Contents

1 Introduction 7
  1.1 Discretization ........................................ 10
  1.2 Conditioning .......................................... 12
  1.3 Error .................................................. 13
    1.3.1 Unavoidable Error ................................ 14
    1.3.2 Error of the Method .............................. 16
    1.3.3 Round-off Error ................................... 16
  1.4 On Methods of Computation .......................... 18

I Interpolation of Functions. Quadratures 23

2 Algebraic Interpolation 27
  2.1 Existence and Uniqueness of Interpolating Polynomial .... 27
    2.1.1 The Lagrange Form of Interpolating Polynomial .... 27
    2.1.2 The Newton Form of Interpolating Polynomial. Divided Differences ................................. 28
    2.1.3 Comparison of the Lagrange and Newton Forms .... 33
    2.1.4 Conditioning of the Interpolating Polynomial ...... 34
    2.1.5 On Poor Convergence of Interpolation with Equidistant Nodes ........................................... 35
  2.2 Classical Piece-Wise Polynomial Interpolation ............ 37
    2.2.1 Definition of Piece-Wise Polynomial Interpolation .... 37
    2.2.2 Formula for the Interpolation Error ............... 37
    2.2.3 Approximation of Derivatives for a Grid Function ... 40
    2.2.4 Estimate of the Unavoidable Error and the Choice of Degree for Piece-Wise Polynomial Interpolation ... 42
    2.2.5 Saturation of Piece-Wise Polynomial Interpolation .... 44
  2.3 Smooth Piece-Wise Polynomial Interpolation (Splines) ... 45
    2.3.1 Local Interpolation of Smoothness s and Its Properties .... 45
    2.3.2 Nonlocal Smooth Piece-Wise Polynomial Interpolation .... 50
    2.3.3 Proof of Theorem 2.11 ........................... 55
  2.4 Interpolation of Functions of Two Variables ............ 59
    2.4.1 Structured Grids ................................... 59
    2.4.2 Unstructured Grids ................................ 60
3 Trigonometric Interpolation 63

3.1 Interpolation of Periodic Functions 64
    3.1.1 An Important Particular Choice of Interpolation Nodes 64
    3.1.2 Sensitivity of the Interpolating Polynomial to Perturbations of the Function Values 69
    3.1.3 Estimate of Interpolation Error 70
    3.1.4 An Alternative Choice of Interpolation Nodes 74

3.2 Interpolation of Functions on an Interval. Relation between Algebraic and Trigonometric Interpolation 75
    3.2.1 Periodization 75
    3.2.2 Trigonometric Interpolation 77
    3.2.3 Chebyshev Polynomials. Relation between Algebraic and Trigonometric Interpolation 77
    3.2.4 Properties of Algebraic Interpolation with Roots of the Chebyshev Polynomial $T_{n+1}(x)$ as Nodes 79
    3.2.5 An Algorithm for Evaluating the Interpolating Polynomial 80
    3.2.6 Algebraic Interpolation with Extrema of the Chebyshev Polynomial $T_n(x)$ as Nodes 81
    3.2.7 More on the Lebesgue Constants and Convergence of Interpolants 82

4 Evaluation of Definite Integrals. Quadratures 93
    4.1 The Trapezoid Rule and Simpson’s Formula 93

II Systems of Algebraic Equations 95

5 Systems of Linear Algebraic Equations. Direct Methods 97
    5.1 Different Forms of Consistent Linear Systems 97
        5.1.1 Canonical Form of a Linear System 97
        5.1.2 Operator Form 97
        5.1.3 Finite-Difference Dirichlet Problem for the Poisson Equation 97
    5.2 Norms 98
    5.3 Gaussian Elimination and Its Tri-Diagonal Version 98

6 Iterative Methods for Solving Linear Systems 99
    6.1 First-Order Linear Stationary Methods 99

III The Method of Finite Differences for the Numerical Solution of Differential Equations 101

9 Numerical Solution of Ordinary Differential Equations 105
    9.1 Examples of Finite-Difference Schemes. Convergence 105
        9.1.1 Examples of Difference Schemes 106
        9.1.2 Convergent Difference Schemes 108
        9.1.3 Verification of Convergence for a Difference Scheme 111
9.2 Approximation of Continuous Problem by a Difference Scheme.
9.2.1 Truncation Error $\delta f^{(h)}$ .................................. 113
9.2.2 Evaluation of the Truncation Error $\delta f^{(h)}$ ............... 114
9.2.3 Accuracy of Order $h^k$ ........................................ 116
9.2.4 Examples ...................................................... 117
9.2.5 Replacement of Derivatives by Difference Quotients ...... 121
9.2.6 Other Approaches to Constructing Difference Schemes .... 121
9.3 Stability of Finite-Difference Schemes .......................... 123
9.3.1 Definition of Stability ....................................... 124
9.3.2 The Relation between Consistency, Stability, and Converge-
9.3.3 Convergent Scheme for an Integral Equation .............. 128
9.3.4 The Effect of Rounding ..................................... 130
9.3.5 General Comments. A-stability ............................. 131
9.4 The Runge-Kutta Methods ...................................... 135
9.4.1 The Runge-Kutta Schemes .................................. 136
9.4.2 Extension to Systems ........................................ 138
9.5 Solution of Boundary Value Problems .......................... 140
9.5.1 The Shooting Method ........................................ 140
9.5.2 Tri-Diagonal Elimination ................................... 142
9.5.3 The Newton Method .......................................... 143
9.6 Saturation of Finite-Difference Methods ......................... 144
9.7 The Notion of Spectral Methods ................................ 152

10 Finite-Difference Schemes for Partial Differential Equations
10.1 Key Definitions and Illustrating Examples ...................... 159
10.1.1 Definition of Convergence ................................ 159
10.1.2 Definition of Consistency .................................. 161
10.1.3 Definition of Stability ..................................... 164
10.1.4 The Courant, Friedrichs, and Lewy Condition .......... 168
10.1.5 The Mechanism of Instability .............................. 171
10.1.6 The Kantorovich Theorem ................................. 172
10.1.7 On the Efficacy of Finite-Difference Schemes .......... 173
10.1.8 Bibliography Comments .................................... 175
10.2 Construction of Consistent Difference Schemes ............... 178
10.2.1 Replacement of Derivatives by Difference Quotients .... 178
10.2.2 The Method of Undetermined Coefficients .............. 185
10.2.3 Other Methods .............................................. 191
10.2.4 Predictor-Corrector Schemes ............................... 194
10.3 Spectral Stability Criterion for Finite-Difference Cauchy Problems
10.3.1 Stability with Respect to Initial Data .................... 199
10.3.2 A Necessary Spectral Condition for Stability ........... 200
10.3.3 Examples .................................................... 202
10.3.4 Stability in $C$ ................................................ 212
10.3.5 Sufficiency of the Spectral Stability Condition in $l_2$ ........ 212
10.3.6 Scalar Equations vs. Systems ........................................ 215
10.4 Stability for Problems with Variable Coefficients ............... 219
  10.4.1 The Principle of Frozen Coefficients .......................... 219
  10.4.2 Dissipation of Finite-Difference Schemes ...................... 221
10.5 Stability for Initial Boundary Value Problems ..................... 226
  10.5.1 The Babenko-Gelfand Criterion .................................. 226
  10.5.2 Spectra of Families of Operators ................................. 234
  10.5.3 The Energy Method ................................................. 251
  10.5.4 A Necessary and Sufficient Condition of Stability ........... 258
10.6 Explicit and Implicit Schemes for the Heat Equation ................ 271
  10.6.1 An Explicit Scheme .............................................. 272
  10.6.2 An Implicit Scheme .............................................. 273

11 Discontinuous Solutions and Methods of Their Computation ....... 275
  11.1 Differential Form of an Integral Conservation Law .............. 276
    11.1.1 Differential Equation in the Case of Smooth Solutions .... 276
    11.1.2 The Mechanism of Formation of Discontinuities .......... 277
    11.1.3 Condition at the Discontinuity ................................ 279
    11.1.4 Generalized Solution of a Differential Problem ........... 281
    11.1.5 The Riemann Problem ........................................... 282
  11.2 Construction of Difference Schemes ................................. 284
    11.2.1 Artificial Viscosity ............................................. 285
    11.2.2 The Method of Characteristics ................................ 286
    11.2.3 Conservative Schemes ......................................... 287

12 Discrete Methods for Elliptic Problems ............................... 293
  12.1 The Notion of Finite Elements ..................................... 293

IV The Methods of Boundary Equations for the Numerical Solution of Boundary Value Problems 295
  13 Boundary Integral Equations and Boundary Element Method ........ 297
    13.1 Reduction of Boundary Value Problems to Integral Equations . 297

List of Figures ................................................................. 299

References ................................................................. 301

Index ................................................................. 305
Chapter 1

Introduction

Modern numerical mathematics provides a theoretical foundation behind the use of electronic computers for solving applied problems. A mathematical approach to any such problem typically begins with the construction of a model for the phenomenon of interest (situation, process, object, device, laboratory/experimental setting, etc.). Classical examples of mathematical models include definite integrals, equation of a pendulum, the heat equation, equations of elasticity, equations of electromagnetic waves, and many other equations of mathematical physics. For comparison, we should also mention here a model used in formal logics — the Boolean algebra.

Analytical methods have always been considered as fundamental means for the study of mathematical models. In particular, these methods allow one to obtain closed form exact solutions for some special cases (for example, tabular integrals). There are also classes of problems for which one can obtain solution in the form of a power series, Fourier series, or some other expansion. In addition, a certain role has always been played by approximate computations. For example, quadrature formulae are used for the evaluation of definite integrals.

The advent of electronic computers in the middle of the twentieth century has drastically increased our capability of performing approximate computations, and has essentially transformed them into a dominant tool for the analysis of mathematical models. Computers have dramatically broadened the applicability range of mathematical methods in many traditional areas, such as mechanics, physics, and engineering. They have also facilitated a rapid expansion of the mathematical methods into various non-traditional fields, such as management, economics, finance, chemistry, biology, psychology, linguistics, ecology, and others.

Computers provide a capability of storing large (but still finite) arrays of numbers, and performing arithmetic operations with these numbers according to a given program that would run with large a (but still finite) execution speed. Therefore, computers are only appropriate for studying those models that are described by no more than finite sets of numbers, and require no more than finite sequences of arithmetic operations, as well as comparisons between numbers, to be performed. (The latter are typically needed for the automated control of subsequent computations.)

In the traditional fields, one frequently employs such mathematical models as functions, derivatives, integrals, and differential equations. To enable the use of computers, these original models must therefore be (approximately) replaced by the new models that would only be based on finite arrays of numbers supplemented by finite sequences of arithmetic operations for their processing (i.e., finite algorithms). For
example, a function can be replaced by a table of its numerical values; the derivative

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

can be replaced by an approximate formula, such as

$$f'(x) \approx \frac{f(x + h) - f(x)}{h},$$

where \(h\) is fixed (and small); a definite integral can be replaced by its integral sum; a boundary value problem for the differential equation can be replaced by the problem of finding its solution at the discrete nodes of some grid, so that by taking a suitable (i.e., sufficiently small) grid size an arbitrary desired accuracy can be achieved. In so doing, among the two methods that could seem equivalent at a first glance, one may produce good results while the other may turn out completely inapplicable, because the approximate solution it generates would not approach the exact solution as the grid size decreases, or because of an overly strong sensitivity of the approximate solution to small round-off errors.

The subject of numerical analysis is precisely the theory of those models and algorithms that are applicable, i.e., that can be efficiently implemented on electronic computers. This theory is intimately connected with many other branches of mathematics: Approximation theory and interpolation of functions, ordinary and partial differential equations, integral equations, complexity theory for functional classes and algorithms, etc., as well as with the theory and practice of programming languages. In general, both the exploratory capacity and the methodological advantages that electronic computers deliver to numerous applied areas are truly unparalled. Modern numerical methods allow, for example, to compute the flow of fluid around a given aerodynamic configuration, e.g., an airplane, which in most cases would present an unsurmountable task for analytical methods (like a non-tabular integral).

Moreover, the use of computers has enabled an entire new scientific methodology known as computational experiment, i.e., computations aimed at verifying the hypotheses, as well as at monitoring the behavior of the model, when it is not known ahead of time what may interest the researcher. In fact, computational experiment may provide a sufficient level of feedback for the original formulation of the problem to be noticeably refined. In other words, numerical computations help accumulate the vital information that eventually allows one to identify the most interesting cases and results in a given area of study. Many remarkable observations, and even discoveries, have been made along this route that empowered the development of the theory and have found important practical applications as well.

Computers have also facilitated the application of mathematical methods to non-traditional areas, for which few or no “compact” mathematical models, such as differential equations, are readily available, but other models can be built that lend themselves to the analysis by means of a computer. A model of this kind can often be interpreted as a direct numerical counterpart (such as encoding) of the object of interest and of the pertinent relations between its elements (e.g., a language or its
Introduction

abridged subset and the corresponding words and phrases). The very possibility of studying such models on a computer prompts their construction, which, in turn, requires that the rules and guiding principles that govern the original object be clearly and unambiguously identified. On the other hand, the results of computer simulations, e.g., a machine translation of the simplified text from one language to another, provide a practical criterion for assessing the adequacy of the theories that constitute the foundation of the corresponding mathematical model (e.g., linguistic theories).

Furthermore, electronic computers have made it possible to analyze the probabilistic models that require large amounts of test computations, as well as the so-called imitation models that describe the object or phenomenon of interest without simplifications (e.g., functional properties of a telephone network).

The variety of problems that can benefit from the use of computers is huge. For solving a given problem, one would obviously need to know enough specific detail. Clearly, this knowledge cannot be obtained ahead of time for all possible scenarios.

Therefore, the purpose of this book is rather to provide a systematic perspective on those fundamental ideas and concepts that span across different applied disciplines and that can be considered established in the field of numerical analysis. Having mastered the material of the book, one should encounter little or no difficulties when receiving subsequent specialized training required for the successful work in a given research or industrial field. The general methodology and principles of numerical analysis are illustrated in the book by “sampling” the methods designed for mathematical analysis, linear algebra, and differential equations. The reason for this particular selection is that the aforementioned methods are most mature, lead to a number of well-known efficient algorithms, and are extensively used for solving various applied problems that are often quite distant from one another.

Let us mention here some of the general ideas and concepts that require the most thorough attention in every particular setting, and that acquire a concrete interpretation and meaning in the context of each specific problem that needs to be solved on a computer. They are discretization of the problem, conditioning of the problem, numerical error, and computational stability of a given algorithm. In addition, comparison of the algorithms along different lines obviously plays a central role when selecting a specific method. The key criteria for comparison are accuracy, storage and operation count requirements, as well as efficiency of utilization of the input information. On top of that, different algorithms may vary in how amenable they are to parallelization — a technique that allows one to conduct computations simultaneously on multi-processor computer platforms. Finally, yet another important criterion is whether or not one can construct a combined analytical/numerical approach that would be efficient and useful in a given setting.

In the rest of the Introduction, we provide a brief overview of the foregoing notions and concepts. It helps create a general perspective on the subject of numerical mathematics, and establishes a foundation for studying the subsequent material.
1.1 Discretization

Let $f(x)$ be a function of the continuous argument $x \in [0, 1]$. Assume that this function provides (some of) the required input data for a given problem that needs to be approximately solved on a computer. The value of the function $f$ at every given $x$ can be either measured or obtained numerically. Then, to store this function in the memory of a computer, one may need to approximately characterize it with the table of values at a finite set of points: $x_1, x_2, \ldots, x_n$. This is an elementary example of discretization: The problem of storing the function defined on the interval $[0, 1]$, which is a continuum of points, is replaced by the problem of storing a table of its discrete values at the subset of points $x_1, x_2, \ldots, x_n$ that all belong to this interval.

Let now $f(x)$ be sufficiently smooth, and assume that we need to calculate its derivative at a given point $x$. The problem of exactly evaluating the expression $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$ that contains a limit can be replaced by the problem of computing an approximate value of this expression using one of the following formulae:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}, \quad (1.1)$$

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}, \quad (1.2)$$

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}. \quad (1.3)$$

Similarly, the second derivative $f''(x)$ can be replaced by the finite formula

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}. \quad (1.4)$$

It is easy to see (Exercise 1) that all these formulae become more and more accurate as $h$ becomes smaller. Moreover, for every fixed $h$, each formula (1.1)–(1.4) will only require a finite set of values of $f$ and a finite number of arithmetic operations. These formulae are examples of discretization for the derivatives $f'(x)$ and $f''(x)$.

Let us now consider a boundary value problem:

$$\frac{d^2 y}{dx^2} - x^2 y = \cos x, \quad 0 \leq x \leq 1,$$

$$y(0) = 2, \quad y(1) = 3, \quad (1.5)$$

where the unknown function $y = y(x)$ is defined on the interval $0 \leq x \leq 1$. To construct a discrete approximation of problem (1.5), let us first partition the interval
Introduction

[0, 1] into \( N \) equal sub-intervals of size \( h = N^{-1} \). Instead of the continuous function \( y(x) \), we will be looking for a finite set of its values \( y_0, y_1, \ldots, y_N \) on the grid \( x_k = kh, k = 0, 1, \ldots, N \). At the interior nodes of this grid: \( x_k, k = 1, 2, \ldots, N - 1 \), we can approximately replace the second derivative \( y''(x) \) by expression (1.4). After substituting into the differential equation of (1.5) this yields:

\[
\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - (kh)^2 y_k = \cos(kh), \quad k = 1, 2, \ldots, N - 1. \tag{1.6}
\]

Furthermore, the boundary conditions at \( x = 0 \) and at \( x = 1 \) from (1.5) translate into:

\[
y_0 = 2, \quad y_N = 3. \tag{1.7}
\]

The system of \( N + 1 \) linear algebraic equations (1.6), (1.7) contains exactly as many unknowns \( y_0, y_1, \ldots, y_N \), and renders a discrete counterpart of the boundary value problem (1.5). One can, in fact, show that the finer the grid, i.e., the larger the \( N \), the more accurate will the approximation be that the discrete solution of problem (1.6), (1.7) provides for the continuous solution of problem (1.5). Later, this fact will be formulated and proven rigorously.

Let us denote the continuous boundary value problem (1.5) by \( M_\infty \), and the discrete boundary value problem (1.6), (1.7) by \( M_N \). Then, one can say that we have associated an infinite sequence of discrete problems \( \{M_N\} \), \( N = 2, 3, \ldots \), with the continuous problem \( M_\infty \). When computing the solution to a given problem \( M_N \) for any fixed \( N \), we only have to work with a finite array of numbers that specify the input data, and with a finite set of unknown quantities \( y_0, y_1, y_2, \ldots, y_N \). It is, however, the entire infinite sequence of finite discrete models \( \{M_N\} \) that plays the central role from the standpoint of numerical mathematics. Indeed, as those models happen to be more and more accurate, one can always choose a sufficiently large \( N \) that would guarantee any desired accuracy of approximation.

In general, there are many different ways of transitioning from a given continuous problem \( M_\infty \) to the sequence \( \{M_N\} \) of its discrete counterparts. In other words, the approximation (1.6), (1.7) of the boundary value problem (1.5) is by no means the only one possible. Let \( \{M_N\} \) and \( \{M'_N\} \) be two sequences of approximations, and let us also assume that the computational costs of obtaining the discrete solutions of \( M_N \) and \( M'_N \) are the same. Then, a better method of discretization would be the one that provides for the same accuracy of approximation with a smaller value of \( N \).

Let us also note that for two seemingly equivalent discretization methods \( M_N \) and \( M'_N \), it may happen that one will approximate the continuous solution of problem \( M_\infty \) with an increasingly high accuracy as \( N \) increases, whereas the other will yield “an approximate solution” that would bear less and less resemblance of the continuous solution of \( M_\infty \). We will encounter situations like that in Part III of the book, where we also discuss how the corresponding difficulties can be partially or fully overcome.

Exercises

1. Let \( f(x) \) have as many bounded derivatives as needed. Show that the approximation error of formulae (1.1), (1.2), (1.3), and (1.4), is \( O(h) \), \( O(h^2) \), \( O(h^3) \), and \( O(h^2) \).
1.2 Conditioning

Speaking in most general terms, for any given problem one can basically identify the input data and the output result(s), i.e., the solution, so that the former determine the latter. The sensitivity of the solution, or some of its key characteristics, to perturbations in the input data may differ strongly for different problems that could otherwise look very similar. If this sensitivity is “low” (weak), then the problem is said to be well conditioned; if, conversely, the sensitivity is “high” then the problem is ill conditioned. The notions of low and high are, of course, problem-dependent. Typically, not only ill conditioned problems require excessively accurate specification of the input data, but also appear more difficult for computations.

Consider, for example, the quadratic equation \( x^2 - 2\alpha x + 1 = 0 \) for \(|\alpha| > 1\). It has two real roots that can be expressed as functions of the argument \( \alpha \): \( x_{1,2} = \alpha \pm \sqrt{\alpha^2 - 1} \). We will interpret \( \alpha \) as the datum in the problem, and \( x_1 = x_1(\alpha) \) and \( x_2 = x_2(\alpha) \) as the corresponding solution. Clearly, the sensitivity of the solution to the perturbations of \( \alpha \) can be characterized by the magnitude of the derivatives \( \frac{dx_{1,2}}{d\alpha} = 1 \pm \frac{\alpha}{\sqrt{\alpha^2 - 1}} \). Indeed, \( \Delta x_{1,2} \approx \frac{dx_{1,2}}{d\alpha} \Delta \alpha \). One can easily see that the derivatives \( \frac{dx_{1,2}}{d\alpha} \) are small for large \(|\alpha|\), but they become large when \( \alpha \) approaches 1. We can therefore conclude that the problem of finding the roots of \( x^2 - 2\alpha x + 1 = 0 \) is well conditioned when \(|\alpha| \gg 1\), and ill conditioned when \(|\alpha| = \mathcal{O}(1)\). We should also note that conditioning can be improved if, instead of the original quadratic equation, we consider its equivalent \( x^2 - \frac{1+\beta^2}{\beta} x + 1 = 0 \), where \( \beta = \alpha + \sqrt{\alpha^2 - 1} \). In this case, \( x_1 = \beta \) and \( x_2 = \beta^{-1} \); the two roots coincide for \(|\beta| = 1\), or equivalently, \(|\alpha| = 1\). However, the problem of evaluating \( \beta = \beta(\alpha) \) is still ill conditioned near \(|\alpha| = 1\).

Our next example involves a simple ordinary differential equation. Let \( y = y(t) \) be the concentration of some substance at the time \( t \), and assume that it satisfies

\[
\frac{dy}{dt} - 10y = 0.
\]

Let us take an arbitrary \( t_0, 0 \leq t_0 \leq 1 \), and measure the concentration \( y_0 = y(t_0) \) approximately at this moment of time, thus obtaining:

\[
y_{t=t_0} = y_0.
\]

Our overall task will be to determine the concentration \( y = y(t) \) at all other moments of time \( t \) from the interval \([0, 1]\).

If we knew the quantity \( y_0 = y(t_0) \) exactly, then we could have used the exact formula available for the concentration:

\[
y(t) = y_0 e^{10(t-t_0)}.
\]

We, however, only know the approximate value \( y_0^* \approx y_0 \) of the unknown quantity \( y_0 \). Therefore, instead of (1.8), the next best thing is to employ the approximate formula:

\[
y^*(t) = y_0^* e^{10(t-t_0)}.
\]
Introduction

Clearly, the error \( y^\ast - y \) of the approximate formula (1.9) is given by

\[
y^\ast(t) - y(t) = (y^\ast_0 - y_0)e^{10(t-t_0)}, \quad 0 \leq t \leq 1.
\]

Assume now that we need to measure \( y^\ast_0 \) the accuracy \( \delta \), \(|y^\ast_0 - y_0| < \delta\), that would be sufficient to guarantee an initially prescribed tolerance \( \varepsilon \) for determining \( y(t) \) everywhere on the interval \( 0 \leq t \leq 1 \), i.e., would guarantee the error estimate:

\[
|y^\ast(t) - y(t)| < \varepsilon, \quad 0 \leq t \leq 1.
\]

It is easy to see that \( \max_{0 \leq t \leq 1} |y^\ast(t) - y(t)| = |y^\ast(1) - y(1)| = |y^\ast_0 - y_0|e^{10(1-t_0)} \). This yields the following constraint that the accuracy \( \delta \) of measuring \( y_0 \) must satisfy:

\[
\delta \leq \varepsilon e^{-10(1-t_0)}.
\]

(1.10)

Let \( y_0 \) be measured at the moment of time \( t_0 = 0 \). Then, inequality (1.10) would imply that this measurement has to be \( e^{10} \) times, i.e., thousands of times, more accurate than the required guaranteed accuracy of the result \( \varepsilon \). In other words, the answer \( y(t) \) appears quite sensitive to the error in specifying the input data \( y_0 \), and the problem is ill conditioned.

On the other hand, if \( y_0 \) were to be measured at \( t_0 = 1 \), then \( \delta = \varepsilon \), and it would be sufficient to conduct the measurement with a considerably lower accuracy than the one required in the case of \( t_0 = 0 \). This problem is well conditioned.

Exercises

1. Consider the problem of computing \( y(x) = \frac{1}{x+1} \) as a function of \( x \), for \( x \in (1/2, 1) \) and also for \( x \in (-1, 0) \). On which of the two intervals is this problem better conditioned with respect to the perturbations of \( x \)?

2. Let \( y = \sqrt{2} - 1 \). Equivalently, one can write \( y = (\sqrt{2} + 1)^{-1} \). Which of the two formulæ is more sensitive to the error when \( \sqrt{2} \) is approximated by a finite decimal fraction?

\[ \text{Hint.} \] Compare absolute values of derivatives for the functions \((x - 1)\) and \((x + 1)^{-1}\).

1.3 Error

In any computational problem, one needs to find the solution given some appropriate input data. If the solution can be obtained with an ideal accuracy, then there is no error. Typically, however, there is a certain error content in every feasible numerical solution. This error may be attributed to (at least) three different mechanisms.

First, the input data are often specified with some degree of uncertainty that, in turn, will generate uncertainty in the corresponding output. Then, the solution to the problem of interest may only be obtained with an error called the \textit{unavoidable error}. 
Second, even if we eliminate the foregoing uncertainty by fixing the input data, and subsequently compute the solution using an approximate method, then we still won’t find the solution that would exactly correspond to the specified data. There will be an error due to the choice of an approximate computational procedure.

Third, the chosen approximate method is not implemented exactly either, because of the round-off errors that arise when performing computations on a real machine.

Therefore, the overall error in the solution consists of the unavoidable error, the error of the method, and the round-off error. We will now illustrate these concepts.

1.3.1 Unavoidable Error

Assume that we need to find the value \( y \) of some function \( y = f(x) \) for a given \( x = x_0 \). The quantity \( x_0 \), as well as the relation \( f \) itself that associates the value of the function with every given value of its argument, are considered the input data of the problem, whereas the quantity \( y = y(x_0) \) will be its solution.

Now let the function \( f(x) \) be known approximately rather than exactly, say, \( f(x) \approx \sin x \), and suppose that \( f(x) \) may differ from \( \sin x \) by no more than a specified \( \varepsilon > 0 \):

\[
\sin x - \varepsilon \leq f(x) \leq \sin x + \varepsilon.
\] (1.11)

Let the value of the argument \( x = x_0 \) be also measured approximately: \( x = x_0^* \), so that regarding the actual \( x_0 \) we can only say that

\[
x_0^* - \delta \leq x_0 \leq x_0^* + \delta,
\] (1.12)

where \( \delta > 0 \) characterizes the accuracy of the measurement.

One can easily see from Figure 1.1 that any point on the interval \([a, b]\) of variable \( y \), where \( a = \sin(x_0^* - \delta) - \varepsilon \) and \( b = \sin(x_0^* + \delta) + \varepsilon \), can serve in the capacity of \( y = y(x_0) \). Clearly, by taking an arbitrary \( y^* \in [a, b] \) as the approximate value of \( y = f(x_0) \), we can always guarantee the error estimate

\[
|y - y^*| \leq |b - a|.
\] (1.13)

For the given uncertainty in the input data, see formulae (1.11) and (1.12), this estimate cannot be considerably improved. In fact, the best error estimate that one can guarantee is obtained by choosing \( y^* \) exactly in the middle of the interval \([a, b]\):

\[
y^* = y_{\text{opt}}^* = (a + b)/2.
\]
From Figure 1.1 we then conclude that
\[ |y - y^*| \leq |b - a|/2. \tag{1.14} \]

This inequality transforms into an exact equality when \( y(x_0) = a \) or when \( y(x_0) = b \).

As such, the quantity \( |b - a|/2 \) is precisely the unavoidable (or irreducible) error, i.e., the minimum error content that will always be present in the solution and that one won’t be able to “dodge” no matter how the approximation \( y^* \) is actually chosen, simply because of the uncertainty that exists in the input data. For the optimal choice of the approximate solution \( y^*_{\text{opt}} \) the smallest error (1.14) can be guaranteed; otherwise, the appropriate error estimate is (1.13).

We see, however, that the optimal error estimate (1.14) is not that much better than the general estimate (1.13). We will therefore stay within reason if we interpret any otherwise, the appropriate error estimate is (1.13).

Along with the simplest illustrative example of Figure 1.1, let us consider another example that would be a little more realistic and would involve one of the most common problem formulations in numerical analysis, namely, that of reconstructing a function of continuous argument given its tabulated values at some discrete set of points. More precisely, let the values \( f(x_k) \) of the function \( f = f(x) \) be known at the equidistant grid nodes \( x_k = kh, h > 0, k = 0, \pm 1, \pm 2, \ldots \). Let us also assume that the first derivative of \( f(x) \) is bounded everywhere: \( |f'(x)| \leq 1 \), and that together with \( f(x_k) \), this is basically all the information that we have about \( f(x) \). We need to be able to obtain the (approximate) value of \( f(x) \) at an arbitrary “intermediate” point \( x \) that does not necessarily coincide with any of the nodes \( x_k \).

A large variety of methods have been developed in the literature for solving this problem. Later, we will consider interpolation by means of algebraic (Chapter 2) and trigonometric (Chapter 3) polynomials. There are other ways of building the approximating polynomials, e.g., the least squares and trigonometric (Chapter 3) polynomials. There are other ways of building the approximating polynomials, e.g., the least squares and trigonometric (Chapter 3) polynomials. There are other ways of building the approximating polynomials, e.g., the least squares and trigonometric (Chapter 3) polynomials. There are other ways of building the approximating polynomials, e.g., the least squares and trigonometric (Chapter 3) polynomials. There are other ways of building the approximating polynomials, e.g., the least squares and trigonometric (Chapter 3) polynomials. Each specific method will obviously have its own accuracy. We, however, are going to show that irrespective of any particular technique used for reconstructing \( f(x) \), there will always be error due to incomplete specification of the input data. This error merely reflects the uncertainty in the formulation, it is unavoidable and cannot be suppressed by any “smart” choice of the reconstruction procedure.

Consider the simplest case \( f(x_k) = 0 \) for all \( k = 0, \pm 1, \pm 2, \ldots \). Clearly, the function \( f_1(x) = 0 \) has the required trivial table of values, and also \( |f_1'(x)| \leq 1 \). Along with \( f_1(x) \), it is easy to find another function that would satisfy the same constraints, e.g., \( f_2(x) = \frac{b}{\pi} \sin \left( \frac{x \pi}{h} \right) \). Indeed, \( f_2(x_k) = 0 \), and \( |f_2'(x)| = |\sin \left( \frac{x \pi}{h} \right)| \leq 1 \). We therefore see that there are at least two different functions that cannot be told apart based on the available information. Consequently, the error \( \max_{x} |f_1(x) - f_2(x)| = O(h) \) is unavoidable when reconstructing the function \( f(x) \), given its tabulated values \( f(x_k) \) and the fact that its first derivative is bounded, no matter what specific reconstruction methodology may be employed.
For more on the notion of the unavoidable error in the context of reconstructing continuous functions from their discrete values see Section 2.2.4 of Chapter 2.

1.3.2 Error of the Method

Let $y^* = \sin x_0^*$. The number $y^*$ belongs to the interval $[a, b]$; it can be considered a non-improvable approximate solution of the first problem analyzed in the previous Section 1.3.1. For this solution, the error satisfies estimate (1.13) and is unavoidable. The point $y^* = \sin x_0^*$ has been selected among other points of the interval $[a, b]$ only because it is given by the formula convenient for subsequent analysis.

To evaluate the quantity $y^* = \sin x_0^*$ on a computer, let us use Taylor’s expansion for the function $\sin x$:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$$

Thus, for computing $y^*$ one can take one of the following approximate expressions:

$$y^* \approx y_1^* = x_0^*,$$

$$y^* \approx y_2^* = x_0^* - \frac{(x_0^*)^3}{3!},$$

$$\ldots$$

$$y^* \approx y_n^* = \sum_{k=1}^{n} (-1)^{k-1} \frac{(x_0^*)^{2k-1}}{(2k-1)!}.$$ (1.15)

By choosing a specific formulae (1.15) for the approximate evaluation of $y^*$, we select our method of computation. The quantity $|y^* - y_n^*|$ is then known as the error of the computational method. In fact, we are considering a family of methods parameterized by the integer $n$. The larger the $n$ the smaller the error, see (1.15); and by taking a sufficiently large $n$ we can always make sure that the associated error will be smaller than any initially prescribed threshold.

It, however, does not make sense to drive the computational error much further down than the level of the unavoidable error. Therefore, the number $n$ does not need to be taken excessively large. On the other hand, if $n$ is taken too small so that the error of the method appears much larger than the unavoidable error, then one can say that the chosen method does not fully utilize the information about the solution that is contained in the input data, or equivalently, loses a part of this information.

1.3.3 Round-off Error

Assume that we have fixed the computational method by selecting a particular $n$ in (1.15), i.e., by setting $y^* \approx y_n^*$. When calculating this $y_n^*$ on a real computer, we will, generally speaking, obtain a different value $\tilde{y}_n^*$ because of the rounding. Rounding is an intrinsic feature of the floating-point arithmetics on electronic computers, as these computers only operate with the numbers that can be represented as finite binary fractions of a given fixed length. As such, all other real numbers (e.g., infinite fractions) may only be stored approximately in the computer memory, and the
corresponding approximation procedure is known as rounding. The error $|y^*_n - \tilde{y}^*_n|$ is called the round-off error.

This error shall not noticeably exceed the error of the computational method. Otherwise, a loss of the overall accuracy will be incurred due to the round-off error.

**Exercises**

1. Assume that we need to calculate the value $y = f(x)$ of some function $f(x)$, while there is an uncertainty in the input data $x^*: x^* - \delta \leq x \leq x^* + \delta$.

   How does the corresponding unavoidable error depend on $x^*$ and on $\delta$ for the following functions:

   a) $f(x) = \sin x$;
   
   b) $f(x) = \ln x$, where $x > 0$?

   For what values of $x^*$, obtained by approximately measuring the “loose” quantity $x$ with the accuracy $\delta$, can one guarantee only a one-sided upper bound for $\ln x$ in problem b)? Find this upper bound.

2. Let the function $f(x)$ be defined by its values sampled on the grid $x_k = kh$, where $h = 1/N$ and $k = 0, \pm 1, \pm 2, \ldots$. In addition to these discrete values, assume that $\max_x |f''(x)| \leq 1$.

   Prove that as the available input data are incomplete, they do not, generally speaking, allow one to reconstruct the function at an arbitrary given point $x$ with the accuracy better than the unavoidable error $\varepsilon(h) = h^2/\pi^2$.

   **Hint.** Show that along with the function $f(x) \equiv 0$, which obviously has all its grid values equal to zero, another function, $\psi(x) = (h^2/\pi^2) \sin(N\pi x)$, also has all its grid values equal to zero, and satisfies the condition $\max_x |\psi''(x)| \leq 1$, while $\max_x |f(x) - \psi(x)| = h^2/\pi^2$.

3. Let $f = f(x)$ be a function, such that the absolute value of its second derivative does not exceed one. Show that the approximation error for the formula:

$$f'(x) \approx \frac{f(x + h) - f(x)}{h}$$

will not exceed $h$.

4. Let $f = f(x)$ be a function that has bounded second derivative: $\forall x:\ |f''(x)| \leq 1$. For any $x$, the value of the function $f(x)$ is measured and comes out to be equal to some $f^*(x)$; in so doing we assume that the accuracy of the measurement guarantees the following estimate:

$$|f(x) - f^*(x)| \leq \varepsilon, \quad \varepsilon > 0.$$ 

   Suppose now that we need to approximately evaluate the first derivative $f'(x)$.

---

\(^1\)Hereafter, we will be using the symbol “*” to indicate the increased level of difficulty for a given problem.
a) How shall one choose the parameter $h$ so that to minimize the guaranteed error estimate of the approximate formula:

$$f'(x) \approx \frac{f'(x + h) - f'(x)}{h}.$$

b) Show that given the existing uncertainty in the input data, the unavoidable error of evaluating $f'(x)$ is at least $O(\sqrt{\varepsilon})$, no matter what specific method is used.

**Hint.** Consider two functions, $f(x) \equiv 0$ and $f^*(x) = \varepsilon \sin(x/\sqrt{\varepsilon})$. Clearly, the absolute value of the second derivative for either of these two functions does not exceed one. Moreover, $\max_x |f(x) - f^*(x)| \leq \varepsilon$. At the same time,

$$\left| \frac{df^*}{dx} - \frac{df}{dx} \right| = \left| \sqrt{\varepsilon} \cos \left( \frac{x}{\sqrt{\varepsilon}} \right) \right| = O(\sqrt{\varepsilon}).$$

By comparing the solutions of sub-problems a) and b), verify that the specific approximate formula for $f'(x)$ given in a) yields the error of the irreducible order $O(\sqrt{\varepsilon})$; and also show that the unavoidable error is, in fact, exactly of order $O(\sqrt{\varepsilon})$.

5. For storing the information about a linear function $f(x) = kx + b$, $\alpha \leq x \leq \beta$, that satisfies the inequalities: $0 \leq f(x) \leq 1$, we use a table with six available cells, such that one of the ten digits: 0, 1, 2,..., 9, can be written into each cell.

What is the unavoidable error of reconstructing the function, if the foregoing six cells of the table are filled according to one of the following recipes?

a) The first three cells contain the first three digits that appear right after the decimal point when the number $f(\alpha)$ is represented as a normalized decimal fraction; and the remaining three cells contain the first three digits after the decimal point in the normalized decimal fraction for $f(\beta)$.

b) Let $\alpha = 0$ and $\beta = 10^{-2}$. The first three cells contain the first three digits in the normalized decimal fraction for $k$, the fourth cell contains either 0 or 1 depending on the sign of $k$, and the remaining two cells contain the first two digits after the decimal point in the normalized decimal fraction for $b$.

c) Show that irrespective of any specific strategy for filling out the aforementioned six-cell table, the unavoidable error of reconstructing the linear function $f(x) = kx + b$ is always at least $0.5 \cdot 10^{-3}$.

**Hint.** Build $10^6$ different functions from the foregoing class, such that the maximum modulus of the difference between any two of them will be at least $10^{-3}$.

### 1.4 On Methods of Computation

Suppose that a mathematical model is constructed for studying a given object or phenomenon, and subsequently this model is analyzed using mathematical and com-
Introduction

Putational means. For example, under certain assumptions the following problem

\[
\frac{d^2y}{dt^2} + y = 0, \quad t \geq 0, \quad y(0) = 0, \quad \frac{dy}{dt} \bigg|_{t=0} = 1, \quad (1.16)
\]

can provide an adequate mathematical model for small oscillations of a pendulum, where \(y(t)\) is the pendulum displacement from its equilibrium at the time \(t\).

A study of harmonic oscillations based on this mathematical model, i.e., on the Cauchy problem (1.16), can benefit from a priori knowledge about the physical nature of the object of study. In particular, one can predict, based on physical reasoning, that the motion of the pendulum will be periodic. However, once the mathematical model (1.16) has been built, it becomes a separate and independent object that can be investigated using any available mathematical tools, including those that have little or no relation to the physical origins of the problem. For example, the numerical value of the solution \(y = \sin t\) to problem (1.16) at any given moment of time \(t = z\) can be obtained by expanding \(\sin z\) into the Taylor series:

\[
\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \ldots,
\]

and subsequently taking its appropriate partial sum. In so doing, representation of the function \(\sin t\) as a power series hardly admits any tangible physical interpretation.

In general, when solving a given problem on the computer, many different methods, or different algorithms, can be used. Some of them may prove far superior compared to the others. In subsequent parts of the book, we are going to describe a number of established, robust and efficient, algorithms for frequently encountered classes of problems in numerical analysis. In the meantime, let us briefly explain how the algorithms may differ.

Assume that for computing the solution \(y\) to a given problem we can employ two algorithms, \(A_1\) and \(A_2\), that yield the approximate solutions \(y_1^* = A_1(X)\) and \(y_2^* = A_2(X)\), respectively, where \(X\) denotes the entire required set of the input data. In so doing, a variety of situations may occur.

1. The algorithm \(A_2\) may be more accurate than the algorithm \(A_1\), i.e.,

\[
|y - y_1^*| \gg |y - y_2^*|.
\]

For example, let us approximately evaluate \(y = \sin x\bigg|_{x=0.1}\) using the expansion:

\[
y_n^* = \sum_{k=1}^{n} (-1)^{k-1} \frac{x^{2k-1}}{(2k-1)!}, \quad (1.17)
\]

The algorithm \(A_1\) will correspond to taking \(n = 1\) in formula (1.17), and the algorithm \(A_2\) will correspond to taking \(n = 2\) in formula (1.17). Then, obviously,

\[
|\sin 0.1 - y_1^*| \gg |\sin 0.1 - y_2^*|.
\]
2. Both algorithms may provide the same accuracy, but the computation of \( y^*_1 = A_1(X) \) may require many more arithmetic operations than the computation of \( y^*_2 = A_2(X) \). Suppose, for example, that we need to find the value of

\[
y = 1 + x + x^2 + \ldots + x^{1023} \quad \text{(clearly, } y = \frac{1 - x^{1024}}{1 - x})
\]

for \( x = 0.99 \). Let \( A_1 \) be the algorithm that would perform the computations directly using the given formula, i.e., by raising 0.99 to the powers 1, 2, \ldots, 1023 one after another, and subsequently adding the results. Let \( A_2 \) be the algorithm that would perform the computations according to the formula:

\[
y = \frac{1 - 0.99^{1024}}{1 - 0.99}.
\]

The accuracy of these two algorithms is the same — both are absolutely accurate provided that there are no round-off errors. However, the first algorithm requires considerably more arithmetic operations, i.e., it is computationally more expensive. Namely, for successively computing

\[
x, \quad x^2 = x \cdot x, \quad \ldots, \quad x^{1023} = x^{1022} \cdot x,
\]

one will have to perform 1022 multiplications. On the other hand, to compute \( 0.99^{1024} \) one only needs 10 multiplications:

\[
0.99^2 = 0.99 \cdot 0.99, \quad 0.99^4 = 0.99^2 \cdot 0.99^2, \quad \ldots, \quad 0.99^{1024} = 0.99^{512} \cdot 0.99^{512}.
\]

3. The algorithms, again, may yield the same accuracy, but \( A_1(X) \) may be computationally unstable, whereas \( A_2(X) \) may be stable. For example, to evaluate \( y = \sin x \) with the prescribed tolerance \( \varepsilon = 10^{-3} \), i.e., to guarantee \( |y - y^*| \leq 10^{-3} \), let us employ the same finite Taylor expansion as in formula (1.17):

\[
y^*_1 = y^*_1(x) = \sum_{k=1}^{n} (-1)^{k-1} \frac{x^{2k-1}}{(2k-1)!}, \quad (1.18)
\]

where \( n = n(\varepsilon) \) shall be chosen so as to make sure that the inequality

\[
|y - y^*_1| \leq 10^{-3}
\]

will hold. The first algorithm \( A_1 \) will compute the result directly according to (1.18). If \( |x| \leq \pi/2 \), then by noticing that the following inequality holds already for \( n = 5 \):

\[
\frac{1}{(2n-1)!} \left( \frac{\pi}{2} \right)^{2n-1} \leq 10^{-3},
\]

we can reduce the sum (1.18) to

\[
y^*_1 = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}.
\]
Introduction

Clearly, the computations by this formula will only be weakly sensitive to round-off errors when evaluating each term on the right-hand side. Moreover, as for $|x| \leq \pi/2$, those terms rapidly decay when the power grows, there is no room for the cancellation of significant digits, and the algorithm $A_1$ will be computationally stable.

Consider now $|x| \gg 1$; for example, $x = 100$. Then, for achieving the prescribed accuracy of $\varepsilon = 10^{-3}$, the number $n$ should satisfy the inequality:

$$\frac{100^{2n+1}}{(2n+1)!} \leq 10^{-3},$$

which yields an obvious conservative lower bound for $n$: $n > 48$. This implies that the terms in sum (1.18) become small only for sufficiently large $n$. At the same time, the first few leading terms in this sum will be very large. A small relative error committed when computing those terms will result in a large absolute error; and since taking a difference of large quantities to evaluate a small quantity $\sin x$ ($|\sin x| \leq 1$) is prone to the loss of significant digits, the algorithm $A_1$ in this case will be computationally unstable.

On the other hand, in the case of large $x$ a stable algorithm $A_2$ for evaluating $\sin x$ is also easy to build. Let us represent a given $x$ in the form $x = l\pi + z$, where $|z| \leq \pi/2$ and $l$ is integer. Then,

$$\sin x = (-1)^l \sin z,$$

$$y^*_2 = A_2(x) = (-1)^l \left(z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!}\right).$$

This algorithm has the same stability properties as the algorithm $A_1$ for $|x| \leq \pi/2$.

4. Finally, the algorithm may be either convergent or divergent. Suppose we need to compute the value of $y = \ln(1 + x)$. Let us employ the power series

$$y = \ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

(1.19)

and set

$$y^*_n(x) \approx y^*_n = \sum_{k=1}^{n} (-1)^{k+1} \frac{x^k}{k}.$$  

(1.20)

In so doing, we will obtain a method of approximately evaluating $y = \ln(1 + x)$ that would depend on $n$ as on a parameter.

If $|x| = q < 1$, then $\lim_{n \to \infty} y^*_n(x) = y(x)$, i.e., the error committed when computing $y(x)$ according to formula (1.20) will be vanishing as $n$ increases. If, however, $x > 1$, then $\lim_{n \to \infty} y^*_n(x) = \infty$, because the convergence radius for the series (1.19) is $r = 1$.

In this case the algorithm based on formula (1.20) diverges, and cannot be used for computations.

In the book, we are going to discuss some other characteristics of numerical algorithms as well. We will see the algorithms that admit easy parallelization, as well as those that are limited to sequential computations; algorithms that automatically
Theoretical Introduction to Numerical Analysis

adopt to specific characteristics of the input data, such as smoothness, as well as those that only partially take it into account; algorithms that have a straightforward logical structure, as well as more elaborate ones.

Exercises

1. Propose an algorithm for evaluating \( y = \ln(1 + x) \) that would also apply to \( x > 1 \).

2. Consider the problem of evaluating the sequence of numbers \( x_0, x_1, \ldots, x_N \) that satisfy the difference equations:

\[
2x_n - x_{n+1} = 1 + n^2/N^2, \quad n = 0, 1, \ldots, N - 1,
\]

and the additional condition:

\[
x_0 + x_N = 1. \tag{1.21}
\]

We introduce two algorithms for computing \( x_n \). First, let

\[
x_n = u_n + cv_n, \quad n = 0, 1, \ldots, N. \tag{1.22}
\]

Then, in the algorithm \( A_1 \) we define \( u_n, n = 0, 1, \ldots, N \), as solution of the system

\[
2u_n - u_{n+1} = 1 + n^2/N^2, \quad n = 0, 1, \ldots, N - 1, \tag{1.23}
\]

subject to the initial condition

\[
u_0 = 0. \tag{1.24}
\]

Consequently, the sequence \( v_n, n = 0, 1, \ldots, N \), is defined by the equalities

\[
2v_n - v_{n+1} = 0, \quad n = 0, 1, \ldots, N - 1, \tag{1.25}
\]

\[
v_0 = 1, \tag{1.26}
\]

and the constant \( c \) of (1.22) is obtained from the condition (1.21). In so doing, the actual values of \( u_n \) and \( v_n \) are computed consecutively using the formulae:

\[
u_{n+1} = 2u_n - (1 + n^2/N^2), \quad n = 0, 1, \ldots,
\]

\[
v_{n+1} = 2^n, \quad n = 0, 1, \ldots.
\]

In the algorithm \( A_2 \), \( u_n, n = 0, 1, \ldots, N \), is still defined as solution to system (1.23), but instead of the condition (1.24) an alternative condition \( u_N = 0 \) is employed. The sequence \( v_n, n = 0, 1, \ldots, N \), is again defined as a solution to system (1.25), but instead of the condition (1.26) we use \( v_N = 1 \).

a) Verify that the second algorithm, \( A_2 \), is stable while the first one, \( A_1 \), is ("violently") unstable.

b) Implement both algorithms on the computer and try to compare their performance for \( N = 10 \) and for \( N = 100 \).
Part I

Interpolation of Functions. Quadratures
One of the key concepts in mathematics is that of a function. In the simplest case, the function \( y = f(x) \), \( a \leq x \leq b \), could be given in the closed form, i.e., defined by means of a finite formula, say, \( y = x^2 \). This formula can subsequently be transformed into a computer code that would calculate the value of \( y = x^2 \) for every given \( x \). In real-life settings, however, the functions of interest are rarely available in the closed form. Instead, a finite array of numbers, commonly referred to as the table, would often be associated with the function \( y = f(x) \). By processing the numbers from the table in a particular prescribed way, one should be able to obtain an approximate value of the function \( f(x) \) at any point \( x \). For instance, a table can contain several leading coefficients of a power series for \( f(x) \). In this case, processing the table would mean calculating the corresponding partial sum of the series.

Let us, for example, take the function

\[
y = e^x, \quad 0 \leq x \leq 1, \quad e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots,
\]

for which the power series converges for all \( x \), and consider the table

\[
1, \frac{1}{1!}, \frac{1}{2!}, \ldots, \frac{1}{n!}
\]

of its \( n+1 \) leading Taylor coefficients, where \( n > 0 \) is given. The larger the \( n \), the more accurately one can reconstruct the function \( f(x) = e^x \) from this table. In so doing, the formula

\[
e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}
\]

is used for processing the table.

In most cases, however, the table that is supposed to characterize \( y = f(x) \) would not contain its Taylor coefficients, and would rather be obtained by sampling the values of this function at some finite set of points \( x_0, x_1, \ldots, x_n \in [a, b] \). In practice, the sampling can be rendered by either measurements or computations. This naturally gives rise to the problem of reconstructing (e.g., interpolating) the function \( f(x) \) at the “intermediate” locations \( x \) that do not necessarily coincide with any of the nodes \( x_0, x_1, \ldots, x_n \).

The two most widely used and most efficient interpolation techniques are algebraic interpolation and trigonometric interpolation. We are going to analyze both of them. In addition, in the current Part I of the book we will also consider the problem of evaluating definite integrals of a given function when the latter, again, is specified by a finite table of its numerical values. The motivation behind considering this problem along with interpolation is that the main approaches to approximate evaluation of definite integrals, i.e., to obtaining the so-called quadrature formulae, are very closely related to the interpolation techniques.
Chapter 2

Algebraic Interpolation

Let \( x_0, x_1, \ldots, x_n \) be a given set of points, and let \( f(x_0), f(x_1), \ldots, f(x_n) \) be values of the function \( f(x) \) at these points (assumed known). The one-to-one correspondence

\[
\begin{array}{cccc}
  x_0 & x_1 & \cdots & x_n \\
  f(x_0) & f(x_1) & \cdots & f(x_n)
\end{array}
\]

will be called a table of values of the function \( f(x) \) at the nodes \( x_0, x_1, \ldots, x_n \). We need to realize, of course, that for actual computer implementations one may only use the numbers that can be represented as finite binary fractions (Section 1.3.3 of the Introduction), whereas the values \( f(x_j) \) do not necessarily have to belong to this class (e.g., \( \sqrt{3} \)). Therefore, the foregoing table may, in fact, contain rounded rather than true values of the function \( f(x) \).

A polynomial \( P_n(x) \equiv P_n(x, f, x_0, x_1, \ldots, x_n) \) of degree no greater than \( n \) that has the form

\[ P_n(x) = c_0 + c_1 x + \ldots + c_n x^n \]

and coincides with \( f(x_0), f(x_1), \ldots, f(x_n) \) at the nodes \( x_0, x_1, \ldots, x_n \), respectively, is called the algebraic interpolating polynomial.

2.1 Existence and Uniqueness of Interpolating Polynomial

2.1.1 The Lagrange Form of Interpolating Polynomial

**THEOREM 2.1**

Let \( x_0, x_1, \ldots, x_n \) be a given set of distinct interpolation nodes, and let the values \( f(x_0), f(x_1), \ldots, f(x_n) \) of the function \( f(x) \) be known at these nodes. There is one and only one algebraic polynomial \( P_n(x) \equiv P_n(x, f, x_0, x_1, \ldots, x_n) \) of degree no greater than \( n \) that would coincide with the given \( f(x_k) \) at the nodes \( x_k \), \( k = 0, 1, \ldots, n \).

**PROOF** We will first show that there may be no more than one interpo-
Assume that there are two algebraic interpolating polynomials, \( P_n^{(1)}(x) \) and \( P_n^{(2)}(x) \). Then, the difference between these two polynomials, \( R_n(x) = P_n^{(1)}(x) - P_n^{(2)}(x) \), is also a polynomial of degree no greater than \( n \) that vanishes at the \( n+1 \) points \( x_0, x_1, \ldots, x_n \). However, any polynomial that is not identically equal to zero has exactly as many roots (counting the multiplicities) as its degree is. Therefore, \( R_n(x) \equiv 0 \), i.e., \( P_n^{(1)}(x) \equiv P_n^{(2)}(x) \), which proves uniqueness.

Let us now introduce the auxiliary polynomials

\[
l_k(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)}
\]

It is clear that each \( l_k(x) \) is a polynomial of degree no greater than \( n \), and that the following equalities hold:

\[
l_k(x_j) = \begin{cases} 1, & x_j = x_k, \\ 0, & x_j \neq x_k, \end{cases} \quad j = 0, 1, \ldots, n.
\]

Then, the polynomial \( P_n(x) \) given by the equality

\[
P_n(x) \overset{\text{def}}{=} P_n(x, f, x_0, x_1, \ldots, x_n) = f(x_0)l_0(x) + f(x_1)l_1(x) + \ldots + f(x_n)l_n(x)
\]

(2.1)

is precisely the interpolating polynomial sought for. Indeed, its degree is no greater than \( n \), because each term \( f(x_j)l_j(x) \) is a polynomial of degree no greater than \( n \). Moreover, it is clear that this polynomial satisfies the equalities \( P_n(x_j) = f(x_j) \) for all \( j = 0, 1, \ldots, n \).

Let us emphasize that not only have we proven Theorem 2.1, but have also written down the interpolating polynomial explicitly using formula (2.1). This formula is known as the Lagrange form of the interpolating polynomial. There are other convenient forms of the unique interpolating polynomial \( P_n(x, f, x_0, x_1, \ldots, x_n) \). The Newton form is used particularly often.

### 2.1.2 The Newton Form of Interpolating Polynomial. Divided Differences

Let \( f(x_a), f(x_b), f(x_c), f(x_d), \ldots \), be values of the function \( f(x) \) at the given nodes \( x_a, x_b, x_c, x_d, \ldots \). A Newton’s divided difference of order zero \( f(x_k) \) of the function \( f(x) \) at the point \( x_k \) is defined as simply the value of the function at this point:

\[
f(x_k) = f(x_k), \quad k = a, b, c, d, \ldots
\]

A divided difference of order one \( f(x_k, x_m) \) of the function \( f(x) \) is defined for an arbitrary pair of points \( x_k, x_m \) (\( x_k \) and \( x_m \) do not have to be neighbors, and we allow
Algebraic Interpolation

$x_k \geq x_m$) through the previously introduced divided differences of order zero:

$$f(x_k, x_m) = \frac{f(x_m) - f(x_k)}{x_m - x_k}.$$  

In general, a divided difference of order $n$: $f(x_0, x_1, \ldots, x_n)$ for the function $f(x)$ is defined through the preceding divided differences of order $n - 1$ as follows:

$$f(x_0, x_1, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n) - f(x_0, x_1, \ldots, x_{n-1})}{x_n - x_0}.$$  

(2.2)

Note that all the points $x_0, x_1, \ldots, x_n$ in formula (2.2) have to be distinct, but they do not have to be arranged in any particular way, say, from the smallest to the largest value of $x_j$ or vice versa.

Having defined the Newton divided differences\(^1\) according to (2.2), we can now represent the interpolating polynomial $P_n(x, f, x_0, x_1, \ldots, x_n)$ in the following Newton form:

$$P_n(x, f, x_0, x_1, \ldots, x_n) = f(x_0) + (x - x_0)f(x_0, x_1) + \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1})f(x_0, x_1, \ldots, x_n).$$  

(2.3)

Formula (2.3) itself will be proven later. In the meantime, we will rather establish several useful corollaries that it brings along.

**COROLLARY 2.1**

The following equality holds:

$$P_n(x, f, x_0, x_1, \ldots, x_n) = P_{n-1}(x, f, x_0, x_1, \ldots, x_{n-1}) + (x - x_0)(x - x_1)\ldots(x - x_{n-1})f(x_0, x_1, \ldots, x_n).$$  

(2.4)

**PROOF** Immediately follows from formula (2.3).

**COROLLARY 2.2**

The divided difference $f(x_0, x_1, \ldots, x_n)$ of order $n$ is equal to the coefficient $c_n$ in front of the term $x^n$ in the interpolating polynomial

$$P_n(x, f, x_0, x_1, \ldots, x_n) = c_n x^n + c_{n-1} x^{n-1} + \ldots + c_0.$$  

In other words, the following equality holds:

$$f(x_0, x_1, \ldots, x_n) = c_n.$$  

(2.5)

**PROOF** It is clear that the monomial $x^n$ on the right-hand side of expression (2.3) is multiplied by the coefficient $f(x_0, x_1, \ldots, x_n)$.

---

\(^1\)Also referred to in the literature as the difference quotients.
COROLLARY 2.3

The divided difference \( f(x_0, x_1, \ldots, x_n) \) may be equal to zero if and only if the quantities \( f(x_0), f(x_1), \ldots, f(x_n) \) are nodal values of some polynomial \( Q_m(x) \) of degree \( m \) strictly less than \( n \): \( m < n \).

PROOF

If \( f(x_0, x_1, \ldots, x_n) = 0 \), then formula (2.3) implies that the degree of the interpolating polynomial \( P_n(x, f, x_0, x_1, \ldots, x_n) \) is less than \( n \), because according to equality (2.5) the coefficient \( c_n \) in front of \( x^n \) is equal to zero. As the nodal values of this interpolating polynomial are equal to \( f(x_j), j = 0, 1, \ldots, n \), we can simply set \( Q_m(x) = P_n(x) \). Conversely, as the interpolating polynomial of degree no greater than \( n \) is unique (Theorem 2.1), the polynomial \( Q_m(x) \) with nodal values \( f(x_0), f(x_1), \ldots, f(x_n) \) must coincide with the interpolating polynomial \( P_n(x, f, x_0, x_1, \ldots, x_n) = c_n x^n + c_{n-1} x^{n-1} + \ldots + c_0 \). As \( m < n \), equality \( Q_m(x) = P_n(x) \) implies that \( c_n = 0 \). Then, according to formula (2.5), \( f(x_0, x_1, \ldots, x_n) = 0 \).

COROLLARY 2.4

The divided difference \( f(x_0, x_1, \ldots, x_n) \) remains unchanged under any arbitrary permutation of its arguments \( x_0, x_1, \ldots, x_n \).

PROOF

Due to its uniqueness, the interpolating polynomial \( P_n(x) \) will not be affected by the order of the interpolation nodes. Let \( x'_0, x'_1, \ldots, x'_n \) be a permutation of \( x_0, x_1, \ldots, x_n \); then, \( \forall x: P_n(x, f, x_0, x_1, \ldots, x_n) = P_n(x, f, x'_0, x'_1, \ldots, x'_n) \). Consequently, along with formula (2.3) one can write

\[
P_n(x, f, x_0, x_1, \ldots, x_n) = f(x'_0) + (x - x'_0)f(x'_0, x'_1) + \ldots + (x - x'_0)(x - x'_1) \ldots (x - x'_{n-1})f(x'_0, x'_1, \ldots, x'_n).
\]

According to Corollary 2.2, one can therefore conclude that

\[
f(x'_0, x'_1, \ldots, x'_n) = c_n.
\]

By comparing formulae (2.5) and (2.6), one can see that \( f(x_0, x_1, \ldots, x_n) = f(x'_0, x'_1, \ldots, x'_n) \).

COROLLARY 2.5

The following equality holds:

\[
f(x_0, x_1, \ldots, x_n) = \frac{f(x_n) - P_{n-1}(x_n, f, x_0, x_1, \ldots, x_{n-1})}{(x_n - x_0)(x_n - x_1) \ldots (x_n - x_{n-1})}.
\]

PROOF

Let us set \( x = x_n \) in equality (2.4); then its left-hand side becomes equal to \( f(x_n) \), and formula (2.7) follows.
THEOREM 2.2
The interpolating polynomial \( P_n(x_0, x_1, \ldots, x_n) \) can be represented in the Newton form, i.e., equality (2.3) does hold.

PROOF We will use induction with respect to \( n \). For \( n = 0 \) (and \( n = 1 \)) formula (2.3) obviously holds. Assume now that it has already been justified for \( n = 1, 2, \ldots, k \), and let us show that it will also hold for \( n = k + 1 \). In other words, let us prove the following equality:

\[
P_{k+1}(x, f, x_0, x_1, \ldots, x_k, x_{k+1}) = P_k(x, f, x_0, x_1, \ldots, x_k)
+ f(x_0, x_1, \ldots, x_k, x_{k+1})(x-x_0)(x-x_1)\ldots(x-x_k).
\]  

(2.8)

Notice that due to the assumption of the induction formula (2.3) is valid for \( n \leq k \). Consequently, the proofs of Corollaries 2.1 through 2.5 that we have carried out on the basis of formula (2.3) will also remain valid for \( n \leq k \).

To prove equality (2.8), we will first demonstrate that the polynomial \( P_{k+1}(x, f, x_0, x_1, \ldots, x_k, x_{k+1}) \) can be represented in the form:

\[
P_{k+1}(x, f, x_0, x_1, \ldots, x_k, x_{k+1}) = P_k(x, f, x_0, x_1, \ldots, x_k)
+ \frac{f(x_{k+1}) - P_k(x_{k+1}, f, x_0, x_1, \ldots, x_k)}{(x_{k+1} - x_0)(x_{k+1} - x_1)\ldots(x_{k+1} - x_k)}(x-x_0)(x-x_1)\ldots(x-x_k).
\]  

(2.9)

Indeed, it is clear that on the right-hand side of formula (2.9) we have a polynomial of degree no greater than \( k + 1 \) that is equal to \( f(x_j) \) at all nodes \( x_j, \ j = 0, 1, \ldots, k + 1 \). Therefore, the expression on the the right-hand side of (2.9) is the interpolating polynomial

\[
P_{k+1}(x, f, x_0, x_1, \ldots, x_k, x_{k+1}).
\]

Next, by comparing formulae (2.8) and (2.9) we see that in order to justify (2.8) we need to establish the equality:

\[
f(x_0, x_1, \ldots, x_k, x_{k+1}) = \frac{f(x_{k+1}) - P_k(x_{k+1}, f, x_0, x_1, \ldots, x_k)}{(x_{k+1} - x_0)(x_{k+1} - x_1)\ldots(x_{k+1} - x_k)}.
\]  

(2.10)

Using the same argument as in the proof of Corollary 2.4, and also employing Corollary 2.1, we can write:

\[
P_k(x, f, x_0, x_1, \ldots, x_k) = P_k(x, f, x_1, x_2, \ldots, x_k, x_0)
= P_{k-1}(x, f, x_1, x_2, \ldots, x_k)
+ f(x_1, x_2, \ldots, x_k, x_0)(x-x_1)(x-x_2)\ldots(x-x_k).
\]  

(2.11)

Then, by substituting \( x = x_{k+1} \) into (2.11), we can transform the right-hand
side of equality (2.10) into:

\[
\frac{f(x_{k+1}) - P_k(x_{k+1}, f, x_0, x_1, \ldots, x_k)}{(x_{k+1} - x_0)(x_{k+1} - x_1) \cdots (x_{k+1} - x_k)} = \frac{1}{x_{k+1} - x_0} \frac{f(x_{k+1}) - P_{k-1}(x_{k+1}, f, x_1, \ldots, x_k)}{(x_{k+1} - x_1) \cdots (x_{k+1} - x_k)} - \frac{f(x_1, x_2, \ldots, x_k, x_0)}{x_{k+1} - x_0}.
\]

(2.12)

By virtue of Corollary 2.5, the minuend on the right-hand side of equality (2.12) is equal to:

\[
\frac{1}{x_{k+1} - x_0} f(x_1, x_2, \ldots, x_k, x_{k+1}),
\]

whereas in the subtrahend, according to Corollary 2.4, one can change the order of the arguments so that it would coincide with

\[
\frac{f(x_0, x_1, \ldots, x_k)}{x_{k+1} - x_0}.
\]

Consequently, the right-hand side of equality (2.12) is equal to

\[
\frac{f(x_1, x_2, \ldots, x_{k+1}) - f(x_0, x_1, \ldots, x_k)}{x_{k+1} - x_0} \overset{\text{def}}{=} f(x_0, x_1, \ldots, x_{k+1}).
\]

In other words, equality (2.12) coincides with equality (2.10) that we need to establish in order to justify formula (2.8). This completes the proof.

**THEOREM 2.3**

Let \(x_0 < x_1 < \ldots < x_n\); assume also that the function \(f(x)\) is defined on the interval \(x_0 \leq x \leq x_n\), and is \(n\) times differentiable on this interval. Then,

\[
n! f(x_0, x_1, \ldots, x_n) = \frac{d^n f}{dx^n} \bigg|_{x=\xi} \equiv f^{(n)}(\xi),
\]

(2.13)

where \(\xi\) is some point from the interval \([x_0, x_n]\).

**PROOF** Consider an auxiliary function

\[
\varphi(x) \overset{\text{def}}{=} f(x) - P_n(x, f, x_0, x_1, \ldots, x_n)
\]

(2.14)

defined on \([x_0, x_n]\); it obviously has a minimum of \(n+1\) zeros on this interval — at the nodes \(x_0, x_1, \ldots, x_n\). Then, according to the Rolle (mean value) theorem, its first derivative vanishes at least at one point in-between every two neighboring zeros of \(\varphi(x)\). Therefore, the function \(\varphi'(x)\) will have a minimum of \(n\) zeros on the interval \([x_0, x_n]\). Similarly, the function \(\varphi''(x)\) vanishes at least at one point in-between every two neighboring zeros of \(\varphi'(x)\), and will therefore have a minimum of \(n-1\) zeros on \([x_0, x_n]\).
By continuing this line of argument, we conclude that the $n$-th derivative $q^{(n)}(x)$ will have at least one zero on the interval $[x_0, x_n]$. Let us denote this zero by $\xi$, so that $q^{(n)}(\xi) = 0$. Next, we differentiate identity (2.14) exactly $n$ times and subsequently substitute $x = \xi$, which yields:

$$0 = q^{(n)}(\xi) = f^{(n)}(\xi) - \frac{d^n}{dx^n} P_n(x, f, x_0, x_1, \ldots, x_n) \bigg|_{x=\xi}. \quad (2.15)$$

On the other hand, according to Corollary 2.2, the divided difference $f(x_0, x_1, \ldots, x_n)$ is equal to the leading coefficient of the interpolating polynomial $P_n$, i.e., $P_n(x, f, x_0, x_1, \ldots, x_n) = f(x_0, x_1, \ldots, x_n)x^n + c_{n-1}x^{n-1} + \ldots + c_0$. Consequently, $\frac{d^n}{dx^n} P_n(x, f, x_0, x_1, \ldots, x_n) = n! f(x_0, x_1, \ldots, x_n)$, and therefore, equality (2.15) implies (2.13).

**THEOREM 2.4**

The values $f(x_0)$, $f(x_1)$, ..., $f(x_n)$ of the function $f(x)$ are expressed through the divided differences $f(x_0)$, $f(x_0, x_1)$, ..., $f(x_0, x_1, \ldots, x_n)$ by the formulae:

$$f(x_j) = f(x_0) + (x_j - x_0)f(x_0, x_1) + (x_j - x_0)(x_j - x_1)f(x_0, x_1, x_2) + (x_j - x_0)(x_j - x_1)\ldots(x_j - x_{n-1})f(x_0, x_1, \ldots, x_n), \quad j = 0, 1, \ldots, n,$$

i.e., by linear combinations of the type:

$$f(x_j) = a_{j0}f(x_0) + a_{j1}f(x_0, x_1) + \ldots + a_{jn}f(x_0, x_1, \ldots, x_n), \quad j = 0, 1, \ldots, n. \quad (2.16)$$

**PROOF** The result follows immediately from formula (2.3) and equalities $f(x_j) = P(x, f, x_0, x_1, \ldots, x_n) \bigg|_{x=x_j}$ for $j = 0, 1, \ldots, n$. 

**2.1.3 Comparison of the Lagrange and Newton Forms**

To evaluate the function $f(x)$ at a point $x$ that is not one of the interpolation nodes, one can approximately set: $f(x) \approx P_n(x, f, x_0, x_1, \ldots, x_n)$.

Assume that the polynomial $P_n(x, f, x_0, x_1, \ldots, x_n)$ has already been built, but in order to try and improve the accuracy we incorporate an additional interpolation node $x_{n+1}$ and the corresponding function value $f(x_{n+1})$. Then, to construct the interpolating polynomial $P_{n+1}(x, f, x_0, x_1, \ldots, x_{n+1})$ using the Lagrange formula (2.1) one basically needs to start from the scratch. At the same time, to use the Newton formula (2.3), see also Corollary 2.1:

$$P_{n+1}(x, f, x_0, x_1, \ldots, x_{n+1}) = P_n(x, f, x_0, x_1, \ldots, x_n) + (x - x_0)(x - x_1)\ldots(x - x_n)f(x_0, x_1, \ldots, x_{n+1})$$

one only needs to obtain the correction

$$(x - x_0)(x - x_1)\ldots(x - x_n)f(x_0, x_1, \ldots, x_{n+1}).$$

Moreover, one will immediately be able to see how large this correction is.
2.1.4 Conditioning of the Interpolating Polynomial

Let all the interpolation nodes \(x_0, x_1, \ldots, x_n\) belong to some interval \(a \leq x \leq b\). Let also the values \(f(x_0), f(x_1), \ldots, f(x_n)\) of the function \(f(x)\) at these nodes be given. Hereafter, we will be using a shortened notation \(P_n(x, f)\) for the interpolating polynomial \(P_n(x) = P_n(x, f, x_0, x_1, \ldots, x_n)\).

Let us now perturb the values \(f(x_j)\) by some quantities \(\delta f(x_j), j = 0, 1, \ldots, n\). Then, the interpolating polynomial \(P_n(x, f)\) will change and become \(P_n(x, f + \delta f)\). One can clearly see from the Lagrange formula (2.1) that \(P_n(x, f + \delta f) = P_n(x, f) + P_n(x, \delta f)\). Therefore, the corresponding perturbation of the interpolating polynomial, i.e., its response to \(\delta f\), will be \(P_n(x, \delta f)\). For a given fixed set of \(x_0, x_1, \ldots, x_n\), this perturbation depends only on \(\delta f\) but not on \(f\) itself. As such, one can introduce the minimum number \(L_n\) such that the following inequality would hold for any \(\delta f\):

\[
\max_{a \leq x \leq b} |P_n(x, \delta f)| \leq L_n \max_j |\delta f(x_j)|. \tag{2.17}
\]

The numbers \(L_n = L_n(x_0, x_1, \ldots, x_n, a, b)\) are called the Lebesgue constants.\(^2\) They provide a natural measure for the sensitivity of the interpolating polynomial to the perturbations \(\delta f(x_j)\) of the interpolated function \(f(x)\) at the nodes \(x_j\). The Lebesgue constants are known to grow as \(n\) increases. Their specific behavior strongly depends on how the interpolation nodes \(x_j\), \(j = 0, 1, \ldots, n\), are located on the interval \([a, b]\).

If, for example, \(n = 1\), \(x_0 = a\), \(x_1 = b\), then \(L_1 = 1\). If, however, \(x_0 \neq a\) and/or \(x_1 \neq b\), then \(L_1 \geq \frac{b-a}{2|x_1-x_0|}\), i.e., if \(x_1\) and \(x_0\) are sufficiently close to one another, then the interpolation may appear arbitrarily sensitive to the perturbations of \(f(x)\). The reader can easily verify the foregoing statements regarding \(L_1\).

In the case of equally spaced interpolation nodes:

\[
x_j = a + j \cdot h, \quad j = 0, 1, \ldots, n, \quad h = \frac{b-a}{n},
\]

one can show that

\[
2^n > L_n > 2^{n-2} \frac{1}{\sqrt{n}} \cdot \frac{1}{n-1/2}. \tag{2.18}
\]

In other words, the sensitivity of the interpolant to any errors committed when specifying the values of \(f(x_j)\) will grow rapidly (exponentially) as \(n\) increases. Note that in practice it is impossible to specify the values of \(f(x_j)\) without any error, no matter how these values are actually obtained, i.e., whether they are measured (with inevitable experimental inaccuracies) or computed (subject to rounding errors).

For a rigorous proof of inequalities (2.18) we refer the reader to the approximation theory literature, in particular, the monographs and texts cited in Section 3.2.7 of Chapter 3. However, an elementary treatment can also be given, and one can easily provide a qualitative argument of why the Lebesgue constants for equidistant nodes

\(^2\)Note that the Lebesgue constant \(L_n\) corresponds to interpolation on \(n+1\) nodes: \(x_0, \ldots, x_n\).
Algebraic Interpolation

Indeed grow exponentially as the grid dimension $n$ increases. From the Lagrange formula (2.1) and definition (2.17) it is clear that

$$L_n = O \left( \max_{a \leq x \leq b} \sum_{k=0}^{n} |l_k(x)| \right)$$

(later, see Section 3.2.7 of Chapter 3, we will prove an even more accurate statement). Take $k \approx n/2$ and $x$ very close to one of the edges $a$ or $b$, say, $x - a = \eta \ll h$. Then,

$$|l_k(x)| = \left| \frac{(x-x_0)(x-x_1)\ldots(x-x_k-1)(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)} \right|$$

$$\approx \frac{\eta \cdot h^{2k-1} \cdot (2k)! / k}{(h^k!)^2} = \frac{\eta \cdot h \cdot (2k)!}{k(k!)^2}$$

$$= \eta \cdot h \cdot \frac{(2 \cdot 4 \cdot 6 \ldots \cdot 2k)((1 \cdot 3 \cdot 5 \ldots \cdot (2k - 1))}{k(k!)^2}$$

$$\approx \eta \cdot h \cdot \frac{(2 \cdot 4 \cdot 6 \ldots \cdot 2k)^2}{(k!)^2} = \eta \cdot h \cdot \frac{2^{2k}(k!)^2}{(k!)^2} \approx \eta \cdot h \cdot 2^n.$$ 

The foregoing estimate for $|l_k(x)|$, along with the previous formula (2.19), do imply the exponential growth of the Lebesgue constants on uniform (equally spaced) interpolation grids. Let now $a = -1, b = 1$, and let the interpolation nodes on $[a, b]$ be rather given by the formula:

$$x_j = -\cos \frac{(2j+1)\pi}{2(n+1)} \quad j = 0, 1, \ldots, n. \quad (2.20)$$

It is possible to show that placing the nodes according to (2.20) guarantees a much better estimate for the Lebesgue constants (again, see Section 3.2.7):

$$L_n \leq \frac{2}{\pi} \ln(n+1) + 1. \quad (2.21)$$

We therefore conclude that in contradistinction to the previous case (2.18), the Lebesgue constants may, in fact, grow slowly rather than rapidly, as they do on the non-equally spaced nodes (2.20). As such, even the high-degree interpolating polynomials in this case will not be overly sensitive to perturbations of the input data. Interpolation nodes (2.20) are known as the Chebyshev nodes. They will be discussed in detail in Chapter 3.

2.1.5 On Poor Convergence of Interpolation with Equidistant Nodes

One should not think that for any continuous function $f(x), x \in [a, b]$, the algebraic interpolating polynomials $P_n(x, f)$ built on the equidistant nodes $x_j = a + j \cdot h, x_0 = a, x_n = b$, will converge to $f(x)$ as $n$ increases, i.e., will deviate from $f(x)$ less and less for larger $n$'s. For example, as has been shown by Bernstein, the sequence of
interpolating polynomials obtained for the function \( f(x) = |x| \) on equally spaced nodes diverges at every point of the interval \([a, b] = [-1, 1]\) except at \((-1, 0, 1)\).

The next example is attributed to Runge. Consider the function \( f(x) = \frac{1}{x + 1/4} \) on the same interval \([a, b] = [-1, 1]\); not only is this function continuous itself, but also has continuous derivatives of all orders. It is, however, possible to show that for the sequence of interpolating polynomials with equally spaced nodes the maximum difference \( \max_{-1 \leq x \leq 1} |f(x) - P_n(x, f)| \) will not approach zero as \( n \) increases.

Moreover, by working on Exercise 4 below, one will be able to see that the areas of no convergence for this function are located next to the endpoints of the interval \([-1, 1]\). For larger intervals the situation may even deteriorate and the sequence \( P_n(x, f) \) may diverge. In other words, the quantity \( \max_{-1 \leq x \leq 1} |f(x) - P_n(x, f)| \) may become arbitrarily large for large \( n \)'s (see, e.g., [IK66]).

Altogether, these convergence difficulties can be accounted for by the fact that on the complex plane the function \( f(z) = \frac{1}{z + 1/4} \) is not an entire function of its argument \( z \), and has singularities at \( z = \pm i/2 \).

On the other hand, if, instead of the equidistant nodes, we were to use Chebyshev nodes (2.20) to interpolate either the Bernstein function \( f(x) = |x| \) or the Runge function \( f(x) = \frac{1}{x + 1/4} \), then in both cases the sequence of interpolating polynomials \( P_n(x, f) \) would have converged to \( f(x) \) uniformly as \( n \) increases (see Exercise 5).

**Exercises**

1. Evaluate \( f(1.14) \) by means of linear, quadratic, and cubic interpolation using the following table of values:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( 1.08 )</th>
<th>( 1.13 )</th>
<th>( 1.20 )</th>
<th>( 1.27 )</th>
<th>( 1.31 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x) )</td>
<td>( 1.302 )</td>
<td>( 1.386 )</td>
<td>( 1.509 )</td>
<td>( 1.217 )</td>
<td>( 1.284 )</td>
</tr>
</tbody>
</table>

Implement the interpolating polynomials in both the Lagrange and Newton form.

2. Let \( x_j = j \cdot h \), \( j = 0, \pm 1, \pm 2, \ldots \), be equidistant nodes with the spacing \( h \). Verify that the following equality holds:

\[
f(x_{k+1} - x_k, x_{k+1}) = \frac{f(x_{k+1}) - 2f(x_k) + f(x_k)}{2h^2}.
\]

3. Let \( a = x_0 \), \( a < x_1 < b \), \( x_2 = b \). Find the value of the Lebesgue constant \( L_2 \) when \( x_1 \) is the midpoint of \([a, b] \): \( x_1 = (a + b)/2 \). Show that if, conversely, \( x_1 \to a \) or \( x_1 \to b \), then the Lebesgue constant \( L_2 = L_2(x_0, x_1, x_2, a, b) \) grows with no bound.

4. Plot the graphs of \( f(x) = \frac{1}{x + 1/4} \) and \( P_n(x, f) \) from Section 2.1.5 (Runge example) on the computer and thus corroborate experimentally that there is no convergence of the interpolating polynomial on equally spaced nodes when \( n \) increases.

5. Use Chebyshev nodes (2.20) to interpolate \( f(x) = |x| \) and \( f(x) = \frac{1}{x + 1/4} \) on the interval \([-1, 1]\), plot the graphs of each \( f(x) \) and the corresponding \( P_n(x, f) \) for several increasing values of \( n = 10, 20, 40, \) and \( 80 \), evaluate numerically the error \( \max_{-1 \leq x \leq 1} |f(x) - P_n(x, f)| \), and make sure that it decreases as \( n \) increases.
2.2 Classical Piece-Wise Polynomial Interpolation

High sensitivity of algebraic interpolating polynomials to the errors in the tabulated values of \( f(x) \), as well as the “iffy” convergence of the sequence \( P_n(x, f) \) on uniform grids, prompt the use of piece-wise polynomial interpolation.

2.2.1 Definition of Piece-Wise Polynomial Interpolation

Let the function \( f(x), x \in [a, b] \), be defined by the table \( \{f(x_0), f(x_1), \ldots, f(x_n)\} \) of its numerical values at the nodes \( \{a = x_0 < x_1 < x_2 < \ldots < x_n = b\} \). To reconstruct this function in-between the nodes \( x_0, x_1, \ldots, x_n \), one can use an auxiliary function that would coincide with a polynomial of a given low degree (say, the first, the second, the third, etc.) between every two neighboring nodes of the interpolation grid. This approach is known as piece-wise polynomial interpolation; in particular, it may be piece-wise linear, piece-wise quadratic, piece-wise cubic, etc.

In the case of piece-wise linear interpolation on the interval \( x_k \leq x \leq x_{k+1} \), one uses the linear interpolating polynomial \( P_1(x, f, x_k, x_{k+1}) \) to approximate the function \( f(x) \). In the case of piece-wise quadratic interpolation on the interval \( x_k \leq x \leq x_{k+1} \), one can use either of the two polynomials: \( P_2(x, f, x_k, x_{k+1}, x_{k+2}) \) or \( P_2(x, f, x_{k-1}, x_k, x_{k+1}) \).

Piece-wise polynomial interpolation of an arbitrary degree \( s \) is obtained similarly. There is always some flexibility in constructing the interpolant, and to approximate the function \( f(x) \) on the interval \( x_k \leq x \leq x_{k+1} \) one can basically use any of the polynomials \( P_s(x, f, x_k, x_{k+1}, \ldots, x_{k+s}) \), where \( j \) is one of the integers \( 0, 1, \ldots, s-1 \). It is, however, desirable that the smaller interval \( [x_k, x_{k+1}] \) be located maximally close to the middle of the larger interval \( [x_{k-j}, x_{k+j}] \) (see Section 2.1.4). For equidistant nodes, the latter requirement translates into choosing \( j \) maximally close to \( s/2 \). In general, once the strategy for selecting \( j \) has been adopted, one can reconstruct \( f(x) \) on \( [a, b] \) in the form of a piece-wise polynomial of degree \( s \). It will be composed of the individual interpolating polynomials that correspond to different intervals \( [x_k, x_{k+1}], k = 0, 1, \ldots, n-1 \). For simplicity, we will hereafter denote the piece-wise polynomial as follows:

\[
P_s(x, f, x_{k-j}, x_k, x_{k+1}, \ldots, x_{k+j}) = P_s(x, f_{k_j}).
\]

2.2.2 Formula for the Interpolation Error

Let us estimate the error

\[
R_s(x) \overset{\text{def}}{=} f(x) - P_s(x, f_{k_j}), \quad x_k \leq x \leq x_{k+1}, \tag{2.22}
\]

that arises when the function \( f(x) \) is approximately replaced by the polynomial \( P_s(x, f_{k_j}) \). To do so, we will need to exploit the following general theorem:
THEOREM 2.5

Let the function $f = f(t)$ be defined on the interval $\alpha \leq t \leq \beta$, and let it have there a continuous derivative of order $s+1$. Let $t_0, t_1, \ldots, t_s$ be an arbitrary set of distinct points that all belong to $[\alpha, \beta]$, and let $f(t_0), f(t_1), \ldots, f(t_s)$ be the values of the function $f(t)$ at these points. Finally, let $p_s(t) \equiv p_s(t, f, t_0, t_1, \ldots, t_s)$ be the algebraic interpolating polynomial of degree no greater than $s$ built for these given points and function values. Then, the interpolation error $R_s(t) = f(t) - P_s(t)$ can be represented on $[\alpha, \beta]$ as follows:

$$R_s(t) = \frac{f^{(s+1)}(\xi)}{(s+1)!}(t - t_0)(t - t_1)\ldots(t - t_s), \quad (2.23)$$

where $\xi = \xi(t)$ is some point from the interval $(\alpha, \beta)$. 

PROOF  We first notice that formula (2.23) does hold for all nodes $t_j, j = 0, 1, \ldots, s$, themselves, because on one hand $\forall t_j: f(t_j) - P_s(t_j) = 0$, and on the other hand, $R_s(t_j) = 0$, where $R_s(t)$ is defined by formula (2.23). Let us now take an arbitrary $\bar{t} \in [\alpha, \beta]$ that does not coincide with any of $t_0, t_1, \ldots, t_s$. To prove formula (2.23) for $t = \bar{t}$, we introduce an auxiliary function:

$$q(t) = f(t) - P_s(t) - k(t - t_0)(t - t_1)\ldots(t - t_s) \quad (2.24)$$

and choose the parameter $k$ so that $q(\bar{t}) = 0$, which obviously implies

$$k = \frac{f(\bar{t}) - P_s(\bar{t})}{(\bar{t} - t_0)(\bar{t} - t_1)\ldots(\bar{t} - t_s)}. \quad (2.25)$$

The numerator in formula (2.25) coincides with the value of the error $R_s(\bar{t})$, therefore, this formula yields:

$$R_s(\bar{t}) = k(\bar{t} - t_0)(\bar{t} - t_1)\ldots(\bar{t} - t_s). \quad (2.26)$$

The auxiliary function $q$ of (2.24) clearly has a minimum of $s + 2$ zeros on the interval $[\alpha, \beta]$ — at the points $\bar{t}, t_0, t_1, \ldots, t_s$. Then, its first derivative $q'(t)$ will have a minimum of $s + 1$ zeros on the (open) interval $(\alpha, \beta)$, because according to the Rolle (mean value) theorem, the derivative $q'(t)$ has to vanish at least once in-between every two neighboring points where $q(t)$ itself vanishes. Similarly, $q''(t)$ will have at least $s$ zeros on $(\alpha, \beta)$, $q^{(3)}(t)$ — at least $s - 1$ zeros, etc., so that finally the derivative $q^{(s+1)}(t)$ will have to have a minimum of one zero on the interval $(\alpha, \beta)$. Let us denote this zero by $\xi \in (\alpha, \beta)$, so that $q^{(s+1)}(\xi) = 0$.

Next, we note that

$$\frac{d^{s+1}}{dt^{s+1}} q^{(s+1)} = (s + 1)!,$$

and that $(t - t_0)(t - t_1)\ldots(t - t_s) = t^{s+1} + Q_s(t)$, where $Q_s(t)$ is a polynomial of degree no greater than $s$. We also note that

$$\frac{d^{s+1}}{dt^{s+1}} P_s(t) \equiv \frac{d^{s+1}}{dt^{s+1}} Q_s(t) \equiv 0.$$
Using the previous two expressions, we differentiate the function \( \varphi(t) \) defined by formula (2.24) \( s+1 \) times and obtain:

\[
\varphi^{(s+1)}(t) = f^{(s+1)}(t) - k(s+1)!
\]

Substituting \( t = \xi \) into the last equality, and recalling that \( \varphi^{(s+1)}(\xi) = 0 \), we arrive at the following expression for \( k \):

\[
k = \frac{f^{(s+1)}(\xi)}{(s+1)!}.
\] (2.27)

Finally, by substituting \( k \) of (2.27) into equality (2.26) we obtain a formula for \( R_s(\bar{t}) \) that would actually coincide with formula (2.23) because \( \bar{t} \in [\alpha, \beta] \) has been chosen arbitrarily.

**THEOREM 2.6**

Under the assumptions of the previous theorem, the following estimate holds:

\[
\max_{\alpha \leq t \leq \beta} |R_s(t)| \leq \frac{1}{(s+1)!} \max_{\alpha \leq t \leq \beta} |f^{(s+1)}(t)|(|\beta - \alpha|)^{s+1}.
\] (2.28)

**PROOF** We first note that \( \forall t \in [\alpha, \beta] \) the absolute value of each expression \( t - t_0, t - t_1, \ldots, t - t_s \) will not exceed \( \beta - \alpha \). Then, we use formula (2.23):

\[
|R_s(t)| = \frac{1}{(s+1)!} |f^{(s+1)}(\xi)(t - t_0)(t - t_1) \cdots (t - t_s)|
\]

\[
\leq \frac{1}{(s+1)!} \max_{\alpha \leq t \leq \beta} |f^{(s+1)}(t)|(|\beta - \alpha|)^{s+1}.
\] (2.29)

As \( t \in [\alpha, \beta] \) on the left-hand side of formula (2.29) is arbitrary, the required estimate (2.28) follows.

Let us emphasize that we have proven inequality (2.28) for an arbitrary distribution of the (distinct) interpolation nodes \( t_0, t_1, \ldots, t_s \) on the interval \( [\alpha, \beta] \). For a given fixed distribution of nodes, estimate (2.28) can often be improved. For example, consider a piece-wise linear interpolation and assume that the nodes \( t_0 \) and \( t_1 \) coincide with the endpoints \( \alpha \) and \( \beta \), respectively, of the interval \( \alpha \leq t \leq \beta \). Then,

\[
|R_1(t)| = \frac{f''(\xi)}{(s+1)!} (t - \alpha)(t - \beta)
\]

\[
\leq \frac{1}{2} \max_{\alpha \leq \xi \leq \beta} |f''(\xi)| \max_{\alpha \leq \xi \leq \beta} (t - \alpha)(t - \beta) = \frac{1}{8} \max_{\alpha \leq \xi \leq \beta} |f''(\xi)(\beta - \alpha)^2|
\]

which yields

\[
\max_{\alpha \leq \xi \leq \beta} |R_1(t)| \leq \frac{1}{8} \max_{\alpha \leq \xi \leq \beta} |f''(\xi)(\beta - \alpha)^2|,
\] (2.30)
whereas estimate (2.28) for $s = 1$ transforms into

$$
\max_{a \leq t \leq b} |R_1(t)| \leq \frac{1}{2} \max_{a \leq t \leq b} |f''(t)|(\beta - \alpha)^2.
$$

We will now use Theorems 2.5 and 2.6 to estimate the error (2.22) of piece-wise polynomial interpolation of the function $f(x)$ on the interval $x_k \leq x \leq x_{k+1}$. First, let

$$
\alpha = x_k, \quad \beta = x_k + s, \quad t_0 = x_k - j, \quad t_1 = x_k - j + 1, \ldots, \quad t_s = x_k + s.
$$

Then, it is clear that

$$
\max_{x_k \leq x \leq x_{k+1}} |R_s(x, f_{kj})| \leq \max_{a \leq x \leq b} |R_s(x, f_{kj})|,
$$

and according to (2.28) we obtain

$$
\max_{x_k \leq x \leq x_{k+1}} |R_s(x, f_{kj})| \leq \frac{1}{(s + 1)!} \max_{s_k - j \leq x \leq s_{k-1} + 1} |f^{(s+1)}(x)|(x_k - x_{k-1})^{s+1}. \quad (2.31)
$$

If the quantity $|f^{(s+1)}(x)|$ undergoes strong variations on the interval $[a, b]$, then, in order for the estimate (2.31) to guarantee some prescribed accuracy, it will be advantageous to have the grid size (distance between the neighboring nodes) and the value of $x_k - x_{k-1}$ smaller in those parts of $[a, b]$ where $|f^{(s+1)}(x)|$ is larger.

In the case of equidistant nodes $x_0, x_1, \ldots, x_n$, estimate (2.31) implies

$$
\max_{x_k \leq x \leq x_{k+1}} |R_s(x, f_{kj})| \leq \frac{1}{(s + 1)!} \max_{x_j - s \leq x \leq x_{j+s} + 1} |f^{(s+1)}(x)|h^{s+1}, \quad (2.32)
$$

where $h = (b - a)/n = x_{k+1} - x_k$ is the size of the interpolation grid. Inequality (2.32) can obviously be recast as

$$
\max_{x_k \leq x \leq x_{k+1}} |R_s(x, f_{kj})| \leq \text{const.} \max_{x_j - s \leq x \leq x_{j+s}} |f^{(s+1)}(x)|h^{s+1}, \quad (2.33)
$$

where the key consideration is that the constant on the right-hand side of (2.33) does not depend on the grid size $h$.

To conclude this section, let us specifically mention the case of piece-wise linear interpolation: $s = 1$, $\alpha = x_k$, and $\beta = x_{k+1}$. Then, according to estimate (2.30):

$$
\max_{x_k \leq x \leq x_{k+1}} |R_1(x)| \leq \frac{1}{8} \max_{x_k \leq x \leq x_{k+1}} |f''(x)|(x_{k+1} - x_k)^2 = \frac{h^2}{8} \max_{x_k \leq x \leq x_{k+1}} |f''(x)|. \quad (2.34)
$$

### 2.2.3 Approximation of Derivatives for a Grid Function

**THEOREM 2.7**

Let the function $f = f(x)$ be defined on the interval $[\alpha, \beta]$, and let it have a continuous derivative of order $s + 1$ on this interval. Let $x_k, x_k - j + 1, \ldots, x_k - j + s$...
be a set of interpolation nodes, such that \( \alpha = x_{k-j} < x_{k-j+1} < \ldots < x_{k-j+s} = \beta \). Then, to approximately evaluate the derivatives

\[
\frac{d^q f(x)}{dx^q}, \quad q = 1, 2, \ldots, s,
\]

of the function \( f(x) \) on the interval \( x_k \leq x \leq x_{k+1} \), one can employ the interpolating polynomial \( P_s(x, f_{kj}) \) and set

\[
\frac{d^q f(x)}{dx^q} \approx \frac{d^q}{dx^q} P_s(x, f_{kj}), \quad x_k \leq x \leq x_{k+1}.
\] (2.35)

In so doing, the approximation error will satisfy the estimate:

\[
\max_{x_k \leq x \leq x_{k+1}} \left| \frac{d^q f(x)}{dx^q} - \frac{d^q}{dx^q} P_s(x, f_{kj}) \right| \leq \frac{1}{(s - q + 1)!} \max_{x_k - j \leq x \leq x_{k-j+s}} |f^{(s+1)}(x)|(x_{k-j+s} - x_{k-j})^{s-q+1}.
\] (2.36)

**PROOF** Consider an auxiliary function \( \varphi(x) \equiv f(x) - P_s(x, f_{kj}) \); it obviously vanishes at all \( s+1 \) interpolation nodes \( x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s} \). Therefore, its first derivative \( \varphi'(x) \) will have a minimum of \( s \) zeros on the interval \( x_{k-j} \leq x \leq x_{k-j+s} \), because according to the Rolle (mean value) theorem, there is a zero of the function \( \varphi'(x) \) in-between any two neighboring zeros of \( \varphi(x) \).

Similarly, the function \( \frac{d^q \varphi(x)}{dx^q} \) will have at least \( s-q+1 \) zeros on the interval \( x_{k-j} \leq x \leq x_{k-j+s} \). This implies that the derivative \( \frac{d^q \varphi(x)}{dx^q} \) and the polynomial \( \frac{d^q}{dx^q} P_s(x, f_{kj}) \) of degree no greater than \( s-q \) coincide at \( s-q+1 \) distinct points.

In other words, the polynomial \( P_s^{(q)}(x, f_{kj}) \) can be interpreted as an interpolating polynomial of degree no greater than \( s-q \) for the function \( f^{(q)}(x) \) on the interval \( x_{k-j} \leq x \leq x_{k-j+s} \), built on some set of \( s-q+1 \) interpolation nodes.

Moreover, the function \( f^{(q)}(x) \) does have a continuous derivative of order \( s-q+1 \) on \([\alpha, \beta] \):

\[
\frac{d^{s-q+1}}{dx^{s-q+1}} f^{(q)}(x) = \frac{d^{s+1}}{dx^{s+1}} f(x).
\]

Consequently, one can use Theorem 2.6 and, by setting \( \alpha = x_{k-j}, \beta = x_{k-j+s} \), obtain the following estimate [cf. formula (2.28)]:

\[
\max_{x_k - j \leq x \leq x_{k-j+s}} \left| f^{(q)}(x) - P_s^{(q)}(x, f_{kj}) \right| \leq \frac{1}{(s - q + 1)!} \max_{x_k - j \leq x \leq x_{k-j+s}} |f^{(s+1)}(x)|(x_{k-j+s} - x_{k-j})^{s-q+1}.
\]

As \( \alpha = x_{k-j} \leq x_k < x_{k+1} \leq x_{k-j+s} = \beta \), it immediately yields (2.36).
2.2.4 Estimate of the Unavoidable Error and the Choice of Degree for Piece-Wise Polynomial Interpolation

Let the function \( f = f(x) \) be defined on the interval \([0, \pi]\), and let its values be known at the nodes of the uniform grid: \( x_k = k\pi/n \equiv kh, \ k = 0, 1, \ldots, n \). Using only the tabulated values of the function \( f(x_0), f(x_1), \ldots, f(x_n) \), one cannot, even in principle, obtain an exact reconstruction of \( f(x) \) in-between the nodes, because different functions may have identical tables, i.e., may coincide at the nodes \( x_k \), \( k = 0, 1, \ldots, n \), and at the same time be different elsewhere. If, for example, in addition to the table of values nothing is known about the function \( f(x) \) except that it is simply continuous, then one can guarantee no accuracy at all when reconstructing \( f(x) \) at \( x \neq x_k, \ k = 0, 1, \ldots, n \).

Assume now that \( f(x) \) has a bounded derivative of the maximum order \( s + 1 \):

\[
\max_x |f^{(s+1)}(x)| \leq M_s = \text{const.} \quad (2.37)
\]

It is easy to find two different functions from the class characterized by \( M_s = 1 \):

\[
f_1(x) = \frac{\sin nx}{n^{s+1}} \quad \text{and} \quad f_2(x) = -\frac{\sin nx}{n^{s+1}},
\]

that would deviate from one another by the value of order \( h^{s+1} \):

\[
\max_{0 \leq x \leq \pi} |f_1(x) - f_2(x)| = \max_{0 \leq x \leq \pi} 2 \left| \frac{\sin nx}{n^{s+1}} \right| = \frac{2}{\pi^{s+1}} h^{s+1}, \quad (2.38)
\]

and for which the tables would nonetheless fully coincide (both will be trivial):

\[
f_1(x_k) = f_2(x_k) = 0, \quad k = 0, 1, \ldots, n.
\]

We therefore conclude that given the tabulated values of the function \( f(x) \), and only estimate (2.37) in addition to that, one cannot, even in theory, reconstruct the function \( f(x) \) on the interval \( 0 \leq x \leq \pi \) with the accuracy better than \( O(h^{s+1}) \). In other words, the error \( O(h^{s+1}) \) is unavoidable when reconstructing the function \( f(x), 0 \leq x \leq \pi \), using its table of values on a uniform grid with size \( h \).

It is also clear that

\[
\max_{0 \leq x \leq \pi} \left| \frac{d^q f_1(x)}{dx^q} - \frac{d^q f_2(x)}{dx^q} \right| = 2 \frac{1}{n^{s-q+1}} = \frac{2}{\pi^{s-q+1}} h^{s-q+1}, \quad (2.39)
\]

which means that the unavoidable error when reconstructing the derivative \( \frac{d^q f(x)}{dx^q} \) is at least \( O(h^{s-q+1}) \).

By comparing equalities (2.38) and (2.39) with estimates of the error obtained in Sections 2.2.2 and 2.2.3 for the piece-wise polynomial interpolation of the function \( f(x) \) and its derivatives, we conclude that the interpolation error and the unavoidable error have the same asymptotic order (of smallness) with respect to the grid size \( h \). If, under the condition (2.37), one still chooses to use interpolating polynomials of
degree $r < s$, then the interpolation error (for the function itself) will be $O(h^{r+1})$. In other words, there will be an additional loss of the order of accuracy, on top of the uncertainty-based unavoidable error $O(h^{s+1})$ that is due to the specification of $f(x)$ through its discrete table of values.

On the other hand, the use of interpolation of a higher degree $r > s$ cannot increase the order of accuracy beyond the threshold set by the unavoidable error $O(h^{s+1})$, and therefore cannot speed up the decay of the error as $h \to 0$. In this respect, the degree $s$ of piece-wise polynomial interpolation is optimal for the functions that satisfy inequality (2.37).

REMARK 2.1 The considerations of the current section pertain primarily to the asymptotic behavior of the error as $h \to 0$. For a given fixed $h > 0$, interpolation of some degree $r < s$ may, in fact, appear more accurate than the interpolation of degree $s$. Besides, in practice the tabulated values $f(x_k)$, $k = 0, 1, \ldots, n$, may only be specified approximately, rather than exactly, with a finite fixed number of decimal (or binary) digits. In this case, the loss of interpolation accuracy due to rounding is going to increase as $s$ increases, because of the growth of the Lebesgue constants (defined by formula (2.18) of Section 2.1.4). Therefore, the piece-wise polynomial interpolation of high degree (higher than the third) is not used routinely.

REMARK 2.2 Error estimate (2.33) does, in fact, imply uniform convergence of the interpolant $P_s(x, f_k)$ (a piece-wise polynomial) to the interpolated function $f(x)$ with the rate $O(h^{r+1})$ as the grid is refined, i.e., as $h \to 0$. Estimate (2.34), in particular, indicates that piece-wise linear interpolation converges uniformly with the rate $O(h^2)$. Likewise, estimate (2.36) in the case of a uniform grid with size $h$ will imply uniform convergence of the $q$-th derivative of the interpolant $P_s^{(q)}(x, f_k)$ to the $q$-th derivative of the interpolated function $f^{(q)}(x)$ with the rate $O(h^{r-q+1})$ as $h \to 0$.

REMARK 2.3 The notion of unavoidable error as presented in this section (see also Section 1.3.1 and Exercises on page 17) illustrates the concept of Kolmogorov diameters for compact sets of functions. This and related concepts play a fundamental role in the modern theory of approximation; in particular, for the analysis of the so-called best approximations, for the analysis of saturation of numerical methods (see Section 2.2.5), as well as in the theory of $\varepsilon$-entropy and related theory of transmission and processing of information.
2.2.5 Saturation of Piece-Wise Polynomial Interpolation

Let \( f(x) \) be defined on the interval \([a, b]\), and let its table of values \( f(x_k) \) be known for the equally spaced nodes \( x_k = a + kh, k = 0, 1, \ldots, n, h = (b - a)/n \). We have seen in Section 2.2.2 that the error of piece-wise polynomial interpolation of degree \( s \) is of order \( \mathcal{O}(h^{s+1}) \), provided that the polynomial \( P_s(x, f_{k_j}) \) is used to approximate \( f(x) \) on the interval \( x_k \leq x \leq x_{k+1} \), and that the derivative \( f^{(s+1)}(x) \) exists and is bounded. If our only additional knowledge about \( f(x) \) (besides the table) is that it has a bounded derivative of the maximum order \( q + 1 \), where \( q < s \), then the unavoidable error of reconstructing \( f(x) \) from its tabulated values is \( \mathcal{O}(h^{q+1}) \). One can also show that when interpolating \( f'(x) \) by means of \( P_s(x, f_{k_j}) \) the order \( \mathcal{O}(h^{q+1}) \) is actually achieved. On the other hand, if \( f(x) \) has a bounded derivative of order \( q + 1 \), where \( q > s \), then the accuracy of the interpolation by means of piece-wise polynomials \( P_s(x, f_{k_j}) \) still remains of order \( \mathcal{O}(h^{q+1}) \). In other words, the order of interpolation error does not react in any way to the additional smoothness of the function \( f(x) \), beyond the required \( s + 1 \) derivatives. This property of piece-wise polynomial interpolation is known as its saturation (by smoothness).

Exercises

1. What size of the grid \( h \) guarantees that the error of piece-wise linear interpolation for the function \( f(x) = \sin x \) will never exceed \( 10^{-6} \)?

2. What size of the grid \( h \) guarantees that the error of piece-wise quadratic interpolation for the function \( f(x) = \sin x \) will never exceed \( 10^{-6} \)?

3. The values of \( f(x) \) can be measured at any given point \( x \) with the accuracy \( |\delta f| \leq 10^{-4} \). What is the optimal grid size for tabulating \( f(x) \), if the function is to be subsequently reconstructed by means of a piece-wise linear interpolation?

   **Hint.** Choosing \( h \) excessively small may make the interpolation error smaller than the perturbation of the interpolating polynomial due to the perturbations in the data, see Section 2.1.4.

4. The same question as in problem 3, but for piece-wise quadratic interpolation.

5. Consider two approximation formulae for the first derivative \( f'(x) \):

\[
\begin{align*}
\delta f' & \approx \frac{f(x + h) - f(x)}{h}, \quad (2.40) \\
\delta f'' & \approx \frac{f(x + h) - f(x - h)}{2h}, \quad (2.41)
\end{align*}
\]

and let \( |f''(x)| \leq 1 \) and \( |f'''(x)| \leq 1 \).

   a) Find \( h \) such that the error of either formula will not exceed \( 10^{-3} \).

   b) Assume that the function \( f \) itself is only known with the error \( \delta \). What is the best accuracy that one can achieve using formulae (2.40) and (2.41), and how one should properly choose \( h \)?

   c) Show that the asymptotic order of the error with respect to \( \delta \), obtained by formula (2.41) with the optimal \( h \), cannot be improved.
2.3 Smooth Piece-Wise Polynomial Interpolation (Splines)

A classical piece-wise polynomial interpolation of any degree \( s \), e.g., piece-wise linear, piece-wise quadratic, etc., see Section 2.2, yields the interpolant that, generally speaking, is not differentiable even once at the interpolation nodes. There are, however, two alternative types of piece-wise polynomial interpolants — local and nonlocal splines — that do have a given number of continuous derivatives everywhere, including the interpolation nodes.

2.3.1 Local Interpolation of Smoothness \( s \) and Its Properties

Assume that the interpolation nodes \( x_j \) and the function values \( f(x_j) \) are given. Let us then specify a positive integer number \( s \) and also fix another positive integer \( j \): \( 0 \leq j \leq s - 1 \). We will associate the interpolating polynomial \( P_j(x, f_{kj}) \equiv P_j(x, f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s})) \) with every point \( x_k \); this polynomial is built on the nodes \( x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s} \) using the function values \( f(x_{k-j}), f(x_{k-j+1}), \ldots, f(x_{k-j+s}) \). A piece-wise polynomial local spline \( q(x, s) \) that has continuous derivatives up to the order \( s \) is defined individually for each segment \([x_k, x_{k+1}]\) by means of the following equalities:

\[
q(x, s) = Q_{2s+1}(x, k), \quad x \in [x_k, x_{k+1}], \quad k = 0, \pm 1, \ldots, \tag{2.42}
\]

where each \( Q_{2s+1}(x, k) \) is a polynomial of degree no greater than \( 2s + 1 \) that satisfies the relations:

\[
\frac{d^m Q_{2s+1}(x, k)}{dx^m} \bigg|_{x=x_k} = \frac{d^m P_j(x, f_{kj})}{dx^m} \bigg|_{x=x_k}, \quad m = 0, 1, 2, \ldots, s, \tag{2.43}
\]

\[
\frac{d^m Q_{2s+1}(x, k)}{dx^m} \bigg|_{x=x_{k+1}} = \frac{d^m P_j(x, f_{k+1,j})}{dx^m} \bigg|_{x=x_{k+1}}, \quad m = 0, 1, 2, \ldots, s. \tag{2.44}
\]

**THEOREM 2.8**

The polynomial \( Q_{2s+1}(x, k) \) of degree no greater than \( 2s + 1 \) defined by means of equalities (2.43) and (2.44) exists and is unique.

**PROOF** Let \( Q_{2s+1}(x, k) = c_{0,k} + c_{1,k}x + \ldots + c_{2s+1,k}x^{2s+1} \). Then, relations (2.43) and (2.44) together can be considered a system of \( 2s + 2 \) linear algebraic equations with respect to the \( 2s + 2 \) unknown coefficients \( c_{0,k}, c_{1,k}, \ldots, c_{2s+1,k} \). This system becomes homogeneous if we replace all the right-hand sides of equalities (2.43) and (2.44) by zeros. Having this system as homogeneous means that the corresponding polynomial \( Q_{2s+1}(x, k) \) would have a root of multiplicity \( s + 1 \) at \( x = x_k \) and another root of multiplicity \( s + 1 \) at \( x = x_{k+1} \).
In other words, \( Q_{2s+1}(x,k) \) would have a total of \( 2s+2 \) roots counting their multiplicities. This is only possible if \( Q_{2s+1}(x,k) \equiv 0 \), because \( Q_{2s+1}(x,k) \) is a polynomial of degree no greater than \( 2s+1 \). Consequently, \( c_{0,k} = c_{1,k} = \ldots = c_{2s+1,k} = 0 \), and we conclude that the homogeneous counterpart of the linear algebraic system (2.43), (2.44) may only have a trivial solution. As such, the original inhomogeneous system (2.43), (2.44) itself will have a unique solution for any given choice of its right-hand sides.

**THEOREM 2.9**

Let \( f(x) \) be a polynomial of degree no greater than \( s \). Then, the interpolant \( \varphi(x,s) \) coincides with this polynomial.

**PROOF** We will prove the identity \( \varphi(x,s) \equiv f(x) \) on the interval \([x_k, x_{k+1}]\) for an arbitrary \( k \), i.e., for all \( x \) in-between any two neighboring interpolation nodes. In other words, we will prove that \( Q_{2s+1}(x,k) \equiv f(x) \). Due to the uniqueness of the interpolating polynomial, we have \( P_s(x,f_{kj}) \equiv P_s(x,f_{k+1,j}) \equiv f(x) \). Then, clearly, the polynomial \( f(x) \) solves system (2.43), (2.44).

**THEOREM 2.10**

The piece-wise polynomial interpolating function \( \varphi(x,s) \) defined by equalities (2.42) assumes the given values \( f(x_k) \) at the interpolation nodes \( x_k \), \( k = 0, \pm 1, \ldots \). Moreover, \( \varphi(x,s) \) has a continuous derivative of order \( s \) everywhere on its domain.

**PROOF** According to equalities (2.43) and (2.44), at any given node \( x_k \) the two functions: \( Q_{2s+1}(x,k-1) \) and \( Q_{2s+1}(x,k) \), have derivatives of orders \( m = 0,1,2,\ldots,s \) that coincide with the corresponding derivatives of one and the same interpolating polynomial \( P_s(x,f_{kj}) \). By virtue of equalities (2.42), this proves the theorem.

Let us now recast the polynomial \( Q_{2s+1}(x,k) \) as

\[
Q_{2s+1}(x,k) = P_s(x,f_{kj}) + R_{2s+1}(x,k),
\]

where \( R_{2s+1}(x,k) \) denotes a correction to the classical interpolating polynomial \( P_s(x,f_{kj}) \). Then, the following theorem holds.

\[\text{\footnotesize Derivative of order zero shall naturally be interpreted as the function itself.}\]
**THEOREM 2.11**

The correction \( R_{2s+1}(x, k) \) defined by (2.45) can be written in the form:

\[
R_{2s+1}(x, k) = (x_{k+1} - x_k)^{s+1} f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) q_{2s+1} \left( \frac{x-x_k}{x_{k+1} - x_k} \right)
\]

where \( f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) \) is a divided difference of order \( s+1 \), and

\[
q_{2s+1}(X, k) = \left( \frac{x_{k-j+s+1} - x_{k-j}}{x_{k+1} - x_k} \right) X
\]

\[
\sum_{r=0}^{s} \left[ \prod_{j=1}^{s} \left( X - \frac{x_{k-j+i} - x_k}{x_{k+1} - x_k} \right) \right]^{(r)} l_r(X), \quad X = \frac{x-x_k}{x_{k+1} - x_k}
\]

(2.47)

In formula (2.47), expression \( [\ldots]^{(r)} \) denotes a derivative of order \( r \) with respect to \( X \) evaluated for \( X = 1 \).

**REMARK 2.4** Representation (2.45) of the local piece-wise polynomial interpolant with \( s \) continuous derivatives can be thought of as Newton’s form of the interpolating polynomial \( P_{s+1}(x, f_k) \):

\[
P_{s+1}(x, f_k) = P_s(x, f_k) + f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) \phi_{s+1}(x, k),
\]

\[
\phi_{s+1}(x, k) = (x-x_{k-j})(x-x_{k-j+1}) \ldots (x-x_{k-j+s})
\]

in which the polynomial \( \phi_{s+1}(x, k) \) has been replaced by another polynomial:

\[
(x_{k+1} - x_k)^{s+1} q_{2s+1} \left( \frac{x-x_k}{x_{k+1} - x_k} \right)
\]

To illustrate Theorem 2.11, we first note that according to formulae (2.45)–(2.48), in the case \( s = 0 \), \( j = 0 \), the local spline \( q(s, k) \) renders a piece-wise linear interpolation.\(^4\) As far as the most interesting case from the standpoint of applications, it is, perhaps, \( s = 2 \), \( j = 1 \), and \( x_{k+1} - x_k = h = \text{const} \) for all \( k \). Then, we have:

\[
R_5(x, k) = P_2(x, f_{k,1}) + \frac{h^3}{2!} f(x_{k+2}) - 3f(x_{k+1}) + 3f(x_k) - f(x_{k-1}) \times
\]

\[
\left( \frac{x-x_k}{h} \right)^3 \frac{x-x_{k+1}}{h} \left( \frac{3}{h} - \frac{2(x-x_k)}{h} \right)
\]

(2.49)

\(^4\)Note that we have originally required that \( s > 0 \) and \( 0 \leq j \leq s-1 \). However, a formal substitution of \( s = 0 \), \( j = 0 \) into (2.45)–(2.48) does yield a piece-wise linear interpolant.
The proof of Theorem 2.11 is given in Section 2.3.3 that concludes our discussion of smooth piece-wise polynomial interpolation.

**THEOREM 2.12**

Let the function \( f(x) \) be defined everywhere on the interval \([x_{k-j}, x_{k-j+s+1}]\) and let its derivative \( f^{(s+1)}(x) \) be bounded on this interval. Then, the approximate equalities

\[
f^{(m)}(x) \approx \frac{d^m q(x,s)}{dx^m}, \quad m = 0, 1, \ldots, s,
\]

hold for all \( x \in [x_k, x_{k+1}] \), and the following error estimates are guaranteed:

\[
\max_{x_k \leq x \leq x_{k+1}} \left| f^{(m)}(x) - \frac{d^m q(x,s)}{dx^m} \right| \leq \text{const} \cdot (x_{k-j+s} - x_{k-j})^{s+1-m} \max_{x_{k-j} \leq x \leq x_{k-j+s+1}} |f^{(s+1)}(x)|.
\]  

**PROOF** First, using formulae (2.42) and (2.45), we can obtain:

\[
\left| f^{(m)}(x) - \frac{d^m q(x,s)}{dx^m} \right| \leq |f^{(m)}(x) - P^m_s(x, f_k)| + |R^m_{2s+1}(x,k)|.
\]  

Next, we recall estimate (2.36) established in Theorem 2.7:

\[
\max_{x_k \leq x \leq x_{k+1}} \left| f^{(m)}(x) - P^m_s(x, f_k) \right| \leq \text{const} \cdot (x_{k-j+s} - x_{k-j})^{s+1-m} \max_{x_{k-j} \leq x \leq x_{k-j+s+1}} |f^{(s+1)}(x)|.
\]

Then, taking into account that \( \frac{d^m}{dx^m} = \frac{1}{(x_{k+1} - x_k)^m} \frac{d^m}{dx^m} \), where \( X = \frac{x-x_k}{x_{k+1}-x_k} \), and using formulae (2.46) and (2.47) we obtain:

\[
\left| \frac{d^m R_{2s+1}(x,k)}{dx^m} \right| = (x_{k+1} - x_k)^{s+1-m} |f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})| \times \left| \frac{x_{k-j+s+1} - x_{k-j}}{x_{k+1} - x_k} \right| \sum_{r=0}^{x_k} \left\{ \left[ \prod_{l=1}^{s} \left( X - \frac{x_{k-j+l} - x_{k-j+l}}{x_{k+1} - x_k} \right) \right]^{(r)} X = 1 \right\} \frac{d^m f(X)}{dX^m}.
\]

Finally, according to Theorem 2.3 we can write:

\[
(s+1)!f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) = f^{(s+1)}(\xi), \quad \xi \in [x_{k-j}, x_{k-j+s+1}].
\]

Then, substitution of (2.55) into (2.54) yields the estimate:

\[
\left| \frac{d^m R_{2s+1}(x,k)}{dx^m} \right| \leq \text{const} \cdot (x_{k+1} - x_k)^{s+1-m} \max_{x_{k-j} \leq x \leq x_{k-j+s+1}} \left| \frac{d^{s+1} f(x)}{dx^{s+1}} \right|,
\]
that appear in formula (2.54). Clearly, \(C_2\) will remain bounded when the grid is refined as long as all ratios (2.57) remain bounded.

Then, the required estimate (2.51) follows from (2.52), (2.53), and (2.56), because

\[
\max_{x_{k-j} \leq x \leq x_{k-j+1}} |f^{(s+1)}(x)| \leq \max_{x_{k-j} \leq x \leq x_{k-j+1}} |f^{(s+1)}(x)| \quad \text{and also} \quad (x_{k+1} - x_k) \leq (x_{k-j+s} - x_{k-j}).
\]

**REMARK 2.5** On an equally spaced grid with size \(h\), error estimate (2.51) means uniform convergence of the local spline \(\varphi(x,s)\) and its derivatives of orders \(m = 1, 2, \ldots, s\), to the interpolated function \(f(x)\) and its respective derivatives, with the rate \(O(h^{s+1-m})\), as the grid is refined, i.e., as \(h \rightarrow 0\).

Let us also make here two remarks on optimality of the local splines (2.42).

**REMARK 2.6** Assuming that \(s+1\) is the maximum order of a bounded derivative that the function \(f(x)\) has, one cannot reconstruct \(f^{(m)}(x), m = 0, 1, 2, \ldots, s\), from the tabulated values \(f(x_k), k = 0, \pm 1, \ldots\), with an accuracy better than \(O(h^{s+1-m})\). This is an unsurmountable constraint, no matter what specific reconstruction methodology may be used. It is due to the lack of information contained in the values \(f(x_k)\) (see Section 2.2.4). At the same time, the best allowable accuracy \(O(h^{s+1-m})\) is achieved by estimate (2.51).

**REMARK 2.7** To evaluate \(\varphi(x,s)\) for every given \(x\), one needs no more than \(s+2\) interpolation nodes. This characteristic reflects on the local nature of formula (2.42), and it cannot be improved in the following sense. If we were to require that no more than \(s+1\) interpolation nodes be used, then in order to achieve convergence of \(\varphi^{(m)}(x)\) to \(f^{(m)}(x)\) with the rate of \(O(h^{s+1-m})\) we would have had to relinquish the condition of continuity even for the first derivative of \(\varphi(x,s)\), and rather resort back to the classical piece-wise polynomial interpolation of degree no greater than \(s\).

Let us also note that previously in this section we have basically used the interpolation grid with no “endpoints,” i.e., the grid that formally contained infinitely many nodes \(x_k, k = 0, \pm 1, \ldots\). If we were to use a finite interpolation grid instead, with a total of \(n+1\) nodes: \(x_0, x_1, \ldots, x_n\), located on some interval \([a,b]\) so that \(a = x_0 < x_1 < \ldots < x_n = b\), then we would have had to modify formula (2.42) near the endpoints of the interval \([a,b]\) and define the new interpolant \(\varphi(x) \equiv \varphi(x,s,a,b)\).
by the following equalities:

\[\varphi(x,s,a,b) = \begin{cases} 
P_s(x, f_{jj}), & \text{if } x_0 \leq x \leq x_j, \\
Q_{2s+1}(x,k), & \text{if } x \in [x_k, x_{k+1}] \text{ and } k = j, j+1, \ldots, n+j-s-1, \\
P_s(x, f_{n+j-s,j}), & \text{if } x_{n+j-s} \leq x \leq x_n. 
\end{cases} \tag{2.58}\]

We recall that the notion of \(P_s(x, f_{kj})\) was introduced in the end of Section 2.2.1.

The interpolant \(\varphi(x,s,a,b)\) defined on the interval \([a,b]\) has all continuous derivatives up to the order \(s\) everywhere on this interval. For this new interpolant, the results similar to those of Theorems 2.8–2.12 can be established.

**REMARK 2.8** Local smooth piece-wise polynomial interpolation has been introduced by Ryaben’kii in 1952 for the functions specified on multidimensional rectangular grids with constant size; the formulae presented in this section were obtained in 1974. The multi-dimensional results and the bibliography can be found in [Rya02, Section 1.1] and in [Rya75].

### 2.3.2 Nonlocal Smooth Piece-Wise Polynomial Interpolation

Local splines with \(s\) continuous derivatives that we have constructed in Section 2.3.1 are realized using polynomials of degree \(2s + 1\). For example, a twice continuously differentiable local spline is built as a piece-wise polynomial function of degree five. A natural question then arises of whether or not the same level of regularity (i.e., smoothness) can be achieved using polynomials of a lower degree.

The answer to this question is, generally speaking, affirmative. Splines \(\psi(x,s)\) of minimal degree \(m\) that would guarantee the desired extent of smoothness \(s\) have been constructed by Schoenberg for all positive integers \(s\). For example, if \(s = 2\), then it turns out that the minimal necessary degree is \(m = 3\), i.e., the spline is realized by a piece-wise cubic polynomial. In the literature, such splines are referred to as Schoenberg’s cubic splines or simply cubic splines. However, the key distinction between Schoenberg’s splines and local splines of Section 2.3.1 is that the reduction of degree of the constituent polynomials for Schoenberg’s splines is basically obtained at the expense of compromising their local nature. In other words, Schoenberg’s splines become nonlocal, so that the coefficients of each of their polynomial “pieces” depend on the function values on the entire interpolation grid, rather than only at a few neighboring nodes.

Assume that a total of \(n+1\) interpolation nodes: \(x_0, x_1, \ldots, x_n\), are given on the interval \([a,b]\) so that \(a = x_0 < x_1 < \ldots < x_n = b\), and assume also that the function values \(f(x_k), k = 0, 1, \ldots, n\), at these nodes are known. The Schoenberg cubic spline \(\psi(x,2) = \psi(x)\) is formally defined on \([a,b]\) as a function composed of individual third-degree polynomials \(P_3(x,k)\) specified on each partition interval \(x_k \leq x \leq x_{k+1}\), i.e., \(\psi(x) \equiv P_3(x,k)\) for \(x \in [x_k, x_{k+1}]\), \(k = 0, 1, \ldots, n-1\). The spline \(\psi\) must coin-
cide with the original function \( f \) at the interpolation nodes: \( \psi(x_k) = f(x_k) \) for all \( k = 0, 1, \ldots, n \). Besides, its first and second derivatives must be continuous everywhere. As \( \psi \) is a piece-wise polynomial function, i.e., for \( x \in [x_k, x_{k+1}] \) we have

\[
\psi(x) = P_3(x, k) = \sum_{l=0}^{3} a_{kl}(x-x_k)^l,
\]

where we denote \( a_{kl} \) the integration constants \( a_{kl} \). Besides, its second derivative is continuous:

\[
\psi''(x_k) = \psi''(x_{k+1}) = \psi''(x_0), \quad k = 1, \ldots, n - 1.
\]

Consider the function \( \psi''(x) \), it is piece-wise linear and continuous on the interval \([a, b]\). Then, we can write that for \( x \in [x_k, x_{k+1}] \):

\[
\psi''(x) = \psi''(x_k) + \frac{\psi''(x_{k+1}) - \psi''(x_k)}{h_k}(x-x_k)
\]

where we denote \( h_k = x_{k+1} - x_k, k = 0, 1, \ldots, n-1 \), hereafter. Moreover, according to the previously adopted notations for \( P_3(x, k) \) we have \( \psi''(x_k) = 2a_{k,2} \), and also since the second derivative of the spline is continuous: \( \psi''(x_{k+1} - 0) = \psi''(x_{k+1} + 0) = 2a_{k+1,2} \). Next, by integrating the previous equality twice we obtain:

\[
\psi(x) = a_{k+1,2} \frac{(x-x_k)^3}{3h_k} + a_{k,2} \frac{1}{3h_k} (x_{k+1} - x)^3 + c_{k,1} \frac{x-x_k}{h_k} + c_{k,2} \frac{x_{k+1} - x}{h_k},
\]

where the integration constants \( c_{k,1} \) and \( c_{k,2} \) are to be obtained from the conditions \( \psi(x_k) = f(x_k) \) and \( \psi(x_{k+1}) = f(x_{k+1}) \), respectively. This yields:

\[
a_{k+1,2} \frac{h_k^3}{3} + c_{k,1} = f(x_{k+1}),
\]

\[
a_{k,2} \frac{h_k^2}{3} + c_{k,2} = f(x_k).
\]

Therefore, the cubic spline for \( x \in [x_k, x_{k+1}] \) is given by the following expression:

\[
\psi(x) = a_{k+1,2} \frac{(x-x_k)^3}{3h_k} + a_{k,2} \frac{1}{3h_k} (x_{k+1} - x)^3 +
\]

\[
\left(f(x_{k+1}) - a_{k+1,2} \frac{h_k^2}{3}\right) \frac{x-x_k}{h_k} + \left(f(x_k) - a_{k,2} \frac{h_k^2}{3}\right) \frac{x_{k+1} - x}{h_k}.
\]

When deriving formula (2.59) we have already used the condition \( \psi(x_k) = f(x_k) \), as well as the condition of continuity of the second derivatives of the spline. We have not used the continuity of the first derivatives yet. To actually do so, we first differentiate formula (2.59) and substitute \( x = x_k \) thus obtaining \( \psi'(x_k + 0) \), and then write down a similar formula for the interval \([x_{k-1}, x_k]\), differentiate it and substitute \( x = x_k \) thus obtaining \( \psi'(x_k - 0) \). By requiring that \( \psi'(x_k - 0) = \psi'(x_k + 0) \), we get:

\[
a_{k-1,2} \frac{h_{k-1}}{3} + a_{k,2} \frac{2h_{k-1} + 2h_k}{3} + a_{k+1,2} \frac{h_k}{3} =
\]

\[
f(x_{k+1}) - f(x_k) \frac{h_k}{h_{k-1}} = f(x_k) - f(x_{k-1}).
\]
Equations (2.60) need to be considered for \( k = 1, 2, \ldots, n - 1 \), which yields a total of \( n - 1 \) linear algebraic equations with respect to \( n + 1 \) unknown coefficients \( a_{0,2}, a_{1,2}, \ldots, a_{n,2} \). We therefore conclude that the cubic spline \( \psi(x) \) is not yet defined unambiguously by the conditions specified previously, and two additional conditions must be provided. For example, we may set \( a_{0,2} = a_{n,2} = 0 \), which corresponds to \( \psi''(x_0) \equiv \psi''(a) = 0 \) and \( \psi''(x_n) \equiv \psi''(b) = 0 \). Then, system (2.60) can be written in the matrix form as follows:

\[
Aa = g. \tag{2.61}
\]

In system (2.61), \( a = [a_{1,2}, a_{2,2}, \ldots, a_{n-1,2}]^T \) is an \( n - 1 \)-dimensional vector of unknowns, \( g = [g_1, g_2, \ldots, g_{n-1}]^T \) is the right-hand side: \( g = Bf \), where \( f = [f(x_0), f(x_1), \ldots, f(x_n)]^T \) is a given vector of dimension \( n + 1 \), \( A \) is a square matrix of dimension \( (n - 1) \times (n - 1) \):

\[
A = \begin{bmatrix}
\frac{2h_0 + 2h_1}{h_1} & \frac{h_1}{h_2} & 0 & \cdots & 0 & 0 \\
\frac{2h_1}{h_2} & \frac{h_2}{h_3} & \frac{2h_1}{h_3} & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \frac{h_{n-2}}{h_{n-1}} & \frac{2h_{n-2} + 2h_{n-1}}{h_{n-1}}
\end{bmatrix}, \tag{2.62}
\]

and \( B \) is a matrix of dimension \( (n - 1) \times (n + 1) \):

\[
B = \begin{bmatrix}
\frac{1}{h_0} & -\frac{1}{h_1} & -\frac{1}{h_2} & 0 & \cdots & 0 & 0 \\
0 & \frac{1}{h_1} & -\frac{1}{h_1} - \frac{1}{h_2} & \frac{1}{h_2} & \cdots & 0 & 0 \\
0 & 0 & \frac{1}{h_2} & -\frac{1}{h_2} - \frac{1}{h_3} & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \cdots & -\frac{1}{h_{n-2}} - \frac{1}{h_{n-1}} & \frac{1}{h_{n-1}}
\end{bmatrix}. \tag{2.63}
\]

Both matrices, \( A \) of (2.62) and \( B \) of (2.63) are tri-diagonal. The matrix \( A \) of (2.62) is strictly diagonally dominant by rows, as defined in Chapter 5 (this matrix is also symmetric). Therefore, \( A \) is non-singular, and as such, system (2.61) always has one and only one solution \( a \) for any right-hand side \( g \). This means that the spline \( \psi(x) \) exists and is unique; it can be obtained from the tabulated function values \( f(x_k), k = 0, 1, \ldots, n \), using formula (2.59). Numerically, the matrix \( A \) of (2.62) can be inverted by a very efficient algorithm known as the tri-diagonal elimination. This algorithm is stable, and has linear computational complexity with respect to the dimension of the matrix, see Section 5.3. Let us also note that once the solution \( a = [a_{1,2}, a_{2,2}, \ldots, a_{n-1,2}]^T \) of system (2.61) has been obtained, the spline \( \psi(x) \) can already be built. If necessary, however, the remaining three coefficients \( a_{k,0}, a_{k,1}, \) and \( a_{k,3} \) of the polynomial \( P_3(x,k) = \sum_{l=0}^{3} a_{kl}(x-x_k)^l \) for every \( k = 0, 1, \ldots, n - 1 \), can also be found explicitly, by merely regrouping the terms in formula (2.59).
Finally, we may consider periodic boundary conditions for the spline: the system, and will not disrupt its solvability by the tri-diagonal elimination either.

As has been shown, the Schoenberg cubic spline basically requires two additional conditions for uniqueness. Of course, there are many other choices for these conditions, besides the previously used $\psi''(a) = \psi''(b) = 0$. For example, the spline may have a non-zero curvature at the endpoints $a$ and $b$; this yields inhomogeneous conditions for its second derivatives instead of the homogeneous ones:

$$\psi''(a) \equiv 2a_{0,2} = \psi''_{0}, \quad \psi''(b) \equiv 2a_{n,2} = \psi''_{n}, \quad (2.64)$$

where $\psi''_{0}$ and $\psi''_{n}$ are given constants. The new boundary conditions (2.64) prompt a minor change in system (2.61); namely, the first and last components of its right-hand side $g$ will now be defined as follows:

$$g_1 = \frac{f(x_2) - f(x_1)}{h_1} - \frac{f(x_1) - f(x_0)}{h_0} - \frac{\psi''_{0}h_0}{6},$$

$$g_{n-1} = \frac{f(x_n) - f(x_{n-1})}{h_{n-1}} - \frac{f(x_{n-1}) - f(x_{n-2})}{h_{n-2}} - \frac{\psi''_{n}h_{n-1}}{6}.$$

Alternatively, one may specify the slope of the spline rather than its curvature at the endpoints $a$ and $b$, which yields two additional conditions for the first derivatives:

$$\psi'(a) \equiv \frac{f(x_1) - f(x_0)}{h_0} - a_{0,2} \frac{2h_0}{3} - a_{1,2} \frac{h_0}{3} = \psi'(0),$$

$$\psi'(b) \equiv \frac{f(x_n) - f(x_{n-1})}{h_{n-1}} + a_{n,2} \frac{2h_{n-1}}{3} + a_{n-1,2} \frac{h_{n-1}}{3} = \psi'(n), \quad (2.65)$$

where, again, $\psi'(0)$ and $\psi'(n)$ are some constants. Relations (2.65) can be added to system (2.60) as its new first and new last equation, thus making the overall dimension of the system $(n + 1) \times (n + 1)$. This modification will not alter the consistency of the system, and will not disrupt its solvability by the tri-diagonal elimination either.

Finally, we may consider periodic boundary conditions for the spline:

$$f(x_0) = f(x_n), \quad \psi'(a) = \psi'(b), \quad \psi''(a) = \psi''(b). \quad (2.66)$$

Then, system (2.60) gets supplemented by one additional equation of the same structure (only one is needed because the other one is simply $a_{n,2} = a_{0,2}$):

$$a_{n-1,2} \frac{h_{n-1}}{3} + a_{0,2} \frac{2h_{n-1} + 2h_0}{3} + a_{1,2} \frac{h_0}{3} =$$

and while its resulting matrix form can still be symbolically written as (2.61), the vector of unknowns will now be $n$-dimensional: $\mathbf{a} = [a_{0,2}, a_{1,2}, \ldots, a_{n-1,2}]^T$, so will
be the vector of the right-hand sides \( g = Bf \); the matrix \( A \) will have dimension \( n \times n \):

\[
A = \begin{bmatrix}
\frac{2h_{n-1}+2h_0}{3} & \frac{h_0}{3} & 0 & \cdots & 0 & h_{n-1} \\
\frac{h_0}{3} & \frac{2h_0+2h_1}{3} & \frac{h_1}{3} & \cdots & 0 & 0 \\
0 & \frac{h_1}{3} & \frac{2h_1+2h_2}{3} & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\frac{h_{n-1}}{3} & 0 & 0 & \cdots & \frac{h_{n-2}}{3} & \frac{2h_{n-2}+2h_{n-1}}{3}
\end{bmatrix}, \quad (2.67)
\]

and the matrix \( B \) will be a square \( n \times n \) matrix as well:

\[
B = \begin{bmatrix}
\left(-\frac{1}{h_{n-1}} - \frac{1}{h_0}\right) & \frac{1}{h_0} & 0 & \cdots & 0 & \frac{1}{h_{n-1}} \\
\frac{1}{h_0} & \left(-\frac{1}{h_0} - \frac{1}{h_1}\right) & \frac{1}{h_1} & \cdots & 0 & 0 \\
0 & \frac{1}{h_1} & \left(-\frac{1}{h_1} - \frac{1}{h_2}\right) & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\frac{1}{h_{n-1}} & 0 & 0 & \cdots & \frac{1}{h_{n-2}} & \left(-\frac{1}{h_{n-2}} - \frac{1}{h_{n-1}}\right)
\end{bmatrix}, \quad (2.68)
\]

In so doing, the vector \( f \) whose components are given tabulated values of the function \( f \) will be \( n \)-dimensional: \( f = [f(x_0), f(x_1), \ldots, f(x_{n-1})]^T \), because \( f(x_n) = f(x_0) \). Notice that both the matrix \( A \) of (2.67) and the matrix \( B \) of (2.68) are symmetric and almost tri-diagonal, except for the non-zero entries in the upper-right and lower-left corners of each matrix. The matrix \( A \) of (2.67) can be efficiently inverted by a special cyclic version of the tri-diagonal elimination.

Other types of boundary conditions for the spline \( \psi(x) \) at the endpoints \( a \) and \( b \) can be used as well.

Let us now assume that the function \( f(x) \) has a bounded derivative of order \( s+1 \) on the interval \( [a, b] \). Then, one can set the required additional conditions so that the nonlocal spline \( \psi(x, s) \) with \( s \) continuous derivatives will keep the same optimal approximation properties that characterize both the classical piece-wise polynomial interpolation of Section 2.2 and the local smooth piece-wise polynomial interpolation of Section 2.3.1. Namely, estimates of type (2.51) will hold for \( m = 0, 1, \ldots, s \):

\[
\max_{a \leq x \leq b} \left| f^{(m)}(x) - \frac{d^m \psi(x, s)}{dx^m} \right| \leq \text{const} \left( \max_k h_k \right)^{s+1-m} \max_{a \leq x \leq b} \left| f^{(s+1)}(x) \right|.
\]

In particular, provided that \( f(x) \) is twice continuously differentiable on \( [a, b] \), the cubic spline \( \psi(x, 2) \equiv \psi(x) \), along with its derivatives of orders \( m = 1, 2 \), will converge to the function \( f(x) \) and its respective derivatives with the rate of \( O(h^{s+1-m}) \).

On the other hand, the undesirable property of saturation by smoothness, which is inherent both for the classical piece-wise polynomial interpolation and for the local splines, is shared by the Schoenberg splines as well, notwithstanding the loss of their local nature. Namely, the nonlocal splines of smoothness \( s \) on a uniform interpolation grid with size \( h \) will guarantee the order of error \( O(h^{s+1}) \) for the functions that have a
maximum of \( s + 1 \) bounded derivatives, and they will only guarantee the same order of error \( O(h^{s+1}) \) for the functions \( f(x) \) that have more than \( s + 1 \) derivatives as well.

We should also mention that even though the coefficients of a nonlocal spline on a given interval \([x_k, x_{k+1}]\) depend on the function values on the entire grid, it is known that in practice the influence exerted by the remote nodes is rather weak. Nonetheless, to actually evaluate the coefficients, one needs to solve the full system (2.61). Therefore, a natural task of improving the accuracy by adding a few more interpolation nodes in a particular local region would basically imply starting from the very beginning for a Schoenberg’s spline, i.e., writing down and then solving a new system of type (2.61). In contradistinction to that, the splines of Section 2.3.1 are particularly well suited for such local grid refinements. In so doing, the additional computational effort is simply proportional to the number of new nodes.

### 2.3.3 Proof of Theorem 2.11

Coefficients of the polynomial

\[
Q_{2s+1}(x,k) = c_{0,k} + c_{1,k}x + \ldots + c_{2s+1,k}x^{2s+1}
\]  

(2.69)

are determined by solving the linear algebraic system (2.43), (2.44). The right-hand sides of the equations that compose sub-system (2.43) have the form

\[
a_0^{(m)} f_{k-j} + a_1^{(m)} f_{k-j+1} + \ldots + a_s^{(m)} f_{k-j+s}, \quad m = 0, 1, \ldots, s,
\]

while those that pertain to sub-system (2.44) have the form

\[
b_0^{(m)} f_{k-j+1} + b_1^{(m)} f_{k-j+2} + \ldots + b_s^{(m)} f_{k-j+s+1}, \quad m = 0, 1, \ldots, s,
\]

where \( a_i^{(m)}, b_i^{(m)}, i = 0, 1, \ldots, s, m = 0, 1, \ldots, s, \) are some numbers that do not depend on \( f_{k-j}, f_{k-j+1}, \ldots, f_{k-j+s+1} \).

Consequently, one can say that the given quantities \( f_{k-j}, f_{k-j+1}, \ldots, f_{k-j+s+1} \) determine the solution \( c_{0,k}, c_{1,k}, \ldots, c_{2s+1,k} \) of system (2.43), (2.44) through the formulae of type

\[
c_{r,k} = a_0^{(r)} f_{k-j} + a_1^{(r)} f_{k-j+1} + \ldots + a_s^{(r)} f_{k-j+s+1}, \quad r = 0, 1, \ldots, 2s + 1,
\]  

(2.70)

where \( a_i^{(r)}, r = 0, 1, \ldots, s + 1, \) are, again, some numbers that do not depend on \( f_{k-j}, f_{k-j+1}, \ldots, f_{k-j+s+1} \). Next, by substituting expressions (2.70) into formula (2.69) and grouping together all the terms that contain each particular \( f_{k-j+l}, l = 0, 1, \ldots, s + 1, \) we can recast the polynomial (2.69) as follows:

\[
Q_{2s+1}(x,k) = \sum_{l=0}^{s+1} f_{k-j+l}p(x,l),
\]  

(2.71)

\(^5\)Hereafter in this section, we will use the notation \( f = f(x) \) for all nodes \( l.\)
where \( p(x, l) \), \( l = 0, 1, \ldots, s + 1 \), are some polynomials of the argument \( x \). Let us now replace the quantities \( f_{k-j+1} \) in formula (2.71) using formulae (2.16), see Theorem 2.4. Then, the polynomial \( Q_{2s+1}(x, k) \) becomes:

\[
Q_{2s+1}(x, k) = \sum_{l=0}^{s} f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})q_l(x, k) + (x_{k+1} - x_k)^{(s+1)}f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})q_{2s+1}(x, k),
\]

(2.72)

where \( q_l(x, k) \), \( l = 0, 1, \ldots, s \), and \( q_{2s+1}(x, k) \) are some polynomials independent of \( f_{k-1}, f_{k-j+1}, \ldots, f_{k-j+s+1} \). We will exploit this independence for the purpose of finding the polynomials \( q_l(x, k) \), \( l = 0, 1, \ldots, s \). Let us specify \( f_{k-1}, f_{k-j+1}, \ldots, f_{k-j+s} \) arbitrarily, and let us specify \( f_{k-j+s+1} \) as follows:

\[
f_{k-j+s+1} = P_s(x, f) \bigg|_{x=x_{k-j+s+1}}.
\]

According to Theorem 2.9, we then have

\[
Q_{2s+1}(x, k) = P_s(x, f),
\]

while Corollary 2.3 from the Newton’s formula (see Section 2.1) implies that the divided difference \( f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) \) of order \( s + 1 \) vanishes. Then, formula (2.72) transforms into:

\[
P_s(x, f) = \sum_{l=0}^{s} f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})q_l(x, k).
\]

(2.73)

By comparing formula (2.73) with formula (2.3), we conclude that as the values \( f_{k-1}, f_{k-j+1}, \ldots, f_{k-j+s} \) have been chosen arbitrarily, then \( q_l(x, k) = \frac{(x-x_k-j)}{(x-k-j)} \cdots \frac{(x-x_k-j)}{(x-k-j-s+1)} \cdot \frac{(x-x_k-j)}{(x-x_k-j+s+1)} \cdot \frac{(x-x_k-j)}{(x-x_k-j+s+1)} \cdot \frac{(x-x_k-j)}{(x-x_k-j+s+1)} \), \( l = 0, 1, \ldots, s \). Equality (2.73) also implies that formula (2.72) can be recast as follows:

\[
Q_{2s+1}(x, k) = P_s(x, f) + (x_{k+1} - x_k)^{(s+1)}f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})q_{2s+1}(x, k).
\]

(2.74)

To obtain the polynomial \( q_{2s+1}(x, k) \) from (2.74), we specify

\[
f_{k-j} = f_{k-j+1} = \ldots = f_{k-j+s} = 0, \text{ and } f_{k-j+s+1} = 1.
\]

(2.75)

Then, equality (2.74) becomes:

\[
Q_{2s+1}(x, k) = (x_{k+1} - x_k)^{(s+1)}f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})q_{2s+1}(x, k).
\]

(2.76)

Note that under conditions (2.75), the Lagrange form of the interpolating polynomial, see formula (2.1), yields

\[
P_{s+1}(x, f, x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) = \prod_{l=0}^{s} \frac{x-x_{k-j-l}}{x_{k-j-s+1}-x_{k-j-s+1}},
\]

(2.77)
while according to its Newton’s form (2.3), we have:

\[
P_{s+1}(x; f, x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) = f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) \prod_{j=0}^{s+1} (x - x_{k-j+1}).
\]  

(2.78)

By comparing (2.77) and (2.78), we conclude that once conditions (2.75) hold:

\[
f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1}) = \left[ \prod_{j=0}^{s+1} (x - x_{k-j+1}) \right]^{-1}.
\]

Then, from formula (2.76) we find that

\[
\tilde{q}_{2s+1}(x, k) = (x_{k+1} - x_{k})^{-1} Q_{2s+1}(x, k) \prod_{j=0}^{s} (x_{k-j+s+1} - x_{k-j+1}),
\]

(2.79)

where the expressions on the right-hand side of (2.79) are evaluated, again, under conditions (2.75).

Let us now transform formula (2.79) to the required form (2.47), (2.48), see Theorem 2.11. For that purpose, we will need to obtain an explicit expression for the polynomial \( Q_{2s+1}(x, k) \) driven by the function values of (2.75). Note that due to condition (2.43), along with (2.75), the value \( x = x_{k} \) appears to be a root of multiplicity \( s + 1 \) for the polynomial \( Q_{2s+1}(x, k) \), so that the function \( Q(x) = Q_{2s+1}(x, k)/(x - x_{k})^{s+1} \) is a polynomial of degree no greater than \( s \). We can then expand \( Q(x) \) with respect to the powers of \( x - x_{k+1} \):

\[
Q_{2s+1}(x, k) = (x - x_{k})^{s+1} Q(x) = (x - x_{k})^{s+1} \frac{Q_{2s+1}(x, k)}{(x - x_{k})^{s+1}} = (x - x_{k})^{s+1} \sum_{r=0}^{s} \frac{1}{r!} \left[ \frac{Q_{2s+1}(x, k)}{(x - x_{k})^{r+1}} \right]_{x=x_{k+1}}^{(r)} (x - x_{k+1})^r.
\]

(2.80)

Next, we can use equality (2.44) and write:

\[
\frac{d^r}{dx^r} \left[ \frac{Q_{2s+1}(x, k)}{(x - x_{k})^{s+1}} \right]_{x=x_{k+1}} = \sum_{l=0}^{r} \binom{r}{l} \frac{d^{r-l}}{dx^{r-l}} P_{s}(x, f, x_{k-j+1}, x_{k-j+2}, \ldots, x_{k-j+s+1}) \bigg|_{x=x_{k+1}} \times
\]

\[
[(x - x_{k})^{s+1}]_{x=x_{k+1}} \end{matrix} \bigg|_{x=x_{k+1}}
\]

(2.81)

where \( \binom{r}{l} = \frac{r!}{l!(r-l)!} \) are the binomial coefficients. However, conditions (2.75) along with the Lagrange formula (2.1) yield [cf. formula (2.77)]:

\[
P_{s}(x, f, x_{k-j+1}, x_{k-j+2}, \ldots, x_{k-j+s+1}) = \prod_{j=0}^{s} \frac{x - x_{k-j+i}}{x_{k-j+s+1} - x_{k-j+i}}.
\]

(2.82)
Theoretical Introduction to Numerical Analysis

Substituting (2.82) into (2.81), we obtain:

\[
\frac{d^r}{dx^r} \left[ \frac{Q_{2k+1}(x,k)}{(x-x_k)^{s+1}} \right]_{x=x_{k+1}} = \left[ \prod_{l=1}^{s} (x_{k-l} + x_{k+l}) \right]^{-1} \times \\
\sum_{l=0}^{r} \binom{r}{l} (x-x_k)^{-r-1-\frac{r}{2}} \prod_{i=1}^{s} (x-x_{k-i})^{r-1} 
\]  

(2.83)

Finally, we substitute (2.83) into (2.80), and subsequently (2.80) into (2.79), thus obtaining an explicit expression for \(q_{2k+1}(x,k)\) as a double sum with respect to \(l\) and \(r\). This sum would depend on \(x\) only as a composite function \(q_{2k+1}(X,k)\) of the argument \(X = (x-x_k)/(x_{k+1} - x_k)\):

\[q_{2k+1}(x,k) = q_{2k+1}(X,k)\]

Then, changing the summation order in the resulting formula for \(q_{2k+1}(X,k)\) obtained after the last substitution into (2.79), we arrive at expressions (2.47), (2.48).

Exercises

1. Let \(x_{k+1} - x_k = h = \text{const.}\), and let \(\psi(x,s)\) be a local spline of given smoothness \(s\) defined by formula (2.45) for \(x_k \leq x \leq x_{k+1}\). Explicitly write down the polynomials \(h_k(x,s)\) that would enable representation of \(\psi(x,s)\) in the form [cf. formula (2.1)]:

\[\psi(x,s) = \sum_{k} f(x_k) h_k(x,s)\]

Analyze the cases: \(s = 0, j = 0; s = 1, j = 0;\) and \(s = 2, j = 1\).

2.* Let \(\|f\| \tag{6} = \max_{k-j \leq l \leq k+j+1} |f_l|\), and let the interpolation grid be uniform. Estimate the value of the Lebesgue constant

\[L_s = \max_{|f| = 1 \text{ for } x_{k+1}} \max_{k-j \leq l \leq k+j+1} |Q_{2k+1}(x,k)|\]

for the local spline with \(s = 1\) and \(j = 0\).

\[\text{Hint.}\] See Lemma 3.3 of Section 3.2.7, Chapter 3.

3. Delineate the concluding part of the proof of Theorem 2.11, given in the last paragraph of Section 2.3.3.

4.* Prove that the nonlocal cubic spline \(\psi(x)\) of (2.59) that satisfies the homogeneous boundary conditions (2.64): \(\psi^{(a)}(a) = \psi^{(b)}(b) = 0\), delivers minimum to the functional

\[\Phi(u) = \int_{a}^{b} [u''(x)]^2 dx\]

on the class of all interpolants \(u = u(x), a \leq x \leq b, u(x_k) = f(x_k), k = 0, 1, \ldots, n,\) that have a square integrable second derivative: \(u \in W^2_2[a,b]\).

5. Let \([a,b] = [-\pi, \pi]\), let \(f_1(x) = |x|\), \(f_2(x) = \sin x\), and let the interpolation grid be uniform with size \(h\). For both functions \(f_1\) and \(f_2\) implement on the computer the local spline (2.42), (2.58) with \(s = 2, j = 1\), and the nonlocal cubic spline (2.59) with any type of boundary conditions: (2.64), (2.65), or (2.66). Demonstrate that in either case the convergence rate is \(O(h^2)\).
2.4 Interpolation of Functions of Two Variables

The problem of reconstructing a function of continuous argument from its discrete table of values can be formulated in the multi-dimensional case as well, for example, when \( f = f(x, y) \), i.e., when there are two independent variables. The principal objective remains the same as in the case of one dimension, namely, to build a procedure for (approximately) evaluating the function in-between the given interpolation nodes. However, in the case of two variables one can consider a much wider variety of interpolation grids. All these grids basically fall into one of the two categories — structured or unstructured.

2.4.1 Structured Grids

Typically, structured grids on the \((x, y)\) plane are composed of rectangular cells. In other words, the nodes \((x_k, y_l), k = 0, \pm 1, \ldots, l = 0, \pm 1, \ldots\), are obtained as intersections of the two families of straight lines: the vertical lines \(x = x_k, k = 0, \pm 1, \ldots\), and the horizontal lines \(y = y_l, l = 0, \pm 1, \ldots\). In so doing, we always assume that \( \forall k : x_{k+1} > x_k \), and \( \forall l : y_{l+1} > y_l \). In the literature, such grids are also referred to as rectangular or Cartesian. The grid sizes \( h_k^{(x)} = x_{k+1} - x_k \) and \( h_l^{(y)} = y_{l+1} - y_l \) may but do not have to be constant. In the case of constant size(s), the grid is called uniform or equally spaced (in the corresponding direction). The simplest example of a uniform two-dimensional grid is a grid with square cells: \( h_k^{(x)} = h_l^{(y)} = \text{const.} \).

To approximately compute the value of the function \( f \) at the point \((\bar{x}, \bar{y})\) that does not coincide with any of the nodes \((x_k, y_l)\) of a structured rectangular grid, one can, in fact, use the apparatus of piece-wise polynomial interpolation for the functions of one variable. To do so, we first select the parameters \( s \) (degree of interpolation) and \( j \), as in Section 2.2. We also need to determine what cell of the grid the point of interest belongs to. Let us assume that \( x_k < \bar{x} < x_{k+1} \) and \( y_l < \bar{y} < y_{l+1} \) for some particular values of \( k \) and \( l \). Then, we interpolate along the horizontal grid lines:

\[
\tilde{f}(\bar{x}, y_{l-j+i}) = P_s(x, f(\cdot, y_{l-j+i}), x_{k-j}, x_{k-j+1}, \ldots, x_{k+j}), \quad i = 0, 1, \ldots, s.
\]

and obtain the intermediate values \( \tilde{f} \). Having done that, we interpolate along the vertical grid lines and obtain the approximate value \( f \):

\[
f(\bar{x}, \bar{y}) \approx P_s(y, \tilde{f}(\bar{x}, \cdot), y_{l-j}, y_{l-j+1}, \ldots, y_{l+j}).
\]

Clearly, the foregoing formulae can be used to approximate the function \( f \) at any point \((\bar{x}, \bar{y})\) inside the rectangular grid cell \( \{(x, y) | x_k < x < x_{k+1}, y_l < y < y_{l+1}\} \). For example, if we choose piece-wise linear interpolation along \( x \) and \( y \), i.e., \( s = 1 \), then

\[
f(\bar{x}, \bar{y}) \approx f(x_k, y_l) \left( \frac{x - x_{k+1}}{x_k - x_{k+1}} \right) \left( y - y_{l+1} \right) + f(x_{k+1}, y_l) \left( \frac{x - x_k}{x_{k+1} - x_k} \right) \left( y - y_{l+1} \right) + f(x_k, y_{l+1}) \left( \frac{x - x_{k+1}}{x_k - x_{k+1}} \right) \left( y - y_l \right) + f(x_{k+1}, y_{l+1}) \left( \frac{x - x_k}{x_{k+1} - x_k} \right) \left( y - y_l \right).
\]

(2.84)
Note, however, that in general the degree of interpolation does not necessarily have to be the same for both dimensions. Also note that the procedure is obviously symmetric. In other words, it does not matter whether we first interpolate along $x$ and then along $y$, as shown above, or the other way around.

The piece-wise polynomial interpolation on the plane, built dimension-by-dimension on a rectangular grid as explained above, inherits the key properties of the one-dimensional interpolation. For example, if the function $f = f(x, y)$ is twice differentiable, with bounded second partial derivatives, then the interpolation error of formula (2.84) on a square-cell grid with size $h$ will be $O(h^2)$. For piece-wise polynomial interpolation of a higher degree, the rate of convergence will accordingly be faster, provided that the interpolated function is sufficiently smooth. On the other hand, similarly to the one-dimensional case, the two-dimensional piece-wise polynomial interpolation is also prone to the saturation by smoothness.

Again, similarly to the one-dimensional case, one can also construct a smooth piece-wise polynomial interpolation in two dimensions. As before, this interpolation may be either local or nonlocal. Local splines that would extend the methodology of Section 2.3.1 can be built on the plane dimension-by-dimension, in much the same way as the conventional piece-wise polynomials outlined previously. Their key properties will be preserved from the one-dimensional case — the relation between the degree and the smoothness, the minimum number of grid nodes in each direction, the convergence rate, and susceptibility to saturation (see [Rya75] for detail).

The construction of nonlocal cubic splines can also be extended to two dimensions; in this case the splines are called bi-cubic. On a domain of rectangular shape, they can be obtained by solving multiple tri-diagonal linear systems of type (2.61) along the $x$ and $y$ coordinate lines of the Cartesian grid. The approximation properties of bi-cubic splines remain the same as those of the one-dimensional cubic splines.

Similar constructions — standard piece-wise polynomials, local splines, and non-local splines — are also available for the interpolation of multivariable functions (more than two arguments). We should emphasize, however, that in general the size of the tables that would guarantee a given accuracy of interpolation for a function of certain smoothness rapidly grows as the number of arguments of the function increases. The corresponding interpolation algorithms also become more cumbersome.

### 2.4.2 Unstructured Grids

Unstructured grids on the $(x, y)$ plane are typically composed of triangular cells. In so doing, it is required that any two triangles have either a common side or a common vertex, or alternatively, do not intersect at all. Unstructured grids offer a lot more flexibility in accommodating irregular geometric shapes, see Figure 2.1. In addition, they may also be more convenient when the interpolated function undergoes a strong variation in some local area.
Let, for example, the function \( f = f(x, y) \) be defined on the domain schematically shown in Figure 2.1. Let us additionally assume that it varies rapidly inside the “bottleneck.” In this case, a rectangular grid will not be particularly well suited for tabulating the function \( f(x, y) \). One reason is that such a grid obviously cannot fit the curvilinear boundary of the domain. Another reason is that a rectangular grid cannot be refined locally, i.e., only in the bottleneck area, where most of the variation of the function \( f(x, y) \) supposedly occurs. On the other hand, the unstructured triangular grid shown in Figure 2.1 is obviously free of these disadvantages. Therefore, the vertices of the triangles can be used in the capacity of interpolation nodes.

However, a key difficulty in using triangular grids for interpolation is that the previously analyzed one-dimensional methods will not directly extend to this case in a dimension-by-dimension fashion. Instead, triangular grids require genuinely two-dimensional approaches to building the interpolating polynomials.

Let, for example, \( f_1, f_2, \) and \( f_3 \) be the values of the function \( f(x, y) \) at the vertices \((x_1, y_1), (x_2, y_2), \) and \((x_3, y_3)\), respectively, of some triangle. Then, inside the triangle the unknown function \( f(x, y) \) can be approximated by the linear function (a first degree polynomial of two variables):

\[
f(x, y) \approx ax + bx + c,
\]

where the coefficients \( a, b, \) and \( c \) must satisfy the equalities:

\[
ax_1 + by_1 + c = f_1,
ax_2 + by_2 + c = f_2,
ax_3 + by_3 + c = f_3,
\]

in order for the interpolant to coincide exactly with the function values at the interpolation nodes. Geometrically, the linear function (2.85) is a plane that crosses through the three points: \((x_1, y_1, f_1), (x_2, y_2, f_2), \) and \((x_3, y_3, f_3)\) in the three-dimensional space. For a twice differentiable function with bounded second derivatives, the interpolation error guaranteed by this approach will be \( O(h^2) \), where \( h \) is the largest side of the triangle. If the linear interpolant (2.85) is built for every triangle of the grid, see Figure 2.1, then the overall interpolating function will be piece-wise linear and continuous on the entire domain.

We note that linear interpolants on triangular grids are very instrumental for building the finite-element approximations of partial differential equations, see Section 12.1 of Chapter 12. We also note that piece-wise polynomial interpolants of higher degrees can be obtained on triangular grids as well.
Exercises

1. Prove the $O\left(h(x)^2 + h(y)^2\right)$ error estimate for the interpolation (2.84), where $h(x)$ and $h(y)$ are the grid sizes (constant in each direction).

2. Prove the $O(h^2)$ error estimate for the interpolation (2.85), where $h$ is the largest side of the triangle.

3. Let $f = f(x, y)$ be twice differentiable, and let all of its second partial derivatives be bounded by one. How shall one economically choose the size(s) of the rectangular interpolation grid on the plane so that to be able to reconstruct the function $f(x, y)$ everywhere inside the square $|x| \leq 1, |y| \leq 1$ with the error not exceeding $10^{-3}$?

4. Let $f = f(x, y)$ be twice differentiable, let all of its second partial derivatives be bounded by one, and let also the circles $x^2 + y^2 = r^2$ be level lines of the function $f = f(x, y)$ for every $r$. How can one economically place the interpolation nodes on the plane so that to achieve the same quality of the reconstructed function as in Problem 3?

5. Answer the question of Problem 4, but instead of having the circles $x^2 + y^2 = r^2$ as level lines of $f$, assume that the function $f = f(x, y)$ can be represented as a product of two single-variable functions: $f = \phi(x)\psi(y)$, where $\phi(x) \neq 0$ and $\psi(y) \neq 0$. 
Chapter 3

Trigonometric Interpolation

Along with the algebraic interpolation described in Chapter 2, one also uses interpolation by means of trigonometric polynomials of the type:

$$Q \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x \right) = \sum_{k=0}^{n} a_k \cos \frac{2\pi k}{L} x + \sum_{k=1}^{n} b_k \sin \frac{2\pi k}{L} x, \quad (3.1)$$

where \( n \) is a positive integer, \( L > 0 \), and \( a_k \) & \( b_k \) are the coefficients. A trigonometric interpolating polynomial \( Q \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x \right) \) that would coincide with the given \( L \)-periodic function \( f(x) \), \( f(x+L) = f(x) \), at the equidistant interpolation nodes:

$$x_m = \frac{L}{N} m + x_0, \quad m = 0, 1, \ldots, N - 1, \quad x_0 = \text{const}, \quad (3.2)$$

can be chosen so that it would have some important advantages compared to the algebraic interpolating polynomial built on the same grid (3.2).

First, the error of the trigonometric interpolation

$$R_N(x, f) \overset{\text{def}}{=} f(x) - Q \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x \right) \quad (3.3)$$

converges to zero uniformly with respect to \( x \) as \( N \rightarrow \infty \) already if the second derivative of \( f(x) \) is piece-wise continuous.\(^1\) Moreover, the rate of this convergence, i.e., the rate of decay of the error (3.3) as \( N \rightarrow \infty \), automatically takes into account the smoothness of \( f(x) \), i.e., increases for those functions \( f(x) \) that have more derivatives. Specifically, we will prove that

$$\max_x |R_N(x, f)| = O \left( \frac{M_{r+1}}{N^{r-1/2}} \right), \quad \text{where} \quad M_{r+1} = \max_x \left| \frac{d^{r+1} f(x)}{dx^{r+1}} \right|. \quad (3.3)$$

Second, it turns out that the sensitivity of the trigonometric interpolating polynomial (3.1) to the errors committed when specifying the function values \( f_m = f(x_m) \) on the grid (3.2) remains “practically flat” (i.e., grows slowly) as \( N \) increases.

The foregoing two properties — automatic improvement of accuracy for smoother functions, and slow growth of the Lebesgue constants that translates into numerical stability — are distinctly different from the properties of algebraic interpolation on

\(^1\)In fact, even less regularity may be required of \( f(x) \), see Section 3.2.7.
uniform grids, see Chapter 2. It, however, turns out that algebraic interpolation of functions on an interval can also possess the same remarkable qualities. To achieve that, one shall choose the Chebyshev interpolation nodes and use Chebyshev interpolating polynomials.

When reading the book for the first time, one can restrict him/herself by what has been said about the contents of Chapter 3, and proceed directly to Chapter 4.

3.1 Interpolation of Periodic Functions

Let \( f \equiv f(x) \) be an \( L \)-periodic function:

\[
\forall x: \quad f(x + L) = f(x), \quad L > 0,
\]

defined on the grid:

\[
x_m = \frac{L}{N} m + x_0, \quad m = 0, \pm 1, \pm 2, \ldots, \quad x_0 = \text{const},
\]

where \( N \) is a positive integer. Introducing the notation \( f_m \equiv f(x_m) \), we obtain by virtue of (3.4) and (3.5):

\[
\forall m: \quad f_{m+N} = f_m.
\]

3.1.1 An Important Particular Choice of Interpolation Nodes

**THEOREM 3.1**

Let \( x_0 = L/(2N), \quad N = 2(n+1), \quad n \) being a positive integer. Let \( f \) be an \( L \)-periodic function, and let \( f_m \) be its values on the grid (3.5). For a given arbitrary set of \( f_m \), there is one and only one trigonometric interpolating polynomial:

\[
Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right) = \sum_{k=0}^{n} a_k \cos \frac{2\pi k}{L} x + \sum_{k=1}^{n+1} b_k \sin \frac{2\pi k}{L} x, \quad (3.6)
\]

that satisfies the equalities:

\[
Q_n \bigg|_{x=x_m} = f_m, \quad m = 0, \pm 1, \pm 2, \ldots
\]
The coefficients of this polynomial are given by the formulae:

\[
a_0 = \frac{1}{N} \sum_{m=0}^{N-1} f_m, \tag{3.7}
\]

\[
a_k = \frac{2}{N} \sum_{m=0}^{N-1} f_m \cos \left( \frac{2\pi}{N} m + \frac{\pi}{N} \right), \quad k = 1, 2, \ldots, n, \tag{3.8}
\]

\[
b_k = \frac{2}{N} \sum_{m=0}^{N-1} f_m \sin \left( \frac{2\pi}{N} m + \frac{\pi}{N} \right), \quad k = 1, 2, \ldots, n, \tag{3.9}
\]

\[
b_{n+1} = \frac{1}{N} \sum_{m=0}^{N-1} f_m (-1)^m. \tag{3.10}
\]

**PROOF** Let us consider a set of all real valued periodic discrete functions:

\[
f_{m+N} = f_m, \quad m = 0, \pm 1, \pm 2, \ldots, \tag{3.11}
\]

defined on the grid \( x_m = -\frac{L}{N} m + \frac{L}{2N} \). We will only be considering these functions on the grid interval \( m = 0, 1, \ldots, N-1 \), because for all other \( m \)'s they can be unambiguously reconstructed by virtue of periodicity (3.11).

The entire set of these functions, supplemented by the conventional operations of addition and multiplication by a real scalar quantity, form a linear space that we will denote \( F_N \). The dimension of this space is equal to \( N \), because the system of \( N \) linearly independent functions (vectors) \( \tilde{\psi}^{(k)} \in F_N \), \( k = 1, 2, \ldots, N \):

\[
\tilde{\psi}^{(k)}_m \overset{\text{def}}{=} \begin{cases} 
0, & \text{if } m \neq k-1, \\
1, & \text{if } m = k-1,
\end{cases}
\]

provides a basis in the space \( F_N \). Indeed, any function \( f \in F_N \), \( f = \{ f_m | m = 0, 1, \ldots, N \} \) always admits a unique representation as a linear combination of the basis functions \( \tilde{\psi}^{(k)} \): \( f = \sum_{k=1}^{N} f_{k-1} \tilde{\psi}^{(k)} \).

Let us now introduce a Euclidean dot product in the space \( F_N \):

\[
(f, g) = \frac{1}{N} \sum_{m=0}^{N-1} f_m g_m, \tag{3.12}
\]

and show that the system of functions:

\[
\xi^{(0)}_m = \cos(0 \cdot x_m) \equiv 1,
\]

\[
\xi^{(k)}_m = \sqrt{2} \cos \left( \frac{2\pi k}{L} x_m \right), \quad k = 1, 2, \ldots, n,
\]

\[
\eta^{(k)}_m = \sqrt{2} \sin \left( \frac{2\pi k}{L} x_m \right), \quad k = 1, 2, \ldots, n,
\]

\[
\eta^{(n+1)}_m = \sin \left( \frac{2\pi (n+1)}{L} x_m \right) = (-1)^m, \tag{3.13}
\]
forms an orthonormal basis in the space $F_N$ (we remind that $n = N/2 − 1$). The overall number of functions defined by formulae (3.13) is equal to $N$. Therefore, it only remains to prove the equalities:

$$(\xi^{(k)}, \xi^{(k)}) = 1, \quad k = 0, 1, \ldots, n,$$  \hspace{1cm} (3.14)

$$(\eta^{(k)}, \eta^{(k)}) = 1, \quad k = 1, 2, \ldots, n + 1,$$  \hspace{1cm} (3.15)

$$(\xi^{(r)}, \xi^{(s)}) = 0, \quad r \neq s, \quad r, s = 0, 1, \ldots, n,$$  \hspace{1cm} (3.16)

$$(\eta^{(r)}, \eta^{(s)}) = 0, \quad r \neq s, \quad r, s = 1, 2, \ldots, n + 1,$$  \hspace{1cm} (3.17)

$$(\xi^{(r)}, \eta^{(s)}) = 0, \quad r = 0, 1, \ldots, n, \quad s = 1, 2, \ldots, n + 1.$$  \hspace{1cm} (3.18)

To prove equalities (3.14)–(3.18), we first notice that for any $N$ and $\gamma$ the following relations hold:

$$\sum_{m=0}^{N-1} \frac{1}{N} N^{-1} 1 = 1,$$  \hspace{1cm} (3.19)

$$\sum_{m=0}^{N-1} \cos \left( \frac{2\pi}{N} m + \gamma \right) = 0, \quad l = 1, 2, \ldots, N - 1,$$  \hspace{1cm} (3.20)

$$\sum_{m=0}^{N-1} \sin \left( \frac{2\pi}{N} m + \gamma \right) = 0, \quad l = 1, 2, \ldots, N - 1.$$  \hspace{1cm} (3.21)

Indeed, (3.19) is obviously true. To verify (3.20), we can write for any $l = 0, 1, \ldots, N - 1$:

$$\sum_{m=0}^{N-1} \cos \left( \frac{2\pi}{N} m + \gamma \right) =$$
$$\frac{1}{2} \sum_{m=0}^{N-1} \left[ \exp \left( i \left( \frac{2\pi m}{N} + l \gamma \right) \right) + \exp \left( -i \left( \frac{2\pi m}{N} + l \gamma \right) \right) \right] =$$
$$\frac{1}{2} e^{i\gamma} \sum_{m=0}^{N-1} \left( \exp \left( i \frac{2\pi m}{N} \right) \right)^m + \frac{1}{2} e^{-i\gamma} \sum_{m=0}^{N-1} \left( \exp \left( -i \frac{2\pi m}{N} \right) \right)^m =$$
$$\frac{1}{2} e^{i\gamma} \frac{1 - \exp \{ i2\pi \gamma \}}{1 - \exp \{ i2\pi / N \}} + \frac{1}{2} e^{-i\gamma} \frac{1 - \exp \{ -i2\pi \gamma \}}{1 - \exp \{ -i2\pi / N \}} = 0 + 0 = 0,$$  \hspace{1cm} (3.21)

where we have used the formula for the sum of a geometric sequence. Equality (3.21) is proven similarly.

We are now ready to show that the basis (3.13) is indeed orthonormal, i.e., that equalities (3.14)–(3.18) do hold. Equality (3.14) for $k = 0$, as well as equality (3.15) for $k = n + 1$, coincide with (3.19). For $k = 1, 2, \ldots, n$ both
equalities (3.14) and (3.15) hold by virtue of (3.20) and (3.21):

\[(\xi^{(k)}, \xi^{(k)}) = \frac{2}{N} \sum_{m=0}^{N-1} \cos^2 \left( \frac{2\pi k}{N} m + \frac{\pi k}{N} \right) = \frac{1}{N} \sum_{m=0}^{N-1} \left[ 1 + \cos \left( \frac{4\pi k}{N} m + \frac{2\pi k}{N} \right) \right] = \frac{1}{N} \sum_{m=0}^{N-1} 1 + \frac{1}{N} \sum_{m=0}^{N-1} \cos 2k \left( \frac{2\pi}{N} m + \frac{\pi}{N} \right) = 1 + 0 = 1, \]

\[(\eta^{(k)}, \eta^{(k)}) = \frac{2}{N} \sum_{m=0}^{N-1} \sin^2 \left( \frac{2\pi k}{N} m + \frac{\pi k}{N} \right) = \frac{1}{N} \sum_{m=0}^{N-1} \left[ 1 - \cos \left( \frac{4\pi k}{N} m + \frac{2\pi k}{N} \right) \right] = 1. \]

To prove equality (3.16), we first notice that for \(r,s = 0,1,\ldots,n\) and \(r \neq s\) we always have \(1 \leq |r \pm s| \leq N - 1\), and then use formula (3.20) to obtain:

\[(\xi^{(r)}, \xi^{(s)}) = \frac{2}{N} \sum_{m=0}^{N-1} \cos r \left( \frac{2\pi m}{N} + \frac{\pi k}{N} \right) \cos s \left( \frac{2\pi m}{N} + \frac{\pi k}{N} \right) = \frac{1}{N} \sum_{m=0}^{N-1} \left[ \cos \left( (r+s) \left( \frac{2\pi m}{N} + \frac{\pi k}{N} \right) \right) + \cos \left( (r-s) \left( \frac{2\pi m}{N} + \frac{\pi k}{N} \right) \right) \right] = 0. \]

Equality (3.17) is proven similarly, except that instead of the trigonometric identity \(2 \cos \alpha \cos \beta = \cos(\alpha + \beta) + \cos(\alpha - \beta)\) that has been used when proving (3.16), one rather needs to employ another identity: \(2 \sin \alpha \sin \beta = \cos(\alpha + \beta) - \cos(\alpha - \beta)\). Finally, yet another trigonometric identity: \(2 \sin \alpha \cos \beta = \sin(\alpha + \beta) + \sin(\alpha - \beta)\) is to be used for proving formula (3.18).

Altogether, we have established that (3.13) is an orthonormal basis in the space \(F_N\). Therefore, every function \(f = \{f_m\} \in F_N\) can be represented as a linear combination of the basis functions (3.13):

\[f_m = \sum_{k=0}^{n} a_k \cos \frac{2\pi k}{L} x_m + \sum_{k=1}^{n+1} b_k \sin \frac{2\pi k}{L} x_m. \]

Calculating the dot products of both the left-hand side and the right-hand side of the previous equality with all the basis functions \(\xi^{(r)}\) and \(\eta^{(s)}\), \(r = 0,1,\ldots,n\), \(s = 1,2,\ldots,n+1\), we arrive at the equalities:

\[a_0 = (f, \xi^{(0)}), \]
\[a_k = \sqrt{2} (f, \xi^{(k)}), \quad k = 1,2,\ldots,n, \]
\[b_k = \sqrt{2} (f, \eta^{(k)}), \quad k = 1,2,\ldots,n, \]
\[b_{n+1} = (f, \eta^{(n+1)}), \]
that, according to definitions (3.13), coincide with formulae (3.7)–(3.10).

The grid \( x_m = \frac{L}{N} m + \frac{L}{2N}, \quad m = 0, \pm 1, \pm 2, \ldots \), that has been used for specifying the discrete functions \( f \in F_N \) in Theorem 3.1 is symmetric with respect to the origin \( x = 0 \), so that along with the point \( x = x_m \) it also contains the point \( x = -x_m = x_{-(m+1)} \). Therefore, one can consider even and odd functions on this grid. The grid function \( f_m \equiv f(x_m) \) is called even if

\[
f_m = f_{-(m+1)}, \quad m = 0, \pm 1, \pm 2, \ldots
\]

(3.22)

Similarly, the grid function \( f_m \) is called odd if

\[
f_m = -f_{-(m+1)}, \quad m = 0, \pm 1, \pm 2, \ldots
\]

(3.23)

**THEOREM 3.2**  
Let \( f_m \) be an even \( N \)-periodic grid function specified at the nodes \( x_m = \frac{L}{N} m + \frac{L}{2N}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad N = 2(n + 1) \). Then, the trigonometric interpolating polynomial (3.6) becomes:

\[
Q_n = \sum_{k=0}^{n+1} a_k \cos \frac{2\pi k}{L} x,
\]

(3.24)

where

\[
a_0 = \frac{1}{n+1} \sum_{m=0}^{n} f_m,
\]

(3.25)

\[
a_k = \frac{2}{n+1} \sum_{m=0}^{n} f_m \cos k \left( \frac{\pi}{n+1} m + \frac{\pi}{2(n+1)} \right), \quad k = 1, 2, \ldots, n.
\]

(3.26)

**PROOF**  
For an even function \( f_m \), see formula (3.22), expressions (3.7) and (3.8) translate into (3.25) and (3.26), respectively, whereas expressions (3.9) and (3.10) imply that \( b_k \equiv 0 \).

**THEOREM 3.3**  
Let \( f_m \) be an odd \( N \)-periodic grid function specified at the nodes \( x_m = \frac{L}{N} m + \frac{L}{2N}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad N = 2(n + 1) \). Then, the trigonometric interpolating polynomial (3.6) becomes:

\[
Q_n = \sum_{k=1}^{n+1} b_k \sin \frac{2\pi k}{L} x,
\]

(3.27)

\(^2\)Note that the origin itself is not a node of this grid.
where

\[ b_k = \frac{2}{n+1} \sum_{m=0}^{n} f_m \sin \left( \frac{\pi}{n+1} m + \frac{\pi}{2(n+1)} \right), \quad k = 1, 2, \ldots, n, \quad (3.28) \]

\[ b_{n+1} = \frac{1}{n+1} \sum_{m=0}^{n} f_m (-1)^m. \quad (3.29) \]

**PROOF** For an odd function \( f_m \), see formula (3.23), formulae (3.7) and (3.8) imply that the coefficients \( a_k, k = 0, 1, \ldots, n \), are equal to zero, expressions (3.9) and (3.10) then transform into (3.28) and (3.29), respectively, and the polynomial (3.6), accordingly, reduces to (3.27).

### 3.1.2 Sensitivity of the Interpolating Polynomial to Perturbations of the Function Values

Let us estimate how sensitive the trigonometric interpolating polynomial (3.6) is to the errors committed when specifying the values of the grid function \( f_m \). Assume that instead of \( f = \{f_m\} \) we actually have a perturbed function \( f + \delta f = \{f_m + \delta f_m\} \). Then, instead of the polynomial (3.6) we obtain a new polynomial:

\[ Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f + \delta f \right) = Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right) + \delta Q_n. \]

From formulae (3.6) and (3.7)–(3.10) one can easily see that the corresponding error is given by:

\[ \delta Q_n = Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, \delta f \right). \]

Therefore, similarly to how it has been done in Section 2.1.4 of Chapter 2, we can introduce the **Lebesgue constants**:

\[ L_n \overset{\text{def}}{=} \sup_{\delta f \in F_n} \frac{\max_x |Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, \delta f \right)|}{\max_m |\delta f_m|}, \quad n = 1, 2, \ldots, \quad (3.30) \]

that would naturally quantify the sensitivity of the trigonometric interpolating polynomial (3.6) to the perturbations of its input data. Obviously,

\[ \max_x |\delta Q_n| \leq L_n \max_m |\delta f_m|. \quad (3.31) \]

**THEOREM 3.4**

The Lebesgue constants of the trigonometric interpolating polynomials (3.6) satisfy the estimates

\[ L_n \leq 2(n+1). \quad (3.32) \]
PROOF Formulae (3.7)–(3.10) imply that
\[ |a_k| \leq 2 \max_m |f_m|, \quad \text{(3.33)} \]
\[ |b_k| \leq 2 \max_m |f_m|. \quad \text{(3.34)} \]
Then, from formula (3.6) and inequalities (3.33), (3.34) applied to the perturbation \( \delta f \) one can easily see that
\[ \left| Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, \delta f \right) \right| \leq \sum_{k=0}^{n} |a_k| + \sum_{k=1}^{n+1} |b_k| \leq 2(n+1) \max_m |\delta f_m|. \quad \text{(3.35)} \]
Estimate (3.35) holds for any \( \delta f \in F_N \) and any \( N \). This implies (3.32).

Let us emphasize that the growth rate of the Lebesgue constants in the case of trigonometric interpolation on equally spaced nodes, see estimate (3.32), is considerably slower than that in the case of algebraic interpolation, see estimates (2.18) in Section 2.1.4 of Chapter 2.

3.1.3 Estimate of Interpolation Error

**THEOREM 3.5**

Let \( f = f(x) \) be an \( L \)-periodic function that has continuous \( L \)-periodic derivatives up to the order \( r > 0 \) and a square integrable derivative of order \( r+1 \):
\[ \int_0^L \left[ f^{(r+1)}(x) \right]^2 dx < \infty. \]

Let \( f \in F_N, \ f = \{f_m\} \), be a table of values of this function sampled at the equally spaced grid nodes
\[ x_m = \frac{L}{N} m + \frac{L}{2N}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad N = 2(n+1), \]
and accordingly, let \( Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right) \) be the trigonometric interpolating polynomial of type (3.6). Then, for the error of the trigonometric interpolation:
\[ R_n(x) \overset{\text{def}}{=} f(x) - Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right), \]
the following estimate holds:
\[ |R_n(x)| \leq \frac{\zeta_n}{n^{r-1/2}}, \quad \text{where} \quad \zeta_n = o(1) \quad \text{as} \quad n \rightarrow \infty. \quad (3.36) \]

**PROOF** Let us represent \( f(x) \) as the sum of its Fourier series:
\[ f(x) = S_n(x) + \delta S_n(x), \]
where
\[ S_n(x) = \frac{\alpha_0}{2} + \sum_{k=1}^{n} \alpha_k \cos \frac{2\pi k x}{L} + \sum_{k=1}^{n+1} \beta_k \sin \frac{2\pi k x}{L} \]  \hspace{1cm} (3.37)

is a partial sum of order \( n \) and
\[ \delta S_n(x) = \sum_{k=n+1}^{\infty} \alpha_k \cos \frac{2\pi k x}{L} + \sum_{k=n+2}^{\infty} \beta_k \sin \frac{2\pi k x}{L} \]  \hspace{1cm} (3.38)
is the corresponding remainder. The coefficients \( \alpha_k \) and \( \beta_k \) of the Fourier series for \( f(x) \) are given by the formulae:
\[ \alpha_k = \frac{2}{L} \int_{0}^{L} f(x) \cos \frac{2\pi k x}{L} \, dx, \quad \beta_k = \frac{2}{L} \int_{0}^{L} f(x) \sin \frac{2\pi k x}{L} \, dx. \]  \hspace{1cm} (3.39)

Let us first prove the following estimate for the remainder \( \delta S_n \) of (3.38):
\[ |\delta S_n(x)| \leq \frac{\zeta_n}{n^{r+\frac{1}{2}}}, \]  \hspace{1cm} (3.40)

where \( \zeta_n \) is a numerical sequence such that \( \zeta_n = o(1) \), i.e., \( \lim_{n \to \infty} \zeta_n = 0 \). Estimate (3.40) is an estimate of the convergence rate of the Fourier series for \( f(x) \). Denote by \( A_k \) and \( B_k \) the Fourier coefficients of the square integrable \( L \)-periodic function \( f^{(r+1)}(x) \):
\[ A_k = \frac{2}{L} \int_{0}^{L} f^{(r+1)}(x) \cos \frac{2\pi k x}{L} \, dx, \quad B_k = \frac{2}{L} \int_{0}^{L} f^{(r+1)}(x) \sin \frac{2\pi k x}{L} \, dx. \]  \hspace{1cm} (3.41)

Then, we can integrate the previous equalities by parts \( r+1 \) times, employ the periodicity, and use definitions (3.39) to obtain \( A_0 = 0 \) and either:
\[ A_k = \pm \left( \frac{2\pi k}{L} \right)^{r+1} \alpha_k, \quad B_k = \mp \left( \frac{2\pi k}{L} \right)^{r+1} \beta_k, \]  \hspace{1cm} (3.42)
or
\[ A_k = \pm \left( \frac{2\pi k}{L} \right)^{r+1} \beta_k, \quad B_k = \mp \left( \frac{2\pi k}{L} \right)^{r+1} \alpha_k, \]  \hspace{1cm} (3.43)
for \( k = 1, 2, \ldots, \) depending on the particular value of \( r \). Moreover, according to the Bessel inequality (see, e.g., [KF75]) we can write:
\[ \sum_{k=1}^{\infty} A_k^2 + B_k^2 \leq \frac{2}{L} \int_{0}^{L} \left[ f^{(r+1)}(x) \right]^2 \, dx. \]  \hspace{1cm} (3.43)
Let us then define the sequence $\mu_k \overset{\text{def}}{=} \left(\frac{k}{2\pi}\right)^{r+1} \sqrt{A_k^2 + B_k^2}$. Inequality (3.43) implies that the series $\sum_{k=1}^{\infty} \mu_k^2$ converges. Now assume for definiteness that equalities (3.41) hold; then we have:

$$|\alpha_k| = \left(\frac{L}{2\pi}\right)^{r+1} \frac{|A_k|}{k^{r+1}} \leq \left(\frac{L}{2\pi}\right)^{r+1} \frac{\sqrt{A_k^2 + B_k^2}}{k^{r+1}} = \frac{\mu_k}{k^{r+1}},$$

and similarly

$$|\beta_k| = \left(\frac{L}{2\pi}\right)^{r+1} \frac{|B_k|}{k^{r+1}} \leq \left(\frac{L}{2\pi}\right)^{r+1} \frac{\sqrt{A_k^2 + B_k^2}}{k^{r+1}} = \frac{\mu_k}{k^{r+1}}.$$  

The same estimates can obviously be obtained when relations (3.42) hold instead of (3.41). Altogether we can therefore say that the Fourier coefficients (3.39) satisfy the following inequalities:

$$|\alpha_k| \leq \frac{\mu_k}{k^{r+1}}, \quad |\beta_k| \leq \frac{\mu_k}{k^{r+1}},$$  

where the sequence $\mu_k = o(1)$ is such that the series $\sum_{k=1}^{\infty} \mu_k^2$ converges.

Then, for the remainder (3.38) of the Fourier series we can write:

$$|\delta S_n(x)| = \left| \sum_{k=n+1}^{\infty} \alpha_k \cos \frac{2\pi kx}{L} + \sum_{k=n+2}^{\infty} \beta_k \sin \frac{2\pi kx}{L} \right| \leq \sum_{k=n+1}^{\infty} |\alpha_k| + \sum_{k=n+2}^{\infty} |\beta_k| \leq \sum_{k=n+1}^{\infty} |\alpha_k| + \sum_{k=n+1}^{\infty} |\beta_k| \leq 2 \sum_{k=n+1}^{\infty} \frac{\mu_k}{k^{r+1}} \leq 2 \sqrt{\sum_{k=n+1}^{\infty} \mu_k^2} \sqrt{\sum_{k=n+1}^{\infty} \frac{1}{k^{2(r+1)}},}$$

where the last estimate in (3.45) is obtained using the classical Hölder inequality (see [KF75]):

$$\sum_{k=1}^{\infty} u_k v_k \leq \sqrt{\sum_{k=1}^{\infty} u_k^2} \sqrt{\sum_{k=1}^{\infty} v_k^2}$$

valid for any two sequences $\{u_k\}$ and $\{v_k\}$, such that $\sum_{k=1}^{\infty} u_k^2 < \infty$ and $\sum_{k=1}^{\infty} v_k^2 < \infty$.

Next, we note that for $\xi \in [k-1,k]$ the following inequality always holds:

$$\frac{1}{k^{2(r+1)}} \leq \frac{1}{\xi^{2(r+1)}}, \quad \text{and consequently,} \quad \frac{1}{k^{2(r+1)}} \leq \int_{k-1}^{k} \frac{d\xi}{\xi^{2(r+1)}} \cdot \text{Therefore,} \quad \int_{k-1}^{k} \frac{d\xi}{\xi^{2(r+1)}} = \int_{1}^{\infty} \frac{d\xi}{\xi^{2(r+1)}} = \frac{1}{(2r+1)n^{2r+1}}.$$
Substituting the latter estimate into inequality (3.45), we obtain:

$$|\delta S_n(x)| \leq \frac{2}{\sqrt{2r+1}} \sqrt{\sum_{k=n+1}^{\infty} \mu_k^2} \frac{1}{\sqrt{n^{2r+1}}}.$$

(3.46)

It only remains to notice that $$\sum_{k=1}^{\infty} \mu_k^2$$ converges. Then, estimate (3.46) does imply (3.40).

Having justified estimate (3.40), we next notice that the partial sum $$S_n(x)$$ is, in fact, a trigonometric polynomial of type (3.6). Due to the uniqueness of the trigonometric interpolating polynomial, see Theorem 3.1, we then have

$$Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, S_n \right) = S_n(x).$$

(3.47)

Moreover, estimates (3.31), (3.32), and (3.40) together yield:

$$\left| Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, \delta S_n \right) \right| \leq L_n |\delta S_n| = L_n \frac{\zeta_n}{n^{r+\frac{1}{2}}} \leq 4 \frac{\zeta_n}{n^{r+\frac{1}{2}}}.$$

(3.48)

Finally, by combining (3.40), (3.47), and (3.48) we obtain the following estimate for the interpolation error $$R_n(x)$$:

$$|R_n(x)| = \left| f(x) - Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right) \right| =$$

$$\left| f(x) - Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, S_n \right) - Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, \delta S_n \right) \right| =$$

$$\left| (f(x) - S_n(x)) - Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, \delta S_n \right) \right| \leq$$

$$|\delta S_n(x)| + L_n |\delta S_n(x)| \leq \frac{\zeta_n}{n^{r+\frac{1}{2}}} + 4 \frac{\zeta_n}{n^{r+\frac{1}{2}}} \leq \text{const} \cdot \frac{\zeta_n}{n^{r+\frac{1}{2}}},$$

which is obviously equivalent to the required estimate (3.36).

We emphasize that the rate of convergence of the trigonometric interpolating polynomials established by estimate (3.36) automatically becomes faster for smoother interpolated functions $$f(x)$$. In other words, unlike in the case of algebraic interpolation (Chapter 2), when the convergence rate is limited by the degree of the polynomial, the trigonometric interpolation of periodic functions is not characterized by the undesirable property of saturation by smoothness.
REMARK 3.1 The estimate for the Lebesgue constants (3.32) that we have used when proving Theorem 3.5 can, in fact, be further improved, see Section 3.2.7. Accordingly, the foregoing estimate for the interpolation error, see (3.36), can be improved as well.

3.1.4 An Alternative Choice of Interpolation Nodes

Another type of trigonometric interpolation is also important for applications.

THEOREM 3.6

Let $f$ be an $L$-periodic function, and let $f_m$ be its values on the grid:

$$x_m = \frac{L}{N}m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad N = 2n.$$  

For a given arbitrary set of $f_m$, there is one and only one trigonometric interpolating polynomial:

$$\tilde{Q}_n \left( \cos \frac{2\pi}{L}x, \sin \frac{2\pi}{L}x, f \right) = \sum_{k=0}^{n} \tilde{a}_k \cos \frac{2\pi k}{L}x + \sum_{k=1}^{n-1} \tilde{b}_k \sin \frac{2\pi k}{L}x, \quad (3.49)$$

that satisfies the equalities:

$$\tilde{Q}_n \bigg|_{x=x_m} = f_m, \quad m = 0, \pm 1, \pm 2, \ldots.$$  

The coefficients of this polynomial are given by the formulae:

$$\tilde{a}_0 = \frac{1}{N} \sum_{m=0}^{N-1} f_m, \quad (3.50)$$

$$\tilde{a}_k = \frac{2}{N} \sum_{m=0}^{N-1} f_m \cos \frac{2\pi km}{N}, \quad k = 1, 2, \ldots, n - 1, \quad (3.51)$$

$$\tilde{a}_n = \frac{1}{N} \sum_{m=0}^{N-1} f_m (-1)^m, \quad (3.52)$$

$$\tilde{b}_k = \frac{2}{N} \sum_{m=0}^{N-1} f_m \sin \frac{2\pi km}{N}, \quad k = 1, 2, \ldots, n - 1. \quad (3.53)$$

PROOF It is very similar to that of Theorem 3.1, and we omit it here. □
For an even grid function, \( f_m = f_{-m} \), formulae (3.50)–(3.53) transform into:

\[
\tilde{a}_0 = \frac{1}{N}(f_0 + f_n) + \frac{2}{N} \sum_{m=1}^{n-1} f_m,
\]
\[
\tilde{a}_k = \frac{2}{N}(f_0 + (-1)^k f_n) + \frac{4}{N} \sum_{m=1}^{n-1} f_m \cos \frac{2\pi km}{N}, \quad k = 1, 2, \ldots, n-1,
\]
\[
\tilde{a}_n = \frac{1}{N}(f_0 + (-1)^n f_n),
\]
\[
\tilde{b}_k = 0, \quad k = 1, 2, \ldots, n-1,
\]

and the polynomial (3.49) reduces to:

\[
\tilde{Q}_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right) = \sum_{k=0}^{n} \tilde{a}_k \cos \frac{2\pi k}{L} x.
\]

Note that the arguments very similar to those used when proving the key properties of the trigonometric interpolating polynomial \( Q_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right) \) in Theorems 3.4 and 3.5, also apply to the polynomial \( \tilde{Q}_n \left( \cos \frac{2\pi}{L} x, \sin \frac{2\pi}{L} x, f \right) \) defined by formulae (3.49)–(3.53). Namely, this polynomial has slowly growing Lebesgue constants and as such, is basically stable with respect to the perturbations of the grid function \( f_m \). Moreover, it converges to the interpolated function \( f(x) \) as \( n \to \infty \) with the rate determined by the smoothness of \( f(x) \).

**REMARK 3.2** If the interpolated function \( f(x) \) has derivatives of all orders, then the rate of convergence of the trigonometric interpolating polynomials to \( f(x) \) will be faster than any inverse power of \( n \). In the literature, this type of convergence is often referred to as *spectral*.

### 3.2 Interpolation of Functions on an Interval. Relation between Algebraic and Trigonometric Interpolation

Let \( f = f(x) \) be defined on the interval \(-1 \leq x \leq 1\), and let it have there a bounded derivative of order \( r+1 \). We have chosen this specific interval \(-1 \leq x \leq 1\) as the domain of \( f(x) \), rather than an arbitrary interval \( a \leq x \leq b \), for the only reason of simplicity and convenience. Indeed, the transformation \( x = \frac{a+b}{2} + \frac{b-a}{2} t \) renders a transition from the function \( f(x) \) defined on an arbitrary interval \( a \leq x \leq b \) to the function \( F(t) \equiv f \left( \frac{a+b}{2} + \frac{b-a}{2} t \right) \) defined on the interval \(-1 \leq t \leq 1\).

#### 3.2.1 Periodization

According to Theorem 3.5 of Section 3.1, trigonometric interpolation is only suitable for the reconstruction of smooth periodic functions from their tables of values.
Therefore, to be able to apply it to the function \( f(x) \) given on \(-1 \leq x \leq 1\), one should first equivalently replace \( f(x) \) by some smooth periodic function. However, a straightforward extension of the function \( f(x) \) from its domain \(-1 \leq x \leq 1\) to the entire real axis may, generally speaking, yield a discontinuous periodic function with the period \( L = 2 \), see Figure 3.1.

Therefore, instead of the function \( f(x) \), \(-1 \leq x \leq 1\), let us consider a new function

\[
F(\varphi) = f(\cos \varphi), \quad x = \cos \varphi. \tag{3.54}
\]

It will be convenient to think that the function \( F(\varphi) \) of (3.54) is defined on the unit circle as a function of the polar angle \( \varphi \). The value of \( F(\varphi) \) is obtained by merely translating the value of \( f(x) \) from the point \( x \in [-1, 1] \) to the corresponding point \( \varphi \in [0, \pi] \) on the unit circle, see Figure 3.2. In so doing, one can interpret the resulting function \( F(\varphi) \) as an even, \( F(-\varphi) = F(\varphi) \), \( 2\pi \)-periodic function of its argument \( \varphi \). Moreover, it is easy to see from definition (3.54) that the derivative \( \frac{d}{d\varphi} F(\varphi) \) exists and is bounded.
3.2.2 Trigonometric Interpolation

Let us choose the following interpolation nodes:

\[ \phi_m = \frac{2\pi}{N} m + \frac{\pi}{N}, \quad m = 0, \pm 1, \ldots, \pm n, -(n+1), \quad N = 2(n+1). \]  (3.55)

According to (3.54), the values \( F_m = F(\phi_m) \) of the function \( F(\phi) \) at the nodes \( \phi_m \) of (3.55) coincide with the values \( f_m = f(x_m) \) of the original function \( f(x) \) at the points \( x_m = \cos \phi_m \). To interpolate a \( 2\pi \)-periodic even function \( F(\phi) \) using its tabulated values at the nodes (3.55), one can employ formula (3.24) of Section 3.1:

\[ Q_n(\cos \phi, \sin \phi, F) = \sum_{k=0}^{n} a_k \cos k\phi. \]  (3.56)

As \( F_m = f_m \) for all \( m \), the coefficients \( a_k \) of the trigonometric interpolating polynomial (3.56) are given by formulae (3.25), (3.26) of Section 3.1:

\[ a_0 = \frac{1}{n+1} \sum_{m=0}^{n} f_m, \]
\[ a_k = \frac{2}{n+1} \sum_{m=0}^{n} f_m \cos k\phi_m, \quad k = 1, 2, \ldots, n. \]  (3.57)

3.2.3 Chebyshev Polynomials. Relation between Algebraic and Trigonometric Interpolation

Let us use the equality \( \cos \phi = x \) and introduce the functions:

\[ T_k(x) = \cos k\phi = \cos(k\arccos x), \quad k = 0, 1, 2, \ldots. \]  (3.58)

THEOREM 3.7

The functions \( T_k(x) \) defined by formula (3.58) are polynomials of degree \( k = 0, 1, 2, \ldots \). Specifically, \( T_0(x) = 1 \), \( T_1(x) = x \), and all other polynomials: \( T_2(x), T_3(x), \ldots \), etc., can be obtained consecutively using the recursion formula

\[ T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x). \]  (3.59)

PROOF It is clear that \( T_0(x) = \cos 0 = 1 \) and \( T_1(x) = \cos \arccos x = x \). Then, we employ a well-known trigonometric identity

\[ \cos(k+1)\phi = 2\cos\phi \cos k\phi - \cos(k-1)\phi, \quad k = 1, 2, \ldots, \]

which immediately yields formula (3.59) when \( \phi = \arccos x \). It only remains to prove that \( T_k(x) \) is a polynomial of degree \( k \); we will use induction with respect to \( k \) to do that. For \( k = 0 \) and \( k = 1 \) it has been proven directly. Let us fix some \( k > 1 \) and assume that for all \( j = 0, 1, \ldots, k \) we have already shown
that $T_j(x)$ are polynomials of degree $j$. Then, the expression on the right-hand side of (3.59), and as such, $T_{k+1}(x)$, is a polynomial of degree $k+1$.

The polynomials $T_k(x)$ have been first introduced and studied by Chebyshev. We provide here the formulae for a first few Chebyshev polynomials, along with their graphs, see Figure 3.3:

$$
T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x.
$$

![Chebyshev polynomials](image)

**FIGURE 3.3:** Chebyshev polynomials

Next, by substituting $\varphi = \arccos x$ into the right-hand side of formula (3.56), we can recast it as a function of $x$, thus obtaining:

$$
Q_n(\cos \varphi, \sin \varphi, F) \equiv P_n(x, f),
$$

where

$$
P_n(x, f) = \sum_{k=0}^{n} a_k T_k(x),
$$

and

$$
a_0 = \frac{1}{n+1} \sum_{m=0}^{n} f_m = \frac{1}{n+1} \sum_{m=0}^{n} f_m T_0(x_m),
$$

$$
a_k = \frac{2}{n+1} \sum_{m=0}^{n} f_m \cos k \varphi_m = \frac{2}{n+1} \sum_{m=0}^{n} f_m T_k(x_m).
$$
We can therefore conclude using formulae (3.60), (3.61) and Theorem 3.7, that \( P_n(x, f) \) is an algebraic polynomial of degree no greater than \( n \) that coincides with the given function values \( f(x_m) = f_m \) at the interpolation nodes \( x_m = \cos \phi_m \). In accordance with (3.55), these interpolation nodes can be defined by the formula [cf. formula (2.20) of Chapter 2]:

\[
x_m = \cos \phi_m = \cos \left( \frac{\pi (2m + 1)}{2(n+1)} \right), \quad m = 0, 1, \ldots, n.
\]

We are also showing these nodes schematically in Figure 3.4.

Note that the points \( \phi_m = \frac{\pi}{n+1} m + \frac{\pi}{2(n+1)}, \quad m = 0, 1, \ldots, n \), defined by formula (3.55) are actually zeros of the function \( \cos(n+1)\phi \). Accordingly, the points \( x_m = \cos \phi_m \) defined by formula (3.63) are roots of the Chebyshev polynomial \( T_{n+1}(x) = \cos(n+1)\phi \). In the literature, this particular choice of nodes for the Chebyshev grid, see Figure 3.4, is often referred to as the Chebyshev-Gauss or simply Gauss nodes (an alternative choice of nodes is discussed in Section 3.2.6).

In other words, the polynomial \( P_n(x, f) \) specified by formulae (3.61), (3.62) renders algebraic interpolation of the function \( f(x) \) based on its values \( f_m \) sampled at the roots \( x_m \), see (3.63), of the Chebyshev polynomial \( T_{n+1}(x) \).

### 3.2.4 Properties of Algebraic Interpolation with Roots of the Chebyshev Polynomial \( T_{n+1}(x) \) as Nodes

Equality \( Q_n(\cos \phi, \sin \phi, F) = P_n(x, f) \) implies that the properties of the trigonometric interpolating polynomial \( Q_n(\cos \phi, \sin \phi, F) \) established by Theorems 3.4 and 3.5 of Section 3.1 do carry over to the algebraic interpolating polynomial \( P_n(x, f) \) defined by formula (3.61). In particular, the Lebesgue constants \( L_n \) that characterize the sensitivity of the polynomial \( P_n(x, f) \) to the perturbations of \( f_m \), satisfy estimate (3.32) from Section 3.1:

\[
L_n \leq 2(n+1),
\]

and the interpolation error

\[
R_n(x) = f(x) - P_n(x, f)
\]

uniformly converges to zero as \( n \to \infty \) with the rate determined by the number of derivatives \( r + 1 \) that \( f(x) \) has [cf. formula (3.36) of Section 3.1]:

\[
\max_{-1 \leq x \leq 1} |R_n(x)| \leq \frac{\zeta_n}{n^{r+1/2}}, \quad \text{where} \quad \zeta_n = o(1), \quad n \to \infty.
\]

\( \zeta_n \)
In other words, similarly to the trigonometric interpolation, see Section 3.1.3, algebraic interpolation on the Chebyshev nodes does not get saturated by smoothness.

In contradistinction to the Chebyshev nodes (3.63), when a uniform grid is used for interpolation, the Lebesgue constants rapidly grow as \( n \) increases, see inequalities (2.18) of Chapter 2, and convergence of the interpolating polynomial to the function \( f(x) \) may break down even for infinitely smooth functions, see Section 2.1.5. These are precisely the considerations that make the algebraic interpolation of high degree inappropriate, and prompt the use of piece-wise polynomial or spline interpolation on uniform or arbitrary non-uniform grids (see Chapter 2).

**REMARK 3.3** Estimate (3.32) that we have proven in Theorem 3.4 of Section 3.1 can be substantially improved. In fact, the following equality holds [see the bibliography quoted in Section 3.2.7, and cf. formula (2.21) of Section 2.1, Chapter 2]:

\[
L_n = \frac{2}{\pi} \ln(n+1) + 1 - \theta_n, \quad \frac{1}{4} \leq \theta_n \leq 1.
\] (3.65)

Likewise, the estimate for the interpolation error can also be improved, and instead of (3.64) we will obtain:

\[
\max_{-1 \leq x \leq 1} |R_n(x)| = o\left(\frac{\ln n}{n^{r+1/2}}\right) \quad \text{as} \quad n \to \infty,
\] (3.66)

where we remind that \( r+1 \) is the maximum number of derivatives that the function \( f(x) \) has. Further improvements of estimate (3.66) can be obtained with the help of the Jackson inequality, see Section 3.2.7.

### 3.2.5 An Algorithm for Evaluating the Interpolating Polynomial

To obtain the coefficients \( a_k \) of the polynomial \( P_n(x, f) \) of (3.61) with the help of formulae (3.62), as well as to actually evaluate this polynomial itself at a given \( x \in [-1, 1] \), one needs to be able to compute the values of the polynomials \( T_k(x) \), \( k = 0, 1, 2, \ldots \), for \(-1 \leq x \leq 1\). We will show that it is appropriate to use formula (3.59) for this purpose. Both its and computational efficacy are apparent. We only need to demonstrate that the computations according to this formula are stable with respect to the round-off errors.

Consider a difference equation of the type:

\[
y_{k+1} = 2xy_k - y_{k-1},
\] (3.67)

where \( y_k \) is the unknown sequence parameterized by the integer quantity \( k \). We will be looking for a solution of equation (3.67) in the form \( y_k = q^k \), where \( q \) is a fixed
Substituting the latter expression into the difference equation (3.67), we obtain the following algebraic equation for $q$:

$$q^2 - 2xq + 1 = 0.$$ 

It is often called the characteristic equation, and it has two roots:

$$q_{1,2} = x \pm \sqrt{x^2 - 1}.$$ 

Due to the linearity of equation (3.67), its general solution can be written in the form

$$y_k = c_1 q_1^k + c_2 q_2^k,$$  \hspace{1cm} (3.68)

where $c_1$ and $c_2$ are arbitrary constants. Let us choose these constants $c_1$ and $c_2$ so that to satisfy the following conditions:

$$y_0 = T_0(x) = 1, \hspace{1cm} y_1 = T_1(x) = x,$$

or equivalently,

$$c_1 + c_2 = 1, \hspace{1cm} c_1 q_1 + c_2 q_2 = x.$$ 

This implies $c_1 = c_2 = 1/2$, and then formula (3.68) yields the following solution of equation (3.67):

$$T_k(x) = \frac{1}{2} \left( x + \sqrt{x^2 - 1} \right)^k + \frac{1}{2} \left( x - \sqrt{x^2 - 1} \right)^k.$$  \hspace{1cm} (3.69)

According to formula (3.59), $T_k(x)$ defined by (3.69) shall be interpreted for a given fixed $x$ as a (discrete) function of $k$ that solves equation (3.67).

Next, note that when $|x| < 1$ the roots $q_{1,2} = x \pm \sqrt{x^2 - 1}$ of the characteristic equation are complex conjugate and have unit moduli. Consequently, the quantities $q_1^k$ and $q_2^k$ will remain equal to one by their absolute value as $k$ increases. An error committed for some $k = k_0$ would cause a perturbation in the values of $c_1$ and $c_2$ that enter into formula (3.68) for $k > k_0$. However, due to the equalities $|q_1^k| = |q_2^k| = 1$, $k = 1, 2, \ldots$, this error will not get amplified as $k$ increases. This implies numerical stability of the computations according to formula (3.59) for $|x| < 1$.

### 3.2.6 Algebraic Interpolation with Extrema of the Chebyshev Polynomial $T_n(x)$ as Nodes

To interpolate the function $F(\varphi) = f(\cos \varphi)$, let us now use the nodes:

$$\tilde{q}_m = \frac{\pi m}{n}, \hspace{1cm} m = 0, 1, \ldots, 2n - 1.$$
In accordance with Theorem 3.6 of Section 3.1, and the discussion on page 74 that follows this theorem, we obtain the trigonometric interpolating polynomial:

\[
\tilde{Q}_n(\cos \varphi, \sin \varphi, F) = \sum_{k=0}^{n} \tilde{a}_k \cos k\varphi,
\]

\[
\tilde{a}_0 = \frac{1}{2n} (f_0 + f_n) + \frac{1}{n} \sum_{m=1}^{n-1} f_m, \quad \tilde{a}_n = \frac{1}{2n} (f_0 + (-1)^n f_n),
\]

\[
\tilde{a}_k = \frac{1}{n} (f_0 + (-1)^k f_n) + \frac{2}{n} \sum_{m=1}^{n-1} f_m \cos k\varphi_m, \quad k = 1, 2, \ldots, n - 1.
\]

Changing the variable to \( x = \cos \varphi \) and denoting \( \tilde{Q}_n(\cos \varphi, \sin \varphi, F) = \tilde{P}_n(x, f) \), we have:

\[
\tilde{P}_n(x, f) = \sum_{k=0}^{n} \tilde{a}_k T_k(x),
\]

\[
\tilde{a}_0 = \frac{1}{2n} (f_0 + f_n) + \frac{1}{n} \sum_{m=1}^{n-1} f_m, \quad \tilde{a}_n = \frac{1}{2n} (f_0 + (-1)^n f_n),
\]

\[
\tilde{a}_k = \frac{1}{n} (f_0 + (-1)^k f_n) + \frac{2}{n} \sum_{m=1}^{n-1} f_m T_k(\tilde{x}_m), \quad k = 1, 2, \ldots, n - 1.
\]

Similarly to the polynomial \( P_n(x, f) \) of (3.61), the algebraic interpolating polynomial \( \tilde{P}_n(x, f) \) built on the grid:

\[
\tilde{x}_m = \cos \tilde{q}_m = \cos \frac{\pi m}{n}, \quad m = 0, 1, \ldots, n,
\]

also inherits the two foremost advantageous properties from the trigonometric interpolating polynomial \( \tilde{Q}_n(\cos \varphi, \sin \varphi, F) \). They are the slow growth of the Lebesgue constants as \( n \) increases (that translates into the numerical stability with respect to the perturbations of \( f_m \)), as well as the absence of saturation, i.e., convergence with the rate that automatically takes into account the smoothness of the function \( f(x) \).

Finally, we notice that the Chebyshev polynomial \( T_n(x) \) reaches its extreme values on the interval \(-1 \leq x \leq 1\) precisely at the interpolation nodes \( \tilde{x}_m \) of (3.70): \( T_n(\tilde{x}_m) = \cos \pi m = (-1)^m, \ m = 0, 1, \ldots, n \). In the literature, the grid nodes \( \tilde{x}_m \) of (3.70) are known as the Chebyshev-Gauss-Lobatto nodes or simply the Gauss-Lobatto nodes.

### 3.2.7 More on the Lebesgue Constants and Convergence of Interpolants

In this section, we discuss the problem of interpolation from the general perspective of approximation of functions by polynomials. Our considerations, in a substantially abridged form, follow those of [LG95], see also [Bab86]. We quote many of the fundamental results without a proof (the theorems of Jackson, Weierstrass, Faber-Bernstein, and Bernstein). The justification of these results, along with a broader and more comprehensive account of the subject, can
be found in the literature on the classical theory of approximation, see, e.g., [Jac94, Ber52, Ber54, Ach92, Nat64, Nat65a, Nat65b, Lor86, Che66, Riv74]. In the numerical analysis literature, some of these issues are addressed in [Wen66]. In these books, the reader will also find references to research articles. The material of this section is more advanced, and can be skipped during the first reading.

The meaning of Lebesgue’s constants $L_n$ introduced in Chapter 2 as minimum numbers that for each $n$ guarantee the estimate [see formula (2.17)]:

$$
\max_{a \leq x \leq b} |P_n(x, \delta f)| \leq L_n \max_j |\delta f(x_j)|
$$

is basically that of an operator norm. Indeed, interpolation by means of the polynomial $P_n(x, f)$ can be interpreted as a linear operator that maps the finite-dimensional space of vectors $[f_0, f_1, \ldots, f_n]$ into the space $C[a, b]$ of all continuous functions $f(x)$ defined on $[a, b]$. The space $C[a, b]$ is equipped with the maximum norm $\|f\| = \max_{a \leq x \leq b} |f(x)|$. Likewise, the space of vectors $\vec{f} = [f_0, f_1, \ldots, f_n]$ can also be equipped with a maximum norm, but discrete rather than continuous: $\|\vec{f}\| = \max_{0 \leq j \leq n} |f_j|$. Then, $L_n$ of (2.17) appears to be the induced norm of the foregoing linear operator:

$$
L_n = \sup_{\|\vec{f}\|=1} \|P_n(x, f)\|.
$$

(3.71)

However, for subsequent analysis it will be more convenient to use a slightly different definition of $L_n$ — as norm of an operator that would rather map $C[a, b]$ onto itself.

**DEFINITION 3.1** The operator $P_n = P_n(x_0, x_1, \ldots, x_n) : C[a, b] \rightarrow C[a, b]$ takes a function $f \in C[a, b]$, samples its values at a given set of nodes $\{x_0, x_1, \ldots, x_n\} \in [a, b]$ thus creating the table $\{f_0, f_1, \ldots, f_n\}$, and subsequently builds the polynomial $P_n(x, f, x_0, x_1, \ldots, x_n) \in C[a, b]$.

**LEMMA 3.1**
The operator $P_n$ introduced by Definition 3.1 is linear and continuous.

**PROOF** The linearity of $P_n$ is obvious. To show the continuity, we use the Lagrange formula (2.1) of Section 2.1, Chapter 2, and obtain:

$$
|P_n[f](x)| = |P_n(x, f, x_0, x_1, \ldots, x_n)| \leq \sum_{k=0}^n |f_k||l_k(x)| \leq \|f\| \sum_{k=0}^n |l_k(x)|.
$$

Next, we introduce a new quantity

$$
\lambda_n \overset{\text{def}}{=} \sup_{[a, b]} \sum_{k=0}^n |l_k(x)| = \max_{[a, b]} \sum_{k=0}^n |l_k(x)|,
$$

(3.72)
where the second equality in (3.72) holds because \([a, b]\) is a compact set, and 
\[
\sum_{k=0}^{n} |l_k(x)|
\]
is a continuous function. Then, clearly,
\[
|\mathcal{P}_n[f](x)| \leq \lambda_n \|f\|.
\]
Consequently, \(\mathcal{P}_n\) is a bounded operator, \(\mathcal{P}_n : C[a, b] \to C[a, b]\), and therefore, it is continuous. Moreover, \(\|\mathcal{P}_n\| \leq \lambda_n\).

**Definition 3.2** The norm of the operator \(\mathcal{P}_n\) introduced by Definition 3.1 is called the Lebesgue constant of the polynomial interpolation based on the nodes \(x_0, x_1, \ldots, x_n\):
\[
L_n = \|\mathcal{P}_n\|. \tag{3.73}
\]
Recall that the operator norm on the right-hand side of formula (3.73) is given by:
\[
\|\mathcal{P}_n\| = \sup_{\|f\|=1} \|\mathcal{P}_n[f](x)\| = \sup_{\|f\|=1} \|P_n(x, f)\|. \tag{3.74}
\]
We have therefore formulated two alternative definitions of the Lebesgue constants — by means of formula (3.71) and by means of formulae (3.73), (3.74). We will now prove that these definitions are, in fact, equivalent. In other words, we will show that the right-hand side of formula (3.74) coincides with the right-hand side of formula (3.71). The difference between these right-hand sides is that in (3.74) the smallest upper bound is taken across the unit sphere in the space \(C[a, b]\) that has infinite dimension, whereas in (3.71) it is taken across the unit sphere in the \(n + 1\)-dimensional space of vectors \(\vec{f} = [f_0, f_1, \ldots, f_n]\).

**Lemma 3.2** The Lebesgue constant defined by formulae (3.73), (3.74) is the same as the Lebesgue constant defined by formula (3.71).

**Proof** For every vector \(\vec{f} = [f_0, f_1, \ldots, f_n]\), consider a piece-wise linear function defined as:
\[
f(x) = f_{j+1} \frac{x-x_j}{x_{j+1}-x_j} + f_j \frac{x_{j+1}-x}{x_{j+1}-x_j} \quad \text{for} \quad x \in [x_j, x_{j+1}], \quad j = 0, 1, \ldots, n-1.
\]
Clearly, \(f(x) \in C[a, b]\), and also if \(\|\vec{f}\| = 1\) then \(\|f\| = 1\). In other words, every unit vector \(\vec{f} = [f_0, f_1, \ldots, f_n]\) gives rise to a continuous (piece-wise linear) function that belongs to the unit sphere in \(C[a, b]\). Therefore, one can say that the smallest upper bound on the right-hand side of (3.74) is taken across a wider set than that on the right-hand side of (3.71). Consequently,
\[
\sup_{\|f\|=1} \|P_n(x, f)\| \geq \sup_{\|f\|=1} \|P_n(x, f)\|. \tag{3.75}
\]
On the other hand, let \( f(x) \in C[a, b] \) be a particular function that realizes the smallest upper bound on the right-hand side of (3.74). By construction, \( \|f\| = 1 \). Let us sample the values of \( f(x) \) at the nodes \( x_0, x_1, \ldots, x_n \). This yields the table \( \{f_0, f_1, \ldots, f_n\} \), or equivalently, the vector \( \vec{f} = [f_0, f_1, \ldots, f_n] \). Assume that \( \|\vec{f}\| < 1 \). Then, denote \( \alpha = \|\vec{f}\|^{-1} > 1 \) and stretch the vector \( \vec{f} \): \( \vec{f} \mapsto \alpha \vec{f} = [\alpha f_0, \alpha f_1, \ldots, \alpha f_n] \) so that \( \|\alpha \vec{f}\| = 1 \). As the interpolation by means of the polynomials \( P_n \) is a linear operator, we obviously have \( P_n(x, \alpha f) = \alpha P_n(x, f) \), and consequently, \( \|P_n(x, \alpha f)\| > \|P_n(x, f)\| \). We have therefore found a unit vector \( \alpha \vec{f} \), for which the norm of the corresponding interpolating polynomial will be greater than the left-hand side of (3.75). The contradiction proves that the two definitions of the Lebesgue constants are indeed equivalent. \( \square \)

**Lemma 3.3**

The Lebesgue constant of (3.73), (3.74) is equal to

\[
L_n = \lambda_n,
\]

where the quantity \( \lambda_n \) is defined by formula (3.72).

**Proof** When proving Lemma 3.1, we have seen that \( \|\mathcal{P}_n\| \leq \lambda_n \). We therefore need to show that \( \lambda_n \leq \|\mathcal{P}_n\| \).

As has been mentioned, the function \( \psi(x) \equiv \sum\limits_{k=0}^{n} |l_k(x)| \) is continuous on \( [a, b] \). Consequently, \( \exists x^* \in [a, b] : \psi(x^*) = \lambda_n \). Let us now consider a function \( f_0 \in C[a, b] \) such that \( f_0(x_k) = \text{sign} l_k(x^*) \), \( k = 0, 1, 2, \ldots, n \), and also \( \|f_0\| = 1 \). For this function we have:

\[
|\mathcal{P}_n[f_0](x^*)| = \left| \sum\limits_{k=0}^{n} f_0(x_k) l_k(x^*) \right| = \left| \sum\limits_{k=0}^{n} (\text{sign} l_k(x^*)) l_k(x^*) \right| = \sum\limits_{k=0}^{n} |l_k(x^*)| = \psi(x^*) = \lambda_n.
\]

On the other hand,

\[
|\mathcal{P}_n[f_0](x^*)| \leq \|\mathcal{P}_n[f_0]\| \leq \|\mathcal{P}_n\| \cdot \|f_0\| = \|\mathcal{P}_n\|,
\]

which implies \( \lambda_n \leq \|\mathcal{P}_n\| \). It only remains to construct a specific example of \( f_0 \in C[a, b] \). This can be done easily by taking \( f_0(x) \) as a piece-wise linear function with the values \( \text{sign} l_k(x^*) \) at the points \( x_k, k = 0, 1, 2, \ldots, n \).

We can therefore conclude that

\[
L_n = \max\limits_{a \leq x \leq b} \sum\limits_{k=0}^{n} |l_k(x)|.
\]

We have used a somewhat weaker form of this result in Section 2.1.4 of Chapter 2.

The Lebesgue constants of Definition 3.2 play a fundamental role when studying the convergence of interpolating polynomials. To actually see that, we will first need to introduce another key new concept and formulate some important results.
**DEFINITION 3.3**  The quantity

\[
\epsilon(f, P_n) = \min_{P_n(x)} \max_{a \leq x \leq b} |P_n(x) - f(x)|
\]  

(3.77)

is called the best approximation of a given function \(f(x)\) by polynomials of degree no greater than \(n\) on the interval \([a, b]\).

Note that the minimum in formula (3.77) is taken with respect to all algebraic polynomials of degree no greater than \(n\) on the interval \([a, b]\), not only the interpolating polynomials. In other words, the polynomials in (3.77) do not, generally speaking, have to coincide with \(f(x)\) at any given point of \([a, b]\). It is possible to show existence of a particular polynomial that realizes the best approximation (3.77). In most cases, however, this polynomial is difficult to obtain constructively. In general, polynomials of the best approximation can only be built using sophisticated iterative algorithms of non-smooth optimization. On the other hand, their theoretical properties are well studied. Perhaps the most fundamental property is given by

**THEOREM 3.8 (Jackson inequality)**

Let \(f = f(x)\) be defined on the interval \([a, b]\), and let it have there \(r - 1\) continuous derivatives so that \(f^{(r-1)}(x)\) is Lipshitz-continuous:

\[
\forall x_1, x_2 \in [a, b] : |f^{(r-1)}(x_1) - f^{(r-1)}(x_2)| \leq M |x_1 - x_2|, \quad M > 0.
\]

Then, for any \(n \geq r\) the following inequality holds:

\[
\epsilon(f, P_n) < C_r \left( \frac{b - a}{2} \right)^r \frac{M}{n^r},
\]

(3.78)

where \(C_r = \left( \frac{\pi r}{2} \right)^r \frac{1}{r!}\) are universal constants that depend neither on \(f\), nor on \(n\), nor on \(M\).

The Jackson inequality [Jac94] reinforces, for sufficiently smooth functions, the result of the following classical theorem established in real analysis:

**THEOREM 3.9 (Weierstrass)**

Let \(f \in C[a, b]\). Then, for any \(\varepsilon > 0\) there is an algebraic polynomial \(P_{\varepsilon}(x)\) such that \(\forall x \in [a, b] : |f(x) - P_{\varepsilon}(x)| \leq \varepsilon\).

The proof of Theorem 3.9 is based on periodization of \(f\) that preserves its continuity (the period should obviously be larger than \([a, b]\)) and then on the approximation by partial sums of the Taylor series that converges uniformly. The Weierstrass theorem implies that for \(f \in C[a, b]\) the best approximation defined by (3.77) converges to zero: \(\epsilon(f, P_n) \rightarrow 0\) when \(n \rightarrow \infty\). This is basically as much as one can tell regarding the behavior of \(\epsilon(f, P_n)\) if nothing else is known about \(f(x)\) except that it
is continuous. On the other hand, the Jackson inequality specifies the rate of decay for the best approximation as a particular inverse power of \( n \), see formula (3.78), provided that \( f(x) \) is smooth.

Let us also note that the value of \( \frac{\pi}{2} \) that enters the expression for \( C_r = \left( \frac{\pi}{2} \right)^r \frac{1}{n} \) in the Jackson inequality (3.78) may, in fact, be replaced by smaller values:

\[
K_0 = 1, \quad K_1 = \frac{\pi}{2}, \quad K_2 = \frac{\pi^2}{8}, \quad K_3 = \frac{\pi^3}{24}, \quad K_4 = \frac{5\pi^4}{384}, \ldots
\]

known as the Favard constants. The Favard constants can be obtained explicitly for all \( r = 0, 1, 2, \ldots \), and it is possible to show that all \( K_r < \frac{\pi}{2} \). The key consideration regarding the Favard constants is that substituting them into (3.78) makes this inequality sharp.

The main result that connects the properties of the best approximation (Definition 3.3) and the quality of interpolation by means of algebraic polynomials is given by the following

**THEOREM 3.10 (Lebesgue inequality)**

Let \( f \in C[a, b] \) and let \( \{x_0, x_1, \ldots, x_n\} \) be an arbitrary set of distinct interpolation nodes on \([a, b] \). Then,

\[
\epsilon(f, P_n) \leq \| f - \mathcal{P}_n[f] \| \leq (L_n + 1) \epsilon(f, P_n). \tag{3.79}
\]

Note that according to Definition 3.1, the operator \( \mathcal{P}_n \) in formula (3.79) generally speaking depends on the choice of the interpolation nodes.

**PROOF** It is obvious that we only need to prove the second inequality in (3.79), i.e., the upper bound. Consider an arbitrary polynomial \( Q(x) \in \{P_n(x)\} \) of degree no greater than \( n \). As the algebraic interpolating polynomial is unique (Theorem 2.1 of Chapter 2), we obviously have \( \mathcal{P}_n[Q] = Q \). Next,

\[
\| f - \mathcal{P}_n[f] \| = \| f - Q + \mathcal{P}_n[Q] - \mathcal{P}_n[f] \|
\leq \| f - Q \| + \| \mathcal{P}_n[Q] - \mathcal{P}_n[f] \| = \| f - Q \| + \| \mathcal{P}_n[Q - f] \|
\leq \| f - Q \| + \| \mathcal{P}_n \| \| f - Q \| = (1 + L_n) \| f - Q \|.
\]

Let us now introduce \( \delta > 0 \) and denote by \( Q_\delta(x) \in \{P_n(x)\} \) a polynomial for which \( \| f - Q_\delta \| < \epsilon(f, P_n) + \delta \). Then,

\[
\| f - \mathcal{P}_n[f] \| \leq (1 + L_n) \| f - Q_\delta \| < (1 + L_n)(\epsilon(f, P_n) + \delta).
\]

Finally, by taking the limit \( \delta \rightarrow 0 \), we obtain the desired inequality (3.79). \( \blacksquare \)

The Lebesgue inequality (3.79) essentially provides an upper bound for the interpolation error \( \| f - \mathcal{P}_n[f] \| \) in terms of a product of the best approximation (3.77) times the Lebesgue constant (3.73). Often, this estimate allows one to judge the
convergence of algebraic interpolating polynomials as \( n \) increases. It is therefore clear that the behavior of the Lebesgue constants is of central importance for the convergence study.

**THEOREM 3.11 (Faber-Bernstein)**

For any choice of interpolation nodes \( x_0, x_1, \ldots, x_n \) on the interval \([a, b]\), the following inequality holds:

\[
L_n > \frac{1}{8\sqrt{\pi}} \ln(n + 1). \quad (3.80)
\]

Theorem 3.11 shows that the Lebesgue constants always grow as the grid dimension \( n \) increases. As such, the best one can generally hope for is to be able to place the interpolation nodes in such a way that this growth will be optimal, i.e., logarithmic.

As far as the problem of interpolation may be concerned, if, for example, nothing is known about the function \( f \in C[a, b] \) except that it is continuous, then nothing can basically be said about the behavior of the error beyond the estimate given by the Lebesgue inequality (3.79). The Weierstrass theorem (Theorem 3.9) indicates that \( \epsilon(f, P_n) \to 0 \) as \( n \to \infty \), and the Faber-Bernstein theorem (Theorem 3.11) says that \( L_n \to \infty \) as \( n \to \infty \). We therefore have the uncertainty \( 0 \cdot \infty \) on the right-hand side of the Lebesgue inequality; and the behavior of this right-hand side is determined by which of the two processes dominates — the decay of the best approximations or the growth of the Lebesgue constants. In particular, if \( \lim_{n \to \infty} L_n \epsilon(f, P_n) = 0 \), then the interpolating polynomials uniformly converge to \( f(x) \).

If the function \( f(x) \) is sufficiently smooth (as formulated in Theorem 3.8), then combining the Lebesgue inequality (3.79) and the Jackson inequality (3.78) we obtain the following error estimate: \(^3\)

\[
\|f - \mathcal{P}_n[f]\| < (L_n + 1)C_r \left( \frac{b - a}{2} \right)^r \frac{M}{n^r}, \quad (3.81)
\]

which implies that the convergence rate (if there is convergence) will depend on the behavior of \( L_n \) when \( n \) increases. If the interpolation grid is uniform (equidistant nodes), then the Lebesgue constants grow exponentially as \( n \) increases, see inequalities (2.18) of Section 2.1, Chapter 2. In this case, the limit (as \( n \to \infty \)) on the right-hand side of (3.81) is infinite for any finite value of \( r \). This does not necessarily mean that the sequence of interpolating polynomials \( \mathcal{P}_n[f](x) \) diverges, because inequality (3.81) only provides an upper bound for the error. It does mean though that in this case convergence of the interpolating polynomials simply cannot be judged using the arguments based on the inequalities of Lebesgue and Jackson.

On the other hand, for the Chebyshev interpolation grid (3.63) the following theorem asserts that the asymptotic behavior of the Lebesgue constants is optimal:

\(^3\)Note that in estimate (3.81) the function \( f(x) \) is assumed to have a maximum of \( r - 1 \) derivatives, whereas in the previous error estimate (3.66) we have used a slightly different notation and the function was assumed to have a maximum of \( r + 1 \) derivatives.
THEOREM 3.12 (Bernstein)

Let the interpolation nodes $x_0, x_1, \ldots, x_n$ on the interval $[-1, 1]$ be given by roots of the Chebyshev polynomial $T_{n+1}(x)$. Then,

$$L_n < 8 + \frac{4}{\pi} \ln(n+1).$$  \hspace{1cm} (3.82)

Therefore, according to (3.81) and (3.82), if the derivative $f^{(r-1)}(x)$ of the function $f(x)$ is Lipshitz-continuous, then the sequence of algebraic interpolating polynomials built on the Chebyshev nodes converges uniformly to $f(x)$ as $n \to \infty$ with the rate $O\left(n^{-r} \ln(n+1)\right)$.

We thus see that the type of interpolation grid may indeed have a drastic effect on convergence, which corroborates our previous observations. For the Bernstein example $f(x) = |x|$, $-1 \leq x \leq 1$ (Section 2.1.5 of Chapter 2), the sequence of interpolating polynomials constructed on a uniform grid diverges. On the Chebyshev grid we have seen experimentally that it converges. Now, using estimates (3.81) and (3.82) we can say that the rate of this convergence is at least $O\left(n^{-1} \ln(n+1)\right)$.

To conclude, let us also note that strictly speaking the behavior of $L_n$ on the Cheby-
shev grid is only asymptotically optimal rather than optimal, because the constants in the lower bound (3.80) and in the upper bound (3.82) are different. Better values of these constants than those guaranteed by the Bernstein theorem (Theorem 3.12) have been obtained more recently, see inequality (3.65). However, there is still a gap between (3.80) and (3.65).

REMARK 3.4 Formula (3.77) that introduces the best approximation according to Definition 3.3 can obviously be recast as

$$\varepsilon(f, P_n) = \min_{P_n(x)} \|P_n(x) - f(x)\|_C,$$

where the norm on the right-hand side is taken in the sense of the space $C[a,b]$. In general, the notion of the best approximation admits a much broader interpretation, when both the norm (currently, $\| \cdot \|_C$) and the class of approximating functions (currently, polynomials $P_n(x)$) may be different. In fact, one can consider the problem of approximating a given element of the linear space by linear combinations of a pre-defined set of elements from the same space in the sense of a selected norm.

For example, consider the space $L^2[a,b]$ of all square integrable functions $f(x), x \in [a,b]$, equipped with the norm:

$$\|f\|_2 = \left[ \int_a^b f^2(x)dx \right]^\frac{1}{2}.$$

This space is known to be a Hilbert space. Let us take an arbitrary $f \in L^2[a,b]$ and consider a set of all trigonometric polynomials $Q_n(x)$ of type (3.6), where
$L = b - a$ is the length of the interval. Similarly to the algebraic polynomials $P_n(x)$ employed in Definition 3.3, the trigonometric polynomials $Q_n(x)$ do not have to be interpolating polynomials. Then, it is known that the best approximation in the sense of $L_2$:

$$
\varepsilon_2(f, Q_n) = \min_{Q_n(x)} \| f(x) - Q_n(x) \|_2
$$

is, in fact, realized by the partial sum $S_n(x)$, see formula (3.37), of the Fourier series for $f(x)$ with the coefficients defined by (3.39). An upper bound for the actual magnitude of the $L_2$ best approximation is then given by estimate (3.40) for the remainder $\delta S_n(x)$ of the series, see formula (3.38):

$$
\varepsilon_2(f, Q_n) \leq \frac{\zeta_n}{n^{r+\frac{1}{2}}}, \quad \text{where} \quad \zeta_n = o(1), \quad n \to \infty,
$$

where $r + 1$ is, again, the maximum smoothness of $f(x)$. Having identified what the best approximation in the sense of $L_2$ is, we can easily see now that both the Lebesgue inequality (Theorem 3.10) and the error estimate for trigonometric interpolation (Theorem 3.5) are, in fact, justified using the same argument. It employs uniqueness of the corresponding interpolating polynomial, the estimate for the best approximation, and the estimate of sensitivity to perturbations given by the Lebesgue constants.

Exercises

1. Let the function $f = f(x)$ be defined on an arbitrary interval $[a, b]$, rather than on $[-1, 1]$. Construct the Chebyshev interpolation nodes for $[a, b]$, and write down the interpolating polynomials $P_n(x, f)$ and $\tilde{P}_n(x, f)$ similar to those obtained in Sections 3.2.3 and 3.2.6, respectively.

2. For the function $f(x) = \frac{1}{x^2 + 1/4}$, $-1 \leq x \leq 1$, construct the algebraic interpolating polynomial $P_n(x, f)$ using roots of the Chebyshev polynomial $T_{n+1}(x)$, see (3.63), as interpolation nodes. Plot the graphs of $f(x)$ and $P_n(x, f)$ for $n = 5, 10, 20, 30,$ and $40$. Do the same for the interpolating polynomial $P_n(x, f)$ built on the equally spaced nodes $x_k = -1 + 2k/n$, $k = 0, 1, 2, \ldots, n$ (Runge example of Section 2.1.5, Chapter 2). Explain the observable qualitative difference between the two interpolation techniques.

3. Introduce the normalized Chebyshev polynomial $\hat{T}_n(x)$ of degree $n$ by setting $\hat{T}_n(x) = 2^{1-n}T_n(x)$.

   a) Show that the coefficient in front of $x^n$ in the polynomial $\hat{T}_n(x)$ is equal to one.

   b) Show that the deviation $\max_{-1 \leq x \leq 1} |\hat{T}_n(x)|$ of the polynomial $\hat{T}_n(x)$ from zero on the interval $-1 \leq x \leq 1$ is equal to $2^{1-n}$. 


c) Show that among all the polynomials of degree $n$ with the leading coefficient (in front of $x^n$) equal to one, the normalized Chebyshev polynomial $T_n(x)$ has the smallest deviation from zero on the interval $-1 \leq x \leq 1$.

d) How shall one choose the interpolation nodes $t_0, t_1, \ldots, t_n$ on the interval $[-1, 1]$, so that the polynomial $(t - t_0)(t - t_1)\ldots(t - t_n)$, which is a part of the formula for the interpolation error (2.23), Chapter 2, would have the smallest possible deviation from zero on the interval $[-1, 1]$?

4. Find a set of interpolation nodes for an even $2\pi$-periodic function $F(\phi)$, see formula (3.54), for which the Lebesgue constants would coincide with the Lebesgue constants of algebraic interpolation on equidistant nodes.
Chapter 4

Evaluation of Definite Integrals. Quadratures

4.1 The Trapezoid Rule and Simpson’s Formula

To be written
Part II

Systems of Algebraic Equations
Chapter 5

Systems of Linear Algebraic Equations. Direct Methods

5.1 Different Forms of Consistent Linear Systems

5.1.1 Canonical Form of a Linear System

5.1.2 Operator Form

5.1.3 Finite-Difference Dirichlet Problem for the Poisson Equation

Let $D$ be a domain of square shape on the Cartesian plane: $D = \{(x,y)|0 < x < 1, \ 0 < y < 1\}$. Assume that a Dirichlet problem for the Poisson equation:

$$
-\Delta u \equiv -\left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right] = f(x,y), \quad (x,y) \in D,
$$

$$
u|_{\partial D} = 0, \quad (x,y) \in \partial D,
$$

needs to be solved on the domain $D$. In formula (5.1), $\partial D$ denotes the boundary of $D$, and $f = f(x,y)$ is a given right-hand side.

To solve problem (5.1) numerically, we introduce a positive integer $M$, specify $h = M^{-1}$, and construct a uniform Cartesian grid on the square $D$:

$$(x_m, y_m) = (m_1h, m_2h), \quad m_1, m_2 = 0, 1, \ldots, M. \quad (5.2)$$

Instead of the continuous solution $u = u(x,y)$ to the original problem (5.1), we will rather be looking for its trace, or projection, $[u]_h$ onto the grid (5.2). To compute the values $u_{m_1,m_2}$ of this projection approximately, we replace the derivatives in the Poisson equation by the second-order difference quotients at every interior node of the grid, and thus obtain the following system of difference equations:

$$
-\Delta^{(h)} u_{m_1,m_2} = - \left(\frac{u_{m_1+1,m_2} - 2u_{m_1,m_2} + u_{m_1-1,m_2}}{h^2} + \frac{u_{m_1,m_2+1} - 2u_{m_1,m_2} + u_{m_1,m_2-1}}{h^2}\right) = f_{m_1,m_2}, \quad m_1, m_2 = 1, 2, \ldots, M - 1. \quad (5.3)
$$
The right-hand side of each equation (5.3) is defined as \( f_{m_1,m_2} \overset{\text{def}}{=} f(m_1 h, m_2 h) \).

Recall, the difference equation (5.3) is only valid at the interior nodes of the grid (5.2), see Figure 5.1(a). For every fixed pair of \( m_1 \) and \( m_2 \) inside the square \( D \), it connects the values of the solution at the five points that are shown in Figure 5.1(b). These points are said to form the stencil of the finite-difference scheme (5.3).

At the boundary points of the grid (5.2), see Figure 5.1(a), we obviously cannot write the finite-difference equation (5.3). Instead, according to (5.1), we specify the homogeneous Dirichlet boundary conditions:

\[
    u_{m_1,M,m_2} = 0, \quad \text{for } m_1 = 0, M \quad \& \quad m_2 = 0, M. \tag{5.4}
\]

Substituting expressions (5.4) into (5.3), we obtain a system of \((M-1)^2\) linear algebraic equations with respect to as many unknowns: \( u_{m_1,m_2}, m_1, m_2 = 1, 2, \ldots, M-1 \). As far as the motivation and justification for constructing the difference counterpart to problem (5.1) in the particular form (5.3), (5.4), we postpone it till Chapter 10.

5.2 Norms

5.3 Gaussian Elimination and Its Tri-Diagonal Version
Chapter 6

Iterative Methods for Solving Linear Systems

6.1 First-Order Linear Stationary Methods
Part III

The Method of Finite Differences for the Numerical Solution of Differential Equations
Perhaps, the most commonly used approach to the numerical solution of both ordinary and partial differential equations is based on the method of finite differences. This method has a large number of modifications and “flavors” tuned for different specific problems and/or classes of problems. In all versions of this method, one first introduces a special collection of discrete nodes in the domain of the solution. This set of nodes is called the finite-difference grid; and the solution of interest is subsequently sought for on the grid. For the unknown grid function one constructs a system of algebraic equations that it is supposed to satisfy. The solution of this system is interpreted as an approximate discrete table of values for the original continuous solution.

A most straightforward approach to obtaining the foregoing system of algebraic equations, which is called the finite-difference scheme, is based on approximately replacing the derivatives contained in the original differential equation, as well as in the initial and/or boundary conditions, by the appropriate difference quotients. This is where the method of finite differences draws its name from.
Chapter 9

Numerical Solution of Ordinary Differential Equations

Hereafter, we will assume that a given ordinary differential equation, or a system of ordinary differential equations, is to be solved on the interval \( a \leq x \leq b \) of the independent variable \( x \), subject to the additional initial or boundary conditions.

9.1 Examples of Finite-Difference Schemes. Convergence

For both initial and boundary value problems to be analyzed in this chapter, we will be using the same generic notation:

\[ Lu = f. \]  

(9.1)

Several simple examples of such problems are provided below.

An initial value (Cauchy) problem for the first order ordinary differential equation:

\[ \frac{du}{dx} + \frac{x}{1 + u^2} = \cos x, \quad 0 \leq x \leq 1, \quad u(0) = 3. \]  

(9.2)

An initial value problem for the second-order ordinary differential equation:

\[ \frac{d^2 u}{dx^2} = (1+x^2)u + \sqrt{x+1}, \quad 0 \leq x \leq 1, \quad u(0) = 2, \quad \left. \frac{du}{dx} \right|_{x=0} = 1. \]  

(9.3)

A boundary value problem for the second-order ordinary differential equation, with the boundary conditions given at the endpoints \( x = 0 \) and \( x = 1 \) of the interval:

\[ \frac{d^2 u}{dx^2} = (1+x^2)u + \sqrt{x+1}, \quad 0 \leq x \leq 1, \quad u(0) = 2, \quad u(1) = 1. \]  

(9.4)
An initial value problem for the system of two first-order ordinary differential equations:

\[
\frac{dv}{dx} + xvw = x^2 - 3x + 1, \quad 0 \leq x \leq 1, \quad v(0) = 1,
\]

\[
\frac{dw}{dx} + \frac{v + w}{1 + x^2} = \cos^2 x, \quad 0 \leq x \leq 1, \quad w(0) = -3.
\]

Solution of this system is a two-dimensional vector function \( u = [v, w]^T \) with the components that satisfy the differential equations of (9.5).

Note that in all these examples — (9.2) through (9.5) — it is only for definiteness and not for any other reason that we are considering each problem on the specific interval \( D = \{0 \leq x \leq 1\} \) rather than on a general interval \( \{a \leq x \leq b\} \).

### 9.1.1 Examples of Difference Schemes

We will always assume that the solution \( u = u(x) \) of problem (9.1) does exist on the interval \( 0 \leq x \leq 1 \). To compute this solution using the method of finite differences, one first needs to specify a finite set of (distinct) points on the interval \( \bar{D} \) that will be called the grid and denoted \( D_h \). Subsequently, instead of the continuous solution \( u(x) \) of problem (9.1), a discrete table \( [u]_h \) of its values on the grid \( D_h \) is to be considered as the unknown function of the problem. It is assumed that the grid \( D_h \) depends on the positive parameter \( h \) that can be arbitrarily small. This parameter is referred to as the grid size, so that the smaller the \( h \) the finer the grid. For example, one can consider \( h = 1/N \), where \( N \) is a positive integer, and compose the grid \( D_h \) of the \( N + 1 \) equally spaced points (also called nodes): \( x_0 = 0, x_1 = h, ..., x_n = nh, ..., x_N = 1 \), with the distance between each two neighboring points being equal to \( h \). Finite-difference grids of this type are called uniform. The value of the unknown grid function \( [u]_h \) at the node \( x_n \) of \( D_h \) will be denoted by \( u_n \).

Consider, for example, problem (9.2). To approximately compute the trace \( [u]_h \) of its solution \( u(x) \) on the grid \( D_h \), one can use the system of equations:

\[
\frac{u_{n+1} - u_n}{h} + \frac{x_n}{1 + u_n^2} = \cos x_n, \quad n = 0, 1, \ldots, N - 1, \quad u_0 = 3,
\]

obtained via replacing the values of the derivative \( du/dx \) at the nodes \( x_n, n = 0, 1, \ldots, N - 1 \), by difference quotients according to the approximate formula:

\[
\frac{du}{dx} \approx \frac{u(x + h) - u(x)}{h}.
\]

Solution \( u_h = \{u_0^{(h)}, u_1^{(h)}, \ldots, u_N^{(h)}\} \) of system (9.6) is defined on the same grid \( D_h \) as the unknown grid function \( [u]_h \) is. The values \( u_1^{(h)}, u_2^{(h)}, \ldots, u_N^{(h)} \) of this solution at the nodes \( x_1, x_2, \ldots, x_N \) are computed consecutively for \( n = 0, 1, \ldots, N - 1 \) from system (9.6), while \( u_0^{(h)} = 3 \) is determined from the initial condition. For simplicity, we are
omitting the superscript “(h)” for all $u_n^{(h)}$ in equations (9.6). Henceforth, we will keep omitting this superscript on similar occasions, whenever it may not cause any confusion. The system of equations (9.6) provides an example of a finite-difference scheme (or difference scheme, or simply scheme) for the initial value problem (9.2).

In the case of problem (9.4), to obtain a grid function $u^{(h)}$ that would approximate the desired exact table of values $[u]_h$, one can employ the following scheme:

$$
\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - (1 + x_n^2)u_n = \sqrt{x_n + 1}, \quad n = 1, \ldots, N - 1,
$$

$$
\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - (1 + x_n^2)u_n = \sqrt{x_n + 1}, \quad n = 1, \ldots, N - 1,
$$

$$
u_0 = 2, \quad u_N = 1.
$$

(9.7)

This scheme arises when the values of the second derivative $d^2u/dx^2$ from the differential equation of (9.4) are replaced at the nodes $x_1, x_2, \ldots, x_{N-1}$ of the grid $D_h$ according to the approximate formula:

$$
\frac{d^2u}{dx^2} \approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}.
$$

To actually compute the solution $u^{(h)}$ of problem (9.7), one can use the algorithm of tri-diagonal elimination described in Section 5.3 of Chapter 5.

Let us also write down a scheme that could be used for approximately computing the solution of problem (9.5):

$$
\frac{v_{n+1} - v_n}{h} + x_n u_n w_n = x_n^2 - 3x_n + 1, \quad n = 0, 1, \ldots, N - 1, \quad v_0 = 1,
$$

$$
\frac{w_{n+1} - w_n}{h} + \frac{v_n + w_n}{1 + x_n^2} = \cos^2 x_n, \quad n = 0, 1, \ldots, N - 1, \quad w_0 = -3.
$$

(9.8)

Here $u_0^{(h)} = \left[ \begin{array}{c} v_0 \\ w_0 \end{array} \right] = \left[ \begin{array}{c} 1 \\ -3 \end{array} \right]$ is given. When $n = 0$, one can find $u_1^{(h)} = \left[ \begin{array}{c} v_1 \\ w_1 \end{array} \right]$ from equations (9.8). In general, if $u_n^{(h)} = \left[ \begin{array}{c} v_n \\ w_n \end{array} \right]$ is known, one can obviously compute the discrete solution at the next grid node: $u_{n+1}^{(h)} = \left[ \begin{array}{c} v_{n+1} \\ w_{n+1} \end{array} \right]$. In doing so, the solution can be obtained on the entire grid $D_h$ consecutively, one node after another. This numerical procedure is known as marching.

In the previous examples the grid $D_h$ was taken uniform. This, however, is not a necessity. Instead of having $N + 1$ equally spaced nodes with $h = 1/N$ on the interval $[0, 1]$, we could have chosen $D_h$ as follows:

$$
x_0 = 0, \quad x_1 = x_0 + h_0, \quad x_2 = x_1 + h_1, \ldots, \quad x_{n+1} = x_n + h_n, \ldots, \quad x_N = 1,
$$

(9.9)

where the grid sizes $h_n > 0$, $n = 0, 1, \ldots, N - 1$, do not have to be equal. We do require, though, that $\max h_n \rightarrow 0$ when $N \rightarrow \infty$. The grids of type (9.9) are called nonuniform. By appropriately choosing the node locations of a nonuniform grid for a given fixed $N$, one can make the desired table $[u]_h$ more detailed in those areas.
where the solution $u(x)$ undergoes stronger variations. Those areas can sometimes be identified ahead of time from the physics considerations, or from the preliminary low-accuracy computations. Besides, the information about the behavior of $u(x)$ also becomes available in the course of computation. In particular, the rate of change of $u(x)$ is readily obtained when the solution is marched: $u_1, u_2, \ldots, u_n, \ldots$. This information can be taken into account every time when choosing the location of the next grid node $x_{n+1}$. Altogether, the strategies of building special nonuniform grids that would accommodate particular features of the computed solution are known as grid adaptation.

We have provided several examples that illustrate the important concept of the grid and that of the grid function we are seeking — the trace, or projection, $[u]_h$ of the continuous solution $u(x)$ onto the grid. More examples will be given as we introduce and study properties of finite-difference schemes.

Note that we are interested in computing the grid function $u^{(h)}$ because when the grid is refined, i.e., when $h \to 0$, we expect it to furnish an increasingly accurate table of values for the unknown solution $u(x)$ and as such, provide its more detailed description altogether. Once the discrete table of values is known, we could use interpolation and reconstruct the continuous solution $u(x)$ everywhere on the domain $D$ with the accuracy that would improve as $h \to 0$. It is clear that for a fixed dimension of the grid $D_h$, i.e., fixed number of nodes $N + 1$, the actual reconstruction accuracy will depend on the location of those nodes, as well as on the additional information available about the continuous solution, such as estimates for its derivatives.

One can find a considerably more detailed discussion of the reconstruction-related issues in Chapters 2 and 3 devoted to the interpolation theory. In this chapter, we will rather restrict ourselves by the foregoing brief comments only, and will focus on the computation of the table $u^{(h)}$ instead. In this context, it will be natural to say that problem (9.1) has been solved exactly already if the grid function $[u]_h$ is obtained.\(^1\) In general, however, we cannot compute the table $[u]_h$ exactly either. Instead, we compute its approximation $u^{(h)}$, which is expected to converge to $[u]_h$ as the grid is refined. It is for computing $u^{(h)}$ that one can employ the finite-difference equations.

### 9.1.2 Convergent Difference Schemes

Our primary objective hereafter will be to study the techniques for the construction and analysis of convergent schemes for ordinary differential equations. However, before we can even approach this agenda, we will need to attach an accurate quantitative interpretation to the very notion of convergence $u^{(h)} \to [u]_h$, which is the key property that we will require of the schemes we build. To do so, let us consider a normed linear space $U_h$ of all the functions defined on the grid $D_h$. As those grid functions are essentially vectors of dimension $N + 1$, the definitions and examples that pertain to vector norms and related concepts, see Section 5.2 of Chapter 5, do apply here. Very briefly, we remind that the norm $\|u^{(h)}\|_{U_h}$ of the grid function

\(^1\)As opposed to the actual continuous solution $u(x)$ that requires reconstruction form the grid.
Numerical Solution of Ordinary Differential Equations

\( u^{(h)} \in U_h \) is a non-negative real number that basically quantifies the extent of its deviation from the identical zero on the grid.

There are many different ways one can define an appropriate norm for the grid function. For example, the least upper bound of all its absolute values at the grid nodes is a norm that is called the maximum norm:

\[
\|u^{(h)}\|_{U_h} = \sup_{n} |u_n| = \max_{n} |u_n|.
\]  

(9.10)

If \( u^{(h)} \) is a vector function, say, \( u^{(h)} = \begin{bmatrix} v_n \\ u_n \end{bmatrix}, n = 0, 1, \ldots, N \), as in scheme (9.8), then the norm similar to (9.10) can be defined as the least upper bound of the absolute values of both functions \( v_n \) and \( w_n \) on the corresponding grid.

If the grid \( D_h \) is uniform, i.e., if the grid functions \( u^{(h)} \in U_h \) are defined at the equally spaced nodes \( x_n = nh, h > 0, n = 0, 1, \ldots, N \), then the following Euclidean norm is often used:

\[
\|u^{(h)}\|_{U_h} = \left[ h \sum_{n=0}^{N} u_n^2 \right]^{1/2}.
\]

This norm is analogous to the continuous \( L_2 \) norm for the square integrable functions:

\[
\|u(x)\| = \left[ \int_{0}^{1} u^2(x)dx \right]^{1/2}
\]

Henceforth, we will always assume (for simplicity) that the maximum norm (9.10) is used, unless explicitly stated otherwise.

Having introduced the normed space \( U_h \), we can now quantify the discrepancy between any two functions in this space. Let \( a^{(h)} \in U_h \) and \( b^{(h)} \in U_h \) be a pair of arbitrary functions defined on the grid \( D_h \). The measure of their deviation from one another is naturally given by the norm of their difference:

\[
\|a^{(h)} - b^{(h)}\|_{U_h}.
\]

The latter quantification finally enables us to give an accurate definition of convergence for finite-difference schemes.

Let us denote by [cf. formula (9.1)]:

\[
L_h u^{(h)} = f^{(h)}
\]

(9.11)

the system of algebraic equations to be used for approximately computing the solution of problem (9.1). In other words, solution of system (9.11) is supposed to yield an approximation to \([u]_h\), which is the discrete table of values for the continuous solution \( u(x) \) of problem (9.1). Specific examples of the systems of type (9.11) are given by the difference schemes (9.6), (9.7), and (9.8) built for problems (9.2), (9.4), and (9.5), respectively. To recast scheme (9.6) in the general form (9.11) on a uniform grid \( x_n = nh, n = 0, 1, \ldots, N \), one can define the operator \( L_h \) and the right-hand
side \( f(h) \) as follows:

\[
L_h u^{(h)} = \begin{cases} 
\frac{u_{n+1} - u_n}{h} + \frac{nh}{1 + u_n^2}, & n = 0, 1, \ldots, N - 1, \\
u_0, & \end{cases}
\]

\[
f^{(h)} = \begin{cases} 
\cos(nh), & n = 0, 1, \ldots, N - 1, \\
3, & \end{cases}
\]

Scheme (9.7) can be written in the form (9.11) once we set:

\[
L_h u^{(h)} = \begin{cases} 
\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - [1 + (nh)^2]u_n, & n = 1, 2, \ldots, N - 1, \\
u_0, \\
u_N, & \end{cases}
\]

\[
f^{(h)} = \begin{cases} 
\sqrt{1 + nh}, & n = 1, 2, \ldots, N - 1, \\
2, & \end{cases}
\]

Scheme (9.8) for a system of differential equations transforms into (9.11) if:

\[
L_h [v^{(h)} w^{(h)}] = \begin{cases} 
\frac{v_{n+1} - v_n}{h} + (nh)v_n w_n, & n = 0, 1, \ldots, N - 1, \\
\frac{w_{n+1} - w_n}{h} + \frac{v_n + w_n}{1 + (nh)^2}, & n = 0, 1, \ldots, N - 1, \\
v_0, \\
w_0, & \end{cases}
\]

\[
f^{(h)} = \begin{cases} 
(nh)^2 - 3nh + 1, & n = 0, 1, \ldots, N - 1, \\
\cos^2(nh), & n = 0, 1, \ldots, N - 1, \\
1, & \end{cases}
\]

As we can see, in general system (9.11) depends on the grid size \( h \). Therefore, it shall be written down for all those values of \( h \), for which we introduce the grid \( D_h \) and the exact solution \([u]_h\). Consequently, the discrete problem (9.11) is to be interpreted as the entire family of algebraic systems parameterized by the quantity \( h \), rather than as one single system of equations. Henceforth, we will assume that for each sufficiently small \( h \) the corresponding system (9.11) has a unique solution \( u^{(h)} \in U_h \).

We will say that solution \( u^{(h)} \) to the finite-difference problem (9.11) converges to the solution \( u(x) \) of the original differential problem (9.1) as the grid is refined, if

\[
||[u]_h - u^{(h)}||_{U_h} \rightarrow 0, \text{ when } h \rightarrow 0, \tag{9.12}
\]

where \([u]_h = \{u(x_n)\}_{n = 0, 1, \ldots, N}\). If, in addition, \( k > 0 \) happens to be the largest integer such that the following inequality holds for all sufficiently small \( h \):

\[
||[u]_h - u^{(h)}||_{U_h} \leq ch^k, \quad c = \text{const}, \tag{9.13}
\]
where $c$ does not depend on $h$, then we say that the convergence rate of scheme (9.11) is $O(h^k)$, or alternatively, that the error of the approximate solution, which is the quantity (rather, function) under the norm on the left-hand side of (9.13), has order $k$ with respect to the grid size $h$. Sometimes we would also say that the order of convergence is equal to $k > 0$.

Convergence is a fundamental requirement that one imposes on the difference scheme (9.11) so that to make it an appropriate tool for the numerical solution of the original differential (also referred to as continuous) problem (9.1). If convergence does take place, then the solution $[u]_h$ can be computed using scheme (9.11) with any initially prescribed accuracy by simply choosing a sufficiently small grid size $h$.

Having rigorously defined the concept of convergence, we have now arrived at the central question of how to construct a convergent difference scheme (9.11) for computing the solution of problem (9.1). The examples of Section 9.1.1 suggest a simple initial idea for building the schemes: One should first generate a grid and subsequently replace the derivatives in the governing equation(s) by appropriate difference quotients. However, for the same continuous problem (9.1) one can obviously obtain a large variety of schemes (9.11) by choosing different grids $D_h$ and different ways to approximate the derivatives by difference quotients. In doing so, it turns out that some of the schemes do converge, while the others do not.

### 9.1.3 Verification of Convergence for a Difference Scheme

Let us therefore reformulate our central question in a somewhat different way. Suppose that the finite-difference scheme $L_h u^{(h)} = f^{(h)}$ has already been constructed, and we expect that it could be convergent:

$$
\| [u]_h - u^{(h)} \|_{L_h} \to 0, \quad \text{when} \quad h \to 0.
$$

How can we actually check whether it really converges or not?

Assume that problem (9.11) has a unique solution $u^{(h)}$, and let us substitute the grid function $[u]_h$ instead of $u^{(h)}$ into the left-hand side of (9.11). If equality (9.11) were to hold exactly upon this substitution, then uniqueness would have implied $u^{(h)} = [u]_h$, which is ideal for convergence. Indeed, it would have meant that solution $u^{(h)}$ to the discrete problem $L_h [u]_h = f^{(h)}$ coincides with the grid function $[u]_h$ that is sought for and that we have agreed to interpret as the unknown exact solution.

However, most often one cannot construct system (9.11) so that the solution $[u]_h$ would satisfy it exactly. Instead, the substitution of $[u]_h$ into the left-hand side of (9.11) would typically generate a residual $\delta f^{(h)}$:

$$
L_h [u]_h = f^{(h)} + \delta f^{(h)}
$$

also known as the truncation error. If the truncation error $\delta f^{(h)}$ tends to zero as $h \to 0$, then we say that the finite-difference scheme $L_h u^{(h)} = f^{(h)}$ is consistent, or alternatively, that it approximates the differential problem $Lu = f$ on the solution $u(x)$ of the latter. This notion indeed makes sense because smaller residuals for
smaller grid sizes mean that the exact solution $[u]_h$ satisfies equation (9.11) with better and better accuracy as $h$ vanishes.

When the scheme is consistent, one can think that equation (9.14) for $[u]_h$ is obtained from equation (9.11) for $u^{(h)}$ by adding a small perturbation $\delta f^{(h)}$ to the right-hand side $f^{(h)}$ ($\delta f^{(h)}$ is small provided that $h$ is small). Consequently, if the solution $u^{(h)}$ of problem (9.11) happens to be only weakly sensitive to the perturbations of the right-hand side, or in other words, if small changes of the right-hand side $f^{(h)}$ may only induce small changes in the solution $u^{(h)}$, then the difference between the solution $u^{(h)}$ of problem (9.11) and the solution $[u]_h$ of problem (9.14) will be small. In other words, consistency $\delta f^{(h)} \rightarrow 0$ will imply convergence:

$$u^{(h)} \rightarrow [u]_h, \text{ as } h \rightarrow 0.$$  

The aforementioned weak sensitivity of the finite-difference solution $u^{(h)}$ to perturbations of $f^{(h)}$ can, in fact, be defined using rigorous terms. This definition will lead us to the fundamental notion of stability for finite-difference schemes.

Altogether, we can now outline our approach to verifying the convergence (9.12). We basically suggest to split this difficult task into two potentially simpler tasks. First, we would need to see whether the scheme (9.11) is consistent. Then, we will need to find out whether the scheme (9.11) is stable. This approach also indicates how one might actually construct convergent schemes for the numerical solution of problem (9.1). Namely, one would first need to obtain a consistent scheme, and then among many such schemes select those that would also be stable.

The foregoing general approach to the analysis of convergence obviously requires that both consistency and stability be defined rigorously, so that a theorem can eventually be proven on convergence as an implication of consistency and stability. The previous definitions of consistency and stability are, however, vague. As far as consistency, we need to be more specific on what the truncation error (residual) $\delta f^{(h)}$ is in the general case, and how to properly define its magnitude. As far as stability, we need to assign a precise meaning to the words “small changes of the right-hand side $f^{(h)}$ may only induce small changes in the solution $u^{(h)}$” of problem (9.11). We will devote two separate sections to the rigorous definitions of consistency and stability.

9.2 Approximation of Continuous Problem by a Difference Scheme. Consistency

In this section, we will accurately define the concept of consistency and explain thoroughly what it actually means when we say that the finite-difference scheme (9.11) approximates the original continuous problem (9.1) on its solution $u = u(x)$. 

9.2.1 Truncation Error $\delta f^{(h)}$

To do so, we will first need to delineate on what the truncation error $\delta f^{(h)}$ is, and how one can sensibly introduce its magnitude. We remind that according to formula (9.14), $\delta f^{(h)}$ is the residual that arises when the exact solution $[u]^h$ is substituted into the left-hand side of (9.11). The decay of the magnitude of $\delta f^{(h)}$ as $h \to 0$ is precisely what we term as consistency of the finite-difference scheme (9.11).

We begin with analyzing the example of a difference scheme for solving the following second-order initial value problem:

$$\frac{d^2 u}{dx^2} + a(x) \frac{du}{dx} + b(x) u = \cos x, \quad 0 \leq x \leq 1,$$

(9.15)

The grid $D_h$ on the interval $[0, 1]$ will be uniform: $x_n = nh, n = 0, 1, \ldots, N, h = 1/N$. To obtain the scheme, we approximately replace all the derivatives in relations (9.15) by difference quotients:

$$\frac{d^2 u}{dx^2} \approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2},$$

(9.16a)

$$\frac{du}{dx} \approx \frac{u(x+h) - u(x-h)}{2h},$$

(9.16b)

$$\left. \frac{du}{dx} \right|_{x=0} \approx \frac{u(h) - u(0)}{h}.$$  

(9.16c)

Substituting expressions (9.16a)–(9.16c) into (9.15), we arrive at the system of equations that can be used for the approximate computation of $[u]^h$:

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + a(x_n) \frac{u_{n+1} - u_{n-1}}{2h} + b(x_n) u_n = \cos x_n,$$

$$n = 1, 2, \ldots, N-1,$$

$$u_0 = 1, \quad \frac{u_1 - u_0}{h} = 2.$$

(9.17)

Scheme (9.17) can be easily converted to form (9.11) if we denote:

$$L_hu^{(h)} = \begin{cases} 
\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + a(x_n) \frac{u_{n+1} - u_{n-1}}{2h} + b(x_n) u_n, \\
u_0, \\
u_1 - u_0 \\
\frac{1}{h}, \\
\cos x_n, \\
1, \\
2.
\end{cases}$$

(9.18)

To estimate the magnitude of the truncation error $\delta f^{(h)}$ that arises when the grid function $[u]^h$ is substituted into the left-hand side of (9.11), we will need to “sharpen” the
approximate equalities (9.16a)–(9.16c). A common approach to doing so is to evaluate the quantities on the right-hand sides of (9.16a)–(9.16c) using Taylor formulae.

For the central difference approximation of the second derivative (9.16a), we will need four terms of the Taylor expansion of \( u \) at the point \( x \):

\[
\begin{align*}
    u(x + h) &= u(x) + hu'(x) + \frac{h^2}{2!}u''(x) + \frac{h^3}{3!}u'''(x) + \frac{h^4}{4!}u''''(\xi_1), \\
    u(x - h) &= u(x) - hu'(x) + \frac{h^2}{2!}u''(x) - \frac{h^3}{3!}u'''(x) + \frac{h^4}{4!}u''''(\xi_2).
\end{align*}
\]

(9.19a)

For the central difference approximation of the first derivative (9.16b), it will be sufficient to use three terms of the expansion:

\[
\begin{align*}
    u(x + h) &= u(x) + hu'(x) + \frac{h^2}{2!}u''(x) + \frac{h^3}{3!}u'''(x), \\
    u(x - h) &= u(x) - hu'(x) + \frac{h^2}{2!}u''(x) - \frac{h^3}{3!}u'''(x),
\end{align*}
\]

(9.19b)

and for the forward difference approximation of the first derivative (9.16c) one would only need two terms of the Taylor formula (to be used at \( x = 0 \)):

\[
    u(x + h) = u(x) + hu'(x) + \frac{h^2}{2!}u''(\xi). \tag{9.19c}
\]

Note that the last term in each of the expressions (9.19a)–(9.19c) is the error of the Taylor formula in the so-called Lagrange form, where \( \xi_1, \xi_2, \xi_5 \in [x, x + h] \) and \( \xi_2, \xi_4 \in [x - h, x] \). Substituting expressions (9.19a), (9.19b), and (9.19c) into formulae (9.16a), (9.16c), and (9.16e), respectively, we obtain:

\[
\begin{align*}
    \frac{u(x + h) - 2u(x) + u(x - h)}{h^2} &= u''(x) + \frac{h^2}{24} \left( u''''(\xi_1) + u''''(\xi_2) \right), \\
    \frac{u(x + h) - u(x) - h}{2h} &= u'(x) + \frac{h^2}{12} \left( u'''(\xi_3) + u'''(\xi_4) \right), \\
    \frac{u(x + h) - u(x)}{h} \bigg|_{x=0} &= u'(x) \bigg|_{x=0} + \frac{h}{2} u''(\xi). \tag{9.20c}
\end{align*}
\]

9.2.2 Evaluation of the Truncation Error \( \delta_f^{(h)} \)

Let us now assume that the solution \( u = u(x) \) of problem (9.15) has bounded derivatives up to the fourth order everywhere on the interval \( 0 \leq x \leq 1 \). Then, according to formulae (9.20a)–(9.20c) one can write:

\[
\begin{align*}
    \frac{u(x + h) - 2u(x) + u(x - h)}{h^2} + a(x) \frac{u(x + h) - u(x) - h}{2h} + b(x)u \\
    &= \frac{d^2u}{dx^2} + a(x) \frac{du}{dx} + b(x)u + \left[ \frac{u(\xi_1)}{24} + \frac{u(\xi_2)}{12} + a(x) \frac{u''(\xi_3)}{2} + b(x)u''(\xi)
\right],
\end{align*}
\]

(9.20d)
where \( \xi_1, \xi_2, \xi_3, \xi_4 \in [x-h, x+h] \). Therefore, expression

\[
L_h[u]_h = \begin{cases} 
\frac{u(x_n+h) - 2u(x_n) + u(x_n-h)}{h^2} + a(x_n) \frac{u(x_n+h) - u(x_n-h)}{2h} + b(x_n)u(x_n), & n = 1, 2, \ldots, N-1, \\
u(0), & \\
u(h) - u(0) \end{cases} 
\]

see formula (9.18), can be re-written as follows:

\[
L_h[u]_h = \begin{cases} 
\cos x_n + h^2 \left[ \frac{u^{(4)}(\xi_1) + u^{(4)}(\xi_2)}{24} + a(x_n) \frac{u'''(\xi_3) + u'''(\xi_4)}{12} \right], & n = 1, 2, \ldots, N-1, \\
1 + 0, & \\
2 + h \frac{u''(\xi_5)}{2}, & 
\end{cases}
\]

where \( \xi_1, \xi_2, \xi_3, \xi_4 \in [x_n-h, x_n+h], n = 1, 2, \ldots, N-1, \) and \( \xi_5 \in [0, h] \). Alternatively, we can write [cf. formula (9.14)]:

\[
L_h[u]_h = f^{(h)} + \delta f^{(h)},
\]

where \( f^{(h)} \) is also defined in (9.18), and

\[
\delta f^{(h)} = \begin{cases} 
\frac{h^2}{24} \left[ \frac{u^{(4)}(\xi_1) + u^{(4)}(\xi_2)}{24} + a(x_n) \frac{u'''(\xi_3) + u'''(\xi_4)}{12} \right], & n = 1, 2, \ldots, N-1, \\
0, & \\
h \frac{u''(\xi_5)}{2}, & 
\end{cases}
\]

To quantify the truncation error, i.e., introduce its magnitude, it will be convenient to assume that \( f^{(h)} \) and \( \delta f^{(h)} \) belong to a linear normed space \( F_h \) that consists of the following elements:

\[
g^{(h)} = \begin{cases} 
q_n, & n = 1, 2, \ldots, N-1, \\
\psi_0, & \\
\psi_1, & 
\end{cases}
\]

where \( q_1, q_2, \ldots, q_{N-1} \), and \( \psi_0, \psi_1 \) is an arbitrary ordered system of numbers. One can think of \( g^{(h)} \) given by formula (9.22) as of a combination of the grid function \( q_n, n = 1, 2, \ldots, N-1, \) and an ordered pair of numbers \((\psi_0, \psi_1)\). Addition of the elements from the space \( F_h \), as well as their multiplication by scalars, are performed
component-wise. It is clear that the foregoing $F_h$ is a linear (vector) space of dimension $N + 1$. It can be supplied with a variety of different norms. If the norm in $F_h$ is introduced as the maximum absolute value of all the components of $g^{(h)}$:

$$\|g^{(h)}\|_{F_h} = \max \left\{ |\psi_0|, |\psi_1|, \max_n |\varphi_n| \right\},$$

then according to (9.21) we obviously have for all sufficiently small grid sizes $h$:

$$\|\delta f^{(h)}\| \leq ch, \quad (9.23)$$

where $c$ is a constant that may, generally speaking, depend on $u(x)$, but should not depend on $h$. Inequality (9.23) guarantees that the truncation error $\delta f^{(h)}$ does vanish when $h \to 0$, and that the scheme (9.17) has first-order accuracy.

If the difference equation (9.17) that we have considered as an example is represented in the form $L_hu^{(h)} = f^{(h)}$, see formula (9.18), then one can interpret $L_h$ as an operator. Given a grid function $v^{(h)} = \{v_n\}$, $n = 0, 1, \ldots, n$, that belongs to the linear space $U_h$ of the functions defined on the grid $D_h$, the operator $L_h$ maps it onto some element $g^{(h)} \in F_h$ of type (9.22) according to the following formula:

$$L_hv^{(h)} = \begin{cases} \frac{v_{n+1} - 2v_n + v_{n-1}}{h^2} + a(x_n)\frac{v_{n+1} - v_{n-1}}{2h} + b(x_n)v_n, & n = 1, 2, \ldots, N - 1, \\ v_0, \\ v_1 - v_0 \end{cases}$$

As for the general finite-difference problem (9.11), it can always be said to comprise a number of individual (scalar) equations. We will be considering the right-hand sides of all these equations as components of the vector $f^{(h)}$ that belongs to some linear normed space $F_h$. In this framework, $L_h$ of (9.11) becomes an operator that maps any given grid function $u^{(h)} \in U_h$ onto some element $f^{(h)} \in F_h$, i.e., $L_h: U_h \to F_h$. Then, expression $L_h[u]_h$ stands for the result of application of the operator $L_h$ to the function $[u]_h$ that is an element of $U_h$; as such, $L_h[u]_h \in F_h$. Consequently, $\delta f^{(h)} \equiv L_h[u]_h - f^{(h)} \in F_h$ as a difference between two elements of the space $F_h$. The magnitude of the residual $\delta f^{(h)}$, i.e., the magnitude of the truncation error, is given by the norm $\|\delta f^{(h)}\|_{F_h}$.

### 9.2.3 Accuracy of Order $h^k$

**DEFINITION 9.1** We will say that the finite-difference scheme $L_hu^{(h)} = f^{(h)}$ approximates the continuous problem $Lu = f$ on its solution $u = u(x)$ if the truncation error vanishes when the grid is refined, i.e., $\|\delta f^{(h)}\|_{F_h} \to 0$ as $h \to 0$. A scheme that does approximate the original continuous problem is called consistent. If, in addition, $k > 0$ happens to be the largest integer that
guarantees the following estimate:
\[ \| \delta f^{(h)} \|_{F_h} \leq c_1 h^k, \quad c_1 = \text{const}, \]
where \( c_1 \) does not depend on the grid size \( h \), then we say that the scheme has order of accuracy \( O(h^k) \) or that its accuracy is of order \( k \) with respect to \( h \).\(^2\)

Note that in Definition 9.1 the function \( u = u(x) \) is considered a solution of problem \( Lu = f \). This assumption can provide useful information about \( u \), e.g., bounds for its derivatives, that can subsequently be exploited when constructing the scheme, as well as when verifying its consistency. This is the reason for incorporating a solution of problem (9.1) into Definition 9.1. Let us emphasize, however, that the notion of approximation of the problem \( Lu = f \) by the scheme \( L_h u^{(h)} = f^{(h)} \) does not rely on the equality \( Lu = f \) for the function \( u \). The central requirement of Definition 9.1 is only the decay of the quantity \( \| L_h[u]_h - f^{(h)} \|_{F_h} \) when the grid size \( h \) vanishes.\(^3\) Therefore, if there is a function \( u = u(x) \) that meets this requirement, then we could simply say that the truncation error of the scheme \( L_h u^{(h)} = f^{(h)} \) has some order \( k > 0 \) with respect to \( h \) on a given function \( u \), without going into detail regarding the origin of this function. In this context, it may often be instrumental to use a slightly different concept of approximation — that of the differential operator \( L \) by the difference operator \( L_h \) (see Section 10.2.2 of Chapter 10 for more detail).

Namely, we will say that the finite-difference operator \( L_h \) approximates the differential operator \( L \) on some function \( u = u(x) \) if \( \| L_h[u]_h - [L u]_h \|_{F_h} \to 0 \) as \( h \to 0 \). In the previous expression, \([u]_h\) denotes the trace of the continuous function \( u(x) \) on the grid as before, and likewise, \([L u]_h\) denotes the trace of the continuous function \( Lu \) on the grid. For example, equality (9.20a) can be interpreted as approximation of the differential operator \( Lu \equiv u'' \) by the central difference on the left-hand side of (9.20a), with the accuracy \( O(h^2) \). In so doing, \( u(x) \) may be any function with the bounded fourth derivative. Consequently, the approximation can be considered on the class of all functions \( u \) with bounded derivatives up to the order four, rather than on one single function \( u \). Similarly, equality (9.20b) can be interpreted as approximation of the differential operator \( Lu \equiv u' \) by the central difference on the left-hand side of (9.20b), with the accuracy \( O(h^2) \), and this approximation holds on the class of all functions \( u = u(x) \) that have bounded third derivatives.

9.2.4 Examples

Example 1

According to formula (9.21) and estimate (9.23), the finite-difference scheme (9.17) is consistent and has accuracy \( O(h) \), i.e., it approximates problem (9.15) with the first order with respect to \( h \). In fact, scheme (9.17) can be easily improved so that

\(^2\)Sometimes also referred to as the order of approximation of the continuous problem by the scheme.

\(^3\)Obviously, the equality \( Lu = f \) may still need to be employed for the analysis of consistency for a particular scheme, as done, e.g., in Section 9.2.2.
it would gain the second order of accuracy. To achieve that, let us first notice that every component of the vector $\delta f^{(h)}$ except for the last one, see (9.21), decays with the rate $O(h^2)$ when $h$ decreases, and the second to last component is even equal to zero exactly. It is only the last component of the vector $\delta f^{(h)}$ that displays a slower rate of decay, namely, $O(h)$. This last component is the residual generated by substituting $u = u(x)$ into the last equation $(u_1 - u_0)/h = 2$ of system (9.17). Fortunately, the slower rate of decay for this component, which hampers the overall accuracy of the scheme, can be easily sped up.

Using the Taylor formula with the Lagrange error form, we can write:

$$u(h) - u(0) = u'(0) + \frac{h}{2} u''(0) + \frac{h^2}{6} u'''(\xi), \quad \text{where } 0 \leq \xi \leq h.$$  

At the same time, the original differential equation, along with its initial conditions, see (9.15), yield:

$$u''(0) = -a(0)u'(0) - b(0)u(0) + \cos 0 = -2a(0) - b(0) + 1.$$  

Therefore, if we replace the last equality of (9.17) with the following:

$$\frac{u_1 - u_0}{h} = 2 + \frac{h}{2} u''(0) = 2 - \frac{h}{2} [2a(0) + b(0) - 1], \quad (9.24)$$

then we obtain a new expression for the right-hand side $f^{(h)}$, instead of the one given in formula (9.18):

$$f^{(h)} = \begin{cases} \cos \alpha_n, & n = 1, 2, \ldots, N - 1, \\ 1, & n = N, \\ 2 - \frac{h}{2} [2a(0) + b(0) - 1]. \end{cases}$$

This, in turn, yields a new expression for the truncation error [cf. formula (9.21)]:

$$\delta f^{(h)} = \begin{cases} h^2 \left[ \frac{u^{(4)}(\xi_1) + u^{(4)}(\xi_2)}{24} + a(\alpha_n) \frac{u''(\xi_3) + u''(\xi_4)}{12} \right], & n = 1, 2, \ldots, N - 1, \\ 0, & n = N. \end{cases} \quad (9.25)$$

Formula (9.25) implies that under the previous assumption of boundedness of all the derivatives up to the order four, we have $\| \delta f^{(h)} \| \leq c h^2$, where the constant $c$ does not depend on $h$. Thus, the accuracy of the scheme becomes $O(h^2)$ instead of $O(h)$.

Let us emphasize that in order to obtain the new difference initial condition (9.24) not only have we used the original continuous boundary conditions of (9.15), but also the differential equation itself. It is possible to say that we have exploited the additional initial condition:

$$u''(x) + a(x)u'(x) + b(x)u(x) \big|_{x=0} = \cos x \big|_{x=0},$$

which is a direct implication of the given differential equation of (9.15) at $x = 0$. 

---

118  

*Theoretical Introduction to Numerical Analysis*
Example 2

Let
\[
L_h u^{(h)}(h) = \begin{cases} 
    \frac{u_{n+1} - u_{n-1}}{2h} + Au_n, & n = 1, 2, \ldots, N - 1, \\
    u_0, & n = 1, \\
    u_1, & n = 2, \ldots, N - 1,
\end{cases}
\]  
(9.26)

\[
f^{(h)} = \begin{cases} 
    1 + x_n^2, & n = 1, 2, \ldots, N - 1, \\
    b, & n = 1, \\
    b, & n = 2, \ldots, N - 1,
\end{cases}
\]

and let us determine the order of accuracy of the scheme:

\[
L_h u^{(h)} = f^{(h)}
\]  
(9.27)

on the solution \( u = u(x) \) of the initial value problem:

\[
\frac{du}{dx} + Au = 1 + x^2, \quad u(0) = b.
\]  
(9.28)

For the exact solution \([u]_h\) we obtain:

\[
L_h[u]_h = \begin{cases} 
    \frac{u(x_n + h) - u(x_n - h)}{2h} + Au(x_n), & n = 1, 2, \ldots, N - 1, \\
    u(0), & n = 1, \\
    u(h), & n = 2, \ldots, N - 1,
\end{cases}
\]

which, according to formula (9.20b), yields:

\[
L_h[u]_h = \begin{cases} 
    \frac{du(x_n)}{dx} + Au(x_n) + \frac{h^2}{12}(u''''(\xi_3) + u''''(\xi_4)), & n = 1, 2, \ldots, N - 1, \\
    u(0), & n = 1, \\
    u(0) + hu'(\xi_0), & n = 2, \ldots, N - 1,
\end{cases}
\]

where \(\xi_0 \in [0, h]\) and for each \(n\): \(\xi_3 \in [x_n, x_n + h]\) and \(\xi_4 \in [x_n - h, x_n]\). As \(u\) is a solution to (9.28), we have:

\[
\frac{du(x_n)}{dx} + Au(x_n) = 1 + x_n^2,
\]

and for the truncation error \(\delta f^{(h)}\) we obtain:

\[
\delta f^{(h)} = \begin{cases} 
    \frac{h^2}{12}(u''''(\xi_3) + u''''(\xi_4)), & n = 1, 2, \ldots, N - 1, \\
    0, & n = 1, \\
    hu'(\xi_0), & n = 2, \ldots, N - 1,
\end{cases}
\]
Consequently, scheme (9.26)–(9.27) altogether has the first order of accuracy with respect to \( h \). It is interesting to note, however, that similarly to the previous Example 1, different components of the truncation error have different orders with respect to the grid size. The system of difference equations:

\[
\frac{u_{n+1} - u_{n-1}}{2h} + Au_n = 1 + x_n^2, \quad n = 1, 2, \ldots, N - 1,
\]

is satisfied by the exact solution \([u]_h\) with the residual \( \frac{h^2}{12}(u''(\xi_3) + u''(\xi_4)) \) of order \( \mathcal{O}(h^2) \). The first initial condition \( u_0 = b \) is satisfied exactly, and only the second initial condition \( u_1 = b \) is satisfied by \([u]_h\) with the residual \( hu'(\xi_0) \) of order \( \mathcal{O}(h) \).

**Example 3**

Finally, let us illustrate the comments we have made right after the definition of consistency, see Definition 9.1, in Section 9.2.3. For simplicity, we will be considering problems with no initial or boundary conditions, defined on the grid: \( x_n = nh, \quad h > 0, n = 0, \pm 1, \pm 2, \ldots \). Let

\[
L_h u^{(h)} = \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + u_n.
\]

Formula (9.20a) immediately implies that \( L_h \) approximates the differential operator:

\[
Lu = u'' + u
\]

with the second order of accuracy on the class of functions with bounded fourth derivatives. Indeed, according to (9.20a) for any such function we can write:

\[
L_h[u]_h - [Lu]_h = \frac{h^2}{24} \left( u^{(4)}(\xi_1) + u^{(4)}(\xi_2) \right),
\]

and consequently, \( ||L_h[u]_h - [Lu]_h|| \leq ch^2 \).

Let us now take \( u(x) = \sin x \) and show that the homogeneous difference scheme \( L_h u^{(h)} = 0 \) is consistent on this function. Indeed,

\[
L_h[u]_h = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + u(x)
\]

\[
= u''(x) + \frac{h^2}{24} \left( u^{(4)}(\xi_1) + u^{(4)}(\xi_2) \right) + u(x)
\]

\[
= -\sin x + \frac{h^2}{24} \left( u^{(4)}(\xi_1) + u^{(4)}(\xi_2) \right) + \sin x
\]

\[
= \frac{h^2}{24} \left( u^{(4)}(\xi_1) + u^{(4)}(\xi_2) \right),
\]

which means that \( ||L_h[u]_h - 0|| \leq \text{const} \cdot h^2/12 \). Yet we note that in the previous argument we have used the consideration that \((\sin x)'' + \sin x = 0\), i.e., that \( u(x) = \sin x \).
is not just a function with bounded fourth derivatives, but a solution to the homogeneous differential equation \( Lu \equiv u'' + u = 0 \). In other words, while the definition of consistency per se does not explicitly require that \( u(x) \) be a solution to the underlying differential problem, this fact may still be needed when actually evaluating the magnitude of the truncation error.

However, for the specific example that we are currently analyzing, consistency can also be established independently. We have:

\[
L_h[u]_h = \frac{\sin(x + h) - 2\sin(x) + \sin(x - h)}{h^2} + \sin(x) = \frac{\sin x \cos h + \cos x \sin h - 2\sin(x) + \sin x \cos h - \cos x \sin h}{h^2} + \sin(x) = \sin x \left( \frac{2(\cos h - 1)}{h^2} + \sin x = \left( 1 - \frac{4}{h^2} \sin^2 \frac{h}{2} \right) \sin x. \right.
\]

Using the Taylor formula, one can easily show that for small grid sizes \( h \) the expression in the brackets above is \( \mathcal{O}(h^2) \), and as \( \sin x = u(x) \) is bounded, we, again, conclude that the scheme \( L_h u^{(h)} = 0 \) has second order of accuracy.

### 9.2.5 Replacement of Derivatives by Difference Quotients

In all our previous examples, in order to obtain a finite-difference scheme we would replace the derivatives in the original continuous problem (differential equation and initial conditions) by the appropriate difference quotients. This approach is quite universal, and for any continuous problem with a sufficiently smooth solution \( u(x) \) it enables construction of a scheme with any prescribed order of accuracy.

### 9.2.6 Other Approaches to Constructing Difference Schemes

However, the replacement of derivatives by difference quotients is by no means the only method of building the schemes, and often not the best one either. In Section 9.4, we describe a family of popular methods of approximation that lead to the widely used Runge-Kutta schemes. Here, we only provide some examples.

The simplest scheme:

\[
\frac{u_{n+1} - u_n}{h} - G(x_n, u_n) = 0, \quad n = 0, 1, \ldots, N - 1, \quad u_0 = a, \quad h = 1/N, \tag{9.29}
\]

known as the forward Euler scheme is consistent and has first-order accuracy with respect to \( h \) on the solutions of the differential problem:

\[
\frac{du}{dx} G(x, u) = 0, \quad 0 \leq x \leq 1, \quad u(0) = a. \tag{9.30}
\]

Solution of the finite-difference equation (9.29) can be found by marching, i.e., it can be computed consecutively one grid node after another using the formula:

\[
u_{n+1} = u_n + hG(x_n, u_n). \tag{9.31}
\]
This formula yields $u_{n+1}$ in the closed form once the values of $u_n$ at the previous nodes are available. That’s why the forward Euler scheme is called explicit.

Let us now take the same formula (9.31) as in the forward Euler method but use it only as a predictor at the preliminary stage of computation. At this stage we obtain the intermediate quantity $\bar{u}$:

$$\bar{u} = u_n + h G(x_n, u_n).$$

Then, the actual value of the finite-difference solution at the next grid node $u_{n+1}$ is obtained by applying the corrector at the final stage:

$$u_{n+1} = u_n + \frac{h}{2} [G(x_n, u_n) + G(x_{n+1}, \bar{u})].$$

The overall resulting predictor-corrector scheme:

$$\bar{u} = u_n + h G(x_n, u_n),$$

$$\frac{u_{n+1} - u_n}{h} = \frac{1}{2} [G(x_n, u_n) + G(x_{n+1}, \bar{u})],$$

$$u_0 = a,$$

(9.32)

can be shown to have second-order accuracy with respect to $h$ on the solutions of problem (9.30). It is also explicit, and is, in fact, a member of the Runge-Kutta family of methods, see Section 9.4. In the literature, scheme (9.32) is sometimes also referred to as the Heun scheme.

Let us additionally note that the corrector stage of scheme (9.32) can actually be considered as an independent single stage finite-difference method of its own:

$$\frac{u_{n+1} - u_n}{h} = \frac{1}{2} [G(x_n, u_n) + G(x_{n+1}, u_{n+1})],$$

$$u_0 = a.$$  

(9.33)

Scheme (9.33) is known as the Crank-Nicolson scheme, it approximates problem (9.30) with second order of accuracy. Solution of the finite-difference equation (9.33) can also be found by marching. However, unlike in the forward Euler method (9.29), equation (9.33) contains $u_{n+1}$ both on its left-hand side and on its right-hand side, as an argument of $G(x_{n+1}, u_{n+1})$. Therefore, to obtain $u_{n+1}$ for each $n = 0, 1, 2, \ldots$, one basically needs to solve an algebraic equation. This may require special techniques, such as the Newton method of Section ??? of Chapter ???, when the function $G$ happens to be nonlinear with respect to its argument $u$. Altogether, we see that for the Crank-Nicolson scheme (9.33) the value of $u_{n+1}$ is not immediately available in the closed form as a function of the previous $u_n$. That’s why the Crank-Nicolson method is called implicit.

Finally, perhaps the simplest implicit scheme for problem (9.30) is the so-called backward Euler scheme:

$$\frac{u_{n+1} - u_n}{h} - G(x_{n+1}, u_{n+1}) = 0, \quad u_0 = a,$$

(9.34)
that has first-order accuracy with respect to $h$.

Nowadays, a number of well established universal and efficient methods for solving the Cauchy problem for ordinary differential equations and systems are available as standard implementations in numerical software libraries. There is also a large number of specialized methods developed for solving particular (often, narrow) classes of problems that may present difficulties of a specific type.

Exercises

1. Verify that the forward Euler scheme (9.29) has first-order accuracy on a smooth solution $u = u(x)$ of problem (9.30).

2. Modify the second initial condition $u_1 = b$ in the scheme (9.26)-(9.27) so that to achieve the overall second-order accuracy.

3.* Verify that the predictor-corrector scheme (9.32) has second-order accuracy on a smooth solution $u = u(x)$ of problem (9.30).

   **Hint.** See the analysis in Section 9.4.1.

4. Verify that the Crank-Nicolson scheme (9.33) has second-order accuracy on a smooth solution $u = u(x)$ of problem (9.30).

5. Verify that the backward Euler scheme (9.34) has first-order accuracy on a smooth solution $u = u(x)$ of problem (9.30).

9.3 Stability of Finite-Difference Schemes

In the previous sections, we have constructed a number of consistent finite-difference schemes for ordinary differential equations. It is, in fact, possible to show that those schemes are also convergent, and that the convergence rate for each scheme coincides with its respective order of accuracy.

One can, however, build examples of consistent yet divergent (i.e., non-convergent) finite-difference schemes. For instance, it is easy to see that the scheme:

\[
\frac{4u_{n+1} - u_{n-1}}{2h} - \frac{3u_{n+1} - u_n}{h} + A u_n = 0, \quad n = 1, 2, \ldots, N - 1, \\
u_0 = b, \quad u_1 = be^{-Ah}
\]

(9.35)

approximates the Cauchy problem:

\[
\frac{du}{dx} + Au = 0, \quad 0 \leq x \leq 1, \\
u(0) = b, \quad A = \text{const},
\]

(9.36)

on its solution $u = u(x)$ with first-order accuracy with respect to $h$. However, the solution $u_h^{(b)}$ obtained with the help of this scheme does not converge to $|[u]|_h$, and does not even remain bounded as $h \to 0$. 
Indeed, the general solution of the difference equation from (9.35) has the form:

\[ u_n = c_1 q_1^n + c_2 q_2^n, \]

where \( c_1 \) and \( c_2 \) are arbitrary constants, and \( q_1 \) and \( q_2 \) are roots of the algebraic characteristic equation: \(-q^2 + (3 - Ah)q - 2 = 0\). Choosing the constants \( c_1 \) and \( c_2 \) to satisfy the initial conditions, we obtain:

\[
\begin{align*}
\frac{u_n}{u_0} &= \frac{q_2^n/q_1^n}{q_2^n/q_1^n} + \frac{q_1^n/q_2^n}{q_1^n/q_2^n} \\
&= \left(1 - \frac{q_1}{q_2}\right)^n + \left(1 - \frac{q_2}{q_1}\right)^n.
\end{align*}
\]

(9.37)

Analysis of formula (9.37) shows that \( \max_{0 \leq nh \leq 1} |u_n| \rightarrow \infty \) as \( h \rightarrow 0 \). Therefore, consistency alone is generally not sufficient for convergence. Stability is also required.

### 9.3.1 Definition of Stability

Consider an initial or boundary value problem

\[ Lu = f, \]

(9.38)

and assume that the finite-difference scheme

\[ L_hu^{(h)} = f^{(h)} \]

(9.39)

is consistent and has accuracy \( O(h^k) \), \( k > 0 \), on the solution \( u \) of problem (9.38). This means that the truncation error \( \delta f^{(h)} \) defined by the formula:

\[ L_h[u]_{h} = f^{(h)} + \delta f^{(h)}, \]

(9.40)

where \([u]_{h}\) is the projection of the exact solution \( u(x) \) onto the grid \( D_h \), satisfies the following inequality (cf. Definition 9.1):

\[ \| \delta f^{(h)} \|_{F_{h}} \leq c_1 h^k, \]

(9.41)

where \( c_1 \) is a constant that does not depend on \( h \).

**DEFINITION 9.2** Scheme (9.39) will be called stable if one can choose two numbers \( h_0 > 0 \) and \( \delta > 0 \) such that for any \( h < h_0 \) and for any \( e^{(h)} \in F_h \), \( \| e^{(h)} \|_{F_{h}} < \delta \), the finite-difference problem:

\[ L_hz^{(h)} = f^{(h)} + e^{(h)} \]

(9.42)

obtained from (9.39) by adding the perturbation \( e^{(h)} \) to the right-hand side, has one and only one solution \( z^{(h)} \) whose deviation from the solution \( u^{(h)} \) of the unperturbed problem (9.39) satisfies the estimate:

\[ \| z^{(h)} - u^{(h)} \|_{F_{h}} \leq c_2 \| e^{(h)} \|_{F_{h}}, \quad c_2 = \text{const}, \]

(9.43)
where \( c_2 \) does not depend on \( h, \varepsilon^{(h)} \).

Inequality (9.43) implies, in particular, that a small perturbation \( \varepsilon^{(h)} \) of the right-hand side of (9.39) may only cause a uniformly small perturbation \( z^{(h)} - u^{(h)} \) of the solution with respect to \( h \). In other words, it implies weak sensitivity of the solution \( u^{(h)} \) of problem (9.39) to perturbations of its right-hand side.

Let us now assume that the operator \( L_h \) that maps the space \( U_h \) onto \( F_h \) is linear.

**DEFINITION 9.3** Scheme (9.39) with the linear operator \( L_h \) will be called stable if for any given \( f^{(h)} \in F_h \) the equation \( L_h u^{(h)} = f^{(h)} \) has a unique solution \( u^{(h)} \) such that

\[
\| u^{(h)} \|_{U_h} \leq c_2 \| f^{(h)} \|_{F_h}, \quad c_2 = \text{const},
\]

(9.44)

where \( c_2 \) does not depend on \( h, f^{(h)} \).

The equivalence of Definitions 9.2 and 9.3 is rigorously proven in Chapter 10 in the context of partial differential equations, see Lemma 10.1.

### 9.3.2 The Relation between Consistency, Stability, and Convergence

The following theorem is of central importance for the entire analysis.

**THEOREM 9.1**

Let the scheme \( L_h u^{(h)} = f^{(h)} \) be consistent with the order of accuracy \( O(h^k) \), \( k > 0 \), on the solution \( u = u(x) \) of the problem \( Lu = f \), and let this scheme also be stable. Then, the finite-difference solution \( u^{(h)} \) converges to the exact solution \( [u]_h \), and the following estimate holds:

\[
\| [u]_h - u^{(h)} \|_{U_h} \leq c_1 c_2 h^k,
\]

(9.45)

where \( c_1 \) and \( c_2 \) are the constants from estimates (9.41) and (9.43).

The result of Theorem 9.45 basically means that a consistent and stable scheme converges with the same rate as its order of accuracy.

**PROOF** Let \( \varepsilon^{(h)} = \delta f^{(h)} \) and \( [u]_h = z^{(h)} \). Then, estimate (9.43) becomes:

\[
\| [u]_h - u^{(h)} \|_{U_h} \leq c_2 \| \delta f^{(h)} \|_{F_h}.
\]

Taking into account (9.41), we immediately obtain (9.45).
Let us now analyze the examples. First, we will prove stability of the forward Euler scheme \([\text{cf. formula (9.29)}]\):

\[
\frac{u_{n+1} - u_n}{h} - G(x_n, u_n) = \varphi_n, \quad n = 0, 1, \ldots, N - 1,
\]

\[u_0 = \psi,\] \hspace{1cm} (9.46)

where \(x_n = nh, n = 0, 1, \ldots, N, \) and \(h = 1/N.\) Scheme (9.46) can be employed for the numerical solution of the initial value problem \([\text{cf. formula (9.30)}]\):

\[
\frac{du}{dx} - G(x, u) = \varphi(x), \quad 0 \leq x \leq 1, \quad u\big|_{x=0} = \psi. \]

\[(9.47)\]

Henceforth, we will assume that the functions \(G = G(x, u)\) and \(\varphi = \varphi(x)\) are such that problem (9.47) has a solution \(u = u(x)\) with the bounded second derivative so that \(|u''(x)| \leq \text{const}.\) We will also assume that \(G(x, u)\) has a bounded partial derivative with respect to \(u:\)

\[
\left| \frac{\partial G}{\partial u} \right| \leq M = \text{const}. \]

\[(9.48)\]

Next, we will introduce the norms \([\text{cf. formula (9.10)}]\):

\[
\|u^{(h)}\|_{L_h} = \max_n |u_n|,
\]

\[
\|f^{(h)}\|_{F_h} = \max \{ |\psi|, \max_n |\varphi_n| \},
\]

and verify that the forward Euler scheme (9.46) is indeed stable. This scheme can be recast in the operator form (9.39) if we define:

\[
L_h u^{(h)} = \begin{cases} \frac{u_{n+1} - u_n}{h} - G(x_n, u_n), & n = 0, 1, 2, \ldots, N - 1, \\ u_0, \end{cases}
\]

\[
f^{(h)} = \begin{cases} \varphi(x_n), & n = 0, 1, 2, \ldots, N - 1, \\ \psi. \end{cases}
\]

At the same time, the perturbed problem (9.42) in detail reads \([\text{cf. formula (9.46)}]\):

\[
\frac{z_{n+1} - z_n}{h} - G(x_n, z_n) = \varphi(x_n) + \varepsilon_n, \quad n = 0, 1, \ldots, N - 1,
\]

\[z_0 = \psi + \varepsilon, \]

\[(9.49)\]

which means that in formula (9.42) we can take:

\[
\varepsilon^{(h)} = \begin{cases} \varepsilon_n, & n = 0, 1, 2, \ldots, N - 1, \\ \varepsilon. \end{cases}
\]

Let us now subtract equations (9.46) from the respective equations (9.49) termwise. In so doing, let us also denote \(z_n - u_n = w_n\) and take into account that

\[
G(x_n, z_n) - G(x_n, u_n) = \frac{\partial G(x_n, z_n)}{\partial u} w_n \equiv M_n^{(h)} w_n,
\]
where $\xi_n$ is some number in between $z_n$ and $u_n$. We will thus obtain the following system of equations with the unknown $w^{(h)} = \{w_0, w_1, \ldots, w_N\}$:

$$\frac{w_{n+1} - w_n}{h} - M_n^{(h)} w_n = \epsilon_n, \quad n = 0, 1, \ldots, N - 1,$$

(9.50)

$$w_0 = \epsilon.$$

According to (9.48), we can claim that $\forall n$: $|M_n^{(h)}| \leq M$. Then, system (9.50) yields:

$$|w_{n+1}| = |(1 + hM_n^{(h)})w_n + h\epsilon_n| \leq (1 + Mh)|w_n| + h|\epsilon_n|$$

$$\leq (1 + Mh)^2|w_{n-1}| + h(1 + Mh)|\epsilon_{n-1}| + h|\epsilon_n|$$

$$\leq (1 + Mh)^2|w_{n-1}| + 2h(1 + Mh)\|\epsilon^{(h)}\|_{F_h}$$

$$\leq (1 + Mh)^3|w_{n-2}| + 3h(1 + Mh)^2\|\epsilon^{(h)}\|_{F_h}$$

$$\cdots$$

$$\leq (1 + Mh)^{n+1}|w_0| + (n + 1)h(1 + Mh)^n\|\epsilon^{(h)}\|_{F_h}$$

$$\leq 2(1 + Mh)^N\|\epsilon^{(h)}\|_{F_h} \leq 2e^M\|\epsilon^{(h)}\|_{F_h},$$

because $(n + 1)h \leq Nh = 1$. Hence:

$$|w_{n+1}| \leq 2e^M\|\epsilon^{(h)}\|_{F_h}.$$

This inequality obviously implies an estimate of type (9.43):

$$\|w^{(h)}\|_{U_h} \leq 2e^M\|\epsilon^{(h)}\|_{F_h},$$

which is equivalent to stability with the constant $c_2 = 2e^M$ in the sense of Definition 9.2. Let us also recall that scheme (9.46) has accuracy $O(h)$ on the solution $u = u(x)$ of problem (9.47), see Exercise 1 of the previous Section 9.2. Theorem 9.1 would therefore guarantee a first-order convergence with respect to $h$ of the forward Euler scheme (9.46) on the interval $0 \leq x \leq 1$.

In our next example, we will analyze convergence of the finite-difference scheme (9.7) for the second-order boundary value problem (9.4). This scheme is consistent and guarantees the second order of accuracy due to formula (9.20a). However, stability of the scheme (9.7) yet remains to be verified.

Due to the linearity, it is sufficient to establish the unique solvability of the problem:

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - (1 + x_n^2)u_n = g_n, \quad n = 1, 2, \ldots, N - 1,$$

$$u_0 = \alpha, \quad u_N = \beta,$$

(9.51)

for arbitrary $\{g_n\}$, $\alpha$, and $\beta$, and to obtain the estimate:

$$\max_n |u_n| \leq c \max\{|\alpha|, |\beta|, \max_n |g_n|\}.$$  

(9.52)
We have, in fact, previously considered a problem of type (9.51) in the context of the tri-diagonal elimination, see Section 5.3 of Chapter 5. For the problem:

\[ a_n u_{n-1} + b_n u_n + c_n u_{n+1} = g_n, \]
\[ u_0 = \alpha, \quad u_N = \beta \]

under the assumption:

\[ \forall n : |b_n| > |a_n| + |c_n| + \delta, \quad \delta > 0, \]

we have proven its unique solvability and the estimate:

\[ \max_n |u_n| \leq \max \left\{ |\alpha|, |\beta|, \frac{2}{\delta} \max_m |g_m| \right\}. \quad (9.53) \]

In the case of problem (9.51), we have:

\[ a_n = \frac{1}{h^2}, \quad c_n = \frac{1}{h^2}, \quad b_n = -\frac{2}{h^2} - 1 - x_n^2. \]

Consequently, \( |b_n| > |a_n| + |c_n| + 1 \), and estimate (9.53) therefore implies estimate (9.52) with \( c = 2 \). We have thus proven stability of the scheme (9.7). Along with the second order of accuracy of the scheme, stability yields its convergence with the rate \( O(h^2) \). Note that stability of scheme (9.7) has been proven in the sense of Definition 9.3, which is equivalent to Definition 9.2 that was used when proving convergence in Theorem 9.1.

Let us also emphasize that the approach to proving convergence of the scheme via independently verifying its consistency and stability is, in fact, quite general. Formula \( Lu = f \) does not necessarily have to denote an initial or boundary value problem for an ordinary differential equation; it can actually be an operator equation from a rather broad class. We outline this general framework in Section 10.1.6 of Chapter 10, where we discuss the Kantorovich theorem. Here we just mention that it is not very important what type of problem the function \( u \) solves. The continuous operator equation \( Lu = f \) is only needed to construct its difference counterpart \( L_h u^{(h)} = f^{(h)} \). In Section 9.3.3, we elaborate on the latter consideration. Namely, we provide an example of a consistent, stable, and therefore convergent scheme for an integral rather than differential equation.

### 9.3.3 Convergent Scheme for an Integral Equation

We will construct and analyze a difference scheme for solving the equation:

\[ Lu \equiv u(x) - \int_0^1 K(x,y)u(y)dy = f(x), \quad 0 \leq x \leq 1, \quad (9.54) \]

while assuming that the kernel in (9.54) is bounded: \( |K(x,y)| \leq \rho < 1 \).
Let us specify a positive integer $N$, set $h = 1/N$, and introduce a uniform grid $D_h$ on $[0, 1]$ as before: $x_n = nh$, $n = 0, 1, 2, \ldots, N$. As usual, our unknown grid function will be the table of values $[u]_h$ of the exact solution $u(x)$ on the grid $D_h$.

To obtain a finite-difference scheme for the approximate computation of $[u]_h$ we will replace the integral in each equality:

$$u(x_n) - \int_0^1 K(x_n, y)u(y)dy = f(x_n), \quad n = 0, 1, 2, \ldots, N,$$

(9.55)

by a finite sum using one of the numerical quadrature formulas discussed in Chapter 4, namely, the trapezoidal rule, see Section 4.1 It says that for an arbitrary twice differentiable function $\varphi = \varphi(y)$ on the interval $0 \leq y \leq 1$, the following approximate equality holds:

$$\int_0^1 \varphi(y)dy \approx h \left( \frac{\varphi_0}{2} + \varphi_1 + \ldots + \varphi_{N-1} + \frac{\varphi_N}{2} \right), \quad h = \frac{1}{N},$$

and the corresponding approximation error is $O(h^2)$. Having replaced the integral in every equality (9.55) with the aforementioned sum, we obtain:

$$u_n - h \left( \frac{K(x_n, 0)}{2}u_0 + K(x_n, h)u_1 + \ldots + K(x_n, (N - 1)h)u_{N-1} \right. \left. + \frac{K(x_n, 1)}{2}u_N \right) = f_n, \quad n = 0, 1, 2, \ldots, N.$$

(9.56)

The system of equations (9.56) can be recast in the standardized form $L_h u^{(h)} = f^{(h)}$ if we set:

$$L_h u^{(h)} = \left\{ \begin{array}{c} g_0, \\ g_1, \\ \vdots \\ g_N, \end{array} \right\} = \left\{ \begin{array}{c} f(0), \\ f(h), \\ \vdots \\ f(Nh) \equiv f(1), \end{array} \right\},$$

where

$$g_n = u_n - h \left( \frac{K(x_n, 0)}{2}u_0 + K(x_n, h)u_1 + \ldots + \frac{K(x_n, 1)}{2}u_N \right),$$

$$n = 0, 1, 2, \ldots, N.$$

The scheme $L_h u^{(h)} = f^{(h)}$ we have just constructed is consistent and has the second order of accuracy on the solution $u = u(x)$ of the integral equation $Lu = f$ of (9.54), because the error of the trapezoidal rule is $O(h^2)$.

Let us now determine whether the scheme (9.56) is also stable. Assume that $u^{(h)} = \{u_0, u_1, \ldots, u_N\}$ is a solution of system (9.56), and let $u_0$ be its component with the maximum absolute value:

$$|u_0| \geq |u_n|, \quad n = 0, 1, 2, \ldots, N.$$
If there are more than one component with the same maximum absolute value, we will take one of those. Then, from the equation with \( n = s \) of system (9.56) we find:

\[
|f(x)| = \left| u_s - h\left(\frac{K(x,0)}{2}u_0 + K(x,h)u_1 + \ldots + \frac{K(x,1)}{2}u_N\right)\right|
\geq |u_s| - h\left(\rho + \rho + \ldots + \rho + \frac{\rho}{2}\right)|u_s| = (1 - Nh\rho)|u_s| = (1 - \rho)|u_s|,
\]

because \( |K(x,y)| \leq \rho < 1 \). Consequently,

\[
\|u^{(h)}\|_{U_{h}} = \max_{n} |u_n| = |u_s| \leq \frac{1}{1 - \rho} |f(x)| \leq \frac{1}{1 - \rho} \|f^{(h)}\|_{f_h}.
\]

In particular, when \( f(x_n) \equiv 0 \), inequality (9.57) implies that the homogeneous counterpart of system (9.56) may only have a trivial solution. Therefore, system (9.56) is uniquely solvable for any given right-hand side \( f_n \equiv f(x_n) \), \( n = 0,1,\ldots,N \). For the solution \( u^{(h)} = \{u_s\} \), inequality (9.57) implies stability in the sense of Definition 9.3, with the constant in estimate (9.44) equal to \( c_2 = 1/(1 - \rho) \). According to Theorem 9.1, solution \( u^{(h)} \) then converges to \( u \) with the second order:

\[
\|u - u^{(h)}\|_{U_{h}} = \max_{n} |u(\xi) - u_n| \leq ch^2, \quad c = \text{const}.
\]

9.3.4 The Effect of Rounding

A consistent and stable finite-difference scheme converges, and can therefore be used, at least in theory, for computing a sequence of increasingly more accurate approximations to the unknown solution when the grid is refined. In practice, however, the originally designed scheme is never realized exactly, due to the round-off errors in specifying its coefficients and right-hand sides.

Convergence with the rate \( O(h^k) \), \( k > 0 \), as \( h \to 0 \) means that we are expecting to obtain the answer with the number of accurate decimal digits on the order of \( \ln(1/h) \). Therefore, it is natural to require that when the grid size decreases, the coefficients of the scheme and its right-hand sides be specified with the increasing accuracy as well. In doing so, one has to maintain a similar number of the correct digits, again, on the order of \( \ln(1/h) \). This requirement is not very stringent, because \( \ln(1/h) \) is a slowly growing function of \( h \).

If the accuracy of specifying the coefficients of a stable scheme is commensurate with the accuracy that the original unperturbed scheme would have on a given solution of the underlying differential problem [say, both are \( O(h^k) \)], then it is possible to show that the actual perturbed scheme will remain stable and will also keep the same order of accuracy, \( O(h^k) \). Therefore, according to Theorem 9.1, its \( k \)-th order of convergence with respect to the grid size \( h \) will not deteriorate.

If one conducts the computations and refines the grid while not synchronously increasing the number of correct decimal digits in the specification of the scheme, then
one should not, generally speaking, expect any increase of the overall accuracy. In practice, however, there is usually no need to continuously re-define and improve the coefficients as $h \to 0$. The reason is that on real computers the number of decimal digits available for the machine representation of numbers is fixed. Most often one uses either single precision given by 23 binary digits or equivalently, about 7 decimal digits, or double precision given by 52 binary digits or equivalently, about 16 decimal digits. If, for example, the coefficients of the scheme and its right-hand sides are specified with double precision, then this accuracy would typically far supersede the magnitude of the truncation error on any realistic grid.

Another, and perhaps even more important issue, is that of the round-off errors that inevitably appear in the course of computation of the difference solution by a specified scheme. These errors may accumulate, which, in turn, may necessitate running the computations with continuously increasing number of correct decimal digits. If this number happens to grow “too fast” as a function of $1/h$, then the corresponding algorithm is deemed unstable and inappropriate for computations. Of course, if we have only been able to conduct the computations with an infinite precision, then we would not have had this problem at all.

For the complicated practical problems, it may generally prove both difficult and cumbersome to analyze the effect of the round-off errors on performance of the actual computational algorithms. Often, we can only make a judgment based on experimental observations. However, for a number of relatively simple cases we can basically see how many spare digits are needed using only the stability of the original scheme and the previous result that allows to specify this scheme approximately rather than exactly. The idea of the argument is that the round-off errors committed in the course of computation can sometimes be interpreted as errors in specifying the right-hand side of the scheme, up to a given factor of $h^m$, $m > 0$. Then the previous result implies that for a stable scheme these errors would not hamper the convergence, provided that the computations are conducted with the number of digits that slowly increases as $c \ln(1/h)$, $c = \text{const}$. In practice, again, the number of digits is fixed ahead of time at a value that would be a priori sufficient.

### 9.3.5 General Comments. A-stability

Consider the Cauchy problem (9.36) for a first-order constant coefficient ordinary differential equation; solution of this problem is $u(x) = b e^{-Ax}$. Problem (9.36) can be approximated by the forward Euler scheme (9.46) with first-order accuracy, here $G(x_n, u_n) = -Au_n$, $q_n = 0$, $n = 0, 1, \ldots, N-1$, and $\psi = b$. Solution of the corresponding finite-difference equation is then given by the formula:

$$u_n = b (1 - Ah)^n.$$  \hfill (9.58)

Let $A < 0$. Then we easily see from (9.58) that $\|u^{(h)}\|_{U_h} = \max_{0 \leq n \leq N} |u_n| = |b| (1 + |A|h)^{1/2}$. We also note that in this simple homogeneous case the discrete right-hand side $f^{(h)}$ consists only of the initial value $b$. As such, $\|u^{(h)}\|_{U_h} = (1 + |A|h)^{1/2} \|f^{(h)}\|_{V_h}$,
so that in the stability inequality [cf. formula (9.44)]:
\[ \|u(h)\|_{\ell_h} \leq c_2 \|f^{(h)}\|_{F_h}, \quad c_2 = \text{const}, \]  
(9.59)
the constant \(c_2\) has a lower bound: \(c_2 \geq (1 + |A|h)^{1/2} \rightarrow e^{|A|} \equiv e^{-A}\), as \(h \rightarrow 0\). At the same time, in our general stability analysis for the forward Euler method, see page 127, we have obtained an upper bound for this constant: \(c_2 \leq 2e^M = 2e^{-A}\).

We therefore see that not only the general upper bound for the stability constant in problem (9.36) with \(A < 0\) has exponential behavior, but also that it is impossible to choose a stability constant that would grow slower than the exponential \(e^{-A}\).

Clearly, for \(|A| \gg 1\) the foregoing stability constant becomes very large. This situation is actually quite general. No convergent scheme that approximates problem (9.36) for \(A < 0, \ |A| \gg 1\), may have a small stability constant. Indeed, for small grid sizes the finite-difference solution is supposed to be close to the continuous solution. However, for this continuous solution itself we have \(\|u\|_{F_h} = \max_{x \in [0,1]} |u(x)| = \max_{0 \leq n \leq N} |u_n| = \max_{0 \leq n \leq N} |b| |1 - Ah|^n = |b| |1 - Ah|^0 = |b| = \|f^{(h)}\|_{F_h}\).

Therefore, we can set \(h_0 = 2/A\) in Definition 9.2 and consider \(c_2 = 1\) in the stability inequality (9.59). For the general inhomogeneous equation, a straightforward modification of the argument given on page 127 will allow us to conclude that \(c_2 \leq 2\).

The foregoing drastic change of behavior for the stability constants is obviously accounted for by the change of behavior of the corresponding continuous solution \(u(x) = be^{-Ax}\). If \(A < 0\), we have \(\max_{0 \leq x \leq 1} |u(x)| = e^{|A|}\), which is large for \(|A| \gg 1\), whereas if \(A > 0\), then \(\max_{0 \leq x \leq 1} |u(x)| = u(0) = 1\) irrespective of \(A\). We therefore see that on one hand, stability in the sense of Definitions 9.2 and 9.3 is an intrinsic property of the scheme concerned with its asymptotic behavior as the grid dimension \(N\) grows, or equivalently, as the grid size \(h\) vanishes, on a finite fixed interval of the independent variable \(x\), say, \(0 \leq x \leq 1\). On the other hand, for a stable scheme the constant \(c_2\) that provides a quantitative measure of stability, see inequality (9.59), does depend on the properties of the continuous solution approximated by the scheme.

One should not think though that any stable scheme that approximates problem (9.36) in the “good” case \(A > 0\) has a small stability constant. For instance, one can show that the central-difference scheme:
\[ \frac{u_{n+1} - u_{n-1}}{2h} + Au_n = 0, \quad n = 1, 2, \ldots, N - 1, \]
\[ u_0 = b, \quad u_1 = b(1 - Ah), \]  
(9.60)
approximates problem (9.36) on its solution \(u(x) = be^{-Ax}\) with second-order accuracy and is stable, but the stability constant \(c_2\) for \(A \gg 1\) is large: \(c_2 \geq e^A\).
For practical computations, not only the property of stability itself, but also the values of the stability constants are important — the smaller the constants the better error estimate one can expect to get. Indeed when proving Theorem 9.1, we have obtained inequality (9.45) for the error. Assume that the truncation error $\sim c_1 h^k$ is small. Then according to (9.45), to have the magnitude of the solution error $\|u_h - u^{(b)}\|_{U_h}$ also small, one may not have a very large stability constant $c_2$.

Let us introduce the relative error $\frac{\|u_h - u^{(b)}\|_{U_h}}{\|u_h\|_{U_h}}$ in the solution. If the stability constant of the scheme happens to be of the same order of magnitude as the norm of the exact solution: $c_2 = O(\|u_h\|_{U_h})$, as in the case of the forward Euler scheme applied to problem (9.36) with $A < 0$, then the foregoing relative error basically appears bounded by the (small) value of the truncation error, which still makes the scheme appropriate for computations. If on the other hand, we have a large stability constant while $\|u_h\|_{U_h} = O(1)$ as in the case of scheme (9.60) for $A \gg 1$, then obtaining the solution with a prescribed accuracy using this scheme may require an overly fine grid and consequently, an excessive and unjustified amount of computations.

On the other hand, we would like to caution the reader from getting a distorted perception of all second-order schemes based on one particular example of scheme (9.60) considered for large values of $A$. Even for the same scheme (9.60), but applied to a problem with $A = O(1)$, one can easily show that it requires far fewer grid points for achieving a comparable accuracy than the “competing” first-order scheme (9.46). Altogether, the advantages and disadvantages of using a given scheme are determined not only by the scheme itself, but also by the problem it is applied to.

Let us now elaborate on the case $A > 0$ just a little further. Every solution of the differential equation $u' + Au = 0$ remains bounded for all $x \geq 0$. It is natural to expect a similar behavior of the finite-difference solution as well. More precisely, consider the same model problem (9.36), but on the semi-infinite interval $x \geq 0$ rather than on $0 \leq x \leq 1$. Let $u^{(b)}$ be its finite-difference solution defined on the grid $x_n = nh, n = 0, 1, 2, \ldots$. The corresponding scheme is called asymptotically stable, or $A$-stable, if for any fixed grid size $h$ the solution is bounded on the entire grid:

$$\sup_{0 \leq n < \infty} |u_n| \leq \text{const.}$$

According to formula (9.58), the explicit forward Euler scheme meets the requirement of $A$-stability if $|1 - Ah| \leq 1$, i.e., when $h \leq 2/A$. In other words, to guarantee $A$-stability, the grid size should be sufficiently small, which makes the forward Euler method conditionally $A$-stable.

The implicit backward Euler scheme:

$$\frac{u_{n+1} - u_n}{h} + Au_{n+1} = 0, \quad u_0 = b,$$

approximates the original Cauchy problem with the same first order of accuracy. For this scheme, however, instead of formula (9.58) we obtain:

$$u_n = \frac{b}{(1 + Ah)^n}.$$
For $A > 0$, formula (9.62) guarantees boundedness of $u_n$ with no constraints on the grid size. That’s why the backward Euler scheme is unconditionally $A$-stable.

To conclude this section, we describe a convenient criterion for the stability analysis of a large class of methods known as linear. The scheme that approximates the Cauchy problem (9.30) is called linear if it can be represented in the form:

\[
\sum_{k=0}^{m} \alpha_k u_{n+k} = h \sum_{k=0}^{m} \beta_k G(x_{n+k}, u_{n+k}),
\]

(9.63)

where it is assumed that $\alpha_m \neq 0$ and $\alpha_0^2 + \beta_0^2 \neq 0$. The forward Euler scheme (9.29) is linear with $m = 1$, $\alpha_0 = -1$, $\alpha_1 = 1$, $\beta_0 = 1$, $\beta_1 = 0$. It is called a single-step method because $m = 1$. The backward Euler scheme (9.34) and the Crank-Nicolson scheme (9.33) are also linear single-step methods. If $m > 1$, the linear method is called multi-step; for example, for the schemes (9.35) and (9.60) we have $m = 2$. The popular Runge-Kutta schemes described in Section 9.4 do not, generally speaking, belong to the class of linear methods.

For an $m$-step method, by appropriately choosing the coefficients $\alpha_k$ and $\beta_k$, $k = 0, \ldots, m$, one can always achieve the order of accuracy $O(h^m)$. The corresponding choice of the coefficients is clearly not unique, as already the examples of the forward and backward Euler schemes demonstrate. If $m > 1$, achieving the overall accuracy $O(h^m)$ may also require additional initial conditions for the scheme, beyond the condition $u_0 = u(0) = a$ given in (9.30), see Exercise 2 after Section 9.2.

Stability of a linear method can be characterized in terms of the roots of a certain polynomial associated with the scheme. The following Theorem that is proven, e.g., in [Sch02] provides a necessary and sufficient condition for stability.

**THEOREM 9.2**

Assume that the function $G = G(x,u)$ has a bounded partial derivative with respect to $u$: \( \left| \frac{\partial G}{\partial u} \right| \leq M = \text{const} \). Define the following polynomial of degree $m$:

\[
p(t) = \sum_{k=0}^{m} \alpha_k t^k.
\]

(9.64)

The scheme (9.63) is stable if and only if all roots $q_j$, $j = 1, \ldots, m$, of the algebraic equation $p(t) = 0$, where $p(t)$ is given by (9.64), belong to the unit disk on the complex plane: $|q_j| \leq 1$, $j = 1, \ldots, m$, and no multiple root (if any) can belong to the unit circle (i.e., for multiple roots we may not have $|q_j| = 1$).

Note that the stability criterion presented by Theorem 9.2 is closely related to the so-called spectral stability theory, which is based on the analysis of the appropriately defined eigenvalues that characterize the scheme, see [GR87]. The spectral theory of stability can be developed for both ordinary and partial differential equations, and we discuss it in the following Chapter 10, see Section 10.3.
Exercises

1. Consider the finite-difference scheme (9.35).
   a) Verify that it has first-order accuracy on the solutions of problem (9.36).
   b) Make sure that formula (9.37) yields solution of equations (9.35).
   c) Prove directly that for \( u_n \) given by (9.37): \( \max_{0 \leq m \leq 1} |u_n| \longrightarrow \infty \) as \( h \longrightarrow 0 \).
   d) Alternatively, employ Theorem 9.2 to analyze stability of the scheme.

2. Analyze stability of the backward Euler scheme (9.34).
   **Hint.** Use the argument from page 127. Alternatively, employ Theorem 9.2.

   **Hint.** Use the argument from page 127. Alternatively, employ Theorem 9.2.
   Also obtain an accurate estimate for the stability constant in the case of a linear constant-coefficient equation \( u' + Au = 0 \). Make sure it is not large for \( A \gg 1 \).

4. * Prove that scheme (9.60) has second-order accuracy and is stable.
   **Hint.** For the stability analysis, use the approach similar to the one outlined in the very beginning of Section 9.3. Show that if \( q_1 \) and \( q_2 \) are roots of the characteristic equation \( q^2 + 2Ahq - 1 = 0 \) that corresponds to scheme (9.60), then the quantities in front of \( u_0 \) and \( u_1 \) in formula (9.37) remain bounded for all \( n = 0, 1, \ldots, N = 1/h \) when \( h \longrightarrow 0 \). Alternatively, employ Theorem 9.2.

5. Show that the Crank-Nicolson scheme (9.33) for the equation \( u' + Au = 0, A > 0 \), is unconditionally A-stable.

9.4 The Runge-Kutta Methods

We will first discuss a class of widely used finite-difference schemes for solving the Cauchy problem for a scalar first-order ordinary differential equation:

\[
\frac{du}{dx} G(x,u) = 0, \quad 0 \leq x \leq 1, \quad u(0) = a.
\]  

(9.65)

Then, we will analyze extensions to first-order systems. This is the most general case as an equation or system of any order can be recast as a first-order system.

Again, let the grid \( D_h \) be uniform on the interval \([0, 1] \):

\[
x_n = nh, \quad n = 0, 1, 2, \ldots, N, \quad h = 1/N.
\]

We will be building schemes for the approximate computation of the table \([u]_h = \{u(x_n) | n = 0, 1, 2, \ldots, N\}\), where \( u = u(x) \) is the exact solution of problem (9.65).

We have already described the simplest explicit scheme for solving problem (9.65), which is the forward Euler scheme [cf. formulae (9.29) & (9.46)]:

\[
\frac{u_{n+1} - u_n}{h} - G(x_n, u_n) = 0, \quad n = 0, 1, \ldots, N - 1, \quad u_0 = a.
\]  

(9.66)
This scheme has first-order accuracy, it also converges with the rate $O(h)$. Computations according to scheme (9.66) can be given a straightforward geometric interpretation. If $u_n$ is already known, then computing $u_{n+1}$ according to the marching formula: $u_{n+1} = u_n + hG(x_n, u_n)$ is basically equivalent to the translation from the point $(x_n, u_n)$ to the point $(x_{n+1}, u_{n+1})$ on the $(x, u)$ plane along the line tangent to the integral curve of the equation $\frac{du}{dx} = G(x, u)$ that crosses through the point $(x_n, u_n)$.

The Runge-Kutta methods yield explicit schemes capable of attaining higher orders of accuracy; they are most popular for solving ordinary differential equations.

9.4.1 The Runge-Kutta Schemes

Assume that the approximate solution $u_n$ is already known at the node $x_n$. We need to find $u_{n+1}$ at the next node $x_{n+1} = x_n + h$. To do so, we first choose a positive integer $l$ and write down the following expressions:

\[ k_1 = G(x_n, u_n), \]
\[ k_2 = G(x_n + \alpha h, u_n + \alpha h k_1), \]
\[ k_3 = G(x_n + \beta h, u_n + \beta h k_2), \]
.................................
\[ k_l = G(x_n + \gamma h, u_n + \gamma h k_{l-1}). \]

Then, we construct the scheme itself:

\[ \frac{u_{n+1} - u_n}{h} - (p_1 k_1 + p_2 k_2 + \ldots + p_l k_l) = 0, \quad n = 0, 1, \ldots, N - 1, \]
\[ u_0 = a. \]

In doing so, the coefficients $\alpha, \beta, \ldots, \gamma, p_1, p_2, \ldots, p_l$ are selected to obtain the maximum possible order of accuracy for a given $l$. Once $u_n$ is known and the coefficients have been determined, then one can first calculate $k_1, k_2, \ldots, k_l$, and subsequently obtain the solution $u_{n+1} = u_n + h(p_1 k_1 + p_2 k_2 + \ldots + p_l k_l)$. Clearly, the computations according to the Runge-Kutta scheme remain explicit.

The simplest Runge-Kutta scheme is the previously analyzed forward Euler scheme (9.66) that corresponds to $l = 1$ and $p_1 = 1$. The predictor-corrector scheme (9.32) is a Runge-Kutta scheme with $l = 2$, $p_1 = p_2 = 1/2$, and $\alpha = 1$.

The Runge-Kutta scheme:

\[ \frac{u_{n+1} - u_n}{h} - \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 0, \quad n = 0, 1, \ldots, N - 1, \]
\[ u_0 = a, \]

has fourth-order accuracy if the additional quantities are defined as:

\[ k_1 = G(x_n, u_n), \quad k_2 = G\left(x_n + \frac{h}{2}, u_n + \frac{k_1 h}{2}\right), \]
\[ k_3 = G\left(x_n + \frac{h}{2}, u_n + \frac{k_2 h}{2}\right), \quad k_4 = G(x_n + h, u_n + k_3 h). \]
The Runge-Kutta scheme:

\[
\frac{u_{n+1} - u_n}{h} = \left( \frac{2\alpha - 1}{2\alpha} k_1 + \frac{1}{2\alpha} k_2 \right), \quad n = 0, 1, \ldots, N - 1, \tag{9.68}
\]

\[u_0 = a,
\]

where \(k_1 = G(x_n, u_n)\) and \(k_2 = G(x_n + \alpha h, u_n + \alpha h k_1)\) has second-order accuracy for any fixed \(\alpha \neq 0\). We will only analyze the accuracy of scheme (9.68); the analysis for scheme (9.67) can be conducted similarly, but it is more cumbersome.

Solution \(u = u(x)\) of the differential equation \(u' = G(x, u)\) satisfies the identities:

\[
\frac{du}{dx} = G(x, u(x)),
\]

\[
\frac{d^2 u}{dx^2} = \frac{d}{dx} G(x, u(x)) = \frac{\partial G}{\partial x} + \frac{\partial G}{\partial u} G.
\]

Therefore, the Taylor formula:

\[
\frac{u(x_n + h) - u(x_n)}{h} = u'(x_n) + \frac{h}{2} u''(x_n) + \mathcal{O}(h^2)
\]

leads to the following equality for the solution \(u(x)\):

\[
\frac{u(x_n + h) - u(x_n)}{h} = \left[ G + \frac{h}{2} \left( \frac{\partial G}{\partial x} + \frac{\partial G}{\partial u} G \right) \right]_{x = x_n, u = u(x_n)} = \mathcal{O}(h^2). \tag{9.69}
\]

Next, we employ the Taylor formula for the function of two variables: \(k_2 = G(x_n + \alpha h, u_n + \alpha h k_1)\), and by retaining the terms up to the first order, obtain:

\[
\frac{2\alpha - 1}{2\alpha} k_1 + \frac{1}{2\alpha} k_2
\]

\[
= \left[ G + \frac{h}{2} \left( \frac{\partial G}{\partial x} + \frac{\partial G}{\partial u} G \right) \right]_{x = x_n, u = u(x_n)} + \mathcal{O}(h^2). \tag{9.70}
\]

Consequently, if we substitute \(u(x_n)\) instead of \(u_n\) and \(u(x_{n+1})\) instead of \(u_{n+1}\) into the left-hand side of the first equality of (9.68), where \(u(x)\) is the solution to (9.65), then we obtain an expression that would coincide with the left-hand side of (9.69) with accuracy \(\mathcal{O}(h^2)\). As such, the left-hand side of the first equality of (9.68) has the second order of accuracy on the solution \(u = u(x)\) of the equation \(u' = G(x, u)\). Since the initial condition \(u_0 = a\) is specified exactly, this completes the proof of the overall second-order accuracy of scheme (9.68).
9.4.2 Extension to Systems

All the schemes that we have introduced in Section 9.4.1 for solving the Cauchy problem for a scalar first-order ordinary differential equation (9.65) can be easily generalized to systems of such equations. To do so, in formula (9.65) we simply need to consider the vector functions \( u = u(x) \) and \( G = G(x, u) \) instead of \( u(x) \) and \( G(x, u) \) respectively, and a fixed vector \( a \) instead of the given scalar quantity \( a \).\(^4\) For example, the system of ordinary differential equations:

\[
\begin{align*}
\frac{dv}{dx} - (x + v^2 + \sin w) &= 0, \\
\frac{dw}{dx} + xvw &= 0, \\
v(0) &= a_1, \quad w(0) = a_2
\end{align*}
\]

can be recast in the form:

\[
\begin{align*}
\frac{du}{dx} - G(x, u) &= 0, \\
u(0) &= a,
\end{align*}
\]

if we set:

\[
\begin{align*}
u(x) &= \begin{bmatrix} v(x) \\ w(x) \end{bmatrix}, \\
G(x, u) &= \begin{bmatrix} x + v^2 + \sin w \\ -xvw \end{bmatrix}, \\
a &= \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.
\end{align*}
\]

The formula for \( u_{n+1} \) in the forward Euler scheme for this system:

\[
\begin{align*}
u_{n+1} &= v_n + hG(x_n, u_n)
\end{align*}
\]

can be written in components as follows:

\[
\begin{align*}
v_{n+1} &= v_n + h(x_n + v_n^2 + \sin w_n), \\
w_{n+1} &= w_n - hx_n v_n w_n.
\end{align*}
\]

All the considerations regarding the order of accuracy for one scalar equation also translate to systems. In so doing, the derivative \( \frac{\partial G}{\partial u} \) in formulae (9.69) and (9.70) is replaced by the Jacobi matrix:

\[
\begin{align*}
\frac{\partial G}{\partial u} &= \begin{bmatrix}
\frac{\partial G_1}{\partial u_1} & \frac{\partial G_1}{\partial u_2} & \cdots & \frac{\partial G_1}{\partial u_n} \\
\frac{\partial G_2}{\partial u_1} & \frac{\partial G_2}{\partial u_2} & \cdots & \frac{\partial G_2}{\partial u_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial G_n}{\partial u_1} & \frac{\partial G_n}{\partial u_2} & \cdots & \frac{\partial G_n}{\partial u_n}
\end{bmatrix}.
\end{align*}
\]

An arbitrary system of ordinary differential equations resolved with respect to the highest-order derivatives can be reduced to a system of first-order equations,

\(^4\)All vectors must be of the same dimension.
Numerical Solution of Ordinary Differential Equations

\[ \frac{du}{dx} = G(x, u), \] by introducing additional unknowns. We illustrate this technique by the following example. The Cauchy problem:

\[
\begin{align*}
\frac{d^2 v}{dx^2} + \sin \left( x \frac{dv}{dx} + v^2 + w \right) &= 0, \\
\frac{dw}{dx} + \sqrt{x^2 + v^2 + \left( \frac{dv}{dx} \right)^2 + w^2} &= 0, \\
v(0) &= a, \quad \frac{dv}{dx} \bigg|_{x=0} = b, \quad w(0) = c
\end{align*}
\]

is transformed into a first-order system if we set: \( u_1 = v, \ u_2 = \frac{dv}{dx}, \) and \( u_3 = w(x). \) Then we obtain:

\[
\begin{align*}
\frac{du_1}{dx} - u_2 &= 0, \\
\frac{du_2}{dx} + \sin (xu_2 + u_1^2 + u_3) &= 0, \\
\frac{du_3}{dx} + \sqrt{x^2 + u_2^2 + u_1^2 + u_3^2} &= 0, \\
u_1(0) &= a, \quad u_2(0) = b, \quad u_3(0) = c.
\end{align*}
\]

**Remark 9.1** Special Runge-Kutta type schemes have been developed that can solve second-order ordinary differential equations directly, i.e., without initially reducing them to first order systems.

**Exercises**

1. Assume that the function \( G = G(x, u) \) has a bounded partial derivative with respect to \( u, \) i.e., that inequality (9.48) holds. Prove stability according to Definition 9.2 of the second-order Runge-Kutta scheme (9.68) built for solving the Cauchy problem (9.65).
   **Hint.** Develop an argument similar to the one used in Section 9.3.2 when proving stability of the forward Euler method. Use the Taylor formula with the error term of order \( O(h) \) in the Lagrange form.

2. Consider a Cauchy problem for the first-order ordinary differential equation: \( u' + u = 0, \) \( u(0) = 1, \) \( 0 \leq x \leq 1. \) To solve this problem numerically, implement on the computer the forward Euler scheme (9.66), the second-order Runge-Kutta scheme (9.68) [more precisely, (9.32)], and the fourth order Runge-Kutta scheme (9.67). Verify experimentally the first-order convergence, the second-order convergence, and the fourth-order convergence, respectively, of the corresponding finite-difference solutions to the exact solution \( u(x) = e^{-x} \) of the differential problem.

3. Consider a Cauchy problem for the second-order ordinary differential equation: \( u'' + u = 0, \) \( u(0) = 0, \) \( u'(0) = 1, \) \( 0 \leq x \leq \pi/2. \) To solve this problem numerically, first reduce it to a system of two first-order equations, and then follow the recipes of Problem 2; the exact solution is \( u(x) = \sin x. \)
9.5 Solution of Boundary Value Problems

We will outline some techniques for the numerical solution of boundary value problems using the following simple example of a boundary value problem for the second-order ordinary differential equation:

\[ y'' = f(x,y,y'), \quad 0 \leq x \leq 1, \]
\[ y(0) = Y_0, \quad y(1) = Y_1, \]

(9.71)

where the boundary conditions are specified at both endpoints of the interval $[0, 1]$.\(^5\)

9.5.1 The Shooting Method

In Section 9.4, we have described a number of numerical methods for solving the Cauchy problem, i.e., a problem of the type:

\[ y'' = f(x,y,y'), \quad 0 \leq x \leq 1, \]
\[ y(0) = Y_0, \quad \frac{dy}{dx} \bigg|_{x=0} = \tan \alpha, \]

(9.72)

where $Y_0$ is the ordinate of the point $(0,Y_0)$ on the $(x,y)$ plane from which the integral curve of the differential equation $y'' = f(x,y,y')$ originates, and $\tan \alpha$ is the slope of the tangent to this integral curve with respect to the horizontal axis, see Figure 9.1.

For a fixed $Y_0$, the solution of problem (9.72) is a function of the independent variable $x$ and of the parameter $\alpha$: $y = y(x,\alpha)$. When $x = 1$, the solution $y = y(x,\alpha)$ depends only on $\alpha$:

\[ y(x,\alpha) \bigg|_{x=1} = y(1,\alpha). \]

Using this consideration, we can reformulate the boundary value problem (9.71) as follows: Find such an angle $\alpha = \alpha^*$, for which the integral curve of the equation $y'' = f(x,y,y')$ would hit exactly the point $(1,Y_1)$:

\[ y(1,\alpha) = Y_1, \]

(9.73)

provided that this curve originates at the point $(0,Y_0)$ and has slope $\tan \alpha$ with respect to the horizontal axis at its origin. For

---

\(^5\)This is in contradistinction to the Cauchy, or initial value problem, for which the initial conditions are always specified only at one endpoint.
this particular $\alpha = \alpha^*$, the solution of the Cauchy problem (9.72) will obviously coincide with the solution of the boundary value problem (9.71) that we are seeking. As such, the overall problem reduces to solving equation (9.73).

Equation (9.73) is a functional equation of the general type $F(\alpha) = 0$, where $F(\alpha) = y(1, \alpha) - Y_1$. The only difference between this equation and the traditional equations is that the function $F = F(\alpha)$ is not specified by an analytic expression and is rather introduced through an algorithm for solving the Cauchy problem (9.72). The essence of the shooting method is precisely the foregoing reduction of the boundary value problem (9.71) to the initial value problem (9.72).

To solve equation (9.73), one can employ any of the conventional rootfinding techniques for scalar algebraic equations. Perhaps the simplest approach employs the well-known idea of bisection. We start with specifying $\alpha_0$ and $\alpha_1$, the initial guesses, so that the differences $y(1, \alpha_0) - Y_1$ and $y(1, \alpha_1) - Y_1$ have opposite signs, see Figure 9.1. Then, we take the mid-point of the interval $[\alpha_0, \alpha_1]$:

$$\alpha_2 = \frac{\alpha_0 + \alpha_1}{2}$$

and evaluate $y(1, \alpha_2)$. The next iteration, $\alpha_3$, is found according to one of the two formulae:

$$\alpha_3 = \frac{\alpha_0 + \alpha_2}{2} \quad \text{or} \quad \alpha_3 = \frac{\alpha_2 + \alpha_1}{2},$$

depending on whether the differences $y(1, \alpha_0) - Y_1$ and $y(1, \alpha_2) - Y_1$ have the opposite signs or the same sign. Then, the quantity $y(1, \alpha_3)$ is computed. The process is continued until the desired accuracy $|y(1, \alpha_n) - Y_1| < \varepsilon$ is reached.

If one chooses to use the secant method for solving equation (9.73), then, having specified the two initial values $\alpha_0$ and $\alpha_1$, one computes the subsequent $\alpha_n$ according to the recursion formula:

$$\alpha_{n+1} = \alpha_n - \frac{F(\alpha_n)}{F(\alpha_n) - F(\alpha_{n-1})} (\alpha_n - \alpha_{n-1}), \quad n = 1, 2, \ldots.$$ 

One can also employ the Newton method for solving equation (9.73), see Section ?? of Chapter ???.

In general, the shooting method that reduces the solution of the boundary value problem (9.71) to the repeated computation of the solution to the initial value problem (9.72), works well if the solution $y = y(x, \alpha)$ is not “overly sensitive” to the value of $\alpha$. Otherwise, the method may become computationally unstable even if problem (9.71) is well conditioned.

Indeed, consider the following example of a linear boundary value problem:

$$y'' - a^2y = 0, \quad 0 \leq x \leq 1,$$

$$y(0) = Y_0, \quad y(1) = Y_1,$$  \hspace{1cm} (9.74)

where $a^2 = \text{const}$. The solution of problem (9.74) is given by the expression:

$$y(x) = \frac{e^{-ax} - e^{-a(2-x)}}{1 - e^{-2a}} Y_0 + \frac{e^{-a(1-x)} - e^{-a(1+x)}}{1 - e^{-2a}} Y_1.$$  \hspace{1cm} (9.75)
Note that the coefficients in front of $Y_0$ and $Y_1$ in formula (9.75) remain bounded functions of their argument $x \in [0, 1]$ uniformly with respect to $a > 0$, namely, both coefficients never exceed one. Consequently, small perturbations of the values of $Y_0$ and $Y_1$ may only cause small perturbations in the solution $y(x)$ given by (9.75), and hence problem (9.74) is well conditioned.

Let us now analyze the Cauchy problem that corresponds to (9.74):

$$y'' - a^2 y = 0, \quad 0 \leq x \leq 1,$$

$$y(0) = Y_0, \quad \frac{dy}{dx} \bigg|_{x=0} = \tan \alpha,$$  \hspace{1cm} (9.76)

The solution of problem (9.76) can be written as:

$$y(x) = aY_0 + \frac{\tan \alpha}{2a} e^{ax} + \frac{aY_0 - \tan \alpha}{2a} e^{-ax}.$$ \hspace{1cm} (9.77)

If we have an error of magnitude $\varepsilon$ when specifying $\tan \alpha$, then the value of the solution $y(x)$ of (9.77) at $x = 1$ gets perturbed by the quantity:

$$\Delta y(1) = \frac{\varepsilon}{2a} e^{a} - \frac{\varepsilon}{2a} e^{-a}.$$ \hspace{1cm} (9.78)

If $a$ is large, then the subtrahend on the right-hand side of formula (9.78) is negligibly small, however, the coefficient in front of $\varepsilon$ in the minuend becomes large. Therefore, the quantity $y(1, \alpha)$ that we need to repeatedly compute in the course of shooting for different values of $\alpha$ may actually appear too sensitive to $\alpha$ itself. This will make the shooting method inappropriate for practical computation when the boundary value problem (9.74) needs to be solved numerically for a large $a$.

### 9.5.2 Tri-Diagonal Elimination

To compute the solution of the linear boundary value problem:

$$y'' - p(x)y = f(x), \quad 0 \leq x \leq 1,$$

$$y(0) = Y_0, \quad y(1) = Y_1,$$

where $p(x) \geq \delta > 0$, one can employ the second-order scheme:

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p(x_n)y_n = f(x_n), \quad n = 1, 2, \ldots, N - 1,$$

$$y_0 = Y_0, \quad y_N = Y_1,$$

and subsequently use the tri-diagonal elimination for solving the resulting system of linear algebraic equations. One can easily see that the sufficient conditions for using tri-diagonal elimination, see Section 5.3 of Chapter 5, will be satisfied.
9.5.3 The Newton Method

As we have seen, the shooting method of Section 9.5.1 may become unstable even if the original boundary value problem is well conditioned. On the other hand, the disadvantage of the tri-diagonal elimination as it applies to solving boundary value problems (Section 9.5.2) is that it can only be used for the linear case.

An alternative is offered by the direct application of the Newton method. Recall, in Section 9.5.1 we have indicated that the Newton method can be used for finding roots of the scalar algebraic equation (9.73) in the context of shooting. This method, however, can also be used directly for solving systems of nonlinear algebraic and differential equations, see Section ??? of Chapter ???. The Newton method is based on linearization, i.e., on reducing the solution of a nonlinear problem to the solution of a sequence of linear problems.

Consider the same boundary value problem (9.71) and assume that there is a function $y_i = y_0(x)$ that satisfies the boundary conditions of (9.71) exactly, and also roughly approximates the unknown solution $y_i = y(x)$. We will use this function $y_i = y_0(x)$ as the initial guess for the Newton iteration. Let

$$y(x) = y_0(x) + v(x),$$

where $v(x)$ is the correction to the initial guess $y_0(x)$. Substitute equality (9.79) into formula (9.71) and linearize the problem around $y_0$ using the following relations:

$$y'(x) = y_0'(x) + v'(x), \quad y''(x) = y_0''(x) + v''(x),$$

$$f(x, y_0 + v, y_0' + v') = f(x, y_0, y_0') + \frac{\partial f(x, y_0, y_0')}{\partial y'} v' + \frac{\partial f(x, y_0, y_0')}{\partial y''} v'' + O(v^2 + |v'|^2).$$

Disregarding the quadratic remainder $O(v^2 + |v'|^2)$, we arrive at the linear problem for the correction $v(x)$:

$$v'' = pv' + qv + \varphi(x), \quad v(0) = v(1) = 0,$$

where

$$p = p(x) = \frac{\partial f(x, y_0, y_0')}{\partial y'}, \quad q = q(x) = \frac{\partial f(x, y_0, y_0')}{\partial y},$$

$$\varphi(x) = f(x, y_0, y_0') - y_0''(x).$$

By solving the linear problem (9.80) either analytically or numerically, we determine the correction $v$ and subsequently define the next Newton iteration as follows:

$$y_1 = y_0 + v.$$

Then, the procedure cyclically repeats itself, and a sequence of iterations $y_1, y_2, y_3, \ldots$ is computed until the difference between the two consecutive iterations becomes smaller than some initially prescribed tolerance, which means convergence of the Newton method.

Note that the foregoing procedure can also be applied directly to the nonlinear finite-difference problem that arises when the differential problem (9.71) is approximated on the grid, see Exercise ??? of Section ???, Chapter ???.

Numerical Solution of Ordinary Differential Equations
Exercises

1. Show that the coefficients in front of $Y_1$ and $Y_0$ in formula (9.75) are indeed bounded by one for all $x \in [0, 1]$ and all $a > 0$.

2. Use the shooting method to solve the boundary value problem (9.74) for a “moderate” value of $a = 1$; also take $Y_0 = 1, Y_1 = -1$.
   a) Approximate the original differential equation with the second-order central differences on a uniform grid: $x_n = nh, n = 0, 1, \ldots, N, h = 1/N$. Set $y(0) = y_0 = 0$, and use $y(h) = y_1 = \alpha$ as the parameter to be determined by shooting.
   b) Reduce the original second-order differential equation to a system of two first-order equations and employ the Runge-Kutta scheme (9.68). The unknown parameter for shooting will then be $y'(0) = \alpha$, exactly as in problem (9.76).

In both cases, conduct the computations on a sequence of consecutively more fine grids (reduce the size $h$ by a factor of two several times). Verify experimentally that the numerical solution converges to the exact solution (9.75) with the second order with respect to $h$.

3. Investigate applicability of the shooting method to solving the boundary value problem:
   $$y'' + a^2 y = 0, \quad 0 \leq x \leq 1,$$
   $$y(0) = Y_0, \quad y(1) = Y_1,$$
which has a “+” sign instead of the “−” in the governing differential equation, but otherwise is identical to problem (9.74).

9.6 Saturation of Finite-Difference Methods

Previously, we explored the saturation of numerical methods by smoothness in the context of algebraic interpolation (piece-wise polynomials, see Section 2.2.5, and splines, see Section 2.3.2). Very briefly, the idea is to see whether or not a given method of approximation fully utilizes all the information available, and thus attains the optimal accuracy limited only by the threshold of the unavoidable error. When the method introduces its own error threshold, which may only be larger than that of the unavoidable error and shall be attributed to the specific design, we say that the phenomenon of saturation takes place. For example, we have seen that the interpolation by means of algebraic polynomials on uniform grids saturates, whereas the interpolation by means of trigonometric polynomials does not, see Section 3.1.3.

In the current section, we will use a number of very simple examples to demonstrate that the approximations by means of finite-difference schemes are, generally speaking, also prone to saturation.

Consider the following boundary value problem for a second-order ordinary differential equation:
   $$u'' = f(x), \quad 0 \leq x \leq 1, \quad u(0) = 0, \quad u(1) = 0,$$  (9.81)
where the right-hand side \( f(x) \) is assumed given.

We introduce a uniform grid on the interval \([0, 1]\):

\[
x_n = nh, \quad n = 0, 1, \ldots, N, \quad Nh = 1,
\]

and approximate problem (9.81) using central differences:

\[
\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} = f_n \equiv f(x_n), \quad n = 1, 2, \ldots, N - 1,
\]

\[
u_0 = 0, \quad u_N = 0.
\]

Provided that the solution \( u = u(x) \) of problem (9.81) is sufficiently smooth, or more precisely, provided that its fourth derivative \( u^{(4)}(x) \) is bounded for \( 0 \leq x \leq 1 \), the approximation (9.82) is second-order accurate, see formula (9.20a). In this case, if the scheme (9.82) is stable, then it will converge with the rate \( O(h^2) \).

However, for such a simple difference scheme as (9.82) one can easily study the convergence directly, i.e., without using Theorem 9.1. A study of that type will be particularly instrumental because on one hand the regularity of the solution may not always be sufficient to guarantee consistency, and on the other hand, it will allow one to see whether or not the convergence accelerates for the functions that are smoother than those minimally required for obtaining \( O(h^2) \).

Note that the degree of regularity of the solution \( u(x) \) to problem (9.81) is immediately determined by that of the right-hand side \( f(x) \). Namely, the solution \( u(x) \) will always have two additional derivatives. It will therefore be convenient to use different right-hand sides \( f(x) \) with different degree of regularity, and to investigate directly the convergence properties of scheme (9.82). In doing so, we will analyze both the case when the regularity is formally insufficient to guarantee the second-order convergence, and the opposite case when the regularity is "excessive" for that purpose. In the latter case we will, in fact, see that the convergence still remains only second-order with respect to \( h \), which implies saturation.

Let us first consider a discontinuous right-hand side:

\[
f(x) = \begin{cases} 0, & 0 \leq x \leq \frac{1}{2}, \\ 1, & \frac{1}{2} < x \leq 1. \end{cases}
\]

On each of the two sub-intervals: \([0, 1/2]\) and \([1/2, 1]\), the solution can be found as a combination of the general solution to the homogeneous equation and a particular solution to the inhomogeneous equation. The latter is equal to zero on \([0, 1/2]\) and on \([1/2, 1]\) it is easily obtained using undetermined coefficients. Therefore, the overall solution of problem (9.81), (9.83) can be found in the form:

\[
u(x) = \begin{cases} c_1 + c_2 x, & 0 \leq x \leq \frac{1}{2}, \\ c_3 + c_4 x + \frac{1}{2} x^2, & \frac{1}{2} < x \leq 1, \end{cases}
\]

(9.84)

where the constants \( c_1, c_2, c_3, \) and \( c_4 \) are to be chosen so that to satisfy the boundary conditions \( u(0) = u(1) = 1 \) and the continuity requirements:

\[
u\left(\frac{1}{2} - 0\right) = u\left(\frac{1}{2} + 0\right), \quad u'\left(\frac{1}{2} - 0\right) = u'\left(\frac{1}{2} + 0\right).
\]

(9.85)
Altogether this yields:
\[ c_1 = 0, \quad c_3 + c_4 + \frac{1}{2} = 0, \]
\[ c_1 + \frac{c_2}{2} - c_3 - \frac{c_4}{2} - \frac{1}{8} = 0, \quad c_2 - c_4 - \frac{1}{2} = 0. \]  
\hspace{1cm} (9.86)

Solving equations (9.86) we find:
\[ c_1 = 0, \quad c_2 = -\frac{1}{8}, \quad c_3 = \frac{1}{8}, \quad c_4 = -\frac{5}{8}. \]  
\hspace{1cm} (9.87)

so that
\[ u(x) = \begin{cases} 
-\frac{1}{8}x, & 0 \leq x \leq \frac{1}{2}, \\
\frac{1}{8} - \frac{5}{8}x + \frac{1}{2}x^2, & \frac{1}{2} < x \leq 1.
\end{cases} \]  
\hspace{1cm} (9.88)

In the finite-difference case, instead of (9.83) we have:
\[ f_n = \begin{cases} 
0, & n = 0, 1, \ldots, \frac{N}{2}, \\
1, & n = \frac{N}{2} + 1, \frac{N}{2} + 2, \ldots, N.
\end{cases} \]  
\hspace{1cm} (9.89)

Accordingly, the solution is to be sought for in the form:
\[ u_n = \begin{cases} 
c_1 + c_2(nh), & n = 0, 1, \ldots, \frac{N}{2} + 1, \\
c_3 + c_4(nh) + \frac{1}{2}(nh)^2, & n = \frac{N}{2}, \frac{N}{2} + 1, \ldots, N,
\end{cases} \]  
\hspace{1cm} (9.90)

where on each sub-interval we have a combination of the general solution to the homogeneous difference equation and a particular solution of the inhomogeneous difference equation (obtained via the undetermined coefficients). Notice that unlike in the continuous case (9.84), the two grid sub-intervals in formula (9.90) overlap across the entire cell \([N/2, N/2 + 1]\) (we are assuming that \(N\) is even). Therefore, the constants \(c_1, c_2, c_3,\) and \(c_4\) in (9.90) are to be determined from the boundary conditions at the endpoints of the interval \([0, 1]\): \(u_0 = u_N = 0\), and from the matching conditions in the middle that are given simply as [cf. formula (9.85)]:
\[ c_1 + c_2 \left( \frac{N}{2}h \right) = c_3 + c_4 \left( \frac{N}{2}h \right) + \frac{1}{2} \left( \frac{N}{2}h \right)^2, \]
\[ c_1 + c_2 \left( \frac{N}{2}h + h \right) = c_3 + c_4 \left( \frac{N}{2}h + h \right) + \frac{1}{2} \left( \frac{N}{2}h + h \right)^2. \]  
\hspace{1cm} (9.91)

Altogether this yields:
\[ c_1 = 0, \quad c_3 + c_4 + \frac{1}{2} = 0, \]  
\[ c_1 + \frac{c_2}{2} - c_3 - \frac{c_4}{2} - \frac{1}{8} = 0, \quad c_2 - c_4 - \frac{1}{2} = 0. \]  
\hspace{1cm} (9.92)
where the last equation of system (9.92) was obtained by subtracting the first equation of (9.91) from the second equation of (9.91) and subsequently dividing by $h$.

Notice that system (9.92) which characterizes the finite-difference case is almost identical to system (9.86) which characterizes the continuous case, except that there is an $\mathcal{O}(h)$ discrepancy in the fourth equation. Accordingly, there is also an $\mathcal{O}(h)$ difference in the values of the constants [cf. formula (9.87)]:

$$c_1 = 0, \quad c_2 = -\frac{1}{8} + \frac{h}{4}, \quad c_3 = \frac{1}{8} + \frac{h}{4}, \quad c_4 = -\frac{5}{8} - \frac{h}{4},$$

so that the solution to problem (9.82), (9.89) is given by:

$$u_n = \begin{cases} -\frac{1}{8} (nh) + \frac{h}{4} (nh), & n = 0, 1, \ldots, \frac{N}{2} + 1, \\ \frac{1}{8} - \frac{5}{8} (nh) + \frac{h}{4} (1 - nh), & n = \frac{N}{2} + 1, \ldots, N. \end{cases} \quad (9.93)$$

By comparing formulae (9.88) and (9.93), where $nh = x_n$, we conclude that

$$||u||_h - u^{(h)}|| = \max_n |u(x_n) - u_n| = \mathcal{O}(h),$$

i.e., that the solution of the finite-difference problem (9.82), (9.89) converges to the solution of the differential problem (9.81), (9.83) with the first order with respect to $h$. Note that scheme (9.82), (9.89) falls short of the second-order convergence because the solution of the differential problem (9.82), (9.89) is not sufficiently smooth.

Instead of the discontinuous right-hand side (9.83) let us now consider a continuous function with discontinuous first derivative:

$$f(x) = \begin{cases} -x, & 0 \leq x \leq \frac{1}{2}, \\ x - 1, & \frac{1}{2} < x \leq 1. \end{cases} \quad (9.94)$$

Solution to problem (9.81), (9.94) can be found in the form:

$$u(x) = \begin{cases} c_1 + c_2 x - \frac{1}{6} x^3, & 0 \leq x \leq \frac{1}{2}, \\ c_3 + c_4 x + \frac{1}{3} x^3 - \frac{1}{2} x^2, & \frac{1}{2} < x \leq 1, \end{cases}$$

where the constants $c_1, c_2, c_3,$ and $c_4$ are again to be chosen so that to satisfy the boundary conditions $u(0) = u(1) = 1$ and the continuity requirements (9.85):

$$c_1 = 0, \quad c_3 + c_4 - \frac{1}{3} = 0, \quad (9.95)$$

$$c_1 + \frac{c_2}{2} - \frac{1}{48} - c_3 - \frac{c_4}{2} + \frac{5}{48} = 0, \quad c_2 - \frac{c_4}{8} + \frac{3}{8} = 0.$$ 

Solving equations (9.95) we find:

$$c_1 = 0, \quad c_2 = \frac{1}{8}, \quad c_3 = -\frac{1}{24}, \quad c_4 = \frac{3}{8}. \quad (9.96)$$
so that

\[ u(x) = \begin{cases} \frac{1}{4}x - \frac{1}{4}x^3, & 0 \leq x \leq \frac{1}{4}, \\ \frac{1}{24} + \frac{3}{5}x + \frac{1}{6}x^3 - \frac{1}{2}x^2, & \frac{1}{2} < x \leq 1. \end{cases} \]  \tag{9.97}

In the discrete case, instead of (9.94) we write:

\[ f_n = \begin{cases} -(nh), & n = 0, 1, \ldots, \frac{N}{2}, \\ (nh) - 1, & n = \frac{N}{2} + 1, \frac{N}{2} + 2, \ldots, N, \end{cases} \]  \tag{9.98}

and then look for the solution \( u_n \) to problem (9.82), (9.98) in the form:

\[ u_n = \begin{cases} c_1 + c_2(nh) - \frac{1}{2}(nh)^3, & n = 0, 1, \ldots, \frac{N}{2} + 1, \\ c_3 + c_4(nh) + \frac{1}{6}(nh)^3 - \frac{1}{2}(nh)^2, & n = \frac{N}{2}, \frac{N}{2} + 1, \ldots, N. \end{cases} \]

For the matching conditions in the middle we now have [cf. formulae (9.91)]:

\[ c_1 + c_2 \left( \frac{N}{2}h \right) - \frac{1}{6} \left( \frac{N}{2}h \right)^3 = c_3 + c_4 \left( \frac{N}{2}h \right) + \frac{1}{6} \left( \frac{N}{2}h \right)^3 - \frac{1}{2} \left( \frac{N}{2}h \right)^2, \]

\[ c_1 + c_2 \left( \frac{N}{2}h + h \right) - \frac{1}{6} \left( \frac{N}{2}h + h \right)^3 = c_3 + c_4 \left( \frac{N}{2}h + h \right) + \frac{1}{6} \left( \frac{N}{2}h + h \right)^3 - \frac{1}{2} \left( \frac{N}{2}h + h \right)^2, \]  \tag{9.99}

and consequently:

\[ c_1 = 0, \quad c_3 + c_4 - \frac{1}{3} = 0, \]  \tag{9.100}

\[ c_1 + c_2 - \frac{1}{48} - c_3 - c_4 + \frac{5}{48} = 0, \quad c_2 - \frac{1}{8} - c_4 + \frac{3}{8} - \frac{h^2}{3} = 0, \]

where the last equation of (9.100) was obtained by subtracting the first equation of (9.99) from the second equation of (9.99) and subsequently dividing by \( h \).

Solving equations (9.100) we obtain [cf. formula (9.96)]:

\[ c_1 = 0, \quad c_2 = \frac{1}{8} + \frac{h^2}{6}, \quad c_3 = -\frac{1}{24} + \frac{h^2}{6}, \quad c_4 = \frac{3}{8} - \frac{h^2}{6}, \]

and

\[ u_n = \begin{cases} \frac{1}{6}(nh) - \frac{1}{6}(nh)^3 + \frac{h^2}{6}(nh), & n = 0, 1, \ldots, \frac{N}{2} + 1, \\ -\frac{2}{24} + \frac{3}{8}(nh) + \frac{1}{6}(nh)^3 - \frac{1}{2}(nh)^2 \quad & + \frac{h^2}{6}(1 - nh), & n = \frac{N}{2}, \frac{N}{2} + 1, \ldots, N. \end{cases} \]  \tag{9.101}

It is clear that the error between the continuous solution (9.97) and the discrete solution (9.101) is estimated as

\[ \|u_h - u_h^{(k)}\| = \max u(x_n) - u_n = O(h^2), \]
which means that the solution of the finite-difference problem (9.82), (9.98) converges to the solution of the differential problem (9.81), (9.94) with the second order with respect to $h$. Note that second-order convergence is attained here even though the degree of regularity of the solution — third derivative is discontinuous — is formally insufficient to guarantee the second-order accuracy (consistency).

In much the same way one can analyze the case when the right-hand side $f(x)$ has one continuous derivative (the so-called $C^1$ space of functions), for example:

$$f(x) = \begin{cases} -(x - \frac{1}{2})^2, & 0 \leq x \leq \frac{1}{2}, \\ (x - \frac{1}{2})^2, & \frac{1}{2} < x \leq 1. \end{cases} \quad (9.102)$$

For problem (9.81), (9.102), it is also possible to prove the second-order convergence, which is the subject of Exercise 1 at the end of the section.

The foregoing examples demonstrate that the rate of finite-difference convergence depends on the regularity of the solution to the underlying continuous problem. It is therefore interesting to see what happens when the regularity increases beyond $C^1$.

Consider the right-hand side in the form of a quadratic polynomial:

$$f(x) = x(x - 1). \quad (9.103)$$

This function is obviously infinitely differentiable ($C^\infty$ space), and so is the solution $u = u(x)$ of problem (9.81), (9.103), which is given by:

$$u(x) = \frac{1}{12} x + \frac{1}{12} x^4 - \frac{1}{6} x^3. \quad (9.104)$$

Scheme (9.82) with the right-hand side

$$f_n = nh(nh - 1), \quad n = 0, 1, \ldots, N, \quad (9.105)$$

approximates problem (9.81), (9.103) with second-order accuracy. The solution of the finite-difference problem (9.82), (9.105) can be found in the form:

$$u_n = c_1 + c_2(nh) + (nh)^2(A(nh)^2 + B(nh) + C), \quad (9.106)$$

where $u_n^{(h)}$ is the general solution to the homogeneous equation and $u_n^{(p)}$ is a particular solution to the inhomogeneous equation. The values of $A$, $B$, and $C$ are to be found using the method of undetermined coefficients:

$$A \left( \frac{(n + 1)h^4 - 2(nh)^4 + ((n - 1)h)^4}{h^2} \right)$$

$$+ B \left( \frac{(n + 1)h^3 - 2(nh)^3 + ((n - 1)h)^3}{h^2} \right)$$

$$+ C \left( \frac{(n + 1)h^2 - 2(nh)^2 + ((n - 1)h)^2}{h^2} \right) = (nh)^2 - (nh),$$
which yields:

\[ A(12(nh)^2 + 2h^2) + B(6nh) + 2C = (nh)^2 - (nh) \]

and accordingly,

\[ A = \frac{1}{12}, \quad B = -\frac{1}{6}, \quad C = -Ah^2 = -\frac{1}{12}h^2. \quad (9.107) \]

For the constants \(c_1\) and \(c_2\) we substitute the expression (9.106) and the already available coefficients \(A, B,\) and \(C\) of (9.107) into the boundary conditions of (9.82) and write:

\[ u_0 = c_1 = 0, \quad u_N = c_2 + A + B + C = 0. \]

Consequently,

\[ c_1 = 0 \quad \text{and} \quad c_2 = \frac{1}{12} + \frac{1}{12}h^2, \]

so that for the overall solution \(u_n\) of problem (9.81), (9.105) we obtain:

\[ u_n = \frac{1}{12}(nh) + \frac{1}{12}h^2(nh) + (nh)^2 \left( \frac{1}{12}(nh)^2 - \frac{1}{6}(nh) - \frac{1}{12}h^2 \right). \quad (9.108) \]

Comparing the continuous solution \(u(x)\) given by (9.104) with the discrete solution \(u_n\) given by (9.108) we conclude that

\[ ||[u]h - u^h|| = \max_n |u(x_n) - u_n| = O(h^2), \]

which implies that notwithstanding the infinite smoothness of the right-hand side \(f(x)\) of (9.103) and that of the solution \(u(x)\), scheme (9.82), (9.105) still shows only second-order convergence. This is a manifestation of the phenomenon of saturation by smoothness. The rate of decay of the approximation error is determined by the specific approximation method employed on a given grid, and does not reach the level of the pertinent unavoidable error.

To demonstrate that the previous observation is not accidental, let us consider another example of an infinitely differentiable \((C^\infty)\) right-hand side:

\[ f(x) = \sin(\pi x). \quad (9.109) \]

The solution of problem (9.81), (9.109) is given by:

\[ u(x) = -\frac{1}{\pi^2} \sin(\pi x). \quad (9.110) \]

The discrete right-hand side that corresponds to (9.109) is:

\[ f_n = \sin(\pi nh), \quad n = 0, 1, \ldots, N, \quad (9.111) \]
Numerical Solution of Ordinary Differential Equations

and the solution to the finite-difference problem (9.82), (9.111) is to be sought for in the form \( u_n = A \sin(\pi nh) + B \cos(\pi nh) \) with the undetermined coefficients \( A \) and \( B \), which eventually yields:

\[
\begin{align*}
  u_n &= -\frac{h^2}{4}\sin(\frac{\pi}{2})
\end{align*}
\]

The error between the continuous solution given by (9.110) and the discrete solution given by (9.112) is easy to estimate provided that the grid size is small, \( h \ll 1 \):

\[
\begin{align*}
  \|u_h - u^{(h)}\| &= \max_n |u(x_n) - u_n| = \left|\frac{1}{\pi^2} - \frac{h^2}{4\sin^2 \frac{\pi}{2}}\right| \\
  \approx \left|\frac{1}{\pi^2} - \frac{h^2}{4\left[\frac{\pi h}{2} - \frac{1}{6}\left(\frac{\pi h}{2}\right)^3\right]^2}\right| &\approx \left|\frac{1}{\pi^2} - \frac{h^2}{4\left[\frac{\pi h}{2} - \frac{1}{6}\left(\frac{\pi h}{2}\right)^4\right]}\right| \\
  \approx \left|\frac{1}{\pi^2} - \frac{h^2}{4\left(\frac{\pi h}{2}\right)^2}\right| &\approx \frac{1}{\pi^2} + \frac{1}{3}\left(\frac{\pi h}{2}\right)^2 = O(h^2).
\end{align*}
\]

This, again, corroborates the effect of saturation, as the convergence of the scheme (9.82), (9.111) is only second-order in spite of the infinite smoothness of the data.

In general, all finite-difference methods are prone to saturation. This includes the methods for solving ordinary differential equations described in this chapter, as well as the methods for partial differential equations described in Chapter 10. There are, however, other methods for the numerical solution of differential equations. For example, the so-called spectral methods described briefly in Section 9.7 do not saturate and exhibit convergence rates that self-adjust to the regularity of the corresponding solution (similarly to how the error of the trigonometric interpolation adjusts to the smoothness of the interpolated function, see Theorem 3.5 on page 70). The literature on the subject of spectral methods is vast, and we can refer the reader, e.g., to the monographs [GO77] and [CHQZ88].

Exercise

1. Consider scheme (9.82) with the right-hand side:

\[
  f_n = \begin{cases} 
    -\left(\frac{n \pi h}{2} - \frac{\pi}{6}\right)^2, & n = 0, 1, 2, \ldots, N \setminus (N/2) \setminus (N/2 + 1, \ldots, N), \\
    \left(\frac{n \pi h}{2} - \frac{\pi}{6}\right)^2, & n = \frac{N}{2}, \frac{N}{2} + 1, \ldots, N.
  \end{cases}
\]

This scheme approximates problem (9.81), (9.102). Obtain the finite-difference solution in closed form and prove second-order convergence.
9.7 The Notion of Spectral Methods

In this section, we only provide one particular example of a spectral method. Namely, we solve a simple boundary value problem using a Fourier-based technique, with the specific goal in mind — to demonstrate that the discrete approximations to differential equations can be obtained that, unlike the finite-difference methods, will not suffer from the saturation by smoothness (see Section 9.6). A comprehensive account of spectral methods can be found, e.g., in the monographs [GO77] and [CHQZ88], as well as in the recent textbook [GHG06]. The material of this section is based on the analysis of Chapter 3 and can be skipped during the first reading.

Consider the same boundary value problem as in Section 9.6 [cf. formula (9.81)]:

\[ u'' = f(x), \quad 0 \leq x \leq 1, \quad u(0) = 0, \quad u(1) = 0, \] (9.113)

where the right-hand side \( f(x) \) is assumed given. Unlike previously, we will not approximate it on the grid using finite differences. We will rather look for an approximate solution to problem (9.113) in the form of a trigonometric polynomial.

Trigonometric polynomials were introduced and studied in Chapter 3. Let us formally extend both the unknown solution \( u = u(x) \) and the right-hand side \( f = f(x) \) to the interval \([-1, 1]\) antisymmetrically, i.e., \( u(-x) = -u(x) \) and \( f(-x) = -f(x) \), so that the resulting functions be odd. We can then represent the solution \( u(x) \) of problem (9.113) approximately as a trigonometric polynomial:

\[ u^{(n)}(x) = \sum_{k=1}^{n+1} B_k \sin(\pi k x) \] (9.114)

with the coefficients \( B_k \) to be determined. Note that according to Theorem 3.3 (see page 68), the polynomial (9.114), which is a linear combination of the sine functions only, is suited specifically for representing the odd functions. Note also that for any choice of the coefficients \( B_k \) the polynomial \( u^{(n)}(x) \) of (9.114) satisfies the boundary conditions of problem (9.113) exactly.

Let us now introduce the same grid (of dimension \( n + 1 \)) as we used in Section 3.1:

\[ x_m = \frac{1}{n+1} m + \frac{1}{2(n+1)}, \quad m = 0, 1, \ldots, n, \] (9.115)

and interpolate the given function \( f(x) \) on this grid by means of the trigonometric polynomial with \( n + 1 \) terms [cf. formula (3.27)]:

\[ f^{(n)}(x) = \sum_{k=1}^{n+1} b_k \sin(\pi k x). \] (9.116)
Numerical Solution of Ordinary Differential Equations

The coefficients of the polynomial (9.116) are given by [cf. formulae (3.28), (3.29)]:

\[ b_k = \frac{2}{n+1} \sum_{m=0}^{n} f(x_m) \sin k \left( \frac{\pi}{n+1} m + \frac{\pi}{2(n+1)} \right), \quad k = 1, 2, \ldots, n, \]  
\[ b_{n+1} = \frac{1}{n+1} \sum_{m=0}^{n} f(x_m) (-1)^m. \]  
(9.117)

To approximate the differential equation \( u'' = f \) of (9.113), we require that the second derivative of the approximate solution \( u^{(n)}(x) \):

\[ \frac{d^2}{dx^2} u^{(n)}(x) = -\pi^2 \sum_{k=1}^{n+1} B_k k^2 \sin(\pi k x) \]  
(9.118)

coincide with the interpolant of the right-hand side \( f^{(n)}(x) \) at every node \( x_m \) of the grid (9.115):

\[ \frac{d^2}{dx^2} u^{(n)}(x_m) = f^{(n)}(x_m), \quad m = 0, 1, \ldots, n. \]  
(9.119)

Note that both the interpolant \( f^{(n)}(x) \) given by formula (9.116) and the derivative \( \frac{d^2}{dx^2} u^{(n)}(x) \) given by formula (9.118) are sine trigonometric polynomials of the same order \( n+1 \). According to formula (9.119), they coincide at \( x_m \) for all \( m = 0, 1, \ldots, n \). Therefore, due to the uniqueness of the trigonometric interpolating polynomial (see Theorem 3.1 on page 64), these two polynomials are, in fact, the same everywhere on the interval \( 0 \leq x \leq 1 \). Consequently, their coefficients are identically equal:

\[ -\pi^2 k^2 B_k = b_k, \quad k = 1, 2, \ldots, n+1. \]  
(9.120)

Equalities (9.120) allow one to find \( B_k \) provided that \( b_k \) are known.

Consider a particular example analyzed in the end of Section 9.6:

\[ f(x) = \sin(\pi x). \]  
(9.121)

The exact solution of problem (9.113), (9.121) is given by:

\[ u(x) = \frac{1}{\pi^2} \sin(\pi x). \]  
(9.122)

According to formulae (9.117), the coefficients \( b_k \) that correspond to the right-hand side \( f(x) \) given by (9.121) are:

\( b_1 = 1 \) and \( b_k = 0, \quad k = 2, 3, \ldots, n+1. \)

Consequently, relations (9.120) imply that

\[ B_1 = -\frac{1}{\pi^2} \quad \text{and} \quad B_k = 0, \quad k = 2, 3, \ldots, n+1. \]

Therefore,

\[ u^{(n)}(x) = -\frac{1}{\pi^2} \sin(\pi x). \]  
(9.123)
By comparing formulae (9.123) and (9.122), we conclude that the approximate method based on enforcing the differential equation $u'' = f$ via the finite system of equalities (9.119) reconstructs the exact solution of problem (9.113), (9.121). The error is therefore equal to zero. Of course, one should not expect that this ideal behavior of the error will hold in general. The foregoing particular result only takes place because of the specific choice of the right-hand side (9.121). However, in a variety of other cases one can obtain a very rapid decay of the error as $n$ increases.

Consider the odd function $f(-x) = -f(x)$ obtained on the interval $[-1, 1]$ by extending the right-hand side of problem (9.113) antisymmetrically from the interval $[0, 1]$. Assume that this function can also be translated along the entire real axis:

$$\forall x \in [2l+1, 2(l+1) + 1]: f(x) = f(x - 2(l+1)), \quad l = 0, 1, \pm 2, \pm 3, \ldots$$

so that the resulting periodic function with the period $L = 2$ be smooth. More precisely, we require that the function $f(x)$ constructed this way possess continuous derivatives of order up to $r > 0$ everywhere, and a square integrable derivative of order $r + 1$:

$$\int_{-1}^{1} \left[f^{(r+1)}(x)\right]^2 dx < \infty.$$ 

Clearly the function $f(x) = \sin(\pi x)$, see formula (9.121), satisfies these requirements. Another example which, unlike (9.121), leads to a full infinite Fourier expansion is $f(x) = \sin(\pi \sin(\pi x))$. Both functions are periodic with the period 2 and infinitely smooth everywhere ($r = \infty$).

Let us represent $f(x)$ as the sum of its sine Fourier series:

$$f(x) = \sum_{k=1}^{\infty} \beta_k \sin(k\pi x), \quad (9.124)$$

where the coefficients $\beta_k$ are defined as:

$$\beta_k = 2 \int_{0}^{1} f(x) \sin(k\pi x) dx. \quad (9.125)$$

The series (9.124) converges to the function $f(x)$ uniformly and absolutely. The rate of convergence was obtained when proving proving Theorem 3.5, see pages 70–73. Namely, if we define the partial sum $S_n(x)$ and the remainder $\delta S_n(x)$ of the series (9.124) as done in Section 3.1.3:

$$S_n(x) = \sum_{k=1}^{n+1} \beta_k \sin(k\pi x), \quad \delta S_n(x) = \sum_{k=n+2}^{\infty} \beta_k \sin(k\pi x), \quad (9.126)$$

then

$$|f(x) - S_n(x)| = |\delta S_n(x)| \leq \frac{\zeta_n}{n^{r+\frac{1}{2}}}, \quad (9.127)$$

where $\zeta_n$ is a numerical sequence such that $\zeta_n = o(1)$, i.e., $\lim_{n \to \infty} \frac{\zeta_n}{n^{r+\frac{1}{2}}} = 0$. Substituting the expressions:

$$f(x_m) = S_n(x_m) + \delta S_n(x_m), \quad m = 0, 1, \ldots, n,$$
Numerical Solution of Ordinary Differential Equations

into the definition (9.117) of the coefficients $b_k$ we obtain:

$$b_k = \frac{2}{n+1} \sum_{m=0}^{n} S_n(x_m) \sin(\pi k x_m) + \frac{2}{n+1} \sum_{m=0}^{n} \delta S_n(x_m) \sin(\pi k x_m), \quad k = 1, 2, \ldots, n,$$

(9.128)

$$b_{n+1} = \frac{1}{n+1} \sum_{m=0}^{n} S_n(x_m) (-1)^m + \frac{1}{n+1} \sum_{m=0}^{n} \delta S_n(x_m) (-1)^m.$$

The first sum on the right-hand side of each equality (9.128) is indeed equal to the genuine Fourier coefficient $\beta_k$ of (9.125), $k = 1, 2, \ldots, n+1$, because due to the uniqueness (Theorem 3.1) the partial sum $S_n(x)$ given by (9.126) coincides with its own trigonometric interpolating polynomial for all $0 \leq x \leq 1$. As for the “corrections” to the coefficients, $\delta \beta_k$, they come from the remainder $\delta S_n(x)$ and their magnitude can be easily estimated using inequality (9.127) and formulae (9.117):

$$|\delta \beta_k| \leq \frac{2}{n^2+1}, \quad k = 1, 2, \ldots, n+1.$$

(9.129)

Let us now consider the exact solution $u = u(x)$ of problem (9.113). Given the assumptions made regarding the right-hand side $f = f(x)$, the solution $u$ is also a smooth odd periodic function with the period $L = 2$. It can be represented as its own Fourier series:

$$u(x) = \sum_{k=1}^{\infty} \gamma_k \sin(k \pi x),$$

(9.130)

where the coefficients $\gamma_k$ are given by:

$$\gamma_k = 2 \int_{0}^{1} u(x) \sin(k \pi x) dx.$$  

(9.131)

Series (9.130) converges uniformly. Moreover, the same argument based on the periodicity and smoothness implies that so do the Fourier series for the derivatives $u'(x)$ and $u''(x)$. Consequently, series (9.130) can be differentiated (at least) twice term-wise:

$$u''(x) = -\pi^2 \sum_{k=1}^{\infty} k^2 \gamma_k \sin(k \pi x).$$

(9.132)

As we must have $u'' = f$, then by comparing the expansions (9.132) and (9.124) and using the orthogonality of the trigonometric system, we have:

$$\gamma_k = -\frac{1}{\pi^2 k^2} \beta_k, \quad k = 1, 2, \ldots.$$ 

(9.133)

---

6The first derivative $u'(x)$ will be an even function rather than odd.
Let us now recall that the coefficients \( B_k \) of the approximate solution \( u^{(n)}(x) \) defined by (9.114) are given by formula (9.120). Using the representation \( b_k = \beta_k + \delta \beta_k \), see formula (9.128), and also employing relations (9.133), we obtain:

\[
B_k = -\frac{1}{\pi^2 k^2} b_k \\
= -\frac{1}{\pi^2 k^2} \beta_k - \frac{1}{\pi^2 k^2} \delta \beta_k \\
= \gamma_k - \frac{1}{\pi^2 k^2} \delta \beta_k, \quad k = 1, 2, \ldots, n + 1.
\]

Formula (9.134) will allow us to obtain an error estimate for the approximate solution \( u^{(n)}(x) \). To do so, we first rewrite the Fourier series (9.130) for the exact solution as its partial sum plus the remainder [cf. formula (9.126)]:

\[
u(x) = \tilde{S}_n(x) + \delta \tilde{S}_n(x) = \sum_{k=1}^{n+1} \gamma_k \sin(k\pi x) + \sum_{k=n+2}^{\infty} \gamma_k \sin(k\pi x), \tag{9.135}\]

and obtain the estimate for the convergence rate [cf. formula (9.127)]:

\[
|u(x) - \tilde{S}_n(x)| = |\delta \tilde{S}_n(x)| \leq \frac{\eta_n}{n^{r+\frac{1}{2}}}, \tag{9.136}\]

where \( \eta_n = o(1) \) as \( n \rightarrow \infty \). Note that according to the formulae (9.136) and (9.127), the series (9.130) converges faster than the series (9.124) — with the rates \( o\left(n^{-(r+\frac{1}{2})}\right) \) and \( o\left(n^{-(r+\frac{1}{2})}\right) \), respectively. The reason is that if the right-hand side \( f = f(x) \) of problem (9.113) has \( r \) continuous derivatives and the derivative of order \( r + 1 \) in \( L_2 \), then the solution \( u = u(x) \) to this problem would normally have \( r + 2 \) continuous derivatives and the derivative of order \( r + 3 \) in \( L_2 \).

Next, using equalities (9.114), (9.134), and (9.135) and estimates (9.127) and (9.136), we can write \( \forall x \in [0, 1] \):

\[
|u(x) - u^{(n)}(x)| = \left| \tilde{S}_n(x) + \delta \tilde{S}_n(x) - \sum_{k=1}^{n+1} B_k \sin(\pi k x) \right| \\
= \left| \tilde{S}_n(x) + \delta \tilde{S}_n(x) - \sum_{k=1}^{n+1} \gamma_k \sin(\pi k x) + \sum_{k=n+2}^{n+1} \delta \beta_k \sin(\pi k x) \right| \\
= \left| \delta \tilde{S}_n(x) + \sum_{k=n+2}^{n+1} \frac{\delta \beta_k}{\pi^2 k^2} \sin(\pi k x) \right| \leq |\delta \tilde{S}_n(x)| + \sum_{k=n+2}^{n+1} |\delta \beta_k| \\
\leq \frac{\eta_n}{n^{r+\frac{1}{2}}} + \frac{\xi_n}{n^{r+\frac{1}{2}}} \leq \frac{\sigma_n}{n^{r+\frac{1}{2}}},
\]

where \( \sigma_n \) is another infinitesimal sequence: \( \sigma_n = o(1) \) as \( n \rightarrow \infty \). The key distinctive feature of the error estimate (9.137) is that it provides for a more rapid convergence when the data \( f(x) \) that drive the problem and, accordingly, its solution
Numerical Solution of Ordinary Differential Equations

157

$u(x)$ have higher regularity. In other words, similarly to the original trigonometric interpolation (see Section 3.1.3), the foregoing method of obtaining an approximate solution to problem (9.113) does not get saturated by smoothness. In particular, if the right-hand side $f(x)$ of problem (9.113) has continuous periodic derivatives of all orders, then according to estimate (9.137) the method will converge with a spectral rate, i.e., faster than any inverse power of $n$. That’s why in the literature the methods of this type are commonly referred to as spectral methods.

Let us note that the simple Fourier-based spectral method that we have outlined in this section will only work for smooth periodic functions, i.e., for the functions that withstand smooth periodic extensions. There are many examples of smooth right-hand sides that do not satisfy this constraint, for example the quadratic function $f(x) = x(x - 1)$ used in Section 9.6 see formula (9.103). However, a spectral method can be built for problem (9.113) with this right-hand side as well. In this case, it will be convenient to look for a solution as a linear combination of Chebyshev polynomials, rather than in the form of a trigonometric polynomial (9.114); this approach is similar to Chebyshev-based interpolations discussed in Section 3.2.

Note also that in this section we enforced the differential equation of (9.113) by requiring that the two trigonometric polynomials $\frac{d^2}{dx^2}u^{(n)}(x)$ and $f^{(n)}(x)$ coincide at the nodes of the grid (9.115), see equalities (9.119). In the context of spectral methods, the points $x_m$ given by (9.115) are often referred to as the collocation points, and the corresponding methods are known as the spectral collocation methods. Alternatively, one can use Galerkin approximations for building the spectral methods. The Galerkin method is a very useful and general technique that has many applications in numerical analysis and beyond; we briefly describe it in Section 12.1 when discussing the finite elements.

Similarly to any other method of approximation, one generally needs to analyze the accuracy and stability when designing the spectral methods. Over the recent years, a number of efficient spectral methods have been developed for solving many different problems. For further detail, we refer the reader to [GO77, CHQZ88, GHG06].

Exercise

1. Solve problem (9.113) with the right-hand side $f(x) = \sin(\pi \sin(\pi x))$ on the computer using the Fourier collocation method described in this section. Alternatively, apply the second-order difference method of Section 9.6. Demonstrate experimentally the difference in convergence rates.
Chapter 10

Finite-Difference Schemes for Partial Differential Equations

In Chapter 9, we defined the concepts of convergence, consistency, and stability in the context of finite-difference schemes for ordinary differential equations. We also formulated the theorem that if the scheme is consistent and stable, then the discrete solution converges to the corresponding continuous solution as the grid is refined. This theorem provides a recipe for constructing convergent schemes. Namely, one should start with building consistent schemes and subsequently select those that are stable among them.

We should emphasize that the definitions of convergence, consistency, and stability, as well as the theorem that establishes the relation between them, are quite general. In fact, these three concepts can be introduced and studied for arbitrary functional equations. In Chapter 9, we illustrated them with examples of the schemes for ordinary differential equations and for an integral equation. In the current chapter, we will discuss the construction of finite-difference schemes for partial differential equations, and also consider approaches to the analysis of their stability. Moreover, we will actually prove the theorem that consistent and stable schemes do converge, and see that the proof, in fact, applies to many different setups.

In the context of partial differential equations, we will encounter a number of important and essentially new circumstances as compared to the case of ordinary differential equations. Namely, there will be an extensive variety of grids and applicable approximation techniques; a consistent scheme picked arbitrarily will most often turn out unstable; stability analysis will be rather elaborate and non-trivial, especially in the case of initial-boundary value problems (as opposed to the Cauchy problem); and finally, there may be difficulties associated with the very computation of the finite-difference solution that may require a special effort to overcome.

10.1 Key Definitions and Illustrating Examples

10.1.1 Definition of Convergence

Assume that a continuous (initial-)boundary value problem:

\[ Lu = f \]  \hspace{1cm} (10.1)
is to be solved on some domain \( D \) with the boundary \( \Gamma = \partial D \). To approximately compute the solution \( u \) of problem (10.1) given the data \( f \), one first needs to specify a discrete set of points \( D_h \subset \{ D \cup \Gamma \} \) that is called the grid. Then, one should introduce a linear normed space \( U_h \) of all discrete functions defined on the grid \( D_h \), and subsequently identify the discrete exact solution \([u]_h\) of problem (10.1) in the space \( U_h \). As \([u]_h\) will be a grid function and not a continuous function, it shall rather be regarded as a table of values for the continuous solution \( u \). The most straightforward way to obtain this table is by merely sampling the values of \( u \) at the nodes \( D_h \); in this case \([u]_h\) is said to be the trace (or projection) of the continuous solution \( u \) on the grid.\(^1\)

Since, generally speaking, neither the continuous exact solution \( u \) nor its discrete counterpart \([u]_h\) are known, the key objective is to be able to compute \([u]_h\) approximately. For that purpose one needs to construct a system of algebraic equations

\[
L_h u^{(h)} = f^{(h)}
\]

with respect to the unknown function \( u^{(h)} \in U_h \), such that the convergence would take place of the approximate solution \( u^{(h)} \) to the exact solution \([u]_h\) as the grid is refined:

\[
\| [u]_h - u^{(h)} \|_{U_h} \longrightarrow 0, \quad \text{as} \quad h \longrightarrow 0.
\]

As has been mentioned, the most intuitive way of building the discrete operator \( L_h \) in (10.2) consists of replacing the continuous derivatives contained in \( L \) by the appropriate difference quotients, see Section 10.2.1. In this case the discrete system (10.2) is referred to as a finite-difference scheme. Regarding the right-hand side \( f^{(h)} \) of (10.2), again, the simplest way of obtaining it is to take the trace of the continuous right-hand side \( f \) of (10.1) on the grid \( D_h \).

The notion of convergence (10.3) of the difference solution \( u^{(h)} \) to the exact solution \([u]_h\) can be quantified. Namely, if \( k > 0 \) is the largest integer such that the following inequality holds for the solution \( u^{(h)} \) of the discrete (initial-)boundary value problem (10.2):

\[
\| [u]_h - u^{(h)} \|_{U_h} \leq ch^k, \quad c = \text{const},
\]

then we say that the convergence rate is \( O(h^k) \) or alternatively, that the magnitude of the error, i.e., that of the discrepancy between the approximate solution and the exact solution, has order \( k \) with respect to the grid size \( h \) in the chosen norm \( \| \cdot \|_{U_h} \). Note that the foregoing definitions of convergence (10.3) and its rate, or order (10.4), are basically the same as those for ordinary differential equations, see Section 9.1.2.

Let us also emphasize that the way we define convergence for finite-difference schemes, see formulae (10.3) and (10.4), differs from the traditional definition of convergence in a vector space. Namely, when \( h \longrightarrow 0 \) the overall number of nodes in the grid \( D_h \) will increase, and so will the dimension of the space \( U_h \). As such, \( U_h \) shall, in fact, be interpreted as a sequence of spaces of increasing dimension parameterized by the grid size \( h \), rather than a single vector space with a fixed dimension.

\(^1\)There are many other ways to define \([u]_h\), e.g., those based on integration over the grid cells.
Accordingly, the limit in (10.3) shall be interpreted as a limit of the sequence of
norms in vector spaces that have higher and higher dimension.

The problem of constructing a convergent scheme (10.2) is usually split into two
sub-problems. First, one obtains a scheme that is consistent, i.e., that approximates
problem (10.1) on its solution \( u \) in some specially defined sense. Then, one should
verify that the chosen scheme (10.2) is stable.

10.1.2 Definition of Consistency

To assign a tangible meaning to the notion of consistency, one should first intro-
duce a norm in the linear space \( F_h \) that contains the right-hand side \( f(h) \) of equation (10.2). Similarly to \( U_h, F_h \) should rather be interpreted as a sequence of spaces of in-
creasing dimension parameterized by the grid size \( h \), see Section 10.1.1. We say that
the finite-difference scheme (10.2) is consistent, or in other words, that the discrete
problem (10.2) approximates the differential problem (10.1) on the solution \( u \) of the
latter, if the residual \( \delta f(h) \) that is defined by the following equality:

\[
L_h[u]_h = f(h) + \delta f(h),
\]

i.e., that arises when the exact solution \( [u]_h \) is substituted into the left-hand side of
(10.2), vanishes as the grid is refined:

\[
\|\delta f(h)\|_{F_h} \rightarrow 0, \quad \text{as} \quad h \rightarrow 0.
\]

If, in addition, \( k > 0 \) happens to be the largest integer such that

\[
\|\delta f(h)\|_{F_h} \leq c_1 h^k,
\]

where \( c_1 \) is a constant that does not depend on \( h \), then the approximation is said
to have order \( k \) with respect to the grid size \( h \). In the literature, the residual \( \delta f(h) \)
of the exact solution \( [u]_h \), see formula (10.5), is referred to as the truncation error.
Accordingly, if inequality (10.7) holds, we say that the scheme (10.2) has order of
accuracy \( k \) with respect to \( h \) (cf. Definition 9.1 of Chapter 9).

Let us, for example, construct a consistent finite-difference scheme for the follow-
ing Cauchy problem:

\[
\begin{align*}
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} &= \phi(x,t), \quad -\infty < x < \infty, \quad 0 < t \leq T, \\
u(x,0) &= \psi(x), \quad -\infty < x < \infty.
\end{align*}
\]

Problem (10.8) can be recast in the general form (10.1) if we define

\[
L u = \begin{cases}
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}, & -\infty < x < \infty, \quad 0 < t \leq T, \\
u(x,0), & -\infty < x < \infty,
\end{cases}
\quad f = \begin{cases}
\phi(x,t), & -\infty < x < \infty, \quad 0 < t \leq T, \\
\psi(x), & -\infty < x < \infty.
\end{cases}
\]
We will use a simple uniform Cartesian grid $D_h$ with the nodes given by intersections of the two families of vertical and horizontal equally spaced straight lines:

$$x = x_m = mh, \quad t = t_p = p\tau, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor,$$

(10.9)

where $h > 0$ and $\tau > 0$ are fixed real numbers, and $\lfloor T/\tau \rfloor$ denotes the integer part of $T/\tau$. We will also assume that the temporal grid size $\tau$ is proportional to the spatial size $h$: $\tau = rh$, where $r = \text{const}$, so that in essence the grid $D_h$ be parameterized by only one quantity $h$. The discrete exact solution $[u]_h$ on the grid $D_h$ will be defined in the simplest possible way outlined in Section 10.1.1 — as a trace, i.e., by sampling the values of the continuous solution $u(x,t)$ of problem (10.8) at the nodes (10.9): $[u]_h = \{u(mh, p\tau)\}$.

We proceed now to building a consistent scheme (10.2) for problem (10.8). The value of the grid function $u^{(h)}$ at the node $(x_m, t_p) = (mh, p\tau)$ of the grid $D_h$ will hereafter be denoted by $u_m^p$. To actually obtain the scheme, we replace the partial derivatives $\frac{\partial u}{\partial t}$ and $\frac{\partial u}{\partial x}$ by the first-order difference quotients:

$$\left. \frac{\partial u}{\partial t} \right|_{(x,t)} \approx \frac{u(x, t + \tau) - u(x, t)}{\tau},$$

$$\left. \frac{\partial u}{\partial x} \right|_{(x,t)} \approx \frac{u(x + h, t) - u(x, t)}{h}.$$  

Then, we can write down the following system of equations:

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_m^{p+1} - u_m^p}{h} = q_m^p, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,$$

$$u_m^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots,$$

(10.10)

where the right-hand sides $q_m^p$ and $\psi_m$ are obtained, again, by sampling the values of the continuous right-hand sides $q(x,t)$ and $\psi(x)$ of (10.8) at the nodes of $D_h$:

$$q_m^p = q(mh, p\tau), \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,$$

$$\psi_m = \psi(mh), \quad m = 0, \pm 1, \pm 2, \ldots.$$  

(10.11)

The scheme (10.10) can be recast in the universal form (10.2) if the operator $L_h$ and the right-hand side $f^{(h)}$ are defined as follows:

$$L_{hh}^{(h)} = \left\{ \begin{array}{l}
\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_m^{p+1} - u_m^p}{h}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1, \\
u_m^0, \quad m = 0, \pm 1, \pm 2, \ldots,
\end{array} \right.$$  

$$f^{(h)} = \left\{ \begin{array}{l}
q_m^p, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1, \\
\psi_m, \quad m = 0, \pm 1, \pm 2, \ldots,
\end{array} \right.$$  

where $q_m^p$ and $\psi_m$ are given by (10.11). Thus, the right-hand side $f^{(h)}$ is basically a pair of grid functions $q_m^p$ and $\psi_m$ such that the first one is defined on the two-dimensional grid (10.9) and the second one is defined on the one-dimensional grid:

$$(x_m, 0) = (mh, 0), \quad m = 0, \pm 1, \pm 2, \ldots.$$
Finite-Difference Schemes for Partial Differential Equations

The difference equation from (10.10) can be easily solved with respect to \( u_m^{p+1} \):

\[
    u_m^{p+1} = (1 - r)u_m^p + ru_m^{p+1} + \tau q_m^p.
\]  

(10.12)

Thus, if we know the values \( u_m^p, m = 0, \pm 1, \pm 2, \ldots \), of the approximate solution \( u^{(h)} \) at the grid nodes that correspond to the time level \( t = p\tau \), then we can use (10.12) and compute the values \( u_m^{p+1} \) at the grid nodes that correspond to the next time level \( t = (p + 1)\tau \). As the values \( u_m^0 \) at \( t = 0 \) are given by the equalities \( u_m^0 = \psi_m \), see (10.10), then we can successively compute the discrete solution \( u_m^p \) one time level after another for \( t = \tau, t = 2\tau, t = 3\tau \), etc., i.e., everywhere on the grid \( D_h \). The schemes, for which solution on the upper time level can be obtained as a closed form expression that contains only the values of the solution on the lower time level(s), such as in formula (10.12), are called explicit. The foregoing process of computing the solution one time level after another is known as the time marching.

Let us now determine what order of accuracy does the scheme (10.10) have. The linear space \( F_h \) will consist of all pairs of bounded grid functions \( f^{(h)} = [\psi^m, \psi_m] \) with the norm:\(^2\)

\[
    ||f^{(h)}||_{F_h} = \max_{m, p} |\psi^p_m| + \max_m |\psi_m|.
\]  

(10.13)

We note that one can use various norms for the analysis of consistency, and the choice of the norm can, in fact, make a lot of difference. Hereafter in this section, we will only be using the maximum norm defined by (10.13).

Assume that the solution \( u(x,t) \) of problem (10.8) has bounded second derivatives. Then, the Taylor formula yields:

\[
    \frac{u(x_m + h, t_p) - u(x_m, t_p)}{h} = \frac{\partial u(x_m, t_p)}{\partial x} + \frac{h}{2} \frac{\partial^2 u(x_m + \xi, t_p)}{\partial x^2},
\]

\[
    \frac{u(x_m, t_p + \tau) - u(x_m, t_p)}{\tau} = \frac{\partial u(x_m, t_p)}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u(x_m, t_p + \eta)}{\partial t^2},
\]

where \( \xi \) and \( \eta \) are some numbers that may depend on \( m, p \), and \( h \), and that satisfy the inequalities \( 0 \leq \xi \leq h \), \( 0 \leq \eta \leq \tau \). These formulae allow us to recast the expression

\[
    L_h[u] = \left\{ \begin{array}{l}
    \frac{u(x_m, t_p + \tau) - u(x_m, t_p)}{\tau} - \frac{u(x_m + h, t_p) - u(x_m, t_p)}{h}, \\
    u(x_m, 0),
    \end{array} \right.
\]

in the following form:

\[
    L_h[u] = \left\{ \frac{\partial u}{\partial t} \right\}_{(x_m, t_p)} + \frac{\tau}{2} \frac{\partial^2 u(x_m, t_p + \eta)}{\partial t^2} - \frac{h}{2} \frac{\partial^2 u(x_m + \xi, t_p)}{\partial x^2}.
\]

\(^2\)If either \( \max |\psi^p_m| \) or \( \max |\psi_m| \) is not reached, then the least upper bound, i.e., supremum, \( \sup |\psi^p_m| \) or \( \sup |\psi_m| \) should be used instead in formula (10.13).
or alternatively [cf. formula (10.5)]:

\[ L_h[u]_h = f^{(h)} + \delta f^{(h)}, \]

where

\[
\delta f^{(h)} = \begin{cases} 
\frac{\tau}{2} \frac{\partial^2 u(x_m, t_p + \eta)}{\partial t^2} - \frac{h}{2} \frac{\partial^2 u(x_m + \xi, t_p)}{\partial x^2}, \\
0.
\end{cases}
\]

Consequently [see (10.13)],

\[
\|\delta f^{(h)}\|_{F_h} \leq \frac{h}{2} \left( \sup \left| \frac{\partial^2 u}{\partial t^2} \right| + \sup \left| \frac{\partial^2 u}{\partial x^2} \right| \right). \tag{10.14}
\]

We can therefore conclude that the finite-difference scheme (10.10) renders the first order of accuracy with respect to \( h \) on the solution \( u(x, t) \) of problem (10.8) that has bounded second partial derivatives.

### 10.1.3 Definition of Stability

**DEFINITION 10.1** The scheme (10.2) is said to be stable if one can find \( \delta > 0 \) and \( h_0 > 0 \) such that for any \( h < h_0 \) and any grid function \( \varepsilon^{(h)} \in F_h \): \( \|\varepsilon^{(h)}\| < \delta \), the finite-difference problem

\[ L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)} \tag{10.15} \]

is uniquely solvable, and its solution \( z^{(h)} \) satisfies the following inequality

\[
\|z^{(h)} - u^{(h)}\|_{U_h} \leq c_2 \|\varepsilon^{(h)}\|_{F_h}, \quad c_2 = \text{const}, \tag{10.16}
\]

where \( u^{(h)} \) is the solution of (10.2), and \( c_2 \) does not depend on \( h \) or \( \varepsilon^{(h)} \).

One can also give an alternative definition of stability.

**DEFINITION 10.2** The finite-difference problem (10.2) is called stable if there is an \( h_0 > 0 \) such that for any \( h < h_0 \) and \( f^{(h)} \in F_h \) it is uniquely solvable, and the solution \( u^{(h)} \) satisfies

\[
\|u^{(h)}\|_{U_h} \leq c_2 \|f^{(h)}\|_{F_h}, \quad c_2 = \text{const}, \tag{10.17}
\]

where \( c_2 \) does not depend either on \( h \) or on \( f^{(h)} \).

Definitions 10.1 and 10.2 reproduce, in the context of partial differential equations, the previous Definitions 9.2 and 9.3, respectively, given in Chapter 9 for ordinary differential equations.
**Lemma 10.1**
For the linear operator \( \mathbf{L}_h \), Definition 10.1 and Definition 10.2 are equivalent.

**Proof** We first assume that the scheme (10.2) is stable in the sense of Definition 10.2. By subtracting equality (10.2) from equality (10.15) we obtain

\[
\mathbf{L}_h \left( z^{(h)} - u^{(h)} \right) = \epsilon^{(h)}.
\]

Then, estimate (10.17) would immediately yield inequality (10.16) for an arbitrary \( \epsilon^{(h)} \in \mathcal{F}_h \), which implies stability in the sense of Definition 10.1.

Conversely, let us now assume that the scheme (10.2) is stable in the sense of Definition 10.1. Then, for some \( h_0 > 0, \delta > 0 \), and for arbitrary \( h < h_0 \), \( \epsilon^{(h)} \in \mathcal{F}_h \): \( \| \epsilon^{(h)} \| < \delta \), both equations

\[
\mathbf{L}_h z^{(h)} = f^{(h)} + \epsilon^{(h)} \quad \text{and} \quad \mathbf{L}_h u^{(h)} = f^{(h)}
\]

have unique solutions. Now let \( w^{(h)} = z^{(h)} - u^{(h)} \), and subtract the foregoing two equations from one another. This yields

\[
\mathbf{L}_h w^{(h)} = \epsilon^{(h)},
\]

whereas estimate (10.16) translates into

\[
\| w^{(h)} \|_{\mathcal{U}_h} \leq c_2 \| \epsilon^{(h)} \|_{\mathcal{F}_h}.
\]

By merely changing the notations for the solution and the right-hand side of equation (10.18), one can reformulate the last result as follows. For arbitrary \( f^{(h)} \in \mathcal{F}_h \): \( \| f^{(h)} \| < \delta \), problem (10.2) has a unique solution \( u^{(h)} \).

This solution satisfies estimate (10.17). To establish stability in the sense of Definition 10.2, we, however, need this result to hold not only for all \( f^{(h)} \) that satisfy \( \| f^{(h)} \| < \delta \), but for all other \( f^{(h)} \in \mathcal{F}_h \) as well.

To complete the proof, we invoke a simple argument based on scaling. Let \( \| f^{(h)} \|_{\mathcal{F}_h} > \delta \), and let us justify the unique solvability and estimate (10.17) in this case. Define the two new functions:

\[
\tilde{u}^{(h)} = \frac{\delta}{2 \| f^{(h)} \|_{\mathcal{F}_h}} u^{(h)} \quad \text{and} \quad \tilde{f}^{(h)} = \frac{\delta}{2 \| f^{(h)} \|_{\mathcal{F}_h}} f^{(h)}.
\]

Then, due to the linearity of the problem, we can write down the following equation for \( \tilde{u}^{(h)} \):

\[
\mathbf{L}_h \tilde{u}^{(h)} = \tilde{f}^{(h)},
\]

Moreover, because of (10.19) we have:

\[
\| \tilde{f}^{(h)} \|_{\mathcal{F}_h} = \frac{\delta}{2 \| f^{(h)} \|_{\mathcal{F}_h}} \| f^{(h)} \|_{\mathcal{F}_h} = \frac{\delta}{2} < \delta,
\]

\[
\| f^{(h)} \|_{\mathcal{F}_h} > \delta.
\]
and consequently, equation (10.20) is uniquely solvable, and the following estimate holds for its solution:

\[ \| \tilde{u}^{(h)} \|_{U_h} \leq c_2 \| \tilde{f}^{(h)} \|_{F_h}. \]

Relations (10.19) then imply that problem (10.2) will have a unique solution for any \( f^{(h)} \in F_h \) that will satisfy inequality (10.17), which means stability in the sense of Definition 10.2.

The property of stability can be interpreted as uniform with respect to \( h \) sensitivity of the solution \( u^{(h)} \) to the perturbations \( \varepsilon^{(h)} \) of the right-hand side of problem (10.2). We emphasize that this (asymptotic) uniformity as \( h \to 0 \) is of foremost importance, and that the concept of stability does go far beyond the simple well-posedness of a given problem (10.2) for a fixed \( h \). Stability, in fact, requires that the entire family of these problems parameterized by the grid size be well-posed, and that the well-posedness constant \( c_2 \) in (10.17) be independent of \( h \).

Let us also emphasize that according to the foregoing Definitions 10.1 and 10.2, stability is an intrinsic property of the finite-difference problem. Its formulation does not involve any direct relation to the original continuous problem. For example, even for a perfectly “nice” continuous problem, i.e., uniquely solvable and well-posed, one can obtain both stable and unstable discrete counterparts.\(^3\) Moreover, the property of stability is also formulated independently of either consistency or convergence. However, the following key theorem establishes a fundamental relation between consistency, stability, and convergence.

**THEOREM 10.1**

*If the finite-difference problem (10.2) is consistent and stable, then the approximate solution \( u^{(h)} \) does converge to the exact solution \( [u]_h \), see (10.3), as the grid is refined. Moreover, in so doing the rate of convergence defined by (10.4) coincides with the order of accuracy in the sense of (10.7) that the scheme (10.2) has on the solution \( u \) of problem (10.1).*

The following proof will basically reproduce that of Theorem 9.1 in Chapter 9.

**PROOF**  In equality (10.15), let us set \( z^{(h)} = [u]_h \), which obviously means \( \varepsilon^{(h)} = \delta f^{(h)} \). Then, the stability estimate (10.16) transforms into:

\[ \| [u]_h - u^{(h)} \|_{U_h} \leq c_2 \| \delta f^{(h)} \|_{F_h}, \]

and along with the consistency in the sense of (10.7), this yields:

\[ \| [u]_h - u^{(h)} \|_{U_h} \leq c_2 c_1 h^k \equiv ch^k, \]

\(^3\)If the original continuous problem is ill-posed, then chances for obtaining a stable discretization are slim.
which is equivalent to convergence with the rate $O(h^k)$, see formula (10.4).

Let us now show that the finite-difference scheme (10.10) is stable for $r = \tau / h \leq 1$. To analyze the stability, we will define the norm $\| \cdot \|_{U_h}$ in the space $U_h$ as follows:

$$\| u^{(h)} \|_{U_h} \overset{\text{def}}{=} \max_{p \in p} \sup_{m} |u^p_m|,$$

and the norm $\| \cdot \|_{F_h}$ in the space $F_h$ as follows [cf. formula (10.13)]:

$$\| f^{(h)} \|_{F_h} \overset{\text{def}}{=} \max_{p \in p} \sup_{m} |q^p_m| + \sup_{m} |\psi_m|.$$

In the finite-difference problem (10.10), the right-hand sides $q^p_m$ and $\psi_m$ shall now be interpreted as arbitrary grid functions rather than traces of the continuous right-hand sides $q(x,t)$ and $\psi(x)$ of (10.8) defined in the sense of (10.11). Of course, with these arbitrary right-hand sides, the scheme (10.10) can still be recast in the form (10.12):

$$u^p_m + 1 = (1 - r)u^p_m + ru^p_{m+1} + \tau q^p_m, \quad u^0_m = \psi_m.$$

As $r \leq 1$, then $1 - r \geq 0$. Consequently, the following estimate holds:

$$|u^{p+1}_m| \leq |u^p_m| + \tau \max_{p \in p} |q^p_m| \leq \max_{m} \{ |u^p_m|, |u^p_{m+1}| \} \leq \sup_{m} |u^p_m|.$$

Using this estimate along with representation (10.23), we have

$$\sup_{m} |u^{p+1}_m| \leq \sup_{m} |u^p_m| + \tau \max_{p \in p} |q^p_m|.$$

Note that if $q^p_m \equiv 0$, then inequality (10.24) implies that $\sup_{m} |u^p_m|$ does not increase as $p$ increases.

The right-hand side of inequality (10.24) does not depend on $m$. Therefore, on its left-hand side we can replace $|u^{p+1}_m|$ with $\sup_{m} |u^{p+1}_m|$ and thus obtain

$$\sup_{m} |u^{p+1}_m| \leq \sup_{m} |u^p_m| + \tau \max_{p \in p} |q^p_m|.$$

Similarly, we can obtain the inequalities:

$$\sup_{m} |u^p_m| \leq \sup_{m} |u^{p-1}_m| + \tau \max_{p \in p} |q^p_m|,$$

$$\cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots

By adding all these inequalities and rearranging the terms, we finally arrive at:

$$\sup_{m} |u^{p+1}_m| \leq \sup_{m} |u^0_m| + \tau (p + 1) \max_{p \in p} |q^p_m|.$$
which immediately yields:

\[ \sup_m |u^{p+1}_m| \leq \sup_m |\psi_m| + T \max_p \sup_m |q^P_m| \leq (1 + T) \|f(h)\|_{F_h}. \]

The previous inequality does hold for any \( p \), therefore it will still hold if on its left-hand side instead of \( \sup_m |u^{p+1}_m| \) we write \( \max_p \sup_m |u^{p+1}_m| = \|u(h)\|_{U_h} \):

\[ \|u(h)\|_{U_h} \leq (1 + T) \|f(h)\|_{F_h}. \quad (10.25) \]

Estimate (10.25) implies stability of the finite-difference scheme (10.10) (in the sense of Definition 10.2), because problem (10.23) is obviously uniquely solvable for arbitrary bounded \( \psi_m \) and \( \psi_m \). In so doing, the quantity \( 1 + T \) plays the role of the constant \( c_2 \) in inequality (10.17).

In the previous example, stability (10.25) of the scheme (10.10) along with its consistency (10.14) are sufficient for convergence according to Theorem 10.1. However, consistency of the scheme (10.2) alone, see formula (10.7), is not, generally speaking, sufficient for convergence in the sense of (10.3). Indeed, in Section 3 of Chapter 9, we have constructed an example of consistent but divergent scheme for an ordinary differential equation. For partial differential equations, instability (and as such, inappropriateness) of an arbitrarily selected consistent scheme shall, in fact, be regarded as a general situation, whereas construction of a stable (and, consequently, convergent) scheme appears to be one of the key tasks of a numerical analyst.

Recall, for example, that we have proven stability of the finite-difference scheme (10.10) under the assumption that \( r \leq 1 \). In the case \( r > 1 \) one can easily see that the scheme (10.10) will remain consistent, i.e., it will still provide the \( \mathcal{O}(h) \) accuracy on the solution \( u(x,t) \) of the differential problem (10.8) that has bounded second derivatives. However, the foregoing stability proof will no longer work. Let us show that for \( r > 1 \) the solution \( u(h) \) of the finite-difference problem (10.10) does not, in fact, converge to the exact solution \([u]_h \) (trace of the solution \( u(x,t) \) to problem (10.8) on the grid \( D_h \) of (10.9)). Therefore, there may be no stability either, as otherwise consistency and stability would have implied convergence by Theorem 10.1.

### 10.1.4 The Courant, Friedrichs, and Lewy Condition

In this section, we will prove that if \( r = \tau / h > 1 \), then the finite-difference scheme (10.10) will not converge for a general function \( \psi(x) \). For simplicity, and with no loss of generality (as far as the no convergence argument), we will assume that \( \phi(x,t) \equiv 0 \), which implies \( \psi^P_m = \psi(mh, p\tau) \equiv 0 \). Let us also set \( T = 1 \), and choose the spatial grid size \( h \) so that the point \((0,1)\) on the \((x,t)\) plane be a grid node, or in other words, so that the number

\[ N = \frac{1}{\tau} = \frac{1}{rh} \]

be integer, see Figure 10.1. Using the difference equation in the form (10.12):

\[ u^{p+1}_m = (1 - r)u^p_m + ru^p_{m+1}, \]
one can easily see that the value \( u_0^{p+1} \) of the discrete solution \( u^{(h)} \) at the grid node \((0,1)\) (in this case \( p + 1 = N \)) can be expressed through the values of \( u_0^p \) and \( u_0^0 \) of the solution at the nodes \((0, 1 - \tau)\) and \((h, 1 - \tau)\), respectively, see Figure 10.1.

Likewise, the values of \( u_0^p \) and \( u_0^0 \) are expressed through the values of \( u_0^{p-1} \), \( u_1^{p-1} \), and \( u_2^{p-1} \) at the three grid nodes: \((0, 1 - 2\tau)\), \((h, 1 - 2\tau)\), and \((2h, 1 - 2\tau)\). Similarly, the values of \( u_0^{p-1} \), \( u_1^{p-1} \), and \( u_2^{p-1} \) are, in turn, expressed through the values of the solution at the four grid nodes: \((0, 1 - 3\tau)\), \((h, 1 - 3\tau)\), \((2h, 1 - 3\tau)\), and \((3h, 1 - 2\tau)\). Eventually we obtain that \( u_0^{p+1} \) is expressed through the values of the solution \( u_m^0 = \psi_m \) at the nodes: \((0,0)\), \((h,0)\), \((2h,0)\), ..., \((hT/\tau,0)\). All these nodes belong to the interval:

\[
0 \leq x \leq \frac{hT}{\tau} = \frac{h}{\tau} = \frac{1}{r}
\]

of the horizontal axis \( t = 0 \), see Figure 10.1, where the initial condition

\[
u(x,0) = \psi(x)\]

for the Cauchy problem (10.8) is specified.

Thus, solution \( u_0^p \) to the finite-difference problem (10.10) at the point \((x,t) = (0,1) \iff (m,p) = (0,N)\) does not depend on the values of the function \( \psi(x) \) at the points \( x \) that do not belong to the interval \( 0 \leq x \leq 1/r \). It is, however, easy to see that solution to the original continuous initial value problem:

\[
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T,
\]

\[
u(x,0) = \psi(x), \quad -\infty < x < \infty,
\]

is given by the function

\[
u(x,t) = \psi(x + t).
\]

This function is constant along every characteristic \( x + t = \text{const} \) of the differential equation \( \frac{\partial \psi}{\partial t} - \frac{\partial \psi}{\partial x} = 0 \). In particular, it maintains a constant value along the straight line \( x + y = 1 \), which crosses through the points \((0,1)\) and \((1,0)\) on the plane \((x,t)\), see Figure 10.1. As such, \( u(x,t)|_{(x,t) = (0,1)} = \psi(1) \). Therefore, if \( r > 1 \), one should not, generally speaking, expect convergence of the scheme (10.10).

Indeed, in this case the interval \( 0 \leq x \leq 1/r < 1 \) of the horizontal axis does not contain the point \((1,0)\). Let us assume for a moment that for some particular \( \psi(x) \) the convergence, by accident, does take place. Then, while keeping the values of \( \psi(x) \)
on the interval $0 \leq x \leq 1/r$ unaltered, we can disrupt this convergence by modifying $\psi(x)$ at and around the point $x = 1$. This is easy to see, because the latter modification will obviously affect the value $u(0, 1)$ of the continuous solution, whereas the value $u_0^N$ of the discrete solution at $(0, 1)$ will remain unchanged, as it is fully determined by $\psi(x)$ on the interval $0 \leq x \leq 1/r$. Moreover, one can modify the function $\psi(x)$ at the point $x = 1$ and in its vicinity so that it will stay sufficiently smooth (twice differentiable). Then, the solution $u(x, t) = \psi(x + t)$ will inherit this smoothness, and the scheme (10.10) will therefore remain consistent, see (10.14). Under these conditions, stability of the scheme (10.10) would have implied convergence. As, however, there is no convergence for $r > 1$, there may be no stability either.

The argument we have used to prove that the scheme (10.10) is divergent (and thus unstable) when $r = \tau/h > 1$ is, in fact, quite general. It can be presented as follows.

Assume that some function $\psi$ is involved in the formulation of the original problem, i.e., that it provides all or part of the required input data. Let $P$ be an arbitrary point from the domain of the solution $u$. Let also the value $u(P)$ depend on the values of the function $\psi$ at the points of some region $G_\psi = G_{\psi}(P)$ that belongs to the domain of $\psi$. This means that by modifying $\psi$ in a small neighborhood of any point $Q \in G_\psi(P)$ one can basically alter the solution $u(P)$. The set $G_\psi(P)$ is referred to as the continuous domain of dependence for the solution $u$ at the point $P$. In the previous example, $G_\psi(P)|_{P=(0,1)} = (1,0)$, see Figure 10.1.

Suppose now that the solution $u$ is computed by means of a scheme that we denote $L_\theta h^{(b)}(u) = f^{(b)}$, and that in so doing the value of the discrete solution $u^{(b)}(P)$ at the point $P^{(b)}$ of the grid closest to $P$ is fully determined by the values of the function $\psi$ on some region $G_\psi^{(b)}(P)$. This region is called the discrete domain of dependence, and in the previous example: $G_\psi^{(b)}(P)|_{P=(0,1)} = \{x|0 \leq x \leq 1/r\}$, see Figure 10.1.

The Courant, Friedrichs, and Lewy condition says that for the convergence $u^{(b)} \rightarrow u$ to take place as $h \rightarrow 0$, the scheme should necessarily be designed in such a way that in any neighborhood of an arbitrary point from $G_\psi(P)$ there always be a point from $G_\psi^{(b)}(P)$, provided that $h$ is chosen sufficiently small.

In the literature, this condition is commonly referred to as the CFL condition. The number $r = \tau/h$ that in the previous example distinguishes between the convergence and divergence as $r \leq 1$, is known as the CFL number or the Courant number.

Let us explain why in general there is no convergence if the foregoing CFL condition does not hold. Assume that it is not met so that no matter how small the grid size $h$ is, there are no points from the set $G_\psi^{(b)}(P)$ in some fixed neighborhood of a particular point $Q \in G_\psi(P)$. Even if for some $\psi$ the convergence $u^{(b)} \rightarrow u$ does incidentally take place, we can modify $\psi$ inside the aforementioned neighborhood of $Q$, while keeping it unchanged elsewhere. This modification will obviously cause a change in the value of $u(P)$. At the same time, for all sufficiently small $h$ the values of $u^{(b)}$ at the respective grid nodes $P^{(b)}$ closest to $P$ will remain the same because the function $\psi$ has not changed at the points of the sets $G_\psi^{(b)}(P)$. Therefore, for the new function $\psi$ the scheme may no longer converge.

Note that the CFL condition can be formulated as a theorem, while the foregoing
arguments can be transformed into its proof. We also re-emphasize that for a consistent scheme the CFL condition is necessary not only for its convergence, but also for stability. If this condition does not hold, then there may be no stability, because otherwise consistency and stability would have implied convergence by Theorem 10.1.

10.1.5 The Mechanism of Instability

The proof of instability for the scheme (10.10) given in the previous Section 10.1.4 for the case \( r = \tau / h > 1 \) is based on using the CFL condition that is necessary for convergence and for stability. As such, this proof is non-constructive in nature. It would, however, be very instrumental to actually see how the instability of the scheme (10.10) manifests itself when \( r > 1 \), i.e., how it affects the sensitivity of the solution \( u^{(h)} \) to the perturbations of the data \( f^{(h)} \). Recall that according to Section 10.1.3 the scheme is called stable if the finite-difference solution is weakly sensitive to the errors in \( f^{(h)} \), and the sensitivity is uniform with respect to \( h \).

Assume for simplicity that for all \( h \) the right-hand sides of equations (10.10) are identically equal to zero: \( \phi_{pm}^p \equiv 0 \) and \( \psi_{pm}^p \equiv 0 \), so that \( f^{(h)} = [\phi_{pm}^p \psi_{pm}^p] = 0 \), and consequently, solution \( u^{(h)} = \{u_{pm}^p\} \) of problem (10.10) is also identically equal to zero: \( u_{pm}^p \equiv 0 \). Suppose now that an error has been committed in the initial data, and instead of \( \psi_{pm} = 0 \) a different function \( \tilde{\psi}_{pm} = (−1)^m \varepsilon \) \( (\varepsilon = \text{const}) \) has been specified, which yields:

\[
\tilde{f}^{(h)} = \begin{bmatrix} 0 \\ \tilde{\psi}_{pm} \end{bmatrix}, \quad \|\tilde{f}^{(h)}\|_{L^2} = \varepsilon,
\]

on the right-hand side of (10.10) instead of \( f^{(h)} = 0 \). Let us denote the corresponding solution by \( \tilde{u}^{(h)} \). In accordance with the finite-difference equation:

\[
\tilde{u}_{m}^{p+1} = (1 - r)\tilde{u}_{m}^{p} + r\tilde{u}_{m+1}^{p}, \quad \tilde{u}_{m}^{0} = (-1)^m \varepsilon,
\]

for \( \tilde{u}_{m}^{p} \) we obtain: \( \tilde{u}_{m}^{1} = (1 - r)\tilde{u}_{m}^{0} + r\tilde{u}_{m+1}^{0} = (1 - 2r)\tilde{u}_{m}^{0} \). We thus see that the error committed at \( p = 0 \) has been multiplied by the quantity \( 1 - 2r \) when advancing to \( p = 1 \). For yet another time level \( p = 2 \) we have:

\[
\tilde{u}_{m}^{2} = (1 - 2r)^2\tilde{u}_{m}^{0},
\]

and in general,

\[
\tilde{u}_{m}^{p} = (1 - 2r)^p\tilde{u}_{m}^{0}.
\]

When \( r > 1 \) we obviously have \( 1 - 2r < -1 \), so that each time step, i.e., each transition between the levels \( p \) and \( p + 1 \), implies yet another multiplication of the initial error \( \tilde{u}_{m}^{0} = (-1)^m \varepsilon \) by a negative quantity with the absolute value greater than one. For \( p = [T/\tau] \) we thus obtain:

\[
|\tilde{u}_{m}^{p}| = |1 - 2r|^{[T/\tau]}|\tilde{u}_{m}^{0}|.
\]
Consequently,
\[ \| \tilde{u}^{(h)} \|_{U_h} = \max_p \sup_m |\tilde{u}_m^p| = |1 - 2r|^{T/(rh)} \sup_m \| \tilde{q}_m \| = |1 - 2r|^{T/(rh)} \| \tilde{f}^{(h)} \|_{F_h}. \]

In other words, we see that for a given fixed \( T \) the error \((-1)^m \epsilon \) originally committed when specifying the initial data for the finite-difference problem, grows at a rapid exponential rate \(|1 - 2r|^{T/(rh)}\) as \( h \to 0 \). This is a manifestation of the instability for scheme (10.10) when \( r > 1 \).

10.1.6 The Kantorovich Theorem

For a wide class of linear operator equations, the theory of their approximate solution (not necessarily numerical) can be developed, and the foregoing concepts of consistency, stability, and convergence can be introduced and studied, in a unified general framework. In so doing, the exact equation and the approximate equation should basically be considered in different spaces of functions — the original space and the approximating space (like, for example, the space of continuous functions and the space of grid functions).\(^4\) However, it will also be very helpful to establish a fundamental relation between the consistency, stability, and convergence for the most basic setup, when all the operators are assumed to act in the same space of functions. The corresponding result is known as the Kantorovich theorem [Kan52]; and a number of more specific implications, such as that of Theorem 10.1, can then be considered its exemplifications for the respective fields. Of course, each particular case will still require some non-trivial constructive steps pertaining primarily to the proper definition of the spaces and operators involved. A detailed description of the corresponding developments can be found in [KA82], including a comprehensive analysis of the case when the approximate equation is formulated on a different space (subspace) of functions. Monograph [RM67] provides for an account oriented more toward computational methods. The material of this section is more advanced, and can be skipped during the first reading.

Let \( U \) and \( F \) be two Banach spaces, and let \( L \) be a linear operator: \( L : U \to F \) that has a bounded inverse, \( L^{-1} : F \to U, \| L^{-1} \| < \infty \). In other words, we assume that the problem
\[ Lu = f \] (10.26)
is uniquely solvable for every \( f \in F \) and well-posed.

Let \( L_h : U \to F \) be a family of operators parameterized by some \( h \) (for example, we may have \( h = 1/n, n = 1, 2, 3, \ldots \)). Along with the original problem (10.26), we introduce a series of its “discrete” counterparts:
\[ L_h u^{(h)} = f, \] (10.27)
where \( u^{(h)} \in U \) and each \( L_h \) is also assumed to have a bounded inverse, \( L_h^{-1} : F \to U, \| L_h^{-1} \| < \infty \).

\(^4\)The approximating space often appears isomorphic to a subspace of the original space.
We say that problem (10.27) is consistent, or in other words, that the operators $L_h$ of (10.27) approximate the operator $L$ of (10.26), if for any $u \in U$ we have

$$\|L_h u - Lu\|_F \longrightarrow 0, \quad \text{as } h \longrightarrow 0. \quad (10.28)$$

Note that any given $u \in U$ can be interpreted as solution to problem (10.26) with the right-hand side defined as $F \ni f \overset{\text{def}}{=} Lu$. Then, the general notion of consistency (10.28) becomes similar to the concept of approximation on a solution introduced in Section 10.1.2, see formula (10.6).

Problem (10.27) is said to be stable if all the inverse operators are bounded uniformly:

$$\|L^{-1}_h\| \leq C = \text{const}, \quad (10.29)$$

which means that $C$ does not depend on $h$. This is obviously a stricter condition than simply having each $L^{-1}_h$ bounded; it is, again, similar to Definition 10.2 of Section 10.1.3.

**THEOREM 10.2 (Kantorovich)**

Provided that properties (10.28) and (10.29) hold, solution $u^{(b)}$ of the approximate problem (10.27) converges to the solution $u$ of the original problem (10.26):

$$\|u - u^{(b)}\|_U \longrightarrow 0, \quad \text{as } h \longrightarrow 0. \quad (10.30)$$

**PROOF** Given (10.28) and (10.29), we have

$$\|u - u^{(b)}\|_U \leq \|L^{-1}_h L_h u - L^{-1}_h f\|_F \leq \|L^{-1}_h\| \|L_h u - f\|_F \leq C \|L_h u - f\|_F = C \|L_h u - Lu + Lu - f\|_F = C \|L_h u - Lu\|_F \longrightarrow 0, \quad \text{as } h \longrightarrow 0,$$

because $Lu = f$ and $L_h u^{(b)} = f$.

The foregoing Theorem 10.2 actually asserts one implication of the complete Kantorovich theorem, namely, that consistency and stability imply convergence, cf. Theorem 10.1. The other implication says that if the approximate problem is consistent and convergent, then it is also stable. Altogether it means, that under the assumption of consistency, the notions of stability and convergence are equivalent [Kan52]. The proof of this second implication is based on the additional result known as the principle of uniform boundedness for families of operators, see [RM67]. This proof is generally not difficult, but we omit it here for the reason of not having to introduce the extra notions and concepts from functional analysis.

**10.1.7 On the Efficacy of Finite-Difference Schemes**

Let us now briefly comment on the approach we have previously adopted for assessing the quality of approximation that a finite-difference scheme renders. In
Section 10.1.2, it is characterized by the norm of the truncation error \( \| \delta f^{(h)} \|_{E_h} \), measured as a power of the grid size \( h \), see formula (10.7). As we have seen in Theorem 10.1, for stable schemes the foregoing characterization, known as the order of accuracy, coincides with the (asymptotic) order of the solution error \( \| u_h - u^{(h)} \|_{U_h} \), also referred to as the convergence rate, see formula (10.4). As the latter basically tells us how good the approximate solution is, it is natural to characterize the scheme by the amount of computations required for achieving a given order of error. In its own turn, the amount of computations is generally proportional to the overall number of grid nodes \( N \). For ordinary differential equations \( N \) is inversely proportional to the grid size \( h \). Therefore, when we say that \( \varepsilon \) is the norm of the error, where \( \varepsilon = O(h^k) \), we actually imply that \( \varepsilon = O(N^{-k/2}) \), i.e., that driving the error down by a factor of two would require roughly \( \sqrt{2} \) times as much of the computational effort. Therefore, in the case of ordinary differential equations, the order of accuracy with respect to \( h \) adequately characterizes the required amount of computations.

For partial differential equations, however, this is no longer the case. In the previously analyzed example of a problem with two independent variables \( x \) and \( t \), the grid is defined by the two sizes: \( h \) and \( \tau \). The number of nodes \( N \) that belong to a given bounded region of the plane \( (x,t) \) is obviously of order \( 1/(\tau h) \). Let \( \tau = rh \). In this case \( N = O(h^{-2}) \), and saying that \( \varepsilon = O(h^k) \) is equivalent to saying that \( \varepsilon = O(N^{-k/2}) \). If \( \tau = rh^2 \), then \( N = O(h^{-3}) \), and saying that \( \varepsilon = O(h^k) \) is equivalent to saying that \( \varepsilon = O(N^{-k/3}) \).

We see that in the case of partial differential equations the magnitude of the error may be more natural to quantify using the powers of \( N^{-1} \) rather than those of \( h \). Indeed, this would allow one to immediately determine the amount of additional work (assumed proportional to \( N \)) needed to reduce the error by a prescribed factor. Hereafter, we will nonetheless adhere to the previous way of characterizing the accuracy by means of the powers of \( h \), because it is more convenient for derivations, and it also has some “historical” reasons in the literature. The reader, however, should always keep in mind the foregoing circumstance when comparing the properties of difference schemes and determining their suitability for solving a given problem.

We should also note that our previous statement on the proportionality of the computational work to the number of grid nodes \( N \) does not always hold either. There are, in fact, examples of difference schemes that require \( O(N^{1+\alpha}) \) arithmetic operations for finding the discrete solution, where \( \alpha \) may be equal to 1/2, 1, or even 2. One may encounter situations like that when solving finite-difference problems that approximate elliptic equations, or when solving the problems with three or more independent variables (e.g., \( u(x,y,t) \)).

In the context of real computations, for the comparative assessment of numerical algorithms one usually takes the code execution time as a natural measure of the algorithm quality. The execution time is not necessarily proportional to the number of the floating point arithmetic operations. It may also be affected by the integer, symbolic, or logical operations. There are other factors that may play a role, for example, the so-called memory bandwidth that determines the rate of data exchange between the computer memory and the CPU. For multiprocessor computer platforms with distributed memory, the overall algorithm efficacy is to a large extent deter-
mined by how efficiently the data can be exchanged between different blocks of the computer memory. All these considerations shall obviously be taken into account when choosing the method and when subsequently interpreting its results.

10.1.8 Bibliography Comments

The notion of stability for finite-difference schemes was first introduced by von Neumann and Richtmeyer in a 1950 paper [VNR50] that discusses the computation of gasdynamic shocks. In that paper, stability was treated as sensitivity of finite-difference solutions to perturbations of the input data, in the sense of whether or not the initial errors will get amplified as the time elapses. No attempt was made at proving convergence of the corresponding approximation.

In work [OHK51], O’Brien, Hyman, and Kaplan studied finite-difference schemes for the heat equation. They used the apparatus of finite Fourier series, see Section ???, to analyze the sensitivity of their finite-difference solutions to perturbations of the input data. In modern terminology, the analysis of [OHK51] can be qualified as a spectral study of stability for a particular case, see Section 10.3.

The first comprehensive systems of definitions for stability and consistency that enabled obtaining convergence as their implication, were proposed by Ryaben’kii in 1952, see [Rya52], and by Lax in 1953, see [RM67, Chapter 3].

Ryaben’kii in work [Rya52] analyzed the Cauchy problem for linear partial differential equations with coefficients that may only depend on \( t \). He derived necessary and sufficient conditions for stability of finite-difference approximations in the sense of the definition that he introduced in the same paper. He also built constructive examples of stable finite-difference schemes for the systems that are hyperbolic or \( p \)-parabolic in the sense of Petrowsky.

Lax, in his 1953 system of definitions, see [RM67, Chapter 3], considered finite-difference schemes for time-dependent operator equations. His assumption was that these schemes operate in the same Banach spaces of functions as the original differential operators do. The central result of the Lax theory is known as the equivalence theorem. It says that if the original continuous problem is uniquely solvable and well-posed, and if its finite-difference approximation is consistent, then stability of the approximation is necessary and sufficient for its convergence.

The system of basic definitions adopted in this book, as well as the form of the key theorem that stability and consistency imply convergence, are close to those proposed by Filippov in 1955, see [Fil55] and also [RF56]. The most important difference is that we use a more universal definition of consistency compared to the one by Filippov, [Fil55]. We emphasize that, in the approach by Filippov, the analysis is conducted entirely in the space of discrete functions. It allows for all types of equations, not necessarily time-dependent, see [RF56]. In general, the framework of [Fil55,RF56] can be termed as somewhat more universal and as such, less constructive, than the previous more narrow definitions by Lax and by Ryaben’kii.

Continuing along the same lines, the Kantorovich theorem of Section 10.1.6 can, perhaps, be interpreted as the “ultimate” generalization of all the results of this type.
Let us also note that in the 1928 paper by Courant, Friedrichs, and Lewy [CFL28], as well as in many other papers that employ the method of finite differences to prove existence of solutions to differential equations, the authors establish inequalities that could nowadays be interpreted as stability with respect to particular norms. However, the specific notion of finite-difference stability has rather been introduced in the context of using the schemes for the approximate computation of solutions, under the assumption that those solutions already exist. Therefore, stability in the framework of approximate computations is usually studied in weaker norms than those that would have been needed for the existence proofs.

In the literature, the method of finite differences has apparently been first used for proving existence of solutions to partial differential equations by Lyusternik in 1924, see his later publication [Lyu40]. He analyzed the Laplace equation.

Exercises

1. For the Cauchy problem (10.8), analyze the following finite-difference scheme:

\[
\frac{u_{m+1}^p - u_m^p}{\tau} + \frac{u_m^p - u_{m-1}^p}{h} = \frac{\varphi_m^p}{}, \quad m = 0, 1, \pm, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,
\]

\[u_m^0 = \psi_m, \quad m = 0, 1, \pm, \ldots,\]

where \(\tau = rh\) and \(r = \text{const.}\)

a) Explicitly write down the operator \(L_h\) and the right-hand side \(f(h)\) that arise when recasting the foregoing scheme in the form \(L_h u^{(h)} = f(h)\).

b) Depict graphically the stencil of the scheme, i.e., the locations of the three grid nodes with respect to one another, for which the difference equation connects the corresponding values of \(u^{(h)}\).

c) Show that the finite-difference scheme is consistent and has first-order accuracy with respect to \(h\) on a solution \(u(x,t)\) of the original differential problem that has bounded second derivatives.

d) Find out whether the scheme is stable for any choice of \(r\).

2. For the Cauchy problem:

\[
\frac{u_t + u_x}{x} = \varphi(x,t), \quad -\infty < x < \infty, \quad u(x,0) = \psi(x), \quad 0 \leq t \leq T,
\]

analyze the following two difference schemes according to the plan of Exercise 1:

\[
\frac{u_{m+1}^p - u_m^p}{\tau} + \frac{u_m^p - u_{m-1}^p}{h} = \frac{\varphi_m^p}{}, \quad u_m^0 = \psi_m, \quad m = 0, 1, \pm, \ldots
\]

\[u_{m+1}^p - u_m^p = \frac{\varphi_m^p}{}, \quad u_m^0 = \psi_m.
\]

3. Consider a Cauchy problem for the heat equation:

\[
\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \varphi(x,t), \quad -\infty < x < \infty, \quad 0 \leq t \leq T,
\]

\[u(x,0) = \psi(x), \quad -\infty < x < \infty, \quad (10.31)
\]
4. Solution of the Cauchy problem (10.31) for the one-dimensional homogeneous heat equation, \( \psi(x,t) \equiv 0 \), is given by the Poisson integral:

\[
\psi(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} \psi(\xi)e^{-\frac{(x-\xi)^2}{4t}} \, d\xi.
\]

See whether or not it is possible that a consistent explicit scheme (10.32) would also be stable and therefore convergent if \( \tau = h, h \rightarrow 0 \).

Hint. Compare the domains of dependence for the continuous and discrete problems, and use the Courant, Friedrichs, and Lewy condition.

5. The acoustics system of equations:

\[
\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x}, \quad \frac{\partial w}{\partial t} = \frac{\partial v}{\partial x}, \quad -\infty < x < \infty, \quad 0 \leq t \leq T,
\]

\[
v(x,0) = \psi(x), \quad w(x,0) = \psi(x), \quad -\infty < x < \infty,
\]

has solutions of the type:

\[
v(x,t) = \frac{\psi(x-t) - \psi(x+t) + \psi(x+t) + \psi(x+t)}{2},
\]

\[
w(x,t) = \frac{-\psi(x-t) + \psi(x-t) + \psi(x+t) + \psi(x+t)}{2}.
\]

Can the following finite-difference scheme converge?

\[
\frac{\nu_{m}^{p+1} - \nu_{m}^{p}}{\tau} + \frac{w_{m+1}^{p} - w_{m}^{p}}{h} = 0, \quad \frac{w_{m+1}^{p+1} - w_{m}^{p}}{\tau} + \frac{v_{m+1}^{p} - v_{m}^{p}}{h} = 0,
\]

\[
m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1,
\]

\[
\nu_{m}^{0} = \psi_{m}, \quad w_{m}^{0} = \psi_{m}, \quad m = 0, \pm 1, \pm 2, \ldots.
\]

Hint. Compare the domains of dependence for the continuous and discrete problems.
The Cauchy problem:
\[
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad t > 0,
\]
\[
u(x,0) = e^{\alpha x}, \quad 0 < \alpha < 2\pi, \quad \alpha = \text{const},
\]
has the following solution:
\[
u(x,t) = e^{\alpha t} e^{\alpha x}.
\]
Solution of the corresponding finite-difference scheme:
\[
u_m^{p+1} - \frac{\nu_m^p - \nu_m^{p-1}}{h} = 0, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots,
\]
\[
u_m^0 = e^{\alpha (mh)}, \quad m = 0, \pm 1, \pm 2, \ldots,
\]
is given by (verify that):
\[
u_m^p = (1 - r + re^{inh})^p e^{\alpha (mh)}.
\]
For \(p = t/\tau\) and \(m = x/h\), this discrete solution converges to the foregoing continuous solution as \(h \rightarrow 0\) for any fixed value of \(r = \tau/h\). At the same time, when \(r > 1\) the difference scheme does not satisfy the Courant, Friedrichs, and Lewy condition, which is necessary for stability. Explain the apparent paradox.

### 10.2 Construction of Consistent Difference Schemes

#### 10.2.1 Replacement of Derivatives by Difference Quotients

Perhaps the simplest approach to constructing finite-difference schemes that would approximate the corresponding differential problems is based on replacing the derivatives in the original equations by appropriate difference quotients. Below, we illustrate this approach with several examples. In these examples, we use the following approximation formulae, assuming that \(h\) is small:

\[
\frac{df(x)}{dx} \approx \frac{f(x+h) - f(x)}{h}, \quad \frac{df(x)}{dx} \approx \frac{f(x) - f(x-h)}{h},
\]
\[
\frac{d^2 f(x)}{dx^2} \approx \frac{f(x+h) - 2f(x) + f(x-h)}{2h^2}, \quad \frac{d^2 f(x)}{dx^2} \approx \frac{f(x) - 2f(x) + f(x-h)}{h^2},
\]

(10.33)

Suppose that the function \(f = f(x)\) has sufficiently many bounded derivatives. Then, the approximation error for equalities (10.33) can be estimated similarly to how it has been done in Section 9.2.1 of Chapter 9. Using the Taylor formula, we can write:

\[
f(x + h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + \frac{h^3}{3!} f'''(x) + \frac{h^4}{4!} f^{(4)}(x) + o(h^4),
\]
\[
f(x - h) = f(x) - hf'(x) + \frac{h^2}{2!} f''(x) - \frac{h^3}{3!} f'''(x) + \frac{h^4}{4!} f^{(4)}(x) + o(h^4).
\]

(10.34)
We will build and analyze three temporal grid size sections of equally spaced horizontal and vertical straight lines. In so doing, the of approximate equalities (10.33), one can obtain the expressions for the error:
\[
\frac{f(x+h) - f(x)}{h} = f'(x) + \frac{h}{2} f''(x) + o(h),
\]
\[
\frac{f(x) - f(x-h)}{h} = f'(x) - \frac{h}{2} f''(x) + o(h),
\]
\[
\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{h^2}{6} f'''(x) + o(h^2),
\]
\[
\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = f''(x) + \frac{h^2}{12} f^{(4)}(x) + o(h^2).
\] (10.35)

The error of each given approximation formula in (10.33) is the term in rectangular brackets on the right-hand side of the corresponding equality (10.35).

It is clear that formulae of type (10.33) and (10.35) can also be used for approximating partial derivatives with difference quotients. For example:
\[
\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x,t+\tau) - u(x,t)}{\tau},
\]
where
\[
\frac{u(x,t+\tau) - u(x,t)}{\tau} = \frac{\partial u(x,t)}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u(x,t)}{\partial t^2} + o(\tau).
\]

Similarly, the following formula holds:
\[
\frac{\partial u(x,t)}{\partial x} \approx \frac{u(x+h,t) - u(x,t)}{h},
\]
and
\[
\frac{u(x+h,t) - u(x,t)}{h} = \frac{\partial u(x,t)}{\partial x} + \frac{h}{2} \frac{\partial^2 u(x,t)}{\partial x^2} + o(h).
\]

**Example 1**

Let us consider again the Cauchy problem (10.8) of Section 10.1.2:
\[
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \varphi(x,t), \quad -\infty < x < \infty, \quad 0 < t \leq T,
\]
\[
u(x,0) = \psi(x), \quad -\infty < x < \infty.
\]

We will build and analyze three finite-difference schemes that approximate this problem. For all these schemes, we will use a uniform Cartesian grid \(D_h\) on the plane \((x,t)\). The grid is defined by formula (10.9), and its nodes are obtained as intersections of equally spaced horizontal and vertical straight lines. In so doing, the temporal grid size \(\tau\) will be assumed proportional to the spatial grid size \(h\): \(\tau = rh\), where \(r\) is a positive constant.
Theoretical Introduction to Numerical Analysis

The simplest scheme has, in fact, already been constructed in Section 10.1.2. It is defined by formulœ (10.10), (10.11):

\[
\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = q_m^p, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,
\]

\[
u_m^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots,
\]

and has been obtained by using the following difference quotients in place of the derivatives:

\[
\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x,t+\tau) - u(x,t)}{\tau}, \quad \frac{\partial u(x,t)}{\partial x} \approx \frac{u(x+h,t) - u(x,t)}{h}.
\]

The truncation error \(\delta f^{(h)}\) of scheme (10.10), i.e., the residual of the exact solution \([u]_h\) once it has been substituted into the finite-difference equation: \(L_h[u]_h = f^{(h)} + \delta f^{(h)}\), is given by the expression (see the analysis on pages 163–164):

\[
\delta f^{(h)} = \begin{cases}
\frac{\tau}{2} \frac{\partial^2 u(x_m,t_p)}{\partial t^2} - \frac{h}{2} \frac{\partial^2 u(x_m,t_p)}{\partial x^2} + O(\tau^2 + h^2), \\
0.
\end{cases}
\]

Let us use the same maximum norm (10.22) for the elements \(f^{(h)}\) of the space \(F_h\), as we have used previously in Section 10.1. Then, it is easy to see that

\[
\|\delta f^{(h)}\|_{F_h} = O(\tau + h) = O(rh + h) = O(h),
\]

and consequently the scheme has first order of accuracy. In the literature, scheme (10.10) is often referred to as the first-order upwind, because the direction of the spatial differencing forward \(\frac{u(x+h,t) - u(x,t)}{h}\) is opposite of the advection governed by the differential equation \(u'_t - u'_x = 0\).

The second scheme differs from the first one precisely in this respect — instead of the spatial differencing forward we use differencing backward:

\[
\frac{\partial u(x,t)}{\partial x} \approx \frac{u(x,t) - u(x-h,t)}{h}.
\]

The overall scheme then becomes (see Exercise 1 of Section 10.1):

\[
\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_m^p - u_{m-1}^p}{h} = q_m^p, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,
\]

\[
u_m^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots,
\]

and it also has first-order accuracy. Naturally, it is sometimes referred to as the first-order downwind scheme.

At a first glance, the difference between the upwind and the downwind schemes is minute. However, it turns out that the downwind scheme cannot be used for computations at all, because it violates the Courant, Friedrichs, and Lewy condition of Section 10.1.4 for any \(r = \tau/h\).
Note also that the notions of upwind and downwind pertain to the particular operator \( L = u_t - u_x \). If we were to change the sign and consider \( L = u_t + u_x \) instead, then the upwind and downwind schemes would accordingly change places.

The third scheme is known as the leap-frog scheme:

\[
\begin{align*}
\frac{u^{p+1}_m - u^{p-1}_m}{2\tau} - \frac{u^p_{m+1} - u^p_{m-1}}{2h} &= q^p_m, & m &= 0, \pm 1, \pm 2, \ldots, \quad p = 1, 2, \ldots, [T/\tau] - 1, \\
u^0_m &= \psi_m, & u^1_m &= \psi_m + \left[ \psi'(x_m) + q^0_m \right] \tau, & m &= 0, \pm 1, \pm 2, \ldots \tag{10.36}
\end{align*}
\]

It has overall second order of accuracy, see Exercise 1. Note that as we are using differences of the second order with respect to time in (10.36), the scheme requires one more initial condition, i.e., specification of \( u^1_m, m = 0, \pm 1, \pm 2, \ldots \), in addition to the specification of \( u^0_m \). Of course, a fair degree of flexibility exists in how we can actually specify \( u^1_m \). In doing so, the second order of accuracy needs to be maintained, and we therefore employ two terms of the Taylor expansion along with the original differential equation and its initial condition:

\[
u(x, \tau) = u(x, 0) + u'(x, 0) \tau + O(\tau^2)
\]

\[
= u(x, 0) + \left[ u'_t(x, 0) + \psi(x, 0) \right] \tau + O(\tau^2)
\]

\[
= \psi(x) + \left[ \psi'(x) + \psi(x, 0) \right] \tau + O(\tau^2).
\]

For a given scheme, the subset of the grid nodes connected by the corresponding finite-difference equation for any fixed \( m \) and \( p \) is called the stencil. It is often convenient to depict the stencils graphically; for the three schemes we have just considered this is done in Figure 10.2.

**Example 2**

Next, we will describe two finite-difference schemes that approximate the Cauchy problem (10.31) for the heat equation (see Exercise 3 of Section 10.1):

\[
\begin{align*}
\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} &= \psi(x, t), & -\infty < x < \infty, \quad 0 \leq t \leq T, \\
u(x, 0) &= \psi(x), & -\infty < x < \infty.
\end{align*}
\]
The simplest scheme is defined by formula (10.32), and in the operator form $L_h^{(1)} u^{(h)} = f^{(h)}$ we have:

$$
L_h^{(1)} u^{(h)} = \begin{cases}
\frac{u_{m+1}^n - u_m^n}{\tau}, & m = 0, \pm 1, \pm 2, \ldots, \\
\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, & p = 0, 1, \ldots, [T/\tau] - 1, \\
u_m^0, & m = 0, \pm 1, \pm 2, \ldots
\end{cases}
$$

$$
f^{(h)} = \begin{cases}
q_{m-1}^0 = \varphi(mh, p\tau), & m = 0, \pm 1, \pm 2, \ldots, \\
q_{m-1}^p = \varphi(mh), & m = 0, \pm 1, \pm 2, \ldots
\end{cases}
$$

Scheme (10.37) is obtained with the help of the approximate formulae:

$$
\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x,t + \tau) - u(x,t)}{\tau},
$$

$$
\frac{\partial^2 u(x,t)}{\partial x^2} \approx \frac{u(x + h,t) - 2u(x,t) + u(x - h,t)}{h^2}.
$$

Alternatively, we could write:

$$
\frac{\partial u(x,t + \tau)}{\partial t} \approx \frac{u(x,t + \tau) - u(x,t)}{\tau},
$$

$$
\frac{\partial^2 u(x,t + \tau)}{\partial x^2} \approx \frac{u(x + h,t + \tau) - 2u(x,t + \tau) + u(x - h,t + \tau)}{h^2}.
$$

These formulae lead to a new scheme $L_h^{(2)} u^{(h)} = f^{(h)}$ for the same Cauchy problem (10.31):

$$
L_h^{(2)} u^{(h)} = \begin{cases}
\frac{u_{m+1}^n - u_m^n}{\tau}, & m = 0, \pm 1, \pm 2, \ldots, \\
\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, & p = 0, 1, \ldots, [T/\tau] - 1, \\
u_m^0, & m = 0, \pm 1, \pm 2, \ldots
\end{cases}
$$

$$
f^{(h)} = \begin{cases}
q_{m-1}^{p+1} = \varphi(mh, (p+1)\tau), & m = 0, \pm 1, \pm 2, \ldots, \\
q_{m-1}^p = \varphi(mh), & m = 0, \pm 1, \pm 2, \ldots
\end{cases}
$$

At a first glance the two schemes, (10.37) and (10.38), are not that much different from one another. In practice, however, they differ substantially.

The solution $u^{(h)}$ of the finite-difference system $L_h^{(1)} u^{(h)} = f^{(h)}$ defined by (10.37) can be easily found one time level after another with the help of the explicit formula:

$$
u_{m+1}^p = (1 - 2r)u_m^p + r(u_{m+1}^p + u_{m-1}^p) + \tau q_{m}^p
$$

that constitutes marching in time; here $r = T/h^2$. This formula is obtained by resolving the finite-difference system $L_h^{(1)} u^{(h)} = f^{(h)}$ with respect to $u_{m+1}^p$. Once the
solution \( u_m^p, m = 0, \pm 1, \pm 2, \ldots \), is known at \( t = t_p = p\tau \), one can immediately compute its values \( u_{m+1}^{p+1}, m = 0, \pm 1, \pm 2, \ldots \), at the next level \( t = t_{p+1} = (p+1)\tau \), etc. For this reason, scheme (10.37) is referred to as explicit.

Solution \( u^{(h)} \) of the second finite-difference system \( L^{(2)}u^{(h)} = f^{(h)} \), which is defined by (10.38), can also be found by time marching, i.e., level \( t_{p+1} \) after level \( t_p \). However, this second scheme does not possess the foregoing convenient property of the first one. Namely, the finite-difference system \( L^{(2)}u^{(h)} = f^{(h)} \) cannot be easily resolved with respect to \( u_{m+1}^p \), and there is no explicit formula that would express the value of \( u_{m+1}^p \) for some fixed \( m \) and \( p \) through the known values of \( u_{m-1}^p, u_m^p, \) and \( u_{m+1}^p \), or perhaps some other known values from the previous time level \( t = t_p \). This is the reason why the scheme \( L^{(2)}u^{(h)} = f^{(h)} \) of (10.38) is referred to as implicit.

The system of finite-difference equations \( L^{(2)}u^{(h)} = f^{(h)} \) cannot be explicitly resolved with respect to \( u_{m+1}^p \) because for a given pair of indices \( m \) and \( p \), the corresponding equation contains not only the unknown \( u_{m+1}^p \), but also two other unknowns: \( u_{m+1}^{p+1} \) and \( u_{m+1}^p \).

Therefore, to determine \( u_{m+1}^p, m = 0, \pm 1, \pm 2, \ldots \), for each \( p \), one would need to solve the entire system of equations with respect to the grid function \( u_{m+1}^p \) of the argument \( m \). Nonetheless, it will later be shown (see Examples 6 and 7 of Section 10.3 and also Section 10.6) that the implicit scheme \( L^{(2)}u^{(h)} = f^{(h)} \) is often more convenient for practical computations compared to the explicit scheme \( L^{(1)}u^{(h)} = f^{(h)} \). The stencils of both schemes are schematically depicted in Figure 10.3.

When \( \tau = rh^2, r = \text{const} \), both schemes have second order of accuracy. For the explicit scheme (10.37), this result is a part of Exercise 3 of Section 10.1. Let us therefore prove second-order accuracy for the implicit scheme (10.38).

Using formulae (10.35), we can write:

\[
L_h^{(2)} [u] = \left\{ \begin{align*}
\frac{\partial u(x_m,t_{p+1})}{\partial t} - \frac{\partial^2 u(x_m,t_{p+1})}{\partial x^2} - \frac{\tau}{2} \frac{\partial^2 u(x_m,t_{p+1})}{\partial t^2} - \frac{\tau}{12} \frac{\partial^4 u(x_m,t_{p+1})}{\partial x^4} + o(\tau + h^2), \\
u(x_m,0).
\end{align*} \right.
\]

Consequently, considering that \( \tau = rh^2 \), we have:

\[
L_h^{(2)} [u] = \left\{ \begin{align*}
\psi(mh,(p+1)\tau) + \mathcal{O}(h^2), \\
\psi(mh) + 0.
\end{align*} \right.
\]
Therefore, for the truncation error \( \delta f^{(h)} = L_h^{(2)} |u|^h_h - f^{(h)} \) we obtain:

\[
\delta f^{(h)} = \begin{cases} 
O(h^2), \\
0,
\end{cases}
\]

which obviously implies that

\[
\| \delta f^{(h)} \|_{F_h} = O(h^2).
\]

**Example 3**

Let us now consider a simple finite-difference scheme that would approximate the inhomogeneous Dirichlet problem of type (5.1) for the Poisson equation on the square \( D = \{(x,y) | 0 < x < 1, 0 < y < 1\} \) with the boundary \( \Gamma = \partial D \):

\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \varphi(x,y), \quad (x,y) \in D,
\]

\[
u|_{\Gamma} = \psi(x,y), \quad (x,y) \in \Gamma.
\]

As we did previously in Section 5.1.3, we will use a uniform Cartesian grid in \( D \) with size \( h \) in either coordinate direction; and we will assume that \( M = 1/h \) is an integer:

\[
D_h = \{(x_m, y_n) | x_m = mh, y_n = nh, m = 0,1,\ldots,M, n = 0,1,\ldots,M\}.
\]

The scheme \( L_h u^{(h)} = f^{(h)} \) is obtained by replacing the derivatives in the Poisson equation by the second-order central differences at every interior node of the grid \( D_h \), i.e., at the nodes with \( m = 1,2,\ldots,M-1 \) and \( n = 1,2,\ldots,M-1 \), see Figure 5.1(a) on page 98. This leads to a scheme written on the five-node symmetric stencil shown in Figure 10.4. As for those nodes of the grid \( D_h \) that happen to be on the boundary \( \Gamma \), we rather impose there the Dirichlet boundary condition, which altogether yields:

\[
L_h u^{(h)} = \begin{cases} 
\frac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{h^2} + \frac{u_{m,n+1} - 2u_{m,n} + u_{m,n-1}}{h^2}, \\
u_{m,n},
\end{cases} \\
(mh,nh) \in D_h, \ (mh,nh) \notin \Gamma,
\]

\[
f^{(h)} = \begin{cases} 
\varphi_{m,n} \equiv \varphi(mh,nh), \quad (mh,nh) \in D_h, \ (mh,nh) \notin \Gamma,
\psi_{m,n} \equiv \psi(mh,nh), \quad (mh,nh) \in \Gamma.
\end{cases}
\]
Then, according to the last formula of (10.35), for the truncation error 
\[ \delta f^{(h)} = L_h[u] - f^{(h)} \] we obtain:

\[
f^{(h)} = \begin{cases} 
\frac{h^2}{12} \left( \frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right) + o(h^2), \\
0.
\end{cases}
\]

Therefore, we can conclude that the proposed central-difference scheme for the Poisson equation 
\[ L_h u^{(h)} = f^{(h)} \] has overall second-order accuracy on those solutions of this equation that possess bounded fourth derivatives.

10.2.2 The Method of Undetermined Coefficients

The schemes that we have introduced and analyzed in the previous Section 10.2.1 have all been obtained by replacing each individual derivative in the corresponding differential equation by an appropriate difference quotient. A more general approach for constructing difference schemes is based on approximating the entire differential operator rather than the individual derivatives. We will illustrate this approach using several examples of the schemes for the Cauchy problem (10.8). Henceforth in this section, we will keep using the notations \( L \) and \( L_h \) for the full differential and difference operators, respectively, i.e., the operators that include the initial and/or boundary conditions. We will also introduce the notations \( \Lambda \) and \( \Lambda_h \) for the corresponding equations by themselves only. As before, the grid \( D_h \) on the region \( \{(x,t) | -\infty < x < \infty, 0 \leq t \leq T\} \) will be assumed Cartesian and uniform:

\[ D_h = \{(x_m,t_p) | x_m = mh, m = 0, \pm 1, \pm 2, \ldots; t_p = p\tau, p = 0, 1, \ldots, \lceil T/\tau \rceil \}. \]

Let us first consider the first-order upwind scheme (10.10), (10.11) on the three-node stencil shown in the left part of Figure 10.2. The finite-difference equation:

\[ \Lambda_h u^{(h)} = \frac{u^p_{m+1} - u^p_m}{\tau} - \frac{u^p_{m+1} - u^p_m}{h} = q^p_m \equiv q(x_m,t_p) \]

employed in this scheme can be re-written with generic notations for the coefficients:

\[ \Lambda_h u^{(h)} = a^0 u^{p+1} + a_0 u^p_m + a_1 u^p_{m+1} = q^p_m. \]

For the moment, we will disregard that we already know the actual scheme (10.10), for which we have:

\[ a^0 = \frac{1}{\tau}, \quad a_0 = \frac{1}{h} - \frac{1}{\tau}, \quad a_1 = -\frac{1}{h}. \]

Instead, we will treat the quantities \( a^0, a_0, \) and \( a_1 \) as undetermined coefficients, and will try and choose them to achieve first-order accuracy, i.e., make sure that the following equality hold:

\[ \Lambda_h[u]_{\mid (x_m,t_p)} = \left( \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} \right)_{\mid (x_m,t_p)} + \mathcal{O}(h). \]
This equality can obviously be re-written as:
\[
\mathbf{A}_h[u]_{h} \big|_{(x_m,t_p)} = \mathbf{A} u \big|_{(x_m,t_p)} + O(h), \quad (10.39)
\]
where
\[
\mathbf{A} u = \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}. \quad (10.40)
\]
According to the Taylor formula, we have:
\[
\begin{align*}
u(x_m,t_p + \tau) &= u(x_m,t_p) + \tau \frac{\partial u(x_m,t_p)}{\partial t} + O(\tau^2), \\
u(x_m + h,t_p) &= u(x_m,t_p) + h \frac{\partial u(x_m,t_p)}{\partial x} + O(h^2).
\end{align*}
\]
Substituting the previous expressions into the right-hand side of the equality:
\[
\mathbf{A}_h[u]_{h} \big|_{(x_m,t_p)} = a^0 u(x_m,t_p + \tau) + a_0 u(x_m,t_p) + a_1 u(x_m + h,t_p),
\]
we obtain:
\[
\mathbf{A}_h[u]_{h} \big|_{(x_m,t_p)} = (a^0 + a_0 + a_1) u(x_m,t_p) + a^0 \tau \frac{\partial u(x_m,t_p)}{\partial t} + a_1 h \frac{\partial u(x_m,t_p)}{\partial x} + O(a^0 \tau^2 + a_1 h^2). \quad (10.41)
\]
Since our objective is to facilitate approximation in the sense of (10.39) by selecting the appropriate coefficients \(a^0\), \(a_0\), and \(a_1\), it will be natural to try and regroup the terms on the right-hand side of equality (10.41) in order to single out the term (10.40). Then, the rest of the terms will altogether yield the remainder of the approximation (i.e., the residual) that is supposed to be small. To single out the term \(\mathbf{A} u\), one can replace either of the derivatives: \(\frac{\partial u}{\partial \tau}\) or \(\frac{\partial u}{\partial x}\), on the right-hand side of (10.41) using the respective identity:
\[
\frac{\partial u}{\partial t} = \mathbf{A} u + \frac{\partial u}{\partial x} \quad \text{or} \quad \frac{\partial u}{\partial x} = \frac{\partial u}{\partial t} - \mathbf{A} u. \quad (10.42)
\]
For definiteness, we will exploit the first formula (10.42).

In addition, we will impose the previously used constraint \(\tau = rh\) on the grid sizes \(\tau\) and \(h\), where \(r\) is a positive constant. Then, equality (10.41) transforms into:
\[
\mathbf{A}_h[u]_{h} \big|_{(x_m,t_p)} = a^0 r h \mathbf{A} u \big|_{(x_m,t_p)} + (a^0 + a_0 + a_1) u(x_m,t_p) + (a^0 r + a_1) h \frac{\partial u(x_m,t_p)}{\partial x} + O(a^0 r^2 h^2 + a_1 h^2). \quad (10.43)
\]
Next, we notice that among all smooth functions \(u = u(x,t)\) one can obviously find those, for which the three quantities \(u, \frac{\partial u}{\partial \tau}, \frac{\partial u}{\partial x}\) will assume arbitrary independent values at any given fixed point \((x,t)\). Consequently, the quantities:
\[
u, \frac{\partial u}{\partial x}, \quad \text{and} \quad \mathbf{A} u \equiv \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \psi(x,t)
\]
By construction, the new coefficients (10.44) instead of the coefficients (10.46) we would have obtained:  

\begin{align*}
\hat{a}^0 &= \frac{1}{h} \left[ \frac{1}{r} + \frac{1}{r} \cdot \mathcal{O}_1(h) \right], \\
\hat{a}_0 &= \frac{1}{h} \left[ \frac{r - 1}{r} + \frac{r - 1}{r} \cdot \mathcal{O}_1(h) - \mathcal{O}_3(h) + h \cdot \mathcal{O}_2(h) \right], \\
\hat{a}_1 &= \frac{1}{h} \left[ -1 - \mathcal{O}_1(h) + \mathcal{O}_3(h) \right].
\end{align*}

By construction, the new coefficients $\hat{a}^0$, $\hat{a}_0$, and $\hat{a}_1$ of (10.46) still guarantee first-order accuracy with respect to $h$ for any $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_3$. At the same time, the relative deviation between the coefficients defined by formulae (10.46) and those given by formulae (10.44) becomes negligibly small as the grid size vanishes:

\[ \lim_{h \to 0} \frac{\hat{a}^0 - a^0}{a^0} = 0, \quad \lim_{h \to 0} \frac{\hat{a}_0 - a_0}{a_0} = 0, \quad \lim_{h \to 0} \frac{\hat{a}_1 - a_1}{a_1} = 0. \]
Therefore, we can basically consider the entire collection of schemes \( (10.45) \) with the coefficients \( \bar{a}_i \), \( \bar{\bar{a}}_0 \), and \( \bar{a}_1 \) given by \((10.46)\), as being equivalent to one and the same first-order upwind scheme \((10.10)\) characterized by a fixed choice of the coefficients according to \((10.44)\). Indeed, any scheme from this collection is consistent with order \( O(h) \), and the smaller the grid size the closer will any such scheme be to the “core” scheme \((10.10)\). Moreover, one can show that the result quoted in Section 9.3.4 of Chapter 9 applies. The differences between the respective coefficients given by formulae \((10.44)\) and \((10.46)\) can be treated as small perturbations that do not ruin the stability and as such, do not affect the convergence.

Hereafter, we will always disregard the foregoing arbitrariness, and instead of considering individual schemes will rather operate with their equivalence classes, making no distinction within the class when all the variation is due to the quantities of type \( \mathcal{O}_1(h), \mathcal{O}_2(h), \) and \( \mathcal{O}_3(h) \) in the definition of the coefficients, see, e.g., \((10.46)\). In doing so, we may not always need to introduce those quantities explicitly, and will set them instead to zero from the very beginning.

Let us now see how one can construct a more general class of schemes \( L_h u^{(h)} = f^{(h)} \) for the same advection problem \((10.8)\). Namely, we will be building explicit schemes on the four-node stencil from Figure 10.3(left) that we have otherwise used for the heat equation. The generic notation for the operator \( L_h \) will now involve four undetermined coefficients [cf. formula \((10.45)\)]:

\[
L_h u^{(h)} = \left\{ \begin{array}{l}
\partial_t \partial_t^0 u_m^{n+1} + a_{-1} u_m^n + a_0 u_m^n + a_1 u_{m+1}^n, \quad \partial_t^3 u_m^n, \\
\psi(x_m), \quad \psi(x_m),
\end{array} \right. \quad f^{(h)} = \left\{ \begin{array}{l}
\partial_t \partial_t^0 u_m^{n+1} + a_{-1} u_m^n + a_0 u_m^n + a_1 u_{m+1}^n, \quad \partial_t^3 u_m^n, \\
\psi(x_m), \quad \psi(x_m),
\end{array} \right.
\]

whereas the right-hand side \( f^{(h)} \) will still be the same as defined by formula \((10.45)\). As before, we will assume that \( r = r h, r = \text{const} \), and will also introduce the notation:

\[
\Lambda_h u^{(h)} = \partial_t^0 u_m^{n+1} + a_{-1} u_m^n + a_0 u_m^n + a_1 u_{m+1}^n.
\]

Then, using the Taylor formula, we can write for any sufficiently smooth function \( u = u(x,t) \):

\[
\Lambda_h [u] \bigg|_{(x_m,t_p)} = \left( \partial_t^0 + a_{-1} + a_0 + a_1 \right) u(x_m,t_p) + \partial_t^0 \partial_t \partial_x u(x_m,t_p) + \left( a_1 - a_{-1} \right) h \partial_x u(x_m,t_p) + \partial_t^2 \partial_x^2 u(x_m,t_p) + \left( a_1 - a_{-1} \right) h \partial_x^2 u(x_m,t_p) + \partial_t^3 \partial_x^3 u(x_m,t_p) + O(h^4). \]

On the right-hand side of the previous equality, we will single out the term \( \Lambda u \) of \((10.40)\) with the help of the first identity of \((10.42)\). This yields:

\[
\Lambda_h [u] \bigg|_{(x_m,t_p)} = \partial_t^0 r \Lambda u \bigg|_{(x_m,t_p)} + \left( \partial_t^0 + a_{-1} + a_0 + a_1 \right) u(x_m,t_p) + \partial_t^0 \partial_t \partial_x u(x_m,t_p) + \left( a_1 - a_{-1} \right) h \partial_x u(x_m,t_p) + \partial_t^2 \partial_x^2 u(x_m,t_p) + \left( a_1 - a_{-1} \right) h \partial_x^2 u(x_m,t_p) + \partial_t^3 \partial_x^3 u(x_m,t_p) + O(h^4). \]

188

*Theoretical Introduction to Numerical Analysis*
Then, assuming that the quantity $\mathcal{O}(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$ is small (this assumption will later be corroborated), we see that in order for the consistency condition (10.39) to hold for the operator $\mathbf{A}_h$ of (10.48), the four coefficients $a^0, a_{-1}, a_0,$ and $a_1$ must satisfy the following three equations:

$$\begin{align*}
    a^0 r h &= 1, \\
    a^0 + a_{-1} + a_0 + a_1 &= 0, \\
    h(a^0 r + a_1 - a_{-1}) &= 0.
\end{align*}$$  \tag{10.50}

The system of linear algebraic equations (10.50) has multiple solutions. One of these solutions:

$$
a^0 = \frac{1}{rh}, \quad a_{-1} = 0, \quad a_0 = \frac{r-1}{rh}, \quad a_1 = -\frac{1}{h}
$$

leads to the previously analyzed upwind scheme (10.10). Another solution

$$
a^0 = \frac{1}{rh}, \quad a_{-1} = \frac{1}{2h}, \quad a_0 = \frac{-1}{rh} = \frac{1}{h} - \frac{1}{r}, \quad a_1 = -\frac{1}{2h}
$$

yields the scheme $L_h u^{(h)} = f^{(h)}$ with the operator and the right-hand side defined as:

$$L_h u^{(h)} = \begin{cases} 
    \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_{m-1}^n}{2h}, & \text{and} \quad f^{(h)} = \begin{cases} 
    q(x_m,t_p), \\
    \psi(x_m).
\end{cases}
\end{cases} \tag{10.51}
$$

Given any solution of system (10.50), one, of course, needs to substitute it into the remainder of formula (10.49) and make sure that it is indeed small. For the previous two solutions that lead to the schemes (10.10) and (10.51), respectively, this substitution yields the remainder:

$$\frac{a^0 r^2 h^2}{2} \frac{\partial^2 u}{\partial t^2} + \frac{a_1 + a_{-1} h^2}{2} \frac{\partial^2 u}{\partial x^2} + \mathcal{O}(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$$  \tag{10.52}

that has order $\mathcal{O}(h)$. Indeed, as the expressions for the coefficients $a^0, a_{-1}, a_0,$ and $a_1$ all contain the grid size $h$ in the denominator, we conclude that the first two terms in the previous sum have order $\mathcal{O}(h)$, while the quantity $\mathcal{O}(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$ is, in fact, of order $\mathcal{O}(h^2)$.

In general, among the smooth functions $u = u(x,t)$ there are obviously polynomials of the second degree, for which the derivatives $\frac{\partial^2 u}{\partial t^2}$ and $\frac{\partial^2 u}{\partial x^2}$ can assume arbitrary independent values at any fixed point $(x,t)$. In so doing, the term $\mathcal{O}(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$ that contains the third derivatives of the polynomials vanishes. Therefore, in order to guarantee that the order of the residual (10.52) [remainder of the approximation (10.49)] be at least $\mathcal{O}(h)$, we must require that the coefficients in front of $\frac{\partial^2 u}{\partial t^2}$ and $\frac{\partial^2 u}{\partial x^2}$ both be of order $h$, independently of one another. On the other hand, from the first equation of (10.50) we always have $a^0 = 1/rh$, and
consequently, the coefficient in front of \( \frac{\partial^2 u}{\partial t^2} \) in the sum (10.52) is equal to \( rh/2 \). As such, the order of the residual with respect to the grid size \( h \) may never exceed one.

Thus, we have established that one cannot construct a consistent scheme \( \mathbf{L}_h \phi(h) = f(h) \) of type (10.47) that would approximate the Cauchy problem (10.8) with the accuracy better than \( \mathcal{O}(h) \). To achieve, for example the accuracy of \( \mathcal{O}(h^2) \), one would have to use a larger stencil, i.e., employ more grid nodes for building the difference operator \( \mathbf{L}_h \).

However, under the additional assumption of \( q(x,t) \equiv 0 \), there exists one and only one scheme \( \mathbf{L}_h \phi(h) = f(h) \) of type (10.47) that approximates problem (10.8) with second order of accuracy with respect to \( h \). Let us actually construct this scheme and in doing so also make sure that it is unique.\(^5\) For that purpose, we first notice that the identity

\[
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0
\]

also implies \( \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} \). Indeed,

\[
\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t} \left( \frac{\partial u}{\partial t} \right) = \frac{\partial}{\partial t} \left( \frac{\partial u}{\partial x} \right) = \frac{\partial}{\partial x} \left( \frac{\partial u}{\partial t} \right) = \frac{\partial}{\partial x} \left( \frac{\partial u}{\partial x} \right) = \frac{\partial^2 u}{\partial x^2}.
\]

Therefore, once the necessary conditions (10.50) for achieving first-order accuracy hold, formula (10.49) transforms into:

\[
\begin{align*}
\mathbf{L}_h[u]_h \bigg|_{(x_m,t_p)} &= a^0 rh \mathbf{d} [u]_h \bigg|_{(x_m,t_p)} \\
&+ \frac{1}{2} \left( a^0 r^2 + a_1 + a_{-1} \right) h^2 \frac{\partial^2 u(x_m,t_p)}{\partial x^2} + \mathcal{O}(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3).
\end{align*}
\]

To obtain a second-order scheme while taking into account that \( a^0 = 1/rh \), we need to require that \( h^2 (a^0 r^2 + a_1 + a_{-1}) = \mathcal{O}(h^2) \). As has been discussed, the most straightforward way to satisfy this requirement is to take:

\[
h^2 (a^0 r^2 + a_1 + a_{-1}) = 0.
\]

The system of four linear equations (10.50), (10.54) has a unique solution:

\[
a^0 = \frac{1}{rh}, \quad a_0 = -\frac{1}{rh} + \frac{r}{h}, \quad a_{-1} = \frac{1-r}{2h}, \quad a_1 = -\frac{1+r}{2h}.
\]

If we were to exercise the existing flexibility and retain the vanishing quantities of order \( \mathcal{O}(h^2) \) on the right-hand side of equations (10.50) and (10.54), then we would have only obtained small corrections to the coefficients (10.55).

As the third term on the right-hand side of (10.53) is \( \mathcal{O}(h^2) \), we conclude that the scheme \( \mathbf{L}_h \phi(h) = f(h) \) of (10.47), (10.55) is the only scheme built on the four-nodes stencil for Figure 10.3(left) that approximates problem (10.8) with \( q(x,t) = 0 \) with second-order accuracy. The corresponding finite-difference equation has the form:

\[
\frac{u^{p+1}}{\tau} - \frac{u^{p}}{\tau} - \frac{u^{p+1}_m - u^{p}_m}{2h} - \frac{r}{2h} (u^{p+1}_{m+1} - 2u^{p+1}_m + u^{p+1}_{m-1}) = 0.
\]

\(^5\) Uniqueness is to be understood in the same sense as on page 187.
In the literature, scheme (10.56) is known as the Lax-Wendroff scheme. In Section 10.2.4, it will be discussed further from a somewhat different perspective. In the meantime, we only notice that according to the Lax-Wendroff example, one does not necessarily have to consider the second-order differences in time, as in the leap-frog scheme (10.36), in order to achieve the overall second order of accuracy. On the other hand, it may still seem a little odd that the second order of accuracy can only be achieved for the homogeneous equation, i.e., when $\phi(x,t) = 0$. This issue is addressed in Exercise 8 of this section.

The methods that have been described can also be applied when constructing the schemes for a wider range of formulations than only constant coefficients and uniform grids. In addition, differential equations with variable coefficients, classes of nonlinear equations, and nonuniform grids can be analyzed. For example, to obtain a difference scheme for the Poisson equation $u_{xx} + u_{yy} = \phi(x,y)$ on a nonuniform Cartesian grid, one can use the following approximation formulae for the derivatives:

$$
\left. \frac{\partial^2 u}{\partial x^2} \right|_{(x_m,y_n)} \approx \frac{2}{\Delta y_m + \Delta y_{m-1}} \left[ \frac{u(x_{m+1},y_n) - u(x_m,y_n)}{\Delta y_m} - \frac{u(x_m,y_n) - u(x_{m-1},y_n)}{\Delta y_{m-1}} \right]
$$

$$
\left. \frac{\partial^2 u}{\partial y^2} \right|_{(x_m,y_n)} \approx \frac{2}{\Delta x_n + \Delta x_{n-1}} \left[ \frac{u(x_m,y_{n+1}) - u(x_m,y_n)}{\Delta x_n} - \frac{u(x_m,y_n) - u(x_m,y_{n-1})}{\Delta x_{n-1}} \right]
$$

Using the method of undetermined coefficients, one can actually make sure that these formulae are unique. More precisely, disregarding the non-essential lower order arbitrariness in the coefficients (see page 187), one can show that there exists one and only one set of coefficients $a_{-1}$, $a_0$, and $a_1$ that would guarantee first-order accuracy for the approximation of the second derivative:

$$
\left. \frac{\partial^2 u}{\partial x^2} \right|_{(x_m,y_n)} = a_{-1} u(x_{m-1},y_n) + a_0 u(x_m,y_n) + a_1 u(x_{m+1},y_n) + O(\max\{\Delta x_m, \Delta x_{m-1}\})
$$

One can also show that unless $\Delta x_m = \Delta x_{m-1}$, one cannot construct a similar three-node approximation of the second derivative that would have second order of accuracy with respect to $\max\{\Delta x_m, \Delta x_{m-1}\}$. To achieve that, one would generally need to incorporate more grid points into the stencil.

### 10.2.3 Other Methods

For a given differential problem, consider the class of finite-difference schemes built on some fixed stencil. Let us require that the desired scheme satisfy two types of constraints — primary and secondary. These constraints may, generally speaking, be incompatible. The objective is to search through all the schemes on the stencil that meet the primary constraint(s), and find the specific scheme (or schemes) among them that would minimize the degree of violation of the secondary constraint(s).

To illustrate the idea, consider the class of explicit schemes:

$$
u_{m+1}^p = \sum_{j=-n}^{j=n} b_j u_{m+j}^p\quad(10.57)$$
where \( j_{\text{left}} \geq 0 \) and \( j_{\text{right}} \geq 0 \) are two given fixed integers, and \( b, j = j_{\text{left}}, \ldots, j_{\text{right}} \), are the coefficients of the scheme. Many schemes that we have studied previously fall into the category (10.57). For example, under the assumption \( \phi(x,t) = 0 \) the scheme (10.47) transforms into (10.57) if we take \( j_{\text{left}} = j_{\text{right}} = 1 \):

\[
up{m+1} = b_{-1} u_{m-1} + b_0 u_m + b_1 u_{m+1},
\]

and set:

\[
b_{-1} = -a_{-1}, \quad b_0 = -a_0, \quad b_1 = -a_1.
\]

Next, with every scheme of type (10.57) we identify the point \( b = \{b_{-j_{\text{left}}}, b_{-j_{\text{left}}+1}, \ldots, b_{j_{\text{right}}} \} \) in the \( J \)-dimensional linear (vector) space, where \( J = j_{\text{left}} + j_{\text{right}} + 1 \). Denote by \( M_{\text{pr}} \) the subset of those schemes (10.57), i.e., points in the \( J \)-dimensional space, that satisfy the given primary constraint(s). Often, the primary constraints are related to the approximation properties of the scheme. For instance, we may require that the scheme have a prescribed order of accuracy on smooth solutions of the corresponding differential problem, e.g., the Cauchy problem:

\[
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad u(x,0) = \psi(x).
\]

Of course, the set \( M_{\text{pr}} \) may appear empty. For example, we have seen in Section 10.2.2 that no scheme built on the three-node upwind stencil from Figure 10.2(left) may have accuracy better than \( O(h) \). Thus, if we were to require second-order accuracy on this stencil, then the set \( M_{\text{pr}} \) would have contained no points. Alternatively, the set \( M_{\text{pr}} \) may consist of one single point that we denote \( b_0 \), which implies uniqueness of the scheme that satisfies the given primary constraint. For example, the Lax-Wendroff scheme (10.56) is the only second-order scheme on the four-node stencil from Figure 10.3(left). Finally, the set \( M_{\text{pr}} \) may contain many points. For example, there are multiple first-order schemes on the aforementioned four-node stencil from Figure 10.3(left).

Clearly, the first two scenarios provide no further room for “maneuvering.” In either case: \( M_{\text{pr}} = \emptyset \) or \( M_{\text{pr}} = b_0 \), regardless of the secondary constraint that we may impose, no improvement can be made unless it is satisfied ahead of time. However, the third scenario opens the possibility of quantifying the degree of violation for the secondary constraint, and subsequently minimizing the resulting quantity.

Consider the so-called plane wave solutions \( u = e^{i(\omega t + kx)} \) of the differential equation \( \partial_t u - \partial_x u = 0 \); here \( \omega \) is the frequency and \( k \) is the wavenumber. Direct substitution yields: \( i\omega u - iku = 0 \), and we therefore arrive at the following simple relation between the two parameters: \( \omega = k \), known as the dispersion relation. When the frequency is directly proportional to the wavenumber, the waves are said to propagate with no dispersion. Let us now substitute the same solution into the homogeneous finite-difference equation that corresponds to the first-order upwind scheme (10.10):

\[
e^{i(\omega \tau + k\Delta x)} \frac{e^{i\omega \tau} - 1}{\tau} = e^{i(\omega \tau + k\Delta x)} \frac{e^{ik\Delta x} - 1}{\Delta x}.
\]
Taking into account that $\tau = rh$, we obtain the following dispersion relation for the scheme:

$$\frac{1}{r}(e^{i\omega rh} - 1) = e^{ikh} - 1. \quad (10.58)$$

The implicit relation between $\omega$ and $k$ given by formula (10.58) is obviously not as simple as $\omega = k$, unless $r = 1$. If $r < 1$, we see that only when $kh \ll 1$ and $\omega \tau \ll 1$ we may claim that $\omega \approx k$, and even this equality is approximate [it is obtained with the help of the Taylor formula for the exponential functions in (10.58)]. The condition $kh \ll 1$ means that the wavelength $\lambda = 2\pi/k$ must be much greater than the grid size $h$. Otherwise, the waves traveling on the grid become dispersive.

On the other hand, in many applications related to the propagation of waves (e.g., acoustic or electromagnetic) it is highly desirable to minimize their dispersion due to the properties of the scheme. Therefore, we may formulate the following problem: Among the schemes (10.57) that have a prescribed order of accuracy on a fixed stencil (primary constraint that defines the set $M_{pr}$), find the one that would have the best dispersion relation. The latter notion can be quantified. To do so, we first notice that in the general case (10.57), the dispersion relation similar to (10.58) reads:

$$e^{i\omega rh} = \sum_{j=-j_{\text{left}}}^{j_{\text{right}}} b_j e^{ihj}. \quad (10.59)$$

Then, we should require that the solution curve $\omega = \omega(k)$ of equation (10.59) be maximally close to the nondispersive straight line $\omega = k$. Moreover, the condition of the minimal deviation between the two curves must be enforced on the largest possible subinterval $0 < k \leq k_0$ of the overall range of wavenumbers $0 < k \leq 2\pi/h$ that corresponds to the range of wavelengths $h \leq \lambda < \infty$.

In other words, the secondary constraint in this case can be introduced as a requirement that the dispersion relation of the scheme coincide with the ideal nondispersive law $\omega = k$. It can, however, be shown that among the schemes (10.57) only the simplest first-order upwind scheme (10.10) on the three-node stencil from Figure 10.2(left) with $r = 1$ may have the ideal dispersion relation. As such, the set $M_{sec}$ of those points in the $J$-dimensional vector space of schemes that satisfy the secondary constraint will only contain one point. Consequently, for the schemes with accuracy higher than $O(h)$, the primary and secondary constraints will indeed be incompatible, i.e., $M_{pr} \cap M_{sec} = \emptyset$. Then, finding the scheme with the best dispersion relation on $M_{pr}$ may require solving a fairly non-trivial optimization problem. Some particular formulations of this type have been studied in [TW93].

In general, if we equip the $J$-dimensional vector space of schemes with some norm $\| \cdot \|$, then we can formulate the problem of finding $b \in M_{pr}$ that would minimize the distance between the two sets $M_{pr}$ and $M_{sec}$, where $\text{dist}(\mathbf{b}, M_{sec}) \triangleq \inf_{x \in M_{sec}} \|\mathbf{b} - x\|$. This is, in fact, a rigorous formulation of the previously outlined problem of finding the scheme that satisfies the primary constraint(s) and minimally violates the secondary constraint(s). Depending on the particular situation, the problem of finding $\mathbf{b}$ that minimizes $\text{dist}(\mathbf{b}, M_{sec})$ can be solved using various numerical optimization methods.
techniques, such as linear programming, quadratic optimization, etc. This method of constructing the schemes has been first introduced by Kholodov. A detailed account of the corresponding developments, along with the analysis of numerous applications, can be found in the monograph [MK88].

10.2.4 Predictor-Corrector Schemes

When constructing difference schemes for time-dependent partial differential equations, one can exploit the same key idea that provides the foundation of Runge-Kutta schemes for ordinary differential equations. This is the idea of introducing the intermediate stages of computation, or equivalently, of employing the predictor-corrector strategy, see Sections 9.2.6 and 9.4 of Chapter 9. This strategy allows one to increase the order of accuracy that one would have obtained if only the original scheme were to be used by itself, i.e., with no intermediate stages. Besides, in the case of quasi-linear differential equations this strategy facilitates design of the so-called conservative finite-difference schemes that will be discussed in Chapter 11.

Let us recall the idea of the predictor-corrector approach using one of the simplest Runge-Kutta schemes as an example; this scheme will be applied to solving the Cauchy problem for a first-order ordinary differential equation:

\[
\frac{dy}{dt} = G(t,y), \quad y(0) = \psi, \quad 0 \leq t \leq T. \quad (10.60)
\]

If the value of the solution \(y_p\) at the moment of time \(t_p = p\tau\) is already computed, then in order to compute \(y_{p+1}\) we first find the auxiliary quantity \(\tilde{y}_{p+1/2}\) using the standard forward Euler scheme in the capacity of a predictor:

\[
\frac{\tilde{y}_{p+1/2} - y_p}{\tau/2} = G(t_p, y_p). \quad (10.61)
\]

Subsequently, we apply the corrector scheme to compute \(y_{p+1}\):

\[
\frac{y_{p+1} - y_p}{\tau} = G(t_{p+1/2}, \tilde{y}_{p+1/2}). \quad (10.62)
\]

The auxiliary quantity \(\tilde{y}_{p+1/2}\) obtained by scheme (10.61) with first-order accuracy helps us approximately evaluate the slope of the integral curve at the midpoint of the interval \([t_p, t_{p+1}]\) and thus obtain \(y_{p+1}\) by formula (10.62) with accuracy higher than that of the Euler scheme (10.61).

We have already mentioned in Section 9.4.2 of Chapter 9 that all these considerations will remain valid if \(y_p, \tilde{y}_{p+1/2}, \) and \(y_{p+1}\) were to be interpreted as finite-dimensional vectors and \(G\), accordingly, was to be thought of as a vector function. However, one can go even further and consider \(y_p, \tilde{y}_{p+1/2}, \) and \(y_{p+1}\) as elements of some functional space, and \(G\) as an operator acting in this space.
For instance, the Cauchy problem:

\[
\frac{\partial u}{\partial t} + A \frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T, \quad u(x,0) = \psi(x), \quad -\infty < x < \infty.
\]  

(10.63)

where \(A = \text{const}\), can be interpreted as a problem of type (10.60) if we set \(y(t) = u(x,t)\) so that for any given \(t\) the quantity \(y\) appears to be a function of the argument \(x\), and the operation \(G\) stands for the differential operator \(-A \frac{\partial}{\partial x}\). Let us now construct an example of a predictor-corrector scheme for problem (10.63).

Example

Assume that the grid function \(u_p^m = \{u_p^m\}, m = 0, \pm 1, \pm 2, \ldots\), has already been computed for some value of \(p\). Introduce the auxiliary grid function \(\tilde{u}_{m+1/2}^p\), \(m = 0, \pm 1, \pm 2, \ldots\), defined for the semi-integer time level \(t_{p+1/2} = (p + 1/2)\tau\) and at the semi-integer spatial points \(x_{m+1/2} = (m + 1/2)h\) (cell midpoints). This auxiliary grid function is to be computed using the following first-order scheme:

\[
\frac{\tilde{u}_{m+1/2}^p - (u_{m+1}^p + u_m^p)/2}{\tau/2} + A \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, \pm 1, \pm 2, \ldots. \tag{10.64}
\]

known as the Lax-Friedrichs scheme. Then, we employ the leap-frog scheme (10.36) in the capacity of a corrector to obtain \(u_{p+1}^m:\)

\[
\frac{u_{m+1}^p - u_m^p}{\tau} + A \frac{\tilde{u}_{m+1/2}^p - u_{m+1/2}^p}{h} = 0, \quad m = 0, \pm 1, \pm 2, \ldots. \tag{10.65}
\]

Eliminating the intermediate quantities \(\tilde{u}_{m+1/2}^p\) from equations (10.64) and (10.65) we arrive at the following scheme:

\[
\frac{u_{m+1}^p - u_m^p}{\tau} + A \frac{u_{m+1}^p - u_{m-1}^p - 2u_m^p + u_{m-1}^p}{2h} = 0, \quad u_0^p = \psi(mh), \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1. \tag{10.66}
\]

This is the Lax-Wendroff scheme that we have previously obtained in Section 10.2.2 for the case \(A = -1\), see formula (10.56). Scheme (10.66) has second-order accuracy with respect to \(h\), provided that \(\tau = rh, \ r = \text{const}\). Consequently, its “parent” predictor-corrector scheme (10.64)-(10.65) also has the second order of accuracy.

Another notable example is the predictor-corrector scheme of MacCormack, see formula (11.21) of Chapter 11; it is popular for fluid flow computations. For the case of a linear homogeneous equation, it reduces to the Lax-Wendroff scheme (10.66).

Exercises

1. Prove that the leap-frog scheme (10.36) has second-order accuracy on the solution \(u = u(x,t)\) of the Cauchy problem (10.8).
2. Prove uniqueness of the Lax-Wendroff scheme (10.56) in the sense outlined on page 187, as the only second-order scheme in its class that approximates problem (10.8) with \( \psi(x,t) = 0 \).

3. For the Cauchy problem in two space dimensions:

\[
\frac{\partial u}{\partial t} - \left( \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} \right) = \psi(x,y,t), \quad -\infty < x, y < \infty, \quad 0 < t \leq T, \]

\[ u(x,y,0) = \psi(x,y), \quad -\infty < x, y < \infty, \]

use a uniform Cartesian grid: \( x_m = mh, y_n = nh, t_p = p\tau \), and construct a consistent finite-difference scheme that would approximate this problem on its smooth solutions.

4. Consider a Cauchy problem for the heat equation with variable coefficients:

\[
\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \psi(x,t), \quad -\infty < x < \infty, \quad 0 < t \leq T, \]

\[ u(x,0) = \psi(x), \quad -\infty < x < \infty, \]

and approximate it on a uniform rectangular grid with the help of the scheme:

\[
\frac{u_{m+1}^{p+1} - u_{m}^{p+1}}{\tau} = \alpha \left( \frac{u_{m+1}^{p+1} - 2u_{m}^{p+1} + u_{m-1}^{p+1}}{h^2} + (1 - \alpha) \frac{u_{m+1}^{p} - 2u_{m}^{p} + u_{m-1}^{p}}{h^2} \right),
\]

\[ a_m^0 = \psi(mh), \]

where \( \alpha \) is a real parameter, \( 0 \leq \alpha \leq 1 \).

a) Show that for any \( \alpha \) the scheme is consistent and has the order of accuracy \( O(\tau + h^2) \) on a smooth solution \( u = u(x,t) \).

b) Find the value of \( \alpha \), for which the order of accuracy would become \( O(\tau^2 + h^2) \).

c) Assuming that \( \tau/h^2 = r = \text{const} \), find the value of \( \alpha \) that yields the approximation of accuracy \( O(h^4) \).

d) For \( \alpha = 0 \), find the value of \( r = \tau/h^2 \) that enables fourth-order accuracy: \( O(h^4) \).

Hint. In items c) and d) use: \( u_m = (u_{xx})_x = (u_x)_{xx} = u_{xxxx} \).

e) For a fixed \( r = \tau/h^2 \), can one choose the value of \( \alpha \) that would yield higher than fourth order of accuracy on any smooth solution of problem (10.67)?

5. Consider a Cauchy problem for the heat equation with variable coefficients:

\[
\frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left( a(x,t) \frac{\partial u}{\partial x} \right) = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T, \]

\[ u(x,0) = \psi(x), \quad -\infty < x < \infty. \]

Use a uniform rectangular grid: \( x_m = mh, t_p = p\tau \), and construct a consistent finite-difference scheme that would approximate this problem on its smooth solutions.

6. Consider a Cauchy problem for the nonlinear heat equation:

\[
\frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left( a(u) \frac{\partial u}{\partial x} \right) = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T, \]

\[ u(x,0) = \psi(x), \quad -\infty < x < \infty. \]

Use a uniform rectangular grid: \( x_m = mh, t_p = p\tau \), and construct a consistent finite-difference scheme that would approximate this problem on its smooth solutions.
7. Consider the Cauchy problem (10.67) for the heat equation, and approximate it with the predictor-corrector scheme designed as follows. The auxiliary grid function \( \tilde{u}_{m+1/2} \) is to be computed with accuracy \( O(\tau + h^2) \) by the implicit method:

\[
\frac{\tilde{u}_{m+1/2}^p - u_m^p}{\tau/2} - \frac{\tilde{u}_{m+1/2}^r - \tilde{u}_{m-1/2}^r}{h^2} = 0, \quad m = 0, \pm 1, \pm 2, \ldots, \]

and the actual solution \( u_m^{p+1} \) at \( t = t_{p+1} \) is to be computed by the scheme:

\[
\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{u_{m+1/2}^p - u_{m-1/2}^p}{h} = 0, \quad u_0^0 = \psi(mh).
\]

Prove that the overall predictor-corrector scheme has the accuracy \( O(\tau^2 + h^2) \) on the smooth solution \( u = u(x, t) \).

8. Consider a modified scheme \( \mathbf{L}_{\delta} u^{(h)} = f^{(h)} \) of type (10.47) (modification of the right-hand side only):

\[
\mathbf{L}_{\delta} u^{(h)} = \begin{cases} 
\partial_0 u_m^{p+1} + a_{-1} u_{m-1}^p + a_0 u_m^p + a_1 u_{m+1}^p, \\
u_m^p 
\end{cases}

f^{(h)} = \begin{cases} 
\frac{\psi(x_m, t_p)}{2}(\phi_0 + \phi_1)(x_m, t_p), \\
\psi(x_m),
\end{cases}
\]

and define the coefficients of the operator \( \mathbf{L}_\delta \) according to formula (10.55) that corresponds to the Lax-Wendroff method. Show that the resulting scheme approximates problem (10.8) with second-order accuracy for an arbitrary sufficiently smooth right-hand side \( \psi(x, t) \) (not necessarily zero).

9. Represent the scheme (10.47) for \( \psi(x, t) = 0 \) in the form:

\[
u_m^{p+1} = b_{-1} u_{m-1}^p + b_0 u_m^p + b_1 u_{m+1}^p, \quad \text{ (10.68)}
\]

where \( b_{-1} = -a_{-1}/a_0, \ b_0 = -a_0/a_0, \) and \( b_1 = -a_1/a_0, \) see Section 10.2.3. The scheme (10.68) is said to be monotone if \( b_j \geq 0, \ j = -1, 0, 1. \) Adopting the terminology of Section 10.2.3, let the primary constraint be the order of accuracy of at least \( \partial_0(h) \), and the secondary constraint be the monotonicity of the scheme. In the three-dimensional space of vectors \( \{b_{-1}, b_0, b_1\} \), describe the set \( M_0 = M_{pr} \cap M_{sec} \) (in this case, the distance between the sets \( M_{pr} \) and \( M_{sec} \) is zero).

**Answer.** If \( \tau = r/h > 1 \), then \( M_0 = 0 \). If \( r = 1 \), then \( M_0 = \{(0, 0, 1)\} \). If \( 0 < r < 1 \), then \( M_0 \) is the interval with the endpoints: \( \left( \frac{1}{1-r}, 0, 0 \right) \) and \( (0, 1-r, 0) \).

10. Prove that the monotone schemes introduced in the sense of Exercise 9 for \( \psi = 0 \) and \( j = j_{\text{left}}, \ldots, j_{\text{right}} \) satisfy the maximum principle from page 167, i.e., that the maximum of the corresponding difference solution will not increase as the time elapses.

11.* (Godunov theorem) Prove that no monotone scheme has order of accuracy higher than \( \partial_0(h) \).

**Hint.** In the J-dimensional space of schemes, show that if the accuracy is higher than \( \partial_0(h) \), then the coefficients \( b_j, \ j = j_{\text{left}}, \ldots, j_{\text{right}} \), may not all be non-negative. A full proof can be found, e.g., in [Str04, pages 71–72].
12. Consider a Cauchy problem for the one-dimensional wave (d’Alembert) equation:

\[
\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = \psi(x,t), \quad -\infty < x < \infty, \quad 0 < t \leq T,
\]
\[
u(x,0) = \psi_0(x), \quad \frac{\partial u(x,0)}{\partial t} = \psi_1(x), \quad -\infty < x < \infty.
\]

Analyze the approximation properties of the scheme \(L_hu^{(h)} = f^{(h)}\) on the smooth solution \(u = u(x,t)\), where:

\[
L_hu^{(h)} = \begin{cases} 
\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{\tau^2} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2}, \\
u_m^0, \\
u_m^1
\end{cases}
\]

\[
f^{(h)} = \begin{cases} 
\psi(x_m, t_p), \\
\psi_0(x_m), \\
\psi_1(x_m)
\end{cases}
\]

Define the norm of \(f^{(h)}\) as \(\|f^{(h)}\|_{F_0} = \max_{p,m} \left\{ \max |\psi_m^p|, \sup_{m}|\psi_m^0|, \sup_{m}|\psi_m^1| \right\}\) and show that when \(\tau = rh, r = \text{const.}\), the accuracy of the scheme is \(O(h^2)\).

How can one specify the values of \(\psi_m^1\), making use of the given functions \(\psi(x,t)\), \(\psi_0(x)\), and \(\psi_1(x)\), so that to improve the order of accuracy and make it \(O(h^2)\)?

13. Consider the process of heat transfer on a finite interval, as opposed to the infinite line. It is governed by the heat equation subject to both initial and boundary conditions, which altogether yield the initial boundary value problem:

\[
\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \psi(x,t), \quad 0 < x < 1, \quad 0 < t \leq T,
\]
\[
u(x,0) = \psi_0(x), \quad 0 < x < 1,
\]
\[
u(0,t) = \psi_1(t), \quad 0 < t \leq T,
\]
\[
u(1,t) = \psi_2(t), \quad 0 < t \leq T.
\]

Use the grid: \(x_m = mh, t_p = pt\), and approximate this problem with the scheme:

\[
\frac{u_{m+1}^p - u_m^p}{\tau} - \frac{u_{m+1}^{p+1/2} - 2u_{m}^{p+1/2} + u_{m-1}^{p+1/2}}{h^2} = \psi(x_m, t_p),
\]
\[
m = 1, 2, \ldots, M - 1, \quad h = 1/M, \quad p = 0, 1, \ldots, [T/\tau] - 1,
\]
\[
u_m^0 = \psi_0(x_m), \quad m = 0, 1, \ldots, M,
\]
\[
u_0^p = \psi_1(t_p), \quad p = 0, 1, \ldots, [T/\tau],
\]
\[
u_M^p = \psi_2(t_p), \quad p = 0, 1, \ldots, [T/\tau].
\]

Define the norm in the space \(F_0\) as:

\[
\|f^{(h)}\|_{F_0} = \max_{p,m} \left\{ \max |\psi(x_m, t_p)|, \max |\psi_0(x_m)|, \max |\psi_1(t_p)|, \max |\psi_2(t_p)| \right\},
\]
and show that when \(\tau = rh^2, r = \text{const.}\), the accuracy of the scheme is \(O(h^2)\).
10.3 Spectral Stability Criterion for Finite-Difference
Cauchy Problems

Perhaps the most widely used approach to the analysis of stability for finite-
difference Cauchy problems has been proposed by von Neumann. In this section,
we will introduce and illustrate it using several examples of difference equations
with constant coefficients. The case of variable coefficients will be addressed in the
next Section 10.4 and the case of initial boundary value problems (as opposed to pure
initial value problems) will be explored in Section 10.5.

10.3.1 Stability with Respect to Initial Data

The simplest example of a finite-difference Cauchy problem is the first-order up-
wind scheme:

\[ L_h u(h) = f(h), \quad (10.69) \]

for which the operator \( L_h \) and the right-hand side \( f(h) \) are given by:

\[
\begin{align*}
L_h u(h) &= \begin{cases} 
\frac{u_{m+1}^p - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h}, & m = 0, \pm 1, \pm 2, \ldots, \\
u_0^0, & p = 0, 1, \ldots, [T/\tau] - 1,
\end{cases} \\
f(h) &= \begin{cases} 
q_m^p, & m = 0, \pm 1, \pm 2, \ldots, \\
\psi_m, & m = 0, \pm 1, \pm 2, \ldots,
\end{cases}
\end{align*}
\]  

(10.70)

We have already encountered this scheme on many occasions. Let us define the
norms \( \|u(h)\|_{U_h} \) and \( \|f(h)\|_{F_h} \) as follows:

\[
\|u(h)\|_{U_h} \overset{\text{def}}{=} \max_p \sup_m |u_m^p|, \quad \|f(h)\|_{F_h} \overset{\text{def}}{=} \max_p \sup_m |q_m^p| + \sup_m |\psi_m|.
\]  

(10.71)

Then for the scheme (10.69)–(10.70), the stability condition (10.17):

\[
\|u(h)\|_{U_h} \leq c \|f(h)\|_{F_h}
\]

given by Definition 10.2 transforms into:

\[
\sup_m |u_m^p| \leq c \left[ \max_p \sup_m |q_m^p| + \sup_m |\psi_m| \right], \quad p = 0, 1, \ldots, [T/\tau];
\]  

(10.72)

where the constant \( c \) is not supposed to depend on \( h \) (or on \( \tau = rh \), \( r = \text{const} \)).

Condition (10.72) must hold for any arbitrary \( \{\psi_m\} \) and \( \{q_m^p\} \). In particular, it
should obviously hold for an arbitrary \( \{\psi_m\} \) and \( \{q_m^p\} \equiv 0 \). In other words, for
stability it is necessary that solution $u^p_m$ to the problem:

\[
\frac{u^p_m - u^p_{m+1}}{\tau} - \frac{u^p_m - u^p_{m+1}}{h} = 0, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1,
\]
\[
\left. u^0_m \right|_{m = 0, \pm 1, \pm 2, \ldots} = \psi_m,
\]

satisfy the inequality:

\[
\sup_m |u^p_m| \leq c \sup_m |\psi_m|, \quad p = 0, 1, \ldots, [T/\tau],
\]

for any bounded grid function $\psi_m$.

Property (10.74), which is necessary for the finite-difference scheme (10.69)–(10.70) to be stable in the sense (10.72), is called stability with respect to perturbations of the initial data, or simply stability with respect to the initial data. It means that if a perturbation is introduced into the initial data $\psi_m$ of problem (10.73), then the corresponding perturbation of the solution will be no more than $c$ times greater in magnitude than the original perturbation of the data, where the constant $c$ does not depend on the grid size $h$.

10.3.2 A Necessary Spectral Condition for Stability

For the Cauchy problem (10.69)–(10.70) to be stable with respect to the initial data it is necessary, in particular, that inequality (10.74) hold for $\psi_m$ being equal to a harmonic function:

\[
\left. u^0_m \right|_{m = 0, \pm 1, \pm 2, \ldots} = \psi_m = e^{i\alpha m}, \quad m = 0, \pm 1, \pm 2, \ldots,
\]

where $\alpha$ is a real parameter. One can easily solve problem (10.73) with the initial condition (10.75); the solution $u^p_m$ can be found in the form:

\[
u^p_m = \lambda^p e^{i\alpha m}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau],
\]

where the quantity $\lambda = \lambda(\alpha)$ is determined by substitution of expression (10.76) into the homogeneous finite-difference equation of problem (10.73):

\[
\lambda(\alpha) = 1 - r + re^{i\alpha}, \quad r = \tau/h = \text{const}.
\]

Solution (10.76) satisfies the equality:

\[
\sup_m |u^p_m| = |\lambda(\alpha)|^p \sup_m |u^0_m| = |\lambda(\alpha)|^p \sup_m |\psi_m|.
\]

Therefore, for the stability condition (10.74) to be true it is necessary that the following inequality hold for all real $\alpha$:

\[
|\lambda(\alpha)|^p \leq c, \quad p = 0, 1, \ldots, [T/\tau].
\]
Equivalently, we can require that

$$|\lambda(\alpha)| \leq 1 + c_1 \tau,$$

(10.78)

where $c_1$ is a constant that does not depend either on $\alpha$ or on $\tau$. Inequality (10.78) represents the necessary spectral condition for stability due to von Neumann. It is called spectral because of the following reason.

Existence of the solution in the form (10.76) shows that the harmonic $e^{i\alpha m}$ is an eigenfunction of the transition operator from time level $t_p$ to time level $t_{p+1}$:

$$u_m^{p+1} = (1 - r)u_m^p + ru_{m+1}^p, \quad m = 0, \pm 1, \pm 2, \ldots$$

According to the finite-difference equation (10.73), this operator maps the grid function $\{u_m^p\}$, $m = 0, \pm 1, \pm 2, \ldots$, defined for $t = t_p$ onto the grid function $\{u_m^{p+1}\}$, $m = 0, \pm 1, \pm 2, \ldots$, defined for $t = t_{p+1}$. The quantity $\lambda(\alpha)$ given by formula (10.77) is therefore an eigenvalue of the transition operator that corresponds to the eigenfunction $\{e^{i\alpha m}\}$. In the literature, $\lambda(\alpha)$ is sometimes also referred to as the amplification factor. The set of all complex numbers $\lambda = \lambda(\alpha)$ obtained when the parameter $\alpha$ sweeps through the real axis forms a curve on the complex plane. This curve is called the spectrum of the transition operator.

Consequently, the necessary stability condition (10.78) can be re-formulated as follows: The spectrum of the transition operator that corresponds to the difference equation of problem (10.73) must belong to the disk of radius $1 + c_1 \tau$ centered at the origin on the complex plane. In our particular example, the spectrum (10.77) does not depend on $\tau$ at all. Therefore, condition (10.78) is equivalent to the requirement that the spectrum $\lambda = \lambda(\alpha)$ belong to the unit disk:

$$|\lambda(\alpha)| \leq 1.$$

(10.79)

Let us now use the criterion that we have formulated, and actually analyze stability of problem (10.69)–(10.70). The spectrum (10.77) forms a circle of radius $r$ centered at the point $(1 - r, 0)$ on the complex plane. When $r < 1$, this circle lies inside inside the unit disk, being tangent to the unit circle at the point $\lambda = 1$. When $r = 1$ the spectrum coincides with the unit circle. Lastly, when $r > 1$ the spectrum lies outside the unit disk, except one point $\lambda = 1$, see Figure 10.5. Accordingly, the necessary stability condition (10.79) is satisfied for $r \leq 1$ and is violated for $r > 1$. Let us now recall that in Section 10.1.3 we have studied the same difference problem and have proven that it is stable when $r \leq 1$ and is unstable when $r > 1$. Therefore, in this particular case the von Neumann necessary stability condition appears sufficiently sensitive to distinguish between the actual stability and instability.

In the case of general Cauchy problems for finite-difference equations and systems, we give the following

**DEFINITION 10.3** The spectrum of a finite-difference problem is given by the set of all those and only those $\lambda = \lambda(\alpha, h)$, for which the corresponding
homogeneous finite-difference equation or system has a solution of the form:

$$u_p^m = \lambda^p [u^0 e^{i \alpha m}], \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau],$$

(10.80)

where $u^0$ is either a fixed number in the case of one scalar difference equation (this number may be taken equal to one with no loss of generality), or a constant finite-dimensional vector in the case of a system of difference equations.

Then, the von Neumann necessary stability condition says that given an arbitrarily small number $\epsilon > 0$, for all sufficiently small grid sizes $h$ the spectrum $\lambda = \lambda(\alpha, h)$ of the difference problem has to lie inside the following disk on the complex plane:

$$|\lambda| \leq 1 + \epsilon.$$  

(10.81)

Note that if for a particular problem the spectrum appears to be independent of the grid size $h$ (and $\tau$), then condition (10.81) becomes equivalent to the requirement that the spectrum $\lambda(\alpha, h) = \lambda(\alpha)$ belong to the unit disk, see (10.79).

If the von Neumann necessary condition (10.81) is not satisfied, then one should not expect stability for any reasonable choice of norms. If, on the other hand, this condition is met, then one may hope that for a certain appropriate choice of norms the scheme will turn out stable.

### 10.3.3 Examples

We will exploit the von Neumann spectral condition to analyze stability of a number of interesting finite-difference problems. First, we will consider the schemes that approximate the Cauchy problem:

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \psi(x,t), \quad -\infty < x < \infty, \quad 0 < t \leq T,$$

$$u(x,0) = \psi(x), \quad -\infty < x < \infty.$$  

(10.82)
Example 1

Consider the first-order downwind scheme:
\[
\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_m^p - u_{m-1}^p}{h} = q_m^p, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,
\]
\[
u_m^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots.
\]

Substituting the solution of type (10.76) into the corresponding homogeneous finite-difference equation, we obtain:
\[
\lambda(\alpha) = 1 + r - re^{-i\alpha}.
\]

Therefore, the spectrum is a circle of radius \(r\) centered at the point \((1 + r, 0)\) on the complex plane, see Figure 10.6. This spectrum does not depend on \(h\). It is also clear that for no value of \(r\) does it belong to the unit circle. Consequently, the stability condition (10.79) may never be satisfied. This conclusion is expected, because the downwind scheme obviously violates the Courant, Friedrichs, and Lewy condition for any \(r = \tau/h\). (see Section 10.1.4).

Example 2

Next, consider the Lax-Wendroff scheme [cf. formula (10.66)]:
\[
\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_m^{p+1} - u_{m-1}^p}{2h} - \frac{\tau}{2h^2}(u_m^p - 2u_m^p + u_m^{p-1}) = q_m^p,
\]
\[
u_m^0 = \psi_m,
\]
that approximates problem (10.82) with second-order accuracy if \(q \equiv 0\), and with first-order accuracy otherwise. For this scheme, the spectrum \(\lambda = \lambda(\alpha, h)\) is determined from the equation:
\[
\frac{\lambda - 1}{\tau} - \frac{e^{i\alpha} - e^{-i\alpha}}{2h} - \frac{\tau}{2h^2}(e^{i\alpha} - 2 + e^{-i\alpha}) = 0.
\]

Let us denote \(r = \tau/h\) as before, and notice that
\[
e^{i\alpha} - e^{-i\alpha} = \sin \alpha,
\]
\[
e^{i\alpha} - 2 + e^{-i\alpha} = \left(\frac{e^{i\alpha/2} - e^{-i\alpha/2}}{2i}\right)^2 = -\sin^2 \frac{\alpha}{2}.
\]
Then,
\[ \lambda(\alpha) = 1 + ir\sin\alpha - 2r^2\sin^2\frac{\alpha}{2}, \]
\[ |\lambda(\alpha)|^2 = \left(1 - 2r^2\sin^2\frac{\alpha}{2}\right)^2 + r^2\sin^2\alpha. \]

The spectrum does not depend on \( h \), and from the previous equality we easily find that
\[ 1 - |\lambda|^2 = 4r^2(1 - r^2)\sin^4\frac{\alpha}{2}. \]

The von Neumann condition is satisfied when the right-hand side of this equality is non-negative, which means \( r \leq 1 \); it is violated when \( r > 1 \).

**Example 3**

Consider the explicit central-difference scheme:
\[
\frac{u_{m+1}^p - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{2h} = q_m^p, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,
\]
\[
u_m^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad (10.84)
\]

for the same Cauchy problem (10.82).

Substituting expression (10.76) into the homogeneous counterpart of the difference equation from (10.84), we have:
\[
\frac{\lambda - 1}{\tau} - \frac{e^{i\alpha} - e^{-i\alpha}}{2h} = 0,
\]
which yields:
\[
\lambda(\alpha) = 1 + i\left(\frac{\tau}{h}\sin\alpha\right).
\]

The spectrum \( \lambda = \lambda(\alpha) \) fills the vertical interval of length \( 2\tau/h \) that crosses through the point \((1,0)\) on the complex plane, see Figure 10.7.

If \( \tau/h = r = \text{const} \), then the spectrum can be said to be independent of \( h \) (and of \( \tau \)). This spectrum lies outside of the unit disk, the von Neumann condition (10.79) is not met, and the scheme is unstable. However, if we require that the temporal size \( \tau \) be proportional to \( h^2 \) as \( h \to 0 \), so that \( r = \tau/h^2 = \text{const} \), then the point of the spectrum, which is most distant from the origin, will be \( \lambda(\pi/2) \) [and also \( \lambda(-\pi/2) \)]. For this point we have:
\[
|\lambda(\pi/2)| = \sqrt{1 + \left(\frac{\tau}{h}\right)^2} = \sqrt{1 + r\tau} < 1 + \frac{r}{2}\tau.
\]
Then, the von Neumann condition (10.81) in the form 
\[ \lambda(\alpha) \leq 1 + c_1 \tau \] is satisfied for \( c_1 = r/2 \). It is clear that the requirement \( \tau = rh^2 \) puts a considerably stricter constraint on the rate of decay of the temporal grid size \( \tau \) as \( h \to 0 \) than the previous requirement \( \tau = rh \) does. Still, that previous requirement was sufficient for the von Neumann condition to hold for the difference schemes (10.69)–(10.70) and (10.83) that approximate the same Cauchy problem (10.82).

We also note that the Courant, Friedrichs, and Lewy condition of Section 10.1.4 allows us to claim that the scheme (10.84) is unstable only for \( \tau/h = r > 1 \), but does not allow us to judge the stability for \( \tau/h = r \leq 1 \). As such, it appears weaker than the von Neumann condition.

**Example 4**

The instability of scheme (10.84) for \( \tau/h = r = \text{const} \) can be fixed by changing the way the time derivative is approximated. Instead of (10.84), consider the scheme:

\[
\frac{u_{m}^{p+1} - (u_{m-1}^{p} + u_{m+1}^{p})/2}{\tau} - \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} = q_{m}^{p},
\]

(10.85)

obtained by replacing \( u_{m}^{p} \) with \( (u_{m-1}^{p} + u_{m+1}^{p})/2 \). This is, in fact, a general approach attributed to Friedrichs, and the scheme (10.85) is known as the Lax-Friedrichs scheme; we have first introduced it in Section 10.2.4. The equation to determine the spectrum for the scheme (10.85) reads:

\[
\frac{\lambda - (e^{i\alpha} + e^{-i\alpha})/2}{\tau} - \frac{e^{i\alpha} - e^{-i\alpha}}{2h} = 0,
\]

which yields:

\[
\frac{\lambda}{\tau} - \frac{i\sin\alpha}{h} = 0
\]

and

\[
\lambda(\alpha) = \cos\alpha + ir\sin\alpha,
\]

where \( r = \tau/h = \text{const} \). Consequently,

\[
|\lambda(\alpha)|^2 = \cos^2\alpha + r^2\sin^2\alpha.
\]

Clearly, the von Neumann condition (10.79) is satisfied for \( r \leq 1 \), because then \( |\lambda|^2 \leq \cos^2\alpha + \sin^2\alpha = 1 \). For \( r > 1 \), the von Neumann condition is violated.

**Example 5**

Finally, consider the leap-frog scheme (10.36) for problem (10.82). The corresponding homogeneous finite-difference equation is written as:

\[
\frac{u_{m}^{p+1} - u_{m}^{p-1}}{2\tau} - \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} = 0,
\]

(10.86)
and for the spectrum we obtain:

\[ \frac{\lambda - \lambda^{-1}}{2 \tau} - \frac{e^{i \alpha} - e^{-i \alpha}}{2 h} = 0. \]

This is a quadratic equation with respect to \( \lambda \):

\[ \lambda^2 - i 2 r \sin \alpha - 1 = 0, \]

where \( r = \tau/h = \text{const.} \) The roots of this equation are given by:

\[ \lambda_{1,2} = ir \sin \alpha \pm \sqrt{1 - r^2 \sin^2 \alpha}. \]

We notice that when \( r \leq 1 \), then \( |\lambda_{1,2}| = r^2 \sin^2 \alpha + (1 - r^2 \sin^2 \alpha) = 1 \), so that the entire spectrum lies precisely on the unit circle and the von Neumann condition (10.79) is satisfied. Otherwise, when \( r > 1 \), we again take \( \alpha = \pi/2 \) and obtain: \( |\lambda_{1,2}| = \sqrt{|i r + i \sqrt{r^2 - 1}|} = r + \sqrt{r^2 - 1} > 1 \), which means that the von Neumann condition is not met. This example illustrates how the von Neumann criterion can be applied to a finite-difference equation, such as equation (10.86), that connects the values of the solution on three, rather than two, consecutive time levels of the grid.

Next, we will consider two schemes that approximate the following Cauchy problem for the heat equation [cf. problem (10.67)]:

\[ \frac{\partial u}{\partial t} - a^2 \frac{\partial^2 u}{\partial x^2} = \phi(x, t), \quad -\infty < x < \infty, \quad 0 < t \leq T, \]

\[ u(x, 0) = \psi(x), \quad -\infty < x < \infty. \]  

(10.87)

**Example 6**

The first scheme is explicit:

\[ \frac{u_m^{p+1} - u_m^p}{\tau} - a^2 \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = q_m^p, \]

\[ u_m^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lceil T/\tau \rceil - 1. \]

It allows us to compute \( \{u_m^{p+1}\} \) in the closed form via \( \{u_m^p\} \):

\[ u_m^{p+1} = (1 - 2ra^2)u_m^p + ra^2(u_{m+1}^p + u_{m-1}^p) + \tau q_m^p, \quad p = 0, 1, \ldots, \lceil T/\tau \rceil - 1, \]

where \( r = \tau/h^2 = \text{const.} \). Substitution of \( u_m^p = \lambda^p e^{i \alpha m} \) into the corresponding homogeneous difference equation yields:

\[ \frac{\lambda - 1}{\tau} - a^2 \frac{e^{-i \alpha} - 2 + e^{i \alpha}}{h^2} = 0. \]

By noticing that

\[ \frac{e^{-i \alpha} - 2 + e^{i \alpha}}{4} = -\sin^2 \frac{\alpha}{2}, \]
we obtain:
\[ \lambda(\alpha) = 1 - 4ra^2 \sin^2 \frac{\alpha}{2}, \quad r = \frac{\tau}{h^2}. \]

When \( \alpha \) varies between 0 and \( 2\pi \), the point \( \lambda(\alpha) \) sweeps the interval \( 1 - 4ra^2 \leq \lambda \leq 1 \) of the real axis, see Figure 10.8.

For stability, it is necessary that the left endpoint of this interval still be inside the unit circle (Figure 10.8), i.e., that \( 1 - 4ra^2 \geq -1 \). This requirement translates into:
\[
r \leq \frac{1}{2a^2}. \tag{10.88}
\]

Inequality (10.88) guarantees that the von Neumann stability condition will hold. Conversely, if we have \( r > 1/(2a^2) \), then the point \( \lambda(\alpha) = 1 - 4ra^2 \sin^2(\alpha/2) \) that corresponds to \( \alpha = \pi \) will be located to the left of the point \(-1\), i.e., outside the unit circle. In this case, the harmonic \( e^{i\pi m} = (-1)^m \) generates the solution:
\[
    u^p_m = (1 - 4ra^2)^p(-1)^m
\]

that does not satisfy inequality (10.74) for any value of \( c \).

**Example 7**

The second scheme is implicit:
\[
    \frac{u_{m+1}^p - u_m^p}{\tau} - a^2 \frac{u_{m+1}^{p+1} - 2u_{m+1}^p + u_{m-1}^{p+1}}{h^2} = q_m^{p+1},
\]
\[
    u_0^p = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1.
\]

For this scheme, \( \{u_{m+1}^{p+1}\} \) cannot be obtained from \( \{u_m^p\} \) by an explicit formula because for a given \( m \) the difference equation contains three, rather than one, unknowns: \( u_{m+1}^{p+1}, u_{m+1}^p, \) and \( u_{m-1}^{p+1} \). As in the previous case, we substitute \( u_m^p = \lambda^p e^{i\pi m} \) into the homogeneous equation and obtain:
\[
    \lambda(\alpha) = \frac{1}{1 + 4ra^2 \sin^2(\alpha/2)}, \quad r = \frac{\tau}{h^2}.
\]

Consequently, the spectrum of the scheme fills the interval:
\[
    (1 + 4ra^2)^{-1} \leq \lambda \leq 1
\]
of the real axis and the von Neumann condition \( |\lambda| \leq 1 \) is met for any \( r \).
Example 8

The von Neumann analysis also applies when studying stability of a scheme in the case of more than one spatial variable. Consider, for instance, a Cauchy problem for the heat equation on the \((x,y)\) plane:

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad -\infty < x, y < \infty, \quad 0 < t \leq T,
\]

\[u(x,y,0) = \psi(x,y), \quad -\infty < x, y < \infty.\]

We approximate this problem on the uniform Cartesian grid:

\[
(x_m, y_n, t_p) = (mh, nh, p\tau).\]

Replacing the derivatives with difference quotients we obtain:

\[
\frac{u_{m,n}^{p+1} - u_{m,n}^p}{\tau} = \frac{u_{m+1,n}^p - 2u_{m,n}^p + u_{m-1,n}^p}{h^2} + \frac{u_{m,n+1}^p - 2u_{m,n}^p + u_{m,n-1}^p}{h^2} - u_{m,n}^0 = \psi_{m,n}, \quad m, n = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1.
\]

The resulting scheme is explicit. To analyze its stability, we specify \(u_{m,n}^0\) in the form of a two-dimensional harmonic \(e^{i(\alpha m + \beta n)}\) determined by two real parameters \(\alpha\) and \(\beta\), and generate a solution in the form:

\[u_{m,n}^p = \lambda^p(\alpha, \beta)e^{i(\alpha m + \beta n)}.\]

Substituting this expression into the homogeneous difference equation, we find after some easy equivalence transformations:

\[\lambda(\alpha, \beta) = 1 - 4r \left( \sin^2 \frac{\alpha}{2} + \sin^2 \frac{\beta}{2} \right).\]

When the real quantities \(\alpha\) and \(\beta\) vary between 0 and \(2\pi\), the point \(\lambda = \lambda(\alpha, \beta)\) sweeps the interval \(1 - 8r \leq \lambda \leq 1\) of the real axis. The von Neumann stability condition is satisfied if \(1 - 8r \geq -1\), i.e., when \(r \leq 1/4\).

Example 9

In addition to the previous Example 5, let us now consider another example of the scheme that connects the values of the difference solution on the three, rather than two, consecutive time levels of the grid.

A Cauchy problem for the one-dimensional homogeneous d’Alembert (wave) equation:

\[
\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T,
\]

\[u(x,0) = \psi^{(0)}(x), \quad \frac{\partial u(x,0)}{\partial t} = \psi^{(1)}(x), \quad -\infty < x < \infty.\]
can be approximated on a uniform Cartesian grid by the following scheme:

\[
\begin{align*}
\frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{\tau^2} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} &= 0, \\
u_0^p &= \psi_0^0, \\
u_1^p &= \psi_1^1, \\
m &= 0, \pm 1, \pm 2, \ldots, \\
p &= 1, 2, \ldots, \lfloor T/\tau \rfloor - 1.
\end{align*}
\]

Substituting a solution of type (10.76) into the foregoing finite-difference equation, we obtain the following quadratic equation for determining \( \lambda = \lambda(\alpha) \):

\[
\lambda^2 - 2 \left( 1 - 2r^2 \sin^2 \frac{\alpha}{2} \right) \lambda + 1 = 0, \quad r = \frac{\tau}{h}.
\]

The product of the two roots of this equation is equal to one. If its discriminant:

\[
D(\alpha) = 4r^2 \sin^2 \frac{\alpha}{2} \left( r^2 \sin^2 \frac{\alpha}{2} - 1 \right)
\]

is negative, then the roots \( \lambda_1(\alpha) \) and \( \lambda_2(\alpha) \) are complex conjugate and both have a unit modulus. If \( r < 1 \), the discriminant \( D(\alpha) \) remains negative for all \( \alpha \in [0, 2\pi) \).

The spectrum for this case is shown in Figure 10.9(a); it fills an arc of the unit circle. If \( r = 1 \), the spectrum fills exactly the entire unit circle. When \( r > 1 \), the discriminant \( D(\alpha) \) may be either negative or positive depending on the value of \( \alpha \). In this case, once the argument \( \alpha \) increases from 0 to \( \pi \) the roots \( \lambda_1(\alpha) \) and \( \lambda_2(\alpha) \) depart from the point \( \lambda = 1 \) and move along the unit circle: One root moves clockwise and the other root counterclockwise, and then they merge at \( \lambda = -1 \). After that one root moves away from this point along the real axis to the right, and the other one to the left, because they are both real and their product \( \lambda_1 \lambda_2 = 1 \), see Figure 10.9(b). The von Neumann stability condition is met for \( r \leq 1 \).
Consider a Cauchy problem for the following first-order hyperbolic system of equations that describes the propagation of acoustic waves:

\[
\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x}, \quad \frac{\partial w}{\partial t} = \frac{\partial v}{\partial x},
\]

\[-\infty < x < \infty, \quad 0 < t \leq T, \quad v(x,0) = \psi^{(1)}(x), \quad w(x,0) = \psi^{(2)}(x).
\] (10.89a)

Let us set:

\[u(x,t) = \begin{bmatrix} v(x,t) \\ w(x,t) \end{bmatrix}, \quad \psi(x) = \begin{bmatrix} \psi^{(1)}(x) \\ \psi^{(2)}(x) \end{bmatrix}.\]

Then, problem (10.89a) can be recast in the matrix form:

\[
\frac{\partial u}{\partial t} - A \frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T,
\]

\[u(x,0) = \psi(x), \quad -\infty < x < \infty, \quad \psi(x) = \begin{bmatrix} \psi^{(1)}(x) \\ \psi^{(2)}(x) \end{bmatrix}.\] (10.89b)

where

\[A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.\]

We will now analyze two difference schemes that approximate problem (10.89b).

**Example 10**

Consider the scheme:

\[
\frac{u_{m+1}^p - u_m^p}{\tau} - A \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad u_0^p = \psi_m.
\] (10.90)

We will be looking for a solution to the vector finite-difference equation (10.90) in the form (10.80):

\[u_m^p = \lambda^p \begin{bmatrix} v_0^p \\ w_0^p \end{bmatrix} e^{i\alpha m}.\]

Substituting this expression into equation (10.90) we obtain:

\[
\frac{\lambda - 1}{\tau} u_0^0 - A \frac{e^{i\alpha} - 1}{h} u_0^0 = 0,
\]

or alternatively,

\[(\lambda - 1)u_0^0 - r(e^{i\alpha} - 1)Au_0^0 = 0, \quad r = \tau/h.\] (10.91)

Equality (10.91) can be considered as a vector form of the system of linear algebraic equations with respect to the components of the vector \(u_0^0\). System (10.91) can be written as:

\[
\begin{bmatrix} \lambda - 1 & -r(e^{i\alpha} - 1) \\ -r(e^{i\alpha} - 1) & \lambda - 1 \end{bmatrix} \begin{bmatrix} v_0^0 \\ w_0^0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.\] (10.92)
System (10.92) may only have a nontrivial solution \( \mathbf{u}^0 = \begin{bmatrix} v_0^0 \\ w_0^0 \end{bmatrix} \) if its determinant turns into zero, which yields the following equation for \( \lambda = \lambda(\alpha) \):

\[
(\lambda - 1)^2 = r^2(e^{i\alpha} - 1)^2.
\]

Consequently,

\[
\lambda_1(\alpha) = 1 - r + re^{i\alpha},
\]

\[
\lambda_2(\alpha) = 1 + r - re^{i\alpha}.
\]

When the parameter \( \alpha \) varies between 0 and \( 2\pi \), the roots \( \lambda_1(\alpha) \) and \( \lambda_2(\alpha) \) sweep two circles of radius \( r \) centered at the locations \( 1 - r \) and \( 1 + r \), respectively. Therefore, the von Neumann stability conditions may never be satisfied; irrespective of any particular value of \( r \).

**Example 11**

Consider the vector Lax-Wendroff scheme:

\[
\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau} - \lambda \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} - \frac{\tau}{2h^2} \lambda^2 (u_{m+1}^{p} - 2u_{m}^{p} + u_{m-1}^{p}) = 0,
\]

\[
u_0^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,
\]

that approximates problem (10.89b) on its smooth solutions with second-order accuracy and that is analogous to the scalar scheme (10.83) for the Cauchy problem (10.82). As in the previous Example 10, the finite-difference equation of (10.93) may only have a non-trivial solution of type (10.80) if the determinant of the corresponding linear system for finding \( \mathbf{u}^0 = \begin{bmatrix} v_0^0 \\ w_0^0 \end{bmatrix} \) turns into zero.

Writing down this determinant and requiring that it be equal to zero, we obtain a quadratic equation with respect to \( \lambda = \lambda(\alpha) \). Its roots can be easily found:

\[
\lambda_1(\alpha) = 1 + ir \sin \alpha - 2r^2 \sin^2 \frac{\alpha}{2},
\]

\[
\lambda_2(\alpha) = 1 - ir \sin \alpha - 2r^2 \sin^2 \frac{\alpha}{2}.
\]

These formulae are analogous to those obtained for the scalar Example 2, and similarly to that example we have:

\[
1 - |\lambda_1(\alpha)|^2 = 4r^2 \sin^4 \frac{\alpha}{2} (1 - r^2).
\]

We therefore see that the spectrum \( \{ \lambda_1(\alpha), \lambda_2(\alpha) \} \) belongs to the unit disk if \( r \leq 1 \).
10.3.4 Stability in $C$

Let us emphasize that the type of stability we have analyzed in the previous Sections 10.3.1–10.3.3 was stability in the sense of the maximum norm (10.71). Alternatively, it is referred to as stability in (the space) $C$. This space contains all bounded numerical sequences. The von Neumann spectral condition (10.78) is necessary for the scheme to be stable in $C$. As for the sufficient conditions of stability in $C$, they turn out to be very delicate, and require quite sophisticated arguments\(^6\) although at a first glance it may seem counterintuitive. Even in the simplest case of a scalar constant coefficient difference equation, and stability with respect to the initial data only, the corresponding analysis would generally go beyond the scope of the current book. Instead, we refer the reader to the work by Fedoryuk [Fed67] (see also his monograph [Fed77, Chapter V, § 4]), and to that of Widlund [Wid66].

10.3.5 Sufficiency of the Spectral Stability Condition in $l^2$

However, a sufficient condition for stability may sometimes be easier to find if we were to use a different norm instead of the maximum norm (10.71). Let, for example,

\[
||u^p||^2 = \sum_{m=-\infty}^{\infty} |u^p_m|^2, \quad ||q^p||^2 = \sum_{m=-\infty}^{\infty} |q^p_m|^2, \quad ||\psi||^2 = \sum_{m=-\infty}^{\infty} |\psi_m|^2,
\]

\[
||u^{(b)}||_{\mathcal{L}_h} = \max_p ||u^p||, \quad ||f^{(b)}||_{\mathcal{F}_h} = \left\| \max_p \||q^p|| F_h = \max\{||\psi||, \max_p ||q^p||\}\right\|.
\]

Relations (10.94) define Euclidean norms for $u^p$, $q^p$, and $\psi$. Accordingly, stability in the sense of the norms given by (10.94) is referred to as stability in (the space) $l^2$. We recall that the space $l^2$ is a Hilbert space of all numerical sequences, for which the sum of squares of absolute values of all their terms is bounded.

Consider a general constant coefficient finite-difference Cauchy problem:

\[
\sum_{j=-j_{\text{left}}}^{j_{\text{right}}} b_j u^p_{m+j} - \sum_{j=-j_{\text{left}}}^{j_{\text{right}}} a_j u^p_{m+j} = q^p_m, \quad u^0_m = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1,
\]

under the assumption that

\[
\sum_{j=-j_{\text{left}}}^{j_{\text{right}}} b_j e^{i\alpha j} \neq 0, \quad 0 \leq \alpha < 2\pi.
\]

Note that all spatially one-dimensional schemes from Section 10.3.3 fit into the category (10.95), even for a relatively narrow range: $j_{\text{left}} = j_{\text{right}} = 1$.

\(^6\)Except for very special circumstances when stability can be proven directly, as done, e.g., in Section 10.1.3 for the first-order upwind scheme.
THEOREM 10.3
For the scheme (10.95) to be stable in $l_2$ with respect to the initial data, i.e., for the following inequality to hold:

$$
\|u^p\| \leq c\|\psi\|, \quad p = 0, 1, \ldots, [T/\tau],
$$

(10.96)

where the constant $c$ does not depend on $h$ [or on $\tau = \tau(h)$], it is necessary and sufficient that the von Neumann condition (10.78) be satisfied, i.e., that the spectrum of the scheme $\lambda = \lambda(\alpha)$ belong to the disk:

$$
|\lambda(\alpha)| \leq 1 + c_1 \tau,
$$

(10.97)

where $c_1$ is another constant that does not depend either on $\alpha$ or on $\tau$.

PROOF We will first prove the sufficiency. By hypotheses of the theorem, the number series $\sum_{m=-\infty}^{\infty} |\psi_m|^2$ converges. Then, the function series of the independent variable $\alpha$:

$$
\sum_{m=-\infty}^{\infty} \psi_m e^{-iam}
$$

also converges in the space $L_2[0, 2\pi]$, and its sum that we denote $\Psi(\alpha)$, $0 \leq \alpha \leq 2\pi$, is a function that has $\psi_m$ as the Fourier coefficients:

$$
\psi_m = \frac{1}{2\pi} \int_{0}^{2\pi} \Psi(\alpha) e^{iam} d\alpha, \quad m = 0, \pm 1, \pm 2, \ldots,
$$

(10.98)

(a realization of the Riesz-Fischer theorem, see, e.g., [KF75]). Moreover, the following relation holds:

$$
\|\psi\|^2 = \sum_{m=-\infty}^{\infty} |\psi_m|^2 = \frac{1}{2\pi} \int_{0}^{2\pi} |\Psi(\alpha)|^2 d\alpha = \frac{1}{2\pi} \|\Psi\|^2_2
$$

known as the Parseval equality.

Consider a homogeneous counterpart to the difference equation (10.95):

$$
\sum_{j=-h_{\text{het}}}^{b_{\text{hoh}}} b_j u_{m+j}^{p+1} - \sum_{j=-h_{\text{het}}}^{a_{\text{hoh}}} a_j u_{m+j}^p = 0,
$$

$m = 0, \pm 1, \pm 2, \ldots, p = 0, 1, \ldots, [T/\tau] - 1$.

For any $\alpha \in [0, 2\pi)$ this equation obviously has a solution of the type:

$$
u_{m}^{p}(\alpha) = \lambda(\alpha)^{p} e^{i am}
$$

(10.99)
for some particular \( \lambda = \lambda(\alpha) \) that can be determined by substitution:

\[
\lambda(\alpha) = \left( \sum_{j=-\infty}^{\infty} a_j e^{\alpha j} \right) \left( \sum_{j=-\infty}^{\infty} b_j e^{\alpha j} \right)^{-1}.
\]

Then, the grid function:

\[
u_m^p = \frac{1}{2\pi} \int_0^{2\pi} \Psi(\alpha) \lambda_p(\alpha) e^{i\alpha m} d\alpha, \quad m = 0, \pm 1, \pm 2, \ldots, \quad (10.100)
\]

provides solution to the Cauchy problem (10.95) for the case \( \phi_m^p = 0 \) because it is a linear combination of solutions \( u_m^p(\alpha) \) of (10.99), and coincides with \( \psi_m \) for \( p = 0 \), see (10.98). Note that the integral on the right-hand side of formula (10.100) converges by virtue of the Parseval equality, because

\[
\int_0^{2\pi} |\Psi(\alpha)|^2 d\alpha < \infty \implies \int_0^{2\pi} |\Psi(\alpha)| d\alpha < \infty.
\]

If the von Neumann spectral condition (10.97) is satisfied, then

\[
|\lambda_p(\alpha)|^p \leq |1 + c_1\tau|^{p/T} \leq e^{c_1T}.
\]

Consequently, using representation (10.100) along with the Parseval equality and inequality (10.101), we can obtain:

\[
\|u_m^p\|^2 = \frac{1}{2\pi} \int_0^{2\pi} |\lambda_p(\alpha)\Psi(\alpha)|^2 d\alpha \leq e^{c_1T} \frac{1}{2\pi} \int_0^{2\pi} |\Psi(\alpha)|^2 d\alpha = e^{c_1T} \|\psi\|^2,
\]

which clearly implies stability with respect to the initial data: \( \|u^p_m\| \leq c\|\psi\| \).

To prove the necessity, we will need to show that if (10.97) holds for no fixed \( c_1 \), then the scheme is unstable. We should emphasize that to demonstrate the instability for the chosen norm (10.94) we may not exploit the unboundedness of the solution \( u_m^p = \lambda_p(\alpha)e^{i\alpha m} \) that takes place in this case, because the grid function \( \{e^{i\alpha m}\} \) does not belong to \( l_2 \).

Rather, let us take a particular \( \Psi(\alpha) \in L_2[0,2\pi] \) such that

\[
\frac{1}{2\pi} \int_0^{2\pi} |\lambda(\alpha)|^2p|\Psi(\alpha)|^2 d\alpha \geq \max_{\alpha} \left( |\lambda(\alpha)|^{2p} - \varepsilon \right) \frac{1}{2\pi} \int_0^{2\pi} |\Psi(\alpha)|^2 d\alpha, \quad (10.102)
\]

where \( \varepsilon > 0 \) is given. For an arbitrary \( \varepsilon \), estimate (10.102) can always be guaranteed by selecting:

\[
\Psi(\alpha) = \begin{cases} 1, & \text{if } \alpha \in [\alpha^*-\delta, \alpha^*+\delta], \\ 0, & \text{if } \alpha \notin [\alpha^*-\delta, \alpha^*+\delta], \end{cases}
\]

where \( \alpha^* = \arg\max_{\alpha} |\lambda(\alpha)| \) and \( \delta > 0 \). Indeed, as the function \( |\lambda(\alpha)|^{2p} \) is continuous, inequality (10.102) will hold for a sufficiently small \( \delta = \delta(\varepsilon) \).
If estimate (10.101) does not take place, then we can find a sequence \( h_k, k = 0, 1, 2, 3, \ldots \), and the corresponding sequence \( \tau_k = \tau(h_k) \) such that

\[
c_k = \left( \max_{\alpha} |\lambda(\alpha, h_k)| \right)^{[T/\tau_k]} \longrightarrow \infty \quad \text{as} \quad k \longrightarrow \infty.
\]

Let us set \( \varepsilon = 1 \) and choose \( \Psi(\alpha) \) to satisfy (10.102). Define \( \psi_m \) as Fourier coefficients of the function \( \Psi(\alpha) \), according to formula (10.98). Then, inequality (10.102) for \( p_k = [T/\tau_k] \) transforms into:

\[
\|u^{p_k}\|^2 \geq (c_k^2 - 1)\|\psi\|^2 \implies \|u^{p_k}\| \geq (c_k - 1)\|\psi\|,
\]

\( c_k \longrightarrow \infty \quad \text{as} \quad k \longrightarrow \infty, \)

i.e., there is indeed no stability (10.96) with respect to the initial data.

Theorem 10.3 establishes equivalence between the von Neumann spectral condition and the \( l_2 \) stability of scheme (10.95) with respect to the initial data. In fact, one can go even further and prove that the von Neumann spectral condition is necessary and sufficient for the full-fledged \( l_2 \) stability of the scheme (10.95) as well, i.e., when the right-hand side \( \phi^p_m \) is not disregarded. One implication, the necessity, immediately follows from Theorem 10.3, because if the von Neumann condition does not hold, then the scheme is unstable even with respect to the initial data. The proof of the other implication, the sufficiency, can be found in [GR87, §25]. This proof is based on using the discrete Green’s functions. In general, once stability with respect to the initial data has been established, stability of the full inhomogeneous problem can be derived using the Duhamel principle. This principle basically says that the solution to the inhomogeneous problem can be obtained as linear superposition of the solutions to some specially chosen homogeneous problems. Consequently, a stability estimate for the inhomogeneous problem can be obtained on the basis of stability estimates for a series of homogeneous problems, see [Str04, Chapter 9].

10.3.6 Scalar Equations vs. Systems

As of yet, our analysis of finite-difference stability has focused primarily on scalar equations; we have only considered a \( 2 \times 2 \) system in Examples 10 and 11 of Section 10.3.3. In Examples 5 and 9 of Section 10.3.3 we have also considered scalar difference equations that connect the values of the solution on more than two consecutive time levels of the grid; they can be reduced to systems of finite-difference equations on two consecutive time levels.

In general, a constant coefficient finite-difference Cauchy problem with vector unknowns (i.e., a system) can be written in the form similar to (10.95):

\[
\sum_{j_{\text{right}}}^{j_{\text{left}}} B_j u_{m+1}^{p+1} - \sum_{j=0}^{j_{\text{left}}} A_j u_{m+1}^p = \phi^p_m, \quad (10.103)
\]

\[
u_m^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1,
\]
under the assumption that the matrices
\[ \sum_{j=-n}^{n} B_j e^{i\alpha j}, \quad 0 \leq \alpha < 2\pi, \]
are non-singular. In formula (10.103), \( u_m^p, \phi_m^p, \) and \( \psi_m \) are grid vector functions of the same fixed dimension, and \( A_j = A_j(h), B_j = B_j(h), j = -n, \ldots, n \), are given square matrices of matching dimension.

Solution of the homogeneous counterpart to the finite-difference equation of (10.103) can be sought for in the form (10.80), where \( u_0 = u_0(\alpha, h) \) and \( \lambda = \lambda(\alpha, h) \) are the eigenvectors and eigenvalues, respectively, of the amplification matrix:

\[ \Lambda = \Lambda(\alpha, h) = \left( \sum_{j=-n}^{n} B_j e^{i\alpha j} \right)^{-1} \left( \sum_{j=-n}^{n} A_j e^{i\alpha j} \right) \tag{10.104} \]

that corresponds to scheme (10.103).

The von Neumann spectral condition (10.97) is clearly necessary for stability of finite-difference systems in all norms. Indeed, if it is not met, then estimate (10.101) will not hold, and the scheme will develop a catastrophic exponential instability that cannot be fixed by any reasonable choice of norms.

Yet the von Neumann condition remains only a necessary stability condition for systems in either \( C \) or \( L_2 \). Regarding \( C \), we have indicated in Section 10.3.4 that the analysis of sufficient conditions becomes very cumbersome already for the scalar case. However, even in the case of \( L_2 \) that is supposedly more amenable for investigation, obtaining the sufficient conditions for systems proves quite difficult.

Qualitatively, the aforementioned difficulties stem from the fact that the amplification matrix (10.104) may have multiple eigenvalues and as a consequence, may not necessarily have a full set of eigenvectors. If a multiple eigenvalue occurs exactly on the unit circle or just outside the unit disk, this may still cause instability even when all the eigenvalues satisfy the von Neumann constraint (10.97). These considerations are, in fact, similar to those that constitute the result of Theorem 9.2.

Of course, if the amplification matrix appears normal (a matrix that commutes with its conjugate) and therefore unitarily diagonalizable, then none of the aforementioned difficulties is present, and the von Neumann condition becomes not only necessary but also sufficient for stability of the vector scheme (10.103) in \( L_2 \).

Otherwise, the question of stability for scheme (10.103) can be equivalently reformulated using the new concept of stability for families of matrices. A family of square matrices (of a given fixed dimension) is said to be stable if there is a constant \( K > 0 \) such that for any particular matrix \( \Lambda \) from the family, and any positive integer \( p \), the following estimate holds: \( ||\Lambda^p|| \leq K \). Scheme (10.103) is stable in \( L_2 \) if and only if the family of amplification matrices \( \Lambda = \Lambda(\alpha, h) \) given by (10.104) is stable in the sense of the previous definition (this family is parameterized by \( \alpha \in [0, 2\pi) \) and \( h > 0 \)). The following theorem, known as the Kreiss matrix theorem, provides some necessary and sufficient conditions for a family of matrices to be stable.
**THEOREM 10.4 (Kreiss)**

Stability of a family of matrices $\Lambda$ is equivalent to any of the following conditions:

1. There is a constant $C_1 > 0$ such that for any matrix $\Lambda$ from the given family, and any complex number $z$, $|z| > 1$, there is a resolvent $(\Lambda - zI)^{-1}$ bounded as:
   \[ \| (\Lambda - zI)^{-1} \| \leq \frac{C_1}{|z| - 1}. \]

2. There are constants $C_2 > 0$ and $C_3 > 0$, and for any matrix $\Lambda$ from the given family there is a non-singular matrix $M$ such that $\| M \| \leq C_2$, $\| M^{-1} \| \leq C_2$, and the matrix $D \equiv M \Lambda M^{-1}$ is upper triangular, with the off-diagonal entries that satisfy:
   \[ |d_{ij}| \leq C_3 \min \{ 1 - \kappa_i, 1 - \kappa_j \}, \]
   where $\kappa_i = d_{ii}$ and $\kappa_j = d_{jj}$ are the corresponding diagonal entries of $D$, i.e., the eigenvalues of $\Lambda$.

3. There is a constant $C_4 > 0$, and for any matrix $\Lambda$ from the given family there is a Hermitian positive definite matrix $H$, such that
   \[ C_4^{-1} I \leq H \leq C_4 I \text{ and } \Lambda^T H \Lambda \leq H. \]

The proof of Theorem 10.4 can be found, e.g., in [RM67, Chapter 4] or [Str04, Chapter 9].

**Exercises**

1. Consider the so-called weighted scheme for the heat equation:
   \[
   \frac{u_m^{p+1} - u_m^p}{\tau} = \frac{u_{m+1}^{p+1} - 2u_m^{p+1} + u_{m-1}^{p+1}}{h^2} + (1 - \sigma) \frac{u_m^{p} - 2u_m^p + u_{m-1}^p}{h^2},
   \]
   where $u_m^0 = \psi_m$, $m = 0, \pm 1, \pm 2, \ldots$, $p = 0, 1, \ldots, [T/\tau] - 1$,
   where the real parameter $\sigma \in [0, 1]$ is called the weight (between the fully explicit scheme, $\sigma = 0$, and fully implicit scheme, $\sigma = 1$). What values of $\sigma$ guarantee that the scheme will meet the von Neumann stability condition for any $r = \tau/h^2 = \text{const}$?

2. Consider the Cauchy problem (10.87) for the heat equation. The scheme:
   \[
   \frac{u_m^{p+1} - u_m^p}{2\tau} - \alpha^2 \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \psi_m^p,
   \]
   where we assume that $\psi(x, 0) \equiv 0$ and define:
   \[
   \psi_m = u(x, 0) + \tau \frac{\partial u(x, 0)}{\partial t} \bigg|_{x=x_m} = u(x, 0) + \tau \alpha^2 \frac{\partial^2 u(x, 0)}{\partial x^2} \bigg|_{x=x_m} = \psi_m + \tau \alpha^2 \psi''(x_m),
   \]
   approximates problem (10.87) on its smooth solutions with the accuracy $O(\tau^2 + h^2)$. Does this scheme satisfy the von Neumann spectral stability condition?
3. For the two-dimensional Cauchy problem:

\[ \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = \psi(x,y,t), \quad -\infty < x, y < \infty, \quad 0 < t \leq T; \]

\[ u(x,y,0) = \psi(x,y), \quad -\infty < x, y < \infty, \]

investigate the von Neumann spectral stability of:

a) The first-order explicit scheme:

\[ \frac{u_{m,n}^{p+1} - u_{m,n}^p}{\tau} - \frac{u_{m+1,n}^{p+1} - u_{m,n}^{p+1}}{h} - \frac{u_{m,n+1}^{p+1} - u_{m,n}^{p+1}}{h} = \psi_{m,n}^p, \]

\[ u_{m,n}^0 = \psi_{m,n}, \quad m, n = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1; \]

b) The second-order explicit scheme:

\[ \frac{u_{m,n}^{p+1} - u_{m,n}^p}{2\tau} - \frac{u_{m+1,n}^{p+1} - u_{m,n}^{p+1}}{2h} - \frac{u_{m,n+1}^{p+1} - u_{m,n}^{p+1}}{2h} = \psi_{m,n}^p, \]

\[ u_{m,n}^0 = \psi_{m,n}, \quad u_{m,n}^1 = \psi_{m,n} + \tau [\psi_1(x_m,y_n) + \psi_2(x_m,y_n) + \psi(x_m, y_n, 0)], \]

\[ m, n = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1. \]

4. Investigate the von Neumann spectral stability of the implicit two-dimensional scheme for the homogeneous heat equation:

\[ \frac{u_{m,n}^{p+1} - u_{m,n}^p}{\tau} - \frac{\frac{1}{\tau^2} [u_{m+1,n}^{p+1} - 2u_{m,n}^{p+1} + u_{m-1,n}^{p+1}]}{h^2} - \frac{\frac{1}{\tau^2} [u_{m,n+1}^{p+1} - 2u_{m,n}^{p+1} + u_{m,n-1}^{p+1}]}{h^2} = \psi_{m,n}^p, \]

\[ u_{m,n}^0 = \psi_{m,n}, \quad m, n = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1. \]

5. Investigate the von Neumann stability of the implicit upwind scheme designed for solving the Cauchy problem (10.82):

\[ \frac{u_{m,n}^{p+1} - u_{m,n}^p}{\tau} - \frac{u_{m+1,n}^{p+1} - u_{m,n}^{p+1}}{h} = \psi_{m,n}^p, \]

\[ u_{m,n}^0 = \psi_{m,n}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1. \]

6. Investigate the von Neumann stability of the implicit downwind scheme designed for solving the Cauchy problem (10.82):

\[ \frac{u_{m,n}^{p+1} - u_{m,n}^p}{\tau} - \frac{u_{m,n}^{p+1} - u_{m-1,n}^{p+1}}{h} = \psi_{m,n}^p, \]

\[ u_{m,n}^0 = \psi_{m,n}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1. \]

7. Investigate the von Neumann stability of the implicit central scheme designed for solving the Cauchy problem (10.82):

\[ \frac{u_{m,n}^{p+1} - u_{m,n}^p}{2\tau} - \frac{u_{m+1,n}^{p+1} - u_{m,n}^{p+1}}{2h} = \psi_{m,n}^p, \]

\[ u_{m,n}^0 = \psi_{m,n}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, [T/\tau] - 1. \]
8. Transform the leap-frog scheme (10.86) of Example 5, Section 10.3.3, and the central-difference scheme for the d’Alembert equation of Example 9, Section 10.3.3, to the schemes written for finite-difference systems, as opposed to scalar equations, but connecting only two, as opposed to three, consecutive time levels of the grid. Investigate the von Neumann stability by calculating spectra of the corresponding amplification matrices (10.104).

**Hint.** Use the difference \( \{u_m^{p+1} - u_m^p\} \) as the second unknown grid function.

### 10.4 Stability for Problems with Variable Coefficients

The von Neumann necessary condition that we have introduced in Section 10.3 to analyze stability of linear finite-difference Cauchy problems with constant coefficients can, in fact, be applied to a wider class of formulations. A simple extension that we describe in this section allows one to exploit the von Neumann condition to analyze stability of problems with variable coefficients (continuous, but not necessarily constant) and even some nonlinear problems.

#### 10.4.1 The Principle of Frozen Coefficients

Introduce a uniform Cartesian grid: \( x_m = mh, m = 0, \pm 1, \pm 2, \ldots, t_p = p\tau, p = 0, 1, 2, \ldots \), and consider a finite-difference Cauchy problem for the homogeneous heat equation with the variable coefficient of heat conduction \( a = a(x, t) \):

\[
\frac{u_m^{p+1} - u_m^p}{\tau} - a(x_m, t_p) \frac{u_m^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \quad u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \pm 2, \ldots, \quad p \geq 0. \tag{10.108}
\]

Next, take an arbitrary point \((\bar{x}, \bar{t})\) in the domain of problem (10.108) and “freeze” the coefficients of problem (10.108) at this point. Then, we arrive at the constant-coefficient finite-difference equation:

\[
\frac{u_m^{p+1} - u_m^p}{\tau} - a(\bar{x}, \bar{t}) \frac{u_m^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p \geq 0. \tag{10.109}
\]

Having obtained equation (10.109), we can formulate the following principle of frozen coefficients. *For the original variable-coefficient Cauchy problem (10.108) to be stable it is necessary that the constant-coefficient Cauchy problem for the difference equation (10.109) satisfy the von Neumann spectral stability condition.*

To justify the principle of frozen coefficients, we will provide an heuristic argument rather than a proof. When the grid is refined, the variation of the coefficient \(a(x, t)\) in a neighborhood of the point \((\bar{x}, \bar{t})\) becomes smaller if measured over any
finite fixed number of grid cells that have size $h$ in space and size $\tau$ in time. This
is true because of the continuity of the function $a = a(x,t)$. Therefore, the finer the
grid, the closer the values of $a(x,t)$ will be to that of $a(\bar{x}, \bar{t})$ anywhere on the grid
— but not farther away from $(\bar{x}, \bar{t})$ than a fixed number of cells. Consequently, if
we were to perturb the solution of problem (10.108) on a fine grid at the moment of
time $t = \bar{t}$ near the space location $x = \bar{x}$, then over the short intervals of time these
perturbations would have evolved pretty much the same way as perturbations of the
solution to the constant-coefficient equation (10.109).

It is clear that the previous argument is quite general. It is not affected by the
number of space dimensions, the number of unknown functions, or the specific type of
the finite-difference equation or system.

In Section 10.3.3, we analyzed a Cauchy problem for the equation of type
(10.109), see Example 6, and found that for the von Neumann stability condition
to hold the ratio $r = \tau/h^2$ must satisfy the inequality [cf. inequality (10.88)]:

$$ r \leq \frac{1}{2a(\bar{x}, \bar{t})} $$

(10.110)

According to the principle of frozen coefficients, stability of scheme (10.108) re-
quires that condition (10.110) be met for any $(\bar{x}, \bar{t})$. Therefore, altogether the ratio
$r = \tau/h^2$ must satisfy the inequality:

$$ r \leq \frac{1}{2\max_{x,t} a(x,t)} . $$

(10.111)

The principle of frozen coefficients can also provide an heuristic argument for the
analysis of stability of nonlinear difference equations. We illustrate this using an
example of the Cauchy problem for the nonlinear heat equation:

$$ \frac{\partial u}{\partial t} - (1 + u^2) \frac{\partial^2 u}{\partial x^2} = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T, $$

$$ u(x,0) = \psi(x), \quad -\infty < x < \infty. $$

We approximate this problem with the explicit scheme:

$$ \frac{u_m^{p+1} - u_m^p}{\tau_p} - (1 + |u_m|^2) \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, $$

(10.112)

$$ u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, 2, \ldots. $$

The scheme is built on a uniform spatial grid: $x_m = mh, m = 0, \pm 1, \pm 2, \ldots$, but with a
non-uniform temporal grid, such that the size $\tau_p = t_{p+1} - t_p$ may vary from one time
level to another. The finite-difference solution can still be obtained by marching.

Assume that we have already marched all the way up the time level $t_p$ and com-
puted the solution $u_m^p, m = 0, \pm 1, \pm 2, \ldots$. To continue marching, we first need to
select the next grid size $\tau_{p+1}$. This can be done by interpreting the finite-difference
equation to be solved at $t = t_p$ with respect to $u_m^{p+1}$ as the linear equation:

$$ \frac{u_m^{p+1} - u_m^p}{\tau_p} - a_m^p \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0 $$

(10.113)
with the given variable coefficient of heat conduction: \( a_p^m \equiv (1 + |u_m|^2) \). Indeed, it is natural to think that the values of the grid function \( u_m^p \) are close to the values \( u(x_m, t_p) \) of the continuous solution \( u(x, t) \). Then, the heat conduction coefficient \( a_p^m \) will be close to the grid projection \( a(x_m, t_p) \) of the continuous function \( a(x, t) = 1 + u(x, t)^2 \). This function will vary only slightly over a few temporal steps.

By applying the principle of frozen coefficients to equation (10.113), we arrive at the constraint (10.111) for the grid sizes that is necessary for stability:

\[
\frac{\tau_p}{h^2} = \frac{1}{2 \max_m a_m^p} \leq \frac{1}{2 \max_m (1 + |u_m|^2)}.
\]

Consequently, when marching equation (10.112) one should select the temporal grid size \( \tau_p \) for each \( p = 0, 1, 2, \ldots \) based on the inequality:

\[
\tau_p \leq \frac{h^2}{2 \max_m (1 + |u_m|^2)}.
\]

Numerical experiments corroborate correctness of these heuristic arguments.

If the necessary stability condition obtained by considering the Cauchy problem with frozen coefficients happens to be violated at some point of the domain, then in general there will be no stability. Let us emphasize, however, that the analysis based on freezing the coefficients has only been conducted for the genuine Cauchy problem, i.e., when the range of the independent variable is \((-\infty, \infty)\). If we were to consider an initial boundary value problem formulated on a finite interval, then the foregoing analysis would not be sufficient. Even if the necessary stability condition based on the principle of frozen coefficients were to hold, the overall problem on the finite interval could still be either stable or unstable depending on the choice of the boundary conditions at the endpoints of the interval. In Section 10.5.1 we discuss the Babenko-Gelfand stability criterion that takes into account the effect of boundary conditions in the case of a problem on an interval.

### 10.4.2 Dissipation of Finite-Difference Schemes

A very useful concept that can lead to a sufficient condition of stability for some problems with variable coefficients is that of dissipation. We will first illustrate it using scalar finite-difference equations with constant coefficients. We will again use the notion of the spectrum \( \lambda = \lambda(\alpha, h) \) of the transition operator introduced in Section 10.3.2, see page 201, and subsequently studied in Section 10.3.3. For convenience, in this section we will assume that the frequency range is \(-\pi < \alpha \leq \pi\), rather than \(0 \leq \alpha < 2\pi\) that we have used previously. Both choices are clearly equivalent, because the harmonics \( e^{i\alpha m} \) are \(2\pi\)-periodic.

**Definition 10.4** A scheme is said to be dissipative of order \( 2d \) if \( d \) is the smallest positive integer for which one can find a constant \( \delta > 0 \) that
would not depend on $h$ (and on $\tau$ either), such that for all $|\alpha| \leq \pi$ and all sufficiently small grid sizes $h$, the following inequality will hold:

$$|\lambda(\alpha, h)| \leq 1 - \delta |\alpha|^{2d}. \quad (10.114)$$

Recall that in the case of constant coefficients we analyze solutions in the form $\lambda e^{i\alpha m}$, see formulae (10.76) and (10.80), and therefore, for a given $\alpha$, the amplification factor $\lambda(\alpha, h)$ indicates how rapidly the harmonic $e^{i\alpha m}$ decays as the number of time steps $p$ increases. From this perspective, by introducing the concept of dissipation, see inequality (10.114), we quantify the rate of this decay for individual harmonics in terms of a positive constant $\delta$ and some positive power $2d$ of the frequency $\alpha$. Alternatively, if instead of inequality (10.114) we were to have $|\lambda(\alpha, h)| = 1$, i.e., if the spectrum were to lie entirely on the unit circle, then there would be no decay of the amplitude for any harmonic $e^{i\alpha m}$ (equivalently, no dissipation), and the scheme would be called non-dissipative, accordingly.

Let us now analyze some typical examples.

**Example 1**

The first-order upwind scheme (see Section 10.3.1):

$$\frac{u_{m+1}^p - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0,$$

$$u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, 2, \ldots, [T/\tau] - 1,$$

for the Cauchy problem:

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T,$$

$$u(x, 0) = \psi(x), \quad -\infty < x < \infty,$$

has its spectrum defined by formula (10.77):

$$\lambda(\alpha) = 1 - r + r e^{i\alpha}, \quad r = \tau/h = \text{const},$$

and consequently,

$$|\lambda(\alpha)|^2 = 1 - 4r(1 - r) \sin^2 \frac{\alpha}{2}.\quad (10.115)$$

For $r > 1$, this spectrum is outside the unit disk, see Figure 10.5(b), and it makes no sense to discuss dissipation according to Definition 10.4. If $r = 1$, then $|\lambda(\alpha)| = 1$, and the scheme is non-dissipative. If $r < 1$, we can use the inequality $|\alpha|/4 \leq |\sin(\alpha/2)|$ for $\alpha \in [-\pi, \pi]$, and by setting $\delta = r(1 - r)/8 > 0$ we obtain:

$$|\lambda(\alpha)|^2 = 1 - 4r(1 - r) \sin^2 \frac{\alpha}{2} \leq 1 - 4r(1 - r) \left(\frac{|\alpha|}{16}\right)^2,$$

$$= 1 - 2\delta |\alpha|^2 \leq 1 - 2\delta |\alpha|^2 + \delta^2 |\alpha|^4 = (1 - \delta |\alpha|^2)^2.$$

Therefore, the first-order upwind scheme for $r < 1$ is dissipative of order $2d = 2$. 


Finite-Difference Schemes for Partial Differential Equations

Example 2

For the Lax-Wendroff scheme [see formula (10.83)]:

\[
\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau} - \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} - \frac{\tau}{2h^2} (u_{m+1}^{p} - 2u_{m}^{p} + u_{m-1}^{p}) = 0,
\]

(10.117)

we determined in Example 2 of Section 10.3.3 that

\[
|\lambda(\alpha)|^2 = 1 - 4r^2 (1 - r^2) \sin^2 \frac{\alpha}{2}, \quad r = \frac{\tau}{h} = \text{const.}
\]

The scheme is, again, non-dissipative when \( r = 1 \). For \( r < 1 \), it is easy to obtain:

\[
|\lambda(\alpha)| \leq 1 - \delta |\alpha|^4,
\]

where \( \delta = r^2 (1 - r^2)/128 \), i.e., the scheme appears dissipative of order \( 2d = 4 \).

Example 3

In Section 9.2.6 of Chapter 9 we introduced the Crank-Nicolson scheme for ordinary differential equations, see formula (9.33). A very similar scheme can be designed and used in the case of partial differential equations. For the Cauchy problem (10.116) we write:

\[
\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau} - \frac{1}{2} \left[ \frac{u_{m+1}^{p+1} - u_{m-1}^{p+1}}{2h} + \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} \right] = 0,
\]

(10.118)

\[
u_{m}^{0} = \psi(x_{m}), \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, 2, \ldots, \frac{T}{\tau} - 1.
\]

Unlike the schemes discussed in the previous Examples 1 & 2, the Crank-Nicolson scheme (10.118) is implicit. It approximates problem (10.116) with the second order of accuracy with respect to \( h \), provided that \( r = \tau/h = \text{const.} \)

For the spectrum of scheme (10.118) we can easily find:

\[
\frac{\lambda - 1}{\tau} - \frac{1}{2} \left[ \frac{\lambda e^{i\alpha} - e^{-i\alpha}}{2h} + \frac{e^{i\alpha} - e^{-i\alpha}}{2h} \right],
\]

which yields:

\[
\lambda(\alpha) = \left( 1 + \frac{ir}{2} \sin \alpha \right) \cdot \left( 1 - \frac{ir}{2} \sin \alpha \right)^{-1}.
\]

Consequently,

\[
|\lambda(\alpha)| = 1
\]

irrespective of the specific value of \( r \). Therefore, the Crank-Nicolson scheme (10.118) is non-dissipative. We also see that it satisfies the von Neumann stability condition (10.78).
The concept of dissipation introduced in Definition 10.4 for scalar constant-coefficient difference equations can be extended to the case of variable coefficients and to the case of systems, including systems in multiple space dimensions [RM67, Chapter 5]. For the reason of simplicity, in this book we will only discuss one-dimensional systems. Instead of the scalar problem (10.116) consider a Cauchy problem for the hyperbolic system:

\[ \frac{\partial u}{\partial t} - A(x) \frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad 0 < t \leq T, \]
\[ u(x,0) = \psi(x), \quad -\infty < x < \infty, \quad (10.119) \]

where \( u = u(x,t) \) is the unknown vector function of a given fixed dimension \( K \). The \( K \times K \) square matrix \( A = A(x) \) is assumed Hermitian for all \( x \).

We will approximate problem (10.119) by an explicit finite-difference scheme of type (10.103):

\[ u_{m+1}^p - \sum_{j=-j_{\text{left}}}^{j_{\text{right}}} A_j(x,h) u_{m+j}^p = 0, \]
\[ u_0^0 = \psi_m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \left[ T/\tau \right] - 1, \quad (10.120) \]

where \( A_j(x,h), \quad j = -j_{\text{left}}, \ldots, j_{\text{right}}, \) are also assumed Hermitian. The amplification matrix (10.104) for scheme (10.120) is obtained by exploiting the same idea as outlined in Section 10.4.1 — i.e., by freezing the matrix coefficients \( A_j \) for each particular location \( x \):

\[ \Lambda = \Lambda(x,\alpha,h) = \sum_{j=-j_{\text{left}}}^{j_{\text{right}}} A_j(x,h)e^{i\alpha j}. \quad (10.121) \]

**DEFINITION 10.5** Scheme (10.120) is said to be dissipative of order \( 2d \) if \( d \) is the smallest positive integer for which there is a constant \( \delta > 0 \) that does not depend on either \( h \) or \( x \), such that for all \( |\alpha| \leq \pi \), all \( x \in (-\infty, \infty) \), and all sufficiently small grid sizes \( h \), the following inequalities hold:

\[ |\lambda_k(x,\alpha,h)| \leq 1 - \delta|\alpha|^{2d}, \quad k = 1, 2, \ldots, K, \quad (10.122) \]

where \( \lambda_k = \lambda_k(x,\alpha,h) \) are eigenvalues of the amplification matrix \( \Lambda = \Lambda(x,\alpha,h) \) of (10.121).

The following theorem due to Kreiss provides a sufficient condition of stability for dissipative finite-difference schemes (see [RM67, Chapter 5] for the proof):

**THEOREM 10.5 (Kreiss)**

Let the matrix \( A(x) \) in equation (10.119) and the matrices \( A_j(x,h), \quad j = -j_{\text{left}}, \ldots, j_{\text{right}}, \) in equation (10.120) be Hermitian, uniformly bounded, and
uniformly Lipschitz-continuous with respect to $x$. Then, if scheme (10.120) is dissipative of order $2d$ in the sense of Definition 10.5, and has accuracy of order $2d - 1$ for some positive integer $d$, then this scheme is stable in $l_2$.

Having formulated the Kreiss Theorem 10.5, we can revisit our previously analyzed examples. The first-order upwind scheme (10.115) of Example 1 is dissipative of order $2d = 2$ when $r = \tau/h < 1$. Its accuracy is $O(h)$, and therefore we can set $d = 1$, apply Theorem 10.5, and conclude that this scheme is stable.

The Lax-Wendroff scheme (10.117) of Example 2 is dissipative of order $2d = 4$ when $r < 1$, i.e., we need to consider $d = 2$. However, this scheme is only second order accurate, while $2d - 1 = 3$. Therefore, Theorem 10.5 does not allow us to immediately judge the stability of scheme (10.117).

The Crank-Nicolson scheme (10.118) of Example 3 is non-dissipative, and therefore Theorem 10.5 does not apply.

On the other hand, let us note that according to Theorem 10.3 (proven in Section 10.3.5), all three foregoing schemes are, in fact, stable in the sense of $l_2$ when $r \leq 1$, i.e., when they satisfy the von Neumann condition. Theorem 10.3 says that the von Neumann condition is necessary and sufficient for stability of a scalar constant-coefficient finite-difference scheme in $l_2$.

As such, we see that for the case of scalar constant-coefficient finite-difference equations, the result of Theorem 10.5 may, in fact, be superseded by that of Theorem 10.3. We know, however, that for systems the von Neumann criterion alone provides only a necessary condition for stability, see Section 10.3.6. Moreover, for the case of variable coefficients, the principle of frozen coefficients, see Section 10.4.1, also provides a necessary condition for stability only. These are the cases when the sufficient condition given by Theorem 10.5 is most instrumental.

Note also that there is a special large group of methods used in particular for the computation of fluid flows, when dissipation (viscosity) is artificially added to the scheme to improve its stability characteristics, see Section 11.2.1.

**Exercises**

1. Prove that the Crank-Nicolson scheme (10.118) has accuracy $O(h^2)$ provided that $r = \tau/h = \text{const.}$

2. Show that the Lax-Friedrichs scheme (10.85) is dissipative of order $2d = 2$, although not strictly in the sense of Definition 10.4. Prove that it rather satisfies inequality (10.114) for all $|\alpha| \leq \pi - \epsilon$, where $\epsilon > 0$ may be arbitrary.

3. Use Theorem 10.5 to study stability of the scheme:

$$
\frac{u^{m+1} - u^{m}}{\tau} - a(x_m) \frac{u^{p+1} - u^{p}}{h} = 0,
$$

$$
u^0_m = \psi(x_m), \quad m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, 2, \ldots, \quad p = 0, 1, 2, \ldots, \quad [T/\tau] - 1.
$$

\footnote{The Theorem can still be applied though, but only after a change of variables, see [RM67, Chapter 5].}
Theoretical Introduction to Numerical Analysis

4. Show that the implicit upwind scheme (10.105) for the Cauchy problem (10.116) is dissipative of order $d = 2$.

5. Show that the implicit downwind scheme (10.106) for the Cauchy problem (10.116) is dissipative of order $d = 2$ when $r > 1$.

6. Show that the implicit central scheme (10.107) for the Cauchy problem (10.116) is dissipative of order $d = 2$ in the same non-strict sense as outlined in Exercise 2.

10.5 Stability for Initial Boundary Value Problems

Instead of the Cauchy problem (10.108), let us now consider an initial boundary value problem for the heat equation formulated on the finite interval $0 \leq x \leq 1$:

\[
\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau} - a(x_{m}, t_{p}) \frac{u_{m+1}^{p} - 2u_{m}^{p} + u_{m-1}^{p}}{h^2} = 0,
\]

\[
u_{0}^{p} = \psi(x_{m}), \quad I_{1}u_{0}^{p+1} = 0, \quad I_{2}u_{M}^{p+1} = 0,
\]

$m = 0, 1, 2, \ldots, M, \quad p \geq 0.$

In formula (10.123), we assume that the grid is uniform: $x_{m} = mh, m = 0, 1, 2, \ldots, M, \quad M = 1/h, t_{p} = pt, p = 0, 1, 2, \ldots$, and denote by $I_{1}$ and $I_{2}$ the operators of the boundary conditions at the left and right endpoints of the interval, respectively.

10.5.1 The Babenko-Gelfand Criterion

To analyze stability of the difference problem (10.123), we will first develop a heuristic argument based on freezing the coefficients; this argument will further extend the previous considerations of Section 10.4.1. In Section 10.4.1, we have noticed that because of the continuity of the coefficient $a = a(x, t)$, its variation in the range of a fixed number of cells from any given point $(\tilde{x}, \tilde{t})$ becomes smaller when the grid is refined. In the context of the initial boundary value problem (10.123), as opposed to the initial value problem (10.108), we supplement this consideration by another obvious observation. If the point $(\tilde{x}, \tilde{t})$ lies inside the domain, then the distance from this point to either of the endpoints, $x = 0$ or $x = 1$, measured in the number of grid cells (of size $h$) will increase with no bound when $h \rightarrow 0$. In other words, on fine grids the point $(\tilde{x}, \tilde{t})$ can be considered to be located far away from the boundaries. Consequently, we can still claim that for small $h$ the perturbations
superimposed on the solution of problem (10.123) at the moment of time \( t = \tilde{t} \) near any interior space location \( x = \tilde{x} \) will evolve similarly to how the perturbations of the solution to the same “old” constant-coefficient equation (10.109) would have evolved. This, in turn, implies that stability of the scheme (10.109) for every \((\tilde{x}, \tilde{t})\) inside the domain is still necessary for the overall stability of scheme (10.123).

The foregoing heuristic argument, however, becomes far less convincing if the point \((\tilde{x}, \tilde{t})\) happens to lie precisely on one of the lateral boundaries: \( x = 0 \) or \( x = 1 \). For example, when we let \( \tilde{x} = 0 \), the distance from \((\tilde{x}, \tilde{t})\) to any fixed location \( x > 0 \) (and in particular, to the right endpoint \( x = 1 \)) measured in the number of grid cells will again increase with no bound as \( h \to 0 \). Yet the number of grid cells to the left endpoint \( x = 0 \) will not change and will remain equal to zero. In other words, the point \((\tilde{x}, \tilde{t})\) will never be far from the left boundary, no matter how fine the grid may be. Consequently, we can no longer expect that perturbations of the solution to problem (10.123) near \( \tilde{x} = 0 \) will behave similarly to perturbations of the solution to equation (10.109), as the latter is formulated on the grid infinite in both directions.

Instead, we shall rather expect that over the short periods of time the perturbations of the solution to problem (10.123) near the left endpoint \( x = 0 \) will develop analogously to perturbations of the solution to the following constant-coefficient problem:

\[
\frac{u_m^{p+1} - u_m^p}{\tau} - a(0, \tilde{t}) \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \quad \text{for } m = 0, 1, 2, \ldots, p \geq 0. \tag{10.124}
\]

Problem (10.124) is formulated on the semi-infinite grid: \( m = 0, 1, 2, \ldots \) (i.e., semi-infinite line \( x \geq 0 \)). It is obtained from the original problem (10.123) by freezing the coefficient \( a(x, t) \) at the left endpoint of the interval \( 0 \leq x \leq 1 \) and by simultaneously “pushing” the right boundary off all the way to \( +\infty \). Problem (10.124) shall be analyzed only for those grid functions \( u^p = \{u_0^p, u_1^p, \ldots\} \) that satisfy:

\[
\lim_{m \to \infty} u_m^p = 0, \quad \text{as } m \to +\infty. \tag{10.125}
\]

Indeed, only in this case will the perturbation be concentrated near the left boundary \( x = 0 \), and only for the perturbations of this type will the problems (10.123) and (10.124) be similar to one another in the vicinity of \( x = 0 \).

Likewise, the behavior of perturbations to the solution of problem (10.123) near the right endpoint \( x = 1 \) should resemble that for the problem:

\[
\frac{u_m^{p+1} - u_m^p}{\tau} - a(1, \tilde{t}) \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \quad \text{for } m = 0, 1, 2, \ldots, M \geq 0, \tag{10.126}
\]

that has only one boundary at \( m = M \). Problem (10.126) is derived from problem (10.123) by freezing the coefficient \( a(x, t) \) at the right endpoint of the interval \( 0 \leq x \leq 1 \) and by simultaneously pushing the left boundary off all the way to \( -\infty \). It should be considered only for the grid functions \( u^p = \{\ldots, u_{-2}^p, u_0^p, u_1^p, \ldots, u_M^p\} \) that satisfy:

\[
\lim_{m \to -\infty} u_m^p = 0, \quad \text{as } m \to -\infty. \tag{10.127}
\]
The three problems: (10.109), (10.124), and (10.126), are easier to investigate than the original problem (10.123), because they are all $h$ independent provided that $r = \tau / h^2 = \text{const}$, and they all have constant coefficients.

Thus, the issue of studying stability for the scheme (10.123), with the effect of the boundary conditions taken into account, can be addressed as follows. One needs to formulate three auxiliary problems: (10.109), (10.124), and (10.126). For each of these three $h$ independent problems, one needs to find all those numbers $\lambda$ (eigenvalues of the transition operator from $u^0$ to $u^{p+1}$), for which solutions of the type

$$u^p_m = \lambda^p u^0_m$$  \hspace{1cm} (10.128)

exist. In doing so, for problem (10.109), the function $u^0 = \{u^0_m\}$, $m = 0, \pm 1, \pm 2, \ldots$, has to be bounded on the grid. For problem (10.124), the grid function $u^0 = \{u^0_m\}$, $m \geq 0$, has to satisfy: $u^0_m \to 0$ as $m \to +\infty$, and for problem (10.124), the grid function $u^0 = \{u^0_m\}$, $m \leq M$, has to satisfy: $u^0_m \to 0$ as $m \to -\infty$. For scheme (10.123) to be stable, it is necessary that the overall spectrum of the difference initial boundary value problem, i.e., all eigenvalues of all three problems: (10.109), (10.124), and (10.126), belong to the unit disk: $|\lambda| \leq 1$, on the complex plane. This is the Babenko-Gelfand stability criterion. Note that problem (10.109) has to be considered for every fixed $\tilde{x} \in (0, 1)$ and all $\tilde{t}$.

**REMARK 10.1** Before we continue to study problem (10.123), let us present an important intermediate conclusion that can already be drawn based on the foregoing qualitative analysis. For stability of the pure Cauchy problem (10.108) that has no boundary conditions it is necessary that finite-difference equations (10.109) be stable in the von Neumann sense $\forall (\tilde{x}, \tilde{t})$. This requirement remains necessary for stability of the initial boundary value problem (10.123) as well. Moreover, when boundary conditions are present, two more auxiliary problems: (10.124) and (10.126), have to be stable in a similar sense. Therefore, adding boundary conditions to a finite-difference Cauchy problem will not, generally speaking, improve its stability. Boundary conditions may either remain neutral or hamper the overall stability if, for example, problem (10.109) appears stable but one of the problems (10.124) or (10.126) happens to be unstable. Later on, we will discuss this phenomenon in more detail.

Let us now assume for simplicity that $a(x, t) \equiv 1$ in problem (10.123), and let us calculate the spectra of the three auxiliary problems (10.109), (10.124), and (10.126) for various boundary conditions $I_1 u^0_{m+1} = 0$ and $I_3 u^0_{m-1} = 0$.

Substituting the solution in the form $u^p_m = \lambda^p u^0_m$ into the finite-difference equation (10.109), we obtain:

$$(\lambda - 1)u_m - r(u_{m+1} - 2u_m + u_{m-1}) = 0, \quad r = \tau / h^2,$$
which immediately yields:

$$u_{m+1} - \frac{\lambda - 1 + 2r}{r} u_m + u_{m-1} = 0. \hspace{1cm} (10.129)$$
This is a second-order homogeneous ordinary difference equation. To find the general solution of equation (10.129) we write down its algebraic characteristic equation:

\[ q^2 - \frac{\lambda - 1 + 2r}{r} q + 1 = 0. \]  
(10.130)

If \( q \) is a root of the quadratic equation (10.130), then the grid function

\[ u_m^p = \lambda^p q^m \]

solves the homogeneous finite-difference equation:

\[ \frac{u_{m+1}^p - u_m^p}{\tau} + \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0. \]  
(10.131)

If \( |q| = 1 \), i.e., if \( q = e^{i\alpha} \), then the grid function

\[ u_m^p = \lambda^p e^{i\alpha m}, \]

which is obviously bounded for \( m \to +\infty \) and \( m \to -\infty \), yields the solution of equation (10.131), provided that

\[ \lambda = 1 - 4r \sin^2 \frac{\alpha}{2}, \quad 0 \leq \alpha < 2\pi, \]

see Example 6 of Section 10.3.3. These \( \lambda = \lambda(\alpha) \) fill the interval \( 1 - 4r \leq \lambda \leq 1 \) of the real axis, see Figure 10.8 on page 207. Therefore, interval \( 1 - 4r \leq \lambda \leq 1 \) is the spectrum of problem (10.109) for \( a(\tilde{x}, \tilde{t}) = 1 \), i.e., of problem (10.131). This problem has no eigenvalues that lie outside of the interval \( 1 - 4r \leq \lambda \leq 1 \), because if the characteristic equation (10.130) does not have a root \( q \) with \( |q| = 1 \), then equation (10.129) may have no solution bounded for \( m \to \pm \infty \).

If \( \lambda \) does not belong to the interval \( 1 - 4r \leq \lambda \leq 1 \), then the absolute values of both roots of the characteristic equation (10.130) differ from one. Their product, however, is still equal to the zero-order term of equation (10.130), i.e., to one. Consequently, the absolute value of the first root of equation (10.130) will be greater than one, while that of the second root will be less than one. Let us denote \( |q_1(\lambda)| < 1 \) and \( |q_2(\lambda)| > 1 \). The general solution of equation (10.129) has the form:

\[ u_m = c_1 q_1^m + c_2 q_2^m, \]

where \( c_1 \) and \( c_2 \) are arbitrary constants. Accordingly, the general solution that satisfies additional constraint (10.125), i.e., that decays as \( m \to +\infty \), is written as

\[ u_m = c_1 q_1^m, \quad |q_1| = |q_1(\lambda)| < 1, \]

and the general solution that satisfies additional constraint (10.127), i.e., that decays as \( m \to -\infty \), is given by

\[ u_m = c_2 q_2^m, \quad |q_2| = |q_2(\lambda)| > 1. \]
To calculate the eigenvalues of problem (10.124), one needs to substitute \( u_m = c_1 q_1^m \) into the left boundary condition \( I_1 u_0 = 0 \) and find those \( q_1 \) and \( \lambda \), for which it is satisfied. The corresponding \( \lambda \) will yield the spectrum of problem (10.124). If, for example, \( I_1 u_0 \equiv u_0 = 0 \), then equality \( c_1 q_1^0 = 0 \) will not hold for any \( c_1 \neq 0 \), so that problem (10.124) has no eigenvalues. (Recall, the eigenfunction must be nontrivial.) If \( I_1 u_0 \equiv u_1 - u_0 = 0 \), then condition \( c_1 (q_1 - q_1^0) = c_1 (q_1 - 1) = 0 \) again yields \( c_1 = 0 \) because \( q_1 \neq 1 \), so that there are no eigenvalues either. If, however, \( I_1 u_0 \equiv 2u_1 - u_0 = 0 \), then condition \( c_1 (2q_1 - q_1^0) = c_1 (2q_1 - 1) = 0 \) is satisfied for \( c_1 \neq 0 \) and \( q_1 = 1/2 < 1 \). Substituting \( q_1 = 1/2 \) into the characteristic equation (10.130) we find that

\[
\lambda = 1 + r \left( q_1 - 2 + \frac{1}{q_1} \right) = 1 + \frac{r}{2}.
\]

This is the only eigenvalue of problem (10.124). It does not belong to the unit disk on the complex plane, and therefore the necessary stability condition is violated.

The eigenvalues of the auxiliary problem (10.126) are calculated analogously. They are found from the equation \( I_2 u_M = 0 \) when

\[
u_m = c_2 q_2^m, \quad |q_2| = |q_2(\lambda)| > 1, \quad m = M, M - 1, M - 2, \ldots .\]

For stability, it is necessary that they all belong to the unit disk on the complex plane.

We can now provide more specific comments following Remark 10.1. When boundary condition \( I_1 u_0 \equiv 2u_1 - u_0 = 0 \) is employed in problem (10.124) then the solution that satisfies condition (10.125) is found in the form \( u_m = \lambda^p q_m^m \), where \( q_1 = 1/2 \) and \( \lambda = 1 + r/2 > 1 \). This solution is only defined for \( m \geq 0 \). If, however, we were to extend it to the region \( m < 0 \), we would have obtained an unbounded function: \( u_m^p \rightarrow \infty \) as \( m \rightarrow -\infty \). In other words, the function \( u_m^p = \lambda^p q_m^m \) cannot be used in the framework of the standard von Neumann analysis of problem (10.109).

This consideration leads to a very simple explanation of the mechanism of instability. The introduction of a boundary condition merely expands the pool of candidate functions, on which the instability may develop. In the pure von Neumann case, with no boundary conditions, we have only been monitoring the behavior of the harmonics \( e^{i\omega_m} \) that are bounded on the entire grid \( m = 0, \pm 1, \pm 2, \ldots \). With the boundary conditions present, we may need to include additional functions that are bounded on the semi-infinite grid, but unbounded if extended to the entire grid. These functions do not belong to the von Neumann category. If any of them brings along an unstable eigenvalue \( |\lambda| > 1 \), such as \( \lambda = 1 + r/2 \), then the overall scheme becomes unstable as well. We therefore re-iterate that if a scheme that approximates some Cauchy problem is supplemented by boundary conditions and thus transformed into an initial boundary value problem, then its stability will not be improved. In other words, if the Cauchy problem was stable, then the initial boundary value problem may either remain stable or become unstable. If, however, the Cauchy problem was unstable, then the initial boundary value problem will not become stable.

Our next example will be the familiar first-order upwind scheme, but built on a finite grid: \( x_m = mh, m = 0, 1, 2, \ldots , M, \; Mh = 1 \), rather than on the infinite grid
Finite-Difference Schemes for Partial Differential Equations

\[ m = 0, \pm 1, \pm 2, \ldots \text{[cf. formula (10.115)]}: \]
\[ \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \]
\[ m = 0, 1, 2, \ldots, M - 1, \quad p = 0, 1, 2, \ldots, [T/\tau] - 1, \quad (10.132) \]
\[ u_m^0 = \psi(x_m), \quad u_M^{p+1} = 0. \]

Scheme (10.132) approximates the following first-order hyperbolic initial boundary value problem [cf. formula (10.116)]:
\[ \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad 0 \leq x \leq 1, \quad 0 < t \leq T, \]
\[ u(x, 0) = \psi(x), \quad u(1, t) = 0, \]
on the interval \( 0 \leq x \leq 1 \). To investigate stability of scheme (10.132), we will employ the Babenko-Gelfand criterion. In other words, we will need to analyze three auxiliary problems: A problem with no lateral boundaries:
\[ \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \]
\[ m = 0, \pm 1, \pm 2, \ldots, \quad (10.133) \]
a problem with only the left boundary:
\[ \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \]
\[ m = 0, 1, 2, \ldots, \quad (10.134) \]
and a problem with only the right boundary:
\[ \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \]
\[ m = M - 1, M - 2, \ldots, 1, 0, -1, \ldots, \quad (10.135) \]
\[ u_M^{p+1} = 0. \]

Note that we do not set any boundary condition at the left boundary in problem (10.134) as we did not have any in the original problem (10.132) either.

We will need to find spectra of the three transition operators from \( u^p \) to \( u^{p+1} \) that correspond to the three auxiliary problems (10.133), (10.134), and (10.135), respectively, and determine under what conditions all the eigenvalues belong to the unit disk \( |\lambda| \leq 1 \) on the complex plane.

Substituting a solution of the type:
\[ u_m^p = \lambda^p u_m \]
into the finite-difference equation:
\[ u_m^{p+1} = (1 - r)u_m^p + ru_{m+1}^p, \quad r = \tau/h, \]
that corresponds to all three problems (10.133), (10.134), and (10.135), we obtain
the following first-order ordinary difference equation for the eigenfunction \( \{u_m\} \):
\[
(\lambda - 1 + r)u_m - ru_{m+1} = 0.
\] (10.136)
Its characteristic equation:
\[
\lambda - 1 + r - rq = 0
\] (10.137)
yields the relation between \( \lambda \) and \( q \), so that the general solution of equation (10.136) can be written as
\[
u_m = cq^m = c \left( \frac{\lambda - 1 + r}{r} \right)^m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad c = \text{const}.
\]
When \( |q| = 1 \), i.e., when \( q = e^{i\alpha}, 0 \leq \alpha < 2\pi \), we have:
\[
\lambda = 1 - r + re^{i\alpha}.
\]
The point \( \lambda = \lambda(q) \) sweeps the circle of radius \( r \) centered at the point \((1-r, 0)\) on the complex plane. This circle gives the spectrum, i.e., the full set of eigenvalues, of the first auxiliary problem (10.133), see Figure 10.11(a). It is clearly the same spectrum as we have discussed in Section 10.3.2, see formula (10.77) and Figure 10.5(a).

As far as the second auxiliary problem (10.134), we need to look for its non-trivial solutions that would decrease as \( m \rightarrow +\infty \), see formula (10.125). Such a solution, \( u_m = c\lambda_p^m q^m \), obviously exists for any \( q: |q| < 1 \). The corresponding eigenvalues \( \lambda = \lambda(q) = 1 - r + rq \) fill the interior of the disk bounded by the circle \( \lambda = 1 - r + re^{i\alpha} \) on the complex plane, see Figure 10.11(b).

Solutions of the third auxiliary problem (10.135) that would satisfy (10.127), i.e., that would decay as \( m \rightarrow -\infty \), must obviously have the form: \( u_m = c\lambda_p^m q^m \), where \( |q| > 1 \) and the relation between \( \lambda \) and \( q \) is, again, given by formula (10.137). The homogeneous boundary condition \( u_{M+1}^p = 0 \) of (10.135) implies that a non-trivial
eigenfunction $u_m = cq^m$ may only exist when $\lambda = \lambda(q) = 0$, i.e., when $q = (r - 1)/r$. The quantity $q$ given by this expression may have its absolute value greater than one if either of the two inequalities holds:

$$\frac{r - 1}{r} > 1 \quad \text{or} \quad \frac{r - 1}{r} < -1.$$  

The first inequality has no solutions. The solution to the second inequality is $r < 1/2$. Consequently, when $r < 1/2$, problem (10.135) has the eigenvalue $\lambda = 0$, see Figure 10.11(c).

In Figure 10.12, we are schematically showing the combined sets of all eigenvalues, i.e., combined spectra, for problems (10.133), (10.134), and (10.135) for the three different cases: $r < 1/2$, $1/2 < r < 1$, and $r > 1$.

FIGURE 10.12: Combined spectra of auxiliary problems for scheme (10.132).

It is clear that the combined eigenvalues of all three auxiliary problems may only belong to the unit disk $|\lambda| \leq 1$ on the complex plane if $r \leq 1$. Therefore, condition $r \leq 1$ is necessary for stability of the difference initial boundary value problem (10.132).

Compared to the von Neumann stability condition of Section 10.3, the key distinction of the Babenko-Gelfand criterion is that it takes into account the boundary conditions for unsteady finite-difference equations on finite intervals. This criterion can also be generalized to systems of such equations. In this case, a scheme that may look perfectly natural and “benign” at a first glance, and that may, in particular, satisfy the von Neumann stability criterion, could still be unstable because of a poor approximation of the boundary conditions. Consequently, it is important to be able to build schemes that are free of this shortcoming.

In [GR87], the spectral criterion of Babenko and Gelfand is discussed from a more general standpoint, using a special new concept of the spectrum of a family of operators introduced by Godunov and Ryaben’kii. In this framework, one can rigorously prove that the Babenko-Gelfand criterion is necessary for stability, and also that when it holds, stability cannot be disrupted too severely. In the next section,
we reproduce key elements of this analysis, while referring the reader to [GR87, Chapters 12 & 13] and [RM67, § 6.6 & 6.7] for further detail.

10.5.2 Spectra of Families of Operators

In this section, we briefly describe a rigorous approach, due to Godunov and Ryaben’kii, to studying stability of evolution-type finite-difference schemes on finite intervals. In other words, we study stability of the discrete approximations to initial boundary value problems for hyperbolic and parabolic partial differential equations. This material is more advanced, and can be skipped during the first reading.

As we have seen previously, for evolution finite-difference schemes the discrete solution \( u^{(h)} = \{ u^p_m \} \), which is defined on a two-dimensional space-time grid:

\[(x_m, t_p) \equiv (mh, pt), \quad m = 0, 1, \ldots, M, \quad p = 0, 1, \ldots, [T/\tau],\]

gets naturally split or “stratified” into a collection of one-dimensional grid functions \( \{ u^p \} \) defined for individual time layers \( t_p, p = 0, 1, \ldots, [T/\tau] \). For example, the first-order upwind scheme:

\[
\frac{u^p_{m+1} - u^p_m}{\tau} - \frac{(1 - r)u^p_m + ru^p_{m+1}}{h} = \phi_{x}^p, \quad m = 0, 1, 2, \ldots, M - 1, \quad p = 0, 1, 2, \ldots, [T/\tau] - 1,
\]

(10.138)

gets defined on a two-dimensional space-time grid:

\[
\begin{align*}
\frac{u^p_{m+1} - u^p_m}{\tau} - \frac{(1 - r)u^p_m + ru^p_{m+1}}{h} &= \phi_{x}^p, \\
&\quad m = 0, 1, 2, \ldots, M - 1, \quad p = 0, 1, 2, \ldots, [T/\tau] - 1, \\
u^0_m &= \psi_m, \quad u^1_M = \chi^{p+1},
\end{align*}
\]

for the initial boundary value problem:

\[
\begin{align*}
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} &= \phi(x, t), \quad 0 \leq x \leq 1, \quad 0 < t \leq T, \\
u(x, 0) &= \psi(x), \quad u(1, t) = \chi(t),
\end{align*}
\]

can be written as:

\[
\begin{align*}
u^{p+1}_m &= [(1 - r)u^p_m + ru^p_{m+1}] + \tau\phi^p_m, \quad m = 0, 1, \ldots, M - 1, \\
u^1_M &= \chi^{p+1}, \quad u^0_m = \psi_m, \quad m = 0, 1, \ldots, M,
\end{align*}
\]

(10.139)

where \( r = \tau/h \). Form (10.139) suggests that the marching procedure for scheme (10.138) can be interpreted as consecutive computation of the grid functions:

\[
u^0, u^1, \ldots, u^p, \ldots, [T/\tau],
\]

defined on identical one-dimensional grids \( m = 0, 1, \ldots, M \) that can all be identified with one and the same grid. Accordingly, the functions \( u^p, p = 0, 1, \ldots, [T/\tau] \), can be considered elements of the linear space \( U^p \) of functions \( u = \{ u_0, u_1, \ldots, u_M \} \) defined on the grid \( m = 0, 1, \ldots, M \). We will equip this linear space with the norm, e.g.,

\[
\| u \|_{L^h}^2 = \max_{0 \leq m \leq M} |u_m| \quad \text{or} \quad \| u \|_{L^h}^2 = \left[ \frac{1}{h} \sum_{m=0}^{M} |u_m|^2 \right]^{1/2}.
\]
We also recall that in the definitions of stability (Section 10.1.3) and convergence (Section 10.1.1) we employ the norm \( \| u(h) \|_{U_h} \) of the finite-difference solution \( u(h) \) on the entire two-dimensional grid. Hereafter, we will only be using norms that explicitly take into account the layered structure of the solution, namely, those that satisfy the equality:

\[
\| u(h) \|_{U_h} = \max_{0 \leq \rho \leq [T/\tau]} \| u^\rho \|_{U_h^\rho}.
\]

Having introduced the linear normed space \( U'_h \), we can represent any evolution scheme, in particular, scheme (10.138), in the canonical form:

\[
\begin{align*}
  u^{p+1} &= R_h u^p + \tau p^p, \\
  u^0 & \text{ is given.}
\end{align*}
\]

In formula (10.140), \( R_h : U'_h \to U'_h \) is the transition operator between the consecutive time levels, and \( p^p \in U'_h \). If we denote \( v^{p+1} = R_h u^p \), then formula (10.139) yields:

\[
v^{p+1}_m = (1 - r) u^p_m + r u^p_{m+1}, \quad m = 0, 1, \ldots, M - 1.
\]

As far as the last component \( m = M \) of the vector \( v^{p+1} \), a certain flexibility exists in the definition of the operator \( R_h \) for scheme (10.138). For example, we can set:

\[
v^{p+1}_M = u^p_M,
\]

which would also imply:

\[
\rho^p_m = q^p_m, \quad m = 0, 1, \ldots, M - 1, \quad \text{ and } \quad \rho^p_M = \frac{\chi^{p+1} - \chi^p}{\tau},
\]

in order to satisfy the first equality of (10.140).

In general, the canonical form (10.140) for a given evolution scheme is not unique. For scheme (10.138), we could have chosen \( v^{p+1}_M = 0 \) instead of \( v^{p+1}_M = u^p_M \) in formula (10.141b), which would have also implied \( \rho^p_M = \frac{\chi^{p+1}}{\tau} \) in formula (10.141c). However, when building the operator \( R_h \), we need to make sure that the following rather natural conditions hold that require a correlation between the norms in \( U'_h \) and \( F_h \):

\[
\begin{align*}
  \| p^p \|_{U'_h} & \leq K_1 \| f^{(h)} \|_{F_h}, \quad p = 0, 1, \ldots, [T/\tau], \\
  \| u^0 \|_{U'_h} & \leq K_2 \| f^{(h)} \|_{F_h}.
\end{align*}
\]

The constants \( K_1 \) and \( K_2 \) in inequalities (10.142) do not depend on \( h \) or on \( f^{(h)} \). For scheme (10.138), if we define the norm in the space \( F_h \) as:

\[
\| f^{(h)} \|_{F_h} = \max_{m, p} \left[ q^p_m + \max_m | \psi_m | + \max_p \left| \frac{\chi^{p+1} - \chi^p}{\tau} \right| \right]
\]

and the norms of \( p^p \) and \( u_0 \) as \( \| p^p \|_{U'_h} = \max_m | p^p_m | \) and \( \| u^0 \|_{U'_h} = \max_m | u^0_m | \), respectively, then conditions (10.142) obviously hold for the operator \( R_h \) and the source term \( p^p \) defined by formulae (10.141a)-(10.141c).
Let us now take an arbitrary $\hat{u}^0 \in U'_h$ and obtain $\hat{u}^1, \hat{u}^2, \ldots, \hat{u}^{[T/\tau]} \in U'_h$ using the recurrence formula $u^{p+1} = R_h u^p$. Denote $\hat{u}^{(h)} = \{ \hat{u}^p \}_{p=0}^{[T/\tau]}$ and evaluate $f^{(h)} \overset{\text{def}}{=} L_h \hat{u}^{(h)}$.

Along with conditions (10.142), we will also require that

$$\| f^{(h)} \| \leq K_3 \| \hat{u}^0 \|_{U'_h},$$

(10.143)

where the constant $K_3$ does not depend on $\hat{u}^0 \in U'_h$ or on $h$.

In practice, inequalities (10.142) and (10.143) prove relatively non-restrictive. These inequalities allow one to establish the following important theorem that provides a necessary and sufficient condition for stability in terms of the uniform boundedness of the powers of $R_h$ with respect to the grid size $h$.

**THEOREM 10.6**

Assume that when reducing a given evolution scheme to the canonical form (10.140) the additional conditions (10.142) are satisfied. Then, for stability of the scheme in the sense of Definition 10.2 it is sufficient that

$$\| R_h^p \| \leq K, \quad p = 0, 1, \ldots, [T/\tau],$$

(10.144)

where the constant $K$ in formula (10.144) does not depend on $h$. If the third additional condition (10.143) is met as well, then estimates (10.144) are also necessary for stability.

Theorem 10.6 is proven in [GR87, § 39].

For scheme (10.138), estimates (10.144) can be established directly, provided that $r \leq 1$. Indeed, according to formula (10.141a), we have for $m = 0, 1, \ldots, M - 1$:

$$|v_m^{p+1}| = |(1 - r)u_m^p + ru_{m+1}^p| \leq (1 - r + r) \max_m |u_m^p| = \| u^p \|_{U'_h},$$

and according to formula (10.141b), we have for $m = M$:

$$|v_M^{p+1}| = |u_M^p| \leq \max_m |u_m^p| = \| u^p \|_{U'_h}.$$

Consequently,

$$\| R_h u^p \|_{U'_h} = \| v_M^{p+1} \|_{U'_h} = \max_m |v_m^{p+1}| \leq \max_m |u_m^p| = \| u^p \|_{U'_h},$$

which means that $\| R_h \| \leq 1$. Therefore, $\| R_h^p \| \leq \| R_h \|^p \leq 1$, and according to Theorem 10.6, scheme (10.138) is stable.

**REMARK 10.2** Previously, we have seen that the notion of stability for a finite-difference scheme can be reformulated as boundedness of the powers

---

The first condition of (10.142) can, in fact, be further relaxed, see [GR87, § 40].
for families of matrices. Namely, in Section 10.3.6 we have discussed stability of finite-difference Cauchy problems for systems of equations with constant coefficients (as opposed to scalar equations). We saw that the finite-difference stability was equivalent to stability of the families of amplification matrices. In fact, the latter is defined as boundedness of their powers, and the Kreiss matrix theorem (Theorem 10.4) provides necessary and sufficient conditions for this property to hold.

Transition operators $R_h$ can also be interpreted as matrices that operate on vectors from the space $U_h'$. In this perspective, inequality (10.144) implies uniform boundedness of all powers or stability of this family of operators (matrices). There is, however, a fundamental difference between the considerations of this section and those of Section 10.3.6. The amplification matrices that appear in the context of the Kreiss matrix theorem (Theorem 10.4) are parameterized by the frequency $\alpha$ and possibly the grid size $h$. Yet the dimension of all these matrices remains fixed and equal to the dimension of the original system, regardless of the grid size. In contradistinction to that, the dimension of the matrices $R_h$ is inversely proportional to the grid size $h$, i.e., it grows with no bound as $h \to 0$. Therefore, estimate (10.144) actually goes beyond the notion of stability for families of matrices of a fixed dimension (Section 10.3.6), as it implies stability (uniform bound on powers) for a family of matrices of increasing dimension.

As condition (10.144) is equivalent to stability according to Theorem 10.6, then to investigate stability we need to see whether inequalities (10.144) hold. Let $\lambda_h$ be an eigenvalue of the operator $R_h$, and let $\nu^{(b)}$ be the corresponding eigenvector so that $R_h \nu^{(b)} = \lambda_h \nu^{(b)}$. Then,

$$\|R_h^p\| \nu^{(b)} \geq \|R_h^p \nu^{(b)}\| = |\lambda_h|^p \|\nu^{(b)}\|$$

and consequently $\|R_h^p\| \geq |\lambda_h|^p$. Since $\lambda_h$ is an arbitrary eigenvalue, we have:

$$\|R_h^p\| \geq \max |\lambda_h|^p, \quad p = 0, 1, \ldots \lfloor T/\tau \rfloor,$$

where $\max |\lambda_h|$ is the largest eigenvalue of $R_h$ by modulus. Hence, for the estimate (10.144) to hold, it is necessary that all eigenvalues $\lambda$ of the transition operator $R_h$ belong to the following disk on the complex plane:

$$|\lambda| \leq 1 + c_1 \tau,$$  \hspace{1cm} (10.145)

where the constant $c_1$ does not depend on the grid size $h$ (or $\tau$). It means that inequality (10.145) must hold with one and the same constant $c_1$ for any given transition operator from the family $\{R_h\}$ parameterized by $h$.

Inequality (10.145) is known as the spectral necessary condition for the uniform boundedness of the powers $\|R_h^p\|$. It is called spectral because as long as the operators $R_h$ can be identified with matrices of finite dimension, the eigenvalues of those matrices yield the spectra of the operators. This spectral condition is also closely
related to the von Neumann spectral stability criterion for finite-difference Cauchy problems on infinite grids that we have studied in Section 10.3, see formula (10.81).

Indeed, instead of the finite-difference initial boundary value problem (10.138), consider a Cauchy problem on the grid that is infinite in space:

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = q_m^p,$$

$$m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, 2, \ldots, [T/\tau] - 1.$$ (10.146)

The von Neumann analysis of Section 10.3.2 has shown that for stability it is necessary that \(r = \tau/h \leq 1\). To apply the spectral criterion (10.145), we first reduce scheme (10.146) to the canonical form (10.140). The operator \(R_h : U_h' \rightarrow U_h', R_h u^p = v^{p+1}\), and the source term \(\rho^p\) are then given by [cf. formulae (10.141)]:

$$v_m^{p+1} = (1-r)u_m^p + ru_{m+1}^p, \quad \rho_m^p = q_m^p,$$

$$m = 0, \pm 1, \pm 2, \ldots.$$ The space \(U_h'\) contains infinite sequences \(u = \{\ldots, u_{-m}, \ldots, u_1, u_0, u_1, \ldots, u_m, \ldots\}\).

We can supplement this space with the \(C\) norm: \(||u|| = \sup_m |u_m|\). The grid functions \(u = \{u_m\} = \{e^{i\alpha m}\}\) then belong to the space \(U_h'\) for all \(\alpha \in [0, 2\pi)\) and provide eigenfunctions of the transition operator:

$$R_h u = (1-r)e^{i\alpha m} + r e^{i\alpha (m+1)} = [(1-r) + re^{i\alpha}]e^{i\alpha m} = \lambda(\alpha) u,$$

where the eigenvalues are given by:

$$\lambda(\alpha) = (1-r) + re^{i\alpha}.$$ (10.147)

According to the spectral condition of stability (10.145), all eigenvalues must satisfy the inequality: \(|\lambda(\alpha)| \leq 1 + c_1 \tau\), which is the same as the von Neumann condition (10.78). As the eigenvalues (10.147) do not explicitly depend on the grid size, the spectral condition (10.145) reduces here to \(|\lambda(\alpha)| \leq 1\), cf. formula (10.79).

Let us also recall that as shown in Section 10.3.5, the von Neumann condition is not only necessary, but also sufficient for the \(l_2\) stability of the two-layer (one-step) scalar finite-difference Cauchy problems, see formula (10.95). If, however, the space \(U_h'\) is equipped with the \(l_2\) norm: \(||u|| = \left(\sum_{m=-\infty}^{\infty} |u_m|^2 \right)^{1/2}\) (as opposed to the \(C\) norm), then the functions \(\{e^{i\alpha m}\}\) no longer belong to this space, and therefore may no longer be the eigenfunctions of \(R_h\). Nonetheless, we can show that the points \(\tilde{\lambda}(\alpha)\) of (10.147) still belong to the spectrum of the operator \(R_h\), provided that the latter is defined as traditionally done in functional analysis.\(^9\) Consequently, if we interpret \(\lambda\) in formula (10.145) as all points of the spectrum rather than just the eigenvalues of

\(^9\)In general, the points \(\lambda = \lambda(\alpha, h)\) given by Definition 10.3 on page 201 will be a part of the spectrum in the sense of its classical definition, see Definition 10.7 on page 241.
the operator $R_h$, then the spectral condition (10.145) also becomes sufficient for the $l_2$ stability of the Cauchy problems (10.95) on an infinite grid $m = 0, \pm 1, \pm 2, \ldots$.

Returning now to the difference equations on finite intervals and grids (as opposed to Cauchy problems), we first notice that one can most easily verify estimates (10.144) when the matrices of all operators $R_h$ happen to be normal: $R_h R_h^* = R_h^* R_h$. Indeed, in this case there is an orthonormal basis in the space $U_h'$ composed of the eigenvectors of the matrix $R_h$, see, e.g., [HJ85, Chapter 2]. Using expansion with respect to this basis, one can show that the spectral condition (10.145) is necessary and sufficient for the $l_2$ stability of an evolution scheme with normal operators $R_h$ on a finite interval. More precisely, the following theorem holds.

**THEOREM 10.7**

Let the operators $R_h$ in the canonical form (10.140) be normal, and let them all be uniformly bounded with respect to the grid: $\|R_h\| \leq c_2$, where $c_2$ does not depend on $h$. Let also all norms be chosen in the sense of $l_2$. Then, for the estimates (10.144) to hold, it is necessary and sufficient that the inequalities be satisfied:

$$\max_n |\lambda_n| \leq 1 + c_1 \tau, \quad c_1 = \text{const},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_N$ are eigenvalues of the matrix $R_h$ and the constant $c_1$ in formula (10.148) does not depend on $h$.

One implication of Theorem 10.7, the necessity, coincides with the previous necessary spectral condition for stability that we have justified on page 237. The other implication, the sufficiency, is to be proven in Exercise 6 of this section. A full proof of Theorem 10.7 can be found, e.g., in [GR87, §41].

Unfortunately, in many practical situations the operators (matrices) $R_h$ in the canonical form (10.140) are not normal. Then, the spectral condition (10.145) still remains necessary for stability. Moreover, we have just seen that in the special case of two-layer scalar constant-coefficient Cauchy problems it is also sufficient for stability and that sufficiency takes place regardless of whether or not $R_h$ has a full system of orthonormal eigenfunctions. However, for general finite-difference problems on finite intervals the spectral condition (10.145) becomes pretty far detached from sufficiency and provides no adequate criterion for uniform boundedness of $\|R_h^p\|$.

For instance, the matrix of the transition operator $R_h$ defined by formulae (10.141a) and (10.141b) is given by:

$$R_h = \begin{bmatrix}
1 - r & r & 0 & \cdots & 0 & 0 \\
0 & 1 - r & r & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 - r & r \\
0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}.$$  \hspace{1cm} (10.149)

Its spectrum consists of the eigenvalues $\lambda = 1$ and $\lambda = 1 - r$ and as such, does not depend on $h$ (or on $\tau$). Consequently, for any $h > 0$ the spectrum of the operator $R_h$
Theoretical Introduction to Numerical Analysis

consists of only these two numbers: \( \lambda = 1 \) and \( \lambda = 1 - r \). This spectrum belongs to the unit disk \( |\lambda| \leq 1 \) when \( 0 \leq r \leq 2 \). However, for \( 1 < r \leq 2 \), scheme (10.138) violates the Courant, Friedrichs, and Lewy condition necessary for stability, and hence, there may be no stability \( \| R_h^p \| \leq K \) for any reasonable choice of norms.

Thus, we have seen that the spectral condition (10.145) that employs the eigenvalues of the operators \( R_h \) and that is necessary for the uniform boundedness \( \| R_h^p \| \leq K \) appears too rough in the case of non-normal matrices. For example, it fails to detect the instability of scheme (10.138) for \( 1 < r \leq 2 \).

To refine the spectral condition we will introduce a new concept. Assume, as before, that the operator \( R_h \) is defined on a normed linear space \( U_h' \). We will denote by \( \{ R_h \} \) the entire family of operators \( R_h \) for all legitimate values of the parameter \( h \) that characterizes the grid.\(^{10}\)

**DEFINITION 10.6** A complex number \( \lambda \) is said to belong to the spectrum of the family of operators \( \{ R_h \} \) if for any \( h_0 > 0 \) and \( \epsilon > 0 \) one can always find such a value of \( h, h < h_0 \), that the inequality

\[
\| R_h u - \lambda u \|_{U_h'} < \epsilon \| u \|_{U_h'}
\]

will have a solution \( u, u \in U_h' \).

The set of all such \( \lambda \) will be called the spectrum of the family of operators \( \{ R_h \} \).

The following theorem employs the concept of the spectrum of a family of operators from Definition 10.6 and provides a key necessary condition for stability.

**THEOREM 10.8 (Godunov-Ryaben’kii)**

If even one point \( \lambda_0 \) of the spectrum of the family of operators \( \{ R_h \} \) lies outside the unit disk on the complex plane, i.e., \( |\lambda_0| > 1 \), then there is no common constant \( K \) such that the inequality

\[
\| R_h^p \| \leq K
\]

will hold for all \( h > 0 \) and all integer values of \( p \) from 0 till some \( p = p_0(h) \), where \( p_0(h) \rightarrow \infty \) as \( h \rightarrow 0 \).

**PROOF** Let us first assume that no such numbers \( h_0 > 0 \) and \( c > 0 \) exist that for all \( h < h_0 \) the following estimate holds:

\[
\| R_h \| \leq c.
\]

(10.150)

This assumption means that there is no uniform bound on the operators \( R_h \) themselves. As such, there may be no bound on the powers \( R_h^p \) either.

\(^{10}\)By the very nature of finite-difference schemes, \( h \) may assume arbitrarily small positive values.
Consequently, we only need to consider the case when there are such \( h_0 > 0 \) and \( c > 0 \) that for all \( h < h_0 \) inequality (10.150) is satisfied.

Let \( |\lambda_0| = 1 + \delta \), where \( \lambda_0 \) is the point of the spectrum for which \( |\lambda_0| > 1 \).

Take an arbitrary \( K > 0 \) and choose \( p \) and \( \varepsilon \) so that:

\[
(1 + \delta)^p > 2K,
\]
\[
1 - (1 + c + c^2 + \ldots + c^{p-1})\varepsilon > \frac{1}{2}.
\]

According to Definition 10.6, one can find arbitrarily small \( h \), for which there is a vector \( u \in U'_h \) that solves the inequality:

\[
\|R_hu - \lambda_0 u\|_{U'_h} < \varepsilon \|u\|_{U'_h}.
\]

Let \( u \) be the solution, and denote:

\[
R_hu = \lambda_0 u + z.
\]

It is clear that \( \|z\| < \varepsilon \|u\| \). Moreover, it is easy to see that

\[
R_h^p u = \lambda_0^p u + (\lambda_0^{p-1}z + \lambda_0^{p-2}R_hz + \ldots + R_h^{p-1}z).
\]

As \( |\lambda_0| > 1 \), we have:

\[
|\lambda_0^{p-1}z + \lambda_0^{p-2}R_hz + \ldots + R_h^{p-1}z| < |\lambda_0|^p(1 + \|R_h\| + \|R_h^2\| + \ldots + \|R_h^{p-1}\|)\varepsilon \|u\|,
\]

and consequently,

\[
\|R_h^p u\| > |\lambda_0|^p[1 - \varepsilon(1 + c + c^2 + \ldots + c^{p-1})]|\|u\|,
\]

\[
> (1 + \delta)^p \frac{1}{2} \|u\| > 2K \frac{1}{2} \|u\| = K \|u\|.
\]

In doing so, the value of \( h \) can always be taken sufficiently small so that to ensure \( p < p_0(h) \).

Since the value of \( K \) has been chosen arbitrarily, we have essentially proven that for the estimate \( \|R_h^p\| < K \) to hold, it is necessary that all points of the spectrum of the family \( \{R_h\} \) belong to the unit disk \( |\lambda| \leq 1 \) on the complex plane.

It will be very instrumental to discuss the concept of the spectrum of a family of operators and the result of Theorem 10.8. First of all, we notice an analogy between Definition 10.6 and the following definition of the spectrum of an operator \( R \) given in functional analysis. Suppose that \( R = R_h \) for some fixed \( h \).

**Definition 10.7** A complex number \( \lambda \) is said to belong to the spectrum of the operator \( R_h : U'_h \rightarrow U'_h \) if for any \( \varepsilon > 0 \) the inequality

\[
\|R_hu - \lambda u\|_{U'_h} < \varepsilon \|u\|_{U'_h}
\]
has a solution $u$, $u \in U'_h$.

The set of all such $\lambda$ is called the spectrum of the operator $R_h$.

At first glance, comparing Definitions 10.6 10.7 may lead one to think that the spectrum of the family of operators $\{R_h\}$ consists of all those and only those points on the complex plane that are obtained by passing to the limit $h \rightarrow 0$ from the points of the spectrum of $R_h$, when $h$ approaches zero along all possible sub-sequences. However, this assumption is, generally speaking, not correct.

Consider, for example, the operator $R_h : U'_h \rightarrow U'_h$ defined by formulae (10.141a) and (10.141b). It is described by the matrix (10.149) and operates in the $M + 1$-dimensional linear space $U'_h$, where $M = 1/h$. Recall that the spectrum of a matrix consists of its eigenvalues, and the eigenvalues of the matrix $R_h$ of (10.149) are $\lambda = 1$ and $\lambda = 1 - r$. These eigenvalues do not depend on $h$ (or on $\tau$) and consequently, the spectrum of the operator $R_h$ consists of only two points, $\lambda = 1$ and $\lambda = 1 - r$, for any $h > 0$. As, however, we are going to see (pages 246-251), the spectrum of the family of operators $\{R_h\}$ contains not only these two points, but also all points of the disk $|\lambda - 1 + r| \leq r$ of radius $r$ centered at the point $(1 - r, 0)$ on the complex plane, see Figure 10.12 on page 233. When $r \leq 1$, the spectrum of the family of operators $\{R_h\}$ belongs to the unit disk $|\lambda| \leq 1$, see Figures 10.12(a) and 10.12(b). However, when $r > 1$, this necessary spectral condition of stability does not hold, see Figure 10.12(c), and the inequality $\|R^p_h\| \leq K$ can not be satisfied uniformly with respect to $h$.

Before we accurately compute the spectrum of the family of operators $\{R_h\}$ given by formula (10.149), let us qualitatively analyze the behavior of the powers $\|R^p_h\|$ for $r > 1$ and also show that the necessary stability criterion given by Theorem 10.8 is, in fact, rather close to sufficient.

We first notice that for any $h > 0$ there is only one eigenvalue of the matrix $R_h$ that has unit modulus: $\lambda = 1$, and that by means of a similarity transformation $S_h^{-1}R_hS_h$, where

$$S_h = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \quad \text{and} \quad S_h^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

this matrix is reduced to the block-diagonal form [cf. formula (10.149)]:

$$S_h^{-1}R_hS_h = \begin{bmatrix} 1 - r & r & 0 & \cdots & 0 & 0 \\ 0 & 1 - r & r & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = B_h.$$
words, the limiting matrix of $B^p_h$ for the powers $p$ approaching infinity has only one non-zero entry equal to one at the lower right corner. Consequently

$$\lim_{p \to \infty} R^p_h = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix},$$

and as such, $\lim_{p \to \infty} \|R^p_h\| = 1$. We can therefore see that regardless of the value of $h$, the norms of the powers of the transition operator $R_h$ approach one and the same finite limit. In other words, we can write $\lim_{p \to \infty} \|R^p_h\| = 1$, and this “benign” asymptotic behavior of $\|R^p_h\|$ for large $p\tau$ is indeed determined by the eigenvalues $\lambda = 1 - r$ and $\lambda = 1$ that belong to the unit disk.

The fact that the spectrum of the family of operators $\{R_h\}$ does not belong to the unit disk for $r > 1$ manifests itself in the behavior of $\|R^p_h\|$ for $h \to 0$ and for moderate (not so large) values of $p\tau$. The maximum value of $\{R_h\}$ on the interval $0 < p\tau < T$, where $T$ is an arbitrary positive constant, will rapidly grow as $h$ decreases, see Figure 10.13. This is what implies the instability, whereas the behavior of $\|R^p_h\|$ as $p\tau \to \infty$, which is related to the spectrum of each individual operator $R_h$, is not important for the stability study.

Let us also emphasize that even though from a technical point of view Theorem 10.8 only provides a necessary condition for stability, this condition, is, in fact, not so distant from sufficient. More precisely, the following theorem holds.

**THEOREM 10.9**

Let the operators $R_h$ be defined on a linear normed space $U'_h$ for each $h > 0$, and assume that they are uniformly bounded with respect to $h$:

$$\|R_h\| \leq c.$$  \hspace{1cm} (10.151)

Let also the spectrum of the family of operators $\{R_h\}$ belong to the unit disk on the complex plane: $|\lambda| \leq 1$. 

Then for any \( \eta > 0 \), the norms of the powers of operators \( \mathbf{R}_h \) satisfy the estimate:
\[
\| \mathbf{R}_h^k \| \leq A(\eta)(1 + \eta)^k, \quad (10.152)
\]
where \( A = A(\eta) \) may depend on \( \eta \), but does not depend on the grid size \( h \).

Theorem 10.9 means that having the spectrum of the family of operators \( \{\mathbf{R}_h\} \) lie inside the unit disk is not only necessary for stability, but it also guarantees us from a catastrophic instability. Indeed, if the conditions of Theorem 10.9 hold, then the quantity \( \max_{1 \leq p \leq |r/T|} \| \mathbf{R}_h^p \| \) either remains bounded as \( h \to 0 \) or increases, but slower than any exponential function, i.e., slower than any \((1 + \eta)^{|r/T|}\), where \( \eta > 0 \) may be arbitrarily small.

**PROOF** Let us first show that if the spectrum of the family of operators \( \{\mathbf{R}_h\} \) belongs to the disk \( |\lambda| \leq \rho \), then for any given \( \tilde{\lambda} \) that satisfies the inequality \( |\lambda| \geq \rho + \eta, \eta > 0 \), there are the numbers \( A = A(\eta) \) and \( h_0 > 0 \) such that \( \forall h < h_0 \) and \( \forall u \in U'_h, u \neq 0 \), the following estimate holds:
\[
\| \mathbf{R}_h u - \tilde{\lambda} u \|_{U'_h} \geq \frac{\rho + \eta}{A(\eta)} \| u \|_{U'_h}. \quad (10.153)
\]
Assume the opposite. Then there exist: \( \eta > 0 \); a sequence of real numbers \( h_k > 0, h_k \to 0 \) as \( k \to \infty \); a sequence of complex numbers \( \lambda_k, |\lambda_k| > \rho + \eta \); and a sequence of vectors \( u_{h_k} \in U'_{h_k} \) such that:
\[
|\mathbf{R}_{h_k} u_{h_k} - \lambda_k u_{h_k} \|_{U'_{h_k}} < \frac{\rho + \eta}{k} \| u_{h_k} \|_{U'_{h_k}}. \quad (10.154)
\]
For sufficiently large values of \( k \), for which \( \frac{\rho + \eta}{k} < 1 \), the numbers \( \lambda_k \) will not lie outside the disk \( |\lambda| \leq c + 1 \) by virtue of estimate (10.151), because outside this disk we have:
\[
\| \mathbf{R}_{h_k} u_{h_k} - \lambda u_{h_k} \|_{U'_{h_k}} \geq (|\lambda| - \| \mathbf{R}_{h_k} \|) \| u_{h_k} \|_{U'_{h_k}} \geq \| u_{h_k} \|_{U'_{h_k}}.
\]
Therefore, the sequence of complex numbers \( \lambda_k \) is bounded and as such, has a limit point \( \tilde{\lambda} \), \( |\tilde{\lambda}| \geq \rho + \eta \). Using the triangle inequality, we can write:
\[
\| \mathbf{R}_{h_k} u_{h_k} - \lambda_k u_{h_k} \|_{U'_{h_k}} \geq \| \mathbf{R}_{h_k} u_{h_k} - \tilde{\lambda} u_{h_k} \|_{U'_{h_k}} - |\lambda_k - \tilde{\lambda}| \| u_{h_k} \|_{U'_{h_k}}.
\]
Substituting into inequality (10.154), we obtain:
\[
\| \mathbf{R}_{h_k} u_{h_k} - \tilde{\lambda} u_{h_k} \|_{U'_{h_k}} < \left[ \frac{\rho + \eta}{k} + |\lambda_k - \tilde{\lambda}| \right] \| u \|_{U'_{h_k}}.
\]
Therefore, according to Definition 10.6 the point \( \tilde{\lambda} \) belongs to the spectrum of the family of operators \( \{\mathbf{R}_h\} \). This contradicts the previous assumption that the spectrum belongs to the disk \( |\lambda| \leq \rho \).
Now let $R$ be a linear operator on a finite-dimensional normed space $U$, $R: U \rightarrow U$. Assume that for any complex $\lambda$, $|\lambda| \geq \gamma > 0$, and any $u \in U$ the following inequality holds for some $a = \text{const} > 0$:

$$\|Ru - \lambda u\| \geq a\|u\|. \quad (10.155)$$

Then,

$$\|R^p\| \leq \frac{\gamma^{p+1}}{a}, \quad p = 1, 2, \ldots. \quad (10.156)$$

Inequality (10.156) follows from the relation:

$$R^p = -\frac{1}{2\pi i} \oint_{|\lambda| = \gamma} \lambda^p (R - \lambda I)^{-1} d\lambda. \quad (10.157)$$

combined with estimate (10.155), because the latter implies that $\|(R - \lambda I)^{-1}\| \leq \frac{1}{a}$. To prove estimate (10.152), we set $R = R_h, \rho = 1$ so that $|\lambda| \geq 1 + \eta = \gamma$, and use (10.153) instead of (10.155). Then estimate (10.156) coincides with (10.152).

It only remains to justify equality (10.157). Define

$$u^{p+1} = Ru^p \quad \text{and} \quad w(\lambda) = \sum_{p=0}^{\infty} \frac{u^p}{\lambda^p}. \quad (10.158)$$

Multiply the equality $u^{p+1} = Ru^p$ by $\lambda^{-p}$ and take the sum with respect to $p$ from $p = 0$ to $p = \infty$. This yields:

$$\lambda w(\lambda) - \lambda u^0 = Rw(\lambda), \quad (10.159)$$

or alternatively,

$$(R - \lambda I)w(\lambda) = -\lambda u^0, \quad w(\lambda) = -\lambda (R - \lambda I)^{-1}u^0. \quad (10.160)$$

From the definition of $w(\lambda)$ it is easy to see that $u^p$ is the residue of the vector-function $\lambda^{p-1}w(\lambda)$:

$$u^p = \frac{1}{2\pi i} \int_{|\lambda| = \gamma} \lambda^{p-1}w(\lambda)d\lambda = \frac{1}{2\pi i} \int_{|\lambda| = \gamma} \lambda^{p}(R - \lambda I)^{-1}u^0 d\lambda. \quad (10.161)$$

As $u^0 = R^0u^0$, the last equality is equivalent to (10.157).

Altogether, we have seen that the question of stability for evolution finite-difference schemes on finite intervals reduces to studying the spectra of the families of the corresponding transition operators $\{R_h\}$. More precisely, we need to find out whether the spectrum for a given family of operators $\{R_h\}$ belongs to the unit disk $|\lambda| \leq 1$. If it does, then the scheme is either stable or, in the worst case scenario, it may only develop a mild instability.
Let us now show how we can actually calculate the spectrum of a family of operators. To demonstrate the approach, we will exploit the previously introduced example (10.141a), (10.141b). It turns out that the algorithm for computing the spectrum of the family of operators \{R_k\} coincides with the Babenko-Gelfand procedure described in Section 10.5.1. Namely, we need to introduce three auxiliary operators: \(\overrightarrow{R}, \overrightarrow{R}, \text{and } \overleftarrow{R}\). The operator \(\overrightarrow{R}, v = \overrightarrow{R}u\), is defined on the linear space of bounded grid functions \(u = \{u_0, u_1, \ldots, u_m, \ldots\}\) according to the formula:

\[
v_m = (1 - r)u_m + ru_{m+1}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad (10.158)
\]

which is obtained from (10.141a), (10.141b) by removing both boundaries. The operator \(\overrightarrow{R}\) is defined on the linear space of functions \(u = \{u_0, u_1, \ldots, u_m, \ldots\}\) that vanish at infinity: \(|u_m| \to 0\) as \(m \to \infty\). It is given by the formula:

\[
v_m = (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, 2, \ldots, \quad (10.159)
\]

which is obtained from (10.141a), (10.141b) by removing the right boundary. Finally, the operator \(\overleftarrow{R}\) is defined on the linear space of functions \(\{u_0, u_1, \ldots, u_m, \ldots\}\) that satisfy: \(|u_m| \to 0\) as \(m \to -\infty\). It is given by the formula:

\[
v_m = (1 - r)u_m + ru_{m+1}, \quad m = \ldots, -1, 0, 1, \ldots, M - 1, \quad (10.160)
\]

which is obtained from (10.141a), (10.141b) by removing the left boundary. Note that the spaces of functions for the operators \(\overrightarrow{R}\) and \(\overleftarrow{R}\) are defined on semi-infinite grids \(m = 0, 1, 2, \ldots\) and \(m = \ldots, -1, 0, 1, \ldots, M\), respectively.

None of the operators \(\overrightarrow{R}, \overleftarrow{R}, \text{or } \overleftarrow{R}\) depend on \(h\). We will show that the combination of all eigenvalues of these three auxiliary operators yields the spectrum of the family of operators \(\{R_k\}\). In Section 10.5.1, we have, in fact, already computed the eigenvalues of the operators \(\overrightarrow{R}\) and \(\overleftarrow{R}\). For the operator \(\overleftarrow{R}\), the eigenvalues are all those and only those complex numbers \(\lambda\), for which the equation \(\overleftarrow{R}u = \lambda u\) has a bounded solution \(u = \{u_m\}, m = 0, \pm 1, \pm 2, \ldots\). According to (10.158), this equation can be written as:

\[
(1 - r - \lambda)u_m + ru_{m+1} = 0, \quad m = 0, \pm 1, \pm 2, \ldots,
\]

and its general solution is \(u_m = cq^m\), where \(q\) is a root of the characteristic equation: \((1 - r - \lambda) + rq = 0\). This solution is bounded as \(|m| \to \infty\) if and only if \(|q| = 1\), i.e., \(q = e^{i\alpha}, \alpha \in [0, 2\pi)\). The corresponding eigenvalues are given by:

\[
\lambda = 1 - r + rq = 1 - r + re^{i\alpha}, \quad \alpha \in [0, 2\pi).
\]

The curve \(\lambda = \lambda(\alpha)\) is a circle of radius \(r\) on the complex plane centered at the point \((1 - r, 0)\), see Figure 10.11(a). We will denote this circle by \(\Lambda\).
The eigenvalues of the operator \( \overline{R} \) are all those and only those complex numbers \( \lambda \), for which the equation \( \overline{R} u = \lambda u \) has a solution \( u = \{u_0, u_1, \ldots, u_m, \ldots\} \) that satisfies \( \lim_{m \to \pm \infty} |u_m| = 0 \). Recasting this equation with the help of formula (10.159), we have:

\[
(1 - r - \lambda)u_m + ru_{m+1} = 0, \quad m = 0, 1, 2, \ldots.
\]

The solution \( u_m = cq^m \) may only be bounded as \( m \to \pm \infty \) if \( |q| < 1 \). The corresponding eigenvalues \( \lambda = 1 - r + rq \) completely fill the interior of the disk of radius \( r \) centered at the point \((1-r,0)\), see Figure 10.11(b). We will denote this set by \( \overline{\Lambda} \).

The eigenvalues of the operator \( \overline{R} \) are computed similarly. Using formula (10.160), we can write equation \( \overline{R} u = \lambda u \) as follows:

\[
(1 - r - \lambda)u_m + ru_{m+1} = 0, \quad m = \ldots, -1, 0, 1, \ldots, M - 1,
\]

\[
(1 - \lambda)u_M = 0.
\]

The general solution of the first equation from this pair is \( u_m = cq^m \), and the relation between \( \lambda \) and \( q \) is \( \lambda = 1 - r + rq \). The solution \( u_m = cq^m \) may only vanish as \( m \to -\infty \) if \( |q| > 1 \). The second equation provides an additional constraint \( (1 - \lambda)q^M = 0 \) so that \( \lambda = 1 \). However, for this particular \( \lambda \) we also have \( q = 1 \), which implies no decay as \( m \to -\infty \). We therefore conclude that the equation \( \overline{R} u = \lambda u \) has no solutions \( u = \{u_m\} \) that satisfy \( \lim_{m \to -\infty} |u_m| = 0 \), i.e., there are no eigenvalues: \( \overline{\Lambda} = 0 \).

The combination of all eigenvalues \( \Lambda = \overline{\Lambda} \cup \overline{\Lambda} \cup \overline{\Lambda} \) is the disk \( |\lambda - (1-r)| \leq r \) on the complex plane; it is centered at \((1-r,0)\) and has radius \( r \). We will now show that the spectrum of the family of operators \( \{R_h\} \) coincides with the set \( \Lambda \). This is equivalent to showing that every point \( \lambda_0 \in \Lambda \) belongs to the spectrum of \( \{R_h\} \) and that this spectrum contains no other points.

According to Definition 10.6, to prove the first implication it is sufficient to demonstrate that for any \( \epsilon > 0 \) the inequality

\[
\|R_h u - \lambda_0 u\|_{L^h} < \epsilon \|u\|_{L^h}
\]

has a solution \( u \in U'_h \) for all sufficiently small \( h > 0 \). As \( \lambda_0 \in \Lambda \), then \( \lambda_0 \in \overline{\Lambda} \) or \( \lambda_0 \in \overline{\Lambda} \), because \( \overline{\Lambda} = 0 \). Note that when \( \epsilon \) is small one may call the solution \( u \) of inequality (10.161) “almost an eigenvector” of the operator \( R_h \), since a solution to the equation \( R_h u - \lambda_0 u = 0 \) is its genuine eigenvector.

Let us first assume that \( \lambda_0 \in \overline{\Lambda} \). To construct a solution \( u \) of inequality (10.161), we recall that by definition of the set \( \overline{\Lambda} \) there exists \( q_0 \): \( |q_0| = 1 \), such that \( \lambda_0 = 1 - r + rq_0 \) and the equation \( (1 - r - \lambda_0)v_m + rv_{m+1} = 0 \), \( m = 0, \pm 1, \pm 2, \ldots \), has a bounded solution \( v_m = q_m^0 \), \( m = 0, \pm 1, \pm 2, \ldots \). We will consider this solution only for \( m = 0, 1, 2, \ldots, M \), while keeping the same notation \( v \). It turns out that the vector:

\[
v = [v_0, v_1, v_2, \ldots, v_M] = [1, q_0, q_0^2, \ldots, q_0^M]
\]
almost satisfies the operator equation $R_h v - \lambda_0 v = 0$ that we write as:

$$
(1 - r - \lambda_0)v_m + rv_{m+1} = 0, \quad m = 0, 1, 2, \ldots, M - 1,
$$

$$(1 - \lambda_0)v_M = 0.
$$

The vector $v$ would have completely satisfied the previous equation, which is an even stronger constraint than inequality (10.161), if it did not violate the last relation $(1 - \lambda_0)v_M = 0$.\footnote{Relation $(1 - \lambda_0)v_M = 0$ is violated unless $\lambda_0 = q_0 = 1$.} This relation can be interpreted as a boundary condition for the difference equation:

$$
(1 - r - \lambda_0)u_m + ru_{m+1} = 0, \quad m = 0, 1, 2, \ldots, M - 1.
$$

The boundary condition is specified at $m = M$, i.e., at the right endpoint of the interval $0 \leq x \leq 1$. To satisfy this boundary condition, let us "correct" the vector $v = [1, q_0, q_0^2, \ldots, q_0^M]$ by multiplying each of its components $v_m$ by the respective factor $(M - m)h$. The resulting vector will be denoted $u = [u_0, u_1, \ldots, u_M]$, $u_m = (M - m)h q_0^m$. Obviously, the vector $u$ has unit norm:

$$
\|u\|_{U_h}^2 = \max_m |u_m|^2 = \max_m |(M - m)h q_0^m|^2 = Mh = 1.
$$

We will now show that this vector $u$ furnishes a desired solution to the inequality (10.161). Define the vector $w \overset{\text{def}}{=} R_h u - \lambda_0 u$, $w = [w_0, w_1, \ldots, w_M]$. We need to estimate its norm. For the individual components of $w$, we have:

$$
|w_m| = |(1 - r - \lambda_0)(M - m)h q_0^m + r(M - m - 1)h q_0^{m+1}|
= |[(1 - r - \lambda_0) + r q_0](M - m)h q_0^m - rh q_0^{m+1}|
= 0 \cdot (M - m)h q_0^m - rh q_0^{m+1} = rh, \quad m = 0, 1, \ldots, M - 1,
$$

$$
|w_M| = |u_M - \lambda_0 u_M| = |0 - \lambda_0 \cdot 0| = 0.
$$

Consequently, $\|w\|_{U_h} = rh$, and for $h < \epsilon / r$ we obtain: $\|w\|_{U_h} < \|R_h u - \lambda_0 u\|_{U_h} < \epsilon = \epsilon \|u\|_{U_h}$, i.e., inequality (10.161) is satisfied. Thus, we have shown that if $\lambda_0 \in \Lambda$, then this point also belongs to the spectrum of the family of operators $\{R_h\}$.

Next, let us assume that $\lambda_0 \in \Lambda$ and show that in this case $\lambda_0$ also belongs to the spectrum of the family of operators $\{R_h\}$. According to (10.159), equation $\overline{R} v - \lambda_0 v = 0$ is written as:

$$
(1 - r - \lambda_0)v_m + rv_{m+1} = 0, \quad m = 0, 1, 2, \ldots.
$$

Since $\lambda_0 \in \Lambda$, this equation has a solution $v_m = q_0^m$, $m = 0, 1, 2, \ldots$, where $|q_0| < 1$. We will consider this solution only for $m = 0, 1, 2, \ldots, M$:

$$
u = [u_0, u_1, u_2, \ldots, u_M] = [1, q_0, q_0^2, \ldots, q_0^M], \quad \|u\|_{U_h} = 1.
$$
As before, define \( w \overset{\text{def}}{=} R_hu - \lambda_0 u \). For the components of the vector \( w \) we have:
\[
\begin{align*}
|w_m| &= |(1 - r - \lambda_0)q_0^m + rq_0^{m+1}| = 0, \quad m = 0, 1, \ldots, M - 1, \\
|w_M| &= |1 - \lambda_0| \cdot |q_0^M|.
\end{align*}
\]
Consequently, \( \|w\|_{L^r} = |1 - \lambda_0| \cdot |q_0|^M = |1 - \lambda_0| \cdot |q_0|^{1/h} \). Since \( |q_0| < 1 \), then for any \( \varepsilon > 0 \) we can always choose a sufficiently small \( h \) so that \( |1 - \lambda_0| \cdot |q_0|^{1/h} < \varepsilon \).

Then, \( \|w\|_{L^r} = \|R_hu - \lambda_0 u\|_{L^r} < \varepsilon = \|u\|_{L^r} \) and the inequality (10.161) is satisfied.

Note that if the set \( \Lambda \) were not empty, then proving that each of its elements belongs to the spectrum of the family of operators \( \{R_h\} \) would have been similar. Altogether, we have thus shown that in our specific example given by equations (10.141) every \( \lambda_0 \in \{\Lambda \cup \bar{\Lambda} \cup \overline{\Lambda} \} \) is also an element of the spectrum of \( \{R_h\} \).

Now we need to prove that if \( \lambda_0 \not\in \{\Lambda \cup \bar{\Lambda} \cup \overline{\Lambda} \} \) then it does not belong to the spectrum of the family of operators \( \{R_h\} \) either. To that end, it will be sufficient to show that there is an \( h \)-independent constant \( A \), such that for any \( u = [u_0, u_1, \ldots, u_M] \) the following inequality holds:
\[
\|R_hu - \lambda_0 u\|_{L^r} \geq A\|u\|_{L^r}.
\] (10.162)

Then, for \( \varepsilon < A \), inequality (10.161) will have no solutions, and therefore the point \( \lambda_0 \) will not belong to the spectrum. Denote \( f = R_hu - \lambda_0 u \), then inequality (10.162) reduces to:
\[
\|f\|_{L^r} \geq A\|u\|_{L^r}.
\] (10.163)

Our objective is to justify estimate (10.163). Rewrite the equation \( R_hu - \lambda_0 u = f \) as:
\[
\begin{align*}
(1 - r - \lambda_0)u_m + ru_{m+1} &= f_m, \quad m = 0, 1, \ldots, M - 1, \\
(1 - \lambda_0)u_M &= f_M,
\end{align*}
\]
and interpret these relations as an equation with respect to the unknown \( u = \{u_m\} \), whereas the right-hand side \( f = \{f_m\} \) is assumed given. Let
\[
u_m = v_m + w_m, \quad m = 0, 1, \ldots, M,
\] (10.164)
where \( v_m \) are components of the bounded solution \( v = \{v_m\}, m = 0, \pm 1, \pm 2, \ldots, \) to the following equation:
\[
(1 - r - \lambda_0)v_m + rv_{m+1} = f_m \overset{\text{def}}{=} \hat{f}_m \begin{cases} 0, & \text{if } m < 0, \\
 f_m, & \text{if } m = 0, 1, \ldots, M - 1, \\
 0, & \text{if } m \geq M.
\end{cases}
\] (10.165)

Then because of the linearity, the grid function \( w = \{w_m\} \) introduced by formula (10.164) solves the equation:
\[
\begin{align*}
(1 - r - \lambda_0)w_m + rw_{m+1} &= 0, \quad m = 0, 1, \ldots, M - 1, \\
(1 - \lambda_0)w_M &= f_M - (1 - \lambda_0)v_M.
\end{align*}
\] (10.166)
Let us now recast estimate (10.163) as \(|u_m| \leq A^{-1} \max_m |f_m|\). According to (10.164), to prove this estimate it is sufficient to establish individual inequalities:

\[
|v_m| \leq A_1 \max_m |f_m|, \quad (10.167a)
\]

\[
|w_m| \leq A_2 \max_m |f_m|, \quad (10.167b)
\]

where \(A_1\) and \(A_2\) are constants. We begin with inequality (10.167a). Notice that equation (10.165) is a first-order constant-coefficient ordinary difference equation:

\[
av_m + bv_{m+1} = \hat{f}_m, \quad m = 0, \pm 1, \pm 2, \ldots
\]

where \(a = 1 - r - \lambda_0\), \(b = r\). Its bounded fundamental solution is given by

\[
G_m = \begin{cases} \frac{1}{a} \left( -\frac{a}{b} \right)^m, & m \leq 0, \\ 0, & m \geq 1, \end{cases}
\]

because \(\lambda_0 \not\in \{ \Lambda \cup \Lambda \cup \Lambda \} \), i.e., \(|\lambda_0 - (1 - r)| > r\), and consequently \(|a/b| > 1\).

Representing the solution \(v_m\) in the form of a convolution:

\[
v_m = \sum_{k=-\infty}^{\infty} G_{m-k} \hat{f}_k
\]

and summing up the geometric sequence we arrive at the estimate:

\[
|v_m| \leq \frac{\max_m |\hat{f}_m|}{|a| - |b|} = \frac{\max_m |f_m|}{|a| - |b|}.
\]

Introducing the distance \(\delta_0\) between the point \(\lambda_0\) and the set \(\{ \Lambda \cup \Lambda \cup \Lambda \} \), we can obviously claim that \(|a| - |b| > \delta_0/2\), which makes the previous estimate equivalent to (10.167a). Estimate (10.167b) can be obtained by representing the solution of equation (10.166) in the form:

\[
w_m = \frac{f_M - (1 - \lambda_0)v_M}{1 - \lambda_0}g_0^{-m-M}, \quad (10.168)
\]

where \(q_0\) is determined by the relation \((1 - r - \lambda_0) + rq_0 = 0\). Our assumption is that \(\lambda_0 \not\in \{ \Lambda \cup \Lambda \cup \Lambda \} \), i.e., that \(\lambda_0\) lies outside of the disk of radius \(r\) on the complex plane centered at \((1 - r, 0)\). In this case \(|q_0| > 1\). Moreover, we can say that \(|1 - \lambda_0| = \delta_1 > 0\), because if \(\lambda_0 = 1\), then \(\lambda_0\) would have belonged to the set \(\{ \Lambda \cup \Lambda \cup \Lambda \} \). As such, using formula (10.168) and taking into account estimate (10.167a) that we have already proved, we obtain the desired estimate (10.167b):

\[
|w_m| = \frac{|f_M - (1 - \lambda_0)v_M|}{|1 - \lambda_0|} \cdot |q_0^{-m-M}| \leq \frac{|f_M|}{|1 - \lambda_0|} + |v_M| \\
\leq \frac{\max_m |f_m|}{\delta_1} + A_1 \max_m |f_m| = A_2 \max_m |f_m|.
\]
We have thus proven that the spectrum of the family of operators \( \{ R_h \} \) defined by formulae (10.141) coincides with the set \( \{ \overline{\Lambda} \cup \Lambda \cup \overline{\Lambda} \} \) on the complex plane.

The foregoing algorithm for computing the spectrum of the family of operators \( \{ R_h \} \) is, in fact, quite general. We have illustrated it using a particular example of the operators defined by formulae (10.141). However, for all other scalar finite-difference schemes the spectrum of the family of operators \( \{ R_h \} \) can be obtained by performing the same Babenko-Gelfand analysis of Section 10.5.1. The key idea is to take into account other candidate modes that may be prone to developing the instability, besides the eigenmodes \( \{ e^{i\alpha m} \} \) of the pure Cauchy problem that are accounted for by the von Neumann analysis.

When systems of finite-difference equations need to be addressed as opposed to the scalar equations, the technical side of the procedure becomes more elaborate. In this case, the computation of the spectrum of a family of operators can be reduced to studying uniform bounds for the solutions of certain ordinary difference equations with matrix coefficients. A necessary and sufficient condition has been obtained in [Rya64] for the existence of such uniform bounds. This condition is given in terms of the roots of the corresponding characteristic equation and also involves the analysis of some determinants originating from the matrix coefficients of the system. For further detail, we refer the reader to [GR87, § 4 & § 43] and [RM67, § 6.6 & § 6.7], as well as to the original journal publication by Ryaben’kii [Rya64].

### 10.5.3 The Energy Method

For some evolution finite-difference problems, one can obtain the \( l_2 \) estimates of the solution directly, i.e., without employing any special stability criteria, such as spectral. The corresponding technique is known as the method of energy estimates. It is instrumental for deriving sufficient conditions of stability, in particular, because it can often be applied to problems with variable coefficients on finite intervals. We illustrate the energy method with several examples.

In the beginning, let us analyze the continuous case. Consider an initial boundary value problem for the first-order constant-coefficient hyperbolic equation:

\[
\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad 0 \leq x \leq 1, \quad 0 < t \leq T,
\]

\[
u(x,0) = \psi(x), \quad u(1,t) = 0.
\]

Note that both the differential equation and the boundary condition at \( x = 1 \) in problem (10.169) are homogeneous. Multiply the differential equation of (10.169) by \( u = u(x,t) \) and integrate over the entire interval \( 0 \leq x \leq 1 \):

\[
\frac{d}{dt} \int_0^1 u^2(x,t) \frac{1}{2} \, dx \, - \int_0^1 \frac{\partial}{\partial x} \frac{u^2(x,t)}{2} \, dx \, = \frac{d}{dt} \left[ \frac{\| u(\cdot, t) \|^2}{2} \right] \, - \frac{u^2(1,t)}{2} \, + \frac{u^2(0,t)}{2} = 0,
\]
where \(\|u(\cdot, t)\|_2 \overset{\text{def}}{=} \left(\int_0^1 u^2(x,t)\,dx\right)^{1/2}\) is the \(L_2\) norm of the solution in space for a given moment of time \(t\). According to formula (10.169), the solution at \(x = 1\) vanishes: \(u(1,t) = 0\), and we conclude that \(\frac{d}{dt}\|u(\cdot, t)\|_2^2 \leq 0\), which means that \(\|u(\cdot, t)\|_2\) is a non-increasing function of time. Consequently, we see that the \(L_2\) norm of the solution will never exceed that of the initial data:

\[
\|u(\cdot, t)\|_2 \leq \|\psi\|_2, \quad t \geq 0. \tag{10.170}
\]

Inequality (10.170) is the simplest energy estimate. It draws its name from the fact that the quantities that are quadratic with respect to the solution are often interpreted as energy in the context of physics. Note that estimate (10.170) holds for all \(t \geq 0\) rather than only \(0 \leq t \leq T\).

Next, we consider a somewhat more general formulation compared to (10.169), namely, an initial boundary value problem for the hyperbolic equation with a variable coefficient:

\[
\frac{\partial u}{\partial t} - a(x,t)\frac{\partial u}{\partial x} = 0, \quad 0 \leq x \leq 1, \quad 0 < t \leq T, \tag{10.171}
\]

\[
u(x,0) = \psi(x), \quad u(1,t) = 0.
\]

We are assuming that \(\forall x \in [0,1]\) and \(\forall t \geq 0\) : \(a(x,t) \geq a_0 > 0\), so that the characteristic speed is negative across the entire domain. Then, the differential equation renders transport from the right to the left. Consequently, setting the boundary condition \(u(1,t) = 0\) at the right endpoint of the interval \(0 \leq x \leq 1\) is legitimate.

Let us now multiply the differential equation of (10.171) by \(u = u(x,t)\) and integrate over the entire interval \(0 \leq x \leq 1\), while also applying integration by parts to the spatial term:

\[
\frac{d}{dt} \int_0^1 \frac{u^2(x,t)}{2}\,dx - \int_0^1 a(x,t)\frac{\partial u^2(x,t)}{\partial x}\,dx
\]

\[
= \frac{d}{dt}\|u(\cdot, t)\|_2^2 - a(1,t)\frac{u^2(1,t)}{2} + a(0,t)\frac{u^2(0,t)}{2} + \int_0^1 a'(x,t)\frac{u^2(x,t)}{2}\,dx = 0.
\]

Using the boundary condition \(u(1,t) = 0\), we find:

\[
\frac{d}{dt}\|u(\cdot, t)\|_2^2 = -a(0,t)\frac{u^2(0,t)}{2} - \int_0^1 a'(x,t)\frac{u^2(x,t)}{2}\,dx = 0.
\]

The first term on the right-hand side of the previous equality is always non-positive. As far as the second term, let us denote \(A = \sup_{(x,t)}[-a'(x,t)]\). Then we have:

\[
\frac{d}{dt}\|u(\cdot, t)\|_2^2 \leq A\|u(\cdot, t)\|_2^2,
\]

which immediately yields:

\[
\|u(\cdot, t)\|_2 \leq e^{At/2}\|\psi\|_2, \quad t \geq 0.
\]
If $A < 0$, the previous inequality implies that the $L_2$ norm of the solution decays as $t \rightarrow +\infty$. If $A = 0$, then the norm of the solution stays bounded by the norm of the initial data. To obtain an overall uniform estimate of $\|u(\cdot, t)\|_2$ for $A \leq 0$ and all $t \geq 0$, we need to select the maximum value of the constant: $\max_x e^{At/2} = 1$, and then the desired inequality will coincide with (10.170). For $A > 0$, a uniform estimate can only be obtained for a given fixed interval $0 \leq t \leq T$, so that altogether we can write:

$$
\|u(\cdot, t)\|_2 \leq \begin{cases} 
\|\psi\|_2, & \text{if } A \leq 0, \ t \geq 0, \\
 e^{AT/2}\|\psi\|_2, & \text{if } A > 0, \ 0 \leq t \leq T.
\end{cases} \tag{10.172}
$$

Similarly to inequality (10.170), the energy estimate (10.172) also provides a bound for the $L_2$ norm of the solution in terms of the $L_2$ norm of the initial data. However, when $A > 0$ the constant in front of $\|\psi\|_2$ is no longer equal to one. Instead, $e^{AT/2}$ grows exponentially as the maximum time $T$ elapses, and therefore estimate (10.172) for $A > 0$ may only be considered on a finite interval $0 \leq t \leq T$ rather than for $t \geq 0$.

In problems (10.169) and (10.171) the boundary condition at $x = 1$ was homogeneous. Let us now introduce yet another generalization and analyze the problem:

$$
\frac{\partial u}{\partial t} - a(x,t) \frac{\partial u}{\partial x} = 0, \quad 0 \leq x \leq 1, \quad 0 < t \leq T, \quad u(x,0) = \psi(x), \quad u(1,t) = \chi(t) \tag{10.173}
$$

that differs from (10.171) by its inhomogeneous boundary condition: $u(1,t) = \chi(t)$. Otherwise everything is the same; in particular, we still assume that $\forall x \in [0,1]$ and $\forall t \geq 0 : a(x,t) \geq a_0 > 0$ and denote $A = \sup_{(x,t)} [-a''(x,t)]$. Multiplying the differential equation of (10.173) by $u(x,t)$ and integrating by parts, we obtain:

$$
\frac{d}{dt} \|u(\cdot, t)\|_2^2 = -a(0,t) \frac{u^2(0,t)}{2} + a(1,t) \frac{\chi^2(t)}{2} - \int_0^1 a''(x,t) \frac{u^2(x,t)}{2} dx = 0.
$$

Consequently,

$$
\frac{d}{dt} \|u(\cdot, t)\|_2^2 \leq A \|u(\cdot, t)\|_2^2 + a(1,t) \chi^2(t).
$$

Multiplying the previous inequality by $e^{-At}$, we have:

$$
\frac{d}{dt} \left[e^{-At}\|u(\cdot, t)\|_2^2\right] \leq e^{-At} a(1,t) \chi^2(t),
$$

which, after integrating over the time interval $0 \leq \theta \leq t$, yields:

$$
\|u(\cdot, t)\|_2^2 \leq e^{At} \|\psi\|_2^2 + e^{At} \int_0^t e^{-A\theta} a(1,\theta) \chi^2(\theta) d\theta.
$$

As in the case of a homogeneous boundary condition, we would like to obtain a uniform energy estimate for a given interval of time. This can be done if we again
distinguish between $A \leq 0$ and $A > 0$. When $A \leq 0$ we can consider all $t \geq 0$ and when $A > 0$ we can only have the estimate on some fixed interval $0 \leq t \leq T$:

$$\|u(\cdot, t)\|_2^2 \leq \begin{cases} \|\psi\|_2^2 + \int_0^\infty a(1, \theta)\chi^2(\theta)d\theta, & A \leq 0, \ t \geq 0, \\ e^{AT} \|\psi\|_2^2 + \int_0^T a(1, \theta)\chi^2(\theta)d\theta, & A > 0, \ 0 \leq t \leq T. \end{cases} \quad (10.174)$$

When deriving inequalities (10.174), we obviously need to assume that the integrals on the right-hand side of (10.174) are bounded. These integrals can be interpreted as weighted $L_2$ norms of the boundary data $\chi(t)$. Clearly, energy estimate (10.174) includes the previous estimate (10.172) as a particular case for $\chi(t) \equiv 0$.

All three estimates (10.170), (10.172), and (10.174) indicate that the corresponding initial boundary value problem is well-posed in the sense of $L_2$. Qualitatively, well-posedness means that the solution to a given problem is only weakly sensitive to perturbations of the input data (such as initial and/or boundary data). In the case of linear evolution problems, well-posedness can, for example, be quantified by means of the energy estimates. These estimates provide a bound for the norm of the solution in terms of the norms of the input data. In the finite-difference context, similar estimates would have implied stability in the sense of $l_2$, provided that the corresponding constants on the right-hand side of each inequality can be chosen independent of the grid. We will now proceed to demonstrate how energy estimates can be obtained for finite-difference schemes.

Consider the first-order upwind scheme for problem (10.169):

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0,$$

$$m = 0, 1, \ldots, M - 1, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1,$$  \quad (10.175)

$$u_m^0 = \psi_m, \quad u_M^p = 0.$$

To obtain an energy estimate for scheme (10.175), let us fist consider two arbitrary functions $\{u_m\}$ and $\{v_m\}$ on the grid $m = 0, 1, \ldots, M$. We will derive a formula that can be interpreted as a discrete analogue of the classical continuous integration by parts. In the literature, it is sometimes referred to as the summation by parts:

$$\sum_{m=0}^{M-1} u_m(v_{m+1} - v_m) = \sum_{m=0}^{M-1} u_m v_{m+1} - \sum_{m=0}^{M-1} u_m v_m = \sum_{m=1}^{M-1} u_{m-1} v_m - \sum_{m=0}^{M-1} u_m v_m$$

$$= - \sum_{m=1}^{M} (u_m - u_{m-1})v_m + u_Mv_M - u_0v_0$$

$$= - \sum_{m=0}^{M-1} (u_{m+1} - u_m)v_{m+1} + u_Mv_M - u_0v_0$$

$$= - \sum_{m=0}^{M-1} (u_{m+1} - u_m)v_{m+1} - \sum_{m=0}^{M-1} (u_{m+1} - u_m)(v_{m+1} - v_m)$$

$$+ u_Mv_M - u_0v_0.$$


Next, we rewrite the difference equation of (10.175) as
\[ u^{p+1}_m = u^p_m + r(u^p_{m+1} - u^p_m), \quad r = \frac{\tau}{h} = \text{const}, \quad m = 0, 1, \ldots, M - 1, \]
square both sides, and take the sum from \( m = 0 \) to \( m = M - 1 \). This yields:
\[
\sum_{m=0}^{M-1} (u^{p+1}_m)^2 = \sum_{m=0}^{M-1} (u^p_m)^2 + r^2 \sum_{m=0}^{M-1} (u^p_{m+1} - u^p_m)^2 + 2r \sum_{m=0}^{M-1} u^p_m (u^p_{m+1} - u^p_m).
\]
To transform the last term on the right-hand side of the previous equality, we apply formula (10.176):
\[
\sum_{m=0}^{M-1} (u^{p+1}_m)^2 = \sum_{m=0}^{M-1} (u^p_m)^2 + r^2 \sum_{m=0}^{M-1} (u^p_{m+1} - u^p_m)^2 + r \sum_{m=0}^{M-1} u^p_m (u^p_{m+1} - u^p_m)
+ r \left[ - \sum_{m=0}^{M-1} (u^p_{m+1} - u^p_m) u^p_m - \sum_{m=0}^{M-1} (u^p_{m+1} - u^p_m)^2 \right] + 2r \sum_{m=0}^{M-1} u^p_m (u^p_{m+1} - u^p_m).
\]
Let us now assume that \( r \leq 1 \). Then, using the conventional definition of the \( l_2 \) norm:
\[ \|u\|_2 = \left[ h \sum_0^M |u_m|^2 \right]^{1/2} \]
and employing the homogeneous boundary condition \( u^p_M = 0 \) of (10.175), we obtain the inequality:
\[
\|u^{p+1}\|_2 \leq \|u^p\|_2, \quad p = 0, 1, 2, \ldots,
\]
which clearly implies the energy estimate:
\[
\|u^p\|_2 \leq \|\psi\|_2, \quad p = 0, 1, 2, \ldots.
\] (10.177)
The discrete estimate (10.177) is analogous to the continuous estimate (10.170).

To approximate the variable-coefficient problem (10.171), we use the scheme:
\[
\frac{u^p_{m+1} - u^p_m}{\tau} - a^p_m \frac{u^p_{m+1} - u^p_m}{h} = 0, \quad m = 0, 1, \ldots, M - 1, \quad p = 0, 1, \ldots, [T/\tau] - 1,
\]
where \( a^p_m \equiv a(x_m, t_p) \). Applying a similar approach, we obtain:
\[
\sum_{m=0}^{M-1} (u^{p+1}_m)^2 = \sum_{m=0}^{M-1} (u^p_m)^2 + r^2 \sum_{m=0}^{M-1} (a^p_m)^2 (u^p_{m+1} - u^p_m)^2 + r \sum_{m=0}^{M-1} a^p_m (u^p_{m+1} - u^p_m)
+ r \left[ - \sum_{m=0}^{M-1} (a^p_{m+1} u^p_{m+1} - a^p_m u^p_m) u^p_m + r \sum_{m=0}^{M-1} (a^p_{m+1} u^p_{m+1} - a^p_m u^p_m) (u^p_{m+1} - u^p_m) + a^p_M (u^p_M)^2 - a^p_0 (u^p_0)^2 \right].
\]
Next, we notice that \( a_{m+1} u_{m+1} = a_m u_m + (a_{m+1} - a_m) u_{m+1} \). Substituting this expression into the previous formula, we again perform summation by parts, which is analogous to the continuous integration by parts, and which yields:

\[
\sum_{m=0}^{M-1} (u_{m+1}^p)^2 = \sum_{m=0}^{M-1} (u_m^p)^2 + r^2 \sum_{m=0}^{M-1} (ap_m^p)^2 (u_{m+1}^p - u_m^p)^2 + r \sum_{m=0}^{M-1} ap_m^p (u_{m+1}^p - u_m^p) \]

Next, we notice that \( 256 \). Theoretical Introduction to Numerical Analysis

Then, using the homogeneous boundary condition \( u_0^p = 0 \) and dropping the a priori non-positive term \( \sum_{m=0}^{M-1} r a_m^p (r a_m^p - 1) (u_{m+1}^p - u_m^p)^2 \), we obtain:

\[
\sum_{m=0}^{M-1} (u_m^{p+1})^2 \leq \sum_{m=0}^{M-1} (u_m^p)^2 + r A \sum_{m=0}^{M-1} h (u_m^p)^2 - r a_0^p (u_0^p)^2 - r A h (u_0^p)^2.
\]

If \( A > 0 \), then for the last two terms on the right-hand side of the previous inequality we clearly have: \(-r (u_0^p)^2 (a_0^p + h) < 0 \). Even if \( A \leq 0 \) we can still claim that \(-r (u_0^p)^2 (a_0^p + h) < 0 \) for sufficiently small \( h \). Consequently, on fine grids the following inequality holds:

\[
\|u^{p+1}\|_2^2 \leq \|u^p\|_2^2 + r A \|u^p\|_2^2 = (1 + A r) \|u^p\|_2^2,
\]

which immediately implies:

\[
\|u^p\|_2^2 \leq (1 + A r)^p \|\psi\|_2^2, \quad p = 1, 2, \ldots.
\]

If \( A \leq 0 \), the norm of the discrete solution will either decay or remain bounded as \( p \) increases. If \( A > 0 \), a uniform estimate of \( \|u^p\|_2 \) can only be obtained for \( p = \).
0, 1, . . . , [T/τ]. Altogether, the solution $u^p = \{u^p_m\}$ to the finite-difference problem (10.178) satisfies the following energy estimate:

$$
\|u^p\|_2 \leq \begin{cases} 
\|\psi\|_2^2, & A \leq 0, \quad p = 0, 1, 2, \ldots, \\
\cdot e^{AT/2} \|\psi\|_2, & A > 0, \quad p = 0, 1, 2, \ldots, [T/τ].
\end{cases}
$$

(10.179)

The discrete estimate (10.179) is analogous to the continuous estimate (10.172). Finally, for problem (10.173) we use the scheme:

$$
\frac{u^p_m - u^0_m}{\tau} - a^p_m \frac{u^p_{m+1} - u^p_m}{h} = 0,
$$

$$
m = 0, 1, \ldots, M - 1, \quad p = 0, 1, \ldots, [T/τ] - 1,
$$

(10.180)

Under the same assumptions that we introduced when deriving estimate (10.179) for scheme (10.178), we can now write for scheme (10.180):

$$
\sum_{m=0}^{M-1} (u^p_{m+1})^2 \leq \sum_{m=0}^{M-1} (u^p_m)^2 + r A \sum_{m=0}^{M} h (u^p_m)^2 + ra^p_m (\chi^p)^2.
$$

Denote $\|u^p\|_2^2 = h \sum_{m=0}^{M-1} |u_m|^2 = \|u\|_2^2 - hu_M^2$. Then the previous inequality implies:

$$
\|u^p+1\|_2^2 \leq (1 + A\tau)\|u^p\|_2^2 + \tau(a^p_M + Ah)(\chi^p)^2, \quad p = 1, 2, \ldots,
$$

and consequently:

$$
\|u^p\|_2^2 \leq \|u^p\|_2^2 + (1 + A\tau)^p \|\psi\|_2^2 + \sum_{k=1}^{p} (1 + A\tau)^{p-k} \tau(a^k_M + Ah)(\chi^{k-1})^2, \quad p = 1, 2, \ldots.
$$

We again need to distinguish between the cases $A \leq 0, p = 0, 1, 2, \ldots$, and $A > 0, p = 0, 1, 2, \ldots, [T/τ]$:

$$
\|u^p\|_2^2 \leq \begin{cases} 
\|\psi\|_2^2 + \sum_{k=1}^{\infty} \tau a^k_M (\chi^{k-1})^2, & A \leq 0, \\
e^{AT/2} \|\psi\|_2^2 + \sum_{k=1}^{[T/τ]} \tau(a^k_M + Ah)(\chi^{k-1})^2, & A > 0.
\end{cases}
$$

(10.181)

The discrete estimate (10.181) is analogous to the continuous estimate (10.174). To use the norms $\| \cdot \|_2$ instead of $\| \cdot \|_2$ in (10.181), we only need to add a bounded quantity $h(\chi^0)^2 + h(\chi^p)^2$ on the right-hand side.

Energy estimates (10.177), (10.179), and (10.181) imply the $l_2$ stability of the schemes (10.175), (10.178), and (10.180), respectively, in the sense of the Definition 10.2 from page 164. Note that since the foregoing schemes are explicit, stability is not unconditional, and the Courant number has to satisfy $r \leq 1$ for scheme (10.175) and $r \leq \left[\sup_{(x,t)} a(x,t)\right]^{-1}$ for schemes (10.178) and (10.180).
In general, direct energy estimates appear instrumental for studying stability of finite-difference schemes. Indeed, they may provide sufficient conditions for those difficult cases that involve variable coefficients, boundary conditions, and even multiple space dimensions. In addition to the scalar hyperbolic equations, energy estimates can be obtained for some hyperbolic systems, as well as for the parabolic equations. For detail, we refer the reader to [GKO95, Chapters 9 & 11], and to some fairly recent journal publications [Str94, Ols95a, Ols95b]. However, there is a key non-trivial step in proving energy estimates for finite-difference initial boundary value problems, which is obtaining the discrete summation by parts rules appropriate for a given discretization [see the example given by formula (10.176)]. Sometimes, this step may not be obvious at all; otherwise, it may require using the alternative norms based on specially chosen inner products.

10.5.4 A Necessary and Sufficient Condition of Stability

In Section 10.5.2, we have shown that for stability of a finite-difference initial boundary value problem it is necessary that the spectrum of the family of transition operators $R_h$ belongs to the unit disk on the complex plane. We have also shown, see Theorem 10.9, that this condition is, in fact, not very far from a sufficient one, as it guarantees the scheme from developing a catastrophic exponential instability. However, it is not a fully sufficient condition, and there are examples of the schemes that satisfy the Godunov and Ryaben’kii criterion of Section 10.5.2, i.e., that have their spectrum of $\{R_h\}$ inside the unit disk, yet they are unstable.

A comprehensive analysis of necessary and sufficient conditions for the stability of evolution-type schemes on finite intervals is rather involved. In the literature, the corresponding series of results is commonly referred to as the Gustafsson, Kreiss, and Sundström (GKS) theory, and we refer the reader to the monograph [GKO95, Part II] for detail. A concise narrative of this theory can also be found in [Str04, Chapter 11]. As often is the case when analyzing sufficient conditions, all results of the GKS theory are formulated in terms of the $l_2$ norm. An important tool used for obtaining stability estimates is the Laplace transform in time.

Although a full account of (and even a self-contained introduction to) the GKS theory is beyond the scope of this text, its key ideas are easy to understand on the qualitative level and easy to illustrate with examples. The following material is essentially based on that of Section 10.5.2 and can be skipped during the first reading.

Let us consider an initial boundary value problem for the first-order constant coefficient hyperbolic equation:

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad 0 \leq x \leq 1, \quad 0 < t \leq T,$$

$$u(x,0) = \psi(x), \quad u(1,t) = 0.$$  \quad (10.182)

We introduce a uniform grid: $x_m = mh, \ m = 0,1,\ldots,M, \ h = 1/M; \ t_p = p\tau, \ p =$
Finite-Difference Schemes for Partial Differential Equations

0, 1, 2, ..., and approximate problem (10.182) with the leap-frog scheme:

\[
\frac{u_{m+1}^{p+1} - u_m^{p-1}}{2\tau} - \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} = 0,
\]

\(m = 1, 2, \ldots, M - 1, \quad p = 1, 2, \ldots, \lceil T/\tau \rceil - 1,\)

(10.183)

\[
u_m^0 = \psi(x_m), \quad u_m^1 = \psi(x_m + \tau), \quad m = 0, 1, \ldots, M,
\]

\(l u_0^{p+1} = 0, \quad u_M^{p+1} = 0, \quad p = 1, 2, \ldots, \lceil T/\tau \rceil - 1.
\]

Notice that scheme (10.183) requires two initial conditions, and for simplicity we use the exact solution, which is readily available in this case, to specify \(u_{m}^{1}\) for \(m = 0, 1, \ldots, M - 1\). Also notice that the differential problem (10.182) does not require any boundary conditions at the “outflow” boundary \(x = 0\), but the discrete problem (10.183) does require an additional boundary condition that we symbolically denote \(l u_0^{p+1} = 0\). We will investigate two different outflow conditions for scheme (10.183):

\[
u_0^{p+1} = u_1^{p+1}
\]

(10.184a)

and

\[
u_0^{p+1} = u_0^{p} + r(u_1^{p} - u_0^{p}),
\]

(10.184b)

where we have used our standard notation \(r = \frac{\tau}{h} = \text{const.}\).

Let us first note that scheme (10.183) is not a one-step scheme. Therefore, to reduce it to the canonical form (10.140) to subsequently investigate the spectrum of the family of operators \(\{R_h\}\), we would formally need to introduce additional variables (i.e., transform a scalar equation into a system) and then consider a one-step finite-difference equation, but with vector unknowns. It is, however, possible to show that even after a reduction of that type has been implemented, the Babenko-Gelfand procedure of Section 10.5.1 applied to the resulting vector scheme for calculating the spectrum of \(\{R_h\}\) will be equivalent to the Babenko-Gelfand procedure applied directly to the multi-step scheme (10.183). As such, we will skip the formal reduction of scheme (10.183) to the canonical form (10.140) and proceed immediately to computing the spectrum of the corresponding family of transition operators.

We need to analyze three model problems that follow from (10.183): A problem with no lateral boundaries:

\[
\frac{u_{m+1}^{p+1} - u_m^{p-1}}{2\tau} - \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} = 0,
\]

\(m = 0, \pm 1, \pm 2, \ldots,\)

(10.185)

a problem with only the left boundary:

\[
\frac{u_{m+1}^{p+1} - u_m^{p-1}}{2\tau} - \frac{u_{m+1}^{p} - u_{m-1}^{p}}{2h} = 0,
\]

\(m = 1, 2, \ldots,\)

\(l u_0^{p+1} = 0,
\)

(10.186)
and a problem with only the right boundary:

\[
\frac{u_m^{p+1} - u_m^{p-1}}{2\tau} - \frac{u_{m+1}^p - u_{m-1}^p}{2h} = 0, \quad m = M - 1, M - 2, \ldots, 1, 0, -1, \ldots, \\
u_M^{p+1} = 0.
\] (10.187)

Substituting a solution of the type:

\[u_m^p = \lambda^p u_m\]

into the finite-difference equation:

\[u_m^{p+1} - u_m^{p-1} = r(u_{m+1}^p - u_{m-1}^p), \quad r = \tau/h,
\]

that corresponds to all three problems (10.185), (10.186), and (10.187), we obtain the following second-order ordinary difference equation for the eigenfunction \(\{u_m\}\):

\[
(\lambda - \lambda^{-1})u_m - r(u_{m+1} - u_{m-1}) = 0.
\] (10.188)

Its characteristic equation:

\[
(\lambda - \lambda^{-1}) - r(q - q^{-1}) = 0 \quad (10.189a)
\]

has two roots: \(q_1 = q_1(\lambda)\) and \(q_2 = q_2(\lambda)\), so that the general solution of equation (10.188) can be written as

\[u_m = c_1 q_1^m + c_2 q_2^m, \quad m = 0, \pm 1, \pm 2, \ldots, \quad c_1 = \text{const}, \quad c_2 = \text{const}.
\]

It will also be convenient to recast the characteristic equation (10.189a) in an equivalent form:

\[
q^2 - \frac{\lambda - \lambda^{-1}}{r} q - 1 = 0.
\] (10.189b)

From equation (10.189b) one can easily see that \(q_1 q_2 = -1\) and consequently, unless both roots have unit magnitude, we always have \(|q_1(\lambda)| < 1\) and \(|q_2(\lambda)| > 1\).

The solution of problem (10.185) must be bounded: \(|u_m| \leq \text{const} \quad \text{for} \quad m = 0, \pm 1, \pm 2, \ldots\). We therefore require that for this problem \(|q_1| = |q_2| = 1\), which means \(q_1 = e^{i\alpha}, \quad 0 \leq \alpha < 2\pi\), and \(q_2 = -e^{-i\alpha}\). The spectrum of this problem was calculated in Example 5 of Section 10.3.3:

\[
\Lambda = \left\{\lambda(\alpha) = ir \sin \alpha \pm \sqrt{1 - r^2 \sin^2 \alpha} \right\} | 0 \leq \alpha < 2\pi \right\}.
\] (10.190)

Provided that \(r \leq 1\), the spectrum \(\Lambda\) given by formula (10.190) belongs to the unit circle on the complex plane.

For problem (10.187), we must have \(u_m \rightarrow 0\) as \(m \rightarrow -\infty\). Consequently, its general solution is given by:

\[u_m^p = c_2 \lambda^p q_2^m, \quad m = M, M - 1, \ldots, 1, 0, -1, \ldots\]
The homogeneous boundary condition $u_{M}^{p+1} = 0$ of (10.183) implies that a nontrivial eigenfunction $u_{M} = c_{2}q_{2}^{m}$ may only exist if $\lambda = 0$. From the characteristic equation (10.189a) in yet another equivalent form $(\lambda^{2} - 1)q - r\lambda(q^{2} - 1) = 0$, we conclude that if $\lambda = 0$ then $q = 0$, which means that problem (10.187) has no eigenvalues:

$$\Lambda = \emptyset.$$  \hspace{1cm} (10.191)

To study problem (10.186), we first consider boundary condition (10.184a), known as the extrapolation boundary condition. The solution of problem (10.186) must satisfy $u_{m} \rightarrow 0$ as $m \rightarrow \infty$. Consequently, its general form is:

$$u_{m}^{p} = c_{1}\lambda^{p}q_{1}^{m}, \hspace{1cm} m = 0, 1, 2, \ldots.$$  \hspace{1cm} (10.192a)

The extrapolation condition (10.184a) implies that a nontrivial eigenfunction $u_{m} = c_{1}q_{1}^{m}$ may only exist if either $\lambda = 0$ or $c_{1}(1 - q_{1}) = 0$. However, we must have $|q_{1}| < 1$ for problem (10.186), and as such, we see that this problem has no eigenvalues either:

$$\Lambda = \emptyset.$$  \hspace{1cm} (10.192b)

Combining formulae (10.190), (10.191), and (10.192), we obtain the spectrum of the family of operators:

$$\Lambda = \Lambda_{1} \cup \Lambda_{2} \cup \Lambda_{3} = \Lambda.$$  \hspace{1cm} (10.192c)

We therefore see that according to formula (10.190), the necessary condition for stability (Theorem 10.8) of scheme (10.183), (10.184a) is satisfied when $r \leq 1$.

Moreover, according to Theorem 10.9, even if there is no uniform bound on the powers of the transition operators for scheme (10.183), (10.184a), their rate of growth will be slower than any exponential function. Unfortunately, this is precisely what happens. Even though the necessary condition of stability holds for scheme (10.183), (10.184a), it still turns out unstable. The actual proof of instability can be found in [GKO95, Section 13.1] or in [Str04, Section 11.2]. We omit it here and only corroborate the instability with a numerical demonstration. In Figure 10.14 we show the results of numerical integration of problem (10.182) using scheme (10.183), (10.184a) with $r = 0.95$. We specify the exact solution of the problem as $u(x, t) = \cos 2\pi(x + t)$ and as such, supplement problem (10.182) with an inhomogeneous boundary condition $u(1, t) = \cos 2\pi(1 + t)$.

In order to analyze what may have caused the instability of scheme (10.183), (10.184a), let us return to the proof of Theorem 10.9. If we were able to claim that the entire spectrum of the family of operators $\{R_{h}\}$ lies strictly inside the unit disk, then a straightforward modification of that proof would immediately yield a uniform bound on the powers $R_{h}^{p}$. This situation, however, is generally impossible. Indeed, in all our previous examples, the spectrum has always contained at least one point on the unit circle: $\lambda = 1$. It is therefore natural to assume that since the points $\lambda$ inside the unit disk present no danger of instability according to Theorem 10.9, then the potential “culprits” should be sought on the unit circle.

Let us now get back to problem (10.186). We have shown that this problem has no nontrivial eigenfunctions in the class $u_{m} \rightarrow 0$ as $m \rightarrow \infty$ and accordingly, it has no
eigenvalues either, see formula (10.192). As such, it does not contribute to the overall spectrum of the family of operators. However, even though the boundary condition (10.184a) in the form $c_1(1 - q_1) = 0$ is not satisfied by any function $u_m = c_1 q_1^m$, where
Consider a special case of real \(q_1\) such that \(|q_1| < 1\), we see that it is “almost satisfied” if the root \(q_1\) is close to one. Therefore, the function \(u_m = c_1 q_1^m\) is “almost an eigenfunction” of problem (10.186), and the smaller the quantity \(|1 - q_1|\), the more of a genuine eigenfunction it becomes.

To investigate stability, we need to determine whether or not the foregoing “almost an eigenfunction” can bring along an unstable eigenvalue, or rather “almost an eigenvalue,” \(|\lambda| > 1\). By passing to the limit \(q_1 \to 1\), we find from equation (10.189a) that \(\lambda = 1\) or \(\lambda = -1\). We should therefore analyze the behavior of the quantities \(\lambda\) and \(q\) in a neighborhood of each of these two values of \(\lambda\), when the relation between \(\lambda\) and \(q\) is given by equation (10.189a).

First recall that according to formula (10.190), if \(|q| = 1\), then \(|\lambda| = 1\) (provided that \(r \leq 1\)). Consequently, if \(|\lambda| > 1\), then \(|q| \neq 1\), i.e., there are two distinct roots: \(|q_1| < 1\) and \(|q_2| > 1\). In particular, when \(\lambda\) is near the point \((1,0)\) on the complex plane, there are still two roots: One with the magnitude greater than one and the other with the magnitude less than one. When \(|\lambda - 1| \to 0\) we will clearly have \(|q_1| \to 1\) and \(|q_2| \to 1\). We, however, don’t know ahead of time which of the two possible scenarios actually takes place:

\[
\lim_{|\lambda| > 1, \lambda \to 1} q_1(\lambda) = 1, \quad \lim_{|\lambda| > 1, \lambda \to 1} q_2(\lambda) = -1 \quad (10.193a)
\]

or

\[
\lim_{|\lambda| > 1, \lambda \to 1} q_1(\lambda) = -1, \quad \lim_{|\lambda| > 1, \lambda \to 1} q_2(\lambda) = 1. \quad (10.193b)
\]

To find out this, let us notice that the roots \(q_1(\lambda)\) and \(q_2(\lambda)\) are continuous (in fact, analytic) functions of \(\lambda\). Consequently, if we take \(\lambda\) in the form \(\lambda = 1 + \eta\), where \(|\eta| \ll 1\), and if we want to investigate the root \(q\) that is close to one, then we can say that \(q(\lambda) = 1 + \zeta\), where \(|\zeta| \ll 1\). From equation (10.189a) we then obtain:

\[2 \eta + \theta'(\eta^2) = 2 r \zeta + \theta'(\zeta^2).\]  

(10.194)

Consider a special case of real \(\eta > 0\), then \(\zeta\) must obviously be real as well. From the previous equality we find that \(\zeta > 0\) (because \(r > 0\), i.e., \(|q| > 1\). As such, we see that if \(|\lambda| > 1\) and \(\lambda \to 1\), then

\[\{q = q(\lambda) \to 1\} \Rightarrow \{|q| > 1\}.
\]

Indeed, for real \(\eta\) and \(\zeta\), we have \(|q| = 1 + \zeta > 1\); for other \(\eta\) and \(\zeta\) the same result follows by continuity. Consequently, it is the root \(q_2\) that approaches \((1,0)\) when \(\lambda \to 1\), and the true scenario is given by (10.193b) rather than by (10.193a).

We therefore see that when a potentially “dangerous” unstable eigenvalue \(|\lambda| > 1\) approaches the unit circle at \((1,0)\): \(\lambda \to 1\), it is the grid function \(u_m = c_2 q_2^m\), \(|q_2| > 1\), that will almost satisfy the boundary condition (10.184a), because \(c_2(1 - q_2) \to 0\). This grid function, however, does not satisfy the requirement \(u_m \to 0\) as \(m \to \infty\), i.e., it does not belong to the class of functions admitted by problem (10.186). On the other hand, the function \(u_m = c_1 q_1^m\), \(|q_1| > 1\), that satisfies \(u_m \to 0\) as \(m \to \infty\), will be very far from satisfying the boundary condition (10.184a) because \(q_1 \to -1\).
Next, recall that we actually need to investigate what happens when \(q_1 \to 1\), i.e., when \(c_1 q_1^n\) is almost an eigenfunction. This situation appears opposite to the one we have analyzed. Consequently, when \(q_1 \to 1\) we will not have such a \(\lambda(q_1) \to 1\) where \(|\lambda(q_1)| > 1\). Qualitatively, this indicates that there is no instability associated with “almost an eigenfunction” \(u_m = c_1 q_1^n\), \(|q_1| > 1\), of problem (10.186). In the framework of the GKS theory, this assertion can be proven rigorously.

Let us now consider the second case: \(\lambda \to -1\) while \(|\lambda| > 1\). We need to determine which of the two scenarios holds:

\[
\lim_{{|\lambda| > 1, \lambda \to -1}} q_1(\lambda) = 1, \quad \lim_{{|\lambda| > 1, \lambda \to -1}} q_2(\lambda) = -1 \quad (10.195a)
\]

or

\[
\lim_{{|\lambda| > 1, \lambda \to -1}} q_1(\lambda) = -1, \quad \lim_{{|\lambda| > 1, \lambda \to -1}} q_2(\lambda) = 1. \quad (10.195b)
\]

Similarly to the previous analysis, let \(\lambda = -1 + \eta\), where \(|\eta| \ll 1\), then also \(q(\lambda) = 1 + \zeta\), where \(|\zeta| \ll 1\) (recall, we are still interested in \(q \to 1\)). Consider a particular case of real \(\eta < 0\), then equation (10.194) yields \(\zeta < 0\), i.e., \(|q| < 1\). Consequently, if \(|\lambda| > 1\) and \(\lambda \to -1\), then

\[
\{ q = q(\lambda) \to 1 \} \implies \{ |q| < 1 \}.
\]

In other words, this time it is the root \(q_1\) that approaches \((1, 0)\) as \(\lambda \to -1\), and the scenario that gets realized is (10.195a) rather than (10.195b). In contradistinction to the previous case, this presents a potential for instability. Indeed, the pair \((\lambda, q_1)\), where \(|q_1| < 1\) and \(|\lambda| > 1\), would have implied the instability in the sense of Section 10.5.2 if \(c_1 q_1^n\) were a genuine eigenfunction of problem (10.186) and \(\lambda\) if were the corresponding genuine eigenvalue. As we know, this is not the case. However, according to the first formula of (10.195a), the actual setup appears to be a limit of the admissible yet unstable situation. In other words, the combination of “almost an eigenfunction” \(u_m = c_1 q_1^n\), \(|q_1| < 1\), that satisfies \(u_m \to 0\) as \(m \to \infty\) with “almost an eigenvalue” \(\lambda = \lambda(q_1)\), \(|\lambda| > 1\), is unstable. While remaining unstable, this combination becomes more of a genuine eigenpair of problem (10.186) as \(\lambda \to -1\). Again, a rigorous proof of the instability is given in the framework of the GKS theory using the technique based on the Laplace transform.

Thus, we have seen that two scenarios are possible when \(\lambda\) approaches the unit circle from the outside. In one case, there may be an admissible root \(q\) of the characteristic equation that almost satisfies the boundary condition, see formula (10.195a), and this situation is prone to instability. Otherwise, see formula (10.193b), there is no admissible root \(q\) that would ultimately satisfy the boundary condition, and as such, no instability will be associated with this \(\lambda\).

In the unstable case exemplified by formula (10.195a), the corresponding limit value of \(\lambda\) is called the \textit{generalized eigenvalue}, see [GKO95, Chapter 13]. In particular, \(\lambda = -1\) is a generalized eigenvalue of problem (10.186). We re-emphasize that it is not a genuine eigenvalue of problem (10.186), because when \(\lambda = -1\) then \(q_1 = 1\).
and the eigenfunction $u_m = cq_1^m$ does not belong to the admissible class: $u_m \to 0$ as $m \to \infty$. In fact, it is easy to see that $\|u\|_2 = \infty$. However, it is precisely the generalized eigenvalues that cause the instability of a scheme, even in the case when the entire spectrum of the family of operators $\{R_h\}$ belongs to the unit disk.

Accordingly, the Kreiss necessary and sufficient condition of stability requires that the spectrum of the family of operators be confined to the unit disk as before, and additionally, that the scheme have no generalized eigenvalues $|\lambda| = 1$. Scheme (10.183), (10.184a) violates the Kreiss condition at $\lambda = -1$ and as such, is unstable. The instability manifests itself via the computations presented in Figure 10.14.

As, however, this instability is only due to a generalized eigenvalue with $|\lambda| = 1$, it is relatively mild, as expected. On the other hand, if we were to replace the marginally unstable boundary condition (10.184a) with a truly unstable one in the sense of Section 10.5.2, then the effect on the stability of the scheme would have been much more drastic. Instead of (10.184a), consider, for example:

$$u_0^{p+1} = 1.05 \cdot u_1^{p+1}.$$ (10.196)

This boundary condition generates an eigenfunction $u_m = c_1q_1^m$ of problem (10.186) with $q_1 = \frac{1}{10} < 1$. The corresponding eigenvalues are given by:

$$\lambda(q_1) = \frac{r}{2} \left( q_1 - \frac{1}{q_1} \right) \pm \sqrt{1 + \left( \frac{r^2}{4} \left( q_1 - \frac{1}{q_1} \right)^2 \right)},$$

and for one of these eigenvalues we obviously have $|\lambda| > 1$. Therefore, the scheme is unstable according to Theorem 10.8. In Figure 10.15, we are showing the results of the numerical solution of problem (10.182) using the unstable scheme (10.183), (10.196). Comparing the plots in Figure 10.15 with those in Figure 10.14, we see that in the case of boundary condition (10.196) the instability develops much more rapidly in time. Moreover, comparing the left column in Figure 10.15 that corresponds to the grid with $M = 100$ cells with the right column in the same figure that corresponds to $M = 200$, we see that the instability develops more rapidly on a finer grid, which is characteristic of an exponential instability.

Let now analyze the second outflow boundary condition (10.184b):

$$u_0^{p+1} = u_0^p + r(u_1^p - u_0^p).$$

Unlike the extrapolation-type boundary condition (10.184a), it is obtained as a one-sided finite-difference approximation of the differential equation itself. To study stability, we again need to investigate three model problems: (10.185), (10.186), and (10.187). Obviously, only problem (10.186) changes due to the new boundary condition, where the other two stay the same. To find the corresponding $\lambda$ and $q$ we need to solve the characteristic equation (10.189a) along with a similar equation that stems from the boundary condition (10.184b):

$$\lambda = 1 - r + rq.$$ (10.197)
Consider first the case $r = 1$, then from equation (10.189a) we find that $\lambda = q$ or $\lambda = -1/q$, and equation (10.197) says that $\lambda = q$. Consequently, problem (10.186) will not have a solution of the type $u_m^p = c_1 \lambda^p q_1^m$, $m = 0, 1, 2, \ldots$, where $|\lambda| > 1$ and $|q_1| < 1$, i.e., will have no eigenfunctions/eigenvalues in the class $u_m \rightarrow 0$ as $m \rightarrow \infty$. Moreover, as $|\lambda| > 1$ is always accompanied by $|q| > 1$, we conclude that there are no generalized eigenvalues with $|\lambda| = 1$ either.

In the case $r < 1$, substituting $\lambda$ from equation (10.197) into equation (10.189a) and subsequently solving for $q$, we find that there is only one solution: $q = 1$. For the corresponding $\lambda$, we then have from equation (10.197): $\lambda = 1$. Consequently, for $r < 1$, problem (10.186) also has no proper eigenfunctions/eigenvalues, which means that we again have $\Lambda = \emptyset$ [cf. formula (10.192)]. As far as the generalized eigenvalues, we only need to check one value of $\lambda$: $\lambda = 1$ (because $\lambda = -1$ does not
satisfy equation (10.197) for \( q = 1 \). Let \( \lambda = 1 + \eta \) and \( q = 1 + \zeta \), where \( |\eta| \ll 1 \) and \( |\zeta| \ll 1 \). We then arrive at the same equation (10.194) that we obtained in the context of the previous analysis and conclude that \( \lambda = 1 \) does not violate the Kreiss condition, because \( |\lambda| > 1 \) implies \( |q| > 1 \). As such, the scheme (10.183), (10.184b) is stable.

**Exercises**

1. For the scalar Lax-Wendroff scheme [cf. formula (10.83)]:

\[
\frac{u_m^{p+1} - u_m^p}{\tau} = \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \quad p = 0, 1, \ldots, |T/\tau| - 1, \quad m = 1, 2, \ldots, M - 1, \quad Mh = 1, \quad u_m^0 = \psi(x_m), \quad m = 0, 1, 2, \ldots, M,
\]

that approximates the initial boundary value problem:

\[
\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}, \quad 0 \leq x \leq 1, \quad 0 < t \leq T,
\]

\[
u(x,0) = \psi(x), \quad u(1,t) = 0,
\]

on a uniform rectangular grid: \( x_m = mh, \ m = 0, 1, \ldots, M, \) \( Mh = 1, \ t_p = p\tau, \ p = 0, 1, \ldots, |T/\tau|, \) find out when the Babenko-Gelfand stability criterion holds.

**Answer.** \( r = \tau/h \leq 1 \).

2. Consider an initial boundary value problem for the acoustics system of equations [cf. formula (10.89a)]:

\[
\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x}, \quad \frac{\partial w}{\partial t} = \frac{\partial v}{\partial x},
\]

\[
0 \leq x \leq 1, \quad 0 < t \leq T, \quad v(x,0) = \psi^{(1)}(x), \quad w(x,0) = \psi^{(2)}(x), \quad v(0,t) = 0, \quad w(1,t) = 0.
\]

To build a finite-difference scheme for it, we first set: \( u(x,t) = \begin{bmatrix} v(x,t) \\ w(x,t) \end{bmatrix} \), \( \psi(x) = \begin{bmatrix} \psi^{(1)}(x) \\ \psi^{(2)}(x) \end{bmatrix} \) and recast the system in the matrix form [cf. formula (10.89b)]:

\[
\frac{\partial u}{\partial t} - A \frac{\partial u}{\partial x} = 0, \quad 0 \leq x \leq 1, \quad 0 < t \leq T,
\]

\[
u(x,0) = \psi(x), \quad 0 \leq x \leq 1, \quad L_1 u(0,t) = 0, \quad L_2 u(1,t) = 0, \quad 0 < t \leq T,
\]
where \( A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \), \( l_1 = [1 0] \), and \( l_2 = [0 1] \). Next, we construct a vector Lax-Wendroff scheme on the same uniform Cartesian grid as used in Exercise 1 [cf. formula (10.93)]:

\[
\begin{align*}
\frac{u_m^{p+1} - u_m^p}{\tau} - A \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2} A^2 \frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{h^2} &= 0, \\
p = 0, 1, \ldots, [T/\tau] - 1, \quad m = 1, 2, \ldots, M - 1, \quad Mh = 1,
\end{align*}
\]

\[
u_0^p = \psi(x_m), \quad m = 0, 1, 2, \ldots, M,
\]

\[
l_1 u_0^{p+1} = 0, \quad l_2 u_M^{p+1} = 0, \quad p = 0, 1, \ldots, [T/\tau] - 1.
\]

To complete the construction of the scheme, we need additional boundary conditions at the left and right endpoints of the interval. To obtain those, we first notice that for any \( \alpha \) and \( \beta \) the original acoustics system implies that the following equalities hold:

\[
\begin{align*}
\frac{\partial (v + \alpha w)}{\partial t} - \frac{\partial (w + \alpha v)}{\partial x} &\bigg|_{x=0} = 0, \\
\frac{\partial (v + \beta w)}{\partial t} - \frac{\partial (w + \beta v)}{\partial x} &\bigg|_{x=1} = 0.
\end{align*}
\]

Then, we set the required additional boundary conditions as follows:

\[
\begin{align*}
\frac{(v_0^{p+1} + \alpha w_0^{p+1}) - (v_0^p + \alpha w_0^p)}{\tau} - \frac{(w_0^{p+1} + \alpha v_0^{p+1}) - (w_0^p + \alpha v_0^p)}{\tau} &= 0, \\
\frac{(v_M^{p+1} + \beta w_M^{p+1}) - (v_M^p + \beta w_M^p)}{\tau} - \frac{(w_M^{p+1} + \beta v_M^{p+1}) - (w_M^p + \beta v_M^p)}{\tau} &= 0.
\end{align*}
\]

Assuming that \( r = \tau/h \leq 1 \), show that:

a) If \( \alpha = 1, \beta = -1 \), then the Babenko-Gelfand stability criterion holds.

b) If \( \alpha = -1 \) then the Babenko-Gelfand stability criterion does not hold irrespective of the specific value of \( \beta \).

c) Find the conditions on \( \alpha \) and \( \beta \) that would guarantee fulfillment of the Babenko-Gelfand criterion.

3.* Prove Theorem 10.6.

a) Prove the sufficiency part.

b) Prove the necessity part.

4.* Approximate the acoustics Cauchy problem:

\[
\begin{align*}
\frac{\partial u}{\partial t} - A \frac{\partial u}{\partial x} &= \varphi(x,t), \quad -\infty \leq x \leq \infty, \quad 0 < t \leq T, \\
u(x,0) &= \psi(x), \quad -\infty \leq x \leq \infty,
\end{align*}
\]

\[
u(x,t) = \begin{bmatrix} v(x,t) \\ w(x,t) \end{bmatrix}, \quad \varphi(x) = \begin{bmatrix} \varphi^{(1)}(x) \\ \varphi^{(2)}(x) \end{bmatrix}, \quad \psi(x) = \begin{bmatrix} \psi^{(1)}(x) \\ \psi^{(2)}(x) \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.
\]
with the Lax-Wendroff scheme:
\[
\frac{u_m^{p+1} - u_m^p}{\tau} - A\frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2} A \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi_m^p, \\
p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1, \quad m = 0, \pm 1, \pm 2, \ldots,
\]
\[
u_m^p = \psi(x_m), \quad m = 0, \pm 1, \pm 2, \ldots.
\]
Set \( u^p = \{ u_m^p \} \) and \( \varphi^p = \{ \varphi_m^p \} \), and define the norms as follows:
\[
\| u^h \|_{U_h} = \max_p \| u^p \|, \quad \| \varphi^h \|_{F_h} = \max \left[ \| \psi \|, \max_p \| \varphi^p \| \right],
\]
where
\[
\| u^p \|^2 = \sum_m \left( |v_m^p|^2 + |w_m^p|^2 \right), \quad \| \psi \|^2 = \sum_m \left( |\psi^{(1)}(x_m)|^2 + |\psi^{(2)}(x_m)|^2 \right), \quad \| \varphi^p \|^2 = \sum_m \left( |\varphi^{(1)}(x_m, t_p)|^2 + |\varphi^{(2)}(x_m, t_p)|^2 \right).
\]

a) Show that when reducing the Lax-Wendroff scheme to the canonical form (10.140), inequalities (10.142) and (10.143) hold.

b) Prove that when \( r = \frac{\tau}{h} \leq 1 \) the scheme is \( l_2 \) stable, and when \( r > 1 \) it is unstable.

**Hint.** To prove estimate (10.144) for the norms \( \| R_n^p \| \), first introduce the new unknown variables (called the Riemann invariants):
\[
i_m^{(1)} = v_m + w_m \quad \text{and} \quad i_m^{(2)} = v_m - w_m,
\]
and transform the discrete system accordingly, and then employ the spectral criterion of Section 10.3.

5. Prove that all complex numbers \( \lambda(\alpha) = 1 - r + re^{i\alpha}, \quad 0 \leq \alpha < 2\pi \) [see formula (10.147)], belong to the spectrum of the transition operator \( R_h \) that corresponds to the difference Cauchy problem (10.146) if the norm in the space \( U_h \) is defined in the sense of \( l_2 \): \( \| u \|_2 = \left[ \sum_{m=-\infty}^{\infty} |u_m|^2 \right]^{1/2} \) and the spectrum is defined according to Definition 10.7.

**Hint.** Construct the solution \( u = \{ u_m \}, m = 0, \pm 1, \pm 2, \ldots \), to the inequality that appears in Definition 10.7 in the form: \( u_m = \begin{cases} q_m^+, & m \geq 0, \\ q_m^-, & m < 0, \end{cases} \) where \( q = (1 - \delta)e^{i\alpha} \), and \( \delta > 0 \) is a small quantity.


**Hint.** Use expansion with respect to an orthonormal basis in \( U \) composed of the eigenvectors of \( R_h \).

7. Compute the spectrum of the family of operators \( \{ R_h \}, v = R_h u \), given by the formulae [cf. formulae (10.141a) & (10.141b)]:
\[
v_m = (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M - 1, \quad v_M = 0.
\]
8. Prove that the spectrum of the family of operators \( \{ R_h \} \), \( v = R_h u \), given by the formulae [cf. formulae (10.141a) & (10.141b)]:

\[
\begin{align*}
v_m &= (1 - r + \gamma h)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M - 1, \\
v_M &= u_M,
\end{align*}
\]

does not depend on the value of \( \gamma \) and coincides with the spectrum computed in Section 10.5.2 for the case \( \gamma = 0 \).

**Hint.** Notice that this operator is obtained by adding \( \gamma h I \) to the operator \( R_h \) defined by formulae (10.141a) & (10.141b), and then use Definition 10.6 directly.

9. \( \star \) Compute the spectrum of the family of operators \( \{ R_h \} \), \( v = R_h u \), given by the formulae:

\[
\begin{align*}
v_m &= (1 - r)u_m + r(u_{m-1} + u_{m+1})/2, \quad m = 1, 2, \ldots, M - 1, \\
v_M &= 0, \quad av_0 + bv_1 = 0,
\end{align*}
\]

where \( a \) and \( b \) are two known fixed numbers. Consider the cases \(|a| > |b|\) and \(|a| < |b|\).

10. Prove that the spectrum of the family of operators \( \{ R_h \} \), \( v = R_h u \), defined by formulae (10.141a) & (10.141b) and analyzed in Section 10.5.2:

\[
\begin{align*}
v_m &= (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M - 1, \\
v_M &= u_M,
\end{align*}
\]

will not change if the \( C \) norm: \( \| u \| = \max_m |u_m| \) is replaced by the \( l_2 \) norm: \( \| u \| = \left( \sum_m u_m^2 \right)^{1/2} \).

11. For the first-order ordinary difference equation:

\[
av_m + bv_{m+1} = f_m, \quad m = 0, \pm 1, \pm 2, \ldots,
\]

the fundamental solution \( G_m \) is defined as a bounded solution of the equation:

\[
aG_m + bG_{m+1} = \delta_m \equiv \begin{cases} 1, & m = 0, \\ 0, & m \neq 0. \end{cases}
\]

a) Prove that if \(|a/b| < 1\), then \( G_m \) equals \( 0 \), \( \frac{1}{2} (-\frac{a}{b})^m \), \( m \leq 0 \), \( \frac{1}{2} \), \( (\frac{a}{b})^m \), \( m \geq 1 \).

b) Prove that if \(|a/b| > 1\), then \( G_m \) equals \( \frac{1}{2} \), \( (\frac{a}{b})^m \), \( m \leq 0 \), \( 0 \), \( m \geq 1 \).

c) Prove that \( v_m = \sum_{k=-\infty}^{\infty} G_{m-k} f_k \).

12. Obtain energy estimates for the implicit first-order upwind schemes that approximate problems (10.169), (10.171), and (10.173).
13.* Approximate problem (10.169) with the Crank-Nicolson scheme supplemented by one-sided differences at the left boundary \( x = 0 \):

\[
\frac{u^{p+1}_m - u^p_m}{\tau} = \frac{1}{2} \left[ \frac{u^{p+1}_{m+1} - u^p_{m+1}}{2h} + \frac{u^p_{m+1} - u^{p+1}_m}{2h} \right] = 0,
\]

\( m = 1, 2, \ldots, M - 1, \quad p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1, \)

\[
\frac{u^p_0 - u^p_0}{\tau} = \frac{1}{2} \left[ \frac{u^{p+1}_1 - u^p_1}{h} + \frac{u^p_1 - u^{p+1}_0}{h} \right] = 0, \quad u^p_M = 0, \quad (10.198)
\]

\( p = 0, 1, \ldots, \lfloor T/\tau \rfloor - 1, \)

\( u_0^0 = \psi_m, \quad m = 0, 1, 2, \ldots, M. \)

a) Use an alternative definition of the \( L^2 \) norm: \( \| u \|^2 = \frac{h}{2}(u_0^2 + u_M^2) + h \sum_{m=1}^{M-1} u_m^2 \) and develop the energy estimate for scheme (10.198).

**Hint.** Multiply the equation by \( u^{p+1}_m + u^p_m \) and sum over the entire range of \( m \).

b) Construct the schemes similar to (10.198) for the variable-coefficient problems (10.171) and (10.173) and obtain the energy estimates.

14. Using the Kreiss condition, show that the leap-frog scheme with the boundary condition:

\[
u^p_0 = u^1_1
\]

is stable, but with the boundary condition:

\[
u^p_0 = u^{p-1}_1 + 2r(u^p_1 - u^0_0)
\]

it is unstable.

15. Reproduce on the computer the results shown in Figures 10.14 and 10.15. In addition, conduct the computations using the leap-frog scheme with the boundary conditions (10.184b), (10.199a), and (10.199b), and demonstrate experimentally the stability and instability in the respective cases.

16.* Using the Kreiss condition, investigate stability of the Crank-Nicolson scheme applied to solving problem (10.182) and supplemented either with the boundary condition (10.184b) or with the boundary condition (10.199a).

10.6 Explicit and Implicit Schemes for the Heat Equation

Consider the following initial boundary value problem for a variable-coefficient homogeneous heat equation:

\[
\frac{\partial^2 u}{\partial t^2} - a(x,t) \frac{\partial^2 u}{\partial x^2} = 0, \quad a(x,t) > 0, \quad 0 \leq x \leq 1,
\]

\[
u(x,0) = \psi(x),
\]

\[
u(0,t) = \phi(t), \quad \nu(1,t) = \chi(t).
\]

\[
(10.200)
\]
To solve problem (10.200) numerically, we can use either an explicit or an implicit finite-difference scheme. We will analyze and compare both schemes. In doing so, we will see that quite often the implicit scheme has certain advantages compared to the explicit scheme, even though the algorithm of computing the solution with the help of an explicit scheme is simpler than that for the implicit scheme. The advantages of using an implicit scheme stem from its unconditional stability, i.e., stability that holds for any ratio between the spatial and temporal grid sizes.

10.6.1 An Explicit Scheme

We introduce a uniform grid on the interval $[0, 1]$:

$$x_m = mh, \quad m = 0, 1, \ldots, M, \quad Mh = 1,$$

and build the scheme on the four-node stencil shown in Figure 10.3 (see page 183):

$$\frac{u_m^{p+1} - u_m^p}{\tau} - a(x_m, t_p) \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0,$$

$$m = 1, 2, \ldots, M - 1,$$

$$u_0^0 = \psi(x_0), \quad u_M^0 = \chi(x_M), \quad p = 0, \quad t_0 = 0,$$

$$u_m^1 = \phi(t_{p+1}), \quad u_M^1 = \chi(t_{p+1}), \quad p \geq 0,$$

$$t_p = t_0 + \tau_1 + \ldots + \tau_{p-1}, \quad p = 1, 2, \ldots.$$

(10.201)

If the solution $u_m^k, \quad m = 0, 1, \ldots, M,$ is already known for $k = 0, 1, \ldots, p$, then, according to (10.201), the values of $u_m^{p+1}$ at the next time level $t = t_{p+1} = t_p + \tau$ can be computed with the help of an explicit formula:

$$u_m^{p+1} = u_m^p + \frac{\tau}{h^2} a(x_m, t_p) (u_{m+1}^p - 2u_m^p + u_{m-1}^p),$$

$$m = 1, 2, \ldots, M - 1.$$

(10.202)

This explains why the scheme (10.201) is called explicit. Formula (10.202) allows one to easily march the discrete solution $u_m^p$ from one time level to another.

According to the principle of frozen coefficients (see Sections 10.4.1 and 10.5.1), when marching the solution by formula (10.202), one may only expect stability if

$$\tau_p \leq \frac{h^2}{2 \max_m a(x_m, t_p)}.$$

(10.203)

Consequently, if the heat conduction coefficient $a(x_m, t_p)$ assumes large values in a neighborhood of some point $(\bar{x}, \bar{t})$, then computing the solution on the time level $t = t_{p+1}$ will necessitate taking a very small time step $\tau = \tau_p$. Therefore, marching the solution all the way through some prescribed value of $t = T$ may require an excessively large number of time steps, which will make the computation impractical.

Let us note that the foregoing restriction on the time step is of a purely numerical nature and has nothing to do with the physics behind problem (10.200). Indeed, this
problem models the propagation of heat in a spatially one-dimensional structure, e.g., a rod, for which the heat conduction coefficient \( a = a(x, t) \) may vary along the rod and also in time. Large values of \( a(x, t) \) in a neighborhood of some point \((\tilde{x}, \tilde{t})\) merely imply that this neighborhood can be removed, i.e., “cut off,” from the rod without changing the overall pattern of heat propagation. In other words, we may think that this part of the rod consists of a material with zero heat capacity.

10.6.2 An Implicit Scheme

Instead of scheme (10.201), we can use the same grid and build the scheme on the stencil shown in Figure 10.3 (right) (see page 183):

\[
\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau_p} - a(x_m, t_p) \frac{u_{m+1}^{p+1} - 2u_{m}^{p+1} + u_{m-1}^{p+1}}{h^2} = 0, \quad m = 1, 2, \ldots, M - 1, \\
u_{0}^{0} = \psi_{0}, \quad m = 0, 1, \ldots, M, \\
u_{0}^{p+1} = \phi(t_{p+1}), \quad u_{M}^{p+1} = \chi(t_{p+1}), \quad p \geq 0, \\
t_0 = 0, \quad t_p = t_0 + \tau_1 + \ldots + \tau_{p-1}, \quad p = 1, 2, \ldots.
\]

(10.204)

Assume that the solution \( u_{m}^{p}, m = 0, 1, \ldots, M, \) at the time level \( t = t_p \) is already known. According to formula (10.204), in order to compute the values of \( u_{m}^{p+1}, m = 0, 1, \ldots, M, \) at the next time level \( t = t_{p+1} = t_p + \tau_p \) we need to solve the following system of linear algebraic equations with respect to \( u_{m}^{p+1} = u_{m}^{p+1} \):

\[
u_0 = \phi(t_{p+1}),
\]

\[
\alpha_m u_{m-1}^{p+1} + \beta_m u_{m}^{p+1} + \gamma_m u_{m+1}^{p+1} = f_m, \quad m = 1, 2, \ldots, M - 1, \\
u_M = \chi(t_{p+1}),
\]

(10.205)

where

\[
\alpha_m = \gamma_m = - \frac{\tau_p}{h^2} a(x_m, t_p), \quad \beta_m = 1 + 2 \frac{\tau_p}{h^2} a(x_m, t_p), \quad f_m = u_{m}^{p}, \\
m = 1, 2, \ldots, M - 1, \\
\gamma_0 = \alpha_M = 0, \quad \beta_0 = \beta_M = 1.
\]

It is therefore clear that

\[
|\beta_m| = |\alpha_m| + |\gamma_m| + \delta, \quad \delta = 1 > 0, \\
m = 0, 1, 2, \ldots, M,
\]

and because of the diagonal dominance, system (10.205) can be solved by the algorithm of tri-diagonal elimination described in Section 5.3. Note that in the case of scheme (10.204) there are no explicit formulae, i.e., closed form expressions such as formula (10.202), that would allow one to obtain the solution \( u_{m}^{p+1} \) at the upper
time level given the solution \( u_m^p \) at the lower time level. Instead, when marching the solution in time one has to repeatedly solve systems of type (10.205), and that is why the scheme (10.204) is called \textit{implicit}.

In Section 10.3.3 (see Example 7), we analyzed an implicit finite-difference scheme for the constant-coefficient heat equation and demonstrated that the von Neumann spectral stability condition holds for this scheme for any ratio \( \tau / h^2 = r \) between the temporal and spatial grid sizes. By virtue of the principle of frozen coefficients (see Section 10.4.1), in the case of a variable heat conduction coefficient \( a(x,t) \) the spectral stability condition will not impose any constraints on the time step \( \tau \) either. Therefore, the implicit scheme (10.204) can be used efficiently even when the coefficient \( a(x,t) \) assumes large values for some \((\tilde{x}, \tilde{t})\). For convenience, when computing the solution of problem (10.200) with the help of the scheme (10.204), one can set a constant, rather than variable, time step \( \tau_p = \tau \).

To conclude this section, let us note that both the stability condition (10.203) for the explicit scheme (10.201) and the unconditional stability of the implicit scheme (10.204) have been established heuristically rather than rigorously. Indeed, we have used the argument based on the principle of frozen coefficients that is heuristic in nature itself. However, the conclusions that we have drawn regarding the properties and applicability of the schemes (10.201) and (10.204) are actually true, and can be proven accurately.

**Exercise**

1. Let the heat conduction coefficient in problem (10.200) be defined as \( a = 1 + u^2 \), so that problem (10.200) becomes nonlinear.
   a) Introduce an explicit scheme and an implicit scheme for this new problem.
   b) Consider the following explicit scheme:

   \[
   \frac{u_m^{p+1} - u_m^p}{\tau_p} = \left[ 1 + \left( u_m^{p+1} \right)^2 \right] \frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{h^2} = 0, \quad m = 1, 2, \ldots, M - 1,
   \]

   \[
   u_m^0 = \psi(x_m) \equiv \psi_m, \quad m = 0, 1, \ldots, M,
   \]

   \[
   u_0^{p+1} = \phi(t_{p+1}), \quad u_M^{p+1} = \chi(t_{p+1}), \quad p \geq 0,
   \]

   \[
   t_0 = 0, \quad t_p = t_0 + \tau_1 + \cdots + \tau_{p-1}, \quad p = 1, 2, \ldots.
   \]

   How should one choose \( \tau_p \), given the values of the solution \( u_m^p \) at the level \( p \)?

   c) Consider an implicit scheme based on the following finite-difference equation:

   \[
   \frac{u_m^{p+1} - u_m^p}{\tau_p} - \left[ 1 + \left( u_m^{p+1} \right)^2 \right] \frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{h^2} = 0.
   \]

   How should one modify this equation in order to employ the tri-diagonal elimination of Section 5.3 for the transition from \( u_m^p, \ m = 0, 1, \ldots, M \) to \( u_m^{p+1}, \ m = 0, 1, \ldots, M \)?
Chapter 11

Discontinuous Solutions and Methods of Their Computation

A number of frequently encountered partial differential equations, such as the equations of hydrodynamics, elasticity, diffusion, and others, are in fact derived from the conservation laws of certain physical quantities (e.g., mass, momentum, energy, etc.). The conservation laws are typically written in an integral form, and the foregoing differential equations are only equivalent to the corresponding integral relations if the unknown physical fields (i.e., the unknown functions) are sufficiently smooth.

In every example considered previously (Chapters 9 and 10) we assumed that the differential initial (boundary) value problems had regular solutions. Accordingly, our key approach to building finite-difference schemes was based on approximating the derivatives using difference quotients. Subsequently, we could establish consistency with the help of Taylor expansions. However, the class of differentiable functions often appears too narrow for adequately describing many important physical phenomena and processes. For example, physical experiments demonstrate that the fields of density, pressure, and velocity in a supersonic flow of inviscid gas are described by functions with discontinuities (known as either shocks or contact discontinuities in the context of gas dynamics). It is important to emphasize that shocks may appear in the solution as time elapses, even in the case of smooth initial data.

From the standpoint of physical conservation laws, their integral form typically makes sense for both continuous and discontinuous solutions, because the corresponding discontinuous functions can typically be integrated. However, in the discontinuous case one can no longer obtain an equivalent differential formulation of the problem, because discontinuous functions cannot be differentiated. As a consequence, one can no longer enjoy the convenience of studying the properties of solutions to the differential equations, and can no longer use those for building finite-difference schemes. Therefore, prior to constructing the algorithms for computing discontinuous solutions of integral conservation laws, it will be very instrumental to generalize the concept of a solution to a differential initial (boundary) value problem. The objective is to make it meaningful and equivalent to the original integral conservation law even in the case of discontinuous solutions.

We will use a simple example below to illustrate the transition from the problem formulated in terms of an integral conservation law to an equivalent differential problem. The latter will only have a generalized (weak) solution, which is discontinuous. We will also demonstrate some approaches to computing this solution.

Assume that in the region (strip) $0 \leq t \leq T$ we need to find the function $u = u(x,t)$
that satisfies the integral equation:

\[
\int_{\Gamma} \frac{u^k}{k} dx - \frac{u^{k+1}}{k+1} dt = 0 
\]  

(11.1a)

for an arbitrary closed contour \( \Gamma \). The quantity \( k \) in formula (11.1a) is a fixed positive integer. We also require that \( u = u(x,t) \) satisfies the initial condition:

\[
u(x,0) = \psi(x), \quad -\infty < x < \infty. 
\]  

(11.1b)

The left-hand side of equation (11.1a) can be interpreted as the flux of the vector field:

\[ \phi(x,t) = \left[ \frac{u^k}{k} \right]_0^{k+1} = \frac{u^k}{k} + \frac{1}{k+1} \]

through the contour \( \Gamma \). The requirement that the flux of this vector field through an arbitrary contour \( \Gamma \) be equal to zero can be thought of as a conservation law written in an integral form.

Problem (11.1a), (11.1b) provides the simplest formulation that leads to the formation of discontinuities albeit smooth initial data. It can serve as a model for understanding the methods of solving similar problems in the context of fluid dynamics.

### 11.1 Differential Form of an Integral Conservation Law

#### 11.1.1 Differential Equation in the Case of Smooth Solutions

Let us first assume that the solution \( u = u(x,t) \) to problem (11.1a), (11.1b) is continuously differentiable everywhere on the strip \( 0 \leq t \leq T \). We will then show that problem (11.1a), (11.1b) is equivalent to the following Cauchy problem:

\[
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0, \quad 0 < t < T, \quad -\infty < x < \infty, 
\]

\[
u(x,0) = \psi(x), \quad -\infty < x < \infty. 
\]  

(11.2)

In the literature, the differential equation of (11.2) is known as the Burgers equation.

To establish the equivalence of problem (11.1a), (11.1b) and problem (11.2), we recall Green’s formula. Let \( \Omega \) be an arbitrary domain on the \( (x,t) \) plane, let \( \Gamma = \partial \Omega \) be its boundary, and let the functions \( \phi_1(x,t) \) and \( \phi_2(x,t) \) have partial derivatives with respect to \( x \) and \( t \) on the domain \( \Omega \) that are continuous everywhere up to the boundary \( \Gamma \). Then, the following Green’s formula holds:

\[
\iint_{\Omega} \left( \frac{\partial \phi_1}{\partial t} + \frac{\partial \phi_2}{\partial x} \right) dx dt = \int_{\Gamma} \phi_1 dx - \phi_2 dt. 
\]  

(11.3)
Identity (11.3) means that the integral of the divergence \( \frac{\partial \phi_1}{\partial t} + \frac{\partial \phi_2}{\partial x} \) of the vector field \( \phi = [\phi_1, \phi_2]^T \) over the domain \( \Omega \) is equal to the flux of this vector field through the boundary \( \Gamma = \partial \Omega \).

Using formula (11.3), we can write:

\[
\int_{\Gamma} \frac{u^k}{k} dx - \frac{u^{k+1}}{k+1} dt = \int_{\Omega} \left[ \frac{\partial}{\partial t} \left( \frac{u^k}{k} \right) + \frac{\partial}{\partial x} \left( \frac{u^{k+1}}{k+1} \right) \right] dx dt. \tag{11.4}
\]

Equality (11.4) implies that if a smooth function \( u = u(x,t) \) satisfies the Burgers equation, see formula (11.2), then equation (11.1a) also holds. Indeed, if the Burgers equation is satisfied, then we also have:

\[
u^{k-1} \left( \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} \right) = \frac{\partial}{\partial t} \left( \frac{u^k}{k} \right) + \frac{\partial}{\partial x} \left( \frac{u^{k+1}}{k+1} \right) = 0. \tag{11.5}
\]

Consequently, the right-hand side of equality (11.4) becomes zero. The converse is also true: If a smooth function \( u = u(x,t) \) satisfies the integral conservation law (11.1a), then at every point \((\tilde{x}, \tilde{t})\) of the strip \(0 < t < T\) equation (11.5) holds, and hence equation (11.2) is true as well. To justify that, let us assume the opposite, and let us, for definiteness, take some point \((\tilde{x}, \tilde{t})\) for which:

\[
\left. \frac{\partial}{\partial t} \left( \frac{u^k}{k} \right) + \frac{\partial}{\partial x} \left( \frac{u^{k+1}}{k+1} \right) \right| _{(\tilde{x}, \tilde{t})} > 0.
\]

Then, by continuity, we can always find a sufficiently small disk \( \Omega \subset \{(x,t)|0 < t < T\} \) centered at \((\tilde{x}, \tilde{t})\) such that

\[
\left. \frac{\partial}{\partial t} \left( \frac{u^k}{k} \right) + \frac{\partial}{\partial x} \left( \frac{u^{k+1}}{k+1} \right) \right| _{(x,t) \in \Omega} > 0.
\]

Hence, combining equations (11.1a) and (11.4) we obtain (recall, \( \Gamma = \partial \Omega \)):

\[
0 = \int_{\Gamma} \frac{u^k}{k} dx - \frac{u^{k+1}}{k+1} dt = \int_{\Omega} \left[ \frac{\partial}{\partial t} \left( \frac{u^k}{k} \right) + \frac{\partial}{\partial x} \left( \frac{u^{k+1}}{k+1} \right) \right] dx dt > 0.
\]

The contradiction we have just arrived at, \(0 > 0\), proves that for smooth functions \( u = u(x,t) \), problem (11.1a), (11.1b) and the Cauchy problem (11.2) are equivalent.

### 11.1.2 The Mechanism of Formation of Discontinuities

Let us first suppose that the solution \( u = u(x,t) \) of the Cauchy problem (11.2) is smooth. Then we introduce the curves \( x = x(t) \) defined by the following ordinary differential equation:

\[
\frac{dx}{dt} = u(x,t). \tag{11.6}
\]
These curves are known as characteristics of the Burgers equation: \( u_t + uu_x = 0 \).
Along every characteristic \( x = x(t) \), the solution \( u = u(x,t) \) can be considered as a function of the independent variable \( t \) only:
\[
u(x,t) = u(x(t), t) = u(t).
\]
Therefore, using (11.6) and (11.2), we can write:
\[
\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} \frac{dx}{dt} = 0.
\]
Consequently, the solution is constant along every characteristic \( x = x(t) \) defined by equation (11.6): \( u(x,t) \big|_{x=x(t)} = \text{const} \). In turn, this implies that the characteristics of the Burgers equation are straight lines, because if \( u = \text{const} \), then equation (11.6) yields:
\[
x = ut + x_0. \quad (11.7)
\]
In formula (11.7), \( x_0 \) denotes the abscissa of the point \((x_0,0)\) on the \((x,t)\) plane from which the characteristic originates, and \( u = \psi(x_0) = \tan \alpha \) denotes its slope with respect to the vertical axis \( t \), see Figure 11.1. Thus, we see that by specifying the initial function \( u(x,0) = \psi(x) \) in the definition of problem (11.2), we fully determine the pattern of characteristics, as well as the values of the solution \( u = u(x,t) \), at every point of the semi-plane \( t > 0 \).

![Figure 11.1: Characteristic (11.7).](image)

![Figure 11.2: Schematic behavior of characteristics of the Burgers equation.](image)
Note that under the assumption of a smooth solution, the characteristics of the Burgers equation cannot intersect. Otherwise, each characteristic would bring its own value of the solution into the intersection point and accordingly, the solution would not be a single-valued function. If the function $\psi = \psi(x)$ monotonically increases, i.e., if $\psi'(x) > 0$, then the angle $\alpha$ (see Figure 11.1) also increases as a function of $x_0$, and the characteristics do not intersect, see Figure 11.2(a). However, if $\psi = \psi(x)$ happens to be a monotonically decreasing function, i.e., if $\psi'(x) < 0$, then the characteristics converge toward one another as time elapses, and their intersections are unavoidable regardless of what regularity the function $\psi(x)$ has, see Figure 11.2(b). In this case, the smooth solution of problem (11.2) ceases to exist and a discontinuity forms, starting at the moment of time $t = \tilde{t}$ when at least two characteristics intersect, see Figure 11.2(b).

The corresponding graphs of the function $u = u(x,t)$ at the moments of time $t = 0$, $t = \tilde{t}/2$, and $t = \tilde{t}$ are schematically shown in Figure 11.3.

### 11.1.3 Condition at the Discontinuity

Assume that there is a curve $L = \{(x,t) | x = x(t)\}$ inside the domain of the solution $u = u(x,t)$ of problem (11.1a), (11.1b), on which this solution undergoes a discontinuity of the first kind, alternatively called the jump. Assume also that when approaching this curve from the left and from the right we obtain the limit values:

$u(x,t) = u_{\text{left}}(x,t)$

and

$u(x,t) = u_{\text{right}}(x,t),$

respectively, see Figure 11.4. It turns out that the values of $u_{\text{left}}(x,t)$ and $u_{\text{right}}(x,t)$...
are related to the velocity of the jump $\dot{x} = dx/dt$ in a particular way, and altogether these quantities are not independent.

Let us introduce a contour $ABCD$ that straddles a part of the trajectory of the jump $L$, see Figure 11.4. The integral conservation law (11.1a) holds for any closed contour $\Gamma$ and in particular for the contour $ABCD$:

$$\int_{ABCD} \frac{u^k}{k} dx - \frac{u^{k+1}}{k+1} dt = 0. \quad (11.8)$$

Next, we start contracting this contour toward the curve $L$, i.e., start making it narrower. In doing so, the intervals $BC$ and $DA$ will shrink toward the points $E$ and $F$, respectively, and the corresponding contributions to the integral (11.8) will obviously approach zero, so that in the limit we obtain:

$$\int_{L'} \left( \frac{u^k}{k} \right) dx - \left( \frac{u^{k+1}}{k+1} \right) dt = 0,$$

or alternatively:

$$\int_{L'} \left( \frac{u^k}{k} \right) \frac{dx}{dt} - \left( \frac{u^{k+1}}{k+1} \right) dt = 0.$$

Here the rectangular brackets: $[z] \equiv z_{\text{right}} - z_{\text{left}}$ denote the magnitude of the jump of a given quantity $z$ across the discontinuity, and $L'$ denotes an arbitrary stretch of the jump trajectory.

Since $L'$ is arbitrary, the integrand in the previous equality must be equal to zero at every point:

$$\left( \frac{u^k}{k} \right) \frac{dx}{dt} - \left( \frac{u^{k+1}}{k+1} \right) \bigg|_{(x,t) \in L} = 0,$$

and consequently,

$$\frac{dx}{dt} = \frac{u^{k+1}}{k+1} \cdot \left[ \frac{u^k}{k} \right]^{-1}. \quad (11.9)$$

Formula (11.9) indicates that for different values of $k$ we can obtain different conditions at the trajectory of discontinuity $L$. For example, if $k = 1$ we have:

$$\frac{dx}{dt} = \frac{u_{\text{left}} + u_{\text{right}}}{2}, \quad (11.10)$$

and if $k = 2$ we can write:

$$\frac{dx}{dt} = \frac{2 u_{\text{left}}^2 + u_{\text{left}}u_{\text{right}} + u_{\text{right}}^2}{u_{\text{left}} + u_{\text{right}}}. \quad (11.11)$$

We therefore conclude that the conditions that any discontinuous solution of problem (11.1a), (11.1b) must satisfy at the jump trajectory $L$ depend on $k$. 

11.1.4 Generalized Solution of a Differential Problem

Let us define a generalized solution to problem (11.2). This solution can be discontinuous, and we simply identify it with the solution to the integral conservation law (11.1a), (11.1b). Often the generalized solution is also called a weak solution.

In the case of a solution that has continuous derivatives everywhere, we have seen (Section 11.1.1) that the weak solution, i.e., solution to problem (11.1a), (11.1b), does not depend on $k$ and coincides with the classical solution to the Cauchy problem (11.2). In other words, the solution in this case is a differentiable function $u = u(x,t)$ that turns the Burgers equation $u_t + uu_x = 0$ into an identity and also satisfies the initial condition $u(x,0) = \psi(x)$. We have also seen that even in the continuous case it is very instrumental to consider both the integral formulation (11.1a), (11.1b) and the differential Cauchy problem (11.2). By staying only within the framework of problem (11.1a), (11.1b), we would make it more difficult to reveal the mechanism of the formation of discontinuity, as done in Section 11.1.2.

In the discontinuous case, the definition of a weak solution to problem (11.2) that we have just introduced does not enhance the formulation of problem (11.1a), (11.1b) yet; it merely renames it. Let us therefore provide an alternative definition of a generalized solution to problem (11.2). In doing so, we will only consider those solutions to problem (11.1a), (11.1b) that have continuous partial derivatives everywhere on the strip $0 < t < T$, except perhaps for a set of smooth curves $x = x(t)$ along which the solution may undergo discontinuities of the first kind (jumps).

**DEFINITION 11.1** The function $u = u(x,t)$ is called a generalized (weak) solution to the Cauchy problem (11.2) that corresponds to the integral conservation law (11.1a) if:

1. The function $u = u(x,t)$ satisfies the Burgers equation [see formula (11.2)] at every point of the strip $0 < t < T$ that does not belong to the curves $x = x(t)$ which define the jump trajectories.

2. Condition (11.9) holds at the jump trajectory.

3. For every $x$ for which the initial function $\psi = \psi(x)$ is continuous, the solution $u = u(x,t)$ is continuous at the point $(x,0)$ and satisfies the initial condition $u(x,0) = \psi(x)$.

The proof of the equivalence of Definition 11.1 and the definition of a generalized solution given in the beginning of this section is the subject of Exercise 1.

Let us emphasize that in the discontinuous case, the generalized solution of the Cauchy problem (11.2) is not determined solely by equalities (11.2) themselves. It also requires that a particular conservation law, i.e., particular value of $k$, be specified that would relate the jump velocity with the magnitude of the jump across the discontinuity, see formula (11.9). Note that the general formulation of problem (11.1a), (11.1b) we have adopted provides no motivation for selecting any preferred value of $k$. However, in the problems that originate from real-world scientific applications,
the integral conservation laws analogous to (11.1a) would normally express the conservation of some actual physical quantities. These conservation laws are, of course, well defined. In our subsequent considerations, we will assume for definiteness that the integral conservation law (11.1a) that corresponds to the value of $k = 1$ holds. Accordingly, condition (11.10) is satisfied at the jump trajectory.

11.1.5 The Riemann Problem

Having defined the weak solutions of problem (11.2), see Definition 11.1, we will now see how a given initial discontinuity evolves when governed by the Burgers equation. In the literature, the problem of evolution of a discontinuity specified in the initial data is known as the Riemann problem.

Consider problem (11.2) with the following discontinuous initial function:

$$
\psi(x) = \begin{cases} 
2, & x < 0, \\
1, & x > 0.
\end{cases}
$$

The corresponding solution is shown in Figure 11.5. The evolution of the initial discontinuity consists of its propagation with the speed $\dot{x} = (2 + 1)/2 = 3/2$. This speed, which determines the slope of the jump trajectory, is obtained according to formula (11.10) as the arithmetic mean of the slopes of characteristics to the left and to the right of the shock. As can be seen, the characteristics on either side of the discontinuity impinge on it. In this case the discontinuity is called a shock; it is similar to the shock waves in the flows of ideal compressible fluid. One can show that the solution from Figure 11.5 is stable with respect to the small perturbations of initial data.

Next consider a different type of initial discontinuity:

$$
\psi(x) = \begin{cases} 
1, & x < 0, \\
2, & x > 0.
\end{cases}
$$

One can obtain two alternative solutions for the initial data (11.11). The solution shown in Figure 11.6(a) has no discontinuities for $t > 0$. It consists of two regions with $u = 1$ and $u = 2$ bounded by the straight lines $\dot{x} = 1$ and $\dot{x} = 2$, respectively, that originate from $(0,0)$. These lines do not correspond to the trajectories of discontinuities, as the solution is continuous across both of them. The region in between these two lines is characterized by a family of characteristics that all originate at the same point $(0,0)$. This structure is often referred to as a fan (of characteristics); in the context of gas dynamics it is known as the rarefaction wave.
The solution shown in Figure 11.6(b) is discontinuous; it consists of two regions $u = 1$ and $u = 2$ separated by the discontinuity with the same trajectory $\dot{x} = (1 + 2)/2 = 3/2$ as shown in Figure 11.5. However, unlike in the case of Figure 11.5, the characteristics in Figure 11.6(b) emanate from the discontinuity and veer away as the time elapses rather than impinge on it; this discontinuity is not a shock.

To find out which of the two solutions is actually realized, we need to include additional considerations. Let us perturb the initial function $\psi(x)$ of (11.11) and consider:

$$
\psi(x) = \begin{cases} 
1, & x < 0, \\
1 + x/\varepsilon, & 0 \leq x \leq \varepsilon, \\
2, & x > \varepsilon.
\end{cases} \quad (11.12)
$$

The function $\psi(x)$ of (11.12) is continuous, and the corresponding solution $u(x,t)$ of problem (11.2) is determined uniquely. It is shown in Figure 11.7. When $\varepsilon$ tends to zero, this solution approaches the continuous fan-type solution of problem (11.2), (11.11) shown in Figure 11.6(a). At the same time, the discontinuous solution of problem (11.2), (11.11) shown in Figure 11.6(b) appears unstable with respect to the small perturbations of initial data. Hence, it is the continuous solution with the fan that should be selected as the true solution of problem (11.2), (11.11), see Figure 11.6(a). As for the discontinuous solution from Figure 11.6(b), its exclusion due to the insta-
bility is similar to the exclusion of the so-called rarefaction shocks that appear as mathematical objects when analyzing the flows of ideal compressible fluid.

Exercises

1. Prove that the definition of a weak solution to problem (11.2) given in the very beginning of Section 11.1.4 is equivalent to Definition 11.1.

2.* Consider the following auxiliary problem [cf. formula (11.2)]:

\[
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \mu \frac{\partial^2 u}{\partial x^2}, \quad 0 < t < T, \quad -\infty < x < \infty, \\
u(x,0) = \psi(x), \quad -\infty < x < \infty,
\]

(11.13)

where \(\mu > 0\) is a parameter that is similar to viscosity in the context of fluid dynamics. The differential equation of (11.13) is parabolic rather than hyperbolic. It is known to have a smooth solution for any smooth initial function \(\psi(x)\). If the initial function is discontinuous, the solution is also known to become smoother as time elapses.

Let \(\psi(x) = 2\) for \(x < 0\) and \(\psi(x) = 1\) for \(x > 0\). Prove that when \(\mu \to 0\), the solution of problem (11.13) approaches the generalized solution of problem (11.2) (see Definition 11.1) that corresponds to the conservation law (11.1a) with \(k = 1\).

11.2 Construction of Difference Schemes

In this section, we will provide examples of finite-difference schemes for computing the generalized solution (see Definition 11.1) of problem (11.2):

\[
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0, \quad 0 < t < T, \quad -\infty < x < \infty, \\
u(x,0) = \psi(x), \quad -\infty < x < \infty,
\]

that corresponds to the integral conservation law (11.1a) with \(k = 1\).

Let us assume for definiteness that \(\psi(x) > 0\). Then, \(u(x,t) > 0\) for all \(x\) and all \(t > 0\). Naturally, the first idea is to consider the simple upwind scheme:

\[
\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau} + \frac{u_{m}^{p} - u_{m-1}^{p}}{h} = 0, \\
m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots, \\
u_{m}^{0} = \psi(mh).
\]

(11.14)

By freezing the coefficient \(u_{m}^{p}\) at a given grid location, we conclude that the resulting constant coefficient difference equation satisfies the maximum principle (see the analysis on page 167) when going from time level \(t_{p}\) to time level \(t_{p+1}\), provided that
the time step \( \tau_p = t_{p+1} - t_p \) is chosen to satisfy the inequality:

\[
\frac{\tau_p}{h} \leq \frac{1}{\sup_m |u_m^p|}.
\]

Then, stability of scheme (11.14) can be expected. Furthermore, if the solution of problem (11.2) is smooth, then scheme (11.14) is clearly consistent. Therefore, it should converge, and indeed, numerical computations of the smooth test solutions corroborate the convergence.

However, if the solution of problem (11.2) is discontinuous, then no convergence of scheme (11.14) to the generalized solution of problem (11.2) is expected. The reason is that no information is built into the scheme (11.14) as to what specific conservation law is used for defining the generalized solution. The Courant, Friedrichs, and Lewy condition is violated here in the sense that the generalized solution given by Definition 11.1 depends on the specific value of \( k \) that determines the conservation law (11.1a), while the finite-difference solution does not depend on \( k \).

Therefore, we need to use special techniques for the computation of generalized solutions. One approach is based on the equation with artificial viscosity \( \mu > 0 \):

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

As indicated previously (see Exercise 2 of Section 11.1), when \( \mu \to 0 \) this equation renders a correct selection of the generalized solution to problem (11.2), i.e., it selects the solution that corresponds to the conservation law (11.1a) with \( k = 1 \). Moreover, it is also known to automatically filter out the unstable solutions of the type shown in Figure 11.6(b). Alternatively, we can employ the appropriate conservation law explicitly. In Sections 11.2.2 and 11.2.3 we will describe two different techniques based on this approach. The main difference between the two is that one technique introduces special treatment of the discontinuities, as opposed to all other areas in the domain of the solution, whereas the other technique employs uniform formulae for computation at all grid nodes.

### 11.2.1 Artificial Viscosity

Consider the following finite-difference scheme that approximates problem (11.13) and that is characterized by the artificially introduced small viscosity \( \mu > 0 \) [cf. scheme (11.14)]:

\[
\frac{u_{m+1}^p - u_m^p}{\tau} + u_m^p \frac{u_m^p - u_{m-1}^p}{h} = \mu \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2},
\]

\[
m = 0, \pm 1, \pm 2, \ldots, \quad p = 0, 1, \ldots,
\]

\[
u_m^0 = \psi(mh).
\]

Assume that \( h \to 0 \), and also that the sufficiently small time step \( \tau = \tau(h, \mu) \) is chosen accordingly to ensure stability. Then the solution \( u^{(h)} = \{u_m^p\} \) of problem (11.15)
converges to the generalized solution of problem (11.2), (11.10). The convergence is uniform in space and takes place everywhere except on the arbitrarily small neighborhoods of the discontinuities of the generalized solution. To ensure the convergence, the viscosity \( \mu = \mu(h) \) must vanish as \( h \to 0 \) with a certain (sufficiently slow) rate. Various techniques based on the idea of artificial dissipation (artificial viscosity) have been successfully implemented for the computation of compressible fluid flows, see, e.g., [RM67, Chapters 12 & 13] or [Tho95, Chapter 7]. Their common shortcoming is that they tend to smooth out the shocks. As an alternative, one can explicitly build the desired conservation law into the structure of the scheme used for computing the generalized solutions to problem (11.2).

11.2.2 The Method of Characteristics

In this method, we use special formulae to describe the evolution of discontinuities that appear in the process of computation, i.e., as the time elapses. These formulae are based on the condition (11.10) that must hold at the location of discontinuity. At the same time, in the regions of smoothness we use the differential form of the conservation law, i.e., the Burgers equation itself: \( u_t + uu_x = 0 \).

The key components of the method of characteristics are the following. For simplicity, consider a uniform spatial grid \( x_m = mh, m = 0, \pm 1, \pm 2, \ldots \). Suppose that the function \( \psi(x) \) from the initial condition \( u(x,0) = \psi(x) \) is smooth. From every point \( (x_m,0) \), we will “launch” a characteristic of the differential equation \( u_t + uu_x = 0 \). In doing so, we will assume that for the given function \( \psi(x) \), we can always choose a sufficiently small \( \tau \) such that on any time interval of duration \( \tau \), every characteristic intersects with no more than one neighboring characteristic. Take this \( \tau \) and draw the horizontal grid lines \( t = t_p = p\tau, p = 0, 1, 2, \ldots \). Consider the intersection points of all the characteristics that emanate from the nodes \( (x_m,0) \) with the straight line \( t = \tau \), and transport the respective values of the solution \( u(x_m,0) = \psi(x_m) \) along the characteristics from the time level \( t = 0 \) to these intersection points.

If no two characteristics intersect on the time interval \( 0 \leq t \leq \tau \), then we perform the next step, i.e., extend all the characteristics until the time level \( t = 2\tau \) and again transport the values of the solution along the characteristics to the points of their intersection with the straight line \( t = 2\tau \). If there are still no intersections between the characteristics for \( \tau \leq t \leq 2\tau \), then we perform yet another step and continue this way until on some interval \( t_p \leq t \leq t_{p+1} \) we find two characteristics that intersect. Assume that these characteristics emanate from the

![FIGURE 11.8: The method of characteristics.](image-url)
nodes \((x_m, 0)\) and \((x_{m+1}, 0)\), see Figure 11.8. Then, we consider the midpoint of
the interval \(Q_{m+1}^{p+1}, Q_{m}^{p+1}\) as the point that the incipient shock originates from. Sub-
sequently, we replace the two different points \(Q_{m}^{p+1}\) and \(Q_{m+1}^{p+1}\) by one point \(Q\) (the
midpoint), and assign two different values of the solution \(u_{\text{left}}\) and \(u_{\text{right}}\) to this point:
\[
u_{\text{left}}(Q) = u(Q_{m}^p), \quad u_{\text{right}}(Q) = u(Q_{m+1}^p).
\]
From the point \(Q\), we start up the trajectory of the shock until it intersects with the
horizontal line \(t = t_p + 2\), see Figure 11.8. The slope of this trajectory with respect to
the time axis is determined from the condition at the discontinuity (11.10):
\[
\tan \alpha = \frac{u_{\text{left}} + u_{\text{right}}}{2}.
\]
From the intersection point of the shock trajectory with the straight line \(t = t_p + 2\),
we draw two characteristics backwards until they intersect with the straight line \(t = t_{p+1}\). The slopes of these two characteristics are \(u_{\text{left}}\) and \(u_{\text{right}}\) from the previous time
level, i.e., from the time level \(t = t_{p+1}\). With the help of interpolation, we find the
values of the solution \(u\) at the intersection points of the foregoing two characteristics
with the line \(t = t_{p+1}\). Subsequently, we use these values as the left and right values of
the solution, respectively, at the location of the shock on the next time level \(t = t_{p+2}\).
This enables us to evaluate the new slope of the shock as the arithmetic mean of the
new left and right values. Subsequently, the trajectory of the shock is extended one
more time step \(\tau\), and the procedure repeats itself.

The advantage of the method of characteristics is that it allows us to track all of
the discontinuities and to accurately compute the shocks starting from their incep-
tion. However, new shocks continually form in the course of the computation. In
particular, the non-essential (low-intensity) shocks may start intersecting so that the
overall solution pattern will become rather elaborate. Then, the computational logic
becomes more complicated as well, and the requirements of the computer time and
memory increase. This is a key disadvantage of the method of characteristics that
singles out the discontinuities and computes those in a non-standard way.

### 11.2.3 Conservative Schemes

Finite-difference schemes that neither introduce the artificial dissipation (Section 11.2.1), nor explicitly use the condition at the discontinuity (11.10) (Section 11.2.2), must rely on the integral conservation law itself.

Consider two families of
straight lines on the plane
\((x, t)\): the horizontal lines \(t = p \tau\), \(p = 0, 1, 2, \ldots\), and the
vertical lines \(x = (m + 1/2)h\), \(m = 0, \pm 1, \pm 2, \ldots\). These
lines partition the plane into rectangular cells. On the

![Grid cell](image-url)
sides of each cell we will mark the respective midpoints, see Figure 11.9, and compose the overall grid \( D_h \) of the resulting nodes (we are not showing the coordinate axes in Figure 11.9).

The unknown function \([u]_h\) will be defined on the grid \( D_h \). Unlike in many previous examples, when \([u]_h\) was introduced as a mere trace of the continuous exact solution \( u(x,t) \) on the grid, here we define \([u]_h\) by averaging the solution \( u(x,t) \) over the side of the grid cell (see Figure 11.9) that the given node belongs to:

\[
[u]_h \bigg|_{(x_m,t_p)} \overset{\text{def}}{=} \bar{u}_m^p = \frac{1}{h} \int_{x_{m-1/2}}^{x_{m+1/2}} u(x,t_p) dx,
\]

\[
[u]_h \bigg|_{(x_m+1/2,t_{p+1/2})} \overset{\text{def}}{=} \hat{U}_{m+1/2}^{p+1} = \frac{1}{\tau} \int_{t_p}^{t_{p+1}} u(x_{m+1/2},t) dt.
\]

The approximate solution \( u^{(h)} \) of our problem will be defined on the same grid \( D_h \). The values of \( u^{(h)} \) at the nodes \((x_m,t_p)\) of the grid that belong to the horizontal sides of the rectangles, see Figure 11.9, will be denoted by \( u_m^p \), and the values of the solution at the nodes \((x_{m+1/2},t_{p+1/2})\) that belong to the vertical sides of the rectangles will be denoted by \( U_{m+1/2}^{p+1/2} \).

Instead of the discrete function \( u^{(h)} \) defined only at the grid nodes \((m,p)\) and \((m+1/2,p+1/2)\), let us consider its extension to the family piece-wise constant functions of a continuous argument defined on the horizontal and vertical lines of the grid. In other words, we will think of the value \( u_m^p \) as associated with the entire horizontal side \( \{ (x,t) | x_{m-1/2} < x < x_{m+1/2}, t = t_p \} \) of the grid cell that the node \((x_m,t_p)\) belongs to, see Figure 11.9. Likewise, the value \( U_{m+1/2}^{p+1/2} \) will be defined on the entire vertical grid interval \( \{ (x,t) | x = x_{m+1/2}, t_p < t < t_{p+1} \} \). The relation between the quantities \( u_m^p \) and \( U_{m+1/2}^{p+1/2} \), where \( m = 0, \pm 1, \pm 2, \ldots \) and \( p = 0, 1, 2, \ldots \), will be established based on the integral conservation law (11.1a) for \( k = 1 \):

\[
\oint_{\Gamma} u dx - \frac{u^2}{2} dt = 0.
\]

Let us consider the boundary of the grid cell from Figure 11.9 as the contour \( \Gamma \):

\[
- \oint_{\Gamma} u^{(h)} dx - \frac{[u^{(h)}]^2}{2} dt = 0. \quad (11.16)
\]

Using the actual values of the foregoing piece-wise constant function \( u^{(h)} \), we can rewrite equality (11.16) as follows:

\[
h[u_m^{p+1} - u_m^p] + \frac{\tau}{2} \left( U_{m+1/2}^{p+1/2} - U_{m-1/2}^{p+1/2} \right)^2 = 0. \quad (11.17)
\]
Formula (11.17) implies that if there were a certain rule for the evaluation of the quantities \( \left( U_{m+1/2}^{p+1/2} \right)^2 \), \( m = 0, \pm 1, \pm 2, \ldots \), given the quantities \( u_m^p, m = 0, \pm 1, \pm 2, \ldots \), then we could have advanced one time step and obtained \( u_m^{p+1}, m = 0, \pm 1, \pm 2, \ldots \). In other words, formula (11.17) would have enabled a marching algorithm. Note that the quantities \( \left( U_{m+1/2}^{p+1/2} \right)^2 \) are commonly referred to as fluxes. The reason is that the Burgers equation can be equivalently recast in the divergence form:

\[
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \frac{\partial u^2}{\partial t} + \frac{\partial F(u)}{\partial x} = 0,
\]

where \( F(u) \) is known as the flux function for the general equation: \( u_t + F_x(u) = 0 \).

The fluxes \( \left( U_{m+1/2}^{p+1/2} \right)^2 \) can be computed using various approaches. However, regardless of the specific approach, the finite difference scheme (11.17) always appears conservative. This important characterization means the following.

Let us draw an arbitrary non-self-intersecting closed contour in the upper semi-plane \( t > 0 \) that would be completely composed of the grid segments, see Figure 11.10. Accordingly, this contour \( \Gamma_h \) encloses some domain \( \Omega_h \) composed of the grid cells. Next, let us perform term-wise summation of all the equations (11.17) that correspond to the grid cells from the domain \( \Omega_h \). Since equations (11.17) and (11.16) are equivalent and the only difference is in the notations, we may think that the summation is performed on equations (11.16). This immediately yields:

\[
- \int_{\Gamma_h} u^{(h)} dx - \frac{\left( u^{(h)} \right)^2}{2} dt = 0.
\]  

(11.18)

Formula (11.18) is easy to justify. The integrals along those sides of the grid rectangles that do not belong to the boundary \( \Gamma_h \) of the domain \( \Omega_h \), see Figure 11.10, mutually cancel out. Indeed, each of these interior grid segments belongs to two neighboring cells. Consequently, the integration of the function \( u^{(h)} \) along each of those appears twice in the sum (11.18) and is conducted in...
the opposite directions, see Figure 11.11. Hence only the contributions due to the exterior boundary \( \Gamma_h \) do not cancel, and we arrive at equality (11.18).

Scheme (11.17) provides an example of what is known as conservative finite-difference schemes. In general, given a scheme, if we perform term-wise summation of its finite-difference equations over the nodes of the grid domain \( \Omega_h \), and only those contributions to the sum remain that correspond to the boundary \( \Gamma_h \), then the scheme is called conservative. Conservative schemes are analogous to the differential equations of divergence type, for example:

\[
\text{div} \varphi = \frac{\partial \varphi_1}{\partial t} + \frac{\partial \varphi_2}{\partial x} = 0.
\]

Once integrated over a two-dimensional domain \( \Omega \), these equations give rise to a contour integral along \( \Gamma = \partial \Omega \), see formula (11.3). Finite-difference scheme (11.14) is not conservative, whereas scheme (11.17) is conservative.

**REMARK 11.1** Let the grid function \( u^{(h)} \) that satisfies equation (11.17) for \( m = 0, \pm 1, \pm 2, \ldots \) and \( p = 0, 1, 2, \ldots \), converge to a piece-wise continuous function \( u(x,t) \) when \( h \rightarrow 0 \) uniformly on any closed region of space that does not contain the discontinuities. Also let \( u^{(h)} \) be uniformly bounded with respect to \( h \). Then, \( u(x,t) \) satisfies the integral conservation law:

\[
\oint_{\Gamma} u \, dx - \frac{\partial^2}{2} dt = 0,
\]

where \( \Gamma \) is an arbitrary piece-wise smooth contour. In other words, \( u^{(h)} \) converges to the generalized solution of problem (11.2). This immediately follows from the possibility to approximate \( \Gamma \) by \( \Gamma_h \), from formula (11.18), and from the convergence that we have just assumed.

For the difference scheme (11.17) to make sense, we still need to define a procedure for evaluating the fluxes \( \left( U_{m+1/2}^{p+1/2} \right)^2 \) given the quantities \( u_{m}^{p} \). To do that, we can exploit the solution to a special Riemann problem. This approach leads to one of the most popular and successful conservative schemes known as the Godunov scheme.

Assume that at the initial moment of time \( t = 0 \) the solution is specified as follows:

\[
u(x,0) = \begin{cases} u_{\text{left}}, & x < 0, \\ u_{\text{right}}, & x > 0, \end{cases}
\]

(11.19)

where \( u_{\text{left}} = \text{const} \) and \( u_{\text{right}} = \text{const} \). We can then obtain the corresponding generalized solution. In Section 11.1.5, we have analyzed specific examples for \( u_{\text{left}} = 2 \), \( u_{\text{right}} = 1 \) and for \( u_{\text{left}} = 1 \), \( u_{\text{right}} = 2 \). In the general case, the weak solution is obtained similarly. To compute the fluxes for the Godunov scheme, we need to know the value \( U = u(0,t) \) of the generalized solution \( u(x,t) \) at \( x = 0 \).
In fact, it is easy to see that for $x = 0$ the solution to the Riemann problem with the data (11.19) may be equal to $u_{\text{left}}$, $u_{\text{right}}$, or 0, depending on the specific values of $u_{\text{left}}$ and $u_{\text{right}}$. In particular, if $u_{\text{left}} > 0$ and $u_{\text{right}} > 0$, then $u(0,t) = u_{\text{left}}$; in doing so, we will have the situation shown in Figure 11.5 if $u_{\text{left}} > u_{\text{right}}$, and the situation shown in Figure 11.6(a) if $u_{\text{left}} < u_{\text{right}}$. Likewise, if $u_{\text{left}} < 0$ and $u_{\text{right}} < 0$, then $u(0,t) = u_{\text{right}}$; again, we will have a shock if $|u_{\text{left}}| > |u_{\text{right}}|$ and a fan (rarefaction wave) if $|u_{\text{left}}| > |u_{\text{right}}|$. If $u_{\text{left}} > 0 > u_{\text{right}}$, then there will be a shock, and depending on whether $(u_{\text{left}} + u_{\text{right}})/2 > 0$ or $(u_{\text{left}} + u_{\text{right}})/2 < 0$ we will have either $u(0,t) = u_{\text{left}}$ or $u(0,t) = u_{\text{right}}$, respectively. If, conversely, $u_{\text{left}} < 0 < u_{\text{right}}$, then the vertical axis $x = 0$ will always be contained inside the fan, and $u(0,t) = 0$.

Accordingly, the quantity $U_{m+1/2}^{p+1}$, $U$ for scheme (11.17) will be determined by solving the Riemann problem for the initial discontinuity at the location $x = x_{m+1/2}$, when the constant values are specified to the left and to the right of the discontinuity: $u_{\text{left}} = u_{m}^{p}$ and $u_{\text{right}} = u_{m+1}^{p}$, respectively. For example, if $u_{m}^{p} > 0$, and $m = 0, \pm 1, \pm 2, \ldots$, then $U_{m+1/2}^{p+1} = u_{\text{left}} = u_{m}^{p}$ for all $m = 0, \pm 1, \pm 2, \ldots$, and scheme (11.17) becomes:

$$\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau} + \frac{1}{h} \left[ \frac{(u_{m}^{p})^2}{2} - \frac{(u_{m-1}^{p})^2}{2} \right] = 0,$$

$$u_{m}^{0} = \frac{1}{h} \int_{x_{m-1/2}}^{x_{m+1/2}} \psi(x) dx,$$

or alternatively:

$$\frac{u_{m}^{p+1} - u_{m}^{p}}{\tau} + \frac{u_{m}^{p} - u_{m-1}^{p}}{2h} + \frac{u_{m}^{p} - u_{m-1}^{p}}{h} = 0.$$

It is easy to see that when

$$r = \frac{\tau}{h} \leq \frac{1}{\sup_{m} |u_{m}^{p}|},$$

the maximum principle holds:

$$\sup_{m} |u_{m}^{p+1}| \leq \sup_{m} |u_{m}^{p}| \leq \ldots \leq \sup_{m} |u_{m}^{0}| \leq \sup_{x} |\psi(x)|.$$

Consequently, if $\tau \leq h [\sup_{x} |\psi(x)|]^{-1}$, we can expect that the resulting finite-difference scheme will be stable for some reasonable choice of norms. Numerical experiments corroborate that when the grid is refined, the solution $u^{(h)}$ of problem (11.17) with piece-wise monotone and piece-wise smooth initial data $\psi(x)$ converges to some function $u(x,t)$ that has a finite number of discontinuities. In doing so, the convergence is uniform outside of any neighborhood of the shocks.

Of course, the Godunov scheme, for which the fluxes are computed based on the solution to a specially chosen Riemann problem, is not the only conservative scheme for problem (11.2), (11.10). Conservative schemes can be obtained using various approaches, in particular, one based on the predictor-corrector idea. For example,
The well-known MacCormack scheme is conservative:

\[
\frac{\bar{u}_m - u_m^p}{\tau} + \frac{1}{h} \left[ \frac{(u_{m+1}^p)^2 - (\bar{u}_m)^2}{2} \right] = 0,
\]

\[
\frac{u_{m+1}^p - (u_{m}^p + \bar{u}_m)/2}{\tau/2} + \frac{1}{h} \left[ \frac{(\bar{u}_m)^2 - (\bar{u}_{m-1})^2}{2} \right] = 0.
\]  

(11.21)

Proving this property is the subject of Exercise 2. Let us note, however, that even though the MacCormack scheme is consistent and conservative, its solutions do not always converge to the correct generalized solution of problem (11.2). Under certain conditions, the solution obtained by scheme (11.21) may contain a non-physical unstable discontinuity of the type shown in Figure 11.6(b), see [Tho99, Section 9.5].

Yet another example of a conservative predictor-corrector scheme consists of the implicit non-conservative predictor stage:

\[
\frac{\bar{u}_m - u_m^p}{\tau} + u_m^p \frac{\bar{u}_m - \bar{u}_{m-1}}{h} = 0
\]

(11.22a)

followed by the corrector stage rendered via scheme (11.17), where

\[
U_{m+1/2}^{p+1} = \frac{1}{2} \left( \bar{u}_m + \bar{u}_{m+1} \right).
\]  

(11.22b)

The properties of scheme (11.22a), (11.22b), (11.17) are to be studied in Exercise 3.

A detailed discussion on numerical solution of conservation laws can be found, e.g., in [Tho99, Chapter 9], as well as in [Lev92, Lev02].

**Exercises**

1. Prove the maximum principle (11.20).

2. Prove that the MacCormack scheme (11.21) is conservative.

   **Hint.** Define \( \left( t_{m+1/2}^{p+1} \right)^2 = \frac{(u_{m+1}^p)^2}{2} + \frac{(\bar{u}_m)^2}{2} \).

3. Investigate the consistency and stability of scheme (11.22a), (11.22b), (11.17).

   a) Prove that this scheme has second-order accuracy on smooth solutions of the Burgers equation.

   b) Prove that the overall scheme (11.22a), (11.22b), (11.17) is conservative.

   c) For the linearized equation with frozen coefficients, show that the scheme is stable in the von Neumann sense for any \( r = \tau/h \).
Chapter 12

Discrete Methods for Elliptic Problems

12.1 The Notion of Finite Elements
Part IV

The Methods of Boundary Equations for the Numerical Solution of Boundary Value Problems
Chapter 13

Boundary Integral Equations and Boundary Element Method

13.1 Reduction of Boundary Value Problems to Integral Equations
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Unavoidable error</td>
<td>14</td>
</tr>
<tr>
<td>2.1</td>
<td>Unstructured grid</td>
<td>61</td>
</tr>
<tr>
<td>3.1</td>
<td>Straightforward periodization</td>
<td>76</td>
</tr>
<tr>
<td>3.2</td>
<td>Periodization according to formula (3.54)</td>
<td>76</td>
</tr>
<tr>
<td>3.3</td>
<td>Chebyshev polynomials</td>
<td>78</td>
</tr>
<tr>
<td>3.4</td>
<td>Chebyshev interpolation nodes</td>
<td>79</td>
</tr>
<tr>
<td>5.1</td>
<td>Finite-difference scheme for the Poisson equation.</td>
<td>98</td>
</tr>
<tr>
<td>9.1</td>
<td>The shooting method</td>
<td>140</td>
</tr>
<tr>
<td>10.1</td>
<td>Continuous and discrete domain of dependence.</td>
<td>169</td>
</tr>
<tr>
<td>10.2</td>
<td>Stencils of the upwind, downwind, and leap-frog schemes.</td>
<td>181</td>
</tr>
<tr>
<td>10.3</td>
<td>Stencils of the explicit and implicit schemes for the heat equation.</td>
<td>183</td>
</tr>
<tr>
<td>10.4</td>
<td>Central-difference scheme for the Poisson equation.</td>
<td>184</td>
</tr>
<tr>
<td>10.5</td>
<td>Spectra of the transition operators for the upwind scheme.</td>
<td>202</td>
</tr>
<tr>
<td>10.6</td>
<td>Spectrum for the downwind scheme.</td>
<td>203</td>
</tr>
<tr>
<td>10.7</td>
<td>Spectrum for scheme (10.84).</td>
<td>204</td>
</tr>
<tr>
<td>10.8</td>
<td>Spectrum of the explicit scheme for the heat equation.</td>
<td>207</td>
</tr>
<tr>
<td>10.9</td>
<td>Spectrum of the scheme for the wave equation.</td>
<td>209</td>
</tr>
<tr>
<td>10.10</td>
<td>Spectrum of scheme (10.90).</td>
<td>211</td>
</tr>
<tr>
<td>10.11</td>
<td>Spectra of auxiliary problems for the upwind scheme (10.132).</td>
<td>232</td>
</tr>
<tr>
<td>10.12</td>
<td>Combined spectra of auxiliary problems for scheme (10.132).</td>
<td>233</td>
</tr>
<tr>
<td>10.13</td>
<td>Schematic behavior of the powers $|R_{\psi}^r|$ for $1 &lt; r &lt; 2$ and $h_3 &gt; h_2 &gt; h_1$.</td>
<td>243</td>
</tr>
<tr>
<td>10.14</td>
<td>Solution of problem (10.182) with scheme (10.183), (10.184a).</td>
<td>262</td>
</tr>
<tr>
<td>10.15</td>
<td>Solution of problem (10.182) with scheme (10.183), (10.196).</td>
<td>266</td>
</tr>
<tr>
<td>11.1</td>
<td>Characteristic (11.7).</td>
<td>278</td>
</tr>
<tr>
<td>11.2</td>
<td>Schematic behavior of characteristics of the Burgers equation.</td>
<td>278</td>
</tr>
<tr>
<td>11.3</td>
<td>Schematic behavior of the solution to the Burgers equation (11.2) in the case $\psi'(x) &lt; 0$.</td>
<td>279</td>
</tr>
<tr>
<td>11.4</td>
<td>Contour of integration.</td>
<td>279</td>
</tr>
<tr>
<td>11.5</td>
<td>Shock.</td>
<td>282</td>
</tr>
<tr>
<td>11.6</td>
<td>Solutions of the Burgers equation with initial data (11.11).</td>
<td>283</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>11.7</td>
<td>Solution of the Burgers equation with initial data (11.12)</td>
<td>283</td>
</tr>
<tr>
<td>11.8</td>
<td>The method of characteristics</td>
<td>286</td>
</tr>
<tr>
<td>11.9</td>
<td>Grid cell</td>
<td>287</td>
</tr>
<tr>
<td>11.10</td>
<td>Grid domain</td>
<td>289</td>
</tr>
<tr>
<td>11.11</td>
<td>Directions of integration</td>
<td>289</td>
</tr>
</tbody>
</table>
References


References


References


References


Index

accuracy, see computational methods
  order of, 117, 161, 164, 174
    first, 116, 117, 120, 164
    second, 118, 121
amplification factor, 201
amplification matrix, 216, 237
approximate
  computations, 7
  evaluation
    of a function, 19
    of integrals, 25
  solution, 8, 11, 15, 16
  value, 14
    of a function, 25
approximation
  accuracy of, 11
  best
    by polynomials, 86
    in linear space, 89
    in the sense of $C$, 89
    in the sense of $L_2$, 90
  by difference scheme, 111, 112,
    117, 161
  order of, 161
  by rounding, 17
  discrete, 11
  error, 11, 14, 17
  of differential operator by difference operator, 117, 185
  of functions, 15
  order of, 117, 161
artificial dissipation, see viscosity, artificial
artificial viscosity, see viscosity, artificial

basis
  in the linear space, 65

orthonormal, 66
Bernstein example, 36, 89

CFL condition, 170
CFL number, 170
characteristics, see equation, Burgers,
  characteristics of
    intersection of, 279, 286
    method of, 286
Chebyshev polynomials, 78
  roots of, 79
  stability of computation of, 81
computational methods, 9, 18–21
  characteristics of, 9
    accuracy, 9
    operation count, 9
    storage requirements, 9
  comparison of, 19–21
    in accuracy, 19
    in convergence, 21
    in efficacy, 20
    in stability, 20–21
  parallelization of, 9, 21
condition at the discontinuity, see discontinuity, condition at
conditioning, 9, 12
  ill conditioned problems, 12–13
  of interpolating polynomial, 34–35
  well conditioned problems, 12–13
conservation law, 275, 276
  integral form of, 275, 276
consistency
  of approximate operator equation, 172
of finite-difference scheme, see
finite-difference scheme, consistency of
contact discontinuity, 275
convergence
of finite-difference scheme, see
finite-difference scheme, convergence of
of finite-difference solution, see
finite-difference solution, convergence of
of Fourier series, see series, Fourier, convergence of
of interpolating polynomials, 36, 85, 88
on Chebyshev grid, 79, 82, 89
of local splines, 49
with derivatives, 49
of nonlocal splines, 54
with derivatives, 54
of piece-wise polynomial interpolation, 43
for derivatives, 43
of solutions to approximate operator equations, 173
of trigonometric interpolation, 63, 73
spectral, 75, 157
Courant number, 170
Courant, Friedrichs, and Lewy condition, 170
difference
central, 114, 117
divided, 28–29, 47
forward, 114
quotient, see quotient, difference
discontinuity, 279
condition at, 280
formation of, 279, 281
of the first kind, 279, 281
tracking of, 287
trajectory of, 280
discretization, 9, 10
comparison of methods, 11
of a boundary value problem, 11
of a function, 10
by table of values, 10
of derivatives, 10
dispersion, 192
relation, 192, 193
domain of dependence
continuous, 170
discrete, 170
dot product, 65
Duhamel principle, 215
elimination
Gaussian, 98
tri-diagonal, 52, 98, 107, 128, 142
cyclic, 54
energy estimate, 251
for a constant-coefficient equation, 252
for a finite-difference scheme, see finite-difference scheme, energy estimate for
for a variable-coefficient equation, 253
with inhomogeneous boundary condition, 254
equality
Parseval, 213
equation
Burgers, 276
characteristics of, 278
divergence form of, 289
d’Alembert, 198, 208
difference, 80
heat, 176, 181, 206
homogeneous, 196
nonlinear, 220
two-dimensional, 208
Poisson, 97, 184
system of
acoustics, 177, 210
wave, 198, 208
error, 9, 13
asymptotic behavior for interpolation, 43
estimate, 14, 40, 41
for algebraic interpolating polynomials, 88
for local splines, 48
for nonlocal splines, 54
for spectral method, 156
for unavoidable error, 42–44
of algebraic interpolation on Chebyshev grid, 79
of finite-difference solution, see finite-difference solution, error of
of piece-wise polynomial interpolation, 37–40
for derivatives, 41
multi-dimensional, 60
of Taylor formula Lagrange form, 114, 118
of the computational method, 14, 16
of trapezoidal rule, 129
of trigonometric interpolation, 63, 70
order of, 174
relative, 21, 133
round-off, 14, 16–17, 130–131
truncation, 111, 113, 161
magnitude of, 114–116
unavoidable, 13–15
due to uncertainty in the data, 14–15
family of matrices, 216, 237
stability of, 216, 237
family of operators, 237, 240
behavior of the powers $R^n_k$, 243
spectrum of, 233, 240, 259
computation using Babenko-Gelfand procedure, 245–251
uniform boundedness of powers, 236
fan, 282, 283, 291
Favard constants, 87
finite-difference scheme, 98, 103, 107, 160
A-stable, 133
conditionally, 133
unconditionally, 134
backward Euler, 123
canonical form of, 235
central difference, 185
conservative, 289, 290
consistency of, 111, 113, 161
consistent, 161, 162
convergence of, 109, 111, 125, 160
order of, 111, 160
convergent, 108, 128, 161
Crank-Nicolson for ordinary differential equations, 122
for partial differential equations, 223
dissipation of, 221
artificial, see viscosity, artificial divergent, 123
energy estimate for constant coefficients, 255
inhomogeneous boundary conditions, 257
variable coefficients, 257
evolution-type, 234
explicit, 122, 136, 163
explicit central, 204
first-order downwind, 180
instability of, 203
first-order upwind, 180
Babenko-Gelfand stability of, 230–233
dissipation of, 222
for initial boundary value problem, 231, 234
von Neumann stability of, 201
for an integral equation, 129
stability of, 130
for systems, 138
for the heat equation, 182
  Babenko-Gelfand stability of, 228–230
  explicit, 183, 206, 272
  implicit, 183, 207, 273
  on a finite interval, 226
  two-dimensional, 208
  weighted, 217
  with variable coefficients, 219
forward Euler, 121, 135, 194
  convergence of, 127
  for systems, 138
  stability of, 127
Godunov, 290
Heun, 122
  implicit, 122
  implicit central, 218
  implicit downwind, 218
  implicit upwind, 218
Lax-Friedrichs, 195
  von Neumann stability of, 205
Lax-Wendroff, 191, 195, 203
  Babenko-Gelfand stability of, 267
  dissipation of, 223
  for acoustics system, 211
  for inhomogeneous equation, 197
  von Neumann stability of, 204
leap-frog, 181, 195
  Kreiss stability of, 258–267
  von Neumann stability of, 206
linear, 134
  multi-step, 134
  single-step, 134
  stability of, 134
MacCormack, 195
  monotone, 197
non-dissipative, 222
predictor-corrector, 122, 194, 291
Runge-Kutta, 136, 194
  for second-order equation, 139
  fourth order, 137
second-order, 137
  saturation of, 144–151
  spectrum of
    for Cauchy problem, 201–211
    for initial boundary value problem, 228–234
  stability constants, 131–133
  stability of, 112, 124, 164
  “almost” sufficient condition
    in terms of the spectrum of a family of operators, 243
  as energy estimate, 257
  as uniform boundedness of $R^p$, 236
Babenko-Gelfand criterion, 228
Godunov-Ryabë'kii necessary condition, 240
  in $C$, 212
  in $l_2$, 212
Kreiss necessary and sufficient condition, 265
  spectral condition for Cauchy problems, 201
  spectral condition for initial boundary value problems, 237
spectral theory, 134, 201
  sufficient conditions in $C$, 212
  unconditional, 271
von Neumann analysis of, 199
  condition of, 202
  with respect to initial data, 200
stable, 161
  stencil of, 98, 181
finite-difference solution
  convergence of, 110, 160
  rate of, 111, 125, 160, 166, 174
  error of, 111, 160
characterization by grid dimension, 174
characterization by grid size, 174
order of, 111, 160
to an integral equation
corvergence of, 130
flux, 289
evaluation of, 290
by solving a Riemann problem, 291
formula
Taylor, 16, 20, 114, 118, 137, 178
frequency, 192
function, 25
Lipschitz-continuous, 86
reconstruction of, 25
from table of Taylor coefficients, 25
from table of values, 25, 42
table of values of, 27, 44, 70
Gauss nodes, see grid, Chebyshev-Gauss
Gauss-Lobatto nodes, see grid, Chebyshev-Gauss-Lobatto
generalized eigenvalue, 264
grid, 11, 15
adaptation of, 108
Cartesian, 59, 97, 162, 179
Chebyshev, 35, 64, 79, 81, 88
Chebyshev-Gauss, 79
Chebyshev-Gauss-Lobatto, 82
dimension of, 108
equally spaced, see grid, uniform
finite-difference, 103, 106, 160
nonuniform, 107
rectangular, 59
refinement of, 43, 49
local, 55, 61
size, 40, 107
structured, 59
triangular, 60, 61
uniform, 35, 42, 59, 63, 64, 97, 106, 109, 110, 113, 129, 135, 162
unstructured, 59, 60
grid function, 103, 106
even, 68
odd, 68
vector, 216
inequality
Bessel, 71
Hölder, 72
Jackson, 86
Lebesgue, 87
instability, 20–21
of finite-difference scheme, 172
interpolating polynomial, 27
algebraic, 27
comparison of different forms, 33
conditioning of, 34–35
corvergence of, 36
divergence of, 36
existence of, 27–28
Lagrange form of, 28, 33
Newton form of, 29, 33
on Chebyshev grid, 79, 81
sensitivity to perturbations, 34
uniqueness of, 27–28
trigonometric, 63, 64, 74
convergence of, 75
for even grid functions, 68, 75, 77
for odd functions, 152–153
for odd grid functions, 68
sensitivity to perturbations, 69
stability of, 75
interpolation, 25
algebraic, 25, 27
no saturation on Chebyshev grid, 80, 82
on Chebyshev grid, 79, 81
multi-dimensional, 58–61
error estimate of, 60, 61
on structured grids, 58–60
on unstructured grids, 60–61
saturation of, 60
piece-wise polynomial, 37–44
choice of degree, 42–44
convergence of, 43
error estimate of, 37–40
of derivatives, 40–41
saturation of, 44
smooth, 45
trigonometric, 25, 63
convergence of, 63, 73, 75
eroof estimate of, 63, 70, 74
no saturation of, 73
of functions on an interval, 77
of periodic functions, 64
sensitivity to perturbations,
63, 69
interpolation grid, 35, 37, 50
effect on convergence, 89

Jacobi matrix, 138
jump, 279, 281
across the discontinuity, 280, 281
trajectory of, 280, 281
velocity of, 280, 281

Kolmogorov diameters, 43
Lebesgue constants, 34, 83, 88
growth of, 88
of algebraic interpolating poly-
nomial, 34–35, 43, 84
on Chebyshev grid, 35, 79, 80, 82, 89
on uniform grid, 35
optimal asymptotic behavior of, 89
of trigonometric interpolating
polynomial, 63, 69, 74, 75
linear combination, 65, 67, 89
loss of significant digits, 21
marching, 107, 121
mathematical model(s), 7–9
for small oscillations, 19
methods of analysis, 7–9
computational, 7–9
maximum principle, 167, 177, 197,
291
method
finite-difference, see
finite-difference scheme
Kholodov’s of constructing
schemes, 194
Newton’s, 143
convergence of, 143
of rootfinding, 141
bisection, 141
Newton’s, 141
secant, 141
of shooting, 141
instability of, 141–142
of undetermined coefficients,
185–191
spectral
Chebyshev, 157
collocation, 157
Fourier, 152–157
Galerkin, 157
no saturation of, 157

norm
Euclidean, 109, 212
maximum, 109, 163, 199
of a grid function, 109
of operator, 83
normal matrix, 216, 239
operation count, see computational
methods
operator
continuous, 83
linear, 83, 125
transition, 201, 235
boundedness of powers of, 236
eigenfunctions of, 201, 238
eigenvalues of, 201, 228, 237, 238
Index

spectrum of, 201, 221, 231, 238
optimization, 193
of dispersion properties, 193
periodization
by cosine transformation, 76
by extension, 76
principle of frozen coefficients, 219
problem
boundary value, 10, 105, 140, 142
Cauchy, 19, 123, 140, 142
difference, 199, 212, 219
difference vector, 215
for a first order advection
equation, 161, 176, 178, 195
for a first order ordinary dif.
eq., 105, 135
for a hyperbolic system, 224
for a system of ordinary dif.
eqs, 139
for acoustics system, 210
for the Burgers equation, 276
for the heat equation, 176, 181, 196, 206, 208
for the nonlinear heat equation, 220
for the wave equation, 208
Dirichlet, 97, 184
ill conditioned, see conditioning
initial boundary value, 198, 226–267
initial value, 105, 106, 113, 119, 126, 169
Riemann, 282, 290
well conditioned, see conditioning
projection
of continuous solution on the
grid, 108, 124, 160
quadrature formulae, 25
quotient
difference, 29, 106, 113, 121
rarefaction wave, 282
residual
of exact solution, 111, 113, 116, 161
Riemann invariants, 269
Runge example, 36
saturation
of finite-difference schemes, see
finite-difference scheme, saturation of
of numerical methods, 44
of piece-wise polynomial inter-
polation, see interpolation, piece-wise polynomial, saturation of
of splines
local, see splines, local, saturation of
nonlocal, see splines, nonlocal, saturation of
scheme
difference, see finite-difference
scheme
finite-difference, see
finite-difference scheme
sensitivity to perturbations
of interpolating polynomial
algebraic, 34, 37
trigonometric, 63, 69
of solution, 12
finite-difference, 112, 125, 166, 171, 175
series
Fourier, 70, 90
coefficients of, 71
convergence of, 71, 154, 156
partial sum of, 71, 90, 154, 156
remainder of, 71, 90, 154, 156
power, 25
Taylor, 19, 86
partial sum of, 19, 86
shock, 175, 275, 282, 291
rarefaction, 284
trajectory of, 287
solution
approximate, 108, 160
classical, 281
continuous, 275
discontinuous, 275, 281
exact, 106, 110, 160
  on the grid, 106, 110, 160
finite-difference, see finite-difference solution
generalized, 281
plane wave, 192
weak, 281
space
  $C[a, b]$, 83, 89
  $C^1$, 149
  $C^\infty$, 149
  $L^2[a, b]$, 89
  $l_2$, 212
Banach, 172
Hilbert, 89
linear, 65, 116
dimension of, 65, 116
normed, 109, 116
spectrum
  of a family of operators, see family of operators, spectrum of
  of an operator, 241
  of finite-difference problem, see finite-difference scheme, spectrum of
  of transition operator, see operator, transition, spectrum of
splines
  local, 45–50
  convergence of, 49
  error estimate for, 48
  optimality of, 49
  saturation of, 54
  multi-dimensional, 60
  nonlocal, 45, 50–55
  convergence of, 54
error estimate for, 54
optimality of, 54
saturation of, 54
Schoenberg’s cubic, 50
stability, 9, 20–21, 75, 81, 82
  of approximate operator equation, 172
  of families of matrices, see family of matrices, stability of
  of finite-difference scheme, see finite-difference scheme, stability of
  summation by parts, 256, 258
theorem
  Bernstein, 89
  convergence of consistent and stable schemes, 125, 166
  existence and uniqueness
    of algebraic interpolating polynomial, 27
    of local spline, 45
    of trigonometric interpolating polynomial, 64, 74
  Faber-Bernstein, 88
  Godunov, 197
  Godunov-Ryaben’kii, 240
  Jackson inequality, see inequality, Jackson
  Kantorovich, 173
  Kreiss
    matrix, 216
    on stability of dissipative schemes, 224
  Lax equivalence, 175
  Lebesgue inequality, see inequality, Lebesgue
  necessary and sufficient condition for stability in $l_2$, 213
  Newton form of interpolating polynomial, 31
  Riesz-Fischer, 213
  Rolle (mean value), 32, 38, 41
  stability of linear schemes for
Index

ordinary differential equations, 134
Weierstrass, 86

theory
GKS, 258
of approximation, 83
time marching, 163, 182, 183

trace
of continuous function on the grid, 117
of continuous right-hand side on the grid, 160
of continuous solution on the grid, 106, 108, 160, 162

trapezoidal rule, 129
error of, 129

viscosity, 284
artificial, 225, 285
von Neumann stability analysis, see finite-difference scheme, stability of, von Neumann analysis of

wave
rarefaction, 282
shock, 282
wavelength, 193
wavenumber, 192
well-posedness, 166, 254