi, \quad r = r_1 + \alpha^m \xi
\]

(16.211)

that yields a reduced equation rich enough to allow this transition. We substitute

\[
\alpha^{n-2m} \frac{d}{(r_1 + \alpha^m \xi)^2} \left( (r_1 + \alpha^m \xi)^2 \frac{d^2 \psi}{d \xi^2} \right) = \frac{\lambda \alpha^n \psi}{\alpha + \alpha^n \psi},
\]

(16.212)

which becomes, under \( \alpha \to 0 \),

\[
\alpha^{n-2m} \frac{d^2 \psi}{d \xi^2} = \frac{\lambda}{\alpha + \alpha^n \psi} \psi.
\]

(16.213)

Apparently, the richest structure (distinguished limit) is found for \( m = \frac{1}{2} \) and \( n = 1 \); thus \( c = \alpha \psi \) and \( r = r_1 + \sqrt{\alpha} \xi \) and

\[
\frac{d^2 \psi}{d \xi^2} = \frac{\lambda}{1 + \psi}. \quad (16.214)
\]

Of remarkable importance now are the matching conditions. First, \( \psi \) has to connect to the zero solution to the left of \( r_1 \), so \( \psi \to 0 \) if \( \xi \to -\infty \). At the right-hand side of \( r_1 \), however, \( \psi \) has to connect to \( c_0(r) \), the behaviour of which is described near \( r_1 \) by

\[
c_0(r) = c_0(r_1) + (r - r_1) c_0'(r_1) + \frac{1}{2} (r - r_1)^2 c_0''(r_1) + \cdots \quad (r \downarrow r_1)
\]

\[
= \sqrt{\alpha} \xi c_0'(r_1) + \frac{1}{2} \alpha \xi^2 c_0''(r_1) + \cdots
\]

\[
= \alpha \psi(\xi) \quad (\xi \to \infty)
\]

\[
\approx \alpha \psi(\xi)
\]

(16.215)
This provides us with the following important information. The $O(\sqrt{\alpha})$ term has to be equal to zero, so $c'_0(r_1) = \frac{1}{3} \lambda r_1 + A_0/r_1^2 = 0$, implying that $A_0 = -\frac{1}{3} \lambda r_1^3$. Physically, this makes sense: if there is no source at the border of the region where the concentration vanishes, then the flux there ($\sim c'$) cannot be anything else but zero. Next we can determine $r_1$ from $c_0(r_1) = 0$, which is now $\frac{1}{3} r_1^3 - \frac{1}{2} r_1^2 + \frac{1}{6} - \lambda^{-1} = 0$ and has the solution \[ r_1 = \frac{1}{2} + \cos \left( \frac{1}{3} \arccos \left( 12 \lambda^{-1} - 1 \right) - \frac{2}{3} \pi \right). \] (16.216)

In Figure 16.24 $r_1$ is plotted as a function of $1/\lambda$ (note that $r_1 = 0$ if $\lambda = 6$).

We are now ready as far as the most important information is concerned. The outer solution $c_0$ is described by

\[ c_0(r) = 1 - \frac{1}{6} \lambda (1 - r^2), \quad \lambda \leq 6, \quad (16.217a) \]

\[ c_0(r) = \begin{cases} 
0 & \text{if } 0 \leq r < r_1 \\
1 - \frac{1}{6} \lambda (1 - r^2) - \frac{1}{3} \lambda r_1^3 \left( 1 - \frac{1}{r} \right) & \text{if } r_1 \leq r \leq 1
\end{cases}, \quad \lambda \geq 6. \quad (16.217b) \]

The mass flux into the pellet is found from $c'_0(1) = \frac{1}{3} \lambda$ if $\lambda \leq 6$ and $c'_0(1) = \frac{1}{3} \lambda (1 - r_1^3)$ if $\lambda \geq 6$.

The boundary layer, or inner solution, near $r = r_1$ is rather complicated so we leave its determining as an exercise. Interestingly enough, the behaviour of the boundary layer
was essential for fixing the outer solution, although the inner problem itself need not be solved!

Some examples of the outer solution (solid line), with a numerical “exact” solution (dotted line) as a reference, are given in Figure 16.25. We see that for \( \alpha = 0.1 \) the approximation isn’t very good, although it is indeed just the expected error of \( O(\alpha) \). For \( \alpha = 0.01 \) the analytical prediction of a transition layer is completely confirmed by the numerical solution.

### 16.8.4 Discussion and Related Problems

- We have considered one particular asymptotic combination of \( \lambda \) and \( \alpha \) because it contained the interesting (steady) reaction-diffusion front if \( \lambda > 6 \). Other combinations may be considered too. We may distinguish the following five cases:

1. \( \lambda \to 0, \quad \alpha = O(1) \) (regular),
2. \( \lambda \to \infty, \quad \alpha = O(1) \) (singular),
3. \( \lambda = O(1), \quad \alpha \to 0 \) (singular and regular),
4. \( \lambda = O(1), \quad \alpha \to \infty \) (regular),
5. \( \lambda \to \infty, \quad \alpha \to \infty \quad \lambda / \alpha = O(1) \) (linearization, regular).
In case 1 there is little reaction, and the diffusion effects dominate. In case 2 it is the other way round. In case 3, considered above, the concentration is relatively large, and it is small in case 4. C is small in case 5 because we take $C_R$ small ($\lambda/\alpha$ is independent of $C_R$). To leading order this comes down to linearizing the equation.

• Suggestions for further research are
  – construction of higher-order corrections of the above-found solution $c_0$ in order to see if the numerical agreement in Figure 16.25 may be improved;
  – generalisations to other type of reactions, i.e., other right-hand sides of (16.195a).

• Related problems are in particular the reaction-diffusion problem of the polymerisation reaction discussed in Section 16.1 (panel production by resin curing) and the combustion problem discussed in Section 16.3 (the thermal explosion).

• One of the interesting features that appeared in the analysis is the transition layer for $\lambda > 6$ separating two regions of the outer solution. A very similar transition layer may be found in the groundwater problem of Section 16.6.

### 16.8.5 Exercises

16.31. Analyse asymptotically the regular perturbation problems of cases 1, 4, and 5.

16.32. Analyse asymptotically the singular (boundary layer type) perturbation problem of case 2.

16.33. Give the discretisation of (16.202) for $j = 0$.


16.35. Determine the inner solution (implicitly, in the form of an integral) given by the problem (16.214). Find the integration constants by matching arguments.

16.36. We may prove uniqueness and positivity of the solution of the problem

$$\nabla^2 c = f(c) \quad \text{for} \quad x \in \Omega, \quad c = b \quad \text{for} \quad x \in \partial \Omega,$$

where $\Omega$ is the unit sphere, $b$ is a nonnegative constant, and $f$ is a sufficiently smooth function with $f(0) = 0$ if $c \leq 0$ and $f'(c) > 0$ for $c > 0$.

(a) Assume there exist two solutions $c_1$ and $c_2$ satisfying the same boundary condition. Let $u = c_2 - c_1$. Apply Green’s first identity to $u \nabla^2 u$ with an integral over $\Omega$. Apply a monotonicity argument to $f(c_2) - f(c_1)$ and show that $c_2 = c_1$.

(b) Observe that $c \equiv 0$ is the unique solution for $b = 0$. Assume a solution $c$ with positive boundary value. If $c(r_1) = 0$ for some $r = r_1 < 1$, argue that $c$ remains zero for $0 \leq r < r_1$. In other words, $c \geq 0$. 


16.9 Outdoor Noise Enhancement by Atmospheric Conditions

16.9.1 Problem Formulation

The propagation of sound waves in the atmosphere is greatly affected by wind (see Figure 16.26). For example, the communication between two people, one downstream and one upstream, is not symmetric. The one upstream is easier to understand for the one downstream than the other way around. This is not because the wind “carries the waves faster,” but it is due to refraction by the wind gradient (the atmospheric boundary layer). It may be compared to the refraction by a sound speed or temperature gradient described in Example 15.37, but the refraction is now directionally dependent. We can study this problem by the following ray model, a generalisation of Example 15.37.

16.9.2 The Model

Consider an inviscid, perfect, non-heat-conducting, subsonic, and steady mean flow with density $\rho_0$, velocity $v_0$, pressure $p_0$, and entropy $s_0$, satisfying the steady version of (7.5), i.e., the equations

\[
\nabla \cdot (\rho_0 v_0) = 0, \tag{16.218a}
\]

\[
\rho_0 v_0 \cdot \nabla v_0 + \nabla p_0 = 0, \tag{16.218b}
\]

\[
v_0 \cdot \nabla s_0 = 0, \tag{16.218c}
\]

\[
s_0 = CV \ln p_0 - CP \ln \rho_0. \tag{16.218d}
\]

We introduce the sound speed $c_0$ given by

\[
c_0^2 = \frac{CP}{CV} \frac{p_0}{\rho_0} = \gamma \frac{p_0}{\rho_0}, \tag{16.218e}
\]

where the gas constants $C_P$ and $C_V$ are the specific heats at constant pressure and constant volume, respectively, and $\gamma = C_P / C_V$. Acoustic perturbations in pressure $p'$, density $\rho'$,
velocity $v'$, and entropy $s'$ are given by the linear perturbations of (7.5), i.e., the equations

\[
\frac{\partial \rho'}{\partial t} + v_0 \cdot \nabla \rho' + v' \cdot \nabla \rho_0 + \rho_0 \nabla \cdot v' + \rho' \nabla \cdot v_0 = 0, \tag{16.219a}
\]

\[
\rho_0 \left( \frac{\partial v'}{\partial t} + v_0 \cdot \nabla v' + v' \cdot \nabla v_0 \right) + \rho' v_0 \cdot \nabla v_0 = -\nabla p', \tag{16.219b}
\]

\[
\frac{\partial s'}{\partial t} + v_0 \cdot \nabla s' + v' \cdot \nabla s_0 = 0, \tag{16.219c}
\]

\[
s' = \frac{C_v}{\rho_0} p' - \frac{C_p}{\rho_0} \rho'. \tag{16.219d}
\]

If $\omega_{ref}$ is a typical frequency of the perturbations and $c_{ref}$ and $\rho_{ref}$ are typical values of mean sound speed and density, we can make (16.218) and (16.219) dimensionless by scaling velocities on $c_{ref}$, densities on $\rho_{ref}$, pressures on $\rho_{ref}$, time on $\omega_{ref}^{-1}$, length on $c_{ref}/\omega_{ref}$, and entropy on $C_p$ (but that is less important). The equations are now of exactly the same form as before, so we will not repeat them here. Thus in the following we will regard system (16.219) as the equations for the dimensionless perturbations.

We assume that the typical wavelength $c_{ref}/\omega_{ref}$ is short compared to the typical length scale $L$ of the mean flow variations. In order to quantify this, we introduce the small parameter $\epsilon := c_{ref}/\omega_{ref} L \ll 1$. We introduce the slow variables $X := \epsilon x$ and $T := \epsilon t$. The mean flow variables depend on $X$ only.

We assume that the perturbations are harmonic in time with frequency $\omega$, equal to or of the order of $\omega_{ref}$. Next we introduce the ray approximation (see Section 15.4.2)

\[
p', \rho', v', s' = P(X; \epsilon), R(X; \epsilon), V(X; \epsilon), S(X; \epsilon) \times e^{i(\omega T - \theta(X, \epsilon))}/\epsilon. \tag{16.220}
\]

### 16.9.3 Asymptotic Analysis

We define the operator $\nabla := \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$ so that $\nabla = \epsilon \nabla$. We introduce the local wave vector $\kappa := \nabla \theta$, $\kappa := |\kappa|$, and the phase plane normal field $n := \nabla \theta / |\nabla \theta| = \kappa / \kappa$. The phase planes $\theta = \omega T = \text{constant}$ describe the propagating wave front. We define the shifted frequency $\Omega := \omega - v_0 \cdot \kappa$. Since

\[
\nabla p' = \left( -i \kappa P + \epsilon \nabla P \right) e^{i(\omega T - \theta)/\epsilon},
\]

etc., we obtain, by substituting (16.220) in (16.219) and dropping the exponentials, the equations

\[
i \Omega R - i \rho_0 V \cdot \kappa + \epsilon (v_0 \cdot \nabla R + V \cdot \nabla \rho_0 + \rho_0 \nabla \cdot V + R \nabla \cdot v_0) = 0, \tag{16.221a}
\]

\[
i \rho_0 \Omega V - i P \kappa + \epsilon (\rho_0 v_0 \cdot \nabla V + \rho_0 V \cdot \nabla v_0 + R v_0 \cdot \nabla v_0 + \nabla P) = 0, \tag{16.221b}
\]

\[
i \Omega S + \epsilon (v_0 \cdot \nabla S + V \cdot \nabla s_0) = 0, \tag{16.221c}
\]

\[
S = \frac{C_v}{\rho_0} P - \frac{C_p}{\rho_0} R. \tag{16.221d}
\]

We expand in the usual way (cf. (15.64)) to obtain

\[
P(X; \epsilon) = P(X) + O(\epsilon), \quad \vartheta(X; \epsilon) = \vartheta(X) + O(\epsilon), \tag{16.222}
\]
etc., where we adopt the convention of using the same notation for both the variable and its leading-order approximation. Then we get for the perturbations to $O(1)$

\begin{align}
\Omega R - \rho_0 V \cdot \kappa &= 0, \\
\rho_0 \Omega V - P \kappa &= 0, \\
\Omega S &= 0.
\end{align}

Using (16.221d), this leads to the relations

\begin{align}
S &= 0, \\
P &= c_0^2 R, \\
V &= \frac{P \kappa}{\rho_0 \Omega} \mathbf{n}.
\end{align}

written as a differential equation for $\vartheta$ this is

\begin{align}
c_0^2 |\nabla \vartheta|^2 &= (\omega - \mathbf{v}_0 \cdot \nabla \vartheta)^2.
\end{align}

In a subsonic flow with $|\mathbf{v}_0| < c_0$ and positive $\omega$ and $\kappa$ we can only have the positive branch of (16.225a), so we may assume that $c_0 \kappa = \Omega$ and thus $V = (P/\rho_0 c_0) \mathbf{n}$.

If we interpret (16.225a) as a dispersion relation $\omega = \omega(\kappa)$, the group velocity is given by

\begin{align}
\mathbf{v}_g = \frac{\partial \omega}{\partial \kappa} = c_0 \mathbf{n} + \mathbf{v}_0.
\end{align}

To rewrite equation (16.225b) in characteristic form, i.e., as a system of ODEs along rays, we apply Theorem 12.6. It is enlightening to make sure that we obtain a parametrization of the rays that corresponds as much as possible to physical time. It is easily verified that (16.225b), rewritten in the form (see Theorem 12.6)

\begin{align}
F(X, \vartheta, \kappa) := \frac{1}{2} c_0^2 \kappa^2 - \frac{1}{2} (\omega - \mathbf{v}_0 \cdot \kappa) = 0,
\end{align}

yields for characteristic variable $s$ the equation

\begin{align}
\frac{d \vartheta}{ds} = \kappa \cdot \frac{\partial F}{\partial \kappa} = \omega;
\end{align}

in other words, $\vartheta = \omega s + \text{constant}$. Since we know that $\vartheta = \omega T + \text{constant}$ along the phase surfaces, it follows that $s$ may be identified with the slow time scale $T$. The ray equations are then

\begin{align}
\vartheta(X_r) = \omega T + \text{constant}
\end{align}

along the ray $X = X_r(T)$ given by

\begin{align}
\frac{dX_r}{dT} = \frac{\partial F}{\partial \kappa} = c_0 \mathbf{n} + \mathbf{v}_0, \\
\frac{d\kappa}{dT} = -\kappa \frac{\partial F}{\partial \vartheta} - \nabla F = -\kappa \nabla c_0 - \kappa \cdot \frac{\partial \mathbf{v}_0}{\partial X},
\end{align}

where $\nabla F$ is the gradient of $F$ with respect to $X$. The evolution of $\mathbf{v}_0$ along the ray is given by

\begin{align}
\frac{d \kappa}{dT} = -\mathbf{v}_0 \cdot \frac{\partial F}{\partial \vartheta} - \nabla F = -\mathbf{v}_0 \cdot \frac{\partial \mathbf{v}_0}{\partial X}.
\end{align}
We have
\[ \nabla \cdot (v_0 R + \rho_0 V) = 0, \]  
\[ \rho_0 (v_0 \cdot \nabla V + V \cdot \nabla v_0) + R v_0 \cdot \nabla v_0 + \nabla P = 0, \]  
\[ V \cdot \nabla \rho_0 = 0. \]

Equation (16.228a), expressed in \( P \) and \( n \), becomes a conservation law for \( P \) equivalent to mass conservation; i.e.,
\[ \nabla \cdot \left( \frac{P(v_0 + c_0 n)}{c_0^2} \right) = 0. \]

This does not, however, include the effect of inertia, i.e., the conservation of momentum. For this we need some rather subtle, albeit technical, manipulations. First, we have from (16.228a) and (16.228b) that
\[ \nabla \cdot (\frac{v_0}{c_0} P + \rho_0 V) = 0, \]  
\[ \rho_0 (v_0 \cdot \nabla V + V \cdot \nabla v_0) + \nabla P - \frac{P}{\rho_0 c_0^2} \nabla \rho_0 = 0. \]

We multiply (16.230a) by \( P/\rho_0 \) and dot-multiply (16.230b) by \( V \) and add. Using the identity \( V \cdot (v_0, \nabla V) = \frac{1}{2} v_0 \cdot \nabla V^2 \), where \( V := |V| \), and the relation \( V \cdot \nabla \rho_0 = c_0^2 V \cdot \nabla \rho_0 \), we get
\[ \frac{1}{2} \rho_0 v_0 \cdot \nabla V^2 + \rho_0 V \cdot (V \cdot \nabla v_0) + V \cdot \nabla P - \frac{P}{\rho_0} V \cdot \nabla \rho_0 + \frac{P}{\rho_0} \nabla \cdot \left( \frac{v_0}{c_0^2} P + \rho_0 V \right) = 0. \]

Next we introduce the (energy-like) auxiliary variable
\[ E := \frac{\rho_0 V^2}{2} + \frac{p^2}{2\rho_0 c_0^2}. \]

Since
\[ \nabla E = \frac{1}{2} \rho_0 \nabla V^2 + \frac{1}{2} V^2 \nabla \rho_0 + \frac{1}{2} \frac{1}{\rho_0 c_0^2} \nabla p^2 + \frac{1}{2} p^2 \nabla \left( \frac{1}{\rho_0 c_0^2} \right), \]
we have
\[ v_0 \cdot \nabla E - \frac{1}{2} E v_0 \cdot \nabla \rho_0 - \frac{1}{2} p^2 v_0 \cdot \nabla \left( \frac{1}{\rho_0 c_0^2} \right) + \rho_0 V \cdot (V \cdot \nabla v_0) + \nabla \cdot (P V) \]
\[ + \frac{p^2}{\rho_0} \nabla \cdot \left( \frac{v_0}{c_0^2} \right) = 0. \]

Rewriting \( P \) and \( V \) in \( E \) and \( n \), i.e., \( P^2 = \rho_0 c_0^2 E \), \( V^2 = E/\rho_0 \), \( P V = c_0 E n \), we obtain
\[ v_0 \cdot \nabla E - \frac{1}{2} E v_0 \cdot \left[ \frac{1}{\rho_0} \nabla \rho_0 + \rho_0 c_0^2 \nabla \left( \frac{1}{\rho_0 c_0^2} \right) \right] + E n \cdot (n \cdot \nabla v_0) + \nabla \cdot (c_0 E n) \]
\[ + c_0^2 E \nabla \cdot \left( \frac{v_0}{c_0^2} \right) = 0. \]
Then we use the relation
\[-\frac{1}{2} \left[ \frac{1}{\rho_0} \nabla \rho_0 + \rho_0 c_0^2 \nabla \left( \frac{1}{\rho_0 c_0^2} \right) \right] = -\frac{1}{2} \nabla (\ln \rho_0 - \ln \rho_0 c_0^2) = \nabla \ln c_0 = \frac{1}{c_0} \nabla c_0 \]
to get
\[v_0 \cdot \nabla E + \frac{E}{c_0} v_0 \cdot \nabla c_0 + En \cdot (n \cdot \nabla v_0) + \nabla \cdot (c_0 En) + c_0^2 E \nabla \cdot \left( \frac{\nabla v_0}{c_0^2} \right) = 0.\]

Next we eliminate the term \( n \cdot (n \cdot \nabla v_0) \). First, we rewrite, by using the vectorial identities (J.5) and (J.6) and the fact that \( \nabla \times \kappa = \nabla \times \nabla \vartheta = 0 \), the following expression in \( \kappa \):
\[\kappa \cdot (\kappa \cdot \nabla v_0) = \kappa \cdot \nabla (\kappa \cdot v_0) - \frac{1}{2} v_0 \cdot \nabla \kappa^2 - v_0 \cdot (\nabla \times \kappa \times \kappa)\]
\[= \kappa \cdot \nabla (\kappa \cdot v_0) - \frac{1}{2} v_0 \cdot \nabla \kappa^2\]
\[= \kappa \cdot \nabla (\omega - \Omega) - \kappa v_0 \cdot \nabla \kappa\]
\[= -\kappa \cdot \nabla \Omega - (\Omega/c_0) v_0 \cdot \nabla (\Omega/c_0).\]

From this we have
\[n \cdot (n \cdot \nabla v_0) = -\frac{c_0}{\Omega} n \cdot \nabla \Omega - \frac{c_0}{\Omega} v_0 \cdot \nabla \left( \frac{\Omega}{c_0} \right) = c_0 \Omega n \cdot \nabla \left( \frac{1}{\Omega} \right) + \frac{\Omega}{c_0} v_0 \cdot \nabla \left( \frac{c_0}{\Omega} \right),\]
which yields
\[v_0 \cdot \nabla E + \frac{E}{c_0} v_0 \cdot \nabla c_0 + Ec_0 \Omega n \cdot \nabla \left( \frac{1}{\Omega} \right) + E \frac{\Omega}{c_0} v_0 \cdot \nabla \left( \frac{c_0}{\Omega} \right) + \nabla \cdot (c_0 En) + c_0^2 E \nabla \cdot \left( \frac{\nabla v_0}{c_0^2} \right) = 0.\]

This may finally be simplified to
\[\Omega \nabla \cdot \left( \frac{E}{\Omega} v_0 \right) + \nabla \cdot \left( \frac{E}{\Omega} \nabla n \cdot \nabla c_0 \right) = \Omega \nabla \cdot \left( \frac{E}{\Omega} (v_0 + c_0 n) \right) = \Omega \nabla \cdot \left( \frac{E}{\Omega} v_0 \right) = 0\]
or
\[\nabla \cdot \left( \frac{P^2 (c_0 n + v_0)}{\rho_0 c_0^2 \Omega} \right) = 0.\]

This is a classic result in ray acoustics known as the Blokhintzev invariant. It allows us to connect the variation of the amplitude with the rate of divergence of the rays.

At time \( T = T_1 \) the initial phase surface is given by \( \vartheta = C_0 \). At this surface we consider a small patch \( A_1 \) (see Figure 16.27). We trace the rays that start from \( A_1 \) until a time \( T = T_2 \), where they form the surface \( A_2 \), the projection of \( A_1 \) on the phase surface \( \vartheta = \alpha (T_2 - T_1) + C_0 \). When \( A_1 \) is small enough to approximate the ray’s amplitude \( P \) as constant over a cross section, this bundle of rays forms a ray tube. When we integrate
(16.232) over the ray tube, closed by $A_1$ and $A_2$, and apply Gauss’s theorem, we obtain

$$\int_{A_1 \cup A_2 \cup \text{tube wall}} \frac{p^2}{\rho_0 c_0^3} (c_0 n + v_0) \cdot dS = 0.$$  

Since $c_0 n + v_0$ is tangential to the ray tube surface, there is no contribution from the tube wall. Since $n$ is normal to any cross section $A$, while $(c_0 n + v_0) \cdot n = c_0 + (v_0 \cdot \kappa) / \kappa = \omega/\kappa$, we thus have

$$\frac{p^2}{\rho_0 c_0^3} \frac{\omega A}{\kappa^2} = \text{constant} \quad (16.233)$$

along the ray tube. A mathematically more satisfactory form [19] may be obtained by taking the limit $|A| \to 0$. Assume a smooth parametrisation $X = \Xi(T, \xi, \eta)$ such that any curve $(\xi, \eta) = \text{constant}$ corresponds to a ray. Then the change of surface elements between surfaces given by $T = \text{constant}$ is equal to the Jacobian $|\partial / \partial \xi \times \partial / \partial \eta|$. As a result we have

$$\frac{p^2}{\rho_0 c_0^3} \frac{\omega}{\kappa^2} \left| \frac{\partial \Xi}{\partial \xi} \times \frac{\partial \Xi}{\partial \eta} \right| = \text{constant}. \quad (16.234)$$

### 16.9.4 Solutions

As an example we compute the solution of the ODE system (16.227b) and (16.227c) for an initially vertical wave front located in the $YZ$ plane. We will consider waves of $108.2$ Hz ($\omega_{\text{ref}} = 680$ Hz) propagating along an atmospheric boundary layer, modeled as uniform shear flow and dimensionally given by $17 \tanh(z/L)$ m/s, where the typical mean flow length scale is $L = 50$ m. The mean sound speed is uniform and equal to $c_{\text{ref}} = 340$ m/s. Hence $\varepsilon = 10^{-2}$. The typical wavelength is then $\lambda = 2\pi c_{\text{ref}}/\omega_{\text{ref}} = 3.14$ m, which is indeed much smaller than $L$. Nondimensionally, we have the mean flow velocity given by

$$v_0(X) = (U_0(Z), 0, 0)^T, \quad U_0(Z) = U_\infty \tanh(Z), \quad (16.235)$$

where $U_\infty = 5 \times 10^{-2}$ and we have a constant sound speed $c_0 = 1$. More specifically, we compute the rays starting from the circle (in terms of the slow variables) of radius $\varepsilon$ and
16.9. Outdoor Noise Enhancement by Atmospheric Conditions

Figure 16.28. A ray tube starting from a circle of radius $\varepsilon$ in the $YZ$ plane, refracting downward in right-running shear flow.

centre $(0, 0, 0.5)$. In summary, we have the initial value problem

$$\begin{align*}
\frac{dX_0}{dT} &= \frac{c_0\kappa X}{\kappa} + U_0(Z), \\
\frac{dY_0}{dT} &= \frac{c_0\kappa Y}{\kappa}, \\
\frac{dZ_0}{dT} &= \frac{c_0\kappa Z}{\kappa}, \\
\frac{d\kappa X}{dT} &= \frac{d\kappa Y}{dT} = 0, \\
\frac{d\kappa Z}{dT} &= -U_0'(Z)\kappa X,
\end{align*}$$

(16.236a)

where $\kappa = (\kappa_X^2 + \kappa_Y^2 + \kappa_Z^2)^{1/2}$ and $\alpha = \alpha_n = n\pi/4$ for $n = 0, 1, 2, \ldots, 7$. The equations for $\kappa_X, \kappa_Y$, and $Y_r$ are easily solved and we find $\kappa_X(T) = 1, \kappa_Y(T) = 0, Y_r(T) = \varepsilon \cos \alpha$.

The ODE system for the other variables reduces to

$$\begin{align*}
\frac{dX_r}{dT} &= U_0(Z_r) + \frac{c_0}{\sqrt{1 + \kappa_r^2}}, \\
\frac{dZ_r}{dT} &= \frac{c_0 \kappa Z_r}{\sqrt{1 + \kappa_r^2}}, \\
\frac{d\kappa Z_r}{dT} &= -U_0'(Z_r).
\end{align*}$$

(16.237)

We computed numerically a solution of (16.237) for $0 \leq T \leq 4.8 (\equiv 4.8/\varepsilon \omega_{ref} = 0.71 \text{ s dimensionally})$ and the result is presented in Figure 16.28. We can see that the phase plane normal $n$ is initially directed horizontally, $n_0 = (1, 0, 0)$, and at the end it is bent down to $n_1 = (0.9785, 0, -0.2061)$.

Clearly, since we have a positive wind shear, i.e., $U_0'(Z) > 0$, the $Z$ component $\kappa_Z = \frac{\partial}{\partial Z} \theta$ decreases. In other words, the rays will bend toward the low wind speed regions. Propagating with the wind (as in the example), the waves bend down and remain near the ground (by repeated reflections, not included here); against the wind they bend up and disappear into free space.

It happens that the tube cross-sectional surface (i.e., the surface enclosed by the bundle of eight rays) increased in size by a factor of 1.4167, while the modulus of the tube wave vector $\kappa$ (i.e., averaged over the rays) increased by a factor of 1.0219. From (16.233) we obtain that the amplitude $P$ decreased by a factor of 0.8586, which is about 1.3 dB.

16.9.5 Discussion and Related Problems

- The above analysis refers to sound propagating in a medium with smooth mean flow and temperature gradients. Scattering at edges is to be included by other ways (see the literature on rays, e.g., [69, 89, 37, 114, 22]). Also, rays that creep along smoothly curved surfaces into the shadow zone are to be treated differently.
Chapter 16. Modeling, Analysing, and Simulating Problems from Practice

- If the temperature or velocity gradient exists only locally, the incident waves are refracted according to Snell’s law of refraction.
- With zero mean flow and a linearly varying sound speed profile, the ray paths are exactly circular.
- Salinity together with temperature and pressure gradients in the ocean may create layers of minimum sound speed where sound rays remain trapped.
- The solution of any problem that can be described by a first order scalar PDE, and can therefore be solved along characteristics, is ray like. In this sense the etching problem of Section 16.2 and the thin-layer flow of Section 16.10 are related problems.

16.9.6 Exercises

16.37. Prove that for subsonic flow the negative branch $\Omega = -c_0 k$ as solution of (16.225b) cannot hold.
16.38. Show the existence of circular ray paths for linearly varying sound speed profiles.
16.39. Derive Snell’s law for a temperature change across a thin layer.

16.10 Thin-Layer Flow Along a Curved Surface

An example of a regular perturbation problem is the flow of a thin layer of water along a curved surface, e.g., in a sink or a toilet [125].

The flow of a moderately thin layer of water of sufficiently high speed along a curved surface, like a sink or a toilet, is presumably dominated by a balance between inertia and gravity, while viscosity and surface tension are less important. Of course, this is not entirely correct. The Reynolds number of water flow (the ratio between inertia and viscous forces), based on a density of $10^3$ kg/m$^3$, a thickness of 1 mm, and a velocity of 1 m/s, is 1000, which is indeed large but may not be large enough to ignore accumulated friction effects. The Weber number (the ratio between inertia and surface tension), based on similar estimates, is around 15, which is again large but may not be large enough in all cases, e.g., to avoid drop formation.

Assuming irrotational steady flow allowing a velocity potential, and assuming a small layer thickness compared to other length scales, we may describe the flow may be described by a regular perturbation in terms of the dimensionless typical layer thickness $\varepsilon$. The resulting leading-order equations are similar to the eikonal and energy equations in ray theory. These hyperbolic equations can be integrated along streamlines, with explicit results for various simple geometries.

In these equations the smoothing back reaction of the pressure is decoupled from the problem, leading to the possibility of singular lines being the envelope of crossing streamlines. This crossing of streamlines seems unphysical. However, in reality crossing streamlines do appear to exist: the streamlines cross at different levels.
16.10. Thin-Layer Flow Along a Curved Surface

16.10.1 The Model

Consider along a surface \( \tilde{S}(\tilde{x}) = 0 \) in a gravity field \(-ge\), a thin layer of incompressible, inviscid, irrotational, and stationary flow (water). At the free surface, given by \( \tilde{F}(\tilde{x}) = 0 \) (we write dimensional variables with a tilde), the pressure is equal to the constant atmospheric pressure \( p_0 \).

Inside the layer we have a potential \( \tilde{\Phi} \), a pressure \( \tilde{p} \), and a constant density \( \rho_0 \) satisfying the equation of mass conservation and Bernoulli’s equation, respectively:

\[
\nabla^2 \tilde{\Phi} = 0, \tag{16.238a}
\]
\[
\tilde{p} + \frac{1}{2} \rho_0 |\nabla \tilde{\Phi}|^2 + \rho_0 g \tilde{z} = \text{constant along streamlines}, \tag{16.238b}
\]
with boundary conditions

\[
\nabla \tilde{\Phi} \cdot \nabla \tilde{S} = 0 \quad \text{at} \quad \tilde{S} = 0, \tag{16.238c}
\]
\[
\nabla \tilde{\Phi} \cdot \nabla \tilde{F} = 0 \quad \text{at} \quad \tilde{F} = 0, \tag{16.238d}
\]
\[
\tilde{p} = p_0 \quad \text{at} \quad \tilde{F} = 0. \tag{16.238e}
\]

When we make lengths dimensionless on a typical length \( L \) related to the size of the curved surface, make velocities dimensionless on a typical velocity \( U \), and write the pressure as \( \tilde{p} := p_0 + \rho_0 U^2 p \), we obtain

\[
\nabla^2 \Phi(x, y, z) = 0, \tag{16.239a}
\]
\[
p + \frac{1}{2} |\nabla \Phi|^2 + \gamma z = \text{constant along streamlines}, \tag{16.239b}
\]
with boundary conditions

\[
\nabla \Phi \cdot \nabla S = 0 \quad \text{at} \quad S = 0, \tag{16.239c}
\]
\[
\nabla \Phi \cdot \nabla F = 0 \quad \text{at} \quad F = 0, \tag{16.239d}
\]
\[
p = 0 \quad \text{at} \quad F = 0, \tag{16.239e}
\]

where the inverse of the squared Froude number

\[
\gamma := \frac{gL}{U^2} \tag{16.240}
\]

is supposed to be finite, i.e., \( O(1) \) in terms of the asymptotic expansion to follow.

16.10.2 Curvilinear Coordinates

We assume that we can define a curvilinear orthogonal coordinate system \((\sigma, \tau, \nu)\) given by

\[
x = x(\sigma, \tau, \nu), \quad y = y(\sigma, \tau, \nu), \quad z = z(\sigma, \tau, \nu) \tag{16.241}
\]

and attached to the surface such that \( \nu = 0 \) corresponds to \( S = 0 \) and \( \nu > 0 \) is the wet side. The free surface \( F = 0 \) corresponds to \( \nu = \varepsilon H(\sigma, \tau) \) and \( \varepsilon \) is small. For simplicity we will write

\[
\Phi(x, y, z) = \Phi(\sigma, \tau, \nu). \tag{16.242}
\]
To rewrite the problem in the new coordinates we utilize the calculus of vectors and tensors in curvilinear orthogonal coordinates [15]:

scale factors: \( h_\alpha^2 = \sum_\xi \left( \frac{\partial \xi}{\partial \alpha} \right)^2 \), \( \alpha = \sigma, \tau, \nu, \ \xi = x, y, z; \)

unit vectors: \( e_\alpha = \frac{1}{h_\alpha} \sum_\xi \frac{\partial \xi}{\partial \alpha} e_\xi, \quad e_\beta \cdot \frac{\partial e_\alpha}{\partial \beta} = \sum_\xi \left( \frac{1}{h_\alpha^2} \frac{\partial \xi}{\partial \beta} \frac{\partial \xi}{\partial \alpha} \right); \)

gradient: \( \nabla = \sum_\alpha \frac{e_\alpha}{h_\alpha} \frac{\partial}{\partial \alpha}. \)

This yields for the Laplace equation (16.239a)

\[
\nabla^2 \Phi = \sum_\alpha \frac{1}{h_\alpha} \frac{\partial}{\partial \alpha} \left( \frac{1}{h_\alpha} \frac{\partial \Phi}{\partial \alpha} \right) + \sum_\alpha \left( \frac{1}{h_\alpha} \frac{\partial \Phi}{\partial \alpha} \sum_\beta \left( \frac{e_\beta}{h_\beta} \cdot \frac{\partial e_\alpha}{\partial \beta} \right) \right) = 0. \quad (16.243)
\]

The normal vectors of the surfaces are given by

\[
\nabla S \propto \nabla \nu = \frac{e_\nu}{h_\nu} \quad \text{at} \quad \nu = 0,
\]

\[
\nabla F \propto \nabla (\nu - \varepsilon H) = \frac{e_\nu}{h_\nu} - \varepsilon \frac{e_\sigma}{h_\sigma} \frac{\partial H}{\partial \sigma} - \varepsilon \frac{e_\tau}{h_\tau} \frac{\partial H}{\partial \tau} \quad \text{at} \quad \nu = \varepsilon H.
\]

As the surface \( F = 0 \) consists of streamlines, we may use Bernoulli’s equation, and we obtain from (16.239c) to (16.239e) that

\[
\frac{1}{h_\nu^2} \frac{\partial \Phi}{\partial \nu} = 0 \quad \text{at} \quad \nu = 0, \quad (16.244a)
\]

\[
\frac{1}{h_\nu^2} \frac{\partial \Phi}{\partial \nu} - \frac{\varepsilon}{h_\sigma^2} \frac{\partial H}{\partial \sigma} \frac{\partial \Phi}{\partial \sigma} - \frac{\varepsilon}{h_\tau^2} \frac{\partial H}{\partial \tau} \frac{\partial \Phi}{\partial \tau} = 0 \quad \text{at} \quad \nu = \varepsilon H, \quad (16.244b)
\]

\[
\frac{1}{h_\nu^2} \sum_\alpha \left( \frac{\partial \Phi}{\partial \alpha} \right)^2 + \gamma \nu = \text{constant} \quad \text{at} \quad \nu = \varepsilon H. \quad (16.244c)
\]

### 16.10.3 Perturbation Analysis

For small \( \varepsilon \) we scale the normal coordinate

\[
\nu = \varepsilon \eta \quad (16.245)
\]
16.10. Thin-Layer Flow Along a Curved Surface

and substitute this in (16.243) and (16.244), resulting in

\[
\nabla^2 \Phi = \frac{1}{h_\sigma} \frac{\partial}{\partial \sigma} \left( \frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \sigma} \right) + \frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \sigma} \sum_a \left( \frac{e_a}{h_a} \cdot \frac{\partial e_a}{\partial \sigma} \right) + \frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \tau} \left( \frac{1}{h_\tau} \frac{\partial \Phi}{\partial \tau} \right) + \frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \sigma} \sum_a \left( \frac{e_a}{h_a} \cdot \frac{\partial e_v}{\partial \sigma} \right)
\]

\[
+ \frac{1}{h_\tau} \frac{\partial \Phi}{\partial \sigma} \sum_a \left( \frac{e_a}{h_a} \cdot \frac{\partial e_v}{\partial \sigma} \right) = 0
\]

and the boundary conditions

\[
\frac{\partial \Phi}{\partial \eta} = 0 \quad \text{at} \quad \eta = 0,
\]

\[
\frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \eta} \cdot \varepsilon^2 \left( \frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \sigma} \right) + \frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \tau} \cdot \varepsilon^2 \frac{1}{h_\tau} \frac{\partial \Phi}{\partial \tau} = 0 \quad \text{at} \quad \eta = H(\sigma, \tau),
\]

\[
\frac{1}{2} \left( \frac{1}{h_\sigma} \frac{\partial \Phi}{\partial \sigma} \right)^2 + \frac{1}{h_\tau} \left( \frac{\partial \Phi}{\partial \tau} \right)^2 + \frac{1}{\gamma} \frac{1}{h_v} \frac{\partial \Phi}{\partial \eta} \cdot \gamma z = \text{constant} \quad \text{at} \quad \eta = H(\sigma, \tau).
\]

It seems reasonable to assume that our rescaling of \( \nu \) yields a regular expansion (see Definition 15.19) for \( \Phi \), i.e.,

\[
\Phi(\sigma, \tau, \eta; \varepsilon) = \Phi_0(\sigma, \tau, \eta) + \varepsilon \Phi_1(\sigma, \tau, \eta) + \varepsilon^2 \Phi_2(\sigma, \tau, \eta) + \cdots
\]

if the surface is sufficiently smooth, such that

\[
h_\sigma = \mathcal{O}(1), \quad \frac{\partial}{\partial \alpha} \left( \frac{e_a}{h_a} \right) = \mathcal{O}(1).
\]

We expand

\[
h_\sigma(\sigma, \tau, \varepsilon \eta) = h_\sigma(\sigma, \tau, 0) + \mathcal{O}(\varepsilon), \quad e_a(\sigma, \tau, \varepsilon \eta) = e_a(\sigma, \tau, 0) + \mathcal{O}(\varepsilon).
\]

Substituting these expansions in the equations and boundary conditions and collecting like powers of \( \varepsilon \), we obtain that \( \Phi_0 \) and \( \Phi_1 \) are functions independent of \( \eta \), as follows:

\[
\varepsilon^0: \quad \frac{\partial^2 \Phi_0}{\partial \eta^2} = 0 \quad \text{while} \quad \frac{\partial \Phi_0}{\partial \eta} = 0 \quad \text{at} \quad \eta = 0 \implies \Phi_0 = \phi(\sigma, \tau),
\]

\[
\varepsilon^1: \quad \frac{\partial^2 \Phi_1}{\partial \eta^2} = 0 \quad \text{while} \quad \frac{\partial \Phi_1}{\partial \eta} = 0 \quad \text{at} \quad \eta = 0 \implies \Phi_1 = \phi_1(\sigma, \tau).
\]

The condition of constant pressure is, to leading order,

\[
\frac{1}{2} \left( \frac{1}{h_\sigma} \left( \frac{\partial \Phi}{\partial \sigma} \right)^2 + \frac{1}{h_\tau} \left( \frac{\partial \Phi}{\partial \tau} \right)^2 \right) + \gamma z(\sigma, \tau) = \text{constant}.
\]
To fully determine the zeroth order problem we go on to the $O(\varepsilon^2)$ term. First we have the equation

$$\frac{1}{\varepsilon_0^2} \frac{\partial^2 \Phi_2}{\partial \eta^2} = - \left[ \frac{1}{h_\sigma} \frac{\partial}{\partial \sigma} \left( \frac{1}{h_\sigma} \frac{\partial \Phi_0}{\partial \sigma} \right) + \frac{1}{h_\sigma} \frac{\partial \Phi_0}{\partial \sigma} \sum \left( \frac{\boldsymbol{e}_a \cdot \partial \boldsymbol{e}_a}{h_\sigma} \right) 
+ \frac{1}{h_\tau} \frac{\partial}{\partial \tau} \left( \frac{1}{h_\tau} \frac{\partial \Phi_0}{\partial \tau} \right) + \frac{1}{h_\tau} \frac{\partial \Phi_0}{\partial \tau} \sum \left( \frac{\boldsymbol{e}_a \cdot \partial \boldsymbol{e}_\tau}{h_\tau} \right) \right], \quad (16.250)$$

which can be integrated immediately because $\Phi_0$, $h_\sigma$, and $\boldsymbol{e}_a$ (evaluated at $\nu = 0$) are independent of $\eta$. With the boundary condition $\frac{\partial}{\partial \eta} \Phi_2 = 0$ at $\eta = 0$ we obtain

$$\frac{1}{\varepsilon_0^2} \frac{\partial \Phi_2}{\partial \eta} = \eta \left( \text{right-hand side of (16.250)} \right).$$

We combine this result with the boundary condition for $\Phi_2$ at the free surface $\eta = H$, i.e.,

$$\frac{1}{\varepsilon_0^2} \frac{\partial \Phi_2}{\partial \eta} = \frac{1}{\varepsilon_0^2} \frac{\partial H \Phi_0}{\partial \sigma} + \frac{1}{\varepsilon_0^2} \frac{\partial H \Phi_0}{\partial \tau} \quad \text{at} \quad \eta = H, \quad (16.251)$$

to eliminate $\frac{\partial}{\partial \eta} \Phi_2$ and arrive at the following equation for $H$ and $\phi$:

$$\frac{1}{\varepsilon_0^2} \frac{\partial}{\partial \sigma} \left( H \frac{\partial \phi}{\partial \sigma} \right) + \left( H \frac{\partial \phi}{\partial \sigma} \right) \sum \left( \frac{\boldsymbol{e}_a \cdot \partial \boldsymbol{e}_a}{h_\sigma} \right) 
+ \frac{1}{\varepsilon_0^2} \frac{\partial}{\partial \tau} \left( H \frac{\partial \phi}{\partial \tau} \right) + \left( H \frac{\partial \phi}{\partial \tau} \right) \sum \left( \frac{\boldsymbol{e}_a \cdot \partial \boldsymbol{e}_\tau}{h_\tau} \right) = 0. \quad (16.252)$$

Note that this is just the expected

$$\nabla_S \cdot (H \nabla_S \phi) = 0, \quad (16.253)$$

where subscript $S$ means restricted to the surface $S = 0$. Equations (16.249) and (16.253) are very similar to the eikonal and transport equations of ray theory (Example 15.37, (♯), (♯♯)). Equation (16.249) is of first order hyperbolic type and may be solved by integration along characteristics or (what they denote here) streamlines. We apply Theorem 12.6 to (16.249). Define

$$q_\sigma := \frac{\partial \phi}{\partial \sigma}, \quad q_\tau := \frac{\partial \phi}{\partial \tau}. \quad (16.254)$$

Consider the curve $(\sigma(t), \tau(t))$ parametrized by $t$ (indeed it will appear to be equivalent to
time). We obtain the system of ODEs

\[
\frac{d\sigma}{dt} = \frac{g_\sigma}{h_\sigma^2}, \quad (16.255a)
\]

\[
\frac{d\tau}{dt} = \frac{g_\tau}{h_\tau^2}, \quad (16.255b)
\]

\[
\frac{d\phi}{dt} = \frac{g_\sigma^2}{h_\sigma^2} + \frac{g_\tau^2}{h_\tau^2}, \quad (16.255c)
\]

\[
\frac{dq_\sigma}{dt} = \frac{1}{h_\sigma^3} \left( \frac{\partial h_\sigma}{\partial \sigma} \right) q_\sigma^2 + \frac{1}{h_\tau^3} \left( \frac{\partial h_\tau}{\partial \sigma} \right) q_\tau^2 - \gamma \frac{\partial z}{\partial \sigma}, \quad (16.255d)
\]

\[
\frac{dq_\tau}{dt} = \frac{1}{h_\sigma^3} \left( \frac{\partial h_\sigma}{\partial \tau} \right) q_\sigma^2 + \frac{1}{h_\tau^3} \left( \frac{\partial h_\tau}{\partial \tau} \right) q_\tau^2 - \gamma \frac{\partial z}{\partial \tau}, \quad (16.255e)
\]

with suitable initial (point source) conditions at \( t = 0 \)

\[
\sigma(0) = \sigma^0, \quad \tau(0) = \tau^0, \quad \frac{d\phi}{d\sigma} = \phi_0, \quad \frac{d\phi}{d\tau} = \phi_0^\phi, \quad \phi = \phi^0. \quad (16.255f)
\]

Note that the curve is indeed a streamline: the local velocity of a point following the curve

\[
\frac{dx}{dt} = \frac{d\sigma}{d\sigma} \frac{\partial x}{\partial \sigma} + \frac{d\tau}{d\tau} \frac{\partial x}{\partial \tau} = \frac{d\sigma}{dt} h_\sigma e_\sigma + \frac{d\tau}{dt} h_\tau e_\tau = \frac{e_\sigma}{h_\sigma} \frac{\partial \phi}{\partial \sigma} + \frac{e_\tau}{h_\tau} \frac{\partial \phi}{\partial \tau} = \nabla \phi \quad (16.256)
\]

is just the flow velocity.

The height \( H \) is found from Gauss’s theorem applied to (16.253). Consider a small line segment of length \( \ell_1 \), where \( \phi = c_1 \) (constant), and connect this via streamlines (\( \parallel \nabla \phi \)) to another line segment of length \( \ell_2 \), where \( \phi = c_2 \) (Figure 16.29). We have now (with \( n \) denoting the in-plane normal vector to \( \ell_1, \ell_2 \) and the streamlines, respectively)

\[
\text{at } \ell_1, \ell_2: \quad \nabla \phi \cdot n = |\nabla \phi|,
\]

along streamlines: \( \nabla \phi \cdot n = 0 \),

so that \( H \) is given by

\[
H |\nabla_\phi| \ell = \text{constant.} \quad (16.257)
\]

Figure 16.29. Mass conservation between streamlines.
In the following we will consider examples of a flow field that issues from a point source, so the streamlines may be parametrized by the azimuthal angle $\theta$ and formally given by $\Sigma(\sigma, t, \theta) = 0$ and $T(\tau, t, \theta) = 0$.

As noted before, there is no mechanism in the model that precludes streamlines from crossing. The envelope of crossing streamlines (caustics) is formed by the locus of the point of intersection of two arbitrarily closely spaced streamlines [169]. This implies that $\Sigma$ is stationary in $\sigma$ for small changes in $t$ and $\theta$. In other words, $\Sigma(\sigma, t + \Delta t, \theta + \Delta \theta) \simeq \Delta t \Sigma_t + \Delta \theta \Sigma_\theta = 0$. The same goes for $T$. This leads to the condition

$$\Sigma_\theta/\Sigma_t = T_\theta/T_t.$$  \hfill (16.258)

### 16.10.4 The Inclined Plane Surface

Probably the simplest example, with a completely analytical solution, is that of a plane surface inclined under an angle $\kappa$ (Figure 16.30). The local coordinates

$$x = \sigma \cos \kappa - \nu \sin \kappa, \quad y = \tau, \quad z = \sigma \sin \kappa + \nu \cos \kappa$$ \hfill (16.259)

have unit scale factors, i.e., $h_\sigma = h_\tau = h_\nu = 1$, so we have the equations

$$\frac{\partial}{\partial \sigma} \left( H \frac{\partial \phi}{\partial \sigma} \right) + \frac{\partial}{\partial \tau} \left( H \frac{\partial \phi}{\partial \tau} \right) = 0, \quad \frac{1}{2} \left( \frac{\partial \phi}{\partial \sigma} \right)^2 + \left( \frac{\partial \phi}{\partial \tau} \right)^2 + \gamma \sigma \sin \kappa = \text{constant},$$ \hfill (16.260)

which are in characteristic form described by

$$\frac{d\sigma}{dt} = q_\sigma, \quad \frac{dq_\sigma}{dt} = -\gamma \sin \kappa, \quad \frac{d\tau}{dt} = q_\tau, \quad \frac{dq_\tau}{dt} = 0.$$ \hfill (16.261)

![Figure 16.30. Inclined plane surface.](image-url)
They have the solution

\[ q_\sigma = \phi_\sigma^0 - \gamma t \sin \kappa, \quad q_t = \phi_t^0, \quad \sigma = \sigma^0 + \phi_\sigma^0 t - \frac{1}{2} \gamma t^2 \sin \kappa, \quad \tau = \tau^0 + \phi_t^0 t. \]

The quadratic relation between \( \sigma \) and \( t \) indicates that a streamline may cross a given point \((\sigma, \tau)\) twice, making the solution double valued which may not be realistic (Figure 16.31). For a point source at the origin, i.e., \( \sigma^0 = \tau^0 = 0 \), of strength \( \phi_\sigma^0 = Q \sin \theta \) and \( \phi_t^0 = Q \cos \theta \), this crossing occurs along the parabola-shaped caustic (see Figure 16.31)

\[ \sigma = \frac{\gamma \sin \kappa}{2Q^2} \left( \frac{Q^4}{\gamma^2 \sin^2 \kappa} - \tau^2 \right). \]

In general it is safe to say that as soon as two or more streamlines cross, the present theory, based on the assumption that the pressure in the flow is completely governed by the outside atmospheric pressure, is not valid any more, and some interaction of the confluent flows is to be taken into account. It seems possible, however, for interaction to be avoided and the
flows to cross each other at two different levels, generating a two-layered flow, in which case the crossing streamlines do have a physical interpretation.

In order to allow an expression for the layer thickness $H$, we have to consider a flow without crossing streamlines. For example, parallel flow in the negative $\sigma$ direction, starting at the coordinate line $\sigma = 0$ as $\phi^0 = -U_0$ and $\phi^0 = 0$, is given by

$$\phi = \frac{1}{3\gamma \sin \kappa}(U_0^2 - 2\gamma\sigma \sin \kappa)^{3/2}, \quad \frac{\partial \phi}{\partial \sigma} = -(U_0^2 - 2\gamma\sigma \sin \kappa)^{1/2},$$

$$H = \frac{H_0 U_0}{(U_0^2 - 2\gamma\sigma \sin \kappa)^{1/2}}. \quad (16.264)$$

### 16.10.5 The Inclined Cylinder

Consider the inside of a cylinder of radius $R$, with its axis in the $(x,z)$ plane making an angle $\kappa$ with the $x$ axis (Figure 16.32). (Assume for simplicity that $\cos \kappa > 0$.)

The coordinate transformation

$$x = \sigma \cos \kappa + (R - \nu) \sin \kappa \cos \tau,$$

$$y = (R - \nu) \sin \tau, \quad z = \sigma \sin \kappa - (R - \nu) \cos \kappa \cos \tau \quad (16.265)$$

has the Jacobi matrix

$$\begin{pmatrix}
\frac{\partial x}{\partial \sigma} & \frac{\partial y}{\partial \sigma} & \frac{\partial z}{\partial \sigma} \\
\frac{\partial x}{\partial \tau} & \frac{\partial y}{\partial \tau} & \frac{\partial z}{\partial \tau} \\
\frac{\partial x}{\partial \nu} & \frac{\partial y}{\partial \nu} & \frac{\partial z}{\partial \nu}
\end{pmatrix} = \begin{pmatrix}
\cos \kappa & 0 & \sin \kappa \\
-(R - \nu) \sin \kappa \sin \tau & (R - \nu) \cos \tau & (R - \nu) \cos \kappa \sin \tau \\
-\sin \kappa \cos \tau & -\sin \tau & \cos \kappa \cos \tau
\end{pmatrix}.$$
16.10. Thin-Layer Flow Along a Curved Surface

This yields

\[ h_\sigma = 1, \quad h_\tau = R - \nu, \quad h_\nu = 1, \]
\[ e_\sigma = e_x \cos \kappa + e_z \sin \kappa, \]
\[ e_\tau = -e_x \sin \kappa \sin \tau + e_y \cos \tau + e_z \cos \kappa \cos \tau, \]
\[ e_\nu = -e_x \sin \kappa \cos \tau - e_y \sin \tau + e_z \cos \kappa \cos \tau, \]
\[ \sum_{\alpha} \left( \frac{e_\alpha}{h_\alpha} \cdot \frac{\partial e_\sigma}{\partial \alpha} \right) = 0, \quad \sum_{\alpha} \left( \frac{e_\alpha}{h_\alpha} \cdot \frac{\partial e_\tau}{\partial \alpha} \right) = 0, \quad \sum_{\alpha} \left( \frac{e_\alpha}{h_\alpha} \cdot \frac{\partial e_\nu}{\partial \alpha} \right) = \frac{-1}{R - \nu} \]

such that

\[ \frac{1}{2} \left( \frac{\partial \phi}{\partial \sigma} \right)^2 + \frac{1}{R^2} \left( \frac{\partial \phi}{\partial \tau} \right)^2 + \gamma (\sigma \sin \kappa - R \cos \kappa \cos \tau) = \text{constant} \] (16.266)

\[ \frac{\partial}{\partial \sigma} \left( H \frac{\partial \phi}{\partial \sigma} \right) + \frac{1}{R^2} \frac{\partial}{\partial \tau} \left( H \frac{\partial \phi}{\partial \tau} \right) = 0. \] (16.267)

This becomes in characteristic form

\[ \frac{d\sigma}{dt} = q_\sigma, \quad \frac{dq_\sigma}{dt} = -\gamma \sin \kappa, \quad \frac{d\tau}{dt} = q_\tau, \quad \frac{dq_\tau}{dt} = -\gamma R \cos \kappa \sin \tau. \] (16.268)

The \( \sigma \) direction has solution

\[ q_\sigma = q_\sigma^0 - \gamma t \sin \kappa, \quad \sigma = \sigma^0 + q_\sigma^0 t - \frac{1}{2} \gamma^2 \nu^2 \sin \kappa. \] (16.269)

The \( \tau \) direction may be combined into the oscillator equation

\[ \frac{d^2 \tau}{dt^2} + \Omega^2 \sin \tau = 0 \quad \text{with frequency} \quad \Omega = \left( \frac{\gamma R \cos \kappa}{\nu} \right)^{1/2}. \] (16.270)

This equation has the energy integral

\[ \frac{1}{2} \left( \frac{d\tau}{dt} \right)^2 - \Omega^2 \cos \tau = E_\tau = \frac{1}{2} \left( \frac{q_\tau^0}{R^2} \right)^2 - \Omega^2 \cos \tau^0. \] (16.271)

We have a left- or right-spiraling flow if the total energy is high enough, i.e., \( E_\tau > \Omega^2 \):

\[ \Omega t = \pm \sqrt{\frac{2}{1 + E_\tau / \Omega^2}} F \left( \frac{1}{2}, \frac{1}{2} \right) \sqrt{1 + E_\tau / \Omega^2} + \text{constant}. \] (16.272)

We have an oscillatory flow if \( -\Omega^2 < E_\tau < \Omega^2 \):

\[ \Omega t = \pm F \left( \arcsin \frac{2 \sin^2 \left( \frac{1}{2} \tau \right)}{1 + E_\tau / \Omega^2}, \frac{\sqrt{1 + E_\tau / \Omega^2}}{\sqrt{2}} \right) + \text{constant}. \] (16.273)
Figure 16.33. Streamlines along an inclined cylinder issuing from a point source in the wall. Right: analytical solution; left: experiments by one of the authors (SWR). Note the realistic but (strictly speaking) not modeled crossing streamlines!

$F(\varphi, k)$ denotes the elliptic integral of the first kind [1]. In the case of the oscillatory solution the sign of $t$ changes each time $\tau$ reaches its maximum or minimum, and the solution continues on the other branch. This is clearly illustrated by the linearized version of the problem (small $\tau$), for which the explicit solution is given by

$$\tau = t^0 \cos \Omega t + \frac{q^0}{R^2 \Omega} \sin \Omega t. \quad (16.274)$$

Again, the situation of crossing streamlines requires special consideration.

An illustration of this oscillatory motion is given in Figure 16.33 (right), to be compared with experimental results (left) from one of the authors (SWR). The qualitative agreement is remarkable.

16.10.6 Discussion and Related Problems

- If inertial effects due to the geometry of the basin become negligible, like the flow in a canal, the Froude number $U/\sqrt{gL}$ should be based on the water depth rather than $L$. In the present limit this would result in a high Froude number flow, which is known as “supercritical.” This Froude number turns out to be the ratio between flow speed
and the speed of the fastest surface waves on the fluid. This means that supercritical flow is faster than the waves: information about downstream disturbances cannot propagate upstream.

- The related problem of thin-layer flow along a horizontal plane is described by the shallow-water equation [150] (see Example 15.25 or Section 12.12.5).

- Wall friction is left out of the present (asymptotic) model. It is possible, however, to include some form of friction (in an ad hoc way) along the streamlines (characteristics).

- The issue of the crossing streamlines is far from resolved. Various mechanisms and competing effects may be proposed.

- Other problems described by similar nonlinear first order hyperbolic equations are ray propagation (Section 16.9) and etching of glass (Section 16.2).

16.10.7 Exercises

16.41. Do a similar analysis as above for the sphere given by

\[ x = (R - \nu) \cos \tau \cos \sigma, \quad y = (R - \nu) \sin \tau \cos \sigma, \quad z = (R - \nu) \sin \sigma. \]

16.42. The present formulation is based on orthogonal coordinates along the surface. This is convenient but not necessary. Try to rewrite the present formulas in a slightly more general way with nonorthogonal curvilinear coordinates \((\sigma, \tau)\) along the surface [143].

16.43. Propose wall friction laws that may be included in the model.

16.44. Derive (16.258).


16.46. Derive the linearised result (16.274).

16.47. Rewrite (16.257) by means of a Jacobian in a form similar to (16.234).

16.11 Forming Container Glass

16.11.1 Problem Formulation

For many years, glass technology based on empirical knowledge and the expertise of craftspeople was sufficient to ensure competitiveness. However, over the last 20 years mathematical modeling of various aspects of glass production has become increasingly important. Two thirds of the glass produced is used for packaging in the form of jars and bottles, one
quarter of the production is in the form of float glass used for panes, and the rest is for
specialist products such as cathode ray tubes (CRTs) and fibers.

Since production costs are dominated by the cost of melting raw material, ever more
stringent rules regarding the weight and thickness of container glass are being imposed for
its production. As a result, modeling has become an essential tool to meet these demands.
In the glass production process, sand grains and additives, such as soda, are heated in a
tank, which can be several tens of meters long and a few meters high and wide. Gas burners
or electric heaters are used to raise the temperature of the material to about 1200°C. The
liquid glass emerges at one end of the tank and is fed to pressing or blowing machines, or
onto a bed of liquid tin, where it spreads out to become float glass. In this section we will
consider the problem of pressing liquid glass in a mould. Often the forming is done in two
stages. After the preform (parison) has been pressed, it is blown or further pressed into its
final shape (bottle, jar, etc.). At each stage the (pre)form is cooled through the walls of the
mould.

16.11.2 Governing Equations

The motion of glass at temperatures above 600°C can be described by the Navier–Stokes
equations for incompressible Newtonian fluids [77], which are given by

\[
\rho \left( \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) = -\nabla p + \mu \nabla^2 \mathbf{v}, \quad \nabla \cdot \mathbf{v} = 0, \tag{16.275}
\]

where \( \rho \), \( \mathbf{v} \), and \( p \) denote density, velocity, and pressure; see (7.3). This expression is
valid for a constant viscosity \( \mu \), which is to be verified first. Since the viscosity depends
significantly on the temperature [157, 10], we have to make sure that the temperature is
constant in the problem. This is not immediately evident. The walls differ in temperature
from the glass and may supply or absorb heat by conduction, while the high viscous forces
may generate heat by friction. Also, heat transport by radiation may play a role.

To make progress we will exclude any significant effect of radiation by assuming
that the walls are highly reflective. Of course, this may vary in other cases [80]. As for
the other effects, a suitable model, describing heat conduction according to Fourier’s law,
heat convection by the incompressible glass flow, and heat generation by viscous friction,
is given by

\[
\rho C_p \left( \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \right) = \kappa \nabla^2 T + 2\mu (\mathbf{D} : \mathbf{D}) \tag{16.276}
\]

(see (6.45h) with relations (6.46b) and (6.46c)). The equations are valid in the region
between the fixed mould and the axially moving plunger that contains the glass (see Fig-
ure 16.34). In view of the geometry of plunger and mould, we choose cylindrical coordinates
(\( r, \phi, z \)), while \( v, w, u \) will denote the \( r, \phi, z \) components of the velocity \( \mathbf{v} \). We assume the
problem to be axisymmetric, so that both \( w \) and all \( \phi \) derivatives vanish and the problem
reduces to a two-dimensional one in the \( (r, z) \) plane and \( \mathbf{v} = ve_r + ue_z \).

\[
r = R_m(z) \text{ is the mould surface, } z = z_p(t) \text{ is the varying position of the plunger top,}
\]

\[
r = R_p(z - z_p) \text{ is the plunger surface, and } z = b(t) \text{ is the position of the glass surface. As}
\]

the plunger moves down, the decreasing volume between mould and plunger presses the
glass upward, so the glass level \( b \) increases in time. If we define \( R_p(z) = 0 \) if \( z \leq 0 \), the
region of definition of our problem is \( 0 \leq z \leq b(t) \) and \( R_p \leq r \leq R_m \).
16.11. Forming Container Glass

Figure 16.34. Sketch of configuration. $R_p = R_p(z - z_p)$ is the plunger surface, $R_m = R_m(z)$ is the mould surface, $z_p(t)$ is the position of the plunger top, and $b(t)$ is the position of the glass surface.

16.11.3 Slender Geometry Approximation

We will concentrate our analysis on the flow in the narrow annular duct between plunger and mould ($z > z_p$; see Figure 16.34). This region is very slender and therefore amenable to asymptotic analysis (the method of slow variation; see Section 15.3.1), while at the same time the flow in the region between mould bottom and plunger top is practically stagnant and therefore less important.

We will make (16.275) dimensionless by a suitable scaling. From the geometry of plunger and mould we have two relevant length scales: the characteristic thickness of the parison ($D$) and the length of the plunger ($L$), with $D \ll L$. Except right near the plunger top, any variation in $z$ scales on $L$ and any variation in $r$ scales on $D$. Therefore we scale $z$ with $L$ and $r$ with $D$ and we introduce the small parameter

$$\varepsilon := \frac{D}{L}. \quad (16.277)$$

The axial velocity $u$ scales on a typical velocity $V$, while from the equation of mass conservation it follows that the radial velocity $v$ has to scale on $\varepsilon V$. As density $\rho$ and viscosity $\mu$ are constant, they are parameters of the problem. Pressure $p$ is to be scaled on $\mu V L / D^2$.
Table 16.4. Typical values of the parison problem parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>typical parison length scale</td>
</tr>
<tr>
<td>$\rho$</td>
<td>glass density</td>
</tr>
<tr>
<td>$V$</td>
<td>typical plunger velocity</td>
</tr>
<tr>
<td>$\mu_g$</td>
<td>glass dynamic viscosity</td>
</tr>
<tr>
<td>$L$</td>
<td>typical plunger length scale</td>
</tr>
<tr>
<td>$T_g$</td>
<td>glass temperature</td>
</tr>
<tr>
<td>$C_p$</td>
<td>glass heat capacity</td>
</tr>
<tr>
<td>$T_m$</td>
<td>mould temperature</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>glass thermal conductivity</td>
</tr>
</tbody>
</table>

(rather than $\rho V^2$) because the glass flow is highly viscous, as we will see below. We have then

$$z := Dz^*, \quad Z := \varepsilon z^*, \quad r := Dr^*, \quad u := Vu^*, \quad v = \varepsilon Vv^*, \quad p = \frac{\mu VL}{D^2} p^*, \quad t = \frac{L}{V} t^*.$$  \hspace{1cm} (16.278)

Note that in order to maintain a transparent bookkeeping of the equations, it is advisable not to scale $z$ on $L$ directly, but via a slow variable $Z = \varepsilon z^*$. Now we substitute the above scalings in (16.275) and henceforth ignore the asterisks * to obtain the dimensionless equations

$$\varepsilon Re \left( \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial Z} + v \frac{\partial u}{\partial r} \right) = - \frac{\partial p}{\partial Z} + \varepsilon^2 \frac{\partial^2 u}{\partial Z^2} + \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right),$$  \hspace{1cm} (16.279a)

$$\varepsilon^3 Re \left( \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial Z} + v \frac{\partial v}{\partial r} \right) = - \frac{\partial p}{\partial r} + \varepsilon^4 \frac{\partial^2 v}{\partial Z^2} + \varepsilon^2 \frac{\partial}{\partial r} \left( \frac{1}{r} \frac{\partial}{\partial r} (rv) \right),$$  \hspace{1cm} (16.279b)

$$\frac{\partial u}{\partial Z} + \frac{1}{r} \frac{\partial}{\partial r} (rv) = 0,$$  \hspace{1cm} (16.279c)

where $Re = \rho V D/\mu$ is the Reynolds number. Typical values of the parameters (Table 16.4) yield $\varepsilon = 0.1$, $Re = 2.5 \times 10^{-4}$, $\varepsilon Re = 2.5 \times 10^{-5}$, and $\varepsilon^3 Re = 2.5 \times 10^{-7}$, and we can ignore the inertia and radial friction terms to obtain

$$\frac{\partial p}{\partial Z} + \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right), \quad \frac{\partial p}{\partial r} = 0, \quad \frac{\partial u}{\partial Z} + \frac{1}{r} \frac{\partial}{\partial r} (rv) = 0,$$  \hspace{1cm} (16.280)

with an assumed error of $O(\varepsilon^2)$ because $Re < \varepsilon$. This set of equations may be recognized as Reynolds’s lubrication flow equations in cylindrical coordinates [47, p. 83].

### 16.11.4 Constant Temperature

To investigate the assumption of a constant viscosity we will analyse the energy equation and estimate the order of magnitude of the various terms. In addition to (16.278) we introduce...
scaled temperature and viscosity

\[ T = T_m + \Delta T T^*, \quad \mu = \mu_g \mu^*, \]  

(16.281)

where \( \Delta T = T_g - T_m \) and \( \mu_g \) denotes the glass viscosity at the bulk temperature. (Note that in the rest of this section we will use a constant viscosity \( \mu = \mu_g \).) We find (ignoring the asterisks \(*\)) that

\[
\varepsilon \left[ \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \right] = \frac{1}{Pe} \left[ \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T}{\partial r} \right) + \varepsilon \frac{\partial^2 T}{\partial z^2} \right] \\
+ \frac{Ec}{Re} \left[ \left( \frac{\partial u}{\partial r} + \varepsilon \frac{\partial v}{\partial z} \right)^2 + 2\varepsilon^2 \left( \left( \frac{\partial v}{\partial r} \right)^2 + \left( \frac{v}{r} \right)^2 + \left( \frac{\partial u}{\partial z} \right)^2 \right) \right],
\]

(16.282)

where the Reynolds number is \( Re = \rho V D/\mu_g \), the Eckert number is \( Ec = V^2/C_p/\Delta T \), and the Péclet number is \( Pe = \rho V D C p/\kappa \). Note that \( Pe \) and \( Re \) are related through the Prandtl number as \( Pe = Re Pr \), while \( Pe Ec/Re = Br \) is called the Brinkman number. When we substitute the values given in Table 16.4, typical of glass at 800°C, we get the corresponding dimensionless numbers

\[
\varepsilon = 10^{-1}, \quad \frac{1}{Pe} = 6.2 \times 10^{-4}, \quad \frac{Ec}{Re} = 1.2 \times 10^{-4}.
\]

(16.283)

Both \( 1/\varepsilon Pe \) and \( Ec/\varepsilon Re \) are very small numbers, so we can ignore heat conduction and thermal production terms, the second and third terms in (16.282), in the bulk of the flow against the convection (first) term. Hence the energy equation simplifies to

\[ \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T = \frac{dT}{dr} = 0, \]

(16.284)

indicating that the temperature is preserved following the flow. So if we start with a uniform temperature field, it will remain uniform everywhere, and it follows that the viscosity also remains constant. Only very close to the wall is this not true. Here the temperature varies from the value at the wall to the bulk temperature, and the conduction term is comparable with the convection term. We will consider this thermal boundary layer small enough to be ignored.

### 16.11.5 Boundary Conditions

At the walls of the plunger and the mould we have to apply boundary conditions. First, we introduce suitably scaled plunger and mould surface functions. In order to make their slowly varying behaviour in the \( z \) direction explicit, we write

\[
R_p(z) = DR_p^* (z/L) = DR_p^* (\varepsilon z^*) = DR_p^*(Z);
\]

(16.285)

\( R_m \) is written likewise. (The asterisks \(*\) will be ignored from now on.) Subscript \( p \) and \( m \) will denote the values at the plunger and mould, respectively. We recall that the moving plunger surface is defined by

\[ r = R_p(Z - Z_p(t)), \]
where \( Z = Z_p(t) = \varepsilon Z_p(t) \) denotes the position of the top of the plunger at time \( t \). (Unless indicated otherwise, we will write here \( R_p = R_p(Z - Z_p) \).) The velocity of a point moving with the plunger is then \( v_p = \dot{z}_p \varepsilon \).

We now return to the physics of the flowing glass. Depending on problem parameters like wall temperature, fluid pressure, surface tension, or the presence of a lubricant like graphite powder \([34, 160, 161, 40]\), the glass flow will slip completely, partly, or not at all along the wall. This means that the tangential component of the glass velocity \( v \) at the wall may differ from the wall velocity \( v_w \), where the difference is called the slip velocity. For a complete model this should be included \([130]\), but for now we will simply assume no slip. In other words, \( v \) is just equal to the wall velocity and we have

\[
\begin{align*}
  u &= u_p, \quad v = 0 \quad \text{at} \quad r = R_p(Z - Z_p), \\
  u &= 0, \quad v = 0 \quad \text{at} \quad r = R_m(Z).
\end{align*}
\]  
(16.286)

As the blob of glass does not initially fill the mould completely (see Figure 16.34), there is a free surface of glass moving into the annular duct between mould and plunger. At the free boundary the normal stress must be equal to the external pressure \( p_0 \), which is assumed to be constant. In the present incompressible model it is no restriction to take \( p_0 = 0 \). In our slender geometry approximation the exact shape of the free surface cannot be determined. In its neighbourhood the flow scales in both the \( r \) and \( Z \) directions on the thickness \( D \) of the annular channel. In other words, the flow is not slowly varying in \( Z \) any more, and the present approximation is not valid.

Within the present approximation (and assumption of circular symmetry), we will deal with the average level \( b \) of the free surface as follows:

\[
p = 0 \quad \text{at} \quad Z = b(t),
\]  
(16.287)

where \( b \) (scaled on \( L \)) is a function of the time-dependent geometry, implicitly defined in such a way that the (incompressible) glass volume between \( Z = 0 \) and \( Z = b(t) \) is constant for all \( t \).

### 16.11.6 Solution

We start by deriving an expression for the axial flux. Consider the flux at some level \( z \) through cross section \( S_1 \) (see Figure 16.35). The value of this flux depends on \( z \), since the plunger goes down and causes the glass to move upward through a varying cross section. We will use this value later to find the pressure gradient.

Since glass is an incompressible fluid, we have with Gauss’s divergence theorem that

\[
0 = \int_\Omega \nabla \cdot \mathbf{v} \, dV = \int_{\partial\Omega} \mathbf{v} \cdot \mathbf{n} \, dS = \int_{S_1} \mathbf{v} \cdot \mathbf{n} \, dS + \int_{S_2} \mathbf{v} \cdot \mathbf{n} \, dS + \int_{S_3} \mathbf{v} \cdot \mathbf{n} \, dS,
\]  
(16.288)

where \( \mathbf{v} = u \mathbf{e}_z + \varepsilon v \mathbf{e}_r \). Since the mould is stationary and impermeable, we have a zero flux through \( S_2 \). For the flux through \( S_1 \) we note that it would make no difference for the amount of glass displaced by the plunger if the plunger were filled with glass, because glass is just as incompressible as the solid plunger. Therefore, instead of control surface \( S_1 \), we can just
as well use surface $S_4$, which yields more easily the result

$$\int_{S_3} v \cdot n \, dS = \int_{S_4} v \cdot n \, dS = \pi u_p R_p^2.$$

It follows that the axial flux, i.e., through $S_1$, is given by

$$\int_{S_1} v \cdot n \, dS = 2\pi \int_{R_p}^{R_m} r u(r, z) \, dr = -\pi u_p R_p^2. \quad (16.289)$$

Now we are ready to solve (16.280) with boundary conditions (16.286). At first, we note that $p$ is a function of $z$ only. Then the equation

$$\frac{dp}{dZ} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right)$$

can be integrated to

$$u = \frac{1}{4} r^2 \frac{dp}{dZ} + A(Z) \ln(r) + B(Z).$$

Using boundary conditions (16.286), we obtain

$$u = \frac{1}{4} \frac{dp}{dZ} \left[ r^2 - \frac{R_m^2 \ln(r/R_p) + R_p^2 \ln(R_m/r)}{\ln(R_m/R_p)} \right] + u_p \frac{\ln(R_m/r)}{\ln(R_m/R_p)}. \quad (16.290)$$
We have the following integral for \( ru \)

\[
\int ru(r, Z) \, dr = \frac{1}{8} r^2 \frac{d}{dZ} \left( \frac{1}{2} r^2 - \frac{R_m^2 \ln(r/R_p) + R_p^2 \ln(R_m/r)}{\ln(R_m/R_p)} + \frac{R_m^2 - R_p^2}{2 \ln(R_m/R_p)} \right) + \frac{1}{2} r^2 u_p \ln(R_m/r) + \frac{1}{2} \frac{d}{dZ} \ln(R_m/R_p). \tag{16.291}
\]

With (16.291) and expression (16.289), we can find the pressure gradient

\[
\frac{dP}{dZ} = \frac{4u_p}{(R_m^2 + R_p^2) \ln(R_m/R_p) - (R_m^2 - R_p^2)}. \tag{16.292}
\]

Finally, the continuity equation in (16.280) may be integrated to yield the expression

\[
v = \frac{1}{r} \frac{d}{dZ} \int_r^{R_m} ru(r, Z) \, dr, \tag{16.293}
\]

which includes, with expression (16.291), an explicit, albeit complicated, expression for \( v \).

### 16.11.7 Example

An example of the pressure gradient (16.292) is plotted in Figure 16.36 for mould and plunger defined by

\[
R_m(Z) = \sqrt{\frac{2\sqrt{Z} + 2.56Z^4}{1 + 4Z - 6.4Z^2 + 2.56Z^4}}, \quad R_p(Z) = 1.2 \sqrt{\frac{Z + Z^4}{1 + 4Z - 6.4Z^2 + 2.56Z^4}},
\]

while \( \varepsilon = 0.2 \). The plunger enters the mould from right to left with (dimensionless) velocity \( u_p = -1 \). The scaled plunger top position \( Z_p \) runs through the values 0.9, \ldots, 0.1 with steps of −0.1.

![Figure 16.36](image_url)

**Figure 16.36.** Dimensionless position (left) and minus pressure gradient (right) corresponding to a plunger entering a mould fully filled with fluid glass.
16.11.8 Discussion and Related Problems

- A major simplification in the present analysis is the adopted no-slip boundary condition. This has, however, no bearing on the approximation applied, and it is indeed possible to include slip conditions (see [130]).

- Details of the flow near the free surface \( z = h(t) \) are known to play a role in certain production faults (local cracks or fractures). The present approach is not able to describe these details and a numerical full three-dimensional simulation is typically our next tool of choice [78, 79].

- In practice the plunger is pushed by a piston. This means that we may assume that it is moving by virtue of a constant force or a constant power (force \( \times \) velocity). As this force is counteracted by the viscous forces in the glass, we need to couple this motion to the glass flow.

- The cooling of the parison becomes important during the so-called dwell. This is the time period after the pressing has come to a stop, while the mould is still closed. In general the heat exchange is a very complicated matter, as the heat flux also comes from radiation. For some further reading about this see, e.g., [80]

16.11.9 Exercises


16.49. Explain why the pressure gradient in Figure 16.36 grows so quickly when the distance between the mould and plunger wall becomes small. (Hint: Write \( R_m = R_p + \delta \) and expand (16.292) for small \( \delta \).)

16.50. A practically important quantity is the total force \( F_{tot}(t) \) on the plunger (see Section 16.11.8 above). Write this as a single integral, consistent with the present approximation. Show that its time integral

\[
\int_{t_0}^{t_1} F_{tot}(t) \, dt
\]

only depends on the end positions \( z_p(t_0) \) and \( z_p(t_1) \) and is independent of the way \( z_p \) actually varies in time.

16.12 Laser Percussion Drilling

16.12.1 Problem Formulation

Lasers are often used to machine materials when conventional methods fail. Laser percussion drilling is used, e.g., to drill cooling holes in fans, that are part of a gas turbine;
such fans are typically made of super alloys. The term “percussion” refers to the repeated operation of the laser in short pulses ($10^{-3}$ s), separated by longer time periods ($10^{-2}$ s). The energy supplied by the laser is bounded, and pulsewise behaviour allows for large bursts of energy. These energy bursts cause local melting and splashing of material and sometimes (unwanted) local recasting. A laser percussion drilling process may in fact be split into three stages. First, a thin region of molten material is formed by absorption of laser energy at the target surface. After some time, the surface of this melt pool reaches vaporisation temperature. The second stage is then the squirting out of the molten material, caused by the sudden expansion of the vapor evaporating from the surface. On its way out some part of this molten material may resolidify at the walls, which is the third stage. For each of these three stages a melting model is needed. These events are depicted in Figure 16.37.

In this section we will only consider the first stage, studying the free boundary evolution between the solid and the liquid material.

16.12.2 The Model

The lasers used in practice to drill holes typically produce a Gaussian intensity distribution, which is, ideally, axisymmetric. So let us take an axisymmetric coordinate system, where $z = 0$ denotes the surface of the irradiated material. The density $\rho$, the specific heat $c$, and the thermal conductivity $\kappa$ of the material are known and assumed to be constant. The temperature $T = T(r, z, t)$ in both the solid and the liquid material is governed by the heat equation. In cylindrical coordinates, and assuming axisymmetry, it is given by

$$
\rho c \frac{\partial T}{\partial t} = \kappa \frac{\partial}{\partial r} \left( r \frac{\partial T}{\partial r} \right) + \kappa \frac{\partial^2 T}{\partial z^2}.
$$

(16.294)
Table 16.5. Parameters for the laser drilling process applied to aluminium using an Nd:YAG-laser.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>$2.7 \times 10^3$</td>
<td>kg/m$^3$</td>
<td>$L_f$</td>
<td>$3.6 \times 10^5$</td>
<td>J/kg</td>
</tr>
<tr>
<td>$c$</td>
<td>$9.0 \times 10^2$</td>
<td>J/(kg·K)</td>
<td>$w_0$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>m</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>$2.3 \times 10^2$</td>
<td>J/(m·s·K)</td>
<td>$I_{ref}$</td>
<td>$1.5 \times 10^{10}$</td>
<td>J/(m$^2$·s)</td>
</tr>
<tr>
<td>$T_m$</td>
<td>$9.3 \times 10^5$</td>
<td>K</td>
<td>$\varepsilon$</td>
<td>$6.4 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>$T_v$</td>
<td>$2.5 \times 10^3$</td>
<td>K</td>
<td>$\lambda_f$</td>
<td>$0.25$</td>
<td></td>
</tr>
</tbody>
</table>

The intensity distribution of the laser beam is denoted as $I = I(r, t)$. The laser energy is supplied at the surface $z = 0$, yielding the boundary condition

$$\kappa \frac{\partial T}{\partial z}(r, 0, t) = -I(r, t).$$  \hspace{1cm} (16.295)

The model has to be completed with conditions at the solid-liquid interface. This will be discussed in the next section.

To make the model dimensionless we introduce some scaling parameters. Let us consider the melting problem for aluminium, which is good enough to test our numerical method. Typical parameters are given in Table 16.5. For the temperature we need the vaporisation temperature $T_v$ and the melting temperature $T_m$. For the radial coordinate we use the waist, denoted by $w_0$, of the (Gaussian) laser beam to define a length scale. Furthermore, let $I_{ref}$ be a typical intensity, say its maximum value. From this we can define the following dimensionless variables (indicated by an asterisk $*$):

$$z =: L z^*, \quad r =: w_0 r^*, \quad t =: \tau t^*, \quad T =: T_m + (T_v - T_m) T^*, \hspace{1cm} (16.296a)$$

where $L$ is a typical length scale in the $z$ direction and $\tau$ is the diffusive time scale, defined by

$$L := \kappa \frac{T_v - T_m}{I_{ref}}, \quad \tau := \frac{\rho c}{\kappa} L^2, \hspace{1cm} (16.296b)$$

respectively. The length scale $L$ follows from balancing the two terms in the boundary condition (16.295). Applying this scaling to (16.294) and (16.295), we obtain (omitting the asterisk $*$ for ease of notation)

$$\frac{\partial T}{\partial t} = \frac{\varepsilon r}{r} \frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2}, \hspace{1cm} (16.297)$$

$$\frac{\partial T}{\partial z}(r, 0, t) = -\frac{J}{\tau}, \hspace{1cm} (16.298)$$

where the parameter $\varepsilon$ and the function $J = J(r, t)$ are defined by

$$\varepsilon := \left( \frac{L}{w_0} \right)^2 = \left( \frac{k(T_v - T_m)}{w_0 I_{ref}} \right)^2, \quad J := \frac{I}{I_{ref}}, \hspace{1cm} (16.299)$$

16.12. Laser Percussion Drilling
respectively. For a typical laser percussion drilling process, the parameter $\varepsilon$ is much smaller than one; see Table 16.5. Thus radial diffusion can be neglected and the initial stage of the laser percussion drilling process can in fact be modeled as a one-dimensional problem.

### 16.12.3 The Stefan Problem

Clearly, the problem at hand is a Stefan problem; see Section 10.5. Let $s(t)$ denote the position of the solid-liquid interface, as sketched in Figure 16.38. Let $\Omega_\ell$ denote the liquid region, $0 \leq z < s(t)$, and $\Omega_s$ the solid region, $s(t) < z < \infty$. The temperature in the liquid and solid regions is governed by the (dimensionless) one-dimensional heat equation

$$\frac{\partial T_p}{\partial t} = \frac{\partial^2 T_p}{\partial z^2}, \quad z \in \Omega_p \quad (p = \ell, s). \tag{16.300}$$

Here the subscripts $\ell$ and $s$ refer to the liquid and solid regions, respectively. At the boundary $z = 0$ where the laser hits the surface we have the boundary condition

$$\frac{\partial T_\ell}{\partial z}(0, t) = -J. \tag{16.301a}$$

The extra heat to melt the material, in addition to the amount needed to raise the temperature, is called the latent heat of fusion $L_f$. In dimensionless form it is defined as

$$\lambda_f := \frac{L_f}{c(T_v - T_m)}. \tag{16.301b}$$

A typical value of $\lambda_f$ for aluminium can be found in Table 16.5. At the liquid-solid interface we need an additional condition that expresses the absorption of heat needed for phase change. So the Stefan condition at $z = s(t)$ is given by (cf. (11.151g))

$$\frac{\partial T_\ell}{\partial z}(s(t)+, t) - \frac{\partial T_\ell}{\partial z}(s(t)-, t) = \lambda_f \frac{ds}{dt}. \tag{16.301c}$$

Moreover, since the temperature is assumed to be continuous across the interface, we have

$$T_\ell(s(t)+, t) = T_\ell(s(t)-, t) = 0. \tag{16.301d}$$
We assume the domain to be large enough so that the solid material far away from the interface has the same temperature as its surrounding, \( T_{\text{amb}} \) say. Hence for some “cutoff” point \( z_b \) we have

\[
T_s(z_b, t) = T_{\text{amb}}. \tag{16.301e}
\]

To complete our model we prescribe a known initial temperature distribution

\[
T(z, 0) = T_0(z). \tag{16.301f}
\]

To obtain a numerical solution we proceed as in Section 11.10. However, contrary to the approach in that section, we now choose the number of grid points on each of the two regions fixed, say, on the domains \( 0 \leq z \leq s(t) \) and \( s(t) \leq z \leq z_b \), so we have \( N_\ell \) and \( N_s \) equispaced points, respectively. So in the liquid domain we define

\[
z_{\ell,j} := j/D_{\ell}, \quad j = 0, 1, \ldots, N_\ell \quad \text{with} \quad D_{\ell} = D_{\ell}(t) := s(t)/N_\ell, \tag{16.302a}
\]

and likewise for the solid we have

\[
z_{s,j} := s(t) + j/D_{s}, \quad j = 0, 1, \ldots, N_s \quad \text{with} \quad D_{s} = D_{s}(t) := (z_b - s(t))/N_s. \tag{16.302b}
\]

Applying the implicit Euler scheme for time integration and central differences for space discretisation, we have

\[
T_n^{p+1} - \frac{\Delta t}{D_{p}} \left( T_{n+1}^{p+1} - 2T_n^{p+1} + T_{n-1}^{p+1} \right) = T_n^{p}, \quad p = \ell, s. \tag{16.303}
\]

We provide these equations with the boundary conditions

\[
T_{\ell,N_\ell}^n = T_{s,0}^n = 0, \quad T_{s,N_s}^n = T_{\text{amb}}.
\]

and a discretised form of (16.301a); see Section 11.10. So we find an approximate temperature at time \( t^{n+1} \). Next we have to update the interface position and recompute the grids in both domains. Let \( s^n \) be the numerical approximation of \( s(t^n) \). Then, using the explicit Euler scheme and one-sided differences for the Stefan condition in (16.301c) and applying the boundary conditions above, we obtain

\[
s^{n+1} = s^n + \frac{\Delta t}{\lambda_f} \left( \frac{1}{D_{\ell}} T_{\ell,N_\ell-1}^n + \frac{1}{D_{s}} T_{s,1}^n \right), \tag{16.304}
\]

where \( D_{p} = D_{p}(t^n) \) (\( p = \ell, s \)). One should note that the time step restriction \( \Delta t = \mathcal{O}(D_p^2) \) (\( p = \ell, s \)) implies that the grid size at the next time level \( t^{n+1} \) is only \( \mathcal{O}((D_p^2)^2) \) away from the grid size at \( t^n \); see (16.302). Hence the approximations at the new grids are still \( \mathcal{O}((D_p^2)^2) \) correct and thus the free boundary is found with the same accuracy.

In a practical implementation we cannot have \( \Delta t \) being constant (unless it is unduly small) because of the stability restriction for explicit Euler. Hence we should use the implicit Euler scheme.
16.12.4 Finding Suitable Initial Conditions

We still have to provide suitable initial conditions. In order to find these, let us look at the premelting problem. In the heating-up stage, the temperature $T$ in the solid material satisfies the heat equation (16.300), the boundary condition $T(z, t) \rightarrow T_{\text{amb}}$ for $z \rightarrow \infty$, and

$$\frac{\partial T}{\partial z}(0, t) = -J,$$

(16.305)

where $J$ is the (dimensionless) energy $I/I_{\text{ref}}$ supplied at the surface. Moreover, for the initial condition we take $T(z, 0) = T_{\text{amb}}$. We can solve this initial boundary value problem using the Laplace transform [70] and find

$$T(z, t) = J \left( 2 \sqrt{\frac{t}{\pi}} \exp \left( -\frac{z^2}{4t} \right) - z \text{erfc} \left( \frac{z}{2 \sqrt{t}} \right) \right) + T_{\text{amb}}.$$

(16.306)

Therefore, when latent heat is neglected, the position $s$ of the liquid-solid interface is found from

$$J \left( 2 \sqrt{\frac{t}{\pi}} \exp \left( -\frac{z^2}{4t} \right) - s \text{erfc} \left( \frac{s}{2 \sqrt{t}} \right) \right) + T_{\text{amb}} = 0$$

(16.307)

for all $t$. In Figure 16.39 this position is sketched for aluminium. This yields the correct initial value of $\frac{ds}{dt}$. From this we may compute $s(\Delta t)$ by an explicit Euler step. Furthermore, substituting $s = 0$ in (16.306), we conclude that the surface starts to melt at time $t_m$ given by

$$t_m = \frac{1}{4} \pi \left( \frac{T_{\text{amb}}}{J} \right)^2.$$

(16.308)

![Figure 16.39](image)

Figure 16.39. The position of the solid-liquid interface for the case of aluminium. The absorption of latent heat is neglected.
In order to compute the initial condition for the temperature distribution, we once more employ the Stefan condition (16.301c), i.e.,

$$\frac{\partial T_s}{\partial z}(s(t)+, t) = \lambda_u \frac{ds}{dt} + \frac{\partial T_L}{\partial z}(s(t)-, t) \approx \lambda_u \frac{ds}{dt} - J =: -J_\varepsilon \quad (16.309)$$

for \(s\) small. Using (16.309) as boundary condition, we obtain, similarly to (16.306), an expression for the temperature. In particular, we find

$$T_{1,s,j} = T_{amb} \exp \left( -\frac{(z_{s,j} - s_j)^2}{2J_\varepsilon^2} \right) - J_\varepsilon \frac{(z_{s,j} - s)}{T_{amb} \sqrt{\pi}} + T_{amb}, \quad j = 1, 2, \ldots, N_s - 1,$$

as the initial temperature in the solid at time level is \(t^1 = \Delta t\). Because \(s(\Delta t)\) is small, we take the temperature distribution in the liquid at \(t^1 = \Delta t\) to be linear; i.e.,

$$T^{1,l}_{j} = J (s(t^1) - z_{l,j}^1), \quad j = 0, 1, \ldots, N_{liq} - 1. \quad (16.311)$$

Clearly this satisfies the boundary conditions at \(z = 0\) and \(z = s(t)\). Note that this approach cannot easily be extended to cover problems with no constant energy supply.

16.12.5 The Enthalpy Formulation

The enthalpy \(H\) is defined as the sum of the sensible and the latent heat in a substance. If a material is liquid, it contains latent heat of fusion per unit mass \(L_f\) in addition to the sensible heat (per unit mass) \(cT\). Figure 16.40 shows the relation between the temperature and the enthalpy for two different types of materials. Figure 16.40(a) shows this relation for pure, i.e., nonalloy, substances with a single melting-point temperature, whereas Figure 16.40(b) illustrates this relation for a material where the phase change takes place over an extended temperature range from the solid temperature \(T_{sol}\) to the liquid temperature \(T_{liq}\). The region

![Figure 16.40](image_url)

(a) Pure crystalline substances  
(b) Glassy substances and alloys

Figure 16.40. Relation between enthalpy and temperature for pure crystalline substances and for glassy substances and alloys.
with temperature between the solid and the liquid temperatures is referred to as the mushy region. In order to nondimensionalise the enthalpy, we introduce two typical parameters. The first is the enthalpy at vaporisation temperature \( H_v \), which is given by

\[
H_v := \rho c T_v + \rho L_f. \tag{16.312}
\]

For materials with a discrete melting point the second number is the enthalpy at melting temperature \( H_m \), given by

\[
H_m := \rho c T_m, \tag{16.313a}
\]

whereas for glassy substances and alloys this second number is the enthalpy at solid temperature

\[
H_{\text{sol}} := \rho c T_{\text{sol}}. \tag{16.313b}
\]

Hence the dimensionless enthalpy \( H^* \) is defined for pure materials by

\[
H^* = \frac{H_m + (H_v - H_m)H^*}{H_v - H_m}, \tag{16.314a}
\]

and for materials with a melting range by

\[
H^* = \frac{H_{\text{sol}} + (H_v - H_{\text{sol}})H^*}{H_v - H_{\text{sol}}}. \tag{16.314b}
\]

For later use we introduce the constants \( D_1 \) and \( D_2 \):

\[
D_1 := \frac{\rho c (T_v - T_m)}{H_v - H_m}, \quad D_2 := \frac{\rho c (T_v - T_{\text{sol}})}{H_v - H_{\text{sol}}}. \tag{16.315}
\]

We will drop the asterisk \( * \) on the dimensionless enthalpy from now on.

The relationship between the dimensionless enthalpy and the dimensionless temperature for pure materials is now given by (see Figure 16.40(a)),

\[
H(T) = \begin{cases} 
D_1 T & \text{if } T < 0, \\
[0, D_1 \lambda_f] & \text{if } T = 0, \\
D_1 T + D_1 \lambda_f & \text{if } T > 0.
\end{cases} \tag{16.316a}
\]

Likewise, this relationship for materials with a melting range is given by (see Figure 16.40(b))

\[
H(T) = \begin{cases} 
D_2 T & \text{if } T \leq 0, \\
D_2 T + D_2 \lambda_f \frac{T}{T_{\text{liq}}} & \text{if } 0 \leq T \leq T_{\text{liq}}, \\
D_2 T + D_2 \lambda_f & \text{if } T_{\text{liq}} \leq T.
\end{cases} \tag{16.316b}
\]

Note that \( T_{\text{liq}} \) denotes the dimensionless liquid temperature and that \( \lambda_f = L_f / (c(T_v - T_{\text{sol}})) \).

Like in the previous subsection we take the material to be in the region \( z \geq 0 \) with the surface at \( z = 0 \). The enthalpy and temperature of the material in this region are governed by the single energy equation in enthalpy form, which, in dimensionless form, is given by

\[
\frac{\partial H}{\partial t} = D_1 \frac{\partial^2 T}{\partial z^2}, \quad z > 0, \tag{16.317}
\]
for materials with a discrete melting point. For materials with a melting range \( D_2 \) takes the role of \( D_1 \).

Our problem to solve is (16.317), with the enthalpy given in (16.316a) or (16.316b), plus the boundary values derived before. Let us discretise (16.317) with the implicit Euler method on a fixed grid with \( \Delta z := z_b/N \) for some \( N \); i.e.,

\[
H_j^{n+1} - \frac{\Delta t}{\Delta z^2} (T_{j+1}^{n+1} - 2T_j^{n+1} + T_{j-1}^{n+1}) = H_j^n. \tag{16.318}
\]

From (16.316a) or (16.316b) we can express the temperature as a function of the enthalpy. Thus we have a relation of the form

\[
T_j^n := T(H_j^n). \tag{16.319}
\]

We use the boundary conditions (16.301a) and (16.301e) to obtain \( H_N^n \) as well as \( H_0^n \). Writing the approximate vector of enthalpies at time \( t^n \) as \( \mathbf{v}^n \) and of temperatures as \( \mathbf{u}^n \), we have a system of equations like

\[
\mathbf{v}^{n+1} - \mathbf{Au}^{n+1} = \mathbf{v}^n.
\]

Again using (16.319) we can rewrite this as the nonlinear system

\[
\mathbf{v}^{n+1} - \mathbf{F} (\mathbf{v}^{n+1}) - \mathbf{v}^n = \mathbf{0}. \tag{16.320}
\]

This can be solved by Newton’s method. If no phase change occurs, the numerical method derived from the enthalpy problem is satisfactory. For phase change problems it is correct on average, since heat balances are fulfilled throughout. However, the calculated positions of the solid-liquid interface, the temperature, and the enthalpy oscillate with a period corresponding to the time the interface needs to travel through a certain grid interval. This phenomenon can be seen in Figure 16.41. The plateau generated propagates to adjacent intervals and smooths out only after the solid-liquid interface has travelled a sufficient distance from the point under consideration. These problems can be overcome by special approaches.

Figure 16.41. Time history plots for temperature at the interface in the enthalpy problem.
16.12.6 Discussion and Related Problems

- The radiation coming from the laser can have a variety of appearances. The form of the bundle is responsible for the shape of the melt pool. Two typical forms are a top hat and a Gaussian. If the melt pool, or even worse, the drilled hole, is not flat, there will be reflections by the walls that may result in secondary strikes. The radiation bundle is most likely divergent; i.e., not all rays are parallel; this may also cause secondary strikes.

- Most materials do not have a single melting point but a mushy region. (For an alloy this can be understood from the fact that the various constituting components have a different melting point.) This requires a special approach. In fact, one needs to resort to the enthalpy formulation entirely.

- Splashing takes place if enough vapor has been formed. Note that this is in fact a three-phase problem. In order to predict the moment of splashing, one needs to calculate the pressure buildup by vaporization.

- One of the problems (in particular for deeper holes) is that the material will resolidify along the walls on its way up. This is a very interesting problem to model. Indeed, there are two free boundaries involved: one being the top of the moving fluid and the other being the top of the recast layer. In fact, shocks may occur.

- Free-boundary problems occur in a large variety of problems. A large class can be found in books on porous media. Here one often encounters two kinds of fluids, like water and oil, or salty and fresh water. Then convection is usually the dominant phenomenon.

- Another important class of Stefan problems is where we have two different chemical substances rather than phase changes. Typical examples are etching processes, where a chemical (like an acid) reacts with a material.

16.12.7 Exercises

16.51. Derive expression (16.306) for the temperature in the heating-up stage.

16.52. One can also use a fixed grid to compute the solution of the Stefan problem (16.300), (16.301). Formulate what the numerical problem will look like on a fixed grid and compute its solution.

16.53. In Section 16.12.5 it is suggested to find the solution of (16.320) by Newton’s method. Given the very form of it, Picard’s iteration (successive substitution) would be quite attractive.

(a) Explain why this is so given the form of $H(T)$.

(b) Comment on the feasibility of this approach.

16.54. Elaborate the numerical scheme for the enthalpy problem outlined in Section 16.12.5. Solve this problem for a number of different grids to see how the staircase phenomenon, noticed in Figure 16.41, depends on the grid size.
16.13 Determining Chemical Composition by Electrophoresis

16.13.1 Introduction

The purpose of analytical chemistry is to separate, quantify, and identify chemical components in complicated mixtures. Important areas of application include the monitoring of environmental pollution and the quality control of products ranging from food and beverages to pharmaceuticals and virtually all materials made by modern industry. For diagnosis of diseases, and biomedical research in general, analytical instrumentation is indispensable. A wide range of instrumentation is available, the choice of which is among other things determined by the properties of the chemical components to be determined. For volatile components gas chromatography is the most important and widespread separation technique, the separation making use of differences in evaporation upon heating of the mixture. Many components in biochemistry, such as proteins and nucleic acids (DNA, RNA), however, are sensitive to heat treatment, but they are not volatile. On the other hand, these molecules, as well as many low molecular weight components, carry one or more charges when dissolved in water. The differences in their charge, size, and shape make it possible to separate them in an electric field using a family of techniques called electrophoresis [102]. This separation technique was originally developed by Arne Tiselius, who received the Nobel Prize for Chemistry for this in 1948. It can be carried out in glass tubes, in capillaries, in gels, and on paper impregnated with an aqueous buffer solution, the pH of which can be used to selectively alter the effective charge of different components and thus enhance separation.

The component property relevant for separation is the mobility \( \mu \) \( \left[ \text{m}^2/(\text{V} \cdot \text{s}) \right] \), which is defined as the ratio between the linear velocity \( v \) [m/s] of the molecule in the solution and the electric field \( E \) [V/m], i.e.,

\[
v = \mu E. \tag{16.321}
\]

Depending on the chemical composition of the buffer solution(s) in the separation system, there are a few variations of the technique. One is called isotachophoresis [39]. Here one uses a leading electrolyte, with a higher mobility than any of the ions in the sample, as well as a terminating electrolyte, with a lower mobility than any of those ions. The sample is placed in between. A buffering counterion is present in the whole system. In the first stage of isotachophoresis the various components will migrate with different velocities and thus separate, and eventually zones with only one component (in addition to the common counterion) will appear. The length (or volume) of the zones is proportional to their respective amounts. This is called a steady state as all components will travel with the same speed (hence its name: \( \iota \sigma \) means equal and \( \tau \alpha \chi \sigma \) means speed). In this steady state the electric field increases, therefore going from the leading electrolyte to the terminating one, and is apparently inversely proportional to the mobility of the respective component. Strictly speaking, isotachophoresis is not a separation technique, but rather the end result of one. However, the intermediate stages leading to the steady state are very interesting for gaining valuable insight into the separation mechanism, not only of isotachophoresis but of electrophoresis in general.
16.13.2 Mathematical Model of Capillary Electrophoresis

The most important issue in electrophoresis is to determine the concentration of various components in a mixture. In this section we consider a simple one-dimensional model of capillary electrophoresis involving \( n \) components. The concentrations \( c_1, c_2, \ldots, c_n \) [mol/m\(^3\)] of these components depend on the spatial coordinate \( x \) along the capillary and the time \( t \) and satisfy the following set of advection-diffusion equations:

\[
\frac{\partial c_i}{\partial t} + \frac{\partial}{\partial x} \left( v_i c_i \right) = D_i \frac{\partial^2 c_i}{\partial x^2} \quad (i = 1, 2, \ldots, n).
\]  

(16.322)

In (16.322) \( v_i \) is the advection velocity of the \( i \)th component and \( D_i \) [m\(^2\)/s] is the diffusion coefficient, assumed to be constant for all components. As noted before, the velocities \( v_i \) are related to the mobility and electric field. In particular, we have the relation (cf. (16.321))

\[
v_i = (\mu_e + \mu_i) E,
\]

(16.323)

with \( \mu_e \) the (constant) electro-osmotic mobility, \( \mu_i \) the effective mobility of the \( i \)th component, and \( E \) the electric field. For this problem we can take \( \mu_e = 0 \). The equations in (16.322) are coupled through the advection terms, since the mobilities \( \mu_i \) and the electric field \( E \) depend in a complicated way on the concentrations \( c_i \), as specified below.

Let component \( i \) have \( n_i \) subspecies; the acidity is responsible for the degree to which they are dissociated. Let us denote the concentration of (ionised) hydrogen by \( c_H \). Then the degree of dissociation \( \alpha_{i,j} [-] \) of subspecies \( j \) in component \( i \) is given by

\[
\alpha_{i,j} = \frac{\prod_{p=1}^{j-1} K_{i,p}}{1 + \sum_{q=1}^{n_i} \prod_{r=1}^{j} K_{i,r}} \quad (j = 1, 2, \ldots, n_i - 1).
\]  

(16.324a)

Here \( K_{i,j} \) [mol/m\(^3\)] is the acidic constant for the latter subspecies. The dissociation degrees then give the mobilities from the relation

\[
\mu_i = \sum_{j=1}^{n_i} \alpha_{i,j} (\mu_{i,j}^0 - \mu_{i,j-1}^0),
\]

(16.324b)

where \( \mu_{i,j}^0 \) is the mobility of the subspecies at infinite dilution and full subspecies dissociation. These mobilities and the concentrations are needed to compute the conductivity \( \kappa \text{ [S/m]} \) from the relation

\[
\kappa = F \left( \mu_H c_H - \frac{\mu_{OH} K_w}{c_H} + \frac{n_i}{c_i} \sum_{j=1}^{n_i} \alpha_{i,j} Z_{i,j}^0 \mu_{i,j}^0 \right),
\]

(16.325a)

where \( F \) is Faraday’s constant; \( \mu_H \) and \( \mu_{OH} \) are the mobilities of hydrogen and hydroxide ions, respectively; \( K_w \) is the equilibrium constant of water; and \( Z_{i,j} [-] \) is the charge number of subspecies \( j \) in component \( i \). These constants have the values \( F = 9.648456 \times 10^4 \text{ C/mol}, \ \mu_H = 3.5 \times 10^{-7} \text{ m}^2/(\text{V} \cdot \text{s}), \ \mu_{OH} = 2 \times 10^{-7} \text{ m}^2/(\text{V} \cdot \text{s}), \)
16.13. Determining Chemical Composition by Electrophoresis

and \( K_w = 10^{-8} \text{ mol}^2/\text{m}^6 \). Given a (constant) current density \( J [\text{A/m}^2] \), the electric field follows from the modified Ohm’s law; i.e.,

\[
E = \frac{J}{\kappa},
\]

(16.325b)

The expressions in (16.325) determine \( E \) as a function of the concentrations \( c_i \) and \( c_H \).

Finally, to compute \( c_H \), and thus pH, we use Gauss’s law from electromagnetism, describing conservation of charge. Assuming electroneutrality, it reduces in our application to

\[
0 = c_H - \frac{K_w}{c_H} + \sum_{i=1}^{n} \sum_{j=1}^{n} Z_{i,j} \alpha_{i,j} c_i.
\]

(16.326)

This relation specifies \( c_H \) as a function of the concentrations \( c_i \).

16.13.3 Numerical Solution Method

In this section we outline a numerical solution method for the electrophoresis problem from Section 16.13.2. We focus on the advection-diffusion equation (16.322). In order to choose the proper discretisation scheme for (16.322), we note the following. First, the diffusion term should be taken implicitly, since an explicit scheme would lead to a severe time step restriction; e.g., the explicit Euler scheme in combination with central difference space discretisation would require an unduly small time step \( \Delta t \) given by \( \Delta t \leq \frac{1}{4} D \Delta x^2 \), with \( \Delta x \) the grid size; cf. Section 13.3.1. Second, since all equations are coupled through the advection term, we have to take it explicitly. Finally, a scheme discretising the advection term by central differences is prone to spurious oscillations when the Péclet number \( Pe \), defined by

\[
Pe := \max_{i,j} |v_{i,j}| \Delta x / D,
\]

(16.327)

is large; cf. Section 13.3.1. In (16.327) \( v_{i,j} \) denotes the numerical approximation of \( v_i \) at grid point \( x_j \). Based on these observations we choose a mixed time integration scheme (explicit advection and implicit diffusion) (see Section 11.8.2), a central difference approximation for the diffusion term, and upwinding for the advection term. If \( v_i > 0 \), we have

\[
c_{i,j}^{n+1} - \frac{D \Delta t}{\Delta x^2} (c_{i,j+1}^{n+1} - 2c_{i,j}^{n+1} + c_{i,j-1}^{n+1}) = c_{i,j}^{n} - \frac{\Delta t}{\Delta x} (v_{i,j}^n c_{i,j}^{n} - v_{i,j-1}^n c_{i,j-1}^{n}),
\]

(16.328)

where \( c_{i,j}^n \) denotes the numerical approximation of \( c_i(x_j, t^n) \), etc. A similar scheme holds for \( v_i < 0 \). The stability condition for this scheme reads

\[
\frac{\Delta t}{\Delta x} \max_{i,j} |v_{i,j}^n| \leq 1,
\]

(16.329)

where the maximum should be taken over all components \( i \) and all grid points \( x_j \). Furthermore, (16.328) leads to a symmetric positive definite tridiagonal system, which can be solved very efficiently using LU-decomposition; see Section 11.3.1.
Once we have computed the concentrations at the new time level \( t^{n+1} = (n + 1) \Delta t \), we have to update the velocities as follows. First, we compute \( c_H \) from (16.326). Next we compute the mobilities \( \mu_i \) from (16.324) and the electric field \( E \) from (16.324a) and (16.325). Substituting these in (16.323), we find the advection velocities \( v_i \).

### 16.13.4 A Practical Example

The Swedish Department of Occupational Safety and Health in Stockholm has been monitoring brief exposure to styrene and xylene of factory workers in a construction plant that manufactures plastic recreational boats, where organic solvents were in common use [144]. This monitoring was carried out by simply analyzing the urine collected during 24 h. In the human body styrene and xylene are metabolized to mandelic acid and methyl-hippuric acid, respectively. These metabolites are subsequently excreted in the urine. Their concentration is therefore a good quantitative indication of styrene/xylene exposure.

The samples were analysed using isotachophoresis. An example of experimental conditions was as follows:

- The leading electrolyte was 10 mol/m\(^3\) hydrochloric acid, buffered to pH = 3.10 using 12.5 mol/m\(^3\) \( \beta \)-alanine as counterion.
- The terminator electrolyte was 10 mol/m\(^3\) propionic acid, buffered to pH = 4 with 5 mol/m\(^3\) \( \beta \)-alanine as counterion.

We have computed numerical solutions of a 30 cm capillary with internal diameter of 75 \( \mu \)m and with a constant applied current of 10 \( \mu \)A. The detection system was a DC conductivity detector recording field strength. A mixture of mandelic acid and methyl-hippuric acid was injected at the interface between leading and terminating electrolyte, and the current was switched on. A solution of 5 mol/m\(^3\) of each of the two components was injected over a length of 1 mm, as a perfectly cylindrical plug. Component data of this experiment are given in Table 16.6, where \( pK^*_{i,j} := -10 \log(K^*_{i,j}) \). Furthermore, the charge number is \( Z_{i,j} = 0 \) if \( \mu_{i,j} = 0 \); otherwise we have \( Z_{i,j} = 1 \).

One can see from the tabulated data that separation at pH > 8 would be easiest. The methyl-hippuric acid molecule would carry two full charges, whereas mandelic acid would carry a single one. Consequently, methyl-hippuric acid would be twice as fast as mandelic acid (49.7 vs. 24.5). But in practice this does not work, because there is too much

<table>
<thead>
<tr>
<th>component</th>
<th>( \mu_{0,i,j} )</th>
<th>( pK_{i,1} )</th>
<th>( \mu_{0,i,j} )</th>
<th>( pK_{i,2} )</th>
<th>( \mu_{0,i,j} )</th>
<th>( D_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 hydrochloric acid</td>
<td>0</td>
<td>-2</td>
<td>-79.0</td>
<td>2.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 propionic acid</td>
<td>0</td>
<td>4.78</td>
<td>-36.9</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 ( \beta )-alanine</td>
<td>36.7</td>
<td>3.55</td>
<td>0</td>
<td>10.24</td>
<td>-30.8</td>
<td>0.94</td>
</tr>
<tr>
<td>4 methyl-hippuric acid</td>
<td>0</td>
<td>2.71</td>
<td>-25.0</td>
<td>7.25</td>
<td>-49.7</td>
<td>0.64</td>
</tr>
<tr>
<td>5 mandelic acid</td>
<td>0</td>
<td>3.37</td>
<td>-24.5</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
interference from the other components present. For the sake of illustration we choose a separation at pH = 3.10 because it takes longer for the separation to complete, and the intermediate stages can be more easily seen.

### 16.13.5 Numerical Simulations

The time development of the locations in the capillary of the two sample components is illustrated in Figure 16.42. In the leftmost part one can observe that the components only partially leave the injection compartment (boundary at 1 mm) and that adjustment of concentration is taking place. A short, pure zone of methyl-hippuric acid can already be seen, migrating in front (the fastest component), followed by a mixed zone. In the middle part, after 3 s, the isotachophoretic train is fully adjusted (it has left the injection compartment). The steep front and rear boundaries move at equal velocities, whereas the length of the mixed zone is slowly decreasing with time. On the right-hand side it can be seen that there is already a steady state after 5 s; the lengths of both zones, along the horizontal distance axis, will remain the same and are directly proportional to the amount originally injected.

The rectangular concentration profiles of both components are clearly seen. Also, the front and rear boundaries are much steeper than the boundary in between the components. This is easily explained by plotting the simulated electric field; see Figure 16.43. We see that the larger the electric field gradient at the boundary, the steeper the corresponding concentration gradient, because the self-correcting properties of the field are much stronger there. One can say that the electric field gradient is really the driving force that counteracts the diffusion of components into each other’s zone. One should also notice the stepwise increase of the local pH going from the leading to the terminating zone.

The method turns out to be very robust. In this particular case, with five different components (two sample components, leading, terminating, and counterion), the Pe number during the separation ranges from 5 to 70 initially, and from 5 (in the leading) to 50 (in the terminating) as soon as the steady state is reached, without oscillations or mass-balance violations.

In contrast to most other chemical separation methods, the antidiffusive power of electromigration in electrophoresis is striking. This is illustrated in the next example, where in the separation system listed in Table 16.6, a steady state 1 mm zone of 5 mol/m$^3$
mandelic acid is depicted in Figure 16.44, a perfect rectangular concentration profile located at 2 mm. Next the driving current is reduced from 10 μA to 0.01 μA. As a result, the zone virtually comes to a standstill and the diffusion will have full play. In the figure the subsequent concentration profiles after 2, 5, 10, and 20 s are shown. The outward diffusion flux at the zone boundaries leads to a rapid smoothing. In the middle of the zone the concentration seems to increase beyond the original value. Being physically impossible, this must be a numerical artifact, probably due to the sudden decrease in driving current, and thus the electric field. When the original current of 10 μA is restored, several phenomena are observed, as illustrated by the sequence depicted at intervals of 200 ms in Figure 16.44. The zone is again moving to the right. The steep concentration profiles on both sides of the mandelic acid zone are quickly restored within 100 ms. A pronounced sinusoidal disturbance occurs, originating from the front (right) of the zone, which slowly dies out to the left. The original concentration profile of the steady state is restored after 1.2 s in the profile on the far right, at 10 mm.

These examples show that numerical simulation is a powerful tool to study isota-chophoresis. In particular, it enables one to investigate the intermediate stages, which cannot be measured.
16.13.6 Discussion and Related Problems

- Deviation from local electroneutrality is often negligible in electrophoresis. However, if we do not assume electroneutrality, the right-hand side of (16.326) contains the gradient of the electric field, thus coupling the computation of \( c_H \) and \( E \); see exercises.

- An additional, locally different, electric field is generated since all components have different diffusion coefficients. This additional field has a strength of at most 10 percent of the external electric field.

- In some applications, the temperature in the capillary has a radial gradient due to heat loss at the wall, thus introducing two-dimensional effects. In this case, the two-dimensional cylindrical symmetric advection-diffusion equation holds. Also, the heat equation should be included.

- To discretize the advection term, we could use higher-order schemes, such as, e.g., slope limiter methods; see Chapter 13.

- Equation (16.322), with \( \nu_i \) given by (16.323), is also referred to as a drift-diffusion equation. These occur often in physics and electrical engineering. An example is the drift-diffusion model describing semiconductor devices [136].

16.13.7 Exercises

16.55. Give the discretization scheme of (16.322) when the upwind discretization of the advection equation is replaced by the slope limiter method from Section 13.8.2.

16.56. Consider the advection-diffusion equation

\[
\frac{\partial u}{\partial t} + b \frac{\partial u}{\partial x} = D \frac{\partial^2 u}{\partial x^2},
\]

where the advection velocity \( b \) is assumed constant.

(a) Give the mixed explicit advection/implicit diffusion time integration scheme. Use central differences for the space derivatives.

(b) Investigate the stability of this scheme.

(c) Investigate when spurious oscillations occur.

16.57. Suppose electroneutrality does not hold. Then (16.326) should be replaced by

\[
\frac{\epsilon}{F} \frac{\partial E}{\partial x} = c_H \frac{K_w}{c_H} + \sum_{i=1}^{n} \sum_{j=1}^{n_i} Z_{i,j} \alpha_{i,j} c_i,
\]

with \( \epsilon \) the dielectric constant. Modify the numerical solution procedure in Section 16.13.3.
16.14 Pulse Tube Refrigerators

16.14.1 Problem Formulation

Many apparatuses produce so much heat that they need cooling to function properly. The market for stand-alone small cooling devices is still increasing, in particular for communication equipment. There exist all sorts of cooling machinery. A relatively new small type of so-called cryocooler is the pulse tube, based on a principle originally designed by Stirling. This device typically reaches temperatures of about 120 K, but temperatures below 2 K have been reached as well. The essential elements of a pulse tube refrigerator are shown in Figure 16.45. The pulse tube employs a cyclic compression and expansion of a gas, say helium. Due to heat exchange between gas, regenerator, tube walls, and the two heat exchangers, a temperature difference arises along the tube. The pressure oscillations in the system are generated by a piston compressor or by switching valves. The aftercooler removes the heat of compression so that the regenerator can work more efficiently. The regenerator acts as a buffer: it absorbs heat from the gas during the compression part of the pressure cycle and it returns heat to the gas during the expansion part. The cold heat exchanger (CHX) is the coldest point of the system. Here the heat is extracted from the load to be cooled. In the tube the compressible gas oscillates. If there is a suitable phase relationship between the pressure and the gas flow, heat will be transported from the cold end to the warm end. The hot heat exchanger (HHX) removes the heat carried through the tube. It is maintained at ambient temperature. In the orifice design (see Figure 16.45b) the basic pulse tube is modified by adding a reservoir and an orifice. The reservoir is large compared to the pulse tube volume. Gas flows through the orifice due to a pressure difference. More gas now contributes to the cooling power, which improves the efficiency of the cooler.

![Figure 16.45](image-url)  
**Figure 16.45.** A schematic picture of the Stirling-type pulse tube refrigerator: (a) the basic pulse tube and (b) the orifice pulse tube. From right to left the basic system consists of a piston, a regenerator, a cold heat exchanger (CHX), a tube, and a hot heat exchanger (HHX). In (b) the system is extended with an orifice and a reservoir (buffer).
16.14. Pulse Tube Refrigerators

16.14.2 The Model

We will model the tube section of the pulse tube refrigerator to study the energy transfer from the cold to the hot end; see, e.g., [100, 165, 166, 66, 5, 6, 141]. Figure 16.46 shows the energy flow in the pulse tube. At the HHX heat $\dot{Q}_h$ is extracted from the system. At the CHX heat $\dot{Q}_c$ is put into the system. There is a net enthalpy flow from the CHX to the HHX. According to the first law of thermodynamics, the average cooling power is equal to the enthalpy flow in the pulse tube. To simplify the analysis we will consider a one-dimensional model. Let $L$ be the length of the tube section of the cooler. Then the domain is the interval $0 < x < L$. We assume that the fluid is Newtonian, the gas is ideal, and the flow is laminar.

Let us furthermore denote the density by $\rho(x, t)$, the velocity by $u(x, t)$, the temperature by $T(x, t)$, the pressure by $p(x, t)$, the dynamic viscosity by $\mu$, the specific heat of the gas at constant pressure by $C_p$, the thermal conductivity of the gas by $\kappa_g$, and the specific ideal gas constant by $R$. The equations for mass, momentum, and energy conservation and the equation of state are then (cf. Chapter 6)

\[
\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \quad \text{(16.330a)}
\]

\[
\rho \left( \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} \right) = -\frac{\partial p}{\partial x} + \frac{4}{3} \frac{\partial}{\partial x} \left( \mu \frac{\partial u}{\partial x} \right), \quad \text{(16.330b)}
\]

\[
\rho C_p \left( \frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + \frac{\partial}{\partial x} \left( \kappa_g \frac{\partial T}{\partial x} \right), \quad \text{(16.330c)}
\]

\[
p = \rho R T. \quad \text{(16.330d)}
\]

16.14.3 Nondimensionalisation

The system of equations (16.330) will first be made dimensionless. The scaling parameters are chosen as follows: The shortest physical time scale of importance is $1/\omega$, where $\omega$ is the angular frequency of the piston movement. We introduce $U$ to be a representative value for the velocity and therefore $U/\omega$ to be a typical length scale. Let $P$ be the amplitude of the pressure oscillations, $p_0$ the average pressure, $T_{\text{amb}}$ an ambient temperature, $\rho_{\text{ref}}$ a typical density, $\mu_{\text{ref}}$ a typical viscosity, and $\kappa_{\text{ref}}$ a typical thermal conductivity of the gas. We take the pressure variation caused by the piston or valves to be $p(t) = p_0 + P p^*(t)$, where
Chapter 16. Modeling, Analysing, and Simulating Problems from Practice

$p^*(t)$ is a known function of time. We introduce the following dimensionless variables (indicated by an asterisk *):

\[ \rho := \rho_{ref} \rho^*, \quad T := T_{amb} T^*, \quad p := p_0 + PP^*, \quad u := U u^*, \quad x := (U/\omega)x^*, \quad t := t^*/\omega, \quad \mu := \mu_{ref} \mu^*, \quad \kappa_g := \kappa_{ref} \kappa_g^*. \] (16.331)

Using this, we find that the governing equations (16.330) become (leaving out the asterisk notation for simplicity’s sake)

\[ \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x} (\rho u) = 0, \] (16.332a)

\[ \rho \left( \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} \right) = -\frac{1}{Ma^2} \frac{\partial p}{\partial x} + 4 \frac{\partial}{3Re} \mu \frac{\partial u}{\partial x}, \] (16.332b)

\[ \rho \left( \frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = Ec \left( \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} \right) + 1 \frac{\partial}{Pe} \kappa_g \frac{\partial T}{\partial x}, \] (16.332c)

\[ (A + p)B = \rho T. \] (16.332d)

Let us, more specifically, consider a pulse tube operating at 20 Hz, with a typical velocity of 1.5 m/s. The average distance that gas particles travel in the tube (displacement length) is 0.012 m. The length of the tube is 0.2 m. From this we can easily derive the following dimensionless numbers with their typical values:

\[ Ma := \frac{U}{\sqrt{P/\rho_{ref}}} \sim 4.6 \times 10^{-3}, \quad Re := \frac{\rho_{ref} U^2}{\mu_{ref} \omega} \sim 4 \times 10^3, \]

\[ Ec := \frac{U^2}{cp T_{amb}} \sim 4.5 \times 10^{-6}, \quad Pe := \frac{\rho_{ref} U^2 cp}{\omega k_{ref} \rho} \sim 2.6 \times 10^3, \]

\[ A := \frac{p_0}{P} \sim 6, \quad B := \frac{P}{\rho_{ref} RT_{amb}} \sim 0.17. \]

The Mach number $Ma$ characterizes the compressibility of the flow, the Reynolds number $Re$ indicates how strong the inertia is with respect to the viscous effects, the Péclet number $Pe$ determines the ratio of advection of heat to conduction, and, finally, the Eckert number $Ec$ gives the kinetic energy against thermal energy.

16.14.4 Derivation of Velocity and Temperature Equations

The momentum equation (16.332b) can be rewritten as

\[ \frac{\partial p}{\partial x} = \frac{4Ma^2}{3Re} \frac{\partial}{\partial x} \left( \mu \frac{\partial u}{\partial x} \right) - Ma^2 \rho \left( \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} \right). \] (16.333)

The first and second terms on the right-hand side represent viscous and inertial forces. Dimensional analysis reveals that the constants $Ma^2/Re$ and $Ma^2$ are of the order of $10^{-9}$ and $10^{-5}$, respectively. This means that the inertial and viscous forces are too small to produce
a significant pressure gradient; so the right-hand side of (16.333) is approximately zero.
Therefore the pressure is uniform in space and the momentum equation can be neglected.
The system (16.332) now simplifies to
\[
\frac{\partial p}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \quad (16.334a)
\]
\[
\frac{\partial p}{\partial x} = 0, \quad (16.334b)
\]
\[
\rho \left( \frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = \frac{Ec}{Ma^2} \frac{dp}{dt} + \frac{1}{Pe} \frac{\partial}{\partial x} \left( \kappa_g \frac{\partial T}{\partial x} \right), \quad (16.334c)
\]
\[
(A + p)B = \rho T. \quad (16.334d)
\]
Adding (16.334a) and (16.334c) and subsequently substituting (16.334d), we obtain the
following equation for the velocity:
\[
(A + p)B \frac{\partial u}{\partial x} = \left( \frac{Ec}{Ma^2} - B \right) \frac{dp}{dt} + \frac{1}{Pe} \frac{\partial}{\partial x} \left( \kappa_g \frac{\partial T}{\partial x} \right). \quad (16.335)
\]
In order to obtain the required equation for the temperature, we eliminate the density in
(16.334c) using (16.334d). Finally, assuming that the gas thermal conductivity \(\kappa_g\) is con-
stant, we have the equations for the velocity and temperature:
\[
\frac{\partial u}{\partial x} = \epsilon(t) \frac{\partial^2 T}{\partial x^2} + s_1(t), \quad (16.336a)
\]
\[
\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} = \epsilon(t) T \frac{\partial^2 T}{\partial x^2} + s_2(t) T, \quad (16.336b)
\]
where \(s_1(t), s_2(t),\) and \(\epsilon(t)\) are defined by
\[
s_1(t) := -1 \frac{1}{\gamma} \frac{1}{A + p(t)} \frac{dp}{dt}, \quad (16.337a)
\]
\[
s_2(t) := \gamma - 1 \frac{1}{\gamma} \frac{1}{A + p(t)} \frac{dp}{dt}, \quad (16.337b)
\]
\[
\epsilon(t) := \frac{1}{(A + p(t))B Pe} \ll 1, \quad (16.337c)
\]
with \(\gamma = c_p/c_v\) the specific heat ratio. The temperature equation (16.336b) is a nonlinear
advection-diffusion equation where the variable velocity \(u(x, t)\) represents the presence
of advection and the diffusion coefficient \(\epsilon(t) T\) represents the presence of diffusion. Note that
the diffusion coefficient is small, as follows from our dimensional analysis. The temperature
equation is then mostly of an advective nature and is therefore close to hyperbolic.
To complete the system of equations we introduce boundary and initial conditions.
We only need one boundary condition for (16.336a), i.e., the velocity at the hot end \(u_h(t)\). To
find the velocity at the hot end we consider the volume flow through the orifice. From [166]
the volume flow in a linear approximation is given by
\[
\dot{V}_h = -C_{or}(p - p_b), \quad (16.338)
\]
where \( p \) is the tube pressure, \( \rho_b \) is the buffer pressure, and \( C_{or} \) is the flow conductance of the orifice. The velocity is then given by

\[
u(0, t) = u_0(t) = -\frac{C_{or}}{A} (p(t) - \rho_b(t)),
\]

(16.339)

where \( A \) is the cross-sectional area of the tube. Special care has to be taken in determining the buffer pressure such that the net mass flow over the hot end is zero. We nondimensionalise (16.339) by taking \( \rho = \rho_0 + P \rho^* \) and \( \rho_b = \rho_0 + P \rho_b^* \), just like in Section 16.14.3. In dimensionless form (asterisks * are omitted) the boundary condition for the velocity is

\[
u_h(t) = -C \left( \rho(t) - \rho_b(t) \right),
\]

(16.340)

Furthermore, for the temperature boundary conditions we take

\[
\begin{align*}
T(0, t) &= T_h \quad \text{if} \quad u(0, t) \geq 0, \\
\frac{\partial T}{\partial x}(0, t) &= 0 \quad \text{if} \quad u(0, t) < 0,
\end{align*}
\]

(16.341)

We only have to specify an initial condition for the temperature, which we assume to be linear:

\[
T(x, 0) = T_h - (T_h - T_c) \frac{x}{L}.
\]

(16.342)

### 16.14.5 Numerical Solution

To solve the problem numerically we have to discretise the governing equations (16.336) in space and time. Thus we introduce a spatial grid \( \{ x_j = j \Delta x, \ j = 0, 1, \ldots, N_x \} \) with grid size \( \Delta x := L/N_x \) and time levels \( t^n = n \Delta t \) for some time step \( \Delta t > 0 \). We denote by \( u^n_j \) the numerical approximation of the velocity and by \( T^n_j \) that of the temperature at the grid point \( (x_j, t^n) \). We use the following scheme to compute the velocity:

\[
u_j^n = u_{j-1}^n + \frac{\epsilon^n}{\Delta x} \left( T_{j+1}^n - 2 T_j^n + T_{j-1}^n \right) + \Delta x s_1(t^n),
\]

(16.343)

with \( \epsilon^n = \epsilon(t^n) \). Before we choose the proper discretisation scheme for the advection-dominated temperature equation, we note the following. First, the diffusion term should be taken implicitly, since an explicit scheme requires a stability condition of the type \( \Delta t = \mathcal{O}(\Delta x^2) \). Second, in order to get a sharp approximation of steep gradients, we use a flux limiter scheme similar to the one from Section 13.7. So if, e.g., \( u^n_j > 0 \), we have for the temperature equation

\[
T_{j+1}^{n+1} = T_j^n - c_j^n \left( 1 + \frac{1}{2} (1 - c_j^n) \left( \Phi_{j+\frac{1}{2}}^{n+1} - \Phi_{j-\frac{1}{2}}^{n+1} \right) \right) (T_j^n - T_{j-1}^n)
\]

\[
+ \frac{\Delta t}{\Delta x^2} T_j^n \left( T_{j+1}^{n+1} - 2 T_j^{n+1} + T_{j-1}^{n+1} \right) + \Delta t s_2(t^n) T_{j+\frac{1}{2}}^{n+1},
\]

(16.344)
where the local Courant number $c^n_j$ is defined by $c^n_j := \Delta t u^n_j / \Delta x$. Furthermore, the limiter $\Phi^n_{j+\frac{1}{2}}$ is given by $\Phi^n_{j+\frac{1}{2}} = \Phi(r^n_{j+\frac{1}{2}})$ with $\Phi(r)$ a suitable limiter function and with the variable $r^n_{j+\frac{1}{2}}$ defined by

$$
 r^n_{j+\frac{1}{2}} := \begin{cases} 
 T^n_j - T^n_{j-1} & \text{if } u^n_j > 0, \\
 T^n_{j+1} - T^n_j & \text{if } u^n_j < 0; 
\end{cases}
$$

(16.345)

cf. Section 13.7. A similar scheme holds if $u^n_j < 0$.

In a practical implementation it is worthwhile to choose a variable time step. To do so we compute $\max_j |u^n_j|$ at each time level $t^n$ and determine $\Delta t$ such that the condition

$$
\max_j |u^n_j| \Delta t \leq \Delta x
$$

(16.346)
is satisfied. We then compute the temperature $T^n_{j+1}$ from (16.344) and subsequently the velocity $u^n_{j+1}$ from (16.343).

The time integration proceeds until the system reaches the periodic steady state. We require that the absolute difference in temperature between two consecutive full cycles at the grid points be less than a predefined tolerance. The parameter values for the typical single-inlet pulse tube we used in our simulation are given in Table 16.7. In our numerical experiments we took a sinusoidal pressure variation, as shown in Figure 16.47, and a constant buffer pressure. Our model now allows us to study the temperature dynamics in the tube. In Figure 16.48 temperature profiles are displayed at times $t = 4\pi, 4.5\pi, 5\pi, \text{and} 5.5\pi$, which correspond to the various parts of the pressure cycle: $4\pi$ for pressure increases, $4.5\pi$ for maximum pressure, $5\pi$ for pressure decreases, and $5.5\pi$ for minimum pressure. As for determining the periodic steady state for $\Delta x = 0.015$ with a tolerance $\text{tol} = 10^{-4}$, this was achieved after nine cycles.

### Table 16.7. Physical data for a typical single-inlet pulse tube (values at 300 K).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{ref}}$</td>
<td>4.7</td>
<td>kg/m$^3$</td>
<td>$c_p$</td>
<td>5.2 x 10$^5$</td>
<td>J/(kg · K)</td>
</tr>
<tr>
<td>$T_{\text{amb}}$</td>
<td>300</td>
<td>K</td>
<td>$\mathcal{R}$</td>
<td>2.1 x 10$^5$</td>
<td>J/(kg · K)</td>
</tr>
<tr>
<td>$P_0$</td>
<td>3 x 10$^6$</td>
<td>Pa</td>
<td>$L$</td>
<td>0.2</td>
<td>m$^2$/s</td>
</tr>
<tr>
<td>$P$</td>
<td>5 x 10$^5$</td>
<td>Pa</td>
<td>$C_{\text{in}}$</td>
<td>10$^{-8}$</td>
<td>m$^3$/s/ Pa</td>
</tr>
<tr>
<td>$U$</td>
<td>1.5</td>
<td>m/s</td>
<td>$A$</td>
<td>2 x 10$^{-3}$</td>
<td>Pa</td>
</tr>
<tr>
<td>$a_0$</td>
<td>1.2566 x 10$^2$</td>
<td>s$^{-1}$</td>
<td>$p_0$</td>
<td>3.025 x 10$^6$</td>
<td>Pa</td>
</tr>
<tr>
<td>$\mu_{\text{ref}}$</td>
<td>2.0 x 10$^{-5}$</td>
<td>Pa·s</td>
<td>$T_\mathcal{R}$</td>
<td>300</td>
<td>K</td>
</tr>
<tr>
<td>$\kappa_{\text{ref}}$</td>
<td>1.58 x 10$^{-1}$</td>
<td>W/(m · K)</td>
<td>$T_\mathcal{C}$</td>
<td>70</td>
<td>K</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 16.47. Dimensionless pressure variation used in the simulation.

Figure 16.48. Temperature profiles for the different parts of the pressure cycle.

16.14.6 Discussion and Related Problems

- There are more parts of the pulse tube that need further analysis, e.g., the regenerator, which plays a crucial role in the entire process. Actually, the flow as such is much
more complicated than modeled here. One should at least include the valve and the reservoir. Cryocoolers often have more reservoirs that may be included as well.

- An important question for any manufacturer is how to design an optimal product. Pulse tubes are no exception. This is a subject outside the scope of this book as it involves optimisation techniques. Shape optimisation requires a way to solve a series of problems, say by Newton’s method. An option is to use a simpler analytical approximating solution to the problem, such as a harmonic balance approach based on Fourier series [165].

- In practice one has to resort to higher-dimensional situations. In particular, the flow is locally turbulent and as such requires a three-dimensional approach. This falls outside the scope of this book. It is still an active area of research as the full heat transfer problem may crucially hinge on the mixing effect of this turbulence.

### 16.14.7 Exercises

16.58. One can also derive a time-dependent buffer pressure by considering expansion and compression as adiabatic. To see this use Poisson’s law to derive the volume flow through the orifice. The resulting equation for the pressure is of Riccati type (and can therefore be solved exactly).

16.59. Repeat the calculations, now with a square wave as the (dimensionless) pressure, i.e., a piecewise constant function being either +1 or −1, which results from instantaneous valve operation.

16.60. Rather than one-sided differences, one may try to use central differences for the velocity in (16.343). Give this scheme. Carry out numerical experiments and compare the results.

16.61. Perform numerical tests to see whether the number of iterations to reach the periodic steady state depends on the accuracy, i.e., on the chosen grid, both in time and space.

### 16.15 Flow in a Glass Oven

#### 16.15.1 Problem Formulation

Glass is manufactured in tank furnaces in which raw materials like sand and soda are melted in a complicated physical and chemical process. The typical length of furnaces varies from 6 m to 30 m, the width from 3 m to 15 m, and the height (of the glass melt) from 0.6 m to 1.5 m. Raw materials, the batch, are put into the furnace via a so-called doghouse. These materials, being lighter than the glass melt, initially stay on top and gradually melt under the influence of heat coming from burners and already molten glass. Melting gradually, the glass changes its viscosity from around 1000 [Pa·s] to 10 [Pa·s], which triggers the flow. The changes in density due to the heating process are less dramatic. In fact, the density changes by about 10 percent of its value, say from 2600 [kg/m³] to 2350 [kg/m³]. However, these density gradients are essential for establishing a (mixing) flow.
Since the density of the molten glass is higher than that of the batch, the glass flow is first directed downward, then it flows forward along the bottom, and somewhere around the middle of the furnace it flows up and toward the batch (backflow). This particular pattern is caused by buoyancy due to the temperature and density differences. The heating process is organized in such a way that the hottest spot on the surface of the melt stays in the middle of the furnace. This particular arrangement ensures the backflow of the glass, which is crucial for the proper melting of the batch. The glass needs to stay in the furnace for a certain period of time (up to 24 h) to melt and mix properly. Due to the viscous nature of the glass melt, its typical velocity is relatively small, say in the range of 0.01 [m/s] to 0.04 [m/s]. It is very important to obtain proper insight into the glass flow as this determines the mixing and thus the quality of the glass.

16.15.2 The Model

In order to model the flow in a glass oven, we employ the conservation equations of mass and momentum. The inflow of raw material and the outflow of the melt can be assumed constant, implying a stationary flow. Therefore these equations read

\[ \nabla \cdot (\rho \mathbf{v}) = 0, \quad (16.347a) \]

\[ \nabla \cdot (\rho \mathbf{v} \mathbf{T}) = \nabla \cdot \mathbf{T} - \rho g e_z, \quad (16.347b) \]

where \( \rho \) is the density, \( \mathbf{v} \) is the velocity, \( \mathbf{T} \) is the stress tensor, and \( g \) is the gravitational acceleration, assumed to be directed in the negative \( z \) direction. To solve the momentum equation (16.347b) we need to relate the stress tensor \( \mathbf{T} \) to the velocity \( \mathbf{v} \) and the pressure \( p \). We assume the fluid to be Newtonian, so

\[ \mathbf{T} = -p \mathbf{I} + 2\mu \mathbf{D}, \quad \mathbf{D} = \frac{1}{2} \left( \nabla \mathbf{v} + \nabla \mathbf{v}^T \right), \quad (16.348) \]

where \( \mathbf{I} \) and \( \mathbf{D} \) are the identity tensor and the rate of deformation tensor, respectively, and where \( \mu \) is the dynamic viscosity. Substituting (16.348) in (16.347b), we find for the momentum equation

\[ \nabla \cdot (\rho \mathbf{v} \mathbf{T}) - \nabla \cdot (\mu \nabla \mathbf{v}) + \nabla p = -\rho g e_z. \quad (16.349) \]

The scaling parameters to make the equations dimensionless are a length scale \( L \) representative of the furnace, a characteristic velocity \( V \), a characteristic density \( \rho_{\text{ref}} \), and a characteristic viscosity \( \mu_{\text{ref}} \). We introduce the following dimensionless variables (indicated with an asterisk *):

\[ x = L x^*, \quad \mathbf{v} = V \mathbf{v}^*, \quad \rho = \rho_{\text{ref}} \rho^*, \quad \mu = \mu_{\text{ref}} \mu^*, \quad p = \rho_{\text{ref}} V^2 p^*. \quad (16.350) \]

Applying this scaling, we find that the governing equations (16.347a) and (16.349) become (omitting the asterisk * for simplicity’s sake)

\[ \nabla \cdot (\rho \mathbf{v}) = 0, \quad (16.351a) \]

\[ \nabla \cdot (\rho \mathbf{v} \mathbf{T}) - \frac{1}{Re} \nabla \cdot (\mu \nabla \mathbf{v}) + \nabla p = -\frac{1}{Fr^2} \rho e_z, \quad (16.351b) \]
where \( Re \) and \( Fr \) are the Reynolds number and Froude number, respectively, given by

\[
Re := \frac{\rho_{\text{ref}} V L}{\mu_{\text{ref}}}, \quad Fr := \frac{V}{\sqrt{gL}}.
\]

In order to assess (16.351b), we use the following typical values of the scaling parameters: \( L = 5 \text{[m]} \), \( V = 0.05 \text{[m/s]} \), \( \rho_{\text{ref}} = 2300 \text{[kg/m}^3\text{]} \), and \( \mu_{\text{ref}} = 10 \text{[Pa \cdot s]} \). Hence the dimensionless numbers are \( Re = 57.5 \) and \( Fr = 7.14 \times 10^{-3} \). The convection part dominates over diffusion and obviously cannot be neglected; i.e., the Stokes approximation (cf. Section 16.11) cannot be used.

We note that the density and the viscosity are temperature dependent. The temperature field is not known a priori and is to be computed from the energy equation. However, this equation makes the model far more complicated and results, among other things, in a nonlinear coupling of all equations. We therefore simplify the model by using simulated (given) data for the temperature instead. The density and the viscosity are then computed using this given temperature field. Having fixed the temperature field as \( T = T(x) \), we treat the density and the viscosity as functions depending on space:

\[
\rho = \rho(x) = \rho(T(x)), \quad \mu = \mu(x) = \mu(T(x)).
\]

The density of glass \( \rho \) is a linear function of the temperature given by

\[
\rho(x) := \rho_0 \left(1 - \beta(T(x) - T_0)\right).
\]

Here \( T_0 \) is the reference temperature, \( \rho_0 \) is the density of the glass at \( T_0 \), and \( \beta \) is the thermal expansion coefficient. The value of the density typically ranges from 2300 [kg/m\(^3\)] to 2500 [kg/m\(^3\)]. Although the density variations are not large, they cause buoyancy effects and thus are the driving force for the glass flow. The gravitational force in the momentum equation (16.351b) changes proportionally with the density. The viscosity of the glass decreases exponentially with increasing temperature according to the Vogel–Fulcher–Tamman relation [157].

### 16.15.3 Defining the Domain

Since the oven has a rectangular structure we can model it as a union of, say, \( S \) nonintersecting blocks; i.e., the domain \( \Omega \) can be written as

\[
\Omega := \bigcup_{\ell=1}^S \Omega^{(\ell)}, \quad \Omega^{(\ell)} := (x_0^{(\ell)}, x_1^{(\ell)}) \times (y_0^{(\ell)}, y_1^{(\ell)}) \times (z_0^{(\ell)}, z_1^{(\ell)}).
\]

The boundary of \( \Omega \), which we denote by \( \Gamma \), can be split into several parts. We distinguish the \textit{walls} of the furnace, the \textit{top layer}, the \textit{inflow} area, and the \textit{outflow} area:

\[
\Gamma = \Gamma_{\text{wall}} \cup \Gamma_{\text{top}} \cup \Gamma_{\text{in}} \cup \Gamma_{\text{out}}.
\]

We now specify the boundary conditions. At any point of the boundary we can write the velocity vector \( \mathbf{v} \) in a coordinate system constituted by the outer unit normal vector \( n \).
and two unit vectors \( t_1 \) and \( t_2 \) in the boundary plane; i.e., \( v = v_n n + v_1 t_1 + v_2 t_2 \) for all \( x \in \Gamma \). We thus have the following situations.

- **Walls.** The walls are impermeable; i.e., \( v_n = 0 \). Moreover, we assume the no-slip condition \( v_{t_1} = v_{t_2} = 0 \). As a result, we have

  \[
  v = 0, \quad x \in \Gamma_{\text{wall}}.
  \]  

  \[ 16.356a \]

- **Top layer.** We model the top layer as a symmetry plane, i.e., \( v_n = 0 \), and the normal derivative of the tangential velocity components is zero as well. The boundary condition is thus given by

  \[
  v_n = 0, \quad \frac{\partial v_{t_1}}{\partial n} = \frac{\partial v_{t_2}}{\partial n} = 0, \quad x \in \Gamma_{\text{top}}.
  \]  

  \[ 16.356b \]

- **Inflow.** The inflow area is characterised by a perpetual supply of glass flowing into the furnace with a constant speed. We assume that the inflow is directed along the inner unit normal, which means that the tangential components of the inflow are equal to zero; i.e.,

  \[
  v_n = v_0 < 0, \quad v_{t_1} = v_{t_2} = 0, \quad x \in \Gamma_{\text{in}}.
  \]  

  \[ 16.356c \]

- **Outflow.** The boundary conditions for the outflow are similar to the ones for the inflow. We assume that the glass melt floats out of the furnace along the outer unit normal with a constant speed. These boundary conditions thus read

  \[
  v_n = v_1 > 0, \quad v_{t_1} = v_{t_2} = 0, \quad x \in \Gamma_{\text{out}}.
  \]  

  \[ 16.356d \]

The outflow velocity \( v_1 \) cannot be chosen arbitrarily; otherwise we would encounter mass loss or mass gain effects in the furnace. Applying Gauss’s law to (16.351a) and taking into account the boundary conditions at the walls and the top layer, we find the following relation between inflow velocity \( v_0 \) and outflow velocity \( v_1 \):

\[
v_0 \int_{\Gamma_{\text{in}}} \rho \, dS + v_1 \int_{\Gamma_{\text{out}}} \rho \, dS = 0.
\]  

\[ 16.356e \]

The domain \( \Omega \) can be enclosed by a single block. Indeed, the corner points of this block are given by

\[
x_\ell := \min_s x_0^{(s)}, \quad y_\ell := \min_s y_0^{(s)}, \quad z_\ell := \min_s z_0^{(s)},
\]

\[
x_r := \max_s x_0^{(s)}, \quad y_r := \max_s y_0^{(s)}, \quad z_r := \max_s z_0^{(s)}.
\]  

\[ 16.357 \]

By construction, the closure of \( \Omega \) is a subset of the closure of the block defined by (16.357); i.e.,

\[
\bar{\Omega} \subset \Omega', \quad \Omega' := [x_\ell, x_r] \times [y_\ell, y_r] \times [z_\ell, z_r].
\]  

\[ 16.358 \]

Note that \( \Omega' \) is the smallest block that encloses \( \Omega \).
Let us define a grid $\Omega_\Delta$ as follows:

$$
\Omega_\Delta := \{(x, y, z) | x = x_\ell + (i - 1)\Delta x, y = y_\ell + (j - 1)\Delta y, z = z_\ell + (k - 1)\Delta z, \quad i = 1, \ldots, N_x, \quad j = 1, \ldots, N_y, \quad k = 1, \ldots, N_z\}, \quad (16.359)
$$

where $x_\ell$, $y_\ell$, and $z_\ell$ are defined in (16.357) and $N_x$, $N_y$, and $N_z$ are the numbers of grid points in the $x$, $y$, and $z$ directions, respectively. The grid sizes $\Delta x$, $\Delta y$, and $\Delta z$ are thus defined by

$$
\Delta x := \frac{x_r - x_\ell}{N_x - 1}, \quad \Delta y := \frac{y_r - y_\ell}{N_y - 1}, \quad \Delta z := \frac{z_r - z_\ell}{N_z - 1}.
$$

In order to facilitate our discussion, we introduce the material property function $m(x)$ for $x \in \Omega_\Delta$ as follows:

$$
m(x) = \begin{cases} 
0 & \text{if } x \notin \bar{\Omega}, \\
\text{wall} & \text{if } x \in \Gamma_{\text{wall}}, \\
\text{inflow} & \text{if } x \in \Gamma_{\text{in}}, \\
\text{top} & \text{if } x \in \Gamma_{\text{top}}, \\
\text{outflow} & \text{if } x \in \Gamma_{\text{out}}, \\
\text{glass} & \text{otherwise}.
\end{cases} \quad (16.360)
$$

We have to construct the discrete equations based on the information provided by (16.360). Thus if, e.g., $m(x) = \text{glass}$, we interpret this as $x \in \Omega \setminus \Gamma$ and use the full set of equations. If, however, $m(x) = \text{inflow}$ or $m(x) = \text{outflow}$, the prescribed boundary values of the velocity should be used at the corresponding point.

### 16.15.4 Discretising the Convection-Diffusion Operator

Rather than writing down the complete discrete momentum equations, we derive the discretisation for the $x$ component of the convection-diffusion operator only. It reads

$$
\mathcal{M}(v_x) := \nabla \cdot (\rho v_x v_x) - \frac{1}{Re} \nabla \cdot (\mu \nabla v_x), \quad (16.361)
$$

We will treat the diffusion and convection parts of (16.361) separately. At a point $x \in \Omega_\Delta$ for which $m(x) = \text{glass}$ we approximate the diffusion part of (16.361), i.e., $-\nabla \cdot (\mu \nabla v_x)/Re$, 

...
by central differences, i.e.

\[- \frac{1}{Re} \nabla \cdot (\mu \nabla v_x)(x)\]

\[\doteq \frac{1}{Re \Delta x^2} \left( \mu \left( x - \frac{1}{2} h_x \right) (v_x(x) - v_x(x - h_x)) \right) + \mu \left( x + \frac{1}{2} h_x \right) (v_x(x) - v_x(x + h_x))\]

\[+ \frac{1}{Re \Delta y^2} \left( \mu \left( x - \frac{1}{2} h_y \right) (v_x(x) - v_x(x - h_y)) \right) + \mu \left( x + \frac{1}{2} h_y \right) (v_x(x) - v_x(x + h_y))\]

\[+ \frac{1}{Re \Delta z^2} \left( \mu \left( x - \frac{1}{2} h_z \right) (v_x(x) - v_x(x - h_z)) \right) + \mu \left( x + \frac{1}{2} h_z \right) (v_x(x) - v_x(x + h_z))\],

(16.362)

where \( h_x := (\Delta x, 0, 0)^T \), etc. For a collocated grid any variable computed at the middle between two neighbouring grid points is approximated by linear interpolation; i.e.,

\[g \left( x + \frac{1}{2} h \right) := \frac{1}{2} (g(x) + g(x + h)),\]

(16.363)

where \( h \) is equal to \( h_x, h_y, \) or \( h_z \).

The discrete diffusion operator on a collocated grid in (16.362) is identical to the discrete diffusion operator on a staggered grid (cf. Figure 9.9). The important difference lies, however, in computing the viscosity \( \mu \) at the midpoints. When a staggered grid is used, no interpolation is required.

For the convection part of (16.361), i.e., \( \nabla \cdot (\rho v_x x) \), we use central differences as well. Thus we have

\[\nabla \cdot (\rho v_x x)\]

\[\doteq \frac{1}{\Delta x} \left( (\rho v_x) (x + \frac{1}{2} h_x) \right) \frac{1}{2} (v_x(x) + v_x(x + h_x))\]

\[- (\rho v_x) (x - \frac{1}{2} h_x) \frac{1}{2} (v_x(x) + v_x(x - h_x))\]

\[+ \frac{1}{\Delta y} \left( (\rho v_y) (x + \frac{1}{2} h_y) \right) \frac{1}{2} (v_x(x) + v_x(x + h_y))\]

\[- (\rho v_y) (x - \frac{1}{2} h_y) \frac{1}{2} (v_x(x) + v_x(x - h_y))\]

\[+ \frac{1}{\Delta z} \left( (\rho v_z) (x + \frac{1}{2} h_z) \right) \frac{1}{2} (v_x(x) + v_x(x + h_z))\]

\[- (\rho v_z) (x - \frac{1}{2} h_z) \frac{1}{2} (v_x(x) + v_x(x - h_z))\],

(16.364)
For this term the mass fluxes $\rho v_x$, $\rho v_y$, and $\rho v_z$ have to be approximated by linear interpolation.

### 16.15.5 Discretising the Pressure Gradient and the Continuity Equation

Discretising the pressure gradient on a collocated grid is rather straightforward. The components of the discrete gradient have to be computed at all points of $\Omega_1$ where the corresponding velocity component is not prescribed. This means that at an arbitrary point $x \in \Omega_1$, where $m(x) = \text{glass}$ we need to discretise three components of the gradient. At a point $x$ on the top layer $\Gamma_{\text{top}}$, assumed to be horizontal, only two components ($x$ and $y$) of the gradient need to be discretised; the third ($z$) will not be used in the solution method because $v_z(x) = 0$ when $m(x) = \text{top}$. If $x$ is a glass point, the discrete gradient there is defined by

$$
\nabla p(x) = \left( \frac{1}{2\Delta x} \left( p(x + h_x) - p(x - h_x) \right), \frac{1}{2\Delta y} \left( p(x + h_y) - p(x - h_y) \right), \frac{1}{2\Delta z} \left( p(x + h_z) - p(x - h_z) \right) \right). \quad (16.365)
$$

Discretising the continuity equation is somewhat more complicated. Ideally, we would like to use central differences for each of the three terms. A problem with central differences, however, appears when the grid point at which we discretise the continuity equation is a boundary point; this means that one of the grid points involved in the central difference approximation lies outside the computational domain. Hence a one-sided difference should be used instead. If $x$ is a point in the interior, i.e.,

$$
m(x - h_x) \neq 0 \quad \text{and} \quad m(x + h_x) \neq 0,
$$

the first term in the continuity equation, $\partial (\rho v_x) / \partial x$, is discretised as

$$
\frac{1}{2\Delta x} \left( (\rho v_x)(x + h_x) - (\rho v_x)(x - h_x) \right).
$$

If, however, $x$ belongs to the left boundary, i.e.,

$$
m(x - h_x) = 0 \quad \text{and} \quad m(x) \neq 0,
$$

no velocity is defined at $x - h_x$; therefore we use the forward difference

$$
\frac{1}{\Delta x} \left( (\rho v_x)(x + h_x) - (\rho v_x)(x) \right). \quad (16.366)
$$

Similarly, we employ the backward difference approximation at the right boundary. Discretising the two other components of the continuity equation, $\partial (\rho v_y) / \partial y$ and $\partial (\rho v_z) / \partial z$, and adding up the result, we end up with the discrete continuity equation. For instance, at an arbitrary glass point $x$ the discrete divergence reads

$$
\frac{1}{2\Delta x} \left( (\rho v_x)(x + h_x) - (\rho v_x)(x - h_x) \right) + \frac{1}{2\Delta y} \left( (\rho v_y)(x + h_y) - (\rho v_y)(x - h_y) \right) + \frac{1}{2\Delta z} \left( (\rho v_z)(x + h_z) - (\rho v_z)(x - h_z) \right) = 0. \quad (16.367)
$$
16.15.6 Simulations

To illustrate the potential of the discretisation technique we consider a model furnace, see Figure 16.49. It measures $10.5 \times 4 \times 1$ and consists of two tanks connected by a small channel: a larger melting tank and a smaller postprocessing tank. The melting tank measures $8 \times 4 \times 1$, while the postprocessing tank measures $2 \times 2 \times 1$.

The inflow area $\Gamma_{\text{in}}$ is located at the left end of the furnace and is of rectangular shape. The outflow area $\Gamma_{\text{out}}$, in turn, is located at the right end of the furnace. These areas measure $3.8 \times 0.2$ and $3.4 \times 0.3$, respectively. The inflow velocity is directed normal to the boundary and is equal in absolute value to 0.01. We can find the velocity for the outflow area from (16.356e).

In order to have a reasonable approximation for the temperature field, the lowest temperature in the furnace occurs near the inflow area and furthermore the temperature is lower near the walls and the hottest spot occurs somewhere in the middle of the top layer. For a unit cube a suitable approximation is then given by the function

$$T = \frac{1}{2} \left( 1 + \sin \left( \frac{1}{2} \pi (3x - 1) \right) \right) + 4 \left( x - \frac{1}{5} \right) \left( 1 - x \right) \left( \frac{6}{5} + \sin \left( \frac{1}{2} \pi (2z - 1) \right) \right).$$

For the realistic problem of an actual furnace we need to scale this relation first.

The computational domain is covered with a grid $\Omega_{\Delta}$ the sizes of which, $\Delta x$, $\Delta y$, and $\Delta z$, are equal to

$$\Delta x = \Delta y =: h, \quad \Delta z = \frac{1}{2} h, \quad h = 0.2.$$

We have applied the pressure correction method [45] to solve the discretised continuity and momentum equations; see also Section 9.7, where we applied the method to the Stokes equations. The computed velocity is shown in Figure 16.50. As can be seen from this picture, the velocity of the flow near the inflow area is directed downward, which is explained by the fact that the cold glass flow is heavier than the warm one. The molten glass flows along the bottom and then up to the top, approximately in the middle of the melting tank. Then the flow splits: part of the melt flows toward the input area, thus forming the backflow, and part flows after a convection loop toward the postprocessing tank. These results reflect the actual behaviour of glass melt in furnaces.
16.15. **Flow in a Glass Oven**

16.15.7 **Discussion and Related Problems**

- The temperature that was left out in the model above, for simplicity’s sake, is, of course, a crucial quantity for obtaining proper insight into the problem itself. Indeed, the differences in density driving the flow are caused by temperature differences. Besides heating the glass from above, one also often employs *boosters*, i.e., electric heaters in the tank to enforce this flow locally. For a more complete model one should also take into account that the material at the inlet is not a fluid yet (not even glass, as the chemical process takes place only gradually). This batch floats on top of the bath to some extent. Besides this shielding by the batch, a foam blanket is often formed that hinders the heat transfer of the burners above.

- The mixing in the tank is enhanced in several ways. Besides using boosters, one may also stir the glass mechanically. Both for this and to decrease the gas content in the melt one uses a bubbling technique in which a regular flow of air bubbles is injected at some spots. The global flow is furthermore directed by a special shape in the tank, such as a threshold somewhere on the bottom.

- As with many problems where confined material flows, there are difficult questions about what is happening at the contacts. In particular, there is always some friction and so we often have to impose some combined stick/slip condition (Robin type). Given the strong dependence of the glass on the temperature it is to be expected that the warmer regions have more slip than the cooler ones.

16.15.8 **Exercises**

16.62. The continuity and momentum equations can also be discretised on a staggered grid, as outlined in Section 9.7.
   (a) Give the discretisation of (16.351b) on a staggered grid.
   (b) Formulate the pressure correction method for the resulting algebraic systems.
   (c) Simplify the setting by ignoring the \( y \) dependency so that the problem becomes two-dimensional. Compute the flow in the oven specified in Section 16.15.6.

16.63. Change the geometry of the furnace by putting a block shaped threshold somewhere on the bottom of the tank. Compute the flow in this tank.
16.16 Attenuation of Sound in Aircraft Engine Ducts

16.16.1 Problem Formulation

Since jet noise has been reduced considerably by the introduction of high-bypass turbofan engines, the primary source of noise of aircraft aeroengines is the aerodynamic interaction of fan and bypass stator (Figure 16.51). This noise radiates upstream via the inlet duct and downstream via the bypass duct, away into the far field. It causes considerable nuisance to the community during approach and takeoff. Aircraft noise is therefore subject to very strict regulations, which already limit the capacity of a number of major airports around the world. One of the many measures taken to reduce the noise is the application of acoustic lining in the inlet and bypass ducts. This lining usually consists of an array of small resonators designed to attenuate the sound waves that pass along the wall. In order to optimize the damping properties, we need to model the sound propagation through the inlet and bypass ducts. This will be done here by the well-established model of inviscid mean flow and acoustic perturbations, with assumed absent vorticity and uniform entropy. The inherent slow variation of the geometry of a typical flow duct provides the basis of an asymptotic analysis favoring analytic solutions. As far as the asymptotic reduction is part of the modeling process, it is an example of asymptotic modeling (Section 7.2).

16.16.2 The Model

Especially in the inlet, both mean flow and acoustic perturbations are quite well modeled by an inviscid, homentropic, irrotational flow (7.5), while the mean flow velocity is nearly uniform. So we can introduce a velocity potential and integrate the momentum equation to obtain Bernoulli’s equation (7.12). We write velocity, density, and pressure (made dimensionless as shown below) as the sum of a steady component and a harmonic perturbation of frequency $\omega$ (dimensionless; also known as the Helmholtz number); i.e.,

$$\begin{align*}
[V, D, P] &+ \text{Re}([\nabla \phi, \rho, p] e^{i\omega t}).
\end{align*}
$$

Figure 16.51. Sketch of an aircraft turbofan engine.
16.16. Attenuation of Sound in Aircraft Engine Ducts

(There is no need to introduce explicitly a mean flow potential.) After linearisation we find for the mean flow the (dimensionless) equations

\[
\frac{1}{2}V^2 + \frac{C^2}{\gamma - 1} = E, \quad \nabla \cdot (DV) = 0, \quad \frac{P}{D} = \text{constant}, \quad C^2 = \gamma \frac{P}{D},
\]

(16.369)

where \( \gamma \) is the specific heat ratio, \( C \) denotes the mean sound speed, and \( E \) is the integration constant in Bernoulli’s equation. For the acoustic perturbations we have

\[
(i \omega + V \cdot \nabla)\rho + \rho \nabla \cdot V + \nabla \cdot (D \nabla \phi) = 0, \quad D(i \omega + V \cdot \nabla)\phi + p = 0, \quad p = C^2 \rho.
\]

(16.370)

By means of the second and third equations we eliminate \( p \) and \( \rho \) from the first. By using \( \nabla \cdot (DV) = 0 \), we obtain

\[
D^{-1} \nabla \cdot (D \nabla \phi) - (i \omega + V \cdot \nabla) \left[ C^{-2} (i \omega + V \cdot \nabla) \phi \right] = 0.
\]

(16.371)

The model is valid inside the slowly varying hollow or annular duct given in cylindrical coordinates \((x, r, \theta)\) by

\[
0 \leq R_1(X) \leq r \leq R_2(X), \quad X = \varepsilon x,
\]

(16.372)

where \( \varepsilon \) is small and \( R_1, R_2 \) do not depend on \( \varepsilon \) other than via \( X \). (Note that technically \( \varepsilon \) is only a bookkeeping parameter. Its size is of the order of the derivative of \( R_i \) with respect to \( x \).) For definiteness we will define the position of the source at \( x = 0 \). Lengths are made dimensionless on a typical outer diameter, usually at the source plane. We may have, e.g., \( R_2(0) = 1 \). Velocities and densities are made dimensionless on a typical value of sound speed and density, e.g., at the source plane.

The duct wall is impermeable for the mean flow, while it is lined by acoustic material of impedance \( Z_1 \) and \( Z_2 \). The acoustic boundary condition is, however, not the regular impedance condition—the ratio between pressure and velocity—but Myers’s impedance condition [98] to include the effect of refraction through the vanishing mean flow shear layer along the wall. This yields the following boundary conditions at \( r = R_i, i = 1, 2 \):

\[
(V \cdot n_i) = 0, \quad i \omega Z_i (\nabla \phi \cdot n_i) = [i \omega + V \cdot \nabla - n_i \cdot (n_i \cdot \nabla V)] p.
\]

(16.373)

The mean flow will be defined by a given mass flux \( F \) and the assumption that it starts far upstream as a uniform flow and varies only through \( X \) (like in Example 15.24). The acoustic field that will be considered is a slowly varying duct mode, i.e., a mode that would be a strict solution of the equations in a straight duct (like in Example 15.36).

16.16.3 Asymptotic Solution

The Mean Flow  The mean flow is found by using the method of slow variation (Section 15.3.1). Its derivation is given in detail in Example 15.24. For now we only need to know that

\[
V = U_0(X) e_x + \varepsilon V_1(X, r)e_r + \mathcal{O}(\varepsilon^2), \quad U_0(X) = F/D_0(X) A(X),
\]

\[
\frac{\partial}{\partial X} (r D_0 U_0) + \frac{\partial}{\partial r} (r D_0 V_1) = 0, \quad \text{and} \quad V_1 = U_0 R'_i \quad \text{along} \quad r = R_i,
\]

(16.374)
Chapter 16. Modeling, Analysing, and Simulating Problems from Practice

where $R_i'$ denotes the derivative of $R_i$ with respect to $X$, $A$ is the surface of a duct cross section, $C = C_0(X) + \mathcal{O}(\varepsilon^2)$, $D = D_0(X) + \mathcal{O}(\varepsilon^2)$, and $D_0$ is the solution of the algebraic equation

$$\frac{F}{2D_0 A^2} + \frac{D_0^{-1}}{\gamma - 1} = E. \quad (16.375)$$

The Acoustic Field  A straight-duct modal wave form is a function of $r$ multiplied by a complex exponential in $\theta$ and $x$. The mode-like wave we are looking for here is obtained by assuming the amplitude and axial and radial wave numbers to be slowly varying, i.e., depending on $X$. So we introduce the so-called WKB ansatz (15.62)

$$\phi(x,r,\theta; \varepsilon) = \frac{i \varepsilon}{\Phi} \left\{ \frac{\partial}{\partial X} \left[ \left( \frac{U_0 \Omega}{C_0^2} + \kappa \right) D_0 \Phi^2 \right] + \frac{1}{r} \frac{\partial}{\partial r} \left[ \frac{r V_1 \Omega}{C_0^2} D_0 \Phi^2 \right] \right\}, \quad (16.376)$$

where $m$ and $\kappa$ denote, respectively, the circumferential periodicity and the slowly varying axial wave number of the mode of interest. Substituting in (16.371), suppressing the exponential, we obtain, and collecting like powers of $\varepsilon$ up to order $\varepsilon^2$,

$$D_0 \mathcal{L} [\Phi] = \frac{i \varepsilon}{\Phi} \left\{ \frac{\partial}{\partial X} \left[ \left( \frac{U_0 \Omega}{C_0^2} + \kappa \right) D_0 \Phi^2 \right] + \frac{1}{r} \frac{\partial}{\partial r} \left[ \frac{r V_1 \Omega}{C_0^2} D_0 \Phi^2 \right] \right\}, \quad (16.377)$$

where

$$\Omega = \omega - \kappa U_0 \quad (16.378)$$

and the operator $\mathcal{L}$ is defined by

$$\mathcal{L} = \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{\Omega^2}{C_0^2} - \kappa^2 - \frac{m^2}{r^2}. \quad (16.379)$$

With $n_i \cdot (n_i \cdot \nabla V) = \varepsilon \frac{\partial}{\partial r} V_1 + \mathcal{O}(\varepsilon^2)$, the boundary conditions (16.373) are, up to $\mathcal{O}(\varepsilon^2)$,

$$i \omega \frac{\partial \Phi}{\partial r} \pm \frac{\Omega^2 D_0 \Phi}{Z_i} = \varepsilon \omega \kappa R_i' \Phi \pm \frac{i \varepsilon}{\Phi} \left[ U_0 \frac{\partial}{\partial X} + V_1 \frac{\partial}{\partial r} \right] \left( \frac{\Omega D_0 \Phi^2}{Z_i} \right), \quad (16.380)$$

where the plus sign relates to $r = R_1$ and the minus sign to $r = R_2$. We now assume

$$\Phi(X, r; \varepsilon) = \Phi_0(X, r) + \varepsilon \Phi_1(X, r) + \cdots. \quad (16.381)$$

Substitution into (16.377) yields for the $\mathcal{O}(1)$ and $\mathcal{O}(\varepsilon)$ terms, respectively,

$$\mathcal{L} [\Phi_0] = 0, \quad (16.382a)$$

$$D_0 \mathcal{L} [\Phi_1] = \frac{i}{\Phi_0} \left\{ \frac{\partial}{\partial X} \left[ \left( \frac{U_0 \Omega}{C_0^2} + \kappa \right) D_0 \Phi_0^2 \right] + \frac{1}{r} \frac{\partial}{\partial r} \left[ \frac{r V_1 \Omega}{C_0^2} D_0 \Phi_0^2 \right] \right\}, \quad (16.382b)$$
subject to the boundary conditions
\[ i \omega \frac{\partial \Phi_0}{\partial r} \pm \frac{\Omega^2 D_0 \Phi_0}{Z_i} = 0, \tag{16.383a} \]
\[ i \omega \frac{\partial \Phi_1}{\partial r} \pm \frac{\Omega^2 D_0 \Phi_1}{Z_i} = \omega k R'_i \Phi_0 \pm i \frac{1}{\Phi_0} \left[ U_0 \frac{\partial}{\partial X} + V_1 \frac{\partial}{\partial r} - \frac{\partial V_1}{\partial r} \right] \left( \frac{\Omega D_0 \Phi_0^2}{Z_i} \right). \tag{16.383b} \]

The leading-order equation (16.382a) is, up to a radial coordinate stretching, Bessel's equation in \( r \), with \( X \) acting only as a parameter. The mode-like solution we are looking for is then
\[ \Phi_0(X, r) = N(X) \psi(X, r), \tag{16.384} \]
where \( \psi \) is given by
\[ \psi(X, r) = a(X) J_m(\alpha(X)r) + b(X) Y_m(\alpha(X)r) \tag{16.385} \]
and \( J_m \) and \( Y_m \) are the \( m \)th order Bessel function of the first and second kind, respectively [1]. \( \alpha \) is defined by the dispersion relation
\[ \alpha^2 = \Omega^2 / C_0^2 - \kappa^2. \tag{16.386} \]

Boundary conditions (16.383a) produce the following equations for the slowly varying eigenvalue \( \alpha \) considered and corresponding amplitudes \( a \) and \( b \):
\[ \frac{\alpha R_2 J_m(\alpha R_1) - \xi_2 J_m(\alpha R_2)}{\alpha R_2 Y_m(\alpha R_1) - \xi_2 Y_m(\alpha R_2)} = \frac{\alpha R_1 J_m(\alpha R_1) + \xi_1 J_m(\alpha R_1)}{\alpha R_1 Y_m(\alpha R_1) + \xi_1 Y_m(\alpha R_1)} = -\frac{b(X)}{a(X)}. \tag{16.387} \]
where \( \xi_i = \Omega^2 D_0 R_i / i \omega Z_i \). \( \psi \) is scaled such that for any \( X \)
\[ \int_{R_i}^{R_i} \psi^2(X, r) r \, dr = 1. \tag{16.388} \]
This can be done explicitly (see [1]) by using the relation
\[ \int \psi^2 r \, dr = \frac{1}{2} (r^2 - (m/\alpha)^2) \psi^2 + \frac{1}{2} (r/\alpha)^2 \left( \frac{\partial}{\partial r} \psi \right)^2, \tag{16.389} \]
where, of course, \( \frac{\partial}{\partial r} \psi \) can be eliminated by substituting the boundary condition (16.383a).

For notational convenience we introduce the reduced axial wave number
\[ \sigma = \sqrt{1 - \left( C_0^2 - U_0^2 \right) \frac{\alpha^2}{\omega^2}}, \tag{16.390} \]
which is roughly equal to \( \kappa \) for a comoving observer and scaled by \( \omega \), such that
\[ \kappa = \frac{\omega}{C_0} \frac{C_0 \sigma - U_0}{C_0^2 - U_0^2}, \quad \frac{U_0 \omega}{C_0^2 - U_0^2} + \kappa = \frac{\omega \sigma}{C_0}, \quad \Omega = \omega C_0 \left( \frac{C_0 - U_0 \sigma}{C_0^2 - U_0^2} \right), \quad \left( C_0^2 - U_0^2 \right) \frac{\alpha^2}{\omega^2} + \sigma^2 = 1. \]
Finally, we have to determine \( N = N(X) \) by the usual solvability argument applied to \( \Phi_1 \).

Since the operator \( r \mathcal{L} \) is self-adjoint in \( r \), we have

\[
\int_{R_1}^{R_2} \Phi_0 \mathcal{L} \Phi_1 r \, dr = R_2 \left[ \Phi_0 \frac{\partial \Phi_1}{\partial r} - \Phi_1 \frac{\partial \Phi_0}{\partial r} \right]_{r=R_2} - R_1 \left[ \Phi_0 \frac{\partial \Phi_1}{\partial r} - \Phi_1 \frac{\partial \Phi_0}{\partial r} \right]_{r=R_1}.
\]

Further evaluation of this expression (using (16.383b) and the corresponding right-hand side of (16.382b)) gives, finally, after a considerable amount of algebra, the adiabatic invariant (see (15.72))

\[
\frac{d}{dX} \left[ D_0 \frac{\omega \sigma}{C_0} \int_{R_1}^{R_2} \Phi_0^2(X, r) r \, dr + \frac{D_0 U_0}{\Omega} \left( \zeta_2 \Phi_0^2(X, R_2) + \zeta_1 \Phi_0^2(X, R_2) \right) \right] = 0. \tag{16.391}
\]

We have made use of the defining equations (16.374) of \( U_0 \) and \( V_1 \), Leibniz’s rule, and the identity \( U_0 \frac{d}{dX} + V_1 \frac{d}{dX} = U_0 \frac{d}{dX} \) along \( r = R_i \).

Equation (16.391) can be integrated immediately with a constant of integration \( Q^2 \).

This constant is determined by the initial amplitude of the mode entering the duct. Since the solution is linear, it is irrelevant here. The remaining integral over \( \Phi_0^2 \) is absorbed by the normalisation of \( \psi \). We eventually obtain for \( N \) the remarkably simple expression

\[
N(X) = \frac{Q}{\sqrt{\omega \sigma D_0} + \frac{D_0^2 \Omega U_0}{\frac{1}{i \omega}} \left( R_2 \phi_2^2(X, R_2) + R_1 \phi_2^2(X, R_1) \right)}. \tag{16.392}
\]

The chosen branch of the square root is not important (\( Q \) is to be chosen anyway), but we have to make sure to remain at the same branch along the duct.

It should be noted that expression (16.392) is quite general. It is equally valid for no flow (\( U_0 = 0 \)), a hard-walled duct \((Z_i \to \infty)\), and a hollow duct \((R_1 \equiv 0)\).

### 16.16.4 Numerical Evaluation of a Typical Example

The found analytic expression is not entirely explicit, since it requires solving two algebraic equations. Numerical evaluation involves the following steps:

(i) We define the duct in terms of the inner \((R_1)\) and outer \((R_2)\) walls, from which we obtain the cross section \( A = \pi (R_2^2 - R_1^2) \).

(ii) We define the mean flow parameters \( E \) and \( F \) and solve (16.375) for \( D_0 \) at any \( X \). This defines the axial mean flow \( U_0 \). In aircraft applications at takeoff or approach we will be interested in a Mach number between 0.2 and 0.5.

(iii) We select the Helmholtz number \( \omega \). This is determined by the shaft rotational speed being such that the fan blade tip speed is just about transonic and depends on the diameter of the duct and the harmonic order of interest (one, two, or sometimes three times the blade passing frequency). Usual values for high-bypass engines vary between 20 and 70.
16.16. Attenuation of Sound in Aircraft Engine Ducts

(iv) We select the prevailing circumferential mode number \( m \). This number is related to the number of fan blades and stator vanes and also depends on the frequency of the fan we are interested in. Typical values vary between 10 and 30.

(v) We select a class of candidate wall impedances. Hard walls correspond to \(|Z| = \infty\) and any passive wall has \( \text{Re}(Z) \geq 0 \), while a purely imaginary wall is not practically feasible. A typical impedance is found in a rectangle of about \( 0 < \text{Re}(Z) < 6, -6 < \text{Im}(Z) < 6 \). Furthermore, if we consider several frequencies, we should take account of \( Z \)'s inherent dependence on \( \omega \). This requires knowledge of the way the lining is actually constructed.

(vi) Finally, we determine (or guess) the source, i.e., the amplitudes of the radial modes. For each relevant mode we have to solve the eigenvalue equation (16.387) and trace \( \alpha \) as a function of \( X \). This gives us the axial wave number \( \kappa \), which has to be integrated with respect to \( X \) for the modal phase (16.376).

We computed numerically a plot (Figure 16.52) of a pressure field (at a given time) for a typical inlet duct given by

\[
\begin{align*}
R_2(x) &= 1 - 0.18453 \xi^2 + 0.10158 (e^{-11(1-\xi)} - e^{-11})/(1 - e^{-11}), \\
R_1(x) &= \max\left[0, 0.64212 - (0.04777 + 0.98234 \xi^2)^{1/2}\right],
\end{align*}
\tag{16.393}
\]

where \( \xi = x/L \) and \( L = 1.86393 \) is the length of the duct. (Note that \( R_i \) is written as a function of \( x \) rather than \( X \).) The corresponding value of \( \varepsilon \) is about 0.1, since \( R'_2 \) varies between \(-0.1\) and \( 0.1 \) over \( 0 \leq x \leq 1.73 \), the larger part of the interval. Although \(|R'_1|\) attains higher values near the spinner top (almost 0.5), this is not as relevant because the sound field practically vanishes near the duct axis due to the usual high values of \( m \), while the mode behaves like \( \sim r^m \) for \( r \to 0 \).

By selecting \( F = -0.2462 \) and \( E = 2.545 \), we obtain a Mach number \( M = -0.3 \) at \( x = 0 \) (in the picture: flow is from right to left). The outer wall impedance is \( Z_2 = 2 - i \) while the inner wall (the spinner) is hard. The acoustic field is given by \( m = 10 \)

\begin{figure}[h]
\centering
\includegraphics{Figure_16.52.png}
\caption{Acoustic pressure field in an aircraft engine inlet duct in the \((x, r)\) plane.}
\end{figure}
and \( \omega = 25 \), while a right-running radial mode is initially selected with the eigenvalue \( \alpha = 14.834 + 0.645i \) and axial mode number \( \kappa = 30.905 - 0.464i \). The total attenuation between source and inlet plane in this example is 9.48 dB.

### 16.16.5 Discussion and Related Problems

- The present example is a shortened version of [127], while the simplification using a normalised \( \psi \) is taken from [132]. The main result (16.392) is to be compared with equation (49) in [132], including the correction of a misprinted minus sign.
- The present solution may be generalised to very challenging problems of nonpotential flows with mean swirl [28], noncircular ducts [132, 113] and nonhomentropic flow [27].
- The present solution breaks down where the mode passes a point where the denominator of (16.392) vanishes. This typically happens for hard-walled ducts \( (Z_i = \infty) \) at a point of “duct resonance” (this is only a real resonance in the case of a straight duct), where \( \sigma = 0 \). As the mode cannot continue and has to reflect, this point is called a turning point. A local analysis is required where incident and reflected mode couple [129, 109, 110].
- The acoustic design of aircraft engines is usually done by fully numerical modeling, allowing a more flexible variation of geometry and mean flow properties. Nevertheless, the present solution has been shown to be a very interesting alternative [128] and served as an independent check for a number of numerical solutions. It also provides a tool to produce a matching procedure to connect computational fluid dynamics type source data to the acoustic field [111].
- For high frequencies that occur in aircraft applications the eigenvalues defined by (16.387) may be classified in two groups [131]. One consists of typical duct modes, with the imaginary parts of \( \alpha \) not large and \( \kappa \) remaining near a corresponding hard-wall value. The other consists of surface waves (for each \( m \), two without mean flow and four with mean flow), where the imaginary part of \( \alpha \) is large and \( \kappa \) strays in the complex plain, detached from any hard-wall mode. The surface waves are usually more difficult to find.

### 16.16.6 Exercises

16.64. Give an explicit, full expression of \( \psi \), i.e., including the normalization, and compare the resulting expression of \( N \) with equations (4.11) and (4.14) of [127].
16.65. Give a detailed derivation of (16.391), the adiabatic invariant.
16.66. The present solution is derived for \( \omega = \mathcal{O}(1) \) but appears to be correct for relatively large \( \omega \). What analysis would result for small \( \omega \)? Compare with Webster’s horn equation described in Exercise 15.6 of Chapter 15.
Appendices

Useful Definitions and Properties

A Asymptotic Order Symbols

A function $f(\varepsilon)$ may be expressed asymptotically for small $\varepsilon$ by another (usually simpler) function $\varphi(\varepsilon)$ as follows.

Definition A.1.

1. $f(\varepsilon) = O(\varphi(\varepsilon))$ as $\varepsilon \to 0$ if there are positive constants $K$ and $\varepsilon_1$ (both independent of $\varepsilon$) such that
   \[ |f(\varepsilon)| \leq K|\varphi(\varepsilon)| \quad \text{for} \quad 0 < \varepsilon < \varepsilon_1. \]
   We say, “$f$ is big-O of $\varphi$ as $\varepsilon$ tends to zero.”

2. $f(\varepsilon) = o(\varphi(\varepsilon))$ as $\varepsilon \to 0$ if for every $\delta > 0$ there is an $\varepsilon_1$ (independent of $\varepsilon$) such that
   \[ |f(\varepsilon)| \leq \delta|\varphi(\varepsilon)| \quad \text{for} \quad 0 < \varepsilon < \varepsilon_1. \]
   We say, “$f$ is little-o of $\varphi$ as $\varepsilon$ tends to zero.”

3. $f(\varepsilon) = O_s(\varphi(\varepsilon))$ as $\varepsilon \to 0$ if $f(\varepsilon) = O(\varphi(\varepsilon))$ and $f(\varepsilon) \neq o(\varphi(\varepsilon))$. Note that quite often $O$ is used where $O_s$ is actually meant. Further, there is no uniform terminology. We could say, “$f$ is big sharp of $\varphi$ as $\varepsilon$ tends to zero.”

Theorem A.2.

1. If $f(\varepsilon) = o(\varphi(\varepsilon))$ as $\varepsilon \to 0$, then also $f(\varepsilon) = O(\varphi(\varepsilon))$.

2. If the limit $\lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{\varphi(\varepsilon)}$ exists as a finite number, then $f(\varepsilon) = O(\varphi(\varepsilon))$ as $\varepsilon \to 0$.

3. If the limit $\lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{\varphi(\varepsilon)}$ exists as a finite nonzero number, then $f(\varepsilon) = O_s(\varphi(\varepsilon))$ as $\varepsilon \to 0$.

4. If $\lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{\varphi(\varepsilon)} = 0$, then $f(\varepsilon) = o(\varphi(\varepsilon))$ as $\varepsilon \to 0$. 

629
Appendices. Useful Definitions and Properties

Proof. The proof is trivial.

Example A.3

\[ \varepsilon \sin(\varepsilon) = O(\varepsilon), \quad \varepsilon \to 0, \]
\[ \varepsilon \cos(\varepsilon) = O(1), \quad \varepsilon \to 0, \]
\[ \varepsilon^n = o(1), \quad \varepsilon \to 0, \quad \text{for any positive } n, \]
\[ e^{-1/\varepsilon} = o(\varepsilon^n), \quad \varepsilon \to 0, \quad \text{for any positive } n. \]

From this last example, \( e^{-1/\varepsilon} \) is called a transcendentally small term or exponentially small term and can be ignored asymptotically against any power of \( \varepsilon \).

\[ \square \]

B Trigonometric Relations

The real or imaginary parts of the binomial series \( (e^{ix} \pm e^{-ix})^n = \sum_{k=0}^{n} \binom{n}{k} e^{i(n-2k)x} \) easily yield trigonometric relations that are useful for recognising resonance terms:

\[ \sin^2 x = \frac{1}{2} (1 - \cos 2x), \quad \cos^2 x = \frac{1}{2} (1 + \cos 2x),\]
\[ \sin x \cos x = \frac{1}{2} \sin 2x, \quad \sin x = \frac{1}{8} (2 \sin 2x + \sin 4x),\]
\[ \cos^3 x = \frac{1}{2} (1 + \cos 2x), \quad \cos^4 x = \frac{1}{8} (3 + 4 \cos 2x + \cos 4x),\]
\[ \sin^3 x = \frac{1}{4} (3 \sin x - \sin 3x), \quad \sin^5 x = \frac{1}{16} (10 \sin x - 5 \sin 3x + \sin 5x),\]
\[ \sin^2 x \cos x = \frac{1}{4} (\cos x - \cos 3x), \quad \sin^4 x \cos x = \frac{1}{16} (2 \cos x - 3 \cos 3x + \cos 5x),\]
\[ \sin x \cos^2 x = \frac{1}{4} (\sin x + \sin 3x), \quad \sin^3 x \cos^2 x = \frac{1}{16} (2 \sin x + \sin 3x - \sin 5x),\]
\[ \cos^3 x = \frac{1}{4} (3 \cos x + \cos 3x), \quad \cos^4 x = \frac{1}{16} (2 \cos x - 3 \cos 3x - \cos 5x),\]
\[ \sin^4 x = \frac{1}{8} (3 - 4 \cos 2x + \cos 4x), \quad \sin x \cos^4 x = \frac{1}{16} (2 \sin x + 3 \sin 3x + \sin 5x),\]
\[ \sin^3 x \cos x = \frac{1}{8} (2 \sin 2x - \sin 4x), \quad \cos^5 x = \frac{1}{16} (10 \cos x + 5 \cos 3x + \cos 5x).\]

C Convergence of Series

The series

\[ S(x) := \sum_{n=0}^{\infty} c_n f_n(x), \quad x \in \Omega, \quad (C.1) \]
is said to converge \textit{pointwise} in \( x \in \Omega \) if we can find for any given \( \varepsilon > 0 \) a sufficiently large number \( N \in \mathbb{N} \) such that the remaining part of the series is smaller than \( \varepsilon \); i.e.,

\[
\left\| S(x) - \sum_{n=0}^{N} c_n f_n(x) \right\| < \varepsilon.
\]

In general, \( N \) depends on \( \varepsilon \) and \( x \). If \( N \) may be chosen independent of any \( x \in \Omega \), the series is said to converge \textit{uniformly} in \( \Omega \). A sufficient condition for uniform convergence is Weierstrass’s M-test. If

\[
\left\| c_n f_n(x) \right\| \leq M_n \quad \text{for all } x \in \Omega
\]

and

\[
\sum_{n=0}^{\infty} M_n
\]

converges, then (C.1) is uniformly convergent.

The concept of uniform convergence is important because of the following properties. Suppose (C.1) is uniformly convergent.

(i) If all \( f_n(x) \) are continuous, then \( S(x) \) is continuous and

\[
\lim_{x \to a} S(x) = \sum_{n=0}^{\infty} c_n \lim_{x \to a} f_n(x).
\]

(ii) If all \( f_n(x) \) are continuous, then \( S(x) \) is integrable and

\[
\int S(x) \, dx = \sum_{n=0}^{\infty} c_n \int f_n(x) \, dx.
\]

(iii) If all \( f_n(x) \) are differentiable, then \( S(x) \) is differentiable and

\[
\frac{d}{dx} S(x) = \sum_{n=0}^{\infty} c_n \frac{d}{dx} f_n(x).
\]

\section{D. Multistep Formulas}

If the grid points \( \xi_i \) are equispaced, a simple relation exists between interpolating polynomials and \textit{backward difference} operators, which in turn give explicit expressions for the coefficients of multistep methods.

The backward difference operators \( \nabla^k \) for \( k = 0, 1, \ldots \) are defined for a set \( f_1, \ldots, f_m \) by

\[
\nabla^0 f_j := f_j, \quad \nabla^l f_j := \nabla^{l-1} f_j - \nabla^{l-1} f_{j-1}, \quad l = 1, 2, \ldots \quad \text{(D.1)}
\]

So we have, e.g.,

\[
\nabla^1 f_j = f_j - f_{j-1}, \quad \nabla^2 f_j = (f_j - f_{j-1}) - (f_{j-1} - f_{j-2}) = f_j - 2f_{j-1} + f_{j-2}.
\]
From this we find
\[ f_{m-l} = \sum_{j=0}^{l} (-1)^j \binom{l}{j} \nabla^j f_m. \] (D.2)

The backward difference notation is the same as the nabla operator. However, no confusion should occur as we only use backward difference operators in this Appendix D.

Let \( \xi_1, \ldots, \xi_m \) be equispaced grid points with grid size \( h \) and let \( f_j \) be a function value at the point \( \xi_j \). Then we can rewrite the interpolation polynomial in terms of the \( \nabla^j f_m \) as
\[ p(\xi) = \sum_{j=0}^{m-1} (-1)^j \binom{-s}{j} \nabla^j f_m, \] (D.3)

where \( s := \frac{\xi - \xi_m}{h} \).

Now consider the ODE
\[ \frac{dx}{dt} = f(x, t). \]

We want to find an approximation of the solution \( x(t_j) \) on a set of equispaced grid points \( t_0, t_1, \ldots, t_i, t_{i+1}, \ldots \). At the point \( t_{i+1} \) we can easily find an approximation of \( \frac{dx}{dt} (t_{i+1}) \) from (D.3) by differentiation. As a result, we find, identifying the points \( \xi_j \) with \( t_{i-m+1+j} \) and denoting the numerical approximation of \( x(t_{i+1}) \) by \( x_{i+1} \), that
\[ \sum_{j=0}^{m-1} \nabla^j x_{i+1} = hf(x_{i+1}, t_{i+1}). \] (D.4a)

The coefficients in (D.4a) have a nice form. One can check that
\[ \gamma_0 = 0, \]
\[ \gamma_j = \frac{1}{j}, \quad j \geq 0. \] (D.4b)

The formulas found in (D.4) are the backward difference formulas (BDF). In standard form (cf. (5.75)) they are given by
\[ \sum_{j=0}^{k} x_{i-j+1} = hf(t_{i+1}, x_{i+1}). \] (D.5)

For \( k = 1, 2, 3, 4 \) the coefficients \( \alpha_j \) are given in Table D.1. Note that for \( k = 1 \) we have the Euler backward method. Also, for other multistep methods, the formulation in terms of backward differences enables one to find the coefficients in a simple way:
\[ x(t_{i+1}) - x(t_i) = \int_{t_i}^{t_{i+1}} f(x(\tau), \tau) \, d\tau. \] (D.6)
D. Multistep Formulas

Table D.1. Coefficients for the BDF:

<table>
<thead>
<tr>
<th>( k )</th>
<th>( a_0 )</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( a_3 )</th>
<th>( a_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3/2</td>
<td>-2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11/6</td>
<td>-3</td>
<td>3/2</td>
<td>-1/3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>25/12</td>
<td>-4</td>
<td>3</td>
<td>-4</td>
<td>1</td>
</tr>
</tbody>
</table>

If we apply (D.3) on the interval \((t_i-k+1, t_i)\), i.e., approximate \( f(x(t), t) \) by a polynomial \( p \) of degree \( k-1 \) there, we obtain

\[
x(t_{i+1}) - x(t_i) = h \sum_{j=0}^{k-1} \hat{\gamma}_j \nabla^j f_i, \tag{D.7a}
\]

where

\[
\hat{\gamma}_j := (-1)^j \int_0^1 \left( -\frac{\tau^j}{j} \right) d\tau. \tag{D.7b}
\]

This is the \( k \)-step Adams–Bashforth formula. The \( \hat{\gamma}_j \) are simply calculated by recursion:

\[
\hat{\gamma}_0 = 1, \\
\hat{\gamma}_j = 1 - \frac{1}{j+1} \hat{\gamma}_0 - \frac{1}{j} \hat{\gamma}_1 - \cdots - \frac{1}{2} \hat{\gamma}_{j-1}, \quad j \geq 1. \tag{D.7c}
\]

In standard form we have

\[
x(t_{i+1}) - x(t_i) = h \sum_{j=1}^{k} b_j f_{i-j+1}. \tag{D.8}
\]

For \( k = 1, 2, 3, 4 \) the coefficients are given in Table D.2. If we apply (D.3) on the interval \((t_i-k+1, t_{i+1})\), i.e., approximate \( p \) by a polynomial of degree \( k \) there, we obtain

\[
x(t_{i+1}) - x(t_i) = h \sum_{j=0}^{k} \bar{\gamma}_j \nabla^j f_{i+1}, \tag{D.9a}
\]

where

\[
\bar{\gamma}_j = (-1)^j \int_{-1}^{0} \left( -\frac{\tau^j}{j} \right) d\tau. \tag{D.9b}
\]

This is the \( k \)-step Adams–Moulton formula.
Table D.2. Coefficients for the Adams–Bashforth formula.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{3}{2}$</td>
<td>$-\frac{1}{2}$</td>
<td>$\frac{3}{12}$</td>
<td>$\frac{16}{12}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{23}{12}$</td>
<td>$\frac{1}{12}$</td>
<td>$\frac{5}{12}$</td>
<td>$\frac{1}{12}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{55}{24}$</td>
<td>$\frac{19}{24}$</td>
<td>$\frac{37}{24}$</td>
<td>$\frac{9}{24}$</td>
</tr>
</tbody>
</table>

From this we find

\[
\bar{\gamma}_0 = 1,
\]

\[
\bar{\gamma}_j = -\frac{1}{j+1} \bar{\gamma}_0 - \frac{1}{j} \bar{\gamma}_1 - \cdots - \frac{1}{2} \bar{\gamma}_{j-1}, \quad j \geq 1. \tag{D.9c}
\]

The coefficients $\hat{\gamma}_j$ and the $\bar{\gamma}_j$ are independent of $k$. They are related by

\[
\sum_{j=0}^{k} \bar{\gamma}_j = \hat{\gamma}_k. \tag{D.10}
\]

We can write the Adams–Moulton formulas in the standard formulation as

\[
x_{i+1}^h - x_i^h = h \sum_{j=0}^{k} b_j f_i - j + 1. \tag{D.11}
\]

For $k = 0, 1, 2, 3$ the coefficients $b_j$ are given in Table D.3. Note that for $k = 0$ we obtain the Euler backward method and for $k = 1$ we find the trapezoidal formula (both one step!).

Table D.3. Coefficients for the Adams–Moulton formula.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{12}$</td>
<td>$\frac{1}{12}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{5}{12}$</td>
<td>$\frac{8}{12}$</td>
<td>$\frac{1}{12}$</td>
<td>$\frac{1}{12}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{0}{24}$</td>
<td>$\frac{19}{24}$</td>
<td>$\frac{5}{24}$</td>
<td>$\frac{1}{24}$</td>
</tr>
</tbody>
</table>
E. Solution of Recursions

Consider the recursion
\[ x_{i+1} = A_i x_i + b_i, \quad (E.1) \]
where \( \{A_i\} \) is a set of square matrices (in particular it may be a set of scalars) and \( \{b_i\} \) is a set of vectors. Then the solution of (E.1) is given by
\[ x_i = \left( \prod_{j=0}^{i-1} A_j \right) x_0 + \sum_{m=0}^{i-1} \left( \prod_{j=m+1}^{i-1} A_j \right) b_j. \quad (E.2) \]
In (E.2) one has to interpret \( \prod_{i-1}^{m+1} A_j \) as \( A_{i-1} A_{i-2} \cdots A_{m+1} \) and as \( \mathbf{I} \) if \( i - 1 < m + 1 \). Now consider the second order scalar recursion
\[ x_{i+1} = a_i x_i + b_i x_{i-1} + c_i. \quad (E.3) \]
If \( c_i \equiv 0 \) and \( a_i \) and \( b_i \) are constant, then the solution is given by
\[ x_i = \alpha \lambda_1^i + \beta \lambda_2^i, \quad \alpha, \beta \in \mathbb{R}, \quad (E.4) \]
if the characteristic equation
\[ \lambda^2 - a \lambda - b = 0 \quad (N.B. \ \forall a_i = a, \ \forall b_i = b) \quad (E.5) \]
has two different roots \( \lambda_1, \lambda_2 \). If it has a double root, the solution can be written as
\[ x_i = (\alpha + \beta i) \lambda_1^i. \quad (E.6) \]
The constants in (E.4) and (E.6) are to be found from initial or boundary values.

In general the homogeneous part of (E.3) has two independent basis solutions \( \{f_i\} \) and \( \{g_i\} \). In terms of these we can give a formal solution of (E.3). We find
\[ x_i = \sum_{j=1}^{i} \left( f_{j-1} g_l - g_{j-1} f_l \right) c_j + \alpha f_i + \beta g_i, \quad (E.7) \]
where \( \alpha \) and \( \beta \) are to be determined from initial or boundary conditions. In particular, for constant coefficients and two different roots \( \lambda_1, \lambda_2 \) we obtain
\[ x_i = \sum_{j=1}^{i} c_j \frac{\lambda_2^{i-j} - \lambda_1^{i-j}}{\lambda_2 - \lambda_1} + \alpha \lambda_1^i + \beta \lambda_2^i. \quad (E.8) \]

F. Eigenvalues and Eigenvectors of a Tridiagonal Matrix

Consider the matrix \( A \in \mathbb{R}^{n \times n} \):
\[ A = \begin{pmatrix} b & c & \cdots & \cdots & a \cr a & b & c & \cdots & \cdots \cr \cdots & \cdots & \ddots & \ddots & \cdots \cr a & b & c \end{pmatrix}. \quad (F.1) \]
Appendices. Useful Definitions and Properties

Let \( \mu \) be an eigenvalue of \( A \) and \( x = (x_1, \ldots, x_N)^T \) be a corresponding eigenvector. Then we find from \( Ax = \mu x \) that

\[
(b - \mu)x_1 + cx_2 = 0, \quad (F.2a)
\]
\[
ax_{j-1} + (b - \mu)x_j + cx_{j+1} = 0, \quad j = 2, \ldots, N - 1, \quad (F.2b)
\]
\[
ax_{N-1} + (b - \mu)x_N = 0. \quad (F.2c)
\]

Let us take \( x_0 = x_{N+1} = 0 \). Then we find from (F.2) that \( \{x_j\}_{0}^{N+1} \) is a solution of the three-term recurrence equation

\[
ax_{j-1} + (b - \mu)x_j + cx_{j+1} = 0, \quad j = 1, \ldots, N, \quad (F.3)
\]

with boundary values

\[
x_0 = x_{N+1} = 0. \quad (F.4)
\]

The recursion (F.3) has the characteristic polynomial

\[
\lambda^2 + (b - \mu)\lambda + c = 0 \quad (F.5)
\]

with roots \( \lambda_1 \) and \( \lambda_2 \) say, so that a solution is formally given by

\[
x_j = \alpha \lambda_1^j + \beta \lambda_2^j, \quad (F.6)
\]

where \( \alpha \) and \( \beta \) can be found from the boundary values (F.4). This gives \( \beta = -\alpha \) and eventually the relation

\[
\left( \frac{\lambda_1}{\lambda_2} \right)^{N+1} = 1; \quad (F.7)
\]

i.e., \( \lambda_1/\lambda_2 \) is an \( (N+1) \)th power unit root. Hence

\[
\frac{\lambda_1}{\lambda_2} = e^{\frac{2\pi i}{N}}, \quad l = 1, \ldots, N. \quad (F.8)
\]

Since the product of the roots equals \( \frac{c}{a} \), we obtain

\[
\lambda_1 = \left( \frac{c}{a} \right)^{1/2} e^{\frac{2\pi i}{N}}, \quad (F.9a)
\]
\[
\lambda_2 = \left( \frac{c}{a} \right)^{1/2} e^{-\frac{2\pi i}{N}}. \quad (F.9b)
\]

We now use the fact that \( \lambda_1 + \lambda_2 = -(b - \mu)/a \), so that (F.9) gives for the eigenvalues \( \mu_l \)

\[
\mu_l = b + \sqrt{ac} \left( e^{\frac{2\pi i}{N}} + e^{-\frac{2\pi i}{N}} \right) = b + 2\sqrt{ac} \cos \left( \frac{l\pi}{N+1} \right), \quad l = 1, \ldots, N. \quad (F.10)
\]

For the \( j \)th component \( x_j \) of the corresponding eigenvector \( x_l \) we apparently have

\[
x_j = \alpha(\lambda_1^j - \lambda_2^j) = \alpha \left( \frac{c}{a} \right)^{j/2} \left( e^{\frac{2\pi i}{N}} + e^{-\frac{2\pi i}{N}} \right) = 2i\alpha \left( \frac{c}{a} \right)^{j/2} \sin \left( \frac{\pi jl}{N+1} \right). \quad (F.11)
\]

Note that \( \alpha \) can be chosen arbitrarily, so, in particular, if \( \text{sign}(a) = \text{sign}(b) \), we may assume \( x_j \) to be real for all \( l \) and \( j \). If, moreover, \( a = c \), we can choose \( \alpha \) such that \( \sum x_j^2 = 1 \), resulting in an orthogonal set of eigenvectors with

\[
x_j = \frac{2}{\sqrt{N}} \sin \left( \frac{\pi jl}{N+1} \right). \quad (F.12)
\]
Remark 1. We still have to check whether the representation (F.6) is correct. Indeed, if \( \lambda_1 = \lambda_2 \) (so we have a double root), the foregoing is not correct. This situation cannot occur, however. Suppose we have \( \lambda_1 = \lambda_2 \). Then we would have, instead of (F.6),

\[
x_j = (\alpha + \beta_j)\lambda_1^j.
\]

(F.13)

From (F.4) we then immediately see that \( \alpha = \beta = 0 \), which is not interesting, of course.

Remark 2. We can find out more generally whether a tridiagonal matrix has (geometrically) multiple eigenvalues and real eigenvectors (so it is diagonisable) or not. Indeed consider instead of (F.1) the matrix

\[
A = \begin{pmatrix}
b_1 & c_1 \\
a_2 & b_2 & c_2 \\ & \ddots & \ddots \\
 & & a_{N-1} & b_{N-1} & c_{N-1} \\
 & & & a_N & b_N
\end{pmatrix}.
\]

(F.14)

Let \( D \) be a diagonal matrix \( D = \text{diag}(d_1, \ldots, d_N) \), with \( d_1, \ldots, d_N \) still to be chosen. We define

\[
\tilde{A} = DAD^{-1}.
\]

(F.15)

We now require \( D \) to be such that \( \tilde{A} \) is a symmetric matrix. This is so if

\[
\left( \frac{d_{j+1}}{d_j} \right)^2 = \frac{c_j}{a_{j+1}}, \quad 1 \leq j \leq N - 1.
\]

(F.16)

Hence if, e.g., \( \text{sign}(c_j) = \text{sign}(a_{j+1}) \), such a matrix \( D \) certainly exists. Apparently, \( \tilde{A} \) has simple eigenvalues. Due to \( A \) and \( \tilde{A} \) being similar, this property carries over to \( A \).

G Norms

Let \( V \) be a linear vector space.

Definition G.1. A norm on \( V \), denoted by \( \| \cdot \| \), satisfies the following four conditions:

(i) \( \| \cdot \| \in \mathbb{R} \) and \( \| \cdot \| \geq 0 \).
(ii) \( \| x \| = 0 \) if and only if \( x = 0 \).
(iii) \( \| \gamma x \| = \| \gamma \| \| x \| \) for any \( \gamma \in \mathbb{R} \).
(iv) \( \| x + y \| \leq \| x \| + \| y \| \) for any \( x, y \in V \) (triangle inequality).

If \( V = \mathbb{R}^n \) and \( x \in \mathbb{R}^N \) denotes a vector with coordinates \( x_1, \ldots, x_N \), say, then we
often use Hölder norms:

\[ \|x\|_1 = \sum_{j=1}^{N} |x_j|, \quad \text{(G.1a)} \]

\[ \|x\|_2 = \left( \sum_{j=1}^{N} |x_j|^2 \right)^{1/2}, \quad \text{(G.1b)} \]

\[ \|x\|_\infty = \max_{1 \leq j \leq N} |x_j|. \quad \text{(G.1c)} \]

The norms \( \| \cdot \|_\alpha \) and \( \| \cdot \|_\beta \) for some \( \alpha \) and \( \beta \) are called equivalent if there exist \( c_1, c_2 \in \mathbb{R}_+ \) such that for all \( x \) \( c_1 \|x\|_\alpha \leq \|x\|_\beta \leq c_2 \|x\|_\alpha \). For \( N < \infty \) all norms are equivalent. In particular, we have

\[ \|x\|_2 \leq \|x\|_1 \leq \sqrt{N} \|x\|_2, \quad \text{(G.2)} \]

\[ \|x\|_\infty \leq \|x\|_2 \leq \sqrt{N} \|x\|_\infty. \]

The bounds above are attainable. Hence there is no longer an equivalence if \( N \to \infty \).

A consequence of equivalence is that a series that converges in one norm also converges in an equivalent norm. If \( V = \mathbb{R}^N \), this implies that convergence considerations are norm independent.

A useful property is the Cauchy–Schwarz inequality: For all \( x, y \in \mathbb{R}^N \) we have

\[ |(x, y)| \leq \|x\|_2 \|y\|_2 \]  \quad (\cdot, \cdot \text{ is the Euclidean inner product}). \quad \text{(G.3)}

Next we consider matrix norms. Let \( V \) be the linear space consisting of matrices. A norm on \( V \) satisfies conditions similar to those in Definition G.1(i), …, (iv). A vector norm induces an associated matrix norm in a natural way as follows:

\[ \|A\| := \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}. \quad \text{(G.4)} \]

As one can easily verify, we have

\[ \|A\| = \max_{\|x\|=1} \|Ax\|. \]

On top of (i), …, (iv), such an associated norm apparently also has a multiplicity property

\( \|AB\| \leq \|A\|\|B\|. \)

The most often used associated matrix norms are

\[ \|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|, \quad \text{(G.5a)} \]

\[ \|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|, \quad \text{(G.5b)} \]

\[ \|A\|_2 = (\rho(A^TA))^{1/2}, \quad \text{(G.5c)} \]
where $\rho(B)$ is the absolutely largest eigenvalue of $B$; see Definition H.3.

The 2-norm is especially interesting because it is orthogonally invariant; i.e., if $Q_1$ and $Q_2$ are orthogonal matrices, then the following holds:

$$\|Q_1AQ_2\|_2 = \|A\|_2.$$ 

Also for matrices one can show equivalence of norms, at least for finite dimension. In particular we have

$$\frac{1}{\sqrt{N}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{N}\|A\|_1, \quad (G.6)$$

$$\frac{1}{\sqrt{N}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{N}\|A\|_\infty.$$ 

If $V$ is the linear space of scalar functions $x(t)$, defined on an interval $[\alpha, \beta]$ say, we can introduce analogues of the Hölder norm for the continuous case:

$$\|x\|_p := \left(\int_\alpha^\beta |x(t)|^p \, dt\right)^{\frac{1}{p}}. \quad (G.7)$$

Clearly, if $V$ is a space of vector functions, one has to replace the modulus by a suitable vector norm in (G.7). For $p = \infty$ we have

$$\|x\|_\infty = \sup_{t \in [\alpha, \beta]} |x(t)|. \quad (G.8)$$

Each norm on a linear space $V$ generates a metric $d$ on $V$ by $d(x, y) := \|x - y\|$, $x, y \in V$. This metric has the property of translation invariance: $d(x + z, y + z) = d(x, y)$.

**H  Similarity**

Let $A$ be a matrix. If

$$Ax = \lambda x \quad (H.1)$$

for some vector $x$ and scalar $\lambda$, then $\lambda$ is called an eigenvalue and $x$ an eigenvector (belonging to $\lambda$) of $A$.

**Property H.1.**

(i) An eigenvalue $\lambda$ is a zero of the characteristic polynomial $\det(A - \lambda I)$.

(ii) The product of the eigenvalues of $A$ is equal to $\det(A)$.

(iii) The sum of the eigenvalues of $A$ is equal to $\sum_{j=1}^n a_{jj}$, the trace of $A$.

For every matrix $A$ a nonsingular matrix $T$ exists such that

$$T^{-1}AT = J, \quad (H.2)$$

where $J$ is a bidirectional matrix consisting of blocks containing the same eigenvalues whose
dimensions correspond to the (algebraic) multiplicity of those eigenvalues; i.e.,

$$J := \begin{pmatrix}
\lambda_1 & 1 & & & & \\
& & \ddots & & & \\
& & 1 & \lambda_1 & & \\
& & & \ddots & \ddots & \\
& & & & \ddots & 1 \\
& & & & & \emptyset & \lambda_p
\end{pmatrix}. \quad (H.3)$$

The matrices $T$ in (H.2) are often chosen such that the “dots” in (H.3) are all equal to one. From a numerical point of view this may not be so meaningful (as it may cause $T$ to be very skew). The form (H.2) is called the Jordan normal form. The geometric multiplicity of an eigenvalue is the dimension of the space of independent eigenvectors. If, in particular, the algebraic and geometric multiplicity are the same for all eigenvalues, then $J$ is a diagonal matrix and each column of $T$ is then an eigenvector.

If $A$ is symmetric, i.e., $A = A^T$, or skew symmetric, i.e., $A = -A^T$, the transformation matrix $T$ is orthogonal.

The transformation (H.2) is a special instance of a matrix similarity transformation. If $S$ is nonsingular, the matrix $S^{-1}AS = B$ is called similar to $A$. Of course it corresponds to viewing a mapping on a different basis.

**Property H.2.** Similar matrices have the same eigenvalues.

**Definition H.3.** The absolute value of the absolutely largest eigenvalue of a matrix is called the spectral radius of $A$ and is denoted as $\rho(A)$.

**Property H.4.** If $A$ is symmetric, then $\|A\|_2 = \rho(A)$. If $A$ is not symmetric, then $\|A\|_2 = (\rho(A^T A))^{1/2}$.

## I Estimates of Eigenvalues and Consequences

The spectral radius of a matrix is a useful means of deciding a number of properties. In particular, it is sometimes crucial to know whether this radius is at least smaller than one. This is why we have the following theorems.

**Theorem I.1.** $\rho(A) < 1$ if and only if $\lim_{i \to \infty} A^i = 0$.

**Proof.** First we prove the “if” direction:

$$(\rho(A))^i = \rho(A^i) \leq \|A^i\|^2.$$ 

Now we deal with the “only if” direction: Consider the Jordan normal form. Then $A^i = T J^i T^{-1}$. By studying a single Jordan block and splitting it into diagonal and codiagonal parts, we can easily complete the proof. \qed
I. Estimates of Eigenvalues and Consequences

**Theorem I.2 (Neumann series).** \( \lim_{n \to \infty} \sum_{i=0}^{n} A^i \) exists if and only if \( \rho(A) < 1 \). The following holds: \( \sum_{i=0}^{\infty} A^i = (I - A)^{-1} \).

**Proof.** If \( \{ \sum_{i=0}^{n} A^i \} \) converges, then \( A^i \to 0 \), which implies that \( \rho(A) < 1 \). If, on the other hand, \( \rho(A) < 1 \), then \( \det(I - A) \neq 0 \), so \( (I - A)^{-1} \) exists. For each \( n \) we have \( (I + A + \cdots + A^n)(I - A) = I - A^{n+1} \).

**Corollary I.3.** The matrix \( I - A \) is nonsingular if for any norm \( \|A\| < 1 \).

**Definition I.4.** An eigenvalue \( \lambda \) is said to have algebraic multiplicity \( p \) if the characteristic polynomial of the matrix has precisely \( p \) roots \( \lambda \). The eigenvalue \( \lambda \) is said to have geometric multiplicity \( q \) if there exists a \( q \)-dimensional subspace of eigenvectors belonging to \( \lambda \). A matrix is called defect if the geometric multiplicity is smaller than the algebraic multiplicity.

Clearly a matrix is defect if and only if it has Jordan blocks of size larger than one.

**Definition I.5.** A matrix is called stable if \( \rho(A) \leq 1 \) and the geometric multiplicity of the eigenvalues with modulus one is equal to the algebraic multiplicity.

Note that a stable matrix may have more than one eigenvalue equal to one. The identity matrix \( I \) is the best-known example.

**Property I.6.** If \( A \) is stable, then there exists a constant \( \kappa \), say, such that for all \( x \in \mathbb{R}^N \|A^i x\| \leq \kappa \|x\| \) (this \( \kappa \) depends on \( N \) and the norm chosen).

Finally we give an important theorem to estimate the eigenvalues.

**Theorem I.7 (Gershgorin).** An eigenvalue of \( A \) lies in at least one of the closed discs with centre point \( a_{ii} \) and radius \( \sum_{j \neq i} |a_{ij}| \). (Note that the discs are in \( \mathbb{C} \).

**Proof.** Let \( \lambda \) be an eigenvalue and \( x = (x_1, \ldots, x_N)^T \) be a corresponding eigenvector. Then

\[
Ax = \lambda x \Rightarrow (\lambda - a_{rr})x_r = \sum_{j \neq r} a_{rj}x_j.
\]

Hence \( |\lambda - a_{rr}| \leq \sum_{j \neq r} |a_{rj}| \) for all \( r \). If \( m \) is such that \( |x_m| = \max_r |x_r| \) (note that \( x_m \neq 0 \), then

\[
|\lambda - a_{mm}| \leq \sum_{j \neq m} |a_{mj}| \frac{|x_j|}{|x_m|} \leq \sum_{j \neq m} |a_{mj}|.
\]

Note that if the matrix is symmetric, the discs are in fact segments of the real axis. \( \square \)
J Theorems from Vector Calculus

Let \( a, b, \text{ and } c \) be three-dimensional vectors, let \( v \) and \( w \) be well-behaved vector functions from \( \mathbb{R}^3 \rightarrow \mathbb{R}^3 \), and let \( \psi \) and \( \psi \) be well-behaved scalar functions mapping \( \mathbb{R}^3 \rightarrow \mathbb{R} \). Let \( \Omega \) be a three-dimensional volume with volume element \( dV \), and let \( \partial \Omega \) be a closed two-dimensional surface bounding \( \Omega \) with area element \( dS \) and associated unit outward vector \( n \). Let \( S \) denote an open surface with the oriented contour \( C \), with line element \( d\ell \) bounding it. The normal \( n \) to \( S \) is defined according to the right-hand screw rule applied to \( C \). Then we have the following vector and integral relations:

\[
\begin{align*}
\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) &= \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}), \\
\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b}), \\
(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) &= (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c}), \\
\nabla (v \cdot w) &= v \cdot \nabla w + w \cdot \nabla v + v \times (\nabla \times w) + w \times (\nabla \times v), \\
v \cdot \nabla v &= \frac{1}{2} \nabla (v \cdot v) + (\nabla \times v) \times v, \\
\mathbf{a} \cdot \nabla (v \cdot \mathbf{w}) &= v \cdot (\nabla \cdot \mathbf{w}) + \mathbf{w} \cdot (\nabla \cdot v), \\
\nabla \cdot (\psi v) &= \psi \nabla \cdot v + v \cdot \nabla \psi, \\
\nabla \cdot (v \times w) &= w \cdot (\nabla \times v) - v \cdot (\nabla \times w), \\
\nabla \times (v \times w) &= v(\nabla \cdot w) - w(\nabla \cdot v) - v \cdot \nabla w + w \cdot \nabla v, \\
\nabla \cdot (\nabla v) &= 0, \\
\nabla \times (\nabla \psi) &= 0, \\
\nabla \times (\nabla \psi) &= \nabla (\nabla \cdot v) - \nabla^2 v.
\end{align*}
\]

We also have the following results:

Gauss’s or divergence theorem:
\[
\int_{\Omega} \nabla \cdot v \, dV = \oint_{\partial \Omega} v \cdot n \, dS, \quad (J.13)
\]

Green’s first identity:
\[
\int_{\Omega} \nabla \psi \, dV = \oint_{\partial \Omega} \psi \, n \, dS, \quad (J.14)
\]

Green’s second identity:
\[
\int_{\Omega} (\nabla^2 \psi + \nabla \psi \cdot \nabla \psi) \, dV = \oint_{\partial \Omega} (\nabla \psi \cdot n) \, dS; \quad (J.15)
\]

Stokes’s theorem:
\[
\int_{S} (\nabla \times \mathbf{v}) \cdot \mathbf{n} \, dS = \oint_{C} \mathbf{v} \cdot d\ell, \quad (J.16)
\]

\[
\int_{S} \mathbf{n} \times \nabla \psi \, dS = \oint_{C} \psi \, d\ell. \quad (J.17)
\]
Let \( q(x, t) \) be a quantity per unit volume of a fluid. Consider a material volume \( \Omega(t) \) moving with the flow. Then we have the following:

\[
\frac{d}{dt} \int_{\Omega(t)} q(x, t) \, dV = \int_{\Omega(t)} \left( \frac{\partial}{\partial t} q(x, t) + \nabla \cdot (q \psi(x, t)) \right) \, dV. \quad (J.20)
\]

**K Cartesian, Cylindrical, and Spherical Coordinates**

**Cartesian.** Let \( e_x, e_y, \) and \( e_z \) be the orthogonal unit vectors associated with the Cartesian \( x, y, \) and \( z \) coordinates, and let \( E = e_x E_x + e_y E_y + e_z E_z \) and \( \phi \) be smooth functions of \((x, y, z)\). Then

\[
(d\ell)^2 = (dx)^2 + (dy)^2 + (dz)^2, \quad (K.1a)
\]

\[
\nabla \phi = e_x \frac{\partial \phi}{\partial x} + e_y \frac{\partial \phi}{\partial y} + e_z \frac{\partial \phi}{\partial z}, \quad (K.1b)
\]

\[
\nabla^2 \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2}, \quad (K.1c)
\]

\[
\nabla \cdot E = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}, \quad (K.1d)
\]

\[
\nabla \times E = \begin{vmatrix}
  e_x & e_y & e_z \\
  \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
  E_x & E_y & E_z
\end{vmatrix}. \quad (K.1e)
\]

**Cylindrical.** Let \( e_r, e_\phi, \) and \( e_z \) be the orthogonal unit vectors associated with the cylindrical \( r, \phi, \) and \( z \) coordinates, and let \( E = e_r E_r + e_\phi E_\phi + e_z E_z \) and \( \phi \) be smooth functions of \((r, \phi, z)\). Then

\[
x = r \cos \phi, \quad y = r \sin \phi, \quad z = z, \quad (K.2a)
\]

\[
(d\ell)^2 = (dr)^2 + (r \, d\phi)^2 + (dz)^2, \quad (K.2b)
\]

\[
\nabla \phi = e_r \frac{\partial \phi}{\partial r} + \frac{1}{r} e_\phi \frac{\partial \phi}{\partial \phi} + e_z \frac{\partial \phi}{\partial z}, \quad (K.2c)
\]

\[
\nabla^2 \phi = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial \phi}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 \phi}{\partial \phi^2} + \frac{\partial^2 \phi}{\partial z^2}, \quad (K.2d)
\]

\[
\nabla \cdot E = \frac{1}{r} \frac{\partial}{\partial r} (r E_r) + \frac{1}{r} \frac{\partial E_\phi}{\partial \phi} + \frac{\partial E_z}{\partial z}, \quad (K.2e)
\]

\[
\nabla \times E = \frac{1}{r} \begin{vmatrix}
  e_r & e_\phi & e_z \\
  \frac{\partial}{\partial r} & \frac{\partial}{\partial \phi} & \frac{\partial}{\partial z} \\
  E_r & E_\phi & E_z
\end{vmatrix}. \quad (K.2f)
Spherical. Let $e_r$, $e_\theta$, and $e_\phi$ be the orthogonal unit vectors associated with the spherical $r$, $\theta$, and $\phi$ coordinates, and let $E = e_r E_r + e_\theta E_\theta + e_\phi E_\phi$ and $\psi$ be smooth functions of $(r, \theta, \phi)$. Then

$$x = r \sin \theta \cos \phi, \quad y = r \sin \theta \sin \phi, \quad z = r \cos \theta,$$

(K.3a)

$$(d\ell)^2 = (dr)^2 + (r \, d\theta)^2 + (r \sin \theta \, d\phi)^2.$$

(K.3b)

$$\nabla \psi = e_r \frac{\partial \psi}{\partial r} + e_\theta \frac{\partial \psi}{\partial \theta} + e_\phi \frac{1}{r \sin \theta} \frac{\partial \psi}{\partial \phi}.$$

(K.3c)

$$\nabla^2 \psi = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \psi}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \psi}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 \psi}{\partial \phi^2}.$$

(K.3d)

$$\nabla \cdot E = \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 E_r) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} ( \sin \theta E_\theta ) + \frac{1}{r \sin \theta} \frac{\partial E_\phi}{\partial \phi}.$$

(K.3e)

$$\nabla \times E = \frac{1}{r^2 \sin \theta} \begin{vmatrix} e_r & e_\theta & e_\phi \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \theta} & \frac{\partial}{\partial \phi} \\ E_r & E_\theta & E_\phi \end{vmatrix}.$$

(K.3f)

## L Tensors

A three-dimensional vector space over the field of real numbers, equipped with an inner, dot, or scalar product $a \cdot b$ and an outer, cross, or vector product $a \times b$ is called a Euclidean vector space.

A tensor—strictly speaking of order two—is a linear transformation of a Euclidean vector space into itself. The identity tensor is denoted by $\mathcal{I}$.

If the tensor $\mathcal{A}$ is written as a $3 \times 3$ matrix $(a_{ij})$ on the standard basis $[e_1, e_2, e_3]$ of $\mathbb{R}^3$, we have the contraction or trace of $\mathcal{A}$, given by $\tr(\mathcal{A}) = a_{11} + a_{22} + a_{33}$, and the determinant of $\mathcal{A}$, given by $\det(\mathcal{A}) = \mathcal{A}_{e_1} \cdot (\mathcal{A} e_2 \times \mathcal{A} e_3)$. Both are invariants of $\mathcal{A}$ under rotation of axes and may therefore be expressed in terms of the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of matrix $(a_{ij})$. Altogether, $\mathcal{A}$ has three invariants.

A (second order) tensor $\mathcal{A}$ is completely determined by its invariants

$$\tr(\mathcal{A}) = \lambda_1 + \lambda_2 + \lambda_3,$$

(L.1a)

$$\frac{1}{2} [\tr(\mathcal{A})^2 - \tr(\mathcal{A}^2)] = \lambda_1 \lambda_2 + \lambda_2 \lambda_3 + \lambda_3 \lambda_1,$$

(L.1b)

$$\det(\mathcal{A}) = \lambda_1 \lambda_2 \lambda_3.$$

(L.1c)

With each $\mathcal{A}$ is associated a traceless tensor (the deviator or deviatoric part of $\mathcal{A}$)

$$\mathcal{A}^{\prime} := \mathcal{A} - \frac{1}{3} \tr(\mathcal{A}) \mathcal{I},$$

(L.2)

with the same invariants except the first one, which is zero.
The divergence of a tensor is a vector whose $i$th component, in orthogonal coordinates $(x_1, x_2, x_3)$, is given by

$$(\nabla \cdot \mathbf{A})_i = \sum_{j=1}^{3} \frac{\partial}{\partial x_j} A_{ij} \quad (i = 1, 2, 3). \quad \text{(L.3)}$$

The inner product of two tensors $\mathbf{A}$ and $\mathbf{B}$ produces a tensor $\mathbf{A} \cdot \mathbf{B}$ whose components are given by

$$(\mathbf{A} \cdot \mathbf{B})_{ij} := \sum_{k=1}^{3} A_{ik} B_{kj}. \quad \text{(L.4)}$$

The double inner product of two tensors $\mathbf{A}$ and $\mathbf{B}$ produces a scalar $\mathbf{A} : \mathbf{B}$, which can be evaluated as the sum of the nine products of the tensor components:

$$\mathbf{A} : \mathbf{B} = \sum_{i=1}^{3} \sum_{j=1}^{3} A_{ij} B_{ij}. \quad \text{(L.5)}$$

The dyadic product of two vectors $\mathbf{a}$ and $\mathbf{b}$ produces a tensor $\mathbf{a} \mathbf{b}^T$ given by

$$(\mathbf{a} \mathbf{b}^T)_{ij} = a_i b_j \quad \text{or} \quad \mathbf{a} \mathbf{b}^T = \begin{pmatrix} a_1 b_1 & a_1 b_2 & a_1 b_3 \\ a_2 b_1 & a_2 b_2 & a_2 b_3 \\ a_3 b_1 & a_3 b_2 & a_3 b_3 \end{pmatrix}. \quad \text{(L.6)}$$
# Dimensionless Numbers

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Formula</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archimedes</td>
<td>$Ar$</td>
<td>$g \Delta \rho L^3 / \rho v^2$</td>
<td>particles, drops, or bubbles</td>
</tr>
<tr>
<td>Arrhenius</td>
<td>$Arr$</td>
<td>$E / RT$</td>
<td>chemical reactions</td>
</tr>
<tr>
<td>Biot</td>
<td>$Bi$</td>
<td>$h L / \kappa$</td>
<td>heat transfer at surface of body</td>
</tr>
<tr>
<td>Biot</td>
<td>$Bi$</td>
<td>$h_\rho L / D$</td>
<td>mass transfer</td>
</tr>
<tr>
<td>Bodenstein</td>
<td>$Bo$</td>
<td>$V L / D_{ax}$</td>
<td>mass transfer with axial dispersion</td>
</tr>
<tr>
<td>Bond</td>
<td>$Bo$</td>
<td>$g \rho L^2 / \sigma$</td>
<td>gravity against surface tension</td>
</tr>
<tr>
<td>Capillary</td>
<td>$Ca$</td>
<td>$\mu V / \sigma$</td>
<td>viscous forces against surface tension</td>
</tr>
<tr>
<td>Dean</td>
<td>$De$</td>
<td>$(V L / \nu)(L / 2r)^{1/2}$</td>
<td>flow in curved channels</td>
</tr>
<tr>
<td>Eckert</td>
<td>$Ec$</td>
<td>$V^2 / C P \Delta T$</td>
<td>kinetic energy against enthalpy difference</td>
</tr>
<tr>
<td>Euler</td>
<td>$Eu$</td>
<td>$\Delta p / \rho V^2$</td>
<td>pressure resistance</td>
</tr>
<tr>
<td>Fourier</td>
<td>$Fo$</td>
<td>$\alpha / L^2$</td>
<td>heat conduction</td>
</tr>
<tr>
<td>Fourier</td>
<td>$Fo$</td>
<td>$D t / L^2$</td>
<td>diffusion</td>
</tr>
<tr>
<td>Froude</td>
<td>$Fr$</td>
<td>$V / (g L)^{1/2}$</td>
<td>gravity waves</td>
</tr>
<tr>
<td>Galileo</td>
<td>$Ga$</td>
<td>$gL^3 \rho^2 / \mu^2$</td>
<td>gravity against viscous forces</td>
</tr>
<tr>
<td>Grashof</td>
<td>$Gr$</td>
<td>$\beta \Delta T g L^3 / \nu^3$</td>
<td>natural convection</td>
</tr>
<tr>
<td>Helmholtz</td>
<td>$He$</td>
<td>$\omega L / c = k L$</td>
<td>acoustic wave number</td>
</tr>
<tr>
<td>Kapitza</td>
<td>$Ka$</td>
<td>$g \mu^4 / \rho \sigma^5$</td>
<td>film flow</td>
</tr>
<tr>
<td>Knudsen</td>
<td>$Kn$</td>
<td>$\lambda / L$</td>
<td>low-density flow</td>
</tr>
<tr>
<td>Lewis</td>
<td>$Le$</td>
<td>$\alpha / D$</td>
<td>combined heat and mass transfer</td>
</tr>
<tr>
<td>Mach</td>
<td>$Ma$</td>
<td>$V / c$</td>
<td>compressible flow</td>
</tr>
<tr>
<td>Nusselt</td>
<td>$Nu$</td>
<td>$h L / \kappa$</td>
<td>convective heat transfer</td>
</tr>
<tr>
<td>Ohnesorge</td>
<td>$Oh$</td>
<td>$\mu / (\rho L \alpha)^{1/2}$</td>
<td>viscous forces, inertia, and surface tension</td>
</tr>
<tr>
<td>Péclet</td>
<td>$Pe$</td>
<td>$V L / \alpha$</td>
<td>forced-convection heat transfer</td>
</tr>
<tr>
<td>Péclet</td>
<td>$Pe$</td>
<td>$V L / D$</td>
<td>forced-convection mass transfer</td>
</tr>
<tr>
<td>Prandtl</td>
<td>$Pr$</td>
<td>$V / \alpha = C_P \mu / \kappa$</td>
<td>convective heat transfer</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>$Ra$</td>
<td>$\beta \Delta T g L^3 / \alpha \nu$</td>
<td>natural convection heat transfer</td>
</tr>
<tr>
<td>Reynolds</td>
<td>$Re$</td>
<td>$\rho V L / \mu$</td>
<td>viscous forces against inertia</td>
</tr>
<tr>
<td>Schmidt</td>
<td>$Sc$</td>
<td>$\nu / D$</td>
<td>convective mass transfer</td>
</tr>
<tr>
<td>Sherwood</td>
<td>$Sh$</td>
<td>$h_\rho L / D$</td>
<td>convective mass transfer</td>
</tr>
<tr>
<td>Stanton</td>
<td>$St$</td>
<td>$h / \rho C_P V$</td>
<td>forced-convection heat transfer</td>
</tr>
<tr>
<td>Stanton</td>
<td>$St$</td>
<td>$h_\rho L / V$</td>
<td>forced-convection mass transfer</td>
</tr>
<tr>
<td>Stokes</td>
<td>$S$</td>
<td>$\nu / f L^2$</td>
<td>viscous damping in unsteady flow</td>
</tr>
<tr>
<td>Strouhal</td>
<td>$Sr$</td>
<td>$f L / V$</td>
<td>hydrodynamic wave number</td>
</tr>
<tr>
<td>Weber</td>
<td>$We$</td>
<td>$\rho V^2 L / \sigma$</td>
<td>film flow, bubble formation, or droplet breakup</td>
</tr>
</tbody>
</table>

In this table we used the following symbols.
### Dimensionless Numbers

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>sound speed</td>
<td>m/s</td>
</tr>
<tr>
<td>$C_P$</td>
<td>specific heat</td>
<td>J/(kg·K)</td>
</tr>
<tr>
<td>$D$</td>
<td>diffusion coefficient</td>
<td>m²/s</td>
</tr>
<tr>
<td>$D_{a_k}$</td>
<td>axial dispersion coefficient</td>
<td>m²/s</td>
</tr>
<tr>
<td>$E$</td>
<td>activation energy</td>
<td>J/mol</td>
</tr>
<tr>
<td>$f$</td>
<td>frequency</td>
<td>1/s</td>
</tr>
<tr>
<td>$g$</td>
<td>gravitational acceleration</td>
<td>m/s²</td>
</tr>
<tr>
<td>$h$</td>
<td>heat transfer coefficient</td>
<td>W/(m²·K)</td>
</tr>
<tr>
<td>$h_D$</td>
<td>mass transfer coefficient</td>
<td>m/s</td>
</tr>
<tr>
<td>$k$</td>
<td>wave number = $\omega/c$</td>
<td>1/m</td>
</tr>
<tr>
<td>$L$</td>
<td>length</td>
<td>m</td>
</tr>
<tr>
<td>$p$, $\Delta p$</td>
<td>pressure</td>
<td>Pa</td>
</tr>
<tr>
<td>$R$</td>
<td>universal gas constant</td>
<td>J/(mol·K)</td>
</tr>
<tr>
<td>$r$</td>
<td>radius of curvature</td>
<td>m</td>
</tr>
<tr>
<td>$T$, $\Delta T$</td>
<td>temperature</td>
<td>K</td>
</tr>
<tr>
<td>$t$</td>
<td>time</td>
<td>s</td>
</tr>
<tr>
<td>$V$</td>
<td>velocity</td>
<td>m/s</td>
</tr>
<tr>
<td>$\alpha = \kappa / \rho C_P$</td>
<td>thermal diffusivity</td>
<td>m²/s</td>
</tr>
<tr>
<td>$\beta$</td>
<td>coefficient of thermal expansion</td>
<td>K⁻¹</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>thermal conductivity</td>
<td>W/(m·K)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>molecular mean free path</td>
<td>m</td>
</tr>
<tr>
<td>$\mu$</td>
<td>dynamic viscosity</td>
<td>Pa·s</td>
</tr>
<tr>
<td>$\nu = \mu / \rho$</td>
<td>kinematic viscosity</td>
<td>m²/s</td>
</tr>
<tr>
<td>$\rho$, $\Delta \rho$</td>
<td>density</td>
<td>kg/m³</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>surface tension</td>
<td>N/m</td>
</tr>
<tr>
<td>$\omega$</td>
<td>circular frequency = $2\pi f$</td>
<td>1/s</td>
</tr>
</tbody>
</table>
Bibliography


Bibliography


Bibliography 657


Index

2-norm, 98
$\theta$ method, 89, 273
absence of free magnetic poles, 123
coustic impedance, 623
coustic liner, 623
Adams methods, 90
Adams–Bashforth, 90
Adams–Moulton, 90
ADI (Alternating Direction Implicit), 213, 294
adiabatic invariants, 492
adjoint operator, 58
advection-diffusion barrier, 375
advection-diffusion equation, 295, 373
air-damped resonator, 488
aircraft engine noise, 622
Alternating Direction Implicit (ADI), 213, 294
Ampère–Maxwell’s law, 122
amplification factor, 261, 365
amplitude error, 371
approximate Riemann solver, 426
artificial diffusion, 378
associated matrix norm, 638
asymptotic approximation, 461
asymptotic expansion, 462
asymptotic sequence, 462
asymptotically equal, 463
averaging, 351
backward difference, 79, 90, 631
backward difference formulae (BDF), 86
backward Euler method, 89
band matrix, 208
barotropic flow, 154
base characteristic, 15
BDF (Backward Difference Formulas), 90, 632
Beam–Warming scheme, 414
block tridiagonal, 208
block-wave function, 37
Blokhintzev invariant, 567
boundary basis vectors, 616
boundary condition, 160
absorbing, 448
nonreflecting, 448
top layer, 616
boundary layer, 92, 476
box scheme, 189, 203, 379
Buckley–Leverett equation, 326
catalytic reaction, 555
Cauchy problem, 334
Cauchy–Schwarz inequality, 638
Cauchy’s equation of motion, 114
causal, 41
causality condition, 41, 136
cell-centered finite volume method, 85
central difference, 79, 80
CFL (Courant–Friedrichs–Lewy) condition, 366
characteristic equations, 15
characteristic polynomial, 639
characteristic strip equations, 316
characteristic variables, 18, 330
characteristics, 15, 308
Charpit’s equations, 316
classical Runge–Kutta, 89
closure, 212
collocation, 78
complementary error function, 144, 236
conductive flux, 85
cone of normals, 315
conservation equation, 109, 311
dimensionless parameter, 138, 646
Dirichlet boundary condition, 160
discrete dispersion relation, 364
discrete Green’s function, 214
discrete pressure gradient, 619
discretisation, 77
backward differences, 619
forward differences, 619
dispersion, 45
displacement vector, 110
distinguished limit, 464, 477
distribution, 54
regular, 54
divergence theorem, 642
DuFort–Frankel scheme, 305
Duhamel integral, 63, 64
Eckert number, 150
diffusion number, 261
development, 89
distinctive limit, 464, 477
discussion, 54
displacement vector, 110
distinctive limit, 464, 477
distribution, 54
displacement vector, 110
distinctive limit, 464, 477
distribution, 54
Euler number, 150
expansion shock, 323
explicit Euler method, 76, 89
explicit Euler scheme, 261
explicit method, 89
explicit midpoint rule, 91
Faraday–Henry’s law, 122
Fick’s law, 118
conservation law, 86
governing equation, 311
conservation of
angular momentum, 113
electric charge, 123
electric field, 123
energy, 115
equivalent norms, 638
linear momentum, 113
critical state, 439
cylindrical symmetry, 83
d’Alembert solution, 350
damped Newton, 218
damping factor, 218
Davidenko equation, 222
decibel, 147
delta function, 54
descent direction, 218
difference equation, 76
difference scheme, 80
diffusion number, 261
dimensionally homogeneous, 138
conservation of
angular momentum, 113
electric charge, 123
energy, 115
linear momentum, 113
conservative, 46
conservative scheme, 390
consistency, 98
explicit Euler scheme, 263
conservative
explicit Euler scheme, 263
d’Alembert solution, 350
damped Newton, 218
damping factor, 218
Davidenko equation, 222
decibel, 147
delta function, 54
descent direction, 218
difference equation, 76
difference scheme, 80
diffusion number, 261
dimensionally homogeneous, 138
dimensionless parameter, 138, 646
Dirichlet boundary condition, 160
discrete dispersion relation, 364
discrete Green’s function, 214
discrete pressure gradient, 619
discretisation, 77
backward differences, 619
forward differences, 619
dispersion, 45
displacement vector, 110
distinguished limit, 464, 477
distribution, 54
regular, 54
divergence theorem, 642
DuFort–Frankel scheme, 305
Duhamel integral, 63, 64
Eckert number, 150
diffusion number, 261
development, 89
distinctive limit, 464, 477
discussion, 54
distinctive limit, 464, 477
distribution, 54
distinctive limit, 464, 477
distribution, 54
Euler number, 150
expansion shock, 323
explicit Euler method, 76, 89
explicit Euler scheme, 261
explicit method, 89
explicit midpoint rule, 91
Faraday–Henry’s law, 122
Fick’s law, 118
conservation equation, 109, 311
dimensionless parameter, 138, 646
Dirichlet boundary condition, 160
discrete dispersion relation, 364
discrete Green’s function, 214
discrete pressure gradient, 619
discretisation, 77
backward differences, 619
forward differences, 619
dispersion, 45
displacement vector, 110
distinguished limit, 464, 477
distribution, 54
regular, 54
divergence theorem, 642
DuFort–Frankel scheme, 305
Duhamel integral, 63, 64
Eckert number, 150
diffusion number, 261
development, 89
distinctive limit, 464, 477
discussion, 54
distinctive limit, 464, 477
distribution, 54
distinctive limit, 464, 477
distribution, 54
Euler number, 150
expansion shock, 323
explicit Euler method, 76, 89
explicit Euler scheme, 261
explicit method, 89
explicit midpoint rule, 91
Faraday–Henry’s law, 122
Fick’s law, 118
finite difference, 73
Fisher’s equation, 252
flux, 85
flux function, 310
flux limiter, 397
flux vector, 331
forward difference, 79
forward Euler method, 89
Fourier
cosine series, 37
discrete transform, 43
mode, 34
series, 34, 35
sine series, 37
transform, 38
Fourier number, 150
Fourier’s law, 117
Fredholm’s alternative, 62
frequency, 34, 261
galloping transmission lines, 531
Gauss’s theorem, 642
Gauss–Jacobi iteration, 209
Gauss–Seidel iteration, 209
generalised eigenvalue problem, 327
generalized Riemann problem, 404
genuinely nonlinear, 337
geometric multiplicity, 640
geopotential, 342
glass oven, 613
glass production, 581
global discretisation error, 193, 263
implied Euler method, 271
vector, 99
Godunov scheme, 388, 423
wave propagation form, 425
Green’s first identity, 642
Green’s function, 61, 174
groundwater flow, 540
group velocity, 45, 565
growth factor, 103
harmonic function, 160
Harten–Hyman entropy fix, 439
heat capacity, 119
heat equation, 24, 233, 260
heat flow in a bar, 469
Helmholtz equation, 8, 159
Helmholtz number, 150
high-resolution flux, 398
high-resolution scheme, 391, 398
homogeneous, 214
Hooke’s law, 120
hyperbolic equation, 308
conservation form, 310
linear, 308
quasi-linear, 310
semilinear, 309
hyperbolic system, 329
characteristic form, 331
conservation form, 331
linear, 328
normal form, 331
quasi-linear, 330
implicit Euler method, 76, 89
implicit Euler scheme, 270
implicit method, 89, 90
incremental form, 392
infinity norm, 98
inherent scaling, 138
initial curve, 15
initial value, 96
inner expansion, 476
integral surface, 13
interior layers, 92
internal layer, 476
interpolation points, 72
isochophoresis, 599
iteration matrix, 208
iterative methods, 208
Jordan normal form, 640
Joukowski transformation, 153
Joule heating, 124
k-rarefaction wave, 339
k-Riemann invariants, 339
k-shock wave, 341
k-simple wave, 339
Kelvin–Voigt model, 121
Kelvin’s theorem, 126
Kutta condition, 154
Lagrangian basis polynomials, 72
Lagrangian interpolation polynomial, 72
Laplace equation, 24, 159
laser drilling, 589
latent heat, 119
Lax–Friedrichs scheme, 414
Lax–Wendroff scheme, 368
layers, 92
leapfrog scheme, 382
left eigenvectors, 327
limiter
  function, 397
  superbee, 402
  van Leer, 402
Lindstedt–Poincaré method, 473
linearly degenerate, 338
local discretisation error, 78, 189, 193,
  263, 366
  \( \vartheta \) method, 274
  Lax–Wendroff scheme, 369
  upwind scheme, 366
  vector, 98
logarithmic switchback, 481
longitudinal, 96
longitudinal method of lines, 265
Lorentz force, 125
lubrication flow, 496
Mach number, 150
matched asymptotic expansions (MAEs),
  475, 477
matching, 477, 483
material surface, 111
material volume, 111
mathematising, 130
matrix
  defect, 641
  similarity transformation, 640
  stable, 641
  trace, 639
matrix method, 276
matrix pencil, 328
regular, 328
singular, 328
Maxwell equations, 123
Maxwell model, 122
median 2-norm, 98
method of lines (MOL), 96
method of Massau, 333
method of strained coordinates, 474, 486
midpoint formula, 75
midpoint rule, 89
minmod slope, 407
mixed boundary conditions, 282
mode, 147
model, 131
  asymptotic, 131
  building block, 132
  canonical, 135
  characteristic, 135
  constructing, 132
  lumped parameter, 132
  quintessential, 135
  reducing, 131
  systematic, 131
modified equation, 421
MOL (method of lines), 96
Monge cone, 315
monocrystalline silicon bar, 547
multiple scales, 483, 484, 486
multiplicity
  algebraic, 641
  geometric, 641
multistep, 77
multistep method, 90
Navier–Stokes equations, 119
Neumann boundary condition, 160
Newton update, 217
Newton’s method, 217
nondissipative scheme, 379
nonlinear first order equation
  characteristic form, 314
numerical boundary condition, 409, 449
numerical diffusion, 377
numerical domain of dependence, 366
numerical flux, 387, 391
numerical group velocity, 384
numerical phase velocity, 371
Index

parasitic root, 91
parasitic solutions, 91
parison, 582
Parseval’s identity, 35
particular, 61
particular solution, 173
Péclet number, 150, 376
pendulum, 474
perfect gas, 119
permafrost, 157
perturbation methods, 464
Pfaffian differential equation, 330
phase angle, 261, 365
phase velocity, 45, 370
physical boundary conditions, 354
physical model, 116
piecewise continuous, 34
piecewise smooth, 34
Poincaré expansion, 466
Poiseuille flow, 151
Poisson equation, 159
Poisson’s formula, 40
Poisson’s ratio, 121
powder blasting, 508
powder erosion, 508
Poynting’s theorem, 124
Prandtl number, 150
Prandtl’s boundary layer, 480, 481
Preface, xxxi
pressure
  mechanical, 120
  thermodynamic, 120
pressure correction algorithm, 224
pressure correction method, 223
pressure Poisson equation, 222
principal curvatures, 120
pulse tube refrigerator, 606
quadrature formula, 75
quasi one-dimensional gas dynamics, 471
quasi-linear, 17
radiation condition, 447
radiative, 244
Rankine–Hugoniot jump condition, 320, 336
rarefaction wave, 323
ray acoustics, 493, 563
ray approximation, 487, 493, 564
ray tube, 567
reality condition, 41
reduced equation, 93
reduced wave equation, 8
refinement path, 98
reflections, 350
refraction
  temperature gradient, 563
  wind shear, 563
regular expansion, 467
relative phase error, 371
resin curing, 501
resonance, 485, 486
Reynolds lubrication equations, 584
Reynolds number, 150
Riemann invariants, 331
Riemann problem, 322
Riemann variables, 331
Robin boundary condition, 160
Roe matrix, 427
Roe numerical flux, 426
Roe scheme, 426
Runge–Kutta method, 89
sawtooth function, 36
secular terms, 474, 475, 485, 487, 489
self-adjoint, 165
shallow-water equations, 342, 471, 473
shock speed, 322
shock wave, 322
signal velocity, 45
significant degeneration, 464, 477
similarity solutions, 25, 143
singular expansion, 467
skew symmetric, 640
slopest limiter, 405
slope limiter scheme, 405
slowly varying duct, 490
slowly varying fast time scale, 487
sound attenuation, 622
sparse matrix, 206
specific energy, 115
specific heat, 119
specific heat ratio, 119
specific internal energy, 115
specific kinetic energy, 115
spectral radius, 640
spherical symmetry, 83
spurious mode, 224
spurious root, 383
stability condition
\( \theta \) method, 273
explicit Euler method, 263
Lax–Wendroff scheme, 369
upwind scheme, 367
stable, 99
stages, 78
staggered grid, 225
stationary elliptic equation, 249
Stefan constant, 246
Stefan problem, 592
stencil, 80
stiff, 95
stiffness, 95
Stokes equations, 152, 178, 222
Stokes flow, 152
Stokeslets, 179
Stokes’s hypothesis, 120
Stokes’s theorem, 642
stretch, 93
stretched coordinate, 476
Strouhal number, 150
subcritical flow, 359
supercritical flow, 358
superposition, 62, 174
superposition principle, 215
support, 53
surface tension, 120, 498
suspended bar, 497
symmetric, 640
telegraph equation, 5
tensor, 644
contraction, 644
deformation velocity, 110
determinant, 644
deviator, 644
divergence, 645
double inner product, 645
dyadic product, 645
inner product, 645
invariants, 644
linear deformation, 110
linear strain, 110
rate of deformation, 110
trace, 644
test function, 318
thin-layer flow, 570
time closure, 212
total variation (TV), 392
total variation diminishing (TVD), 392
traffic flow equation, 324
transient method, 221
transport equation, 311
transport theorem, 111, 643
transversal, 96
trapezoidal formula, 75
trapezoidal rule, 89
travelling wave, 25
travelling-wave solution, 253
triangle inequality, 637
TV (total variation), 392
TVD region, 399
uniform, 77
uniform asymptotic approximation, 461
upwind scheme, 364, 419
van der Pol equation, 498
vanishing viscosity solution, 314
variable step, variable order methods, 91
vertex-centered finite volume method, 86
virtual grid points, 189, 283
viscoelasticity, 121, 523
viscosity
bulk, 120
dynamic, 120
Index

expansion, 120
kinematic, 120
second, 120
viscous stress tensor, 118

wave action, 493, 494
wave energy, 493
wave equation, 24, 349
wave number, 34, 261
wave speed, 25, 45
wave strength, 428
weakly diagonally dominant, 202
Webster’s horn equation, 496
weights, 73, 75
WKB (Wentzel–Kramers–Brillouin) hypothesis, 487

Young’s modulus, 121