STUDIES IN MATHEMATICS AND ITS APPLICATIONS

19

DIFFERENCE SCHEMES

S.K. Godunov
V.S. Ryabenkii

NORTH-HOLLAND

DIFFERENCE SCHEMES

An Introduction to the Underlying Theory

STUDIES IN MATHEMATICS AND ITS APPLICATIONS

VOLUME 19

NORTH-HOLLAND-AMSTERDAM • NEW YORK • OXFORD • TOKYO

DIFFERENCE SCHEMES

An Introduction to the Underlying Theory

S. K. GODUNOV
Mathematics Institute
Novosibirsk, U.S.S.R.

and

V.S. RYABENKII
Institute of Applied Mathematics
Academy of Sciences of the U.S.S.R.
Moscow, U.S.S.R.

English translation by:

E. M. GELBARD
Argonne National Laboratory
Argonne, Illinois
U.S.A.

1987

NORTH-HOLLAND-AMSTERDAM • NEW YORK • OXFORD • TOKYO

ISBN: 0 444 70233 4

A translation of:

Raznostnie Skhemi
Second Edition, Revised
©The Copyright Agency of the U.S.S.R.
Moscow, U.S.S.R., 1977

Publishers:

ELSEVIER SCIENCE PUBLISHERS B.V.
P.O. Box 1991
1000 BZ Amsterdam
The Netherlands

Sole distributors for the U.S.A. and Canada:

ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE
NEW YORK, N.Y. 10017
U.S.A.

Library of Congress Cataloging-in-Publication Data

Godunov, S. K. (Sergeĭ Konstantinovich)
Difference schemes.

(Studies in mathematics and its applications ; v. 18)
Bibliography: p. 475-
Includes index.
1. Difference equations. 2. Differential equations, Linear. I. Ryaben'kiĭ, V. S. (Viktor Solomonovich) II. Title. III. Series.
QA431.G58513 1987 515'.625 87-6867
ISBN 0-444-70233-4

PRINTED IN THE NETHERLANDS

PREFACE

Much applied and theoretical research in natural sciences leads to boundary-value problems stated in terms of differential equations. So as to solve these problems on electronic computers the differential problems are replaced approximately by difference schemes.

This book is intended to serve as a first introduction to the theory of difference schemes; it is written as a textbook for students of technical universities, of the Moscow Physico-Technical and Moscow Engineering-Physics Institutes, and for students in university physics and mathematics departments. In addition, some sections of the book will probably be of interest to computations specialists. Differences in the interests of readers in the above-named categories have been reflected in the structure of this book.

This book consists of five Parts and a small Appendix. Any desired number (two or more) of the first Parts may be taken as a sort of self-contained introduction to the subject. In addition the volume of material studied may be controlled by including more or less of the material in small print,* and by the selection of problems to be solved. At the end of the book we have suggested literature for a deeper study of many questions relating to the theory and application of difference schemes, and for bibliographical investigations. A shorter introduction to the theory of difference schemes can be found in the book listed as Ref. [11].

In the text, below, direct references to original work will appear only in those few cases where auxiliary results are cited without proof.

Contemporary computational techniques and accumulated experience allow us, with the aid of difference schemes, to compute approximate solutions of problems which are very complicated, and are not amenable to study by other methods. Assurance that the solution is computed correctly is attained: by applying the same computational schemes to the solution of those few problems for which exact solutions are available; by comparing computational results with the results of physical experiments in the range of

---

*A section in small print, in the original Russian appears, here in translation, as a section set off by horizontal rows of asterisks. Each such section is preceded by a row of six asterisks, and followed by a row of three.

parameters for which experiments are possible; and through the aid of other methods which cannot be considered mathematically rigorous. But an understanding of essentials, necessary for the construction of appropriate difference schemes, is achieved by consideration of a series of properly chosen model problems; problems simple enough for detailed study on some accepted level of mathematical rigor, but nevertheless capturing one or another of those features of the original problem which interest us, while this original problem is unaccessible to rigorous study either because of its complexity, or for lack of time.

Stressing a mathematically rigorous treatment of model problems, we have tried at the same time to give the reader a correct picture of the relation between theory, on the one hand and, on the other, computational experiments on electronic computers, using difference schemes created for practical computations.

The appearance of this book was made possible by earlier work by the authors on [10], and also by the work of one of them on lecture courses which he presented for several years at the Moscow Physico-Technical Institute. The set-up of these courses was strongly influenced by many fruitful discussions with O. M. Bielotserkovskii (through whose initiative these courses were started), V. F. Dyachenko, O. V. Lokutsievskii, R. P. Fedorenko, L. A. Chudov and E. E. Schnol. Many useful comments were made by N. S. Bakhvalov and B. L. Rozhdestvenskii after reading the book in manuscript.

We are sincerely grateful to all of them.

The authors

PREFACE TO THE SECOND EDITION

The second edition differs from the first in: the inclusion of Chapter 12 on variational-difference schemes; of §47 on the stability of iterative processes for the solution of non-selfadjoint difference equations; and of Sect. 10 of the Appendix containing some considerations on the computational use of the method of internal boundary conditions. In addition some typographical errors and inaccuracies have been eliminated and the bibliography has been brought up to date.

The authors

It should be noted that, in the text below, and in the section entitled "Bibliographical Commentaries", citations in Russian journals have been changed, wherever possible, to corresponding citations in English translations of these journals. In particular, references to the Russian journals

Akademiia Nauk SSR, Doklady;
Uspekhi Matematicheskhikh Nauk; and
Zhurnal Vychislitelnoi Matematiki I Matematicheskoi Fiziki

have been replaced by corresponding references to their translations

Soviet Mathematics, Doklady;
Russian Mathematics Surveys; and
U.S.S.R Computational Mathematics and Mathematical Physics,

respectively.

Translator

This Page Intentionally Left Blank

TABLE OF CONTENTS

PART 2

DIFFERENCE SCHEMES FOR ORDINARY DIFFERENTIAL EQUATIONS

PART 3

DIFFERENCE SCHEMES FOR PARTIAL DIFFERENTIAL EQUATIONS. BASIC CONCEPTS

PART 4

PROBLEMS WITH TWO SPACE VARIABLE

This Page Intentionally Left Blank

INTRODUCTION

Consideration of the problems, both applied and theoretical, of contemporary natural sciences often leads to differential equations, and the study of such problems can be considered finished only after these equations have been solved. In some cases it is possible to write their solutions in terms of well-known elementary functions. As a rule, however, this is in principle impossible, so that the construction of a solution in terms of an explicit closed-form expression cannot be considered a standard method for solving differential equations. One cannot say that this analytic approach has completely lost its value. It remains a necessary and very powerful instrument for the study of simplified, so-called "model", problems. The study of carefully selected model problems allows one to draw some conclusions as to the nature of the behavior of the unsimplified, original, problem.

But, together with this analytic approach, various numerical methods are being more and more widely used for the solution of differential equations. Widespread use of these methods has been made possible by the appearance of fast computers which can store large arrays of numbers, upon which they can perform arithmetic operations in accordance with some given program. So as to take advantage of the capabilities of these machines the computational method makes a transition, from the required solution, to a certain numerical table one needs to construct, and to a sequence of arithmetic operations for the computation of the numbers in this table. One might, for example, set out to find some of the leading coefficients in an expansion of the solution in a power series, or a trigonometric series. Here we develop the theory of differential equation solution-methods based on finite differences. The essence of this most versatile numerical method consists in that one puts, in the role of the desired set of numbers, a table of values of the solution at the points of a certain set, ordinarily called a "net". For computation of the required table one makes use of algebraic equations which approximate, and take the place of, the differential equation.

For the sake of clarity, consider the simplest example of a difference scheme for the numerical solution of the equation

$$u'(x) + Au(x) = 0,$$

with the initial condition $u(0) = 1$. We choose an $h > 0$, and set out to obtain, in place of the function $u(x)$, a table of its values

$$u(0),\ u(h),\ u(2h),\ \ldots,\ u(nh),\ \ldots$$

We now replace the derivative by the difference approximation

$$\frac{u(x + h) - u(x)}{h},$$

which is permissible if the step-width in the table is taken sufficient small. After introduction of this difference approximation we get, in place of the differential equation, the difference equation

$$\frac{u(x + h) - u(x)}{h} + Au(x) = 0,$$

which approximates it, and which can be used for an approximate computation of the required table. To implement this computation we rewrite the difference equation in the form of a recursion relation

$$u(x + h) = (1 - Ah)u(x).$$

Sequentially taking $x = 0, h, 2h, \ldots$, we find that

$$\begin{aligned} u(h) &= (1 - Ah), \\ u(2h) &= (1 - Ah)^2, \\ &\cdots\cdots\cdots \\ u(Nh) &= (1 - Ah)^N. \\ &\cdots\cdots\cdots \end{aligned}$$

Setting $h = 1/N$ we get

$$u(1) = \left(1 - \frac{A}{N}\right)^N$$

in place of the exact solution

$$u(1) = e^{-A}.$$

But, as is well known from a standard course in mathematical analysis, for h sufficiently small or, correspondingly, for N large enough, $(1 - A/N)^N$ differs very little from e^{-A}. Thus we see that the approximate solution

gotten via this difference scheme and depending on the step-size, h, converges, as this step-size decreases, to the exact solution of the differential equation.

Another example of a difference equation approximating the same differential equation

$$u'(x) + Au(x) = 0,$$

is obtained by replacing the derivative with the difference expression

$$\frac{u(x + h) - u(x - h)}{2h}.$$

This equation takes the form

$$\frac{u(x + h) - u(x - h)}{2h} + Au(x) = 0.$$

For the differential equation

$$u''(x) + Au'(x) + Bu(x) = f(x)$$

one can construct a difference analog by replacing $u''(x)$, for example, with the following approximate expression:

$$\frac{\frac{u(x + h) - u(x)}{h} - \frac{u(x) - u(x - h)}{h}}{h} = \frac{u(x + h) - 2u(x) + u(x - h)}{h^2}.$$

The first derivative may be replaced by one of the difference expressions already used. After such substitutions, and using the centered expression for the first derivative, we get the difference equation

$$\frac{u(x + h) - 2u(x) + u(x - h)}{h^2} + A\,\frac{u(x + h) - u(x - h)}{2h} + Bu(x) = f(x).$$

The construction of difference equations is no more difficult in the case of differential equations with variable coefficients. If, for example, one wants to compute the solution of the equation

$$u'(x) + A(x)u(x) = 0,$$

where the coefficient, A, is a function of x, this can be done with the aid of the difference equation

$$\frac{u(x + h) - u(x)}{h} + A(x)u(x) = 0.$$

Difference schemes can treat non-linear equations just as easily. For example, the equation

$$u'(x) + \sin(xu(x)) = 0$$

can be solved approximately via the scheme

$$\frac{u(x + h) - u(x)}{h} + \sin(xu(x)) = 0.$$

From the above examples one may form the impression that the construction of difference schemes, and the solution of differential equations through use of such schemes, are matters presenting no difficulties. This is a deceptive impression.

Already in the simplest cases, even in solving linear equations with constant coefficients, it happens frequently that a seemingly plausible difference scheme has a solution which does not converge, as the net is refined, to the desired solution of the differential equation. Of course, with such a scheme one cannot compute the desired function with unlimited precision.

Further, after a convergent scheme is constructed, it is necessary to compute the solution of the resulting system of algebraic equations for a large number of values of the unknown function at the knots of the net. This, in many important cases, is not at all easy. Sometimes it is possible to circumvent this difficulty by choosing a convergent difference scheme of different construction, such that the resulting system of linear equations is easy to solve exactly; in certain other cases methods have been developed for the approximate computation of the solution of difference problems to any prescribed level of accuracy.

Everyone who is engaged in the numerical solution of differential equations should be aware of the difficulties involved in the construction and use of difference schemes, and should know how to overcome these difficulties.

Part 1

ORDINARY DIFFERENCE EQUATIONS

Chapter 1
Difference Equations of First and Second Order. Examples of Difference Schemes

§ 1. Simplest difference equations

1. Difference equations. For differential equations of first order

$$u'(x) + Au(x) = f(x)$$

we constructed, in the Introduction, two difference schemes:

$$\frac{u(x+h) - u(x)}{h} + Au(x) = f(x),$$

$$\frac{u(x+h) - u(x-h)}{2h} + Au(x) = f(x),$$

which may be written, respectively, as

$$-\left(\frac{1 - Ah}{h}\right) u(x) + \frac{1}{h} u(x+h) = f(x), \tag{1}$$

$$-\frac{1}{2h} u(x-h) + Au(x) + \frac{1}{2h} u(x+h) = f(x). \tag{2}$$

For the differential equation of second order

$$u''(x) + Au'(x) + Bu(x) = f(x)$$

we constructed the difference equation

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + A\,\frac{u(x+h) - u(x-h)}{2h} + Bu(x) = f(x),$$

which one can rewrite in the form

$$\frac{1}{h^2}\left(1 - \frac{Ah}{2}\right)u(x-h) - \frac{1}{h^2}\left(2 - Bh^2\right)u(x) + \frac{1}{h^2}\left(1 + \frac{Ah}{2}\right)u(x+h) = f(x). \tag{3}$$

The above examples of difference equations, approximating the simplest differential equations, each belong to one of the two classes:

$$au(x) + bu(x + h) = f(x), \tag{1'}$$

$$au(x - h) + bu(x) + cu(x + h) = f(x). \tag{2'}$$

If the sequence of points, dividing the x axis into intervals of length h, is numbered from left to right so that $x_n = x_{n-1} + h$, and we define $u_n = u(x_n)$, $f_n = f(x_n)$, then our difference scheme can be rewritten in the form

$$au_n + bu_{n+1} = f_n, \tag{4}$$

$$au_{n-1} + bu_n + cu_{n+1} = f_n. \tag{5}$$

In §§1-4 we will be engaged in the study of difference equations of forms (4) and (5), but will not ask whether these equations constitute difference schemes for any differential equations.

In equations (4) and (5) the unknowns, u_n, form a sequence $\{u_n\}$:

$$\ldots, u_{-3}, u_{-2}, u_{-1}, u_0, u_1, u_2, u_3, \ldots$$

We will often put this sequence into one-to-one correspondence with the sequence of points numbered by the integers

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots,$$

a set of points sometimes referred to as a "net".

The sequence $\{u_n\}$ may be regarded as a function, u, given at the points of the net. In this case u_k is the value of the net function, u, at the point numbered k. In Fig. 1 we have drawn the graph of a net function, u. This graph consists of the totality of points (x_k, u_k) on the plane Oxu.

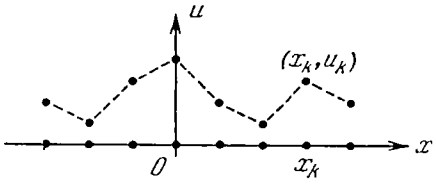


Fig. 1

Since we have chosen not to consider the connection between difference and differential equations, we are in no way obliged to take the distance between neighboring points to be equal to h. We may choose this distance as we like and, for example, could set it equal to to unity, taking x_0 to be at the origin. Then the net function, u, will be defined at the points with coordinates $x_k = k$.

We will assume for simplicity, that the coefficients a, b and c in Eqs. (4) and (5) are constant. In saying that the equations in question are equations with constant coefficients we mean that the coefficients are independent of the index, n; for example, the equation

$$u_{n-1} + 5\sqrt{n}\, u_n + u_{n+1} = 0$$

is not an equation with constant coefficients.

We will consider only such equations (4) for which a and b are different from zero. In (5) a and c will be assumed to be different from zero. The sequence $\{f_n\}$ will be called the "right-hand side" of these equations.

If we postulate that the sequence $\{u_n\}$ is defined at all whole-numbered points n, $-\infty < n < \infty$, and put no further restrictions on this sequence, then it is easy to see that Eqs. (4) and (5) have many solutions. Thus, for example, the equation $qu_n - u_{n+1} = 0$ has, as a solution, $u_n \equiv 0$, as well as the solution $u_n = q^n$.

In order to single out a unique solution of Eq. (4)

$$au_n + bu_{n+1} = f_n,$$

it is sufficient to fix the value of this solution at any single whole-numbered point m, that is to fix u_m. In fact Eq. (4) can be written as a recursion relation

$$u_{n+1} = \frac{1}{b}(f_n - au_n),$$

from which, for $n = m, m + 1, \ldots$, one can sequentially define u_{m+1}, u_{m+2}, ..., i.e., all u_n for $n > m$. Writing the equation in the other recursive form

$$u_{n-1} = \frac{1}{a}(f_n - bu_n),$$

we can, in just the same way, define u_n for $n < m$.

To single out a unique solution of Eq. (5)

$$au_{n-1} + bu_n + cu_{n+1} = f_n$$

it is sufficient to assign, arbitrarily, values of u at any two adjacent whole-numbered points, i.e., for example, to fix the values of u_{m-1} and u_m. That this is true immediately follows from the fact that the cited equation can be rewritten in the following two recursive forms:

$$u_{n+1} = \frac{1}{c}(f_n - bu_n - au_{n-1}),$$

$$u_{n-1} = \frac{1}{a}\left(f_n - bu_n - cu_{n+1}\right).$$

2. Order of difference equations. We will repeat once more the results obtained above, and then formulate the concept of order for difference equations (4) and (5).

In order to single out a unique solution of Eq. (4)

$$au_n + bu_{n+1} = f_n$$

it is sufficient to fix the value of u at one point. Such an equation is called an "equation of first order". To single out a unique solution of Eq. (5)

$$au_{n-1} + bu_n + cu_{n+1} = f_n$$

it suffices to assign values to the solution at two adjacent points. For this reason such an equation is called an "equation of second order".

One might, in fact, apply the same considerations to the simplest equation,

$$au_n = f_n, \quad a \neq 0,$$

the solution of which is uniquely defined without the imposition of any auxiliary restrictions on the sequence $\{u_n\}$. It is natural to call such an equation an "equation of zeroeth order".

The simplest difference scheme (1) for the differential equation of first order, $u' + Au = f$, is a difference equation of first order. The scheme (3) for the second-order differential equation, $u'' + Au' + Bu = f$, is of second order.

Scheme (2)

$$-\frac{1}{2h}u(x-h) + Au(x) + \frac{1}{2h}u(x+h) = f(x)$$

for the equation $u' + Au = f$ shows that the order of the difference scheme may be greater than the order of the differential equation. In this example the differential equation is first order, the corresponding differential equation -- second.

3. General solution of difference equations. We will now describe the structure of the solutions of the above difference equations. First we consider the homogeneous equation

$$a\bar{u}_n + b\bar{u}_{n+1} = 0. \tag{6}$$

Let Y_n be the solution of Eq. (6) satisfying the initial condition $Y_0 = 1$. Clearly $\bar{u}_n = \alpha Y_n$ will also be a solution of the homogeneous equations for any choice of the constant, α. It isn't difficult to show that any solution of the homogeneous equation (6) can be represented in this form. In fact each solution is uniquely determined by it's value at $n = 0$. But the solution, $\bar{u}_n$, taking on the given value $\bar{u}_0$, may be obtained from the expression $\bar{u}_n = \alpha Y_n$ if we take the factor α to be equal to $\bar{u}_0$.

Consider, now, the inhomogeneous equation (4)

$$au_n + bu_{n+1} = f_n.$$

Let $\{\tilde{u}_n\}$ and $\{u_n^*\}$ be any two of it's solutions. Subtracting one of the equations

$$a\tilde{u}_n + b\tilde{u}_{n+1} = f_n,$$

$$au_n^* + bu_{n+1}^* = f_n,$$

from the other we see that the difference $\tilde{u}_n - u_n^* = \bar{u}_n$ satisfies the homogeneous equation (6) $a\bar{u}_n + b\bar{u}_{n+1} = 0$. Therefore any solution $\{\tilde{u}_n\}$ may be written in the form

$$\tilde{u}_n = u_n^* + \bar{u}_n = u_n^* + \alpha Y_n$$

with an appropriate choice of the constant, α. It can easily be verified that, on the other hand, for any arbitrary choice of α the expression $u_n = u_n^* + \alpha Y_n$ represents one solution of the inhomogeneous equation:

$$au_n + bu_{n+1} = a\left(u_n^* + \alpha Y_n\right) + b\left(u_{n+1}^* + \alpha Y_{n+1}\right) =$$

$$= \left(au_n^* + bu_{n+1}^*\right) + \alpha(aY_n + bY_{n+1}) = f_n + \alpha \cdot 0 = f_n.$$

Thus we have shown that the general solution of the homogeneous equation (6)

$$a\bar{u}_n + b\bar{u}_{n+1} = 0$$

takes the form

$$\bar{u}_n + \alpha Y_n,$$

where Y_n is a particular solution of this equation satisfying the initial condition $Y_0 = 1$, and α is an arbitrary constant. The general solution of the inhomogeneous equation (4)

$$au_n + bu_{n+1} = f_n$$

can be represented in the form

$$u_n = u_n^* + \alpha Y_n,$$

where u_n^* is any particular solution of this inhomogeneous equation, and α is again an arbitrary constant.

By analogous arguments one can prove an analogous assertion also for difference equations of second order. We will not carry through these arguments (the reader can construct them without difficulty), but only formulate the final result.

The general solution of the homogeneous difference equation

$$a\bar{u}_{n-1} + b\bar{u}_n + c\bar{u}_{n+1} = 0 \tag{7}$$

may be represented in the form

$$\bar{u}_n = \alpha Y_n + \beta Z_n,$$

where Y_n and Z_n are particular solutions of Eq. (7), satisfying the initial conditions

$$Y_0 = 1,\ Y_1 = 0,$$

$$Z_0 = 0,\ Z_1 = 1,$$

while α and β are arbitrary constants.

The general solution of the inhomogeneous equation (5)

$$au_{n-1} + bu_n + cu_{n+1} = f_n$$

can be represented in the form

$$u_n = u_n^* + \alpha Y_n + \beta Z_n,$$

where u_n^* is any particular solution of this inhomogeneous equation.

All of the results of this section could be repeated verbatim for difference equations with variable coefficients, but we will not do this so as not to encumber our presentation with unessential details.

PROBLEMS

1. Prove that the general solution of the homogeneous difference equation

$$a_n u_n + b_n u_{n+1} = 0$$

with variable coefficients $a_n \neq 0$, $b_n \neq 0$, can be written in the form $u_n = \alpha y_n$, where y_n is any particular solution not identically zero for all n, and α is an arbitrary constant.

2. Prove that the general solution of the homogeneous difference equation of second order

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = 0$$

with variable coefficients $a_n \neq 0$, $c_n \neq 0$, may be written in the form

$$u_n = \alpha y_n + \beta z_n,$$

where y_n and z_n are any two particular solutions of this equation for which the determinant

$$\begin{vmatrix} y_0 & y_1 \\ z_0 & z_1 \end{vmatrix}$$

is not equal to zero.

3. Let y_n and z_n be any two particular solutions of the second-order difference equation of problem 2. Prove that the determinant

$$\begin{vmatrix} y_n & y_{n+1} \\ z_n & z_{n+1} \end{vmatrix} = y_n z_{n+1} - z_n y_{n+1}$$

either vanishes for each n, or is different from zero for all n.

4. At how many consecutive points must one specify values of the solution of the difference equation

$$au_n + bu_{n+1} + cu_{n+2} + du_{n+3} = f_n,$$

$a \neq 0$, $d \neq 0$, so that there will exist one and only one solution, $\{u_n\}$ taking on the specified values at these points? What must we take as the order of this equation?

§ 2. Difference equation of first order

In this section we will derive expressions for the general solution of the difference equation of first order with constant coefficients

$$au_n + bu_{n+1} = f_n$$

imposing fairly weak restrictions on f_n.

As shown in §1, the general solution can be represented in the form

$$u_n = u_n^* + \alpha Y_n = u_n^* + \alpha\left(-\frac{a}{b}\right)^n,$$

where u_n^* is any particular solution, and α is an arbitrary constant.

Thus the problem of finding the general solution has reduced to the problem of finding any one particular solution u_n^*.

1. Fundamental solution. First we will construct the solution for one particular special form of the given right-hand side

$$f_n = \begin{cases} 0, & n \neq 0, \\ 1 & n = 0. \end{cases}$$

To designate such a function one normally uses the Kronecker symbol

$$\delta_k^n = \begin{cases} 0, & n \neq k, \\ 1, & n = k. \end{cases}$$

Then $f_n = \delta_0^n$.

The solution of the equation

$$au_n + bu_{n+1} = \delta_0^n$$

we designate as G_n:

$$aG_n + bG_{n+1} = \delta_0^n. \tag{1}$$

The solution G_n is called a *fundamental solution* of the equation

$$au_n + bu_{n+1} = f_n,$$

because, as we will see on page 14, in terms of G_n one can write particular solutions of this equation for different, fairly arbitrary, right-hand sides f_n.

Thus, we want to find any solution of the following three groups of equations:

$$\text{I.} \quad aG_n + bG_{n+1} = 0 \qquad \text{for } n \leq -1.$$

$$\text{II.} \quad aG_0 + bG_1 = 1.$$

$$\text{III.} \quad aG_n + bG_{n+1} = 0 \qquad \text{for } n \geq 1.$$

Let $G_n = 0$ for $n \leq 0$. Then all equations of Group I will be satisfied. From Eq. II we find that $G_1 = 1/b$. The equations of Group III may be rewritten as a recursion equation, $G_{n+1} = -(a/b)G_n$, from which we find, sequentially,

$$G_2 = \frac{1}{b}\left(-\frac{a}{b}\right) = -\frac{1}{a}\left(-\frac{a}{b}\right)^2,$$

$$G_3 = \frac{1}{b}\left(-\frac{a}{b}\right)^2 = -\frac{1}{a}\left(-\frac{a}{b}\right)^3,$$

. . . . . .

$$G_n = -\frac{1}{a}\left(-\frac{a}{b}\right)^n \quad \text{for } n \geq 1.$$

We now write out a summary of equations determining G_n:

$$G_n = \begin{cases} 0 & \text{for } n \leq 0, \\ -\frac{1}{a}\left(-\frac{a}{b}\right)^n & \text{for } n > 1. \end{cases} \tag{2}$$

This is one solution of Eq. (1). Adding to it the general solution $A(-a/b)^n$ of the corresponding homogeneous equation $au_n + bu_{n+1} = 0$, we get the general solution of Eq. (1):

$$G_n = \begin{cases} A\left(-\frac{a}{b}\right)^n & \text{for } n \leq 0, \\ \left(A - \frac{1}{a}\right)\left(-\frac{a}{b}\right)^n & \text{for } n \geq 1. \end{cases} \tag{3}$$

The fundamental solution (2) falls out of the general Eq. (3) when $A = 0$.

2. Conditions governing the boundedness of the fundamental solution. If $|a/b| = 1$ then, for any value of the constant A, we get a fundamental solution, G_n, bounded in absolute value both as $n \to +\infty$ and $n \to -\infty$. Let us extract, from the general expression (3), a bounded fundamental solution, G_n, in the case $|a/b| \neq 1$. If $|a/b| < 1$, $|-a/b|^n$ grows without bound as $n \to -\infty$. Therefore one gets a bounded solution only for $A = 0$ (Fig. 2, a). It is given by Eq. (2).

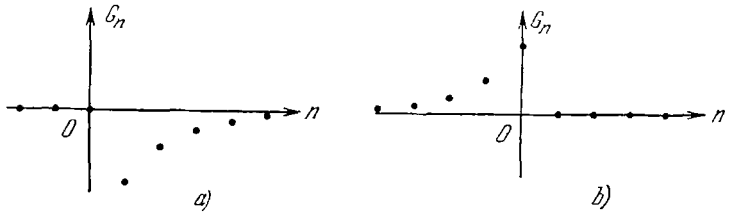


Fig. 2.

If $|a/b| > 1$, a bounded solution is obtained only for $A = 1/a$ (Fig. 2, b):

$$G_n = \begin{cases} \frac{1}{a}\left(-\frac{a}{b}\right)^n, & n \leq 0, \\ 0, & n \geq 1. \end{cases} \tag{4}$$

3. Particular solution. A particular solution of the equation

$$au_n + bu_{n+1} = f_n \tag{5}$$

with arbitrary right-hand side may be written in the form of the series

$$u_n = \sum_{k=-\infty}^{\infty} G_{n-k} f_k, \tag{6}$$

where G_n is any fundamental solution, so long as the series converges.

Let us show this using the equation

$$aG_{n-k} + bG_{n-k+1} = \delta_0^{n-k} \left(= \delta_k^n\right),$$

which is obtained from Eq. (1) if, in (1), we everywhere replace n by n-k. Substituting the convergent series (6) in the left-hand side of Eq. (5) we get

$$au_n + bu_{n+1} = a\sum_{k=-\infty}^{\infty} G_{n-k} f_k + b\sum_{k=-\infty}^{\infty} G_{n-k+1} f_k =$$

$$= \sum_{k=-\infty}^{\infty} \left(aG_{n-k} + bG_{n-k+1}\right) f_k = \sum_{k=-\infty}^{\infty} \delta_k^n f_k = f_n.$$

Series (6) may turn out to be divergent if we make no assumptions as to the behavior of the right-hand side, f_k, of the difference equation. In fact, if $f_k = (-a/b)^k$, then

$$G_{n-k}f_k = \begin{cases} A\left(-\frac{a}{b}\right)^n & \text{for } n \leq k, \\ \left(A - \frac{1}{a}\right)\left(-\frac{a}{b}\right)^n & \text{for } n \geq k+1. \end{cases}$$

and series (6) for fixed n contains an infinite number of identical terms, all different from zero.

Theorem. *Let $|a/b| \neq 1$, let G_n be a bounded fundamental solution and f_k bounded in modulus, i.e., $|f_k| < F$. Then the series*

$$u_n = \sum_{k=-\infty}^{\infty} G_{n-k}f_k$$

certainly converges.

Proof. We shall only deal with the case $|a/b| > 1$. Afterwards the reader can, without difficulty, consider the opposite case.

Under our assumptions each term of the series

$$u_n = \sum_{k=-\infty}^{\infty} G_{n-k}f_k = \sum_{k=n}^{\infty} \left[\frac{1}{a}\left(-\frac{a}{b}\right)^{n-k}\right] f_k$$

can be bounded above, in absolute value, by a term of the convergent geometric progression

$$\left|-\frac{1}{a}\left(-\frac{a}{b}\right)^{n-k} f_k\right| \leq \frac{F}{|a|}\left|\frac{a}{b}\right|^{n-k}.$$

From this follows the convergence of series (6), as well as the estimate

$$|u_n| \leq \frac{F}{|a|} \sum_{k=n}^{\infty} \left|\frac{b}{a}\right|^{k-n} = \frac{F}{|a| - |b|}, \tag{7}$$

which shows that the solution (6) is bounded.

Other bounded solutions of the equation

$$au_n + bu_{n+1} = f_n$$

do not exist, since any solution may be obtained from (6) through the addition of a solution, $\bar{u}_n = \alpha(-a/b)^n$, of the corresponding homogeneous equation. The solution $\{u_n\}$ must be bounded, since it is the difference of two bounded solutions; but this is possible only for $\alpha = 0$.

PROBLEMS

1. Find the general solution of the equation

$$2u_n - u_{n+1} = 5^n.$$

Solution. The general solution of the corresponding homogeneous equation $2\bar{u}_n - \bar{u}_{n+1} = 0$ has the form $\bar{u}_n = \alpha 2^n$. We will look for a particular solution, u_n^*, of the form $u_n^* = C5^n$ with undetermined coefficient. Substituting $u_n^* = C5^n$ into the equation we get

$$(2 \cdot 5^n - 5^{n+1})C = 5^n; \quad C = -1/3.$$

Thus

$$u_n = -\frac{5^n}{3} + \alpha 2^n.$$

(Note that, to write the particular solution u_n^* in the form of series (6) is impossible, since its general term does not tend to zero, and the series diverges.)

2. Find a particular solution u_n^* of the equation

$$2u_n - u_{n+1} = 2^n.$$

Hint. Look for a solution of the form $u_n^* = Cn \cdot 2^n$.

3. Find particular solutions u_n^* of the equation

$$2u_n - u_{n+1} = f_n$$

in the case where the right-hand side has the following special form:

a) $f_n = 1$, b) $f_n = n$, c) $f_n = n^2$, d) $f_n = 1 + 2n - n^2$.

4. Find particular solutions u_n^* of the equation

$$u_n - u_{n+1} = f_n,$$

if the right-hand side has the following special form:

a) $f_n = 1$, b) $f_n = n$, c) $f_n = n^2$.

§ 3. Difference equation of second order.

In this section we will derive expressions for the general solution of the inhomogeneous equation with constant coefficients

$$au_{n-1} + bu_n + cu_{n+1} = f_n. \tag{1}$$

In §1 it was shown that the general solution has the form

$$u_n = u_n^* + \bar{u}_n, \tag{2}$$

where u_n^* is some particular solution of the given inhomogeneous equation, and

$$\bar{u}_n = \alpha Y_n + \beta Z_n$$

is the general solution of the corresponding homogeneous equation

$$au_{n-1} + bu_n + cu_{n+1} = 0. \tag{3}$$

First we will find an expression for the general solution of the homogeneous equation (3), and then a fundamental and particular solution of the inhomogeneous equation.

1. General solution of the homogeneous equation. Recalling that in the case of the first-order difference equation there exists a solution of the form $u_n = q^n$, let us try here also to find a particular solution in the form of a geometric progression. Substituting the expression $u_n = q^n$ into the difference equation, we convince ourselves that it really will be a solution if q is a root of the quadratic

$$a + bq + cq^2 = 0, \tag{4}$$

called the *characteristic equation*. The roots of this equation may be distinct or multiple. Let us consider these two cases consecutively. If the roots q_1 and q_2 of the characteristic equation are different, we can find, in the form of a geometric progression, not one, but two independent particular solutions:

$$u_n^{(1)} = q_1^n, \qquad u_n^{(2)} = q_2^n.$$

The linear combination

$$\bar{u}_n = \alpha u_n^{(1)} + \beta u_n^{(2)} = \alpha q_1^n + \beta q_2^n \tag{5}$$

of these two solutions with arbitrary coefficients α and β also will be a solution of the homogeneous equation. Let us show that it is the general solution.

In fact any arbitrary particular solution, $\bar{u}_n$, of the homogeneous equation, taking on at $n = 0$ and $n = 1$ any prescribed values $\bar{u}_0$ and $\bar{u}_1$, may be written in this form. To accomplish this it is sufficient to define α and β via the equations

$$\alpha + \beta = \bar{u}_0,$$

$$\alpha q_1 + \beta q_2 = \bar{u}_1,$$

i.e., to set

$$\alpha = \frac{\bar{u}_0 q_2 - \bar{u}_1}{q_2 - q_1}, \qquad \beta = \frac{\bar{u}_1 - \bar{u}_0 q_1}{q_2 - q_1}.$$

In particular, Y_n and Z_n, defined in §1 as the solutions of the homogeneous equation satisfying the conditions

$$Y_0 = 1, \quad Y_1 = 0$$

$$Z_0 = 0, \quad Z_1 = 1,$$

have the form

$$\left.\begin{aligned} Y_n &= \frac{q_2}{q_2 - q_1} q_1^n - \frac{q_1}{q_2 - q_1} q_2^n, \\ Z_n &= -\frac{1}{q_2 - q_1} q_1^n + \frac{1}{q_2 - q_1} q_2^n. \end{aligned}\right\} \tag{6}$$

From Eqs. (6) we see that these equations are inapplicable in the case of multiple roots $q_1 = q_2$. Let us now consider this case.

When $q_1 = q_2$ one particular solution can again be written in the form $u_n = q_1^n$. To find a second, let us make, in Eq. (3), the substitution $u_n = y_n q_1^n$, after which we get for y_n the equation

$$a y_{n-1} + b q_1 y_n + c q_1^2 y_{n+1} = 0.$$

As is well known, a/c is equal to the product, and b/c the sum with reversed sign, of the roots of the characteristic equation (4). Since both of these roots are equal to q_1,

$$\frac{a}{c} = q_1^2, \quad \frac{b}{c} = -2q_1,$$

as a consequence of which the difference equation may be rewritten thus:

$$cq_1^2 y_{n-1} - 2cq_1^2 y_n + cq_1^2 y_{n+1} = 0,$$

or more simply:

$$y_{n-1} - 2y_n + y_{n+1} = 0.$$

Rewriting this equation in the form

$$y_{n-1} - y_n = y_n - y_{n+1},$$

we see that the difference $y_{n-1} - y_n$ does not change with n. Thus any arbitrary arithmetic progression is a solution. For us it is sufficient to find any single solution, and we take, as this solution, the arithmetic progression $y_n = n$. Recalling that we were seeking a u_n in the form $u_n = y_n q_1^n$ we find that, among the solutions of the equation $au_{n-1} + bu_n + cu_{n+1} = 0$, there is a solution

$$u_n^{(2)} = nq_1^n.$$

Thus, in the case of multiple roots $q_1 = q_2$, supplementing the particular solution $u_n^{(1)} = q^n$ we have found another, independent, particular solution $u_n^{(2)} = nq_1^n$.

The linear combination

$$\bar{u}_n = \alpha q_1^n + \beta n q_1^n$$

with arbitrary constant coefficients is also a solution of the homogeneous equation, and in fact any arbitrary particular solution may be obtained from this equation, appropriately selecting α and β. In particular, the solutions Y_n and Z_n, in the case of multiple roots, take the form

$$\left.\begin{aligned} Y_n &= q_1^n - nq_1^n, \\ Z_n &= \frac{1}{q_1} nq_1^n = nq_1^{n-1}. \end{aligned}\right\} \qquad (7)$$

It is interesting to note that Eqs. (7) can be gotten from Eqs. (6), the expressions for Y_n and Z_n when the characteristic equation has unequal roots. In that case we had, for Y_n and Z_n, the equations

$$Y_n = \frac{q_2}{q_2 - q_1} q_1^n - \frac{q_1}{q_2 - q_1} q_2^n = q_1 q_2 \frac{q_1^{n-1} - q_2^{n-1}}{q_2 - q_1},$$

$$Z_n = -\frac{1}{q_2 - q_1} q_1^n + \frac{1}{q_2 - q_1} q_2^n = \frac{q_2^n - q_1^n}{q_2 - q_1}.$$

Let us now make q_1 approach q_2. Then the expressions

$$\frac{q_2^{n-1} - q_1^{n-1}}{q_2 - q_1} \text{ and } \frac{q_2^n - q_1^n}{q_2 - q_1}$$

tend to certain limits, i.e., respectively, to $(n-1)q_1^{n-2}$ and nq_1^{n-1}. Thus we see that, in the case of multiple roots, Y_n and Z_n take the form (7).

We have, then, constructed the solutions, Y_n and Z_n, in all cases which may arise when a and c differ from zero. In the process we have shown that it is always possible to write out, in explicit form, any solution of the homogeneous second-order difference equation in question.

It's interesting to consider in more detail the case where, for real coefficients a, b, and c, the equation $a + bq + cq^2 = 0$ has complex conjugate roots q_1 and q_2. We will show that, in this case, the general solution of the homogeneous difference equation (3) may be written in the following form

$$u_n = \gamma_1 \left(\sqrt{\frac{a}{c}}\right)^n \cos n\phi + \gamma_2 \left(\sqrt{\frac{a}{c}}\right)^n \sin n\phi, \tag{8}$$

where ϕ is determined by the equation

$$\cos \phi = -\frac{b}{2\sqrt{ac}},$$

and γ_1 and γ_2 are arbitrary constants.

We get, for q_1 and q_2, the explicit expressions

$$q_1 = \frac{\sqrt{a}}{\sqrt{c}} \left[-\frac{b}{2\sqrt{ac}} + i\sqrt{1 - \left(\frac{b}{2\sqrt{ac}}\right)^2} \right],$$

$$q_2 = \sqrt{\frac{a}{c}}\left[-\frac{b}{2\sqrt{ac}} - i\sqrt{1-\left(\frac{b}{2\sqrt{ac}}\right)^2}\right].$$

In our case of complex roots, $a/c > 0$, $\left|b/(2\sqrt{ac})\right| < 1$. For this reason we may write

$$-\frac{b}{2\sqrt{ac}} = \cos\phi, \quad \sqrt{1-\left(\frac{b}{2\sqrt{ac}}\right)^2} = \sin\phi,$$

after which q_1 and q_2 take the form:

$$q_1 = \sqrt{\frac{a}{c}}\,(\cos\phi + i\sin\phi),$$

$$q_2 = \sqrt{\frac{a}{c}}\,(\cos\phi - i\sin\phi).$$

We now substitute these values of q_1 and q_2 in Eq. (5).

For $\alpha = \beta = 1/2$ we get the particular solution

$$u_n^{(1)} = \left(\sqrt{\frac{a}{c}}\right)^n \cos n\phi,$$

and, for $\alpha = 1/(2i)$, $\beta = -1/(2i)$, the particular solution

$$u_n^{(2)} = \left(\sqrt{\frac{a}{c}}\right)^n \sin n\phi.$$

A linear combination of these particular solutions, with arbitrary constant coefficients γ_1 and γ_2, gives the general solution, (8), above. (The fact that it is possible to write, in this form, the particular solution taking on, for $n = 0$ and $n = 1$, any prescribed values can easily be verified by the reader independently.)

2. General solution of the inhomogeneous equation. Fundamental solution. Now let us study the inhomogeneous difference equation

$$au_{n-1} + bu_n + cu_{n+1} = f_n, \tag{9}$$

limiting ourselves to the case (important below) where, among the roots of the characteristic equation (4), there are none equal to unity in modulus: $|q_1| \neq 1$, $|q_2| \neq 1$. First we will look for a solution of the inhomogeneous equation (9) with right-hand side f_n of the special form

$$f_n = \delta_0^n = \begin{cases} 0, & n \neq 0, \\ 1, & n = 0. \end{cases}$$

This solution will be designated as G_n and called "fundamental". We will look for a *bounded* fundamental solution, i.e., a bounded solution of the following group of equations:

$$\text{I.} \quad aG_{n-1} + bG_n + cG_{n+1} = 0 \quad \text{for} \quad n \leq -1.$$

$$\text{II.} \quad aG_{-1} + bG_0 + cG_1 = 1.$$

$$\text{III.} \quad aG_{n-1} + bG_n + cG_{n+1} = 0 \quad \text{for} \quad n \geq 1.$$

Consider, first the case of non-multiple roots, $q_1 \neq q_2$. In this case the general solution of the homogeneous equation (3) has the form

$$u_n = \alpha q_1^n + \beta q_2^n.$$

For this reason each particular solution of the homogeneous equation I can be written in the form

$$G_n = \alpha' q_1^n + \beta' q_2^n \quad \text{for} \quad n \leq 0,$$

where α' and β' are appropriately chosen constants. So also the particular solution G_n, $n \geq 0$, of the homogeneous equation III may be writen in the form

$$G_n = \alpha'' q_1^n + \beta'' q_2^n \quad \text{for} \quad n \geq 0$$

with corresponding constants α'' and β''.

In the above case $q_1 \neq q_2$, $|q_1| \neq 1$, $|q_2| \neq 1$ the following variants are possible:

a) $|q_1| < 1, \quad |q_2| > 1;$

b) $|q_1| < 1, \quad |q_2| < 1;$

c) $|q_1| > 1, \quad |q_2| < 1;$

d) $|q_1| > 1, \quad |q_2| > 1.$

We now construct the bounded fundamental solution G_n in case a). From the boundedness condition on G_n for $n \to -\infty$ it will be seen that $\alpha' = 0$, and from the boundedness condition on G_n for $n \to \infty$ it follows that $\beta'' = 0$. Therefore

$$G_n = \begin{cases} \beta' q_2^n & \text{for} \quad n \leq 0, \\ \alpha'' q_1^n & \text{for} \quad n \geq 0. \end{cases}$$

For $n = 0$ both of the last equations must give one and the same value G_0. Hence $\beta' = \alpha''$. We choose β' so as to satisfy II:

$$a\beta' q_2^{-1} + b\beta' + c\beta' q_1 = 1,$$

$$\beta' = \frac{1}{aq_2^{-1} + b + cq_1}.$$

The denominator of this fraction is different from zero:

$$aq_2^{-1} + b + cq_1 = (aq_2^{-1} + b + cq_2) + c(q_1 - q_2) = c(q_1 - q_2) \neq 0.$$

Thus,

$$G_n = \begin{cases} \dfrac{1}{aq_2^{-1} + b + cq_1} q_2^n, & n \leq 0, \\ \dfrac{1}{aq_2^{-1} + b + cq_1} q_1^n, & n \geq 0. \end{cases}$$

We have constructed the bounded fundamental solution in case a) (Fig. 3,a).

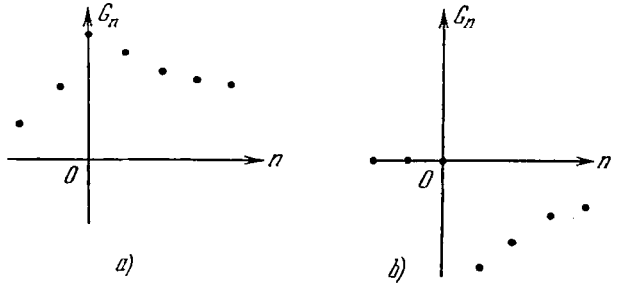


Fig. 3.

It should be noted for future reference that, under the conditions

$$\left.\begin{aligned} &\max(|a|, |b|, |c|) \geq B > 0, \\ &|q_1| < 1 - \frac{\theta}{2}, \quad \left|q_2^{-1}\right| < 1 - \frac{\theta}{2}, \end{aligned}\right\} \tag{10}$$

where $B > 0$ and $\theta > 0$ are certain prescribed numbers, we have the bound

$$|G_n| \leq \frac{4}{B\theta}\left(1 - \frac{\theta}{2}\right)^{|n|}. \tag{11}$$

* * * * * *

To derive bound (11) we note that, by virtue of the first condition (10), it must be true that either $|a| > B/4$, $|c| > B/4$ or $\sqrt{b^2 - 4ac} \geq \sqrt{B^2 - B^2/4} > B/2$. Clearly also

$$aq_2^{-1} + b + cq_1 = c(q_1 - q_2) = a(q_2^{-1} - q_1^{-1}) = \sqrt{b^2 - 4ac},$$

$$|q_1 - q_2| \geq |q_2| - |q_1| \geq \frac{1}{(1 - \theta/2)} - (1 - \theta/2) = \theta\,\frac{(2 - \theta/2)}{(2 - \theta)} > 0,$$

$$\left|q_2^{-1} - q_1^{-1}\right| > \theta.$$

From these relations one gets the bound

$$\left|aq_2^{-1} + b + cq_1\right| > \frac{B\theta}{4}$$

and Eq. (11).

* * *

In case b) it follows from the boundedness condition on G_n, for $n \to -\infty$, that $\alpha' = \beta' = 0$, so that

$$G_n = \begin{cases} 0 & \text{for } n \leq 0, \\ \alpha'' q_1^n + \beta'' q_2^n & \text{for } n \geq 0. \end{cases}$$

The condition $G_0 = 0$ implies that $\alpha'' = -\beta''$. We choose the coefficient α'' so as to satisfy equation II:

$$\alpha'' = \frac{1}{c(q_1 - q_2)}.$$

The bounded fundamental solution (Fig. 3,b) in case b) thus has the form

$$G_n = \begin{cases} 0 & \text{for } n \leq 0, \\ \dfrac{1}{c(q_1 - q_2)}\left(q_1^n - q_2^n\right) & \text{for } n \geq 0. \end{cases}$$

In case c), by analogy with case a) the bounded fundamental solution has the form

$$G_n = \begin{cases} \dfrac{1}{aq_1^{-1} + b + cq_2}\, q_1^n & \text{for } n \leq 0, \\ \dfrac{1}{aq_1^{-1} + b + cq_2}\, q_2^n & \text{for } n \geq 0. \end{cases}$$

Case d) is analogous to case b).

If the roots are multiple, $q_1 = q_2$, then, in the construction of the bounded fundamental solution, instead of the equation

$$u_n = \alpha q_1^n + \beta q_2^n$$

one uses the equation

$$u_n = \alpha q_1^n + \beta n q_1^n.$$

In the case $|q_1| < 1$ we get, for G_n,

$$G_n = \begin{cases} 0 & \text{for } n \leq 0, \\ \frac{1}{c}\, n q_1^{n-1} & \text{for } n \geq 0, \end{cases}$$

and in the case $|q_1| > 1$ we get

$$G_n = \begin{cases} -\frac{1}{a}\, n q_1^{n+1} & \text{for } n \leq 0, \\ 0 & \text{for } n \geq 0. \end{cases}$$

Thus we have treated all the variants one may encounter in the case $|q_1| \neq 1$, $|q_2| \neq 1$, $a \neq 0$, $c \neq 0$, and have found that a bounded fundamental solution exists. From the expression exhibited above one sees that this solution decreases exponentially for $n \to \pm\infty$:

$$|G_n| < G\rho^{|n|}, \tag{12}$$

where $G > 0$ and $0 < \rho < 1$ are constants. The constant ρ may be assigned any value satisfying the inequality.

$$\rho > \max\left[\min\left(|q_1|, \frac{1}{|q_1|}\right), \quad \min\left(|q_2|, \frac{1}{|q_2|}\right)\right].$$

We have examined the question of the existence and form of the fundamental solution, i.e., the solution of the inhomogeneous equation (9). For an arbitrary right-hand side $\{f_n\}$ a particular solution u_n^* may be written as the sum of a series,

$$u_n^* = \sum_{k=-\infty}^{\infty} G_{n-k} f_k, \tag{13}$$

so long as the series converges. This can be verified exactly in the same way as the analogous fact for first-order difference equations in §2. From bound (12) it follows that series (13) certainly converges if the right-hand side $\{f_k\}$ is bounded, $|f_k| < F$. In this case

$$\left|u_n^*\right| = \left|\sum_{k=-\infty}^{\infty} G_{n-k} f_k\right| \leq \sum_{k=-\infty}^{n} \left|G_{n-k} f_k\right| + \sum_{k=n+1}^{\infty} \left|G_{n-k} f_k\right| \leq$$

$$\leq GF\left[\sum_{k=-\infty}^{n} \rho^{n-k} + \sum_{k=n+1}^{\infty} \rho^{k-n}\right] \leq \frac{2G}{1-\rho} F. \tag{14}$$

For Eq. (9), where $|q_1| \neq 1$ and $|q_2| \neq 1$, the solution $\{u_n^*\}$, given by Eq. (13), is the only bounded solution for the given right-hand side. If this were not the case any second bounded solution would be obtained by the addition of some bounded solution, $\{\bar{u}_n\}$, of the homogeneous equation (3). But, from the expression for the general solution of this equation, one sees that for $|q_1| \neq 1$, $|q_2| \neq 1$, the unique solution bounded for $-\infty < n < \infty$ is $\bar{u}_n \equiv 0$. In particular, the bounded fundamental solution G_n for $|q_1| \neq 1$, $|q_2| \neq 1$ is also unique.

We note that, if condition (10) is satisfied then, using bound (11), from (13) it is easy to derive

$$\left|u_n^*\right| \leq \frac{16}{B\vartheta^2} \sup_m |f_m|. \tag{15}$$

3. Estimate of the fundamental solution in terms of the coefficients of the difference equation. In Sect. 2 we have seen that the character of the behavior of the fundamental solution G_n of Eq. (9) depends crucially on the location, in the complex plane, of the roots, q_1 and q_2, of the characteristic equation

$$P(q) \equiv a + bq + cq^2 = 0 . \tag{4}$$

Especially important in practice is the case where a, b and c are real while one of the roots, q_1 or q_2, is greater than, and the other less than one in modulus:

$$|q_1| < \rho, \quad \left|q_2^{-1}\right| < \rho, \quad 0 < \rho < 1. \tag{16}$$

Here we will point out a convenient necessary and sufficient criterion for such a disposition of the roots, applicable without their explicit computation.

Theorem. *Of the two roots, q_1 and q_2, of Eq. (4) with real coefficients, one is greater than, and the other less than one in modulus if and only if*

$$\frac{|b| - |a + c|}{|b| + |a| + |c|} \geq \vartheta > 0 , \tag{17}$$

for some θ*; and further for any* θ *such that (17) is satisfied*

$$|q_1| < 1 - \frac{\theta}{2}, \qquad \left|q_2^{-1}\right| < 1 - \frac{\theta}{2}. \tag{18}$$

Proof. We note that

$$P(1) \cdot P(-1) = (a + c + b)(a + c - b) = |a + c|^2 - b^2 =$$

$$= (|a + c| - |b|)(|a + c| + |b|).$$

If (17) is not satisfied for any $\theta > 0$, then the numerator of the fraction (17) is either equal to zero or negative.

In the first case $P(1)P(-1) = 0$, i.e., either 1 or -1 is a root of Eq. (4), and (16) is not satisfied.

In the second case $P(1)P(-1) > 0$, i.e., at the points $q = -1$ and $q = 1$ the polynomial $P(q)$ takes on values of the same sign. Thus the polynomial $P(q)$ cannot have, on the interval $-1 \leq q \leq 1$, just one root; there must be either two or none.

If there are two, then both are less than one in modulus, and (16) is not satisfied. If, on the interval $[-1, 1]$, there are no roots, then either there are no real roots at all, but only complex conjugate roots of equal modulus, or else both real roots have modulus greater than one, and (16) is again not fulfilled.

If, for some $\theta > 0$, condition (17) is satisfied, then $P(-1)P(1) < 0$, and the values of $P(q)$ at the ends of the interval $[-1, 1]$ have different signs so that, on this interval, there is precisely one root. Then the other root, also real, lies outside this interval, so that for some $\rho < 1$ (16) is satisfied. We will now sharpen this last result and, in fact, will get just the cited bound (18).

From (17) it follows that

$$|b| - |a + c| \geq \theta(|b| + |a| + |c|) > \frac{\theta}{2}|b| + 0 \cdot |a| + \left[\theta - \left(\frac{\theta}{2}\right)^2\right] \cdot |c|.$$

Therefore

$$|b| \cdot \left(1 - \frac{\theta}{2}\right) > |a + c| + \left[\theta - \left(\frac{\theta}{2}\right)^2\right] \cdot |c| \geq$$

$$\geq \left|a + c\left\{1 - \theta + \left(\frac{\theta}{2}\right)^2\right\}\right| = \left|a + c\left(1 - \frac{\theta}{2}\right)^2\right|.$$

Thus it is clear that the expressions

$$P\left(1 - \frac{\theta}{2}\right) = a + c\left(1 - \frac{\theta}{2}\right)^2 + b\left(1 - \frac{\theta}{2}\right),$$

$$P\left[-\left(1 - \frac{\theta}{2}\right)\right] = a + c\left(1 - \frac{\theta}{2}\right)^2 - b\left(1 - \frac{\theta}{2}\right)$$

have different signs so that the polynomial P(q), on the interval $-(1 - \theta/2) \leq q \leq 1 - \theta/2$, has a root q_1, $|q_1| < 1 - \theta/2$. Clearly the quantities

$$q_1' = \frac{1}{q_1}, \qquad q_2' = \frac{1}{q_2},$$

inverses of the roots of Eq. (4), obey the equation

$$a' + b'q' + c'(q')^2 = 0$$

with coefficients $a' = c$, $b' = b$, $c' = a$, satisfying the same condition (17):

$$\frac{|b'| - |a' + c'|}{|b'| + |a'| + |c'|} = \frac{|b| - |a + c|}{|b| + |a| + |c|} \geq \theta > 0.$$

Therefore one of the roots q_1', q_2' satisfies the inequality $|q'| < 1 - \theta/2$. This root can only be $q_2' = 1/q_2$, which completes the proof of bound (18).

For equations with real coefficients subject to condition (17), condition (10) and thus also bound (15) are automatically satisfied for the bounded particular solution, u_n^*, of the inhomogeneous difference equation (9).

PROBLEMS

1. Write the general solutions of equations

$$u_{n-1} - 5u_n + 6u_{n+1} = 0, \qquad n = 0, \pm 1, \ldots,$$

$$u_{n-1} - \frac{5}{2}u_n + u_{n+1} = 0, \qquad n = 0, \pm 1, \ldots,$$

$$9u_{n-1} + 3u_n + u_{n+1} = 0, \qquad n = 0, \pm 1, \ldots$$

2. Find a solution of the equation

$$u_{n-1} - \frac{5}{2}u_n + u_{n+1} = 0,$$

which is bounded for $n \to +\infty$ and takes on the value $u_0 = 1$.

3. Write out the thousandth term of the sequence $u_0, u_1, u_2, \ldots$, the first two terms of which are equal to one, $u_0 = 1$, $u_1 = 1$, while the following terms are defined by the recurrence relation

$$u_{n+1} = u_{n-1} + u_n, \qquad n = 1, 2, \ldots$$

4. Find the necessary and sufficient conditions which one must impose on the roots of the characteristic equation so that the difference equation

$$au_{n-1} + bu_n + cu_{n+1} = 0, \qquad n = 0, \pm 1, \pm 2, \ldots,$$

will have at least one nontrivial bounded solution. (The solution $u_n \equiv 0$ is called "trivial".)

5. Find the conditions which must be satisfied by the roots of the characteristic equation, necessary and sufficient to guarantee that all solutions of the equation

$$au_{n-1} + bu_n + cu_{n+1} = 0, \qquad n = 0, \pm 1, \ldots,$$

will be bounded.

6. What must be true of the roots of the characteristic equation if all solutions of the equation $au_{n-1} + bu_n + cu_{n+1} = 0$ are to tend to zero as $n \to \infty$?

7. Find any particular solution of the inhomogeneous difference equation

$$u_{n-1} - \frac{5}{2} u_n + u_{n+1} = f_n, \qquad n = 0, \pm 1, \ldots,$$

if the right hand side has the following special form:

a) $f_n = 1$. Hint. Look for a solution of the form $u_n^* = A$.

b) $f_n = n$. Hint. Look for a solution of the firm $u_n^* = A + Bn$.

c) $f_n = 3^n$. Hint. Look for a solution of the form $u_n^* = A \cdot 3^n$.

d) $f_n = \cos n$. Hint. Look for a solution of the form $u_n^* = A \sin n + B \cos n$.

8. Construct any bounded fundamental solution of the equation

$$u_{n-1} + u_n + u_{n+1} = f_n.$$

Do there exist unbounded fundamental solutions of this equation?

9. Construct any fundamental solution of the equation

$$u_{n-1} - 2u_n + u_{n+1} = f_n.$$

Is there any bounded fundamental solution?

10. Under what conditions on the roots of the characteristic equation does the difference equation

$$au_{n-1} + bu_n + cu_{n+1} = f_n$$

not have bounded fundamental solutions?

11. Using the bounded fundamental solution, write out that solution $(u_0, u_1, \ldots, u_N)$, of the equation

$$u_{n-1} - \frac{5}{2} u_n + u_{n+1} = f_n, \qquad n = 1, 2, \ldots, N-1,$$

which satisfies the condition $u_0 = \phi$, $u_n = \Psi$, where ϕ and Ψ are given numbers.

12. Find all the eigenvalues, ρ, and the corresponding eigenvectors $\Psi = \{\Psi_m\}$, $m = 0, 1, \ldots, M$, of the operator Λ_{xx},

$$\Lambda_{xx}\psi = \rho\psi ,$$

where Λ_{xx} is the operator which maps each net function, $u = \{u_m\}$, into the net function $v = \{v_m\}$, via the relations

$$v_m = \frac{1}{h^2}\left(u_{m+1} - 2u_m + u_{m-1}\right) \qquad 0 < m < M,$$

$$v_0 = v_M = 0, \qquad Mh = 1.$$

Answer:

$$\rho_k = -\frac{4}{h^2} \sin^2 \frac{\pi k}{2M} , \quad \psi_m^{(k)} = \sin \frac{k\pi m}{M} , \quad k = 1, 2, \ldots, M-1.$$

Chapter 2

Boundary-Value Problems for Equations of Second Order

Boundary-value problems of the form considered here arise when difference schemes are used for the numerical solution of ordinary and partial differential equations.

§4. Formulation of the problem. Good-conditioning criteria.

1. Formulation of the problem. The simplest boundary-value problem consists in the construction of a net function $\{u_n\}$, $n = 0, 1, \ldots, N$, satisfying the difference equation

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n, \qquad n = 1, 2, \ldots, N-1, \tag{1}$$

at the internal points $0 < n < N$ of the net interval $0 \leq n \leq N$, and taking on the given values

$$u_0 = \phi, \qquad u_N = \psi \tag{2}$$

on its boundaries. A boundary-value problem for systems of difference equations will be formulated in Section 7.

Studying the equation $a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n$, $a_n \neq 0$, $c_n \neq 0$, we remarked that, for any arbitrary choice of values of $\{u_n\}$ at any two adjacent points, for example for an arbitrary choice of u_0 and u_1, a solution $\{u_n\}$ is determined and, moreover, a unique solution.

It's interesting to consider whether one can uniquely define a solution if its values are given at two, not necessarily adjacent, points as in the boundary-value problem (1), (2). The following example shows that problem (1), (2) may turn out to be unsolveable.

Consider the boundary-value problem

$$u_{n-1} - u_n + u_{n+1} = 0, \qquad n = 1, 2, \ldots, 299, \tag{3}$$

$$u_0 = 0, \qquad u_{300} = 1 \tag{4}$$

The general solution of Eq. (3), as shown in §3, can be written in the form

$$u_n = \gamma_1 \cos \frac{n\pi}{3} + \gamma_2 \sin \frac{n\pi}{3} .$$

From the condition $u_0 = 0$ it follows that $\gamma_1 = 0$. To satisfy the condition $u_{300} = 1$ one must fix γ_2 via the equation

$$u_{300} = \gamma_2 \sin \frac{300\pi}{3} = 1.$$

But this equation is unsolveable since, for any γ_2, the left hand side is equal to zero, not one.

If, instead of the condition $u_{300} = 1$ we were to set $u_{300} = 0$ (leaving, as before, $u_0 = 0$), then again we would have to take $\gamma_1 = 0$, while γ_2 in this case would be arbitrary:

$$\gamma_2 \sin \frac{300\pi}{3} = \gamma_2 \cdot 0 = 0.$$

We see that the boundary-value problem (1), (2) may, in general, not have any solution, or the solution may turn out not to be unique. But, be that as it may, boundary-value problems are often encountered.

It turns out that there is a rather wide class of difference equations for which the boundary-value problem (1), (2), not only has always one and only one solution, but is also only weakly sensitive to rounding errors for given right-hand sides ϕ, ψ and $\{f_n\}$, i.e., the problem is "well-conditioned".

2. Definition of a well-conditioned problem. Ordinarily in studying difference schemes for the approximate solutions of differential boundary-value problems one considers not a single, isolated problem, but a whole family of such problems, arising for smaller and smaller net step-sizes. The number, N, can then be considered a parameter upon which this family depends. Refinement of the net corresponds to an increase in N.

We will say that the difference boundary-value problem (1), (2) with coefficients a_n, b_n, c_n, bounded in totality, $|a_n|$, $|b_n|$, $|c_n| < K$, is well-conditioned if for all large enough N it has one and only one solution, $\{u_n\}$, for arbitrary right-hand sides ϕ, ψ and $\{f_n\}$, and if the numbers $u_0, u_1, \ldots, u_N$, constituting the solution, satisfy the bound

$$|u_n| \leq M \max \{|\phi|, |\psi|, \max_m |f_m|\}, \tag{5}$$

where M is a number not depending on N.

* * * * * *

Sometimes one adjoins to the class of well-conditioned problems also those problems for which M cannot be taken to be constant, but is allowed to increase no faster than some given power of N, e.g., $M = CN$ or $M = CN^2$.

Our definition of good conditioning is equivalent to one which is customary in the theory of systems of linear equations, where the measure of conditioning of a system of equations $Ax = g$ with matrix A is taken to

be the quantity $||A|| \cdot ||A^{-1}||$, the product of the norms of the matrices A and A^{-1}.

* * *

Fulfillment of inequality (5) indicates that the sensitivity of the solution $\{u_n\}$ to errors (for example measurement or rounding errors) occurring in the given right-hand sides ϕ, ψ or $\{f_n\}$, does not grow with increasing N. In fact if, instead of ϕ, ψ and $\{f_n\}$, one were given, respectively, $\phi + \Delta\phi$, $\psi + \Delta\psi$ and $\{f_n + \Delta f_n\}$, then the solution would change by $\{\Delta u_n\}$. This change, because of the linearity of problem (1), (2), is the solution of the problem

$$\left.\begin{aligned} a_n \Delta u_{n-1} + b_n \Delta u_n + c_n \Delta u_{n+1} &= \Delta f_n, \qquad 0 < n < N, \\ \Delta u_0 = \Delta\phi, \qquad \Delta u_N &= \Delta\psi \end{aligned}\right\}$$

and by virtue of (5) satisfies the bound

$$|\Delta u_n| \leq M \max\{|\Delta\phi|, |\Delta\psi|, \max_m |\Delta f_m|\}.$$

By far not every boundary-value problem (1) possessing a unique solution is well-conditioned. For example if, to the right-hand-side of the equations

$$\left.\begin{aligned} u_{n+1} - 5u_n + 6u_{n-1} &= f_n, \qquad 0 < n < N, \\ u_0 = \phi, \quad u_N &= \psi \end{aligned}\right\}$$

one adds the increments

$$\Delta f_n \equiv 0, \qquad \Delta\psi = 0, \qquad \Delta\phi = \varepsilon,$$

the solution $\{u_n\}$ will change by the increment

$$\Delta u_n = 2^n \frac{1 - (2/3)^{N-n}}{1 - (2/3)^N} \Delta\phi, \qquad n = 0, 1, \ldots, N.$$

Hence

$$\Delta u_{N-1} \geq 2^{N-1} \cdot \frac{1}{3} \varepsilon.$$

The perturbation ε for given ϕ has induced, in the solution, a perturbation which grows rapidly with increasing N. The quantity M in inequality (5) clearly cannot be taken to grow more slowly than the exponential $(1/3) \cdot 2^{N-1}$.

3. Sufficient condition for a well-conditioned problem.

Theorem. *If the coefficients a_n, b_n and c_n satisfy the condition*

$$|b_n| \geq |a_n| + |c_n| + \delta, \qquad \delta > 0, \tag{6}$$

the problem (1), (2) is well-conditioned and, moreover, the solution $\{u_n\}$ satisfies the bound

$$|u_n| \leq \max \{|\phi|, |\psi|, \frac{1}{\delta} \max_m |f_m|\} . \tag{7}$$

Proof. We first assume that, for given ϕ, ψ and $\{f_n\}$, problem (1), (2) has a solution $\{u_n\}$, and establish that this solution satisfies (7). Suppose that the largest of the quantities $|u_n|$, $n = 0, 1, \ldots, N$ is $|u_k|$. If $k = 0$ or $k = N$ inequality (7) is obvious, since $u_0 = \phi$, $u_N = \psi$. It remains to consider the case $0 < k < N$, $|u_k| \geq |u_n|$. In this case, taking account of (6), we may write

$$|b_k| \cdot |u_k| = |-a_k u_{k-1} - c_k u_{k+1} + f_k| \leq$$

$$\leq |a_k| \cdot |u_{k-1}| + |c_k| \cdot |u_{k+1}| + |f_k| \leq (|a_k| + |c_k|)|u_k| + |f_k| ,$$

$$|u_n| \leq |u_k| \leq \frac{|f_k|}{|b_k| - |a_k| - |c_k|} \leq \frac{|f_k|}{\delta} ,$$

and here also (7) is satisfied.

It remains to show that problem (1), (2) has one and only one solution $\{u_n\}$ for any given right-hand sides ϕ, ψ and $\{f_n\}$.

Problem (1), (2) may be regarded as a system of $N + 1$ linear equations for precisely the same number of unknown quantities $u_0, u_1, \ldots, u_N$. Therefore it is necessary to establish that the determinant of this system is different from zero. As we know from algebra, the determinant of a system is different from zero if and only if the corresponding homogeneous system has only an identically vanishing solution. But for the system (1), (2) the homogeneous system is obtained by setting $\phi = \psi = f_m \equiv 0$. From bound (7), which has been proven for every solution $\{u_n\}$, it will be seen that in this case there exists only the trival solution $u_n \equiv 0$.

The following condition, also, is sufficient to guarantee that problem (1), (2) is well-conditioned:

$$\frac{|b_n| - |a_n| - |c_n|}{|b_n| + |a_n| + |c_n|} \geq \theta > 0, \qquad \max \{|a_n|, |b_n|, |c_n|\} \geq B > 0, \tag{8}$$

where θ and B are constants not depending on N or n. In fact from (8) we get (6) with the constant

$$\delta = \theta(|b_n| + |a_n| + |c_n|) \geq \theta B > 0.$$

For this reason (7) takes the form

$$|u_n| \leq \max\{|\phi|, |\psi|, \frac{1}{\theta B} \max_m |f_m|\}. \tag{9}$$

4. Criterion for a well-conditioned boundary-value problem with constant coefficients.

Theorem. *In order that the boundary-value problem*

$$\left.\begin{aligned} au_{n-1} + bu_n + cu_{n+1} &= f_n, \qquad 0 < n < N, \\ u_0 &= \phi, \qquad u_N = \psi \end{aligned}\right\} \tag{10}$$

with constant coefficients be well-conditioned it is necessary and sufficient that one of the roots, q_1 and q_2, of the characteristic equation

$$a + bq + cq^2 = 0 \tag{11}$$

should be greater than, and the other smaller than one in modulus, i.e., that they should satisfy an inequality of the form

$$|q_1| \leq 1 - \frac{\theta}{2}, \qquad |q_2^{-1}| \leq 1 - \frac{\theta}{2}, \tag{12}$$

where θ is some positive constant.

If the coefficients a, b and c are real the criterion for good conditioning, Eq. (2), by virtue of what has been shown in 3§3, can be put into the convenient form:

$$\frac{|b| - |a + c|}{|b| + |a| + |c|} \geq \theta > 0. \tag{13}$$

The convenience of criterion (13) consists in that fulfillment of this criterion can be checked without computing the roots q_1 and q_2.

Criterion (12) will be derived in 4§6, below.

5. Criterion for a well-conditioned problem with variable coefficients. Criterion (12), which guarantees a well-conditioned boundary-value problem for difference equations with constant coefficients, the criterion formulated in the preceding section, can be generalized to cover the problem

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n, \qquad 0 < n < N, \tag{1}$$

$$u_0 = \phi, \qquad u_N = \psi \tag{2}$$

with variable coefficients so long as these coefficients vary sufficiently "smoothly". We will formulate this generalization exactly assuming that the coefficients of (1) are bounded in totality, $|a_n| < M$, $|b_n| < M$, $|c_n| < M$, and that its coefficients a_n, b_n and c_n do not become small, simultaneously, for any n:

$$d_n = \max \{|a_n|, |b_n|, |c_n|\} \geq B > 0.$$

The constants M and B , above, are not to depend on N or n.

Theorem. *Suppose that the coefficients of problem (1), (2) satisfy the conditions*

$$\left.\begin{aligned} |a_k - a_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad |b_k - b_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \\ |c_k - c_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad D > 0, \qquad \omega > 0. \end{aligned}\right\} \tag{14}$$

Then, to guarantee that problem (1), (2) is well-conditioned, it is necessary and sufficient that the roots, q_1 and q_2, of the quadratic

$$a_n + b_n q + c_n q^2 = 0, \qquad 0 < n < N, \tag{15}$$

satisfy a condition of the form

$$|q_1| < 1 - \frac{\theta}{2}, \qquad \left|q_2^{-1}\right| < 1 - \frac{\theta}{2}, \tag{16}$$

where $\theta > 0$ is some number not depending on N or n.

Conditions (14) express the requirement that the coefficients be smooth. They are fulfilled, for example, if

$$a_n = a(n/N), \qquad b_n = b(n/N), \qquad c_n = c(n/N),$$

where a(x), b(x) and c(x) are any functions defined on the interval $0 \leq x \leq 1$, and satisfying the Hölder conditions:

$$\begin{aligned} |a(x) - a(x')| &\leq D|x - x'|^\omega, \\ |b(x) - b(x')| &\leq D|x - x'|^\omega, \\ |c(x) - c(x')| &\leq D|x - x'|^\omega. \end{aligned}$$

Equation (15) is the characteristic equation constructed for the difference equation

$$au_{s-1} + bu_s + cu_{s+1} = 0$$

with constant coefficients a, b and c, coinciding in value with the variable coefficients a_n, b_n and c_n for some fixed n, i.e., $a = a_n$, $b = b_n$, $c = c_n$.

If a_n, b_n and c_n are real coefficients then, by virtue of 3§3, condition (16) may be replaced by the easily verifiable condition

$$\frac{|b_n| - |a_n + c_n|}{|b_n| + |a_n| + |c_n|} \geq \theta > 0, \tag{17}$$

where θ does not depend on N or n.

The validity of criterion (14), (16) or (14), (17) will be proven in §6. There also it will be shown that smoothness conditions (14) must not be ignored.

Note that if $|a_n + c_n| = |a_n| + |c_n|$, condition (17) is identically the same as condition (8) and guarantees good conditioning even without the assumed smoothness and reality of the coefficients.

6. Justification of the criterion for a well-conditioned boundary-value problem with constant coefficients. We will now prove the validity of the criterion, derived in part 4, for good conditioning of the boundary-value problem

$$\left.\begin{aligned} au_{n-1} + bu_n + cu_{n+1} &= f_n, \qquad 0 < n < N, \\ u_0 &= \phi, \qquad u_N = \psi, \end{aligned}\right\} \tag{10}$$

i.e., more specifically we prove the following assertion. In order that problem (10) be well-conditioned it is necessary and sufficient that the roots of the characteristic equation

$$a + bq + cq^2 = 0 \tag{11}$$

satisfy inequalities of the form

$$|q_1| \leq 1 - \frac{\theta}{2}, \qquad \left|q_2^{-1}\right| \leq 1 - \frac{\theta}{2}, \tag{12}$$

where θ is some positive constant.

Sufficiency. We represent the solution of problem (1) as the sum of two net functions, writing

$$u_n = \bar{u}_n + \tilde{u}_n, \tag{18}$$

where $\{u_n\}$ is the solution of the problem

$$\left.\begin{aligned} a\bar{u}_{n-1} + b\bar{u}_n + c\bar{u}_{n+1} &= f_n, \qquad 0 < n < N, \\ \bar{u}_0 &= \phi, \qquad \bar{u}_N = \psi, \end{aligned}\right\} \tag{19}$$

and $\{\tilde{u}_n\}$ the solution of

$$\left.\begin{aligned} a\tilde{u}_{n-1} + b\tilde{u}_n + c\tilde{u}_{n+1} &= f_n, \qquad 0 < n < N, \\ \tilde{u}_0 &= 0, \qquad \tilde{u}_N = 0. \end{aligned}\right\} \tag{20}$$

The solution of problem (19) has the form

$$\bar{u}_n = Aq_1^n + Bq_2^n,$$

where A and B are determined via the conditions $\bar{u}_0 = \phi$, $\bar{u}_N = \psi$:

$$\bar{u}_n = \frac{\phi - \psi q_2^{-N}}{1 - (q_1 q^{-1})^N} q_1^n + \frac{\psi - \phi q_1^N}{1 - (q_1 q_2^{-1})^N} q_2^{n-N}. \tag{21}$$

Defining $\rho = 1 - \theta/2$ we get, from (21),

$$|\bar{u}_n| \le 2 \frac{\max(\rho^n \rho^{N-n})}{1 - \rho^{2N}} \max(|\phi|, |\psi|). \tag{22}$$

Therefore for all $N \ge 2$ and $n = 0, 1, \ldots, N$,

$$|\bar{u}_n| \le \frac{2}{1 - \rho} \max(|\phi|, |\psi|) = \frac{4}{\theta} \max(|\phi|, |\psi|). \tag{23}$$

If n and N - n are taken large enough the coefficients in inequality (22) can be made arbitrarily small. For example for $n > 6/\theta$, $N - n > 6/\theta$

$$\rho^{6/\theta} = \left[\left(1 - \frac{\theta}{2}\right)^{2/\theta}\right]^3 < \left(\frac{4}{9}\right)^3.$$

Here we have used the well known inequality*

$$\left(1 - \frac{1}{a}\right)^a \left(1 + \frac{1}{b}\right)^b \le \left[\frac{a(1 - a^{-1}) + b(1 + b^{-1})}{a + b}\right]^{a+b} = 1$$

* A simple proof of the inequality $(1 - a^{-1})^a(1 + b^{-1})^b < 1$, a, b, > 1, may be outlined as follows: $\ln[(1 - a^{-1})^a] = a \cdot \ln(1 - a^{-1}) < a \cdot (-a)^{-1} = -1$, $(1 - a^{-1})^a < e^{-1}$. Similarly $(1 + b^{-1})^b < e$. (Translator's note.)

for $a = 2/\theta$, $b = 2$. Thus

$$\frac{\max(\rho^n, \rho^{N-n})}{1 - \rho^{2N}} \leq \left(\frac{4}{9}\right)^3 \frac{1}{1 - (4/9)^6} < \frac{1}{10},$$

so that from (22), for $n > 6/\theta$, $N - n > 6/\theta$ we get

$$|\tilde{u}_n| < \frac{1}{5} \max(|\phi|, |\psi|). \tag{24}$$

We now bound the solution $\{\tilde{u}_n\}$ of problem (2). First we represent $\tilde{u}_n$ as the sum

$$\tilde{u}_n = u_n^* + u_n', \qquad 0 \leq n \leq N,$$

of the solutions of two problems - the problem

$$au_{n-1}^* + bu_n^* + cu_{n+1}^* = \begin{cases} f_n, & 0 < n < N, \\ 0, & n \leq 0 \text{ or } n \geq N, \end{cases} \tag{25}$$

and the problem

$$\left.\begin{aligned} au_{n-1}' + bu_n' + cu_{n+1}' &= 0, \qquad 0 < n < N, \\ u_0' = -u_0^*, &\qquad u_N' = -u_N^*. \end{aligned}\right\} \tag{26}$$

A bounded solution $\{u_n^*\}$ of problem (25) exists, is unique, and is subject to bound (15) §3:

$$|u_n^*| \leq \frac{16}{B\theta^2} \max_m |f_m|, \tag{27}$$

where $B = \max(|a|, |b|, |c|)$.

In particular

$$\left.\begin{aligned} |u_0^*| &\leq \frac{16}{B\theta^2} \max_m |f_m|, \\ |u_N^*| &\leq \frac{16}{B\theta^2} \max_m |f_m|. \end{aligned}\right\} \tag{27'}$$

For a bound on the solution $\{u_n'\}$ of problem (26), a problem of the same form as (19), we use Eq. (21) and bound (23), simply substituting $-u_0^*$ and $-u_N^*$ for ϕ and ψ:

$$|u_n'| \leq \frac{4}{\theta} \max(|u_0^*|, |u_N^*|).$$

Now in addition taking note of (27'):

$$|u_n'| \leq \frac{64}{B\theta^3} \max_m |f_m|. \tag{28}$$

Combining bounds (27) and (28), taking into account that $\theta < 2$, we get

$$\left|\tilde{u}_n\right| \leq \frac{128}{B\theta^3} \max_m |f_m|. \tag{29}$$

Consequently, for the solution $\{u_n\}$ of the original problem, combining bounds (23) and (29), we get

$$|u_n| \leq \left|\bar{u}_n\right| + \left|\tilde{u}_n\right| \leq \frac{128}{B\theta^3} \max_m |f_m| + \frac{4}{\theta} \max(|\phi|, |\psi|). \tag{30}$$

Bound (30) guarantees good conditioning, $|u_n| \leq M \max(|\phi|, |\psi|, \max|f_m|)$, where one may take for M

$$M = \frac{128}{B\theta^3} + \frac{4}{\theta}.$$

In the case $n > 6/\theta$, $N - n > 6/\theta$, one can sharpen bound (30) using, in place of inequality (23), inequality (24):

$$|u_n| \leq \frac{128}{B\theta^3} \max_m |f_m| + \frac{1}{5} \max(|\phi|, |\psi|) \tag{31}$$

or

$$|u_n| < M_1 \max_m |f_m| + \frac{1}{5} \max(|\phi|, |\psi|), \tag{31'}$$

where M_1 depends only on θ and B, not on N. Estimate (31) will be used in §6.

Necessity. We note, first, that if condition (12) is not fulfilled for any positive θ, than the roots of the characteristic equation

$$P(q) \equiv a + bq - cq^2 = 0$$

are, in modulus, either both less than one, or both greater than one, or at least one of them is equal to one:

1) $$|q_1| < \rho < 1, \quad |q_2| < \rho < 1, \tag{32}$$

2) $$|q_1| > \rho > 1, \quad |q_2| > \rho > 1, \tag{33}$$

3) $$|q_1| = 1. \tag{34}$$

We will show that, in all three cases, good conditioning is absent.

* * * * * *

For this purpose in all three cases we construct functions, $\{u_n\}$, which solve a problem of the form

$$\left.\begin{aligned} au_{n-1} + bu_n + cu_{n+1} &= f_n, \qquad 0 < n < N, \\ u_0 &= u_N = 0 \end{aligned}\right\} \tag{35}$$

and satisfy the inequalities

$$\max_n |u_n| > M_N \max_m |f_m|, \tag{36}$$

where M_N is a quantity growing without bound as $N \to \infty$.

In case (32), assuming for the sake of definiteness that $q_1 \neq q_2$,* we postulate that

$$u_n = \begin{cases} q_1^n - q_2^n, & 0 \leq n \leq N-1, \\ 0, & n = N. \end{cases}$$

Then

$$\max_n |u_n| \geq |u_1| = |q_1 - q_2| > 0. \tag{37}$$

The right-hand side $\{f_n\}$ of problem (35) is

$$f_n = au_{n-1} + bu_n + cu_{n+1} = \begin{cases} 0, & \text{for } n \neq N-1. \\ c(q_1^N - q_2^N), & \text{for } n = N-1. \end{cases}$$

Hence

$$\max_m |f_m| = |f_{N-1}| \leq 2|c|\rho^N. \tag{38}$$

Comparing (37) and (38), we see that in (36) we must take

$$M_N = \frac{|q_1 - q_2|}{2|c|\rho^N} = O\left(\frac{1}{\rho^N}\right),$$

so that M_N grows exponentially with increasing N. Case (33) is analogous to (32).

---

* The case $q_1 = q_2 = q$ can be treated by setting $u_n = nq^n$, $0 \leq n \leq N-1$, $u_N = 0$, with corresponding modifications in the following steps. (Translator's note.)

If (34) is satisfied we set

$$u_n = q_1^n \sin \frac{n\pi}{N}, \qquad 0 \le n \le N$$

Then, clearly,

$$\max_n |u_n| \ge \frac{1}{2}. \tag{39}$$

For $|f_n|$ we get the bound

$$|f_n| = |au_{n-1} + bu_n + cu_{n+1}| =$$

$$= \left|\left(aq_1^{n-1} + bq_1^n + cq_1^{n+1}\right) \sin \frac{n\pi}{N} + aq_1^{n-1}\left(\sin \frac{(n-1)\pi}{N} - \sin \frac{n\pi}{N}\right) +\right.$$

$$\left. + cq_1^{n+1}\left(\sin \frac{(n+1)\pi}{N} - \sin \frac{n\pi}{BN}\right)\right| = \left|aq_1^{n-1}\left(\sin \frac{(n-1)\pi}{N} - \sin \frac{n\pi}{N}\right) +\right.$$

$$\left. + cq_1^{n+1}\left(\sin \frac{(n+1)\pi}{N} - \sin \frac{n\pi}{N}\right)\right| \le (|a| + |c|) \frac{\pi}{N}. \tag{40}$$

From (39) and (40) it follows that inequality (36) is satisfied if

$$M_N = \frac{N}{2\pi(|a| + |c|)}.$$

Thus good conditioning is absent, if we require of a well-conditioned problem that M be independent of N in inequality (5).

7. General boundary-value problem for a system of difference equations. Problem (1), (2) is only the simplest boundary-value problem for an equation of second order. We now state without proof necessary and sufficient conditions for a well-conditioned general boundary-value problem involving systems of difference equations on a net interval (V. S. Ryaben'kii, Computational Mathematics and Mathematical Physics **4**, 2, p. 43 (1964)).

A boundary value problem is, basically, a search for a vector-function $\{u_n\}$, n = 0, 1, 2, 3, ..., N, satisfying the conditions

$$\sum_{k=-k_0}^{k_0} A_{k,n} u_{n+k} = f_n, \qquad k_0 \le n \le N - k_0, \tag{1'}$$

$$\sum_{i=0}^{2k_0} \alpha_i u_i = \phi, \qquad \sum_{i=0}^{2k_0} \beta_i u_{N-i} = \psi. \tag{2'}$$

Here $A_{k,n}$ is a square matrix of some order $m \geq 1$; u_n and f_n are vectors of this same dimensionality; the α_i are matrices, each with m columns and $r \geq 0$ rows; the β_i are matrices with m columns and $s \geq 0$ rows; ϕ is a given r-dimensional vector; ψ is a given s-dimensional vector.

Problem (1′), (2′) is well-conditioned if it has a solution for arbitrary $\{f_n\}$, ϕ and ψ, with

$$\max_n \left|\left|u_n\right|\right| \leq M \max\left\{\left|\left|\phi\right|\right|, \left|\left|\psi\right|\right|, \max_j \left|\left|f_j\right|\right|\right\},$$

where M does not depend on N.

With respect to coefficients $A_{k,n}$, we will postulate that

$$A_{k,n} = A_k\left(\frac{n}{N}\right),$$

where $A_k(x)$ is a matrix, defined on the interval $0 \leq x \leq 1$, and satisfying on this interval the smoothness condition

$$\left|\left|A_k(x) - A_k(x')\right|\right| \leq D\left|x - x'\right|^{\omega}, \quad D > 0, \quad \omega > 0. \qquad (14')$$

Further, we will assume that

$$d(x) = \max_k \left|\left|A_k(x)\right|\right| \geq B > 0.$$

Given these restrictions then, to guarantee that problem (1′), (2′) is well-conditioned it is necessary and sufficient that each of the following conditions, 1°-3°, should be satisfied:

1° Among the roots μ and ν of the equations

$$\det \sum_{k=-k_0}^{k_0} A_k(x)\mu^{k_0+k} = 0$$

$$\det \sum_{k=-k_0}^{k_0} A_k(x)\nu^{k_0-k} = 0$$

none are equal to one in modulus, while each of the roots μ and ν of these equations satisfies one of the following four inequalities:

$$|\mu| < 1 - \frac{\theta}{2}, \qquad |\nu| < 1 - \frac{\theta}{2},$$

$$|\mu^{-1}| < 1 - \frac{\theta}{2}, \qquad |\nu^{-1}| < 1 - \frac{\theta}{2},$$

where $\theta > 0$ does not depend on x.

2° The dimensionality, r, of the matrices α_i, is equal to the number of roots μ the moduli of which are less than unity, and the dimensionality, s, of matrices β_i is equal to the number of roots ν the moduli of which are less than unity.

3° Among the solutions $\{u_n\}$, $n \geq 0$, of the problem

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} A_k(0)u_{n+k} = 0, \qquad k_0 \leq n < \infty, \\ &\sum_{i=0}^{2k_0} \alpha_i u_i = 0 \end{aligned}\right\}$$

and among the solutions $\{u_n\}$, $n \geq 0$, of the problem

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} A_k(1)u_{n+k} = 0, \qquad -\infty < n \leq N - k_0, \\ &\sum_{i=0}^{2k_0} \beta_i u_{N-i} = 0 \end{aligned}\right\}$$

none are bounded except the trivial solution.

This last condition, 3°, can be put into the form of a requirement that certain determinants, with elements independent of N, must not vanish.

We illustrate the above criteria by studying their application to the problem

$$\left.\begin{aligned} &0 \cdot u_{n-1} - 2u_n + u_{n+1} = f_n, \qquad 0 < n < N, \\ &u_1 - \alpha u_0 = \phi, \qquad u_N - \beta u_{N-1} = \psi, \end{aligned}\right\}$$

where α and β are given; here $m = 1$, $r = 1$, $s = 1$, $k_0 = 1$. The roots of the equations

$$0 - 2\mu + \mu^2 = 0 \qquad \text{and} \qquad 0 \cdot \nu^2 - 2\nu + 1 = 0$$

are equal to

$$\mu_1 = 0, \qquad \mu_2 = 2, \qquad \nu_1 = 1/2 \qquad (\nu_2 = \infty).$$

None are equal to unity in modulus, and condition 1° is satisfied.

Condition 2° is also satisfied, since the number of scalar boundary conditions on the left- and right-hand boundaries are equal, $r = s = 1$, and equal to the number of roots μ and ν which are less than one in modulus.

Let us now determine for what values of α the problem

$$\left.\begin{aligned} 0 \cdot u_{n-1} - 2u_n + u_{n+1} &= 0, \qquad n \geq 1, \\ \alpha u_0 - u_1 &= 0, \end{aligned}\right\}$$

has no nontrivial, bounded solution. The general form of the solution of the problem

$$0 \cdot u_{n-1} - 2u_n + u_{n+1} = 0, \qquad n > 0$$

is

$$u_n = c_1\mu_1^n + c_2\mu_2^n, \qquad n \geq 0, \qquad \mu_1^0 = 1.$$

From the boundedness condition we find that $c_2 = 0$. Therefore

$$u_n = c_1\mu_1^n = \begin{cases} c_1, & \text{for } n = 0, \\ 0, & \text{for } n > 0. \end{cases}$$

Taking account of the condition $\alpha u_0 - u_1 = 0$, we see that for $\alpha \neq 0$ there are no nontrivial solutions, while nontrivial solutions do exist for $\alpha = 0$.

Now we determine for which β the problem

$$\left.\begin{aligned} 0 \cdot u_{n-1} - 2u_n + u_{n+1} &= 0, \qquad n < N, \\ u_N - \beta u_{N-1} &= 0 \end{aligned}\right\}$$

has no bounded nontrivial solutions as $n \to -\infty$. The general solution of $0 \cdot u_{n-1} - 2u_n + u_{n+1} = 0$, $n < N$, is $u_n = c_1\nu^{-n} = c_1(1/2)^{-n} = c_1 2^n$. It is bounded for $n \to -\infty$. From the boundary condition $u_N - \beta u_{N-1} = 0$ we see that

$$c_1 2^N - \beta c_1 2^{N-1} = c_1 2^{N-1}(2 - \beta) = 0$$

and a nontrivial solution, $c_1 \neq 0$, exists only for $\beta = 2$.

Thus the above boundary-value problem is well-conditioned for any $\alpha \neq 0$ and $\beta \neq 2$. If $\alpha = 0$ or $\beta = 2$ the problem is not well-conditioned.

* * *

PROBLEMS

The difference boundary-value problem

$$\left.\begin{aligned} au_{n-1} + bu_n + cu_{n+1} &= f_n, \qquad 0 < n < N, \\ u_0 - \alpha u_1 &= \phi, \qquad u_N - \beta u_{N-1} = \psi \end{aligned}\right\} \qquad (*)$$

will be called "well-conditioned" if it has one and only solution for any N, and if the quantities $u_0, u_1, \ldots, u_N$, constituting the solution $\{u_n\}$, satisfy the inequality $|u_n| < M \max\left(|\phi|, |\psi|, \max_m |f_m|\right)$, where M does not depend on N.

1. If both roots, q_1 and q_2, of the characteristic equation $a + bq + cq^2 = 0$ are less than (greater than) unity in modulus the difference boundary-value problem (*) cannot be well-conditioned. For simplicity take $q_1 \neq q_2$. Prove.

2. If at least one of the roots, q_1, q_2, of the characteristic equation is equal to one in modulus, then the difference boundary-value problem (*) cannot be well conditioned. Prove.

3. If $|q_1| < 1$, $|q_2| > 1$, but

$$1 - \alpha q_1 = 0 \quad \text{or} \quad 1 - \beta q_2 = 0,$$

then problem (*) cannot be well-conditioned. Prove.

4. To guarantee that the difference boundary-value problem (*) is well-conditioned it is necessary and sufficient that one root of the characteristic equation be smaller than one in modulus, $|q_1| < 1$, while the second is greater than one, and that $1 - \alpha q_1 \neq 0$, $1 - \beta q_2 \neq 0$. Prove.

5. The problem with constant (complex) coefficients

$$au_{n-1} + bu_n + cu_{n+1} = f_n, \qquad n = 0, \pm 1, \ldots$$

with arbitrary periodic right-hand side

$$f_{n+N} = f_n$$

has, for all sufficiently large N, the periodic solution $\{u_n\}$, $u_{n+N} = u_n$, satisfying the bound

$$|u_n| \leq M \max_m |f_m|,$$

where M does not depend on N or $\{f_n\}$, if neither of the roots of the characteristic equation, $a + bq + cq^2$, is equal to one in modulus. Prove.

§5. Algorithm for the solution of boundary-value problems - forward elimination, back substitution (FEBS).

1. Description of forward elimination, back substitution (FEBS).* We now describe a simple, convenient method for the solution of the difference boundary-value problem of the form considered in §4:

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad 0 < n < N, \\ u_0 &= \phi, \qquad u_N = \psi. \end{aligned}\right\} \tag{1}$$

It is one variant of Gauss elimination and is called "forward elimination, back substitution (FEBS)".

Let us write the equation $u_0 = \phi$ of system (1) in the form

$$u_0 = L_{1/2}u_1 + K_{1/2},$$

where $L_{1/2} = 0$ and $K_{1/2} = \phi$. From the equation

$$a_1 u_0 + b_1 u_1 + c_1 u_2 = f_1,$$

corresponding to the system (1) equation with $n = 1$, we eliminate u_0 with the aid of the equation $u_0 = L_{1/2}u_1 + K_{1/2}$. We then write the result in the form solved for u_1,

$$u_1 = L_{3/2}u_2 + K_{3/2},$$

introducing the notation

$$L_{3/2} = \frac{-c_1}{b_1}, \qquad K_{3/2} = \frac{a_1\phi - f_1}{-b_1}.$$

The relation $u_1 = L_{3/2}u_2 + K_{3/2}$ can now be used to eliminate u_1 from the equation

$$a_2 u_1 + b_2 u_2 + c_2 u_3 = f_2,$$

---

* This method is often referred to, alternatively, as "Cholesky factorization". However, because of the way the method is presented here, this name seems inappropriate. One is then forced back to the akward name "forward elimination, back substitution" which, because it will be used so frequently below, it seems adviseable to abbreviate. (Translator's note.)

corresponding to n = 2. We again write the result of this elimination in a form explicit with respect to u_2,

$$u_2 = L_{5/2}u_3 + K_{5/2}.$$

The above elimination process can be continued for n = 3, 4, ...

Substituting

$$u_{n-1} = L_{n-1/2}u_n + K_{n-1/2}$$

in the equation

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n,$$

we get

$$u_n = \frac{-c_n}{b_n + a_n L_{n-1/2}} u_{n+1} + \frac{f_n - a_n K_{n-1/2}}{b_n + a_n L_{n-1/2}}.$$

Hence it is clear that the coefficients obtained in the course of the elimination process

$$u_n = L_{n+1/2}u_{n+1} + K_{n+1/2}$$

can be computed via the recurrence relations

$$\left.\begin{aligned} L_{n+1/2} &= \frac{-c_n}{b_n + a_n L_{n-1/2}}, \\ K_{n+1/2} &= \frac{f_n - a_n K_{n-1/2}}{b_n + a_n L_{n-1/2}}. \end{aligned}\right\} \qquad (2)$$

The last of the relations obtained in this way has the form

$$u_{N-1} = L_{N-1/2}u_N + K_{N-1/2}.$$

Since $u_N = \psi$, it is now possible to compute u_{N-1}:

$$u_{N-1} = L_{N-1/2}\psi + K_{N-1/2}.$$

The other unknowns u_{N-2}, u_{N-3}, etc., are determined, respectively, from the equations

$$u_{N-2} = L_{N-3/2}u_{N-1} + K_{N-3/2},$$

$$u_{N-3} = L_{N-5/2}u_{N-2} + K_{N-5/2},$$

and so on, until u_1 is determined.

Let us review, briefly, the basic features of the process just described. First one calculates the coefficients $L_{n+1/2}$, $K_{n+1/2}$ in order of increasing n (forward elimination) via the recurrence relation (2), with $L_{1/2} = 0$ and $K_{1/2} = \phi$ given. Then the computation of the unknowns, u_n, is carried out, also recurrently, in order of decreasing n (back substitution), through use of the equations

$$\left.\begin{aligned} u_N &= \psi, \\ u_n &= L_{n+1/2}u_{n+1} + K_{n+1/2}, \qquad n = N-1, N-2, \ldots, 1. \end{aligned}\right\} \tag{3}$$

Note that to compute, via FEBS, the solution $u_0, u_1, \ldots, u_N$ of system (1), consisting of N+1 equations, one must execute arithmetic operations whose number is larger only by a finite factor than the number of unknowns. To solve an arbitrary linear system of N equations with N unknowns by Gauss elimination ordinarily requires a number of arithmetic operations of order N^3. Such a reduction in the number of arithmetic operations, through solution of (1) via FEBS, has been attained by successful exploitation of the detailed structure of this system.

In §7 it will be shown that when solving, via FEBS, a boundary-value problem (1) satisfying one of the good-conditioning criteria

$$|b_n| \geq |a_n| + |c_n| + \delta, \qquad \delta > 0, \tag{4}$$

or

$$\frac{|b_n| - |a_n| - |c_n|}{|b_n| + |a_n| + |c_n|} \geq \theta > 0, \qquad d_n = \max(|a_n|, |b_n|, |c_n|) \geq B > 0,$$

or

$$\frac{|b_n| - |a_n + c_n|}{|b_n| + |a_n| + |c_n|} \geq \theta > 0, \qquad d_n \geq B > 0,$$

$$|a_k - a_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad |b_k - b_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega,$$

$$|c_k - c_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad D > 0, \qquad \omega > 0,$$

discussed in §4, the expression $b_n + a_n L_{n-1/2}$, which must be used as a divisor, cannot vanish; and, further, the computational errors don't accumulate and don't produce errors increasing with increasing N in the computed solution.

These two noteable properties of FEBS - the small number of arithmetic operations required and the weak sensitivity to computational errors, make FEBS a very useful computational algorithm.

2. Example of a computationally unstable algorithm. For the solution of a well-conditioned difference boundary-value problem (1) various algorithms could be used. We have described the FEBS algorithm, which has the advantages that it requires a small number of arithmetic operations and is computationally stable. We now describe another, still simpler, algorithm which, however, is computationally unstable and practically unuseable for large N.

Given $U_0^{(1)} = \phi$, $U_1^{(1)} = 0$, we find the solution $U^{(1)} = \{U_n^{(1)}\}$, $n = 0, 1, \ldots, N$, of difference equation (1). Naturally, in general $U_N^{(1)} \neq \psi$. Given $U_0^{(2)} = \phi$, $U_1^{(2)} = 1$, we compute the solution $U^{(2)} = \{u_n^{(2)}\}$. This solution also does not satisfy the right-hand boundary condition. Now we postulate that

$$u_n = \sigma U_n^{(1)} + (1 - \sigma)U_n^{(2)}, \qquad n = 0, 1, \ldots, N. \tag{5}$$

Clearly for any σ the condition $u_0 = \phi$ is obeyed and Eq. (1) is satisfied. We now choose σ such as to satisfy the condition

$$u_N = \sigma U_N^{(1)} + (1 - \sigma)U_N^{(2)} = \psi,$$

that is we set

$$\sigma = \frac{\psi - U_N^{(2)}}{U_N^{(1)} - U_N^{(2)}},$$

and get the required solution of (1) from Eq. (5).

If the calculation were carried out on an ideal (necessarily imaginary) machine exactly, then this would be a good algorithm. But we now show that it's sensitivity to rounding errors for a well-conditioned problem (1) grows rapidly as $N \to \infty$. For this purpose we take as an example $a_n \equiv 1$, $b_n \equiv -26/5$, $c_n \equiv 1$, $f_n \equiv 0$.

Condition (4) for a well-conditioned problem is satisfied. In this case the exact solution of the difference boundary-value problem is given by the expression

$$u_n = \frac{5^{N-n} - 5^{n-N}}{5^N - 5^{-N}}\phi + \frac{5^n - 5^{-n}}{5^N - 5^{-N}}\psi. \tag{6}$$

For $U_n^{(1)}$ and $U_n^{(2)}$, by virtue of (5) §3, we have

$$U_n^{(1)} = -\frac{\phi}{24}\,5^n + \frac{\phi}{24}\,5^{2-n},$$

$$U_n^{(2)} = \frac{5-\phi}{24}\,5^n + \left[5 - \frac{25}{24}(5-\phi)\right]5^{-n}.$$

Note that the values of $\max_n \left|U_n^{(1)}\right|$ and $\max_n \left|U_n^{(2)}\right|$ grow like 5^N. For this reason, for large N, in the computation of $U_n^{(1)}$ and $U_n^{(2)}$ the calculated numbers will go out of the allowed range. But suppose this didn't happen, and that we have computed $\{U_n^{(1)}\}$ and $\{U_n^{(2)}\}$ and σ exactly. Suppose that the only rounding error is an error, ε, incurred in computing $1 - \sigma$. Then via Eq. (5) we get, in place of $\{u_n\}$

$$\{u_n + \Delta u_n\},$$

where $\Delta u_n = \varepsilon U_n^{(2)}$.

The error $\{\Delta u_n\}$ for $n \sim N$ will have the form

$$\Delta u_n \sim 5^N \varepsilon$$

and for a fixed relative error ε, committed in the computation of $1 - \sigma$, will quickly grow and "swamp" the exact solution $\{u_n\}$ which, according to Eq. (6), remains bounded.

The method just described is called the "shooting method".* In other situations (see §20) it may turn out to be stable and completely effective.

PROBLEMS

1. How must one change the FEBS algorithm so as to use it for computation of the solution $\{u_n\}$, $0 \le n \le N$, of the difference equation

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n, \qquad 0 < n < N,$$

with boundary conditions of the form

$$u_0 - \alpha u_1 = \phi, \qquad u_N - \beta u_{N-1} = \psi,$$

---

* The same method is often referred to as "marching". (Translators note.)

if α and β are different from zero?

2. In computing the solution of the problem

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad 0 < n < N \\ u_0 = \phi, \qquad u_N &= \psi \end{aligned}\right\}$$

it would have been possible to carry out the elimination in the direction of decreasing n. Write out the recursion relations for the computation of the coefficients $\tilde{K}_{n+1/2}$, $\tilde{L}_{n+1/2}$ of the corresponding elimination-substitution relations

$$u_{N+1} = \tilde{L}_{n+1/2} u_n + \tilde{K}_{n+1/2}, \qquad n = N-1, N-2, \ldots, 0.$$

3. Subjecting the coefficients a_n, b_n and c_n of the difference equation to the constraints $a_n > 0$, $c_n > 0$, $-b_n > a_n + c_n + \delta$, show that the FEBS coefficient $L_{n-1/2}$, occurring in the solution of the problem

$$\left.\begin{aligned} u_0 &= \alpha u_1 + \phi, \qquad 0 < \alpha < 1, \\ a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad 0 < n < N, \\ u_{N-1} &= \beta u_N + \psi, \end{aligned}\right\}$$

satisfies the inequality $0 \leq L_{n-1/2} \leq 1$. How does this fact influence error-accumulation in the back substitution? Is it possible, here, that a denominator in the forward elimination recursion relations will vanish?

4. Which variant of FEBS should one choose for the computation of the solution of the preceding problem if $\alpha = 10$, $\beta = -0.5$? In answering, consider the danger of dividing by zero in the recursive computation of the coefficients in the FEBS equations.

Chapter 3

Basis of the FEBS Method*

§6. Properties of well-conditioned boundary-value problems.

Here we prove the criterion, formulated in 5§4, for good conditioning of a difference boundary-value problem of the form

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad 0 < n < N, \\ u_0 = \phi, \qquad u_N &= \psi \end{aligned}\right\} \tag{1}$$

and establish several properties of well-conditioned difference boundary-value problems, so as to use these properties in §7 to provide a foundation for the FEBS algorithm.

1. Bound for the solution of a boundary-value problem with perturbed coefficients. Consider a problem of form (1)

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad p < n < q, \\ u_p = \phi, \qquad u_q &= \psi, \end{aligned}\right\} \tag{1'}$$

where p and q $\geq$ p + 2 are integers. The fact that we number the components of the solution from p to q, and not 0 to N, is not essential, but turns out to be convenient later. As regards the coefficients, we assume that they are bounded in totality: $|a_n|, |b_n|, |c_n| < M_1$, and M_1 does not depend on N or n.

Suppose that problem (1') is solveable for arbitrary ϕ, ψ and $\{f_n\}$, while the quantities $u_p, u_{p+1}, \ldots, u_q$, constituting the solution, satisfy the inequality

$$|u_n| \leq M_1 \max_m |f_m| + M_2 \max(|\phi|, |\psi|), \tag{2}$$

where M_1 and M_2 are positive constants, $M_1 \geq M_2$, $M_1 \geq 1$.

---

* The material of Chapter 3 is not used in the following chapters, and may be omitted on first reading.

Consider the problem

$$\begin{aligned} &\tilde{a}_n\tilde{u}_{n-1} + \tilde{b}_n\tilde{u}_n + \tilde{c}_n\tilde{u}_{n+1} = f_n, \qquad p < n < q, \\ &\tilde{u}_p = \phi, \qquad \tilde{u}_q = \psi. \end{aligned} \tag{3}$$

If we postulate that the perturbations in the coefficients, $\tilde{a}_n - a_n$, $\tilde{b}_n - b_n$, $\tilde{c}_n - c_n$, are not too great, or more precisely

$$\left.\begin{aligned} \left|\tilde{a}_n - a_n\right| &< \varepsilon < \frac{1}{6M_1}, \\ \left|\tilde{b}_n - b_n\right| &< \varepsilon < \frac{1}{6M_1}, \\ \left|\tilde{c}_n - c_n\right| &< \varepsilon < \frac{1}{6M_1}, \end{aligned}\right\} \tag{4}$$

then the perturbed system (3) will have the following four properties:

1° Problem (3) will have a solution $\{\tilde{u}_n\}$ for any right-hand side.

2° The solution $\{\tilde{u}_n\}$ will satisfy a bound of form (2), but with $2M_1$ and $2M_2$, respectively, in place of M_1 and M_2:

$$\left|\tilde{u}_n\right| \le 2M_1 \max_m |f_m| + 2M_2 \max(|\phi|, |\psi|). \tag{5}$$

3° The coefficients $\tilde{a}_n$, $\tilde{b}_n$ and $\tilde{c}_n$ will satisfy the bounds

$$\left|\tilde{a}_n\right| < |a_n| + \frac{1}{6M_1}, \quad \left|\tilde{b}_n\right| < |b_n| + \frac{1}{6M_1}, \quad \left|\tilde{c}_n\right| < |c_n| + \frac{1}{6M_1}.$$

4° The solutions $\{u_n\}$ and $\{\tilde{u}_n\}$ will differ only slightly from each other, and more precisely

$$\left|\tilde{u}_n - u_n\right| \le \varepsilon\left[6M_1^2 \max_m |f_m| + 6M_1M_2 \max(|\phi|, |\psi|)\right]. \tag{6}$$

Property 3° is obvious. We will prove 2° and, from it, derive 1°. Suppose that system (3) is solvable for some right-hand sides. For these given right-hand sides we will define

$$\mu = \max_k \left|\tilde{u}_k\right|$$

and will get, for μ, the inequality

$$\mu \le 2M_1 \max_m |f_m| + 2M_2 \max(|\phi|, |\psi|). \tag{7}$$

For this purpose we rewrite (3) as follows:

$$\left.\begin{aligned} &\tilde{a}_n\tilde{u}_{n-1} + \tilde{b}_n\tilde{u}_n + \tilde{c}_n\tilde{u}_{n+1} = \tilde{f}_n + (a_n - \tilde{a}_n)\tilde{u}_{n-1} + \\ &+ (b_n - \tilde{b}_n)\tilde{u}_n + (c_n - \tilde{c}_n)\tilde{u}_{n+1}, \qquad 0 < n < N, \\ &\tilde{u}_0 = \phi, \qquad \tilde{u}_N = \psi. \end{aligned}\right\} \tag{8}$$

From this expression, and from bounds (2) and (4), comes the inequality

$$\mu \le M_1\left(\max_m |f_m| + \frac{3}{6M_1}\mu\right) + M_2 \max(|\phi|, |\psi|) \le$$

$$\le \frac{1}{2}\mu + M_1 \max_m |f_m| + M_2 \max(|\phi|, |\psi|).$$

Solving this latter inequality with respect to μ, we get (7) and, hence, (5).

From inequality (5) it follows that the homogeneous system corresponding to problem (3), and obtained from it by setting $\phi = \psi = f_n \equiv 0$, has only the vanishing solution $\tilde{u}_n \equiv 0$. Thus the determinant of coefficients of (3) is different from zero, and problem (3) has one and only one solution for any arbitrary right-hand sides. Properties 1° and 2° are proven. It remains only to prove property 4°, i.e. inequality (6).

Subtracting, term by term, Eq. (1) from Eq. (8), we get

$$\left.\begin{aligned} &a_n(\tilde{u}_{n-1} - u_{n-1}) + b_n(\tilde{u}_n - u_n) + c_n(\tilde{u}_{n+1} - u_{n+1}) = \\ &= (a_n - \tilde{a}_n)\tilde{u}_{n-1} + (b_n - \tilde{b}_n)\tilde{u}_n + (c_n - \tilde{c}_n)\tilde{u}_{n+1}, \qquad 0 < n < N, \\ &\tilde{u}_0 - u_0 = 0, \qquad \tilde{u}_N - u_N = 0. \end{aligned}\right\}$$

Applying (2)

$$\left|\tilde{u}_n - u_n\right| \le M_1 \max_m \left|(\tilde{a}_m - a_m)\tilde{u}_{m-1} + (\tilde{b}_m - b_m)\tilde{u}_m + (\tilde{c}_m - c_m)\tilde{u}_{m+1}\right|,$$

from which, now applying (4) and (5), we derive

$$\left|\tilde{u}_n - u_n\right| \le M_1\varepsilon[3 \cdot 2M_1 \max_m |f_m| + 3 \cdot 2M_2 \max(|\phi|, |\psi|)],$$

i.e., inequality (6).

Now consider the problem obtained from (1′) by perturbing, not only the coefficients, but also the right-hand-sides

$$\left.\begin{aligned} &\tilde{a}_n\tilde{u}_{n-1} + \tilde{b}_n\tilde{u}_n + \tilde{c}_n\tilde{u}_{n+1} = \tilde{f}_n, \qquad p < n < q, \\ &\tilde{u}_p = \tilde{\phi}, \qquad \tilde{u}_q = \tilde{\psi}. \end{aligned}\right\} \tag{9}$$

One can show that

$$\left|\tilde{u}_n - u_n\right| \leq \varepsilon\left[6M_1^2 \max_m \left|\tilde{f}_m\right| + 6M_1M_2 \max(|\tilde{\phi}|, |\tilde{\psi}|)\right] +$$

$$+ M_2 \max\left(|\tilde{\phi} - \phi|, |\tilde{\psi} - \psi|\right) + M_1 \max_m \left|\tilde{f}_m - f_m\right| . \qquad (10)$$

We will only sketch an outline of the proof, which can easily be carried out following this outline.

First changing only the right-hand sides and leaving the coefficients unaltered we see, with the aid of (2), that each u_n changes by no more than

$$M_1 \max_m \left|\tilde{f}_m - f_m\right| + M_2 \max\left(|\tilde{\phi} - \phi|, |\tilde{\psi} - \psi|\right).$$

Then changing the coefficients in the equation system with the altered right-hand sides we find that, by virtue of property 4°, the components u_n change by an additional amount not exceeding

$$\varepsilon\left[6M_1^2 \max_m \left|\tilde{f}_m\right| + 6M_1M_2 \max_m \left(|\phi|, |\psi|\right)\right],$$

which, indeed, leads to bound (10).

We now derive, from those consequences of inequality (2) which have already been discussed, one further consequence. Suppose that for the solution of system (1′) we have, for some $\lambda > 0$, $p + \lambda < n < q - \lambda$, the bound

$$|u_n| < M_1 \max_m |f_m| + M_2' \max_m \left(|\phi|, |\psi|\right).$$

Then the solution of the perturbed system

$$\left.\begin{aligned} \tilde{a}_n\tilde{u}_{n-1} + \tilde{b}_n\tilde{u}_n + \tilde{c}_n\tilde{u}_{n+1} &= f_n, \qquad p < n < q, \\ \tilde{u}_p = \phi, \qquad \tilde{u}_q &= \psi, \end{aligned}\right\}$$

subject to the conditions

$$\left|\tilde{a}_n - a_n\right|, \left|\tilde{b}_n - b_n\right|, \left|\tilde{c}_n - c\right| < \varepsilon < \frac{1}{24M_1^2} < \frac{1}{6M_1}, \qquad (11)$$

satisfies, also for $p + \lambda < n < q - \lambda$, the inequality

$$\left|\tilde{u}_n\right| \leq 2M_1 \max_m |f_m| + \left(M_2' + \frac{1}{f}\right) \max\left(|\phi|, |\psi|\right). \qquad (12)$$

To convince ourselves of this we define the auxiliary net function $\{v_n\}$ as the solution of the system

$$a_n v_{n-1} + b_n v_n + c_n v_{n+1} = 0, \qquad p < n < q,$$

$$v_p = \phi, \qquad v_q = \psi.$$

For $p + \lambda < n < q - \lambda$

$$|v_n| < M_2' \max(|\phi|, |\psi|). \tag{13}$$

Next we use, for a bound on $\left|\tilde{u}_n - v_n\right|$, inequality (10), from which it follows, taking account of (11), that

$$\left|\tilde{u}_n - v_n\right| \leq$$

$$\leq \varepsilon\left[6M_1^2 \max_m |f_m| + 6M_1M_2 \max(|\phi|, |\psi|)\right] + M_1 \max_m |f_m| \leq$$

$$\leq \frac{1}{4}\max(|\phi|, |\psi|) + 2M_1 \max_m |f_m|.$$

Now, through use of (13), we immediately get inequality (12).

Note. It is important to stress that the quantity ε in bound (4), defining the limits within which one can perturb the coefficients without violating solveability, and also the coefficients in bound (5) on the solution of the perturbed problem, and in bounds (6) and (10) on the deviation between the solutions of the perturbed and unperturbed problems - all these quantities depend only on the coefficients M_1 and M_2 in bound (2). The specific values of the coefficients of the difference equation, and the number of points $q - p + 1$, in themselves play no role: their influence acts only through the agency of the constants M_1 and M_2 which render bound (2) valid.

2. Proof of the criterion for good-conditioning. In 5§4 we formulated criteria for a well-conditioned problem (1) with coefficients satisfying the smoothness conditions

$$\left.\begin{aligned} |a_k - a_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad |b_k - b_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \\ |c_k - c_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad D > 0, \qquad \omega > 0, \end{aligned}\right\} \tag{14}$$

along with the conditions

$$\left.\begin{aligned} d_n = \max(|a_n|, |b_n|, |c_n|) \geq B > 0, \\ |a_n| \leq M_1, \qquad |b_n| \leq M_1, \qquad |c_n| \leq M_1. \end{aligned}\right\} \tag{14'}$$

To guarantee that problem (1) is well-conditioned, given (14) and (14′), it is necessary and sufficient that the roots of the quadratic

$$a_n + b_n q + c_n q^2 = 0$$

satisfy the inequalities

$$|q_1| \leq 1 - \frac{\theta}{2}, \qquad \left|q_2^{-1}\right| \leq 1 - \frac{\theta}{2}, \tag{15}$$

where $\theta > 0$ does not depend on N or n.

The necessity of this criterion can be proven by roughly the same methods as were used, in 4§4, when dealing with the case of constant coefficients, and we will not stop to consider this question further.

To prove sufficiency we will use the criterion, discussed in 6§4, for a well-conditioned difference boundary-value problem

$$\left.\begin{aligned} au_{n-1} + bu_n + cu_{n+1} &= f_n, \qquad p < n < q, \\ u_p = \phi, \qquad u_q &= \psi \end{aligned}\right\} \tag{16}$$

with constant coefficients, where p and q, $q \geq p + 2$, are arbitrary integers. In contrast to what was done in §4 we number the components of the solution $\{u_n\}$, not with n = 0, 1, ..., N, but with the numbers n = p, p + 1, ..., q, which changes nothing essential. Problem (16) always has a solution and, moreover, for all n such that $p \leq n \leq q$, bound (30) §4 is valid, i.e.

$$|u_n| \leq M_1 \max_m |f_m| + M_2 \max(|\phi|, |\psi|), \qquad p \leq n \leq q, \tag{17}$$

and, for n such that $p + 6/\theta < n < q - 6/\theta$, we have bound (31) §4:

$$|u_n| \leq M_1 \max_m |f_m| + M_2' \max(|\phi|, |\psi|), \tag{18}$$

where

$$M_1 = \frac{128}{B\theta^2}, \qquad M_2 = \frac{4}{\theta}, \qquad M_2' = \frac{1}{5}.$$

We will choose ε such that

$$\varepsilon = \frac{1}{24M_1^2} \tag{19}$$

and take N large enough so that the inequality

$$D\left|\frac{24}{N\theta}\right|^{\omega} < \varepsilon, \qquad \text{i.e.} \qquad N > \frac{24}{\theta}\left|\frac{D}{\varepsilon}\right|^{1/\omega} \tag{20}$$

is satisfied.

We proceed, now, to prove that problem (1) is well-conditioned. Consider a boundary-value problem of the form

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad p < n < q, \\ u_p = \phi, \qquad u_q &= \psi, \end{aligned}\right\} \tag{21}$$

where p and q are arbitrary, given, integers such that $0 \leq p,\ q \leq N$, $q \geq p + 2$. In the special case $p = 0$, $q = N$ this problem becomes identical with problem (1), and in general it is obtained from problem (1) by "truncation" -- i.e. by discarding the equations for $n \leq p$ and $n \geq q$, and fixing u_p and u_q. We will show that, given any N obeying condition (20), problem (21) has one and only one solution for any arbitrary right-hand sides, and further that the quantities $\{u_n\}$, $p \leq n \leq q$, satisfy a bound of the form

$$|u_n| \leq M \max\left(|\phi|, |\psi|, \max_m |f_m|\right), \tag{22}$$

where M is some constant depending on B and θ, but not on N, p or q.

We consider separately the case $q - p \leq 24/\theta$ and the case $q - p > 24/\theta$.

If $q - p \leq 24/\theta$, then the coefficients of problem (21), for any k and ℓ such that $p \leq k$, $\ell \leq q$, will (by virtue of the smoothness conditions, (14), and the fact that N obeys (20)) satisfy the bounds

$$|a_k - a_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega \leq D\left|\frac{q-p}{N}\right|^\omega \leq D\left|\frac{24}{\theta N}\right|^\omega < \varepsilon,$$

$$|b_k - b_\ell| < \varepsilon, \qquad |c_k - c_\ell| < \varepsilon.$$

These coefficients are "almost" constant, and differ by no more than ε from the coefficients of problem (16) choosing, as a, b and c in (16), the coefficients a_{p+1}, b_{p+1} and c_{p+1}. The solution of problem (16) satisfies bound (17). Here ε is chosen according to Eq. (19), satisfying requirement (4). Therefore to bound the solution of problem (21) one can use inequality (5):

$$|u_n| \leq \frac{256}{B\theta^3} \max_m |f_m| + \frac{8}{\theta} \max(|\phi|, |\psi|). \tag{23}$$

We consider now the case $q - p > 24/\theta$ for example for $p = 0$, $q = N$. Suppose that, for some fixed θ, ψ and $\{f_n\}$, there exists a solution $\{u_n\}$, $p \leq n \leq q$. Choose a sequence of integers, $p = N_0 < N_1 < \ldots < N_r = q$, such that the inequality

$$\frac{6}{\theta} < N_{k+1} - N_k < \frac{12}{\theta} \tag{24}$$

will be satisfied. The solution of the problem with constant coefficients

$$\left.\begin{aligned} av_{n-1} + bv_n + cv_{n+1} &= f_n, \qquad N_{k-1} < n < N_{k+1}, \\ v_{N_{k-1}} &= \phi, \qquad v_{N_{k+1}} = \psi, \end{aligned}\right\} \tag{25}$$

where

$$a = a_{N_k}, \qquad b = b_{N_k}, \qquad c = c_{N_k},$$

for $n = N_k$, by virtue of the inequality

$$N_{k-1} + \frac{6}{\theta} < N_k < N_{k+1} - \frac{6}{\theta}$$

satisfies bound (18)

$$\left|v_{N_k}\right| \leq M_1 \max_m |f_m| + M_2' \max(|\phi_k|, |\psi_k|),$$

where

$$M_1 = \frac{128}{B\theta^3}, \qquad M_2' = \frac{1}{5}.$$

The problem

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad N_{k-1} < n < N_{k+1}, \\ u_{N_{k-1}} &= \phi, \qquad u_{N_{k+1}} = \psi \end{aligned}\right\} \tag{26}$$

can be considered a perturbed version of problem (25), while the coefficients of (26), by virtue of the inequality $N_{k+1} - N_{k-1} \leq 24/\theta$, differ by no more than ε from the coefficients of problem (25). Bound (12) can now be used for the solution of the perturbed problem. For $n = N_k$ one gets

$$\left|u_{N_k}\right| \leq 2M_1 \max_m |f_m| + \left(M_2' + \frac{1}{4}\right) \max\left(\left|u_{N_{k-1}}\right|, \left|u_{N_{k+1}}\right|\right) \leq$$

$$\leq 2M_1 \max_m |f_m| + \frac{1}{2} \max\left(\left|u_{N_{k-1}}\right|, \left|u_{N_{k+1}}\right|\right).$$

Consequently

$$\max_{0 < k < r} \left|u_{N_k}\right| \leq 2M_1 \max_m |f_m| + \frac{1}{2} \max\left[|\phi|, |\psi|, \max_{0 < k < r} \left|u_{N_k}\right|\right] \leq$$

$$\leq 2M_1 \max_m |f_m| + \frac{1}{2} \max_{0 < k < r} \left|u_{N_k}\right| + \frac{1}{2} \max(|\phi|, |\psi|).$$

Hence

$$\max_{0 < k < r} \left|u_{N_k}\right| \leq 4M_1 \max_m |f_m| + \max(|\phi|, |\psi|).$$

Now, for any arbitrary n, we find N_{k-1} and N_{k+1} between which it lies and use bound (23):

$$|u_n| \leq 2M_1 \max_m |f_m| + 2M_2 \max\left(\left|u_{N_{k-1}}\right|, \left|u_{N_{k+1}}\right|\right) \leq$$

$$\leq 2M_1 \max_m |f_m| + 2M_2[4M_1 \max_m |f_m| + \max(|\phi|, |\psi|)] \leq$$

$$\leq (2M_1 + 8M_1M_2) \max_m |f_m| + 2M_2 \max(|\phi|, |\psi|). \qquad (27)$$

Bound (27), obtained for $q - p > 24/\theta$, by virtue of (23) remains valid also for $q - p \leq 24/\theta$. Problem (21) is solveable for any arbitrary right-hand sides since, as can be seen from bound (27), for $\phi = \psi = f_m \equiv 0$ there exists only the trivial solution.

We have completed the proof that, given the smoothness conditions (14) and conditions (14′), condition (15) is a criterion for good conditioning of problem (1). The following example shows that the smoothness condition (14) cannot be ignored.

It's easy to verify that the difference boundary-value problem

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} = 0, \qquad 0 < n < N, \\ u_0 = 0, \qquad u_N = 0, \end{aligned}\right\}$$

where $a_n \equiv 1$, $b_n \equiv (-1)^n$, $c_n \equiv 1$ and $N = 6N_1$ has, for any positive integer N_1, the nontrivial solution

$$u_n = \begin{cases} \sin \dfrac{n\pi}{6}, & \text{if } n \text{ is even,} \\ -\cos \dfrac{n\pi}{6}, & \text{if } n \text{ is odd.} \end{cases}$$

Consequently this boundary-value problem is not well-conditioned, despite the fact that

$$\frac{|b_n| - |a_n + c_n|}{|b_n| + |a_n| + |c_n|} = \frac{1}{3}, \qquad |a_n| = |b_n| = |c_n| = 1,$$

i.e.

$$|q_1| < 1 - \frac{1}{6}, \qquad \left|q_2^{-1}\right| < 1 - \frac{1}{6}.$$

3. Properties of a well-conditioned problem. We now formulate the results obtained in §4 and in section 2 above, on good conditioning of problem (1), in a form convenient for use in the investigation of FEBS in §7.

In order that the difference boundary-value problem (1) be well-conditioned it is sufficient that one of the following three criteria be satisfied:

first criterion:

$$|b_n| \geq |a_n| + |c_n| + \delta, \qquad \delta > 0;$$

second criterion:

$$\frac{|b_n| - |a_n| - |c_n|}{|b_n| + |a_n| + |c_n|} \geq \theta > 0, \qquad d_n = \max(|a_n|, |b_n|, |c_n|) \geq B > 0;$$

third criterion:

$$\frac{|b_n| + |a_n + c_n|}{|b_n| + |a_n| + |c_n|} \geq \theta > 0, \qquad M \geq d_n \geq B > 0,$$

where it is assumed that the coefficients are real and satisfy the smoothness conditions (14)

$$|a_k - a_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad |b_k - b_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega,$$

$$|c_k - c_\ell| \leq D\left|\frac{k-\ell}{N}\right|^\omega, \qquad D > 0, \qquad \omega > 0.$$

If either of the first two criteria is satisfied problem (1) is solveable for $N \geq 2$ and for arbitrary right-hand sides, and if the third is satisfied then problem (1) is solveable for all large enough N and arbitrary right-hand sides. Also, for these same large N, in addition to problem (1) all truncated boundary-value problems of form (21) are also solevable.

The solution $\{u_n\}$ of the original problem and the solutions $\{u_n\}$ of all truncated problems satisfy the bound

$$|u_n| \leq M \max\left(|\phi|, |\psi|, \max_m |f_m|\right), \qquad p \leq n \leq q,$$

where M does not depend on N, p, or q.

§7. Basis for the FEBS method in well-conditioned boundary-value problems.

Now we are ready for the study of the FEBS method, described in §5. Suppose one is required to calculate the solution of the difference boundary-value problem

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad 0 < n < N, \\ u_0 &= \phi, \qquad u_n = \psi, \\ |a_n|, \ |b_n|, \ |c_n| &< M. \end{aligned}\right\} \tag{1}$$

With respect to this problem we postulate that it itself, and all problems derived from it through truncation

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \qquad p < n < q, \\ u_p &= \phi, \qquad u_q = \psi, \end{aligned}\right\}$$

have a solution for any arbitrary right-hand sides, and moreover

$$|u_n| \leq M \max\left(|\phi|, \ |\psi|, \ \max_m |f_m|\right). \tag{2}$$

In studying the FEBS algorithm we will use the fact that, by virtue of the bounds (4) and (5) of 1§6 the difference problem with perturbed coefficients

$$\left.\begin{aligned} \tilde{a}_n \tilde{u}_{n-1} + \tilde{b}_n \tilde{u}_n + \tilde{c}_n \tilde{u}_{n+1} &= f_n, \qquad 0 < n < N, \\ \tilde{u}_0 &= \phi, \qquad \tilde{u}_N = \psi, \\ \left|\tilde{a}_n - a_n\right|, \quad \left|\tilde{b}_n - b_n\right|, \quad \left|\tilde{c}_n - c_n\right| &\leq \varepsilon < \frac{1}{6M}, \end{aligned}\right\} \tag{3}$$

as well as all problems derived from (3) by truncation, have solutions $\{\tilde{u}_n\}$ for arbitrary right-hand sides, and further

$$\left|\tilde{u}_n\right| \leq 2M \max\left(|\phi|, \ |\psi|, \ \max_m |f_m|\right). \tag{4}$$

1. Bounds on the FEBS coefficients. Here we show that, in computing the FEBS coefficients, one never is led to divide by zero, and we arrive at bounds on the FEBS coefficients, bounds valid for the original problem (1), as well as the perturbed problem (3). For this purpose it is sufficient to consider only the perturbed problem, since the original problem is a special case of the perturbed problem (for $\varepsilon = 0$).

Consider the following truncated system

$$\left.\begin{aligned} \tilde{a}_n\tilde{u}_{n-1} + \tilde{b}_n\tilde{u}_n + \tilde{c}_n\tilde{u}_{n+1} &= f_n, \qquad 0 < n < \ell, \\ \tilde{u}_0 = \phi, \qquad \tilde{u}_\ell &= \psi. \end{aligned}\right\}$$

It is solveable. From it we deduce $\tilde{u}_{\ell-1}$. From Kramer's rule for the solution of a system of linear algebraic equations it follows that $\tilde{u}_{\ell-1}$ may be written in the form

$$\tilde{u}_{\ell-1} = L\psi + \sum_{i=1}^{\ell-1} \Lambda_i f_i + \Lambda_o\phi = L\tilde{u}_\ell + K, \tag{5}$$

where L and Λ_i depend only on $\tilde{a}_n$, $\tilde{b}_n$ and $\tilde{c}_n$. As a consequence of bound (4) (valid for any ϕ, ψ and $\{f_m\}$, and therefore for $\phi = 0$, $f_i \equiv 0$, $\psi = 1$) it follows that

$$|L| = \left|\tilde{u}_{\ell-1}\right| \leq 2M,$$

and taking $\tilde{u}_\ell = \psi = 0$ it follows that

$$|K| = \left|\tilde{u}_{\ell-1}\right| \leq 2M \max\left(|\phi|, \max_m |f_m|\right).$$

It is convenient to assign to L and K the index $\ell - \frac{1}{2}$ and to write the above relations and inequalities in the form

$$\left.\begin{aligned} \tilde{u}_{\ell-1} &= L_{\ell-1/2}\tilde{u}_\ell + K_{\ell-1/2}, \\ |L_{\ell-1/2}| \leq 2M, \qquad |K_{\ell-1/2}| &\leq 2M \max\left(|\phi|, \max_m |f_m|\right). \end{aligned}\right\} \tag{6}$$

A relation of this form was obtained in the development of FEBS in §5. From Kramer's rule (5) it will be seen that $L_{\ell-1/2}$ is uniquely determined by $\tilde{a}_n$, $\tilde{b}_n$ and $\tilde{c}_n$, while $K_{\ell-1/2}$ is uniquely determined by ϕ, f_n, $\tilde{a}_n$, $\tilde{b}_n$ and $\tilde{c}_n$ $(0 < n < \ell)$. Hence it follows that the coefficients $L_{\ell-1/2}$ and $K_{\ell-1/2}$ coincide with the FEBS coefficients obtained in §5 where we wrote out, for these coefficients, the recursion relations

$$\left.\begin{aligned} L_{1/2} &= 0, \qquad K_{1/2} = \phi. \\ L_{\ell+1/2} &= \frac{-\tilde{c}_\ell}{\tilde{b}_\ell + \tilde{a}_\ell L_{\ell-1/2}}, \qquad K_{\ell+1/2} = \frac{f_\ell - \tilde{a}_\ell K_{\ell-1/2}}{\tilde{b}_\ell + \tilde{a}_\ell L_{\ell-1/2}}. \end{aligned}\right\} \tag{7}$$

Of course this last assertion is valid only if the recurrence formulae are meaningful, i.e. only if none of the denominators in these expressions vanishes.

We now show that, in fact, none of the denominators does vanish.

Suppose that we have shown that it is possible, via Eqs. (7), to compute

$$\left.\begin{array}{l} L_{1/2},\ L_{3/2},\ \ldots,\ L_{\ell-1/2} \\ K_{1/2},\ K_{3/2},\ \ldots,\ K_{\ell-1/2}, \end{array}\right\} \quad \ell \geq 1;$$

we now verify the validity of these equations also for $L_{\ell+1/2}$ and $K_{\ell+1/2}$. For this purpose it is sufficient to show that

$$\left|\tilde{b}_\ell + \tilde{a}_\ell L_{\ell-1/2}\right| \geq \frac{1}{2M}. \tag{8}$$

Consider the system of equations

$$\left.\begin{array}{rl} \tilde{u}_0 = 0, & \\ \tilde{a}_i\tilde{u}_{i-1} + \tilde{b}_i\tilde{u}_i + \tilde{c}_i\tilde{u}_{i+1} = 0, & i = 1, 2, \ldots, \ell-1, \\ \tilde{a}_\ell\tilde{u}_{\ell-1} + \tilde{b}_\ell\tilde{u}_\ell + \tilde{c}_\ell\tilde{u}_{\ell+1} = 1, & \\ \tilde{u}_{\ell+1} = 0. & \end{array}\right\} \tag{9}$$

As concerns the solution of this system, we know that it exists. From the first ℓ (homogeneous) equations it follows that $\tilde{u}_{\ell-1} = L_{\ell-1/2}\tilde{u}_\ell$. From (4) it follows that $|\tilde{u}_\ell| \leq 2M$. From the only inhomogeneous equation contained in system (9) it follows that

$$(\tilde{a}_\ell L_{\ell-1/2} + \tilde{b}_\ell)\tilde{u}_\ell = 1.$$

For this reason

$$\frac{1}{\left|\tilde{b}_\ell + \tilde{a}_\ell L_{\ell-1/2}\right|} \leq 2M,$$

which indeed proves bound (8), together with the fact that recurrence relations (8) and bound (6) are meaningful.

2. Estimate of the influence on computational results of rounding errors committed in the course of the calculation. We will now solve problem (1) by the FEBS method. In real computations at each step of the computational process one commits computational errors induced by roundoff. For this reason the real computational process is governed by the equations

$$
\left.
\begin{aligned}
L_{1/2} &= 0, \qquad K_{1/2} = \phi + \kappa_{1/2}, \\
L_{\ell+1/2} &= \frac{-c_\ell}{a_\ell L_{\ell-1/2} + b_\ell} + \lambda_{\ell+1/2}, \qquad \ell = 1, 2, \ldots, N-1, \\
K_{\ell+1/2} &= \frac{f_\ell - a_\ell K_{\ell-1/2}}{a_\ell L_{\ell-1/2} + b_\ell} + \kappa_{\ell+1/2}, \qquad \ell = 1, 2, \ldots, N-1, \\
u_N &= \psi + \nu_N, \\
u_\ell &= L_{\ell+1/2} u_{\ell+1} + K_{\ell+1/2} + \nu_\ell, \qquad \ell = N-1, N-2, \ldots, 1.
\end{aligned}
\right\} \quad (10)
$$

Suppose that all computational errors are subject to the bounds

$$|\kappa_{\ell+1/2}| < \delta, \qquad |\lambda_{\ell+1/2}| < \delta, \qquad |\nu_\ell| < \delta$$

with sufficiently small δ so that

$$\delta < \frac{1}{6M^2(2M+1)}.$$

We will show that in this case in the FEBS equations (10) none of the denominators will vanish, and will estimate to what extent these errors will distort the computational results.

We introduce the notation

$$K_{1/2} = \tilde{K}_{1/2}; \qquad K_{\ell+1/2} + \nu_\ell = \tilde{K}_{\ell+1/2}, \qquad \ell > 0.$$

Clearly the collection of equations (10) may be rewritten thus:

$$
\left.
\begin{aligned}
&L_{1/2} = 0, \qquad \tilde{K}_{1/2} = \phi + \kappa_{1/2},\\
&L_{\ell+1/2} = -\frac{c_\ell - (a_\ell L_{\ell-1/2} + b_\ell)\lambda_{\ell+1/2}}{a_\ell L_{\ell-1/2} + b_\ell}, \qquad \ell = 1, 2, \ldots, N-1,\\
&\tilde{K}_{\ell+1/2} = \frac{f_\ell - a_\ell(\tilde{K}_{\ell-1/2} - \nu_{\ell-1})}{a_\ell L_{\ell-1/2} + b_\ell} + \kappa_{\ell+1/2} + \nu_\ell =\\
&\quad = \frac{f_\ell + a_\ell \nu_{\ell-1} + (a_\ell L_{\ell-1/2} + b_\ell)(\kappa_{\ell+1/2} + \nu_\ell) - a_\ell \tilde{K}_{\ell-1/2}}{a_\ell L_{\ell-1/2} + b_\ell},\\
&\qquad\qquad \ell = 1, 2, \ldots, N-1,\\
&u_N = \psi + \nu_N,\\
&u_\ell = L_{\ell+1/2} u_{\ell+1} + \tilde{K}_{\ell+1/2}, \qquad \ell = N-1, N-2, \ldots, 1,
\end{aligned}
\right\} \tag{10'}
$$

and these equations may be regarded as the basis for a computational process designed for the solution of the difference boundary-value problem

$$
\tilde{a}_n \tilde{u}_{n-1} + \tilde{b}_n u_n + \tilde{c}_n \tilde{u}_{n+1} = \tilde{f}_n, \qquad 0 < n < N,
$$

$$
\tilde{u}_0 = \phi, \qquad \tilde{u}_N = \psi
$$

with the following perturbed right-hand sides and coefficients:

$$
\left.
\begin{aligned}
&\tilde{\phi} = \phi + \kappa_{1/2},\\
&\tilde{f}_\ell = f_\ell + a_\ell \nu_{\ell-1} + (a_\ell L_{\ell-1/2} + b_\ell)(\kappa_{\ell+1/2} + \nu_\ell),\\
&\tilde{\psi} = \psi + \nu_N,\\
&\tilde{a}_\ell = a_\ell, \quad \tilde{b}_\ell = b_\ell, \quad \tilde{c}_\ell = c_\ell - (a_\ell L_{\ell-1/2} + b_\ell)\lambda_{\ell+1/2}.
\end{aligned}
\right\} \tag{11}
$$

We will show that

$$
|\tilde{c}_\ell - c_\ell| \le M(2M+1)\delta < \frac{1}{6M}. \tag{12}
$$

The proof is by induction on ℓ. For $\ell = 1$

$$|\tilde{c}_1 - c_1| = |(a_1 L_{1/2} + b_1)\lambda_{3/2}| = |(a_1 \cdot 0 + b_1)\lambda_{3/2}| \leq M\delta \leq$$

$$\leq M(2M + 1)\delta < \frac{1}{6M} .$$

Suppose that for $k = 1, 2, \ldots, \ell-1$ inequality (12) has already been proven. In the computation of $L_{1/2}, L_{3/2}, \ldots, L_{\ell-1/2}$ one uses only $\tilde{a}_i = a_i$, $\tilde{b}_i = b_i$, and $\tilde{c}_i$ for $i = 1, 2, \ldots, \ell-1$. Therefore we can affirm, by virtue of (6), that $|L_{\ell-1/2}| < 2M$ and that, consequently,

$$|\tilde{c}_\ell - c_\ell| = |-(a_\ell L_{\ell-1/2} + b_\ell)\lambda_{\ell+1/2}| \leq (M \cdot 2M + M)\delta < \frac{1}{6M} .,$$

This completes the induction.

Thus it has been shown that, if $\delta \leq 1/[6M^2(2M+1)]$, then the inequalities

$$|\tilde{a}_n - a_n| = 0 < \frac{1}{6M} , \qquad |\tilde{b}_n - b_n| = 0 < \frac{1}{6M} , \qquad |\tilde{c}_n - c_n| < \frac{1}{6M} ,$$

are satisfied and, thus, bounds (6) and (8),

$$\left.\begin{aligned} |L_{\ell-1/2}| &< 2M, \\ |\tilde{a}_\ell L_{\ell-1/2} + \tilde{b}_\ell| &= |a_\ell L_{\ell-1/2} + b_\ell| \geq \frac{1}{2M} , \end{aligned}\right\} \qquad (13)$$

are valid. We see that, in executing the computational process implied by (10), we are never called upon to divide by zero.

Now from Eq. (11) for $\tilde{\phi}$, $\tilde{f}_\ell$, $\tilde{\psi}$, and from bound (13), it follows that

$$|\tilde{\phi} - \phi| < \delta, \qquad |\tilde{\psi} - \psi| < \delta,$$

$$|\tilde{f}_\ell - f_\ell| \leq M\delta + (M \cdot 2M + M)2\delta = M(4M + 3)\delta.$$

Thus, committing at each step of the computational process an error no larger than δ, $\delta < 1/[6M^2(2M + 1)]$, we can, by the process described, solve the problem with perturbed coefficients and right-hand sides.

These perturbations do not exceed $M^*\delta$, where

$$M^* = \max\{2, (4M + 3)M\}$$

depends only on M while, in addition, the perturbations in the coefficients don't exceed $1/(6M)$.

Such perturbations of the coefficients and right-hand sides lead, as is shown by bound (10) of §6, to errors in u_n not exceeding $M^{**}\delta$. Here M^{**} once again depends only on M. (If $M \approx N^r$, then $M^* \approx N^{2r}$, $M^{**} \approx N^{3r}$, so that the error in the solution will be $N^{3r}\delta$.)

If M, and thus also M**, is independent of N then making, in the course of the FEBS computations, an error of order δ at each step (the number of such steps being proportional to N) we get in the final solution an error no greater than const $\cdot \delta$.

Thus the influence, on the result, of an error committed in any single step of the calculation does not grow with increasing N. Further even the cumulative influence of all errors committed during all steps of the computation also does not grow.

This noteable property of FEBS has, indeed, been a main reason for its wide use.

This Page Intentionally Left Blank

Part 2
DIFFERENCE SCHEMES FOR ORDINARY DIFFERENTIAL EQUATIONS

Part 2 of this book is devoted to the construction and the study of difference schemes for ordinary differential equations. In the course of this study we introduce the concepts of convergence, approximation and stability, basic in the theory of difference schemes and general in character. Familiarity with these concepts, acquired in connection with ordinary differential equations, will permit us later, in the study of difference schemes for partial differential equations, to concentrate on the numerous peculiarities and difficulties characteristic of this most variegated class of problems.

Chapter 4
Elementary Examples of Difference Schemes

In this chapter we consider introductory examples of difference schemes, intended only to give the reader a preliminary acquaintance with the basic concepts of the theory.

§8. The concept of order of accuracy and of approximation

1. Order of accuracy of a difference scheme. This section is devoted to the question of the convergence of solutions of difference equations, with refinement of the net, to the solutions of the differential equations which they approximate. We limit ourselves, here, to the study of two difference schemes for the numerical solution of the problem

$$\left.\begin{aligned} &\frac{du}{dx} + au = 0, \qquad 0 \leq x \leq 1, \\ &u(0) = b. \end{aligned}\right\} \tag{1}$$

Let us begin with the simplest difference scheme, based on the use of the difference equation

$$\frac{u(x+h) - u(x)}{h} + Au(x) = 0. \tag{2}$$

We now subdivide the interval [0, 1] into steps of length h. It is convenient to take h = 1/N, where N is an integer. The points of

subdivision will be numbered from left to right, so that $x_n = nh$, $n = 0, 1, \ldots, N$. The value of u obtained, via the difference scheme, at point x_n will be denoted as u_n. We fix an initial value, u_0. Suppose that $u_0 = b$. From difference equation (2) one gets the relation

$$u_n = (1 - Ah)u_{n-1},$$

from which we find the solution of Eq. (2) subject to the initial condition $u_0 = b$:

$$u_n = (1 - Ah)^n b = (1 - Ah)^{(x_n/h)} b. \tag{3}$$

The *exact* solution of problem (1) has the form $u(x) = b \exp(-Ax)$. It takes on, at point x_n, the value

$$u(x_n) = be^{-Ax_n}. \tag{4}$$

We now estimate the error in the approximate solution (3). At point x_n this error is

$$\delta(x_n) = [(1 - Ah)^{(x_n/h)} - e^{-Ax_n}] b. \tag{5}$$

We are interested in the rate at which $\delta(x_n)$ decreases as the number of subdivision points increases or, equivalently, as one decreases the step-width, $h = 1/N$, of the difference net. To bring this out we represent $(1 - Ah)^{(x_n/h)}$ in the form

$$(1 - Ah)^{\frac{x_n}{h}} = e^{\frac{x_n}{h}\ln(1 - Ah)} = e^{\frac{x_n}{h}\left[-Ah + \frac{A^2h^2}{2} + O(h^3)\right]} =$$

$$= \left[e^{-Ax_n}\right]\left[e^{A^2\frac{h}{2}x_n}\right] e^{O(h^2)} = e^{-Ax_n}\left[1 + \frac{A^2hx_n}{2} + O(h^2)\right][1 + O(h^2)] =$$

$$= e^{-Ax_n} + h\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2).$$

Thus Eq. (3) takes the form

$$u_n = be^{-Ax_n} + hb\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2), \tag{3'}$$

so that

$$\delta(x_n) = hb\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2) = O(h), \tag{6}$$

i.e. the error (5) tends to zero as $h \to 0$, and the magnitude of the error is of the order of the first power of the step-size.

On this basis one says that *the difference scheme has first-order accuracy* (which is not to be confused with the order of the difference equation, defined in §1).

Let us now solve problem (1) with the aid of the difference equation

$$\frac{u(x+h) - u(x-h)}{2h} + Au(x) = 0. \tag{7}$$

This is not as simple as it may seem to be at first glance. The problem is that the above scheme is a difference equation of second-order, i.e. it requires the assignment of two initial conditions ($u(x_0) = 0$ and $u(x_1) = u(h)$); while the equation to be integrated, Eq. (1), is an equation of first order, and for it we need only the condition $u(0) = b$. It is natural, also in the difference scheme, to set $u_0 = b$.

It isn't clear, however, how one should choose u_1. To shed some light on this question we use the explicit form of the solution of Eq. (7) (see §3 Eq. (6)):

$$u_n = u_0\left[\frac{q_2}{q_2 - q_1} q_1^n - \frac{q_1}{q_2 - q_1} q_2^n\right] + u_1\left[-\frac{1}{q_2 - q_1} q_1^n + \frac{1}{q_2 - q_1} q_2^n\right] =$$

$$= \frac{q_2 u_0 - u_1}{q_2 - q_1} q_1^n - \frac{q_1 u_0 - u_1}{q_2 - q_1} q_2^n, \tag{8}$$

where

$$\left.\begin{aligned} q_1 &= \sqrt{1 + A^2h^2} - Ah = 1 - Ah + \frac{A^2h^2}{2} + O(h^4), \\ q_2 &= (-1)\left(1 + Ah + \frac{A^2h^2}{2}\right) + O(h^4). \end{aligned}\right\} \tag{9}$$

The Taylor expansions, (9), of the roots of the characteristic equation allow one to develop an approximate representation of q_1^n and q_2^n. We carry out a detailed derivation of such a representation for q_1^n:

$$q_1^n = q_1^{(x_n/h)} = \left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right]^{(x_n/h)} = e^{\frac{x_n}{h}\ln\left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right]}$$

Since $\ln(1 + z) = z - z^2/2 + z^3/3 + O(z^4)$,

$$\ln\left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right] = -Ah + \frac{A^3h^3}{6} + O(h^4).$$

Therefore

$$q_1^n = e^{\frac{x_n}{h}\left[-Ah + \frac{A^3h^3}{6} + O(h^4)\right]} = e^{-Ax_n}\left[1 + h^2\frac{A^3x_n}{6}\right] + O(h^3). \qquad (10)$$

We will not carry out the completely analogous computation for q_2^n, but go directly to the result:

$$q_2^n = (-1)^n e^{Ax_n} + O(h^2) \qquad (11)$$

Putting the approximate expressions for q_1^n and q_2^n into Eq. (8) we get

$$u_n = \frac{q_2u_0 - u_1}{q_2 - q_1}q_1^n - \frac{q_1u_0 - u_1}{q_2 - q_1}q_2^n =$$

$$= \frac{q_2u_0 - u_1}{q_2 - q_1}\left[e^{-Ax_n} + h^2\frac{A^3x_n}{6}e^{-Ax_n} + O(h^3)\right] -$$

$$- \frac{q_1u_0 - u_1}{q_2 - q_1}(-1)^n\left[e^{Ax_n} + O(h^2)\right]. \qquad (12)$$

All further conclusions will be obtained through study of this expression.

We note that if the coefficient $(q_2u_0 - u_1)/(q_2 - q_1)$ tends to the finite limit b as $h \to 0$ then the first term, $(q_2u_0 - u_1)q_1^n/(q_2 - q_1)$, on the right-hand side of Eq. (12) tends to the desired solution of problem (1).

Since

$$(-1)^n\left[e^{Ax_n} + O(h^2)\right] \xrightarrow[h \to 0]{} \begin{cases} e^{Ax_n} & \text{for n even,} \\ -e^{Ax_n} & \text{for n odd,} \end{cases}$$

i.e. does not converge to a definite limit, then to guarantee convergence to a limit, as $h \to 0$, of the second term on the right-hand side of (12),

$$\frac{q_1u_0 - u_1}{q_2 - q_1}(-1)^n\left[e^{Ax_n} + O(h^2)\right], \qquad (13)$$

it is necessary to require that the expression $(q_1u_0 - u_1)/(q_2 - q_1)$ tend to zero as $h \to 0$.

Let us, then summarize what has been said.

So that the solution of the difference equation

$$\frac{u(x + h) - u(x - h)}{2h} + Au(x) = 0$$

should converge to the solution $u = b \exp(-Ax)$ of the boundary-value problem (1), it is necessary that the conditions

$$\frac{q_1u_0 - u_1}{q_2 - q_1} \to 0, \qquad \frac{q_2u_0 - u_1}{q_2 - q_1} \to b. \tag{14}$$

be satisfied. Recall, further, that we chose to set u_0 equal to b. Condition (14) gives us a hint as to how we can assign u_1. It turns out that it is sufficient that $u_1 \to u_0$ as $h \to 0$. In fact $q_1 \to +1$, and $q_2 \to -1$ as $h \to 0$ and, therefore, as $h \to 0$

$$\frac{q_1u_0 - u_1}{q_2 - q_1} \to 0, \qquad \frac{q_2u_0 - u_1}{q_2 - q_1} \to b.$$

2. Speed of convergence of the solution of the difference equation. We now go on to a study of the speed of convergence for different specific choices of $u_1 \approx u(h)$.

To determine u(h) it is natural to make use of the Taylor series expansion of the solution of the differential equation $u' + Au = 0$. Using the fact that $u' = -Au$, we rewrite the Taylor series expression thus:

$$u(x_1) = u(0) - hAu(0) + O(h^2) = u(0)(1 - Ah) + O(h^2).$$

This equation is satisfied by the exact solution of the differential equation. In the approximate solution, limiting ourselves to two terms of this expansion, we can set

$$u_1 = u_0(1 - Ah).$$

If we have decided to take only one term we let

$$u_1 = u_0.$$

In the first case we commit, in the initial value u_1, an error of order h^2, in the second -- an error of order h.

Let use examine the speed of convergence in each of these two cases, for each of these assignments of initial values.

Assume

$$u_0 = b, \qquad u_1 = (1 - Ah)b. \tag{15}$$

Then (see Eq. (9))

$$\frac{q_1u_0 - u_1}{q_2 - q_1} = \frac{[1 - Ah + O(h^2)]b - (1 - Ah)b}{-2 + O(h^2)} = O(h^2),$$

$$\frac{q_2u_0 - u_1}{q_2 - q_1} = \frac{[-1 - Ah - \frac{A^2h^2}{2} + O(h^4)]b - (1 - Ah)b}{-2 + O(h^4)} = b + O(h^2). \tag{16}$$

Returning to Eq. (12), we easily come to the conclusion which has been our goal

$$u_n = be^{-Ax_n} + O(h^2). \tag{17}$$

This conclusion may be stated as follows. If the initial value u_1 is given correctly to order h^2, then the error in the solution will be order h^2, i.e. the difference scheme will be accurate to second order.

It can be shown that, even if we take, for u_1, its exact value $b \cdot \exp(-Ax_1)$, accuracy greater than order h^2 cannot be attained in the solution. We advise the reader to prove this assertion as an exercise. It is easy to show also that if, for u_0, we take not precisely b, but any quantity of the form $b + O(h^2)$, the speed of convergence will still be second order.

We now proceed to consider the second formulation of initial conditions we have set out to study. Suppose

$$u_1 = u_0 = b.$$

Now

$$\frac{q_1u_0 - u_1}{q_2 - q_1} = \frac{[1 - Ah + O(h^2)]b - b}{-2 + O(h^2)} = \frac{1}{2}Ahb + O(h^2),$$

$$\frac{q_2u_0 - u_1}{q_2 - q_1} = \frac{-[1 + Ah + O(h^2)]b - b}{-2 + O(h^2)} = b + \frac{1}{2}Ahb + O(h^2)$$

and, consequently,

$$u_n = \frac{q_2u_0 - u_1}{q_2 - q_1}[e^{-Ax_n} + O(h^2)] - \frac{q_1u_0 - u_1}{q_2 - q_1}(-1)^n[e^{Ax_n} + O(h^2)] =$$

$$= [b + \frac{1}{2}Ahb + O(h^2)][e^{-Ax_n} + O(h^2)] -$$

$$- (-1)^n[\frac{1}{2}Ahb + O(h^2)][e^{Ax_n} + O(h^2)] =$$

$$= be^{-Ax_n} + Ab\frac{e^{-Ax_n} - (-1)^n e^{Ax_n}}{2}h + O(h^2).$$

Thus if the error in initial values is of order h, then the error in the solution will also be of order h.

Let us now summarize what has been said. We have seen that the difference scheme examined earlier,

$$\frac{u(x + h) - u(x - h)}{2h} + Au(x) = 0,$$

as compared to the scheme

$$\frac{u(x + h) - u(x)}{h} + Au(x) = 0,$$

can give faster convergence, and more precisely convergence with remainder terms of order h^2, rather than order h as in the second of these schemes. In order to attain second order accuracy one must, having taken an exact u_0, choose a u_1 differing from the exact solution of the differential equation at point $x = x_0 + h$ by a quantity of order h^2. It can be shown that u_0 also need not be given exactly, but also may contain an error of order h^2. The speed of convergence is not thereby diminished. Refining the initial values up to order h^3 and higher does not result in an increase in the accuracy of the solution.

If the initial values are given with errors of order h, then the solution will contain an error of this same order.

3. Order of approximation. It is interesting to consider just what it is that renders the scheme

$$\frac{u(x + h) - u(x)}{h} + Au(x) = 0$$

less accurate than the scheme

$$\frac{u(x + h) - u(x - h)}{2h} + Au(x) = 0.$$

These schemes differ in the approximate expressions

$$\frac{u(x + h) - u(x)}{h} \quad \text{and} \quad \frac{u(x + h) - u(x - h)}{2h}$$

used for the derivative, du/dx, at point x. It is natural to assume that in the first scheme the derivative has been replaced by a less accurate expression than in the second. And this is, in fact, true. Let us substitute, for u(x + h) and u(x - h), their Taylor series expansions

$$u(x + h) = u(x) + u'(x)h + u''(x)\frac{h^2}{2} + u'''(x)\frac{h^3}{6} + O(h^4),$$

$$u(x - h) = u(x) - u'(x)h + u''(x)\frac{h^2}{2} - u'''(x)\frac{h^3}{6} + O(h^4).$$

Using these expressions we get

$$\frac{u(x+h)-u(x)}{h} = u'(x) + u''(x)\frac{h}{2} + O(h^2),$$

$$\frac{u(x+h)-u(x-h)}{2h} = u'(x) + u'''(x)\frac{h^2}{6} + O(h^4),$$

i.e. in the first case we have an approximation to the derivative of only first-order accuracy, and in the second -- of second order.

The examples we have considered might lead one to think that the order of convergence of solutions of difference equations can be taken to be equal to the order of approximation of the derivatives in the differential equation. It turns out, however, that in such a very general form, this hypothesis is untrue.

On those difference schemes for which it's validity will be proven it will be necessary to impose an essential restriction -- the requirement of stability. The necessity of this requirement will become clear as we consider the examples in the following section.

§ 9. Unstable difference schemes

1. Techniques for approximating the derivative. We now again consider difference schemes for the approximate integration of the simplest differential equation $u' + Au = 0$. As we have already seen, to construct a difference scheme approximating this equation it suffices to replace the derivative, u', by some sort of approximating difference expression. Thus we have examined schemes in which the derivative u' was replaced by

$$\frac{u(x+h)-u(x)}{h} \quad \text{or} \quad \frac{u(x+h)-u(x-h)}{2h}.$$

It is clear than any expression of the form

$$\mu\frac{u(x+h)-u(x-h)}{2h} + (1-\mu)\frac{u(x+h)-u(x)}{h}$$

will also approximate $u'(x)$. In fact let us substitute into this expression the Taylor series for $u(x+h)$ and $u(x-h)$:

$$u(x+h) = u(x) + u'(x)h + O(h^2),$$

$$u(x-h) = u(x) - u'(x)h + O(h^2).$$

We then get

$$\mu\frac{u(x+h)-u(x-h)}{2h} + (1-\mu)\frac{u(x+h)-u(x)}{h} =$$

$$= \mu\frac{[u(x)+u'(x)h+O(h^2)]-[u(x)-u'(x)h+O(h^2)]}{2h} +$$

$$+ (1-\mu)\frac{[u(x)+u'(x)h+O(h^2)]-u(x)}{h} = u'(x) + O(h).$$

Using this sort of approximation for the derivative one can derive a whole family of difference schemes depending on a numerical parameter.

These schemes will have the form

$$\mu \frac{u(x+h) - u(x-h)}{2h} + (1-\mu)\frac{u(x+h) - u(x)}{h} + Au(x) = 0. \qquad (1)$$

To each value of the parameter μ corresponds one such scheme. It was the study of those particular schemes for which $\mu = 0$ and $\mu = 1$ to which §8 was devoted.

2. Example of an unstable difference scheme. We now consider one more scheme of this form, obtained from (1) with $\mu = 4$:

$$4\frac{u(x+h) - u(x-h)}{2h} - 3\frac{u(x+h) - u(x)}{h} + Au(x) = 0. \qquad (2)$$

This scheme may be rewritten thus:

$$-2u(x-h) + (3 + Ah)u(x) - u(x+h) = 0. \qquad (2')$$

As in the examples considered earlier, we compute the solution on the interval $[0,1]$, subdivided by the points of the difference net into N equal steps, each of length $h = 1/N$. The coordinate, x_n, of a point of the net is defined as $x_n = nh = n/N$.

The solution of the difference equation may be written in the explicit form

$$u_n = u_0\left[\frac{q_2}{q_2 - q_1}q_1^n - \frac{q_1}{q_2 - q_1}q_2^n\right] + u_1\left[-\frac{1}{q_2 - q_1}q_1^n + \frac{1}{q_2 - q_1}q_2^n\right], \qquad (3)$$

where q_1 and q_2 are roots of the characteristic equation

$$-2 + (3 + Ah)\,q - q^2 = 0.$$

Let us compute q_1 and q_2:

$$\left.\begin{aligned} q_1 &= \frac{3 + Ah - \sqrt{1 + 6Ah + A^2h^2}}{2} = 1 - Ah + 2A^2h^2 + O(h^3), \\ q_2 &= \frac{3 + Ah + \sqrt{1 + 6Ah + A^2h^2}}{2} = 2(1 + Ah) + O(h^2). \end{aligned}\right\} \qquad (4)$$

We will use, for q_1^n and q_2^n, the approximate expressions

$$\left.\begin{aligned} q_1^n &= [1 - Ah + O(h^2)]^n = [1 - Ah + O(h^2)]^{(x_n/h)} = e^{-Ax_n} + O(h), \\ q_2^n &= [2(1 + Ah) + O(h^2)]^n = [2(1 + Ah) + O(h^2)]^{(x_n/h)} = \\ &= 2^{(x_n/h)}[e^{Ax_n} + O(h)]. \end{aligned}\right\} \quad (5)$$

Substituting Eq. (5) into (3) we get

$$u_n = \frac{q_2u_0 - u_1}{q_2 - q_1}[e^{-Ax_n} + O(h)] + \frac{q_1u_0 - u_1}{q_1 - q_2}[e^{Ax_n} + O(h)]\,2^{(x_n/h)}. \quad (6)$$

Before considering what limit u_n will tend to as $h \to 0$ we must indicate how we will fix the initial values, u_0 and u_1, of the difference solution.

Just as in §8 we will look for a solution satisfying the condition $u(0) = b$, and take as difference starting values $u_0 = b$ and $u_1 = b(1 - Ah)$. We substitute these starting values into Eq. (6) and simplify each term separately.

The first and second terms, respectively, take the forms

$$\frac{q_2u_0 - u_1}{q_2 - q_1}[e^{-Ax_n} + O(h)] =$$

$$= \frac{[2 + O(h)]b - (1 - Ah)b}{[2 + O(h)] - [1 - O(h)]}[e^{-Ax_n} + O(h)] = be^{-Ax_n} + O(h),$$

$$\frac{q_1u_0 - u_1}{q_1 - q_2}[e^{Ax_n} + O(h)]\,2^{(x_n/h)} =$$

$$= \frac{[1 - Ah + 2A^2h^2 + O(h^3)]b - b(1 - Ah)}{[1 + O(h)] - [2 + O(h)]}[e^{Ax_n} + O(h)]\,2^{(x_n/h)} =$$

$$= -2A^2h^2b[e^{Ax_n} + O(h)]\,2^{(x_n/h)}.$$

Thus we get

$$u_n = [be^{-Ax_n} + O(h)] + [-2A^2be^{Ax_n} + O(h)]h^2\,2^{(x_n/h)}.$$

For $x_n = x = \text{const}$, as $h \to 0$ the first term of this expression tends to $b\exp(-Ax)$, i.e. to the desired solution. Therefore if the whole expression for u_n is to converge to this solution it is necessary that the second term should go to zero: but as $h \to 0$ this term tends, not to zero, but to infinity. In fact $-2A^2b\exp(Ax_n) + O(h)$ tends to the finite,

nonvanishing, limit $-2A^2b\exp(Ax)$, and $h^2\,2^{(x_n/h)}$ tends to infinity faster than any positive power of $1/h$.

We have shown that a difference scheme approximating the differential equation can have a solution not converging, as $h \to 0$, to the solution of the differential equation. One might think that the fault lies, here, in an insufficiently accurate choice of u_1. But we will now show that there will be no convergence even if we take u_1 to be exactly equal to the solution of the differential equation at $x_1 = x_0 + h$, that is if we set $u_1 = u_0 \exp(-Ah) = b \exp(-Ah)$. Let us begin by simplifying the expressions occurring in Eq. (6):

$$\frac{q_2u_0 - u_1}{q_2 - q_1} = \frac{[2 + O(h)]b - be^{-Ah}}{[2 + O(h)] - [1 + O(h)]} = b + O(h),$$

$$\frac{q_1u_0 - u_1}{q_1 - q_2} = \frac{[1 - Ah + 2A^2h^2 + O(h^3)]b - be^{-Ah}}{[1 + O(h)] - [2 + O(h)]} = -\frac{3}{2}A^2h^2[b + O(h)].$$

Substituting these expressions into (6) we get

$$u_n = [be^{-Ax_n} + O(h)] - [\tfrac{3}{2}A^2be^{Ax_n} + O(h)]h^2\,2^{(x_n/h)}. \qquad (7)$$

The second term on the right-hand side of this equation again tends to infinity, while the first remains bounded. Therefore the whole solution of the difference equation also tends to infinity.

The reason that difference scheme (2) doesn't converge as $h \to 0$, as we have seen, is the fact that it can have solutions which grow quickly as the step-size h decreases, even if the starting values are completely reasonable.

Such difference schemes are called "unstable". Naturally, they are unsuitable for the numerical solution of differential equations.

This Page Intentionally Left Blank

Chapter 5

Convergence of the Solutions of Difference Equations as a Consequence of Approximation and Stability

In Chapter 4 we showed by example what is meant by the approximation of a differential problem by a difference problem, and what constitutes convergence, thanks to which the solution of the differential equation can be calculated approximately through use of the difference scheme. We became familiar with the phenomenon of instability, which can render the difference scheme divergent and useless for computation. Analysis of the behavior of the solutions in these elementary introductory examples, intended only to give the reader a preliminary acquaintance with fundamental concepts, was based on explicit expressions for the solutions. Such a display of explicit solutions was made possible only by a special choice of examples.

In this chapter we give rigorous definitions of convergence, approximation and stability. We show that proofs of convergence need not be based on the analysis of explicit expressions for solutions. Such proofs can be split into the verification of approximation of the differential problem by the difference problem, and verification of the stability of the difference problem.

§ 10. Convergence of a difference scheme

1. Concept of a net and a net function. Suppose that a differential boundary-value problem is given on some interval, D. This means that one is given a differential equation (or system of equations) which the solution must satisfy in the interval, D, and auxiliary conditions on u at one or both ends of this interval. The differential boundary-value problem will be written in the symbolic form

$$Lu = f, \tag{1}$$

Where L is a given differential operator, and f is a given right-hand side. Thus, for example, to write the problem

$$\left.\begin{aligned} \frac{du}{dx} + \frac{x}{1+u^2} &= \cos x, \qquad 0 \le x \le 1, \\ u(0) &= 3, \end{aligned}\right\} \tag{2}$$

in form (1) we need only take

$$Lu \equiv \begin{cases} \dfrac{du}{dx} + \dfrac{x}{1+u^2}, & 0 \le x \le 1, \\ u(0), \end{cases}$$

$$f \equiv \begin{cases} \cos x, & 0 \le x \le 1, \\ 3. \end{cases}$$

The problem

$$\left.\begin{aligned} \frac{d^2u}{dx^2} - (1+x^2)u &= \sqrt{x}, \qquad 0 \le x \le 1, \\ u(0) &= 2, \\ \frac{du(0)}{dx} &= 1 \end{aligned}\right\} \tag{3}$$

can be written in form (1) if we set

$$Lu \equiv \begin{cases} \dfrac{d^2u}{dx^2} - (1+x^2)u, & 0 \le x \le 1, \\ u(0), \\ \dfrac{du(0)}{dx}, \end{cases}$$

$$f \equiv \begin{cases} \sqrt{x}, & 0 \le x \le 1, \\ 2, \\ 1. \end{cases}$$

To put into form (1) the problem

$$\left.\begin{aligned} \frac{d^2u}{dx^2} - (1+x^2)u &= \sqrt{x+1}, \qquad 0 \le x \le 1, \\ u(0) &= 2, \\ u(1) &= 1, \end{aligned}\right\} \tag{4}$$

with boundary conditions at both ends of the interval $0 \le x \le 1$ one must take

$$Lu \equiv \begin{cases} \dfrac{d^2u}{dx^2} - (1+x^2)u, & 0 \le x \le 1, \\ u(0), \\ u(1), \end{cases}$$

$$f \equiv \begin{cases} \sqrt{x+1}, & 0 \leq x \leq 1, \\ 2, \\ 1. \end{cases}$$

The boundary-value problem for the system of differential equations

$$\left.\begin{aligned} \frac{dv}{dx} + xvw &= x^2 - 3x + 1, \qquad 0 \leq x \leq 1, \\ \frac{dw}{dx} + \frac{1}{1+x^2}(v+w) &= \cos^2 x, \qquad 0 \leq x \leq 1, \\ v(0) &= 1, \\ w(0) &= -3 \end{aligned}\right\} \tag{5}$$

can be written in form (1) if one takes u to be a vector function, $u = (v,w)^T$,* and sets

$$Lu \equiv \begin{cases} \dfrac{dv}{dx} + xvw, & 0 \leq x \leq 1, \\ \dfrac{dw}{dx} + \dfrac{1}{1+x^2}(v+w), & 0 \leq x \leq 1, \\ v(0), \\ w(0), \end{cases}$$

$$f \equiv \begin{cases} x^2 - 3x + 1, & 0 \leq x \leq 1, \\ \cos^2 x, & 0 \leq x \leq 1, \\ 1, \\ -3. \end{cases}$$

In all these examples we have considered problems formulated on the interval $0 \leq x \leq 1$, and not on some other interval, only for the sake of definiteness.

We will assume that the solution, u(x), of problem (1) on the interval $0 \leq x \leq 1$, exists. In order to calculate this solution by the method of finite differences, we must first of all choose, on the interval D, a finite set of points which, in totality, we will call a "net" and designate by the symbol D_h; then we set out to find, not the solution, u(x), of problem (1), but a table, $[u]_h$ of values of the solution at the points of the net D_h. It is assumed that the net D_h depends on a parameter, $h > 0$, which can take on positive values as small as desired. As the "step-size"

*Here and below the superscript T designates the transpose of a vector.

h goes to zero the net becomes steadily "finer". For instance one might set $h = 1/N$, where N is some positive integer, and take, as the net D_h, the totality of points $x_0 = 0$, $x_1 = h$, $x_2 = 2h$, ..., $x_N = 1$. The desired net function $[u]_h$, in this case takes on, at the points $x_n = nh$ of the net D_h, the values u(nh) which, for brevity, we denote as u_n.

For the approximate computation of the table, $[u]_h$, of solution-values, in the case of problem (2), one could use, for example, the system of equations

$$\left.\begin{aligned} \frac{u_{n+1} - u_n}{h} + \frac{x_n}{1 + u_n^2} &= \cos x_n, \qquad n = 0, 1, \ldots, N-1, \\ u_0 &= 3, \end{aligned}\right\} \tag{6}$$

obtained by substituting, for the derivative du/dx at the points of the net, the difference approximation

$$\frac{du}{dx} \approx \frac{u(x + h) - u(x)}{h} .$$

The solution, $u^{(h)} = (u_0^{(h)}, u_1^{(h)}, \ldots, u_N^{(h)})$, of system (6) is defined on the same net as the desired net function $[u]_h$. It's values $u_1^{(h)}$, $u_2^{(h)}$, ..., $u_N^{(h)}$, at the points $x_1, x_2, \ldots, x_N$ are consecutively calculated from (6) for n = 0, 1, ..., N-1. For the sake of brevity, in Eq. (6) we omit the superscript h on $u_n^{(h)}$ and, as a rule, will also do this in analogous situations everywhere below.

In the case of problem (4), in order to determine a net function, $u^{(h)}$, approximating the table of solution values $[u]_h$, one can use the difference scheme

$$\left.\begin{aligned} \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - (1 + x_n^2)u_n &= \sqrt{x_n + 1}\ , \\ n &= 1, 2, \ldots, N-1, \\ u_0 = 2, \quad u_N &= 1. \end{aligned}\right\} \tag{7}$$

This scheme is obtained by substituting, at the net-points, for the derivative d^2u/dx^2 occurring in the differential equation, the difference approximation

$$\frac{d^2u}{dx^2} \approx \frac{u(x + h) - 2u(x) + u(x - h)}{h^2} . \tag{8}$$

To compute the solution $u^{(h)}$ of problem (7) one may use the FEBS algorithm described in §5.

Still another difference scheme which might be used to compute the solution of problem (5) takes the form

$$\left.\begin{aligned} &\frac{v_{n+1}-v_n}{h}+x_n v_n w_n = x_n^2 - 3x_n + 1,\\ &\frac{w_{n+1}-w_n}{h}+\frac{1}{1+x_n^2}(v_n+w_n)=\cos^2 x_n, \quad n=0,1,\ldots,N-1,\\ &v_0=1,\\ &w_0=-3. \end{aligned}\right\} \tag{9}$$

Here $u_0^{(h)} = (v_0^{(h)}, w_0^{(h)})^T = (1, -3)^T$ is given. For n = 0, from Eq. (9) one can determine $u_1^{(h)} = (v_1^{(h)}, w_1^{(h)})^T$. In general, knowing $u_k^{(h)} = (v_k^{(h)}, w_k^{(h)})^T$, k = 0, 1, ..., n, one can, taking k = n, compute $u_{n+1}^{(h)} = (v_{n+1}^{(h)}, w_{n+1}^{(h)})^T$.

In the above examples the net, D_h, consists of points separated from each other by a distance h. Clearly one could have disposed the N+1 points of the net D_h, $h \equiv 1/N$, on the interval [0,1] not uniformly, but in such a way that $x_0 = 0$, $x_1 = x_0 + h_0$, $x_2 = x_1 + h_1$, ..., $x_N = 1$, where the h_n, n = 0, 1, ..., N-1, are not all equal, but $\max h_n \to 0$ as $h = 1/N \to 0$. The knots of D_h could be so distributed that the desired table, $[u]_h$, of the solution u(x) would be most detailed for fixed N (or $h \equiv 1/N$) in those subintervals where u(x) varies most rapidly. These subintervals are sometimes known beforehand from physical considerations, or from preliminary crude calculations. Information on the rate of change of u(x) is also generated in the course of the sequential calculation of $u_1^{(h)}$, $u_2^{(h)}$, ..., $u_n^{(h)}$, and this information may be taken into account in choosing the next net-point x_{n+1}.

We confine ourselves to the examples already discussed as illustrations of the concept of a net, and of an unknown net function (or vector-function) -- a table of values of the solution $[u]_h$. In addition we note only that, in the role of the desired table, $[u]_h$, of solution-values it isn't necessary to consider a net-function coinciding with the solution, u, at the net-points. It is possible to establish a correspondence between the function and the function-table in other ways. For example one may take, as the required table of u(x), $0 < x < 1$, the net function $[u]_h$ defined at the points x = h/2, 3h/2, ..., 1 - h/2, by the equation

$$[u]_h = \frac{1}{h}\int_{x-h/2}^{x+h/2} u(\xi)\,d\xi.$$

This way to set up a correspondence is convenient in the case where u(x) is not a continuous function, but it is known that it's integral over any interval exists. Such a situation may occur for example, if one is dealing with a generalized, discontinuous, solution for which the integral

$$\int_0^1 u^2(x)\,dx$$

exists.

Everywhere below, barring statements to the contrary, we will assume that u is a continuous function and take $[u]_h$ to be the net function coinciding with u at the net-points.

We are concerned with the computation of the net function $[u]_h$ because, as the net is refined, i.e. as $h \to 0$, it becomes a more and more detailed table of the desired solution u, of which it gives us an increasingly more complete representation. Via interpolation it is possible, with increasing accuracy as $h \to 0$, to construct the solution everywhere within D. Clearly the accuracy with which this can be done, for a given number and distribution of points of the net D_h, depends on additional facts concerning the solution (like, for example, bounds on its derivatives), and also on the distribution of the points of net D_h.

We confine ourselves to such passing comments on the construction of the function, u, from the table $[u]_h$. More detailed consideration of the construction of a function from tabular values constitutes the subject matter of the theory of interpolation. We will concern ourselves only with the construction of the table $[u]_h$ and, by convention, consider that problem (1) has been solved exactly if the net-function $[u]_h$ has been determined. But, of course, we will not succeed in computing $[u]_h$ exactly. Instead of the net function, $[u]_h$, we will look for another net function $u^{(h)}$, which "converges" to $[u]_h$ as the net is refined. For this purpose one can make use of difference equations.

2. Convergent difference schemes. We will be concerned with methods for the constuction and study of convergent difference schemes throughout all of this chapter. But first of all we must give a precise meaning to the requirement that $u^{(h)} \to [u]_h$, the convergence requirement that we will impose on difference schemes. For this purpose we consider a linear normed space of functions defined on the net D_h. *The norm* $||u_h||_{U_h}$ *of a net function* u_h in U_h is a non-negative number which measures the deviation of the function u_h from $u \equiv 0$. We recall that the linear space, R, is said to be "normed" if each element, $\mathbf{x}$, of this space is put into correspondence with a non-negative number $||\mathbf{x}||$ and, moreover, the following three norm-axioms are valid:

1° $\quad ||\mathbf{x}|| \geq 0, \quad \mathbf{x}$ in R;

2° $\quad ||\lambda\mathbf{x}|| = |\lambda| \cdot ||\mathbf{x}||$, where $\mathbf{x}$ is in R and λ is an arbitrary number;

3° $\quad ||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$, for $\mathbf{x}$, $\mathbf{y}$ in R.

The norm can be defined in various ways. One can, for example, take as the norm of a function the exact upper bound of the moduli of it's values at the net points, i.e.

$$||u_h||_{U_h} = \sup_n |u_h(x_n)|. \tag{10}$$

If $u^{(h)}$ is a pair of functions, as in (9), then as a norm, in analogy with (10), one can take the upper bound of the moduli of both functions on their respective nets.

If $u^{(h)}$ consists of functions defined on the net $x = 0, h, 2h, \ldots, 1$, then one frequently uses a norm defined by the equation

$$||u^{(h)}||_{U_h} = \left(h \sum_{n=0}^{N} |u_n|^2\right)^{1/2}.$$

This norm is analagous to the norm

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

for functions, $u(x)$, square-integrable on the interval $0 \leq x \leq 1$.

Everywhere below, if nothing is said to the contrary, we will use norm (10).

After the introduction of a normed space, U_h, the concept of a deviation between one function and another becomes meaningful. If $a^{(h)}$ and $b^{(h)}$ are two arbitrary net functions in U_h, then the *measure* of their *deviation* from each other is taken to be the norm of their difference, i.e. the quantity

$$||a^{(h)} - b^{(h)}||_{U_h}.$$

Now we can proceed to a rigorous definition of a convergent difference scheme.

Suppose that, for the approximate computation of the solution of the differential boundary-value problem (1), i.e. for the approximation computation of the net function $[u]_h$ via Eq. (1), we have constructed a system of equations which we will write symbolically, by analogy with Eq. (1), in the form

$$L_h u^{(h)} = f^{(h)}. \tag{11}$$

Difference schemes (6), (7) and (9), for differential boundary-value problems (2), (4) and (5) respectively, may be taken as examples of this differencing process.

To write scheme (6) in form (11) we may set

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1} - u_n}{h} + \dfrac{nh}{1 + u_n^2}, & n = 0, 1, \ldots, N-1, \\ u_0, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \cos nh, & n = 0, 1, \ldots, N-1, \\ 3. \end{cases}$$

Scheme (7) may be written in form (11) if we take

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + [1 - (nh)^2]u_n, & n = 1, 2, \ldots, N-1, \\ u_0, \\ u_N, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \sqrt{1 + nh}, & n = 1, 2, \ldots, N = 1, \\ 2, \\ 1. \end{cases}$$

Finally we write (9) in form (11), taking

$$L_h u^{(h)} \equiv L_h \begin{pmatrix} v^{(h)} \\ w^{(h)} \end{pmatrix} =$$

$$= \begin{cases} \dfrac{v_{n+1} - v_n}{h} + (nh) v_n w_n, & n = 0, 1, \ldots, N-1, \\ \dfrac{w_{n+1} - w_n}{h} + \dfrac{1}{1 + (nh)^2}(v_n + w_n), & n = 0, 1, \ldots, N-1, \\ v_0, \\ w_0, \end{cases}$$

$$f^{(h)} = \begin{cases} (nh)^2 - 3nh + 1, & n = 0, 1, \ldots, N-1, \\ \cos^2 nh, & n = 0, 1, \ldots, N-1, \\ 1, \\ -3. \end{cases}$$

We see that system (11) depends on h, and must be treated separately for each h corresponding to each of the nets, D_h, and net-functions $[u]_h$, which are of interest to us. Thus a difference boundary-value problem is not a single system, but a family of systems depending on a parameter, h.

It will be assumed that, for each sufficiently small h, there exists a solution, $u^{(h)}$, of problem (11), belonging to the space U_h.

We will say that *the solution $u^{(h)}$ of the difference boundary-value problem (11) converges,* as the net is refined, *to the solution of boundary-value problem (1)*, if

$$\| [u]_h - u^{(h)} \|_{U_h} \to 0 \quad \text{as} \quad h \to 0. \tag{12}$$

If, in addition, the inequality

$$||[u]_h - u^{(h)}||_{U_h} < ch^k, \tag{13}$$

is satisfied, where $c > 0$ and $k > 0$ are constants not depending on h, we will say that *convergence is of order* h^k, or that *the difference scheme has k'th order accuracy*.

In §8 we considered two difference schemes for the problem

$$\left.\begin{aligned} \frac{du}{dx} + Au = 0, \qquad 0 \leq x \leq 1, \\ u(0) = b. \end{aligned}\right\}$$

The estimates obtained there for the difference, $\delta(x) = u(x_k) - u_k^{(h)}$, between the exact and approximate solutions show that the first of these schemes converges with order h, while for the second convergence is order h^2.

The requirement that it be convergent is the fundamental requirement which will be imposed on difference scheme (11) for the numerical solution of the differential boundary-value problem (1). When this requirement is met then, with the aid of difference scheme (11), the solution u can be computed to any prescribed accuracy, if h is taken small enough. We have rigorously formulated the concept of convergence and have come up to the central question: i.e., how does one construct a convergent difference scheme (11) for computation of the solution of differential boundary-value problem (1)? The above examples supplement the considerations of Chapter 1, and give some idea as to the simplest method for the construction of such schemes: one must choose a net, and substitute difference expressions for the derivatives. However, as we have seen, for one and the same differential boundary-value problem one can get different difference schemes (11), choosing different nets D_h, and replacing the derivatives by various difference approximations. We have already seen in §9, through the example of the simplest ordinary differential equation, that a difference scheme may be unsuitable for computation.

3. Proof of convergence of a difference scheme. For the moment we will not concern ourselves with the problem of constructing difference schemes, but will pose a slightly different problem. Suppose that a difference scheme $L_h u^{(h)} = f^{(h)}$, which we have reason to hope is convergent, so that

$$||[u]_h - u^{(h)}||_{U_h} \to 0 \quad \text{as} \quad h \to 0$$

has, somehow or another, already been constructed. How can one test whether it is, in fact, convergent or not?

Let us assume that difference scheme (11) has a unique solution $u^{(h)}$ in U_h. If, on substituting the net function $[u]_h$ ($[u]_h$ in U_h), in place of $u^{(h)}$, into the left-hand side of (11), it turns out that (11) is satisfied exactly then, in view of the uniqueness of the solution, we would have $[u]_h = u^{(h)}$, i.e. ideal convergence. The solution, $u^{(h)}$, of the difference problem $L_h u^{(h)} = f^{(h)}$ would then, in other words, coincide with the required net function $[u]_h$, which we have agreed to consider the exact solution.

However, as a rule one will not succeed in constructing (11) in such a way as to be exactly satisfied by $[u]_h$. When $[u]_h$ is substituted into Eq. (11) some sort of residual will form:

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}. \qquad (14)$$

If this residual $\delta f^{(h)}$ "tends to zero" as $h \to 0$, so that $[u]_h$ satisfies Eq. (11) more and more closely, then we will say that the difference scheme $L_h u^{(h)} = f^{(h)}$ approximates the boundary-value problem $Lu = f$ on the solution, u, of this latter problem.

In case of approximation, i.e. if the difference scheme approximates the boundary-value problem, one may suppose that Eq. (14), which is satisfied by $[u]_h$, is gotten from (11) through the addition of some (small for small h) increment, $\delta f^{(h)}$, to the right-hand side $f^{(h)}$. Therefore, if the solution, $u^{(h)}$, of problem (11) is stable with respect to perturbations of the right-hand side $f^{(h)}$, i.e. changes little for small changes of the right-hand side, then the solution $u^{(h)}$ of problem (11) and the solution $[u]_h$ of problem (14) will differ little from each other, so that from approximation

$$\delta f^{(h)} \to 0 \quad \text{as} \quad h \to 0$$

follows convergence

$$u^{(h)} \to [u]_h \quad \text{as} \quad h \to 0.$$

The approach we have indicated, by which to test the convergence of (12), consists in that one splits this difficult problem into two which are simpler: first, test whether problem (1) is approximated by (11), and then determine whether problem (11) is stable. But here is, in fact, an indication as to how one might construct a convergent difference scheme for the solution of problem (1): one must construct an approximating difference scheme; from among the many possible methods of approximation one must choose one such that the difference scheme turns out to be stable.

The above general plan for the study of convergence, naturally, assumes the introduction of rigorous concepts of approximation and stability, such that one can prove a theorem stating that, from approximation and stability, follows convergence. The above definitions of approximation and stability are not rigorous. To define approximation one must first state more precisely what is the residual, $\delta f^{(h)}$, in the general case, and what is meant by its magnitude; and to define stability one must give a precise meaning to the assertion that "to a small perturbation of the right-hand side corresponds a small perturbation of the solution of the difference problem $L_h u^{(h)} = f^{(h)}$".

Strict definitions of approximation and stability will be the principle topics of §11 and §12, respectively.

PROBLEMS

1. Divide the interval [0,1] into N parts, separated by the points $x_0 = 0, x_1, x_2, \cdots x_{N-1}, x_N = 1$, in such a way that

$$\frac{x_{n+1} - x_n}{x_n - x_{n-1}} = q,$$

and determine whether it is possible to use a sequence of such nets with $N \to \infty$ (where q is a constant not depending on N) for the approximate solution of the problem

$$\left.\begin{aligned} u' - u &= 0 \\ u(0) &= 1 \end{aligned}\right\}$$

with the aid of the difference scheme

$$\left.\begin{aligned} &\frac{u^{(h)}(x_{n+1}) - u^{(h)}(x_n)}{x_{n+1} - x_n} - u^{(h)}(x_n) = 0 \qquad \left(h = \frac{1}{N}\right) \\ &u^{(h)}(x_0) = 1. \end{aligned}\right\}$$

Does the maximum of the step-sizes $x_{n+1} - x_n$ tend to zero as $N \to \infty$?

Hint. It is simplest to consider the case $q > 1$, and to convince oneself that

$$\lim_{N\to\infty} u^{(1/N)}(x_N) = \infty .$$

§11. Approximation of a differential boundary-value problem by a difference scheme

1. The residual $\delta f^{(h)}$. We now give a precise meaning to the concept of approximation of boundary value problem (1) §10

$$Lu = f, \tag{1}$$

on the solution u, by difference scheme (11) §10

$$L_h u^{(h)} = f^{(h)}. \tag{2}$$

For this purpose one must state more precisely what is meant by the residual $\delta f^{(h)}$

$$L_h [u]_h = f^{(h)} + \delta f^{(h)}, \tag{3}$$

which forms when the net function $[u]_h$, the table of values of the required solution u, is substituted into Eq. (2); and one must make a precise statement as to its magnitude.

Convergence of the magnitude of $\delta f^{(h)}$ to zero, as $h \to 0$, we then take as the definition of approximation.

We start with the consideration of an example of a difference scheme for the numerical solution of the differential boundary-value problem

$$\left.\begin{aligned} &\frac{d^2u}{dx^2} + a(x)\frac{du}{dx} + b(x)u = \cos x, \qquad 0 \le x \le 1, \\ &u(0) = 1, \\ &u'(0) = 2. \end{aligned}\right\} \tag{4}$$

As our net D_h we take, as before, the set of points $x_n = nh$, $n = 0, 1, \ldots, N$; $h = 1/N$. As a difference scheme for the approximate computation of $[u]_h$ we use the equation-set

$$\left.\begin{aligned} &\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + a(x_n)\frac{u_{n+1} - u_{n-1}}{2h} + b(x_n)u_n = \\ &= \cos x_n, \quad n = 1, 2, \ldots, N-1, \\ &u_0 = 1, \\ &\frac{u_1 - u_0}{h} = 2, \end{aligned}\right\} \tag{5}$$

obtained by substituting, for the derivatives in (4), the approximations

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \approx \frac{d^2u(x)}{dx^2},$$

$$\frac{u(x+h) - u(x-h)}{2h} \approx \frac{du(x)}{dx}, \tag{6}$$

$$\frac{u(h) - u(0)}{h} \approx \frac{du(0)}{dx}.$$

The difference scheme (5) takes the form (2) if one defines

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + a(nh)\dfrac{u_{n+1} - u_{n-1}}{2h} + b(nh)u_n, \\ u_0, \\ \dfrac{u_1 - u_0}{h}, \end{cases}$$

$$f^{(h)} = \begin{cases} \cos nh, \\ 1, \\ 2. \end{cases} \tag{7}$$

To compute and bound the magnitude of the residual, $\delta f^{(h)}$, which arises when $[u]_h$ is substituted into (2), we refine Eqs. (6).

By Taylor's formula we have

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(\xi_1),$$

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(\xi_2),$$

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_3),$$

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_4),$$

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(\xi_5).$$

Here ξ_1, ξ_2, ξ_3, ξ_4 and ξ_5 are certain points in the interval [x-h, x+h].

Hence

$$\left.\begin{aligned} \frac{u(x+h)-u(x-h)}{2h} &= u'(x) + \frac{h^2}{12}\left[u'''(\xi_1) + u'''(\xi_2)\right], \\ \frac{u(x+h)-2u(x)+u(x-h)}{h} &= u''(x) + \frac{h^2}{24}\left[u^{(4)}(\xi_3) + u^{(4)}(\xi_4)\right], \\ \frac{u(x+h)-u(x)}{h} &= u'(x) + \frac{h}{2}u''(\xi_5). \end{aligned}\right\} \quad (8)$$

2. Computation of the residual. We will assume that the solution, $u(x)$, of problem (4) has bounded derivatives up through the fourth. By virtue of (8) one can write

$$\frac{u(x+h)-2u(x)+u(x-h)}{h^2} + a(x)\,\frac{u(x+h)-u(x-h)}{2h} +$$

$$+ b(x)u(x) = \frac{d^2u(x)}{dx^2} + a(x)\,\frac{du(x)}{dx} + b(x)u(x) +$$

$$+ h^2\left[\frac{u^{(4)}(\xi_3) + u^{(4)}(\xi_4)}{24} + a(x)\,\frac{u'''(\xi_1) + u'''(\xi_2)}{12}\right].$$

Therefore the expression

$$L_h[u]_h =$$

$$= \begin{cases} \dfrac{u(x_n+h) - 2u(x_n) + u(x_n-h)}{h^2} + \\ \qquad + a(x_n)\,\dfrac{u(x_n+h)-u(x_n-h)}{2h} + b(x_n)u(x_n), \quad n = 1, 2, \ldots, N-1, \\ u(0), \\ \dfrac{u(h)-u(0)}{h} \end{cases}$$

can be rewritten thus:

$$L_h[u]_h = \begin{cases} \cos x_n + h^2 \left[\dfrac{u^{(4)}(\xi_3) + u^{(4)}(\xi_4)}{24} + \right. \\ \qquad \left. + a(x_n) \dfrac{u'''(\xi_1) + u'''(\xi_2)}{12} \right], \quad n = 1, 2, \ldots, N-1, \\ 1 + 0, \\ 2 + h \dfrac{u''(\xi_5)}{2} \end{cases}$$

or

$$L_h[u]_h = f^{(h)} + \delta f^{(h)},$$

where

$$\delta f(h) = \begin{cases} h^2 \left[\dfrac{u^{(4)}(\xi_3) + u^{(4)}(\xi_4)}{24} + \dfrac{u'''(\xi_1) + u'''(\xi_2)}{12} \right], \\ 0, \\ h \dfrac{u''(\xi_5)}{2}. \end{cases} \tag{9}$$

It is convenient to regard $f^{(h)}$ and $\delta f^{(h)}$, given by Eqs. (7) and (9), as belonging to a linear normed space F_h, which consists of elements of the form

$$g^{(h)} = \begin{cases} \phi_n \qquad (n = 1, 2, \ldots, N-1), \\ \psi_0, \\ \psi_1, \end{cases} \tag{10}$$

where $\phi_1, \phi_2, \ldots, \phi_{n-1}$, and also ψ_0 and ψ_1 are arbitrary ordered sets of numbers; one can take $g^{(h)}$ to be the totality of net functions ϕ_n, $n = 1, 2, \ldots, N-1$, along with the ordered pairs of numbers ψ_0 and ψ_1. The summation of two elements of the space F_h, and the multiplication of the elements of $g^{(h)}$ by a common factor, are carried out term by term. Clearly in the example under consideration F_h is an (N+1)-dimensional linear space. The norm of F_h can be introduced in many ways. If we introduce a norm in F_h via the equation

$$\|g^{(h)}\|_{F_h} = \max(|\psi_0|, |\psi_1|, \max_n |\phi_n|),$$

i.e. take as a norm the maximum absolute value of all components of the vector $g^{(h)}$ then, by virtue of (9), we get

$$||\delta f^{(h)}||_{F_h} \leq Ch, \tag{11}$$

where C is some constant depending on u(x), but not on h.

From this equation it follows that $f^{(h)}$ tends to zero as $h \to 0$.

In the equation $L_h u^{(h)} = f^{(h)}$ which we have taken as an example, (and which is written out in detail in Eqs. (5)) L_h can be regarded as an operator. This operator maps each net function $v^{(h)} = v_n$, $n = 0, 1, \ldots, N$, in the linear space of functions defined on the net D_h, into an element $g^{(h)}$ of form (10), also in the linear space F_h, via the equation

$$L_h v^{(h)} \equiv \begin{cases} \dfrac{v_{n+1} - 2v_n + v_{n-1}}{h^2} + a(x_n)\dfrac{v_{n+1} - v_{n-1}}{2h} + b(x_n)v_n, \\ v_0 \\ \dfrac{v_1 - v_0}{0}. \end{cases}$$

Also in the *general* difference boundary-value problem (2) we will adopt the convention that the right-hand sides of these scalar equations, which we have written collectively in the symbolic form

$$L_h u^{(h)} = f^{(h)},$$

are components of a vector $f^{(h)}$ in some linear normed space F_h. Then we can regard L_h as an operator mapping each net function, $u^{(h)}$ in U_h, into some element $f^{(h)}$ of F_h. In this case the symbol $L_h[u]_h$ is meaningful, and represents the result of the operation L_h on the net function $[u]_h$ of U_h, an operation yielding an element of the space F_h.

The residual $\delta f^{(h)} = L_h[u_h] - f^{(h)}$ belongs to the space F_h, being the difference of two elements of this space. By the "magnitude of the residual" is meant $||\delta f^{(h)}||_{F_h}$.

3. Approximation of order h^k.

Definition. We will say that the difference scheme $L_h u(h) = f^{(h)}$ *approximates* the problem Lu = f on the solution u if $||\delta f^{(h)}||_{F_h} \to 0$ as $h \to 0$. If, moreover, the inequality

$$||\delta f^{(h)}||_{F_h} \leq ch^k,$$

is satisfied, where $c > 0$ and $k > 0$ are constants, then we will say that *the approximation is of order h^k, or order k with respect to the magnitude of h.*

The fact that u is a solution of problem (1) gives information about the function u, information which one can use for the *construction* of system (2), and also to verify approximation. For this reason in the definition of approximation we refer to problem (1). We stress, however, that the above definition of approximation of the problem $Lu = f$ on the solution u, by the difference equation $L_h u^{(h)} = f^{(h)}$, does not rely on the equation, $Lu = f$, which determines the function u. One might have said simply that the scheme $L_h u^{(h)} = f^{(h)}$ agrees to order h^k with function u, making no mention of the origin of this function. In particular if the function u is, simultaneously, the solution of two completely different problems, $L^{(1)}u = f^{(1)}$ and $L^{(2)}u = f^{(2)}$, of form (1), then one and the same difference scheme, $L_h u^{(h)} = f^{(h)}$ simultaneously either does or does not approximate each of these problems on their common solution u.

4. Examples.

Example 1. Difference scheme (5), in view of bound (11), approximates (4) to first order in h. Scheme (5) can easily be refined, however, so that the approximation becomes order h^2. To accomplish this we note that all components of $\delta f^{(h)}$ except the last tend to zero like h^2 (and the next to last is actually exactly equal to zero).

Only the last component of the vector $\delta f^{(h)}$ (i.e., the residual arising from the substitution of $[u]_h$ into the last equation, $(u_1 - u_0)/h = 2$, of system (5)) tends to zero more slowly and is, in fact, first order in h. This annoying circumstance is easily eliminated. By Taylor's formula

$$\frac{u(h) - u(0)}{h} = u'(0) + \frac{h}{2}u''(0) + \frac{h^2}{6}u'''(\xi) =$$

$$= 2 + \frac{h}{2}u''(0) + \frac{h^2}{6}u'''(\xi), \qquad 0 < \xi < h.$$

But from differential equation (4) we find that

$$u''(0) = -a(0)u'(0) - b(0)u(0) + \cos 0 = -2a(0) - b(0) + 1.$$

Therefore, replacing the last equation of (5) by the equation

$$\frac{u_1 - u_0}{h} = 2 - \frac{h}{2}\left[2a(0) + b(0) - 1\right], \tag{12}$$

we get for $f^{(h)}$, in place of (7), the expression

$$f^{(h)} \equiv \begin{cases} \cos x_n, \\ 1, \\ 2 - \frac{h}{2}\left[2a(0) + b(0) - 1\right]. \end{cases}$$

It then turns out that

$$\delta f^{(h)} \equiv \begin{cases} \dfrac{h^2}{12}\left[\dfrac{u^{(4)}(\xi_3) + u^{(4)}(\xi_4)}{2} + \left(u'''(\xi_1) + u'''(\xi_2)\right)\right], \\ 0, \\ \dfrac{h^2}{6}\, u'''(\xi) \end{cases}$$

and $\|\delta f^{(h)}\|_{F_h} < C_1 h^2$, where C_1 is some constant not depending on h. The approximation now becomes second-order in h.

We stress that, for the construction of difference boundary condition (12), we used not only the boundary condition of problem (4), but also the differential equation itself. One may say that we have, in effect, used the boundary condition

$$u''(x) + a(x)u'(x) + b(x)u(x)\big|_{x=0} = \cos x\big|_{x=0},$$

which is a consequence of the differential equation.

Example 2. We examine the order of approximation of the difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1} - u_{n-1}}{2h} + Au_n = 1 + x_n^2, \quad n = 1, 2, \ldots, N-1, \\ u_0 = b, \\ u_1 = b \end{cases} \tag{13}$$

on the solution, u, of the problem

$$\left.\begin{aligned} \frac{du}{dx} + Au = 1 + x^2, \\ u(0) = b. \end{aligned}\right\} \tag{14}$$

A similar scheme was considered in §8 even before the introduction of a rigorously-defined concept of approximation.

Here the role of $f^{(h)}$ is played by

$$f^{(h)} = \begin{cases} 1 + x_n^2, \quad n = 1, 2, \ldots, N-1, \\ b, \\ b. \end{cases}$$

Further

$$L_h[u]_h \equiv \begin{cases} \dfrac{u(x_n + h) - u(x_n - h)}{2h} + Au(x_n), & n = 1, \ldots, N-1, \\ u(0), \\ u(h) \end{cases}$$

or

$$L_h[u]_h \equiv \begin{cases} \left[\dfrac{du(x_n)}{dx} + Au(x_n)\right] + \dfrac{h^2}{6} u'''(\xi_n), & n = 1, \ldots, N-1, \\ u(0), \\ u(0) + h \dfrac{du(\xi_0)}{dx}. \end{cases}$$

Since the solution, u, satisfies the equation

$$\frac{du(x_n)}{dx} + Au(x_n) = 1 + x_n^2,$$

the residual $\delta f^{(h)}$ takes the form

$$\delta f^{(h)} = \begin{cases} \dfrac{h^2}{6} u'''(\xi_n), & n = 1, \ldots, N-1, \\ 0, \\ hu'(\xi_0) \end{cases}$$

The approximation of problem (14) by scheme (13) is of first order in h. One immediately sees that the components of the residual, as in example 1, are of different order in h. The difference equation

$$\frac{u_{n+1} - u_{n-1}}{2h} + Au_n = 1 + x_n^2, \quad n = 1, 2, \ldots, N-1, \tag{15}$$

upon substitution of $[u]_h$, is satisfied with residual $h^2u'''(\xi_n)/6$, a term of order h^2. The first boundary condition

$$u_0 = b \tag{16}$$

on substituting $[u]_h$, is satisfied exactly, and the second

$$u_1 = b \tag{17}$$

-- with residual, $hu'(\xi_0)$, of first order in h. The error of the approximation is estimated via

$$\max_{0 \le x \le 1} |u'''(x)|, \qquad \max_{0 \le x \le 1} |u'(x)|.$$

In the example under consideration the exact solution

$$u(x) = u(0)e^{-Ax} + \frac{1 + x^2 - e^{-Ax}}{A} - \frac{2x}{A^2}$$

allows us to estimate these maxima in terms of u(0) and A:

$$\max_{0 \le x \le 1} |u'(x)| = \max_{0 \le x \le 1} \left| Au(0)e^{-Ax} + \frac{2x + Ae^{-Ax}}{A} - \frac{2}{A^2} \right| \le$$

$$\le |u(0)|\ |A|(1 + e^{-A}) + \frac{2}{|A|} + \frac{2}{A^2} + (1 + e^{-A}),$$

$$\max_{0 \le x \le 1} |u'''(x)| = \left| [u(0) + \tfrac{1}{A}]A^3 e^{-Ax} \right| \le$$

$$\le [|u(0)|\ |A|^3 + |A|^2](1 + e^{-A}).$$

In more complicated cases it is necessary to limit oneself to coarse bounds on these derivatives, based on the theory of differentiability of the solutions of ordinary differential equations with smooth right-hand sides.

5. Splitting of difference schemes into subsystems. For a detailed description of the character of approximation it turned out to be convenient to talk, not about the whole difference scheme (13) of form (2),

$$L_h u^{(h)} = f^{(h)},$$

all at once, but separately about subsystems (15), (16), and (17). These subsystems (of which the latter two each consist of a single equation) can be put, respectively, into the following symbolic forms:

$$\ell_h^0 u^{(h)} = f_0^{(h)}, \tag{18}$$

$$\ell_h^1 u^{(h)} = f_1^{(h)}, \tag{19}$$

$$\ell_h^2 u^{(h)} = f_2^{(h)}. \tag{20}$$

In order to accomplish this one must take

$$\ell_h^0 u^{(h)} = \frac{u_{n+1} - u_{n-1}}{2h} + Au_n, \quad n = 1, 2, \ldots, N-1,$$

$$\ell_h^1 u^{(h)} \equiv u_0,$$

$$\ell_h^2 u^{(h)} \equiv u_1,$$

$$f_0^{(h)} = 1 + x_n^2,$$

$$f_1^{(h)} = b,$$

$$f_2^{(h)} = b.$$

For convenience, also in the general case difference scheme (2) is often split into two or several subsystems

$$\left.\begin{array}{l} \ell_h^0 u^{(h)} = f_0^{(h)}, \\ \cdots\cdots\cdots \\ \ell_h^r u^{(h)} = f_r^{(h)}, \\ \cdots\cdots\cdots \\ \ell_h^R u^{(h)} = f_R^{(h)}, \end{array}\right\} \tag{21}$$

so that

$$L_h u^{(h)} \equiv \left\{\begin{array}{l} \ell_h^0 u^{(h)}, \\ \cdots \\ \ell_h^r u^{(h)}, \\ \cdots \\ \ell_h^R u^{(h)}, \end{array}\right. \qquad f^{(h)} = \left\{\begin{array}{l} f_0^{(h)}, \\ \cdots \\ f_r^{(h)}, \\ \cdots \\ f_R^{(h)}. \end{array}\right.$$

It is convenient to consider the right-hand side, $f_r^{(h)}$, of each subsystem $\ell_h^r u^{(h)} = f_r^{(h)}$, as an element of the normed space $F_h^{(r)}$. And it is also convenient that the norms in space F_h and in spaces $F_h^{(1)}$, $F_h^{(2)}$, ..., $F_h^{(R)}$ should be coordinated so that

$$\left\|f^{(h)}\right\|_{F_h} = \max_r \left\|f_r^{(h)}\right\|_{F_h^{(r)}}. \tag{22}$$

Splitting (2) into subsystems (21), we will always assume that (22) is satisfied.

The convenience of splitting $L_h u^{(h)} = f^{(h)}$ into subsystems consists in the fact that one can then consider separately the order to which each

subsystem approximates the solution of problem (1), $Lu = f$. This order is taken to be the order with which the norm, $\|\delta f_r^{(h)}\|_{F_h^{(r)}}$, of the residual $\delta f_r^{(h)}$,

$$\ell_h^{(r)}[u]_h = f_r^{(h)} + \delta f_r^{(h)},$$

decreases as $h \to 0$. Thanks to the coordinated choice of norms (22), the order of approximation of the whole difference scheme, $L_h u^{(h)} = f^{(h)}$, on the solution u of the problem $Lu = f$, is equal to the order of decrease of the norm

$$\|\delta f^{(r_0)}\|_{F_h^{(r_0)}}$$

of the residual $\delta f_{r_0}^{(h)}$, where r_0 is that r for which the norm decreases most slowly.

In example 2, when system (13) is split into subsystems (15)-(17), or (18)-(20), the space $F_h^{(0)}$ consists of the net function $f_0^{(h)} = \{f_n\}$ with norm $\|f_0^{(h)}\| = \max |f_n|$, defined at the points $x_n = nh$, $n = 1, 2, \ldots, N-1$, while spaces $F_h^{(1)}$ and $F_h^{(2)}$ are one dimensional and consist of numbers with norm $\|a\| = |a|$. Equation (18)

$$\ell_h^{(0)} u^{(h)} = f_0^{(h)},$$

agrees with problem (14), on the solution u, to second order, the equation $\ell_h^{(1)} u^{(h)} = f_1^h$ corresponds exactly to (14) while the equation $\ell_h^{(2)} u^{(h)} = f_2^h$ is correct to first order. To raise the order of approximation of difference scheme (13) from first to second order in h, it suffices to "improve" only the boundary condition $\ell_h^{(2)} u^{(h)} = b$. We note that

$$\ell_h^{(2)}[u]_h = u(h) = u(0) + hu'(0) + \frac{h^2}{2} u''(\xi).$$

We now take into account that $u(0) = b$ and that, by virtue of (14),

$$u'(0) = -Au(0) + 1 = -Ab + 1.$$

Setting

$$\ell_h^2 u^{(h)} = u_1 = b - hAb + h, \quad \text{i.e.} \quad f_2^{(h)} = b - hAb + h,$$

we achieve satisfaction of the boundary condition

$$\ell_h^{(2)}[u]_h = u(h) = f_2^{(h)} + O(h^2),$$

i.e. we attain agreement, to second order in h^2, with the boundary condition

$$\ell_h^{(2)} u^{(h)} = f_2^{(h)} \qquad \left(f_2^{(h)} = b - hAb + h\right) \tag{23}$$

of problem (14), on the solution u. Thus the difference scheme (15), (18), (23), approximates problem (14) to second order in h.

The splitting of difference scheme (2) into subsystems (21) is simply a convention, adopted solely to facilitate discussion. Thus, for example, system (13) could have been split into two subsystems with difference equation (15) assigned, as before, to the first, and both boundary conditions (16) and (17), to the second. We would then write, symbolically

$$\left.\begin{aligned} \ell_h^{(0)} u^{(h)} &= f_0^{(h)}, \\ \ell_h^{(1)} u^{(h)} &= f_1^{(h)}, \end{aligned}\right\}$$

where

$$\ell_h^{(1)} = \begin{cases} u_0, \\ \\ u_1, \end{cases} \qquad f_1^{(h)} = \begin{pmatrix} b \\ b \end{pmatrix}.$$

With this splitting, however, as opposed to the splitting (15)-(17), or (18)-(20), we would have lost the ability to refer concisely to the fact that the first boundary condition, upon substitution of $[u]_h$, is exactly satisfied, and the second — only to first order in h.

6. Replacement of derivatives by difference expressions. In the above examples we constructed difference schemes by replacing derivatives, in the differential equation, with difference expressions. This is a perfectly general approach which allows one to construct, for any differential boundary-value problem with a smooth enough solution u(x), a difference scheme with any prescribed order of approximation.

* * * * * *

In fact, let us show that the derivative d^k/dx^k, of any arbitrary order k, can be replaced by a difference expression such that the error induced by this replacement, for a smooth enough function u(x), will be of any prescribed order, p, in the step-width, h, of the difference net. For this purpose we will use the method of undetermined coefficients,

We will write an equation of the form

$$\frac{d^k(x)}{dx^k} = h^{-k} \sum_{s=-s_1}^{s_2} a_s u(x + sh) + O(h^p) \tag{24}$$

and try to choose the undetermined coefficients, a_s, $s = -s_1, -s_1 + 1, \ldots, s_2$ (independent of h) in such a way that the equation will be valid. The limits of summation, $s_1 \geq 0$ and $s_2 \geq 0$, can be chosen arbitrarily provided that the order, $s_1 + s_2$, of the difference expression $h^{-k} \sum a_s u(x + sh)$ satisfies the inequality $s_1 + s_2 \geq k + p - 1$. By Taylor's formula

$$u(x + sh) = u(x) + sh \frac{du(x)}{dx} + \frac{(sh)^2}{2!} \frac{d^2u(x)}{dx^2} + \ldots$$

$$\ldots + \frac{(sh)^{k+p-1}}{(k + p - 1)!} \frac{d^{k+p-1}u(x)}{dx^{k+p-1}} + \frac{(sh)^{k+p}}{(k + p)!} \frac{d^{k+p}u(\xi)}{dx^{k+p}} .$$

Let us substitute this expression, in place of u(x + sh), into (24) and collect like terms. We then get

$$\frac{d^k u(x)}{dx^k} = h^{-k}\Big[u(x) \sum a_s + \frac{du(x)}{dx} \frac{h}{1!} \sum s a_s + \ldots$$

$$\ldots + \frac{d^{k+p-1}u(x)}{dx^{k+p-1}} \frac{h^{k+p-1}}{(k + p - 1)!} \sum s^{k+p-1} a_s\Big] +$$

$$+ \frac{h^p}{(k + p)!} \sum s^{k+p} a_s \frac{d^{k+p}u(\xi_3)}{dx^{k+p}} .$$

Equating coefficients of like powers h^s, $s = -k, -k+1, \ldots, p-1$, on the left- and right-hand sides of this equation, we get the following system of equations for the a_s:

$$\left.\begin{aligned} \sum a_s &= 0, \\ \sum s a_s &= 0, \\ &\cdots\cdots \\ \sum s^{k-1} a_s &= 0, \\ \sum s^k a_s &= k!, \\ \sum s^{k+1} a_s &= 0, \\ &\cdots\cdots \\ \sum s^{k+p-1} a_s &= 0. \end{aligned}\right\} \qquad (25)$$

If $s_1 + s_2 = k + p - 1$, then the above $k + p$ equations form a linear system of this same number of unknowns a_s. The determinant of this system

$$\begin{vmatrix} 1 & 1 & \dots & 1 \\ -s_1 & -s_1 + 1 & \dots & s_2 \\ \dots & \dots & \dots & \dots \\ (-s_1)^{k+p-1} & (-s_1 + 1)^{k+p-1} & \dots & s_2^{k+p-1} \end{vmatrix}$$

is the well-known Vandermonde determinant, and is different from zero. Thus there is a unique set of coefficients, a_s, satisfying system (25). If $s_1 + s_2 \geq k + p$ then, clearly, there will be many such systems of coefficients, a_s.

Thus, for example, there is a unique first-order difference expression of the form

$$h^{-1}[a_0 u(x) + a_1 u(x + h)],$$

approximating du/dx, accurate to first order in h. It is given by the relation

$$\frac{du}{dx} = \frac{u(x + h) - u(x)}{h} + O(h).$$

Likewise there is a unique first-order difference expression of the form

$$h^{-1}[a_{-1} u(x - h) + a_0 u(x)],$$

approximating du/dx to first order in h:

$$\frac{du}{dx} = \frac{u(x) - u(x - h)}{h} + O(h).$$

Among second-order difference expressions of the form

$$h^{-1}[a_{-1} u(x - h) + a_0 u(x) + a_1 u(x + h)]$$

there are infinitely many which approximate du/dx to first order in h, but only one is of second order accuracy. Solving system (25) in this case we see that, for $a_1 = 1/2$, $a_0 = 0$, $a_{-1} = -1/2$

$$\frac{du}{dx} = \frac{u(x + h) - u(x - h)}{2h} + O(h^2).$$

If we want to approximate d^2u/dx^2 to order h^2, then $k = 2$, $p = 2$, and it is necessary that $s_1 + s_2 \geq 3$. Therefore among difference expressions of the form

$$h^{-2}(a_{-1} u(x - h) + a_0 u(x) + a_1 u(x + h) + a_2 u(x + 2h)) \qquad (26)$$

there is only one that has the desired properties. Solving system (21) for the coefficients a_{-1}, a_0, a_1 and a_2 we find that

$$a_{-1} = a_1 = 1, \qquad a_0 = -2, \qquad a_2 = 0,$$

i.e we get the equation (already often used above)

$$\frac{d^2u(x)}{dx^2} = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^2).$$

* * *

7. Other methods for constructing difference schemes. The replacement of derivatives by difference expressions is not the only, and often not the best, method for constructing difference schemes. Later we will devote §19 to some other methods, leading to the most widely-used difference schemes. Here we limit ourselves to a discussion of examples.

The simplest difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1} - u_n}{h} - G(x_n, u_n) = 0, \quad n = 0, 1, \ldots, N-1, \\ u_0 = a. \end{cases}$$

called "Euler's scheme," approximates the problem

$$\left.\begin{aligned} \frac{du}{dx} - G(x,u) = 0, \quad 0 \le x \le 1, \\ u(0) = a \end{aligned}\right\} \tag{27}$$

to first order in h. For given u_n, u_{n+1} is computed from the expression $u_{n+1} = u_n + hg(x_n, u_n)$. The scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1} - u_n}{h} - \dfrac{1}{2}[G(x_n, u_n) + G(x_{n+1}, \bar{u})] = 0, \\ u_0 = a, \end{cases}$$

where $\bar{u}_n = u_n + hG(x_n, u_n)$, is called the "predictor-corrector Euler scheme". It is in fact, one of the Runge-Kutta schemes, with second-order approximation, which will be discussed in detail in §19. If u_n is already computed then, in this scheme, by Euler's method we compute the value

$$\bar{u} = u_n + hG(x_n, u_n),$$

and then carry out a refinement of this $\bar{u}$, setting

$$u_{n+1} = u_n + \frac{h}{2}[G(x_n, u_n) + G(x_{n+1}, \bar{u})],$$

PROBLEMS

1. Verify that the predictor-corrector Euler scheme approximates problem (27), on a smooth solution u(x), to second order in h.

§12. Definition of stability of a difference scheme. Convergence as a consequence of approximation and stability

1. Definition of stability. Suppose that, for the approximate solution of the boundary-value problem

$$Lu = f \tag{1}$$

we have constructed the difference scheme

$$L_h u^{(h)} = f^{(h)}, \tag{2}$$

which approximates problem (1) on the solution u to some order h^k. This means that the residual $\delta f^{(h)}$

$$L_h [u]_h = f^{(h)} + \delta f^{(h)},$$

which appears when the table, $[u]_h$, of values of the solution u, is substituted into Eq. (2), satisfies a bound of the form

$$||\delta f^{(h)}||_{F_h} \leq C_1 h^k, \tag{3}$$

where C_1 is some constant not depending on h. It is easy to verify that the difference scheme

$$L_h u^{(h)} \equiv \begin{cases} 4\,\dfrac{u_{n+1} - u_{n-1}}{2h} - 3\,\dfrac{u_{n+1} - u_n}{h} + Au_n = 0, \\ \qquad n = 1, 2, \ldots, N-1, \\ u_0 = b \end{cases}$$

approximates

$$\frac{du}{dx} + Au = 0,$$

$$u(0) = b,$$

on the solution u, to first order in h. However, as was shown in §9, the solution $u^{(h)}$, obtained via this difference scheme, does not tend to $[u]_h$ as $h \to 0$.

Thus, generally, approximation is not sufficient for convergence. Stability is needed in addition.

Definition 1. We will call difference scheme (2) *stable* if there exists numbers, $h_0 > 0$ and $\delta > 0$, such that for any $h < h_0$ and any $\varepsilon^{(h)}$ in F_h, $||\varepsilon^{(h)}||_{F_h} < \delta$, the difference problem

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}, \tag{4}$$

obtained from problem (2) through the addition to the right-hand side of a perturbation $\varepsilon^{(h)}$, has one and only one solution $z^{(h)}$; and moreover, this solution deviates from the solution, $u^{(h)}$, of the unperturbed problem (2) by a net function $z^{(h)} - u^{(h)}$, satisfying the bound

$$||z^{(h)} - u^{(h)}||_{U_h} \leq C||\varepsilon^{(h)}||_{F_h}, \tag{5}$$

where C is some constant not depending on h.

Inequality (5) signifies that a small perturbation, $\varepsilon^{(h)}$, of the right-hand side of difference scheme (2) evokes a perturbation, $z^{(h)} - u^{(h)}$, in the solution which is uniformly small with respect to h.

Suppose the operator mapping U_h into F_h is linear. Then the above definition of stability is equivalent to the following:

Definition 2. We will call the difference scheme (2), with linear operator L_h *stable*, if for any $f^{(h)}$ in F_h, the equation $L_h u^{(h)} = f^{(h)}$ has a unique solution $u^{(h)}$ in U_h, and

$$||u^{(h)}||_{U_h} \leq C||f^{(h)}||_{F_h}, \tag{6}$$

where C is some constant not depending on h. We now prove the equivalence of both definitions of stability for a linear operator L_h.

First we establish that, from the stability of difference scheme (2) in the sense of definition 2, follows stability in the sense of definition 1. Suppose the linear problem (2), for all $h < h_0^*$ and arbitrary $f^{(h)}$ in F_h, has a unique solution satisfying bound (6). Subtracting Eq. (2) from Eq. (4) we get

---

*Apparently the requirement $h < h_0$ is implicitly assumed in definition 2. (Translator's note.)

$$L_h(z^{(h)} - u^{(h)}) = \varepsilon^{(h)},$$

from which (by virtue of (6)) (5) follows for any $\varepsilon^{(h)}$ in F_h, and from (5) follows stability in the sense of definition 1.

Now we will show that stability in the sense of definition 1 implies stability in the sense of definition 2. By definition 1, for some $h_0 > 0$ and $\delta > 0$, and for arbitrary $h < h_0$, with $\varepsilon^{(h)}$ in $F^{(h)}$ such that $||\varepsilon^{(h)}||_{F_h} < \delta$, there exist unique solutions of the equations

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)},$$

$$L_h u^{(h)} = f^{(h)}.$$

Set $w^{(h)} \equiv z^{(h)} - u^{(h)}$, and subtract the above equations term by term. We then get

$$L_h w^{(h)} = \varepsilon^{(h)},$$

where, moreover, from (5)

$$||w^{(h)}||_{U_h} \leq C||\varepsilon^{(h)}||_{F_h}.$$

It is clear that, if we change the notation for the solution and right hand side of the equation $L_h w^{(h)} = \varepsilon^{(h)}$, this last result can be stated thus: for arbitrary $h < h_0$ and $f^{(h)}$ in F_h, $||f^{(h)}||_{F_h} < \delta$, problem (2) has a unique solution $u^{(h)}$. This solution satisfies bound (6). But then it must be true that Eq. (2) has a unique solution $u^{(h)}$, and that bound (6) is satisfied, not only for all $f^{(h)}$ such that $||f^{(h)}||_{F_h} < \delta$, but also in general for all $f^{(h)}$ in $F^{(h)}$, i.e. we have stability in the sense of definition 2.

In fact let $||f^{(h)}||_{F_h} \geq \delta$. Let us demonstrate existence and uniqueness of a solution, as well as the validity of (6), in this case. Let

$$u^{(h)} = \frac{2||f^{(h)}||_{F_h}}{\delta} \tilde{u}^{(h)}, \qquad f^{(h)} = \frac{2||f^{(h)}||_{F_h}}{\delta} \tilde{f}^{(h)}$$

For $\tilde{u}^{(h)}$ we get the equation

$$L_h \tilde{u}^{(h)} = \tilde{f}^{(h)},$$

with

$$||\tilde{f}^{(h)}||_{F_h} = \frac{\delta}{2||f^{(h)}||_{F_h}} ||f^{(h)}||_{F_h} = \frac{\delta}{2} < \delta.$$

Therefore $L_h\tilde{u}^{(h)} = \tilde{f}^{(h)}$ has a unique solution, and moreover

$$||\tilde{u}^{(h)}||_{U_h} \leq C||\tilde{f}^{(h)}||_{F_h}.$$

By virtue of the equations establishing the relation between $u^{(h)}$ and $\tilde{u}^{(h)}$, and between $f^{(h)}$ and $\tilde{f}^{(h)}$, it follows from the above inequality that problem (2) has a unique solution, and that bound (6) is valid for arbitrary $f^{(h)}$ in F_h.

2. Connection between approximation, stability and convergence. We show, now, that from approximation and stability follows convergence.

Theorem. *Suppose that the difference scheme* $L_h u^{(h)} = f^{(h)}$ *approximates the problem* $Lu = f$ *on the solution* u *to order* h^k, *and is stable. Then the solution,* $u^{(h)}$, *of the difference problem* $L_h u^{(h)} = f^{(h)}$ *converges to* $[u]_h$, *satisfying the bound*

$$||[u]_h - u^{(h)}||_{U_h} \leq (CC_1)h^k, \tag{7}$$

where C and C_1 *are the numbers entering into bounds (3) and (5).*

Proof. Define $\varepsilon^{(h)} \equiv \delta f^{(h)}$, $[u]_h \equiv z^{(h)}$. Then bound (5) takes the form

$$||[u_h] - u^{(h)}||_{U_h} \leq C||\delta f^{(h)}||_{F_h}.$$

Taking note of (3) we easily get Eq. (7), which was to be demonstrated.

As an illustrative example, we prove the stability of Euler's difference scheme

$$\left.\begin{aligned} \frac{u_{n+1} - u_n}{h} - G(x_n, u_n) &= \phi_n, \qquad n = 0, 1, \ldots, N-1, \\ u_0 &= \psi. \end{aligned}\right\} \tag{8}$$

$x_n = nh$, $h = 1/n$, for the numerical solution of the differential boundary-value problem

$$\left.\begin{aligned} \frac{du}{dx} - G(x, u) &= \phi(x), \qquad 0 \leq x \leq 1, \\ u_0 &= \psi. \end{aligned}\right\} \tag{9}$$

We will assume that the function, $G(x, u)$, of two arguments, and the function $\phi(x)$, are such that there exists a solution, $u(x)$, with bounded

second derivative. In addition, we suppose that G(x,u) has a bounded u-derivative

$$\left|\frac{\partial G}{\partial u}\right| < M. \tag{10}$$

We suggest that the reader verify that the difference scheme (8) approximates (9) on the solution u(x) to first order in h. (The difference equation represents the differential equation to first order, and the boundary conditions $u_0 = \psi$ is exact.) We define the norms

$$||u^{(h)}||_{U_h} = \max_n |u_n|, \qquad ||f^{(h)}||_{F_h} = \max\{|\psi|, \max_m |\phi(x_m)|\}$$

and proceed to verify the stability of difference scheme (8). Let us write this scheme in form (2), setting

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - G(x_n, u_n), & n = 0, 1, \ldots, N-1, \\ u_0, \end{cases}$$

$$f^{(h)} = \begin{cases} \phi(x_n), & n = 0, 1, \ldots, N-1, \\ \psi. \end{cases}$$

The problem

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}$$

has the explicit form

$$\left.\begin{aligned} \frac{z_{n+1} - z_n}{h} - G(x_n, z_n) &= \phi(x_n) + \varepsilon_n, \qquad n = 0, 1, \ldots, N-1, \\ z_0 &= \psi + \varepsilon, \end{aligned}\right\} \tag{11}$$

where

$$\varepsilon^{(h)} = \begin{cases} \varepsilon_n, & n = 0, 1, \ldots, N-1, \\ \varepsilon. \end{cases}$$

Let us subtract, from Eqs. (11), the corresponding Eqs. (8), term by term. We define

$$z_n - u_n = w_n$$

and note that

$$G(x_n, z_n) - G(x_n u_n) = \frac{\partial G(x_n, \xi_n)}{\partial u} w_n = M_n^{(h)} w_n,$$

where ξ_n is some number between z_n and u_n. We then get the following system of equations determining $w^{(h)} = (w_0, w_1, \ldots, w_n, .., w_N)$:

$$\left.\begin{aligned} \frac{w_{n+1} - w_n}{h} - M_n^{(h)} w_n &= \varepsilon_n, \qquad n = 0, 1, \ldots, N-1, \\ w_0 &= \varepsilon. \end{aligned}\right\} \tag{12}$$

Taking account of the fact that $M_n^{(h)} \leq M$ by virtue of (10), and that $nh \leq Nh = 1$, we get

$$\begin{aligned} |w_{n+1}| &= |(1 + hM_n^{(h)})w_n + h\varepsilon_n| \leq (1 + Mh)|w_n| + h|\varepsilon_n| \leq \\ &\leq (1 + Mh)^2|w_{n-1}| + h(1 + Mh)|\varepsilon_{n-1}| + h|\varepsilon_n| \leq \\ &\leq (1 + Mh)^2|w_{n-1}| + 2h(1 + Mh)||\varepsilon^{(h)}||_{F_h} \leq \\ &\leq (1 + Mh)^3|w_{n-2}| + 3h(1 + Mh)^2||\varepsilon^{(h)}||_{F_h} \leq \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\leq (1 + Mh)^{n+1}|w_0| + (n+1)h(1 + Mh)^n||\varepsilon^{(h)}||_{F_h} \leq \\ &\leq (1 + Mh)^{n+1}||\varepsilon^{(h)}||_{F_h} + (1 + Mh)^n||\varepsilon^{(h)}||_{F_h} \leq \\ &\leq 2(1 + Mh)^N||\varepsilon^{(h)}||_{F_h} \leq 2e^M||\varepsilon^{(h)}||_{F_h}. \end{aligned}$$

From the demonstrated inequality

$$|w_{n+1}| \leq 2e^M||\varepsilon^{(h)}||_{F_h}$$

follows a bound of form (6)

$$||w^{(h)}||_{U_h} \leq 2e^M||\varepsilon^{(h)}||_{F_h},$$

signifying stability with constant $C = 2\exp(M)$. By virtue of the above theorem, difference scheme (8) is convergent to first order in h.

Now let us study the convergence of difference scheme (7) §10

$$\left.\begin{aligned} &\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - (1 + x_n^2)u_n = \sqrt{1 + x_n}, \\ &\qquad n = 1, 2, \ldots, N-1, \\ &\qquad u_0 = 2, \\ &\qquad u_N = 1 \end{aligned}\right\} \tag{13}$$

for the differential boundary-value problem (4) §10. That problem (13) approximates (4) §10 to second order in h is obvious,* since

$$\frac{u(x + h) - 2u(x) + u(x - h)}{h^2} = u''(x) + \frac{h^2}{12} u^{(4)}(\xi)$$

Next we set out to verify stability. The problem under consideration is linear. Therefore proof of stability consists in that one establishes existence of a unique solution of the problem

$$\left.\begin{aligned} &\frac{u_{n+1} - 2u_n + u_{n+1}}{h^2} - (1 + x_n^2)u_n = g_n, \quad n = 1, 2, \ldots, N-1, \\ &\qquad u_0 = \alpha, \\ &\qquad u_N = \beta \end{aligned}\right\} \tag{14}$$

for any $\{g_n\}$, α and β, and derives the bound

$$\max_n |u_n| \leq C(\max_n |g_n|, |\alpha|, |\beta|). \tag{15}$$

A problem of form (14) was considered in §4 (see p. 34). There, for the problem

$$\left.\begin{aligned} &a_n u_{n-1} + b_n u_n + c_n u_{n+1} = g_n, \\ &\qquad u_0 = \alpha, \\ &\qquad u_N = \beta \end{aligned}\right\}$$

*We note that, if we were dealing with the solution of the equation

$$u'' - (1 + x^2)u = \sqrt{x},$$

differing very little from the one considered here, then we would not be able to deduce approximation, since $|u'''(x)|$ would, in the given case, be unbounded (prove, rigorously, that $u'''(x)$ is unbounded).

on the assumption that

$$|b_n| \geq |a_n| + |c_n| + \delta, \qquad \delta > 0$$

it was shown that a unique solution exists, and that

$$|u_n| < \max\left[|\alpha|, |\beta|, \frac{1}{\delta} \max_m |g_m| \right]. \tag{16}$$

In the case of problem (14)

$$a_n = \frac{1}{h^2}, \quad c_n = \frac{1}{h^2}, \quad |b_n| = \frac{2}{h^2} + 1 + x_n^2 > |a_n| + |c_n| + 1.$$

Therefore bound (16) implies bound (15) with C = 1. Stability has been proven.

We note one detail which can be useful in proving convergence through verification of approximation and stability. Suppose difference scheme (2) is split into the two subsystems

$$\ell_h^{(0)} u^{(h)} = f_0^{(h)}, \tag{17}$$

$$\ell_h^{(1)} u^{(h)} = f_1^{(h)} \tag{17'}$$

so that

$$L_h u^{(h)} \equiv \begin{cases} \ell_h^{(0)} u^{(h)}, \\ \ell_h^{(1)} u^{(h)}, \end{cases} \qquad f^{(h)} = \begin{cases} f_0^{(h)}, \\ f_1^{(h)}, \end{cases} \qquad \delta f^{(h)} = \begin{cases} \delta f_0^{(h)}, \\ \delta f_1^{(h)}. \end{cases}$$

Assume, further, that difference scheme (2) approximates problem (1) to order h^k, i.e. Eq. (3) is satisfied. Suppose that, in addition, subsystem (17') agrees with problem (1), on the solution u, exactly, i.e. $\delta f_1^{(h)} = 0$, with 0 in $F_h^{(1)}$:

$$\delta f^{(h)} = \begin{cases} \delta f_0^{(h)}, \\ 0. \end{cases} \tag{18}$$

In such a case, for convergence of the solution, $u^{(h)}$, of problem (2) to the required net function $[u]_h$ (i.e. for the validity of bound (7)), it is sufficient that (5) be satisfied, not for all arbitrary $\varepsilon^{(h)}$ in F_h, but only for all $\varepsilon^{(h)}$ of the form

$$\varepsilon^{(h)} = \begin{cases} \varepsilon_0^{(h)}, \\ 0, \end{cases} \tag{19}$$

where 0 is in $F_1^{(h)}$. A proof of this conclusion coincides verbatim with the proof of the above convergence theorem. The reader can easily verify that, for a linear, operator, L_h, the requirement that bound (5) be satisfied only for all $\varepsilon^{(h)}$ of form (19) is equivalent to the requirement that bound (6) be satisfied for all $f^{(h)}$ of the same special form

$$f^{(h)} = \begin{cases} f_0^{(h)}, \\ 0, \end{cases}$$

with 0 in $F_h^{(1)}$.

For example, in proving convergence of difference scheme (13), it would have been possible to make use of the fact that both boundary conditions,

$$\ell_h^{(1)} u^{(h)} \equiv \left\{ \begin{matrix} u_0 = 2, \\ u_N = 1 \end{matrix} \right\} \equiv f_h^{(1)},$$

upon substitution of the tabulated values, $[u]_h$, of the solution of problem (4) §10, are satisfied exactly:

$$\ell_h^{(1)} [u]_h = \begin{cases} u(0) = 2, \\ u(1) = 1. \end{cases}$$

Therefore the proof of inequality (15), signifying the stability of difference scheme (13), could have been carried out, not for an arbitrary right-hand side

$$f^{(h)} = \begin{cases} g_n, & n = 1, 2, \ldots, N-1, \\ \alpha, \\ \beta, \end{cases}$$

but only for a right-hand side of form

$$f^{(h)} = \begin{cases} g_n, & n = 1, 2, \ldots, N-1, \\ 0, \\ 0, \end{cases}$$

where we have taken $\alpha = 0$ and $\beta = 0$.

In problem (13) we dealt with the proof of the stability-inequality even without taking account of this simplification. In more complicated problems (for equations with partial derivatives) the above considerations will sometimes prove useful.

To end this section we underline the fact that the overall plan of the proof of convergence of the solution of the problem $L_h u^{(h)} = f^{(h)}$, to the solution of $Lu = f$, via verification of approximation and stability is very general in character. In the role of the equation $Lu = f$ we can put any functional equation, not only a boundary value problem for an ordinary differential equation. In itself it is not important what sort of problem is solved by the function u. The equation $Lu = f$ is used only as a basis for the construction of the difference equation $L_h u^{(h)} = f^{(h)}$. These considerations will be clarified below, in 3.

* * * * * *

3. Convergent difference scheme for an integral equation. We now construct and study a difference scheme for computing the solution of the integral equation

$$Lu \equiv u(x) - \int_0^1 K(x,y)\, u(y)\, dy = f(x).$$

We will assume that $|K(x,y)| < \rho < 1$.

For a given N we set $h = 1/N$ and seek to obtain a table of the solution, $[u]_h$, on the net $x_n = nh$, $n = 0, 1, \ldots, N$. To arrive at a difference scheme we approximate the integral in the equation

$$u(x_n) - \int_0^1 K(x_n, y)\, u(y)\, dy = f(x_n), \qquad n = 0, 1, \ldots, N,$$

by a sum, using the trapezoidal integration formula. We recall the structure of this formula: for any function, $\phi(y)$, twice differentiable on the interval $0 \le y \le 1$, we may write

$$\int_0^1 \phi(y)\, dy \approx h\left(\frac{\phi_0}{2} + \phi_1 + \phi_2 + \ldots + \phi_{N-1} + \frac{\phi_N}{2}\right), \qquad h = \frac{1}{N},$$

where the error is $O(h^2)$. After the above replacement of the integral we get

$$u_n^{(h)} - h\left(\frac{K(x_n, 0)}{2} u_0^{(h)} + K(x_n, h)\, u_1^{(h)} + \ldots\right.$$

$$\left. \ldots + K(x_n, y_{N-1}) u_{N-1}^{(h)} + \frac{K(x_n, 1)}{2} u_N^{(h)}\right) = f(x_n), \qquad n = 0, 1, \ldots, N. \quad (20)$$

This system of equations can be written in the form $L_h u^{(h)} = f^{(h)}$ if we define

$$L_h u^{(h)} = \begin{cases} g_0 \\ g_1 \\ \cdots \\ g_N \end{cases} \qquad f^{(h)} = \begin{cases} f(0) \\ f(h), \\ \cdots \\ f(1), \end{cases}$$

where

$$g_n = u_n^{(h)} - h\left[\frac{K(x_n, 0)}{2} u_0^{(h)} + K(x_n, h)u_1^{(h)} + \ldots + \frac{K(x_n, 1)}{2} u_N^{(h)}\right],$$

$$n = 0, 1, \ldots, N.$$

This scheme $L_h u^{(h)} = f^{(h)}$ approximates the problem $Lu = f$, on the solution u, to second order in the step-size h, since the trapezoidal quadrature formula is accurate to second order. We now verify stability. Let $u^{(h)} = (u_0, u_1, \ldots, u_N)$ be any solution of system (20) and let u_s be one of those components of the solution whose modulus is no less than that of any other:

$$|u_s| = \max_m |u_m|.$$

From the equation of system (20) numbered $n = s$, it follows that

$$|f(x_s)| = \left|u_s^{(h)} - h\left(\frac{K(x_s, 0)}{2} u_0^{(h)} + K(x_s, h)u_1^{(h)} + \ldots + \frac{K(x_s, 1)}{2} u_N^{(h)}\right)\right| \geq$$

$$\geq \left|u_1^{(h)}\right| - h\left(\frac{\rho}{2} + \rho + \ldots + \rho + \frac{\rho}{2}\right)\left|u_s^{(h)}\right| = (1 - Nh\rho)\left|u_s^{(h)}\right| = (1 - \rho)\left|u_s^{(h)}\right|.$$

Therefore

$$\|u^{(h)}\|_{U_h} = \max_m \left|u_m^{(h)}\right| = \left|u_s^{(h)}\right| \leq \frac{1}{1 - \rho} |f(x_s)| \leq \frac{1}{1 - \rho} \|f^{(h)}\|_{F_h}. \quad (21)$$

From this it follows, in the special case $f(x_n) \equiv 0$, that system (20) has no nontrivial solution, and therefore has one and only one solution for any right-hand side $f^{(h)}$. Inequality (21) signifies stability, since it is equivalent to (6) with constant $C = 1/(1 - \rho)$. The solution, $u^{(h)}$, of the problem $L_h u^{(h)} = f^{(h)}$, by virtue of the convergence theorem, satisfies the inequality

$$\|[u]_h - u^{(h)}\|_{U_h} = \max_m \left|u(mh) - u_m^{(h)}\right| \leq Ah^2,$$

where A is some constant.

* * *

§13. On the choice of a norm

The concepts of convergence, approximation and stability, introduced in §§10-12, are meaningful if, in one way or another, norms have been introduced in the spaces U_h and F_h, to which belong, respectively, the solution, $u^{(h)}$, and the right-hand side, $f^{(h)}$, of the difference scheme

$$L_h u^{(h)} = f^{(h)}$$

for the approximate computation of the solution, u, of the differential boundary-value problem Lu = f.

We now discuss to what extent the choice of norms, in the spaces U_h and F_h, is arbitrary. We begin with the norm $\|\cdot\|_{U_h}$, whose value measures the deviation of the approximate solution, $u^{(h)}$, from the net function $[u]_h$, i.e. from the tabulated values of the solution u.

In all the examples we have considered we have used the norm defined by the equation

$$\|z^{(h)}\|_{U_h} = \max_k \left|z_k^{(h)}\right| . \tag{1}$$

The maximum is taken over all points of the net, D_h, on which the net function $z^{(h)}$ in U_h is defined. We could, of course, have taken

$$\|z^{(h)}\|_{U_h} = h \max_k \left|z_k^{(h)}\right| , \tag{2}$$

or

$$\|z^{(h)}\|_{U_h} = \frac{1}{h} \max_k \left|z_k^{(h)}\right| , \tag{3}$$

or even

$$\|z^{(h)}\|_{U_h} = 2^{-1/h} \max_k \left|z_k^{(h)}\right| .$$

This latter norm may seem to be useful since, using this norm, the scheme

$$\left.\begin{aligned} 4\,\frac{u_{n+1} - u_{n-1}}{2h} - 3\,\frac{u_{n+1} - u_n}{h} + Au_n = 0, \\ n = 0, 1, \ldots, N-1, \\ u_0 = a, \\ u_1 = ae^{-Ah} \end{aligned}\right\}$$

for solving the problem

$$\left.\begin{aligned} u' + Au = 0, \qquad 0 \le x \le 1, \\ u(0) = \alpha, \end{aligned}\right\}$$

cited in §9 as an example of an unuseable scheme, is now convergent. In fact by virtue of the equation

$$u(nh) - u_n = ae^{-Ax_n} - u_n = -\left[\frac{3}{2} A^2 ae^{Ax_n} + O(h)\right] h 2^{x_n/h} + O(h),$$

which follows from (7) §9, the quantity

$$\|[u]_h - u^{(h)}\|_{U_h} = 2^{-1/h} \max_m \left|u(mh) - u_m^{(h)}\right|$$

tends to zero as the net is refined. But it is clear that the approach of this quantity to zero does not, by any reasonable interpretation, imply that the error, $z^{(h)} = [u]_h - u^{(h)}$, tends to zero, insofar as the difference, $u(nh) - u_n$, is allowed to increase very rapidly (almost like $2^{1/h}$), as it does, in fact, in this example. Norms (2) and (3) also are not to be recommended since they also inadequately characterize the error $[u]_h - u^{(h)}$.

It is customary to choose a norm in the space U_h in such a way that, as h tends to zero, it will go over into some norm for functions given on the whole interval, i.e. so that

$$\lim_{h\to 0} \|[u]_h\|_{U_h} = \|u\|_U, \tag{4}$$

where $\|\cdot\|_U$ is a norm in that space of functions, on the given interval, to which u(x) belongs. The norm

$$\|z^{(h)}\|_{U_h} = \max_m \left|z_m^{(h)}\right|$$

satisfies this requirement, if we take as U the space of continuous functions in which

$$\|u\|_U = \max_{0\le x\le 1} |u(x)|,$$

and let the net function $[u]_h$ coincide with u(x) at the points of the net.

The norm

$$\|z^{(h)}\|_{U_h} = \sqrt{h \sum_m \left|z_m^{(h)}\right|^2} \tag{5}$$

is also reasonable. It satisfies condition (4) if we take, as U, the space of continuous functions with norm

$$||u||_U = \sqrt{\int_0^1 u^2(x)\,dx},$$

and define the net function $[u]_h$, as before, to coincide with u(x) at net-points.

In the case of a discontinuous solution u(x) which, however, is square-integrable, we may take as U the space of square-integrable functions with the norm

$$||u||_U = \sqrt{\int_0^1 u^2(x)dx},$$

but define the value, u_n, of the net function $[u]_h$, not by the equation $u_n = u(nh)$ (which may not be meaningful), but by the expression

$$u_n = \frac{1}{h}\int_{x_n-h/2}^{x_n+h/2} u(x)dx.$$

Then also for the discontinuous function we will have

$$\lim_{h\to 0} ||[u]_h||_{U_h} = \sqrt{\int_0^1 u^2 dx}.$$

It is clear that convergence

$$\lim_{h\to 0} ||[u]_h - u^{(h)}||_{U_h} = 0$$

in the sense of norm (1), i.e. uniform convergence, implies convergence in the sense of norm (5), i.e. convergence in the mean, but uniform convergence does not follow from convergence in the mean. Therefore, from among the various reasonable norms satisfying condition (4), one chooses that one in which one can prove the convergence of the particular difference scheme under consideration. For this choice there is no general prescription.

In the case of ordinary differential equations and the corresponding difference equations, which we are studying in this chapter, it is generally satisfactory to use norms (1) or (5), or a norm of the type

$$||z^{(h)}||_{U_h} = \max\left[\max_m \left|z_m^{(h)}\right|,\ \max_m \frac{\left|z_{m+1}^{(h)} - z_m^{(h)}\right|}{h}\right]. \qquad (6)$$

which takes account of the change in the net function from point to point. Equation (4) is satisfied for this norm, if, as U, we take the space of continuous and differentiable functions with norm

$$||u||_U = \max\left[\max_{0\le x\le 1}|u(x)|,\ \max_{0\le x\le 1}|u'(x)|\right].$$

In the case of partial differential equations and the corresponding difference schemes it is sometimes convenient to use quite contrived norms, designed for specific problems.

Let us proceed, now, to consider the choice of a norm in the space F_h, which contains the right-hand side of the difference equation $L_h u^{(h)} = f^{(h)}$. We underscore that convergence of the difference scheme $||[u]_h - u^{(h)}||_{U_h} \to 0$, for the selected norm $||\cdot||_{U_h}$ does not depend on the choice of a norm $||\cdot||_{F_h}$, nor is it even relevant whether **any** such norm has been chosen. It is necessary to consider F_h as a linear normed space only in order to reduce the convergence proof, and verification of the order of accuracy of the difference scheme, to a verification of some order of approximation, and verification of stability.

We will discuss the choice of a norm in F_h assuming linearity of the difference scheme $L_h u^{(h)} = f^{(h)}$. This will be done only to avoid unessential complications.

Suppose that, for some choice of norm, $||\cdot||^{(1)}$, the difference scheme $L_h u^{(h)} = f^{(h)}$ approximates the problem $Lu = f$ on the solution u to some order h^k, and is stable. Then by virtue of the convergence theorem the difference scheme $L_h u^{(h)} = f^{(h)}$ is convergent, with order of accuracy h^k:

$$||[u]_h - u^{(h)}||_{U_h} < Ch^k. \tag{7}$$

Recall that approximation means the satisfaction of an inequality of the form

$$||L_h[u]_h - f^{(h)}||^{(1)}_{F_h} \le C_1 h^k. \tag{8}$$

Stability means that the problem $L_h u^{(h)} = f^{(h)}$ has a unique solution for any $f^{(h)}$ in F_h, and moreover

$$||u^{(h)}||_{U_h} \le C_2 ||f^{(h)}||^{(1)}_{F_h}. \tag{9}$$

If we choose another norm, $||\cdot||_{F_h}$, defining

$$||f^{(h)}||^{(2)}_{F_h} = h||f^{(h)}||^{(1)}_{F_h}, \tag{10}$$

then, obviously, inequalities (8) and (9) are replaced, respectively, by the inequalities

$$\left.\begin{aligned} ||L_h[u]_h - f^{(h)}||^{(2)}_{F_h} &\le Ch^{k+1}, \\ ||u^{(h)}||_{U_h} &\le \frac{C}{h}\,||f^{(h)}||^{(2)}_{F_h}. \end{aligned}\right\} \tag{11}$$

Thus approximation will no longer be of order k with respect to the step-size h, but of one higher order, k+1. Judging by these facts, one might mistakenly conclude that the order of accuracy of the difference scheme is not h^k, but h^{k+1}. The trouble is that inequality (9) no longer signifies stability which, for the new choice of norm, is generally lost.

If instead of (10) we had introduced the norm $||\cdot||^{(2)}_{F_h}$ via the equation

$$||f^{(h)}||^{(2)}_{F_h} = \frac{1}{h}\,||f^{(h)}||^{(1)}_{F_h},$$

then, in place of (8) and (9), we would have gotten, respectively,

$$||L_h[u]_h - f^{(h)}||^{(2)}_{F_h} \le C_1 h^{k-1}, \tag{12}$$

$$||u^{(h)}||_{U_h} \le C_2 h\,||f^{(h)}||^{(2)}_{F_h}. \tag{13}$$

Inequality (13) guarantees stability since $C_2 h$ can be replaced by a constant, C_2, not depending on h, thereby only strengthening the inequality. Inequality (12) indicates approximation of order k-1 with respect to the step-size h.

Thus, having chosen the norm $||\cdot||^{(2)}_{F_h}$, we would only be able, on the basis of the convergence theorem, to guarantee k-1'st order accuracy for the difference scheme $L_h u^{(h)} = f^{(h)}$, one order lower than is guaranteed by inequality (7). This loss of information on order of accuracy has occurred because of a poor choice of norm in the space F_h.

So as to determine, correctly, the order of accuracy of a difference scheme one must choose the norm, $||\cdot||_{F_h}$, in such a way that the order of approximation is as high as possible, while stability is still not lost. For this choice of norm there is no general rule.* Further, it is not always possible to choose a norm in such a way as to give both approximation and stability; otherwise, contrary to what was shown via the example in §9, every difference scheme would be convergent.

---

*We have in mind, here, also the case of difference schemes for partial differential equations.

We will present, however, one general consideration which may help us choose a norm correctly in the linear space F_h. In choosing the norm $||\cdot||_{F_h}$ one must take account of the nature of the continuous dependence of the solution of the differential boundary-value problem, on which the difference scheme $L_h u^{(h)} = f^{(h)}$ was based, on the right-hand side f.

For example, in the case of the problem

$$\frac{du}{dx} + Au = \phi(x), \qquad u(0) = a, \qquad 0 \leq x \leq 1$$

when one introduces the increments $\delta\phi(x)$ and δa, into the right-hand side and the boundary condition, respectively, the solution u(x) changes by an amount, $\delta u(x)$, of the same order of magnitude.

Now let us consider the difference schemes

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} + Au_n = \phi(x_n), & n = 0, 1, \ldots, N-1, \\ u_0 = a, \end{cases}$$

such that

$$f^{(h)} = \begin{cases} \phi(x_n), & n = 0, 1, \ldots, N-1, \\ a. \end{cases}$$

The norm in U_h, as usual, will be given by the equation

$$||u^{(h)}||_{U_h} = \max_m \left|u_m^{(h)}\right| .$$

Stability can be expected only if the norm

$$||f^{(h)}||_{F_h} = \left|\left|\begin{matrix}\phi(x_n) \\ a\end{matrix}\right|\right|$$

depends in some substantial way on both $\phi(x_n)$ and a. The norm, for example may have the form

$$||f^{(h)}||_{F_h} = \max[|a|, \max_m |\phi_m|]. \tag{14}$$

Stability in this norm was proven in §12, where a more general, nonlinear, problem was considered.

One cannot expect stability if a norm is chosen, let us say, according to the equation

$$||f^{(h)}||_{F_h} = \max[h|a|, \max_m |\phi_m|],$$

where a enters more and more weakly as h decreases.

Stability in the sense of this norm would signify that $u^{(h)}$ depends more weakly on a than does the solution, u, of the differential equation. On the other hand for small h, by virtue of convergence (and convergence would follow from stability, since we already have approximation) the solution of the difference equation differs little from the solution of the differential equation; it must change, therefore, when the initial value, a, changes, by about the same amount as the solution u(x).

More concisely: for the given choice of norm the problem

$$\left.\begin{aligned} \frac{u_{n+1} - u_n}{h} + Au_n &= \phi_n, \qquad n = 0, 1, \ldots, N-1, \\ u_0 &= 0 \end{aligned}\right\}$$

approximates the problem

$$\frac{du}{dx} + Au = \phi(x), \qquad u(0) = a$$

on the solution u(x) for any a. Thus, given stability, the function $u^{(h)}$, not depending on a, would have to converge to the solution u(x) whatever the value of a. But $u^{(h)}$ cannot converge simultaneously to different functions u(x).

In the case of the difference scheme

$$\left.\begin{aligned} \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + A\frac{u_{n+1} - u_{n-1}}{2h} + Bu_n &= \phi_n, \\ n &= 1, \ldots, N, \\ u_0 &= a, \\ \frac{u_1 - u_0}{h} &= b. \end{aligned}\right\} \qquad (15)$$

for the problem

$$\left.\begin{aligned} \frac{d^2u}{dx^2} + A\frac{du}{dx} + Bu &= \phi(x), \\ u(0) &= a, \\ \frac{du(0)}{dx} &= b \end{aligned}\right\}$$

from these same considerations the norm

$$||f^{(h)}||_{F_h} = \left|\left|\begin{matrix}\phi(x_m)\\ a\\ b\end{matrix}\right|\right|_{F_h}$$

must depend, in some essential way, on ϕ, a and b. It may have the form

$$||f^{(h)}||_{F_h} = \max[|a|, |b|, \max_m |\phi_m|], \tag{16}$$

but one cannot expect stability if one chooses, as a norm, let us say, the quantity

$$||f^{(h)}||_{F_h} = \max[|a|, h|b|, \max_m |\phi_m|].$$

Let us now rewrite (15) in a somewhat different form:

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + A\,\dfrac{u_{n+1} - u_{n-1}}{2h} + Bu_n = \phi(x_n) \\ u_0 = a, \\ u_1 = a + bh, \end{cases} \tag{17}$$

so that

$$f^{(h)} = \begin{cases} \phi(x_n), & n = 1, 2, \ldots, N-1, \\ a, \\ a + bh. \end{cases}$$

The norm in F_h must now be introduced, for any given $g^{(h)}$,

$$g^{(h)} \equiv \begin{cases} \psi_m \\ \alpha \\ \beta \end{cases},$$

by an equation of the type

$$||g^{(h)}||_{F_h} = \max[|\alpha|, \frac{|\beta - \alpha|}{h}, \max_m|\psi_m|], \tag{18}$$

where $|\beta - \alpha|$ enters with increasing weight $1/h$ as $h \to 0$. In fact a change in α or β of order h is equivalent to a change in u_0 or u_1 of order h; but then $|u_1 - u_0|/h$ changes by a quantity of order 1. Such a change, if the scheme is stable, implies a change in the solution of the equation

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + A\,\frac{u_{n+1} - u_{n-1}}{2h} + Bu_n = \phi_n$$

by a quantity of order 1; clearly an O(1) change in $|u_1 - u_0|/h$ is analagous to a change in the right-hand side of the condition $du(0)/dx = b$, a boundary condition of the differential problem, by a quantity of order 1.

One cannot expect stability if the norm is defined by

$$||g^{(h)}||_{F_h} = \max[|\alpha|, |\beta|, \max_m|\psi_m|],$$

i.e. as it was defined earlier when we used the space F_h in conjunction with difference scheme (15). The order of approxiamtion of schemes (15) and (17), with norms (16) and (18), respectively, is the same for both schemes -- first order in h. The stability of schemes (15) and (17), with norms (16) and (18), will be proven in §14.

§14. Sufficient condition for stability of difference schemes for the solution of the Cauchy problem

Below we will show how to study the stability of difference schemes $L_h u^{(h)} = f^{(h)}$ for the solution of differential problems with initial conditions (Cauchy problems). We do this via consideration of typical examples of difference schemes approximating the problems

$$Lu = \begin{cases} \dfrac{du}{dx} + Au = \phi(x), & 0 \le x \le 1, \\ u(0) = a, \end{cases} \tag{1}$$

$$Lu \equiv L\begin{pmatrix} v \\ w \end{pmatrix} = \begin{cases} \dfrac{dv}{dx} + Av + Bw = p(x), & 0 \le x \le 1, \\ \dfrac{dw}{dx} + Cv + Dw = q(x), & 0 \le x \le 1, \\ v(0) = a, \\ w(0) = b, \end{cases} \tag{2}$$

$$Lu \equiv \begin{cases} \dfrac{d^2u}{dx^2} + A\dfrac{du}{dx} + Bu = \phi(x), & 0 \le x \le 1, \\ u(0) = a, \\ \dfrac{du(0)}{dx} = b. \end{cases} \tag{3}$$

If the concept of stability of the difference scheme $L_h u^h = f^{(h)}$ is to have any meaning one must define the linear normed spaces U_h and F_h. The first space contains the table $[u]_h$ in U_h, which we are to calculate, i.e. the table of the function, u, which solves the differential problem; to the

second space belongs the right hand side, $f^{(h)}$ in F_h, of the difference scheme.

We recall that the difference scheme $L_h u^{(h)} = f^{(h)}$, with linear operator L_h, is said to be "stable" if the problem $L_h u^{(h)} = f^{(h)}$ has a unique solution $u^{(h)}$ in U_h for any $f^{(h)}$ in F_h and, moreover, the condition

$$||u^{(h)}||_{U_h} \leq C||f^{(h)}||.$$

is satisfied.

In solving the Cauchy problem the net function, $u^{(h)}$, is ordinarily computed in moving, sequentially, from one point of the difference net to another, neighboring, point. If we can get a bound on the growth of the solution, $u^{(h)} \equiv \{u_n^{(h)}\}$, after each such move (or, as it is commonly called, each "step" of the computational process), we will have at our disposal one of the most widely used methods for the study of stability. It is this method which we will develop here.

1. Introductory example. We begin with the simplest, and by now thoroughly familiar, difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1} - u_n}{h} + Au_n = \phi_n, \\ n = 0, 1, \ldots, N-1, \quad (h = 1/N), \\ u_0 = a \end{cases} \tag{4}$$

for the solution of problem (1). This scheme may be written in the recursive form

$$\left.\begin{aligned} u_{n+1} &= (1 - Ah)u_n + h\phi_n, \qquad n = 0, 1, \ldots, N-1, \\ u_0 &= a, \end{aligned}\right\} \tag{5}$$

from which it follows that

$$\left.\begin{aligned} u_1 &= (1 - Ah)u_0 + h\phi_0, \\ u_2 &= (1 - Ah)^2 u_0 + h[(1 - Ah)\phi_0 + \phi_1], \\ u_3 &= (1 - Ah)^3 u_0 + h[(1 - Ah)^2\phi_0 + (1 - Ah)\phi_1 + \phi_2], \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ u_n &= (1 - Ah)^n u_0 + h[(1 - Ah)^{n-1}\phi_0 + (1 - Ah)^{n-2}\phi_1 + \ldots + \phi_{n-1}]. \end{aligned}\right\} \tag{6}$$

We will define norms in the spaces U_h and F_h, respectively via the equations

$$||u^{(h)}||_{U_h} = \max_{0 \le n \le N} \left| u_n^{(h)} \right| , \tag{7}$$

$$||f^{(h)}||_{F_h} = \left|\left| \begin{matrix} \phi_m \\ a \end{matrix} \right|\right|_{F_h} = \max[|a|, \max_{0 \le m < N} |\phi_m|]. \tag{8}$$

Now we use the fact that the expression $(1 - Ah)^n$ is bounded for $n \le N = 1/h$,

$$|(1 - Ah)^n| < C_1. \tag{9}$$

From Eq. (6) for u_n, with the aid of inequality (9), we conclude that

$$|u_n| \le C_1|u_0| + hNC_1 \max_m |\phi_m| =$$

$$= C_1|a| + C_1 \max_m |\phi_m| \le 2C_1||f^{(n)}||_{F_h}. \tag{10}$$

Since n is arbitrary, n = 0, 1, ..., N, it follows from (10) that

$$||u^{(h)}||_{U_h} \le 2C_1||f^{(h)}||_{F_h} , \tag{11}$$

and stability is proven.

2. Canonical form of a difference scheme. At this point we introduce new notation, setting

$$u_n = y_n, \qquad R_h = (1 - Ah), \qquad \rho_n = \phi_n. \tag{12}$$

Now inequality (5) may be rewritten in the form

$$\left.\begin{aligned} y_{n+1} &= R_h y_n + h\rho_n, \\ y_0 &\text{ given.} \end{aligned}\right\} \tag{13}$$

Using notation (12) we repeat all the above calculations. Equations (6) now take the form

$$\left.\begin{aligned} y_1 &= R_h y_0 + h\rho_0, \\ y_2 &= R_h^2 y_0 + h[R_h\rho_0 + \rho_1], \\ y_3 &= R_h^3 y_0 + h[R_h^2\rho_0 + R_h\rho_1 + \rho_2], \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n &= R_h^n y_0 + h[R_h^{n-1}\rho_0 + R_h^{n-2}\rho_1 + \ldots + \rho_{n-1}]. \end{aligned}\right\} \tag{6'}$$

Hence

$$\max_n |y_n| \leq \max_n \left|R_h^n\right| \cdot [|y_0| + hN \max_n |\rho_n|].$$

The norms $||\cdot||_{U_h}$ and $||\cdot||_{F_h}$ are given, now, by the equations

$$||u^{(h)}||_{U_h} = \max_n |y_n|, \tag{7'}$$

$$||f^{(h)}||_{F_h} = \max [|y_0|, \max_n |\rho_n|]. \tag{8'}$$

Thus, noting that $Nh = 1$, one may write

$$||u^{(h)}||_{U_h} \leq \max_n \left|R_h^n\right| \cdot 2\, ||f^{(h)}||_{F_h}.$$

The proof of stability will be complete if one establishes the boundedness, uniform in h, of the totality of numbers $\left|R_h^n\right|$, i.e. if one proves that

$$\left|R_h^n\right| \leq C, \qquad n = 1, 2, \ldots, N. \tag{9'}$$

But

$$\left|R_h^n\right| \leq (1 - Ah)^n \leq (1 + |A|\, h)^N \leq e^{|A|} = C,$$

which completes the proof.

Writing the difference scheme in form (13) made it possible to reduce the stability proof to the computation of a bound for $\left|R_h^n\right|$. This is convenient. Indeed we will put all other difference schemes for the solution of initial-value problems into the canonical form (13), taking for y_n, ρ_n and R_n the different expression appropriate to each problem.

For example let us write in form (13) the difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{v_{n+1} - v_n}{h} + Av_n + Bw_n = \rho_n, & n = 0, 1, \ldots, N-1, \\ \dfrac{w_{n+1} - w_n}{h} + Cv_n + Dw_n = q_n, & n = 0, 1, \ldots, N-1, \\ v_0 = a, \\ w_0 = b, \end{cases} \tag{14}$$

approximating Cauchy problem (2) for the given set of differential equations. Here

$$f^{(h)} = \begin{cases} p_n, & n = 0, 1, \ldots, N-1, \\ q_n, & n = 0, 1, \ldots, N-1, \\ a \\ b \end{cases}$$

We will write the difference scheme (14) in the form

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{\begin{bmatrix} v_{n+1} \\ w_{n+1} \end{bmatrix} - \begin{bmatrix} v_n \\ w_n \end{bmatrix}}{h} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{bmatrix} v_n \\ w_n \end{bmatrix} = \begin{bmatrix} p_n \\ q_n \end{bmatrix}, \quad n = 0, 1, \ldots, N-1, \\ \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}, \end{cases}$$

where

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

is a 2x2 matrix. We now cast this vector difference equation into the form of a recursion relation

$$\left.\begin{aligned} \begin{bmatrix} v_{n+1} \\ w_{n+1} \end{bmatrix} &= \begin{pmatrix} 1 - Ah & -Bh \\ -Ch & 1 - Dh \end{pmatrix} \begin{bmatrix} v_n \\ w_n \end{bmatrix} + h \begin{bmatrix} p_n \\ q_n \end{bmatrix}, \\ \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} &= \begin{bmatrix} a \\ b \end{bmatrix}. \end{aligned}\right\}$$

If we define

$$y_n = u_n = \begin{bmatrix} v_n \\ w_n \end{bmatrix}, \qquad R_h = \begin{pmatrix} 1 - Ah & -Bh \\ -Ch & 1 - Dh \end{pmatrix}, \qquad \rho_n = \begin{bmatrix} p_n \\ q_n \end{bmatrix},$$

then the above recursion relation takes on the required form (13).

3. Stability viewed as the boundedness of the norms of powers of the transition operator. We first make a remark which is equally applicable to all equations of form (13), regardless of the dimensionality of the linear space, Y, which contains the vectors y_n and ρ_n, and of the form of the linear operator R_h: Eq. (6′) follows from (13).

If, in the space Y containing ρ_n and y_n, one introduces some norm $\|\cdot\|_Y$, then Eq. (6′) implies the bound

$$\|y_n\|_Y \le \left\|R_h^n\right\|_Y \cdot \|y_0\|_Y + h\left[\left\|R_h^{n-1}\right\|_Y \cdot \|\rho_0\|_Y + \ldots + \|\rho_{n-1}\|_Y\right]. \tag{15}$$

* * * * * *

We recall that the *norm*, $\|T\|$, of the linear operator T, mapping some linear normed space Y into itself, is defined by the relations

$$\|T\| = \sup_{x \text{ in } Y} \frac{\|Tx\|}{\|x\|} \quad \text{or} \quad \|T\| = \sup_{\|x\|=1,\ x \text{ in } Y} \|Tx\|.$$

From these relations, and from the properties of the norm of a vector, it follows that

$$\|Tx\| \le \|T\| \cdot \|x\|,$$

$$\|\lambda T\| = |\lambda| \cdot \|T\|, \quad \text{where } \lambda \text{ is any arbitrary number,}$$

$$\|T^m\| \le \|T\|^m.$$

The first two of these equations have been used to get bound (15).

* * *

From (15), clearly, it follows that

$$\max_{0\le n\le N} \|y_n\|_Y \le \max_{0\le n\le N} \left\|R_h^n\right\|_Y \left[\|y_0\|_Y + Nh \max_{0\le n\le N} \|\rho_n\|_Y\right]. \tag{16}$$

Suppose that the difference scheme $L_h u^{(h)} = f^{(h)}$ *has been cast into the canonical form (13), and assume that the norms introduced in the spaces* U_h, F_h *and* Y *are chosen such as to satisfy the inequalities*

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &\le C_2 \max_{0\le n\le N} \|y_n\|_Y, \\ \|y_0\|_Y &\le C_2 \|f^{(h)}\|_{F_h}, \\ \|\rho_n\|_Y &\le C_2 \|f^{(h)}\|_{F_h}. \end{aligned}\right\} \tag{17}$$

Then for stability

$$\|u^{(h)}\|_{U_h} \leq C \|f^{(h)}\|_{F_h} \tag{18}$$

it is sufficient that the norms, $\left\|R_h^m\right\|_Y$, *of the powers of the operator* R_h, *be uniformly bounded with respect to h, i.e. that*

$$\left\|R_h^n\right\|_Y \leq C_3, \qquad n = 1, 2, \ldots, N.$$

Moreover, as the constant C entering into the definition of stability, Eq. (18), one can take the quantity

$$C = 2C_2^2C_3.$$

The proof of this assertion consists of the following chain of obvious inequalities, written so as to take into account conditions (17) and (18), as well as the fact that Nh = 1:

$$\|u^{(h)}\|_{U_h} \leq C_2 \max_n \|y_n\|_Y \leq C_2 \max_n \left\|R_h^n\right\| [C_2 + C_2] \|f^{(h)}\|_{F_h} \leq$$

$$\leq C_2C_3[C_2 + C_2] \|f^{(h)}\|_{F_h},$$

or

$$\|u^{(h)}\|_{U_h} \leq C \|f^{(h)}\|_{F_h}.$$

4. Examples of investigations of stability.

Example 1. We turn now to an analysis of the stability of difference scheme (14) for a system of differential equations. Norms in U_h and F_h will be introduced via the equations

$$\|u^{(h)}\|_{U_h} = \left\|\begin{Bmatrix} v_n \\ w_n \end{Bmatrix}\right\|_{U_h} = \max[\max_n |v_n|, \max_n |w_n|],$$

$$\|f^{(h)}\|_{F_h} = \left\|\begin{Bmatrix} p_n \\ q_n \\ a \\ b \end{Bmatrix}\right\|_{F_h} = \max[|a|, |b|, \max_n |p_n|, \max_n |q_n|].$$

As we have seen, after introduction of the notation

$$y_n = \begin{bmatrix} v_n \\ w_n \end{bmatrix}, \qquad R_h = \begin{pmatrix} 1 - Ah & -Bh \\ -Ch & 1 - Dh \end{pmatrix}, \qquad \rho_n = \begin{bmatrix} p_n \\ q_n \end{bmatrix}, \qquad y_0 = \begin{bmatrix} a \\ b \end{bmatrix}$$

this system of difference equations takes on the canonical form (13).

We introduce a norm, in the two-dimensional space, Y, to which y_n and ρ_n belong, setting

$$||y_n||_Y = \left|\left|\left| \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \right|\right|\right|_Y = \max \left[\left|y_n^{(1)}\right|, \left|y_n^{(2)}\right| \right].$$

The norms in U_h, F_h and Y turn out to satisfy condition (17). Therefore to verify stability it is sufficient to show that

$$\left|\left|\left| R_h^n \right|\right|\right|_Y \leq M, \qquad n = 1, 2, \ldots, N, \qquad M = \text{const}.$$

Note that, for the vector norms we have chosen in Y, the norm of any linear operator

$$T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$$

is given by the equation

$$||T|| = \max \left[|t_{11}| + |t_{12}|, \; |t_{21}| + |t_{22}| \right], \tag{19}$$

since

$$\max_{||x||=1} ||Tx||_Y = |||T|||_y$$

is attained for at least one of the two vectors

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{or} \qquad x = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

By virtue of Eq. (19) for $||T||$ we get

$$||R_h||_Y = \left|\left|\begin{pmatrix} 1 - Ah & -Bh \\ -Ch & 1 - Dh \end{pmatrix}\right|\right|_Y \leq$$

$$\leq \max\{|1 - Ah| + |Bh|,\ |1 - Dh| + |Ch|\} = 1 + C'h.$$

Consequently,

$$\left|\left|R_h^n\right|\right|_Y \leq ||R_h||_Y^n \leq (1 + C'h)^N \leq e^{C'} = M, \qquad n = 1, 2, \ldots, N,$$

and stability is demonstrated.

Example 2. Consider the scheme

$$\left.\begin{aligned} \frac{u_{n+1} - u_{n-1}}{2h} + Au_n &= \phi_n, \qquad n = 1, 2, \ldots, N-1, \\ u_0 &= \alpha, \\ u_1 &= \beta, \end{aligned}\right\} \tag{20}$$

which, for $\alpha = a$, $\beta = (1 - Ah)a + h\phi_0$, approximates the Cauchy problem (1) to second order in h. We introduce the norms $||\cdot||_{U_h}$ and $||\cdot||_{F_h}$ via the equations

$$||u^{(h)}||_{U_h} = \max_n |u_n|,$$

$$||f^{(h)}||_{F_h} = \left|\left|\begin{bmatrix} \phi_n \\ \alpha \\ \beta \end{bmatrix}\right|\right|_{F_h} = \max\{|\alpha|,\ |\beta|,\ \max_n |\phi_n|\}.$$

In order to study stability we will try to put the difference scheme into form (13) so as to reduce the stability proof to the derivation of a bound $||R_h^n||_Y \leq C$. Let us first rewrite difference equation (20) in the form

$$u_{n+1} = u_{n-1} - 2Ahu_n + 2h\phi_n.$$

What prevents us from rewriting it in form (13) is the fact that it connects not two, but three successive values: u_{n-1}, u_n, u_{n+1}. In order to overcome this difficulty we will set

$$y_n = \begin{bmatrix} u_n + 1 \\ u_n \end{bmatrix}.$$

Now the pair of equations

$$\left.\begin{aligned} u_{n+1} &= u_{n-1} - 2Ahu_n + 2h\phi_n, \\ u_n &= u_n \end{aligned}\right\} \tag{21}$$

gives the components of vector y_n in terms of the components of vector y_{n-1}:

$$\begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix} = \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix} \begin{bmatrix} u_n \\ u_{n-1} \end{bmatrix} + h \begin{bmatrix} 2\phi_n \\ 0 \end{bmatrix}.$$

We have now written (20) in form (13), where

$$R_h = \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix}, \qquad \rho_n = \begin{bmatrix} 2\phi_n \\ 0 \end{bmatrix},$$

$$y_0 = \begin{bmatrix} (1 - Ah)a + h\phi_0 \\ a \end{bmatrix}.$$

Let us introduce a norm in the two-dimensional space Y, to which y_n and ρ_n belong, via the equation

$$\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_Y = \max(|\alpha|, |\beta|).$$

Then the norms

$$\|u^{(h)}\|_{U_h}, \qquad \|f^{(h)}\|_{F_h}, \qquad \|\rho_n\|_Y, \qquad \|y_0\|_Y,$$

as can easily be seen, satisfy condition (17). Therefore the bound

$$\left\|R_h^n\right\|_Y \leq \left\|R_h\right\|_Y^n = \left\| \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix} \right\|_Y^n \leq$$

$$\leq (1 + 2|Ah|)^n \leq e^{2|A|}, \qquad n = 1, 2, \ldots, N,$$

proves stability.

Example 3. Let us study the stability of the difference scheme

$$\left.\begin{aligned} &\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + A\,\frac{u_{n+1} - u_{n-1}}{2h} + Bu_n = \phi_n, \\ &\qquad n = 1, 2, \ldots, N - 1. \\ &\qquad u_0 = a, \\ &\qquad \frac{u_1 - u_0}{h} = b, \end{aligned}\right\} \tag{22}$$

which, for a natural choice of norms, approximates the Cauchy problem (3). The norms $\|u^{(h)}\|_{U_h}$ and $\|f^{(h)}\|_{F_h}$ will be defined by the equations

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &= \max_n |u_n|, \\ \|f^{(h)}\|_{F_h} &= \left\| \begin{matrix} \phi_n \\ a \\ b \end{matrix} \right\|_{F_h} = \max\left[|a|, |b|, \max_n |\phi_n|\right]. \end{aligned}\right\} \tag{23}$$

So as to bring the scheme in question into the canonical form (13) we set, as in example 2,

$$y_n = \begin{bmatrix} u_n + 1 \\ u_n \end{bmatrix}. \tag{24}$$

Then the components of the vector

$$y_n = \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix}$$

are uniquely determined by the components of y_{n-1}, by virtue of the given difference scheme, through the relations

$$\left.\begin{aligned} y_{n+1}^{(1)} &= \frac{2}{2 + Ah}\left[(2 - Bh)^2 y_n^{(1)} - \frac{2 - Ah}{2}\, y_n^{(2)} + h^2\phi_{n+1}\right], \\ y_{n+1}^{(2)} &= y_n^{(1)}. \end{aligned}\right\} \tag{25}$$

Thus

$$y_{n+1} = R_h y_n + h\rho_n, \qquad n = 0, 1, \ldots, N-2,$$

where

$$R_h = \begin{pmatrix} \frac{4 - 2Bh^2}{2 + Ah} & -\frac{2 - Ah}{2 + Ah} \\ 1 & 0 \end{pmatrix}, \qquad \rho_n = \begin{bmatrix} \frac{2h^2}{2 + Ah}\,\phi_{n+1} \\ 0 \end{bmatrix}. \tag{26}$$

Through use of the condition $u_0 = a$, $(u_1 - u_0)/h = b$ (see (22)) we calculate the vector y_0:

$$y_0 = \begin{bmatrix} a + bh \\ a \end{bmatrix}, \tag{27}$$

and thus complete reduction of the given difference scheme to form (13).

It is easy to see that, if the norm of the vector $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ is defined as $\max(|\alpha|, |\beta|)$, it will not be such a simple matter to prove stability with this operator R_h, since $\|R_h\| \approx 2$, and $\|R_h\|^n \to \infty$. For this reason the norm in space Y will not be defined as in example 2. In fact we will take

$$\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y_h} = \max \left[|\alpha|, \left| \frac{\beta - \alpha}{h} \right| \right].$$

We have attached the subscript "h" to Y so as to stress that the norm now depends on h. For this choice of norm the quantities $\|u^{(h)}\|_{U_h}$, $\|f^{(h)}\|_{F_h}$, $\|\rho_n\|_{Y_h}$ and $\|y_0\|_{L_h}$ satisfy relations (17). It remains to show that the conditions $\|R_h^n\|_{Y_h} \leq C$, $n = 1, 2, N$, are satisfied. We are already familiar with Eq. (19), relating the norm of an operator to the elements of its matrix if the norm in Y is given by

$$\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y} = \max \left[|\alpha|, |\beta| \right].$$

Let us now reduce the computation of the norm in Y_h to the computation in Y:

$$\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y_h} = \left\| \begin{pmatrix} 1 & 0 \\ 1/h & -1/h \end{pmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y} = \left\| S \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y},$$

where

$$S = \begin{pmatrix} 1 & 0 \\ 1/h & -1/h \end{pmatrix}.$$

Next we show that, for any arbitrary linear transformation, T, acting in space Y, we have the equality $\|T\|_{Y_h} = \|STS^{-1}\|_{Y}$. In fact,

$$\left\| T \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y_h} = \left\| ST \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y} = \left\| STS^{-1} S \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y}.$$

Further

$$||T||_{Y_h} = \max_x \frac{||Tx||_{Y_h}}{||x||_{Y_h}} = \max_{x \text{ in } y} \frac{||STS^{-1}Sx||_Y}{||Sx||_Y} =$$

$$= \max_{v \text{ in } Y} \frac{||STS^{-1}v||_Y}{||v||_Y} = ||STS^{-1}||_Y.$$

Now we note that

$$\left|\left|R_h^n\right|\right|_{Y_h} = \left|\left|SR_h^nS^{-1}\right|\right|_Y = \left|\left|\left(SR_hS^{-1}\right)^n\right|\right|_Y \leq \left|\left|SR_hS^{-1}\right|\right|_Y^n .$$

Since

$$S^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & -h \end{pmatrix},$$

then

$$SR_hS^{-1} = \begin{pmatrix} 1 - \frac{2B}{2 + Ah}\,h & \frac{2 - Ah}{2 + Ah}\,h \\ -\frac{2B}{2 + Ah}\,h & 1 - \frac{2A}{2 + Ah}\,h \end{pmatrix}.$$

Therefore

$$\left|\left|SR_hS^{-1}\right|\right|_Y \leq 1 + Ch,$$

where C is some constant, independent of h, chosen to satisfy the condition

$$1 + Ch \geq \max\left[\left|1 - \frac{2B}{2 + Ah}\,h\right| + \left|\frac{2 - Ah}{2 + Ah}\,h\right|,\right.$$

$$\left.\left|\frac{2B}{2 + Ah}\,h\right| + \left|1 - \frac{2A}{2 + Ah}\,h\right|\right].$$

In particular, for small enough h this condition is obviously satisfied by the quantity $C = 1 + 2|A| + 2|B|$.

Thus

$$\left|\left|R_h^n\right|\right|_{Y_h} \leq \left|\left|SR_hS^{-1}\right|\right|_Y^n \leq (1 + Ch)^N \leq e^C, \qquad n = 1, 2, \ldots, N,$$

which guarantees stability of the given difference scheme.

* * * * * *

5. Non-uniqueness of the canonical form.

The reduction of a difference scheme to the canonical form (13) can be accomplished in many ways. Setting $y_n' = Ty_n$ where T is an arbitrary linear transformation in the space, Y, to which y_n and ρ_n belong, we go over to the new notation

$$\left.\begin{aligned} y_{n+1}' &= R_h' y_n' + h\rho_n'. \\ y_0' &\text{ given.} \end{aligned}\right\} \qquad (13')$$

Here $R_h' = TR_hT^{-1}$, $\rho_n' = T\rho_n$, $y_0' = Ty_0$.

If, in example 3, instead of taking $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$, we had defined

$$y_n = \begin{bmatrix} u_n \\ \dfrac{u_{n+1} - u_n}{h} \end{bmatrix},$$

we would have arrived at a version of the difference scheme in form (13) with

$$R_h = \begin{pmatrix} 1 & h \\ -\dfrac{2hB}{2 + hA} & \dfrac{2 - hA - 2h^2B}{2 + hA} \end{pmatrix}, \qquad \rho_n = \begin{bmatrix} 0 \\ \dfrac{2}{2 + hA}\,\phi_{n+1} \end{bmatrix}.$$

For the choice of norm in Y given by the equation

$$\left|\left|\begin{bmatrix} \alpha \\ \beta \end{bmatrix}\right|\right|_Y = \max\,(|\alpha|,\ |\beta|)$$

Eq. (17) would have been satisfied. The boundedness of $\left|\left|R_h^n\right|\right|_Y$ is obvious:

$$\left|\left|R_h^n\right|\right|_Y \le ||R_h||_Y^n \le (1 + Ch)^N < e^C,$$

where C is chosen from the condition

$$1 + Ch \ge \max\left(1 + h,\ \left|\frac{2hB}{2 + hA}\right| + \left|\frac{2 - hA - 2h^2B}{2 + hA}\right|\right) =$$

$$= \max\left(1 + h,\ 1 + \frac{2(|A| + |B|h)}{2 - |A|h}\,h\right).$$

There is also some freedom in the choice of the dimensionality of space Y. It would have been possible, in place of $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$, to take, let us say,

$$y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix},$$

but in the given example this would not have simplified the study of stability.

* * *

Let us now summarize the above considerations. From the examples we have considered one sees that, to investigate stability of the difference scheme $L_h u^{(h)} = f^{(h)}$ for the solution of the Cauchy problem with constant coefficients, it is convenient to put this scheme into form (13);

$$\left.\begin{aligned} y_{n+1} &= R_h y_n + h\rho_n, \qquad n = 0, 1, \ldots, \\ y_0 &\text{ given.} \end{aligned}\right\}$$

If, in the space to which y_n and ρ_n *belong, one has introduced a norm such that the conditions*

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &\le C_2 \max_n \|y_n\|, \\ \|\rho_n\| &\le C_2 \|f^{(h)}\|_{F_h}, \\ \|y_0\| &\le C_2 \|f^{(h)}\|_{F_h}, \end{aligned}\right\} \tag{28}$$

are satisfied, then it is sufficient for stability that the norms of the powers of the operator R_h *be bounded uniformly in h,*

$$\left\|R_h^n\right\| \le C_3, \qquad n = 1, 2, \ldots, N.$$

For this to be true it is sufficient, clearly, that the inequality

$$\|R_h\| < 1 + C'h,$$

be satisfied with C' independent of h. In this case the constant C in the definition of stability

$$\|u^{(h)}\|_{U_h} \le C\|f^{(h)}\|_{F_h}$$

can be taken in the form

$$C = 2C_2^2 C_3. \tag{29}$$

PROBLEMS

1. Prove the stability of the following difference schemes for the solution of the problem $u' + Au = \phi(x)$, $u(0) = a$. Find the constant, C, in the definition $\|u^{(h)}\|_{U_h} \le C\|f^{(h)}\|_{F_h}$ of stability.

a) $$\left.\begin{aligned} &\frac{u_{n+1} - u_n}{h} + A(nh)\, u_n = \phi_n, \qquad n = 0, 1, 2, \ldots, N-1, \\ &u_0 = a, \end{aligned}\right\}$$

if $|A(x)| \le M = \text{const}$, and norms are introduced via the equations

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|, \qquad \|f^{(h)}\|_{F_h} = \max\Big[|a|, \max_n |\phi_n|\Big].$$

b) $$\left.\begin{aligned} &\frac{u_{n+1} - u_n}{h} + Au_{n+1} = \phi_n, \\ &u_0 = a. \end{aligned}\right\}$$ Norms — as in a.

c) $$\left.\begin{aligned} &\frac{u_{n+1} - u_n}{h} + A\,\frac{u_n + u_{n+1}}{2} = \phi\Big[\Big(n + \tfrac{1}{2}\Big)h\Big], \qquad n = 0, \ldots, N-1, \\ &u_0 = a, \end{aligned}\right\}$$

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|, \qquad \|f^{(h)}\|_{F_h} = \max\Big(|a|, \max_n \Big|\phi\Big[\Big(n + \tfrac{1}{2}\Big)h\Big]\Big|\Big).$$

2. Solve problem 1 under the assumption that

$u_n = \begin{bmatrix} u_n^{(1)} \\ u_n^{(2)} \end{bmatrix}$ is a vector;

$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ is a matrix;

$a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ and $\phi_n = \begin{bmatrix} \phi_n^{(1)} \\ \phi_n^{(2)} \end{bmatrix}$ are vectors.

Norms are given in the form

$$\|u^{(h)}\|_{U_h} = \max_n\left(\left|u_n^{(1)}\right| + \left|u_n^{(2)}\right|\right),$$

$$\|f^{(h)}\|_{F_h} = \max\left[|a_1| + |a_2|, \max_n\left(\left|\phi_n^{(1)}\right| + \left|\phi_n^{(2)}\right|\right)\right], \quad |A_{ij}(x)| \le M.$$

3. Bring into the canonical form $y_{n+1} = R_h y_n + h\rho_n$, y_0 given, the difference equation

$$\frac{u_{n+2} - 2u_{n+1} + 3u_n - 4u_{n-1}}{h} - 5u_n = \phi_n, \qquad n = 1, 2, \ldots,$$

setting

$$y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix}.$$

§15. Necessary spectral criterion for stability

In §14 we showed that the reduction of a difference scheme for the solution of the Cauchy problem with constant coefficients

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

to the form

$$\left.\begin{aligned} y_{n+1} &= R_h y_n + h\rho_n, \qquad n = 0, 1, \ldots, \\ y_0 &\text{ given} \end{aligned}\right\} \tag{2}$$

can be used to prove stability: under certain conditions (conditions (17) §14) the bound

$$\left\|\left|R_h^n\right|\right\|_Y \le C, \qquad n = 1, 2, \ldots, N, \tag{3}$$

is sufficient for stability.

Here we will show that this bound (3), under certain natural conditions, is necessary for stability. We will also show that, regardless of the choice of norm, for the validity of bound (3) it is necessary that the spectrum of the matrix R_h, i.e. the set of all roots of the equations

$$\det(R_h - \lambda E) = 0, \tag{4}$$

lie in the circle

$$|\lambda| < 1 + Ch, \tag{5}$$

where C does not depend on h.

We proceed now to carry out the indicated program.

1. Boundedness of the norms of the powers of the transition operator necessary for stability. The methods we have described for reduction of difference equations to the canonical form (2) are such, that if the right-hand sides of the difference equations vanish, then ρ_n is also identically equal to zero.

Suppose the constants $M_1 = M_1(h) > 0$ and $M_2 = M_2(h) > 0$ are so chosen that

$$||u^{(h)}||_{U_h} \geq M_1 \max_n ||y_n||, \tag{6}$$

and, under the condition that $\rho_n \equiv 0$,

$$||y_0|| \geq M_2 ||f^{(h)}||_{F_h}. \tag{7}$$

Then, for vanishing right-hand sides of the difference equation (or system of difference equations) Eq. (2) takes the form

$$y_{n+1} = R_h y_n,$$

and therefore

$$y_n = R_h^n y_0. \tag{8}$$

Further, by virtue of (6) and (8)

$$||u^{(h)}||_{U_h} \geq M_1 \max_n \left|\left|R_h^n y_0\right|\right|. \tag{9}$$

From the definition of the norm of a linear operator it follows that, in a finite-dimensional space, one can always choose a vector, y_0, so that for given n $\left|\left|R_h^n y_0\right|\right| = \left|\left|R_h^n\right|\right| \cdot \left|\left|y_0\right|\right|$. Therefore for some y_0 (depending on h),

$$\max_n \left|\left|R_h^n y_0\right|\right| = \max_n \left|\left|R_h^n\right|\right|_Y \cdot \left|\left|y_0\right|\right|.$$

For this choice of y_0 by virtue of (9) and (10) we get

$$||u^{(h)}||_{U_h} \geq M_1 \max_n \left|\left|R_h^n\right|\right| \cdot ||y_0|| \geq M_1 M_2 \max_n \left|\left|R_h^n\right|\right| \cdot ||f^{(h)}||_{F_h}. \tag{6'}$$

From the latter bound it follows that, if difference scheme (1) is stable, the constant C′ in the definition of stability

$$||u^{(h)}||_{U_h} < C'||f^{(h)}||_{F_h}$$

certainly must satisfy the bound

$$C' \geq M_1 M_2 \max_n \left|\left|R_h^n\right|\right|. \qquad (6'')$$

Hence it is clear that, *if the norms* $||u^{(h)}||_{U_h}$, $||f^{(h)}||_{F_h}$ *and* $||y_n||$ *are so coordinated that conditions (6) and (7) are satisfied, then condition (3) is necessary for stability.* Condition (3) is equivalent to the statement that the solution, $\{y_n\}$, of the homogeneous equation $y_{n+1} = R_h y_n$ satisfies, for any y_0, the inequality

$$||y_n|| \leq C\,||y_0||, \qquad n = 1, 2, \ldots, N. \qquad (11)$$

In examples 1 and 2 of §14 it was possible to take the numbers M_1 and M_2 independent of h (in fact equal to 1), as the reader can easily verify. This fact indicates the naturalness of the formulation chosen there.

In example 3 of §14, for difference scheme (22), using Eq. (24) and norms (23), the condition $||u^{(h)}||_{U_h} \geq M_1 \max_n ||y_n||$ can be satisfied only if $M_1 \leq h/2$. But if we change the choice of norm $||u^{(h)}||_{U_h}$, defining

$$||u^{(h)}||_{U_h} = \max\left[\max_n |u_n|, \quad \max_n \left|\frac{u_{n+1} - u_n}{h}\right|\right], \qquad (12)$$

then we can set $M_1 = 1$ and $M_2 = 1$, and bound (3) is necessary for stability. With this change in norm condition (17) of §14, under which bound (3) suffices for stability, is still satisfied.

2. Spectral criterion for stability. To bound $\left|\left|R_h^n\right|\right|$ it is possible to use the eigenvalues of the matrix R_h, i.e. the roots, λ, of the equation

$$\det||R_h - \lambda E|| = 0.$$

If λ is an eigenvalue, then there exists an eigenvector, y, such that $R_h y = \lambda y$. Therefore

$$R_h^n y = \lambda^n y, \qquad \left|\left|R_h^n y\right|\right| = |\lambda|^n ||y||, \qquad \left|\left|R_h^n\right|\right| \geq |\lambda|^n.$$

Thus if $\left|\left|R_h^n\right|\right|$ is to be bounded it is necessary that the powers of the eigenvalues, $|\lambda|^n$, n = 1, 2, ..., N, should be bounded. In turn if this is to be true all the eigenvalues must lie in the circle

$$|\lambda| \leq 1 + ch \qquad (13)$$

in the complex plane, where c does not depend on h. In the contrary case, for an arbitrary c and some sufficiently small h

$$\left|\left|R_h^N\right|\right| \geq |\lambda|^N > (1 + ch)^{1/h} = e^{(1/h)\ln(1+ch)} \geq e^{c(1-(ch/2))} > e^{c/2}.$$

The above criterion for the boundedness of the norms of powers, $\left|\left|R_h^n\right|\right|$, in terms of the location of the spectrum (i.e. the totality of eigenvalues) of the operator R_h, clearly does not depend on the choice of norm in the space on which R_h operates.

The spectral stability criterion (13) also does not depend on the means by which scheme (1) is put into form (2). If this reduction is performed differently, i.e., $y'_{n+1} = R'_h y'_n + h\rho'_n$ with $y' = Ty$, $R'_h = TR_hT^{-1}$, where T is an arbitrary nonsingular linear operator, then the spectra of R_h and R'_h will coincide. In fact

$$\det(R'_h - \lambda E) = \det(TR_hT^{-1} - \lambda E) = \det[T(R_h - \lambda E)T^{-1}] =$$

$$= \det T \det(R_h - \lambda E)\det T^{-1} = \det(R_h - \lambda E).$$

Therefore the equations $\det(R_h - \lambda E) = 0$ and $\det(R'_h - \lambda E) = 0$ have the same roots λ.

* * * * * *

3. Discussion of the spectral stability criterion. Above it was shown that, if norms are chosen in accordance with conditions (6) and (7), the location of the spectrum of the operator R_h in the circle

$$|\lambda| \leq 1 + ch, \tag{13}$$

is necessary for the boundedness of $\left|\left|R_h^n\right|\right|$ and, moreover, also necessary for stability.

Suppose condition (13) is grossly violated so that for a sufficiently small $h > 0$ there is an eigenvalue, λ, substantially greater than 1 in modulus, let us say

$$|\lambda| > 1 + h^{1-\varepsilon},$$

where $\varepsilon > 0$ does not depend on x. Then the difference scheme (1) is unstable for any reasonable choice of norm $||u^{(h)}||_{U_h}$ and $||f^{(h)}||_{F_h}$, even if one doesn't limit the freedom of choice of these norms via conditions (6) and (7).

This assertion cannot be called a theorem, if for no other reason than the fact that it is based on the term "reasonable", which has not been precisely defined. But we will now explain what we mean.

For any reasonable choice of norm, $||u^{(h)}||_{U_h}$, one can choose a positive k_1 such that for all sufficiently small h

$$||u^{(h)}||_{U_h} \geq h^{k_1} \max_n |u_n|. \tag{14}$$

In the contrary case it would not be possible to satisfy Eq. (4) §13:

$$\lim_{h\to 0} \left|\left|[u]_h\right|\right|_{U_h} = ||u||_U.$$

Further, for any reasonable choice of norm $||f^{(h)}||_{F_h}$, one can so choose a $k_2 > 0$ that, for all sufficiently small h,

$$||f^{(h)}||_{F_h} < h^{-k_2} F, \tag{15}$$

where F denotes the maximum modulus of the components of element $f^{(h)}$ of space F_h. In the contrary case difference scheme (1) cannot approximate the problem Lu = f on the solution u: indeed we have seen that the components of the residual $\delta f^{(h)}$, which develops when $[u]_h$ is substituted into the left-hand side of the approximating difference scheme (1), tend to zero no faster than some power of the step-width, h.

We now bring difference scheme (1) into form (2), defining for this purpose

$$y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}, \qquad \left|\left| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right|\right|_Y = \max [|\alpha|, |\beta|]. \tag{16}$$

For the sake of definiteness we assume that the difference scheme under consideration connects three consecutive points, u_{n-1}, u_n and u_{n+1}.

If the right-hand side of the difference equation, on which scheme (1) is based, is taken to be equal to zero, then for some $r > 0$ we will have

$$\max [|u_0|, |u_1|] = ||y_o||_Y \geq h^r F, \tag{17}$$

since the relation connecting u_1 and u_0, and entering into the difference scheme, has the form

$$\left.\begin{aligned} u_0 &= a, \\ u_1 &= b, \end{aligned}\right\} \qquad \text{or} \qquad \left.\begin{aligned} u_0 &= a, \\ \frac{u_1 - u_0}{h} &= b, \end{aligned}\right\}$$

or something similar.

It is now clear that we can always make the inequalities (6) and (7) valid by setting $M_1(h) = h^{k_1}$, $M_2(h) = h^{r+k_2}$. In fact (see also (14) and (17))

$$||u^h||_{U_h} \geq h^{k_1} \max_n |u_n| = h^{k_1} \max_n ||y_n||.$$

$$||y_0|| \geq h^r F = h^r h^{k_2}(h^{-k_2} F) \geq h^{r+k_2} ||f^{(h)}||_{F_h}.$$

Thus inequality (6′) takes the form

$$||u^{(h)}||_{U_h} \geq h^{r+k_1+k_2} \max_n \left|\left|R_h^n\right|\right| \cdot ||f^{(h)}||_{F_h} \geq$$

$$\geq h^{r+k_1+k_2} \left(1 + h^{1-\varepsilon}\right)^{1/h} ||f^{(h)}||_{F_h}.$$

This implies instability since, for any r, k_1, k_2 and $\varepsilon > 0$, as one can easily see,

$$h^{r+k_1+k_2}\left(1 + h^{1-\varepsilon}\right)^{1/h} \to \infty \quad \text{as} \quad h \to 0.$$

With this we conclude our exposition of arguments showing that if, among the eigenvalues of the matrix R_h there are roots obeying the inequality $|\lambda| > 1 + h^{1-\varepsilon}$, then the difference scheme is unstable for any reasonable choice of norms.

* * *

Let us now use the necessary spectral criterion for stability (13) to show that the scheme considered in §9 is really unstable. In §9 a rigorous investigation of instability could not be carried out, if for no other reason than the fact that there we still did not have at our disposal any precise definitions.

The difference scheme under consideration approximates the problem

$$\left.\begin{aligned} u' + Au = 0, \qquad 0 \leq x \leq 1, \\ u(0) = a \end{aligned}\right\} \tag{18}$$

and has the form

$$\left.\begin{aligned} 4\,\frac{u_{n+1} - u_{n-1}}{2h} - 3\,\frac{u_{n+1} - u_n}{h} + Au_n = 0, \qquad n = 1, 2, \ldots, N+1, \\ u_0 = a, \\ u_1 = (1 - Ah)a. \end{aligned}\right\} \tag{19}$$

setting $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$, we bring scheme (19) into form (2), where

$$R_h = \begin{pmatrix} 3 + Ah & -2 \\ 1 & 0 \end{pmatrix}, \qquad \rho_n \equiv 0.$$

The eigenvalues of the matrix R_h are the roots of the quadratic equation $\det(R_h - \lambda E) = 0$:

$$\lambda_{1,2} = \frac{3 + Ah}{2} \pm \sqrt{\left(\frac{3 + Ah}{2}\right)^2 - 2}\,.$$

The first root $\lambda_1(h)$ tends to 2 as $h \to 0$, so that for small h

$$|\lambda_1| > \frac{3}{2} > 1.$$

Therefore it is impossible to expect stability for any reasonable choice of norm.

In particular, if we introduce norms via the equations

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|,$$

$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \phi_n \\ \alpha \\ \beta \end{matrix} \right\|_{F_h} = \max\left[|\alpha|, |\beta|, \max_n |\phi_n|\right],$$

$$\|y_n\|_Y = \left\| \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \right\|_Y = \max\left[\left|y_n^{(1)}\right|, \left|y_n^{(2)}\right|\right],$$

we satisfy both conditions (6) and (7), thereby making (3) an inequality necessary for stability. But $\left\|R_h^n\right\| > (3/2)^n \to \infty$ if $n = 1/h$, $h \to 0$, and stability is absent.

As we have seen, gross violation of the necessary spectral stability condition (13)

$$|\lambda| \leq 1 + ch,$$

for example the presence of an eigenvalue, λ^*, of the operator R_h, satisfying the bound

$$|\lambda^*| > 1 + h^{1-\varepsilon},$$

testifies to an instability which cannot be corrected by any choice of norms.

We must stress, however, that location of the spectrum of the operator R_h in the circle $|\lambda| < 1 + ch$ still does not guarantee stability. Stability in this case may depend on a successful choice of norms, as we see by the example of the following difference scheme, which was already considered in §14 from a slightly different point of view.

The difference scheme for the solution of the problem $u'' = \phi(x)$, $u(0) = a$, $u'(0) = b$ will be written as follows

$$\left.\begin{aligned} &\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} = \phi_n, \qquad n = 1, 2, \ldots, N-1, \\ &u_0 = a, \\ &\frac{u - u_0}{h} = b. \end{aligned}\right\}$$

Setting $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$ we put this scheme into form (2), where

$$R_h = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}, \qquad \rho_n = \begin{bmatrix} h\phi_n \\ 0 \end{bmatrix}, \qquad y_0 = \begin{bmatrix} a + hb \\ a \end{bmatrix}.$$

Both eigenvalues of the matrix R_h are equal to one. In the case $\phi_n \equiv 0$ the solution, $\{u_n\}$, of this problem has the form

$$u_n = u_0 + (u_1 - u_0)n, \qquad n = 0, 1, 2, \ldots, N.$$

We now use two sets of norms:

1) $$\|y_n\|_Y = \max\left(\left|y_n^{(1)}\right|, \left|y_n^{(2)}\right|\right),$$

$$\|u^{(h)}\|_{U_h} = \max_n \|y_n\|_Y,$$

$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \phi_n \\ a \\ b \end{matrix} \right\|_{F_h} = \max\left[\|y_0\|_Y, \max_m |\phi_m|\right];$$

2) $$\|y_n\|_{Y_h} = \max\left(\left|y_n^{(1)}\right|, \left|\frac{y_n^{(2)} - y_n^{(1)}}{h}\right|\right),$$

$$\|u^{(h)}\|_{U_h} = \max_n \|y_n\|_{Y_h},$$

$$\|f^{(h)}\|_{F_h} = \max\left[\|y_0\|_{Y_h}, \max_m |\phi_m|\right].$$

The reader will easily convince himself that, in both cases, conditions (6) and (7) are satisfied, as well as (28) §14, which has the effect that stability is equivalent to the bound

$$\left\|R_h^n\right\| \leq C, \qquad n = 1, 2, \ldots, N-1.$$

If one chooses norms according to prescription 1) this bound is violated. Thus, for example, taking $y_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\|y_0\| = 1$, we get

$$y_n = \begin{bmatrix} n+1 \\ n \end{bmatrix}, \qquad \|y_n\| = n + 1, \qquad \left\|R_h^n\right\| \geq n + 1 \to \infty$$

for $n = 1/h$, $h \to 0$.

Choosing norms by prescription 2) we do have stability: for any arbitrary $y_0 = \begin{bmatrix} u_1 \\ u_0 \end{bmatrix}$ we have

$$\|y_n\|_{Y_h} = \left\|R_h^n y_0\right\| = \left\| \begin{matrix} u_0 + (u_1 - u_0)(n+1) \\ u_0 + (u_1 - u_0)n \end{matrix} \right\|_{Y_h} =$$

$$= \max\left[|u_0 + (u_1 - u_0)(n+1)|, \left|\frac{u_1 - u_0}{h}\right|\right].$$

But $n + 1 \leq 1/h$, and therefore

$$||y_n||_{Y_h} = ||R_h^n y_0||_{Y_h} \leq |u_0| + \left|\frac{u_1 - u_0}{h}\right| \leq 2||y_0||_{Y_h}$$

and

$$||R_h^n|| < 2, \qquad n = 1, 2, \ldots, N-1.$$

In practice one often limits oneself to a check as to whether the necessary spectral stability criterion is satisfied. If it is satisfied, further tests of the utility of the scheme are carried out by running experimental computations using this scheme, not troubling oneself with the explicit construction of norms. More will be said about this approach in §18.

PROBLEMS

1. Suppose that the second-order difference equation $au_{n-1} + bu_n + cu_{n+1} = \phi_n$ has been reduced to the form $y_{n+1} = R_h y_n + h\phi_n$ via the substitution

$$y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}.$$

Show that the roots of the characteristic equation $a + b\lambda + c\lambda^2 = 0$ and the eigenvalues of the matrix R_h coincide.

2. Write the second-order difference equation $au_{n-1} + bu_n + cu_{n+1} = \phi_n$ in the form $y_{n+1} = R_h y_n + h\rho_n$ with the aid of the substitution

$$y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix}.$$

Is this reduction unique? Show that the eigenvalues of the matrix R_h are the roots of the characteristic equation $a + b\lambda + c\lambda^2 = 0$, plus the root $\lambda = 0$, so that satisfaction of the spectral stability criterion $|\lambda| \leq 1 + ch$ does not depend on the choice

$$y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix} \quad \text{or} \quad y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix}.$$

3. Suppose the eigenvectors $v^{(1)}$ and $v^{(2)}$ of the 2x2 matrix R_h, corresponding to the eigenvalues λ_1 and λ_2 respectively,

$$R_h v^{(1)} = \lambda_1 v^{(1)}, \qquad R_h v^{(2)} = \lambda_2 v^{(2)},$$

tend as $h \to 0$ to different, noncollinear, orientations. Then the conditions $|\lambda_1| < 1 + ch$, $|\lambda_2| < 1 + ch$ are not only necessary, but also sufficient for a bound of the form $\left|\left|R_h^n\right|\right| < C$, $n = 1, 2, \ldots, N$ if

$$\left|\left|\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\right|\right|_Y = \max\,[|\alpha|, |\beta|].$$

Prove.

§16. Roundoff errors

1. Errors in the coefficients.

If the difference scheme

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

approximates the problem $Lu = f$ on the solution u and is stable, then we have convergence. But whatever difference scheme we have in mind, no matter how carefully it is designed, it is never implemented exactly because of roundoff errors in the given coefficients and the right hand sides.

Suppose, for example, that one is required to solve the problem

$$u' + Au = \cos x, \quad 0 \le x \le 1,$$

$$u(0) = a$$

via the difference scheme

$$\left.\begin{aligned} \frac{u_{n+1} - u_n}{h} + Au_n &= \cos x_n, \qquad n = 0, 1, \ldots, N-1,\\ u_0 &= a \end{aligned}\right\} \tag{2}$$

Values of $\cos x_n$, a and A, and of the coefficient $1/h$, will be given with roundoff errors of one sort or another. In the general case we are dealing, not with (1), but with the difference scheme

$$L_h v^{(h)} + (\Delta^{(h)} L_h) v^{(h)} = f^{(h)} + \Delta^{(h)} f^{(h)}, \tag{3}$$

where $\Delta^{(h)} L_h$ and $\Delta^{(h)} f^{(h)}$ are errors in the assigned values of the operator L_h and right-hand side $f^{(h)}$, induced by roundoff. For scheme (2) the operator $\Delta^{(h)} L_h$ has the form

$$(\Delta^{(h)}L_h)v^{(h)} = \begin{cases} \Delta^{(h)}(\frac{1}{h})(v_{n+1} - v_n) + (\Delta^{(h)}A)v_n, & n = 0, 1, \ldots, N-1, \\ 0 \cdot v_0. \end{cases}$$

The error $\Delta^{(h)}f^{(h)}$ is given by the expression

$$\Delta^{(h)}f^{(h)} = \begin{cases} \Delta^{(h)} \cos x_n, & n = 0, 1, \ldots, N-1, \\ \Delta^{(h)}a. \end{cases}$$

Here $\Delta^{(h)}M$ is the error committed in determining the quantity M.

So as to avoid purely technical difficulties, we limit ourselves to the case where L_h and $\Delta^{(h)}L_h$ are linear, and the space U_h is finite-dimensional, as in the above scheme (2). Under these assumptions we ask what sort of roundoff errors are permissable, and how the precision with which one specifies the difference scheme must increase as the net is refined, i.e. as h tends to zero.

Theorem. *If a stable difference scheme (1) approximates the problem $Lu = f$ on the solution u to some order h^k:*

$$||L_h[u]_h - f^{(h)}||_{F_h} \leq ch^k,$$

then under the conditions

$$\left.\begin{aligned} ||\Delta^{(h)}L_h v^{(h)}||_{F_h} &\leq c_1 h^k ||v^{(h)}||_{U_h}, \\ ||\Delta^{(h)}f^{(h)}||_{F_h} &\leq c_2 h^k \end{aligned}\right\} \qquad (4)$$

difference scheme (3) also approximates the problem $Lu = f$ to order h^k, and is also stable.

Thus, under conditions (4), the order of accuracy of the difference scheme (3), by which the computation is actually carried out, is h^k and coincides with the accuracy of the intended scheme (1).

Assuming that the norm $||\cdot||_{U_h}$ is chosen according to condition (4) §13, i.e. so that

$$\lim_{h\to 0} ||[u_h]||_{U_h} = ||u||_U,$$

the quantity $||[u]_h||_{U_h}$ remains bounded as $h \to 0$, $||[u]_h||_{U_h} \leq P < \infty$. We will define

$$\tilde{L}_h u^{(h)} \equiv L_h u^{(h)} + (\Delta^h L_h)u^{(h)},$$

$$\tilde{f}^{(h)} \equiv f^{(h)} + \Delta^{(h)}f^{(h)}$$

and convince ourselves that the scheme $\tilde{L}_h u^{(h)} = \tilde{f}^{(h)}$ is of order h^k. In fact we have

$$||\tilde{L}_n[u]_h - \tilde{f}^{(h)}||_{F_h} = ||L_h[u]_h - f^{(h)} + (\Delta^{(h)}L_h[u]_h - \Delta^{(h)}f^{(h)})||_{F_h} \leq$$

$$\leq ||L_h[u]_h - f^{(h)}||_{F_h} + ||\Delta^{(h)}L_h[u]_h||_{F_h} + ||\Delta^{(h)}f^{(h)}||_{F_h} \leq$$

$$\leq ch^k + c_1 Ph^k + c_2 h^k \leq \tilde{c}h^k.$$

To prove the above theorem, we will make use of the following well-known

Lemma. *Let A and B be two linear operators mapping some finite-dimensional linear normed space X into another linear normed space G. Suppose, further, that for every g in G there exists a solution x in X of the equation*

$$Ax = g,$$

with

$$||x||_X \leq c||g||_G, \tag{5}$$

and also that for any $\tilde{x}$ in X we have the inequality

$$||Bx||_G \leq \frac{q}{c}||x||_X, \tag{6}$$

with some c and q, $c > 0$, $0 < q < 1$. Then the equation

$$(A + B)\tilde{x} = g$$

has a unique solution for any g in G, and

$$||\tilde{x}||_X \leq \frac{c}{1 - q}||g||_G. \tag{7}$$

Proof. Note that X and G have the same dimensionality, since otherwise Ax = g would not be solveable for every g in G. Further, if x_0 is any solution of the equation

$$(A + B)x_0 = g,$$

then

$$Ax_0 = g - Bx_0,$$

$$x_0 = A^{-1}g - A^{-1}Bx_0,$$

where $A^{-1}g$ and $A^{-1}Bx_0$ are solutions of the equations $Ax = g$ and $Ax = Bx_0$,

$$||x_0||_X \leq ||A^{-1}g||_X + ||A^{-1}(Bx_0)||_X \leq$$

$$\leq c||g||_G + c||Bx_0||_G \leq c||g||_G + c\frac{q}{c}||x_0||_X .$$

Hence

$$||x_0||_X \leq \frac{c}{1-q}||g||_G.$$

From the latter inequality it follows, that if $g = 0$, the equation $(A + B)x = g$ has only the trivial solution $x_0 = 0$; thus there exists a unique solution for arbitrary g in G, and bound (7) is valid.

Proof of the theorem. We will use the lemma and take as operators A and B, respectively, L_h and $\Delta^{(h)}L_h$. The existence of a solution of the problem $Ax = g$, together with bound (5), are equivalent to the stability of scheme (1). Bound (6) is valid, by virtue of (4), for any positive q so long as h is small enough.

The solveability of the equation $(A + B)x = g$ for any g in G, jointly with bound (7), are exactly equivalent to the stability of difference scheme (3).

We note that the restriction (4) on roundoff errors is perfectly reasonable for a stable difference scheme: if, on decreasing h, we want to obtain a solution accurate to h^k, i.e. with a number of significant decimal digits of order $k\ln(1/h)$, then also the coefficients of the difference scheme will have to be given more and more accurately, increasing the number of figures to which they are given also at a rate of order $k\ln(1/h)$. Such a rate of increase is ordinarily perfectly attainable, since $k\ln(1/h)$ is a slowly growing function. If one decreases the step-size, not increasing the number of significant figures with which the coefficients and right-hand sides are given, then there will be no improvement at all in the accuracy obtained.

2. Computational errors. After the difference scheme is given it is still necessary to compute its solution, $u^{(h)}$. Suppose we can solve the difference equations exactly. Then, if the difference scheme we are using approximates the differential equation and is stable, we know that for a small enough step-size the solution $u^{(h)}$ will differ little from the desired exact solution $[u]_h$. Moreover it is completely immaterial by what sequence of actions (or "algorithm") the computation of $u^{(h)}$ is carried out, since the outcome of the computation does not depend on details of this sequence.

But, in reality, having chosen some algorithm for the computation of the solution $u^{(h)}$ we will, at each step of the implementation of this algorithm, commit roundoff errors which will influence the results of subsequent computational steps. For fixed h and a finite-dimensional space U_h the algorithm consists of a finite sequence of arithmetic operations. The result of each arithmetic operation (the computation of a sum, difference, product or quotient) depends continuously on the quantities on which the operation is performed. Therefore, carrying out the computations with a "large enough" number of significant figures, we can calculate $u^{(h)}$ to any prescribed number of decimal places. The number of "spare" figures which must be carried in the computation so as to get a prescribed number of figures in $u^{(h)}$ depends both on the algorithm chosen, and on h. Thus, for example, it was shown in §7 that, when solving a well-conditioned boundary-value problem by FEBS, the number of required extra significant figures does not increase at all as $h \to 0$. Sometimes a seemingly reasonable algorithm for the solution of a stable problem may require a rapidly increasing number of spare figures, a number proportional to 1/h. An example of such an algorithm was presented in 2§5. With decreasing h this number will, generally, have to grow. An algorithm in which it grows too rapidly is considered unstable and, from a practical point of view, unuseable for computation. The study of the stability of algorithms is complicated. An example of such a study is the establishment of a basis for the FEBS method in §7. But in the simplest cases one can manage to understand how many spare figures are required, relying only upon information on the stability of the difference scheme, and on the theorem proved in section 1, above, dealing with the possibility of specifying the difference scheme approximately.

Suppose, for example, that we carry out a calculation according to the scheme

$$\frac{u^{(h)}(x + h) - u^{(h)}(x)}{h} + Au^{(h)}(x) = f^{(h)}(x).$$

Determining $u^{(h)}(x + h)$ from the recurrence relation

$$u^{(h)}(x + h) = u^{(h)}(x)(1 - Ah) + hf^{(h)}(x)$$

and computing with a finite number of significant figures we may have allowed, into $u^{(h)}(x + h)$, some error δ. It is convenient to suppose that the error was introduced, not in the value of $u^{(h)}(x + h)$, but in the right-hand side, $f^{(h)}$, used in the computation; i.e. to consider that we calculated $u^{(h)}(x + h)$ exactly but used, in place of $f^{(h)}(x)$, the quantity $f^{(h)}(x) + \delta/h$. Since such errors are committed at each point x, the value of δ must be taken to depend on x, so that $\delta = \delta(x)$. Thus in this example the computational roundoff error can be thought of as an error, $\delta(x)/h$, in the specification of the right-hand side. The difference scheme under

consideration is a first-order approximation and is stable. Therefore if we are not to spoil the order-h convergence, we must perform the computation with increasing accuracy and, in fact, in such a way that

$$\Delta^{(h)} f^{(h)} = \frac{\delta(x)}{h}$$

be of order h.

This requires that $\delta(x)$ be of order h^2. Such an accuracy may be attained by computing $u^{(h)}$ with a number of extra significant figures increasing, as $h \to 0$, like $\ell n(1/h)$.

Through this example we have shown that, in simple cases, roundoff errors committed in computing $u^{(h)}$ to an accuracy proportional to h^m can be considered errors in specification of the right-hand side $f^{(h)}$. From the theorem proven above it follows that, for stable schemes, these errors do not prevent convergence, and convergence without loss of order of accuracy, if the number of significant figures carried in the computation slowly grows, like $c\ \ell n(1/h)$ where c is some constant.

§17. Quantitative aspects of stability

We begin by considering the familiar example of difference scheme

$$\left.\begin{aligned} \frac{u_{n+1} - u_n}{h} + Au_n &= 0, \\ u_0 &= 1 \end{aligned}\right\} \qquad (1)$$

for the differential boundary-value problem

$$u' + Au = 0, \qquad u(0) = 1.$$

Its solution has the form

$$u_n = e^{-Ax_n} + h\frac{A^2 x_n}{2} e^{-Ax_n} + O(h^2)$$

(see (3′) of §8; we are taking b = 1). Expression (6) of §8

$$\delta(x_n) = h\frac{A^2 x_n}{2} e^{-Ax_n} + O(h^2)$$

represents the remainder term, i.e. the error committed in replacing the value, $\exp(-Ax_n)$, of the exact solution of the differential equation by the solution, $u_n^{(h)}$, of the difference problem. The remainder term tends to zero like the first power of h; this scheme is accurate to first order. The choice of a step-width, h, depends on the accuracy we want to attain. Clearly the modulus of the ratio of the error to the exact solution, $|\delta(x_n)/u(x_n)|$, must in any case be less than unity if the approximate solution is to be considered accurate at all.

Let us consider for what values of h this condition is satisfied. In the expression for $\delta(x_n)$ we will neglect the term $O(h^2)$ and examine the ratio of the error, $\delta(x_n)$, at point x_n, to the exact solution

$$\frac{\delta(x_n)}{u(x_n)} \approx \frac{h\,\frac{A^2x_n}{2}\,e^{-Ax_n}}{e^{-Ax_n}} = h\,\frac{A^2x_n}{2}.$$

We will take A = 20 and examine this ratio at the point $x_n = 1$. Then from the condition $|\delta(1)/u(1)| < 1$ we get

$$h < 0.2 \times 10^{-3}.$$

Now we determine what step-sizes are required for the integration of this same problem, $u' + Au = 0$, using a scheme of second-order accuracy

$$\left.\begin{aligned} &\frac{u_{n+1} - u_{n-1}}{2h} + Au_n = 0 \\ &u_0 = 1, \\ &u_1 = 1 - Ah, \end{aligned}\right\} \qquad (2)$$

if, again, A = 20 and we again set it as our goal to satisfy the condition

$$\left|\frac{\delta(1)}{u(1)}\right| < 1 \qquad (3)$$

The solution of this problem has the form (see Eq. (12) of §8 for b = 1)

$$u_n = e^{-Ax_n} + h^2\left[\frac{2Ax_n - 3}{12}\,A^2e^{-Ax_n} + (-1)^n\,\frac{A^2}{4}\,e^{Ax_n}\right] + O(h^3).$$

The error, therefore, has the form

$$\delta(x_n) = h^2\left[\frac{2Ax_n - 3}{12}\,A^2e^{-Ax_n} + (-1)^n\,\frac{A^2}{4}\,e^{Ax_n}\right] + O(h^3).$$

Let us neglect the term $O(h^3)$, write out the ratio of the error, $\delta(x_n)$, to the exact solution $u(x_n) = \exp(-Ax_n)$, and determine the step-size, h, from condition (3). This step-size will turn out to be so small that, if we arbitrarily take a second as a computing time for scheme (1), the required time for scheme (2) will be four days!

The point is that an evaluation of the practical utility of this or that scheme for the solution of a given problem must be made, not solely on the basis of the power of h in the expression for the error, but also considering the coefficient of this power of h.

Now we will try to understand how one can judge the utility of some given difference scheme, $L_h u^{(h)} = f^{(h)}$, from a study of its stability. For the sake of brevity we will take the operator L_h to be linear. We recall (see §12) that a difference scheme is called "stable" if, for any $f^{(h)}$ in F_h, it has a unique solution $u^{(h)}$ in U_h, satisfying the bound

$$||u^{(h)}||_{U_h} \leq C||f^{(h)}||_{F_h}.$$

Proving, in §12, a theorem stating that approximation and stability imply convergence we got, for the error $z^{(h)} = [u]_h - u^{(h)}$, the inequality

$$||z^{(h)}||_{U_h} \leq CC_1 h^k,$$

in which $C_1 h^k$ represents a bound on the approximation error:

$$||L_h[u]_h - f^{(h)}||_{F_h} \leq C_1 h^k.$$

Suppose the approximation error $C_1 h^k$ is small. From the bound for $||z^{(h)}||_{U_h}$ one can see that, if $||[u]_h - u^{(h)}||_{U_h}$ is to be small, it is still necessary that the coefficient C, characterizing stability, should not be too large.

Therefore, if we wish to determine the utility of this or that difference scheme for the solution of some particular problem it does not suffice to know that the scheme is stable. We must also know the approximate value of the coefficient C, of which one can form some idea either by the methods indicated in §§14 and 15, or by experimental computations, or by some indirect approach.

Let us calculate, for example, the coefficient C implicit in difference schemes (1) and (2) for the solution of the problem $u' + Au = \phi(x)$, $u(0) = 1$, a problem which we discussed at the beginning of this section. First we consider the scheme

$$\frac{u_{n+1} - u_n}{h} + Au_n = \phi_n, \qquad n = 0, 1, \ldots, N-1,$$

$$u_0 = a$$

with norms

$$||u^{(h)}||_{U_h} = \max_n |u_n|, \qquad ||f^{(h)}||_{F_h} = \max[|a|, \max_n |\phi_n|].$$

We reduce this scheme to the form

$$\left.\begin{aligned} y_{n+1} &= R_h y_n + h\rho_n, \\ y_0 &\text{ given,} \end{aligned}\right\}$$

setting $y_n = u_n$, $R_h = (1 - Ah)$, $\rho_n = \phi_n$. Let $||y_n|| = |y_n|$. Then condition (17) of §14 is satisfied:

$$\left.\begin{aligned} ||u^{(h)}||_{U_h} &\le C_2 \max_n ||y_n||, \\ ||\rho_n|| &\le C_2 ||f^{(h)}||_{F_h}, \\ ||y_0|| &\le C_2 ||f^{(h)}||_{F_h}, \end{aligned}\right\} \tag{5}$$

where, in fact, we can let $C_2 = 1$.

Further, obviously $\left|\left|R_h^n\right|\right| = (1 - Ah)^n$. For this reason we can set $C = 2 \max\{1, (1 - Ah)^n\}$. Hence

$$C = \begin{cases} 2, & \text{if} \quad A > 0, \\ 2(1 - Ah)^N, & \text{if} \quad A \le 0. \end{cases}$$

We now show that the quantity C cannot be taken substantially smaller. The norms have been chosen such, as to satisfy conditions (6) and (7) of § 15:

$$||u^{(h)}||_{U_h} \ge M_1 \max_n ||y_n||, \tag{6}$$

and for $\phi_n = 0$ $(\rho_n = 0)$ also

$$||y_0|| \ge M_2 ||f^{(h)}||_{F_h}, \tag{7}$$

where we can set $M_1 = M_2 = 1$. Therefore the constant C must, as was established in §15, satisfy the bound $C \ge M_1 M_2 \left|\left|R_h^n\right|\right|$:

$$C \ge \begin{cases} 1, & \text{if} \quad A > 0, \\ (1 - Ah)^N, & \text{if} \quad A \le 0. \end{cases}$$

Now we evaluate the constant C, in the definition of stability $||u^{(h)}||_{U_h} \le C||f^{(h)}||_{F_h}$, for difference scheme (2). Let us first write this scheme in the form

$$y_{n+1} = R_h y_n + h\rho_n, \qquad n = 0, 1, \ldots,$$
$$y_0 \text{ given},$$

setting, for this purpose,

$$y_n = \begin{bmatrix} u_n + 1 \\ u_n \end{bmatrix}, \qquad \rho_n = \begin{bmatrix} 2\phi_n \\ 0 \end{bmatrix},$$

$$R_h = \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix}, \qquad y_0 = \begin{bmatrix} (1 - Ah) \\ 1 \end{bmatrix}.$$

We choose the norms

$$||u^{(h)}||_{U_h} = \max_n |u_n|,$$

$$||f^{(h)}|| = \left|\left|\left| \begin{bmatrix} \phi_n \\ \alpha \\ \beta \end{bmatrix} \right|\right| = \max \left[|\alpha|, |\beta|, \max_n |\phi_n| \right],$$

$$||y_n|| = \left|\left|\left| \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \right|\right| = \max \left| y_n^{(1)} \right|, \left| y_n^{(2)} \right| .$$

Conditions (5) - (7) are then satisfied, with $C_2 = M_1 = M_2 = 1$. Therefore by virtue of what has been said in section 3§14 we may take, as the constant C, the quantity $C = 2C_2^2 \max_n \left|\left|R_h^n\right|\right| = 2 \max_n \left|\left|R_h^n\right|\right|$ but, by (6´´) of §15, we cannot decrease this value of C by more than a factor of 2: certainly it must be true that

$$C \geq M_1 M_2 \max_n \left|\left|R_h^n\right|\right| = \max_n \left|\left|R_h^n\right|\right|.$$

An upper bound on the value of $\max \left|\left|R_h^n\right|\right|$ was obtained in §14:

$$\left|\left|R_h^n\right|\right| \leq \left|\left|\begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix}\right|\right|^N \leq (1 + 2|A|h)^N.$$

Thus we may set

$$C = 2e^{2|A|} \geq 2(1 + 2|A|h)^{1/h}.$$

A lower bound for $\max\left\|R_h^n\right\|$ can be gotten from the condition $\left\|R_h^n\right\| \geq |\lambda|^n$, where λ is the larger (in modulus) of the two eigenvalues of the matrix R_h. Solving the equation $\det(R_h - \lambda E) = 0$ we find the eigenvalues

$$\lambda_1 = 1 - Ah + \frac{A^2h^2}{2} + o(h^2) = 1 - Ah + O(h^2),$$

$$\lambda_2 = -1 - Ah - \frac{A^2h^2}{2} + o(h^2) = -1 - Ah + O(h^2),$$

so that

$$\max\left(|\lambda_1|, |\lambda_2|\right) = 1 + |A|h + O(h^2),$$

$$\max_n \left\|R_h^n\right\| \geq (1 + |A|h)^{1/h} + O(h).$$

Therefore the above constant $C = 2e^{2|A|}$ certainly cannot be replaced by a number smaller than $(1 + |A|h)^{1/h} \approx e^{|A|}$, i.e. it cannot be decreased substantially.

For $A = 20$ we see that, for the first scheme, $C = 2$, but for the second $C \geq e^{20} \geq 10^8$.

For $A \approx 1$ or $A < 0$ the two schemes do not differ fundamentally in their stability properties; the constant C is approximately the same for both schemes. It is easy to understand the mechanism by which, for $A \gg 1$, the constant C for the second scheme becomes much larger than unity, whereas for the first $C = 2$.

The general solution of the homogeneous equation $u_{n+1} - (1 - Ah)u_n = 0$, corresponding to scheme (1), is $\bar{u}_n = \alpha q^n$, where q is the root of the characteristic equation $q - (1 - Ah) = 0$, $q = 1 - Ah$ (Fig. 4). The general solution of the homogeneous equation

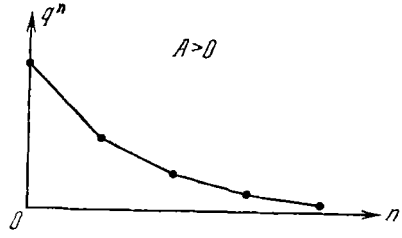


Fig. 4

$$u_{n+1} + 2Ahu_n - u_{n-1} = 0,$$

corresponding to scheme (2), is

$$u_n = \alpha q_1^n + \beta q_2^n,$$

where q_1 and q_2 are the roots of the characteristic equation

$$q^2 + 2Ahq - 1 = \det(R_h - qE) = 0,$$

$$q_1 = 1 - Ah + \frac{A^2h^2}{2} + o(h^2),$$

$$q_2 = -1 - Ah - \frac{A^2h^2}{2} + o(h^2).$$

The root q_1 is "similar" to the root $1 - Ah$, and to it corresponds the solution q_1^n, similar to the solution q^n of the first equation. But the "parasitic" root, $q_2 = -1 - Ah + O(h^2)$, produces a quickly-growing "parasitic" solution q_2^n (Fig. 5), which gives rise to a large value of C.

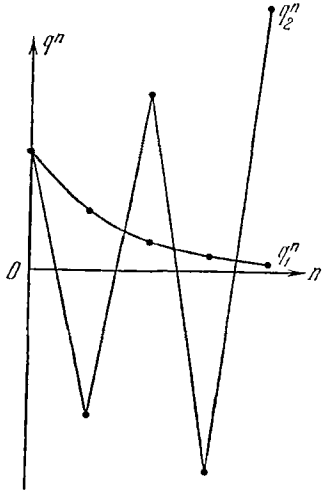


Fig. 5.

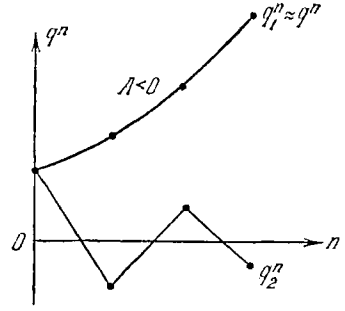


Fig. 6

For negative A we have $q > 1$, $q_1 > 1$, $|q_2| < 1$. The solutions q^n and q_1^n, corresponding to the roots q and q_1, grow about equally fast, while the parasitic solution q_2^n is damped, not influencing the stability properties of the second scheme (Fig. 6).

We note that, for $A \ll 0$, a large value of C is unavoidable in any difference scheme approximating the problem $u' + Au = 0$, $u(0) = a$. In fact, for small h the solution of a stable difference problem is similar to the solution of the differential problem to which it converges as $h \to 0$. But the solution of the differential problem, $u = u_0 \exp(-Ax)$, is such that $\max|u(x)| = |u_0| \exp(-Ax)$, i.e. $\max |u(x)|$ exceeds the modulus, $|u_0|$, of the starting value u_0, by the very large factor $\exp(-A)$.

We must also note that a large coefficient C not only makes it necessary to compute with a small step-size, but also to carry out the calculation with a large number of significant figures.

In fact we showed, in §16, that roundoff errors may be treated as errors in the specification of the right-hand sides, errors whose magnitude

is given by terms of the form $C_1 h^k$. An increase in these errors induces an increase in the coefficient C_1, which for large C can (by virtue of (4)) have a catastrophic effect on the accuracy of the result.

Before concluding this section we would like to warn the reader against any misleading impressions about difference schemes of second-order accuracy, misleading impressions which may have been generated through consideration of the above example. It was not at all our intention to condemn all such schemes in describing the inadequacies of one of them. The reader will find it very useful to study, independently, the scheme of second order accuracy

$$\frac{u_{n+1} - u_n}{h} + A\,\frac{u_{n+1} + u_n}{2} = 0,$$

$$u_0 = 1.$$

If one attempts, for A = 1, to achieve an accuracy such that the error, $\delta(1)$, is smaller than $u(1) = \exp(-A)$, one will find that this scheme puts much weaker restrictions on the step-size, h, than the first-order-accurate scheme (1).

In addition we suggest that the reader calculate what step-size is required to integrate the problem $u' + u = 0$, $u(0) = 1$, so as to compute $u(1)$ with an error no greater than 10^{-5}. If one carries out this calculation for schemes (1) and (2), considered at the beginning of this section, it will be seen that the first-order-accurate scheme (1) requires a significantly smaller step-size than second-order-accurate scheme (2).

Thus the effectiveness or ineffectiveness of this or that scheme will depend, not only on the scheme itself, but also on the problem to which it is applied.

§18. Method for studying stability of nonlinear problems.

The methods developed above, in §§14 and 15, for the study of stability, were specifically designed for difference schemes with constant coefficients. Therefore it may seem that it is impossible to use the material presented in these preceding sections for the analysis of schemes to integrate even the simple equation $du/dx = G(x,u)$, for a fairly general function G. This is, however, not true.

Suppose the desired integral curve of the equation

$$\frac{du}{dx} = G(x,u) \tag{1}$$

passes through the point with coordinates $x = x_0$, $u = u_0$. Near these points we have

$$G(x,u) \approx \frac{\partial G(x,u)}{\partial u}(u - u_0) + \frac{\partial G(x,u)}{\partial x}(x - x_0) + G(x_0, u_0), \tag{2}$$

and therefore Eq. (1), to a certain accuracy, may be replaced by the simpler

$$\frac{du}{dx} - Au = \phi(x), \tag{3}$$

where

$$A = \left.\frac{\partial G(x,u)}{\partial u}\right|_{\substack{x=x_0\\u=u_0}},$$

$$\phi(x) = G(x_0, u_0) + \left.\frac{\partial G(x,u)}{\partial x}\right|_{\substack{x=x_0\\u=u_0}} (x - x_0) - u_0 \left.\frac{\partial G(x, u)}{\partial u}\right|_{\substack{x=x_0\\u=u_0}}.$$

It is plausible that schemes which we propose to use to solve Eq. (1) should satisfactorily integrate Eq. (3), approximating Eq. (1) close to some point which lies on the integral curve. Of course for different points of this curve the values of the coefficient A, obtained from the original equation by the linearization methods just described, will differ from each other. Therefore, after choosing one difference scheme or another, we must test it on Eq. (3), not with only one value of A, but with a whole set of such values, adequately sampling the range of variation of $\partial G/\partial u$ along the integral curve. In the overwhelming majority of cases encountered in practice, such an investigation turns out to be good enough to bring out all the scheme's weaknesses and strengths, having some bearing on the character of the convergence of the approximate solutions which it produces.

Precisely the same method of constructing model problems can be applied also to systems of equations, and to equations of higher order.

In practice the solution of the Cauchy problem, for ordinary differential equations with no special peculiarities, is accomplished by one or two, fairly general, well-tested schemes for which, on present-day computers, there are standard programs. If it becomes necessary to solve, with very high precision, a problem of a special type, then one uses one of the many special schemes adapted specifically for such special problems, resorting to the more general schemes when one is concerned with a different problem area.

This Page Intentionally Left Blank

Chapter 6

Widely-Used Difference Schemes

§19. Runga-Kutta and Adams Schemes

Here we present some widely-used difference schemes for the solution of the Cauchy problem defined by the first-order differential equation

$$\left.\begin{aligned} \frac{du}{dx} - G(x, u) = 0, \qquad 0 \leq x \leq 1, \\ u(0) = a. \end{aligned}\right\} \tag{1}$$

Below in section 4 these schemes will be generalized to systems of first-order equations, to which one can reduce the general case of equations and systems of any order.

We will take, on the segment $0 \leq x \leq 1$, the net of points

$$0 = x_0 < x_1 < x_2 < \ldots < x_{N-1} < x_N = 1, \quad x_n = nh, \quad h = 1/N,$$

and construct difference schemes for the approximate determination of the table, $[u]_h$, of the solution-values on this net.

The simplest scheme in widespread use is one we have already met. This is the Euler scheme

$$L_h u^{(h)} \equiv \left\{\begin{aligned} \frac{u_{n+1} - u_n}{h} - G(x_n, u_n) = 0, \qquad n = 0, 1, \ldots, N-1, \\ u_0 = a, \end{aligned}\right. \tag{2}$$

possessing first-order approximation (and accuracy). Computation via this scheme has a simple geometric interpretation. If u_n has already been computed, then the computation

$$u_{n+1} = u_n + hG(x_n, u_n)$$

is equivalent to a shift from point (x_n, u_n) to point (x_{n+1}, u_{n+1}), in plane Oxu, along the tangent to the integral curve, $u = u(x)$, of the differential equation $u' = G(x, u)$, passing through the point (x_n, u_n).

Among the schemes with higher-order approximation, the most widely used are the different variants of the Runga-Kutta and Adams schemes, which we describe and compare.

1. Runga-Kutta scheme. Suppose the value, u_n, of the approximate solution at point x_n has already been found, and one is required to compute u_{n+1} at point $x_{n+1} = x_n + h$. We choose an integer ℓ and write the expressions

$$k_1 = G(x_n, u_n),$$
$$k_2 = G(x_n + \alpha h, u_n + \alpha h k_1),$$
$$k_3 = G(x_n + \beta h, u_n + \beta h k_2),$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$k_\ell = G(x_n + \gamma h, u_n + \gamma h k_{\ell-1}).$$

Then we set

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - (p_1 k_1 + \ldots + p_\ell k_\ell) = 0, & n = 0, 1, \ldots, N-1, \\ u_0 = a. \end{cases}$$

The coefficients α, β, ..., γ, p_1, p_2, ..., p_ℓ will be chosen such as to give, for the given ℓ, approximation of the highest possible order. Knowing u_n one can compute k_1, ..., k_ℓ, and then

$$u_{n+1} = u_n + h(p_1 k_1 + \ldots + p_\ell k_\ell).$$

The simplest Runga-Kutta scheme is the Euler scheme ($\ell = 1$). The Runga-Kutta scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - \dfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 0, \\ \qquad\qquad n = 0, 1, \ldots, N-1, \\ u_0 = a, \end{cases} \tag{3}$$

where

$$k_1 = G(x_n, u_n),$$
$$k_2 = G\left(x_n + \frac{h}{2}, u_n + \frac{k_1 h}{2}\right),$$
$$k_3 = G\left(x_n + \frac{h}{2}, u_n + \frac{k_2 h}{2}\right),$$
$$k_4 = G(x_n + h, u_n + k_3 h),$$

has fourth-order approximation.

The Runga-Kutta scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - \left[\dfrac{2\alpha - 1}{2\alpha} k_1 + \dfrac{1}{2\alpha} k_2\right] = 0 \\ \qquad\qquad n = 0, 1, \ldots, N-1 \\ u_0 = a, \end{cases} \tag{4}$$

where

$$k_1 = G(x_n, u_n), \qquad k_2 = G(x_n + \alpha h, u_n - \alpha h k_1),$$

for any given α has second-order approximation.

We prove only the assertion about scheme (4). The proof of the assertion about scheme (3) is analogous, but more complicated.

* * * * * *

The solution, $u(x)$, of the equation $u' = G(x, u)$ satisfies the identities

$$\frac{du}{dx} \equiv G(x, u(x)),$$

$$\frac{d^2u}{dx^2} \equiv \frac{d}{dx} G(x,u) = \frac{\partial G}{\partial x} + \frac{\partial G}{\partial u} g.$$

Therefore it follows from the Taylor formula

$$\frac{u(x_n + h) - u(x_n)}{h} = u'(x_n) + \frac{h}{2} u''(x_n) + O(h^2)$$

for the solution $u(x)$, that

$$\frac{u(x_{n+1}) - u(x_n)}{h} - \left[G + \frac{h}{2}\left(\frac{\partial G}{\partial x} + \frac{\partial G}{\partial u} G\right)\right]_{\substack{x=x_n \\ u=u(x_n)}} = O(h^2). \tag{5}$$

But, expanding the functions of two variables in powers of h by Taylor's formula, and retaining only terms of first order, we get

$$\frac{2\alpha - 1}{2\alpha} k_1 + \frac{1}{2\alpha} k_2\Big|_{\substack{x=x_n \\ u=u(x_n)}} = \frac{2\alpha - 1}{2\alpha} G + \frac{1}{2\alpha} G(x + \alpha h,\ u + \alpha hG)\Big|_{\substack{x=x_n \\ u=u(x_n)}} =$$

$$= \frac{2\alpha - 1}{2\alpha} G + \frac{1}{2\alpha} \left[G + \frac{\partial G}{\partial x} \alpha h + \frac{\partial G}{\partial u} \alpha hG + O(h^2)\right]_{\substack{x=x_n \\ u=u(x_n)}} =$$

$$= G + \frac{h}{2} \left(\frac{\partial G}{\partial x} + \frac{\partial G}{\partial u} G\right)\Big|_{\substack{x=x_n \\ u=u(x_n)}} + O(h^2). \qquad (6)$$

Therefore if one puts into the left-hand side of (4), in place of u_n and u_{n+1}, respectively, the values $u(x_n)$ and $u(x_{n+1})$ of the solution $u(x)$, one gets an expression which agrees with the left side of Eq. (5) up to terms $O(h^2)$. Therefore this expression, (4), is of second order with respect to h. Since the initial value $u_0 = a$ is given exactly we have now proven that scheme (4) has second-order approximation.

* * *

To obtain u_{n+1} by the Runga-Kutta scheme, with u_n given, one must evaluate the function $G(x, u)$ ℓ times. The computed ℓ values are then not used any further.

2. Adams schemes. In the Adams schemes, one variant of which we will now describe, computation of the next value, u_{n+1}, requires the evaluation of $G(x, u)$ only at one point, regardless of the order of approximation. In addition it is necessary to carry out a small number of subtractions and additions which require much less time than one evaluation of even a slightly complicated function $G(x,u)$.

We adopt the notation

$$\nabla f_n = f_n - f_{n-1},$$

$$\nabla^2 f_n = \nabla(\nabla f_n) = \nabla f_n - \nabla f_{n-1} = f_n - 2f_{n-1} + f_{n-2},$$

$$\nabla^3 f_n - \nabla\nabla^2 f_n = f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}$$

and write $G_n = G(x_n, u_n)$. Let us write out explicitly some of the difference equations used in Adams schemes for the computation of u_{n+1}, if u_n, u_{n-1}, ... have already been computed:

$$\frac{u_{n+1} - u_n}{h} - G_n = 0, \qquad n = 0, 1, \ldots, N-1, \tag{7}$$

$$\frac{u_{n+1} - u_n}{h} - G_n - \frac{1}{2}\nabla G_n = 0, \qquad n = 1, 2, \ldots, N-1, \tag{8}$$

$$\frac{u_{n+1} - u_n}{h} - G_n - \frac{1}{2}\nabla G_n - \frac{5}{12}\nabla^2 G_n = 0, \qquad n = 2, 3, \ldots, N-1, \tag{9}$$

$$\frac{u_{n+1} - u_n}{h} - G_n - \frac{1}{2}\nabla G_n - \frac{5}{12}\nabla^2 G_n - \frac{3}{8}\nabla^3 G_n - 0 \qquad n = 3, 4, \ldots, N-1. \tag{10}$$

The first of these equations is the difference equation of Euler. If one substitutes into the left-hand sides of Eqs. (7)-(10), in place of u_{n+1}, u_n, u_{n-1},... the values u((n+1)h), u(nh), ... of the exact solution then residuals will appear, in Eqs. (7)-(10), of order h, h^2, h^3 and h^4 respectively.

* * * * * *

The Adams formulae may be obtained as follows. Suppose u(x) is the solution of the equation

$$\frac{du}{dx} = G(x, u).$$

Define

$$G\big(x, u(x)\big) \equiv F(x).$$

Then

$$u(x_n + h) - u(x_n) = \int_{x_n}^{x_n+h} u' \, dx = \int_{x_n}^{x_n+h} F(x) \, dx.$$

From the theory of interpolation it is known that there is one and only one polynomial, $P_k(x, F)$, of order no higher than k, taking on at the k+1 points x_n, x_{n-1}, ..., x_{n-k} the given values $F(x_n)$, $F(x_{n-1})$, ..., $F(x_{n-k})$ respectively. This polynomial $P_k(x, F)$, for a sufficiently smooth function F(x), deviates from F(x) on the interval $x_n \le x \le x_{n+h}$ by a quantity of order h^{k+1}, so that

$$\max|P_k(x, F) - F(x)| = O(h^{k+1}). \tag{11}$$

The Adams difference formula has the form

$$\frac{u_{n+1} - u_n}{h} - \frac{1}{h}\int_{x_n}^{x_n+h} P_k(x, F)dx = 0. \tag{12}$$

Inserting into the left-hand side, in place of

$$u_n,\ u_{n+1},\ G(x_{n-s},\ u_{n-s})$$

the corresponding values

$$u(x_n),\ u(x_{n+1}),\ G(x_{n-s},\ u(x_{n-s}))$$

we get a residual of order h^{k+1}:

$$\left|\frac{u(x_n + h) - u(x_n)}{h} - \frac{1}{h}\int_{x_n}^{x_n+h} P_k(x, F)dx\right| =$$

$$= \left|\left[\frac{u(x_n + h) - u(x_n)}{h} - \frac{1}{h}\int_{x_n}^{x_{n+h}} F(x)dx\right] + \frac{1}{h}\int_{x_n}^{x_{n+h}} [F(x) - P_k(x,F)]dx\right| \leq$$

$$\leq 0 + \max|F(x) - P_k(x, F)| = O(h^{k+1})$$

For $k = 0$ the interpolating polynomial

$$P_0(x, F) = G(x_n, u_n) = \text{const}$$

and Eq. (12) transforms into (7).

For $k = 1$

$$P_1(x, F) = \frac{1}{h}\left[(x - x_{n-1})G_n - (x - x_n)G_{n-1}\right].$$

Further

$$\frac{1}{h}\int_{x_n}^{x_n+h} P_1(x, F)dx = \frac{1}{h^2}\frac{(x - x_{n-1})^2}{2}\Bigg|_{x_n}^{x_n+h} G_n - \frac{1}{h^2}\frac{(x - x_n)^2}{2}\Bigg|_{x_n}^{x_n+h} G_{n-1} =$$

$$= \frac{1}{h^2}\left(\frac{4h^2}{2} - \frac{h^2}{2}\right)G_n - \frac{1}{h^2}\frac{h^2}{2}G_{n-1} = G_n + \frac{1}{2}\nabla G_n.$$

Thus Eq. (12) becomes (8). Analogously for k = 2 and k = 3 we get, from (12), Eqs. (9) and (10) respectively.

* * *

To use scheme (7) it suffices to know $u_0 = a$. To start computing via scheme (8) one must know, beforehand, not only $u_0 = a$, but also u_1. Scheme (9) requires the use of u_0, u_1 and u_2, while for scheme (10) we need four values, i.e., u_0, u_1, u_2 and u_3. These values may be found by the Runga-Kutta method; or by Euler's scheme with small step-sizes; or perhaps by expansion of the solution in a Taylor series about the point x = 0. The need for special starting procedures is one of the disadvantages of the Adams schemes, as compared with the Runga-Kutta schemes. The advantages of the Adams schemes, already noted earlier, is the fact that in the computation of u_{n+1}, given the values of G_s, ∇G_s, ..., $\nabla^k G_s$ already found in the calculation of u_n, u_{n-1}, ..., one needs to compute only one value of the function G, i.e. $G_n = G(x_n, u_n)$, and to carry out a few subtractions involved in the evaluation of ∇G_n, ..., $\nabla^k G_n$.

Thus the advantage of the Adams methods over the Runga-Kutta methods consists in the smaller computational effort required for each step. The basic disadvantages are: the need for special starting methods, and the fact that one cannot (without complicating the computational equations) change the step size h, $x_{n+1} = x_n + h$, in the course of the computation, starting from some point x_n. This latter fact is important in those cases where the solution and its derivatives on some parts of the interval change quickly, changing slowly on other parts.

If such a situation develops during the computation a Runga-Kutta subroutine, for example, might be brought into play to decrease the step-size automatically, or to increase the step-size over smooth parts of the solution-curve so as not to do unnecessary work. Evidently the most sensible approach is to use both the Runga-Kutta *and* Adams methods, automatically switching from one to the other during the computation. Using this approach one must start via the Runga-Kutta scheme. The computer program must contain provisions for automatic control of the step-size, which will be adjusted so as to maintain the required accuracy. Moreover a certain degree of conservatism must be incorporated into the step-size control mechanism; one must call for a change in step-size only when there is a very pressing need for such a change. If it turns out that, after computation of several successive values of u_n by the Runga-Kutta scheme, no step-size change occurs, then it is appropriate to switch automatically to the more economical Adams method. As soon as it again becomes necessary to change the step-size the computational program must again go over to the Runga-Kutta scheme, etc.

So as to monitor the adequacy of the step-size one ordinarily carries out, in parallel, computations with some given step-size, and with another

half as large. Within the required accuracy limits the solutions must coincide. Otherwise the step-size must be decreased. It is also necessary to provide some sort of test which will determine whether it is possible to increase the step-size.

3. Note on stability. For the problem $u' + Au = 0$, linear and with constant coefficient A, the Runga-Kutta equations turn out, after elimination of $k_1, k_2, \ldots$, to be first-order difference equations,

$$u_{n+1} - a(h)u_n = 0.$$

The root of the characteristic equation $\lambda - a(h) = 0$ is $\lambda = a(h)$.

In the case $u_n = u(x_n)$ one gets a value of u_{n+1} which agrees with the exact solution $u(x_n + h)$ up to order h^{p+1}, where p is the order of approximation. Since

$$u(x_n + h) = u(x_n)e^{-Ah} = u(x_n)\left(1 - Ah + \frac{A^2h^2}{2} - \ldots\right),$$

and

$$u_{n+1} = a(h)u_n,$$

then

$$\lambda = a(h) = e^{-Ah} + O(h^{p+1}).$$

Thus

$$|\lambda(h)| < 1 + ch.$$

The powers $\lambda^n(h)$ behave "correctly": they grow if $A < 0$ and the solution of the differential equation grows. They decrease if $A > 0$ and the solution $\exp(-Ax)$ decreases.

In the case of the Adams scheme (8)

$$\frac{u_{n+1} - u_n}{h} + Au_n + \frac{A}{2}(u_n - u_{n-1}) = 0 \qquad (13)$$

the characteristic equation has the form

$$\lambda - \left(1 - \frac{3Ah}{2}\right)\lambda - \frac{Ah}{2} = 0.$$

Therefore

$$\lambda = \frac{1}{2}\left(1 - \frac{3Ah}{2}\right) \pm \frac{1}{2}\sqrt{\left(1 - \frac{3Ah}{2}\right)^2 + 2Ah},$$

$$\lambda_1 = 1 - Ah + O(h^2),$$

$$\lambda_2 = O(h),$$

Thus the solution $u_n = \lambda_1^n$ behaves, as $h \to 0$, like $u(x_n) = \exp(-Anh)$, while the "parasitic solution" λ_2^n, which enters because of the use of a second-order difference equation, tends to zero since $|\lambda_2| = O(h)$, and thus does not affect stability.

It will be useful for the reader to compare scheme (13) with the second-order scheme (2) of §17:

$$\frac{u_{n+1} - u_{n-1}}{2h} + Au_n = 0.$$

For it

$$\lambda_1 = 1 - Ah + \frac{A^2h^2}{2} + O(h^3), \qquad \lambda_2 = -1 - Ah + O(h^2).$$

The "parasitic root", λ_2, for positive A is greater in modulus than the root λ_1, and it is just for this reason that a large constant appears in the stability bound for this scheme, and that the scheme (as established in §17) is not applicable for large A.

4. Generalization to systems of equations. All the above schemes for the numerical solution of the Cauchy problem for first order differential equations (1) automatically generalize to systems of first-order equations. To see this, in the notation of (1)

$$\left.\begin{aligned} \frac{du}{dx} - G(x, u) = 0, \\ u(0) = a \end{aligned}\right\}$$

we must interpret $u(x) = \bar{u}(x)$ and $G(x, u) = \bar{G}(x, \bar{u})$ as vector functions, and $a = \bar{a}$ as a given vector. In this notation, then, the Runga-Kutta schemes (3) and (4) and the Adams schemes (7)-(10), preserve their meaning and applicability.

For example the system of equations

$$\left.\begin{aligned} \frac{dv}{dx} - (x + v^2 + \sin w) = 0, \\ \frac{dw}{dx} + xvw = 0, \end{aligned}\right\}$$

$$\left.\begin{aligned} v(0) = a_1, \\ w(0) = a_2 \end{aligned}\right\}$$

may be written in the form

$$\left.\begin{aligned} \frac{d\bar{u}}{dx} - \bar{G}(x, \bar{u}) = 0, \\ \bar{u}(0) = \bar{a}, \end{aligned}\right\}$$

if we take

$$\bar{u}(x) = \begin{pmatrix} v(x) \\ w(x) \end{pmatrix},$$

$$\bar{G}(x,\bar{u}) = \begin{pmatrix} x + v^2 + \sin w \\ -xvw \end{pmatrix},$$

$$\bar{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

The equation for $\bar{u}_{n+1}$ in the Euler scheme

$$\bar{u}_{n+1} = \bar{u}_n + h\bar{G}(x_n, \bar{u}_n)$$

may be written out in detail thus:

$$\left.\begin{aligned} v_{n+1} = v_n + h(x_n + v_n^2 + \sin w_n), \\ w_{n+1} = w_n + h(-x_n v_n w_n). \end{aligned}\right\}$$

All the arguments about order of approximation, presented between the asterisks on pp. 173-175, also preserve their validity. In (6), however, we must take, as the derivative of the vector $G(G_1, \ldots, G_k)$ by the vector $u(u_1, \ldots, u_k)$, i.e. $\partial G/\partial u$, the matrix

$$\begin{pmatrix} \frac{\partial G_1}{\partial u_1} & \cdots & \frac{\partial G_1}{\partial u_k} \\ \cdots & \cdots & \cdots \\ \frac{\partial G_k}{\partial u_1} & \cdots & \frac{\partial G_k}{\partial u_k} \end{pmatrix}.$$

Any arbitrary system of differential equations, solved for the leading derivative, may be reduced to the system of first-order equations

$$\frac{d\bar{u}}{dx} = \bar{G}(x, \bar{u})$$

via changes in the dependent variables. How this can be accomplished is clear from the following example. The system

$$\left.\begin{aligned} &\frac{d^2v}{dx^2} + \sin(xv' + v^2 + w) = 0,\\ &\frac{dw}{dx} + \sqrt{x^2 + v^2 + (v')^2 + w^2} = 0,\\ &v(0) = a,\\ &v'(0) = b,\\ &w(0) = c \end{aligned}\right\}$$

will take the required form if we set

$$u_1(x) = v(x),$$

$$u_2(x) = \frac{dv}{dx},$$

$$u_3(x) = w(x).$$

We then get

$$\left.\begin{aligned} &\frac{du_1}{dx} - u_2 = 0,\\ &\frac{du_2}{dx} + \sin(xu_2 + u_1^2 + u_3) = 0,\\ &\frac{du_3}{dx} + \sqrt{x^2 + u_1^2 + u_2^2 + u_3^2} = 0,\\ &u_1(0) = a,\\ &u_2(0) = b,\\ &u_3(0) = c. \end{aligned}\right\}$$

* * * * * *

Note. Runga-Kutta difference schemes have been developed which can be applied directly to second-order equations, without preliminary reduction of these equations to systems of first order.

* * *

§ 20. Methods of solution of boundary-value problems

One example of a boundary-value problem is the problem

$$\left.\begin{aligned} &y'' = f(x, y, y'), && 0 \le x \le 1, \\ &y(0) = Y_0, && y(1) = Y_1 \end{aligned}\right\} \tag{1}$$

with boundary conditions on both sides of the interval $0 \le x \le 1$, the interval on which we must determine the solution $y = y(x)$. Using this example we will systematically develop some methods for the numerical solution of boundary-value problems.

1. The shooting method. In §19 we pointed out some convenient methods for the numerical solution of the Cauchy problem, e.g. a problem of the form

$$\left.\begin{aligned} &y'' = f(x, y, y'), && 0 \le x \le 1, \\ &y(0) = Y_0, && \left.\frac{dy}{dx}\right|_{x=0} = \tan\alpha, \end{aligned}\right\} \tag{2}$$

where Y_0 is the ordinate of the point $(0, Y_0)$ from which the integral curve emerges, while α is the angle which the integral curve makes with the Ox axis as it leaves the point $(0, Y_0)$ (Fig. 7,a). For fixed Y_0 problem (2) takes the form $y = y(x, \alpha)$. At $x = 1$ the solution $y(x, \alpha)$ depends only on α:

$$y(x, \alpha)|_{x=1} = y(1, \alpha).$$

Using what has just been said about the solution of the Cauchy problem (2), we can now reformulate problem (1) as follows: find the angle, $\alpha = \alpha^*$, such that the integral curve emerging from point $(0, Y_0)$, at an angle α from the abscissa, will arrive at the point $(1, Y_1)$:

$$y(1, \alpha) = Y_1. \tag{3}$$

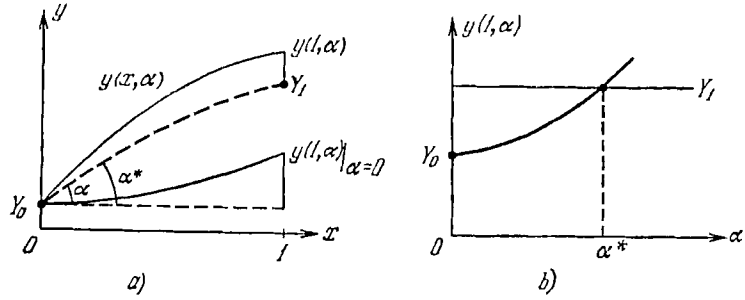


Fig. 7.

The solution of problem (2) for this $\alpha = \alpha^*$ coincides with the desired solution of problem (1). The whole problem reduces, then, to the solution of Eq. (3) (Fig. 7,b). Equation (3) is an equation of the form $F(\alpha) = 0$, where $F(\alpha) = y(1, \alpha) - Y_1$. It differs from the ordinary equation only in that the function F(a) is given, not as an analytic expression, but via an algorithm for the solution of problem (2).

Just this reduction of the process of solution of boundary-value problem (1) to the solution of Cauchy problem (2) constitutes the essential feature of the shooting method.

For the solution of (3) one may use the method of interval-halving, the chord method, the tangent method (i.e. Newton's method), etc. For example, using the method of interval-halving we find values of α_0 and α_1 such that the differences

$$y(1, \alpha_0) - Y_1 \qquad \text{and} \qquad y(1, \alpha_1) - Y_1$$

have opposite signs. We then take

$$\alpha_2 = \frac{\alpha_0 + \alpha_1}{2},$$

and compute $y(1, \alpha_2)$. Next we calculate α_3 from one of the expressions

$$\alpha_3 = \frac{\alpha_1 + \alpha_2}{2} \qquad \text{or} \qquad \alpha_3 = \frac{\alpha_0 + \alpha_2}{2}$$

depending on whether the differences

$$y(1, \alpha_2) - Y_1 \qquad \text{or} \qquad y(1, \alpha_1) - Y_1$$

respectively, have different or identical signs. Then we compute $y(1, \alpha_3)$. This process continues until the required accuracy, $|y(1, \alpha_n) - Y_1| < \varepsilon$, has been attained.

Using the chord method we would start with α_0 and α_1, computing successive α_i by the recurrence relation

$$\alpha_{n+1} = \alpha_n - \frac{F(\alpha_n)}{F(\alpha_n) - F(\alpha_{n-1})}(\alpha_n - \alpha_{n-1}), \qquad n = 1, 2, \ldots$$

The shooting method, which reduces the process of solution of boundary-value problem (1) to the computation of the solution of Cauchy problem (2), works well in cases where the solution $y(x, \alpha)$ doesn't depend "too strongly" on α. In the contrary case it becomes computationally unstable, even if the solution of problem (1) depends on the given data "reasonably".

Let us clarify what is meant by the words in quotation marks via the example of the following boundary-value problem:

$$\left.\begin{aligned} &y'' - a^2 y = 0, && 0 \le x \le 1, \\ &y(0) = Y_0, && y(1) = Y_1 \end{aligned}\right\} \qquad (1')$$

with constant a^2. This problem has the solution

$$y(x) = \frac{e^{-ax} - e^{-a(2-x)}}{1 - e^{-2a}} Y_0 + \frac{e^{-a(1-x)} - e^{-a(1+x)}}{1 - e^{-2a}} Y_1.$$

The coefficients of Y_0 and Y_1, with increasing a, remain bounded functions on the interval $0 \le x \le 1$; for all $a > 0$ they are never greater than one. Therefore small errors in the assignments of Y_0 and Y_1 lead to equally small errors in the solution. Let us now consider the Cauchy problem

$$\left.\begin{aligned} &y'' - a^2 y = 0, && 0 \le x \le 1, \\ &y(0) = Y_0, && y'(0) = \tan \alpha. \end{aligned}\right\} \qquad (2')$$

Its solution has the form

$$y(x) = \frac{aY_0 + \tan \alpha}{2a} e^{ax} + \frac{aY_0 - \tan \alpha}{2a} e^{-ax}.$$

If in fixing $\tan \alpha$ we make an error ε, then the value of the solution at $x = 1$ will increase by

$$\Delta y(1) = \frac{\varepsilon}{2a} e^{a} - \frac{\varepsilon}{2a} e^{-a}. \qquad (4)$$

For large a the subtracted term in Eq. (4) is negligibly small, but the coefficient of ε in the first term, $\exp(a)/(2a)$, becomes large. Therefore the shooting method as applied to the solution of (1'), although a formally valid procedure, for large a becomes practically unuseable. This brings to mind the considerations of 2§5, where we presented an example of a computationally unstable algorithm for the solution of a difference boundary-value problem.

2. The FEBS method. For the solution of the boundary-value problem

$$\left.\begin{aligned} &y'' - p(x)y = f(x), && 0 \le x \le 1, \\ &y(0) = Y_0, && y(1) = Y_1 \end{aligned}\right\}$$

when $p(x) \gg 1$ one can use the difference scheme

$$\left.\begin{aligned} &\frac{y_{m+1} - 2y_m + y_{m-1}}{h^2} - p(x_m)y_m = f(x_m), \\ &0 < m < M, \qquad Mh = 1, \\ &y_0 = Y_0, \qquad y_M = Y_1 \end{aligned}\right\}$$

and solve the difference problem by FEBS. If $p(x) > 0$ the conditions for applicability of FEBS are satisfied, as the reader can easily verify.

3. Newton's method. The shooting method, applied to the solutions of well-set boundary-value problems may, as we have seen, turn out to be unsuitable because of numerical instability. But the FEBS method, even formally, can be used only for the solution of linear problems.

Newton's method reduces the solution of a nonlinear problem to that of a series of linear problems, as follows. Suppose we know some function $y_0(x)$, satisfying boundary condition (1) and roughly equal to the desired solution $y(x)$. Let

$$y(x) = y_0(x) + v(x), \tag{5}$$

where v is a correction to the zeroeth approximation $y_0(x)$. We substitute (5) into Eq. (1) and linearize the problem, setting

$$y''(x) = y_0''(x) + v''(x),$$

$$f(x, y_0 + v, y_0' + v') =$$

$$= f(x, y_0, y_0') + \frac{\partial f(x, y_0, y_0')}{\partial y} v + \frac{\partial f(x, y_0, y_0')}{\partial y'} v' + O(v^2 + |v'|^2).$$

Discarding the remainder term $O(v^2 + |v'|^2)$, we get a linear problem for the correction $\bar{v}(x)$:

$$\left.\begin{aligned} &\bar{v}'' = p(x)\bar{v}' + q(x)\bar{v} + \phi(x), \\ &\bar{v}(0) = \bar{v}(1) = 0, \end{aligned}\right\} \tag{6}$$

where

$$p(x) = \frac{\partial f(x, y_0, y_0')}{\partial y'}, \qquad q(x) = \frac{\partial f(x, y_0, y_0')}{\partial y},$$

$$\phi(x) = f(x, y_0, y_0') - y_0''.$$

Solving the linear problem (6) analytically, or by some numerical method, we find an approximate correction $\bar{v}$, and take

$$y_1 \equiv y_0(x) + \bar{v}$$

as the next approximation.

The above procedure may be applied to a nonlinear difference boundary-value problem, generated as an approximation to problem (1).

This Page Intentionally Left Blank

Part 3

DIFFERENCE SCHEMES FOR PARTIAL DIFFERENTIAL EQUATIONS. BASIC CONCEPTS

Above, in connection with difference schemes for ordinary differential equations, we defined the concepts of convergence, approximation and stability. We proved a theorem stating that, if the difference boundary-value problem approximates the differential problem and is stable then, as the net is refined, the solution of the difference problem converges to the solution of the differential problem. In this theorem we have an indication as to how one can develop a convergent difference scheme for the numerical solution of a differential boundary-value problem: one must first construct approximating difference schemes and then, from among them, select those that are stable.

The definition of convergence, approximation and stability, and the theorem connecting these concepts, are general in character. They are equally meaningful for any functional equations. We illustrated them via examples of difference schemes for ordinary differential equations and for an integral equation. Here we illustrate some basic methods for constructing difference schemes, and testing their stability, taking as examples difference schemes for partial differential equations. Study of these examples will reveal many important and basically new circumstances not encountered in the case of ordinary differential equations. Principle among these are: the great variety of possible difference nets and methods of approximation, the instability of most randomly-chosen approximating schemes, the complexity of stability investigations, and the difficulties involved in the computational solution of difference boundary-value problems, difficulties which can only be overcome by substantial special effort.

Chapter 7
Simplest Examples of the Construction and Study of Difference Schemes

§21. Review and Illustrations of Basic Definitions

1. Definition of convergence. Suppose one is required to compute an approximate solution, u, of the differential boundary-value problem

$$Lu = f, \tag{1}$$

posed in some domain D with boundary Γ. One must then, for this purpose, choose a discrete set of points D_h (i.e. a net) contained in $D + \Gamma$; introduce a linear normed space, U_h, of functions defined on the net D_h; and, finally, establish a correspondence between the solution u and the function $[u]_h$ in U_h, the required table of the solution u. For the approximate computation of the table $[u]_h$, which we have agreed to treat as the exact solution of problem (1), we must, on the basis of problem (1), construct a system of equations

$$L_h u^{(h)} = f^{(h)} \tag{2}$$

for the function $u^{(h)}$ of U_h, such that we will get convergence

$$\|[u]_h - u^{(h)}\|_{U_h} \to 0 \quad \text{for} \quad h \to 0. \tag{3}$$

If the solution of the difference boundary-value problem (2) satisfies the inequality

$$\|[u]_h - u^{(h)}\|_{U_h} \leq Ch^k,$$

then we say that convergence is of order k with respect to h.

The problem of the construction of a convergent difference scheme (2) can be split into two parts: the construction of a difference-scheme (2) approximating problem (1) on the solution, u, of this latter problem, and the verification of stability of scheme (2).

2. Definition of approximation. Let us recall the definition of approximation. If this concept is to have meaning one must introduce a norm in the space, F_h, containing the right-hand side $f^{(h)}$ of Eq. (2). By definition, difference scheme (2) approximates problem (1) on the solution u if, in the equation

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}$$

the residual, $\delta f^{(h)}$, which develops when $[u]_h$ is substituted into the difference boundary-value problem (2), tends to zero as $h \to 0$

$$\|\delta f^{(h)}\|_{F_h} = \|L_h[u]_h - f^{(h)}\|_{F_h} \to 0.$$

If

$$\|\delta f^{(h)}\|_{F_h} \leq Ch^k,$$

where C does not depend on h, then the approximation is of order k with respect to h.

Let us construct, for example, for the Cauchy problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \phi(x, t), \qquad -\infty < x < \infty, \qquad 0 \le t \le T, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\} \tag{4}$$

one possible approximating difference scheme. Problem (4) can be written in form (1) if we set

$$Lu \equiv \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x}, & -\infty < x < \infty, \quad 0 \le t \le T \\ u(x, 0), & -\infty < x < \infty, \end{cases}$$

$$f = \begin{cases} \phi(x, t), & -\infty < x < \infty, \quad 0 \le t \le T. \\ \psi(x), & -\infty < x < \infty. \end{cases}$$

As the net D_h (Fig. 8) we take the set of intersection points of the lines

$$x = mh, \quad t = n\tau, \quad m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau],$$

where $h > 0$ and $\tau > 0$ are given numbers, and $[T/\tau]$ is the integral part of the fraction T/τ. We will assume that the step-size τ is connected to step-size h via the relation $\tau = rh$, where $r = \text{const}$, so that the net D_h depends only on the single parameter h. The desired net function is the table $[u]_h = \{u(mh, n\tau)\}$ of values of the solution $u(x, t)$ of problem (4) at the points of the net D_h.

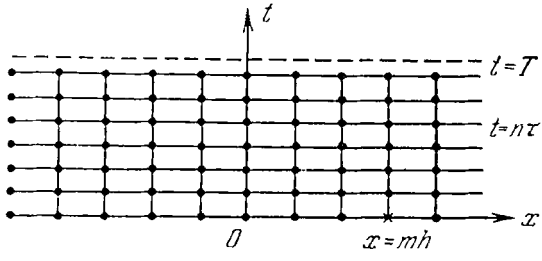


Fig. 8.

Let us now proceed to the construction of a difference scheme (2) approximating problem (4). The value of the net function $u^{(h)}$ at the point $(x_m, t_n) = (mh, n\tau)$ of net D_h will be denoted as u_m^n. We arrive at a scheme (2) by approximating the derivatives $\partial u/\partial t$ and $\partial u/\partial x$ by the difference relations

$$\left.\begin{aligned} \left.\frac{\partial u}{\partial t}\right|_{x,t} &\approx \frac{u(x,\ t+\tau) - u(x,\ t)}{\tau}, \\ \left.\frac{\partial u}{\partial x}\right|_{x,t} &\approx \frac{u(x+h,\ t) - u(x,\ t)}{h}, \end{aligned}\right\} \qquad (4')$$

This scheme has the form

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = \phi(mh,\ n\tau), \\ &m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots \end{aligned}\right\} \qquad (5)$$

The operator L_h and the right-hand side $f^{(h)}$ for scheme (5) are given, respectively, by the equations

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_m^n}{h}, \\ \qquad m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0, \qquad m = 0, \pm 1, \ldots, \end{cases}$$

$$f^{(h)} = \begin{cases} \phi(mh,\ n\tau), & m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ \psi(mh), & m = 0, \pm 1, \ldots \end{cases}$$

Thus $f^{(h)}$ consists of the pair of net functions $\phi(mh,\ n\tau)$ and $\psi(mh)$, one of which is given on the two-dimensional net

$$(x_m,\ t_n) = (mh,\ n\tau), \qquad m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau]-1$$

(see Fig. 8), and the other on the one-dimensional net

$$(x_m,\ 0) = (mh,\ 0), \qquad m = 0, 1, \ldots$$

Difference equation (4) can be solved for u_m^{n+1}, giving

$$u_m^{n+1} = (1 - r)u_m^n + ru_{m+1}^n + \tau\phi(mh,\ n\tau). \qquad (6)$$

Thus, knowing the values u_m^n, $m = 0, \pm 1, \ldots$, of the solution $u^{(h)}$ at the net-points for which $t = n\tau$, one can calculate u_m^{n+1} at the points for which

$t = (n + 1)\tau$. Since the values u_m^0 at $t = 0$ are given by the equation $u_m^0 = \psi(mh)$ we can, step by step, compute the values of the solution $u^{(h)}$ at the net-points on the lines $t = \tau$, $t = 2\tau$, etc., i.e. everywhere on D_h.

We will now go on to find the order of approximation attained by scheme (5). As F_h we can take the linear space of all pairs of bounded functions $g^{(h)} = (\phi_m^n, \psi_m)^T$, defining

$$\|g^{(h)}\|_{F_h} = \max_{m,n} |\phi_m^n| + \max_m |\psi_m|^*.$$

As has already been noted in §13, the norm used in the treatment of approximation can be chosen in many ways, and the choice is not inconsequential. At this point it will suffice to take as a norm the upper bound of the modulus of each of the components making up the elements $g^{(h)}$, of the space F_h. It is just this norm which we will use everywhere below.

Let us assume that the solution $u(x, t)$ of problem (4) has bounded second derivatives. Then by Taylor's formula

$$\left.\begin{aligned} \frac{u(x_m + h, t_n) - u(x_m, t_n)}{h} &= \frac{\partial u(x_m, t_n)}{\partial x} + \frac{h}{2}\frac{\partial^2 u(x_m + \xi, t_n)}{\partial x^2}, \\ \frac{u(x_m, t_n + \tau) - u(x_m, t_n)}{\tau} &= \frac{\partial u(x_m, t_n)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x_m, t_n + \eta)}{\partial t^2}, \end{aligned}\right\} \quad (7)$$

where ξ and η are certain numbers, depending on m, n and h, and satisfying the inequalities $0 < \xi < h$, $0 < \eta < \tau$.

With the aid of Eq. (7) the expression

$$L_h[u]_h \equiv \begin{cases} \dfrac{u(x_m\ t_n + \tau) - u(x_m, t_n)}{\tau} - \dfrac{u(x_m + h, t_n) - u(x_m, t_n)}{h} \\ u(x_m, 0) \end{cases}$$

can be rewritten in the form

$$L_h[u]_h = \begin{cases} \left(\dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x}\right)_{x_m, t_n} + \dfrac{\tau}{2}\dfrac{\partial^2 u(x_m, t_n + \eta)}{\partial t^2} - \dfrac{h}{2}\dfrac{\partial^2 u(x_m + \xi, t_n)}{\partial x^2}, \\ u(x_m, 0) + 0 \end{cases}$$

---

*) If the $\max|\phi_m^n|$ or $\max|\psi_m|$ is not attained, then we take, here, the least upper bound $\sup|\phi_m^n|$ or $\sup|\psi_m|$.

or

$$L_h[u]_h = f^{(h)} + \delta f^{(h)},$$

where

$$\delta f^{(h)} = \begin{cases} \frac{\tau}{2} \frac{\partial^2 u(x_m, t_n + \eta)}{\partial t^2} - \frac{h}{2} \frac{\partial^2 u(x_m + \xi, t_n)}{\partial x^2} \\ 0. \end{cases}$$

Therefore

$$||\delta f^{(h)}||_{F_h} \leq \left(\sup \left| \frac{\partial^2 u}{\partial t^2} \right| \cdot \frac{r}{2} + \sup \left| \frac{\partial^2 u}{\partial x^2} \right| \cdot \frac{1}{2} \right) h.$$

Thus the above difference scheme (5) has first-order approximation with respect to h on a solution, u(x, t), with bounded second derivatives.

3. Definition of stability. We now review and illustrate the definition of stability. Difference boundary-value problem (2), by definition, is stable if there exists numbers $\delta > 0$ and $h_0 > 0$ such, that for any $h < h_0$, and any $\delta f^{(h)}$ in F_h satisfying the inequality $||\delta f^{(h)}||_{F_h} < \delta$, the difference boundary-value problem

$$L_h z^{(h)} = f^{(h)} + \delta f^{(h)}$$

has one and only one solution which, moreover, fulfills the condition

$$||z^{(h)} - u^{(h)}||_{U_h} \leq C||\delta f^{(h)}||_{F_h},$$

where C is some constant, independent of h.

In §12, where the concept of stability was introduced, it was shown that, for a linear operator L_h, the above definition is equivalent to the following:

Definition. *Difference boundary-value problem (2) is stable if there exists an $h_0 > 0$ such, that for $h < h_0$ and any $f^{(h)}$ in F_h, it has a unique solution and, moreover*

$$||u^{(h)}||_{U_h} \leq C||f^{(h)}||_{F_h}, \tag{8}$$

where C is some constant not on h or on $f^{(h)}$.

The property of stability may be regarded as a uniform-in-h sensitivity of the solution of the difference boundary-value problem (2) to a perturbation $\delta f^{(h)}$ of the right hand side.

We stress that in view of the above definition stability is an *internal* property of the difference boundary-value problem. The definition is formulated independently of any connection with a differential boundary-

value problem, and in particular, with no reference to approximation or convergence.

However, *if the difference boundary-value problem approximates a differential boundary-value problem on the solution u, and the difference scheme is stable, then we have convergence, i.e. (3). Further, the order in h of the rate of convergence coincides with the order of approximation.*

The proof of this important theorem was presented in §12.

Let us now show that difference scheme (5), for $r < 1$, is stable. The norm $\|\cdot\|_{U_h}$ will be defined by the equation

$$\|u^{(h)}\|_{U_h} = \sup_{m,n} \|u_m^n\| = \max_n \sup_m |u_m^n| .$$

while the norm $\|\cdot\|_{F_h}$ will be interpreted as above: for $g^{(h)}$ in F_h,

$$g^{(h)} = \begin{cases} \phi_m^n, & m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau], \\ \psi_m, & m = 0, \pm 1, \ldots, \end{cases}$$

we take

$$\|g^{(h)}\|_{F_h} = \max_{m,n} |\phi_m^n| + \max_m |\psi_m| = \max_n \left[\max_m |\phi_m^n| + \max_m |\psi_m|\right].$$

The difference problem

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = \phi_m^n, && m = 0, \pm 1, \ldots; \\ &&& n = 0, 1, \ldots, [T/\tau], \\ &u_m^0 = \psi_m, && m = 0, \pm 1, \ldots, \end{aligned}\right\} \qquad (5')$$

which differs from problem (5) only in that ϕ_m^n and ψ_m are arbitrary right-hand sides which, generally, do not coincide with $\phi(mh, n\tau)$ and $\psi(mh)$, will now be rewritten in the form

$$\left.\begin{aligned} u_m^{n+1} &= (1 - r)u_m^n + ru_{m+1}^n + \tau\phi_m^n, \\ u_m^0 &= \psi_m. \end{aligned}\right\} \qquad (6')$$

Since $r \leq 1$, $(1 - r) \geq 0$. In this case we have the bound

$$\left|(1 - r)u_m^n + ru_{m+1}^n\right| \leq \left[(1 - r) + r\right]\max\left(\left|u_m^n\right|, \left|u_{m+1}^n\right|\right) =$$

$$= \max\left(\left|u_m^n\right|, \left|u_{m+1}^n\right|\right) \leq \max_m \left|u_m^n\right|.$$

Using this bound we derive, from (6′), the inequality

$$\left|u_m^{n+1}\right| \le \max_m \left|u_m^n\right| + \tau \max_m \left|\phi_m^n\right| \le \max_m \left|u_m^n\right| + \tau \max_{m,n} \left|\phi_m^n\right|. \qquad (6'')$$

Note that, in the case $\phi_m^n \equiv 0$, it follows from (6′′) that $\max_m \left|u_m^n\right|$ does not increase with increasing n. This property of the difference scheme is conventionally called the "maximum principle". For the sake of brevity we will sometimes use this name for the whole inequality

$$\left|u_m^{n+1}\right| \le \max_m \left|u_m^n\right| + \tau \max_{m,n} \left|\phi_m^n\right| .$$

The right-hand side of this inequality does not depend on m, so that on the left-hand side one may write $\max_m \left|u_m^{n+1}\right|$, in place of $\left|u_m^{n+1}\right|$, thus arriving at the inequality

$$\max_m \left|u_m^{n+1}\right| \le \max_m \left|u_m^n\right| + \tau \max_{m,n} \left|\phi_m^n\right| .$$

Similarly we get the inequalities

$$\max_m \left|u_m^n\right| \le \max_m \left|u_m^{n-1}\right| + \tau \max_{m,n} \left|\phi_m^n\right|,$$

. . . . . .

$$\max_m \left|u_m^1\right| \le \max_m \left|u_m^0\right| + \tau \max_{m,n} \left|\phi_m^n\right| .$$

Adding these inequalities term by term, and finally combining like terms, we get

$$\max_m \left|u_m^{n+1}\right| \le \max_m \left|u_m^0\right| + (n+1)\tau \max_{m,n} \left|\phi_m^n\right| .$$

from which immediately follows

$$\max_m \left|u_n^{n+1}\right| \le \max_m \left|\psi_m\right| + T \max_{m,n} \left|\phi_m^n\right| \le$$

$$\le \|f^{(h)}\|_{F_h} + T\|f^{(h)}\|_{F_h} = (1+T)\|f^{(h)}\|_{F_h} .$$

The inequality we have just derived

$$\max_m \left|u_m^{n+1}\right| \le (1+T)\|f^{(h)}\|_{F_h}$$

is valid for all n, so that it remains valid if, in place of $\max_m \left|u_m^{n+1}\right|$, we write $\max_n \max_m \left|u_m^n\right| = \|u^{(h)}\|_{U_h}$:

$$||u^{(h)}||_{U_h} \leq (1 + T)||f^{(h)}||_{F_h}. \qquad (9)$$

This inequality, (9), implies the stability of linear problem (5), since, obviously, the solution of (6′) for arbitrary bounded ϕ_m^n and ψ_m, exists and is unique. The role of the constant C in inequality (8) is taken on, here, by the number 1 + T.

One must not think that, in itself, approximation of the differential boundary-value problem (1) by difference boundary-value problem (2) guarantees the stability, and therefore the convergence, of (3). We convinced ourselves of this in §9 with the aid of a specially constructed example of an approximating, but divergent, difference scheme.

In the case of partial differential equations failure of randomly chosen approximating difference schemes is the rule, and the choice of a stable (and therefore convergent) difference scheme is the constant concern of the computations specialist.

We recall, for example, that the proof of the stability of difference scheme (5) was carried out under the assumption that $\tau/h \equiv r \leq 1$. In the case $r > 1$ the difference problem (5) still approximates (4), but our stability proof fails. We now show that in this case the solution, $u^{(h)}$, of the difference problem (5) does not converge to the solution, u(x, t), of the differential problem (4), which means that the difference scheme cannot be stable since stability would imply convergence.

Suppose, for the sake of definiteness, that $\phi(x, t) \equiv 0$, so that also $\phi(mh, n\tau) = 0$; further, let $T = 1$. The step-size h will be chosen such, that the point (0, 1) in the plane Oxt belongs to the net, i.e. such that the number

$$N = \frac{1}{\tau} = \frac{1}{rh}$$

will be an integer (Fig. 9). From the difference equation we get

$$u_m^{n+1} = (1 - r)u_m^n + ru_{m+1}^n.$$

The value $u_0^{n+1} = u_0^N$ of the solution $u^{(h)}$ at the point (0, 1) of the net is expressed, via the difference equation, in terms of the values u_0^n and u_1^n of the solution at the points (0, 1-τ) and (h, 1-τ) of the net. The two values u_0^n and u_1^n, are expressed in terms of the values u_0^{n-1}, u_1^{n-1} and u_2^{n-1} of the solution

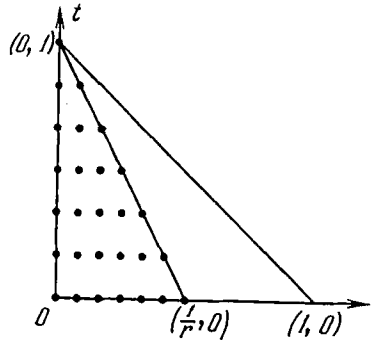


Fig. 9.

at the three net-points $(0, 1-2\tau)$, $(h, 1-2\tau)$ and $(2h, 1-2\tau)$. The values of the solution u_0^{n-1}, u_1^{n-1} and u_2^{n-1}, in turn, are given in terms of the solution-values at the four points $(0, 1-3\tau)$, $(h, 1-3\tau)$, $(2h, 1-3\tau)$ and $(3h, 1-3\tau)$, etc. Finally the value u_0^{n+1} may be expressed in terms of the values, u_m^0, of the solution at the net-points $(0, 0)$, $(h, 0)$, $(2h, 0)$, ..., $(h/\tau, 0) = (Nh, 0)$. All these points lie on the interval

$$0 \leq x \leq \frac{h}{\tau} = \frac{1}{r}$$

of the line $t = 0$ (see Fig. 9), where we are given the initial condition

$$u(x, 0) = \psi(x)$$

for the differential equation. Thus the solution of the difference equation at the point $(0, 1)$ of the net does not depend on the values of the function $\psi(x)$ at points, x, lying outside the interval

$$0 \leq x \leq \frac{1}{r}.$$

Further, the solution of the problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \qquad -\infty < x < \infty, \quad t > 0, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\}$$

as one can easily verify, is the function

$$u(x, t) \equiv \psi(x + t).$$

This function is constant on each characteristic $x + t = \text{const}$; and, in particular, on the line $x + t = 1$, which passes through the points $(0, 1)$ and $(1, 0)$ (see Fig. 9). At the point $(1, 0)$ it takes on the value $\psi(1)$. Thus it is clear that, in the case $r > 1$, convergence, generally, cannot occur. In fact in this case the segment of the axis with abcissas

$$0 \leq x \leq \frac{1}{r} < 1$$

does not contain the point $(1, 0)$. If, for some value of the function $\psi(x)$, convergence were to take place accidentally then, without changing the value of $\psi(x)$ on the interval

$$0 \leq x \leq \frac{1}{r}$$

and, thus, not changing the solution of the difference equation at the point $(0, 1)$, we could eliminate convergence by altering $\psi(x)$ at and near

the point $x = 1$, a change which would, in turn, change the value, $u(0, 1) = \psi(1)$, of the solution of the differential equation. The change in $\psi(x)$ at and near $x = 1$ could be managed in such a way as not to negate the existence of second derivatives of the function $\psi(x)$, or of the solution $u(x, t) = \psi(x + t)$, so that approximation on the solution $u(x, t)$ remains in effect. Under these conditions stability of scheme (5) would imply convergence. But since for $r > 1$ we cannot have convergence, we cannot have stability either.

The proof we have given of the instability of difference scheme (5) is indirect in character. It is interesting to examine directly how the instability of difference scheme (5) for $r > 1$ is reflected in the sensitivity of the solution, $u^{(h)}$, to errors in the specification of $f^{(h)}$. After all, it is precisely the uniformity, with respect to h, of the sensitivity of the solution to errors in $f^{(h)}$ which was defined, above, as stability.

Suppose that, identically for all h, $\phi(mh, n\tau) \equiv 0$ and $\psi(mh) \equiv 0$, so that

$$f^{(h)} = \begin{Bmatrix} \phi_m^n \\ \psi_m \end{Bmatrix} = 0$$

and the solution $u^{(h)} = \{u_m^n\}$ of problem (5) is identically zero, $u_m^n \equiv 0$. Suppose, further, that, in specifying initial conditions an error has occurred so that, instead of $\psi_m = 0$, we are given $\tilde{\psi}_m = (-1)^m\varepsilon$, $\varepsilon = \text{const}$, and instead of

$$f^{(h)} = \begin{Bmatrix} \phi_m^n \\ \psi_m \end{Bmatrix} = 0$$

we have

$$\tilde{f}^{(h)} = \begin{Bmatrix} 0 \\ \tilde{\psi}_m \end{Bmatrix}, \qquad \|\tilde{f}^{(h)}\|_{F_h} = \varepsilon.$$

We will call the resulting solution $\tilde{u}^{(h)}$. From the equations

$$\tilde{u}_m^{n+1} = (1 - r)\tilde{u}_m^n + r\tilde{u}_{m+1}^n,$$

$$\tilde{u}_m^0 = (-1)^m\varepsilon$$

we get, for $\tilde{u}_m^1$,

$$\tilde{u}_m^1 = (1 - r)\tilde{u}_m^0 + r\tilde{u}_{m+1}^0 =$$

$$= (1 - r)(-1)^m\varepsilon + r(-1)^{m+1}\varepsilon = (1 - 2r)(-1)^m\varepsilon = (1 - 2r)\tilde{u}_m^0.$$

We see that the error committed at $n = 0$ has been multiplied by $(1 - 2r)$. On proceeding to $\tilde{u}_m^2$ we get

$$\tilde{u}^2_m = (1 - r)\tilde{u}^1_m + r\tilde{u}^1_{m+1} = (1 - 2r)\tilde{u}^1_m = (1 - 2r)^2\,\tilde{u}^0_m .$$

In general

$$\tilde{u}^n_m = (1 - 2r)^n\,\tilde{u}^0_m = (1 - 2r)^n(-1)^m\,\varepsilon.$$

For $r > 1$ we have $1 - 2r < -1$, so that the error

$$\tilde{u}^0_m = (-1)^m\,\varepsilon$$

on stepping from one level $t = n\tau$ of the net to the next, is multiplied by a negative number exceeding one in modulus. For $n = [T/\tau]$

$$\left|\tilde{u}^n_m\right| = |1 - 2r|^{[T/\tau]}\left|\tilde{u}^0_m\right|$$

so that

$$||\tilde{u}^{(h)}||_{U_h} = |1 - 2r|^{[T/(rh)]}\left|\tilde{u}^0_m\right| = |1 - 2r|^{[T/(rh)]}\max|\tilde{\psi}_m| =$$

$$= |1 - 2r|^{[T/(rh)]}||\tilde{f}^{(h)}||_{F_h}.$$

In a fixed time, T, an error $(-1)^m\varepsilon$ in initial values increases by the factor $|1 - 2r|^{[T/(rh)]}$, a factor which grows very rapidly as $h \to 0$.

We pause now for a brief critique of the method by which we have chosen to evaluate the quality of approximation; i.e, a method based on a comparison of the norm of the residual $||\delta f^{(h)}||$, with this or that power of h. As we know, for stable schemes the order of approximation coincides with the order of the error, $[u]_h - u^{(h)}$, in the solution. It is natural to judge the quality of a scheme by the amount of computational effort which is required for the attainment of a given accuracy. This amount of computational effort, generally speaking, is proportional to the number of points, N, used in the difference net. For ordinary differential equations N is inversely proportional to the step-width, h. Therefore, when we say that the error $\varepsilon \approx h^p$ we are, at the same time, asserting that $\varepsilon \approx 1/N^p$, i.e. that halving the error will require that we increase the expended effort by a factor $\sqrt[p]{2}$. Thus, in the case of ordinary differential equations, the order of approximation with respect to h characterizes the volume of computational effort.

For partial differential equations the situation is different. In the above example of a problem in two variables, x and t, the net is specified by the two step-sizes τ and h. The number, N, of net-points, located in a bounded region of the plane Oxt is of order $1/(\tau h)$. This number also can be taken as a measure of the amount of work expended in solving the

difference equations. Suppose $\tau = rh$. In this case $N \approx 1/h^2$ and the assertion that $\varepsilon \approx h^p$ is equivalent to the statement that $\varepsilon \approx N^{p/2}$. IF $\tau = rh^2$, then $N \approx 1/h^3$ and the assertion that $\varepsilon \approx h^p$ is equivalent to $\varepsilon \approx 1/N^{p/3}$.

We see that, in the case of partial differential equations, it would be more natural to measure the order of the error, not in powers of h, but in powers of 1/N. We will, nevertheless, settle on the method described above, in which approximation is evaluated in powers of h, since this is more convenient for computational purposes. The reader should, however, in judging the quality of difference schemes, keep in mind the above considerations.

We must note, further, that the assertion that the computational work is proportional to the number, N, of net-points is also not always true. One can cite examples of difference schemes whose use requires, in the solution process, $\approx N^{1+q}$ arithmetic operations, where $q = 1/2$ or even 2. One encounters such schemes in the solution of difference boundary-value problems approximating elliptic equations, or in solving problems in three or more independent variables (e.g., $u = u(t, x, y)$). In the multidimensional case the construction of difference schemes such that the solution process entails $\approx N$ arithmetic operations is a nontrivial problem, about which more will be said in §§31, 32.

For real calculations on electronic computers it is common to take machine time as a measure of quality, for the purpose of comparing algorithms. Machine time is not necessarily proportional to the number of arithmetic operations.

The time required to transfer information from one block of computer memory to another may also play a significant, sometimes even a predominant role. And the time expended on logical operations must also be considered.

PROBLEMS

1. For Cauchy problem (4) study the following difference scheme:

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_m^n - u_{m-1}^n}{h} = \phi(mh, n\tau), \\ &m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

where $\tau = rh$, $r = \text{const}$. More precisely

a) Write out in detail the operator, L_h, and right-hand side $f^{(h)}$, which appear when this scheme is put into the form $L_h u^{(h)} = f^{(h)}$.

b) Sketch the relative locations of three net-points, such that the values $u^{(h)}$ at these points are connected by the difference equation for fixed m and n.

c) Show that the difference scheme approximates the differential problem to first order in h on a solution, u(x, t), having bounded second derivatives.

d) Determine whether the difference scheme in question is stable for some choice of r, $\tau = rh$.

2. For the Cauchy problem $u_t + u_x = \phi(x,t)$, $u(x, 0) = \psi(x)$, $-\infty < x < \infty$, $0 \le t \le T$, investigate, following the outline laid out in problem 1, above, each of the following difference schemes:

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{u_m^n - u_{m-1}^n}{h} = \phi(mh, n\tau), \\ &m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots; \end{aligned}\right\}$$

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{u_{m+1}^n - u_m^n}{h} = \phi(mh, n\tau), \\ &m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots; \end{aligned}\right\}$$

§22. Simplest methods for the construction of approximating difference schemes

1. Replacement of derivatives by difference relations. The simplest method for the construction of difference boundary-value problems, approximating differential boundary-value problems, consists in the replacement of derivatives by corresponding difference relations. We will present several examples of difference schemes obtained in this way. In these examples we will use the approximate expressions

$$\left.\begin{aligned} &\frac{df(z)}{dz} \approx \frac{f(z + \Delta z) - f(z)}{\Delta z}, \\ &\frac{df(z)}{dz} \approx \frac{f(z) - f(z - \Delta z)}{\Delta z}, \\ &\frac{df(z)}{dz} \approx \frac{f(z + \Delta z) - f(z - \Delta z)}{2\Delta z}, \\ &\frac{d^2 f(z)}{dz^2} \approx \frac{f(z + \Delta z) - 2f(z) + f(z + \Delta z)}{\Delta z^2}. \end{aligned}\right\} \quad (1)$$

Assuming a function $f(z)$ having sufficiently many bounded derivatives, it is possible to write out expressions for the remainder terms in these approximations. By Taylor's formula

$$\left.\begin{aligned} f(z+\Delta z) &= f(z) + \Delta z f'(z) + \frac{(\Delta z)^2}{2!} f''(z) + \\ &\quad + \frac{(\Delta z)^3}{3!} f'''(z) + \frac{(\Delta z)^4}{4!} f^{(4)}(z) + o[(\Delta z)^4], \\ f(z-\Delta z) &= f(z) - \Delta z f'(z) + \frac{(\Delta z)^2}{2!} f''(z) - \\ &\quad - \frac{(\Delta z)^3}{3!} f'''(z) + \frac{(\Delta z)^4}{4!} f^{(4)}(z) + o[(\Delta z)^4]. \end{aligned}\right\} \quad (2)$$

Using expansions (2), one can get expressions for the remainder terms in the approximate Eqs. (1). Specifically, one finds that

$$\left.\begin{aligned} \frac{f(z+\Delta z) - f(z)}{\Delta z} &= f'(z) + \left[\frac{\Delta z}{2} f''(z) + o(\Delta z)\right], \\ \frac{f(z) - f(z-\Delta z)}{\Delta z} &= f'(z) + \left[-\frac{\Delta z}{2} f''(z) + o(\Delta z)\right], \\ \frac{f(z+\Delta z) - f(z-\Delta z)}{2\Delta z} &= f'(z) + \left[\frac{(\Delta z)^2}{3} f'''(z) + o\left((\Delta z)^2\right)\right], \\ \frac{f(z+\Delta z) - 2f(z) + f(z-\Delta z)}{\Delta z^2} &= f''(z) + \left[\frac{(\Delta z)^2}{12} f^{(4)}(z + o\left((\Delta z)^2\right)\right]. \end{aligned}\right\} \quad (3)$$

The remainder terms in the approximations (1) enter into the corresponding Eqs. (3) in the form of the expressions in square brackets.

Clearly Eqs. (1), as well as the expressions for the remainder terms written out explicitly in (3), can also be used to replace partial derivatives by difference relations. For example

$$\frac{\partial u(x, t)}{\partial t} \approx \frac{u(x, t+\Delta t) - u(x, t)}{\Delta t},$$

since

$$\frac{u(x, t+\Delta t) - u(x, t)}{\Delta t} = \frac{\partial u(x, t)}{\partial t} + \left[\frac{\Delta t}{2} \frac{\partial^2 u(x, t)}{\partial t^2} + o(\Delta t)\right].$$

Equally

$$\frac{\partial u(x,\ t)}{\partial x} \approx \frac{u(x + \Delta x,\ t) - u(x,\ t)}{\Delta x}$$

and, in this case,

$$\frac{u(x + \Delta x,\ t) - u(x,\ t)}{\Delta x} = \frac{\partial u(x,\ t)}{\partial x} + \left[\frac{\Delta x}{2}\frac{\partial^2 u(x,\ t)}{\partial x^2} + o(\Delta x)\right]$$

etc.

Example 1. We return, here, to Cauchy problem (4) of §21:

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \phi(x,\ t), \qquad -\infty < x < \infty, \quad 0 \leq t \leq T,\\ &u(x,\ 0) = \psi(x), \qquad -\infty < x < \infty. \end{aligned}\right\} \tag{4}$$

To approximate this Cauchy problem we construct three schemes. In all these schemes we use the net, D_h, formed by those points of intersection of the lines $x = mh$, $t = n\tau$, falling inside the strip $0 \leq t \leq T$. The values of τ and h we take to be connected by the relation $\tau = rh$, where r is some positive constant. The simplest of these schemes has the form of (5) §21:

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_n^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_m^n}{h} = \phi(mh,\ n\tau), \\ u_m^0 = \psi(mh), \end{cases} \tag{5}$$

and is obtained by replacing the derivatives $u_t = \partial u/\partial t$ and $u_x = \partial u/\partial x$ by the approximate expressions

$$u_t(x,\ t) \approx \frac{u(x,\ t + \tau) - u(x,\ t)}{\tau},$$

$$u_x(x,\ t) \approx \frac{u(x + h,\ t) - u(x,\ t)}{h}.$$

We have studied this scheme in detail in §21. In this case the residual, $\delta f^{(h)}$, which develops when the solution, $[u]_h$, of the differential problem is substituted into the left-hand side of the difference problem

$$L_h[u]_h = f^{(h)} + \delta f^{(h)},$$

has the form

$$\delta f^{(h)} = \begin{cases} \left(\dfrac{\tau}{2} u_{tt} - \dfrac{h}{2} u_{xx}\right)_m^n + o(\tau + h), \\ 0. \end{cases}$$

In this section we take, as the norm of the element $f^{(h)}$ of space F_h, the maximum of all components of this element. Then, obviously

$$||\delta f^{(h)}||_{F_h} = O(\tau + h) = O(rh + h) = O(h),$$

and the approximation turns out to be of first order.

The second scheme results from the substitution of another expression for $\partial u/\partial x$:

$$\frac{\partial u(x, t)}{\partial x} \approx \frac{u(x, t) - u(x - h, t)}{h},$$

This scheme has the form

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_m^n - u_{m-1}^n}{h} - \phi(mh, n\tau), \\ u_m^0 = \psi(mh). \end{cases}$$

Here

$$\delta f^{(h)} \equiv \begin{cases} \left(\dfrac{\tau}{2} u_{tt} + \dfrac{h}{2} u_{xx}\right)_m^n + o(\tau + h), \\ 0, \end{cases}$$

$$||\delta f^{(h)}||_{F_h} = O(h),$$

and approximation again turns out to be first order.

The second scheme, it would seem, differs only insignificantly from the first. Below we will see, however, that this second scheme is completely unsuitable for computation: it is unstable for any $\tau/h = r = \text{const}$.

The third scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - \dfrac{u_{m+1}^n + u_{m-1}^n}{2}}{\tau} - \dfrac{u_{m+1}^n - u_m^n}{h} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh), \end{cases}$$

is obtained by replacement of the derivatives by difference relations via the approximate expressions

$$\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x,\ t+\tau) - \dfrac{u(x+h,\ t) + u(x-h,t)}{2}}{\tau},$$

$$\frac{\partial u(x,\ t)}{\partial x} \approx \frac{u(x+h,\ t) - u(x,\ t)}{h}.$$

With the aid of the Taylor expansions (2), for a sufficiently smooth solution, $u(x, t)$, of problem (1) we get

$$\frac{u(x,\ t+\tau) - \dfrac{u(x+h,\ t) + u(x-h,\ t)}{2}}{\tau} - \frac{u(x+h,\ t) - u(x,\ t)}{h} =$$

$$= \left[\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} - \frac{h^2}{2\tau}\frac{\partial^2 u}{\partial x^2} + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}\right]_{x,t} + O\left(\tau^2 + h^2 + \frac{h^4}{\tau}\right) =$$

$$= \phi(x,\ t) + \left[-\frac{h}{2r}\,u_{xx} + \frac{\tau}{2}\,u_{tt} + O(h^2)\right]_{x,t}$$

Therefore

$$L_h[u]_h = \begin{cases} \phi(mh,\ nh) + \left[-\dfrac{h}{2r}\,u_{xx} + \dfrac{\tau}{2}\,u_{tt} + O(h^2)\right], \\ \psi(mh) + 0, \end{cases}$$

so that $\delta f^{(h)}$ in the equation

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}$$

has the form

$$\delta f^{(h)} \equiv \begin{cases} -\dfrac{h}{2r}\,u_{xx} + \dfrac{\tau}{2}\,u_{tt} + O(h^2), \\ 0. \end{cases}$$

Thus $\|\delta f^{(h)}\|_{F_h} = O(h)$ and we again have first order approximation, as in the two first examples.

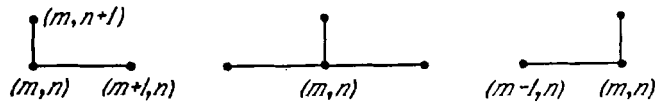


Fig. 10.

Let us now consider the case where the connection between the mesh widths is given, not by the relation $\tau = rh$ as above, but by the equation

$$\tau = rh^2, \qquad r = \text{const},$$

presupposing a more rapid refinement in τ than in h. In this case

$$L_h[u]_h = \begin{cases} \left[\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} - \frac{1}{2r}\frac{\partial^2 u}{\partial x^2}\right]_{x_m, t_n} + O(h^2), \\ u(mh, n\tau), \end{cases}$$

from which it is clear that the above difference scheme approximates the problem

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} - \frac{1}{2r}\frac{\partial^2 u}{\partial x^2} = \phi(x, t),$$

$$u(x, 0) = \psi(x),$$

not at all the same as the Cauchy problem (4) which we set out to approximate.

We have, thus, stumbled onto the fact that one and the same difference scheme may, for different functional relations $\tau = \tau(h)$, approximate different differential problem as $h \to 0$. Such difference schemes are called "rigid".

For heuristic purposes it is common to associate a difference scheme with a sketch (or "stencil") representing the relative positions of the net points at which (for some fixed m and n) solution values are directly connected by the difference equations. For the above three schemes these sketches are displayed in Fig. 10.

Example 2. We now present two difference schemes approximating the Cauchy problem for the heat equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial u^2} = \phi(x, t), \qquad -\infty < x < \infty, \quad 0 < t < T,$$

$$u(x, 0) = \psi(x), \qquad -\infty < x < \infty.$$

The simplest of these

$$L_h^{(1)} u^{(h)} \equiv \begin{cases} \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh), \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \phi(mh, n\tau), \\ \psi(mh), \end{cases}$$

is obtained by replacement of the derivatives u_t and u_{xx} by difference relations via the equations

$$u_t(x, t) \approx \frac{u(x, t + \tau) - u(x, t)}{\tau},$$

$$u_{xx}(x, t) \approx \frac{u(x + h, t) - 2u(x, t) + u(x - h, t)}{h^2}.$$

If, for the replacement of $u_{xx}(x, t)$, one were to use another expression:

$$u_{xx}(x, t) \approx \frac{u(x + h, t + \tau) - 2u(x, t + \tau) + u(x - h, t + \tau)}{h^2},$$

one would arrive at a different scheme for the same equation:

$$L_h^{(2)}u(h) \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh). \end{cases}$$

To distinguish the two operators L_h of these two schemes we have numbered them, writing $L_h^{(1)}u^{(h)} = f^{(h)}$ and $L_h^{(2)}u^{(h)} = f^{(h)}$. The stencils corresponding to both difference schemes are shown in Fig. 11.

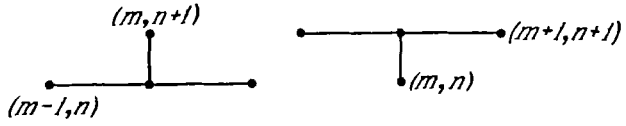


Fig. 11.

These schemes are basically different. Computation of the solution by the first scheme presents no difficulties, and is carried out by use of the explicit relation

$$u_m^{n+1} = (1 - 2r)u_m^n + r(u_m^n + u_{m+1}^n) + \tau\phi(mh, n\tau),$$

where $r = \tau/h^2$. This expression is obtained from the difference equation by solving for u_m^{n+1}. Knowing the value of the solution, u_m^n, $m = 0, \pm 1, \ldots$, at the level $t = t_n (= n\tau)$ of the net, we can compute its value u_m^{n+1} at the next level $t = t_{n+1}$.

In the second scheme $L_h^{(2)}u^{(h)} = f^{(h)}$ this convenient property has been lost. For this reason the scheme is said to be "implicit". In this case the difference equation, written for fixed m and n, cannot be solved explicitly for u_m^{n+1}, expressing this quantity in terms of the known values of u_{m+1}^n, u_m^n, u_{m-1}^n, from the preceding level. The problem is that this equation contains not only the unknown u_m^{n+1}, but also the other unknowns

u_{m-1}^{n+1} and u_{m+1}^{n+1}. Therefore, to determine u_m^{n+1}, $m = 0, \pm 1, \ldots$, it is necessary to solve the difference equation for the whole net function, u_m^{n+1}, of the argument m. Nevertheless it will be shown below that the scheme $L_h^{(2)} u^{(h)} = f^{(h)}$ is, as a rule, more convenient than the scheme $L_h^{(1)} u^{(h)} = f^{(h)}$.

For $\tau = rh^2$, $r = \text{const}$, both schemes have second order approximation with respect to h. We calculate the residual $\delta f^{(h)}$ and evaluate the order of approximation of the second scheme. Using Eq. (3) one can write

$$L_h^{(2)}[u]_h \equiv$$

$$\equiv \begin{cases} \left(u_t - u_{xx}\right)_{\substack{x=mh \\ t=(n+1)\tau}} - \dfrac{\tau}{2} u_{tt}(x, t_{n+1}) - \dfrac{h^2}{12} u_{xxxx}(x, t_{n+1}) + o(\tau + h^2), \\ u(mh, 0). \end{cases}$$

It follows, since $\tau = rh^2$, that

$$L_h^{(2)}[u]_h \equiv \begin{cases} \phi(x_m, t_{n+1}) + O(h^2), \\ \psi(mh) + 0, \end{cases}$$

$$\delta f^{(h)} = \begin{cases} \phi(x_m, t_n) - \phi(x_m, t_{n+1}) + O(h^2), \\ 0. \end{cases}$$

But

$$\phi(x_m, t_{n+1}) = \phi(x_m, t_n) + \left[\phi(x_m, t_{n+1}) - \phi(x_m, t_n)\right] =$$

$$= \phi(x_m, t_n) + O(\tau) = \phi(x_m, t_n) + O(h^2).$$

Therefore

$$\|\delta f^{(h)}\|_{F_h} = O(h^2).$$

Example 3. We now consider the simplest difference scheme approximating the Dirichlet problem for Poisson's equation in the square D ($0 < x < 1$, $0 < y < 1$) with boundary Γ (Fig. 12,a):

$$u_{xx} + u_{yy} = \phi(x, y), \qquad (x, y) \text{ in } D,$$

$$u|_\Gamma = \psi(x,y), \qquad (x, y) \text{ in } \Gamma.$$

We construct the net, D_h, assigning to it those points $(x_m, t_n) = (mh, nh)$, which fall inside the square or on its boundary. The step-width,

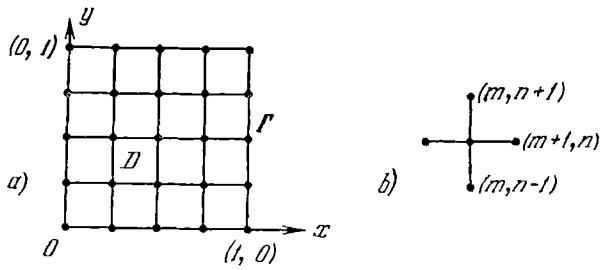


Fig. 12.

h, we assume to be chosen such that 1/h is an integer and the difference scheme, $L_h u^{(h)} = f^{(h)}$ will be given by the equations

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2} + \dfrac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2} = \\ \qquad = \phi(mh, nh), \qquad (mh, nh) \text{ in } D, \\ u_{mn} = \psi(mh, nh), \qquad (mn, nh) \text{ in } \Gamma \end{cases}$$

$$f^{(h)} = \begin{cases} \phi(mh, nh), & \text{if } (mh, nh) \text{ in } D, \\ \psi(mh, nh), & \text{if } (mh, nh) \text{ in } \Gamma. \end{cases}$$

By virtue of Eq. (3) the residual $\delta f^{(h)}$, $L_h[u]_h = f^{(h)} + \delta f^{(h)}$, has the form

$$\delta f^{(h)} = \begin{cases} \dfrac{h^2}{12}\left(u_{xxxx} + u_{yyyy}\right)\Big|_{x_m, y_n} + o(h^2), \\ 0, \end{cases}$$

so that approximation is of second order. The five-point stencil, corresponding to the given difference equation, is pictured in Fig. 12,b.

We have, above, constructed difference schemes by replacing each derivative in the differential equation by a difference relation of one sort or another.

2. The method of undetermined coefficients. A more general method of constructing difference schemes is, not to replace each derivative separately, but to replace the whole differential operator at once. We explain this method by way of examples of difference schemes for the Cauchy problem (4). First we consider a first-order approximation, scheme (5). This scheme connects the values of the required function at three points, as shown in the left-hand panel of Fig. 10. The difference equation

$$\Lambda_h u^{(h)} \equiv \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = \phi(mh, n\tau),$$

used in this scheme has the form

$$\Lambda_h u^{(h)} \equiv a^0 u_m^{n+1} + a_0 u_m^n + a_1 u_{m+1}^n = \phi(mh, n\tau).$$

Let us forget, for the moment, that we already know about difference scheme (5), for which

$$a^0 = \frac{1}{\tau}, \qquad a_0 = \frac{1}{h} - \frac{1}{\tau}, \qquad a_1 = -\frac{1}{h},$$

and, considering these coefficients undetermined, try to choose them in such a way that

$$\Lambda_h [u]_h \Big|_{\substack{x=mh,\\ t=n\tau}} = \left(\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}\right)\Big|_{\substack{x=mh,\\ t=n\tau}} + O(h)$$

or

$$\Lambda_h [u]_h \Big|_{\substack{x=mh,\\ t=n\tau}} = \Lambda u \Big|_{\substack{x=mh,\\ t=n\tau}} + O(h), \tag{6}$$

where

$$\Lambda u \equiv \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}. \tag{7}$$

For this purpose we make use of Taylor's formula:

$$u[mh, (n+1)\tau] = u(mh, n\tau) + \tau u_t'(mh, n\tau) + O(\tau^2),$$

$$u[(m+1)h, n\tau] = u(mh, n\tau) + h u_x'(mh, n\tau) + O(h^2).$$

Substituting this expression into the right-hand side of the equation

$$\Lambda_h [u]_h \Big|_{\substack{x=mh,\\ t=n\tau}} \equiv a^0 u[mh, (n+1)\tau] + a_0 u(mh, n\tau) + a_1 u[(m+1)h, n\tau]$$

we get

$$\Lambda_h [u]_h \Big|_{\substack{x=mh,\\ t=n\tau}} = (a^0 + a_0 + a_1) u(mh, n\tau) +$$

$$+ a^0 \tau \frac{\partial u(mh, n\tau)}{\partial t} + a_1 h \frac{\partial u(mh, n\tau)}{\partial x} + O(a^0\tau^2, a_1 h^2). \tag{8}$$

Since it is our goal to choose the coefficients a^0, a_0 and a_1 so as to fulfill the condition of approximation (6), it is natural, preliminarily,

to group terms on the right-hand side of Eq. (8) in such a way as to separate out term (7). Then the remaining terms will constitute the remainder term of the approximation, a term which must be small. To single out the term Λu one may replace, in the right-hand side of (8), the derivatives $\partial u/\partial t$ or $\partial u/\partial x$ using, respectively, one of the two relations

$$\frac{\partial u}{\partial t} \equiv \Lambda u + \frac{\partial u}{\partial x} \qquad \text{or} \qquad \frac{\partial u}{\partial x} \equiv \frac{\partial u}{\partial t} - \Lambda u.$$

For the sake of definiteness we use the first of these.

In addition we connect the step-widths τ and h by the relation $\tau = rh$, with some constant r. After these manipulations Eq. (8) takes the following form:

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = a^0 rh\Lambda u\Big|_{\substack{x=mh,\\ t=n\tau}} + (a^0 + a_0 + a_1)u(mh, n\tau) +$$

$$+ (a^0 r + a_1)hu_x(mh, n\tau) + O(a^0 r^2 h^2, a_1 h^2). \tag{9}$$

Among all smooth functions $u(x, t)$ one can find a subset for which u, $\partial u/\partial x$ and $\partial u/\partial t$ will take on, at any prescribed point, any mutually independent values. Therefore the quantities

$$u, \frac{\partial u}{\partial x} \quad \text{and} \quad \Lambda u = \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \phi(x, t)$$

also may be considered mutually independent. In view of this fact, it follows from (9) that if, for any right hand $\phi(x, t)$ side of problem (4), we are to fulfill the approximation condition

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = (\Lambda u)_{\substack{x=mh,\\ t=n\tau}} + O(h)$$

it is necessary that

$$a^0 rh = 1 + O_1(h),$$

$$a^0 + a_0 + a_1 = 0 + O_2(h),$$

$$(a^0 r + a_1)h = 0 + O_3(h),$$

where $O_1(h)$, $O_2(h)$ and $O_3(h)$ are some arbitrary quantities of order h. Suppose that $O_1(h) = O_2(h) = O_3(h) = 0$. The resulting system

$$a^0 rh = 1,$$

$$a^0 + a_0 + a_1 = 0,$$

$$a^0 r + a_1 = 0$$

has the unique solution

$$a^0 = \frac{1}{rh} = \frac{1}{\tau},$$

$$a_0 = \frac{r - 1}{rh} = \frac{1}{h} - \frac{1}{\tau},$$

$$a_1 = -\frac{1}{h},$$

which takes us back to the already familiar scheme (5).

Now, however, we have learned that, among difference schemes of the form

$$L_h u^{(h)} \equiv \begin{cases} a^0 u_m^{n+1} + a_0 u_m^n + a_1 u_{m+1}^n = \phi(mh, n\tau), \\ u_m^0 = \psi(mh) \end{cases}$$

this is the only one approximating the given Cauchy problem. In considering uniqueness we neglect the degree of arbitrariness resulting from the free choice of the functions $0_1(h)$, $0_2(h)$ and $0_3(h)$. Everywhere in the examples below we will also neglect a similar sort of obvious arbitrariness and, in fact, will not always introduce, explicitly, arbitrary quantities analogous to $0_1(h)$, $0_2(h)$ and $0_3(h)$, assuming from the start that they are zero.

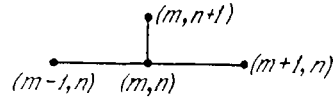


Fig. 13

The reader will easily convince himself that, in the present example, the introduction of these quantities would have lead to the following insignificant change in results

$$a^0 = \frac{1}{h}\left[\frac{1}{r} + O(h)\right],$$

$$a_0 = \frac{1}{h}\left[\frac{r - 1}{r} + O(h)\right],$$

$$a_1 = \frac{1}{h}\left[-1 + O(h)\right].$$

The situation will be much the same also in the other examples we will encounter.

Let us now consider how one can construct, for problem (4), the difference scheme

$$L_h u^{(h)} \equiv \begin{cases} a^0 u_m^{n+1} + a_0 u_m^n + a_{-1} u_{m-1}^n + a_1 u_{m+1}^n = \phi(mh, n\tau). \\ u_m^0 = \psi(mh) \end{cases} \tag{10}$$

of more general form, connecting the values of the unknown function at four points, as shown in Fig. 13.

Again we connect the step-widths by the equation $\tau = rh$, $r = \text{const}$, and introduce the notation Λ_h, defining

$$\Lambda_h u^{(h)} \equiv a^0 u_m^{n+1} + a_0 u_m^n + a_{-1} u_{m-1}^n + a_1 u_{m+1}^n. \tag{11}$$

For every sufficiently smooth function $u(x,t)$ we may write, with the aid of Taylor's formula,

$$\Lambda_h [u]_h \Big|_{\substack{x=mh, \\ t=n\tau}} = (a^0 + a_0 + a_1 + a_{-1}) u(mh, n\tau) +$$

$$+ a^0 r h u_t(mh, n\tau) + (a_1 - a_{-1}) h u_x(mh, n\tau) + \frac{1}{2} a^0 r^2 h^2 u_{tt}(mh, n\tau) +$$

$$+ \frac{1}{2} (a_1 + a_{-1}) h^2 u_{xx}(mh, n\tau) + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3). \tag{12}$$

We now separate out, in the right-hand side of this equation, the term $\Lambda u \equiv (\partial u/\partial t) - (\partial u/\partial x)$, using for this purpose the identity $u_t = u_x + \Lambda u$. As a result we get

$$\Lambda_h [u]_h \Big|_{\substack{x=mh, \\ t=n\tau}} = a^0 r h \Lambda u \Big|_{\substack{x=mh, \\ t=n\tau}} + (a^0 + a_0 + a_1 + a_{-1}) u(mh, n\tau) +$$

$$+ (a^0 r + a_1 - a_{-1}) h u_x(mh, n\tau) + \frac{1}{2} a^0 r^2 h^2 u_{tt}(mh, n\tau) +$$

$$+ \frac{1}{2} (a_1 + a_{-1}) h^2 u_{xx}(mh, n\tau) + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3).$$

If we assume that the quantity $O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$ is sufficiently small, an assumption which will later be confirmed, then in order to fulfill the approximation requirement

$$(\Lambda_h [u]_h)_{\substack{x=mh, \\ t=n\tau}} = (\Lambda u)_{\substack{x=mh, \\ t=n\tau}} + O(h)$$

it is necessary that the four numbers, a^0, a_0, a_1 and a_{-1}, satisfy the three equations:

$$a^0 rh = 1 + O_1(h),$$

$$a^0 + a_0 + a_1 + a_{-1} = 0 + O_2(h),$$

$$(a^0 r + a_1 - a_{-1})h = 0 + O_3(h).$$

Suppose, according to our convention, that the arbitrary quantities $O_1(h)$, $O_2(h)$ and $O_3(h)$ of order h, are equal to zero. We then get the system of equations

$$\left.\begin{aligned} a^0 rh &= 1, \\ a^0 + a_0 + a_1 + a_{-1} &= 0, \\ a^0 r + a_1 - a_{-1} &= 0. \end{aligned}\right\} \tag{13}$$

If condition (13) is fulfilled, then

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = \Lambda u\Big|_{\substack{x=mh,\\ t=n\tau}} + \frac{1}{2}\, a^0 r^2 h^2 u_{tt}(mh, t\tau) +$$

$$+ \frac{1}{2}\,(a_1 + a_{-1})h^2 u_{xx}(mh, n\tau) + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3).$$

System (13) has many solutions, in fact a family of solutions depending on one parameter. One of these solutions

$$a^0 = \frac{1}{rh}, \qquad a_{-1} = 0, \qquad a_0 = \frac{r-1}{rh}, \qquad a_1 = -\frac{1}{h},$$

leads to the above scheme (5). The solution

$$a^0 = \frac{1}{rh}, \qquad a_0 = -\frac{1}{rh}, \qquad a_{-1} = \frac{1}{2h}, \qquad a_1 = -\frac{1}{2h}$$

corresponds to the scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_{m-1}^n}{2h} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh). \end{cases}$$

Having chosen some solution of system (13), one must substitute this solution into the remainder term and confirm that it is small. For the above two solutions substitution of the quantities a^0, a_0, a_1 and a_{-1} gives the remainder terms

$$\frac{a^0 r^2 h^2}{2}\frac{\partial^2 u}{\partial t^2} + \frac{a_1 + a_{-1}}{2} h^2 \frac{\partial^2 u}{\partial x^2} + O(a^0 r^3 h^3,\ a_1 h^3,\ a_{-1} h^3)$$

of order $O(h)$.

Among the smooth functions $u(x, t)$ are second-order polynomials for which $\partial^2 u/\partial t^2$ and $\partial^2 u/\partial x^2$ take on, at any given point, any arbitrary prescribed values. Moreover the term $O(a^0 r^3 h^3,\ a_1 h^3,\ a_{-1} h^3)$ containing the third derivatives of the polynomial $u(x, t)$ will vanish. Therefore, if the remainder term is to be of order higher than h, it is necessary that the coefficients of $\partial^2 u/\partial t^2$ and $\partial^2 u/\partial x^2$ should each, separately, be of higher order. Since, from the first of Eqs. (13), we have $a^0 = 1/(rh)$, the coefficient of $\partial^2 u/\partial t^2$ is $rh/2$ and the remainder term is never of order higher than first.

We have established that is is impossible to construct a difference scheme of form (10) which approximates the problem

$$Lu \equiv \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x} = \phi(x, t) \\ u(x, 0) = \psi(x) \end{cases}$$

to order h^2. To raise the order of approximation it would be necessary to increase the number of net-points used in constructing the scheme.

But we will now point out some methods which, nevertheless, permit the construction of a difference scheme with order h^2 approximation, using the four indicated points of the difference net. The method of raising the order of approximation, which we now present by way of examples, is general in character. It turns out that one can choose the coefficients in such a way that the equation

$$\Lambda_h [u]_h \equiv a^0 u(mh, (n+1)\tau) + a_{-1} u((m-1)h, n\tau) +$$

$$+ a_0 u(mh, n\tau) + a_1 u((m+1)h, n\tau) =$$

$$= \Lambda u + \frac{rh}{2}\left[(\Lambda u)_t + (\Lambda u)_x\right]\Big|_{\substack{x=mh,\\ t=n\tau}} + O(h^2) = P_h \Lambda u\Big|_{(x_m,\ t_n)} + O(h^2),$$

will be satisfied, where

$$P_h = E + \frac{rh}{2}\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right),$$

and E is the identity operator. Then, since $\Lambda u = u_t - u_x = \phi(x, t)$, the difference scheme

$$L_h u^{(h)} = \begin{cases} a^0 u_m^{n+1} + a_{-1} u_{m-1}^n + a_0 u_m^n + a_1 u_{m+1}^n = \phi_m^n, \\ u_m^0 = \psi(mh), \end{cases}$$

with

$$\phi_m^n = (P_h\phi)_{\substack{x=mh,\\ t=n\tau}} = \phi + \frac{rh}{2}(\phi_t + \phi_x)\Big|_{\substack{x=mh,\\ t=n\tau}},$$

will approximate the given differential problem, on the solution $u(x, t)$, to second order in h.

The coefficients a^0, a_{-1}, a_0 and a_1 again may be chosen by the method of undetermined coefficients. They turn out to have the following values:

$$a^0 = \frac{1}{rh}, \qquad a_0 = -\frac{1}{rh} + \frac{r}{h}, \qquad a_{-1} = \frac{1 - r}{2h}, \qquad a_1 = -\frac{1 + r}{2h}.$$

With these values the operator Λ_h takes the form

$$\Lambda_h u^{(h)} = \frac{1}{\tau}(u_m^{n+1} - u_m^n) - \frac{1}{2h}(u_{m+1}^n - u_{m-1}^n) - \frac{r}{2h}(u_{m+1}^n - 2u_m^n + u_{m-1}^n).$$

By the method of undetermined coefficients one can not only choose coefficients a^0, a_{-1}, a_0 and a_1 for which

$$\Lambda_h[u]_h = a^0u(x, t + \tau) + a_{-1}u(x - h, t) +$$

$$+ a_0u(x, t) + a_1u(x + h, t) = P_h\Lambda u + O(h^2)$$

with the above defined operator P_h, but one can also construct all operators, P_h, for which the above equation can be satisfied.

* * * * * *

Let us now show how this can be done.

Taking

$$\Lambda_h u^{(h)} \equiv a^0u_m^{n+1} + a_{-1}u_{m-1}^n + a_0u_m^n + a_1u_{m+1}^n,$$

and using Taylor's formula, we get

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = (a^0 + a_0 + a_1 + a_{-1})u(mh, n\tau) +$$

$$+ a^0rhu_t(mh, n\tau) + (a_1 + a_{-1})hu_x(mh, n\tau) + \frac{1}{2}a^0r^2h^2u_{tt}(mh, n\tau) +$$

$$+ \frac{1}{2}(a_1 + a_{-1})h^2u_{xx}(mh, n\tau) + O(a^0r^3h^3, a_1h^3, a_{-1}h^3). \quad (14)$$

This equation will, next, be put into a somewhat different form. We start with a derivation of the identity

$$\frac{\partial^2u}{\partial t^2} \equiv \frac{\partial^2u}{\partial x^2} + (\Lambda u)_t + (\Lambda u)_x$$

which follows from the definition of Λu:

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} + \Lambda u.$$

The proof consists of a chain of obvious identities:

$$\frac{\partial^2 u}{\partial t^2} = \left(\frac{\partial u}{\partial x} + \Lambda u\right)_t = \frac{\partial^2 u}{\partial x \partial t} + (\Lambda u)_t \equiv \frac{\partial}{\partial x} u_t + (\Lambda u)_t \equiv$$

$$\equiv \frac{\partial}{\partial x}\left(u_x + \Lambda u\right) + (\Lambda u)_t \equiv u_{xx} + (\Lambda u)_t + (\Lambda u)_x.$$

Using these identities one can rewrite Eq. (14) in the following, equivalent, form

$$\Lambda_h [u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = a^0 rh(\Lambda u)^n_m + \frac{1}{2} a^0 r^2 h^2 \left[(\Lambda u)_t + (\Lambda u)_x\right]^n_m +$$

$$+ \left(a_0 + a^0 + a_1 + a_{-1}\right)u(mh, n\tau) + \left(a^0 r + a_1 - a_{-1}\right)hu_x(mh, n\tau) +$$

$$+ \left[\frac{1}{2} a^0 r^2 + \frac{1}{2}\left(a_1 + a_{-1}\right)\right]h^2 u_{xx}(mh, n\tau) + O\left(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3\right). \quad (15)$$

We now construct the operator satisfying the condition $\Lambda_h u = P_h \Lambda u + O(h^2)$. The terms containing Λ_u, $(\Lambda u)_x$ and $(\Lambda u)_t$ may be included in the expression $P_h \Lambda u$, since the definition of $P_h \Lambda u$ is at our disposal. All the other terms

$$\left(a_0 + a^0 + a_{-1} + a_1\right)u(mh, n\tau),$$

$$\left(a^0 r + a_1 - a_{-1}\right)hu_x(mh, n\tau),$$

$$\frac{a^0 r^2 + a_1 + a_{-1}}{2} h^2 u_{xx}(mh, n\tau),$$

$$O\left(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3\right),$$

must be constituents of the remainder term of the equation

$$\Lambda_h [u]_h = P_h \Lambda u + \text{remainder term},$$

no matter how we try to choose the operator P_h. The validity of this assertion is proven by the fact that there exist functions, $u(x, t)$, for which u, u_x, u_{xx}, Λu, $(\Lambda u)_x$ and $(\Lambda u)_t$ take on, at any given point (x_0, t_0) any mutually independent prescribed values u^0, u^0_x, u^0_{xx}, $(\Lambda u)^0$, $(\Lambda u)^0_x$ and $(\Lambda u)^0_t$. One such function, for example, is the polynomial

$$P(x, t) = u^0 + u^0_x(x - x_0) + [(\Lambda u)^0 + u^0_x](t - t_0) + \frac{1}{2} u^0_{xx}(x - x_0)^2 +$$

$$+ \frac{1}{2}[u^0_{xx} + (\Lambda u)^0_x + (\Lambda u)^0_t](t - t_0)^2 + [u^0_{xx} + (\Lambda u)^0_x](x - x_0)(t - t_0).$$

In view of the independence of the values u, u_x, u_{xx}, Λu, $(\Lambda u)_t$ and $(\Lambda u)_x$ it is necessary, in order to achieve second-order approximation, that each term, individually, entering into the remainder should be of order h^2. This requirement may be written in the form

$$\left.\begin{aligned} a_0 + a^0 + a_1 + a_{-1} &= 0, \\ (a_0 r + a_1 - a_{-1})h &= 0. \\ (a_0 r^2 + a_1 + a_{-1})h^2 &= 0. \end{aligned}\right\} \tag{16}$$

The solution of system (16) is determined to within an arbitrary factor. We will supplement this system by the equation

$$a_0 rh = 1, \tag{17}$$

which constitutes a natural, though not necessary, constraint on the choice of the operator P_h: i.e the coefficient of (Λu) in the expression for $P_h \Lambda u$ is taken to be unity.

On the right-hand side of equations (6) and (7) it would be possible to add arbitrary terms $O_1(h^2)$, $O_2(h^2)$, $O_3(h^2)$ and $O_4(h^2)$ but we have, in conformance with our earlier conventions, set these terms to zero.

Solving the system of equations (16), (17), we get the coefficients, a^0, a_{-1}, a_0 and a_1, already given earlier:

$$a^0 = \frac{1}{rh}, \qquad a_0 = -\frac{1}{rh} + \frac{r}{h}, \qquad a_{-1} = \frac{1 - r}{2h}, \qquad a_1 = -\frac{1 + r}{2h}.$$

With these values for the coefficients the remainder term in Eq. (15)

$$\Lambda_h[u]_h = \Lambda u + \frac{rh}{2}(\Lambda u)_t + \frac{rh}{2}(\Lambda u)_x + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h3) \equiv$$

$$\equiv P_h \Lambda u + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$$

satisfies the bound

$$|O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)| \leq A(r^2 h^2 + h^2)$$

where A is some constant depending only on the maximum absolute value of the third-order derivatives of the function u(x, t). Correspondingly we may also write

$$|\Lambda_h[u]_h - P_h\phi|_m^n \leq A(r^2 + 1)h^2.$$

Thus, we have established that, ignoring insignificant variations, only one difference scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_{m-1}^n}{2h} - \dfrac{\tau}{2}\dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = \\ \qquad = \left[\phi + \dfrac{rh}{2}(\phi_t + \phi_x)\right]_{\substack{x=mh,\\ t=n\tau}} \\ u_m^0 = \psi(mh) \end{cases} \qquad (17')$$

among all difference schemes of the form

$$L_h u^{(h)} = \begin{cases} a^0 u_m^{n+1} + a_{-1}u_{m-1}^n + a_0 u_m^n + a_1 u_{m+1}^n = P_h\phi|_m^n, \\ u_m^0 = \psi(mh) \end{cases}$$

approximates differential boundary-value problem (4) on its solution $u(x, t)$ to second order in h.

* * *

In all the examples of difference schemes, $L_h u^{(h)} = f^{(h)}$, presented so far in this chapter, the operator L_h, mapping the space U_h into space F_h, is given by explicit equations. But one often has use for difference schemes in which the operator L_h is specified in some other, more complicated, way. Below we will encounter problems for which such schemes will evolve naturally.

The above methods for constructing difference schemes remain applicable also for problems with variable coefficients, for nonlinear problems and for nets with variable step-size. For example, in the case of the non-uniform net shown in Fig. 14, one can construct a difference scheme for the equation $u_{xx} + u_{yy} = \phi(x, y)$ by substituting, for the derivatives in this equation, the difference relations

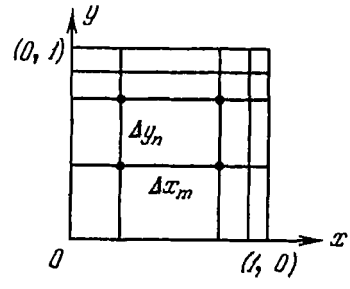


Fig. 14.

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_n, y_n)} = \frac{\dfrac{u(x_{m+1}, y_n) - u(x_m, y_n)}{\Delta x_m} - \dfrac{u(x_m, y_n) - u(x_{m-1}, y_n)}{\Delta x_{m-1}}}{\dfrac{\Delta x_m + \Delta x_{m-1}}{2}} +$$

$$+ \frac{1}{3}(\Delta x_m - \Delta x_{m-1}) u_{xxx} + O\left[(\Delta x_m + \Delta x_{m-1})^2\right],$$

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_{(x_n, y_n)} = \frac{\dfrac{u(x_m, y_{n+1}) - u(x_m, y_n)}{\Delta y_n} - \dfrac{u(x_m, y_n) - u(x_m, y_{n-1})}{\Delta y_{n-1}}}{\dfrac{\Delta y_n + \Delta y_{n-1}}{2}} +$$

$$+ \frac{1}{3}(\Delta y_n - \Delta y_{n-1}) u_{yyy} + O\left((\Delta y_n + \Delta y_{n-1})^2\right),$$

discarding the remainder terms. The validity of the above equations may be confirmed with the aid of Taylor expansion (2). By the method of undetermined coefficients we may convince ourselves of the uniqueness of the equations: to within unessential variations there is only one set of coefficients a_{-1}, a_0, a_1, for which we may write, given any sufficiently smooth function $u(x, t)$, the expression

$$\frac{\partial^2 u(x_m, y_n)}{\partial x^2} = a_{-1} u(x_{m-1}, y_n) + a_0 u(x_m, y_n) +$$

$$+ a_1 u(x_{m+1}, y_n) + O\left[\max(\Delta x_{m-1}, \Delta x_m)\right]$$

with a remainder term which is small to first order with respect to $\max[\Delta x_{m-1}, \Delta x_m]$.

Equations of the form

$$\frac{\partial^2 u(x_m, y_n)}{\partial x^2} = a_{-1} u(x_{m-1}, y_n) + a_0 u(x_m, y_n) +$$

$$+ a_1 u(x_{m+1}, y_n) + O\left(\left[\max(\Delta x_{m-1}, \Delta x_m)\right]^2\right),$$

with remainder terms of second order, do not exist for $\Delta x_{m-1} \neq \Delta x_m$.

To achieve greater accuracy via the replacement of derivatives by difference expressions it would be necessary to involve more than three net-points.

3. Schemes with recomputation, or "predictor-corrector" schemes. To construct difference schemes approximating time-dependent problems one can use the same idea which underlies the construction of the Runge-Kutta scheme for ordinary differential equations, the idea of "recalculation". Recalculation allows one to raise the order of approximation attained, by use of the initial scheme, before recalculation. In addition, in the case of quasilinear differential equations recalculation allows us to construct so-called "divergence" schemes, about which more will be said in §30.

We recall the idea of recalculation via the example of the simplest Runge-Kutta scheme for the numerical solution of the Cauchy problem

$$\frac{dy}{dt} = f(t, y), \qquad y(0) = \psi, \qquad 0 < t < T. \tag{18}$$

If the value y_p at the point $t_p = p\tau$ has already been calculated then, to compute y_{p+1}, we determine an auxiliary quantity, $\tilde{y}_{p+1/2}$, using the simplest Euler scheme (the "predictor" scheme)

$$\frac{\tilde{y}_{p+1/2} - y_p}{\tau/2} = f(t_p, y_p), \tag{19}$$

and then carry out the corrective recalculation

$$\frac{y_{p+1} - y_p}{\tau} = f(t_{p+1/2}, \tilde{y}_{p+1/2}). \tag{20}$$

The auxiliary quantity $\tilde{y}_{p+1/2}$, computed by use of a scheme with first-order accuracy, allows us to determine, approximately, the inclination of the integral curve at the midpoint of the interval $[t_p, t_{p+1}]$, and to get y_{p+1}, using Eq. (20), more accurately than by Euler's scheme (19).

We have already noted, in 4§19, that all our considerations remain valid if y, y_p and $\tilde{y}_{p+1/2}$ are finite-dimensional vectors and f is a vector function. But one can go still further, considering y, y_p and $\tilde{y}_{p+1/2}$ as elements of a function space, and f an operator in this space. For example the Cauchy problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0, \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\} \tag{21}$$

$A = \text{const}$, can be thought of as a problem of form (18) if we set $y(t) = u(x, t)$ so that, for each t, y is taken to be a function of the argument x; the operation f is interpreted to mean the application of the operator $-A\partial/\partial x$. Let us take as an example a difference scheme, with recalculation, for problem (21).

Example. Suppose that the net function $u^p = \{u_m^p\}$, $m = 0, \pm 1, \ldots$, for a given p, has already been computed. We first determine the auxiliary net function $\tilde{u}^{p+1/2} = \{\tilde{u}_{m+1/2}^{p+1/2}\}$, $m = 0, \pm 1, \ldots$, relating to time $t_{p+1/2} = (p + 1/2)\tau$ and to point $x_{m+1/2} = (m + 1/2)h$, using the following (first-order accurate) scheme:

$$\frac{\tilde{u}_{m+1/2}^{p+1/2} - \frac{u_{m+1}^p + u_m^p}{2}}{\tau/2} + A\frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad m = 0, \pm 1, \ldots \tag{22}$$

Then we perform the correction, and find u^{p+1} using the scheme

$$\frac{u_m^{p+1} - u_m^p}{\tau} + A\,\frac{\tilde{u}_{m+1/2}^{p+1/2} - \tilde{u}_{m-1/2}^{p+1/2}}{h} = 0, \qquad m = 0, \pm 1, \ldots \tag{23}$$

Eliminating $\tilde{u}^{p+1/2}$ from Eqs. (22) and (23), we get the scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} + A\,\frac{u_{m+1}^p - u_{m-1}^p}{2h} - A^2\,\frac{\tau}{2}\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ &u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \ldots; \qquad p = 0, 1, \ldots, [T/\tau] - 1. \end{aligned}\right\} \tag{24}$$

This latter scheme, for $A = -1$, coincides with scheme (17′), and the case $A \neq -1$ differs only insignificantly from the one already discussed. Scheme (24), and thus also the scheme with recalculation, i.e. (22), (23), has second-order approximation in h when $\tau = rh$, $r = \text{const}$.

4. On other examples. We now mention, briefly two more extremely important and widely-used methods for the construction of difference schemes. The first of these is based on the formulation of the original differential equation, the equation which is to be differenced, as an "integral conservation law". The need for the use of this method arises naturally in the computation of so-called "generalized solutions," functions which do not have full sets of derivatives, or may even be discontinuous. Difference schemes developed in this way are called "divergence schemes" or "conservative schemes". Methods for constructing divergence schemes are described in Chapter 9.

The second method is based on the use of some variational formulation of the differential boundary-value problem whose solution is to be computed. This method is often called the method of finite elements, and the corresponding difference schemes are referred to as "variational-difference" or "projective-difference" schemes. This method allows the construction of difference schemes on irregular nets, finer in regions where the solution changes more quickly. Chapter 12 will be devoted to the discussion of such schemes.

PROBLEMS

1. For the solution of the Cauchy problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = \phi(x, t), \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty. \end{aligned}\right\}$$

use the net $x_m = mh$, $t_n = n\tau$, $h = \tau$, and construct a difference scheme of the form

$$L_h u^{(h)} = \begin{cases} a^0 u_m^{n+1} + a^1 u_{m+1}^{n+1} + a_0 u_m^n + a_1 u_{m+1}^n = \phi_m^n \\ u_m^0 = \psi(mh). \end{cases}$$

How must one define a^0, a^1, a_0, a_1 and ϕ_m so as to achieve order h^2 approximation?

2. For the Cauchy problem

$$\frac{\partial u}{\partial t} - \left(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y}\right) = \phi(x, y, t), \qquad -\infty < x,y < \infty, \quad 0 < t < T,$$

$$u(x,y,0) = \psi(x,y), \qquad -\infty < x,y < \infty$$

use the net $x_m = mh$, $y_n = n\tau$, $t_p = p\tau$, and construct any approximating difference scheme.

3. For the heat conduction problem

$$\left.\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \qquad -\infty < x < \infty, \quad 0 \le t \le T, \\ u(x, 0) &= \psi(x), \qquad -\infty < x < \infty \end{aligned}\right\} \tag{25}$$

consider the difference scheme

$$\left.\begin{aligned} \frac{u_m^{n+1} - u_m^n}{\tau} &= \sigma \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2} + (1 - \sigma) \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}, \\ u_m^0 &= \psi(mh), \end{aligned}\right\}$$

where σ is a parameter and u_m^n the value of the desired function at the point $(x_m = mh, t_n = n\tau)$ of the net.

a. Show that, for any σ, the differential equation is approximated on a smooth solution $u(x, t)$ to order $O(\tau + h^2)$.

b. Choose a σ such that approximation will be of order $O(\tau^2 + h^2)$.

c. Taking the step sizes to be connected by the relation $\tau/h^2 \equiv r = \text{const}$, choose σ so as to get an approximation of order h^4.

d. For $\sigma = 0$ choose the number $r = \tau/h^2$ so that approximation will be of order h^4.

e. Is it possible, through the choice of σ for given $r = \tau/h^2$, to achieve approximation, on any smooth solution, of order higher than fourth?

4. For the heat-conduction problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(x, t) \frac{\partial u}{\partial x}\right], \qquad -\infty < x < \infty, \quad 0 < t < T,$$

$$u(x, 0) = \psi(x), \qquad -\infty < x < \infty,$$

using the net $x_m = mh$, $t_n = n\tau$, construct an approximating difference scheme.

5. For the nonlinear heat-conduction problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(u)\,\frac{\partial u}{\partial x}\right], \qquad -\infty < x < \infty, \quad 0 < t < T,$$

$$u(x,0) = \psi(x), \qquad -\infty < x < \infty,$$

using the net $x_m = mh$, $t_n = n\tau$, construct an explicit approximating difference scheme. Write out equations for the computation of $u^{(h)}$ by this scheme.

6. Prove that, for a bounded net function $u^p = \{u_m^p\}$, there exists a unique bounded net function $u^{p+1} = \{u_m^{p+1}\}$ defined by the difference scheme

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^{p+1} - u_{m-1}^{p+1}}{2h} = 0, \qquad m = 0, \pm 1, \ldots$$

7. Prove that the predictor-corrector scheme for problem (25), in which the solution-values $\{\tilde{u}_m^{p+1/2}\}$ at the intermediate level are given by the implicit scheme with order of approximation $O(\tau + h^2)$

$$\frac{\tilde{u}_m^{p+1/2} - u_m^p}{\tau/2} - \frac{\tilde{u}_{m+1}^{p+1/2} - 2\tilde{u}_m^{p+1/2} + \tilde{u}_{m-1}^{p+1/2}}{h^2} = 0, \qquad m = 0, \pm 1, \ldots,$$

and the solution $\{u_m^{p+1}\}$ is defined by the scheme

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{\tilde{u}_{m+1}^{p+1/2} - 2\tilde{u}_m^{p+1/2} + \tilde{u}_{m-1}^{p+1/2}}{h^2} = 0, \quad u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \ldots$$

has approximation of order $O(\tau^2 + h^2)$ on a smooth solution u.

§23. Examples of the formulation of boundary conditions in the construction of difference schemes

The examples of §22 were so selected that questions relating to the construction of difference boundary conditions did not arise. These could easily be obtained from the differential boundary conditions and formulated in such a way that, upon substitution of $[u]_h$, they would be satisfied exactly. Here we consider examples which, as regards boundary conditions, are more complicated.

Example 1. In the construction of a difference scheme for the problem

$$Lu = \begin{cases} u_t - u_x = \phi(x, t), \\ u(x, 0) = \psi(x) \end{cases} \tag{1}$$

we will use the difference equation

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} - \frac{u_{m+1}^{n} - u_{m-1}^{n}}{2h} = \phi(mh, n\tau), \tag{2}$$

$$n = 1, 2, \ldots; \qquad m = 0, \pm 1, \ldots; \qquad \tau = rh.$$

To calculate the solution of Eq. (2) we must fix not only u_m^0,

$$u_m^0 = \psi(mh), \qquad m = 0, \pm 1, \ldots, \tag{3}$$

but also u_m^1, $m = 0, \pm 1, \ldots$ Then from difference equation (2) for $n = 1, 2, \ldots$, one can, next, compute u_m^2, $m = 0, \pm 1, \ldots$, then u_m^3, $m = 0, \pm 1, \ldots$, etc. The value assigned to u_m^1 must be close to

$$u(mh, \tau) = u(mh, 0) + \tau u_t(mh, 0) + O(\tau^2).$$

Since $u_t = u_x + Au$, $Au \equiv u_t - u_x = \phi(x, t)$, $u(x, 0) = \psi(x)$,

$$u(mh, \tau) = u(mh, 0) + \tau\left[u_x + Au\right]_{\substack{x=mh,\\ t=0}} + O(\tau^2) =$$
$$= \psi(mh) + \tau\left[\psi'(mh) + \phi(mh, 0)\right] + O(\tau^2).$$

Thus, discarding the term $O(\tau^2)$ we may write

$$u_m^1 = \psi(mh) + \tau\left[\psi'(mh) + \phi(mh, 0)\right]. \tag{4}$$

Clearly the difference scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - u_m^{n-1}}{2\tau} - \dfrac{u_{m+1}^{n} - u_{m-1}^{n}}{2h} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh), \\ u_m^1 = \psi(mh) + \tau\left[\psi'(mh) + \phi(mh, 0)\right] \end{cases} \tag{5}$$

approximates the differential boundary-value problem (1) to order h^2. The complication in this scheme consists in the fact that difference equation (2) is second-order in t, while the differential equation is first-order. For this reason it was necessary to construct a second difference boundary condition (4), not arising directly from the given boundary condition for the differential problem.

Let us now consider another example in which the construction of difference boundary conditions is not trivial.

Example 2. Consider the differential boundary-value problem

$$Lu = \begin{cases} u_t - u_x = \phi(x, t), & 0 < x < 1, \quad 0 < t < T, \\ u(0,x) = \psi_0(x), & 0 < x < 1, \\ u(t, 1) = \psi_1(t), & 0 < t < T. \end{cases} \tag{6}$$

Any solution of the differential equation of problem (6) is uniquely defined if we know its values at one point on each of the lines $x + t = \text{const}$. In fact along such a line

$$\frac{du}{dt} = u_t + u_x \frac{dx}{dt} = u_t - u_x = \phi(x, t),$$

so that $u(x, t)$ is the integral, along the line

$$x + t = \text{const}$$

of $\phi(x, t)$. The value of the integration constant is determined by the value of u at the given point.

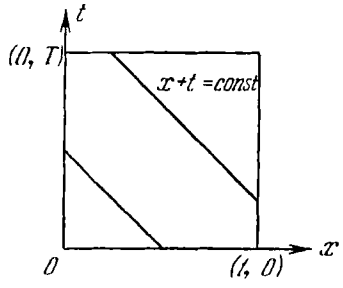


Fig. 15

In Fig. 15 we depict the rectangle, $0 \leq x \leq 1$, $0 \leq t \leq T$, in which we intend to look for the solution, and show two lines of the family of parallel lines $x + t = \text{const}$. Each line of this family intersects, at a single point, either the segment $0 \leq x \leq 1$ of the x-axis, or the segment $0 \leq t \leq T$ of the line $x = 1$, where $u(x, t)$ is given. Thus problem (6) has a unique solution.

We now proceed to the construction of a difference scheme for the computation of the solution of problem (6). Suppose h is given such that $Mh = 1$, and assume that $\tau = rh$, where M is a positive integer and $r = \text{const}$. As a net, D_h, we use the set of points $(mh, n\tau)$, $m = 0, 1, \ldots, M$; $n = 0, 1, \ldots, [T/\tau]$. With each point of D_h not lying on the upper boundary or the sides of the rectangle, we associate an equation

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_{m-1}^n}{2h} - \frac{\tau}{2} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = \phi_m^n, \tag{7}$$

where

$$\phi_m^n = \left[\phi(x, t) + \frac{rh}{2}(\phi_t + \phi_x)\right]_{\substack{x=mh, \\ t=n\tau}}. \tag{8}$$

The derivation of this equation was described in detail in §22.

The values u_m^0 and u_M^n will be given by the equations

$$\left.\begin{aligned} u_m^0 &= \psi_0(mh), \qquad m = 0, 1, \ldots, M-1, \\ u_M^n &= \psi_1(n\tau), \qquad n = 0, 1, \ldots, N, \quad N \equiv [T/\tau], \end{aligned}\right\} \tag{9}$$

which are analogous to the boundary conditions of the given differential problem. But Eqs. (9) do not suffice to determine the solution u_m^n everywhere on D_h. The value of u_0^{n+1}, at the left hand boundary of the rectangle, is still undefined. For this reason we supplement the difference boundary conditions as follows:

$$\frac{u_0^{n+1} - u_0^n}{\tau} - \frac{u_1^n - u_0^n}{h} = \phi(0, n\tau), \qquad n = 0, 1, \ldots, N-1. \tag{10}$$

This conditions results if we substitute for the derivatives in the equation

$$\frac{\partial u(0, t)}{\partial t} - \frac{\partial u(x, t)}{\partial x} = \phi(0, t),$$

which follows from Eq. (6), appropriate difference relations.

Thus we have constructed the difference scheme $L_h u^{(h)} = f^{(h)}$:

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_{m-1}^n}{2h} - \dfrac{\tau}{2} \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \\ \qquad m = 1, 2, \ldots, M-1; \quad n = 0, 1, \ldots, N-1, \\ u_m^0, \quad m = 0, 1, \ldots, M-1, \\ u_M^n, \quad m = 0, 1, \ldots, N, \\ \dfrac{u_0^{n+1} - u_0^n}{\tau} - \dfrac{u_1^n - u_0^n}{h}, \qquad n = 0, 1, \ldots, N-1, \end{cases}$$

$$f^{(h)} = \begin{cases} \left[\phi + \dfrac{rh}{2}(\phi_t + \phi_x)\right]_{\substack{x=mh, \\ t=n\tau}}, \\ \psi_0(mh), \qquad m = 0, 1, \ldots, M-1, \\ \psi_1(n\tau), \qquad n = 0, 1, \ldots, N, \\ \phi(0, n\tau), \qquad n = 0, 1, \ldots, N-1. \end{cases}$$

Let us now determine the order of approximation of this scheme. Taking account of the considerations of §22 it is clear that the residual $\delta f^{(h)}$, which develops when $[u]_h$ is substituted into the difference scheme $L_h[u]_h = f^{(h)} + \delta f^{(h)}$, assuming a sufficiently smooth solution $u(x, t)$, has the form

$$\delta f^{(h)} \equiv \begin{cases} O_{mn}(h^2) = O(h^2), & m = 1, 2, \ldots, M-1; \quad n = 0, 1, \ldots, N-1, \\ 0, & m = 0, 1, \ldots, M-1, \\ 0, & n = 0, 1, \ldots, N, \\ \frac{\tau}{2} u_{tt}(0, n\tau + \xi_1\tau) + \frac{h}{2} u_{xx}(\xi_2 h, n\tau), & n = 0, 1, \ldots, N-1, \\ & 0 < \xi_1 < 1, \quad 0 < \xi_2 < 1. \end{cases}$$

If we introduce a norm in F_h, assuming that for any element $g^{(h)}$ in F_h

$$g^{(h)} = \begin{cases} A_m^n, \\ a_m, \\ b^n, \\ c^n, \end{cases}$$

$$||g^{(h)}||_{F_h} = \max_{m,n} \left|A_m^n\right| + \max_m |a_m| + \max_n |b^n| + \max_n |c^n|,$$

then $||\delta f^{(h)}||_{F_h} = O(h)$, and approximation turns out to be only of first order in h. From the expression for $\delta f^{(h)}$ it is clear that the approximation is first-order because of the residual $(\tau/2)u_{tt} + (h/2)u_{xx} = O(h)$, resulting from the substitution of $[u]_h$ into the auxiliary boundary condition which we have artificially constructed, and imposed at the left-hand boundary.

The magnitude of the remainder term, in the norm $||\cdot||_{F_h}$ which we are now using, is determined only by the second derivatives of the solution; i.e. this norm does not allow us, in studying the boundary conditions, to take advantage of the same degree of smoothness which we had to assume in the solution to get second-order approximation at interior points.

We now introduce a norm $||\cdot||_{F_h}$ for which the above difference scheme has second order approximation on a sufficiently smooth solution $u(x, t)$:

$$\|g^{(h)}\|_{F_h} = h \max_n |c^n| + h \max_n \left|\frac{c^{n+1} - c^n}{\tau}\right| +$$

$$+ \left(h \sum_{m=0}^{N} |a_m|^2\right)^{1/2} + \max_n |b^n| + \max_n \left|\frac{b^{n+1} - b^n}{\tau}\right| + \max_{m,n} |A_m^n| .$$

For this scheme, as is easily seen,

$$\|\delta f^{(h)}\|_{F_h} \leq A(rh^2 + h^2), \quad r = \tau/h.$$

Further the constant A depends on derivatives up to and including the third.

Smoothness is accounted for, in this norm, via the terms

$$\left|\frac{c^{n+1} - c^n}{\tau}\right|, \qquad \left|\frac{b^{n+1} - b^n}{\tau}\right|.$$

The reader has probably noticed that some of the terms in the formula defining the new norm in F_h differ from the corresponding terms in the old norm through the presence of a factor h. It is clear that if one arbitrarily multiplies terms by h, and by various powers of h, then one can achieve any desired order of approximation. But in §13 we have already discussed the question of the choice of norms in connection with ordinary differential equations and we know that only those norms are useful in which the difference scheme simultaneously approximates the differential boundary-value problem, and is stable.

The stability of the above scheme, using norms in which it has second-order approximation, will be proven in §42.

Example 2 is very instructive. It shows that, to verify approximation in any reasonable sense, one must choose a norm correctly. In studying different possible schemes it is necessary to test many norms. In each norm one must try to carry out a study of stability which by itself, at least at present, often requires inventiveness and labor.

In practice in most cases, instead of studying the real problem which concerns us, one investigates a simplified, so-called "model" problem, after which one carries out test calculations using the proposed difference scheme for the original, unsimplified problem.

PROBLEMS

1. For the Cauchy problem

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = \phi(x,t), \qquad -\infty < x < \infty, \qquad 0 \leq t \leq T,$$

$$u(x, 0) = \psi_1(x), \qquad -\infty < x < \infty,$$

$$\frac{\partial u(x, 0)}{\partial t} = \psi_2(x), \qquad -\infty < x < \infty,$$

study the order of approximation, on a sufficiently smooth solution $u(x, t)$, of the difference scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2} - \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = \phi(mh, n\tau), \\ u_m^0 = \psi_1(mh), \\ \dfrac{u_m^1 - u_m^0}{\tau} = (\psi_2)_m. \end{cases}$$

if $(\psi_2)_m = \psi_2(mh)$. Take, as the norm $\|f^{(h)}\|_{F_h}$, the maximum of the moduli of all components of the element

$$f^{(h)} = \begin{cases} \phi_m^n, \\ \psi_1(mh), \\ (\psi_2)_m. \end{cases}$$

Show that the approximation is first-order in h; $\tau = rh$, $r = \text{const}$.

How must one assign the value of $(\psi_2)_m$, using the given functions $\phi(x, t)$, $\psi_1(x)$ and $\psi_2(x)$, so that approximation will be second-order?

2. For the heat-conduction problem on a line-interval,

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \phi(x, t), \qquad 0 < x < 1, \qquad 0 < t < T,$$

$$u(x, 0) = \psi_0(x), \qquad 0 < x < 1,$$

$$\frac{\partial u(0, t)}{\partial x} = \psi_1(t), \qquad 0 < t < T,$$

$$u(1, t) = \psi_2(t)$$

consider the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} = \phi(mh, n\tau), \\ &m = 1, 2, \ldots, M-1; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi_0(mh), \qquad m = 0, 1, \ldots, M, \\ &\frac{u_1^n - u_0^n}{h} = (\psi_1)^n, \qquad n = 1, 2, \ldots, [T/\tau], \\ &u_M^n = \psi_2(n\tau), \qquad n = 1, 2, \ldots, [T/\tau]. \end{aligned}\right\}$$

As the norm $\|\cdot\|_{F_h}$, take the maximum of the absolute values of the right-hand sides of all equations which, collectively, make up the given difference scheme. Assume that the step-sizes, τ and h, are connected by the relation $\tau = rh^2$, $r = \text{const}$. Show that, setting $(\psi_1)^n = \psi_1(nh)$, we get a scheme with first-order approximation on a smooth solution. What sort of expression must one use to define $(\psi_1)^n$ in order to get an approximation of second order?

§24. The Courant-Friedrichs-Levy condition, necessary for convergence

In §21 we proved that the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = 0, \\ &u_m^0 = \psi(mh), \end{aligned}\right\} \tag{1}$$

approximating the Cauchy problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \qquad 0 < t < T, \\ &u(x, 0) = \psi(x), \end{aligned}\right\} \tag{2}$$

cannot be convergent for an arbitrary function $\psi(x)$ if $\tau/h > 1$ (see Fig. 9 on p. 193). In the course of the proof we used a principle, general in character and first formulated, in connection with another example, by Courant, Friedrichs and Levy. This principle is often useful in the construction and study of difference schemes. It may be stated as in the following section.

1. The Courant-Friedrichs-Levy condition. Suppose that the formulation of a differential problem involves some function, ψ (see (2), for example). Choose an arbitrary point, P, belonging to the domain of definition of the solution u. Suppose that the value of the solution u(P) de-

pends on values of the function ψ at the points of some set, $G_\psi = G_\psi(P)$, belonging to the domain of definition of ψ, i.e. a change in the value of ψ in a small neighborhood of any point Q of $G_\psi(P)$ can evoke a change in the value of the solution of u(P). Suppose further that, for the computation of the solution u, one uses a difference scheme, $L_h u^{(h)} = f^{(h)}$, such that the value of the solution, $u^{(h)}$, at the net-point closest to P, is completely determined by values of the function ψ on some set $G_\psi^{(h)} = G_\psi^{(h)}(P)$.

In order that it be convergent, so that $u^{(h)} \to u$ as $h \to 0$, the difference scheme must be so constructed that, as $h \to 0$, an arbitrary neighborhood of any point of $G_\psi(P)$ must, for sufficiently small h, contain a point of the set $G_\psi^{(h)} = G_\psi^{(h)}(P)$.

Let us explain why, if the Courant-Friedrichs-Levy condition is not satisfied, one cannot expect convergence. Suppose the condition is violated so that, in some fixed neighborhood of a point Q of $G_\psi(P)$, for all sufficiently small h, there is no point of the set $G_\psi^{(h)\psi} = G_\psi^{(h)}(P)$. If convergence $u^{(h)} \to u$ does occur (accidentally!) for the given function ψ, then we change ψ in the indicated neighborhood of point Q in such a way as to change u(P), leaving ψ unchanged outside this neighborhood. Convergence $u^{(h)} \to u$ for the new function ψ is impossible: the value u(P) has changed, while $u^{(h)}$ at the net-point closest to P has, for small h, remained unchanged, since there has been no change in ψ at points of the net $G_\psi^{(h)} = G_\psi^{(h)}(P)$.

The Courant-Friedrichs-Levy condition can easily be put into the form of a theorem, and the above arguments converted into a proof, but we shall not do this.

Next we consider several examples where the above considerations permit us to prove the divergence and unsuitability of a proposed difference scheme, and to feel our way to a stable and convergent difference scheme. Of course proof of convergence must be carried out separately, since fulfillment of the Courant-Friedrichs-Levy condition is only necessary, and not sufficient, for convergence. We note further that, given approximation, the Courant-Friedrichs-Levy condition is also necessary for stability, since approximation and stability imply convergence.

2. Examples of difference schemes for the Cauchy problem. We now use the Courant-Friedrichs-Levy condition for the analysis of several difference schemes approximating the Cauchy problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} + a(t)\frac{\partial u}{\partial x} = \psi_0(x, t), \qquad -\infty < x < \infty, \quad 0 < t < t, \\ &u(x, 0) = \psi_1(x), \qquad -\infty < x < \infty, \end{aligned}\right\} \tag{3}$$

where $\psi_0(x, t)$ and $\psi_1(x)$ are given "input data" for problem (3), and

$$a(t) \equiv -1 - 2t.$$

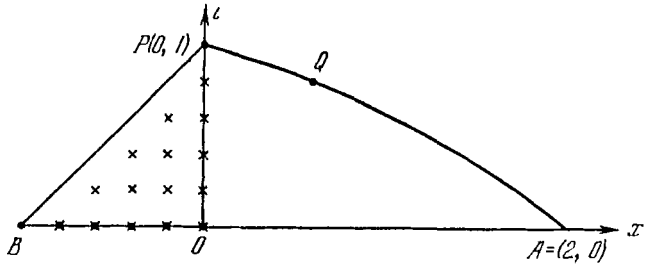


Fig. 16

The solution of problem (3) at any point, (x_p, t_p), depends on the values of the functions $\psi_0(x, t)$ and $\psi_1(x)$ at all points transversed by by the characteristic segment which, emerging from some point, A, of the x axis, ends at point P.

* * * * * *

In fact the characteristics, here, are the integral curves of the differential equation

$$\frac{dx}{dt} = a(t),$$

i.e. the parabolas $x = -t^2 - t + C$. Along each characteristic

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = \frac{\partial u}{\partial t} + a(t)\frac{\partial u}{\partial x} = \psi_0(x, t).$$

Therefore the value of the solution $u(x_p, t_p)$, at some point $P = (x_p, t_p)$, is given by the expression

$$u(x_p, t_p) = \psi_1(A) + \int_0^{t_p} \psi_0[x(t), t]dt = \psi_1(A) + \int_{AQP} \psi_0(x, t)dt.$$

where A is a point on the x axis, and AQP a segment of the characteristic.

* * *

In Fig. 16 we show the characteristic $x = 2 - t - t^2$, emerging from point A = (2, 0), and entering point P = (0, 1). We see that the value $u(P) = u(x_p, t_p)$ of the solution of problem (3) depends on the value of the function $\psi_1(x)$ at point A, so that $A = G_{\psi_1}(P)$. Further, u(P) depends on the values of $\psi_0(x, t)$ on the characteristic segment AQP. This segment AQP is, then, $G_{\psi_0}(P)$.

Consider the difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} + a(t_n)\dfrac{u_m^n - u_{m-1}^n}{h} = \psi_0(x_m, t_n) \\ u_m^0 = \psi_1(x_m), \quad m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [1/\tau]-1, \end{cases} \tag{4}$$

or

$$\left.\begin{aligned} u_m^{n+1} &= [1 + a(t_n)r]u_m^n - a(t_n)ru_{m-1}^n + \tau\psi_0(x_m, t_n), \\ u_m^0 &= \psi_1(x_m), \end{aligned}\right\} \tag{5}$$

where $x_m = mh$, $t_n = n\tau$, $r = \tau/h$, $a(t) \equiv -1 - 2t$. We will show that this scheme cannot be convergent for any step-size ratio r, since for any r it violates the Courant-Friedrichs-Levy condition.

Let us take, as point P, the point (0, 1). The net will be defined so that $N\tau = 1$. The value of the solution $u^{(h)} = u^{(h)}(P)$ at the point $P = (0, 1)$, i.e. u_0^N, by virtue of difference equation (5), is given in terms of $\psi_0(0, 1-\tau)$, and in terms of u_{-1}^{N-1}, u_0^{N-1}. These two values, in turn, are determined by $\psi_0(-h, 1 - 2\tau)$ and $\psi_0(0, 1-2\tau)$ and through the three values u_{-2}^{N-2}, u_{-1}^{N-2}, and u_0^{N-2}, etc. In the final analysis the value u_0^N can be expressed in terms of the values of the function $\psi_0(x, t)$ at the net points designated, in Fig. 16, by crosses, and in terms of the values of $u_{-N}^0 = \psi_1(x_{-N})$, $u_{-N+1}^0 = \psi_1(x_{-N+1})$, $\ldots$, $u_0^0 = \psi_1(x_0)$ of the function $\psi_1(x)$ at the points x_{-N}, x_{-N+1}, $\ldots$, x_0 on the x axis. Thus the set $G_{\psi_0}^{(h)}(P)$ consists of the net-points marked with crosses, and the set $G_{\psi_1}^{(h)}$ of the points x_{-N}, x_{-N+1}, $\ldots$, x_0 on the x axis (and it will be noted that these sets have points on the x axis in common). Obviously any point Q of the set $G_{\psi}(P)$ has a neighborhood which does not contain a point of the set $G_{\psi_0}^{(h)}(P)$, no matter how small we take h. The difference scheme (4) does not satisfy the Courant-Friedrichs-Levy condition, necessary for convergence.

We consider now, for problem (3), the difference scheme (Fig. 17)

$$L_h u^{(h)} \equiv \left\{ \begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} + a(t_b)\frac{u_{m+1}^n - u_m^n}{h} = \psi_0(x_m, t_n), \\ &u_m^0 = \psi_1(x_m), \quad m = 0, \pm 1, \ldots; \; n = 0, 1, \ldots, 1/\tau-1, \end{aligned} \right\} \tag{6}$$

or

$$\left.\begin{aligned} u_m^{n+1} &= [1 + a(t_n)r]u_m^n - a(t_n)ru_{m+1}^n + \tau\psi_0(x_m, t_n), \\ u_m^0 &= \psi_1(x_m), \end{aligned}\right\} \tag{7}$$

where $r = \tau/h$.

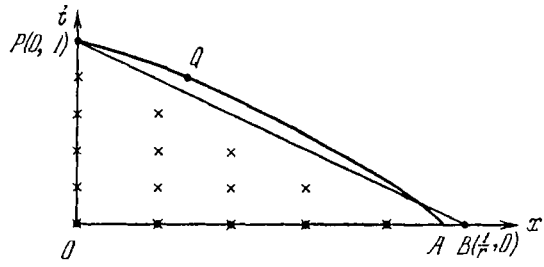


Fig. 17.

The step-size τ will be chosen to satisfy the condition $N\tau = 1$, with N a positive integer, so that the point P = (0, 1) will belong to the net. The value of the solution $u^{(h)}$ at this point, i.e. u_0^N, is expressed through Eq. (7) in terms of $\psi(0, 1-\tau)$, and of the two values u_0^{N-1} and u_1^{N-1}. These two values, in turn, through (7), are expressed in terms of $\psi_0(0, 1-2\tau)$, $\psi_0(h, 1-2\tau)$, and of the three values u_0^{N-2}, u_1^{N-2}, u_2^{N-2}. Finally u_0^N can be expressed in terms of the values of $\psi_0(x_m, t_n)$ at net points indicated in Fig. 17 by crosses, and of the values $u_0^0 = \psi_1(0)$, $u_1^0 = \psi_1(x_1)$, ..., $u_N^0 = \psi_1(x_N)$ at the points x_0, x_1, ..., x_N of the x axis. Thus $G_{\psi_0}^{(h)}(P)$, in this case, is the set of points marked by crosses, while $G_{\psi_1}^{(h)}(P)$ is the set of points x_0, x_1, ..., x_N on the axis. Clearly, if $r = \tau/h > 1/2$ (which is <u>not</u> the case depicted in the figure) the point B = (1/r, 0) lies to the left of the point $A = G_{\psi}(P)$. Therefore there exists a neighborhood of point A in which, as $h \to 0$, there are no points of $G_{\psi_1}^{(h)}(P)$. The Courant-Friedrichs-Levy condition is violated, and one cannot expect convergence.

So that it will be possible for scheme (6) to be convergent it is necessary that $r \leq 1/2$. But this is not enough. Suppose that $r < 1/2$, but that some point Q on the characteristic AQP lies above the line BP, as in Fig. 17. Then, again, one cannot expect convergence. The value of the function $\psi_0(x, t)$ at the point Q exerts an influence on the value u(0, 1) of the solution of the differential problem, i.e. Q belongs to the set $G_{\psi_0}(P)$. But the value $\psi_0(x, t)$ at point Q (like the values $\psi_0(x,t)$ on the whole segment QP of the characteristic) does not affect the value $u^{(h)}(P)$ of the solution of the difference equation at the point P: there exists a neighborhood of point Q into which, as $h \to 0$, no points of the set $G_{\psi_0}^{(h)}(P)$ will fall. The Courant-Friedrichs-Levy condition is not satisfied.

If r has been taken so small that the triangle OPB contains, not only the point A = (2, 0), but also the whole characteristic AQP, then it is already possible to prove the stability (and convergence) of difference scheme (6). To choose r in this way we note (since the differential equa-

tion of the characteristic is dx/dt = a(t)) that $-1/a(t) = \tan\theta$, where θ is the angle between the x axis and the tangent to the characteristic. It is easy to see that the characteristic AQP will lie in the triangle BOP if

$$r \leq \frac{1}{\max\limits_{0\leq t\leq 1} |a(t)|} = \frac{1}{3}, \qquad \tau \leq \frac{h}{3}, \tag{8}$$

and then the Courant-Friedrichs-Levy condition will be fulfilled.

Let us show that, under condition (8), difference scheme (6) which approximates Cauchy problem (3) is stable and, consequently, converges. For this purpose we define norms by means of the equations

$$||u^{(h)}||_{U_h} = \max_{m,n} \left|u_m^n\right| ,$$

$$||f^{(h)}||_{F_h} = \max_{m,n} |\psi_0(x_m, t_n)| + \max_n |\psi_1(x_m)|.$$

Noting that, from condition (8)

$$1 + a(t_n)r \geq 1 - \frac{2t_n + 1}{3} \geq 0, \qquad 0 \leq t_n \leq 1,$$

we get from (7)

$$\left|u_m^{n+1}\right| \leq \left[1 - \frac{2t_n + 1}{3} r + \frac{2t_n + 1}{3} r\right] \max_m \left|u_m^n\right| + \tau \max_{m,n} |\psi_0(x_m, t_n)|$$

$$\leq \max_m \left|u_m^n\right| + \tau \max_{m,n} |\psi_0(x_m, t_n)| \leq$$

$$\leq \max_m \left|u_m^{n-1}\right| + 2\tau \max_{m,n} |\psi_0(x_m, t_n)| \leq$$

. . . . . .

$$\leq \max_m \left|u_m^0\right| + (n + 1)\tau \max_{m,n} |\psi_0(x_m, t_n)| \leq$$

$$\leq \max_m |\psi_1(x_m)| + 1 \cdot \max_{m,n} |\psi_0(x_m, t_n)| = ||f^{(h)}||_{F_h}.$$

Since the final equation

$$\left|u_m^{n+1}\right| \leq ||f^{(h)}||_{F_h}$$

is valid for any $m = 0, \pm 1, \ldots$ and any n, $(n+1)\tau \leq 1$,

$$||u^{(h)}||_{U_h} \leq ||f^{(h)}||_{F_h},$$

and the stability of scheme (6) under condition (8) has been proven. The bound (8) on the step-size τ for given h, $\tau \leq 1/3\ h$, can be weakened without violating the Courant-Friedrichs-Levy condition if one takes τ to

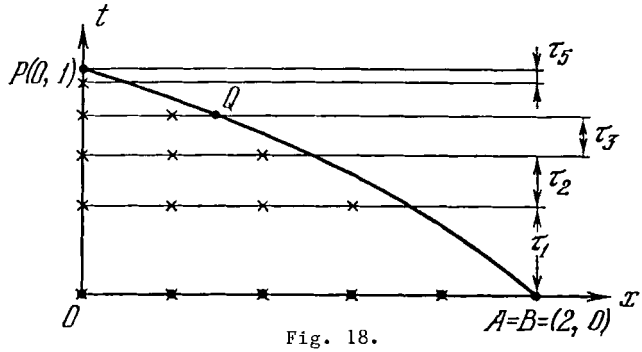


Fig. 18.

be variable, $t_{n+1} = t_n + \tau_n$, and chooses it, in the transition from t_n to t_{n+1}, taking into account the slope of the characteristic close to the point $t = t_n$, i.e. chooses it from the condition

$$r_n \equiv \frac{\tau_n}{h} \leq \frac{1}{|a(t_n)|} = \frac{1}{2t_n + 1}, \qquad n = 0, 1, \ldots \tag{9}$$

Thus modified, scheme (6) takes the form

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau_n} + a(t_n)\dfrac{u_{m+1}^n - u_m^n}{h} = \psi_0(x_m, t_n) \\ u_m^0 = \psi_1(x_m) \end{cases} \tag{10}$$

or

$$\left.\begin{aligned} u_m^{n+1} &= [1 + a(t_n)r_n]u_m^n = a(t_n)r_n u_{m+1}^n + \tau_n \psi_0(x_m, t_n), \\ u_m^0 &= \psi_1(x_m). \end{aligned}\right\} \tag{11}$$

The limitation on the step-size, τ_n, imposed by Eq. (9) is less severe than that which is required when using scheme (6) with constant step-size. For small n one uses the step-size $\tau_n \approx h$, and only when $u^{(h)}$ approaches $t = 1$ is it necessary to take $\tau_n = h/3$ (see Fig. 18). The proof of stability of scheme (10) under condition (9) differs only insignificantly from the proof of stability of (6) under condition (8); using the inequality $1 + a(t_n)r_n \geq 0$ we get from (11)

$$\left|u_m^{n+1}\right| \leq \max_m \left|u_m^n\right| + \tau_n \max_{m,n} |\psi_0(x_m, t_n)| \leq$$

$$\leq \max_m \left|u_m^{n-1}\right| + (\tau_{n-1} + \tau_n) \max_{m,n} |\psi_0(x_m, t_n)| \leq$$

$$\leq \max_m \left|u_m^0\right| + t_{n+1} \max_{m,n} |\psi_0(x_m, t_n)| \leq \|f^{(h)}\|_{F_h}.$$

It follows that

$$\|u^{(h)}\|_{U_h} \leq \|f^{(h)}\|_{F_h},$$

signifying stability.

3. Examples of difference schemes for the Dirichlet problem. Let us now use the Courant-Friedrichs-Levy condition for the analysis of two dif-

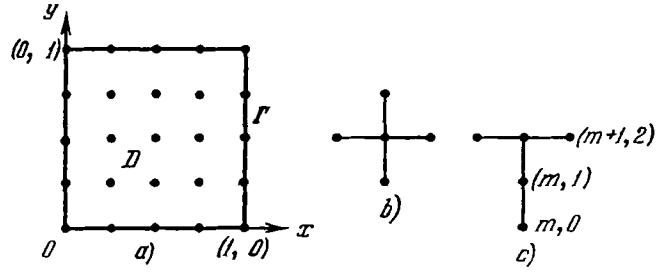


Fig. 19.

ference schemes approximating the following Dirichlet problem for the Poisson equation:

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= \phi(x, y), \qquad 0 \leq x, \quad y \leq 1, \\ u|_\Gamma &= \psi(x, y), \qquad (x, y) \text{ in } \Gamma, \end{aligned}\right\} \tag{12}$$

in the square region $D = (0 \leq x, y \leq 1)$, with boundary Γ. We construct the net $x_m = mh$, $y_n = nh$, where $h = 1/M$ with M an integer (Fig. 19,a). To the net D_h we assign those points, (x_m, y_n), which fall inside the square D, or on its boundary. Consider the following difference scheme, approximating problem (12):

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2} + \dfrac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2} \\ \qquad = \phi(mh, nh), \quad \text{if } (mh, nh) \text{ is in } D, \\ u_{mn} = \psi(mh, nh), \quad \text{if } (mh, nh) \text{ is in } \Gamma. \end{cases} \tag{13}$$

Scheme (13) is obtained by replacing the derivatives u_{xx} and u_{yy} by difference relations, and there can be no doubt about approximation. We will prove its stability in §34 and discuss methods for computing its solution $u^{(h)}$ in §§35-37. But we point out that computation of this solution is not a trivial matter, since the system of equations $L_h u^{(h)} = f^{(h)}$ which determines the values of the net-function, $u^{(h)}$, is rather complicated when h is small. This very complexity leads one to consider whether it is possible to construct a scheme such that the numerical solution process will be simple. At first glance it appears that one can use the scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m-1,n} - 2u_{mn} + u_{m+1,n}}{h^2} + \dfrac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2} = \phi(mh, nh), \\ \qquad m = 1, 2, \ldots, M-1; \quad n = 1, 2, \ldots, M-2, \\ \dfrac{u_{m-1,2} - 2u_{m2} + u_{m+1,2}}{h^2} + \dfrac{u_{m2} - 2u_{m1} + u_{m0}}{h^2} = \phi(mh, h), \\ \qquad m = 1, 2, \ldots, M-1, \\ u_{mn} = \psi(mh, nh), \quad (x, y) \text{ in } \Gamma. \end{cases} \tag{14}$$

Obviously this scheme does approximate the differential problem, since it is obtained by replacing derivatives with difference relations and the boundary conditions are represented exactly. Each equation of the first group connects the values of the solution at the five net-points shown in Fig. 19,b. The second group of equations, for fixed m, connects the solution-values at the five net-points shown in Fig. 19,c.

Consider the set of equations of the first group corresponding to a fixed n, in fact to $n = 1$, together with the whole second group of equations. The resulting system of equations connects the quantities u_{m1}, u_{m2} and u_{m0}, while u_{m0}, u_{01}, u_{02}, u_{M1} and u_{M2} are given by the boundary conditions. This system can be solved for u_{m1} and u_{m2}, $m = 1, 2, \ldots, M-1$. Then we use the difference equation from the first group for $n = 2$, and determine u_{m3} via the explicit formula obtained by solving this equation for the only unknown quantity which it contains, i.e. u_{m3}. Proceeding level by level from u_{mn} to $u_{m,n+1}$ we compute, via the equations of the first group, the solution $u^{(h)}$ at all interior points of the net. Of course the values at boundary points are known from the start.

However this, at first glance seemingly convenient, scheme is completely unuseable. We know that the solution of the Dirichlet problem for the Laplace equation depends, at each point, on the values $\psi(x,y)|_\Gamma$ everywhere on the boundary. In contrast, in the computational scheme we have constructed the computation of the solution $u^{(h)}$ at all internal points proceeds without using the values $\psi(x,y)$ on the upper surface of the square. This difference scheme cannot be convergent. The complexities of scheme (13) are essential to the problem.

* * * * * *

In conclusion we stress again that the Courant-Friedrichs-Levy condition is not a sufficient condition for stability. In §25 we will show, in particular, that the difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_{m-1}^n}{2h} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh) \end{cases}$$

is unstable for any $r = \tau/h = \text{const}$. This scheme approximates the Cauchy problem

$$u_t - u_x = \phi(x, t),$$

$$u(x, 0) = \psi(x),$$

for which we have already considered several other schemes. It is easy to verify that, for $r < 1$, it also satisfies the necessary condition for stability.

In order to do this we again consider, for the sake of definiteness, the point (0, 1) in the x,t plane, assuming that it belongs to the net D_h for all h, so that $N\tau = 1$, where N is an integer. The value u_0^N is computed from the values u_{-1}^{N-1}, u_0^{N-1} and u_1^{N-1}. These three values are then computed from five values at the preceding level $t = (N-2)\tau$, etc. Ultimately u_0^N is computed, then, in terms of the values $u_m^0 = \psi(mh)$, $m = -N, -N+1, \ldots, -1, 0, 1, \ldots, N$, on the net points which belong to the interval $-1/r \leq x \leq 1/r$ of the x axis. If $r = \tau/h < 1$, then this interval contains the point $x = 1$ where the solution value is defined by $u(0,1)$, $u(0, 1) = \psi(1)$. Thus for $r \leq 1$ the Courant-Friedrichs-Levy condition is satisfied.

* * *

PROBLEMS

1. The solution of the heat-conduction problem $u_t = u_{xx}$, $-\infty < x < \infty$, $t > 0$ has the form

$$u(x, t) = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{t}} u(\xi, 0)\, e^{-\frac{(x-\xi)^2}{4t}}\, d\xi.$$

Does there exist a convergent difference scheme approximating this problem, and having the form

$$\frac{u_m^{p+1} - u_m^p}{\tau} = \frac{1}{h^2}\left(\alpha_{-2}u_{m-2}^p + \alpha_{-1}u_{m-1}^p + \alpha_0 u_m^p + \alpha_1 u_{m+1}^p + \alpha_2 u_{m+2}^p\right),$$

(where the α_i are constants) if $\tau = h$?

2. The system of acoustic equations

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} + \frac{\partial w}{\partial x} = 0, \\ &\frac{\partial w}{\partial t} + \frac{\partial v}{\partial x} = 0, \qquad t > 0, \quad -\infty < x < \infty, \\ &v(x, 0) = \phi(x), \; w(x, 0) = \psi(x) \end{aligned}\right\}$$

has a solution of the form

$$\left.\begin{aligned} v(x, t) &= \frac{\phi(x - t) + \psi(x - t) + \phi(x + t) - \psi(x + t)}{2}, \\ w(x, t) &= \frac{\phi(x - t) + \psi(x - t) - \phi(x + t) + \psi(x + t)}{2}. \end{aligned}\right\}$$

Can there be a convergent difference scheme of the form

$$\left.\begin{aligned} &\frac{v_m^{p+1} - v_m^p}{\tau} + \frac{w_{m+1}^p - w_m^p}{h} = 0, \quad p \geq 0, \quad m = 0, \pm 1, \ldots, \\ &\frac{w_m^{p+1} - w_m^p}{\tau} + \frac{v_{m+1}^p - v_m^p}{h} = 0 \\ &v_m^0 = \phi(x_m), \quad w_m^0 = \psi(x_m)? \end{aligned}\right\}$$

Compare the domain of influence of starting values for the difference and differential problems.

3. The Cauchy problem

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \qquad t > 0, \quad -\infty < x < \infty,$$

$$u(x, 0) = e^{i\alpha x}, \qquad -\infty < x < \infty,$$

has the solution

$$u(x, t) = e^{i\alpha t} e^{i\alpha x}.$$

The corresponding difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad p = 0, 1, \ldots, \\ &u_m^0 = e^{i\alpha h m}, \qquad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

has the solution

$$u_m^p = \left[1 - r + re^{i\alpha h}\right]^p e^{i\alpha h m},$$

which, for $p = t/\tau$, $m = x/h$, tends to the solution of the differential problem as $h \to 0$, whatever the preassigned, fixed, value of $r = \tau/h$. Nevertheless for $r > 1$ the difference scheme does not satisfy the Courant-Friedrichs-Levy condition. Explain this apparent paradox.

This Page Intentionally Left Blank

Chapter 8

Some Basic Methods for the Study of Stability

§25. Spectral analysis of the Cauchy difference problem

Here we develop the Von Neumann method, useful in a wide range of circumstances for the study of difference problems with initial conditions. In this section we limit our discussion to the case of the Cauchy difference problem with constant coefficients, and in §26 we partially extend our results to the case of variable coefficients.

1. Stability with respect to starting values. As the simplest example of a Cauchy difference scheme we take the problem, often considered above,

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - u_m^p}{h} = \phi_m^p, \quad p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0 = \psi_m, \quad m = 0, \pm 1, \ldots \end{cases} \tag{1}$$

Setting

$$f^{(h)} = \begin{cases} \phi_m^p, & p = 0, 1, \ldots, [T/t]-1, \\ \psi_m, & m = 0, \pm 1, \ldots, \end{cases}$$

we write problem (1) in the form

$$L_h u^{(h)} = f^{(h)}. \tag{2}$$

We will define the norms $\|u^{(h)}\|_{U_h}$ and $\|f^{(h)}\|_{F_h}$ via the equations

$$\|u^{(h)}\|_{U_h} = \max_p \max_m \left|u_m^p\right|, \quad \|f^{(h)}\|_{F_h} = \max_m |\psi_m| + \max_{m,p} \left|\phi_m^p\right|.$$

The stability condition for problem (2)

$$\|u^{(h)}\|_{U_h} \leq c \|f^{(h)}\|_{F_h} \tag{3}$$

then takes the form

$$\max_m \left|u_m^p\right| \le c\left[\max_m |\psi_m| + \max_{m,k} \left|\phi_m^k\right|\right], \qquad p = 0, 1, \ldots, [T/\tau], \tag{4}$$

where c does not depend on h (nor on $\tau = rh$, $r = \text{const}$). Condition (4) must be satisfied for arbitrary $\{\psi_m\}$ and $\{\phi_m^p\}$. In particular, for stability it is necessary that it be fulfilled for arbitrary $\{\psi_m\}$ and $\phi_m^p \equiv 0$, i.e. that the solution of the problem

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad p = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi_m, \qquad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{5}$$

satisfy the condition

$$\max_m \left|u_m^p\right| \le c \max_m \left|u_m^0\right|, \qquad p = 0, 1, \ldots, [T/\tau], \tag{6}$$

for any arbitrary, bounded, function $u_m^0 = \psi_m$.

Property (6), necessary for the stability (4) of problem (1), is called *stability of problem (1) with respect to perturbations in starting values*. It indicates that a perturbation in $\{u_m^0\}$, the starting values of problem (1), induces a perturbation in $\{u_m^p\}$, the solution of problem (1) which, by virtue of (6), is no greater than c times greater than the perturbation in starting values, where c does not depend on h.

2. Necessary spectral condition for stability. For the stability of problem (1) with respect to starting data it is necessary that condition (6) be fulfilled, in particular, when $\{u_m^0\}$ is any harmonic

$$u_m^0 = e^{i\alpha m}, \qquad m = 0, \pm 1, \ldots, \tag{7}$$

where α is a real parameter. But the solution of problem (5) for initial conditions (7) has the form

$$u_m^p = \lambda^p e^{i\alpha m}, \tag{8}$$

where $\lambda = \lambda(\alpha)$ is determined by substituting expression (8) into the homogeneous difference equation of problem (5):

$$\lambda(\alpha) = 1 - r + re^{i\alpha}, \qquad r = \frac{\tau}{h} = \text{const}. \tag{9}$$

For the solution (8) we may write

$$\max_m \left|u_m^p\right| = |\lambda(\alpha)|^p \max_m \left|u_m^0\right|.$$

Therefore, if condition (6) is to be satisfied, it is necessary that, for all real α, we have

$$|\lambda(\alpha)|^p \leq c, \qquad p = 0, 1, \ldots, [T/\tau],$$

or

$$|\lambda(\alpha)| \leq 1 + c_1\tau, \tag{10}$$

where c_1 is some constant not depending on α or τ. Precisely this is the *necessary spectral condition of Von Neumann* as applied to the example under consideration. It is called a "spectral" condition for the following reason. The existence of a solution of the form (8) shows that the harmonic $\{e^{i\alpha m}\}$ is a proper function of the transition operator

$$u_m^{p+1} = (1 - r)u_m^p + ru_{m+1}^p, \qquad m = 0, \pm 1, \ldots,$$

which, according to difference equation (5), maps the net function $\{u_m^p\}$, $m = 0, \pm 1, \ldots$, defined on level $t_p = p\tau$ of the net, into the function $\{u_m^{p+1}\}$, $m = 0, \pm 1, \ldots$, defined on the level $t_{p+1} = (p + 1)\tau$. The number $\lambda(\alpha) = 1 - r + re^{i\alpha}$ is the eigenvalue of the transition operator corresponding to the harmonic $\{e^{i\alpha m}\}$. The curve described in the complex plane by the point, $\lambda(\alpha)$, when α traverses the real axis, consists entirely of eigenvalues, and is the spectrum of the transition operator.

Thus the necessary condition for the stability of (10) can be stated as follows: the spectrum of the transition operator corresponding to difference problem (5) must lie in a circle of radius $1 + c_1\tau$ in the complex plane. In our example the spectrum of (9) does not depend on τ. For this reason condition (10) is equivalent to the requirement that the spectrum, $\lambda(\alpha)$, lie inside the unit circle

$$|\lambda(\alpha)| \leq 1. \tag{11}$$

Let us now use the above-formulated criterion to analyze the stability of problem (1). The spectrum (9) constitutes a circle, with center at the point $1 - r$ and radius r, in the complex plane. In the case $r < 1$ this region lies inside the unit circle (and is tangent to it at the point $\lambda = 1$); for $r = 1$ it coincides with the unit circle, and for $r > 1$ lies outside the unit circle (Fig. 20). Correspondingly the necessary condition for stability

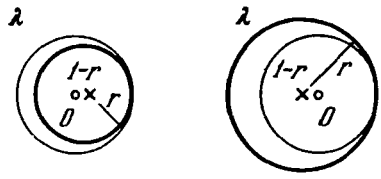


Fig 20.

(11) is fulfilled for $r \leq 1$, and not fulfilled when $r > 1$. In 3§21 we studied this same difference scheme and showed that, for $r \leq 1$, it is stable, and is unstable for $r > 1$. Thus the necessary Von Neumann stability condition turns out to be sensitive enough, in this particular case, so as to separate precisely the region of stability from the region of instability.

In the case of the general Cauchy problem for difference equations, or systems of difference equations, the necessary Von Neumann condition for stability consists in that the spectrum $\lambda = \lambda(\alpha, h)$ of the difference problem, for all sufficiently small h, must lie in the circle

$$|\lambda| \leq 1 + \varepsilon \tag{12}$$

of the complex plane no matter how small the previously-specified positive ε.

Note that if, for the given difference problem the spectrum turns out not to depend on h (or on τ), then condition (12) is equivalent to the requirement that the spectrum, $\lambda = \lambda(\alpha, h) = \lambda(\alpha)$, must lie in the unit circle

$$|\lambda(\alpha)| \leq 1. \tag{12'}$$

By the "spectrum" of the difference problem, referred to in (12), is meant the totality of all $\lambda = \lambda(\alpha, h)$ for which the corresponding homogeneous difference equation (or system of equations) has a solution of the form

$$u_m^p = [\lambda(\alpha, h)]^p [u^0 e^{i\alpha m}], \qquad m = 0, \pm 1, \ldots, \tag{13}$$

where u^0 is a number (unity) if we are dealing with a scalar difference equation, and is a vector if the equation in question is a vector difference equation, i.e. a system of scalar difference equations.

If the necessary Von Neumann condition (12) is not satisfied then one cannot expect stability for <u>any</u> reasonable choice of norms, and if it is satisfied one may hope to achieve stability for <u>some</u> reasonably defined norms. A similar point regarding the indifference of the spectral stability criterion to the choice of norms has already been discussed, in connection with difference schemes for ordinary differential equations, in §15.

3. Examples. We will now consider a series of interesting Cauchy difference problems, and will use the Von Neumann spectral criterion to analyze stability. We start with difference schemes approximating the Cauchy differential problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \phi(x, t), \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty. \end{aligned}\right\} \tag{14}$$

Example 1. Consider the difference scheme

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_m^p - u_{m-1}^p}{h} = \phi(\psi_m, t_p), \qquad p = 0, 1, \ldots, [T/\tau]-1,$$

$$u_m^0 = \psi(x_m), \qquad m = 0, \pm 1, \ldots$$

Substituting an expression of form (8) into the corresponding homogeneous equation we get, after some simple manipulations

$$\lambda(\alpha) = 1 + r - re^{-i\alpha}.$$

It follows that the spectrum constitutes the perimeter of a circle, centered at the point 1 + r, with radius r (Fig. 21). There is no r for which the spectrum lies in the unit circle. Stability criterion (12′) is never satisfied.

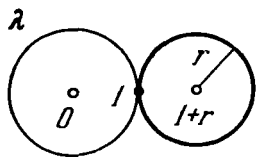


Fig. 21.

In §24 it has already been established that, for any r, the Courant-Friedrichs-Levy necessary condition for convergence (and stability) is violated.

Example 2. Consider the following difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2h^2}\left(u_{m-1}^p - 2u_m^p + u_{m+1}^p\right) = \phi(x_m, t_p), \\ &u_m^0 = \psi_m, \end{aligned}\right\} \tag{15}$$

approximating problem (14) to second order in h (§22). For this scheme $\lambda = \lambda(\alpha)$ is determined by the equation

$$\frac{\lambda - 1}{\tau} - \frac{e^{i\alpha} - e^{-i\alpha}}{2h} - \frac{\tau}{2h^2}\left(e^{i\alpha} - 2 + e^{-i\alpha}\right) = 0.$$

As before, let $r = \tau/h$. Noting that

$$\frac{e^{i\alpha} - e^{-i\alpha}}{2i} = \sin\alpha,$$

$$\frac{e^{i\alpha} - 2 + e^{-i\alpha}}{4} = -\left(\frac{e^{i\alpha/2} - e^{-i\alpha/2}}{2i}\right)^2 = -\sin^2\frac{\alpha}{2},$$

we get

$$\lambda = 1 + ir \sin \alpha - 2r^2 \sin^2 \frac{\alpha}{2}, \tag{16}$$

$$|\lambda(\alpha)|^2 = \left(1 - 2r^2 \sin^2 \frac{\alpha}{2}\right)^2 + r^2 \sin^2 \alpha.$$

After some simple manipulation

$$1 - |\lambda|^2 = 4r^2 \sin^4 \frac{\alpha}{2} (1 - r^2). \tag{17}$$

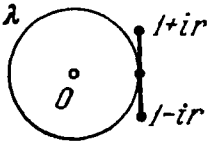
The Von Neumann condition is satisfied if the right-hand side is non-negative, $r \leq 1$, and is not satisfied for $r > 1$.

Example 3. Consider the following difference scheme

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - u_{m-1}^p}{2h} = \phi(x_m, t_p), \\ u_m^0 = \psi(x_m) \end{cases} \tag{18}$$

for the same Cauchy problem (14).

Substituting expression (8) into Eq. (18), after some simplification we get an equation for λ:



or

$$\frac{\lambda - 1}{\tau} - \frac{e^{i\alpha} - e^{-i\alpha}}{2h} = 0$$

$$\lambda(\alpha) = 1 + i\left(\frac{\tau}{h} \sin \alpha\right).$$

Fig. 22. The spectrum $\lambda = \lambda(a)$ fills a vertical segment of length $2\tau/h$, passing through the point $\lambda = 1$ (Fig. 22).

If $\tau/h = r = \text{const}$, then condition (12′) is not satisfied; the spectrum does not lie in the unit circle. If, as $h \to 0$, the step-size τ varies like h^2, so that $\tau = rh^2$, then the point $\lambda(\alpha)$ farthest from $\lambda = 0$ represents an eigenvalue of modulus

$$|\lambda(\alpha)|_{\alpha=\pi/2} = \sqrt{1 + \left(\frac{\tau}{h}\right)^2} = \sqrt{1 + \tau r} \leq 1 + \frac{r}{2}\tau.$$

The condition $|\lambda(\alpha)| \leq 1 + c\tau$ is satisfied, in this case, with $c = r/2$.

Clearly the requirement $\tau = rh^2$ puts a much more severe condition on the reduction of the time step-size, τ, as h tends to zero, then does the requirement $\tau = rh$, $r \leq 1$, which was sufficient to guarantee satisfaction of the Von Neumann condition for schemes (5) and (15), approximating the same Cauchy problem (14).

Note that the Courant-Friedrichs-Levy criterion (as shown at the end of §24) allows us to ascertain the instability of the scheme under consid-

eration only for $\tau/h > 1$, while for $\tau/h \leq 1$ it is inconclusive, so that it turns out, here, to be weaker than the Von Neumann criterion.

Next we consider two difference schemes, constructed in §22, approximating the Cauchy problem for the heat equation

$$\left.\begin{aligned} u_t - a^2 u_{xx} &= \phi(x, t), \qquad -\infty < x < \infty, \quad 0 < t < T, \\ u(x, 0) &= \psi(x), \qquad -\infty < x < \infty . \end{aligned}\right\} \tag{19}$$

Example 4. The explicit difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - a^2 \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau]-1, \end{cases}$$

(on substitution of the expression $u_m^p = \lambda^p e^{i\alpha m}$ into the corresponding homogeneous difference equation) leads to the relation

$$\frac{\lambda - 1}{\tau} - a^2 \frac{e^{-i\alpha} - 2 + e^{i\alpha}}{h^2} = 0.$$

Noting that

$$\frac{e^{-i\alpha} - 2 + e^{i\alpha}}{4} = -\left(\frac{e^{i\alpha/2} - e^{-i\alpha/2}}{2i}\right)^2 = -\sin^2\frac{\alpha}{2},$$

we get

$$\lambda(\alpha) = 1 - 4ra^2 \sin^2\frac{\alpha}{2}, \qquad r = \frac{\tau}{h^2}.$$

As α varies the quantity $\lambda(\alpha)$ traverses the segment $1 - 4ra^2 \leq \lambda \leq 1$ of the real axis (Fig. 23). For stability it is necessary that the left end of this segment lie in the unit circle so that $1 - 4ra^2 \geq -1$, or

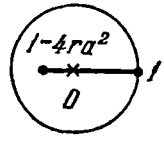


Fig. 23.

$$r \leq \frac{1}{2a^2}. \tag{20}$$

If $r > 1/2a^2$, the point $\lambda(\alpha) = 1 - 4ra^2 \sin^2(\alpha/2)$ corresponding to $\alpha = \pi$ lies to the left of the point -1. The harmonic $\exp(i\pi m) = (-1)^m$ gives rise to the solution

$$u_m^p = (1 - 4a^2 r)^p (-1)^m,$$

not satisfying condition (6) for any constant c.

Example 5. We come now to the second scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - a^2 \dfrac{u_{m+1}^{p+1} - 2u_m^{p+1} + u_{m-1}^{p+1}}{h^2} = \phi(mh, n\tau), \\ u_m^0 = \psi(mh), \\ m = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau]-1. \end{cases} \tag{21}$$

Analogous calculations lead to the expression

$$\lambda(a) = \frac{1}{1 + 4ra^2 \sin^2 \frac{a}{2}}, \qquad r = \frac{\tau}{h^2}. \tag{22}$$

The spectrum for this problem fills the segment

$$\left[1 + 4ra^2 \sin^2 \frac{\alpha}{2}\right]^{-1} \leq \lambda \leq 1$$

of the real axis, and the condition $|\lambda| \leq 1$ is satisfied for all r.

The Von Neumann spectral criterion may be used for the study of the Cauchy difference problem also in the case where there are two or more space variables.

Example 6. For the problem

$$\left.\begin{aligned} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \qquad t > 0, \\ u(x, y, t) = \psi(x, y) \end{aligned}\right\}$$

we take the net $(x_m, y_n, t_p) = (mh, nh, p\tau)$. Replacing derivatives by difference relations we construct the difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{mn}^{p+1} - u_{mn}^p}{\tau} - \dfrac{u_{m+1,n}^p - 2u_{mn}^p + u_{m-1,n}^p}{h^2} - \\ \qquad - \dfrac{u_{m,n+1}^p - 2u_{mn}^p + u_{m,n-1}^p}{h^2} = 0, \\ u_{mn}^0 = \psi(mh, nh). \end{cases} \tag{23}$$

Taking $u_{mn}^0 = \exp[i(\alpha m + \beta n)]$, i.e. postulating a solution in the form of a two-dimensional harmonic depending on the two real parameters α and β, we get a solution of the form

$$u_{mn}^p = \lambda^p(\alpha, \beta) e^{i(\alpha m + \beta n)}.$$

Substituting this expression into the difference equation, after some simplification and identity-transformations we find that

$$\lambda(\alpha, \beta) = 1 - 4r \sin^2 \frac{\alpha}{2} - 4r \sin^2 \frac{\beta}{2}.$$

As the real α and β vary the point $\lambda = \lambda(\alpha, \beta)$ traverses the segment

$$1 - 8r \leq \lambda \leq 1$$

of the real axis. The stability condition is satisfied if $1 - 8r \geq -1$, $r \leq 1/4$.

Now we present an example illustrating the application of the Von Neumann criterion to difference equations connecting the values of the unknown function, not at two, but at three time levels.

Example 7. The Cauchy problem for the wave equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \qquad -\infty < x < \infty, \quad 0 < t < T,$$

$$u(x, 0) = \psi_1(x), \qquad \frac{\partial u(x, 0)}{\partial t} = \psi_2(x), \qquad -\infty < x < \infty,$$

will be approximated by the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{\tau^2} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ &\qquad p = 1, 2, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi_1(x_m), \quad \frac{u_m^1 - u_m^0}{\tau} = \psi_2(x_m), \qquad m = 0, \pm 1, \ldots \end{aligned}\right\} \tag{24}$$

Substituting, into the difference equation, a solution of form (8) we get, after simple transformations, the following equation for determining λ:

$$\lambda^2 - 2\left(1 - 2r^2 \sin^2 \frac{\alpha}{2}\right)\lambda + 1 = 0, \qquad r = \frac{\tau}{h}.$$

The product of the roots of this equation is equal to one. If the discriminant

$$d(\alpha) = 4r^2 \sin^2\alpha \left(r^2 \sin^2 \frac{\alpha}{2} - 1\right)$$

of the quadratic equation is negative, then the roots, $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$, will be complex conjugates,

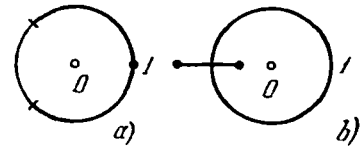


Fig. 24.

and equal to one in modulus. When $r < 1$ the discriminant remains negative for all α. In Fig. 24,a we show the spectrum in this case. It fills part of the circumference of the unit circle. In the case r=1 the spectrum fills all of this circumference. For $r > 1$, as α increases from 0 to π the roots $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ move, from the end point $\lambda = 1$, along the circumference of the unit circle, one in the clockwise and the other in the counterclockwise direction, until they meet at the point $\lambda = -1$; then one of the roots moves along the real axis from the point $\lambda = -1$ to the left, the other to the right, since both are real and $\lambda_1\lambda_2 = 1$ (Fig. 24,b). The stability condition is satisfied for $r \leq 1$.

Let us consider the Cauchy problem for the following hyperbolic system of differential equations, describing the propagation of sound:

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x}, \\ &\frac{\partial w}{\partial t} = \frac{\partial v}{\partial x}, \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &v(x, 0) = \psi_1(x), \quad w(x, 0) = \psi_2(x), \quad -\infty < x < \infty. \end{aligned}\right\} \tag{25}$$

We set

$$u(x, t) = \begin{pmatrix} v(x, t) \\ w(x, t) \end{pmatrix}, \qquad \psi(x) = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \end{pmatrix}$$

and write (25) in the vector form

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - A\frac{\partial u}{\partial x} = 0, \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\} \tag{25'}$$

where

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

We will study two difference schemes approximating problem (25').

Example 8. Consider the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - A\frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad p = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi(x_m), \qquad m = 0, \pm 1, \ldots \end{aligned}\right\} \tag{26}$$

We seek a solution of the homogeneous vector equation of the form (13):

$$u_m^p = \lambda^p(u^0 e^{i\alpha m}) = \lambda^p\begin{pmatrix} v^0 \\ w^0 \end{pmatrix} e^{i\alpha m}.$$

Substituting this equation into difference equation (26) we arrive at the equation

$$\frac{\lambda - 1}{\tau} u^0 - A \frac{e^{i\alpha} - 1}{h} u^0 = 0,$$

or

$$(\lambda - 1)u^0 - r(e^{i\alpha} - 1)Au^0 = 0, \quad r = \frac{\tau}{h}, \tag{27}$$

which one may regard as a system of linear equations, in vector notation, for the determination of the components of the vector u^0.

Let us write the system (27) in expanded form:

$$\begin{pmatrix} \lambda - 1 & -r(e^{i\alpha} - 1) \\ -r(e^{i\alpha} - 1) & \lambda - 1 \end{pmatrix} \begin{pmatrix} v^0 \\ w^0 \end{pmatrix} = 0. \tag{28}$$

The system of linear equations (28) has a nontrivial solution, $u^0 = (v^0, w^0)^T$, only for those $\lambda = \lambda(\alpha)$ for which the determinant of system (28) vanishes:

$$(\lambda - 1)^2 = r^2(e^{i\alpha} - 1)^2.$$

Therefore

$$\lambda_1(\alpha) = 1 - r + re^{i\alpha},$$

$$\lambda_2(\alpha) = 1 + r - re^{i\alpha}.$$

The roots $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ move along circles of radius r, centered at the points $1 - r$ and $1 + r$, respectively (Fig. 25). The Von Neumann stability condition is not satisfied for any r.

Example 9. Consider the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - A \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2h^2} A^2(u_{m+1}^p - 2u_m^p + u_{m-1}^p) = 0, \\ &\qquad p = 0, 1, \ldots, [T/\tau]-1; \quad m = 0, \pm 1, \ldots, \\ &\qquad u_m^0 = \psi(x_m), \qquad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{29}$$

approximating problem (25′) to second order, and analogous to scheme (15) for the scalar case (14). The condition for existence of a nontrivial

solution, in form (13), of vector Eq. (29) consists (as in example 8) in that the determinant of the system which fixes $u_0 = (v^0, w^0)^T$ must vanish. Setting this determinant to zero we get a quadratic equation for $\lambda = \lambda(\alpha)$, from which we find that

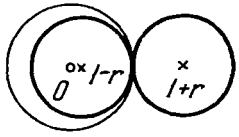


Fig. 25.

$$\left.\begin{aligned} \lambda_1 &= 1 + ir \sin\alpha - 2r^2 \sin^2\frac{\alpha}{2}, \\ \lambda_2 &= 1 - ir \sin\alpha - 2r^2 \sin^2\frac{\alpha}{2}. \end{aligned}\right\} \tag{30}$$

These expressions are analogous to (16), and as in (17) we get

$$1 - |\lambda_{1,2}(\alpha)|^2 = 4r^2 \sin^4\frac{\alpha}{2}(1 - r^2).$$

The spectrum given by Eqs. (30) lies in the unit circle for $r \leq 1$.

4. Integral representation of the solution.* Consider the Cauchy problem of the form

$$\left.\begin{aligned} b_{-1}u_{m-1}^{p+1} + b_0 u_m^{p+1} + b_1 u_{m+1}^{p+1} - \\ - \left(a_{-1}u_{m-1}^{p} + a_0 u_m^{p} + a_1 u_{m+1}^{p}\right) = \tau\phi_m^p, \\ p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0 = \psi_m, \qquad m = 0, \pm1, \ldots, \end{aligned}\right\} \tag{31}$$

with constant coefficients, assuming that

$$b_{-1}e^{-i\alpha} + b_0 + b_1 e^{i\alpha} \neq 0, \qquad 0 \leq \alpha \leq 2\pi. \tag{32}$$

Difference schemes (1), (15), (18) and (21) take on this form if both sides of the difference equations involved in these schemes are multiplied by τ.

We note first of all that, for arbitrary bounded net functions $\{\phi_m^p\}$ and $\{\psi_m\}$, problem (31) has one and only one bounded solution. In fact if it is already known that $\{u_m^p\}$, for a given fixed p, exists and is bounded, then Eq. (31) becomes an ordinary difference equation of second order

*Results obtained in this section are not used later.

$$b_{-1}u_{m-1}^{p+1} + b_0 u_m^{p+1} + b_1 u_{m+1}^{p+1} = \tau\phi_m^p + \left(a_{-1}u_{m-1}^p + a_0 u_m^p + a_1 u_{m+1}^p\right) \tag{33}$$

for $\{u_m^{p+1}\}$, with bounded right-hand side. The corresponding characteristic equation $b_{-1} + b_0 q + b_1 q^2 = 0$, thanks to (32), has no roots, $q = e^{i\alpha}$, equal to one in modulus. Therefore, as shown at the end of 2§3, it has a unique bounded solution $\{u_m^{p+1}\}$. But $\{u_m^0\} = \{\psi_m\}$ is given and bounded; therefore Eq. (33) sequentially and uniquely determines bounded functions $\{u_m^1\}$, $\{u_m^2\}$, etc.

Below we will need the following well-known fact about Fourier series.

Each sequence of numbers c_m, $m = 0, \pm 1, \ldots$, for which $\sum |c_m| < \infty$ corresponds to a convergent (in the mean-square sense) Fourier series

$$C(\alpha) = \frac{1}{\sqrt{2\pi}} \sum c_m e^{-i\alpha m}, \tag{34}$$

the sum of which is a function, $C(\alpha)$, square-integrable on the interval $0 \leq \alpha \leq 2\pi$,

$$\int_0^{2\pi} |C(\alpha)|^2 \, d\alpha < \infty.$$

Conversely, every function, $C(\alpha)$, square-integrable on the interval $0 \leq \alpha \leq 2\pi$, is expandable in a unique Fourier series (34), with coefficients c_m given by the equation

$$c_m = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} C(\alpha) e^{i\alpha m} d\alpha. \tag{35}$$

Further, these coefficients satisfy Parseval's equation

$$\int_0^{2\pi} |C(\alpha)| d\alpha = \sum |c_m|^2 . \tag{36}$$

Theorem 1. *Suppose that, in Problem (31)*

$$\max_p \sum_m \left|\phi_m^p\right|^2 < \infty, \qquad \sum_m |\psi_m|^2 < \infty.$$

Then the bounded solution of this problem admits the integral representation

$$u_m^p = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} U^p(\alpha) e^{i\alpha m} d\alpha, \tag{37}$$

where the square-integrable function $U^p(\alpha)$ is defined by the recurrence relation

$$U^{p+1}(\alpha) = \lambda(\alpha)U^p(\alpha) + \frac{\tau}{b_{-1}e^{i\alpha} + b_0 + b_1 e^{-i\alpha}} \bar{\phi}^p(\alpha), \quad p = 0, 1, \ldots \tag{38}$$

Here

$$\bar{\phi}^p(\alpha) = \frac{1}{\sqrt{2\pi}} \sum_m \phi_m^p e^{-i\alpha m}; \qquad U^0(\alpha) \equiv \bar{\psi}(\alpha) = \frac{1}{\sqrt{2\pi}} \sum_m \psi_m e^{-i\alpha m},$$

and the function

$$\lambda(\alpha) = \frac{a_1 e^{-i\alpha} + a_0 + a_{-1} e^{i\alpha}}{b_1 e^{-i\alpha} + b_0 + b_{-1} e^{i\alpha}}$$

is so chosen that, for each α, $0 \le \alpha \le 2\pi$, *the net function* $u_m^p = \lambda^p(\alpha) \exp(i\alpha m)$ *satisfies the homogeneous equation corresponding to Eq. (31).*

We may prove this theorem by direct substitution of expression (37) into Eqs. (31), making use of Eqs. (34) and (35).

Consequences. If, in (31), the function $\phi_m^p \equiv 0$, then $\bar{\phi}^p(\alpha) = 0$; by virtue of (38) we have $U^p(\alpha) = \lambda^p(\alpha)\,\bar{\psi}(\alpha)$, and from (37) it follows that

$$u_m^p = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \lambda^p(\alpha)\,\bar{\psi}(\alpha) e^{i\alpha m} d\alpha. \tag{39}$$

The integral representations, (37) and (39), can be used to analyze the properties of difference scheme (31).

We define norms via the equations

$$\left.\begin{aligned} &||u^p||^2 = \sum_m \left|u_m^p\right|^2; \qquad ||u^{(h)}||_{U_h} = \max_p ||u^p||, \\ &||\phi^p||^2 = \sum_m \left|\phi_m^p\right|^2; \qquad ||\psi||^2 = \sum |\psi_m|^2, \\ &||f^{(h)}||_{F_h} = \left|\left|\begin{matrix}\phi_m^p \\ \psi_m\end{matrix}\right|\right|_{F_h} = ||\psi|| + \max_p ||\phi^p||. \end{aligned}\right\} \tag{40}$$

Theorem 2. *For the stability of difference scheme (31) with respect to initial conditions, i.e. to guarantee the validity of the inequality*

$$||u^p|| \le c||u^0||, \qquad p = 0, 1, \ldots, \left[\frac{T}{\tau}\right],$$

for arbitrary $u_m^0 = \psi_m$, $||\psi|| < \infty$ *and* $\phi_m^p \equiv 0$, *with constant* c *not depending on* h *(or on* $\tau = \tau(h)$*), it is necessary and sufficient that the spectrum* $\lambda = \lambda(\alpha)$ *lie in the circle (10):*

$$|\lambda(\alpha)| \le 1 + c_1\tau, \tag{41}$$

where c_1 *does not depend on* h *(or on* τ*).*

Proof. First we establish sufficiency. Given (41), clearly

$$|\lambda(\alpha)|^p \leq |1 + c_1\tau|^{T/\tau} \leq e^{c_1 T}. \qquad (42)$$

From representation (39), using the Parseval inequality and inequality (42), we get

$$||u^p|| = \left[\int_0^{2\pi} |\lambda^p(\alpha)U(\alpha)|^2 d\alpha\right]^{1/2} \leq e^{c_1 T}\left[\int_0^{2\pi} |U(\alpha)|^2 d\alpha\right]^{1/2} =$$

$$= e^{c_1 T}||u^0|| = c||u^0||.$$

Necessity. We now show that violation of (41) for all fixed c_1 implies instability. The fact that, in this case, the solution

$$u_m^p = \lambda^p(\alpha)e^{i\alpha m}, \qquad p = 0, \ldots, [\tfrac{T}{\tau}],$$

is unbounded as $\tau \to 0$, cannot be used as a proof of instability, given norm (40), since $\{\exp(i\alpha m)\}$ does not belong to the space of net functions such that the sums of squares of the moduli of function values, over all net-points, is bounded.

To prove instability we note, first, that, one can always choose a square-integrable function, $U(\alpha)$, in such a way as to satisfy the inequality

$$\frac{1}{2\pi}\int_0^{2\pi} |\lambda(\alpha)|^{2p}|U(\alpha)|^2 d\alpha \geq \max_\alpha\left[|\lambda(\alpha)|^{2p} - \varepsilon\right] \cdot \frac{1}{2\pi}\int_0^{2\pi} |U(\alpha)|^2 d\alpha, \qquad (43)$$

where $\varepsilon > 0$ is arbitrary. In fact if $\max_\alpha|\lambda(\alpha)| = |\lambda(\alpha^*)|$, we may take

$$U(\alpha) = \begin{cases} 1, & \text{if } \alpha \text{ is in } [\alpha^* - \delta, \alpha^* + \delta], \\ 0, & \text{if a is not in } [\alpha^* - \delta, \alpha^* + \delta]. \end{cases}$$

Because of the continuity of the function $|\lambda(\alpha)|^{2p}$, (43) will be satisfied if $\delta = \delta(\varepsilon)$ is taken small enough. If (42) is not satisfied, then one can find a sequence of h_k's, and a corresponding sequence $\tau_k = \tau(h_k)$, for which

$$c_k \equiv \left[\max_\alpha |\lambda(\alpha, h_k)|\right]^{[T/\tau_k]} \to \infty \quad \text{as} \quad k \to \infty.$$

Let us set $\varepsilon = 1$ and choose a $U(\alpha)$ such that (43) is satisfied. As our sequence $\{u_m^0\}$ we take the sequence of Fourier coefficients of the function $U(\alpha)$. Then (43), with $p = [T/\tau]$, takes the form

$$||u^{[T/\tau]}|| \geq (c_k - 1)||u^0||, \qquad c_k \to \infty, \quad \text{as} \quad h \to 0,$$

which indeed signals instability with respect to initial data.

Theorem 3. *For stability of the Cauchy difference problem (31), given norms (40), it is necessary and sufficient that the spectral stability criterion (41) be satisfied.*

Proof. Necessity is obvious, since violation of this criterion implies, via Theorem 2, that there is no stability with respect to initial data.

To prove sufficiency we will establish that for every $k \geq 0$ we have the inequality

$$||u^{k+1}|| \leq (1 + c_1\tau)||u^k|| + c_2\tau \max_n ||\phi^n||, \tag{44}$$

from which, clearly, it follows that, for all j such that $p \geq j \geq 0$,

$$(1 + c_1\tau)^j||u^{p+1-j}|| \leq (1 + c_1\tau)^{j+1}||u^{p-j}|| + c_2\tau(1 + c_1\tau)^p \cdot \max_n ||\phi^n||.$$

Summing the left- and right-hand sides of the inequality for j = 0, 1, ..., p term by term, and discarding identical terms on the left- and right-hand sides, we find we can write

$$||u^{p+1}|| \leq (1 + c_1\tau)^{n+1}||u^0|| + c_2\tau p(1 + c_1\tau)^p \max_n||\phi^n|| \leq$$
$$\leq (1 + c_1\tau)e^{c_1T} \cdot ||u^0|| + c_2T\, e^{c_1T} \max_n ||\phi^n|| \leq \text{const}\cdot||f^{(h)}||_{F_h}$$

which, in view of the arbitrariness of p, p = 0, 1, ..., [T/τ]-1, implies stability.

To prove (44) we use the integral representation of the solution, (37), and the recurrence relation (38), from which

$$u_m^{k+1} = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} U^{k+1}(\alpha)e^{i\alpha m}d\alpha =$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \lambda(\alpha)U^k(\alpha)e^{i\alpha m}d\alpha + \frac{\tau}{\sqrt{2\pi}} \int_0^{2\pi} \frac{\phi^k(\alpha)}{b_1e^{-i\alpha} + b_0 + b_{-1}e^{i\alpha}} e^{i\alpha m}d\alpha. \tag{45}$$

Thus the net-function $\{u_m^{k+1}\}$ of the argument m has been represented as the sum of two net functions, written in the form of integrals over the parameter α. Using Parseval's equation we may write, for the norms of these two net-functions

$$\left\|\frac{1}{\sqrt{2\pi}}\int_0^{2\pi}\lambda(\alpha)U^k(\alpha)e^{i\alpha m}d\alpha\right\| = \left[\int|\lambda(\alpha)U^k(\alpha)|^2d\alpha\right]^{1/2} \leq$$

$$\leq \max_\alpha|\lambda(\alpha)|\left[\int_0^{2\pi}|U^k(\alpha)|^2d\alpha\right]^{1/2} \leq (1+c_1\tau)\cdot\|u^k\|;$$

$$\left\|\frac{\tau}{\sqrt{2\pi}}\int_0^{2\pi}\frac{\phi^k(\alpha)}{b_1e^{-i\alpha}+b_0+b_{-1}e^{i\alpha}}e^{i\alpha m}d\alpha\right\| =$$

$$= \tau\left[\int_0^{2\pi}\left|\frac{\phi^k(\alpha)}{b_1e^{-i\alpha}+b_0+b_{-1}e^{i\alpha}}\right|^2d\alpha\right]^{1/2} \leq$$

$$\leq \frac{\tau}{\min_\alpha\left|b_1e^{-i\alpha}+b_0+b_{-1}e^{i\alpha}\right|}\left[\int_0^{2\pi}|\phi^k(\alpha)|^2d\alpha\right]^{1/2} =$$

$$= \frac{\tau}{\min_\alpha\left|b_1e^{-i\alpha}+b_0+b_{-1}e^{i\alpha}\right|}\cdot\|\phi^k\| \leq \tau c_2\max_n\|\phi^n\|.$$

From these last two bounds on the norms of the terms on the right-hand side of Eq. (45) we get bound (44), completing the proof.

One can show that if, in place of the norm in (40), one takes

$$\|u^p\| = \sup_m\left|u_m^p\right|,$$

then the spectral criterion $|\lambda(\alpha)| < 1 + c_1\tau$ will no longer be a sufficient condition for stability. For the Cauchy difference problem for a system of equations this criterion is again only a necessary condition for stability.

* * * * * *

The integral representation (37) of the solution of the Cauchy difference problem is used, not only in the study of stability, but also to bring out other properties of a difference scheme.

If, for example, the spectrum $\lambda = \lambda(\alpha)$, for $\alpha \neq 0$, lies strictly inside the unit circle, then the solutions $u_m^p = \lambda^p(\alpha)\exp(i\alpha m)$ for which $\alpha \neq 0$ are damped, from level to level, by the factor $\lambda(\alpha)$. From Eq. (39) it is clear that, for $[T/\tau] = p$, one gets a net-function, corresponding to the function $\lambda^p(\alpha)\psi(\alpha)$, which is concentrated in the long-wave-length region ($\alpha \approx 0$). The difference scheme "smooths" the initial data.

5. Smoothing of the difference solution as a result of approximational viscosity. We have seen that the spectrum of the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, \pm 1, \ldots \\ &\qquad\qquad p = 0, 1, \ldots, [T/t]-1, \\ &u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{46}$$

approximating the Cauchy problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &u(x, 0) = \psi(x) \quad -\infty < x < \infty, \end{aligned}\right\} \tag{47}$$

in the circle $\lambda = 1 - r + r\exp(i\alpha)$, $0 \leq \alpha \leq 2\pi$. For $r < 1$ each point for which $\alpha \neq 0$ corresponds to a point of the spectrum $\lambda(\alpha)$ such that $|\lambda(\alpha)| < 1$. This means that every harmonic $u_m^0 = \exp(i\alpha mh)$, specified via initial data, is damped, being multiplied by $\lambda(\alpha)$ at each step from one level to another; in the course of time the solution is smoothed, since for small αh (i.e. for low-frequency harmonics) the damping is weaker. Note that the solution of the differential problem (47), $u(x, t) = \psi(x + t)$, does not become smoother with time; it is obtained from the initial data, as time progresses, by shifting these data to the left. Thus the solution of problem (47), corresponding to the initial condition $u(x, 0) = \exp(i\alpha x)$, is $u(x, t) = \exp(i\alpha t)\exp(i\alpha x)$, and the factor $\exp(i\alpha t)$ is equal to one in modulus. The computational smoothing of the solution, which occurs when one uses difference scheme (46), may be understood as the manifestation of an "approximation viscosity", characteristic of this scheme. Let us explain what we mean by approximational viscosity. If the equation

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0$$

is taken as the simplest model of an equation of motion for a non-viscous gas, then it is natural to take the equation

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \mu \frac{\partial^2 u}{\partial x^2} \tag{48}$$

as a model equation for the motion of a gas with a viscosity, $\mu > 0$, smoothing the solution. With the initial conditions $u(x, 0) = \exp(i\alpha x)$ the solution of Eq. (48) has the form

$$u(x, t) = e^{-\mu\alpha^2 t + i\alpha t} e^{i\alpha x} \equiv \tilde{\lambda}(\alpha, t) e^{i\alpha x}.$$

For $\mu = O(\tau)$ and $t = \tau$ the factor, $\tilde{\lambda}(\alpha, t)$, damping the harmonic $\exp(i\alpha x)$, may be written thus:

$$\tilde{\lambda}(\alpha, \tau) = 1 - \mu\alpha^2\tau + i\alpha\tau - \frac{\alpha^2\tau^2}{2} + o(\tau^2). \qquad (49)$$

We will assume that the solution $u^{(h)}$ of the difference problem can be defined in such a way, via auxiliary conditions outside the net, that the resulting smooth function $u^{(h)}(x, t)$ will be bounded uniformly in h, together with its derivatives up to fourth order. Then at the net-points, (x, t), using Taylor's formula we may write

$$0 = \frac{u^{(h)}(x, t+\tau) - u^{(h)}(x, t)}{\tau} - \frac{u^{(h)}(x+h, t) - u^{(h)}(x, t)}{h} =$$
$$= \frac{\partial u^{(h)}(x, t)}{\partial t} - \frac{\partial u^{(h)}(x, t)}{\partial x} + \frac{\tau}{2}\frac{\partial^2 u^{(h)}(x, t)}{\partial t^2} -$$
$$- \frac{h}{2}\frac{\partial^2 u^{(h)}}{\partial x^2} + h^2\varepsilon_1^{(h)}(x, t). \qquad (50)$$

Here and below $\varepsilon_1^{(h)}$, $\varepsilon_2^{(h)}$ and $\varepsilon_3^{(h)}$ are functions uniformly bounded in h, together with their derivatives.

From (50) it follows, in particular, that

$$\frac{\partial u^{(h)}}{\partial t} = \frac{\partial u^{(h)}}{\partial x} + h\varepsilon_2^{(h)}(x, t).$$

Differentiating this identity with respect to t we get

$$\frac{\partial^2 u^{(h)}}{\partial t^2} = \frac{\partial}{\partial x}\left(\frac{\partial u^{(h)}}{\partial t}\right) + h\frac{\partial \varepsilon_2^{(h)}}{\partial t} = \frac{\partial^2 u^{(h)}}{\partial x^2} + h\frac{\partial \varepsilon_2^{(h)}}{\partial x} + h\frac{\partial \varepsilon_2^{(h)}}{\partial t} = \frac{\partial^2 u^{(h)}}{\partial x^2} + h\varepsilon_3^{(h)}.$$

Inserting the above expression for $\partial^2 u^{(h)}/\partial t^2$ into Eq. (50), and discarding terms small to second order, we get a differential equation of form (48),

$$\frac{\partial U^{(h)}}{\partial t} - \frac{\partial U^{(h)}}{\partial x} = \frac{h - \tau}{2}\frac{\partial^2 U^{(h)}}{\partial x^2}, \qquad (51)$$

which we will consider, not just on the net, but everywhere for $t > 0$.

Thus difference equation (26) has turned out to be "basically equivalent" to the *differential approximation* (51), which is an equation of form (48) with small viscosity $\mu = (h - \tau)/2$. This viscosity is called *approximational* since it arises as a result of the approximation of problem (47) by difference problem (46). Differential equation (51) smooths the initial data basically in the same way as scheme (46). In fact if $U(x, 0) = \exp(i\alpha x)$ then at time $t = \tau$ this harmonic, according to Eq. (49), is multiplied by

$$\lambda(\alpha,\tau) = 1 - \frac{h-\tau}{2}\alpha^2\tau + i\alpha\tau - \frac{\alpha^2\tau^2}{2} + o(\tau^2) = 1 + i\alpha\tau - \frac{h}{2}\alpha^2\tau + o(\tau^2). \qquad (52)$$

For $u_m^0 = \exp(i\alpha x)|_{x=mh} = \exp(i\alpha mh)$, by difference scheme (46) we get, at $t = \tau$, this same harmonic multiplied by the factor

$$\lambda(\alpha) = 1 - r + re^{i\alpha h} = 1 - r + r\left(1 + i\alpha h - \frac{\alpha^2 h^2}{2}\right) + o(h^2) =$$

$$= 1 + i\alpha\tau - \frac{h}{2}\alpha^2\tau + O(\tau^2),$$

which agrees with the factor (52) with an accuracy up to small terms of second order in τ (or h).

* * *

PROBLEMS

1. For what values of the parameter $\sigma > 0$ does the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} = \sigma \frac{u_{m+1}^{p+1} - 2u_m^{p+1} + u_{m-1}^{p+1}}{h^2} + (1 - \sigma)\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2}, \\ &u_m^0 \text{ given}, \qquad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

approximating the Cauchy problem for the heat equation, satisfy the Von Neumann spectral stability criterion for all $r = \tau/h^2$?

2. Does the following difference scheme satisfy the spectral stability criterion:

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^{p-1}}{2\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \phi(x_m, t_p), \qquad p \geq 1, \\ &\left.\begin{aligned} u_m^0 &= \psi_m, \\ u_m^1 &= \psi_m^{(1)}, \end{aligned}\right\} \quad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

where

$$\psi_m^{(1)} = u(x, 0) + \tau\frac{\partial u(x, 0)}{\partial t} = u(x_m, 0) + \tau\frac{\partial^2 u(x, 0)}{\partial x^2} = \psi(x_m) + \tau\psi''(x_m)?$$

This difference scheme approximates Cauchy problem (19) for the heat equation to order $O(\tau^2 + h^2)$.

3. Show that the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} + A\frac{u_{m+1}^{p+1} - u_{m-1}^{p+1}}{2h} = 0, \qquad m = 0, \pm 1, \ldots, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad p = 0, 1, \ldots, \\ &u_m^0 = \psi(x_m), \qquad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

approximating the Cauchy problem

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0, \qquad -\infty < x < \infty, \quad t > 0,$$

$$u(x, 0) = \psi(x), \qquad -\infty < x < \infty,$$

to order $O(\tau + h^2)$, satisfies the spectral stability criterion for any $r = \tau/h$ and any constant A.

4. Study the following predictor-corrector difference scheme for the solution of the Cauchy problem $u_t + Au_x = 0$, $u(x, 0) = \phi(x)$:

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} + A\frac{u_{m+1/2}^{p+1/2} - u_{m-1/2}^{p+1/2}}{h} = 0, && m = 0, \pm 1, \ldots, \\ &&& p = 0, 1, \ldots, \\ &u_m^0 = \psi_m, && m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

$A = \text{const}$, where the intermediate net function $u^{p+1/2} = \{u_{m+1/2}^{p+1/2}\}$ is determined from $u^p = \{u_m^p\}$ in two stages: first one calculates $v^p = \{v_m^p\}$ as the solution of the difference problem

$$\frac{v_m^{p+1/2} - u_m^p}{\tau/2} + A\frac{v_{m+1}^{p+1/2} - v_{m-1}^{p+1/2}}{2h} = 0, \qquad m = 0, \pm 1, \ldots,$$

and then $u^{p+1/2} = \{u_{m+1/2}^{p+1/2}\}$ via the expression

$$u_{m+1/2}^{p+1/2} = (1 - \alpha)\frac{v_{m+1}^{p+1/2} + v_m^{p+1/2}}{2} + \alpha\frac{v_{m+2}^{p+1/2} + v_{m-1}^{p+1/2}}{2}.$$

Show that if the interpolation parameter α lies in the interval $0 \le \alpha \le .25$ then, for any $r = \tau/h = \text{const}$ the spectral stability criterion is satisfied. For $\alpha = 0$ the whole spectrum lies on the unit circle, and for $0 < \alpha \le 0.25$ it is located in the unit circle and touches this circle only for $\lambda = 1$. The eigenvalue $\lambda = 1$ corresponds to the eigenfunction $u_m = (\pm 1)^m$.

26. Principle of frozen coefficients

Here we present a method which greatly expands the class of time-dependent difference schemes which may be studied through use of the spectral stability criterion. This necessary condition for stability, developed in §25 for the study of the Cauchy difference problem with constant coefficients, can be used also in the case of Cauchy difference schemes with "continuous", but not constant coefficients, and, as well, for problems in bounded regions when the boundary conditions are given not only

at the t = 0 time-boundary, but also on the side boundaries. This method can also be used for the study of nonlinear problems.

1. Frozen coefficients at interior points. We will formulate the principle of frozen coefficients using, as an example, the following difference boundary-value problem:

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - a(x_m, t_p)\,\frac{u_{m-1}^p - 2u_m^p + u_{m+1}^p}{h^2} = 0, \\ &\qquad p = 0, 1, \ldots, [T/\tau] - 1, \\ &u_m^0 = \psi_m, \qquad m = 0, 1, \ldots, M; \quad Mh = 1, \\ &\qquad \ell_1 u_0^{p+1} = 0, \qquad \ell_2 u_M^{p+1} = 0. \end{aligned}\right\} \tag{1}$$

In this difference boundary-value problem $\ell_1 u_0^{p+1} = 0$ and $\ell_2 u_M^{p+1}$ are certain conditions given, respectively on the left and right boundaries of the net segment $0 \leq m \leq M$; further $a(x, t) > 0$.

Now we take an arbitrary interior point, $(\tilde{x}, \tilde{t})$, of the region $0 \leq x \leq 1$, $0 \leq t \leq T$, in which problem (1) is to be treated, and "freeze" the coefficients of problem (1) at that point.

The difference equation with constant coefficients

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - a(\tilde{x}, \tilde{t})\,\frac{u_{m-1}^p - 2u_m^p + u_{m+1}^p}{h^2} = 0, \\ &p = 0, 1, \ldots, [T/\tau]-1; \quad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{2}$$

generated in this way, will be considered now, not only for $0 < m < M$, but for all integral m. We will now formulate: The principle of frozen coefficients. *For the stability of problem (1) it is necessary that the Cauchy problem for equation (2) with constant coefficients should satisfy the necessary Von Neumann spectral stability criterion.*

To provide a rationale for the principle of frozen coefficients we present the following heuristic argument.

As we refine the mesh the coefficient a(x, t) in the neighborhood of the point $(\tilde{x}, \tilde{t})$, for any fixed number of step-widths h in space, and τ in time, changes less and less, and differs less and less from the value $a(\tilde{x}, \tilde{t})$. In addition, the distance from the point $(\tilde{x}, \tilde{t})$ to the boundaries x = 0 and x = 1 of the interval, as measured in numbers of net-steps, tends to infinity. Therefore for a fine net the perturbations induced in the solution of problem (1), at the instant $t = \tilde{t}$, in the neighborhood of point $x = \tilde{x}$, develops (for short times) in approximately the same way as in problem (2).

It will be seen that this reasoning is general in character. It does not depend on the number of space variables or the number of unknown

functions; nor on the form or the difference equation or the system of difference equations.

In §25 we considered the Cauchy problem for an equation of form (2) and found that, to satisfy the Von Neumann criterion, the ratio $r = \tau/h^2$ of the net mesh-widths must satisfy the condition

$$r \leq \frac{1}{2a(\tilde{x}, \tilde{t})}.$$

Since, by the principle of frozen coefficients, it is necessary for stability that this condition be satisfied for any $(\tilde{x}, \tilde{t})$ the ratio, $r = \tau/h^2$, of step-widths must satisfy the condition

$$r \leq \frac{1}{2 \max\limits_{\tilde{x},\tilde{t}} a(\tilde{x}, \tilde{t})}. \tag{3}$$

The principle of frozen coefficients gives us some guidance, at a heuristic level of rigor, also in the investigation of nonlinear problems. Consider, for example, the following nonlinear problem:

$$u_t - (1 + u^2)u_{xx} = 0, \qquad 0 < x < 1,$$

$$u(x, 0) = \psi_0(x), \qquad 0 < x < 1,$$

$$u(0, t) = \psi_1(t), \qquad u(1, t) = \psi_2(t), \qquad 0 < t < T.$$

To treat this problem we use the following difference scheme:

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau_p} - \left[1 + \left(u_m^p\right)^2\right] \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ 0 < m < M, \qquad p = 0, 1, \ldots, [T/\tau]-1, \\ u^0 = \psi_0(mh), \qquad 0 \leq m \leq M, \\ u_0^{p+1} = \psi_1(t_p + \tau_p), \qquad p = 1, 2, \ldots, [T/\tau], \\ u_M^{p+1} = \psi_2(t_p + \tau_p), \end{cases}$$

allowing the step-width τ_p to change from level to level. This scheme allows one to compute sequentially, level after level, the unknowns u_m^1, $m = 0, \ldots, M$, then u_m^2, $m = 0, 1, \ldots, M$, etc.

Suppose that we have already gotten to the level $t = t_p$, have computed u_m^p, $m = 0, 1, \ldots, M$, and want to continue the calculation. How should we choose the next step-width $\tau = \tau_p$? We can imagine that we are required to compute the solution of the linear difference equation

$$\frac{u_m^{p+1} - u_m^p}{\tau_p} - a(x_m, t_p)\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0$$

with the given variable coefficient $a(x_m, t_p) \equiv 1 + (u_m^p)^2$. In fact it is natural to assume that the values u_m^p are close to the values $u(x_m, t_p)$ of the smooth solution, $u(x, t)$, of the differential problem. The coefficient is then close to the continuous function $a(x, t) \equiv 1 + u^2(x, t)$ which, in the course of several time-steps, changes very little.

The application of the Von Neumann criterion to the equation with variable coefficients $a(x_m, t_p)$ yields the restriction (3) on the relation between step widths,

$$\frac{\tau_p}{h^2} = r_p \leq \frac{1}{2\max\limits_x |a(x, t_p)|} = \frac{1}{2\max\limits_m \left|1 + (u_m^p)^2\right|},$$

as a necessary condition for stability. On this basis it seems plausible that the next step width should satisfy the condition

$$\tau_p \leq \frac{1}{2\max\limits_m \left|1 + (u_m^p)^2\right|} h^2 .$$

Computer experiments confirm the validity of this heuristic reasoning.

If the necessary conditions for stability, derived via consideration of the Cauchy problem with coefficients frozen at any arbitrary net-point, turn out to be violated, then one cannot expect stability for any boundary conditions. We stress, however, that the principle of frozen coefficients does not take into account the influence of boundary conditions. If the necessary stability conditions, flowing from the principle of frozen coefficients, are fulfilled, then we may have stability for some, and not for other boundary conditions. Now we develop the criterion of K. I. Babenko and I. M. Gelfand, which take into account the effect of boundary conditions for problems posed on intervals.

2. Criterion of Babenko and Gelfand. In considering problem (1), we postulated that perturbations communicated to the solution of problem (1) in the neighborhood of some arbitrary internal point $(\tilde{x}, \tilde{t})$ develop, for a fine mesh, approximately in the same way as the same perturbations communicated to the solution of Cauchy problem (2), with coefficients frozen at point $(\tilde{x}, \tilde{t})$. In justifying this principle we argued that the distance from the interior point $(\tilde{x}, \tilde{t})$ to the boundary, measured in numbers of net-steps, increases without bound as the net is refined. But if the point $(\tilde{x}, \tilde{t})$ lies on the side-boundary $\tilde{x} = 0$ or $\tilde{x} = 1$, then this heuristic argument loses its force. Suppose, for example, that $\tilde{x} = 0$. Then the distance from point $\tilde{x}$ to any fixed point $x > 0$ (and, in particular, to the right interval-boundary $x = 1$), as measured in step widths will, as before,

grow without bound as $h \to 0$; but the number of steps to the left boundary, $x = 0$, does not change and remains equal to zero.

For this reason a perturbation of the solution of problem (1) near the left-hand boundary $x = 0$ must develop, in short time-spans, like a perturbation of the solution of the problem

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - a(0, \tilde{t})\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \qquad m = 1, 2, \ldots, \\ &\ell_1 u_0^{p+1} = 0. \end{aligned}\right\} \tag{4}$$

This problem is obtained from the original problem (1) by freezing the coefficient $a(x, \tilde{t})$ at the left interval-boundary $\tilde{x} = 0$ and simultaneously moving the right-hand boundary to $+\infty$. It is natural to consider problem (4) only for those functions $u^p = \{u_0^p, u_1^p, u_2^p, \ldots\}$ for which

$$u_m^p \to 0 \quad \text{as} \quad m \to +\infty.$$

It is only in this case that the perturbation is concentrated near the boundary $x = 0$, and it is only with respect to perturbations of this concentrated form that problems (1) and (4), near the left-hand boundary $x = 0$, are similar to each other.

So too, the development of perturbations of the solutions of problem (1) near the right-hand boundary $x = 1$ must be similar to the development of the same sorts of perturbations in the problem

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - a(1, \tilde{t})\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ &m = \ldots, -2, -1, 0, 1, \ldots, M-1, \\ &\ell_2 u_M^{p+1} = 0 \end{aligned}\right\} \tag{5}$$

with only a right-hand boundary. This problem was obtained from the original problem (1) by freezing the coefficient, $a(x, t)$, at the right-hand boundary $\tilde{x} = 1$, and removing the left-hand boundary to $-\infty$. Problem (5) must be considered for net functions $u^p = \{\ldots, u_{-2}^p, u_{-1}^p, u_0^p, u_1^p, \ldots, u_M^p\}$, satisfying the condition $u_m^p \to 0$ as $m \to -\infty$.

Problems (2), (4), and (5) are simpler than the original problem (1) in the sense that, for fixed $r = \tau/h^2$, they do not depend on h, and are problems with constant coefficients.

Thus the procedure for studying stability, taking into account the effect of the boundary, as applied to problem (1) takes the following form. One formulates three auxiliary problems (2), (4) and (5). For each of these three problems, not depending on h, one finds all those λ (the eigenvalues of the transition operator from u^p to u^{p+1}) for which there exists a solution of the form

$$u_m^p = \lambda^p u_m^0.$$

For problem (2) $u^0 = \{u_m^0\}$, $m = 0, \pm 1, \ldots$, must be bounded. In the case of problem (4) $u^0 = \{u_0^0, u_1^0, \ldots u_m^0, \ldots\}$, $u_m^0 \to 0$ as $m \to +\infty$, while for problem (5)

$$u^0 = \{\ldots, u_{-2}^0, u_{-1}^0, u_0^0, u_1^0, \ldots, u_M^0\},$$

$$u_m^0 \to 0 \quad \text{as} \quad m \to -\infty.$$

If problem (1) is to be stable the totality of eigenvalues of the three problems (2), (4), and (5) must lie in the unit circle $|\lambda| \leq 1$. *Problem (2) is to be considered for each fixed* $\tilde{x}$, $0 < \tilde{x} < 1$.

Let us continue to examine problem (1). We will assume, hereafter, that $a(x, t) \equiv 1$, and compute the spectra for all three problems (2), (4), and (5), with different boundary conditions $\ell_1 u_0^{p+1} = 0$ and $\ell_2 u_M^{p+1} = 0$.

Substituting the solution $u_m^p = \lambda^p u_m$ into difference equation (2) we get

$$(\lambda - 1)u_m - r(u_{m+1} - 2u_m + u_{m-1}) = 0, \qquad r = \frac{\tau}{h^2}$$

or

$$u_{m+1} - \left(\frac{-2r + 1 - \lambda}{r}\right) u_m + u_{m-1} = 0. \tag{6}$$

This is a second order difference equation. Similar equations have been investigated in Chapter 1. To find the general solution of Eq. (6) we first construct the characteristic equation

$$q^2 + \left(2 + \frac{\lambda - 1}{r}\right)q + 1 = 0. \tag{7}$$

If q is a root of this equation the net function

$$u_m^p = \lambda^p q^m$$

is one of the solutions of the equation

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0.$$

If $|q| = 1$, i.e. $q = \exp(i\alpha)$, then the net function

$$u_m^p = \lambda^p e^{i\alpha m},$$

bounded as $m \to +\infty$ and $m \to -\infty$ is a solution, as we saw in §25, for

$$\lambda = 1 - 4r\sin^2\frac{\alpha}{2}, \qquad 0 \leq \alpha \leq 2\pi.$$

These $\lambda = \lambda(\alpha)$ fill the interval $1 - r \leq \lambda \leq 1$ on the real axis. Precisely this interval is, then, the spectrum of problem (2). There are no eigenvalues, λ, of problem (2) which do not lie in this interval since, if characteristic equation (7) has no roots, q, equal to unity in modulus, problem (6) has no solutions bounded as $m \to \pm\infty$.

If λ does not lie on the interval $1 - 4r \leq \lambda \leq 1$, then both roots of characteristic equation (7) are different from one in modulus, but their product is equal to the constant term of Eq. (7), i.e. is equal to one. Therefore one of the roots of (7) is greater than, and the other less than unity. Suppose, for the sake of definiteness, that $|q_1| < 1$ and $|q_2| > 1$. Then the general solution of (6), decreasing in modulus as $m \to +\infty$, has the form

$$u_m = c[q_1(\lambda)]^m,$$

and the general solution of (6) tending to zero as $m \to -\infty$ has the form

$$u_m = c[q_2(\lambda)]^m.$$

To determine the eigenvalues of problem (4) one must substitute $u_m = cq_1^m(\lambda)$ in the left-hand boundary condition $\ell_1 u = 0$ and find all those λ for which it is satisfied. These will, then, be all the eigenvalues of problem (4). If, for example

$$\ell_1 u_0 \equiv u_0 = 0,$$

then the condition $cq_1^0 = 0$ is not fulfilled for any $c \neq 0$, so that there are no eigenvalues.

If $\ell_1 u = u - u_0 = 0$ then the condition $cq_1^1 - cq_1^0 = c(q_1 - 1) = 0$, in view of the fact that $q_1 \neq 1$, leads to $c = 0$ so that, again, there are no eigenvalues.

If $\ell_1 u = 2u - u_0 = 0$, then the condition $c(2q_1 - 1) = 0$ is fulfilled for $c \neq 0$ if $q_1 = 1/2$. From Eq. (7) we find that, for $q_1 = 1/2$, λ is given by

$$\lambda = 1 + r\left(q_1 - 2 + \frac{1}{q_1}\right) = 1 + r\,\frac{1 - 4 + 4}{2} = 1 + \frac{r}{2}.$$

This is, in fact, the only eigenvalue of problem (4), and it lies outside the unit circle, since $\lambda = 1 + r/2 > 1$. The eigenvalues of problem (5) are computed analogously. They are gotten from the equation

$$\ell_2 u_M = 0,$$

with

$$u_m = q_2^m, \quad q_2 = q_2(\lambda), \quad m = M,\ M-1,\ M-2,\ \ldots$$

We take as another example the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad p = 0, 1, \ldots, [T/t]-1, \\ &\qquad\qquad\qquad\qquad\qquad\qquad m = 0, 1, \ldots, M-1, \\ &\qquad\qquad Mh = 1, \\ &u_m^0 = \psi(x_m), \qquad m = 0, 1, \ldots, M, \\ &\qquad\qquad u_M^{p+1} = 0, \end{aligned}\right\} \tag{8}$$

approximating the problem

$$\left.\begin{aligned} &u_t - u_x = 0, \qquad 0 < x < 1, \quad 0 < t < T, \\ &\qquad u(x, 0) = \psi(x), \\ &\qquad u(1, t) = 0. \end{aligned}\right\}$$

Let us examine the stability of this problem via the Babenko-Gelfand criterion. We associate, with scheme (8), three related problems: the problem without side-boundaries

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad m = 0, \pm 1, \ldots, \tag{9}$$

the problem with only a left-hand boundary

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^0}{h} = 0, \qquad m = 0, 1, \ldots, \tag{10}$$

and the problem with only a right-hand boundary

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad m = M-1, M-2, \ldots, \\ &\qquad\qquad u_M^{p+1} = 0, \qquad p = 0, 1, \ldots, [T/\tau]-1. \end{aligned}\right\} \tag{11}$$

In the case of problem (10) with only a left-hand boundary there is no boundary condition, since there was no left-hand boundary condition in the original problem (8).

We must now compute all the eigenvalues of the three transition operators from u^p to u^{p+1}, corresponding to each of the three auxiliary problems (9), (10) and (11), and determine under what conditions they lie in the circle $|\lambda| \leq 1$.

The solution of the form

$$u_m^p = \lambda^p u_m^0$$

under the substitution

$$u_m^{p+1} = (1 - r)u_m^p + ru_{m+1}^p, \qquad r = \frac{\tau}{h},$$

generates the following ordinary first order difference equation for the eigenfunctions:

$$(\lambda - 1 + r)u_m - ru_{m+1} \geq 0. \tag{12}$$

The corresponding characteristic equation

$$(\lambda - 1 + r) - rq = 0 \tag{13}$$

provides a connection between λ and q. The general solution of Eq. (12) is

$$u_m = cq^m = c\left(\frac{\lambda - 1 + r}{r}\right)^m, \qquad m = 0, \pm 1, \ldots$$

For $|q| = 1$, $q = \exp(i\alpha)$, $0 \leq \alpha \leq 2\pi$

$$\lambda = (1 - r) + re^{i\alpha}.$$

The point $\lambda = \lambda(\alpha)$ traverses the perimeter of a circle with center at point $1 - r$ and radius r. This perimeter is, then, the eigenvalue spectrum of

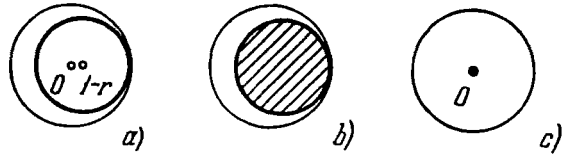


Fig. 26.

problem (9) (see Fig. 26,a). The nontrivial solution decreasing as $m \to +\infty$

$$u_m^p = \lambda^p u_m = c_0 \lambda^p q^m$$

exists for any q, $|q| < 1$. Corresponding eigenvalues $\lambda = 1 - r + rq$, clearly, fill the disc bounded by the circle $\lambda = (1 - r) + r\exp(i\alpha)$ (Fig. 26,b).

Finally, the solution of problem (11), $u_m^p = \lambda^p u_m$, decreasing as $m \to -\infty$, must have the form

$$u_m^p = c\lambda^p q^m, \qquad |q| > 1,$$

where λ and q are connected by Eq. (13). From the boundary condition $u_M^p = 0$ it follows that a nontrivial solution ($c \neq 0$) exists only for $\lambda = \lambda(q) = 0$, i.e. $q = (r - 1)/r$. This value of q is greater than one in modulus if either $(r - 1)/r > 1$ or $(r - 1)/r < -1$. The first equation has no solution: the solution of the second is $r < 1/2$.

Thus for $r < 1/2$ problem (10) has the eigenvalue $\lambda = 0$ (Fig. 26,c). In Fig. 27,a,b and c are represented the totality of eigenvalues of the

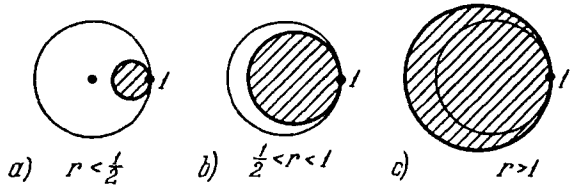


Fig. 27.

three problems for $r < 1/2$, $1/2 < r < 1$, and $r > 1$. Clearly the set of all eigenvalues of all three problems lies in the circle $|\lambda| < 1 + c\tau$, where c does not depend on h, if and only if $r \leq 1$.

The above stability criterion for nonstationary problems on an interval, taking into account the influence of boundary conditions, is applicable to boundary value problems on an interval also for systems of difference equations. In this case seemingly natural difference schemes, satisfying the Von Neumann criterion, often turn out to be unstable because of an unsatisfactory approximation to the boundary conditions, and it is important to know how to choose schemes free from this defect.

In Chapter 14 we will return to a discussion of the Babenko-Gelfand spectral criterion taking a broader point of view. In particular we will demonstrate rigorously that fulfillment of this condition is necessary for stability, and that if it is satisfied there cannot be a "gross" instability.

PROBLEMS

1. Show under what conditions the spectral stability criterion is satisfied for the difference scheme

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2}\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0$$

$$m = 1, 2, \ldots, M-1,$$

$$u_m^0 = \psi(x_m), \qquad m = 0, 1, \ldots, M,$$

$$\left.\begin{aligned} u_M^{p+1} &= 0, \\ \frac{u_0^{p+1} - u_0^p}{\tau} - \frac{u_1^p - u_0^p}{h} &= 0 \end{aligned}\right\} \quad p = 0, 1, \ldots, [T/\tau]-1,$$

approximating the differential problem

$$\left.\begin{aligned} &u_t - u_x = 0, \qquad 0 < x < 1, \quad 0 < t < T, \\ &u(x, 0) = \psi(x), \\ &u(0, t) = u(1, t) = 0 \end{aligned}\right\}$$

on a smooth solution $u(x, t)$ to second order in h.

Answer: $\tau/h \leq 1$.

2. To construct a difference scheme approximating the following boundary-value problem for the hyperbolic system of differential equations

$$\left.\begin{aligned} \frac{\partial v}{\partial t} &= \frac{\partial w}{\partial x}, \\ \frac{\partial w}{\partial t} &= \frac{\partial v}{\partial x}, \end{aligned}\right\} \qquad 0 \leq x \leq 1, \quad 0 \leq t \leq T,$$

$$v(x, 0) = \psi_1(x), \qquad w(x, 0) = \psi_2(x),$$

$$v(0, t) = w(1, t) = 0,$$

we set $u(x, t) = \big(v(x, t), w(x, t)\big)^T$, and write the equation in matrix form

$$\frac{\partial}{\partial t} u - A \frac{\partial}{\partial x} u = 0,$$

$$u(x, 0) = \psi(x),$$

$$v(0, t) = w(1, t) = 0,$$

where

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Take, as a net, $(x_m, t_n) = (mh, n\tau)$, $h = 1/M$, M a positive integer. Set

$$\frac{u_m^{p+1} - u_m^p}{\tau} - A\frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2}A^2\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0,$$

$$m = 1, 2, \ldots, M-1,$$

$$u_m^0 = \psi(x_m),$$

$$v_0^{p+1} = w_M^{p+1} = 0.$$

To complete construction of this scheme it is necessary to impose additional boundary conditions on the left and right side-boundaries. Noting that, for any α and β, it follows from the system of differential equations that

$$\frac{\partial(v + \alpha w)}{\partial t} - \frac{\partial(w + \alpha v)}{\partial x}\Big|_{x=0} = 0,$$

$$\frac{\partial(v + \beta w)}{\partial t} - \frac{\partial(w + \beta v)}{\partial x}\Big|_{x=1} = 0,$$

we pose the supplementary difference boundary-conditions setting

$$\frac{(v_0^{p+1} + \alpha w_0^{p+1}) - (v_0^p + \alpha w_0^p)}{\tau} - \frac{(w_1^p + \alpha v_1^p) - (w_0^p + \alpha v_0^p)}{h} = 0,$$

$$\frac{(v_M^{p+1} + \beta w_M^{p+1}) - (v_M^p + \beta w_M^p)}{\tau} - \frac{(w_M^p + \beta v_M^p) - (w_{M-1}^p + \beta v_{M-1}^p)}{h} = 0.$$

Under the condition $r = \tau/h \leq 1$ show that:

a) if $\alpha = 1$, $\beta = -1$ the spectral stability criterion is satisfied;

b) if $\alpha = -1$ then, regardless of the choice of β the spectral stability criterion is not satisfied.

c) Find the conditions which α and β must obey so that the spectral stability criterion will be satisfied, taking account of the influence of boundary conditions.

§27. Representation of the solution of some model problems in the form of finite Fourier series.

Here we present examples of model problems whose solutions can be represented in the form of finite Fourier series. Such representations are of great value since they allow us to understand the properties of the given model problems, and thus of the class of problems they model.

First we must explain what is meant by a "Fourier series" for a net function.

1. Fourier series for net functions. Let us consider the set of all real functions $v = \{v_m\}$, defined at the points $x_m = mh$, $m = 0, 1, \ldots, M$,

$Mh = 1$, and vanishing at $m = 0$ and $m = M$. The set of such functions, along with the ordinary operations of addition, and of multiplication by real factors, forms a linear space. The dimension of this space is M-1, since the system of functions

$$\tilde{\psi}_m^{(k)} = \begin{cases} 0, & \text{if } m \neq k, \\ 1, & \text{if } m = k, \end{cases} \qquad k = 1, 2, \ldots, M-1,$$

clearly forms a basis. In fact each function $v = (v_0, v_1, \ldots, v_M)$, $v_0 = v_M = 0$ can be represented uniquely as a linear combination of the functions $\tilde{\psi}^{(1)}, \tilde{\psi}^{(2)}, \ldots, \tilde{\psi}^{(M-1)}$:

$$v = v_1\tilde{\psi}^{(1)} + \ldots + v_{M-1}\tilde{\psi}^{(M-1)}.$$

We introduce, in the space under consideration, a scalar product defined by the relation

$$(v, w) = h \sum_{m=0}^{M} v_m w_m. \tag{1}$$

Let us now show that the system of functions

$$\psi^{(k)} = \left\{\sqrt{2} \sin \frac{k\pi m}{M}\right\}, \qquad k = 1, 2, \ldots, M-1, \tag{2}$$

forms an orthonormal basis in this space, i.e. that

$$\left(\psi^{(k)}, \psi^{(r)}\right) = \begin{cases} 0, & k \neq r, \\ 1, & k = r, \end{cases} \tag{3}$$

$$k, r = 1, 2, \ldots, M-1.$$

For this purpose we note that

$$\sum_{m=0}^{M-1} \cos \frac{\ell\pi m}{M} = \frac{1}{2} \sum_{m=0}^{M-1} \left(e^{i\ell\pi m/M} + e^{-i\ell\pi m/M}\right) =$$

$$= \frac{1}{2} \frac{1 - e^{i\ell\pi}}{1 - e^{i\ell\pi/M}} + \frac{1}{2} \frac{1 - e^{-i\ell\pi}}{1 - e^{-i\ell\pi/M}} = \begin{cases} 0, & \text{if } \ell \text{ is even and } 0 < \ell < 2M, \\ 1, & \text{if } \ell \text{ is odd.} \end{cases}$$

It follows that, for $k \neq r$

$$\left(\psi^{(k)}, \psi^{(r)}\right) = 2h \sum_{m=0}^{M} \sin \frac{k\pi m}{M} \sin \frac{r\pi m}{M} = 2h \sum_{m=0}^{M-1} \sin \frac{k\pi m}{M} \sin \frac{r\pi m}{M} =$$

$$= h \sum_{m=0}^{M-1} \cos \frac{(k - r)\pi m}{M} - h \sum_{m=0}^{M-1} \cos \frac{(k + r)\pi m}{M} = 0,$$

while for k = r

$$(\psi^{(k)}, \psi^{(k)}) = h \sum_{m=0}^{M-1} \cos 0 - h \sum_{m=0}^{M-1} \cos \frac{2k\pi m}{M} = hM - h \cdot 0 = 1.$$

Using this orthonormal basis any net function $v = (v_0, v_1, \ldots, v_M)$ may be expanded in the sum

$$v = c_1 \psi^{(1)} + \ldots + c_{M-1} \psi^{(M-1)},$$

or

$$v_m = \sqrt{2} \sum_{k=1}^{M-1} c_k \sin \frac{k\pi m}{M}, \tag{4}$$

where

$$c_k = (v, \psi^{(k)}) = \sqrt{2}\, h \sum_{m=0}^{M} v_m \sin \frac{k\pi m}{M}.$$

Clearly, because of the orthonormality of the basis (2)

$$(v, v) = c_1^2 + c_2^2 + \ldots + c_{M-1}^2. \tag{5}$$

It is sum (4) which is the finite Fourier expansion of the net function $v = \{v_m\}$, and (5) is the exact analog of the Parseval equation in the ordinary theory of Fourier series.

In exactly the same way one can consider the finite Fourier series for functions on a square net. Define the net

$$x_m = mh, \quad y_n = nh, \quad 0 \leq mh \leq 1, \quad 0 \leq nh \leq 1,$$

where h = 1/M, for M a positive integer. The set of all real functions, $v = \{v_{mn}\}$, defined at the points of the net and vanishing at points lying on the boundary of the square, forms a linear space. Introduce, in this space, the scalar product

$$(v, w) = h^2 \sum_{n=0}^{M} \sum_{m=0}^{M} v_{mn} w_{mn}.$$

In the given linear space of dimension $(M-1)^2$ the system of functions

$$\psi^{(k,\ell)} = 2 \sin \frac{k\pi m}{M} \sin \frac{\ell\pi n}{M}, \qquad \begin{matrix} k = 1, 2, \ldots, M-1, \\ \ell = 1, 2, \ldots, M-1, \end{matrix}$$

forms an orthonormal basis

$$\left(\psi^{(k,\ell)},\ \psi^{(r,s)}\right) = \begin{cases} 0, & \text{for } k \neq r \text{ or } \ell \neq s, \\ 1, & \text{for } k = r \text{ and } \ell = s. \end{cases}$$

This follows from (3) if we note that

$$\left(\psi^{(k,\ell)},\ \psi^{(r,s)}\right) = \left(2 \sum_{m=0}^{M} \sin \frac{k\pi m}{M} \sin \frac{r\pi m}{M}\right)\left(2 \sum_{n=0}^{M} \sin \frac{\ell\pi n}{M} \sin \frac{s\pi n}{M}\right) =$$

$$= \left(\psi^{(k)},\ \psi^{(r)}\right)\left(\psi^{(\ell)},\ \psi^{(s)}\right),$$

Any function, $v = \{v_{mn}\}$, which vanishes on the boundary of the square may be expanded in a finite, two-dimensional, Fourier series

$$v_{mn} = 2 \sum_{k,\ell=1}^{M-1} c_{k\ell} \sin \frac{k\ell m}{M} \sin \frac{\ell\pi n}{M}, \tag{6}$$

where

$$c_{k\ell} = \left(v,\ \psi^{(k,\ell)}\right),$$

and the coefficients satisfy Parseval's equation

$$(v,\ v) = \sum_{k,\ell=1}^{M-1} c_{k\ell}^2. \tag{7}$$

In all the examples of difference boundary-value problems whose solutions we will write using finite Fourier series we encounter the expression

$$\Lambda_{xx} v_m \equiv \frac{1}{h^2}\left(v_{m+1} - 2v_m + v_{m-1}\right), \qquad m = 1, 2, \ldots, M-1. \tag{8}$$

Note that

$$\Lambda_{xx} \sin \frac{k\pi m}{M} = \frac{1}{h^2}\left[\sin \frac{k\pi(m+1)}{M} - 2 \sin \frac{k\pi m}{M} + \sin \frac{k\pi(m-1)}{M}\right] =$$

$$= \frac{2}{h^2}\left(\cos \frac{k\pi}{M} - 1\right) \sin \frac{k\pi m}{M} = \mu_k \sin \frac{k\pi m}{M}, \qquad m = 1, 2, \ldots, M-1,$$

where $\mu_k = -(4/h^2) \sin^2 (k\pi/2M)$.

In other words the basis (2) consists of eigenfunctions of the operator Λ_{xx}, which maps functions $v = \{v_m\}$ from the space of functions vanishing at $m = 0$ and $m = M$ into functions of the same space via the relations

$$w_m = \frac{1}{h^2}\left(v_{m+1} - 2v_m + v_{m-1}\right), \qquad m = 1, 2, \ldots, M-1.$$

To the eigenfunction $\psi^{(k)} = \sqrt{2} \sin(k\pi m/M)$ corresponds the eigenvalue

$$\mu_k = -\frac{4}{h^2}\sin^2\frac{k\pi}{2M}, \qquad k = 1, 2, \ldots, M-1. \tag{9}$$

2. Representation of the solutions of difference schemes for the heat equation on an interval. As a first example in which one can represent the solution in a finite Fourier series, consider the simplest difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ &m = 1, 2, \ldots, M-1, \qquad p = 0, 1, \ldots, [T/\tau]-1, \\ &u_0^p = u_M^p = 0, \\ &u_m^0 = \psi(mh) \end{aligned}\right\} \tag{10}$$

for the heat equation on an interval

$$\left.\begin{aligned} &u_t - u_{xx} = 0, \qquad 0 \le t \le T, \quad 0 \le x \le 1, \\ &u(0, t) = u(1, t) = 0, \qquad 0 \le t \le T, \\ &u(x, 0) = \psi(x), \qquad 0 \le x \le 1. \end{aligned}\right\} \tag{11}$$

Problem (10) may be rewritten thus:

$$\left.\begin{aligned} &u_m^{p+1} = u_m^p + \tau\Lambda_{xx}u_m^p = \left(E + \tau\Lambda_{xx}\right)u_m^p, \\ &u_m^0 = \psi(mh). \end{aligned}\right\} \tag{12}$$

Here E is the identity transformation: $Eu_m^p = u_m^p$, while $E + \tau\Lambda_{xx}$ is the operator effecting transitions from u^p to u^{p+1}, i.e. the level-to-level transition operator. As regards the net functions $u^p \equiv \{u_m^p\}$ of argument m we assume that, for each fixed p, they belong to the space under consideration, i.e $u_0^p = u_M^p = 0$.

We will look for a solution of Eq. (12) in the form

$$\lambda_k^p \psi(k) \equiv \lambda_k^p\left(\sqrt{2}\sin\frac{k\pi m}{M}\right).$$

Substituting this expression into the equation, dividing $\lambda_k^p\sqrt{2}\sin(n\pi m/M)$ out of both sides, and making use of (9), we get the following expression for λ_k:

$$\lambda_k = 1 + \tau\mu_k = 1 - \frac{4\tau}{h^2}\sin^2\frac{k\pi}{2M}, \qquad k = 1, 2, \ldots, M-1.$$

Because of the linearity of Eq. (12) the expression

$$u^p = \sum_{k=1}^{M-1} c_k \lambda_k^p \psi^{(k)} \tag{13}$$

is a solution for any arbitrary constants c_k. For $p = 0$ we get

$$u_m^0 = \sum_{k=1}^{M-1} c_k \left(\sqrt{2} \sin \frac{k\pi m}{M}\right).$$

Let us take, as the constants c_k, the coefficients in the expansion of the given function $u_m^0 = \psi(mh)$ in a finite Fourier series, i.e. set

$$c_k = (\psi, \psi^{(k)}) = h \sum_{m=0}^{M} \psi(mh)\left(\sqrt{2} \sin \frac{k\pi m}{M}\right).$$

Then the solution (13),

$$u_m^p = \sum_{k=1}^{M-1} c_k \left(1 - \frac{4\tau}{h^2} \sin^2 \frac{\pi k}{2M}\right)^p \left(\sqrt{2} \sin \frac{k\pi m}{M}\right), \tag{14}$$

will satisfy the given initial condition $u_m^0 = \psi(mh)$. Equation (14) then is the required representation of the solution of this problem in a finite Fourier series.

For fixed p, the coefficients, $c_k^{(p)}$, of the expansion

$$u_m^p = \sum c_k^{(p)} \psi^{(k)}$$

of the function $u_m^{(p)}$, of argument m, in the orthonormal basis functions $\psi^{(k)} = \sqrt{2} \sin(k\pi m/M)$, have the form

$$c_k^{(p)} = c_k \lambda_k^p.$$

Therefore, taking note of Parseval's equality, we have

$$(u^{p+1}, u^{p+1}) = \sum_{k=1}^{M-1} \left|c_k^{(p+1)}\right|^2 = \sum_{k=1}^{M-1} \left|c_k \lambda_k^{p+1}\right|^2 \leq$$

$$\leq \max_k |\lambda_k|^2 \sum_{k=1}^{M-1} \left|c_k \lambda_k^p\right|^2 = \max_k |\lambda_k|^2 (u^p, u^p)$$

where, moreover, the strict equality $(u^{p+1}, u^{p+1}) = \max_k |\lambda_k|^2 (u^p, u^p)$ is attained if, as u^0, one takes that $\psi^{(k)}$ for which $|\lambda_k|$ is greatest.

If $\max_k |\lambda_k|^2 < 1$, then

$$(u^{p+1}, u^{p+1}) \leq (u^p, u^p). \tag{15}$$

The positive-definite quadratic form

$$(Au^p, u^p),$$

where A is a square matrix, brings to mind the expression for energy in the equations of mathematical physics. Therefore an inequality of the form

$$(Au^{p+1}, u^{p+1}) \leq (Au^p, u^p)$$

for the solution of a difference boundary-value problem is commonly called an *energy inequality*.

* * * * * *

Thus bound (15) is the simplest energy inequality. When some sort of energy inequality exists it is natural to relate the norms $\|\cdot\|_{U_h}$ and $\|\cdot\|_F$ to the form (Au^p, u^p) and, in particular, to take $\|u^{(h)}\|_{U_h} = \max_p (Au^p, u^p)^{1/2}$. Such norms are called "energy norms".

* * *

The inequality $\max_k |\lambda_k|^2 \leq 1$ is satisfied, as one easily can see, if

$$r = \frac{\tau}{h^2} \leq \frac{1}{2}.$$

For

$$r = \text{const} > \frac{1}{2}$$

and for small enough h, there will be a λ_k for which $|\lambda_k| > 1$. We cannot then have stability for any reasonable* choice of norms.

Consider the difference scheme of more general form

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \left[(1 - \sigma)\Lambda_{xx}u^p + \sigma\Lambda_{xx}u^{p+1}\right] = 0,$$

$$u_m^0 = \psi(mh)$$

for this same heat-conduction problem (11). Here σ is some free parameter.

We look for a solution of the form

*See §13.

$$u_m^p = \lambda_k^p \sqrt{2} \sin \frac{k\pi m}{M}, \qquad k = 1, 2, \ldots, M-1,$$

where λ_k remains to be determined.

Substituting this expresion into the difference equation we get an equation which λ_k must satisfy:

$$\lambda_k = 1 + \tau(1 - \sigma)\mu_k + \tau\sigma\lambda_k\mu_k.$$

Thus

$$\lambda_k = \frac{1 - \dfrac{4(1-\sigma)\tau}{h^2} \sin^2 \dfrac{k\pi}{2M}}{1 + \dfrac{4\sigma\tau}{h^2} \sin^2 \dfrac{k\pi}{2M}}, \qquad k = 1, 2, \ldots, M-1.$$

As before

$$(u^{p+1}, u^{p+1}) \leq \max_k |\lambda_k|^2 (u^p, u^p).$$

Energy inequality (15) is satisfied if

$$\max|\lambda_k| \leq 1$$

or

$$\left|1 - 4(1-\sigma)r \sin^2 \frac{k\pi}{2M}\right| \leq \left|1 + 4\sigma r \sin^2 \frac{k\pi}{2M}\right|, \qquad r = \frac{\tau}{h^2}.$$

It is clear that for $1 \geq \sigma \geq 1/2$ this inequality, and also energy inequality (15), are satisfied for any r. If $\sigma = 0$ the difference scheme takes the form of the explicit scheme, already considered above and, as we have seen, if energy inequality (15) is then to be satisfied for all h it is necessary that $r \leq 1/2$.

3. Representation of the solution of difference schemes for the two-dimensional heat-conduction problem. We now consider the two-dimensional heat-conduction problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \\ &u(x, y, 0) = \psi(x, y), \\ &u(x, y, t)|_\Gamma = 0, \end{aligned} \right\} \quad \begin{aligned} &0 \leq x \leq 1, \quad 0 \leq y \leq 1, \\ &0 \leq t \leq T. \end{aligned} \qquad (16)$$

Here Γ is the lateral surface of the parallelopiped $0 \leq x, y \leq 1$, $0 \leq t \leq T$.

We construct the net $(x_m, y_n, t_p) = (mh, nh, p\tau)$ with $h = 1/M$ for some positive integer M. As the set D_h we take those points of the net inside and on the boundary of the parallelopiped $0 \leq x, y \leq 1$, $0 \leq t \leq T$.

Let us now introduce the notation

$$\Lambda_{xx}u^p_{mn} = \frac{u^p_{m+1,n} - 2u^p_{mn} + u^p_{m-1,n}}{h^2},$$

$$\Lambda_{yy}u^p_{mn} = \frac{u^p_{m,n+1} - 2u^p_{mn} + u^p_{m,n-1}}{h^2}.$$

The operators Λ_{xx} and Λ_{yy} are perfectly analogous, except that the first acts on the variable m, with n and p treated as parameters, while the second acts on n, in its turn treating m and p as parameters.

The simplest difference scheme for problem (16) is

$$\left.\begin{aligned} &\frac{u^{p+1}_{mn} - u^p_{mn}}{\tau} = \Lambda_{xx}u^p_{mn} + \Lambda_{yy}u^p_{mn}, \qquad 0 \leq mn, \quad nh \leq 1, \\ &u^0_{mn} = \psi(mh, nh), \qquad 0 \leq p\tau < T-\tau, \\ &u^p_{mn}\Big|_\Gamma = 0. \end{aligned}\right\} \tag{17}$$

We will look for solutions of the difference equation, under the condition $u^p_{mn}\Big|_\Gamma = 0$, of the form

$$u^p_{mn} = \lambda^p_{k\ell}\psi^{(k,\ell)}_{mn}.$$

Note that

$$\Lambda_{xx}\psi^{(k,\ell)} = \Lambda_{xx}\left(\psi^{(k)}_m\psi^{(\ell)}_n\right) = \psi^{(\ell)}\Lambda_{xx}\psi^{(k)} = \mu^{(k)}\psi^{(k)}\psi^{(\ell)} = \mu^{(k)}\psi^{(k,\ell)},$$

$$\Lambda_{yy}\psi^{(k,\ell)} = \Lambda_{yy}\left(\psi^{(k)}_m\psi^{(\ell)}_n\right) = \mu^{(\ell)}\psi^{(k,\ell)}.$$

Therefore we get, for $\lambda_{k\ell}$, the expression

$$\frac{\lambda_{k\ell} - 1}{\tau} = \left[\mu^{(k)} + \mu^{(\ell)}\right]$$

or

$$\lambda_{k\ell} = 1 - \frac{4\tau}{h^2}\left(\sin^2\frac{k\pi}{2M} + \sin^2\frac{\ell\pi}{2M}\right).$$

The solution

$$u^p = \sum_{k,\ell=1}^{M-1} c_{k\ell}\lambda_{k\ell}^p \psi^{(k,\ell)} \tag{18}$$

satisfies the conditions on the lateral boundary for any choice of the constants $c_{k\ell}$. For $p = 0$ this solution takes the form

$$u^0 = \sum c_{k\ell}\psi^{(k,\ell)}.$$

If the initial condition

$$u_{mn}^0 = \psi(mh, nh) = \sum c_{k\ell}\psi_{mn}^{(k,\ell)},$$

is to be satisfied, then the $c_{k\ell}$ must be taken to be the Fourier coefficients of the function $\psi(mh, n\tau)$, i.e.

$$c_{k\ell} = h^2 \sum_{m,n=0}^{M} \psi(mh, nh)\left(2 \sin \frac{k\pi m}{M} \sin \frac{\ell\pi m}{M}\right). \tag{19}$$

According to Eq. (18) the coefficient of $\psi^{(k,\ell)}$ in the Fourier expansion of u^p is equal to $c_{k\ell}\lambda_{k\ell}^p$. Therefore

$$(u^p, u^p) = \sum_{k,\ell} \left|c_{k\ell}\lambda_{k\ell}^p\right|^2.$$

For any given p we may, therefore, write

$$(u^{p+1}, u^{p+1}) = \sum_{k,\ell-1}^{M-1} \left|c_{k\ell}\lambda_{k\ell}^{p+1}\right|^2 \leq$$

$$\leq \max_{k,\ell} |\lambda_{k\ell}|^2 \sum_{k,\ell} \left|c_{k\ell}\lambda_{k\ell}^p\right|^2 = \max_{k,\ell} |\lambda_{k\ell}|^2 (u^p, u^p).$$

Equality is attained if we take, as $\psi(mh, nh)$, that eigenfunction of the operator $E + \tau(\Lambda_{xx} + \Lambda_{yy})$ (i.e. the operator which effects transitions from level $t = p\tau$ to level $p = (n + 1)\tau$) with eigenvalue, $\lambda_{k\ell}$, largest in modulus.

If $\max_{k,\ell}|\lambda_{k\ell}| \leq 1$ we have the energy inequality

$$(u^{p+1}, u^{p+1}) \leq (u^p, u^p). \tag{20}$$

As k and ℓ run through the values $k,\ell = 1, 2, \ldots, M-1$, the eigenvalues run through a finite set of points, on the real axis, lying to the left of the point $\lambda = 1$. The leftmost point is reached for $k = \ell = M-1$:

$$\lambda_{M-1,M-1} = 1 - 8r \sin^2 \frac{(M-1)\pi}{2M} =$$

$$= 1 - 8r \cos^2 \frac{\pi}{2M} = 1 - 8r + O\left(\frac{1}{M^2}\right).$$

Therefore the inequality $\max|\lambda_{k\ell}| \leq 1$ is satisfied for $-1 \leq 1 - 8r$, $r \leq 1/4$

For the implicit scheme

$$\left.\begin{aligned} \frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} &= \Lambda_{xx} u_{mn}^{p+1} + \Lambda_{yy} u_{mn}^{p+1}, \\ u_{mn}^{0} &= \psi(mh, nh), \\ u_{mn}^{p}\Big|_{\Gamma} &= 0 \end{aligned}\right\}$$

the solution has the form

$$u_{mn}^{p} = \sum_{k,\ell} c_{k\ell} \lambda_{k\ell}^{p} \psi_{mn}^{(k,\ell)},$$

where

$$\lambda_{k\ell} = \frac{1}{1 + 4r\left(\sin^2 \frac{k\pi}{2M} + \sin^2 \frac{\ell\pi}{2M}\right)},$$

and the coefficients $c_{k\ell}$ are determined, as before, by Eq. (19). Here $0 < \lambda_{k\ell} < 1$, and the energy inequality (20) holds for any value of $r = \tau/h^2$.

* * * * * *

4. Representation of the solution of a difference scheme for the vibrating string problem. Consider the example of the three-level scheme $L_h u^{(h)} = f^{(h)}$, approximating the problem of the vibrating string with fixed ends:

$$\left.\begin{aligned} &u_{tt} - u_{xx} = 0, && 0 \leq x \leq 1, \quad 0 \leq t \leq T, \\ &u(0, t) = (1, t) = 0, && 0 \leq t \leq T, \\ &u(x, 0) = \psi_0(x), && 0 \leq x \leq 1, \\ &u_t(x, 0) = \psi_1(x), && 0 \leq x \leq 1. \end{aligned}\right\}$$

Define

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{\tau^2} - \Lambda_{xx} u_m^n = 0, \\ u_0^p = u_M^p = 0, \\ u_m^0 = \psi_0(mh), \\ u_m^1 = \tilde{\psi}_m, \end{cases}$$

where

$$\tilde{\psi}_m = u(mh,0) + \tau u_t(mh,0) + \frac{\tau^2}{2} u_{tt}(mh,0) = \psi_0(mh) + \tau\psi_1(mh) + \frac{\tau^2}{2}\psi_0''(x).$$

We will look for a solution of the difference equation satisfying the condition $u_0^p = u_M^p = 0$, and of the form

$$u_m^p = \lambda^p \sqrt{2} \sin \frac{k\pi m}{M} \left(\equiv \lambda^p \psi^{(k)} \right), \tag{21}$$

ignoring, for the moment, the initial conditions $u_m^0 = \psi_0(mh)$ and $u_m^1 = \tilde{\psi}_m$.

For λ we get the following equation

$$\frac{\lambda - 2 + \frac{1}{\lambda}}{\tau^2} - \mu_k = 0, \qquad \mu_k = -\frac{4}{h^2} \sin^2 \frac{k\pi}{2M},$$

$$\lambda^2 - 2\left(1 - 2r^2 \sin^2 \frac{k\pi}{2M}\right) + 1 = 0, \qquad r = \frac{\tau}{h},$$

$$\lambda_1(k) = 1 - 2r^2 \sin^2 \frac{k\pi}{2M} + \sqrt{\left(1 - 2r^2 \sin^2 \frac{k\pi}{2M}\right)^2 - 1},$$

$$\lambda_2(k) = 1 - 2r^2 \sin^2 \frac{k\pi}{2M} - \sqrt{\left(1 - 2r^2 \sin^2 \frac{k\pi}{2M}\right)^2 - 1}.$$

Thus there are two solutions of the desired form (21):

$$\lambda_1^p(k)\psi^{(k)} \text{ and } \lambda_2^p(k)\psi^{(k)}.$$

Because of the linearity of the problem the expression

$$u_m^p = \sum_{k=1}^{M-1} \left[\alpha_k \lambda_1^p(k) + \beta_k \lambda_2^p(k) \right] \psi_m^{(k)}$$

is a solution for any choice of the numbers α_k and β_k, k = 1, 2, ..., M-1. For p = 0 and p = 1 one gets, respectively,

$$u_m^0 = \psi_0(mh) = \sum_{k=1}^{M-1} (\alpha_k + \beta_k)\psi^{(k)},$$

$$u_m^1 = \tilde{\psi}_m = \sum_{k=1}^{M-1} [\alpha_k \lambda_1(k) + \beta_k \lambda_2(k)]\psi^{(k)} .$$

These relations determine the values of α_k and β_k. The sum $\alpha_k + \beta_k$ must be the Fourier coefficient of the expansion $\psi_0(mh)$ in the functions $\{\overset{k}{\psi}{}^{(h)}\}$, i.e.

$$\alpha_k + \beta_k = h \sum_{m=0}^{M} \psi_0(mh)\left(\sqrt{2} \sin \frac{k\pi m}{M}\right).$$

Similarly

$$\alpha_k \lambda_1(k) + \beta_k \lambda_2(k) = h \sum_{m=0}^{M} \tilde{\psi}_m\left(\sqrt{2} \sin \frac{k\pi m}{M}\right).$$

* * *

Expansion of the solution of a difference equation in a finite Fourier series is a device used not only to determine under what conditions an energy inequality is satisfied. Below we will often use such expansions for various purposes in the qualitative study of model problems.

It must be noted, however, that the representation of a solution as a finite Fourier series is rarely used directly for the computation of the solution. A good computational method must be useful for a wide range of problems. The above difference schemes may easily be generalized to treat problems with variable coefficients, and regions with curved boundaries. Further, we can expect that, in the modified problems, such properties as the validity of an energy inequality will be preserved. But any such change in the problem eliminates the possibility that its solution can be represented as a finite Fourier series: we generally cannot find eigenfunctions of the level-to-level transition operator and cannot compute corresponding eigenvalues.

PROBLEMS

1. For the two-dimensional heat-conduction problem in a square region, with a solution vanishing on the boundary, consider the difference scheme

$$\left.\begin{aligned} &\frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} = \sigma\left[\Lambda_{xx}u_{mn}^{p+1} + \Lambda_{yy}u_{mn}^{p+1}\right] + \\ &\quad + (1 - \sigma)\left[\Lambda_{xx}u_{mn}^{p} + \Lambda_{yy}u_{mn}^{p}\right], \qquad 0 < mh, \quad nh < 1, \\ &u_{mn}^{p}\Big|_{\Gamma} = 0, \qquad u_{mn}^{0} = \psi(mh, nh) \end{aligned}\right\}$$

(in the notation introduced in the text of the above section). Write out the solution of this problem in the form of a finite Fourier series. Determine for what values of σ, $0 \leq \sigma \leq 1$, the energy inequality $(u^{p+1}, u^{p+1}) \leq (u^{p}, u^{p})$ is valid for any choice of $r = \tau/h^2$.

For which σ can we write, given any $u^{p} \neq 0$, the strict inequality $(u^{p+1}, u^{p+1}) < (u^{p}, u^{p})$, regardless of the choice of r or of the step-width, h?

2. Write the solutions of the differential problem

$$u_t - u_{xx} = 0, \qquad 0 \leq x \leq 1, \quad 0 \leq t \leq T,$$

$$u|_{\Gamma} = 0, \qquad u(x, 0) = \psi(x)$$

and the difference problem

$$\left.\begin{aligned} &\frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} - \Lambda_{xx}u_{mn}^{p} = 0, \\ &u_{mn}^{p}\Big|_{\Gamma} = 0, \qquad u_{mn}^{0} = \psi(mh, nh) \end{aligned}\right\}$$

respectively, in a Fourier series and finite Fourier series. Prove by comparing these series for $r \leq 1/2$, assuming the boundedness of $\psi''(x)$, that the solution of the difference scheme converges to the solution of the differential problem. Prove that for $r > 1/2$ there is, in general, no covergence

3. Write out in a finite Fourier series the solution of the Dirichlet problem for the Poisson equation in the square region $0 \leq x, y \leq 1$:

$$\Lambda_{xx}u_{mn} + \Lambda_{yy}u_{mn} = \phi(mh, nh), \qquad 0 < mh, \quad nh < 1,$$

with the boundary condition

a) $$u_{mn}\Big|_{\Gamma} = 0;$$

b)
$$u_{mn} = \begin{cases} nh, & \text{if } m = 0 \\ 1 + mh, & \text{if } n = M, \\ mh, & \text{if } n = 0, \\ 1 + nh, & \text{if } m = M. \end{cases}$$

Hint for 3b: $u_{mn} = mh + nh + Z_{mn}$, where Z_{mn} satisfies homogeneous conditions on the boundary.

§28. The maximum principle

We have already seen by way of examples, in §§21 and 24, how to prove stability with the aid of the maximum principle. Here we analyze two more interesting examples where one can prove stability via this method: implicit and explicit difference schemes approximating the boundary-value problem for the heat equation

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - a^2(x, t)\frac{\partial^2 u}{\partial x^2} = \phi(x, t), \qquad 0 < x < 1, \quad 0 \le t \le T, \\ &u(x, 0) = \psi_0(x), \qquad 0 < x < 1, \\ &u(0, t) = \psi_1(t), \qquad 0 \le t \le T, \\ &u(1, t) = \psi_2(t), \qquad 0 \le t \le T. \end{aligned}\right\} \quad (1)$$

1. Explicit difference scheme. Let us consider the explicit difference scheme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - a^2(mh, n\tau)\dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = \phi(mh, n\tau), \\ \qquad m = 1, 2, \ldots, M-1; \quad n = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0 = \psi_0(mh), \qquad m = 0, 1, \ldots, Mh, \\ u_0^n = \psi_1(n\tau), \qquad n = 1, 2, \ldots, [T/\tau], \\ u_M^n = \psi_2(n\tau), \qquad n = 1, 2, \ldots, [T/\tau], \end{cases} \quad (2)$$

where $M = 1/h$ is a positive integer.

The Von Neumann spectral criterion, together with the principle of frozen coefficients leads, as we saw in §26, to the necessary stability condition

$$\frac{\tau}{h^2} \le \frac{1}{2 \max\limits_{x,t} a^2(x, t)}. \quad (3)$$

We will now show that, given this condition, the above scheme really is stable if norms are defined via the equations

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &= \max_n \max_n \left|u_m^n\right| , \\ \|f^{(h)}\|_{F_h} &= \max\Big(\max_m|\psi_0(mh)|,\ \max_n|\psi_1(n\tau)|, \\ &\qquad \max_n|\psi_2(n\tau)|,\ \max_{m,n}|\phi(x_m,\ t_n)|\Big). \end{aligned}\right\} \tag{4}$$

Let us first establish the validity of the inequality (the "maximum principle")

$$\max_m\left|u_m^{n+1}\right| \le \max\Big[\max_n|\psi_1(t_n)|,\ \max_n|\psi_2(t_n)|, \max_m\left|u_m^n\right| + \tau \max_{m,n}|\phi(x_m,\ t_n)|\Big]. \tag{5}$$

For this purpose we rewrite the difference equation on which scheme (2) is based, casting it in the form

$$u_m^{n+1} = \big(1 - 2ra^2(x_m,\ t_n)\big)u_n^n + ra^2(x_m,\ t_n)\big(u_{m-1}^n + u_{m+1}^n\big) + \tau\phi(x_m,\ t_n), \quad m = 1,\ \ldots,\ M-1. \tag{6}$$

If condition (3) is satisfied the expression $1 - 2ra^2(x_m,\ t_n)$ is non-negative. Therefore we may write

$$\left|u_m^{n+1}\right| \le \big|1 - 2ra^2(x_m,\ t_n)\big|\max_k\left|u_k^n\right| + ra^2(x_m,\ t_n)\left(\max_k\left|u_k^n\right| + \max_k\left|u_k^n\right|\right) + \tau \max_{m,n}|\phi(x_m,\ t_n)| = \max_k\left|u_k^n\right| + \tau \max_{m,n}|\phi(x_m,\ t_n)|, \quad m = 1,\ 2,\ \ldots,\ M-1. \tag{7}$$

Noting that

$$u_0^{n+1} = \psi_1\big[(n+1)\tau\big], \qquad u_M^{n+1} = \psi_2\big[(n+1)\tau\big], \tag{8}$$

we arrive at the maximum principle, (5).

Next we split the solution, $u^{(h)}$, of the problem $L_h u^{(h)} = f^{(h)}$ into two terms, $u^{(h)} = v^{(h)} + w^{(h)}$, defining $v^{(h)}$ and $w^{(h)}$, respectively, as the solutions of the following equations

$$L_h v^{(h)} = \begin{cases} 0 \\ \psi_0(x_m) \\ \psi_1(t_n) \\ \psi_2(\tau_n) \end{cases} \qquad L_h w^{(h)} = \begin{cases} \phi(x_m, t_n), \\ 0 \\ 0 \\ 0. \end{cases} \tag{9}$$

By virtue of bound (5)

$$\max_m \left|v_m^{n+1}\right| \le \max\left[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m \left|v_m^n\right|\right],$$

$$\max_m \left|v_m^{n}\right| \le \max\left[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m \left|v_m^{n-1}\right|\right],$$

$$\max_m \left|v_m^{n-1}\right| \le \max\left[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m \left|v_m^{n-2}\right|\right],$$

. . . . . .

$$\max_m \left|v_m^{1}\right| = \max\left[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m |\psi_0(x_m)|\right].$$

Similarly, again by virtue of bound (5),

$$\max_m \left|w_m^{n+1}\right| \le \max_m \left|w_m^{n}\right| + \tau \max_{m,k} |\phi(x_m, t_k)| \le$$

$$\le \max_m \left|w_m^{n-1}\right| + 2\tau \max_{m,k} |\phi(x_m, t_k)| \le$$

. . . . . .

$$\le \max_m \left|w_m^{0}\right| + (n+1)\tau \max_{m,k} |\phi(x_m, t_k)| \le T \max_{m,k} |\phi(x_m, t_k)|.$$

From the bounds on v_m^{n+1} and w_m^{n+1} it follows that

$$\max_m \left|u_m^{n+1}\right| = \max_m \left|v_m^{n+1} + w_m^{n+1}\right| \le \max_m \left|v_m^{n+1}\right| + \max_m \left|w_m^{n+1}\right| \le$$

$$\le \max\left[\max_k |\psi_1(t_k)|,\ \max_k |\psi_2(t_k)|,\ \max_m |\psi_0(x_m)|\right] +$$

$$+ T \max_{m,n} |\phi(x_m, t_n)| \le c\|f^{(h)}\|_{F_h}, \tag{10}$$

where

$$c = 2 \max(1, T). \tag{11}$$

This inequality is valid for all n. Therefore

$$\|u^{(h)}\|_{U_h} \le c\|f^{(h)}\|_{F_h}, \tag{12}$$

and the scheme is stable.

2. Implicit difference scheme. Now let us consider the implicit difference scheme

$$L_h u^h = \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - a^2(x_m, t_n)\dfrac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2} = \phi(x_m, t_n), \\ u_m^0 = \psi_0(x_m), \\ u_0^n = \psi_1(t_n), \\ u_M^n = \psi_2(t_n). \end{cases} \tag{13}$$

In order to compute the u_m^{n+1}, given u_m^n, m = 0, 1, ..., N, one must solve the problem

$$\left.\begin{aligned} \frac{u_m^{n+1}}{\tau} - a^2(x_m, t_n)\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2} = \frac{u_m^n}{\tau} + \phi(x_m, t_n), \\ u_0^{n+1} = \psi_1(t_{n+1}), \qquad u_M^{n+1} = \psi_2(t_{n+1}). \end{aligned}\right\} \tag{14}$$

After both sides of the above difference equation have been multiplied by τ this problem takes the form

$$\left.\begin{aligned} a_m v_{m-1} + b_m v_m + c_m v_{m+1} = g_m, \qquad m = 1, 2, \ldots, M-1, \\ v_0 = \alpha, \qquad v_M = \beta, \end{aligned}\right\} \tag{15}$$

where

$$v_m = u_m^{n+1}, \qquad a_m = a^2(x_m, t_n)r,$$

$$b_m = -2a^2(x_m, t_n)r - 1, \qquad c_m = a^2(x_m, t_n)r,$$

$$g_m = -u_m^n - \tau\phi(x_m, t_n), \qquad \alpha = \psi_1(t_{n+1}), \qquad \beta = \psi_2(t_{n+1}).$$

The coefficients a_m, b_m, c_m satisfy the conditions

$$a_m > 0, \quad c_m > 0, \quad |b_m| > a_m + c_m + \phi \qquad (\delta > 0).$$

Therefore, as shown in §§4 and 5, the problem has a unique solution

$$(v_0, v_1, \ldots, v_M) = (u_0^{n+1}, u_1^{n+1}, \ldots, u_M^{n+1}),$$

which can be computed by FEBS.

To prove stability we must still demonstrate the validity of inequality (12). For this purpose we first prove inequality (5) (the maximum principle), from which bounds (10) and (12) may be derived exactly, word for word, as in the case of explicit scheme (2).

Of all the quantities, u_m^{n+1}, equal in modulus to $\max_m |u_m^{n+1}|$, select that one whose index, m, has the smallest value $m = m^*$. If $m^* = 0$ or $m^* = M$ then, in view of (8), the validity of inequality (5) is obvious. Suppose $m^* \neq 0$ and $m^* \neq M$. Let us write out Eq. (14) for $m = m^*$:

$$ra^2(x_{m^*}, t_n)u_{m^*-1}^{n+1} - (1 + 2ra^2(x_{m^*}, t_n))u_{m^*}^{n+1} + ra^2(x_{m^*}, t_n)u_{m^*+1}^{n+1} = -u_{m^*}^n - \tau\phi(x_{m^*}, t_n).$$

Suppose, for the sake of definiteness, that $u_{m^*}^{n+1} > 0$. Then the left-hand side of this equation can be bounded thus:

$$ra^2(x_{m^*}, t_n)u_{m^*-1}^{n+1} - (1 + 2ra^2(x_{m^*}, t_n))u_{m^*}^{n+1} + ra^2(x_{m^*}, t_n)u_{m^*+1}^{n+1} = ra^2(x_{m^*}, t_n)[(u_{m^*-1}^{n+1} - u_{m^*}^{n+1}) + (u_{m^*+1}^{n+1} - u_{m^*}^{n+1})] - u_{m^*}^{n+1} \leq -u_{m^*}^{n+1}.$$

Therefore

$$-u_{m^*}^{n+1} \geq -u_{m^*}^n - \tau\phi(x_{m^*}, t_n),$$

$$\max_m \left|u_m^{n+1}\right| = u_{m*}^{n+1} \leq \left|u_{m*}^{n} + \tau\phi(x_{m*}, t_n)\right| \leq$$

$$\leq \max_m \left|u_m^n\right| + \tau \max_{m,n} |\phi(x_m, t_n)|.$$

3. Comparison of the explicit and implicit difference schemes. Thus we have proven inequality (5), and also the maximum principle implied by (5). At the same time we have also proven the stability of implicit difference scheme (14) in norms (4).

We stress the essential difference between the explicit and implicit schemes (2) and (14). The first requires for stability the step-size limitation

$$\tau < \frac{1}{2 \max a^2(x, t)} h^2,$$

which becomes very restrictive if the coefficient $a^2(x, t)$ takes on large values even in the small neighborhood of some single point. The second, implicit, difference scheme remains stable for any arbitrary relation between the step-sizes h and τ.

Difference schemes which, like the implicit scheme (14), remain stable for any arbitrary relation between net step-sizes are called *absolutely stable* or *unconditionally stable*. Explicit scheme (2) is not unconditionally stable.

This Page Intentionally Left Blank

Chapter 9

Difference Scheme Concepts in the Computation of Generalized Solutions

§29 The Generalized Solution

In all the examples so far considered we have assumed that there existed "sufficiently smooth" solutions of the differential boundary-value problem, and based the construction of difference schemes on the approximate replacement of derivatives, in a differential equation, by difference relations. But differentiable functions do not suffice for the description of many physics processes. Thus, for example, experiments show that the distributions of pressure, density and temperature in the supersonic flow of a non-viscous gas are described by functions with jump-dicontinuities, discontinuities called "shock waves." Discontinuities may develop, in the course of time, even from smooth initial conditions.

The corresponding differential boundary-value problems do not have smooth solutions. It will be necessary for us to broaden the concept of a solution and, in some natural way, to introduce generalized solutions which can be discontinuous. There are two basically different ways to do this.

The first approach is to write the physical conservation laws (conservation of mass, momentum, energy, etc.) not in differential, but in integral form. Then they are meaningful even for discontinuous functions which cannot be differentiated but can be integrated.

The second consists in that one artificially introduces into the differential equations terms such that the resulting equations will have smooth solutions. These artificially introduced terms may, in the case of gas dynamics problems, be intepreted as small viscosity terms which smooth the discontinuities in the solution. Eventually the coefficients of these "viscous" terms tend to zero, and the limit approached by the solution is taken to be the generalized solution of the original problem.

We clarify the definition of the generalized solution, and the computational methods which may be needed to compute this solution, via the example of the following Cauchy problem

$$\begin{aligned} &\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \qquad 0 < t < T, \quad -\infty < x < \infty, \\ &u(x, 0) = \psi(x), \qquad\qquad -\infty < x < \infty, \end{aligned} \tag{1}$$

which is the simplest model gas dynamics problem among all those in which discontinuous solutions develop from smooth initial data.

1. Mechanism generating discontinuities. Let us assume, first, that problem (1) has the smooth solution u(x, t). Draw the lines, x = x(t), defined by the equation

$$\frac{dx}{dt} = u(x, t). \tag{2}$$

These lines are called *characteristics* of the equation $u_t + uu_x = 0$.

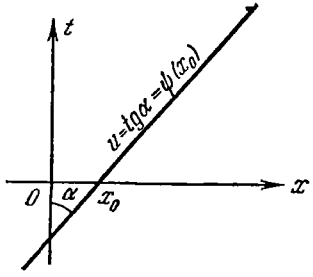


Fig. 28.

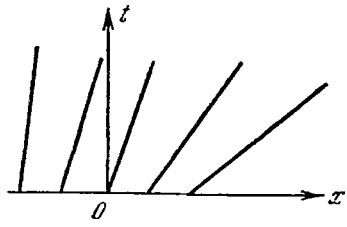


Fig 29.

Along each characteristic x = x(t) the solution u(x, t) may be considered a function of t alone:

$$u(x, t) - u[x(t), t] = u(t).$$

Clearly

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0.$$

Therefore along each characteristic the solution is constant, u(x, t) = const. But, by virtue of Eq. (2), it follows from u = const that each characteristic is a straight line $x = ut + x_0$. Here x_0 is the abscissa of the point $(x_0, 0)$ from which the characteristic emerges, and $u = \psi(x_0)$ is the tangent of the angle which it makes with the t axis. The assignment of initial values u(x, 0) = ψ(x) thus determines, in a manner easily visualised, both the pattern of characteristics, and the value of the solution at each point of the half-plane t > 0 (Fig. 28).

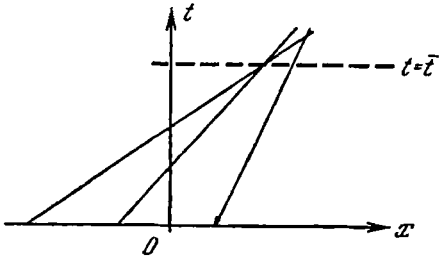


Fig. 30.

Let us note at once that, under the assumption that there exists a smooth solution u(x, t), the char-

acteristics cannot intersect; otherwise each characteristic would bring to the intersection point its own solution value, and the solution would not be a single-valued function. For a monotonically increasing function $\psi(x)$ the angle α increases with increasing x_0, and the characteristics cannot intersect (Fig. 29). But in the case where $\psi(x)$ decreases with increasing x_0 the characteristics converge and must intersect, regardless of the smoothness of $\psi(x)$. A smooth solution of problem (1) ceases to exist at the moment $t = \bar{t}$, when at least two characteristics intersect (Fig. 30).

Graphs of the functions $u = u(x, t)$ at $t = 0$, $\frac{1}{2}\bar{t}$ and $\bar{t}$ are shown in Fig. 31.

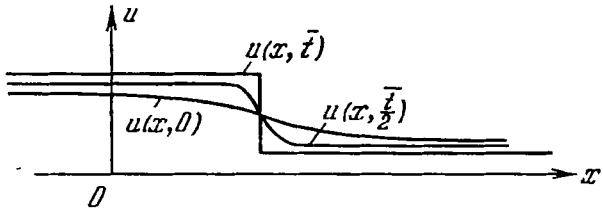


Fig. 31.

2. Definition of the generalized solution. We recall Green's formula, which we will use to determine the generalized solution of problem (1). Let D be an arbitrary region, with boundary Γ, on the xt plane, and suppose that $\Phi_1(x, t)$ and $\Phi_2(x, t)$ have, in region D, partial derivatives which are continuous up to the boundary. Then one can derive the following equation

$$\int_D \int \left(\frac{\partial \Phi_1}{\partial t} + \frac{\partial \Phi_2}{\partial x} \right) dx\, dt = \oint_\Gamma \left(\Phi_1 dx - \Phi_2 dt \right), \tag{3}$$

due to Green. The expression $\left(\partial \Phi_1 / \partial t\right) + \left(\partial \Phi_2 / \partial t\right)$ is the *divergence* of the vector $\Phi = \left(\Phi_1, \Phi_2\right)^T$. Green's formula (3) states that the integral of the divergence of the vector field Φ over the region D is equal to the current of vector Φ across the boundary, Γ, of that region.

We go on, now, to define the concept of a generalized solution. First we will write the differential equation of problem (1) in divergence form:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) = 0. \tag{4}$$

Integrating both sides of Eq. (4) over any arbitrary region, D, lying in the half-plane $t \geq 0$, we get

$$0 = \int_D \int \left[\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{u^2}{2} \right) \right] dx\, dt = \oint_\Gamma \left(u\, dx - \frac{u^2}{2}\, dt \right).$$

Thus each differentiable solution of Eq. (4) satisfies the integral relation

$$\oint_{\Gamma}\left(u\,dx - \frac{u^2}{2}\,dt\right) = 0, \tag{5}$$

where Γ is an arbitrary contour lying in the half-plane $t > 0$. Equation (5) expresses a certain conservation law: i.e. the current of the vector $(u, u^2/2)^T$ across any closed contour vanishes.

Let us now prove that, conversely, if a smooth function satisfies the integral conservation law (5) for every contour Γ, then at each point (x_0, t_0), $t_0 > 0$, Eq. (4) is satisfied. Assume the contrary and suppose, for the sake of definiteness, that at some point (x_0, t_0)

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)\Bigg|_{\substack{x = x_0,\\ t = t_0}} > 0.$$

Then by continuity one can find a circle D, with center at (x_0, t_0) and perimeter Γ, small enough so that everywhere within it $u_t + (u^2/2)_x > 0$. Thus

$$0 = \oint_{\Gamma}\left(u\,dx - \frac{u^2}{2}\,dt\right) = \iint_D \left[\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)\right]dx\,dt > 0.$$

The contradiction $0 > 0$ proves that, for a smooth function $u(x, t)$, (5) implies (4), so that (4) and (5) are equivalent. But in the case of a discontinuous function $u(x, t)$ the differential equation (1) or (4), on a line of discontinuity, will lose its meaning, while the integral condition (5) will not. Therefore any piecewise-differentiable function which satisfies conditions (5), for every arbitrary contour Γ in the half-plane $t \geq 0$, will be called a "generalized solution" of Eq. (4).

3. Condition on a line of discontinuity of a solution. Suppose that, within a region where we seek a solution, there is a line $x \equiv x(t)$, on which the generalized solution has a first-order discontinuity. Suppose that, on approaching from the left or right we get on this line, respectively,

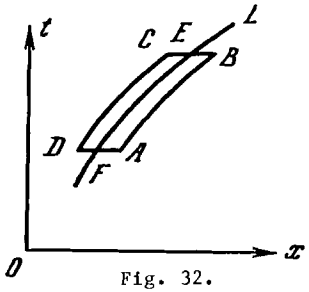


Fig. 32.

$$u(x, t) = u_{left}(x, t),$$

$$u(x, t) = u_{right}(x, t).$$

It turns out that the values $u_{left}(x, t)$ and $u_{right}(x, t)$ and the speed, $\dot{x} = dx/dt$, with which a point of discontinuity moves cannot be arbitrary: these various quantities are interconnected by certain relations.

Suppose L is the line of discontinuity (Fig. 32). The integral $\int_{ABCDA}\left(u\,dx - \frac{1}{2}u^2 dt\right)$ on the contour ABCDA, as on any other contour, vanishes. When the segments BC and DA shrink to the points E and F,

respectively, the integrals along these segments vanish and we get the equation

$$\int_{L'} \left([u]dx - \left[\frac{u^2}{2}\right]dt\right) = 0,$$

or

$$\int_{L'} \left([u]\,\frac{dx}{dt} - \left[\frac{u^2}{2}\right]\right) dt = 0,$$

where $[z] = z_{right} - z_{left}$ is the step-jump in the value of z along the line of discontinuity, and L' is an arbitrary segment of this line.

In view of the arbitrariness of segment L' we conclude that, at each point of line L, the integrands in the above equations must vanish:

$$[u]\,\frac{dx}{dx} - \left[\frac{u^2}{2}\right] = 0.$$

Therefore

$$\frac{dx}{dt} = \left[\frac{u^2}{2}\right] \div [u] = \frac{u^2_{right} - u^2_{left}}{2\left(u_{right} - u_{left}\right)} = \frac{u_{left} + u_{right}}{2}$$

so that

$$\frac{dx}{dt} = \frac{u_{left} + u_{right}}{2}. \tag{6}$$

If we had written the equation $u_t + uu_x = 0$ in another divergence form, e.g.

$$\frac{\partial}{\partial t}\left(\frac{u^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{u^3}{3}\right) = 0, \tag{7}$$

we would have arrived, by a similar route, to another integral relation, in this case to

$$\oint_{\Gamma} \left(\frac{u^2}{2}\,dx - \frac{u^3}{3}\,dt\right) = 0, \tag{8}$$

and to another condition on the line of discontinuity:

$$\frac{dx}{dt} = \frac{2}{3}\,\frac{u^2_{left} + u_{left}u_{right} + u^2_{right}}{u_{left} + u_{right}}. \tag{9}$$

The slope (9) of the discontinuity line (or the speed of the shock wave) does not coincide with the slope (6), corresponding to the first divergence form (4). Thus it is clear that the nature of the generalized solution depends on precisely what conservation law underlies the differential equation (1). In the problems of mathematical physics the conservation laws have a perfectly well defined physical meaning.

For smooth functions u all five expressions

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

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0,$$

$$\frac{\partial}{\partial t}\left(\frac{u^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{u^3}{3}\right) = 0,$$

$$\oint_\Gamma \left(u\,dx - \frac{u^2}{2}\,dt\right) = 0,$$

$$\oint_\Gamma \left(\frac{u^2}{2}\,dx - \frac{u^3}{3}\,dt\right) = 0$$

are equivalent. Below, in discussing Cauchy problem (1), we will be assuming integral conservation law (5), and discontinuity condition (6) which flows from it.

4. Decay of an arbitrary discontinuity. Suppose we are given the discontinuous initial conditions

$$u = \begin{cases} 2 \text{ for } x < 0 \\ 1 \text{ for } x > 0. \end{cases}$$

The solution constructed from these initial conditions is shown in Fig. 33.

The slope of the line of discontinuity $(dx/dt) = (2 + 1)/2 = 3/2$ is the arithmetic average of the slopes of the characteristics on either side of it.

We now assign initial conditions with a different discontinuity:

$$u = \begin{cases} 1 \text{ for } x < 0, \\ 2 \text{ for } x > 0. \end{cases}$$

From Fig. 34 one sees that it is now possible to construct solutions in two ways. The first gives us a continuous solution, while the second

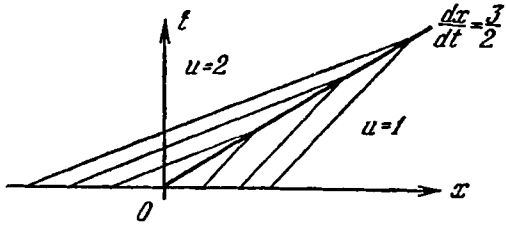


Fig. 33.

yields a solution discontinuous for $t > 0$. Here it is necessary to give preference to the continuous solution. In favor of this conclusion one may argue as follows. If the initial conditions are slightly changed, so that we are given

$$u = \begin{cases} 1 & \text{for } x \leq 0, \\ 2 & \text{for } x \geq \varepsilon, \\ 1 + x/\varepsilon & \text{for } 0 \leq x \leq \varepsilon, \end{cases}$$

then the solution u, shown in Fig. 35, is determined uniquely. As ε tends to zero this solution goes over to the continuous solution drawn in Fig. 34a. The impossibility of the solution depicted in 34b, because of

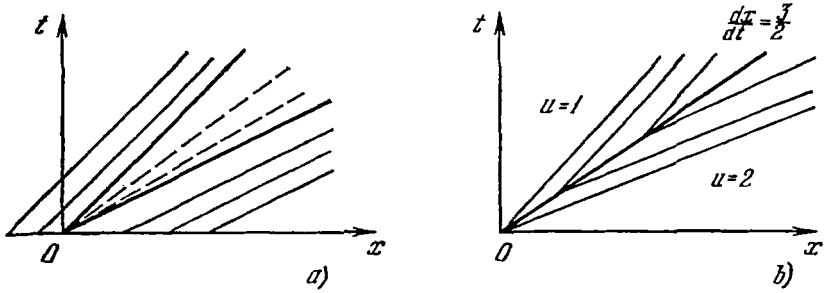


Fig. 34.

its instability with respect to perturbations in initial conditions, is analogous to the impossibility of rarification shock waves in the mathematical description of the flow of ideal gases.

5. Other definition of the generalized solution. One may formulate the generalized-solution concept through consideration of the auxiliary problem

$$\left.\begin{aligned} \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} &= \mu \frac{\partial^2 u}{\partial x^2}, \\ u(x, 0) &= \psi(x). \end{aligned}\right\} \qquad (10)$$

Here the differential equation is no longer hyperbolic, but of parabolic type. It's solutions preserve smoothness if $\psi(x)$ is a smooth function:

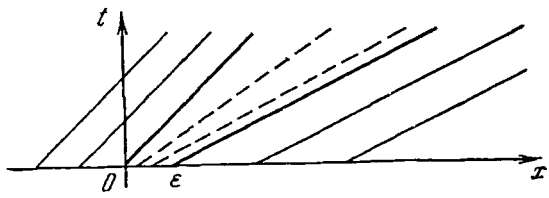


Fig. 35.

and if u(x, 0) is discontinuous, then the discontinuity is smoothed. The parameter $\mu > 0$ plays the same role here as viscosity in gas dynamics. As $\mu \to 0$ the solution of problem (10) tends to a limit which we can take to be the generalized solution of problem (1). One can show that, for problem

(1), this latter definition of the generalized solution is equivalent to the definition based on conservation law (5).

§30. The construction of difference schemes

Let us proceed, now, to discuss the construction of difference schemes for the problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = 0, \\ &u(x,\,0) = \psi(x). \end{aligned}\right\} \tag{1}$$

We will assume, for the sake of definiteness, that $\psi(x) > 0$. Then $u(x, t) > 0$. It may seem natural, at first glance, to consider the use of the difference schemes

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} + u_m^p\,\frac{u_m^p - u_{m-1}^p}{h} = 0, \qquad \begin{matrix} p = 0,\ 1,\ \ldots, \\ m = 0,\ \pm 1,\ \ldots, \end{matrix} \\ &u_m^p = \psi(x_m). \end{aligned}\right\} \tag{2}$$

Freezing the coefficient u_m^p at the point $m = m_0$ we see that, for the resulting equation with constant coefficients, in the transition to the level $t = (p + 1)\tau$ the maximum principle is fulfilled if the step-size, $\tau = \tau_p$, is chosen so as to satisfy the condition

$$r_p = \frac{\tau_p}{h} \le \frac{1}{\max\limits_m |u_m^p|}.$$

Thus we may expect stability. If the solution of problem (1) is smooth, then there is little reason to doubt that problem (2) approximates problem (1). And, in fact, in this case experimental computations of solutions known, beforehand, to be smooth confirm convergence.

However, if problem (1) has a discontinuous solution, then convergence to the generalized solution determined, let us say, by the integral conservation law

$$\oint_\Gamma \left(u\,dx - \frac{u^2}{2}\,dt\right) = 0, \tag{3}$$

cannot be expected on any reasonable grounds. Indeed, no information has been built into the proposed difference scheme (2) as to just what sort of conservation law ((8) §29, or (3), or perhaps some other) we have taken as a basis for the generalized solution.

Therefore in constructing a difference scheme one must use either the integral conservation law corresponding to the desired generalized solution, say law (3), or the equation with artificial viscosity (10) §29;

$$u_t + uu_x = \mu u_{xx}, \tag{4}$$

which accomplishes, as $\mu \to 0$, the selection of the generalized solution which interests us.

1. Schemes with artificial viscosity. Let us point out at once that the difference scheme with artificial viscosity

$$\frac{u_m^{p+1} - u_m^p}{\tau} + u_m^p \frac{u_m^p - u_{m-1}^p}{h} = \mu \frac{u_{m-1}^p - 2u_m^p + u_{m+1}^p}{h^2},$$

$$u_m^0 = \psi(x_m)$$

has a solution, $u^{(h)} = \{u_m^p\}$ which, for sufficiently small $\tau = \tau(h, \mu)$, converges uniformly as $h \to 0$ to the desired generalized solution outside any prescribed neighborhood, however small, of the line of discontinuity of this solution. It must be assumed, here, that $\mu = \mu(h)$ tends to zero sufficiently slowly. Various schemes using artificial viscosity are applied successfully in gas dynamics calculations. Their weakness is the smearing of discontinuities.

We turn now to consider, in detail, the construction of difference schemes based on conservation law (3).

It is possible to distinguish two approaches. In the first one uses, not only the conservation law (3) itself, but also the discontinuity condition

$$\frac{dx}{dt} = \frac{u_{\text{left}} + u_{\text{right}}}{2}. \tag{5}$$

which this law implies. In the second the discontinuities are not singled out, and the computation is governed by the same equations at all points of the computational net.

2. The method of characteristics. The idea of singling out the discontinuity in computing the generalized solution is embodied in its clearest form in the method of characteristics, which may be considered as one of the variants of the finite difference method. The development of discontinuities in the course of the computation, i.e. with increasing time t, is taken into account via special equations, making use of relation (5) on the discontinuity. Away from the discontinuity statements of the differential equations in all the forms we have encountered are equivalent. Therefore, in constructing computational formulas at points where the solution is smooth, we may take as our starting point the conservation law in differential form, i.e. the differential equation

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0.$$

In its main outlines one of the variants of the method of characteristics, a variant applicable to our example, may be sketched as follows. Mark out on the x axis the points $x_m = mh$. We will assume, for the sake of definiteness, that the initial condition, $u(x, 0) = \psi(x)$, is given by the smooth function $\psi(x)$. From each point $(x_m, 0)$ draw a characteristic of the equation $u_t + uu_x = 0$.

Suppose, in order not to complicate our presentation, that for the given function, $\psi(x)$, one can choose such a small τ that during any time interval of length τ a characteristic can intersect no more than one of its neighboring characteristics. Select such a τ and draw the lines $t = t_p = p\tau$. Locate the points of intersection of the characteristic emerging from the point $(x_m, 0)$ with the line $t = \tau$, and translate to these points, along the characteristics, the values of the solution $u(x_m, 0) = \psi(x_m)$.

If, in the interval $0 \leq t \leq \tau$, no two characteristics have intersected we take the following step; we extend the characteristics up to their intersections with the line $t = 2\tau$ and translate the solution values along the characteristics, to the points of intersection. If, during the time $\tau < t < 2\tau$, there is again no intersections of the characteristics we take the following step, etc., until on some segment $t_p < t < t_{p+1}$ two characteristics, emerging for example from the points $(x_m, 0)$ and $(x_{m+1}, 0)$, have intersected (Fig. 36). Then the midpoint of the segment Q^{p+1}_{m+1}, Q^{p+1}_n will be taken as the point from which a discontinuity originates. The points Q^{p+1}_m and Q^{p+1}_{m+1} will be replaced by the single point Q, to which we ascribe two solution values, u_{left} and u_{right}, taking for these quantities the values

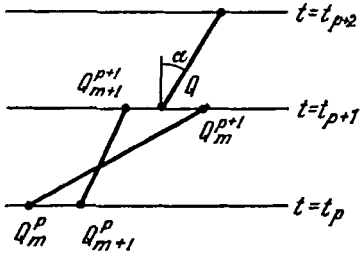


Fig. 36.

$$u_{left} = u(Q^p_m) \quad \text{and} \quad u_{right} = u(Q^p_{m+1}).$$

From the point Q we draw the discontinuity line up to its intersection with the line $t = t_{p+2}$. The slope of the discontinuity line is defined by the discontinuity condition

$$\tan \alpha = \frac{u_{left} + u_{right}}{2}.$$

From the point of intersection of the discontinuity line with $t = t_{p+2}$ we draw characteristics back to their intersections with the line $t = t_{p+1}$, giving them the slopes u_{left} and u_{right}, the values of u assigned at the previous level. At the points of intersection of these characteristics with the line $t = t_{p+1}$ we find values of u by interpolation in x, and take these as the left- and right-hand values of the solution at the point of discontinuity lying on the line t_{p+1}. In this way we can now define a

new slope of the discontinuity line as the arithmetic average of the newly-found left- and right-hand values, and continue this line still another timestep τ.

The advantage of the method of characteristics is that it allows one to track the discontinuity and to compute it accurately. But in the computational process more and more new discontinuities develop and, in fact, unimportant discontinuities may intersect, so that in time the picture becomes more and more complicated. The computational logic becomes more complicated, the demands on computer storage and computing time increase.

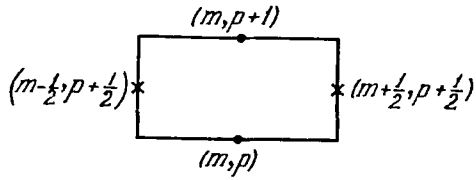


Fig. 37.

This constitutes the disadvantage of the method of characteristics, in which the discontinuities are singled out and treated in a special manner.

3. Divergence difference schemes. Difference schemes which do not use artificial viscosity, and do not use discontinuity conditions must, as noted earlier, rely on integral conservation laws.

On the xt plane let us draw a net of lines $t = p\tau$, $x = (m + 1/2)h$, $m = 0, \pm 1, \ldots$ We next mark off the midpoints of the sides of the thus-formed net rectangles (Fig. 37: coordinate axes not shown) and add these midpoints to the net D_h.

The function $[u]_h$ which we would like to calculate we take to be the net function defined, at each point of D_h, by averaging the solution $u(x, t)$ along that side of the net rectangle to which the point belongs:

$$[u]_h\Big|_{\substack{x=x_m,\\ t=t_p}} \equiv \bar{u}_m^p = \frac{1}{h}\int_{x_{m-1/2}}^{x_{m+1/2}} u(x, t_p)\,dx,$$

$$[u]_h\Big|_{\substack{x=x_{m+1/2},\\ t=t_{p+1/2}}} \equiv \bar{U}_{m+1/2}^{p+1/2} = \frac{1}{\tau}\int_{t_p}^{t_{p+1}} u(x_{m+1/2}, t)\,dt.$$

The approximate solution of this problem is defined on the same net D_h. Values of $u^{(h)}$ at the points, (x_m, t_p), lying on the horizontal sides of the rectangle will be designated u_m^p, and those at points $(x_{m+1/2}, t_{p+1/2})$ of the vertical sides by $U_{m+1/2}^{p+1/2}$.

The values u_m^p may be considered to pertain to the whole rectangle-side, $t = t_p$, $x_m < x < x_{m+1}$, to which the point (x_m, t_p) belongs. Analogously, we will consider that $U_{m+1/2}^{p+1/2}$ is defined on the whole vertical interval

$$x = x_{m+1/2}, \qquad t_p < t < t_{p+1}.$$

Thus $u^{(h)}$ will be a function defined on the lines $x = mh$, $t = p\tau$. The connection between the quantities u_m^p and $U_{m+1/2}^{p+1/2}$, $p = 0, 1, \ldots, m = 0, \pm 1, \ldots$, will be established by a process starting from the integral conservation law

$$\oint_\Gamma \left(u\,dx - \frac{u^2}{2}\,dt\right) = 0.$$

Let us take, as contour Γ, an elementary net-rectangle, setting

$$-\oint_\Gamma \left(u^{(h)}\,dx - \frac{\left(u^{(h)}\right)^2}{2}\,dt\right) = 0, \tag{6}$$

or, in expanded form

$$h\left[u_m^{p+1} - u_m^p\right] + \frac{\tau}{2}\left[\left(U_{m+1/2}^{p+1/2}\right)^2 - \left(U_{m-1/2}^{p+1/2}\right)^2\right] = 0. \tag{7}$$

If we can now set down a rule by which one can compute the quantities $U_{m+1/2}^{p+1/2}$, $m = 0, \pm 1, \ldots$, from already-known values of u_m^p, $m = 0, \pm 1, \ldots$, then Eq. (7) will allow us to compute the quantities u_m^{p+1}, $m = 0, \pm 1, \ldots$, i.e. to move ahead one timestep. But, by whatever specific method we compute $U_{m+1/2}^{p+1/2}$, a difference scheme of form (7) has the *divergence property*, which consists in the following. Draw, in the half-plane $t \geq 0$, any closed contour, which does not intersect itself, and consists entirely of the sides of net-rectangles (Fig. 38). This contour g_h encloses some region, G_h, made up of net rectangles.

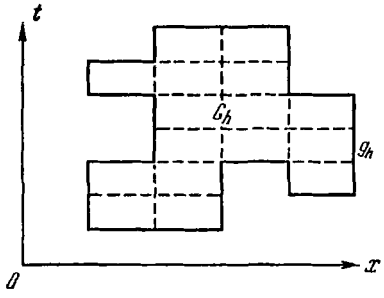


Fig. 38.

We now add, term-by-term, all Eqs. (7) pertaining to the rectangles constituting region G_h. Equations (6) and (7) differ only in notation. Therefore we may, as well, sum Eqs. (7), and we then get

$$-\oint_{g_h} \left(u^{(h)}\,dx - \frac{\left(u^{(h)}\right)^2}{2}\,dt\right) = 0. \tag{8}$$

Integrals in (6), over rectangle-sides which do not lie on the boundary, g_h, of region G_h, will cancel after the summation. In fact each such side belongs to two neighboring rectangles so that integration of $u^{(h)}$ over each is encountered twice, with the two integrations carried out in opposite directions (Fig. 39).

Schemes based on difference equations which, when summed over points of the net-region G_h, involve only algebraic sums of unknowns, or functions

of unknowns, on the region-boundary are called "divergence schemes" or "conservative schemes". Such schemes are analogous to differential equations in divergence form

$$\operatorname{div} \Phi = \frac{\partial \Phi_1}{\partial t} + \frac{\partial \Phi_2}{\partial x} = 0,$$

which, when integrated term by term over the two-dimensional region D yield, on the left-hand side, the contour integral (3) §29. Scheme (3) is not in divergence form, scheme (7) is.

Note the following. Suppose the net function, $u^{(h)}$, satisfying Eq. (7), converges uniformly as $h \to 0$ to some piecewise-continuous function $u(x, t)$ in each closed region not containing a discontinuity line, and let $u^{(h)}$ be uniformly bounded in h. Then $u(x, t)$ satisfies the integral conservation law

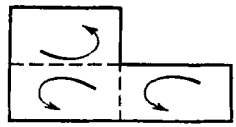


Fig. 39.

$$\oint_g \left(u\,dx - \frac{u^2}{2}\,dt\right) = 0,$$

where g is an arbitrary piecewise-smooth contour.

This follows immediately from the fact that one can approximate contour g by a contour g_h, together with Eq. (8) and the assumption of convergence* $u_h \to u$.

If scheme (7) is to take on meaning one must indicate a method for computing $U^{p+1/2}_{m+1/2}$ from known u^p_n's. In the scheme of S. K. Godunov, which we will use to illustrate the concept of a divergence scheme, one computes $U^{p+1/2}_{m+1/2}$ via the following "discontinuity decay" problem. Suppose that at $t = 0$ the solution $u(x, 0)$ is given by the conditions

$$u(x, 0) = \begin{cases} u_{\text{left}} & \text{for } x < 0, \\ u_{\text{right}} & \text{for } x > 0, \end{cases}$$

where $u_{\text{left}} = \text{const}$ and $u_{\text{right}} = \text{const}$. It is then possible to construct the corresponding generalized solution. How this can be done we have seen in §29, first for the example $u_{\text{left}} = 1$, $u_{\text{right}} = 2$, and then for $u_{\text{left}} = 2$,

*The function $u = u(x, t)$ is defined almost everywhere, while $u^{(h)} = u^{(h)}(x, t)$ is defined only on a network of lines. One can circumvent this formal inconsistency, first, by postulating that, as h decreases, each new net is gotten by subdividing the last net, and, second, by treating convergence at the points of one of the possible nets, constructed for any one possible fixed h.

$u_{right} = 1$. An important step is the determination of the value $U = u(0,t)$ of the solution $u(x, t)$ for $x = 0$.

The reader, having constructed sketches like Figs. 33 and 34, representing the solution $u(x, t)$, will easily verify that on the line $x = 0$ the solution will take on the values u_{left}, u_{right} or 0, depending on the initial data, and can easily determine, for any specific pair of numbers u_{left} and u_{right}, precisely which of these values it will have. For example if $u_{left} > 0$, $u_{right} > 0$, then $u(0, t) \equiv u_{left}$, and if $u_{left} < 0$ and $u_{right} < 0$ then $u(0, t) = u_{right}$.

The quantity $U^{p+1/2}_{m+1/2}$ $(= U)$ in scheme (7) will be determined via the analysis of a discontinuity decay problem at the boundary, $x = x_{m+1/2}$, between two segments where we are given the constant values u^p_m $(= u_{left})$ and u^p_{m+1} $(= u_{right})$.

If, for example, $u^p_m > 0$, $m = 0, \pm 1, \ldots$, then

$$U^{p+1/2}_{m+1/2} = u_{left} = u^p_m, \qquad m = 0, \pm 1, \ldots,$$

and scheme (7) takes the form

$$\frac{u^{p+1}_m - u^p_m}{\tau} + \frac{1}{h}\left[\frac{\left(u^p_m\right)^2}{2} - \frac{\left(u^p_{m-1}\right)^2}{2}\right] = 0,$$

$$u^0_m = \frac{1}{h}\int_{x_{m-1/2}}^{x_{m+1/2}} \psi(x)dx$$

or

$$\frac{u^{p+1}_m - u^p_m}{\tau} + \left(\frac{u^p_{m-1} + u^p_m}{2}\right)\frac{u^p_m - u^p_{m-1}}{h} = 0.$$

One can easily see that, for

$$r = \frac{\tau}{h} \le \frac{1}{\max\limits_m \left|u^p_m\right|}$$

the maximum principle holds

$$\max_m \left|u^{p+1}_m\right| \le \max_m \left|u^p_m\right| \le \ldots \le \max_m \left|u^0_m\right| \le \max_x |\psi(x)|.$$

Clearly then, if $\tau = h/\max|\psi(x)|$, we have reason to hope that the above difference scheme will be stable for some reasonable choice of norms. We will not, however, specifically point out such norms: numerical experiments confirm that, as the net is refined, the solution $u^{(h)}$ of problem

(7), with piecewise-monotonic and piecewise-smooth initial values $\psi(x)$, converges to some function, $u(x, t)$, with a finite number of discontinuities; and, further, outside any neighborhood of these discontinuities, convergence is uniform.

Scheme (7), with $U_{m+1/2}^{p+1/2}$ computed via decay of discontinuities, is not, of course, the only divergence scheme for problem (1). There is, for example, a still simpler scheme based on the predictor-corrector idea. This idea was formulated in 3§22. For simplicity we limit our discussion to the case $\psi(x) > 0$.

First we will determine auxiliary quantities $\bar{u}$ from the non-divergence implicit difference scheme

$$\frac{\bar{u}_m^{p+1/2} - u_m^p}{\tau/2} + u_m^p \frac{\bar{u}_m^{p+1/2} - \bar{u}_{m-1}^{p+1/2}}{h} = 0.$$

The value of the coefficient of u_x in the equation $u_t + uu_x = 0$ has been replaced here by u_m^p, and not by $\bar{u}_m^{p+1/2}$, so that the resulting scheme should be linear with respect to the quantities to be computed.

Next we let

$$U_{m+1/2}^{p+1/2} = \frac{1}{2}\left(\bar{u}_m^{p+1/2} + \bar{u}_m^{p+1/2}\right) \tag{9}$$

and use scheme (7), (9). The divergence scheme so derived has second-order approximation on a smooth solution.

A heuristic analysis using the Von Neumann spectral criterion, after linearization and the freezing of coefficients, suggests stability for arbitrary $r = \tau/h$. Let us now carry out this analysis.

As a result of linearization, and of freezing coefficients, we get a scheme of the form

$$\left.\begin{aligned} &\frac{\bar{u}_m^{p+1/2} - u_m^p}{\tau/2} + a\frac{\bar{u}_m^{p+1/2} - \bar{u}_{m-1}^{p+1/2}}{h} = 0, \\ &\frac{u_m^{p+1} - u_m^p}{2} + \frac{a}{h}\left[\frac{\bar{u}_{m+1}^{p+1/2} + \bar{u}_m^{p+1/2}}{2} - \frac{\bar{u}_m^{p+1/2} + \bar{u}_{m-1}^{p+1/2}}{2}\right] = 0. \end{aligned}\right\}$$

Given the initial conditions

$$u_m^p = e^{i\alpha m}$$

we get

$$\bar{u}_m^{p+1/2} = \mu e^{i\alpha m},$$

where

$$\mu = \frac{1}{1 + a\frac{r}{2} - a\frac{r}{2}e^{-i\alpha}}.$$

Further

$$u_m^{p+1} = \lambda e^{i\alpha m},$$

where

$$\frac{\lambda - 1}{\tau} + \frac{a\mu}{h}\left(\frac{e^{i\alpha} - e^{-i\alpha}}{2}\right) = 0,$$

$$\lambda(\alpha) = \frac{2 + ar - are^{+i\alpha}}{2 + ar - are^{-i\alpha}}, \qquad |\lambda(\alpha)| = 1.$$

Part 4
PROBLEMS WITH TWO SPACE VARIABLE

Chapter 10
The Concept of Difference Schemes with Splitting

Difference schemes with splitting belong among our important tools for computing solutions of the multidimensional time-dependent problems of mathematical physics.

§31. Construction of splitting schemes

At a descriptive level the idea of the construction of splitting schemes may be presented as follows.

Consider a differential problem of the form

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = Au, \qquad 0 < t < T, \\ &u\big|_{t=0} \text{ given}, \end{aligned}\right\} \tag{1}$$

where A is some operator in the space variables such as, for example,

$$Au \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

The value $u(x, y, t_{p+1})$ can be expressed in terms of already known values $u(x, y, t_p)$, $t_p = p\tau$, by means of the relation

$$u(x, y, t_p + \tau) = (x, y, t_p) + \tau \frac{\partial u}{\partial t} + O(\tau^2) =$$

$$= (x, y, t_p) + \tau Au(x, y, t_p) + O(\tau^2) = (E + \tau A)u(x, y, t_p) + O(\tau^2).$$

(where E is the identity operator, $Ev = v$).

Suppose that the right hand side of Eq. (1) has the form

$$Au \equiv A_1 u + A_2 u.$$

We will, then, split Eq. (1)

$$\frac{\partial u}{\partial t} = A_1 u + A_2 u$$

into the following two equations:

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} = A_1 v, \qquad t_p \le t \le t_{p+1}, \\ &v(x, y, t_p) = u(x, y, t_p), \end{aligned}\right\} \tag{2}$$

$$\left.\begin{aligned} &\frac{\partial w}{\partial t} = A_2 w, \qquad t_p \le t \le t_{p+1}, \\ &w(x, y, t_p) = v(x, y, t_{p+1}). \end{aligned}\right\} \tag{3}$$

Note that

$$w(x, y, t_{p+1}) = u(x, y, t_{p+1}) + O(\tau^2). \tag{4}$$

In fact

$$v(x, y, t_{p+1}) = (E + \tau A_1)v(x, y, t_p) + O(\tau^2) =$$

$$= (E + \tau A_1)u(x, y, t_p) + O(\tau^2).$$

Further, taking account of the last equation, we have

$$w(x, y, t_{p+1}) = (E + \tau A_2)w(x, y, t_p) + O(\tau^2) =$$

$$= (E + \tau A_2)v(x, y, t_{p+1}) + O(\tau^2) =$$

$$= (E + \tau A_2)(E + \tau A_1)u(x, y, t_p) + O(\tau^2) =$$

$$= [E + \tau(A_1 + A_2)]u(x, y, t_p) + \tau^2 A_1 A_2 u(x, y, t_p) + O(\tau^2) =$$

$$= (E + \tau A)u(x, y, t_p) + O(\tau^2) = u(x, y, t_{p+1}) + O(\tau^2).$$

On the basis of Eq. (4) we can now, in each time interval $t_p \le t \le t_{p+1}$, solve Eqs. (2) and (3) sequentially in place of problem (1).

In actuality, to solve (2) and (3) we approximate these equations by difference equations of some sort. We then get a difference-splitting-scheme, $L_h u^{(h)} = f^{(h)}$, which allows us to compute u^{p+1}, in two stages, from the already-known u^p (in the first stage computing v^{p+1} from the given $v^p = u^p$ and, in the second, computing $u^{p+1} = w^{p+1}$ from $w^p = v^{p+1}$, using the v^{p+1} already computed in the first stage).

The above considerations are heuristic in character. After some sort of difference-splitting-scheme

$$L_h u^{(h)} = f^{(h)} \tag{5}$$

for computing the solution of problem (1) has been constructed we must still, somehow, verify approximation and stability.

In the case of the Cauchy problem for the two-dimensional heat equation

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \qquad 0 < t < T, \quad -\infty < x, y < \infty, \\ &u(x, y, 0) = \psi(x, y) \end{aligned}\right\} \tag{6}$$

we may take, for example, as the system (2), (3)

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}, \qquad v(x, y, t_p) = u(x, y, t_p), \\ &\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial y^2}, \qquad w(x, y, t_p) = v(x, y, t_{p+1}). \end{aligned}\right\} \tag{7}$$

This splitting of the two dimensional equation of problem (6) into two one-dimensional equations (7) can be interpreted as an approximate replacement of the process by which heat spreads in the xy plane, in time $t_p \leq t \leq t_{p+1}$, by two processes. In the first of these, described by the first of Eqs. (7), one introduces (conceptually) non-heat-conducting

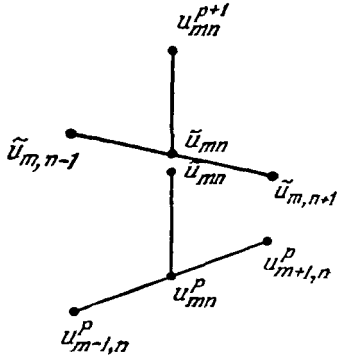


Fig. 40.

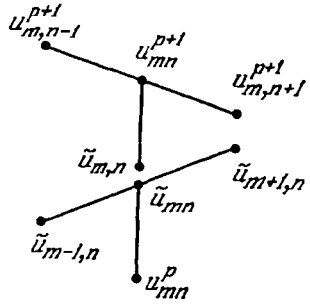


Fig. 41.

partitions preventing the spread of heat in the y-direction. Then, during time-interval τ, instead of those partitions one introduces others preventing the spread of heat in the x direction. This spread of heat, again in time-interval τ, is described by the second equation.

We now choose the net $(x_m, y_n, t_p) = (mh, nh, p\tau)$.

A difference-splitting-scheme based on (7) can be constructed in many ways. We consider two possibilities, i.e.,

$$\left.\begin{aligned} &\frac{\tilde{u}_{mn} - u^{p}_{mn}}{\tau} = \Lambda_{xx} u^{p}_{mn}, \\ &\frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} = \Lambda_{yy} \tilde{u}_{mn}, \\ &u^{0}_{mn} = \psi(x_m, y_n) \end{aligned}\right\} \qquad (8)$$

and

$$\left.\begin{aligned} &\frac{\tilde{u}_{mn} - u^{p}_{mn}}{\tau} = \Lambda_{xx} \tilde{u}_{mn}, \\ &\frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} = \Lambda_{yy} u^{p+1}_{mn}, \\ &u^{0}_{mn} = \psi(x_m, y_n). \end{aligned}\right\} \qquad (9)$$

In both splitting schemes we set

$$\tilde{u}_{mn} \equiv v^{p+1}_{mn} \equiv w^{p}_{mn}, \qquad u^{p+1}_{mn} \equiv w^{p+1}_{mn}.$$

Let us recall that, in notation defined earlier

$$\Lambda_{xx} u_{mn} = \frac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2},$$

$$\Lambda_{yy} u_{mn} = \frac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2}.$$

Scheme (8) is represented schematically in Fig. 40, scheme (9) in Fig. 41.

The splitting of problem (6) itself is also not unique. One can, for example, write the problem in the form

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \\ &u(x, y, 0) = \psi(x, y), \end{aligned}\right\} \qquad (6')$$

constructing, correspondingly, in the interval $t_p \le t \le t_{p+1}$, the following two systems:

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} = \frac{1}{2}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right), \qquad t_p \le t \le t_{p+1}, \\ &v(x, y, t_p) = u(x, y, t_p) \end{aligned}\right\} \qquad (10)$$

and

$$\left.\begin{aligned} &\frac{\partial w}{\partial t} = \frac{1}{2}\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right), \qquad t_p \le t \le t_{p+1}, \\ &w(x, y, t_p) = v(x, y, t_{p+1}). \end{aligned}\right\} \tag{11}$$

Such a splitting is not a physically-based splitting like scheme (7). We now choose a difference scheme as follows (Fig. 42):

$$\left.\begin{aligned} &\frac{\tilde{u}_{mn} - u^p_{mn}}{\tau} = \frac{1}{2}\left(\Lambda_{xx} u^p_{mn} + \Lambda_{yy}\tilde{u}_{mn}\right), \\ &\frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} = \frac{1}{2}\left(\Lambda_{xx} u^{p+1}_{mn} + \Lambda_{yy}\tilde{u}_{mn}\right), \\ &u^0_{mn} = \psi(x_m, y_n). \end{aligned}\right\} \tag{12}$$

To calculate u^{p+1} by *alternating-direction scheme (12)* one must, first, for each fixed m, solve the implicit equation for $\tilde{u}_{mn}$, in which m occurs as a parameter. Then to compute u^{p+1}_{mn} it is necessary to solve the second equation (12), implicit with respect to u^{p+1}_{mn}, in which n occurs as a parameter.

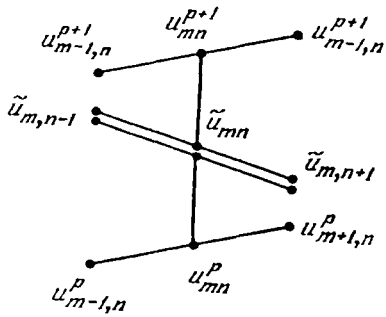


Fig. 42.

Scheme (8) can be written in form (5) if one sets

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} - \Lambda_{yy}\tilde{u}_{mn}, \\ u^0_{mn}. \end{cases}$$

where $\tilde{u}_{mn} = u^p_{mn} + \tau\Lambda_{xx} u^p_{mn}$ is determined from the first of Eqs. (8). Then

$$f^{(h)} \equiv \begin{cases} 0 & m, n = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau]-1 \\ \psi(x_m, y_n), & m, n = 0, \pm 1, \ldots \end{cases}$$

We propose to the reader than he write schemes (9) and (12) in form (5).

The reader may verify that the Von Neumann spectral stability criterion (i.e. the requirement that solutions of the form $u^p_{mn} = \lambda^p \exp[i(\alpha m + \beta n)]$ should remain bounded) is satisfied for scheme (8) with any $r = \tau/h^2 \le \frac{1}{2}$, and for schemes (9) and (12) is satisfied for any r. We will not pause here to study the stability conditions, or to prove approximation, for schemes (8), (9) and (12).

PROBLEMS

1. Determine for which $r = \tau/h^2$ the Von Neumann spectral criterion is satisfied for difference-splitting-schemes (8), (9) and (12), introduced in the text above.

2. Verify that scheme (8) approximates problem (6) on a sufficiently smooth bounded solution $u(x, y, t)$.

3. Repeat problem 2, but for difference-splitting-schemes (9) and (12).

§32. Economical difference schemes

We will now consider and study examples of difference-splitting-schemes for the heat-conduction problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, && 0 \le x, y \le 1, \quad 0 \le t \le T, \\ &u(x, y, 0) = \psi(x, y), && 0 \le x, y \le 1, \\ &u(x, y, t)\big|_{\Gamma} = 0 \end{aligned}\right\} \tag{1}$$

in the rectangular region $0 \le x, y \le 1$ with boundary Γ, using the usual net $(x_m, t_n, t_p) = (mh, nh, p\tau)$, $m, n = 1, 2, \ldots, N$ and $h = 1/N$.

The difference-splitting-scheme which we now introduce has, in certain respects basic advantages over the simplest explicit

$$\left.\begin{aligned} &\frac{u^{p+1}_{mn} - u^{p}_{mn}}{\tau} = \Lambda_{xx} u^{p}_{mn} + \Lambda_{yy} u^{p}_{mn}, \\ &u^{0}_{mn} = \psi(x_m, y_n), \\ &u^{p}_{mn}\Big|_{\Gamma} = 0 \end{aligned}\right\} \tag{2}$$

and the simplest implicit

$$\left.\begin{aligned} &\frac{u^{p+1}_{mn} - u^{p}_{mn}}{\tau} = \Lambda_{xx} u^{p+1}_{mn} + \Lambda_{yy} u^{p+1}_{mn}, \\ &u^{0}_{mn} = \psi(x_m, y_n), \\ &u^{p}_{mn}\Big|_{\Gamma} = 0 \end{aligned}\right\} \tag{3}$$

difference schemes.

Computation by explicit scheme (2) is very simple. To proceed from the already known u^p to the unknown $u^{p+1} = \{u_{mn}^{p+1}\}$ we must execute arithmetic operations whose number is proportional to the number, $(N-1)^2$, of unknowns $\{u_{mn}^{p+1}\}$. In this sense the explicit scheme is the best possible. Difference schemes in which the number of arithmetic operations involved in the step from u^p to $u^{p+1} = \{u_{mn}^{p+1}\}$ is proportional to the number of unknowns are called "economical". On the other hand, although it is economical, the explicit scheme is stable only under the very stringent condition $\tau \leq h^2/4$ on the timestep τ. The above "simplest implicit difference scheme" (3), as we already know (see 3§27) is absolutely stable. But it is far from economical. To calculate the unknowns $\{u_{mn}^{p+1}\}$ one has to solve a complicated (non-separable) system of linear equations. As we know from the analysis of numerical methods, this requires that we perform numerical operations whose number is proportional, not to the first power of the number of unknowns as in economical schemes, but to the cube of the number of unknowns, if one uses some sort of elimination method.

* * * * * *

We note that a search is now in progress for more economical ways to solve general linear systems. Strassen has pointed out an algorithm requiring a number of operations proportional, not to the third, but to the $\log_2(7)$'th power of the number of unknowns.

* * *

The difference-splitting-scheme which we will now construct is economical and unconditionally stable, i.e. it unites the advantages of explicit scheme (2) and implicit scheme (3).

As regards the solution $u(x, y, t)$ of problem (1), we will assume that it has derivatives continuous right up to the boundary Γ, of all orders required in the course of our work. Note that, on the boundary Γ, all even-order space derivatives (up to all orders for which they exist and are continous) will vanish

$$u_{xx}\Big|_\Gamma = u_{xxxx}\Big|_\Gamma = u_{xxyy}\Big|_\Gamma = 0. \tag{4}$$

Thus, along the side $x = 0$ of the boundary Γ of the square $0 \leq x, y \leq 1$, the derivatives $\partial u/\partial t$ and $\partial^2 u/\partial y^2$ both vanish. Therefore, since $u_t = u_{xx} + u_{yy}$, also $u_{xx} = 0$. Differentiating the equation twice by y we get

$$\frac{\partial u_{yy}}{\partial t} = u_{xxyy} + u_{yyyy}.$$

But on the side $x = 0$ of boundary Γ we have

$$u_{yy} = 0, \qquad u_{yyyy} = 0, \qquad \frac{\partial u_{yy}}{\partial t} = 0,$$

and, therefore, it follows from the differential equation that also $u_{xxyy} = 0$.

Now let us proceed to construct a difference-splitting-scheme for problem (1). In parallel with problem (1), on the interval $t_p \leq t \leq t_{p+1}$ we pose the two problems

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}, \\ &v|_\Gamma = 0, \qquad v(x, y, t_p) = u(x, y, t_p), \end{aligned}\right\} \tag{5}$$

$$\left.\begin{aligned} &\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial y^2}, \\ &w|_\Gamma = 0, \qquad w(x, y, t_p) = v(x, y, t_{p+1}). \end{aligned}\right\} \tag{6}$$

The net function

$$u^{(h)} = \{u^p_{mn}\}, \qquad u^0_{mn} = \psi(x_m, y_p), \qquad u^p_{mn}\Big|_\Gamma = 0$$

will be determined, sequentially, from the equations

$$\left.\begin{aligned} &\frac{\tilde{u}_{mn} - u^p_{mn}}{\tau} = \Lambda_{xx}\tilde{u}_{mn}, \qquad m, n = 1, 2, \ldots, N-1 \\ &\tilde{u}_{mn}\Big|_\Gamma = 0; \end{aligned}\right\} \tag{7}$$

$$\left.\begin{aligned} &\frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} = \Lambda_{yy}u^{p+1}_{mn}, \qquad m, n = 1, 2, \ldots, N-1, \\ &u^{p+1}_{mn}\Big|_\Gamma = 0. \end{aligned}\right\} \tag{8}$$

Problem (7) is analogous to problem (5), while (8) is analogous to (6). Here

$$v^p_{mn} = u^p_{mn}, \qquad w^p_{mn} = v^{p+1}_{mn} = \tilde{u}_{mn}, \qquad w^{p+1}_{mn} = u^{p+1}_{mn}.$$

Using (7) and (8) one first computes the auxiliary function $\tilde{u}_{mn}$ from $u^p = \{u^p_{mn}\}$ and then, from (8), computes $u^{p+1} = \{u^{p+1}_{mn}\}$.

Note that difference scheme (7) for $\tilde{u}_{mn}$, for each fixed n, n = 1, ..., N-1, exactly coincides with the implicit difference scheme

$$\frac{v_{mn}^{p+1} - v_{mn}^{p}}{\tau} = \Lambda_{xx} v_{mn}^{p+1}$$

for the one-dimensional heat equation on the interval $0 \leq x \leq 1$, in which y enters only as a parameter.

Difference problem (7), for each fixed n, is solved by FEBS in the direction of the x axis. In precisely the same way difference scheme (8), for each fixed m, is solved by FEBS in the direction of the y axis. We note that, by virtue of the properties of the FEBS algorithm, the total number of arithmetic operations required for the computation $u^{p+1} = \{u_{mn}^{p+1}\}$ turns out to be proportional to the number of unknowns, i.e. difference scheme (7), (8) is economical.

So as to formulate the concepts of approximation and stability exactly we write difference scheme (7), (8) in the form we have taken as standard throughout this book,

$$L_h u^{(h)} = f^{(h)}. \tag{9}$$

For this purpose we set

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{mn}^{p+1} - \tilde{u}_{mn}}{\tau} - \Lambda_{yy} u_{mn}^{p+1}, & m, n = 1, 2, \ldots, N-1, \\ u_{mn}^{p+1}\Big|_{\Gamma}, \\ u_{mn}^{0}, \end{cases} \tag{10}$$

where $\tilde{u}_{mn}$ is the solution of the auxiliary problem

$$\left.\begin{aligned} \frac{\tilde{u}_{mn} - u_{mn}^{p}}{\tau} &= \Lambda_{xx} \tilde{u}_{mn}, \quad m, n = 1, 2, \ldots, N-1, \\ \tilde{u}_{mn}\Big|_{\Gamma} &= 0. \end{aligned}\right\} \tag{11}$$

In this case we must then take, as $f^{(h)}$,

$$f^{(h)} \equiv \begin{cases} 0, & m, n = 1, 2, \ldots, N-1, \\ 0, & (x_m, y_n) \text{ in } \Gamma, \\ \psi(x_m, y_n). \end{cases} \tag{12}$$

As a norm in U_h we take

$$||u^{(h)}||_{U_h} = \max_{m,n,p} \left|u^p_{mn}\right| .$$

Elements in F_h will have the form

$$g^{(h)} = \begin{cases} \phi^p_{mn}, \\ 0, \\ \psi_{mn} \end{cases}$$

and the norm in F_h will be defined by the equation

$$||g^{(h)}||_{F_h} = \max_{m,n,p} \left|\phi^p_{mn}\right| + \max_{m,n} |\psi_{mn}| .$$

First let us demonstrate the unconditional stability of difference scheme (9), defined by Eqs. (10) and (12), and approximation will be proven later. In view of the linearity of difference scheme (9), to prove stability one will have to show that the problem $L_h z^{(h)} = g^{(h)}$ has a solution for any $g^{(h)}$,

$$g^{(h)} = \left\{ \begin{array}{l} \phi^p_{mn} \\ 0 \\ \psi_{mn} \end{array} \right\} \text{ in } F_h,$$

and, moreover,

$$||z^{(h)}||_{U_h} \leq c||g^{(h)}||_{F_h},$$

where c does not depend on h.

Let us now write the problem $L_h z^{(h)}$ in expanded form:

$$\left.\begin{array}{l} \dfrac{z^{p+1}_{mn} - \tilde{z}_{mn}}{\tau} - \Lambda_{yy} z^{p+1}_{mn} = \phi^{p+1}_{mn}, \qquad m, n = 1, 2, \ldots, N-1, \\ z^{p+1}_{0n} = z^{p+1}_{Nn} = 0, \end{array}\right\} \tag{13}$$

where $\tilde{z}_{mn}$ is the solution of the auxiliary problem

$$\left.\begin{array}{l} \dfrac{\tilde{z}_{mn} - z^p_{mn}}{\tau} = \Lambda_{xx} \tilde{z}_{mn}, \qquad m, n = 1, 2, \ldots, N-1, \\ \tilde{z}_{m,0} = \tilde{z}_{m,N} = 0, \end{array}\right\} \tag{14}$$

while

$$z^{0}_{mn} = \psi_{mn}.$$

By the maximum principle, proven in §28 for the one-dimensional, implicit, difference scheme approximating the one-dimensional heat equation, it follows from (13) that

$$\max_{m,n} \left|z^{p+1}_{mn}\right| \leq \max_{m,n} \left|\tilde{z}_{mn}\right| + \tau \max_{m,n,p} \left|\phi^{p}_{mn}\right| .$$

But from (14)

$$\max_{m,n} \left|\tilde{z}_{mn}\right| \leq \max_{m,n} \left|z^{p}_{mn}\right| .$$

Therefore

$$\max_{m,n} \left|z^{p+1}_{mn}\right| \leq \max_{m,n} \left|z^{p}_{mn}\right| + \tau \max_{m,n,p} \left|\phi^{p}_{mn}\right| ,$$

so that

$$\max_{m,n} \left|z^{p+1}_{mn}\right| \leq \max_{m,n} \left|z^{p-1}_{mn}\right| + 2\tau \max \left|\phi^{p}_{mn}\right| \leq$$

. . . . . .

$$\leq \max_{m,n} |\psi_{mn}| + T \max_{m,n,p} \left|\phi^{p}_{mn}\right| \leq$$

$$\leq (1 + T)\Big(\max_{m,n} |\psi_{mn}| + \max_{m,n,p} \left|\phi^{p}_{mn}\right|\Big) = (1 + T)||g^{(h)}||_{F_h}.$$

The above inequality

$$\max_{m,n} \left|z^{p+1}_{mn}\right| \leq (1 + T)||g^{(h)}||F_h$$

is valid for any p, and therefore

$$||z^{(h)}||_{U_h} \leq (1 + T)||g^{(h)}||_{F_h}$$

for any arbitrary relation between step-sizes τ and h. This means that the splitting scheme under consideration is unconditionally stable.

Let us now proceed to verify approximation. It will be assumed, as usual, that problem (1) has a sufficiently smooth solution u(x, y, t). We

will compute the residual $\Lambda f^{(h)}$, $L_h[u]_h = f^{(h)} + \Lambda f^{(h)}$, which develops when $[u]_h$ is substituted into the left-hand side of Eq. (9), and show that $||\Lambda f^{(h)}||_{F_h} = O(\tau + h^2)$.

By the definition of L_h we have

$$L_h[u]_h = \begin{cases} \dfrac{u(x_m, y_n, t_{p+1}) - \tilde{u}_{mn}}{\tau} - \Lambda_{yy} u(x_m, y_n, t_{p+1}), \\ \qquad m, n = 1, \ldots, N-1, \\ 0 \text{ at points on } \Gamma, \\ u(x_m, y_n, 0), \qquad m, n = 1, \ldots, N-1, \end{cases} \tag{15}$$

where $\tilde{u}_{mn}$ is the solution of the auxiliary problem

$$\left.\begin{aligned} &\frac{\tilde{u}_{mn} - u(x_m, y_n, t_p)}{\tau} - \Lambda_{xx}\tilde{u}_{mn} = 0, \qquad m, n = 1, 2, \ldots, N-1, \\ &\tilde{u}_{mn}\big|_\Gamma = 0. \end{aligned}\right\} \tag{16}$$

The solution $\tilde{u}_{mn}$ of auxiliary problem (16), as we will show below, has the form

$$\left.\begin{aligned} &\tilde{u}_{mn} = u(x_m, y_n, t_p) + \tau\Lambda_{xx} u(x_m, y_n, t_p) + O(\tau^2), \\ &\qquad m, n = 1, 2, \ldots, N-1, \\ &\tilde{u}_{mn}\big|_\Gamma = u(x_m, y_n, t_p)\big|_\Gamma = 0. \end{aligned}\right\} \tag{17}$$

Inserting this expression for $\tilde{u}_{mn}$ into (15) we get

$$\frac{u(x_m, y_n, t_{p+1}) - \tilde{u}_{mn}}{\tau} - \Lambda_{yy} u(x_m, y_n, t_{p+1}) =$$

$$= \frac{u(x_m, y_n, t_{p+1}) - [u(x_m, y_n, t_p) + \tau\Lambda_{xx} u(x_m, y_n, t_p) + O(\tau^2)]}{\tau} -$$

$$- \Lambda_{yy} u(x_m, y_n, t_{p+1}) = \frac{u(x_m, y_n, t_{p+1}) - u(x_m, y_n, t_p)}{\tau} -$$

$$- \Lambda_{xx} u(x_m, y_n, t_p) - \Lambda_{yy} u(x_m, y_n, t_{p+1}) + O(\tau) =$$

$$= \left(\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2}\right)_{x_m, y_n, t_p} + O(\tau + h^2) = O(\tau + h^2).$$

Thus

$$L_h[u]_h \equiv \left\{ \begin{array}{ll} 0 & + O(\tau + h^2) \\ 0 & + 0 \\ \psi(x_m, y_n) & + 0 \end{array} \right\} = f^{(h)} + \Lambda f^{(h)},$$

where

$$\Lambda f^{(h)} = \left\{ \begin{array}{ll} O(\tau + h^2) \text{ at points on } (x_m, y_n, t_p), & \\ 0, & p = 0, 1, \ldots, [T/\tau]-1, \\ 0, & m, n = 1, 2, \ldots, N-1. \end{array} \right.$$

Therefore

$$||\Lambda f^{(h)}||_{F_h} = O(\tau + h^2).$$

It remains for us to prove the approximate representation (17) for the solution $\tilde{u}_{mn}$, of problem (16). First we bring out some heuristic considerations suggesting representation (17). It is clear that, for small τ, we may write the approximate equation

$$\tilde{u}_{mn} \approx u(x_m, y_n, t_p).$$

If, on this basis, in Eq. (16) we were to replace the expression $\Lambda_{xx}\tilde{u}_{mn}$ by the expression $\Lambda_{xx}u(x_m, y_n, t_p)$, we would get the equation

$$\frac{\tilde{u} - u}{\tau} - \Lambda_{xx}u = 0,$$

from which follows the equation $\tilde{u} = u + \tau\Lambda_{xx}u$, which differs from (17) only in the remainder term $O(\tau^2)$. Let us now proceed to prove the validity of (17).

First, to complete the definition of $\Lambda_{xx}u(x_m, y_n, t_p)$, we set $\Lambda_{xx}u\big|_\Gamma = 0$. Substituting

$$w_{mn} \equiv u(x_m, t_n, t_p) + \tau\Lambda_{xx}u(x_m, y_n, t_p)$$

in place of $\tilde{u}_{mn}$, into Eq. (16) we get

$$\frac{w_{mn} - u(x_m, y_n, t_p)}{\tau} - \Lambda_{xx} w_{mn} \equiv$$

$$\equiv \frac{\{u(x_m, y_n, t_p) + \tau\Lambda_{xx} u(x_m, y_n, t_p)\} - u(x_m, y_n, t_p)}{\tau} -$$

$$- \Lambda_{xx} u(x_m, y_n, t_p) - \tau\Lambda_{xx}\Lambda_{xx} u(x_m, y_n, t_p) =$$

$$= - \tau\Lambda_{xx}\Lambda_{xx} u(x_m, y_n, t_p).$$

Assuming that $\partial^4 u/\partial x^4$ is continuous and bounded, and taking into account that $\left.\frac{\partial^2 u}{\partial x^2}\right|_\Gamma = 0$, it is easy to see that $\Lambda_{xx}\Lambda_{xx} u(x_m, y_n, t_p)$ is bounded. Therefore

$$\frac{w_{mn} - u(x_m, y_n, t_p)}{\tau} - \Lambda_{xx} w_{mn} = O(\tau),$$

$$w_{mn}\Big|_\Gamma = 0.$$

Subtracting Eq. (16) term by term from these equations we get, for the difference $z_{mn} = w_{mn} - \tilde{u}_{mn}$,

$$\left.\begin{aligned} z_{mn} - \tau\Lambda_{xx} z_{mn} &= O(\tau^2), \\ z_{mn}\Big|_\Gamma &= 0, \end{aligned}\right\} \tag{18}$$

or in expanded form

$$\left.\begin{aligned} rz_{m-1,n} - 2\left(r + \frac{1}{2}\right) z_{mn} + rz_{m+1,n} &= O(\tau^2), \\ m, n &= 1, \ldots, N-1, \\ z_{0n} = z_{Mn} &= 0, \quad r = \tau/h^2. \end{aligned}\right\} \tag{18'}$$

But this problem for (z_{mn}) has the form

$$a_m u_{m-1} + b_m u_m + c_m u_{m+1} = g_m, \quad m = 1, \ldots, N-1,$$

$$u_0 = u_N = 0,$$

$$a_m > 0, \quad c_m > 0, \quad |b_m| \geq a_m + c_m + \delta, \quad \delta = 1.$$

In §4 it was shown that, in such a case

$$\max |u_m| \leq c \max|g_m|,$$

where c depends only on δ. Therefore $z_{mn} = O(\tau^2)$, i.e.

$$\tilde{u}_{mn} = w_{mn} - z_{mn} = w_{mn} + O(\tau^2) =$$

$$= u(x_m, y_n, t_p) + \tau \Lambda_{xx} u(x_m, y_n, t_p) + O(\tau^2).$$

which coincides with representation (17), as was to be proven.

PROBLEMS

1. For the differential boundary-value problem (1), relating to the propagation of heat in the square region $0 \leq x, y \leq 1$, propose and investigate a difference-splitting-scheme, analogous to the explicit splitting scheme (8) of §31 for the Cauchy problem.

2. For the differential boundary-value problem (1) propose a difference scheme analogous to the alternating-direction scheme (12) of §31. Prove that approximation is of order $O(\tau + h^2)$.

3. To solve the problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \qquad 0 \leq t \leq T, \quad (x, y) \text{ in } D, \\ &u(x, y, t)\big|_\Gamma = \psi(x, y, t) \qquad u(x, y, 0) = \psi(x, y) \end{aligned}\right\}$$

in the case of a region, D, with curved boundaries, propose a difference scheme analogous to the difference-splitting-scheme considered for problem (1) in the above text.

§33. Splitting by physical factors

The idea of splitting is used not only as a basis for development of economical and absolutely stable schemes. Sometimes one splits a complicated problem into simpler problems so as to separate, in each small time-interval $t_p < t < t_{p+1}$, the action of various factors which influence the process under study. For the resulting, relatively simple, problem it then becomes easier to construct satisfactory difference schemes which, together, constitute a difference-splitting-scheme for the entire problem.

By way of example we cite the method of super-particles, developed by O. M. Belotserkovskii and Yu. M. Davidov (U.S.S.R. Comp. Math. and Math. Phys. **11**, #1 (1971)) intended for the computation of gas flow with strong deformation of the medium and large density oscillations. This method, like Harlow's particle-in-cell method can, as pointed out by N. N. Yanenko, be treated as a certain difference-splitting-scheme for the gas-dynamics equations. The whole medium is split up, by a net of stationary lines (and

it should be noted that we are considering, here, the two dimensional problem), into cells. The material contained in a cell at time t_p is a "super-particle". To it is ascribed a momentum and total energy. Next one constructs a difference scheme modeling the change in speed, momentum and total energy of the super-particles under the influence of pressure alone, without taking into account those terms, in the system of gas dynamics equations, which describe the transport of matter, momentum and energy. This is the first step in the difference-splitting-scheme. In the second step one recomputes intermediate values, resulting from the first step, by a difference scheme which treats the remaining terms in the gas-dynamics equations, i.e., treats only the flow of matter from each cell to its neighbors, and the corresponding flow of momentum and energy. Thus one produces super-particles, with their corresponding momentum and energy, at time $t_{p+1} = t_p + \tau$.

Chapter 11
Elliptic Problems

§34. Simplest difference scheme for the Dirichlet problem

Here we will confirm that the simplest difference scheme (13) §24 approximates the Dirichlet problem (12) §24 to second order in h, and is stable, so that it may be used for the approximate computation of the solution of the Dirichlet problem.

The Dirichlet problem for the Poisson equation in the square region $D = (0 \leq x, y \leq 1)$, with boundary Γ, will be written in the form

$$\left.\begin{aligned} &\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \phi(x, y), \qquad 0 \leq x, y \leq 1, \\ &u|_{\Gamma} = \psi(s), \end{aligned}\right\} \tag{1}$$

where s is the arc length along the boundary Γ, and the functions $\phi(x, y)$ and $\psi(s)$ are given.

The set of net-points $(x_m, y_n) = (mh, nh)$ (h = 1/M, M a positive integer) falling in the square or on its boundary will be denoted by D_h. The points of D_h lying strictly inside the square D will be considered "internal points" of the net-square D_h: the set of all internal points we call D_h^0. The points of D_h lying on the boundary Γ of square D will be considered "boundary points" of net-region D_h, and the set of boundary points will be denoted by Γ_h. Difference scheme (13) §24

$$L_h u^{(h)} = f^{(h)} \tag{2}$$

we now write in the form

$$L_h u^{(h)} \equiv \begin{cases} \Lambda_{xx} u_{mn} + \Lambda_{yy} u_{mn} = \phi(x_m, y_n), & (x_m, y_n) \text{ in } D_h^0, \\ u_{mn} = \psi(s_{mn}), & (x_m, y_n) \text{ in } \Gamma_h, \end{cases} \tag{3}$$

where $\psi(s_{mn})$ is the value of the function $\psi(s)$ at the point (x_m, y_n) belonging to Γ_h.

1. Approximation. The right-hand side, $f^{(h)}$, of difference scheme (2) has the form

$$f^{(h)} = \begin{cases} \phi(x_m, y_n), & (x_m, y_n) \text{ in } D_h^0, \\ \psi(s_{mn}), & (x_m, y_n) \text{ in } \Gamma_h. \end{cases} \tag{4}$$

Assuming that the solution, $u(x, y)$, of problem (1) has bounded fourth derivatives one can, with the aid of Taylor's formulas, derive the equation

$$\Lambda_h u \equiv \Lambda_{xx} u + \Lambda_{yy} u = \\ = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{h^2}{24}\left(\frac{\partial^4 u(x + \xi h, y)}{\partial x^4} + \frac{\partial^4 u(x, y + \eta h)}{\partial y^4}\right). \tag{5}$$

Therefore for the solution $u(x, y)$ of problem (1) we may write

$$L_h[u]_h = \begin{cases} \phi(x_m, y_n) + O(h^2), & (x_m, y_n) \text{ in } D_h^0, \\ \psi(s_{mn}) + 0, & (x_m, y_n) \text{ in } \Gamma_h. \end{cases} \tag{6}$$

Thus the residual, $\delta f^{(h)}$, which develops when $[u]_h$ is substituted into the left-hand side of difference scheme (2) has the form

$$\delta f^{(h)} = \begin{cases} O(h^2), & (x_m, y_n) \text{ in } D_h^0, \\ 0, & (x_m, y_n) \text{ in } \Gamma_h. \end{cases} \tag{7}$$

In the space F_h, composed of elements of the form

$$f^{(h)} = \begin{cases} \phi_{mn}, & (x_m, y_n) \text{ in } D_h^0, \\ \psi_{mn}, & (x_m, y_n) \text{ in } \Gamma_h, \end{cases}$$

we introduce the norm

$$\|f^{(h)}\|_{F_h} = \max_{(mh,nh) \text{ in } D_h^0} |\phi_{mn}| + \max_{(mh,nh) \text{ in } \Gamma_h} |\psi_{mn}|. \tag{8}$$

Then

$$\|\delta f^{(h)}\|_{F_h} = O(h^2).$$

Thus difference boundary-value problem (3) approximates Dirichlet problem (1) to second order with respect to h.

2. Stability. Let us define a norm in the space, U_h, of functions given on the net D_h, setting

$$||u^{(h)}||_{U_h} = \max_{(mh,nh) \text{ in } D_h} |u_{mn}|. \tag{9}$$

To prove stability of difference scheme (3), as is our present goal, we will have to establish that, in accordance with the definition of stability, problem (2) has a unique solution for any arbitrary right-hand side $f^{(h)}$ (a property which does not depend on choice of norms), and that

$$||u^{(h)}||_{U_h} \leq c||f^{(h)}||_{F_h}, \tag{10}$$

where c depends neither on h nor on $f^{(h)}$.

Lemma 1. *Suppose that the function* $v^{(h)} = \{v_{mn}\}$ *is defined on the net* D_h *and at internal points,* $(x_m, y_n) = (mh, nh)$ *in* D_h^0, *satisfies the condition*

$$\Lambda_h v^{(h)}\Big|_{(mh,\, nh)} \geq 0, \qquad (mh, nh) \text{ in } D_h^0. \tag{11}$$

Then the maximum of $v^{(h)}$ *over the set* D_h *is attained at one or more points of* Γ_h.

Proof. Assume the opposite. Choose, among those points of D_h at which $v^{(h)}$ attains it's maximum, any single point (x_m, y_n) having the largest abcissa. By our assumption (x_m, y_n) is an internal point and, further, v_{mn} is strictly larger than $v_{m+1,n}$. At point m,n then we have

$$\Lambda_h v^{(h)}\Big|_{(mh,\, nh)} \equiv$$

$$\equiv \frac{(v_{m+1,n} - v_{mn}) + (v_{m,n+1} - v_{mn}) + (v_{m-1,n} - v_{mn}) + (v_{m,n-1} - v_{mn})}{h^2} < 0,$$

since, in the numerator, the first expression in parentheses is negative, and the others are non-positive. But this conclusion contradicts (11).

Lemma 2. *Suppose the function* $v^{(h)} = \{v_{mn}\}$ *is defined on the net* D_h *and, at interior points* (mh, nh) *in* D_h^0, *satisfies the condition*

$$\Lambda_h v^{(h)}\Big|_{(mh,\, nh)} \leq 0, \qquad (mh, nh) \text{ in } D_h^0. \tag{12}$$

Then the minimum value of $v^{(h)}$ *on the net* D_h *is taken on at least at one point of the boundary.*

The proof of Lemma 2 is analogous to that of Lemma 1.

Theorem ("maximum principle"). *Every solution of the difference equation*

$$\Lambda_h v^{(h)}\Big|_{(mh,\ nh)} = 0, \qquad (mh,\ nh) \text{ in } D_h^0. \tag{13}$$

attains its maximum and minimum values at points of Γ_h.

A proof may be constructed by combining what is asserted in Lemmas 1 and 2.

This property of the solution of difference equation (13) is analogous to the corresponding property of the solution, $v(x,y)$, of the Laplace equation $v_{xx} + v_{yy} = 0$, a solution which also takes on its least and greatest values on the boundary of the domain where it is defined.

From the maximum principle it follows that the problem

$$L_h u^{(h)} = \begin{cases} \Lambda_h v^{(h)}\Big|_{(mh,nh)} = 0, & (mh,nh) \text{ in } D_h^0, \\ \\ u^{(h)}\Big|_{(mh,nh)} = 0, & (mh,\ nh) \text{ in } \Gamma_h, \end{cases} \tag{14}$$

has only the vanishing solution $u^{(h)} = 0$, since the greatest and smallest values of this solution are taken on at points of Γ_h, where $u_{mn} = 0$. Therefore the determinant of the system of linear equations (3) is different from zero, and difference boundary-value problem (2) has a unique solution for any arbitrary right-hand side $f^{(h)}$.

Let us now go on to a proof of bound (10). By virtue of Eq. (5), for every polynomial $P(x,y)$ of second (or even third) order

$$P(x,y) = ax^2 + bxy + cy^2 + dx + ey + f$$

we have the equality

$$\Lambda_h P = \frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2}, \tag{15}$$

since the fourth derivatives of $P(x,y)$, which appear in the remainder terms of Eq. (5), all vanish.

Using the functions ϕ_{mn} and ψ_{mn} which form the right-hand sides of system (3), and taking $R > \sqrt{2}$, we construct the auxiliary function

$$P^{(h)}(x,\ y) = \frac{1}{4}\left[R^2 - (x^2 + y^2)\right] \max_{(mh,nh) \text{ in } D_h^0} |\phi_{mn}| + \max_{(mh,nh) \text{ in } \Gamma_h} |\psi_{mn}|,$$

which we will consider only at the points of D_h. This latter fact is indicated by the superscript h in the notation $P^{(h)}(x,y)$. From (15) we see that, at all points of D_h^0

$$\Lambda_h P^{(h)}\Big|_{\substack{x=mh,\\ y=nh}} \equiv -\max_{(rh,sh)\text{ in }D_h^0} |\phi_{rs}|, \qquad (mh,nh) \text{ in } D_h^0.$$

Therefore, at the points of D_h^0, the difference between the solution $u^{(h)}$, of problem (3) and the function $P^{(h)}$ satisfies the relations

$$\Lambda_h\left(u^{(h)} - P^{(h)}\right) = \Lambda_h u^{(h)} - \Lambda_h P^{(h)} = \phi_{mn} + \max_{r,s} |\phi_{rs}| \geq 0.$$

By lemma 1 the difference $u^{(h)} - P^{(h)}$ takes on its greatest value on the boundary Γ_h. But there this difference

$$u^{(h)}|_{\Gamma_h} - P^{(h)}|_{\Gamma_h} = \left[\psi_{mn} - P^{(h)}\right]_{mn} =$$

$$= \left[\psi_{mn} - \max_{(rh,sh)\text{ in }\Gamma_h} |\psi_{rs}|\right] + \frac{1}{4}\left[\left(x^2 + y^2\right) - R^2\right] \max_{(rh,sh)\text{ in }D_h^0} |\phi_{rs}|$$

is nonpositive since, everywhere in square D, $x^2 + y^2 < R^2$ and both square brackets on the right-hand side are nonpositive. Since the greatest value of $u^{(h)} - P^{(h)}$ is nonpositive, then everywhere on D_h

$$u^{(h)} - P^{(h)}|_{(mh,nh)} \leq 0, \qquad \text{or} \qquad u^{(h)} \leq P^{(h)}.$$

Similarly, for the function $u^{(h)} + P^{(h)}$ at the points of D_h^0 we have

$$\Lambda_h\left(u^{(h)} + P^{(h)}\right) \leq 0,$$

and at the points of Γ_h the sum $u^{(h)} + P^{(h)}$ is non-negative. By lemma 2, everywhere on D_h

$$u^{(h)} + P^{(h)} \geq 0, \qquad \text{or} \qquad -P^{(h)} \leq u^{(h)}.$$

Thus, everywhere on D_h

$$|u_{mn}| \leq |P_{mn}^{(h)}| \leq \frac{1}{4} R^2 \max_{(rh,sh)\text{ in }D_h^0} |\phi_{rs}| + \max_{(rh,sh)\text{ in }\Gamma_h} |\psi_{rs}|,$$

from which we get inequality (10):

$$\max|u_{mn}| = \|u^{(h)}\|_{U_h} \leq c\|f^{(h)}\|_{F_h} =$$

$$= \Big(\max_{(rh,sh) \text{ in } D_h^0} |\phi_{rs}| + \max_{(rh,sh) \text{ in } \Gamma_h} |\psi_{rs}| \Big),$$

where

$$c = \max\Big[1, \frac{R^2}{4}\Big],$$

completing the proof of stability.

In the case of the Dirichlet problem for the elliptic equation with variable coefficients

$$\frac{\partial}{\partial x}\Big[k_1(x,y)\frac{\partial u}{\partial x}\Big] + \frac{\partial}{\partial y}\Big[k_2(x,y)\frac{\partial u}{\partial y}\Big] = \phi(x,y), \qquad (x, y) \text{ in } D,$$

$$u|_\Gamma = \psi(s),$$

where $k_1(x,y)$ and $k_2(x,y)$ are positive, smooth functions in the rectangle D, one can construct difference equations analogously. Replacing the derivatives $\partial/\partial x\big(k_1\, \partial u/\partial x\big)$ and $\partial/\partial y\big(k_2\, \partial u/\partial y\big)$, at interior points of the net D_h^0, by difference expressions via the approximations

$$\frac{\partial}{\partial x}\Big[k_1(x,y)\frac{\partial u(x,y)}{\partial x}\Big] \approx \tilde{\Lambda}_{xx}u(x,y) \equiv$$

$$\equiv \frac{1}{h}\Big[k_1(x + h/2,\, y)\frac{u(x + h,\, y) - u(x,y)}{h} -$$

$$- k_1(x - h/2,\, y)\frac{u(x,y) - u(x - hy)}{h}\Big],$$

$$\frac{\partial}{\partial y}\Big[k_2(x,y)\frac{\partial u(x,y)}{\partial y}\Big] \approx \tilde{\Lambda}_{yy}u(x,y) \equiv$$

$$\equiv \frac{1}{h}\Big[k_2(x,\, y + h/2)\frac{u(x,\, y + h) - u(x,y)}{h} -$$

$$- k_2(x,\, y - h/2)\frac{u(x,y) - u(x,\, y - h)}{h}\Big],$$

we get a difference scheme (2) of the form

$$L_h u^{(h)} = \begin{cases} \tilde{\Lambda}_{xx}^{(h)} u^{(h)} + \tilde{\Lambda}_{yy}^{(h)} u^{(h)} = \phi(mh, nh), & (mh, nh) \text{ in } D_h^0, \\ u|_{\Gamma_h} = \phi(s_{mn}), & (mh, nh) \text{ in } \Gamma_h. \end{cases}$$

Using Taylor's formula one can convince oneself that approximation is of second order. It is possible to prove stability of this scheme after overcoming some additional difficulties which do not appear in the simpler examples we have chosen to consider.

In practice in treating specific problems one limits oneself, ordinarily, to very fundamental theoretical considerations, based on the analysis of model problems like those above. Concrete error estimates are obtained, as a rule, not from theoretical bounds, but from intercomparison of the results of computations carried out with various stepwidths, h.

After a difference boundary-value problem, approximating a given differential problem, is constructed one still needs to specify a method of solution which is not "too difficult". After all, for small h problem (2) is a system of scalar equations of very high order. In the example we have chosen the solution of the difference equations is a complex and interesting problem, but we defer consideration of this problem to §§35, 36.

PROBLEMS

1. Show that if, at the interior points of domain D_h, the function $u^{(h)}$ satisfies the equation

$$\Lambda_h u^{(h)}\Big|_{(mh,nh)} = 0, \qquad m, n = 1, 2, \ldots, M-1,$$

$$Mh = 1,$$

then either $u^{(h)}$ takes on, everywhere in D_h, one and the same value, or the greatest and least values of $u^{(h)}$ are not attained at <u>any</u> interior point of the set D_h (the "strengthened maximum principle").

2. If, at all internal points of the domain D_h, the condition $\Lambda_h u^{(h)} \geq 0$ is satisfied and, moreover, strict inequality holds at least at one point, then $u^{(h)}$ does not attain its maximum at any interior point.

3. Consider a difference scheme $L_h u^{(h)} = f^{(h)}$ of the form

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m+1,n} + u_{m,n+1} + u_{m-1,n} + u_{m,n-1} - 4u_{mn}}{h^2} = \\ \qquad = \psi(mh,nh), \qquad (mh, nh) \text{ in } D_h^0, \\ u_{mn} = \psi_1(s_{mn}) \text{ on } \Gamma_h^{(1)}, \\ \dfrac{u_{1,n} - u_{0,n}}{h} = \psi_2(nh), \qquad n = 1, \ldots, M-1. \end{cases}$$

This difference scheme approximates the problem (see Fig. 43)

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \phi(x,y) \qquad (x, y) \text{ in } D,$$

$$u(x,y)\big|_{\Gamma^{(1)}} = \psi_1(s), \quad (x,y) \text{ in } \Gamma^{(1)},$$

$$\left.\frac{\partial u}{\partial n}\right|_{\Gamma^{(2)}} = \psi_2(s), \qquad (x, y) \text{ in } \Gamma^{(2)}.$$

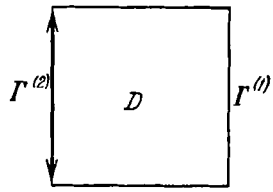


Fig. 43.

a) Prove that for any $\phi(mh,nh)$, $\psi_1(s_{mn})$ and $\psi_2(nh)$ the problem $L_h u^{(h)} = f^{(h)}$ has a unique solution.

b) Prove that if $\phi(mh,nh)$ is non-negative, and $\psi_1(s_{mn})$ and $\psi_2(nh)$ are nonpositive, then $u^{(h)}$ is nonpositive.

c) Prove that for any ϕ, ψ_1 and ψ_2 there exists a bound of form

$$\max_{m,n} |u_{mn}| \leq c\Big(\max_{m,n}|\phi_{mn}| + \max_{(mh,nh) \text{ in } \Gamma_1^{(h)}} |\psi_1(s_{mn})| + \max_n |\psi_2(nh)|\Big),$$

where c is some constant not depending on h. Compute c.

§35. Method of time-development

1. Idea of the method of time-development. To calculate the solutions of many of the stationary problems of mathematical physics, describing various equilibrium states, one considers these equilibria as the results of the approach-to-steady-state of processes developing in time, whose computational treatment is simpler than the direct calculation of the equilibrium state itself.

We illustrate the use of the method of time-development via the example of an algorithm for the computational solution of the Dirichlet problem

$$\left.\begin{aligned} \Lambda_{xx} u_{mn} + \Lambda_{yy} u_{mn} &= \phi(x_m, y_n), \qquad m, n = 1, 2, \ldots, M-1, \\ u_{mn}\big|_{\Gamma} &= \psi(s_{mn}), \end{aligned}\right\} \tag{1}$$

approximating the differential Dirichlet problem

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= \phi(x,y), \qquad 0 \leq x, \quad y \leq 1, \\ u|_{\Gamma} &= \psi(s). \end{aligned}\right\} \tag{2}$$

In the case of problem (1), which we will consider here, it will be possible to carry out a theoretical analysis of various time-development algorithms with the aid of finite Fourier series. Note that, for the solution of elliptic difference problems like problem (1), much more effective iterative methods have been developed. Some of these will be described in §§36, 37. Methods for the exact solution of problem (1), capable of generalization to the case of variable coefficients and domains with curvilinear boundaries (like the Gauss elimination method) for M at all large become very inconvenient, and tend not to be used.

We present, first, some introductory, orientational considerations. The solution u(x,y) of problem (2) can be taken to be the time independent temperature at point (x,y) of a plate in thermal equilibrium. Functions $\phi(x,y)$ and $\psi(s)$ are in this case, respectively, the distributions of heat sources and the temperature on the boundary.

Consider the auxiliary nonstationary heat-flow problem

$$\left.\begin{aligned} \frac{\partial U}{\partial t} &= \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} - \phi(x,y), \\ U|_{\Gamma} &= \psi(s), \\ U(x, y, 0) &= \psi_0(x, y), \end{aligned}\right\} \tag{3}$$

where ϕ and ψ are the same as in problem (2), and $\psi_0(x,y)$ is arbitrary.

Since the distribution of heat sources $\phi(x,y)$, and the boundary temperature $\psi(s)$, are time-independent, it is natural to expect that the solution, U(x,y,t) will change more and more slowly with time, and that the temperature distribution U(x,y,t), in the limit as $t \to \infty$, will evolve into the equilibrium temperature distribution u(x,y) characterized by problem (2). Therefore instead of stationary problem (2) one can solve nonstationary problem (3) out to the time, t, when the solution stops changing within the accuracy we require. This is the idea behind the solution of stationary problems by the "method of time-development".

In accordance with these considerations we will solve problem (3) instead of (2), and instead of difference scheme (1) for problem (2), we consider and compare three different difference schemes for problem (3).

First we will consider the simplest explicit difference scheme

$$\left.\begin{aligned} \frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} &= \Lambda_{xx}u_{mn}^{p} + \Lambda_{yy}u_{mn}^{p} - \phi(x_m, y_n), \\ u_{mn}^{p+1}\Big|_{\Gamma} &= \psi(s_{mn}), \\ u_{mn}^{0} &= \psi_0(x_m, y_n). \end{aligned}\right\} \tag{4}$$

Then we examine also the simplest implicit difference scheme

$$\left.\begin{aligned} \frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} &= \Lambda_{xx}u_{mn}^{p+1} + \Lambda_{yy}u_{mn}^{p+1} - \phi(x_m, y_n), \\ u_{mn}^{p+1}\Big|_{\Gamma} &= \psi(s_{mn}), \\ u_{mn}^{0} &= \psi_0(x_m, y_n). \end{aligned}\right\} \tag{5}$$

Finally we will study the alternating-direction scheme (12) §31:

$$\left.\begin{aligned} \frac{\tilde{u}_{mn} - u_{mn}^{p}}{\tau} &= \frac{1}{2}\left[\Lambda_{xx}\tilde{u}_{mn} + \Lambda_{yy}u_{mn}^{p} - \phi(x_m, y_n)\right], \\ \frac{u_{mn}^{p+1} - \tilde{u}_{mn}}{\tau} &= \frac{1}{2}\left[\Lambda_{xx}\tilde{u}_{mn} + \Lambda_{yy}u_{mn}^{p+1} - \phi(x_m, y_n)\right], \\ u_{mn}^{p+1}\Big|_{\Gamma} &= \tilde{u}_{mn}\Big|_{\Gamma} = \psi(s_{mn}), \\ u_{mn}^{0} &= \psi_0(x_m, y_n). \end{aligned}\right\} \tag{6}$$

It will be assumed that $\psi_0(x_m, y_n)$ is so defined that on the boundary

$$\psi_0\Big|_{\Gamma} = \psi(s_{mn}). \tag{7}$$

The computation of $u^{p+1} = \{u_{mn}^{p+1}\}$, given $u^{p} = \{u_{mn}^{p}\}$, is accomplished in scheme (4) by the use of explicit equations.

The computation of $u^{p+1} = \{u_{mn}^{p+1}\}$, given $u^{p} = \{u_{mn}^{p}\}$, by scheme (5) requires the solution of the problem

$$\left.\begin{aligned} \Lambda_{xx}u^{p+1}_{mn} + \Lambda_{yy}u^{p+1}_{mn} - \frac{u^{p+1}_{mn}}{\tau} &= \phi(x_m, y_n) - \frac{u^{p}_{mn}}{\tau}, \\ u^{p+1}_{mn}\Big|_{\Gamma} &= \psi(s_{mn}). \end{aligned}\right\}$$

This problem is in no way simpler than the original problem (1). Therefore it makes no sense to use this scheme for the approximate solution of (1). Finally, the computation of $u^{p+1} = \{u^{p+1}_{mn}\}$ from given $u^p = \{u^p_{mn}\}$ via scheme (6) is accomplished by the FEBS method used, first, in the direction of the x axis for the solution $\{\tilde{u}_{mn}\}$ of one-dimensional problems for each fixed n, and then in the y direction for computation of the solution $\{u^{p+1}_{mn}\}$ of one-dimensional problems for each fixed m.

The number of arithmetic operations is then proportional to the number of unknowns. For each of the two schemes, (4) and (6), which we set aside for further study, we will consider the difference

$$\varepsilon^p_{mn} \equiv u^p_{mn} - u_{mn} \tag{8}$$

between the net function $u^p = \{u^p_{mn}\}$ and the exact solution, $u = \{u_{mn}\}$, of problem (1), whose existence was demonstrated in §34.

We will determine under what conditions the error ε^p_{mn} in the solution u^p_{mn} of the nonstationary problem tends to zero with increasing p, and also the character of this convergence towards zero; we then choose an optimal τ and evalute the volume of computational work required to decrease the norm of the original error

$$\varepsilon^0_{mn} = \psi_0(x_m, y_n) - u_{mn}$$

by a given amount.

2. Analysis of the explicit time-development scheme. The solution $\{u_{mn}\}$ of problem (1), obviously, satisfies the equations

$$\left.\begin{aligned} \frac{u_{mn} - u_{mn}}{\tau} &= \Lambda_{xx}u_{mn} + \Lambda_{yy}u_{mn} - \phi(x_m, y_n), \\ u_{mn}\Big|_{\Gamma} &= \psi(s_{mn}), \\ u_{mn} &= u_{mn}. \end{aligned}\right\}$$

Subtracting these equations from Eq. (4) term by term, we get for the error, ε^p_{mn}, the following difference problem:

$$\left.\begin{aligned} &\frac{\varepsilon^{p+1}_{mn} - \varepsilon^{p}_{mn}}{\tau} = \Lambda_{xx}\varepsilon^{p}_{mn} + \Lambda_{yy}\varepsilon^{p}_{mn}, \\ &\varepsilon^{p+1}_{mn}\Big|_{\Gamma} = 0, \\ &\varepsilon^{0}_{mn} = \psi_0(x_m, y_n) - u_{mn}. \end{aligned}\right\} \tag{9}$$

Note that the net function ε^{p}_{mn} for each p, p = 0, 1, ..., vanishes on the points of Γ. This function may be considered an element of the linear space of functions, defined on the net $(x_m, y_n) = (mh, nh)$, m, n = 0, 1, ..., N, and vanishing on Γ. A norm in this space will be defined, as in §27, by the equation

$$||\varepsilon^{p}|| = \left(\sum_{m,n} \left|\varepsilon^{p}_{mn}\right|^2\right)^{1/2}.$$

In §27 we arrived at a representation of the solution of problem (17) in the form of a finite Fourier series. This problem differs from difference scheme (9) for the error $\varepsilon^{p} = \{\varepsilon^{p}_{mn}\}$ only in the designation of the unknown function. Therefore

$$\varepsilon^{p} = \sum_{r,s} (c_{rs}\lambda^{p}_{rs})\psi^{(r,s)}, \tag{10}$$

where the c_{rs} are coefficients in the expansion of the initial error, $\varepsilon^0 = \{\varepsilon^0_{mn}\}$, in a finite Fourier series, and the λ_{rs} are given by the expression

$$\lambda_{rs} = 1 - \frac{4\tau}{h^2}\left(\sin^2\frac{r\pi}{2M} + \sin^2\frac{s\pi}{2M}\right). \tag{11}$$

The quantities $c^{p}_{rs} \equiv c_{rs}\lambda^{p}_{rs}$ are the coefficients in expansions of the error, $\varepsilon^{p} = \{\varepsilon^{p}_{mn}\}$, in Fourier series with the orthonormal basis $\psi^{(r,s)}$. Therefore

$$||\varepsilon^{p}|| = \left(\sum \left|c_{rs}\lambda^{p}_{rs}\right|^2\right)^{1/2}, \qquad ||\varepsilon^{0}|| = \left(\sum |c_{rs}|^2\right)^{1/2}. \tag{12}$$

Clearly, then,

$$\frac{||\varepsilon^{p}||}{||\varepsilon^{0}||} \le \{\max_{r,s}|\lambda_{rs}|\}^{p}. \tag{13}$$

Further, one can always choose an ε^0 such that strict equality is attained. For this purpose one need only take $\varepsilon^0 = \psi^{(r',s')}$, where

(r',s') is that pair of indices for which

$$\max_{r,s}|\lambda_{rs}| = |\lambda_{r's'}|.$$

Thus, if $||\varepsilon^p||/||\varepsilon^0||$ is to tend to zero as $p \to \infty$, it is necessary that

$$\max_{r,s}|\psi_{rs}| < 1.$$

The error will decrease most rapidly if τ is so chosen that $\max_{r,s}|\lambda_{r,s}|$ will take on its smallest possible value. From Eq. (11) we find the left-most and right-most points λ_{rs}:

$$\lambda_{left} = 1 - \frac{8\tau}{h^2}\cos^2\frac{\pi}{2M},$$

$$\lambda_{right} = 1 - \frac{8\tau}{h^2}\sin^2\frac{\pi}{2M}$$

(Fig. 44). Increasing τ, starting from $\tau = 0$, we cause a leftward shift of both these points. For that value of τ for which the points are symmetrically placed with respect to the point $\lambda = 0$,

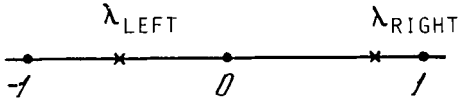


Fig. 44.

$$-\lambda_{left} = \lambda_{right}, \qquad (14)$$

any further increase in τ is harmful. In fact if τ increases further the right-hand point, λ_{right}, will continue to approach zero, but in return the left (which becomes largest in modulus, $\max_{rs}|\lambda_{rs}| = -\lambda_{left}$) moves further from zero.

For that τ for which $\lambda_{left} = -1$, and for larger τ's, the error ε^p will not tend to zero at all.

Thus the optimum $\tau = h^2/4$ is determined from condition (14). Moreover

$$\max_{r,s}|\lambda_{rs}| = 1 - 2\sin^2\frac{\pi}{2M}.$$

Therefore, to reduce the norm of the original error $\varepsilon^0 = \{\varepsilon^0_{mn}\}$ by a factor e it is necessary to carry out a number, p, of steps of iterative process (4), such that

$$\left(1 - 2\sin^2\frac{\pi}{2M}\right)^p < e^{-1}.$$

Therefore

$$p \geq -\frac{1}{\ln\left(1 - 2\sin^2\frac{\pi}{2M}\right)} \approx \frac{2M^2}{\pi^2}.$$

Let us now estimate how many arithmetic operations are required to reduce the error by e. For each transition from u^p to u^{p+1} one needs cM^2 arithmetic operations. Therefore the total number is $cpM^2 = O(M^4)$.

3. The alternating-direction scheme. We now turn to a study of the behavior of the error $\varepsilon^p = \{\varepsilon^p_{mn}\}$ for scheme (6).

As before we find that the error, ε^p, in this case is determined by the difference boundary-value problem

$$\left.\begin{aligned} \frac{\tilde{\varepsilon}_{mn} - \varepsilon^p_{mn}}{\tau/2} &= \Lambda_{xx}\tilde{\varepsilon}_{mn} + \Lambda_{yy}\varepsilon^p_{mn}, \\ \frac{\varepsilon^{p+1}_{mn} - \tilde{\varepsilon}_{mn}}{\tau/2} &= \Lambda_{xx}\tilde{\varepsilon}_{mn} + \Lambda_{yy}\varepsilon^{p+1}_{mn}, \\ \tilde{\varepsilon}_{mn}\Big|_\Gamma &= \varepsilon^p_{mn}\Big|_\Gamma = 0, \\ \varepsilon^0_{mn} &= \psi_0(x_m, y_n) - u_{mn}. \end{aligned}\right\} \tag{15}$$

The solution of problem (15) was written out in the form of a finite Fourier series in §27. As also for problem (9), it has form (10):

$$\varepsilon^p = \sum (c_{rs}\lambda^p_{rs})\psi^{(r,s)},$$

where the c_{rs} are coefficients in the expansion of the initial error

$$\varepsilon^0 = \sum c_{rs}\psi^{(r,s)}$$

in a finite Fourier series, but the quantities, λ_{rs}, by which the harmonics $\psi^{(r,s)}$ are multiplied in the transition from ε^p to ε^{p+1}, are now different:

$$\lambda_{rs} = \frac{\left(1 - 2\tau M^2 \sin^2 \frac{\pi r}{2M}\right)\left(1 - 2\tau M^2 \sin^2 \frac{\pi s}{2M}\right)}{\left(1 + 2\tau M^2 \sin^2 \frac{\pi r}{2M}\right)\left(1 + 2\tau M^2 \sin^2 \frac{\pi s}{2M}\right)}. \tag{16}$$

As in the analysis of the convergence of scheme (4), Eq. (13) is satisfied:

$$\frac{||\varepsilon^p||}{||\varepsilon^0||} \leq \{\max_{r,s}|\lambda_{rs}|\}^p,$$

and, moreover, equality is achieved for some special choice of $\varepsilon^0 = \{\varepsilon^0_{mn}\}$.

From expression (16) for λ_{rs} one can see that, for any τ, the condition $|\lambda_{rs}| < 1$ is satisfied and, consequently, $||\varepsilon^p||$ tends to zero. Further, $\lambda_{rs} = \lambda_r \cdot \lambda_s$, where

$$\lambda_k = \frac{1 - 2\tau M^2 \sin^2 \frac{\pi k}{2M}}{1 + 2\tau M^2 \sin^2 \frac{\pi k}{2M}}, \qquad k = 1, 2, \ldots, M-1.$$

Therefore $\max_{r,s} |\lambda_{rs}|$ is attained for $r = s = r'$, where r' is the index for which $|\lambda_{r'}|$ is maximum. Clearly the function $\lambda = (1 - x)/(1 + x)$ is monotonic. Therefore

$$\lambda_s = \frac{1 - 2\tau M^2 \sin^2 \frac{\pi s}{2M}}{1 + 2\tau M^2 \sin^2 \frac{\pi s}{2M}}$$

lies between the points

$$\lambda_{left} = \frac{1 - 2\tau M^2 \cos^2 \frac{\pi}{2M}}{1 + 2\tau M^2 \cos^2 \frac{\pi}{2M}}$$

and

$$\lambda_{right} = \frac{1 - 2\tau M^2 \sin^2 \frac{\pi}{2M}}{1 + 2\tau M^2 \sin^2 \frac{\pi}{2M}}$$

on the real axis. Increasing τ shifts the points λ_{left} and λ_{right} to the left. Therefore the quantity $\max_s |\lambda_s|$ will be smallest for the τ for which $-\lambda_{left} = \lambda_{right}$, i.e. for $\tau \approx 1/\sqrt{2}\pi M$. In this case

$$\max |\lambda_{rs}| = 1 - \frac{\sqrt{2}\,\pi}{M} + O\left(\frac{1}{M^2}\right).$$

To make the norm of the error $||\varepsilon^p||$ *smaller by a factor e than the original error-norm* $||\varepsilon^0||$ *the number of steps, p, must satisfy the condition* $[1 - (\pi\sqrt{2}/M)]^p \leq e^{-1}$, *so that*

$$p \approx \frac{M}{\pi\sqrt{2}} = O(M).$$

Each transition from u^p to u^{p+1} requires cM^2 arithmetic operations. Therefore the total number of arithmetical operations required to decrease the error e times is $O(M^3)$, and the number of operations needed to decrease the error k times is $O(M^3 \ln k)$.

We see that, for large M, the second of the time-development processes considered here, using the alternating directions scheme, yields a prescribed decrease in error at a smaller cost in arithmetic operations than the time-development method based on the use of the simplest explicit scheme ((4): for sufficiently large values of M (i.e. for fine nets) the alternating directions scheme turns out to be the more efficient of the two.

4. Choice of accuracy. We now make some remarks on the accuracy which must be attained in solving problem (1) by time-development, or some other method yielding a sequence of approximations, u^1, u^2, ..., u^p. Difference scheme (1) approximates problem (2) on a smooth solution $u(x,y)$ to order $h^2 = 1/M^2$. Therefore the exact solution $u^{(h)}$ of problem (1) differs from the desired table $[u]_h$ by a quantity of order $1/M^2$. Thus it makes no sense to calculate the solution $u^{(h)}$ of problem (1) with any greater accuracy. If we suppose that the zeroeth approximation $u^0 = \psi_0$ is given with an error of order 1, than the number, k, by which we want to decrease the norm of the error should be taken of order M^2. To decrease the original error by more than $O(M^2)$ would be a useless expenditure of computational effort.

In computations on a specific, fixed, net one iterates, in practice, until the sequence of approximations u^p, u^{p+1}, ..., stops changing within prescribed limits of accuracy.

5. Limits of applicability of methods. Difference scheme (4), as well as our analysis of error reduction, can be generalized to difference schemes approximating other boundary-value problems for elliptic equations with variable coefficients, in regions with curvilinear boundaries. Here it is important only that the operator $\tilde{\Lambda}_h$, analogous to the operator $-\Lambda_h \equiv -(\Lambda_{xx} + \Lambda_{yy})$ of scheme (1), ranging over net functions satisfying homogeneous boundary conditions, be selfadjoint and that its eigenvalues μ_j be of one sign:

$$0 < \mu_{min} < \mu_j < \mu_{max}.$$

In this case one uses for analysis in finite Fourier series, not the functions

$$\psi^{(r,s)} = 2 \sin \frac{\pi r}{M} \sin \frac{\pi s}{M},$$

but an orthonormal system of eigenfunctions of the selfadjoint operator $\tilde{\Lambda}_h$. It is known that such a system of eigenfunctions exists and is complete, and the specific form of these functions does not enter the general argument.

Alternating directions difference scheme (6) withstands generalization to the case of variable coefficients in domains with curvilinear boundaries (although the Fourier analysis then becomes impossible). For boundary conditions of the form $\alpha u + \beta \partial u/\partial n|_\Gamma = \psi$ a direct generalization of scheme (6) does not lead to an algorithm which separates into two one-dimensional FEBS calculations.

PROBLEMS

1. Write, in analogy to the above schemes (4) and (6), explicit and implicit time-development schemes for the solution of the Dirichlet problem

a) for the Laplace equation with variable coefficients:

$$\frac{\partial}{\partial x}\left[k_1(x,y)\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[k_2(x,y)\frac{\partial u}{\partial y}\right] = 0, \qquad 0 \le x, y \le 1$$

$$u|_\Gamma = \psi(x,y)|_\Gamma,$$

b) for the quasilinear equation

$$\frac{\partial}{\partial x}\left[k_1(u)\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[k_2(u)\frac{\partial u}{\partial y}\right] = 0, \qquad 0 \le x, y \le 1,$$

$$u|_\Gamma = \psi(x,y)|_\Gamma.$$

2. Show that, in the alternating direction method for the iterative solution of the Dirichlet problem

$$\left.\begin{aligned} &\Lambda_{xx}u_{mn} + \Lambda_{yy}u_{mn} = \phi(x_m, y_n), \\ &m, n = 1, 2, \ldots, M-1;\ Mh = 1, \\ &u_{mn}\big|_\Gamma = \psi(x,y)|_\Gamma \end{aligned}\right\}$$

one can choose an iteration parameter τ such that, after the very first iteration, the finite Fourier series for the error ε^p will not contain some single, prescribed, harmonic $\psi^{(r,s)}$.

§ 36. Iteration with variable step-size

1. The idea of Richardson. The convergence mechanism for the simplest time-development scheme (4) §35

$$\left.\begin{aligned} &\frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} = \Lambda_{xx}u_{mn}^{p} + \Lambda_{yy}u_{mn}^{p} - \phi(x_m, y_n), \\ &u_{mn}^{p+1}\big|_\Gamma = \psi(s_{mn}), \\ &u_{mn}^{0} = \psi_0(x_m, y_n) \end{aligned}\right\} \qquad (1)$$

consists, as we have seen, in the damping of each harmonic, $\psi^{(r,s)}$, contained in the Fourier series expansion of the error, $\varepsilon^0_{mn} = u_{mn} - u^0_{mn}$, in the zeroeth approximation. If

$$\varepsilon^p = \sum_{r,s} c^p_{rs}\psi^{(r,s)}, \tag{2}$$

then the Fourier coefficients of the error in the next approximation

$$\varepsilon^{p+1} = \sum_{r,s} c^{p+1}_{rs}\psi^{(r,s)}$$

can be expressed in terms of c^p_{rs} via the equations (see (10) and (11) §35):

$$c^{p+1}_{rs} = (1 - \tau\mu_{rs})c^p_{rs}, \quad \text{where} \quad \mu_{rs} = 4M^2\left(\sin^2\frac{\pi r}{2M} + \sin^2\frac{\pi s}{2M}\right). \tag{3}$$

For a given, fixed, τ not all the harmonics damp equally fast. The harmonics $\psi(r,s)$ which damp most quickly are those for which the damping factors $\lambda_{rs} \equiv 1 - \tau\mu_{rs}$ are closest to zero, i.e those for which $\mu_{rs} \approx 1/\tau$. This suggests that, step after step, we change the parameter τ so as to damp all the harmonics $\psi^{(r,s)}$ effectively in sequence, with the result *that, after several steps, all the harmonics will have damped more or less* uniformly.

This constitutes Richardson's idea for the solution of selfadjoint linear systems of equations, with matrices all of whose eigenvalues have the same sign.

2. The Chebyshev set of parameters. Richardson's iteration process is given by the equations

$$\left.\begin{aligned} u^{p+1}_{mn} &= u^p_{mn} + \tau_{p+1}\left[\Lambda_h u^p_{mn} - \phi(x_m, y_n)\right], \\ &\qquad m, n = 1, 2, \ldots, M-1, \\ u^{p+1}_{mn}\Big|_\Gamma &= \psi(s_{mn}); \quad \{u^0_{mn}\} \text{ given} \end{aligned}\right\} \tag{4}$$

with iteration parameters, τ_{p+1}, depending on the iteration number. Richardson indicated a useful, but not optimal, set of parameters $\{\tau_p\}$. We now present results on the optimum choice of iteration parameters $\{\tau_p\}$, and estimate the rate of decrease of the norm of the error $||\varepsilon^p||$. From Eq. (3) it is clear that the Fourier coefficients, c^k_{rs}, of the error ε^k in the k'th step, can be expressed in terms of the Fourier coefficients c^0_{rs} of the original error ε^0 by the equation

$$c^k_{rs} = c^0_{rs}\prod_{j=1}^{k}(1 - \tau_j\mu_{rs}), \qquad r, s = 1, 2, \ldots, M-1.$$

Let us now introduce the notation $Q_k(\mu)$, setting

$$Q_k(\mu) \equiv \prod_{j=1}^{k} (1 - \tau_j \mu). \tag{5}$$

Then

$$||\varepsilon^k||^2 = \sum_{r,s} |c_{rs}^k|^2 = \sum_{r,s} |Q_k(\mu_{rs}) c_{rs}^0|^2 \leq$$

$$\leq \max_{r,s} |Q_k(\mu_{rs})| \sum_{r,s} |c_{rs}^2|^2 = \max_{r,s} |Q_k(\mu_{rs})| \cdot ||\varepsilon^0||^2.$$

It is clear that the inequality

$$||\varepsilon^k|| \leq \max_{r,s} |Q_k(\mu_{rs})| \cdot ||\varepsilon^0||$$

becomes an exact equality for some ε^0. The quantities μ_{rs}, given by Eq. (3), are distributed over the interval

$$a = \mu_{min} \leq \mu \leq \mu_{max} = b, \tag{6}$$

where

$$\left.\begin{aligned} a = \mu_{min} &= 8M^2 \sin^2 \frac{\pi}{2M} \approx 2\pi^2, \\ b = \mu_{max} &= 8M^2 \cos^2 \frac{\pi}{2M} \approx 8M^2. \end{aligned}\right\} \tag{6'}$$

We will not rely on knowledge of the actual values of the numbers μ_{rs}, since this is an accidental circumstance, particular to our example. Instead we use only the fact that we know the boundaries, a and b, of the interval (6) on which they lie. Therefore, given k, we ask how one can define the iteration parameters $\tau_1, \tau_2, \ldots, \tau_k$, so that, among all polynomials, $Q_k(\mu)$, satisfying the condition

$$Q(0) = 1, \tag{7}$$

polynomial (5) will, on the interval $a \leq \mu \leq b$, deviate least from zero:

$$Q_k^* = \max_{a \leq \mu \leq b} |Q_k(\mu)| \text{ minimal.} \tag{8}$$

This problem in the theory of approximations was solved in 1892 by A. A. Markov. The desired polynomial $Q_k(\mu) \equiv \tilde{T}_k(\mu)$ may be expressed in

terms of the Chebychev polynomial (see for example V. L. Goncharov,* "Theory of Iteration and Approximation of Functions," 1954, in Russian)

$$T_k(x) \equiv \cos k \arccos x \equiv \frac{1}{2}\left(x + \sqrt{x^2 - 1}\right)^k + \left(x - \sqrt{x^2 - 1}\right)^k,$$

which, among all polynomials of order k with coefficient of x^k equal to one, deviates least from zero on the interval $-1 \leq x \leq 1$. In fact if one makes the linear transformation

$$x = \frac{b + a - 2\mu}{b - a}, \tag{9}$$

mapping the interval $a \leq \mu \leq b$ into the interval $-1 \leq x \leq 1$, and the point $\mu = 0$ into $x_0 = (b + a)/(b - a) > 1$, then

$$\tilde{T}_k(\mu) = \frac{T_k(x)}{T_k(x_0)} = \frac{\left(x + \sqrt{x^2 - 1}\right)^k + \left(x - \sqrt{x^2 - 1}\right)^k}{\left(x_0 + \sqrt{x_0^2 - 1}\right)^k + \left(x_0 - \sqrt{x_0^2 - 1}\right)^k}. \tag{10}$$

The set of iteration parameters $\tau_1, \tau_2, \ldots, \tau_k$, which generate polynomial (10), are defined through the condition that the zeroes, $\mu_j = 1/\tau_j$, of the polynomial $\tilde{T}_k(\mu)$, under transformation (9), should be mapped into the zeroes x_j of the Chebyshev polynomial $T_k(x)$:

$$\left.\begin{aligned} \tau_j &= \frac{2}{b + a - (b - a)x_j}, \\ x_j &= \cos\frac{\pi(2j - 1)}{2k}, \qquad j = 1, 2, \ldots, k. \end{aligned}\right\} \tag{11}$$

Let us now evaluate the maximum, Q_k^*, of the deviation from zero of polynomial $Q_k(\mu) \equiv \tilde{T}(\mu)$ on the interval $a \leq \mu \leq b$. As is known from the theory of approximation, the Chebyshev polynomial $T_k(x)$ takes on its maximum-modulus value, on the interval $-1 \leq x \leq 1$, at k + 1 points, including the ends of this interval. Therefore it follows from (10) that

$$Q_k^* = \frac{T_k(1)}{T_k(x_0)} = \frac{2}{\left(x_0 + \sqrt{x_0^2 - 1}\right)^k + \left(x_0 - \sqrt{x_0^2 - 1}\right)^k}. \tag{12}$$

Further, from (9) we get

$$\left.\begin{aligned} x_0 &= \frac{b + a}{b - a} = \frac{1 + \eta}{1 - \eta} = 1 + 2\eta + O(\eta^2), \\ \eta &= \frac{a}{b} = \frac{\mu_{\min}}{\mu_{\max}} \approx \frac{\pi^2}{4M^2}. \end{aligned}\right\} \tag{13}$$

*Or R. S. Varga, "Matrix Iterative Analysis" Prentice Hall 1962. (Translator's note).

Therefore for large M

$$x_0 \pm \sqrt{x_0^2 - 1} = 1 \pm 2\sqrt{\eta} + O(\eta),$$

from which, taking note of (13), it follows that

$$Q_k^* = \frac{2}{[1 + 2\sqrt{\eta} + O(\eta)]^k + [1 - 2\sqrt{\eta} + O(\eta)]^k} =$$

$$= 2 \div \{e^{k \ln(1 + 2\sqrt{\eta} + O(\eta))} + e^{k \ln(1 - 2\sqrt{\eta} + O(\eta))}\} \approx$$

$$\approx 2 \div \{e^{k\pi/M} + e^{-k\pi/M}\}.$$

Considering the fact that the norm of the initial error ε^0 is of order unity, $||\varepsilon^0|| \approx 1$, and noting the comments in 4§35 as to the accuracy which it is reasonable to achieve in the iterative solution process, we conclude that k should be obtained from the condition $Q_k^* \approx M^{-2}$, i.e.

$$k \approx \frac{2 \ln M + \ln 2}{\pi} M \approx \frac{2 \ln M + \ln 2}{2\sqrt{\eta}}. \qquad (14)$$

To reduce the initial error by a factor e^{-1}, *k must be obtained from the condition* $Q_k^* \leq e^{-1}$, i.e.

$$k \approx \frac{1 + \ln 2}{\pi} M \approx \frac{1 + \ln 2}{2\sqrt{\eta}} = O(M). \qquad (15)$$

Having chosen k in this way one can then consider the first k iterations as the first iteration cycle, and repeat the whole cycle with the same set of parameters $\tau_1, \tau_2, \ldots, \tau_k$. To decrease the norm of the error by a factor M^2, the number of cycles, ν, must be taken such that $\exp(-\nu) \sim 1/M^2$, $\nu \sim 2 \ln M$. The net number of elementary steps of the iterative process in ν cycles will be

$$k\nu \approx \left(\frac{1 + \ln 2}{\pi} M\right) 2 \ln M = O(M \ln M).$$

This exceeds only by the finite factor

$$2 \frac{1 + \ln 2}{2 + \ln 2/(\ln M)} \leq 1 + \ln 2$$

the number (14) of elemenary iterative steps required without cycling. Thus the use of cycling with $k \approx (1 + \ln 2)/(2\sqrt{\eta})$ achieves some simplification without substantially increasing the number of iterations.

The use of cycles of length $k \ll 1/(2\sqrt{\eta})$ is inadviseable. For example for k = 1 the Richardson process (4) transforms into simple iteration (1) with optimally chosen τ. The number of steps required, via this process,

to reduce the norm, $||\varepsilon^0||$, of the original error by a factor e is $\approx 2M^2/\pi^2$, as shown in 2§35. This number is O(M) times larger than the number of steps required to achieve this same reduction when the cycle-length is taken in accordance with (15).

3. Numbering of iteration parameters. Equation (11) gives the optimum set of iteration parameters $\tau_1, \tau_2, \ldots, \tau_k$ (for given fixed k). Suppose we now rearrange, somehow, the members of the sequence $\tau_1, \tau_2, \ldots, \tau_k$, into a new sequence $\kappa^{(k)} = (\kappa_1, \kappa_2, \ldots, \kappa_k)$, and iterate according to the formula

$$u^{p+1} = u^p + \tau_{\kappa_{p+1}}(\Lambda_h u^p - \phi), \qquad u^{p+1}|_\Gamma = \psi, \quad u^0 \text{ given}. \tag{16}$$

If algorithm (16) is realized exactly the result of the final, k'th, iteration will not depend on the details of the chosen sequence $\kappa^{(k)} = (\kappa_1, \kappa_2, \ldots, \kappa_k)$. But in real calculations, carried out on a machine with a finite number of significant figures, this ordering is extremely important. For large k it very strongly influences the sensitivity of the computed result to rounding errors committed in intervening steps of the process, i.e. the computational stability of the algorithm. Before introducing acceptable orderings $\kappa^{(k)} = (\kappa_1, \kappa_2 \ldots, \kappa_k)$, we note that the original algorithm (4) corresponding to the ordering $\kappa^{(k)} = (1, 2, \ldots, k)$ is useless, from a practical point of view.

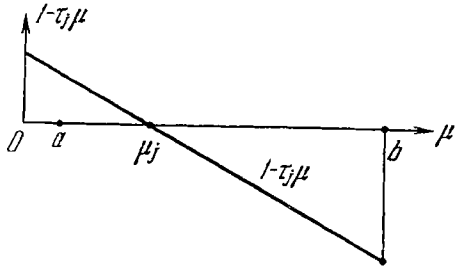


Fig. 45.

We now analyze the mechanism which gives rise to instability in this case. Suppose the original error ε^0 has the form $\varepsilon^0 = \sum c^0_{rs}\psi^{(r,s)}$, $c^0_{rs} \approx 1$, and the computation is carried out exactly, without roundoff error. Then the coefficients of the error in the ℓ'th approximation, $\varepsilon^\ell = \sum c^\ell_{rs}\psi^{(r,s)}$, are given by the expression

$$c^\ell_{rs} = \prod_{j=1}^{\ell}(1 - \tau_j\mu_{rs})c^0_{rs}.$$

Let us follow the evolution of c^ℓ_{rs} with increasing ℓ for r = M-1, s = M-1. In this case

$$\left.\begin{aligned} \mu_{rs} &= \mu_{M-1,M-1} = \mu_{max} = b \sim M^2, \\ \xi^\ell &\equiv \frac{c^\ell_{M-1,M-1}}{c^0_{M-1,M-1}} = \prod_{j=1}^{\ell}(1 - \tau_j b). \end{aligned}\right\} \tag{17}$$

Consider the linear functions $1 - \tau_j\mu$, $j = 1, 2, \ldots, k$, whose zeroes, $\mu_j = 1/\tau_j$, are determined by Eqs. (11). From these equations it is clear that, for $(2j - 1)/2k < 1/3$, or $j < (2k + 3)/6$ (and large enough M), we have $\mu_j < b/2$, and therefore (see Fig. 45)

$$|1 - \tau_j b| > 1 \quad \text{for} \quad j < \frac{2k + 3}{6}. \tag{18}$$

If $k \sim 1/(2\sqrt{\eta}) \sim M$ and $j \sim 1$, then $\mu_j \sim a + (1/M^2)$, $\tau_j \sim (1/a) - (1/M^2)$ and therefore

$$|1 - \tau_j b| \sim \frac{b}{a} \sim M^2.$$

Thus the value ξ^ℓ, defined by Eq. (17), increases initially by about a factor of M^2 per step, and then more slowly. We see from (18) that this growth continues at least as long as $\ell \leq (2k + 3)/6$, so that for $\ell \sim k/3$ the quantity $c^\ell_{M-1,M-1}$, and therefore also $||\epsilon^\ell||$, will be very large, increasing with k. In fact the order of magnitude of the value of the approximation $u^\ell = \{u^\ell_{mn}\}$ may exceed the limits allowed by the computer even for reasonable values of k, $k \ll M$.

If, hypothetically, this didn't happen, and the computation were continued with infinite precision, then up to step $\ell = k$ the quantity $c^\ell_{M-1,M-1}$ would decrease so that $\xi^k \leq Q^*_k$.

But the point is that, even if an overflow did not occur at $\ell \sim k/3$, then unavoidable, relatively small roundoff errors, at $\ell \sim k/3$ will be very large in absolute value. These errors are random, so that in their finite-Fourier series expansions all terms will be present and, in particular, the term of form

$$\tilde{c}_{1,1}\psi^{(1,1)},$$

where $\tilde{c}_{1,1}$ is a quantity which isn't small in absolute value.

We now show that, in later iterative steps, the error $\tilde{c}_{1,1}\psi^{(1,1)}$ introduced into the harmonic $\psi^{(1,1)}$ by roundoff in the step $\ell \sim k/3$ is not substantially damped, and distorts the computationally result unacceptably. The contribution, $\tilde{c}^k_{1,1}\psi^{(1,1)}$, of this error to the approximation u^k obtained at the last, k'th, step, is given by the expression

$$\tilde{c}^k_{1,1} = \Big[\prod_{j=\ell+1}^{k} (1 - \tau_j\mu_{11})\Big]\tilde{c}_{1,1} = \Big[\prod_{j=1+1}^{k} (1 - \tau_j\alpha)\Big]\tilde{c}_{1,1}.$$

But for $j > (2k + 3)/6$, clearly

$$\mu_j > \frac{1}{2}\Big[b + a - \frac{b - a}{2}\Big] > \frac{b}{4} \sim M^2.$$

Therefore

$$1 - \tau_j a \sim 1 - \frac{1}{M^2}, \qquad \left| \prod_{j=\ell+1}^{k} (1 - \tau_j a) \right| \sim \left(1 - \frac{1}{M^2}\right)^{k-1} \sim 1,$$

so that $\tilde{c}^{k}_{1,1} \sim \tilde{c}_{1,1}$, and the roundoff error has not been damped. Thus we have shown that use of the parameter sequence $\kappa^{(k)} = (1, 2, \ldots, k)$ is impractical.

If in the ℓ'th step of process (16) one has introduced a roundoff error

$$c_{rs}\psi^{(r,s)},$$

then at the k'th step this error evolves into

$$\left[\prod_{j=\ell+1}^{k} (1 - \tau_j \mu_{rs}) \right] c_{rs}\psi^{(r,s)}.$$

For this reason it seems desirable to try to achieve an ordering $\kappa^{(k)} = (\kappa_1, \kappa_2, \ldots, \kappa_k)$, for which

$$\max_{a \le \mu \le b} \left| \prod_{j=\ell+1}^{k} (1 - \tau_j \mu) \right| < A \tag{19}$$

with some moderate value of A.

Suppose $c^0_{rs}\psi^{(r,s)}$ is a component of the error ε^0 in the zeroth approximation. By the ℓ'th step this error develops into

$$\left[\prod_{j=1}^{\ell} (1 - \tau_j \mu_{rs}) \right] c^0_{rs}\psi^{(r,s)}.$$

If the norm of this function is large then roundoff will give contributions, large in absolute value, to all harmonics. It can then happen that a contribution to some harmonic will not be damped by further iteration, and will strongly distort the computed result. Therefore it is plausible to look for an ordering, $\kappa^{(k)} = (\kappa_1, \kappa_2, \ldots, \kappa_k)$, for which

$$\max_{a \le \mu \le b} \left| \prod_{j=1}^{\ell} (1 - \tau_j \mu) \right| < B \tag{20}$$

with some moderate value of B.

In the work of V. I. Lebedev and S. A. Finogenov, (U.S.S.R. Comp. Math. and Math. Phys. **11**, #2 (1971)), and of A. A. Samarskii [23], the authors describe various useful methods of ordering parameters, and shed some light on the history of this question. Here we present some results of V. I. Lebedev and S. A. Finogenov. In their work they assume that k is a power of 2, i.e. $k = 2^i$, and give a recurrence formula for the construction of $\kappa^{(k)}$.

Specifically, for i = 1

$$\kappa^{(k)} = (1, 2).$$

If for $k = 2^{i-1}$ the ordering $\kappa^{(2^{i-1})}$ has already been defined

$$\kappa^{(2^{i-1})} = (\kappa_1, \kappa_2, \ldots, \kappa_{2^{i-1}}),$$

then one sets

$$\kappa^{(2^i)} = (\kappa_1, 2^i + 1 - \kappa_1, \kappa_2, 2^i + 1 - \kappa_2, \ldots, 2^i + 1 - \kappa_{2^{i-1}}) \qquad (21)$$

In particular for $i = 2$, $i = 3$ and $i = 4$ we get, sequentially

$$(1, 4, 2, 3); \quad (1, 8, 4, 5, 2, 7, 3, 6);$$

$$(1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11).$$

By the indicated method (21) of ordering iteration parameters, the numbers A and B in inequalities (19) and (20) can be taken to be independent of k and ℓ.

The parameter-ordering algorithm presented by A. A. Samarskii has a somewhat more complicated formulation, but in return does not necessarily require that the order be a power of 2. The number k can have the form $k = (2j + 1)\cdot 2^i$. Instead of (19) and (20) the author establishes other bounds which, in some sense, guarantee stability.

4. The Douglas-Rachford method. In the alternating direction method (6) §35 we will take the iteration parameter, τ, to depend on the step-number, setting

$$\frac{\tilde{u}_{mn} - u^p_{mn}}{\tau_{p+1}} = \frac{1}{2}\left[\Lambda_{xx}\tilde{u}_{mn} + \Lambda_{yy}u^p_{mn} - \phi(x_m, y_n)\right],$$

$$\frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau_{p+1}} = \frac{1}{2}\left[\Lambda_{xx}\tilde{u}_{mn} + \Lambda_{yy}u^{p+1}_{mn} - \phi(x_m, y_n)\right],$$

$$u^{p+1}_{mn}\Big|_\Gamma = \tilde{u}_{mn}\Big|_\Gamma = \psi(s_{mn}), \quad m, n = 1, 2, \ldots, M-1.$$

For the error $\varepsilon^k = u^k - u$ we get the expression

$$\varepsilon^k = \sum_{r,s=1}^{M-1} c_{rs}^k \psi^{(r,s)}, \qquad c_{rs}^k = \left[\prod_{j=1}^{k} \lambda_r(\tau_j)\lambda_s(\tau_j)\right] c_{rs}^0,$$

where

$$\lambda_i(\tau) \equiv \frac{1 - 2\tau M^2 \sin^2 \frac{\pi i}{2M}}{1 + 2\tau M^2 \sin^2 \frac{\pi i}{2M}}, \qquad i = 1, 2, \ldots, k.$$

For a given k the optimum set of τ's is the set, $\tau_1, \tau_2, \ldots, \tau_k$, for which the quantity

$$\max_{r,s} \left| \prod_{j=1}^{k} \lambda_r(\tau_j)\lambda_s(\tau_j) \right|$$

takes on its smallest value. If one does not make use of the exact values of $\lambda_r(\tau)$ and $\lambda_s(\tau)$, but only of the boundaries within which these values lie, one gets involved in a Chebyshev-type problem like the problem for polynomials in section 2, above, but for products of rational fractions, each linear in numerator and denominator.

The statement of this problem and, as well, the proposal to solve the Poisson equation by the process of time-development using an alternating direction scheme, is due to Douglas, Peaceman and Rachford.* In the Douglas-Rachford work of 1956† which is presented here, this problem is solved approximately. For their choice of iteration parameters the number of iterative steps required to decrease the error by a factor e is $O(\ell nM)$, and the number of arithmetic operations is $O(M^2 \ell nM)$.

We show first that, given an arbitrary positive q, $q < 1$, one can choose iteration parameters $\tau_1, \tau_2, \ldots, \tau_k$, with $k = O(\ell nM)$, so as to satisfy the inequality

$$|[\lambda_r(\tau_1)\lambda_s(\tau_1)][\lambda_r(\tau_2)\lambda_s(\tau_2)] \ldots [\lambda_r(\tau_k)\lambda_s(\tau_k)]| < q, \qquad (22)$$

$$r, s = 1, 2, \ldots, M-1.$$

Then $||\varepsilon^k|| \leq q||\varepsilon^0||$. If one carries out the first k iterations with iteration parameters $\tau_1, \tau_2, \ldots, \tau_k$, and the next k iterations again

---

*See, for example, the discussion in R. S. Varga, "Matrix Iterative Analysis," Prentice Hall 1962. (Translator's note.)

†Douglas, J. Jr. and Rachford, H. H., Jr., "On the numerical solution of heat conduction problems in two and three space variables," Trans. Amer. Math. Soc. **82**, 421-439 (1956).

using $\tau_1, \tau_2, \ldots, \tau_k$, then to decrease the norm of the error by a factor e one will require, clearly, some number independent of M), each cycle made up of $k = O(\ell nM)$ iterations.

Let us now justify Eq. (22) and, in the process, explain how one can choose the parameters $\tau_1, \tau_2, \ldots, \tau_k$. It is clear that

$$|\lambda_i(\tau)| < 1, \qquad i = 1, 2, \ldots, M-1, \quad t > 0.$$

Therefore if inequality (22) is to be satisfied for any r,s = 1, 2, ..., M-1, it is sufficient that for each i = 1, 2, ..., M-1 at least one of the k factors $\lambda_i(\tau_p)$, p = 1, 2, ..., k should satisfy the inequality

$$|\lambda_i(\tau_p)| = \left|\frac{1 - 2\tau_p M^2 \sin^2 \frac{\pi i}{2M}}{1 + 2\tau_p M^2 \sin^2 \frac{\pi i}{2M}}\right| \leq \sqrt{q}\ . \tag{23}$$

All the quantities $2M^2\sin^2(\pi i/2M)$, i = 1, 2, ..., M-1 belong to the interval

$$a \leq 0.5\pi^2 \leq \mu \leq 2M^2 = b. \tag{24}$$

Thus to satisfy (22) it is sufficient, in view of (23), that for each μ in the interval (24) the inequality

$$-\sqrt{q} < \frac{1 - \tau_p \mu}{1 + \tau_p \mu} < \sqrt{q}\ ,$$

be satisfied for at least one τ, $\tau = \tau_1, \tau_2, \ldots, \tau_k$; and it is all the more sufficient that the inequality

$$-\sqrt{q} \leq 1 - \tau_p \mu \leq \sqrt{q}$$

be satisfied.

For this to be true it is necessary that, for each μ in the interval (24), there is a τ_p, p = 1, 2, ..., k, for which

$$1 - \sqrt{q} \leq \tau_p \mu \leq 1 + \sqrt{q}\ . \tag{25}$$

Let us define μ_p and τ_p, respectively, via the relations

$$\mu_p = \left(\frac{1 + \sqrt{q}}{1 - \sqrt{q}}\right)^{p-1} a, \qquad p = 1, 2, \ldots, k,$$

$$\tau_p = \frac{1 - \sqrt{q}}{\mu_p}\ , \qquad p = 1, 2, \ldots, k. \tag{26}$$

Then as μ increase from μ_p to μ_{p+1}, the product $\tau_p\mu$ traverses the interval (25).

Clearly, if we take k to satisfy the condition $\mu_k \geq b$, i.e.

$$\left.\begin{aligned} k \geq A \ln \frac{b}{a} + 1 = A\left(2 \ln M + \ln \frac{4}{\pi^2}\right) + 1, \\ A = \frac{1}{\ln \dfrac{1 + \sqrt{q}}{1 - \sqrt{q}}}, \end{aligned}\right\} \tag{27}$$

we do indeed get, from Eq. (25), the desired sequence $\tau_1, \tau_2, \ldots, \tau_k$.

PROBLEMS

1. Is it possible to choose iteration parameters $\tau_1, \tau_2, \ldots, \tau_k$ such that a finite number of iteration steps of process (4) will yield the exact solution of the Dirichlet difference problem?

How many iterations would be required? Can such a method permit generalization to the case where the exact eigenvalues μ_{rs} are unknown.

2. Explain the mechanism for the development of computational instability in computations via Eq. (16), with

$$\kappa^{(k)} = (k, k-1, k-2, \ldots, 2, 1),$$

for large k and M. Which harmonics $\psi^{(r,s)}$ will dominate in the Fourier series for the error, ε^k, in a calculation with $\kappa^{(k)} \approx (k, k-1, k-2, \ldots, 1)$, and in the presence of roundoff errors?

3. Suppose A is a selfadjoint operator whose eigenvalues lie on the interval $0 < \mu_{min} < \mu < \mu_{max}$. What constraint must be satisfied by the condition-number $\eta = \mu_{min}/\mu_{max}$, if, for the equation $Ax = \phi$, the Richardson process

$$x^{p+1} = x^p - \tau_{p+1}(Ax^p - \phi)$$

is to converge, and be computationally stable, for any arbitrary choice of $\tau_j = 1/\mu_j$, $\mu_{min} < \mu_j < \mu_{max}$, $j = 1, 2, \ldots, k$, and arbitrary k?

4. Why is it that, in the Douglas-Rachford scheme, the ordering of the parameters $\tau_1, \tau_2, \ldots, \tau_k$, has no substantial influence on the computational stability of the iterative process?

5. Assuming the the machine time required for one step of the Douglas-Rachford process is twenty times greater than for one step of the Richardson scheme estimate, through use of Eqs. (15) and (27) for what value of M the superiority of the Douglas-Rachford method first becomes apparent.

§37. The Federenko Method

In the work of R. P. Federenko, U.S.S.R. Comp. Math. and Math. Phys. 1, #5 (1961), the author presents an iteration method* for the solution of elliptic difference problems, a method which he calls relaxational. To decrease the norm of the original error by a factor e this method requires, in all, cM^2 arithmetic operations, where M is the number of net-steps in one direction and c is some constant not depending on M. We note that the most rapidly convergent of the above (and generally of all other known) methods, the Douglas-Rachford method requires, for the same error reduction, $O((\ell nM)M^2)$ arithmetic operations.

The range of applicability of the Fedorenko method is almost the same as that of the simplest time-development method. An additional limitation is the requirement of "smoothness" of the lowest-order eigenvectors, a requirement ordinarily fulfilled for elliptic problems.

In simple examples the computational speed of this method, as compared with the most quickly-convergent iterative methods of other types, is already convincingly demonstrated for $M \approx 50$. It must be kept in mind that the logical organization of the relaxational method is substantially more complicated, as we shall see, than the logic of, let us say, the Richardson scheme. Therefore the computer time depends very strongly on the quality of the computer program.

The simplest estimate of convergence rate (for a difference approximation to the Poisson equation in a square region, on a square net, with given boundary-values) was obtained by R. P. Fedorenko, U.S.S.R. Comp. Math. and Math. Phys. 4, #3 (1964).

In the work of N. S. Bakhvalov (U.S.S.R. Comp. Math. and Math. Phys. 6, #5 (1966)) the author studied the convergence of the Fedorenko method, and got precisely the same result for the difference analogue of the first boundary-value problem in a rectangle, for a general elliptic equation with smooth coefficients

$$a_{11}\frac{\partial^2 u}{\partial x^2} + 2a_{12}\frac{\partial^2 u}{\partial x \partial y} + a_{22}\frac{\partial^2 u}{\partial y^2} + a_1\frac{\partial u}{\partial x} + a_2\frac{\partial u}{\partial y} + au = f_0.$$

Finally, G. P. Astrakhantsev (U.S.S.R. Comp. Math. and Math. Phys. 11, #2 (1971)) got analogous results for a difference approximation to the third boundary-value problem for a selfadjoint difference equation in an arbitrary two-dimensional region with smooth boundaries.

Since the derivation is very involved we limit ourselves to a qualitative description of the idea of the method, and of the Fedorenko algorithm itself, referring the reader to proofs in the original work and, to a review article by R. P. Fedorenko, Uspekhi. Mat. Nauk**28, Vol. 2 (1973).

* * * * * *

*An early variant of the multi-grid method. For a more extensive presentation of multi-grid methods see, for example, "Multi-Grid Methods and Applications", W. Hackbusch, Springer-Verlag (1985). (Translator's note).

**Title of translated journal is Russian Math. Surveys. (Translator's note.)

1. Idea of the method. To arrive at an iterative solution of the problem

$$\left.\begin{aligned} \Lambda_h u_{mn} - \phi(x_m, y_n) = 0, \qquad m, n = 1, 2, \ldots, M-1, \\ u_{mn}\big|_\Gamma = \psi(s_{mn}) \end{aligned}\right\} \tag{1}$$

we set out from the simplest time-development process (4) §35

$$\left.\begin{aligned} u_{mn}^{p+1} = u_{mn}^{p} + \tau(\Lambda_h u_{mn}^{p} - \phi(x_m, y_n)), \qquad m, n = 1, 2, \ldots, M-1, \\ u_{mn}^{p+1}\big|_\Gamma = \psi(s_{mn}), \qquad \{u_{mn}^{0}\} \text{ given}, \end{aligned}\right\} \tag{2}$$

which, on the whole, converges very slowly but uniformly in the various harmonics. The error $\varepsilon^p = u^p - u$, as in (10) §35, may be written in the form of a finite Fourier series

$$\varepsilon^p = \sum_{r,s=1}^{M-1} [\lambda_{rs}(\tau)]^p \, c_{rs}^0 \psi^{(r,s)}, \tag{3}$$

where the c_{rs}^0 are expansion coefficients of the error, $\varepsilon^0 = u^0 - u$, in the zeroeth approximation, and

$$\lambda_{rs} = 1 - 4\tau M^2\left(\sin^2 \frac{\pi r}{2M} + \sin^2 \frac{\pi s}{2M}\right).$$

The quantities λ_{rs} lie on the interval $\lambda_{left} \leq \lambda \leq \lambda_{right}$, where

$$\lambda_{left} = \lambda_{M-1,M-1} \approx 1 - 8\tau M^2,$$

$$\lambda_{right} = \lambda_{1,1} \approx 1 - 2\pi^2\tau.$$

Suppose that

$$\tau = \frac{3}{16M^2}. \tag{4}$$

If, under this condition, at least one of the numbers r or s is greater than M/2, then

$$|\lambda_{rs}| < 0.6. \tag{5}$$

Therefore the contribution of the high-frequency harmonics $\psi^{(r,s)}$, $r > M/2$ or $s > M/2$, to the error (3) decreases, in one iterative step, almost in half, and quickly becomes small. After several iterations via Eq. (2) the error ε^p will contain, essentially only a smooth component (harmonics $\psi^{(r,s)}$, $r < M/2$, $s < M/2$), because the low-frequency harmonics $\psi^{(r,s)}$ are multiplied by factors $\lambda_{r,s}$ which are closer to unity. The contribution of the first harmonic $\psi^{(1,1)}$ damps very slowly; for the given choice of τ

$$\lambda_{1,1} \approx 1 - \frac{3\pi^2}{8M^2} \quad (\approx 1). \tag{6}$$

Let us designate by U the approximation u^p, obtained by iterative process (2), and let v by the error $\varepsilon^p = u^p - u = U - u$. If we knew the error v we could find the desired solution $u = U - v$. But all we know about v is that it satisfies the equation

$$\Lambda_h v = \xi, \qquad v|_\Gamma = 0, \tag{7}$$

where ξ, a known net function, is the residual obtained when one substitutes $u^p = U$ into Eq. (1):

$$\xi = \Lambda_h u^p - \phi = \Lambda_h u^p - \Lambda_h u = \Lambda_h(u^p - u) = \Lambda_h v.$$

Problem (7) which determines the correction v is simpler than the original problem (1) only in the sense that we know that v is a smooth net function. Therefore to determine v we can take as an approximation this same problem posed on a net twice as coarse which (for even M) is contained in the original net:

$$\Lambda_{2h} v^* = \xi^*, \qquad v^*|_{\Gamma^*} = 0. \tag{1*}$$

Here the asterisk designates quantities pertaining to the coarsened net. We will solve problem (1*) by the iterative process

$$\left.\begin{aligned} (v^*_{mn})^{p+1} &= (v^*_{mn})^p + \tau^*[\Lambda_{2h}(v^*_{mn})^p - \xi^*_{mn}], \\ m, n &= 1, 2, \ldots, M^*-1, \\ (v^*_{mn})^{p+1}\big|_{\Gamma^*} &= 0 \end{aligned}\right\} \tag{2*}$$

taking as a zeroeth approximation $(v^*_{mn})^0 \equiv 0$. Here $M^* = M/2$, $\tau^* = 3/[16(M^*)^2] = 4\tau$.

Each step of iterative process (2*) requires only a quarter as much work as a step of (2), because there are only a quarter as many points in the computational mesh. Further, thanks to the fact that $\tau^* = 4\tau$, the attenuation of the most slowly damped error component proceeds more quickly. Corresponding to (6) we now have

$$\lambda^*_{1,1} = 1 - \frac{3\pi^2}{8(M^*)^2} = 1 - 4\,\frac{3\pi^2}{8M^2} < \lambda_{1,1},$$

and, to attenuate the contribution $\psi^{(1,1)}$ by a factor e, one needs a fourth as many iterations. Let us designate by V* the result of iteration by Eq. (2*). We next interpolate V* onto the original net (linearly). Smooth components will be obtained almost exactly. The error induced in the smooth function by interpolation will be small relative to the interpolated smooth function, but (since the error due to interpolation is jagged

because of slope-discontinuities at the interpolation points) the Fourier expansion of the error will contain all components. In addition the non-smooth component of V*, which bears no relation to the desired correction, upon interpolation also gives a random contribution to the non-smooth component of the function V, produced by the interpolation process. Thus the smooth component of the difference U-V is close to the smooth component of the desired solution u = U-v, but the non-smooth component is not very small and has random features.

Therefore it is necessary to execute a few more steps of the original iterative process (2), taking U-V as an initial approximation. In this way one quickly damps the non-smooth error components, introduced by the interpolation process, and attenuated by process (2) almost by a factor of 2 in a single step.

2. Description of the algorithm. The convergence-acceleration, achieved by use of the coarsened mesh and process (2*), may prove inadequate. For large M (i.e. a fine net) problem (1*) on the coarsened net may still be difficult. Therefore to solve this second problem it may be worthwhile to carry out still another mesh-width-doubling, and to solve the problem on the quadrupled mesh one may again double the mesh-width, again doubling τ, etc. In the experiments of R. P. Fedorenko the net step-sizes were not doubled, but tripled. For $M \approx 100$ two coarsenings turn out to be sufficient. We will assume for simplicity that $M = 2^k$, i.e. M is some power of two.

On the original net we take several steps of iteration (2) to "smooth" the error. This error is unknown to us and, therefore, we monitor the iterative process by keeping track of the residual, $\Delta_h u^p - \phi$, which also undergoes smoothing. The result of the calculation $U = u^p$ is stored for later use. Next, to calculate the correction v, we treat the problem on the coarsened mesh, performing some iterations (2*) so as to smooth the "correction to the correction" and storing the result $\tilde{V}^*$ (which occupies only a fourth as much storage space as U). To calculate the correction to $\tilde{V}^*$ we consider the problem on a net again coarsened by doubling, and do several iterations with a step-size $\tau^{**} = 4\tau^* = 16\tau$, storing the result $\tilde{V}^{**}$. This process of computing corrections to corrections, on nets coarsened by doubling, is repeated k times until one gets to the coarsest net and to the correction $\tilde{V}^{(k^*)}$.

Next one starts to move back to the fine net. First, from the coarsest net, one interpolates the last-computed correction $\tilde{V}^{(k^*)}$ onto the next twice-as-fine, net, and inserts the interpolated correction into $\tilde{V}^{((k-1)^*)}$ performing several iterations to damp errors introduced by the interpolation. The results of these iterations are then interpolated onto the next twice-as-fine net: through use of this interpolated function, one refines the stored correction, $\tilde{V}^{((k-2)^*)}$, pertaining to this net, performs several iterations, and carries out the next interpolation. On the next-to-last step, after correcting $\tilde{V}^*$ and iterating, one gets the correction V*, which is then interpolated onto the original net. Then, performing some iterations (2) on U-V, the final result is obtained.

Chapter 12
The Concept of Variational-Difference and Projection-Difference Schemes

In this chapter we present a method for constructing difference schemes, based on the use of one or another variational or projective formulation of the boundary-value problem whose solution we wish to evaluate numerically. This method, sometimes called the finite element method, allows one to construct effective difference schemes on irregular nets, and with a minimum of assumptions as to the smoothness of the desired solution, or of the coefficients of the equation. Thanks to the resulting freedom in our choice of nets the net-points may be distributed more densely in those parts of the domain of definition of the desired function where its form is particularly complicated, or where we are interested in the finer details of its behavior.

Our ability to distribute points appropriately allows us to attain a desired accuracy with a minimum number of net-points.

The method of finite elements may be interpreted as one of the possible realizations of the classical variational methods for the solution of boundary-value problems. For this reason we begin (§38) with a description of the classical variational and projection methods, and then (§39) discuss variational-difference schemes.

§38. Variational and projection methods

1. Variational formulation of boundary-value problems. Many differential boundary-value problems of mathematical physics admit natural variational formulations. We limit ourselves to a consideration of two simple examples of such problems and their variational formulations which, however, illustrate what is essential here. In these examples we will be discussing various boundary-value problems for the Poisson equation in a certain bounded domain D of the xy plane, with a piecewise-smooth boundary Γ.

We designate by W the linear space of all functions continuous in domain D and on its boundary Γ and possessing, in addition, bounded first derivatives which may have discontinuities only on a finite set of lines (a set which may be different for each of the functions, $w(x,y)$, in space W). We will introduce a norm, in space W, setting

$$||w||_W = \left[\int_D \int w^2 dxdy + \int_D \int \left[\left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2\right] dx\, dy\right]^{1/2} \tag{1}$$

for each of its member functions w.

* * * * * *

Completion of the space W leads to the complete Sobolev space W_2^1.

* * *

Let us now turn to a consideration of examples.

Example 1. Consider the first boundary-value problem (the Dirichlet problem)

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= f(x, y), \qquad (x, y) \text{ in } D, \\ u|_\Gamma &= \phi(s), \end{aligned}\right\} \tag{A}$$

where s is the arc-length along the boundary Γ of domain D. Further, f(x,y) and $\phi(s)$ are given functions, functions satisfying all conditions which are needed to assure that the solution, u(x,y), of problem (A) will have continuous second derivatives everywhere in D, and on its boundary Γ.

Theorem 1. *Among the functions, w in W, satisfying the boundary condition*

$$w|_\Gamma = \phi(s), \tag{2}$$

the solution u(x,y) of problem (A) gives to the expression (or "functional")

$$I(w) \equiv \int_D \int \left[\left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 + 2fw\right] dxdy \tag{3}$$

its minimum numerical value.

Proof. Let w(x,y) in W, $w|_\Gamma = \phi(s)$, be some given fixed function. Introduce the notation $\xi(x,y) \equiv w(x,y) - u(x,y)$, so that

$$w(x,y) = u(x,y) + \xi(x,y).$$

Since u(x,y) has continuous second derivatives, and w(x,y) is in W, then also $\xi(x,y)$ is in W and, moreover, $\xi|_\Gamma = 0$. We now prove that

$$I(w) \equiv I(u + \xi) = I(u) + \int_D \int \left[\left(\frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial \xi}{\partial y}\right)^2\right] dxdy, \tag{4}$$

from which the theorem follows since, in the case $w(x,y) \neq u(x,y)$, the function $\xi(x,y)$ doesn't vanish identically, so that the second term on the right-hand side of Eq. (4) is strictly positive, and $I(w) > I(u)$. Clearly

$$I(u+\xi) \equiv \int_D \int \left[\left(\frac{\partial u}{\partial x} + \frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y} + \frac{\partial \xi}{\partial y}\right)^2 + 2f(u+\xi)\right] dxdy =$$

$$= I(u) + \int_D \int \left[\left(\frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial \xi}{\partial y}\right)^2\right] dxdy +$$

$$+ 2 \int_D \int \left(\frac{\partial u}{\partial x}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \xi}{\partial y} + f\xi\right) dxdy.$$

It remains to be shown that the third term on the right-hand side vanishes. In fact, from the obvious identities

$$\frac{\partial u}{\partial x}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \xi}{\partial y} \equiv \frac{\partial}{\partial x}\left(\xi \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\xi \frac{\partial u}{\partial y}\right) -$$

$$- \xi\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \equiv \frac{\partial}{\partial x}\left(\xi \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\xi \frac{\partial u}{\partial y}\right) - \xi f \tag{5}$$

it follows that

$$\int_D \int \left(\frac{\partial u}{\partial x}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \xi}{\partial y} + f\xi\right) dxdy =$$

$$= \int_D \int \left[\frac{\partial}{\partial x}\left(\xi \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\xi \frac{\partial u}{\partial y}\right)\right] dxdy = \int_\Gamma \xi \frac{\partial u}{\partial n}\, ds = 0, \tag{6}$$

where $\partial u/\partial n$ is the derivative along the inward-directed normal.

In the next-to-last link in the chain of equations (6) we used the theorem of vector analysis which states that the integral of the divergence of a vector field over a region is equal to the flux of this vector field across the region-boundary. In the given case this flux $\int_\Gamma \xi(\partial u/\partial n)ds$ vanishes, since $\xi|_\Gamma = 0$. The theorem is proven.

Thus problem (A) admits the following variational formulation: *among all the functions of class W satisfying condition (2), find the one that minimizes the functional I(w), defined by Eq. (3).*

Example 2. Consider the third boundary value problem

$$\left.\begin{aligned} \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} &= f(x,y), \qquad (x, y) \text{ in } D, \\ \frac{\partial v}{\partial n} + \sigma(s)\, v|_\Gamma &= \phi(s), \end{aligned}\right\} \tag{B}$$

where $f(x,y)$, $\phi(s)$ and $\sigma(s) \geq \sigma_0 > 0$ are given functions, and $\partial v/\partial n$ is the derivative along the inward-directed normal.

Theorem 2. *Among all functions w in W the solution v of problem (B) minimizes the functional*

$$J(w) \equiv \int_D \int \left[\left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 + 2fw\right]dxdy +$$

$$+ \int_\Gamma \left[\sigma(s)w^2 - 2\phi(s)w\right]ds \tag{7}$$

Proof. Let w in W be some given function, while

$$\eta(x,y) \equiv w(x,y) - v(x,y).$$

We now prove the equality

$$J(w) \equiv J(v + \eta) + J(v) +$$

$$+ \left\{\int_D \int \left[\left(\frac{\partial \eta}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial y}\right)^2\right]dxdy + \int_\Gamma \sigma(s)\eta^2 ds\right\}, \tag{8}$$

from which it follows that for $w \neq v$, i.e. $\eta \neq 0$, we have the inequality $J(w) > J(v)$, whose validity is asserted in the theorem.

Clearly,

$$J(w) \equiv J(v + \eta) =$$

$$= \int_D \int \left[\left(\frac{\partial v}{\partial x} + \frac{\partial \eta}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y} + \frac{\partial \eta}{\partial y}\right)^2 + 2f(v + \eta)\right]dxdy +$$

$$+ \int_\Gamma \left[\sigma(s)(v + \eta)^2 - 2\phi(s)(v + \eta)\right]ds =$$

$$= J(v) + \left\{\int_D \int \left[\left(\frac{\partial \eta}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial y}\right)^2\right]dxdy + \int_\Gamma \sigma(s)\eta^2 ds\right\} +$$

$$+ 2\left\{\int_D \int \left[\frac{\partial v}{\partial x}\frac{\partial \eta}{\partial x} + \frac{\partial v}{\partial y}\frac{\partial \eta}{\partial y} + f\eta\right]dxdy + \int_\Gamma \left[\sigma(s)v - \phi(s)\right]\eta ds\right\}. \tag{9}$$

It remains for us to show that the expression on the right-hand side of (9), in the second pair of curly brackets, vanishes. In fact transforming the double integral in this expression as in (6) we get

$$\int_D \int \left[\frac{\partial v}{\partial x}\frac{\partial \eta}{\partial x} + \frac{\partial v}{\partial y}\frac{\partial \eta}{\partial y} + f\eta\right]dxdy + \int_\Gamma \left[\sigma(s)v - \phi(s)\right]\eta ds =$$

$$= \int_\Gamma \frac{\partial v}{\partial n} \cdot \eta ds + \int_\Gamma \left[\sigma(s)v - \phi(s)\right]\eta ds = \int_\Gamma \eta\left[\frac{\partial v}{\partial n} + \sigma v - \phi\right]ds = 0,$$

since $\partial v/\partial n + \sigma v|_\Gamma \equiv \phi$. The theorem has been proven.

Thus the third boundary-value problem for Poisson equation (B) allows the following variational formulation: *among all functions w in W, find that one which minimizes the functional J(w)* introduced in Eq. (7).

We direct the reader's attention to the fact that the difference between the variational formulations of boundary-value problems (A) and (B) lies not only in the difference between the functionals I(w) and J(w). In minimizing the functional J(w) we are allowed, as trial functions, all functions w in W, while in minimizing the functional I(w) trial functions are admissible only if they satisfy the boundary condition, $w|_\Gamma = \phi(s)$, of problem (A).

It is because of this difference that one calls the boundary condition of problem (B) "natural": in the variational formulation it imposes no limitation on the class of admissable functions.

2. Convergence of minimizing sequences. The exact solution of problem (A), as we have seen, is that function $w(x,y) \equiv u(x,y)$ which, among all *allowable* functions (i.e. functions w in W satisfying the condition $w|_\Gamma = \phi(s)$, minimizes the functional I(w). The numerical solution of the problem of finding u(x,y) consists in the construction of a function, w in W, $w_\Gamma = \phi(s)$, which gives the functional, if not its minimum value, then at least a value "close" to this minimum. More precisely, for computational purposes one must designate a method for constructing the terms of a sequence of admissable functions, $w_N(x,y)$ in W, $w_N\big|_\Gamma = \phi(s)$, for which

$$\lim_{N\to\infty} I(w_N) = I(u).$$

Such a sequence of allowable functions is called a "minimizing sequence". Choosing a term, $w_N(x,y)$, of the minimizing sequence with large enough N, one can attain a functional value, $I(w_N)$, as close as one likes to I(u).

Completely analogously, for the variational formulation of problem (B) a minimizing sequence of allowable functions is any sequence of functions, $w_N(x,y)$ in W, for which

$$\lim_{N\to\infty} J(w_N) = J(v).$$

Methods for constructing minimizing sequences for variational problems (the Ritz method, and variational difference schemes) will be pointed out, below, in this chapter.

Here we will prove only that minimizing sequences converge in the mean-square sense, together with their first derivatives, to the solutions u and v of the respective variational problems, so that their terms may be considered as approximations to these solutions. More precisely, we will prove the following two assertions.

Theorem 3. *Suppose that w is in W,* $w|_\Gamma = \phi(s)$. *Then*

$$||w - u||_W^2 \leq \alpha[I(w) - I(u)], \tag{10}$$

where is some constant completely determined by the form of the domain D, and not depending on the function w

Theorem 4. *Let w be an arbitrary function in W. Then*

$$||w - v||_W^2 \leq \beta[J(w) - J(v)], \tag{11}$$

where the constant $\beta > 0$ depends only on the form of the domain D, and on the quantity min $\sigma(A) = \sigma_0 > 0$, but not on w.

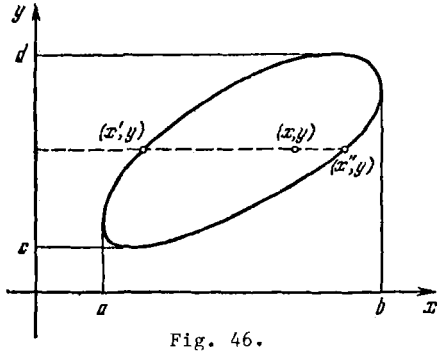


Fig. 46.

Equations (10) and (11), clearly, imply the convergence of minimizing sequences, for the variational formulation of boundary-value problems A and B, to their solutions, u and v respectively: when w is replaced by a member w_N, of the corresponding minimizing sequence, the right- and therefore also the left-hand sides of Eqs. (10) and (11) tend to zero as $N \to \infty$.

The proof of theorems 3 and 4 is based on the following lemma.

Lemma. *Suppose that $\eta(x,y)$ is in W, and $\sigma(s) \geq \sigma_0 > 0$. Then we may write the following inequality:*

$$\iint_D \eta^2 dxdy \leq \tilde{\beta}\left\{\iint_D \left[\left(\frac{\partial \eta}{\partial x}\right)^2\right] + \left(\frac{\partial \eta}{\partial y}\right)^2\right]dxdy + \int_\Gamma \sigma(s)\eta^2 ds\Big\}. \tag{12}$$

Here $\tilde{\beta}$ is a constant whose value is completely defined by the domain D and the number σ_0, and does not depend on the function $\eta(x,y)$ in W.

* * * * * *

We prove inequality (12) under the additional assumption that each line y = const intersects the boundary Γ of domain D in at most two points. This assumption is in no way essential, but considerably shortens the proof.

Suppose x and y are points interior to domain D (Fig. 46). Then

$$\eta(x,y) = \eta(x',y) + \int_{x'}^{x} \frac{\partial \eta(t,y)}{\partial t}\, dt. \tag{13}$$

Let us now square both sides of inequality (13) and use the obvious inequality $2AB \leq A^2 + B^2$, valid for any two numbers A and B:

$$\eta^2(x,y) \leq 2\left[\eta^2(x',y) + \left(\int_{x'}^{x} \frac{\partial \eta(t,y)}{\partial t}\, dt\right)^2\right]. \tag{14}$$

We now apply the Bunyakovsky inequality*

$$\left(\int_{x'}^{x} \frac{\partial\eta(t,y)}{\partial t}\,dt\right)^2 \le \left(\int_{x'}^{x} 1^2\cdot dt\right)\cdot\left[\int_{x'}^{x}\left(\frac{\partial\eta(t,y)}{\partial t}\right)^2 dt\right] \le$$

$$\le (b-a)\int_{x'}^{x}\left(\frac{\partial\eta(x,y)}{\partial t}\right)^2 dx. \qquad (15)$$

Combining (14) and (15) we get

$$\eta^2(x,y) \le 2\left[\eta^2(x',y) + (b-a)\int_{x'}^{x''}\left(\frac{\partial\eta(x,y)}{\partial x}\right)^2 dx\right]. \qquad (16)$$

Integrating both sides of (16) over x from $x = x' = x'(y)$ to $x = x'' = x''(y)$, and using the fact that the right hand side does not depend on x:

$$\int_{x'}^{x''}\eta^2(x,y)dx \le 2(x''-x')\left[\eta^2(x',y) + (b-a)\int_{x'}^{x''}\left(\frac{\partial\eta(x,y)}{\partial x}\right)^2 dx\right] \le$$

$$\le 2(b-a)\left[\eta^2(x',y) + (b-a)\int_{x'}^{x''}\left(\frac{\partial\eta(x,y)}{\partial x}\right)^2\right]. \qquad (17)$$

Now we integrate both sides of inequality (17) over y, from $y = c$ to $y = d$ and get

$$\iint_D \eta^2 dxdy \le 2(b-a)\left[\int_c^d \eta^2(x',y)dy + (b-a)\iint_D\left(\frac{\partial\eta}{\partial x}\right)^2 dx\,dy\right]. \qquad (18)$$

Clearly

$$\int_c^d \eta^2(x',y)dy \le \int_\Gamma \eta^2(s)dx \le \frac{1}{\sigma_0}\int_\Gamma \sigma(s)\eta^2 ds; \qquad (19)$$

$$\iint_D\left(\frac{\partial\eta}{\partial x}\right)^2 dx\,dy \le \iint_D\left[\left(\frac{\partial\eta}{\partial x}\right)^2 + \left(\frac{\partial\eta}{\partial y}\right)^2\right]dx\,dy. \qquad (20)$$

From (18) and (19) it follows that

---

*Usually called "Schwarz's inequality" in English. (Translator's note.)

$$\iint_D \eta^2\,dx\,dy \le$$

$$\le 2(b-a)\Big[\frac{1}{\sigma_0}\int_\Gamma \sigma\eta^2 ds + (b-a)\iint_D \Big[\Big(\frac{\partial\eta}{\partial x}\Big)^2 + \Big(\frac{\partial\eta}{\partial y}\Big)^2\Big]dx\,dy\Big\} \le$$

$$\le 2(b-a)\max\Big[\frac{1}{\sigma_0},\ b-a\Big]\cdot$$

$$\cdot \Big\{\iint_D \Big[\Big(\frac{\partial\eta}{\partial x}\Big)^2 + \Big(\frac{\partial\eta}{\partial y}\Big)^2\Big]dx\,dy + \int_\Gamma \sigma(s)\eta^2 ds\Big\},$$

which is, in fact, inequality (12) with $\tilde{\beta}$ taken as $\tilde{\beta} = 2(b-a)\max\big[(1/\sigma_0),b-a\big]$.

* * *

Inference. *Suppose* $\xi(x,y)$ *is in* W, *and* $\xi|_\Gamma = 0$. *Then the Friedrichs inequality*

$$\iint_D \xi^2 dxdy \le \tilde{\alpha}\iint_D \Big[\Big(\frac{\partial\xi}{\partial x}\Big)^2 + \Big(\frac{\partial\xi}{\partial y}\Big)^2\Big]dx\,dy, \tag{21}$$

is valid, with $\tilde{\alpha} = 2(b-a)^2$.

* * * * * *

Proof. Set $\sigma_s \equiv \sigma_0 = 1/(b-a)$. For a function $\xi(x,y) \equiv \eta(x,y)$ in W, satisfying the auxiliary condition $\xi|_\Gamma = y|_\Gamma = 0$, inequality (12) takes the form (21), where $\tilde{\alpha} = \tilde{\beta} = 2(b-a)^2$.

* * *

Proof of Theorem 3. For each function w in W, $w|_\Gamma = \phi(s)$, the function $\xi = w - u$ satisfies the conditions which allow us to deduce Eq. (21) as an inference from the lemma. Taking account of (4)

$$\iint_D \Big[\Big(\frac{\partial\xi}{\partial x}\Big)^2 + \Big(\frac{\partial\xi}{\partial y}\Big)^2\Big]dx\,dy = I(w) - I(u), \tag{4'}$$

one can write Eq. (21) in the form

$$\iint_D (w-u)^2 dx\,dy \le \tilde{\alpha}[I(w) - I(u)]. \tag{21'}$$

Adding Eq. (4') to inequality (21') one gets Eq. (10), with $\alpha = \tilde{\alpha} + 1$. Theorem 3 is proven.

Proof of Theorem 4. For any function w in W, $\eta = w - v$ satisfies the conditions of the lemma, and thus also inequality (12). Taking into account Eq. (8)

$$\int\int_D \left[\left(\frac{\partial \eta}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial y}\right)^2\right]dx\,dy + \int_\Gamma \sigma(s)\eta^2 ds = J(w) - J(v), \qquad (8')$$

Equation (12) can be written in the form

$$\int\int_D (w - v)^2 dx\,dy \leq \tilde{\beta}[J(w) - J(v)]. \qquad (12')$$

Adding (8′) and (12′) term by term, and discarding from the left-hand side the non-negative term $\int_\Gamma \sigma(s)\eta^2 ds$, we get inequality (11) with the constant $\beta = \tilde{\beta} + 1$. Theorem 4 is proven.

3. The variational method of Ritz. From theorems 3 and 4, by virtue of inequalities (10) and (11), it follows that one may take, as approximations to solutions u and v of boundary-value problems (A) and (B) those functions which, among all admissable (w in W and $w|_\Gamma = \phi(s)$ for problem (A) and w in W for problem (B)), give to the functionals I(w) and J(w) values close to the minima, I(u) and J(u), over the corresponding classes of allowable functions.

For the actual determination of approximate solutions Ritz proposed, in 1908, a method which we present, first, as applied to problem (A). For convenience in presentation we will assume that, in the boundary condition of A, $\phi(s) = 0$, so that $u|_\Gamma = 0$. The general case, $\phi(s) \neq 0$, reduces to this one if we go over to a new unknown function $\tilde{u}$, $u = \tilde{u} + h$, where $h(x,y)$ is any twice differentiable function satisfying the boundary condition $h|_\Gamma = \phi(s)$. The formal scheme for finding an approximate solution by the Ritz method consists of the following steps. Designate by $\overset{0}{W}$ the linear space of all functions w in W satisfying the boundary condition $w|_\Gamma = 0$. Choose a positive integer N, and any N linearly independent function

$$\omega_1^N(x,y),\ \omega_w^N(x,y),\ \ldots,\ \omega_N^N(x,y), \qquad (22)$$

satisfying the condition

$$\omega_n^N\Big|_\Gamma = 0, \qquad n = 1, 2, \ldots, N. \qquad (23)$$

Consider, now, the N-dimensional linear space, $\overset{0}{W}{}^N$, of all possible linear combinations of functions (22)

$$w_N(x, y, a_1, \ldots, a_N) = \sum_{n=1}^{N} a_n \omega_n^N,$$

where $a_1, \ldots, a_N$ are arbitrary real numbers.

We now seek, in place of a function $w \equiv u$ minimizing the functional $I(w)$ over the space W, a function $w_N(x, y, a_1, \ldots, a_N)$ minimizing the functional $I(w)$ on the set of all functions in the N-dimensional space ${}_0W_N$. It is this function $w_N(x, y, a_1, \ldots, a_N) \equiv \bar{w}_N(x,y)$ which we will take as the approximate solution for the given choice of N basis functions (22). The problem of determining the function $\bar{w}_N(x,y)$ is incomparably simpler than that of evaluating the desired exact solution $u(x,y)$.

In fact

$$I[w_N(x, y, a_1, \ldots, a_N)] \equiv \int\int_D \left[\left(\frac{\partial}{\partial x}\sum_{n=1}^{N} a_n w_n^N\right)^2 + \right.$$

$$\left. + \left(\frac{\partial}{\partial y}\sum_{n=1}^{N} a_n \omega_n^N\right)^2\right] dx\, dy + 2\int\int_D f \sum_{n=1}^{N} a_n \omega_n^N \, dx\, dy, \qquad (24)$$

and we are now looking for N numbers $a_1, \ldots, a_N$, which minimize the function $I[w_N(x, y, a_1, \ldots, a_N)]$ of N variables. We show next that such a set of numbers $a_1, \ldots, a_N$ exists. The first term on the right-hand side of expression (24) is a quadratic form in $a_1, \ldots, a_N$. In view of the linear independence of function system (22), this form for $a_N \neq 0$ must be strictly positive, since in the contrary case it would, for some set of $\tilde{\alpha}_N \neq 0$, be equal to zero and we would have, by virtue of (21)

$$\int\int_D \left(\sum_{n=1}^{N} \tilde{a}_n \omega_n^N\right)^2 dx\, dy \leq$$

$$\leq \tilde{\alpha}\int\int_D \left[\frac{\partial}{\partial x}\left(\sum_{n=1}^{N} \tilde{a}_n \omega_n^N\right)^2 + \frac{\partial}{\partial y}\left(\sum_{n=1}^{N} \tilde{a}_n \omega_n^N\right)^2\right] dx\, dy = 0,$$

from which it follows, despite the linear independence of the ω_n^N's, that

$$\sum_{n=1}^{N} \tilde{a}_n \omega_n^N(x,y) \equiv 0.$$

Because of the proven positive-definiteness of the quadratic form expression (24) has a unique minimum. This minimum is attained for those values $a_n = \tilde{a}_n$, $n = 1, \ldots, N$, for which

$$\frac{\partial I[w_N(x, y, a_1, \ldots, a_N)]}{\partial a_n} = 0, \qquad n = 1, \ldots, N. \qquad (25)$$

In detail, the linear system of equations (25) for the numbers $a_1, \ldots, a_N$ can be written in the form

$$\sum_{i=1}^{N} a_i \int\int_D \left[\frac{\partial \omega_i^N}{\partial x}\frac{\partial w_n^N}{\partial x} + \frac{\partial \omega_i^N}{\partial y}\frac{\partial w_n^N}{\partial y}\right] dx\, dy =$$

$$= -\int\int_D f\omega_n^N\, dx\, dy, \qquad n = 1, \ldots, N. \tag{26}$$

So as to abbreviate notation, and for convenience below we consider, along with the normed space W, the linear space $\tilde{W}$ consisting of the same functions as in W, but with the scalar product (w', w'')

$$(w', w'') \equiv \int\int_D \left(\frac{\partial w'}{\partial x}\frac{\partial w''}{\partial x} + \frac{\partial w'}{\partial y}\frac{\partial w''}{\partial y}\right)dx\, dy + \int_\Gamma \sigma(s) w' w''\, ds, \tag{27}$$

where $\sigma(s) \geq \sigma_0 > 0$ is some given function. This scalar product induces a norm $||w||_{\tilde{W}}$ in the space $\tilde{W}$ via the expression

$$||w||_{\tilde{W}}^2 = (w,w). \tag{28}$$

We designate by $\overset{0}{\tilde{W}}$ the subspace of functions w in $\tilde{W}$ satisfying the condition $w|_\Gamma = 0$.

After the introduction of the scalar product system (26), thanks to the condition $w_n^N\big|_\Gamma = 0$, takes the form

$$\sum_{i=1}^{N} a_i (\omega_i^N, \omega_n^N) = -\int\int_D f\omega_n^N dx\, dy, \qquad n = 1, \ldots, N. \tag{29}$$

Note that the matrix of system (29)

$$\omega^N = \left|\left| \begin{matrix} (\omega_1^N, \omega_1^N) & \cdots & (\omega_1^N, \omega_N^N) \\ \cdots & \cdots & \cdots \\ (\omega_N^N, \omega_1^N) & \cdots & (\omega_N^N, \omega_N^N) \end{matrix} \right|\right|$$

is the Gram matrix of a system of linearly independent functions (22). From a standard course in linear algebra we know that the determinant of this matrix is different from zero.

The solution $a_n = \tilde{a}_n$, $n = 1, \ldots, N$, of system (29) now provides the function

$$\bar{w}_N(x,y) = w_N(x, y, \tilde{a}_1, \ldots, \tilde{a}_N),$$

which one takes as the approximate solution. This function $\bar{w}_N$ allows a simple geometric interpretation.

From (4) and (27) we have

$$I(w_N) - I(u) = (w_N - u,\ w_N - u).$$

Further

$$I(\bar{w}_N) - I(u) = \min_{w \text{ in } \overset{0}{W}{}^N_2} [I(w) - I(u)] =$$

$$= \min_{w_N \text{ in } \overset{0}{W}{}^N_2} (w - u,\ w - u) = (\bar{w}_N - u,\ \bar{w}_N - u).$$

Thus $\bar{w}_N$ is that element of the linear N dimensional space w_N which, of all elements erected on the basis (22) (i.e. in the "span" of the basis functions), deviates least from u in the sense of norm (28), i.e. $\bar{w}_N$ is the projection of the solution u onto the subspace $\overset{0}{W}{}^N_2$ in the sense of the scalar product (27). At this point we have completed our formal presentation of the Ritz scheme for determining approximate solutions.

Let us now see what determines how close the approximate solution

$$\bar{w}_N = w_N(x,\ y,\ \bar{a}_1,\ \ldots,\ \bar{a}_N),$$

computed by the Ritz method, comes to the exact solution of problem (A), in which we have adopted the assumption that $\phi(s) = 0$. Naturally the quantity $||\bar{w}_N - u||_W$ depends upon the choice of basis functions (22). If, for example, the basis functions (22) had been chosen in such a way (by an improbable accident) that the function u turned out to be one of the functions of the N dimensional space $\overset{0}{W}{}^N_2$ lying in the span of the basis (22), then the approximate solution, $\bar{w}_N$, would coincide with the exact solution u. In fact

$$I(\bar{w}_N) - I(u) = \min_{w \text{ in } \overset{0}{W}{}^N_2} (I(w) - I(u)) = I(u) - I(u) = 0$$

and by theorem 3

$$||\bar{w}_N - u||_W \leq \alpha(I(\bar{w}_N) - I(u)) = 0.$$

But the function u is unknown to us, and we know only certain of its properties, properties not peculiar to it alone but to a whole class U of functions. Suppose, for example, we know that the second derivatives of the function u are continuous, and bounded by the constant M. Then the class U consists of all twice-continuously-differentiable functions the second derivatives of which don't exceed M, and which satisfy the condition $u|_\Gamma = 0$.

We recall that, for the solution u and any w in $\overset{0}{W}$

$$I(w) - I(u) = (w - u,\ w - u) = ||w - u||^2_{\widetilde{W}}$$

and that by theorem 3

$$||w - u||^2_{\widetilde{W}} \leq \alpha ||w - u||^2_{\widetilde{W}}.$$

Therefore, insofar as possible, the basis functions must be chosen in such a way that, for each function v in U, with U in $\overset{0}{\widetilde{W}}$, there should be a function w_N in $\overset{0}{\widetilde{W}}^N$ close to it, i.e., a function for which $||w_N - v||_{\widetilde{W}}$ is small. Then, in particular, the quantity

$$||\bar{w}_N - u||^2_{\widetilde{W}} = \min_{w \text{ in } \overset{0}{\widetilde{W}}^N} [I(w) - I(u)] = \min_{w \text{ in } \overset{0}{\widetilde{W}}^N} (w - u,\ w - u),$$

will be small and, along with it, also the quantity $||\bar{w}_N - u||_W$:

$$||\bar{w}_N - u||^2_W \leq \alpha ||\bar{w}_N - u||^2_{\widetilde{W}}.$$

More precisely, the best set of functions (22) would be one for which the quantity

$$K_N = K_N(U, \overset{0}{\widetilde{W}}^N) = \sup_{v \text{ in } U} \min_{w \text{ in } \overset{0}{\widetilde{W}}^N} ||w - v||_{\widetilde{W}} \tag{30}$$

is as small as possible. We designate by $\kappa_N(U, \overset{0}{\widetilde{W}})$ the number

$$\kappa_N(U, \overset{0}{\widetilde{W}}) = \inf_{\overset{0}{\widetilde{W}}^N \text{ in } \overset{0}{W}} K_N(U, \overset{0}{\widetilde{W}}^N) = \inf_{\omega_1^N, \ldots, \omega_N^N} \sup_{v \text{ in } U} \min_{w \text{ in } \overset{0}{W}^N} ||w - v||_{\widetilde{W}}. \tag{31}$$

This number is called *the N-dimensional Kolmogorov diameter of the class of functions U with respect to the normed space* $\overset{0}{\widetilde{W}}$ *in* $\widetilde{W}$. Clearly the optimum choice of functions (22) would be one for which the quantity (30) coincides with $\kappa_N(U, \overset{0}{\widetilde{W}})$, the diameter of A. N. Kolmogorov. For any $\varepsilon > 0$ there exists, obviously, a set of basis functions (22) for which

$$I(\bar{w}_N) - I(u) = ||\bar{w}_B - u||^2_{\widetilde{W}} \leq$$

$$\leq \sup_{\tilde{u} \text{ in } U} \inf_{w \text{ in } \overset{0}{\widetilde{W}}^N} ||w - \tilde{u}||^2_{\widetilde{W}} = K_N^2(U, \overset{0}{\widetilde{W}}^N) \leq \kappa_N^2(U, \overset{0}{\widetilde{W}}) + \varepsilon$$

* * * * * *

The N-dimensional Kolmogorov diameter, $\kappa_N(X,Y)$, of a set X lying in linear normed space Y, with respect to this space, is defined by the expression

$$\kappa_N(X,Y) = \inf_{Y^N \text{ in } Y} \sup_{x \text{ in } X} \min_{y \text{ in } Y^N} ||y - x||_Y,$$

where Y_N is an arbitrary, given, N-dimensional linear manifold ("hyperplane").

Diagonals have been computed in many cases. In particular, it is known that, for the class of all functions v, $v|_\Gamma = 0$, whose second derivatives, in some domain, are continuous, and all bounded by one single constant,

$$\kappa_N(U, \overset{0}{W}) = O\left(\frac{1}{\sqrt{N}}\right), \tag{32}$$

$$\kappa_N(U, \tilde{W}) = O\left(\frac{1}{\sqrt{N}}\right). \tag{33}$$

Taking into account additional information about the desired solution u, obtained in a preliminary analysis of the problem, or as a result of experience in solving similar problems, one may narrow the class U, and as a result the diameter $\kappa_N(U, \overset{0}{\tilde{W}})$, N = 1, 2, ... can only decrease.

Thus the skill and experience of the analyst are manifested here through his ability to choose a narrow class, U, containing the required solution u and then, for a given N, to choose basis functions (22) in such a way that the number $K_N(U, \overset{0}{\tilde{W}}^N)$, introduced via Eq. (30), will not be much larger than the N-dimensional diameter $\kappa_N(U, \overset{0}{\tilde{W}})$. Then on the right-hand side of the equation

$$||\bar{w}_N - u||^2_W \leq \alpha[I(\bar{w}_N) - I(u)] = \alpha||\bar{w}_N - u||^2_W \leq \alpha K^2_N(U, \overset{0}{\tilde{W}}^N)$$

we will have a number, close to $\kappa^2_N(U, \overset{0}{\tilde{W}})$, which tends to zero all the more quickly, with increasing N, the narrower the class U. If one takes full enough account of special features of the solution u, known prior to the computation, and then, correspondingly, makes a good choice of basis functions, then a sufficiently accurate solution may be obtained even for a small N. But the volume of computational work, work which consists in the computation of coefficients and solution of system (26), depends precisely on N. Thus the computational algorithm will then be very efficient.

Let us now illustrate the Ritz method with still another example: consider problem (B). After a system of basis functions (22) has been chosen, we look for an approximate solution

$$w_N(x, y, a_1, \ldots, a_N) = \sum_{n=1}^{N} a_n \omega_n^N(x,y)$$

in the space W^N of all linear combinations, choosing constants such that the expression

$$J[w_N(x, y, a_1, \ldots, a_N)]$$

be minimized. Minimizing constants $a_1, \ldots, a_N$ must be determined from the system of equations

$$\frac{\partial J[w_N(x, y, a_1, \ldots, a_N)]}{\partial a_n} = 0, \qquad n = 1, \ldots, N. \tag{34}$$

We will assume that in the definition (27) of scalar multiplication the function $\sigma(s)$ coincides with the corresponding function which appears in the boundary condition of problem (B). Then the system of equations (34) takes the form

$$\sum_{i=1}^{N} a_i(\omega_i^N, \omega_N^N) = -\int\int_D f\omega_n^N dx\, dy + \int_\Gamma \phi(s)\omega_n^N ds, \quad n = 1, \ldots, N. \tag{35}$$

The solution of this system $a_n = \bar{a}_n$, $n = 1, \ldots, N$ then gives exactly the desired approximate solution $\bar{w}_N$,

$$\bar{w}_N(x, y) = \sum_{n=1}^{N} \bar{a}_n \omega_n^N(x,y).$$

For a function $\bar{w}_N$ in W, and the solution v of problem (B), we get from Eq. (8)

$$J(\bar{w}_N) - J(v) = (\bar{w}_N - v, \bar{w}_N - v) \leq \max_{\tilde{v} \text{ in } U} \min_{w \text{ in } W^N} (w - \tilde{v}, w - \tilde{v}),$$

where U is the class of functions containing the solution v of problem (B). From the last inequality it is clear that the basis functions $\omega_1^N, \ldots, \omega_N^N$ must be chosen in such a way that the right-hand side of this inequality will be as small as possible. In this case it is not necessary, as it was in the previous example, to subject the basis functions to any sort of boundary conditions.

4. Projection method of Galerkin. B. G. Galerkin, in 1916, proposed a computational method for the solution of boundary-value problems, a method which did not require that the problem to be treated should have a known variational formulation. We present this method via the example of boundary-value problem (A) assuming, moreover, as in Sect. 3 above, that

$$\left.\begin{aligned} &\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - f(x,y) = 0, \qquad (x, y) \text{ in } D, \\ &u\big|_\Gamma = 0. \end{aligned}\right\} \tag{36}$$

Again we choose a system of basis functions (22), but now we (temporarily) stipulate that the functions $\omega_n^N(x,y)$ have continuous second derivatives. Again we seek an approximate solution in the form of a linear combination

$$w_N(x, y, a_1, \ldots, a_N) = \sum_{n=1}^{N} a_n \omega_n^N(x,y). \tag{37}$$

Now we substitute Eq. (37) into the left-hand side of the equation and boundary condition of (36) and get

$$\frac{\partial^2}{\partial x^2}\left[w_N(x, y, a_1, \ldots, a_N)\right] + \frac{\partial^2}{\partial y^2}\left[w_N(x, y, a_1, \ldots, a_N)\right] - f(x, y) = \delta_N(x, y, a_1, \ldots, a_N)$$

$$w_N\big|_\Gamma = 0$$

where $\delta_N(x, y, a_1, \ldots, a_N)$ is the resulting residual.

Let us now introduce in space $\tilde{W}$, along with the scalar product (w', w'') defined above, another scalar product

$$[w', w''] = \int_D \int w'w'' \, dx \, dy.$$

If it turned out that δ_N were orthogonal to all the functions in $\tilde{W}$, in the sense of this scalar product, then $\delta_N(x, y, a_1, \ldots, a_N)$ would vanish identically, and w_N would be the exact solution. But the number of parameters $a_1, \ldots, a_N$ is too small to allow us to construct an exact solution by adjusting these constants. Therefore we will choose them, instead, from the condition that the projection of the residual on all the ω_n^N, $n = 1, \ldots, N$, be equal to zero, i.e. that the residual be orthogonal to all the basis functions (22)

$$[\delta_N, \omega_n^N] = 0, \qquad n = 1, \ldots, N. \tag{38}$$

In expanded form the system of equations (38) for the numbers $a_1, \ldots, a_N$, may be written thus:

$$\int_D \int \left(\frac{\partial^2 w_N}{\partial x^2} + \frac{\partial^2 w_N}{\partial y^2}\right)\omega_n^N \, dx \, dy = \int_D \int f\omega_n^N \, dx \, dy, \qquad n = 1, \ldots, N. \tag{39}$$

Integrating by parts we see that, thanks to the condition $\omega_n^N\big|_\Gamma = 0$,

$$\iint_D \left(\frac{\partial^2 w_N}{\partial x^2} + \frac{\partial^2 w_N}{\partial y^2} \right) \omega_n^N \, dx \, dy =$$

$$- \iint_D \left(\frac{\partial w_N}{\partial x} \frac{\partial \omega_n^N}{\partial x} + \frac{\partial w_N}{\partial y} \frac{\partial \omega_n^N}{\partial y} \right) dx \, dy =$$

$$= - \sum_{i=1}^{N} a_i \iint_D \left(\frac{\partial \omega_i^N}{\partial x} \frac{\partial \omega_n^N}{\partial x} + \frac{\partial \omega_i^N}{\partial y} \frac{\partial \omega_n^N}{\partial y} \right) dx \, dy =$$

$$= - \sum_{i=1}^{N} a_i \left(\omega_i^N, \omega_n^N \right).$$

Thus system (39) may be rewritten in the form

$$\sum_{i=1}^{N} a_i \left(\omega_i^N, \omega_n^N \right) = - \left[f, \omega_n^N \right], \qquad n = 1, \ldots, N \tag{40}$$

and, for the given choice of scalar product $[\cdot]$, exactly agrees with the system (29) obtained by the Ritz method.

The additional assuption that the basis functions have second derivatives can here be dropped, since the Galerkin equations (40) retain their meaning even without this requirement.

5. Methods for solving the algebraic system. For not very large N(N ~ 100), the Ritz or Galerkin equation-sets can be solved exactly by existing standard codes for systems of linear equations. Further the matrix ω^N of Ritz system (29) in our example (and this is typical) is a Gram matrix for the basis system ω_n^N, n = 1, ..., N. Obviously it is symmetric, and it is known to be positive definite. Therefore to compute the solution of Ritz system (29) one can make use of any one of a number of iterative methods like, for example, iteration with Chebyshev parameters.

Iterative methods become considerably simpler if only a few of the elements of the matrix ω^N are different from zero. We will see that, in the variational-difference and projection-difference schemes precisely this is true.

6. Computational stability. We have seen that the accuracy of the approximate solution, for a given number N of basis functions ω_n^N, n = 1, ..., N, depends on how well one can approximate the solution with elements of the N dimensional linear space formed by the span of these basis functions. Thus the accuracy depends on the choice of an approximating space, but not on the basis used in this space.

The stability, i.e the conditioning properties, of the equation system (29) of the Ritz method, or of Galerkin-method system (40), depends on the conditioning of the matrix ω^N of the equation system. From the point of

view of stability it would be ideal if the functions ω_n^N, $n = 1, \ldots, N$, formed an orthonormal basis. In that case the matrix $\boldsymbol{\omega}^N$ would be unitary.

PROBLEMS

1. Show that the solution of the following first boundary-value problem for elliptic equations with variable coefficients

$$\frac{\partial}{\partial x}\left[a(x,y)\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[b(x,y)\frac{\partial u}{\partial y}\right] = f(x,y),$$

$$u\big|_\Gamma = \phi(s),$$

$$a(x,y) \geq a_0 > 0, \quad b(x,y) \geq b_0 > 0$$

minimizes the functional

$$I(w) = \iint_D \left[a\left(\frac{\partial w}{\partial x}\right)^2 + b\left(\frac{\partial w}{\partial y}\right)^2 + 2fw\right]dx\,dy$$

over the class of all functions w in W satisfying the auxiliary boundary condition $w\big|_\Gamma = \phi(s)$.

Assume that the solution u(x, y) has continuous second derivatives.

2. Given a system of basis functions $\omega_1^N, \ldots, \omega_N^N$, write out the system of Ritz equations for computation of the solution, u(x, y), of the above problem 1 with $\phi(s) \equiv 0$.

3. Show that the solution of the third boundary-value problem for the elliptic equation with variable coefficients

$$\frac{\partial}{\partial x}\left[a(x,y)\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[b(x,y)\frac{\partial u}{\partial y}\right] = f(x,y)$$

$$\frac{\partial u}{\partial n} + \sigma(s)u\Big|_\Gamma = \phi(s)$$

minimizes the functional

$$J(w) = \iint_D \left[a\left(\frac{\partial w}{\partial x}\right)^2 + b\left(\frac{\partial w}{\partial y}\right)^2 + 2wf\right]dx\,dy + \int_\Gamma (\sigma w^2 - 2\phi w)ds$$

over the set of all functions w in W.

4. Given a system of basis functions $\omega_1^N, \ldots, \omega_N^N$, write out the system of Ritz equations for computation of the solution u(x,y) of problem 3.

5. Given a system of basis functions $\omega_1^N, \ldots, \omega_N^N$, write out the system of Galerkin equations for the first boundary-value problem.

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + a(x,y)\frac{\partial u}{\partial x} + b(x,y)\frac{\partial u}{\partial y} - c^2(x,y)u = f,$$

$$u|_\Gamma = 0.$$

§39. Construction and properties of variational-difference and projection-difference schemes

1. Definition of variational-difference and projection-difference schemes. Suppose that, in the closed region, D, in which we want to solve some variational problem for each N of a monotone increasing sequence of natural numbers, we are given N points P_1^N, P_2^N, ..., P_N^N. The totality of these points will be called the "net corresponding to the given N." Suppose, further, that the Ritz method for solving the variational problem makes use of a system of basis functions

$$\omega_1^N(x,y),\ \omega_2^N(x,y),\ \ldots,\ \omega_n^N(x,y),\ \ldots,\ \omega_N^N(x,y),$$

the n'th member of which (i.e. $\omega_n^N(x,y)$) takes on, at P_n^N, the value unity, vanishing at all other points of the net:

$$\omega_k^N(P_k^N) = \delta_n^k, \quad n, k = 1, 2, \ldots, N. \tag{1}$$

In this case the linear combination

$$w_N(x, y, a_1, \ldots, a_N) = \sum_{n=1}^{N} a_n \omega_n^N(x,y)$$

takes on at point P_n^N the value $w_N(P_n^N) = a_n$, n = 1, ..., N. Therefore one may write

$$w_N(x,y) = \sum_{n=1}^{N} w_N(P_n^N)\omega_n^N(x,y).$$

The system of Ritz equations for the determination of coefficients, a_1, ..., a_N, such that this linear combination minimizes the variational functional over the linear space generated by the basis functions ω_1^N, ..., ω_N^N will, thus, connect the values, $w_N(P_n^N)$, n = 1, ..., N, of the solution function itself at the points of the given net P_1^N, ..., P_N^N: i.e. the Ritz equations turn out to be a sort of difference scheme. This difference scheme, in conformity with the method by which it was constructed, is called a "variational-difference" scheme.

Correspondingly if, to implement the Galerkin projection method, one uses basis functions ω_1^N, ..., ω_N^N, satisfying condition (1), then the Galerkin method becomes a sort of difference scheme, which it is natural to call a "projection-difference" scheme.

So that the reader will more easily visualize what has been done it may be worthwhile to make the following remarks. For given values $w_N(P_n^N)$, n = 1, ..., N, the linear combination

$$w_N(x,y) = \sum_{n=1}^{N} w_N(P_n^N)\omega_n^N(x,y)$$

can be understood as an expression which completes the definition of, or "fills in" the function $w_N(x,y)$ everywhere in domain D, according to its values $w_N(P_n^N)$, n = 1, ..., N, at the net-points. Clearly neither the choice of the net P_n^N, n = 1, ..., N for a given N, nor the choice of a system of basis functions ω_1^N, ..., ω_N^N, satisfying condition (1) and defining a method for filling-in the net function, is unique. Thus, for example, in the one-dimensional case the function might be completed in the interval, according to its net-point values, piecewise-linearly, or quadratically, or by Lagrange interpolation, etc. On the choice of net P_n^N, and of the basis functions, depends the form and properties of the resulting variational-difference or projection-difference scheme for the given variational or differential boundary-value problem.

Let us now consider examples of variational-difference schemes for problem (A) and (B) of §38. We will assume, here, that region D, in which we want solutions, is convex. (A region, D, is called "convex" if, for any two points P and P′ in D, all points on the line, PP′, with end-points at P and P′, also belong to D).

The assumption that D is convex is not at all essential, but simplifies our presentation.

2. Example of a variational-difference scheme for the first boundary-value problem. Choose a positive integer N. Next inscribe in contour Γ, bounding region D, a non-intersecting polygon $Q_1^N Q_2^N \ldots Q_m^N Q_1^N$ with vertices at points Q_1^N, ..., Q_m^N. Call this polygon D_N. Divide the polygon D_N into triangles in such a way that (a) each segment of its broken-line perimeter will be a side of one of the triangles, (b) that each pair of triangles either has no points in common, shares a vertex, or shares a side, and (c) that the total number of vertices P_1^N, ..., P_N^N of these triangles lying inside the polygon D_N should be equal to N. The set of points P_1^N, ..., P_N^N will, then, serve as our net (Fig. 47). Now we construct a system of basis functions ω_1^N, ..., ω_N^N. We define the basis function, $\omega_n^N(x,y)$, n = 1, ..., N as follows. First we assign function-values at the net points via Eq. (1)

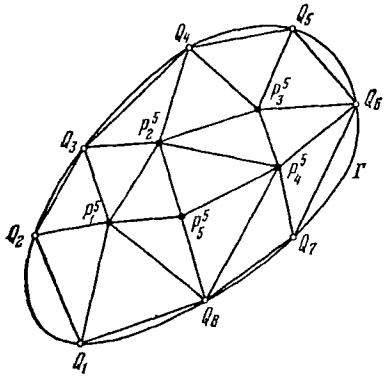


Fig. 47

$$\omega_n^N(P_k^n) = \delta_n^k, \qquad n, k = 1, \ldots, N.$$

Then we fix these values at the points $Q_1^N, \ldots, Q_m^N$, assuming them all to vanish at these points. Thus each function is already defined at the vertices of all triangles constituting the decomposition D_N. In each of these triangles we then complete the definition of the set of basis functions, taking them all to be linear. It remains for us to define them in the region $D \setminus D_N$, where we will set them to zero.

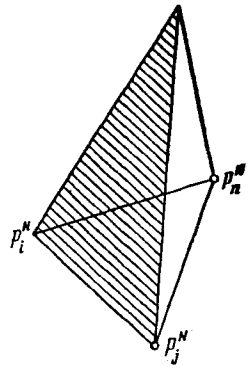


Fig. 48.

Note that, for any triangle that has no point P_n^N as any of its vertices the function $\omega_n^N(x,y)$, as we have constructed it, will vanish. In a triangle with a vertex at point P_n^N the function $\omega_n^N = w_n^N(x,y)$ appears, in the space xyw, as a section of a plane (Fig. 48) passing through the side lying opposite the vertex P_n^N, and upraised to unit height above the point P_n^N. The system of Ritz equations, (29) §38, determining the coefficients $a_n = \bar{w}_N(P_n^N)$ in the approximate solution

$$w_N = \sum_{i=1}^{N} w_N(P_i^N)\omega_i^N(x,y),$$

has the form

$$\sum_{i=1}^{N} w_N(P_i^N)(\omega_i^N, \omega_n^N) = -\int\int_D f\omega_n^N \, dx\, dy, \qquad (2)$$

$$n = 1, \ldots, N.$$

This is precisely the variational-difference scheme corresponding to the above choice of net and of basis functions.

The matrix of this difference scheme

$$\boldsymbol{\omega}^N = \left|\left|(\omega_n^N, \omega_i^N)\right|\right|, \qquad n, i = 1, \ldots, N,$$

has, as elements, the quantities

$$(\omega_n^N, \omega_i^N) = \int\int_D \left(\frac{\partial\omega_n^N}{\partial x}\frac{\partial\omega_i^N}{\partial x} + \frac{\partial\omega_n^N}{\partial y}\frac{\partial\omega_i^N}{\partial y}\right) dx\, dy. \qquad (3)$$

Obviously only those products (ω_n^N, ω_i^N) can differ from zero for which the points P_n^N and P_i^N are vertices of one and the same decomposition triangle. In fact if P_n^N and P_i^N are not, in this sense, neighboring net-points then

regions in which $\omega_n^N \not\equiv 0$ and $\omega_i^N \not\equiv 0$ will not intersect, and therefore the integrand in Eq. (3) will vanish identically everywhere in the range of integration.

Thus the n'th equation of the set which constitutes variational-*difference scheme* (2) *connects values of the unknown function at point* P_n^N with other values of this function only at neighboring points.

The computation of coefficients via Eq. (3) presents no difficulties. In fact the coefficients (ω_n^N, ω_i^N) are integrals of the quantity

$$\frac{\partial \omega_n^N}{\partial x}\frac{\partial \omega_i^N}{\partial x} + \frac{\partial \omega_n^N}{\partial y}\frac{\partial \omega_i^N}{\partial y}. \tag{4}$$

over a pair of subdivision triangles, triangles with the segment $P_n^N P_i^N$ as their common side. Further, the integral over any one of these triangles is completely determined by the lengths of its sides, and does not depend on its orientation or location. In fact the quantity (4), constant over the triangle, is the product of the lengths of the vectors grad ω_n^N and grad ω_i^N, multiplied by the cosine of the angle between these vectors, and therefore may be expressed in the form

$$\frac{1}{h_n h_i}\cos(\bar{h}_n, \bar{h}_i). \tag{5}$$

Here h_n and h_i are the lengths of perpendiculars drawn from the vertices P_n^N and P_i^N respectively to the surfaces $\omega = \omega_n^N(x,y)$ and $\omega = \omega_i^N(x,y)$, while $\bar{h}_n$ and $\bar{h}_i$ are unit vectors directed along these perpendiculars, towards the respective vertices, like the vectors grad ω_n^N and grad ω_i^N. The integral over the triangle is obtained by taking the product of the quantity (5) with the area of the triangle.

The construction of variational-difference scheme (2), for a given choice of points Q_i^N and P_i^N, is completed. Clearly, however, one cannot *expect that for every choice of these points, uniquely defining a system of* basis functions ω_n^N, n = 1, ..., N, the corresponding approximate solution

$$\bar{w}_N = w_N(x, y, \bar{w}_n^N(P_1^N), \ldots, \bar{w}^N(P_N^N))$$

will be a "good approximation" to the exact solution, u(x,y), of the original problem. In fact if, for, example, all the points $Q_1^N, \ldots, Q_m^N$ and $P_1^N, \ldots, P_N^N$ were distributed over one "half" of the region, with not a single net-point in the "other half", then the resulting approximation couldn't turn out to be good. To make a good choice of boundary-points $Q_1^N, \ldots, Q_m^N$, and net-points $P_1^N, \ldots, P_N^N$, one must take account of relations discussed in §38, and reproduced here:

$$||\bar{w}_N - u||^2_{\tilde{W}} \leq \alpha||\bar{w}_N - u||^2_{\tilde{\tilde{W}}} \leq \alpha K_N^2(U, \overset{0}{\tilde{W}}{}^N) =$$

$$= \alpha \sup_{v \text{ in } U} \operatorname{Inf}_{w \text{ in } \overset{0}{\tilde{W}}{}^N} ||w - v||^2_{\tilde{\tilde{W}}} . \qquad (6)$$

In (6) $\overset{0}{\tilde{W}}{}^N$ is the N dimensional linear space generated by all possible linear combinations of the basis functions, and U is a set of functions containing the exact solution.

More precisely, it will be seen from (6) that it is adviseable to choose the points Q_1^N, ..., Q_m^N and P_1^N, ..., P_N^N with the properties of class U in mind, so that the quantity $K_N(U, \overset{0}{\tilde{W}}{}^N)$ will be "as small as practically possible", and so that, as $N \to \infty$, the sequence $K_N(U, \overset{0}{\tilde{W}}{}^N)$ will go to zero "as fast as possible".

Always $K_N(U, \overset{0}{\tilde{W}}{}^N) \geq \kappa_N(U, \overset{0}{\tilde{W}})$, where $\kappa_N(U, \overset{0}{\tilde{W}})$ is the N-dimensional Kolmogorov diameter of the set U with respect to normed space $\overset{0}{\tilde{W}}$ (see 3§38). Therefore for a choice of points to be "good" it is sufficient that $K_N(U, \overset{0}{\tilde{W}}{}^N)$ be close to $\kappa_N(U, \overset{0}{\tilde{W}})$.

* * * * * *

Generally speaking, however, it is not true that, for each set of functions U, there exists a net for which $K_N(U, \overset{0}{\tilde{W}}{}^N)$ does not "greatly" exceed $\kappa_N(U, \overset{0}{\tilde{W}}{}^N)$, so that as $N \to \infty$ the quantities $K_N(U, \overset{0}{\tilde{W}}{}^N)$ and $\kappa_N(U, \overset{0}{\tilde{W}})$ will be small of the same order in N^{-1}. The problem is that, in particular, the piecewise linear basis functions which we are using in this section to complete the net functions will, for each choice of points, generate spaces $\overset{0}{\tilde{W}}{}^N$, of piecewise-linear functions, spaces which do not exhaust all possible N dimensional subspaces of space $\overset{0}{\tilde{W}}$, and among which there may not be a subset $\overset{0}{\tilde{W}}{}^N$ which constitutes a good approximation to set U.

* * *

Let us now analyze in detail a case where all the a priori information which we have as to the properties of the solution, u, permits us to conclude only that u belongs to the class, U, of all functions, vanishing on the boundary, whose second derivatives do not exceed some number, M.

In this case we will show how one must distribute the points Q_1^N, ..., Q_m^N, P_1^N, ..., P_N^N so that, as N increases, $K_N(U, \overset{0}{\tilde{W}})$ will be of order $O(1/\sqrt{N})$. Then, thanks to Eq. (6) for the error, $\bar{w}_N - u$, in the approximate solution $\bar{w}_N$, we are assured that

$$\left.\begin{aligned} ||\bar{w}_N - u||_{\tilde{W}} &\le \frac{c}{\sqrt{N}}, \\ ||\bar{w}_N - u||_{W} &\le \frac{\alpha c}{\sqrt{N}}, \end{aligned}\right\} \tag{7}$$

where c is some constant.

* * * * * *

Note that, because of equations (32) and (33) §38 for the diameters

$$\left.\begin{aligned} \kappa_N(U, \overset{0}{\tilde{W}}) &= O\left(\frac{1}{\sqrt{N}}\right), \\ \kappa_N(U, \overset{0}{W}) &= O\left(\frac{1}{\sqrt{N}}\right), \end{aligned}\right\} \tag{8}$$

these bounds are best possible in the following sense. If we look for an approximate solution in the form of a linear combination of some given functions $\psi_1^N(x,y), \ldots, \psi_N^N(x,y)$,

$$w_N = \sum_{k=1}^{N} c_k \psi_k^N,$$

then one cannot, for any choice of functions $\psi_k^N(x,y)$, nor for any method of computing the coefficients c_k given a right-hand side f(x,y), achieve bounds of the form $||w_N - u||_{\overset{0}{W}} = o(1/\sqrt{N})$ and $||w_N - u||_{\overset{0}{\tilde{W}}} = o(1/\sqrt{N})$, valid for any f(x,y) for which the solution, u, belongs to our class U.

* * *

Theorem 1. *Let U be the set of all functions whose second derivatives are continuous and do not exceed some number, M, in modulus, functions which vanish on the boundary, Γ. Suppose that, for each N of some increasing sequence of natural numbers, one has selected points* $Q_1^N, Q_2^N, \ldots, Q_m^N$ *m = m(N), and a decomposition of the polynomial* $D_N = Q_1^N Q_2^N, \ldots, Q_m^N$ *into triangles generating, as described above, the net* $P_1^N, P_2^N, \ldots, P_N^N$. *Assume, further, that the following conditions are fulfilled:*

1°. The length ℓ, of any side of a subdivision triangle satisfies the inequality

$$\ell \le C_1 h,$$

where

$$h = \left[\frac{\text{area of } D}{N}\right]^{1/2},$$

and C_1 *is some positive number not depending on* h.

2°. *The area of the region* $D \setminus D_N$ *satisfies the bound*

$$S_N \leq C_2 h^2, \qquad C_2 = \text{const}. \tag{9}$$

3°. *Each angle* α *of any of the subdivision triangles of region* D_N *satisfies the bound*

$$\alpha > \alpha_0 = \text{const} > 0. \tag{10}$$

Under the given conditions we have, for the quantity $K_N(U, \overset{0}{\tilde{W}}{}^N)$:

$$K_N^2(U, \overset{0}{\tilde{W}}{}^N) = \sup_{v \text{ in } U} \operatorname{Inf}_{w \text{ in } \overset{0}{\tilde{W}}{}^N} \iint_D \left\{\left[\frac{\partial(w - v)}{\partial x}\right]^2 + \left[\frac{\partial(w - v)}{\partial y}\right]^2\right\} dx\, dy \tag{11}$$

the bound

$$K_N(U, \overset{0}{\tilde{W}}{}^N) \leq C_3 h, \tag{12}$$

where C_3 *is some constant.*

* * * * * *

Proof. It is sufficient to show that, for each function v in U the following function

$$w(x,y) = \sum_{k=1}^{N} v(P_k^N)\omega_k^N(x,y), \quad w \text{ in } W, \tag{13}$$

satisfies the bound

$$\iint_D \left\{\left[\frac{\partial(w - v)}{\partial x}\right]^2 + \left[\frac{\partial(w - v)}{\partial y}\right]^2\right\} dx\, du \leq C_4 h^2, \tag{14}$$

since, obviously, in this case bound (12) is valid.

The integral (14) may be written as the sum of (non-negative) integrals over the polygon D_N, inscribed in region D, and its complement $D \setminus D_N$ in the whole region D:

$$\iint_D \left\{\left[\frac{\partial(w - v)}{\partial x}\right]^2 + \left[\frac{\partial(w - v)}{\partial y}\right]^2\right\} dx\, dy =$$

$$= \iint_{D_N} \left\{\left[\frac{\partial(w - v)}{\partial x}\right] + \left[\frac{\partial(w - v)}{\partial y}\right]\right\} dx\, dy +$$

$$+ \iint_{D \setminus D_N} \left\{\left[\frac{\partial(w - v)}{\partial x}\right]^2 + \left[\frac{\partial(w - v)}{\partial y}\right]^2\right\} dx\, dy. \tag{15}$$

Let us now estimate each of the two terms on the right-hand side of (15), establishing the bounds

$$\int\int_{D_N} \{\ldots\}dx\ dy \leq A_1 h^2, \tag{16}$$

$$\int\int_{D\ D_N} \{\ldots\}dx\ dy \leq A_2 h^2, \tag{17}$$

where A_1 and A_2 are constants not depending on the function v in U, nor on h. Clearly, by virtue of (15), Eqs. (16) and (17) imply (14) with constant $C_4 = A_1 + A_2$.

To prove bound (16) it is sufficient to show that, inside each triangle making up the decomposition of region D, we have the inequality

$$\left|\frac{\partial(w - v)}{\partial x}\right| \leq Bh, \qquad \left|\frac{\partial(w - v)}{\partial y}\right| \leq Bh, \tag{18}$$

where B is a constant which depends neither on the function v in U, nor on h. Then, clearly bound (16) will be valid if, for A, we take $A_1 = 2B^2 \cdot$(area of D). Thus to complete the proof of bound (16) we need to establish bound (18), which we now set out to do. The proof of (18) will be divided into two stages. First we show that the derivative $d(w - v)/d\ell$, in any direction ℓ, can change inside a triangle by no more than $A_3 h$, where A_3 = const, so that for any two points (x',y') and (x'',y'') belonging to the triangle we may write

$$\left|\left[\frac{d(w - v)}{d\ell}\right]_{(x'',y'')} - \left[\frac{d(w - v)}{d\ell}\right]_{(x',y')}\right| \leq A_3 h. \tag{19}$$

Next we choose any two sides of this triangle forming acute angle α, and show that everywhere in the triangle the derivatives $d(w - v)d\ell_1$ and $d(w - v)/d\ell_2$, in the directions of these sides, satisfy the bounds

$$\left|\frac{d(w - v)}{d\ell_1}\right| \leq A_3 h, \qquad \left|\frac{d(w - v)}{d\ell_2}\right| \leq A_3 h. \tag{20}$$

Then we use the equations

$$\left.\begin{aligned} \frac{d(w - v)}{d\ell_1} &= \frac{\partial(w - v)}{\partial x}\cos\alpha_1 + \frac{\partial(w - v)}{\partial y}\sin\alpha_1, \\ \frac{d(w - v)}{d\ell_2} &= \frac{\partial(w - v)}{\partial x}\cos\alpha_2 + \frac{\partial(w - v)}{\partial y}\sin\alpha_2, \end{aligned}\right\} \tag{21}$$

where α_1 and α_2 are the angles which the directions ℓ_1 and ℓ_2 make with the x axis. Considering Eqs. (21) as a system of equations for $\partial(w - v)/\partial x$ and $\partial(w - v)/\partial y$, we find that

$$\left.\begin{aligned}\frac{\partial(w-v)}{\partial x} &= \frac{\sin\alpha_2}{\sin(\alpha_2-\alpha_1)}\frac{d(w-v)}{d\ell_1} - \frac{\sin\alpha_1}{\sin(\alpha_2-\alpha_1)}\frac{d(w-v)}{d\ell_2},\\ \frac{\partial(w-v)}{\partial y} &= -\frac{\cos\alpha_2}{\sin(\alpha_2-\alpha_1)}\frac{d(w-v)}{d\ell_1} + \frac{\cos\alpha_1}{\sin(\alpha_2-\alpha_1)}\frac{d(w-v)}{d\ell_2}.\end{aligned}\right\} \quad (22)$$

But by Eq. (10) the angle $\alpha = \alpha_2 - \alpha_1 \geq \alpha_0 > 0$ and $\alpha \leq \pi - 2\alpha_0$, so that $\sin\alpha \geq \sin\alpha_0 = \text{const} > 0$.

From Eq. (22) and inequality (20) one derives the bounds

$$\left|\frac{\partial(w-v)}{\partial x}\right| \leq \frac{2}{\sin\alpha_0}A_3h, \qquad \left|\frac{\partial(w-v)}{\partial y}\right| \leq \frac{2}{\sin\alpha_0}A_3h,$$

which take the form (18) if we define $B = (2/\sin\alpha_0)A_3$.

To complete the proof of bound (18) and, thus, of (16), it remains for use to prove bounds (19) and (20) on which we have relied. Let us first prove (19). Designate by S the direction from point (x',y') to point (x'',y''). On the interval joining these points any function $\psi(x,y)$ can be considered as a function of s, where s is the distance from point (x',y'). By the theorem on finite increments

$$\psi(x'',y'') - \psi(x',y') = \sqrt{(x''-x')^2 + (y''-y')^2}\,\frac{d\psi(\xi,\eta)}{ds},$$

where (ξ, η) is some point of the interval connecting points (x',y') and (x'',y''). If

$$\psi(x,y) \equiv \frac{dv(x,y)}{d\ell},$$

then

$$\frac{dv(x'',y'')}{d\ell} - \frac{dv(x',y')}{d\ell} = \sqrt{(x''-x')^2 + (y''-y')^2}\,\frac{d}{ds}\left(\frac{dv(\xi,\eta)}{d\ell}\right). \quad (23)$$

Designate the angles that directions ℓ and s, respectively, make with the x axis as α and β. Then we may write the symbolic equality

$$\frac{d}{d\ell} = \cos\alpha\frac{\partial}{\partial x} + \sin\alpha\frac{\partial}{\partial y},$$

$$\frac{d}{ds} = \cos\beta\frac{\partial}{\partial x} + \sin\beta\frac{\partial}{\partial y}.$$

Clearly

$$\frac{d}{ds}\left(\frac{d}{d\ell}\right) = \cos\alpha\cos\beta\frac{\partial^2}{\partial x^2} + [\cos\alpha\sin\beta + \sin\alpha\cos\beta]\frac{\partial^2}{\partial x\partial y} + \sin\alpha\sin\beta\frac{\partial^2}{\partial y^2}.$$

Therefore

$$\left|\frac{d}{ds}\left[\frac{dv(\xi,\eta)}{d\ell}\right]\right| =$$

$$= \left|\cos\alpha\cos\beta\,\frac{\partial^2 v(\xi,\eta)}{\partial x^2} + \sin(\alpha+\beta)\,\frac{\partial^2 v(\xi,\eta)}{\partial x\partial y} + \sin\alpha\sin\beta\,\frac{\partial^2 v(\xi,\eta)}{\partial y^2}\right| \leq 3M$$

Since $\sqrt{(x'' - x')^2 + (y'' - y')^2} \leq 2c_1 h$ we get, from (23), the inequality

$$\left|\frac{dv(x'',y'')}{d\ell} - \frac{dv(x',y')}{d\ell}\right| \leq 6Mc_1 h,$$

which coincides with (19) if we take $A_3 = 6Mc_1$. To prove the first of inequalities (20) we note that on the side of the triangle having direction ℓ_1 there is a point where $d(w - v)/d\ell_1 = 0$. In fact on the ends of this side $w - v$ vanishes by construction and, therefore, by Rolle's theorem there is an intermediate point where the derivative vanishes. We now designate this point as point (x', y') and use (19), in which we take direction ℓ to coincide with direction ℓ_1. In this way we get the first inequality. The second is obtained analogously. Having completed the proofs of inequalities (19) and (20) we have thus also completed the proof of inequality (16). To complete the proof of the whole theorem it remains for us to establish inequality (17).

Note, first of all, that each function v in U satisfies the conditions

$$\left|\frac{dv}{dx}\right| \leq ML, \qquad \left|\frac{\partial v}{\partial y}\right| \leq ML, \tag{24}$$

where M is the maximum modulus of the second derivatives of function $v(x,y)$ in domain D, and L is the diagonal of any square containing D. Suppose that the line $y = \text{const}$ intersects domain D. Since at the ends of the interval of intersection, where this line crosses Γ, by our assumption $v(x,y)$ vanishes, then at some interior point (x_0, y) of this interval, by Rolle's theorem, $\partial v(x_0,y)/\partial x = 0$. At any other point of the interval

$$\left|\frac{\partial v(x,y)}{\partial x}\right| = \left|\frac{\partial v(x,y)}{\partial x} - \frac{\partial v(x_0,y)}{\partial x}\right| = |x - x_0| \cdot \left|\frac{\partial^2 v(\xi,y)}{\partial x^2}\right| \leq LM.$$

The second of inequalities (24) is proven analogously. From the structure of the basis functions ω_n^N, it follows that the functions $w(x,y) = \sum_{n=1}^{N} v(P_n^N)\omega_n^N(x,y)$, in the region $D \setminus D_N$ over which one carries out the integration the left-hand side of (17), is identically zero. Therefore, by bound (24), the integrand on the left-hand side of (17) does not exceed the bound $2M^2L^2$, and the integral itself does not exceed the quantity

$$2M^2L^2 \cdot S_N \leq 2M^2L^2C_2h^2$$

Thus inequality (17) is valid if one takes $A_2 = 2M^2L^2C_2$. The theorem is proven.

* * *

3. An example of a variational-difference scheme for the third boundary-value problem. Consider, now, third boundary-value problem (B)§38:

$$\left.\begin{aligned} &\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = f(x,y), \\ &\frac{\partial u}{\partial n} + \sigma(s)u = \phi(s). \end{aligned}\right\} \tag{25}$$

Suppose that, for some N belonging to a given increasing sequence of natural numbers, we have chosen a net $P_1^N, \ldots, P_N^N$, and a system of basis functions $\omega_1^N, \ldots, \omega_N^N$ satisfying condition (1)

$$\omega_n^N(P_k^N) = \delta_n^k, \qquad n, k = 1, 2, \ldots, N.$$

Then the system of Ritz equations for the coefficients in the linear combination

$$w_N(x,y) = \sum_{n=1}^{N} w_N(P_n^N)\omega_n^N(x,y),$$

minimizing the functional J(w) over the class of all functions $w = a_1\omega_1^N + \ldots + a_N\omega_N^N$, can be written in the form of the following variational-difference scheme:

$$\sum_{i=1}^{N} w_N(P_i^N)(\omega_n^N, \omega_i^N) = -\int\int_D f\omega_n^N dx\, dy + \int_\Gamma \phi(s)\omega_n^N ds,$$

$$n = 1, \ldots, N \tag{26}$$

Let us now state somewhat more specifically how we will choose our net and basis functions. For a given positive integer N we inscribe, in contour Γ, a closed non-intersecting broken-line figure $Q_1^N Q_2^N \ldots Q_m^N Q_1^N$, with vertices at points $Q_1^N, \ldots, Q_m^N$, bounding the polygon D_N. We then decompose this polygon into triangles in such a way that any two either have no points in common, or have a common vertex or a common side, and that the total number of vertices of these triangles, including the vertices $Q_1^N, \ldots, Q_m^N$, is equal to N. The totality of all these vertices will be taken as our net. We label the net points $P_1^N, P_2^N, \ldots, P_N^N$, for the sake of definiteness taking $P_n^N = Q_n^N$ for n = 1, 2, ..., m. Next we define the basis function $\omega_n^N(x,y)$, n = 1, ..., N, as follows. First we specify the function at the net points in accordance with condition (1):

$$\omega_n^N(P_k^N) = \delta_n^k, \qquad n, k = 1, 2, \ldots, N. \tag{27}$$

Next we define it in each decomposition triangle in such a way that it is a linear function in each triangle, taking on values at each vertex given by Eq. (27). Thus the function $\omega_n^N(x,y)$ is already defined everywhere in the polygon D_N. Let us now define it in the region $D \setminus D_N$ and on the boundary Γ. The region $D \setminus D_N$ consists of sectors, each of which is bounded by one of the sides of the broken-line figure $Q_1^N \ldots Q_m^N Q_1^N$, the side forming a chord of an arc of contour Γ. Now single out any one of these sectors and consider that subdivision-triangle of D_N for which the chord of this sector is one of its sides. In this triangle the function $\omega_n^N(x,y)$ is already defined, and is a linear function (albeit perhaps identically zero). Let us now define $\omega_n^N(x,y)$ inside the sector and on its boundary in such a way that $\omega_n^N(x,y)$ remains a linear function in the region formed by the union of the sector and triangle (the ruled area in Fig. 49). With this auxiliary definition in each of the sectors we have completed the construction of the functions $\omega_n^N(x,y)$.

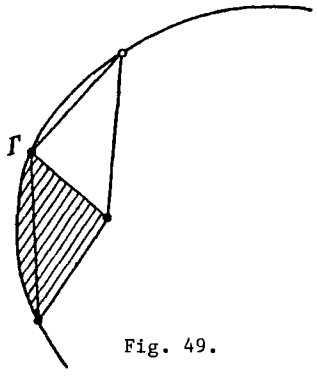


Fig. 49.

Now the coefficients and right-hand sides of variational-difference scheme (26) have taken on definite numerical values. Note that, if points P_n^N and P_i^N are not vertices of one and the same subdivision triangles, then the corresponding coefficient, (ω_n^N, ω_i^N), of scheme (26) will vanish.

Let us now discuss the question of the accuracy of the approximate solution obtained via scheme (26). By theorem 4§38

$$||w_N - v||_W^2 \leq \beta[J(w_N) - J(v)]. \tag{28}$$

Further, in view of (8) §38 and (28)

$$||w_N - v||_W^2 \leq \beta[J(w_N) - J(v)] =$$

$$= \beta\left(\iint_D \left\{\left[\frac{\partial(w_N - v)}{\partial x}\right]^2 + \left[\frac{\partial(w_N - v)}{\partial y}\right]^2\right\} dx\, dy + \right.$$

$$\left. + \int_\Gamma \sigma(s)(w_N - v)^2 ds\right) \equiv \beta||w_N - v||_{\tilde{W}}. \tag{29}$$

Suppose we know nothing about the exact solution except that it belongs to some set of functions V. Then by (29) we are only guaranteed that the difference $w_N - v$ satisfies the following bounds:

$$\left.\begin{aligned} ||\bar{w}_N - v||^2_{\tilde{W}} &\le K_N^2(V, \tilde{W}^N), \\ ||\bar{w}_N - v||^2_{W} &\le \beta K_N^2(V, \tilde{W}^N), \end{aligned}\right\} \tag{30}$$

where

$$K_N^2(V, \tilde{W}^N) = \operatorname*{Sup}_{v \text{ in } V} \operatorname*{Inf}_{w \text{ in } \tilde{W}^N} (w - v, w - v) \tag{31}$$

and $\tilde{W}^N$ is the linear N-dimensional space spanned by our system of basis functions $\omega_1^N(x,y)$, ..., $\omega_N^N(x,y)$. Consider, now, the case where V consists of all functions with continuous second derivatives not exceeding some given number in magnitude.

In theorem 2, below, are formulated additional requirements on the net which, when fulfilled, have the effect that

$$K_N(V, \tilde{W}^N) \le \frac{A}{\sqrt{N}} . \tag{32}$$

Theorem 2. *Suppose V is the set of all functions having continuous second derivatives not exceeding some number, M, in modulus. Suppose, further, that the net* P_n^N, n = 1, ..., N, *constructed above, is subjected to the two additional requirements:*

1°. *The length,* ℓ, *of each side of any of the subdivision-triangles of* D_N *satisfies the bound*

$$\ell \le c_1 h_1, \qquad h = \left(\frac{\text{area of } D}{N}\right)^{1/2},$$

where c_1 *is some constant.*

2°. *Each angle,* α, *of any of the subdivision triangles satisfies the bound*

$$\alpha > \alpha_0 > 0,$$

where α_0 *is some constant not depending on N.*

Then bound (32) is valid.

Proof. From definition (31) of the quantity $K_N(V, \tilde{W}^N)$ it follows that, to prove bound (32), it is sufficient to construct, for each function u(x,y) in V, a function $w_N(x,y)$ which satisfies the inequality

$$\iint_D \left\{\left[\frac{\partial(w - u)}{\partial x}\right]^2 + \left[\frac{\partial(w - u)}{\partial y}\right]^2\right\} dx\, dy +$$

$$+ \int_\Gamma \sigma(s)(w - u)^2 ds \le \frac{A^2}{N} \tag{33}$$

with constant A not depending on u or h. We show that we may take, as this function w, the function

$$w(x,y) = \sum_{n=1}^{N} u(P_n^N)\omega_n^N(x,y). \tag{34}$$

In view of the structure of the left-hand side of inequality (33) it is sufficient to prove the following inequality:

$$\left|\frac{\partial(w-u)}{\partial x}\right| \leq B_1 h, \quad \left|\frac{\partial(w-u)}{\partial y}\right| \leq B_1 h \text{ everywhere in } D, \tag{35}$$

$$|w-u| \leq B_2 h \text{ on } \Gamma, \tag{36}$$

where B_1 and B_2 are constants. Inequality (35) may be proven in the same way, almost word for word, as inequality (18), established above for polygon D_N. To prove inequality (36) we note that, by virtue of inequality (35) which remains valid on the boundary Γ, the derivative

$$\frac{d(w-u)}{ds} = \cos\gamma \frac{\partial(w-u)}{\partial x} + \sin\gamma \frac{\partial(w-u)}{\partial y}$$

of the function $w - u$ along the boundary does not exceed $2B_1 h$ in magnitude. Here γ is the angle between the x axis and the direction of the boundary at the given point. Further, at the points $P_n^N = Q_n^N$, $n = 1, 2, \ldots, m$, we have the equation $w - u = 0$. Therefore at any point Q on the boundary

$$|w-u|_Q = \left|\int_{Q_n^N}^{Q} \frac{d(w-u)}{ds}\, ds\right| \leq S_{QQ_n^N} \cdot 2B_1 h \leq 2 \text{ (length of } \Gamma) \cdot B_1 \cdot h,$$

where $S_{QQ_n^N}$ is the distance from point Q to the closest point Q_n^N measured along the boundary Γ. The theorem is proven.

4. On the method for proving convergence. To analyze variational-difference schemes it was not necessary for us to split the convergence proof into separate studies of stability and approximation, as we did in all other chapters. In carrying out the variational-difference computations stability, which should be understood to mean the good conditioning of the relevant system of equations, as before plays an important role: not, however, as a factor guaranteeing convergence but only as a property which permits us to disregard the influence of roundoff errors on the final result. The concept of approximation, in the sense understood everywhere in other chapters, no longer plays a role. It is replaced by approximation of the set of functions U by linear combinations of basis functions.

Incidentally, however, a variational-difference scheme on a regular net may turn out to be the same as some ordinary difference scheme (see the problem at the end of this section), and then the variational approach to a

study of this scheme can be supplemented by the methods used to study ordinary difference schemes, so as to get additional information as to the properties of the approximate solutions.

5. Comparison of variational-difference schemes with general variational and ordinary difference schemes. Variational-difference schemes are syntheses of variational methods with ordinary difference schemes. One of the basic advantages of the Galerkin-Ritz method is the great freedom it gives us in the choice of basis functions. If it is known a priori that the desired solution, u, belongs to some specific, narrow, class of functions, U, with a rapidly decreasing sequence of N-dimensional diameters κ_N, then in principle one can choose basis functions so as to achieve good accuracy even for small N and, consequently, for small computational effort. This fact made it possible for the skilled analyst to solve selected problems numerically even before the appearance of fast computing machines. But the actual construction of basis functions with good properties is a difficult problem.

In the variational-difference method one's freedom to choose basis functions is limited to the choice of net structure which results from decomposition of the given region into a set of polygons whose vertices serve as net-points, and the choice of methods by which the definition of net functions will be extended over the whole domain. This limitation in our freedom to choose basis functions brings with it, however, a degree of automation in their construction. And we can still, to a certain extent, take into account the special features of the class of functions, U, containing the solution, by use of unequal polygons; or by taking advantage of our freedom to fill in the basis function in each of the decomposition-polygons, accomplishing this process (like the decomposition itself) with the aid of a priori information on the behavior of the solution in this polygon.

On the other hand the variational-difference scheme retains the convenience of ordinary difference schemes resulting from the simple structure of coefficient matrices containing many zero elements. This structure is obtained through use of basis functions each of which differs from zero only in a small neighborhood adjacent to one of the net-points. Further we retain here the simple, visualizeable interpretation of ordinary difference schemes, where the unknowns are the values of the function of interest at the net points, and not some auxiliary system of numbers with no immediately visualizeable significance. At the same time variational-difference schemes enable us to overcome the difficulties which arise through the use of difference schemes on irregular nets, or in dealing with boundary conditions on curvilinear boundaries.

PROBLEMS

Suppose the decomposition of region D_N into triangles is accomplished in such a way that the net-point P_n^N, for a given N, is the vertex of right-angled isosceles triangles with sides of length h, shown as hatched in Fig. 50.

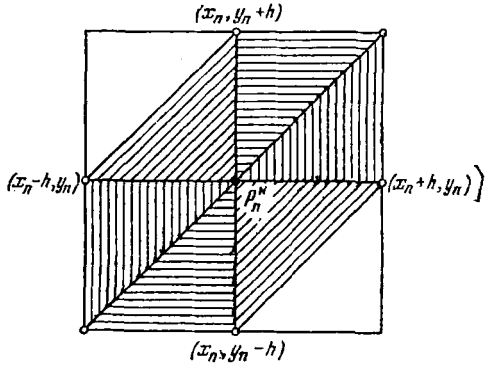


Fig. 50.

Show that the equation

$$\sum_{i=1}^{N} w_N(P_i^N)(\omega_n^N, \omega_i^N) = -\int\int_D f\omega_n^N dx\, dy,$$

corresponding to the net-point P_n^N in variational-difference scheme (2), in this case takes the form

$$-h^2\left[\frac{w_N(x_n + h, y_n) - 2w_N(x_n, y_n) + w_N(x_n - h, y_n)}{h^2} + \right.$$

$$\left. + \frac{w_N(x, y + h) - 2w(x,y) + w(x, y - h)}{h^2}\right] = -\int\int_D f\omega_n^N dx\, dy,$$

where (x_n, y_n) are the coordinates of the point P_n^N.

Part 5

Stability of Evolutional Boundary-Value Problems Viewed as the Boundedness of Norms of Powers of a Certain Operator

In the preceding parts of this book much attention was devoted to study of the stability of difference boundary-value problems $L_h u^{(h)} = f^{(h)}$. We studied, in particular, the stability of certain difference schemes approximating the Dirichlet problem for the Poisson equation. This is a stationary problem; its solution does not depend on time. But we took as fundamental the evolutional problem corresponding to time-dependent processes such as, for example, the propagation of heat or of waves. Methods for the study of evolutional difference boundary-value problems are better developed than those designed for stationary problems. This situation may be explained, in part, by the fact that, in many cases, the stationary state may be regarded as the result of the stabilization of processes evolving in time.

In studying the stability of evolutional difference problems we applied the maximum principle, energy inequalities, spectral criteria, as well as other principles. In all these approaches we used, implicitly, the special structure of the evolutional difference scheme, in which the solution $u^{(h)}$ is given on one or several initial time-levels of the net, and is then calculated step-by-step on succeeding time-levels. Here we will express the layered character of the evolutional difference scheme directly in writing the scheme, setting up a corresponding linear operator, R_h, which acts to effect the transition, from the already-known solution on a given time-level, to still unknown values of this solution on the next level. This operator may be chosen in various ways. We will construct it in such a form that the stability of the difference scheme turns out to be equivalent to the boundedness of the norms of its powers. This approach will permit us to look at the already-encountered methods for studying the stability of evolutional difference boundary-value problems from a unified point of view, regarding them as methods for studying the properties of the operator R_h: it permits us also to formulate the concept of the spectrum of a family of difference operators, and a spectral criterion for the stability of a non-selfadjoint difference boundary-value problem.

Chapter 13
Construction of the Transition Operator

§40. Level structure of the solution of evolutional problems

In all the above examples of evolutional difference schemes

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

we were given the value of the solution $u^{(h)}$ on one or several initial levels of the net. The value of $u^{(h)}$ on the following levels was determined, step by step, from the equations constituting difference boundary-value problem (1). By the term "level" we mean the totality of all points of the net D_h lying on the line (or plane) t = const. Below we will assume that difference scheme (1), under consideration here, has the indicated level-structure.

Example 1. Consider the difference scheme

$$\left.\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} = \Lambda_{xx} u_m^p + \phi(x_m, t_p), \\ u_0^{p+1} = \psi_1(t_{p+1}), \quad u_M^{p+1} = \psi_2(t_{p+1}), \quad u_m^0 = \psi(x_m), \\ m = 0, 1, \ldots, M; \quad p = 0, 1, \ldots, [T/\tau]-1, \end{gathered}\right\} \tag{2}$$

approximating the heat-conduction problem

$$\left.\begin{gathered} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \phi(x,t), \qquad 0 < x < 1, \qquad 0 < t < T, \\ u(0,t) = \psi_1(t), \quad u(1,t) = \psi_2(t), \quad u(x,0) = \psi(x), \\ 0 < x < 1. \end{gathered}\right\} \tag{3}$$

Knowing the value of the solution $u^{(h)}$ at the points of the level $t = t_p = p\tau$, i.e. knowing the net function

$$u^p = \{u_m^p\}, \qquad m = 0, 1, \ldots, M, \tag{4}$$

of argument m we can calculate, sequentially, the values of the net functions $u^{p+1} = \{u_m^{p+1}\}$, u^{p+2}, etc., using the equations

$$u_m^{p+1} = (1 - 2r)u_m^p + r(u_{m-1}^p + u_{m+1}^p) + \tau\phi(x_m, t_p). \tag{5}$$

The net function $u^0 = \{u_m^0\} = \{\psi(x_m)\}$ is given.

Thus the solution $u^{(h)}$, defined on the two-dimensional net

$$(x_m, t_p) = (mh, p\tau), \quad m = 0, 1, \ldots, M; \quad p = 0, 1, \ldots, [T/\tau] \tag{6}$$

in the x-t plane has, in a very natural way, split into layers, having been replaced by the sequence of functions

$$u^0, u^1, \ldots, u^p, \quad p = [T/\tau], \tag{7}$$

defined on one-dimensional nets. The one-dimensional nets, on which the u^p are defined for $p = 0, 1, \ldots, p$, are the same for all p (Fig. 51,a), so that one may consider them as various exemplars of one and the same net. This one dimensional net is represented in Fig. 51,b.

Consider the linear space, U^h, of functions defined on the one-dimensional net of Fig. 51,b. The net functions u^p, $p = 0, 1, \ldots, p$, in particular, belong to this space. We assume that the linear space is normed. For example the norm of the element $u = \{u_0, u_1, \ldots, u_M\}$ might be given by one of the equations

$$\|u\| = \max_m |u_m|,$$
$$\|u\| = \left(h \sum_{m=0}^{M} |u_m|^2 \right)^{1/2}. \tag{8}$$

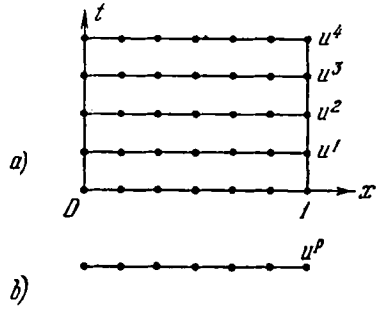


Fig. 51.

In the definition of stability and convergence one encounters the norm, $\|u^{(h)}\|_{U_h}$, of the solution of difference boundary-value problem (1). We will use only such norms $\|u^{(h)}\|_{U_h}$ which take into account the level-character of the solution $u^{(h)}$, more specifically those for which

$$\|u^{(h)}\|_{U_h} = \max_p \|u^p\|, \tag{9}$$

where p takes on the values $p = 0, 1, \ldots, [T/\tau]$, i.e. all those values for which the region of definition of the net function $u^p = \{u_m^p\}$ belongs to the two-dimensional domain of definition of the solution $u^{(h)}$.

Example 2. Consider the difference equation

$$\left.\begin{aligned} &\frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{\tau^2} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \phi(x_m, t_p), \\ &m = 0, \pm 1, \ldots; \quad p = 1, 2, \ldots, [T/\tau]-1, \end{aligned}\right\} \tag{10}$$

which is the difference analogue of the differential equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = \phi(x,t), \qquad 0 < t < T, \quad -\infty < x < \infty. \tag{11}$$

In contrast to example 1, the solution $u^{(h)}$ of this difference equation is not determined by its values at the net-points of one level $t = p\tau$. Here it is necessary to know the values $u^{(h)}$ at the net-points of two levels: $t = p\tau$ and $t = (p+1)\tau$, i.e. values of the vector-function (Fig. 52,a)

$$u^p = \begin{pmatrix} \cdots & u_{-1}^{p+1} & u_0^{p+1} & u_1^{p+1} & \cdots \\ \cdots & u_{-1}^{p} & u_0^{p} & u_1^{p} & \cdots \end{pmatrix}$$

From the values of u^p, through use of Eq. (10), one can define, sequentially, u^{p+1}, u^{p+2}, etc. In accordance with these considerations we take, as space U_h', the space of vector-functions (Fig. 52,b)

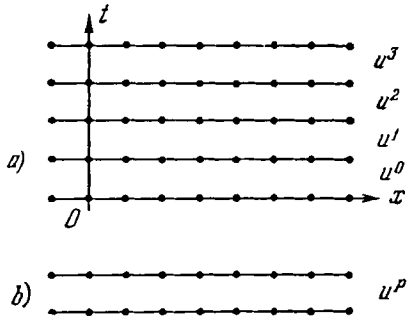


Fig. 52.

$$u = \begin{pmatrix} \cdots & b_{-1} & b_0 & b_1 & \cdots \\ \cdots & a_{-1} & a_0 & a_1 & \cdots \end{pmatrix}$$

with some norm $||u||$. Concerning this norm we make the following remarks.

The solution of differential equation (11) is determined by two functions:

$$u(x, t_0) \quad \text{and} \quad \frac{\partial u(x,t)}{\partial t},$$

whose difference analogues are, respectively, the net functions

$$\ldots, u_{-1}^p, u_0^p, u_1^p, \ldots$$

and

$$\ldots \frac{u_{-1}^{p+1} - u_{-1}^p}{\tau}, \quad \frac{u_0^{p+1} - u_0^p}{\tau}, \quad \frac{u_1^{p+1} - u_1^p}{\tau}, \ldots$$

Therefore any natural norm in space U_h' must depend on both these net functions. We may, for example, take

$$||u|| = \sup_m |a_m| + \sup_m \left|\frac{b_m - a_m}{\tau}\right|$$

or

$$||u|| = \left[h \sum_m (|a_m| + \frac{1}{\tau^2} |b_m - a_m|^2)\right]^{1/2}.$$

After the introduction of a norm in space U_h' we automatically get a norm, via Eq. (9), in the space, U_h, of net functions defined on the two-dimensional net:

$$||u^{(h)}||_{U_h} = \max_p ||u^p||.$$

Here p runs through those values p = 0, 1, ..., [T/τ], for which the region of definition of the net vector-function belongs to the two-dimensional domain of definition of the net function $u^{(h)}$.

Since, by the convention we've adopted, all our norms must be of form (9), the inequality

$$||u^{(h)}||_{U_h} \leq c||f^{(h)}||_{F_h},$$

which, for a linear operator L_h signifies stability, is equivalent to the inequality

$$||u^p|| \leq c||f^{(h)}||_{F_h}$$

for all those p for which the function u^p is defined. This turns out to be convenient for the study of stability.

PROBLEMS

1. Define the space U_h' for the difference scheme

$$\left.\begin{array}{l} \dfrac{u^{p+1}_{mn} - u^p_{mn}}{\tau} = \Lambda_{xx} u^p_{mn} + \Lambda_{yy} u^p_{mn} + \phi(x_m, y_n, t_p), \\ u^p_{mn}\Big|_\Gamma = 0, \quad u^0_{mn} = \psi(x_m, y_n), \\ m, n = 1, 2, \ldots, M-1; \quad Mh = 1, \\ p = 0, 1, \ldots, [T/\tau]-1; \end{array}\right\}$$

Here $(x_m, y_n, t_p) = (mh, nh, p\tau)$, and Γ consists of those net-points which lie on the lateral boundaries of the parallelopiped $0 \leq x, y \leq 1$, $0 \leq t \leq T$.

2. Define the space U_h' for the difference-splitting-scheme

$$\left.\begin{aligned} \frac{u_{mn}^{p+1} - \tilde{u}_{mn}}{\tau} &= \Lambda_{xx} u_{mn}^{p+1} + \phi(x_m, y_n, t_p), \\ \frac{\tilde{u}_{mn} - u_{mn}^{p}}{\tau} &= \Lambda_{yy} \tilde{u}_{mn}, \\ u_{mn}^{p}\Big|_{\Gamma} = \tilde{u}_{mn}\Big|_{\Gamma} &= 0, \quad u_{mn}^{0} = \psi(x_m, y_n), \end{aligned}\right\}$$

$m, n = 1, 2, \ldots, M-1$; $Mh = 1$; $p = 0, 1, \ldots, [T/\tau]-1$;

Γ is the lateral boundary of the parallelopiped $0 \leq x, y \leq 1$, $0 \leq t \leq T$.

§41. Statement of the difference boundary-value problem in the form $u^{p+1} = R_h u^p + \tau\rho^p$

1. **Canonical form.** We will write the difference scheme

$$\left.\begin{aligned} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} &= \phi_m^p, \quad m = 0, \pm 1, \ldots, \\ u_m^0 &= \psi_m, \quad p = 0, 1, \ldots, [T/\tau]-1, \end{aligned}\right\}$$

for the problem

$$\left.\begin{aligned} u_t - u_x &= \phi(x,t), \qquad 0 < t < T, \\ u(x,0) &= \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\}$$

in the form

$$\left.\begin{aligned} u^{p+1} &= \{(1 - r)u_m^p + ru_{m+1}^p\} + \tau\phi_m^p, \\ u_m^0 &= \psi_m, \qquad r = \tau/h. \end{aligned}\right\} \tag{1}$$

If we set

$$v_m^{p+1} = (1 - r)u_m^p + ru_{m+1}^p, \qquad \rho_m^p = \phi_m^p, \tag{2}$$

Eq. (1) can be rewritten in the form

$$u^{p+1} = v^{p+1} + \tau\rho^p, \qquad u_m^0 = \psi_m.$$

The term v^{p+1} is completely determined by $u^p = \{u_m^p\}$, so that we may write

$$v^{p+1} = R_h u^p,$$

where R_h is an operator which maps each net function u^p in U_h' into a net function v^{p+1} in U_h' via Eq. (2). In this notation Eq. (1) takes the form

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau\rho^p, \\ u^0 &\text{ given.} \end{aligned}\right\} \tag{3}$$

In this section we show also by other examples how one can reduce an evolutional difference boundary-value problem

$$L_h u^{(h)} = f^{(h)} \tag{4}$$

to form (3). Further, we establish that if, in this reduction, certain natural requirements are satisfied, then stability of problem (4) on the interval $0 \leq t \leq T$ is equivalent to fulfillment of the inequality

$$\left|\left|R_h^p\right|\right| < K, \qquad p = 1, 2, \ldots, [T/\tau], \tag{5}$$

where K is some constant independent of h: thus we reduce the study of stability to the establishment of bounds on the quantities $||R_h^p||$, i.e. the norms of powers of the *transition operator* R_h.

Analogous constructions and considerations were presented in §§15 and 16. We recall that in §41 the study of stability was reduced to the consideration of the inequality

$$||u^p|| \leq c||f^{(h)}||_{F_h}. \tag{6}$$

Specifically, it was shown that stability is equivalent to the existance of a number c, independent of h and $f^{(h)}$ in F_h, such that inequality (6) is satisfied for all p, $p = 1, 2, \ldots, [T/\tau]$.

Now we set out to implement the proposed plan, starting with an example of the reduction of a difference scheme to form (3).

Consider the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \phi_m^p, \\ &p = 0, 1, \ldots, [T/\tau]-1; \quad m = 1, 2, \ldots, M-1, \\ &u_0^p = \psi_1(t_p), \quad u_M^p = \psi_2(t_p), \quad p = 0, 1, \ldots, [T/\tau] \\ &u_m^0 = \psi(mh), \quad m = 0, 1, \ldots, M. \end{aligned}\right\} \tag{7}$$

Clearly it is necessary to satisfy, here, the consistency condition $\psi_1(0) = \psi(0)$, $\psi_2(0) = \psi(1)$. By the conditions of the problem $u^0 = \{u_m^0\}$ is given, and the functions $u^1, u^2, \ldots,$ can be computed consecutively. To carry out this computation one must rewrite the difference equation of scheme (7) in the form

$$u_m^{p+1} = (1 - 2r)u_m^p + r(u_{m-1}^p + u_{m+1}^p) + \tau\phi_m^p,$$

$$r = \frac{\tau}{h^2}, \quad m = 1, 2, \ldots, M-1; \quad p = 0, 1, \ldots, [T/\tau]-1,$$

and make use of the equation

$$u_0^{p+1} = \psi_1(t_{p+1}), \quad u_M^{p+1} = \psi_2(t_{p+1}).$$

Let us, then, take as $U_h^{\vee}$ the space of net functions

$$u = \{u_0, u_1, \ldots, u_M\}$$

with norm

$$||u|| = \max_m |u_m|.$$

We now write the difference boundary-value problem in the form

$$\left.\begin{aligned} &u^{p+1} = R_h u^p + \tau\rho^p, \\ &u^0 \text{ given}, \end{aligned}\right\} \tag{8}$$

denoting by R_h the operator mapping each element $u = \{a_m\}$ of the space $U_h^{\vee}$ into an element $b = \{b_m\}$ of the same space via the equations

$$\left.\begin{aligned} &b_0 = a_0, \\ &b_m = (1 - 2r)a_m + r(a_{m-1} + a_{m+1}), \quad m = 1, 2, \ldots, M-1, \\ &b_M = a_M. \end{aligned}\right\} \tag{9}$$

With this choice of the operator R_h the net function ρ^p in U_h'

$$\rho^p = \{\rho_0^p, \rho_1^p, \ldots, \rho_M^p\},$$

is defined by the equation

$$\rho^p = \left(\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}, \quad \phi_1^p, \ldots, \phi_{M-1}^p, \quad \frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau} \right)$$

$$p = 0, 1, \ldots, [T/\tau]-1.$$

We have now completed the reduction of the above scheme (7) to form (3). Next we propose to use this form of the difference boundary-value problem to study stability. But if inequality (6), signifying stability, is to have any meaning, one must define the norm $||f^h||_{F_h}$. In our example difference boundary-value problem (7) may be written in form (4) if we set

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2}, & m = 1, 2, \ldots, M-1, \\ u_0^{p+1}, & p = 0, 1, \ldots, [T/\tau]-1, \\ u_M^{p+1}, & p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0, & m = 0, 1, \ldots, M, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \phi(x_m, t_p), & m = 1, 2, \ldots, M-1, \\ \psi_1(t_{p+1}), & p = 0, 1, \ldots, [T/\tau]-1, \\ \psi_2(t_{p+1}), & p = 0, 1, \ldots, [T/\tau]-1, \\ \psi(x_m), & m = 0, 1, \ldots, M. \end{cases}$$

We define the norm $||f^{(h)}||$ by the equation

$$||f^{(h)}||_{F_h} = \max_{m,p} |\phi(x_m, t_p)| + \max_m |\psi(x_m)| + \max_p \left| \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} \right| + \max_p \left| \frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau} \right|. \tag{10}$$

2. Stability as the uniform boundedness of the norms of powers of R_h. We now formulate two conditions which, if satisfied in the reduction of any difference scheme (4) to form (3), allow one to affirm that inequality (5) implies stability.

Condition 1°. *The inequality*

$$||\rho^p|| \leq K_1 ||f^{(h)}||_{F_h},$$

is valid for some K_1 *independent of* h *and* $f^{(h)}$, *and for all* p *for which* ρ^p *is meaningful.*

Condition 2°. *The bound*

$$||u^0|| \leq K_2 ||f^{(h)}||_{F_h},$$

holds for some K_2 *independent of* h *and* $f^{(h)}$.

Conditions 1° and 2° require a certain compatibility between the choice of norms in spaces U_h and F_h, and the definition of the operator R_h (since the form of the vector ρ^p is uniquely determined by the choice of R_h). We note that in the above example, where scheme (7) is reduced to form (3), these conditions are fulfilled. To convince oneself of this it suffices to compare the norms of net functions ρ^p and u^0

$$||\rho^p|| = \max_m |\rho_m^p| = \max\left(\left| \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} \right|, \quad |\phi(x_1, t_p)|, \ \ldots \right.$$
$$\left. \ldots, |\phi(x_{M-1}, t_p)|, \quad \left| \frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau} \right| \right),$$

$$||u^0|| = \max_m |u_m^0| = \max_m |\psi(x_m)|$$

with the norms $||f^{(h)}||_{F_h}$ defined by Eq. (10). The numbers K_1 and K_2, in this example, may be set equal to one.

Let us now prove that, if in the reduction of difference boundary-value problem (4) to form (3), conditions 1° and 2° are satisfied, then the validity of bound (5) is sufficient for the stability of difference scheme (4). We have to show that the bound

$$||u^p|| \leq K_3 ||f^{(h)}||_{F_h},$$

where K_3 is some number independent of h and $f^{(h)}$, is satisfied for all p, $p = 0, 1, \ldots, p_0$ for which the domain of definition of the net function u^p belongs to the region of definition of the solution $u^{(h)}$.

From the equation

$$u^{p+1} = R_h u^p + \tau\rho^p, \quad p = 0, 1, \ldots, p_0-1,$$

it follows that

$$u^p = R_h u^{p-1} + \tau\rho^{p-1} = R_h\left(R_h u^{p-2} + \tau\rho^{p-2}\right) + \tau\rho^{p-1} =$$

. . . . . .

$$= R_h^p u^0 + \tau\left(R_h^{p-1}\rho^0 + R_h^{p-2}\rho^1 + \ldots \rho^{p-1}\right), \quad p = 0, 1, \ldots, p_0. \tag{11}$$

By assumption

$$||R_H^\ell|| \leq K, \quad \ell = 0, 1, \ldots, [T/\tau].$$

From this inequality and Eq. (11) we get the bound

$$||u^p|| \leq K||u^0|| + \tau K(||\rho^0|| + ||\rho^1|| + \ldots + ||\rho^{p-1}||). \tag{12}$$

Taking account of conditions 1° and 2°, by virtue of which we may write

$$||\rho^\ell|| \leq K_1||f^{(h)}||_{F_h}, \quad \ell = 0, 1, \ldots, p_0-1,$$
$$||u^0|| \leq K_2||f^{(h)}||_{F_h},$$

bound (12) can be replaced by the following

$$||u^p|| \leq (KK_2 + \tau\rho KK_1)||f^{(h)}||_{F_h} \leq K(K_1 + TK_1)||f^{(h)}||_{F_h} = c||f^{(h)}||_{F_h},$$

where the constant $c = K(K_2 + TK_1)$ does not depend on h or $f^{(h)}$. Thus we have proven stability.

Let us now use the just-established sufficient condition for stability to show that, for $r = \tau/h^2 \leq 1/2$, difference scheme (7) is stable.

In fact we will convince ourselves that, for the operator R_h, which we introduced via Eq. (9) in the reduction of scheme (7) to canonical form (3), we have the inequality $||R_h|| \leq 1$ and, therefore, also $||R_h^p|| \leq ||R_h||^p \leq 1$. For $r \leq 1/2$

$$|b_0| = |a_0| \leq \max_m|a_m| = ||a||,$$

$$|b_m| = |(1 - 2r)a_m + r(a_{m-1} + a_{m+1})| \leq$$
$$\leq (1 - 2r + 2r)\max_m|a_m| = ||a||,$$

$$|b_m| = |a_M| \leq \max_m|a_m| = ||a||.$$

From these bounds it follows that $||b|| = ||R_h a|| \leq ||a||$, i.e. $||R_h|| \leq 1$. Thus for $r \leq 1/2$ the sufficient condition for stability is satisfied. One can show that, if the constant $r = \tau/h^2 > 1/2$, the sufficient condition for stability is not satisfied. It is, then, natural to ask whether, also in the general case, stability is lost when the inequality $||R_h^p|| < K$, $p = 1, 2, \ldots, [T/\tau]$ is no longer valid. It turns out that in fact the validity of the inequality $||R_h^p|| < K$ is necessary for stability under one, additional, condition 3° which we now state in general form, and which is satisfied in the example just considered above.

Condition 3°. Suppose the difference boundary-value problem (4) is reduced to form (3). Take any function $\bar{u}^0$ in U_h' and construct the net function $\bar{u}^1, \bar{u}^2, \ldots \bar{u}^p, \ldots$ by the recurrence relation $\bar{u}^{p+1} = R_h \bar{u}^p$. The set of net functions $\{\bar{u}^p\}$, $p = 0, 1, \ldots, [T/\tau]$, each of which belongs to U_h', forms some net function $\bar{u}^{(h)}$ in space U_h. Let us now compute the corresponding $\bar{f}^{(h)}$,

$$\bar{f}^{(h)} \equiv L_h \bar{u}^{(h)}.$$

We will say that, *in the reduction of difference scheme (4) to canonical form (3), condition 3° is satisfied if there exists a bound of the form*

$$||\bar{f}^{(h)}||_{F_h} \leq K_3 ||\bar{u}^0||,$$

where the constant K_3 *does not depend on* $\bar{u}^0$ *in* U_h' *and does not depend on* h.

Let us now convince ourselves that in the reduction, just described above, of difference scheme (7) to canonical form (3), condition 3° is fulfilled. In fact, given an arbitrary function $\bar{u}^0 = \{\bar{u}_m^0\}$, we get

$$\bar{\phi}_m^p \equiv 0, \quad \bar{\psi}_1^p = \bar{u}_0^0, \quad \bar{\psi}_2^p = \bar{u}_M^0, \quad \bar{\psi}_m \equiv \bar{u}_m^0.$$

With our choice of norms

$$||\bar{f}^{(h)}||_{F_h} = ||\bar{u}^0||.$$

We now show that if, in the reduction of difference scheme (4) to canonical form (3), condition 3° is fulfilled then, for this scheme to be stable on the interval $0 < t < T$, it is necessary that the transition operator satisfy bound (5):

$$||R_h^p|| < K, \quad p = 1, 2, \ldots, [T/\tau],$$

where K is a constant not depending on h.

If the indicated criterion is not satisfied then, for any K, one can find an h and P_0, and a net function $\bar{u}^0$, such that

$$\left|\left|R_h^{p_0}\bar{u}^0\right|\right| > K||u^0|| \quad \left(\left|\left|\bar{u}^{p_0}\right|\right| > K||\bar{u}^0||\right).$$

Having constructed the vectors $\bar{u}^p\left(\left|\left|\bar{u}^{p_0}\right|\right| > K||\bar{u}^0||\right)$ from $\bar{u}^0$ and, from the $\bar{u}^p$'s, formed the net function $\bar{u}^{(h)}$, we conclude that

$$||\bar{f}^{(h)}|| \leq K_3||\bar{u}^0|| \quad (\bar{f}^{(h)} = L_h\bar{u}^{(h)}).$$

At the same time

$$||\bar{u}^{(h)}||_{U_h} = \max_p||\bar{u}^p|| \geq \left|\left|\bar{u}^{p_0}\right|\right| > K||\bar{u}^0||.$$

It is clear, therefore, that

$$||\bar{u}^{(h)}||_{U_h} > \frac{K}{K_3}||\bar{f}^{(h)}||_{F_h}.$$

This inequality, because of the arbitrariness of K, does indeed signify instability.

Now let us summarize the considerations of this section. We have shown that, after reducing the difference scheme $L_h u^{(h)} = f^{(h)}$ to form (3)

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau\rho^p, \\ u^0 &\text{ given,} \end{aligned}\right\}$$

one can use the operator R_h to study stability. More precisely we have proven the following

Theorem. *If, in the reduction of difference scheme (4) to form (3), condition 3° has been satisfied then, for stability, it is necessary that*

$$||R_h^p|| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{13}$$

where K *is some constant not depending on h. If the reduction to form (3) has been carried out in accordance with conditions 1° and 2°, then bounds (13) are sufficient for stability.*

We call to the reader's attention the fact that, ordinarily, the splitting of $u^{(h)}$ into levels, and the reduction of the difference scheme to canonical form (3), may be accomplished in several different ways. However, we will not pause to discuss this point in detail (see §14, where the same question was discussed in the case of difference schemes for ordinary differential equations).

3. Example. In concluding this section we consider the implicit scheme

$$\left.\begin{gathered}\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m-1}^{p+1}-2u_m^{p+1}+u_{m+1}^{p+1}}{h^2}=\phi_m^p,\\ m=1,\ 2,\ \ldots,\ M-1;\quad p=0,\ 1,\ \ldots,\ [T/\tau]-1,\\ u_m^0=\psi_m,\quad m=0,\ \ldots,\ M,\\ u_0^{p+1}=\psi_1(t_{p+1}),\quad u_M^{p+1}=\psi_2(t_{p+1}),\quad p=0,\ 1,\ \ldots,\ [T/\tau]-1,\end{gathered}\right\}\tag{14}$$

for the heat-conduction problem

$$\left.\begin{gathered}\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=\phi(x,t),\\ u(x,0)=\psi(x),\\ u(0,t)=\psi_1(t),\quad u(1,t)=\psi_2(t),\\ 0<t<T,\quad 0<x<1.\end{gathered}\right\}\tag{15}$$

This scheme was considered in detail in §28.

We take as vector u^p the vector $u^p=(u_0^p,\ u_1^p,\ \ldots,\ u_M^p)$ with norm $\|u^p\|=\max_m|u_m^p|$. The solution at the (p+1)'st level will be written in the form of a sum,

$$u^{p+1}=v^{p+1}+\tau\rho^p,$$

where

$$v^{p+1}=(v_0^{p+1},\ v_1^{p+1},\ \ldots,\ v_M^{p+1})\quad\text{and}\quad \rho^p=(\rho_o^p,\ \rho_1^p,\ \ldots,\ \rho_M^p)$$

are, in turn, solutions of the auxiliary systems of equations

$$\left.\begin{gathered}v_0^{p+1}=u_0^p=\psi_1(t_p),\\ rv_{m+1}^{p+1}-(1+2r)v_m^{p+1}+rv_{m-1}^{p+1}=-u_m^p,\\ m=1,\ 2,\ \ldots,\ M-1,\\ v_M^{p+1}=u_M^p=\psi_2(t_p),\end{gathered}\right\}\tag{16}$$

$$\left.\begin{aligned} \rho_0^p &= \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} \\ r\rho_{m+1}^p - (1 + 2r)\rho_m^p + r\rho_{m-1}^p &= \phi_m^p, \quad m = 1, 2, \ldots, M-1, \\ \rho_M^p &= \frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau}. \end{aligned}\right\} \quad (17)$$

The first of these systems can be taken as the definition of the operator R_h, so that we may write

$$v^{p+1} = R_h u^p.$$

If $||f^h||_{F_h}$ is defined, as before, by Eq. (10), then satisfaction of condition 1°

$$||\rho^p|| \leq K_1 ||f^{(h)}||_{F_h}$$

follows from the bound

$$||\rho^p|| \leq \max\left[\left|\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}\right|, \quad \left|\frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau}\right|, \quad \max_m|\phi_m^p|\right],$$

valid for the solution $\{\rho_m^p\}$ of system (17) by virtue of bound (7) §4. Further, $K_1 = 1$.

Condition 2°

$$||u^0|| \leq K_2 ||f^{(h)}||_{F_h}$$

is also satisfied, by virtue of (10) with $u_m^0 = \psi_m$, and here, clearly, we can set $K_2 = 1$. Further, in §28 we proved the bound

$$|v_m^{p+1}| \leq \max_m|u_m^p|,$$

which may be interpreted as the inequality

$$||R_h u|| \leq ||u||, \quad ||R_h|| \leq 1,$$

from which it follows that

$$||R_h^p|| \leq K = 1.$$

We have here followed the same general plan as in the stability proof of §28, showing that it and the proof presented here are the same. The above example is interesting also in that it makes use of a rather complicated method for constructing the vector ρ^p.

PROBLEMS

1. For the system of acoustic equations

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = \phi(x,t), \quad -\infty < x < \infty, \quad 0 < t < T,$$

$$u(x,0) = \psi(x), \quad -\infty < x < \infty,$$

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad u(x,t) = \begin{pmatrix} v(x,t) \\ w(x,t) \end{pmatrix},$$

$$\phi(x,t) = \begin{pmatrix} \phi_1(x,t) \\ \phi_2(x,t) \end{pmatrix}, \quad \psi(x) = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \end{pmatrix}$$

reduce to canonical form (3) the scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} + A\frac{u_{m+1}^p - u_{m-1}^p}{2h} - \\ &- \frac{\tau}{2h^2}A^2\left(u_{m+1}^p - 2u_m^p + u_{m-1}^p\right) = \phi(x_m, t_p), \\ &u_m^0 = \psi(x_m), \end{aligned}\right\} \qquad (*)$$

taking $u^p = \{u_m^p\}$. Verify that, if norms are defined via the equation

$$||u^{(h)}||_{U_h} = \max_p ||u^p||, \quad ||f^{(h)}||_{F_h} = \max\left[||\psi||, \max_p ||\phi^{(p)}||\right],$$

where

$$||u^p||^2 = \sum_m \left(|v_m^p|^2 + |w_m^p|^2\right).$$

$$||\psi||^2 = \sum_m \left(|\psi_1(mh)|^2 + |\psi_2(mh)|^2\right),$$

$$||\phi^{(p)}||^2 = \sum_m \left(|\phi_1(x_m,t_p)|^2 + |\phi_2(x_m,t_p)|^2\right),$$

conditions 1°-3° are satisfied.

Prove that for $r \leq 1$ difference scheme (*) is stable, and for $r > 1$ is unstable.

Hint. To bound the norms $||R_h^p||$ go over to the variables

$$I_m^{(1)} = v_m + w_m, \quad I_m^{(2)} = v_m - w_m,$$

called "Riemann invariants" and use the spectral criterion of section 4§25.

2. Bring the difference scheme

$$\left.\begin{aligned} &\frac{v_m^{p+1} - 2v_m^p + v_m^{p-1}}{\tau^2} - \frac{v_{m+1}^p - 2v_m^p + v_{m-1}^p}{h^2} = \phi_m^p, \\ &p = 1, \ldots, [T/\tau]-1, \\ &v_m^0 = \psi_0(x_m), \quad v_m^1 = \tau\psi_1(x_m) + \psi_0(x_m), \quad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

approximating the Cauchy problem

$$\frac{\partial^2 v}{\partial t^2} - \frac{\partial^2 v}{\partial x^2} = \phi(x,t), \quad -\infty < x < \infty, \quad 0 < t < T,$$

$$v(x,0) = \psi_0(x), \quad \frac{\partial v(x,0)}{\partial t} = \psi_1(x), \quad -\infty < x < \infty,$$

to canonical form (3) setting

$$u^p = \begin{Bmatrix} v_m^{p+1} \\ v_m^p \end{Bmatrix},$$

$$||u^p||^2 = \sum_m |v_m^p|^2 + \sum_m \left|\frac{v_m^{p+1} - v_m^p}{\tau}\right|^2,$$

$$||\phi^p||^2 = \sum_m |\phi_m^p|^2, \quad ||\psi_k||^2 = \sum |\psi_k(x_m)|^2,$$

$$||u^{(h)}||_{U_h} = \max_p ||u^p||.$$

$$||f^{(h)}||_{F_h} = ||\psi_0|| + ||\psi_1|| + \max_p ||\phi^p||.$$

a) Verify that conditions 1°-3° are satisfied.

b) Prove that for $\tau/h = r \leq 1$ the scheme is stable, while for $\tau/h = r > 1$ it is unstable.

§42. Use of particular solutions in the construction of the transition operator

In §41 we talked about the reduction of the difference boundary-value problem

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

to the form

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau\rho^p, \\ u^0 &\text{ given.} \end{aligned}\right\} \tag{2}$$

In this reduction the operator R_h can be chosen in various ways. The reduction to form (2) has the purpose that, by bounding the values of $||R_h^p||$, we be able to draw conclusions regarding stability. It was shown that the bound

$$||R_h^p|| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{3}$$

ensures stability if only the operator R_h, and the required norms, are chosen in such a way as to satisfy the conditions:

1° $$||\rho^p|| \leq K_1 ||f^{(h)}||_{F_h},$$

where p runs through all values for which ρ^h is defined; and

2° $$||u^0|| \leq K_2 ||f^{(h)}||_{F_h}.$$

In the examples considered in §41 the operator R_h could be taken to be fairly simple, at the same time still satisfying conditions 1° and 2°. But one can also encounter examples (and one will be considered in this section) where condition 1° is too strict, so that the operator R_h, constructed so as to take account of this condition, cannot be as simple as one would like.

Below in this section it will be shown that bound (3) remains sufficient for stability if condition 1° is replaced by the less restrictive condition 1*. Thanks to the substition of 1* for 1° the operator R_h may be taken simpler and simpler the more one knows about the solutions of the above difference problem (1). In particular, the structure of the operator, R_h, which appears in the reduction of the difference scheme to form (2), becomes simpler if we know certain particular solutions of the difference equations entering into the formulation of problem (1). Correspondingly, one can simplify the proof of inequality (3), which implies

stability. All this will be demonstrated later by way of examples. Now we turn to the formulation of condition 1*.

Suppose that z^h is some net function in U_h depending, generally, on $f^{(h)}$. In the reduction of difference scheme (1) to canonical form (2) we split the net function $u^{(h)}$ into functions u^p of U_h'. In the same way we now split the net function $z^{(h)}$, by levels, into functions z^p of U_h', and postulate that the z^p satisfy inequalities of the form

$$||z^p|| \leq \tilde{K}||f^{(h)}||_{F_h}, \tag{4}$$

where $\tilde{K}$ is some constant, and p runs through those values p = 0, 1, ..., p_0 for which z^p is defined.

Condition 1*. *There exists a function* $z^{(h)}$, *satisfying inequality (4), and such that*

$$\left|\left|\rho^p - \frac{1}{\tau}(z^{p+1} - R_h z^p)\right|\right| \leq K_1||f^{(h)}||_{F_h}.$$

If one can take, as $z^{(h)}$, $z^{(h)} \equiv 0$, then not only condition 1* is satisfied, but also the stricter condition 1°.

Theorem. *If difference problem (1) may be written in canonical form (2), while satisfying both conditions 1* and 2°, then bound (3) implies the inequality*

$$||u^p|| \leq c\,||f^{(h)}||_{F_h}, \tag{5}$$

which, in turn, implies stability. As the constant c one may take

$$c = K(2K_2 + 2\tilde{K} + TK_1) + \tilde{K}.$$

Proof. Define the function $w^{(h)} = u^{(h)} - z^{(h)}$. From the equation

$$u^{p+1} = R_h u^p + \tau\rho^p$$

it follows that

$$w^{p+1} = R_h w^p + \tau\tilde{\rho}^p, \tag{6}$$

where

$$\tilde{\rho}^p = \rho^p - \frac{1}{\tau}(z^{p+1} - R_h z^p).$$

By condition 1* we have

$$||\tilde{\rho}^p|| \leq K_1 ||f^{(h)}||_{F_h}.$$

Using Eq. (6) and inequality (3) we find without difficulty, as we have many times before, that

$$||w^p|| \leq K||w^0|| + TK \max_p ||\tilde{\rho}^p|| \leq K||w^0|| + TKK_1||f^{(h)}||_{F_h}. \tag{7}$$

Further

$$||w^0|| \leq 2(K_2 + \tilde{K})||f^{(h)}||_{F_h}. \tag{8}$$

This follows from the inequalities

$$||w^0|| = ||u^0 - z^0|| \leq ||u^0|| + ||z^0||,$$

$$||u^0|| \leq K_2||f^{(h)}||_{F_h}, \quad ||z^0|| \leq \tilde{K}||f^{(h)}||_{F_h},$$

of which the second coincides with condition 2°, and the third with inequality (4) for p = 0.

Substituting bound (8) for $||w^0||$ into (7) we see that

$$||w^p|| \leq [2(K_2 + \tilde{K})K + TKK_1]||f^{(h)}||_{F_h}.$$

It only remains, now, to note that

$$||u^p|| = ||w^p + z^p|| \leq ||w^p|| + ||z^p|| \leq$$

$$\leq \{[2K(K_2 + \tilde{K}) + TKK_1] + \tilde{K}\}||f^{(h)}||_{F_h} \leq$$

$$\leq [K(2K_2 + 2\tilde{K} + TK_1) + \tilde{K}]||f^{(h)}||_{F_h} = c||f^{(h)}||_{F_h}.$$

Thanks to the replacement of condition 1° by condition 1* one can now, in the investigation of stability, apportion the difficulties between the construction of an operator, R_h, whose norms are not too difficult to bound and the proof of the existence of a function $z^{(h)}$. In demanding from the start that condition 1* be satisfied with $z^{(h)} = 0$, i.e. that condition 1° should be satisfied, we are imposing the strictest limitations on the choice of the operator R_h. It may turn out that any operator, R_h, which we manage to construct under condition 1° will have a very complicated form, so that bounding the norms of its powers will be too difficult. On the other hand if we make the operator R_h extremely simple, equal to one let us say, and not in any way connected to the difference problem, we transfer all the difficulties to the verification of condition 1*, i.e. to the computation of the necessary bound for the function $z^{(h)}$ which in this

case, it is most natural to take equal to $u^{(h)}$. The introduction of such an operator R_h, and such a function $z^{(h)}$, would in no way advance our investigation of stability.

One must try to choose an operator, R_h, which is as simple as possible. On the other hand, R_h must so faithfully reflect the properties of the difference problem $L_h u^{(h)} = f^{(h)}$, that fulfillment of condition 1*, i.e. existence of the required function, $z^{(h)}$, will be fairly clear. It is often possible to use the freedom we have in the choice of R_h, thanks to the replacement of condition 1° by the less restrictive condition 1*, to make it easier to prove stability. For this purpose one takes, as $z^{(h)}$, functions constructed from the solutions of difference problems with right-hand sides $f^{(h)}$ of some special form.

We will now show, by way of examples, how to use the proposed method.

Example 1. Consider a difference boundary-value problem (1) of the form

$$\left.\begin{aligned} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} &= \phi_m^p, \quad m = 0, 1, \ldots, M-1;\ Mh = 1, \\ u_m^0 &= \psi(x_m), \quad m = 0, 1, \ldots, M. \\ u_M^p &= \psi_1(t_p), \quad p = 1, 2, \ldots, [T/\tau]. \end{aligned}\right\} \tag{9}$$

This difference problem approximates the problem

$$\left.\begin{aligned} u_t - u_x &= \phi(x,t), \quad 0 < x < 1, \quad 0 < t < T, \\ u(x,0) &= \psi(x), \quad 0 < x < 1, \\ u(1,t) &= \psi_1(t), \quad 0 < t < T, \end{aligned}\right\} \tag{10}$$

for the following choice of norms

$$||u^{(h)}||_{U_h} = \max_p \max_m |u_m^p|,$$

$$||f^{(h)}||_{F_h} = \max_p \max_m |\phi_m^p| + \max_m |\psi(x_m)| + \max_p |\psi_1(t_p)| + \max_p \left|\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}\right|.$$

To reduce problem (9) to canonical form (2) we set

$$u^p = (u_0^p, u_1^p, \ldots, u_M^p), \quad ||u|| = \max_m |u_m|.$$

The operator R_h, $b = R_h a$, carrying the element $a = (a_0, a_1, \ldots, a_M)$ of space U_h° into element $b = (b_0, b_1, \ldots, b_M)$ of the same space, will be defined by the equations

$$\left.\begin{aligned} b_m &= (1 - r)a_m + ra_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ b_M &= 0, \qquad\qquad\qquad r = \tau/h. \end{aligned}\right\}$$

Then, obviously,

$$\rho^p = \left(\phi_0^p, \phi_1^p, \ldots, \phi_{M-1}^p, \frac{\psi_1(t_{p+1})}{\tau}\right).$$

It is clear that condition 1°

$$||\rho^p|| \leq K_1 ||f^{(h)}||_{F_h}$$

is not satisfied because the last component of the vector ρ^p is $\psi_1(t_{p+1})/\tau$, which grows as $\tau \to 0$. (In this problem it would have been easy to formulate an operator, R_h, such that condition 1° would be satisfied. To do this it would suffice, in the definition of R_h, to replace the equation $b_M = 0$ by the equation $b_M = a_M$.) On the other hand it isn't difficult to show that the condition 1*

$$\left|\left|\rho^p - \frac{1}{\tau}\left(z^{p+1} - R_h z^p\right)\right|\right| \leq \tilde{K}||f^{(h)}||_{F_h}$$

is satisfied. The left-hand side of this inequality can be written in the form

$$\left|\left|\left(\phi_0^p, \phi_1^p, \ldots, \phi_{M-1}^p, \frac{\psi_1(t_{p+1})}{\tau}\right) - \frac{1}{\tau}\left(z^{p+1} - R_h z^p\right)\right|\right|.$$

Therefore to prove that 1* is fulfilled it is sufficient to construct a function, $z^{(h)}$, satisfying the equation

$$\frac{1}{\tau}\left(z^{p+1} - R_h z^p\right) = \left(0, 0, \ldots, 0, \frac{\psi_1(t_{p+1})}{\tau}\right)$$

which may be written in the form

$$z^{p+1} = R_h z^p + \tau\left(0, 0, \ldots, 0, \frac{\psi_1(t_{p+1})}{\tau}\right)$$

or

$$\left.\begin{aligned} &\frac{z_m^{p+1} - z_m^p}{\tau} - \frac{z_{m+1}^p - z_m^p}{h} = 0, \quad m = 0, 1, \ldots, M-1, \\ &z_M^{p+1} = \psi_1(t_{p+1}). \end{aligned}\right\} \tag{11}$$

In the case where $\psi_1(t_{p+1})$ does not depend on t this problem has the stationary (i.e. p-independent) solution

$$z_m^p \equiv \underset{m}{\psi_1} = \text{const}.$$

In the general case $\psi_1^p = \psi_1(t_p)$ depends on p but, for a bounded norm $||f^{(h)}||_{F_h}$ (containing the term $|\psi_1(t_{p+1}) - \psi_1(t_p)|/\tau$) it cannot vary very rapidly. Therefore the function $z^{(h)}$, defined by the equation

$$z_m^p \equiv \underset{m}{\psi_1}(t_p),$$

although not a stationary solution (nor a solution at all) of problem (11), "almost" satisfies (11). In fact

$$\frac{1}{\tau}(z^{p+1} - R_h z^p) = \left(\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}, \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}, \ldots\right.$$

$$\left. \ldots, \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}, \frac{\psi_1(t_{p+1})}{\tau}\right).$$

Therefore

$$\left|\left|\rho_p - \frac{1}{\tau}(z^{p+1} - R^h z^p)\right|\right| = \left|\left|\phi_0^p - \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau},\right.\right.$$

$$\left.\left.\phi_1^p - \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}, \ldots, \phi_{M-1}^p - \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}, 0\right|\right| \leq$$

$$\leq ||(\phi_0^p, \phi_1^p, \ldots, \phi_{M-1}^p, 0)|| +$$

$$+ \left|\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}\right| ||(1, 1, \ldots, 1, 0)|| \leq ||f^{(h)}||_{F_h}$$

Condition 1* is satisfied; $\widetilde{K} = 1$ and $z^{(h)}$ satisfies the inequality

$$||z^p|| = \max_m |z_m^p| \leq ||f^{(h)}||_{F_h}.$$

Condition 2°

$$||u^0|| \leq K_2||f^{(h)}||_{F_h}$$

is also satisfied:

$$||u^0|| = \max_m ||u_m^0|| = \max_m |\psi(x_m)| \leq ||f^{(h)}||_{F_h}.$$

To prove stability, which is present for $\tau \leq h$, it is sufficient to show that $||R_h^p|| \leq 1$. The validity of this inequality follows from the bound $||R_h|| \leq 1$:

$$||R_h a|| = \max_m |a_m(1 - r) + ra_{m+1}| \leq \max_m |a_m| = ||a||$$

Example 2. We will take, as a more complicated example, a different difference scheme for the same differential boundary-value problem (10):

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2}\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \phi_m^p, \\ &\qquad p = 0, 1, \ldots, [T/\tau]-1; \quad m = 1, 2, \ldots, M-1, \\ &\qquad u_m^0 = \psi(x_m), \quad m = 0, 1, \ldots, M, \\ &\qquad u_M^{p+1} = \psi_1(t_{p+1}), \quad p = 0, 1, \ldots, [T/\tau]-1, \\ &\frac{u_0^{p+1} - u_0^p}{\tau} - \frac{u_1^p - u_0^p}{h} = \phi_0^p, \quad p = 0, 1, \ldots, [T/\tau]-1. \end{aligned}\right\} \quad (12)$$

The difference equation which occurs in this scheme is of second order in x, while the corresponding differential equation (10) is first order. Therefore at the left-hand boundary x = 0 (m = 0) we have added the condition

$$\frac{u_0^{p+1} - u_o^p}{\tau} - \frac{u_1^p - u_0^p}{h} = \phi_0^p,$$

which we will use in the form

$$u_0^{p+1} = (1 - r)u_0^p + ru_1^p + \tau\phi_0^p,$$

Difference scheme (12) has already been considered in §23, where we discussed the question of approximation. Norms in the space F_h were introduced, there, as follows: if

$$f^{(h)} = \begin{cases} a_m \text{ at points } (mh, 0), & m = 0, 1, \ldots, M \\ b^p \text{ at points } (1, p\tau), & p = 0, 1, \ldots, [T/\tau]. \\ c^p \text{ at points } (0, p\tau), & p = 0, 1, \ldots, [T/\tau]-1, \\ \phi_m^p \text{ at points } (mh, p\tau), & m = 1, 2, \ldots, M-1; \\ & p = 0, 1, \ldots, [T/\tau]-1, \end{cases}$$

then

$$||f^{(h)}||_{F_h} = h \max_p |c^p| + h \max_p \left|\frac{c^{p+1} - c^p}{\tau}\right| +$$

$$+ \left(h \sum_{m=0}^{M} |a_m|^2\right)^{1/2} + \max_p |b^p| + \max_p \left|\frac{b^{p+1} - b^p}{\tau}\right| + \max_p \max_m |\phi_m^p|.$$

As was shown in §23 approximation, in this case, is of order h^2. Let us now show that, if we define the norm $||u^{(h)}||_{U_h}$ by the equation

$$||u^{(h)}||_{U_h} = \max_p \left(\frac{h}{2} |u_0^p|^2 + h \sum_{m=1}^{M} |u_m^p|^2\right)^{1/2}$$

with $r \leq 1$ then, along with approximation, we also have stability.

We verify stability, first having brought difference scheme (12) into form (2). For this purpose we set

$$u^p = (u_0^p, u_1^p, \ldots, u_M^p)$$

with norm

$$||u^p|| = \left(\frac{h}{2} |u_0^p|^2 + h \sum_{m=1}^{M} |u_m^p|^2\right)^{1/2}.$$

The operator R_h will be defined via the following equations:

If $a = (a_0, a_1, \ldots, a_M)$, $b = (b_0, b_1, \ldots, b_M)$ and $b = R_h a$, then

$$\left.\begin{aligned} b_0 &= (1 - r)a_0 + ra_1, \\ b_m &= \left(-\frac{r}{2} + \frac{r^2}{2}\right)a_{m-1} + (1 - r^2)a_m + \left(\frac{r}{2} + \frac{r^2}{2}\right)a_{m+1}, \\ & \qquad m = 1, 2, \ldots, M-1, \\ b_M &= 0. \end{aligned}\right\} \qquad (13)$$

In this case

$$\rho^p = \left[\phi_0^p,\ \phi_1^p,\ \ldots,\ \phi_{M-1}^p,\ \frac{\psi_1(t_{p+1})}{\tau}\right].$$

Clearly, with our choice of norms condition 1° is not satisfied. In fact, if, for example

$$f^{(h)} = \begin{cases} 0 \text{ at points } (mh,\ 0), & m = 0,\ 1,\ \ldots,\ M \\ 0 \text{ at points } (1,\ p\tau), & p = 0,\ 1,\ \ldots,\ [T/\tau]. \\ 1 \text{ at points } (0,\ p\tau), & p = 0,\ 1,\ \ldots,\ [T/\tau]-1, \\ 0 \text{ at points } (mh,\ p\tau), & m = 1,\ 2,\ \ldots,\ M-1; \\ & p = 0,\ 1,\ \ldots,\ [T/\tau]-1, \end{cases}$$

then

$$\rho^p = (1,\ 0,\ \ldots,\ 0), \quad \|\rho^p\| = \frac{\sqrt{h}}{\sqrt{2}}, \quad \|f^{(h)}\|_{F_h} = h,$$

so that there can be no K_1 for which the inequality

$$\|\rho^p\| \leq K_1 \|f^{(h)}\|_{F_h}.$$

will hold for all h.

For our choice of the space U_h', consisting of the vectors $u^p = (u_0^p,\ u_1^p,\ \ldots,\ u_M^p)$, and for our choice of norms it is, apparently, impossible to find an operator, R_h, such that condition 1° will be satisfied, but to satisfy condition 1* is possible.

* * * * * *

Before proving this last assertion we note that, if we change the norm $\|f^{(h)}\|_{F_h}$ setting

$$\|f^{(h)}\|_{F_h} = \sqrt{h}\max|c^p| + \left(h\sum_{m=0}^{M}|a_m|^2\right)^{1/2} + \max\left|\frac{b^p}{\tau}\right| + \max_p \max_m |\phi_m^p|,$$

then the operator R_h defined by Eq. (13), will satisfy condition 1°, but the order of approximation (instead of h^2) will be only $h^{3/2}$. We can, without changing the norms, bring the difference boundary-value problem (12) to canonical form (2), while conforming to conditions 1° and 2°, if we take as U_h' the set of vector functions

$$u^p = \begin{bmatrix} u_m^{p+1} \\ \\ u_m^p \end{bmatrix}, \quad m = 0,\ 1,\ \ldots,\ M.$$

But we then complicate the construction of the operator, R_h, and the estimation of norms of its powers. For this reason we will not consider such a reduction process.

* * *

Let us show that, for our choice (13) of the operator R_h, there exists a z^p, satisfying condition 1*:

$$\left|\left|\rho^p - \frac{1}{\tau}\left(z^{p+1} - R_h z^p\right)\right|\right| < K_1 ||f^{(h)}||_{F_h}.$$

To construct the function $z^{(h)}$ we proceed very much as in example 1, writing out the stationary (p-independent) solution of the problem

$$\left.\begin{aligned} &\frac{z_m^{p+1} - z_m^p}{\tau} - \frac{z_{m+1}^p - z_{m-1}^p}{2h} - \frac{\tau}{2}\frac{z_{m+1}^p - 2z_m^p + z_{m-1}^p}{h^2} = 0, \\ &\qquad\qquad m = 1, 2, \ldots, M-1, \\ &\frac{z_0^{p+1} - z_0^p}{\tau} - \frac{z_1^p - z_0^p}{h} = \phi_0^p, \\ &z_M^p = \psi_1^p, \end{aligned}\right\} \qquad (14)$$

postulating that ϕ_0^p and ψ_1^p are fixed, and do not depend on p.

This solution has the form

$$z_m^p = \phi_0^p\, h\, \frac{1+r}{2}\left[\left(\frac{r-1}{r+1}\right)^m - \left(\frac{r-1}{r+1}\right)^M\right] + \psi_1^p.$$

The function $\{z_m^p\}$ satisfies the bound

$$||z^p|| \le \tilde{K}||f^{(h)}||_{F_h}, \quad K = 2.$$

In fact

$$||z^p|| = \left(\frac{h}{2}|z_0^p|^2 + h\sum_{m=1}^{M}|z_m^p|^2\right)^{1/2} \le$$

$$\le 2h \max_m|\phi_m^p| + \max_p|\psi_1(t_p)| \le 2||f^{(h)}||_{F_h}.$$

Let

$$\xi^p \equiv z^{p+1} - R_h z^p.$$

Since z^p is the solution of a time-independent problem we may write

$$z^p = R_h z^p + \tau\delta^p,$$

where

$$\delta^p = \left(\phi_0^p,\ 0,\ 0,\ \ldots,\ 0,\ \frac{\psi_1(t_p)}{\tau}\right).$$

Therefore

$$\xi^p = z^{p+1} - z^p + \tau\delta^p,$$

so that

$$\xi_m^p = \begin{cases} h(\phi_0^{p+1} - \phi_0^p)\,\dfrac{1+r}{2}\left[1 - \left(\dfrac{r-1}{r+1}\right)^M\right] + \\ \qquad + (\psi_1(t_{p+1}) - \psi_1(t_p)) + \tau\phi_0^p, & m = 0. \\ h(\phi_0^{p+1} - \phi_0^p)\,\dfrac{1+r}{2}\left[\left(\dfrac{r-1}{r+1}\right)^m - \left(\dfrac{r-1}{r+1}\right)^M\right] + \\ \qquad + (\psi_1(t_{p+1}) - \psi_1(t_p)), & 1 < m < M, \\ 0, & m = M. \end{cases}$$

Therefore the coordinates of the vector $\tilde{\rho}^p = \rho^p - \xi^p/\tau$ have the form

$$\tilde{\rho}_m^p = \begin{cases} -\dfrac{\psi_1(t_{p+1}) + \psi_1(t_p)}{\tau} - \dfrac{h}{\tau}(\phi_0^{p+1} - \phi_0^p)\,\dfrac{1+r}{2}\left[1 - \left(\dfrac{r+1}{r+1}\right)^M\right], \\ \qquad m = 0, \\ -\dfrac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} - \dfrac{h}{\tau}(\phi_0^{p+1} - \phi_0^p)\,\dfrac{1+r}{2}\left[\left(\dfrac{r-1}{r+1}\right)^m - \right. \\ \qquad \left. - \left(\dfrac{r-1}{r+1}\right)^M\right] + \phi_m^p, & 0 < m < M, \\ 0. & m = M. \end{cases}$$

The inequality constituting condition 1*

$$||\tilde{\rho}^p|| = \left|\left|\rho^p - \frac{1}{\tau}\left(z^{p+1} - R_h z^p\right)\right|\right| \le K_1 ||f^{(h)}||_{F_h},$$

is satisfied:

$$||\tilde{\rho}^p|| = \left(\frac{h}{2}|\tilde{\rho}_0^p|^2 + h\sum_{m=1}^{M-1}|\tilde{\rho}_m^p|^2\right)^{1/2}$$

$$\leq \left|\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}\right| + \frac{h|\phi_0^{p+1} - \phi_0^p|}{\tau} + \max_m|\phi_m^p| \leq ||f^{(h)}||_{F_h}.$$

Condition 2° is satisfied,

$$||u^0|| \leq K_2||f^{(h)}||_{F_h}$$

since, clearly

$$||u^0|| = \left(\frac{h}{2}|u_0^2| + h\sum_{m=1}^{M}|u_m^0|^2\right)^{1/2} \leq \left(h\sum_{m=0}^{M}|u_m^0|^2\right)^{1/2} \leq ||f^{(h)}||_{F_h}.$$

To prove that the proposed scheme is stable, which it is for $r \leq 1$, it is still necessary to prove that, under this condition,

$$||R_h^p|| \leq K, \quad p = 1, 2, \ldots, [T/\tau], \tag{15}$$

where K is some constant independent of h. We will prove later that for any vector, $u = (u_0, u_1, \ldots, u_M)$, whose last component u_M is equal to zero, we have the inequality

$$||R_h u|| \leq ||u||. \tag{16}$$

Applying the operator R_h to the vector $u = (0, 0, \ldots, 1)$ we get the vector $(0, 0, \ldots, 0, 1/2r + 1/2r^2, 0)$, whose norm does not exceed $\sqrt{h}$. Therefore for an arbitrary vector $u = (u_0, u_1, \ldots, u_M)$, whose component u_M is not necessarily zero we may write (taking account of inequality (16), valid for a vector of the form $u = (u_0, u_1, \ldots, u_{M-1}, 0)$),

$$||R_h u|| = ||R_h(u_0, u_1, \ldots, u_{M-1}, 0) + u_M R_h(0, 0, \ldots, 0, 1)|| \leq$$

$$\leq ||R_h(u_0, u_1, \ldots, u_{M-1}, 0)|| + |u_M|\sqrt{h} \leq$$

$$\leq ||(u_0, u_1, \ldots, u_{M-1}, 0)|| + ||(0, 0, \ldots, u_M)|| \leq 2||u||,$$

$$||R_h|| \leq 2. \tag{17}$$

Now we prove inequality (15). In view of the definition of the operator R_h the vector $v = R_h u$ has the vanishing component v_M, $v_M = 0$. Therefore, using (16) and (17), we get

$$||R_h^p u|| = ||R_h^{p-1}(R_h u)|| = ||R_h^{p-1} v|| \le ||v|| =$$

$$= ||R_h u|| \le 2||u||, \quad ||R_h^p|| \le 2 \equiv K.$$

It remains for us to justify inequality (16) upon which we have relied, i.e. to prove the following proposition.

Suppose $u = (u_0, u_1, \ldots, u_{M-1}, 0)$ is an arbitrary vector whose last component, u_M, is equal to zero, and let $v \equiv R_h u$. Then $||v|| \le ||u||$; i.e.

$$\left(\frac{h}{2} v_0^2 + h \sum_{m=1}^{M} v_m^2\right)^{1/2} \le \left(\frac{h}{2} u_0^2 + h \sum_{m=1}^{M} u_m^2\right)^{1/2}. \tag{18}$$

Recall that, by virtue of definition (13) of the operator R_h

$$\left.\begin{aligned} v_0 &= (1 - r)u_0 + ru_1, \\ v_m &= \left(-\frac{r}{2} + \frac{r^2}{2}\right)u_{m-1} + (1 - r^2)u_m + \left(\frac{r}{2} + \frac{r^2}{2}\right)u_{m+1}, \\ &\qquad\qquad m = 1, 2, \ldots, M-1, \\ v_M &= 0. \end{aligned}\right\}$$

We note the inequality

$$v_m^2 \le \left[\left(-\frac{r}{2} + \frac{r^2}{2}\right)u_{m-1} + (1 - r^2)u_m + \left(\frac{r}{2} + \frac{r^2}{2}\right)u_{m+1}\right]^2 +$$

$$+ \frac{r^2(1 - r^2)}{4}(u_{m-1} - 2u_m + u_{m+1})^2 \equiv$$

$$\equiv \frac{r^2(1 - r)}{2} u_{m-1}^2 + (1 - r^2)u_m^2 + \frac{r^2(1 + r)}{2} u_{m+1}^2 -$$

$$- r(1 - r^2)u_{m-1}u_m + r(1 - r^2)u_m u_{m+1},$$

which is satisfied for $r \le 1$, and also the obvious identity

$$\frac{r^2(1 - r)}{2} + 1 - r^2 + \frac{r^2(1 + r)}{2} = 1.$$

Now for $r \le 1$ it is easy to verify, step by step, the validity of the following chain of inequalities, without requiring that $u_M = 0$:

$$\frac{1}{2}v_0^2 + \sum_{m=1}^{M} v_m^2 \leq \frac{1}{2}\left[u_0(1-r) + u_1 r\right]^2 +$$

$$+ \sum_{m=1}^{M-1}\left[\frac{r^2(1-r)}{2}u_{m-1}^2 + (1-r^2)u_m^2 + \frac{r^2(1+r)}{2}u_{m+1}^2 -\right.$$

$$\left. - r(1-r^2)u_{m-1}u_m + r(1-r^2)u_m u_{m+1}\right] =$$

$$= \frac{1}{2}\left[u_0(1-r) + u_1 r\right]^2 + \sum_{m=1}^{M} u_m^2 +$$

$$+ \left[\frac{r^2(1-r)}{2}u_0^2 - \frac{r^2(1+r)}{2}u_1^2 - r(1-r)u_0u_1\right] - \frac{r^2(1-r)}{2}u_{M-1}^2 =$$

$$= \left[\frac{1}{2}u_0^2 + \sum_{m=1}^{M} u_m^2\right] - \frac{r}{2}\left\{u_0^2 + \left[(1-r)u_0 + ru_1\right]^2\right\} - \frac{r^2(1-r)}{2}u_{M-1}^2 \leq$$

$$\leq \frac{1}{2}u_0^2 + \sum_{m=1}^{M} u_m^2.$$

The resulting energy inequality

$$\frac{1}{2}v_0^2 + \sum_{m=1}^{M} v_m^2 \leq \frac{1}{2}u_0^2 + \sum_{m=1}^{M} u_m^2$$

is stronger than the inequality (18) which we set out to prove.

Thus we have established the stability of scheme (12) for $r \leq 1$. For $r > 1$ we do not have stability for any reasonable choice of norm, since the necessary stability condition of Courant, Friedrichs and Levy is violated.

§43. Some methods for bounding norms of powers of operators

In §§41 and 42 it was shown that evolutional difference schemes

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

ordinarily can be brought to the form

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau\rho^p, \\ u^0 &\text{ given} \end{aligned}\right\} \tag{2}$$

such that stability will be equivalent to the boundedness, uniform in h, of the norms of powers of the transition operator

$$||R_h^p|| < K, \quad p = 1, 2, \ldots, [T/\tau]. \tag{3}$$

Since condition (3) is equivalent to stability, it follows that any method for studying stability is also a method for testing whether or not inequality (3) is satisfied.

Here we present some approaches to the study of stability already encountered in Chapter 8 (but now viewed as methods for bounding the norms of powers of operators) and bring out new aspects of these approaches.

1. Necessary spectral conditions for the boundedness of $||R_h^p||$. Suppose λ_h is any eigenvalue of the operator R_h and $u^{(h)}$ is the corresponding eigenvector, $R_h u^{(h)} = \lambda_h u^{(h)}$. Then

$$||R_h^p u^{(h)}|| = |\lambda_h|^p ||u^{(h)}||,$$

and therefore $||R_h^p|| \geq |\lambda_h|^p$. Since λ_h is an arbitrary eigenvalue, then

$$||R_h^p|| \geq [\max|\lambda_h|]^p, \quad p = 1, 2, \ldots, [T/\tau], \tag{4}$$

where $\max|\lambda_h|$ is the largest absolute value of the eigenvalues of operator R_h. From (4) it is obvious (see §15) that, for (3) to be satisfied, there must be a circle

$$|\lambda| \leq 1 + c\tau \tag{5}$$

in the complex plane, with constant c independent of h, containing all the eigenvalues of operator R_h.

The above considerations do not become essentially more complicated, and the results remain unchanged, if we take as λ_h not only the eigenvalues of the operator R_h, but all the points of its spectrum. If U_h is a finite-dimensional space the spectrum of the operator R_h does not depend on one's choice of norm, and consists entirely of eigenvalues. This is the most important case, which arises naturally in the approximation of differential boundary-value problems in bounded domains by difference problems on a net, D_h, consisting of a finite set of points. In this case condition (5) is necessary for the validity of (3), independent of the choice of norm. If the necessary spectral criterion for stability is not satisfied the problem is hopelessly unstable, and the situation cannot be corrected by any reasonable choice of norms. An analogous situation was investigated in detail for the case of ordinary difference equations in §15.

Let us now clarify the connection between the Von Neumann spectral criterion for the stability of the Cauchy difference problem, considered in §25, and the spectral criterion (5) for the uniform boundedness, (3), of the norms of powers of the operator R_h. We may use for this purpose, for example, the difference scheme

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = \phi(x_m, t_p), \\ &u_m^0 = \psi(x_m), \\ &m = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau]-1, \end{aligned}\right\} \tag{6}$$

approximating the Cauchy problem

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \phi(x,t), \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &u(x,0) = \psi(x). \end{aligned}\right\}$$

Stability of this difference scheme was studied, with the aid of the Von Neumann criterion, in §25.

We now write the above scheme in canonical form (2), defining R_h, $v = R_h u$, and ρ^p via the expressions

$$\left.\begin{aligned} &v_m = (1 - r)u_m + ru_{m+1}, \qquad r = \tau/h, \\ &\rho^p = \phi^p\{\Xi\, \phi(x_m, t_p)\}. \end{aligned}\right\}$$

Define the norm in U_h' by the equation $||u|| = \sup_m |u_m|$. Then the functions $u = \{u_m\} = \{\exp(i\alpha m)\}$, for any real α, belong to the space U_h' and are eigenfunctions of the operator R_h:

$$R_h u = (1 - r)e^{i\alpha m} + re^{i\alpha(m+1)} = [(1 - r) + re^{i\alpha}]e^{i\alpha m} = \lambda(\alpha)u,$$

where

$$\lambda(\alpha) = 1 - r + re^{i\alpha} \tag{7}$$

are the eigenvalues. The stability condition (5), in view of the fact that $\lambda(\alpha)$ is independent of τ, reduces to the requirement that $|\lambda(\alpha)| \leq 1$, which is satisfied for all real α if $r \leq 1$.

As shown in §25 condition (5), in the case of the Cauchy problem for a two-level difference scheme in one net function, is not only necessary, but also sufficient for stability if the norm is defined by the equation

$$||u|| = \left(h \sum_{m=-\infty}^{\infty} u_m^2\right)^{1/2}.$$

(in this case the functions $\{\exp(i\alpha m)\}$ do not belong to the space $U_h^{\prime}$ and, consequently, are not eigenfunctions, but the points (7) still belong to the spectrum of the operator R_h).

2. Spectral criterion for the boundedness of powers of a selfadjoint operator. Suppose that the M-dimensional linear space $U_h^{\prime}$ consists of functions defined at the points $P_1, P_2, \ldots, P_M$ of a net (a net which, for our purposes, may equally well lie on an interval or surface, or in a space) and that one has introduced in $U_h^{\prime}$ a scalar product which, for any pair of functions u and v in $U_h^{\prime}$, we designate as (u,v). Suppose, further, that the operator R_h is bounded, uniformly in h, by some constant c_1:

$$||R_h|| < c_1, \tag{8}$$

and maps space $U_h^{\prime}$ onto some subspace $\tilde{U}_h^{\prime}$ of $U_h^{\prime}$, of dimension $N \leq M$, while on the subspace $\tilde{U}_h^{\prime}$ operator R_h is selfadjoint, i.e. $(R_h u,v) = (u,R_h v)$ for any pair of functions u and v in $\tilde{U}_h^{\prime}$. As is known from linear algebra, in subspace $\tilde{U}_h^{\prime}$ there exists, in this case, an orthonormal basis

$$\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(N)}, \tag{9}$$

consisting of the eigenvectors of operator R_h. Designate by $\lambda_1, \lambda_2, \ldots, \lambda_N$ the corresponding (real) eigenvalues:

$$R_h\psi^{(k)} = \lambda_k\psi^{(k)}, \quad k = 1, 2, \ldots, N. \tag{10}$$

Theorem 1. *For bound (3) to be satisfied it is necessary and sufficient that*

$$\max_k|\lambda_k| \leq 1 + c_2\tau, \quad c_2 = \text{const}. \tag{11}$$

Proof. Necessity has been proven in Sect. 1 above. Let us now prove sufficiency. Suppose u is in $U_h^{\prime}$. Expand the vector $R_h u = v$ (where v is in $\tilde{U}_h^{\prime}$) in basis vectors (9):

$$v = \sum \alpha_k\psi^{(k)}.$$

Then, by (10),

$$R_h^p u = R_h^{p-1}v = \sum_{k=1}^{M} (\lambda_k^{p-1}\alpha_k)\psi^{(k)},$$

$$||R_h^p u|| = \left(\sum_{k=1}^{M} |\alpha_k \lambda_k^{p-1}|^2 \right)^{1/2} \leq$$

$$\leq \max_k |\lambda_k^{p-1}| \left(\sum_{k=1}^{M} \alpha_k^2 \right)^{1/2} = \max_k |\lambda_k^{p-1}| \cdot ||v||. \tag{12}$$

Noting that, by Eq. (8),

$$||v|| = ||R_h u|| \leq c_1 ||u||,$$

and taking account of condition (11), from bound (12) we deduce (3):

$$||R_h^p|| \leq c_1 \max_k |\lambda_k^{p-1}| \leq c_1 (1 + c_2\tau)^{p-1} \leq c_1 (1 + c_2\tau)^{T/\tau} < c_1 e^{c_2 T} = K.$$

Below we establish some criteria for the selfadjointness of operator R_h, and point out some methods for bounding eigenvalues.

3. Selfadjointness criteria. We now introduce the notation

$$[u,v] = \frac{1}{M} \sum u(P_k) v(P_k) \tag{13}$$

and assume that the scalar product in the space U_h' is defined by the equation

$$(u,v) = [u,v]. \tag{14}$$

Suppose, further, that the operator R_h, $b = R_h a$, is given via the expressions

$$b(P_k) = \sum_{P_s} \alpha_{ks} a(P_s),$$

where P_s and P_k run through whole set of net-points. Operator R_h is selfadjoint if and only if

$$\alpha_{ks} \equiv \alpha_{sk}. \tag{15}$$

In an interpretation independent of the numbering of points, this criterion means that, in the computation of b(P) at any arbitrary point P of the net, the value a(Q) at another arbitrary point, Q, must enter with the same coefficient with which the value of a(P) occurs in the expression for b(Q).

If, among the net points $\{P_k\}$, one has singled out some subset Γ_h (the boundary of the region) and the operator R_h is given by the expressions

$$\left.\begin{aligned} b(P_k) &= \sum_{P_s} \alpha_{ks} a(P_s), && P_k \text{ not in } \Gamma_h, \\ b(P_k) &= 0, && P_k \text{ in } \Gamma_h, \end{aligned}\right\} \tag{16}$$

then on the subspace $\tilde{U}_h'$ Eq. (16) is equivalent to the following:

$$\left.\begin{aligned} b(P_k) &= \sum_{P_s \text{ not in } \Gamma_h} \alpha_{ks} a(P_s), && \text{if } P_k \text{ is not in } \Gamma_h, \\ b(P_k) &= 0, && \text{if } P_k \text{ is in } \Gamma_h. \end{aligned}\right\}$$

The condition for selfadjointness of the operator R_h *on the subspace* U_h' then consists, as one can easily see, of the equations

$$\alpha_{ks} = \alpha_{sk}, \quad P_s \text{ not in } \Gamma_h, \quad P_k \text{ not in } \Gamma_h. \tag{17}$$

Thus, for example, the operator $b = R_h a$,

$$\left.\begin{aligned} b_k &= (1 - 2r)a_k + r(a_{k-1} + a_{k+1}), \quad k = 1, 2, \ldots, M-1, \\ b_0 &= b_M = 0, \end{aligned}\right\}$$

which occurs when the difference analog of the heat equation on an interval is reduced to canonical form (2), satisfies condition (17), but not condition (15).

4. Bounds on the eigenvalues of operator R_h. In certain cases the eigenvalues can be written out exactly, as was done in §27 for the operator Λ_{xx} acting on functions given at the points of a net-segment, and vanishing at its endpoints; and also for the operator $\Lambda_{xx} + \Lambda_{yy}$ on functions defined over a net rectangle, and vanishing on its sides.

To deal with selfadjoint difference operators one can make use of variational methods. It is known that, in this case

$$\min_{u' \text{ in } U_h'} \frac{(R_h u', u')}{(u', u')} = \lambda'_{\min}, \qquad \max_{u' \text{ in } U_h'} \frac{(R_h u', u')}{(u', u')} = \lambda'_{\max}. \tag{18}$$

Suppose, for example, the operator $\Lambda_{xx} + \Lambda_{yy}$ acts on net functions of space U_h', functions which are defined, not on a square, but in a more complicated region composed of squares, and which vanish on the boundary of this region. Let us put the region into a square net-region large enough to contain it, and consider the operator $\Lambda_{xx} + \Lambda_{yy}$ acting on the functions of U_h'', defined in the net-square and vanishing on its boundary.

We now extend the definition of each function u' of U_h' so that it becomes a function u'' of U_h'', setting the extended functions identically equal to zero on all those points of the net-square which do not belong to the original region. It is easy to see that for each such function, because of the fact that it vanishes on the boundary of the original region, we may write the inequality

$$\frac{(R_h u', u')}{(u', u')} = \frac{(R_h u'', u'')}{(u'', u'')}, \qquad R_h = \Lambda_{xx} + \Lambda_{yy}.$$

Therefore, in going from Eqs. (18) to the equations

$$\min_{u'' \text{ in } U_h''} \frac{(R_h u'', u'')}{(u'', u'')} = \lambda''_{\min},$$

$$\max_{u'' \text{ in } U_h''} \frac{(R_h u'', u'')}{(u'', u'')} = \lambda''_{\max}$$

we get numbers, $\lambda''_{\min}$ and $\lambda''_{\max}$, which satisfy the bounds

$$\lambda''_{\min} \leq \lambda'_{\min}, \qquad \lambda'_{\max} \leq \lambda''_{\max}. \tag{19}$$

But in the case of the square region the eigenvalues are known, so that $\lambda''_{\min}$ and $\lambda''_{\max}$ are known, and we get bounds (19) on the boundaries of the spectrum of the operator $\Lambda_{xx} + \Lambda_{yy}$ acting on the functions of U'_h, defined in the original region.

In many cases we may bound eigenvalues through use of variational methods analogous to those for differential equations. For example the first eigenvalue of the problem

$$\tilde{R}_h u^{(h)} = \lambda u^{(h)}, \qquad u^{(h)}\big|_{\Gamma_h} = 0,$$

where Γ_h is the boundary of net-region D_h and where, at interior points,

$$\tilde{R}_h u_{mn} \equiv \frac{1}{h^2}\left\{a\left(x_m + \frac{h}{2}, y_n\right)(u_{m+1,n} - u_{mn}) - a\left(x_m - \frac{h}{2}, y_n\right)(u_{mn} - u_{m-1,n})\right\} + \frac{1}{h^2}\left\{b\left(x_m, y_n + \frac{h}{2}\right)(u_{m,n+1} - u_{mn}) - b\left(x_m, y_n - \frac{h}{2}\right)(u_{mn} - u_{m,n-1})\right\}, \qquad a(x,y) > 0, \quad b(x,y) > 0,$$

can only decrease when the variable coefficients, a(x,y) and b(x,y) are replaced by the constants

$$a = \max_{x,y} a(x,y), \qquad b = \max_{x,y} b(x,y)$$

This may be shown by exactly the same methods by which one reaches the analogous conclusion for differential equations (Ref. 19).

In the case of constant coefficients one may go from the original region to a square, and get bounds similar to bounds (19). Eigenvalues of the operator $a\Lambda_{xx} + b\Lambda_{yy}$ in the square region are easy to calculate exactly.

5. Choice of a scalar product. Suppose that the operator R_h, $v = R_h u$, is given by the equation

$$B_h v = \tilde{A}_h u, \tag{20}$$

and, for some particular choice of scalar product $(u,v) = [u,v]$ not necessarily given by Eqs. (13) and (14), the operators $\tilde{A}_h$ and B_h are selfadjoint

$$[\tilde{A}_h u,\ v] \equiv [u,\ \tilde{A}_h v], \qquad [B_h u,\ v] \equiv [u,\ B_h v].$$

Suppose, further, that $B_h > 0$:

$$[B_h u,\ u] > 0, \quad \text{if } u \neq 0.$$

Then the operator $R_h = B_h^{-1}\tilde{A}_h$ is selfadjoint in the sense of the scalar product

$$(u,v)_{B_h} \equiv [B_h u,\ v]. \tag{21}$$

In fact

$$(B_h^{-1}\tilde{A}_h u,\ v)_{B_h} = [B_h(B_h^{-1}\tilde{A}_h)u,\ v] = [\tilde{A}_h u,\ v] =$$
$$= [u,\ \tilde{A}_h v] = [B_h^{-1}B_h u,\ \tilde{A}_h u] = [B_h u,\ B_h^{-1}\tilde{A}_h v] = (u,\ B_h^{-1}\tilde{A}_h v)_{B_h}.$$

The above identity in u and v

$$(B_h^{-1}\tilde{A}_h u,\ v)_{B_h} \equiv (u,\ B_h^{-1}\tilde{A}_h v)_{B_h}$$

means, precisely, that the operator R_h is selfadjoint.

Thus the choice of scalar product via Eq. (21) allows us to use the spectral criterion of Sect. 2 for the boundedness of norms of powers of selfadjoint operators. Specifically, one can affirm that the operator R_h, defined by Eq. (20), has real eigenvalues λ_k, and a complete system of eigenvectors $\psi^{(k)}$:

$$\lambda_k B_h \psi^{(k)} = \tilde{A}_h \psi^{(k)}, \tag{22}$$

and that disposition of all the eigenvalues λ_k on the segment $-1 \leq \lambda \leq 1$ is necessary and sufficient for the validity of the inequality

$$||R_h^p||_{B_h} \leq 1, \tag{23}$$

where the norm of the operator is defined with the aid of the scalar product (21).

6. The stability criterion of Samarskii. A. A. Samarskii, in his stability theory for a wide class of difference schemes in Hilbert space (Refs. 23 and 24), presents necessary and sufficient conditions for stability in terms of linear inequalities between the operator-coefficients in these schemes, and also discusses other results. We give, here, only two results of this theory.

Suppose $U_h^{\wedge}$ is a Euclidian space with some scalar product $(u,v) \equiv [u,v]$, and let the operator R_h, $v = R_h u$, u, v in $U_h^{\wedge}$, be given by the equation

$$B_h \frac{v - u}{\tau} + A_h u = 0, \tag{24}$$

where A_h and B_h are selfadjoint operators with $B_h > 0$. Define an energy norm $||u||_{B_h}$ in space $U_h^{\wedge}$, setting

$$||u||_{B_h}^2 = [B_h u, u] \equiv (u, u)_{B_h}. \tag{25}$$

One may now affirm the following

Theorem 2. *The condition*

$$0 \leq A_h \leq \frac{2}{\tau} B_h \tag{26}$$

is necessary and sufficient for the validity of the inequality

$$||R_h^p|| \leq 1, \qquad p \geq 0. \tag{27}$$

Proof. Let us define the selfadjoint operator $\tilde{A}_h$, $\tilde{A}_h \equiv B_h - \tau A_h$. Then (24) is equivalent to (20), and condition (26) is equivalent to the condition $-B_h \leq \tilde{A}_h \leq B_h$, i.e. to the condition

$$-[B_h u, u] \leq [\tilde{A}_h u, u] \leq [B_h u, u]. \tag{28}$$

As shown in Sect. 5, above, the operator R_h is selfadjoint in the sense of the scalar product (21), and the assertion of the theorem is equivalent to the assertion that all the eigenvalues, λ_k, of the operator R_h lie on the interval $-1 \leq \lambda \leq 1$ if and only if condition (28) is satisfied. Let us now prove this last assertion.

Suppose condition (28) is satisfied. Computing the scalar product of Eq. (22) with the eigenvector $\psi^{(k)}$ of operator R_h, we get

$$\lambda_k [B_h \psi^{(k)}, \psi^{(k)}] = [\tilde{A}_h \psi^{(k)}, \psi^{(k)}].$$

from which

$$|\lambda_k| = \left| \frac{[\tilde{A}_h \psi^{(k)}, \psi^{(k)}]}{[B_h \psi^{(k)}, \psi^{(k)}]} \right| \leq 1.$$

Conversely, let $\max|\lambda_k| \leq 1$. We now show that condition (28) is satisfied. Let $u = \Sigma c_k \psi^{(k)}$ be the expansion of any arbitrary element u, (u in U_h') in the basis $\{\psi^{(k)}\}$, orthonormal in the sense of the scalar product (21). Then

$$|[\tilde{A}_h u, u]| = |[\tilde{A}_h \sum c_k \psi^{(k)}, \sum c_k \psi^{(k)}]| = |[\sum c_k \tilde{A}_h \psi^{(k)}, \sum c_k \psi^{(k)}]| =$$

$$= |[\sum c_k \lambda_k B_h \psi^{(k)}, \sum c_k \psi^{(k)}]| = |[B_h \sum c_k \lambda_k \psi^{(k)}, \sum c_k \psi^{(k)}]| =$$

$$= \left| \left(\sum c_k \lambda_k \psi^{(k)}, \sum c_k \psi^{(k)} \right)_{B_h} \right| = \sum c_k^2 |\lambda_k| \leq$$

$$\leq \sum c_k^2 = (u,u)_{B_h} = [B_h u, u].$$

Therefore $[B_h u, u] \geq |[\tilde{A}_h u, u]|$, which is equivalent to condition (28). The theorem is proven.

Note that verification of condition (28) is equivalent to a determination as to whether or not all the eigenvalues of the operators $B_h - \tilde{A}_h$ and $B_h + \tilde{A}_h$ (selfadjoint in the sense of the scalar product $[u,v]$) are nonnegative.

* * * * * *

Finally we introduce, without proof, still another stability criterion, applicable to difference schemes (24) with $B_h > 0$, $A_h = A_h^* > 0$. Let us introduce, in space U_h', an energy norm $||u||_{A_h}$, setting $||u||_{A_h} \equiv [A_h u, u]$.

Theorem 3. *The condition* $B_h \geq \frac{1}{2} \tau A_h$ *is necessary and sufficient for the validity of the inequality* $||R_h||_{A_h} \leq 1$.

Theorem 3 is contained in Sect. 4§1 of Chapter 6, Ref. 23, and may be proven without the help of the spectral approach, here inapplicable because the operator B_h is not (necessarily) selfadjoint.

* * *

PROBLEMS

1. Suppose the operator R_h, $b = R_h a$, is given by the equations

$$\left.\begin{aligned} b_m &= (1 - r)a_m + ra_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ b_M &= 0. \end{aligned}\right\}$$

Prove that in space U_h of net functions $\{a_m\}$, $m = 0, 1, \ldots, M$, it is impossible to define a scalar product such that the operator R_h will become selfadjoint.

This Page Intentionally Left Blank

Chapter 14
Spectral Criterion for the Stability of Nonselfadjoint Evolutional Boundary-Value Problems

Here we show that, from the spectrum of a nonselfadjoint operator R_h, one cannot judge stability of a difference boundary-value problem in a bounded region; we introduce the concept of a of operators $\{R_h\}$ and consider the spectral formulation of the question of stability, which remains meaningful also in the case of nonselfadjoint boundary-value problems in bounded regions. We will point out a necessary, and close to sufficient spectral criterion for stability.

§44. Spectrum of a family of operators $\{R_h\}$.

1. Need for improvement in the spectral stability criterion. In Chapter 13 it was shown that, ordinarily, evolutional difference boundary-value problems may be brought to the form

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau\rho^p, \\ u^0 &\text{ given,} \end{aligned}\right\} \tag{1}$$

so that stability on the time-interval $0 \leq t \leq T$ will be equivalent to the uniform (in h) boundedness of the norms of powers of the transition operator R_h, i.e. equivalent to the bound

$$\|R_h^p\| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{2}$$

where τ is the net timestep, $\tau=\tau(h)$.

It was established that confinement of the eigenvalues of the operator R_h, inside the circle

$$|\lambda| < 1 + c\tau \tag{3}$$

in the complex plane is necessary for the validity of (2), i.e. for stability. In §43 it was shown that, in the case of a selfadjoint operator R_h, condition (3) is not only a necessary, but also a sufficient condition for the uniform boundedness (2) of the norms of powers of the operator

R_h. This same fact was established in §25 for the Cauchy difference problem with constant coefficients, for two-level difference schemes in one unknown function, with no reference to selfadjointness. But, in the general case of a nonselfadjoint difference boundary-value problem in a bounded region, the necessary criterion (3) is very far from sufficient, and is totally inadequate in dealing with the question of uniform boundedness of norms, $\|R_h^p\|$, of powers of the operator R_h. This may be shown by the following example.

Example. For the difference boundary-value problem

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = \phi(x_m, t_p), \\ &u_M^p = 0, \qquad p = 0, 1, \ldots, [T/\tau], \\ &u_m^0 = \psi(x_m), \qquad m = 0, 1, \ldots, M;\ Mh = 1, \end{aligned}\right\} \tag{4}$$

approximating the problem

$$u_t - u_x = \phi(x,t),$$

$$u(1,t) = 0, \quad 0 < x < 1, \quad 0 < t < T,$$

$$u(x,0) = \psi(x)$$

the natural reduction to canonical form (1) leads to an operator R_h, $v = R_h u$, given by the equations

$$v_m = (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M-1,$$

$$v_M = 0, \quad r = \tau/h.$$

In matrix form

$$R_h = \begin{bmatrix} 1-r & r & 0 \ldots & 0 & 0 \\ 0 & 1-r & r \ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \ldots & 1-r & r \\ 0 & 0 & 0 \ldots & 0 & 0 \end{bmatrix}. \tag{5}$$

The spectrum of the matrix consists of its eigenvalues, i.e. of the roots of the equation

$$\det(R_h - \lambda E) = 0 \quad \text{or} \quad -\lambda(1 - r - \lambda)^M = 0.$$

Thus the roots of this equation, $\lambda = 0$ and $\lambda = 1$, form the spectrum of the operator R_h for any h. This spectrum lies inside the unit circle $|\lambda| = 1$ for $0 < r \leq 2$. Nevertheless, for scheme (4) with $1 < r \leq 2$ the Courant-Friedrichs-Levy condition is not satisfied, so that stability, $||R_h^p|| < K$, is impossible in any reasonable norm.

* * * * * *

In fact we will show that, in the case $T \geq 1$ with norm $||u|| = \max_m |u_m|$ we have the inequality

$$\max_{p=1,\ 2,\ \ldots,\ [T/\tau]} ||R_h^p|| \geq |1 - 2r|^{1/h} = \rho^{1/h}. \tag{6}$$

For $r > 1$ also $\rho > 1$, so that as $h \to 0$ and $\tau = rh \to 0$ the quantity $\max_p ||R_h^p||$ increases exponentially and the condition $||R_h^p|| < K$ is grossly violated. To prove inequality (6) we note that, in the case $u_m^0 = (-1)^m$, $m = 0, 1, \ldots, M$, the values u_m^p of the function

$$u^p = R_h^p u^0, \quad p = 1, 2, \ldots, M \text{ and } m = 0, 1, \ldots, M-p$$

are given by the equations

$$u_m^p = (-1)^m (1 - 2r)^p, \quad m = 0, 1, \ldots, M-p.$$

Therefore

$$||R_h^p u^0|| \geq |1 - 2r|^p ||u^0||, \quad p \leq M,$$

so that for these values of p, $p = 1, 2, \ldots, M$,

$$||R_h^p|| \geq |1 - 2r|^p, \quad p = 1, 2, \ldots; \ M = 1/h,$$

and inequality (6) has been proven.

* * *

Thus it has been established that the necessary spectral criterion (3) for uniform boundedness $||R_h^p|| < K$, using the eigenvalues of the operator R_h, is too coarse when the operator R_h is nonselfadjoint: in our example it does not detect the instability that occurs for $1 < r \leq 2$.

2. Definition of the spectrum of a family of operators. Suppose the linear operator R_h is defined on a linear normed space U_h. We designate by

$\{R_h\}$ the set of operators R_h for all values taken on by the parameter, h, characterizing the density of the net. By the nature of difference schemes h can have positive values as small as we like.

The complex number λ will be called a "point of the spectrum of the family of operators $\{R_h\}$" if, for any positive h_0 and ε, one can find an h, $h < h_0$, for which the inequality

$$||R_h u - \lambda_u|| < \varepsilon||u||$$

has some solution u, u in U_h'.

The set of all such numbers λ we will call the "spectrum of the family of operators $\{R_h\}$".

3. Necessary condition for stability.

Theorem 1. *Suppose that at least one point λ_0 of the spectrum of the family of operators $\{R_h\}$ lies outside the unit circle in the complex plane, so that $|\lambda_0| > 1$. In this case it is impossible to find one constant K, the same for all h, such that*

$$||R_h^p|| < K, \tag{7}$$

in which p runs through the integral values from 0 to $p_0(h)$, where $p_0(h) \to \infty$ as $h \to 0$.

Proof. Let us first assume that there does not exist an $h_0 > 0$ and a $c > 0$ such that, for all $h < h_0$, we have the bound

$$||R_h|| < c. \tag{8}$$

Under this assumption the assertion to be proven is obvious. Therefore we need to consider only the case where there exist values $h_0 > 0$ and $c > 0$ such that, for $h < h_0$, inequality (8) is valid.

Suppose $|\lambda_o| = 1 + \delta$, where λ_0 is that point of the spectrum for which $|\lambda_0| > 1$. Given an arbitrary number K, we choose p and ε such, as to satisfy the inequalities

$$(1 + \delta)^p > 2K,$$

$$1 - (1 + c + c^2 + \ldots + c^{p-1})\varepsilon > \frac{1}{2}.$$

By the definition of a point of the spectrum of a family of operators $\{R_h\}$, one can find an arbitrarily small positive h for which there exists a vector, u in U_h, which is a solution of the equation

$$||R_h u - \lambda_0 u|| < \varepsilon||u||. \tag{9}$$

Let

It is clear that $||z|| < \varepsilon||u||$. Further, from (10) one can conclude that

$$R_h^p u = \lambda_0^p u + (\lambda_0^{p-1} z + \lambda_0^{p-2} R_h z + \ldots + R_h^{p-1} z).$$

Given that $|\lambda_0| > 1$,

$$||\lambda_0^{p-1} z + \lambda_0^{p-2} R_h z + \ldots + R_h^{p-1} z|| \leq$$

$$\leq |\lambda_0|(1 + ||R_h|| + ||R_h^2|| + \ldots + ||R_h^{p-1}||)\varepsilon||u||,$$

and consequently

$$||R_h^p u|| \geq |\lambda_0|^p[1 - \varepsilon(1 + c + c^2 + \ldots + c^{p-1})]||u|| >$$

$$> (1 + \delta)^p \frac{1}{2} ||u|| \geq 2K \frac{1}{2} ||u|| = K||u||.$$

The number h, throughout this construction, can be considered small enough so that p will be smaller than $p_0(h)$.

Since K was arbitrary we have now proven our assertion that *disposition of all the points of the spectrum of the family of operators* $\{R_h\}$ *within or on the boundary of the unit circle* $|\lambda| \leq 1$ *is necessary for the validity of the bound* $||R_h^p|| < K$.

4. Discussion of the concept of the spectrum of a family of operators $\{R_h\}$. We begin by turning the reader's attention to the analogy between the definition of a point of the spectrum of a family of operators $\{R_h\}$, and the following definition of a point of the spectrum of any operator R (a definition commonly introduced in courses on functional analysis). As the operator R, we take the operator R_h for some fixed h.

The point λ in the complex plane is called *a point in the spectrum of the operator* R_h if, for any positive ε, the inequality

$$||R_h u - \lambda u|| < \varepsilon||u||$$

has a solution, u, belonging to the space U_h', the space on which the operator R_h is defined.

On comparing the definitions of a point in the spectrum of a *family of operators* $\{R_h\}$, and a point in the spectrum of the operator R_h, one may get the impression that the spectrum of the family $\{R_h\}$ consists of those points of the complex plane which are obtained by passage to the limit $h \to 0$ of the points of the spectrum of R_h, where the limit $h \to 0$ is approached by all possible subsequences. But, generally speaking, this impression is erroneous.

Consider the operator R_h, $v = R_h u$, given by the equations

$$\left.\begin{aligned} v_m &= (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ v_M &= 0, \qquad\qquad\qquad Mh = 1. \end{aligned}\right\} \tag{11}$$

The operator (11) acts in a $(M + 1)$ dimensional linear space, and is characterized by matrix (5). It is known that the spectrum of a matrix consists of its eigenvalues, i.e. of the roots λ of the equation $\det(R_h - \lambda E) = 0$. We computed these eigenvalues in Sect. 1; they are $\lambda = 0$ and $\lambda = 1 - r$. Thus the spectrum of the operator R_h, for any h, consists of the two points 0 and $1 - r$, independent of h. But the spectrum of the family of operators $\{R_h\}$, as will be shown in §45, consists not only of these two points as, perhaps, one might expect but, in addition, of all the points of the circle $|\lambda - 1 + r| < r$ of radius r, with center at point $\lambda = 1 - r$ (Fig. 27, p. 270). For $r \leq 1$ the spectrum of the family of operators $\{R_h\}$ lies in the unit circle $|\lambda| < 1$, but for $r > 1$ this necessary condition for stability is not satisfied: the inequality $\|R_h^p\| < K$ cannot hold uniformly in h.

In Fig. 53 we show plots of the dependence of the values of $\|R_h^p\|$ on $p\tau = prh$ in the case $r = 3/2$ for various values of h. In this case the spectrum of each operator R_h consists of both points $\lambda = 0$ and $\lambda = -1/2$, thus lying in the unit circle. This fact predetermines the behavior of the graph $\|R_h^p\|$ for large values of $p\tau$. The value of $\|R_h^p\|$ tends to zero as $p\tau \to \infty$, i.e. the horizontal axis is an asymptote (and in detailed algebra courses it is proven that the norms of powers of a matrix tend to zero as the exponent increases if all the eigenvalues of the matrix are smaller than one in absolute value).

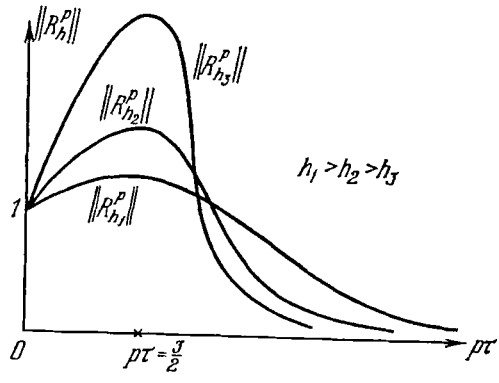


Fig. 53.

The fact that the spectrum of a family of operators $\{R_h\}$ is not totally contained in the unit circle makes its influences felt on the behavior of the values of $\|R_h^p\|$ as $h \to 0$ if $p\tau$ is not too large. The largest value of $\|R_h^p\|$ on the interval $0 < p\tau < T$ (where T is an arbitrary positive constant) grows quickly as h decreases. But this signals instability on the interval $0 < t < T$, while at the same time the behavior of $\|R_h^p\|$ as $p\tau \to \infty$, connected with the behavior of the spectrum of each individual operator R_h, is of no consequence at all in the study of stability.

5. Nearness of the necessary stability criterion to sufficiency

Theorem 2. *Suppose the operator* R_h *is defined on the normed space* U_h, *finite-dimensional for each* h, *and is bounded uniformly in* h *by some constant* c:

$$||R_h|| < c. \tag{12}$$

Suppose, further, that the spectrum of the family of operators $\{R_h\}$ *lies completely inside the closed unit circle* $|\lambda| \leq 1$.

Then for any $\varepsilon > 0$ *the norms of the powers of the operators* R_h^p *satisfy the bound*

$$||R_h^p|| \leq A(\varepsilon)(1 + \varepsilon)^p, \tag{13}$$

where $A = A(\varepsilon)$ *depends only on* ε, *and not on* h.

This theorem means that disposition of the spectrum of the family of operators $\{R_h\}$ in the unit circle is not only necessary for stability, but also guarantees against "gross" instability. If the conditions of the theorem are satisfied the quantity

$$\max_{1 \leq p \leq [T/\tau]} ||R_h^p||$$

either remains bounded as $h \to 0$, or grows more slowly than $\rho^{[T/\tau]}$ for any base, $\rho = 1 + \varepsilon$, greater than unity.

Proof. We show, preliminarily, that if the spectrum of the family of operators $\{R_h\}$ lies in the circle $|\lambda| \leq \rho$ then, for any λ satisfying the inequality $|\lambda| \geq \rho + \varepsilon$, $\varepsilon > 0$, there exists a number $A = A(\varepsilon)$, and a $h_0 > 0$ such that for any $h < h_0$ and any u in U_h, $u \neq 0$, we may write

$$||R_h u - \lambda u|| > \frac{\rho + \varepsilon}{A(\varepsilon)} ||u||. \tag{14}$$

Assume the contrary. Then one can find an $\varepsilon > 0$; a sequence of numbers $h_k > 0$, $h_k \to 0$; of complex numbers λ_k, $|\lambda_k| > \rho + \varepsilon$; and of vectors u_{h_k} in U_{h_k} such that

$$\left|\left|R_{h_k} u_{h_k} - \lambda_k u_{h_k}\right|\right| < \frac{\rho + \varepsilon}{k} \left|\left|u_{h_k}\right|\right|. \tag{15}$$

For large enough values of k, for which $(\rho + \varepsilon)/k < 1$, by virtue of (12) the numbers λ_k cannot lie outside the circle $|\lambda| \leq c + 1$, since outside this circle

$$\left|\left|R_{h_k} u - \lambda u\right|\right| \geq \left(|\lambda| - \left|\left|R_{h_k}\right|\right|\right)||u|| \geq ||u||.$$

Thus the sequence λ_k is bounded, and therefore has a limit point $\tilde{\lambda}$, $|\tilde{\lambda}| \geq \rho + \varepsilon$. One can easily see from (15) that the point $\tilde{\lambda}$ belongs to the spectrum of the family of operators $\{R_h\}$, contradicting our assumption that the spectrum lies in the circle $|\lambda| \leq \rho$.

Suppose, now, that R is a linear operator carrying some finite-dimensional normed space U into itself. And suppose that for any complex λ, $|\lambda| \geq r > 0$, any u in U and some a = const > 0, we may write the inequality

$$||Ru - \lambda u|| \geq a||u||. \tag{16}$$

Then

$$||R^p|| \leq \frac{r^{p+1}}{a}, \quad p = 1, 2, \ldots \tag{17}$$

Inequality (17) follows from the following well-known equality:

$$R^p = -\frac{1}{2\pi i} \oint_{|\lambda|=r} \lambda^p (R - \lambda E)^{-1} \, d\lambda, \tag{18}$$

and from condition (16) which implies that $||(R - \lambda E)^{-1}|| \leq 1/a$. To prove inequality (13) we set $a = (\rho + \varepsilon)/A(\varepsilon)$, $r = \rho + \varepsilon$, $\rho = 1$ and $R = R_h$. Then (17) coincides with (13).

* * * * * *

In conclusion we indicate a proof of Eq. (18). Set

$$u^{p+1} = Ru^p, \quad U(\lambda) = \sum_{p=0}^{\infty} \frac{u^p}{\lambda^p}.$$

Multiply both sides of the equation $u^{p+1} = Ru^p$ by λ^{-p}, then sum over p from p = 0 to p = ∞. One then gets

$$\lambda U(\lambda) - \lambda u^0 = RU(\lambda),$$

or

$$(R - \lambda E)U(\lambda) = -\lambda u^0, \quad U(\lambda) = -\lambda (R - \lambda E)^{-1} u^0.$$

From the definition of $U(\lambda)$ it is clear that u^p is the residue of the vector function $\lambda^{p-1}U(\lambda)$:

$$u^p = \frac{1}{2\pi i} \oint_{|\lambda|=r} \lambda^{p-1} U(\lambda) d\lambda = -\frac{1}{2\pi i} \oint_{|\lambda|=r} \lambda^p (R - \lambda E)^{-1} u^0 \, d\lambda.$$

But $u^p = R^p u^0$, so that the last equation is equivalent to operator equation (18).

* * *

In this section we have stated the spectral formulation of the problem of the stability of evolutional difference schemes, a formulation which is meaningful for any evolutional difference scheme which can be put into the form

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau\rho^p, \\ u^0 &\text{ given} \end{aligned}\right\}$$

in such a way that satisfaction of the condition

$$||R_h^p|| < K, \quad p = 1, 2, \ldots, [T/\tau],$$

would be equivalent to stability. The schemes referred to here may be two-level or multilevel schemes, splitting schemes, etc., for problems on an interval, in multi-dimensional or composite regions.

This spectral formulation requires that one determine whether or not the spectrum of the family of operators $\{R_h\}$ lies in the unit circle $|\lambda| \leq 1$.

§45. Algorithm for the computation of the spectrum of a family of difference operators on net functions in an interval

In this section we describe an algorithm for computing the spectrum of a family of difference operators $\{R_h\}$ on the space of net functions (or vector-functions) defined over an interval. As the norm of the function (or vector-function) we take the maximum of the absolute values taken on by the function (or the components of the vector-function).

1. Typical example. The family of operators $\{R_h\}$, $v = R_h u$, will be defined by the equations

$$\left.\begin{aligned} v_m &= (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ v_M &= 0. \end{aligned}\right\} \tag{1}$$

This operator R_h occurs in the straightforward reduction of the difference boundary-value problem

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = \phi(x_m, t_p), \\ &p = 0, 1, \ldots, [T/\tau]-1, \\ &u_M^{p+1} = 0, \quad u_m^0 = \psi(x_m), \quad m = 0, 1, \ldots, M-1, \end{aligned}\right\} \tag{2}$$

to the form

$$u^{p+1} = R_h u^p + \tau\rho^p, \ u^0 \text{ given.}$$

Equations (2) constitute a difference analogue of the differential boundary-value problem

$$u_t - u_x = \phi(x,t), \quad 0 \leq x \leq 1, \quad 0 \leq t \leq T,$$

$$u(x,0) = \psi(x), \quad u(1,t) = 0.$$

We have already considered difference scheme (2) in 2§26 as an example illustrating the application of the Babenko-Gelfand criterion. It should be recalled that, in using this criterion, the investigation of the original problem, given on an interval, must be split into the study of three auxiliary problems: a problem without lateral boundaries, a problem with only a left-hand boundary and one with only a right boundary, for each of which one must find all the eigenvalues of the transition operator from u^p to u^{p+1}.

It turns out that the algorithm for computing the spectrum of a family of operators $\{R_h\}$ coincides with the Babenko-Gelfand procedure.

In order to describe the algorithm for computing the spectrum of a family of operators $\{R_h\}$, defined by Eqs. (1), we consider three auxiliary operators: $\overleftrightarrow{R}$, $\overrightarrow{R}$ and $\overleftarrow{R}$. The operator $\overleftrightarrow{R}$, $v = \overleftrightarrow{R}u$, is given on the linear space of bounded functions $u = \{\ldots, u_{-1}, u_0, u_1, \ldots\}$ defined on the whole net-line $-\infty < mh < \infty$, by the expression

$$v_m = (1 - r)u_m + ru_{m+1}, \quad m = 0, \pm 1, \ldots \tag{3}$$

This expression is obtained from Eq. (1) by removing the left-hand boundary to $-\infty$ and the right-hand boundary to $+\infty$, a fact reflected in the two-sided arrow of the designation of the operator: $\overleftrightarrow{R}$. The operator $\overrightarrow{R}$, $v = \overrightarrow{R}u$, is given on the linear space of net-functions $u = (u_0, u_1, \ldots, u_m, \ldots)$ defined on the net half-line $x_m = mh$, $m = 0, 1, 2, \ldots$, and tending to zero as $m \to \infty$. It is defined by the equations

$$v_m = (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots \tag{4}$$

These equations are obtained from Eq. (1) by moving the right-hand boundary to $+\infty$, as indicated by the mnemonic sign $\rightarrow$ in the notation for the operator: $\overrightarrow{R}$.

Finally, the operator $\overleftarrow{R}$, $v = \overleftarrow{R}u$, on the functions

$$u = (\ldots, u_m, \ldots, u_{M-1}, u_M), \quad u_m \to 0 \text{ as } m \to -\infty,$$

defined on the net half-line $x_m = mh$, $m = \ldots, -2, -1, 0, 1, \ldots, M$, will be given by the equations

$$\left.\begin{aligned} v_m &= (1 - r)u_m + ru_{m+1}, \quad m = \ldots, -1, 0, 1, \ldots, M-1, \\ v_M &= 0. \end{aligned}\right\} \tag{5}$$

These equations were gotten from (1) by moving the left boundary to $-\infty$, as indicated by the notation for the operator: $\overleftarrow{R}$.

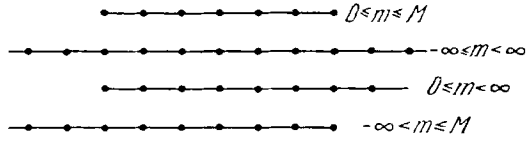


Fig. 54.

We see that the operators $\ddot{R}$, $\overrightarrow{R}$, and $\overleftarrow{R}$ do not depend on h. The domains of definition of the functions $u = \{u_m\}$ for operators (1), (3), (4) and (5) are depicted in Fig. 54. It will be shown that *the set of all eigenvalues of all three operators constitutes the spectrum of the family of operators* $\{R_h\}$.

The eigenvalues of the operators $\ddot{R}$, $\overrightarrow{R}$ and $\overleftarrow{R}$ have already been computed in §26, but we reproduce this computation here because, before going on to a proof of the above assertion, we must have clearly in mind the structure of the eigenfunctions of the operators $\ddot{R}$, $\overrightarrow{R}$, and $\overleftarrow{R}$.

First of all we examine the nature of the set of points, λ, in the complex plane, for which the equation

$$\ddot{R}u - \lambda u = 0$$

has a bounded solution $u = \{u_m\}$, $m = 0, \pm 1, \ldots$ These numbers λ are precisely the eigenvalues of the operator $\ddot{R}$. In our example the equation $\ddot{R}u - \lambda u = 0$ has the form

$$(1 - r - \lambda)u_m + ru_{m+1} = 0, \quad m = 0, \pm 1, \ldots$$

Each solution of this ordinary first-order difference equation, as follows from §1, can differ only by a constant factor from the net function $u = q^m$, $m = 0, \pm 1, \ldots$, where q is a root of the characteristic equation $(1 - r - \lambda) + rq = 0$. The relation between λ and q can also be written in the form

$$\lambda = 1 - r + rq.$$

The solution $q_m = q^m$ is bounded as $m \to +\infty$ and $m \to -\infty$ only if $|q| = 1$, $q = \exp(i\alpha)$, $0 \leq \alpha \leq 2\pi$. Therefore the set of those values of λ for which the solution $u_m = q^m$ is bounded may be obtained from the expression

$$\lambda = 1 - r + rq = 1 - r + re^{i\alpha},$$

when $q = \exp(i\alpha)$ moves over the whole circumference of the circle $|q| = 1$ in the complex plane. The point λ then moves around the circle $\vec{\Lambda}$, with radius r and center at $1 - r$ (Fig. 26a, p. 269).

Let us now compute the eigenvalues of the operator $\vec{R}$, i.e. those λ for which the equation

$$\vec{R}u - \lambda u = 0$$

has the solution $u = (u_0, u_1, \ldots, u_m, \ldots)$ tending to 0 as $m \to +\infty$.

The equation $\vec{R}u - \lambda_u = 0$ may be written in expanded form as follows:

$$(1 - r - \lambda)u_m + ru_{m+1} = 0, \quad m = 0, 1, \ldots$$

Its solution $u_m = q^m$, $m = 0, 1, \ldots$, tends to 0 as $m \to +\infty$ if $|q| < 1$. The corresponding eigenvalues $\lambda = 1 - r + rq$, in this case, fill the interior of the circle $\vec{\Lambda}$, of radius r, centered at point $(1 - r)$ (Fig. 26,b).

The algorithm for computing the eigenvalues of the operator $\vec{R}$ is analogous to that for computing the eigenvalues of $\vec{R}$. The equation $\vec{R}$ is written expanded:

$$\begin{aligned} &(1 - r - \lambda)u_m + ru_{m+1} = 0, \quad m = \ldots, -1, 0, 1, \ldots, M-1, \\ &-\lambda u_M = 0. \end{aligned} \tag{6}$$

Each net function $u = \{u_m\}$, $m = M, M-1, \ldots$, satisfying the first of these relations, to within a constant factor has, as before, the form $u_m = q^m$, with λ and q still connected by the equation $\lambda = 1 - r + rq$. The solution $u_M = q^m$, $m = M, M-1, \ldots$, tends to zero as $m \to -\infty$ if $|q| > 1$. The second of Eqs. (6), i.e. the equation $-\lambda u_M = 0$, imposes on the solution $u_m = -q^m$ the auxiliary requirement $-\lambda u_M = -\lambda q^M = 0$, or $\lambda = 0$. If the point $\lambda = 0$ lies outside the circle of radius r and center at $1 - r$ (shown in

Fig. 26c) i.e. if $r < 1/2$, then to this λ there corresponds a q such that $|q| > 1$. The set, $\overleftarrow{\Lambda}$, of those λ's for which the equation $\overleftarrow{R}u - \lambda u = 0$ has a solution tending to 0 as $m \to -\infty$, consists only of this point $\lambda = 0$. In the case $r \geq 1/2$, as follows from the preceding analysis, the equation $\overleftarrow{R}u - \lambda u = 0$ has no solution tending to zero as $m \to -\infty$ for any complex (or real) λ.

The union of eigenvalues of the operators $\ddot{R}$, $\vec{R}$, and $\overleftarrow{R}$ is shown for the case $r < 1/2$ in Fig. 27,a; and for the case $r > 1/2$ in Figs. 27b and 27c.

We now proceed to prove that the spectrum of the family of operators $\{R_h\}$ coincides with the union, Λ, of the sets $\ddot{\Lambda}$, $\vec{\Lambda}$, and $\overleftarrow{\Lambda}$, of the eigenvalues of the auxiliary operators $\ddot{R}$, $\vec{R}$, and $\overleftarrow{R}$. We need to show that each point in Λ belongs to the spectrum of the family of difference operators $\{R_h\}$, and that the spectrum contains no other points.

Let us show first that each point λ_0 in Λ belongs to the spectrum of the family of difference operators. For this purpose it is sufficient to establish that, for any $\varepsilon > 0$, the inequality

$$||R_h u - \lambda_0 u|| < \varepsilon ||u|| \tag{7}$$

has a solution, u, for all sufficiently small positive values of h. The solution $u = (u_0, u_1, \ldots, u_M)$ might be called a "near-eigenvector" of the operator R_h, insofar as the solution of the equation $R_h u - \lambda u = 0$ is, in linear algebra, commonly called an "eigenvector".

The construction used in the proof depends on the set, $\ddot{\Lambda}$, $\vec{\Lambda}$, or $\overleftarrow{\Lambda}$, to which the point λ_0 belongs. Let us begin with the case λ_0 in $\ddot{\Lambda}$. We will show that, for any $\varepsilon > 0$ and all sufficiently small h, inequality (7) has a solution u.

We turn, now, to the construction of this function $u = (u_0, u_1, \ldots, u_M)$. By definition of the set $\ddot{\Lambda}$ there exists a q_0, $|q_0| = 1$, such that $\lambda_0 = (1 - r) + rq_0$, and the equation $(1 - r - \lambda_0)v_m + rv_{m+1} = 0$, $m = 0, \pm 1, \ldots$, has the bounded solution $v_m = q_0^m$, $m = 0, \pm 1, \ldots$ We will consider this solution only for $m = 0, 1, \ldots, M$, retaining the designation v. The vector

$$v = (v_0, v_1, \ldots, v_M) = (1, q_0, \ldots, q_0^M),$$

clearly, would satisfy the equation $R_h v - \lambda_0 v = 0$, which in expanded form consists of the relations

$$(1 - r - \lambda_0)v_m + rv_{m+1} = 0, \quad m = 0, 1, \ldots, M-1,$$

$$-\lambda_0 v_M = 0,$$

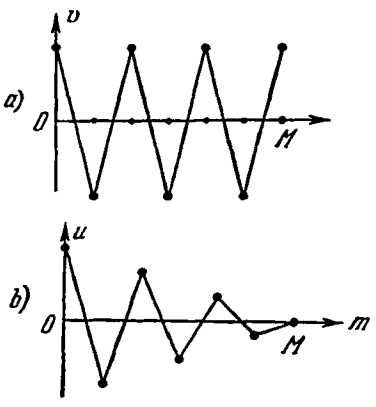


Fig. 55.

if not for the fact that the last of these relations is violated. The relation $-\lambda_0 v_M = 0$ may be considered as a boundary condition for the solution of the ordinary difference equation

$$(1 - r - \lambda_0)u_m + ru_{m+1} = 0,$$

$$m = 0, 1, \ldots, M-1.$$

To satisfy this boundary condition at $m = M$, i.e. at the right end of the interval $0 \le x \le 1$, we "touch-up" the vector $v = (1, q_0, \ldots, q_0^M)$, multiplying each of its components, v_m, by the factor $(M - m)h$. The vector thus obtained we call u, $u = (u_0, u_1, \ldots, u_M)$, $u_m = (M - m)hq_0^m$.

In Fig. 55 we have plotted the function $v = \{v_m\}$ and $u = \{u_m\}$ in the case $q_0 = -1$. The norm of the vector u is equal to one:

$$||u|| = \max_m |u_m| = \max_m |(M - m)hq_0^m| = Mh = 1.$$

Let us now evaluate the norm of the vector $w = (w_0, w_1, \ldots, w_M)$, defined by the equation $w \equiv R_h u - \lambda_0 u$. For the coordinates of the vector w we get the following expressions:

$$|w_m| = |(1 - r - \lambda_0)(M - m)hq_0^m + r(M - m - 1)hq_0^{m+1}| =$$

$$= |[(1 - r - \lambda_0) + rq_0](M - m)hq_0^m - rhq_0^{m+1}| =$$

$$= |0 \cdot (M - m)hq_0^m - rhq_0^{m+1}| = rh, \quad m = 0, 1, \ldots, M-1,$$

$$|w_M| = 0 - \lambda_0 \cdot 0 = 0.$$

Thus it is clear that $||w|| = rh$, and for $h < \varepsilon/r$ the inequality $||w|| = ||R_h u - \lambda_0 u|| < \varepsilon||u||$ is satisfied. This completes the proof that the point λ_0 in $\overleftrightarrow{\Lambda}$ belongs to the spectrum of the family of operators $\{R_h\}$.

Now we show that, if the point λ_0 belongs to one of the sets $\overleftarrow{\Lambda}$ or $\overrightarrow{\Lambda}$, then it is a point of the spectrum of the family of operators $\{R_h\}$. Suppose, for concreteness, that λ_0 is in $\overrightarrow{\Lambda}$. Then by definition of the set $\overrightarrow{\Lambda}$ the equation $\vec{R}v - \lambda_0 v = 0$, which in expanded form consists of the equations

$$(1 - r - \lambda_0)v_m + rv_{m+1} = 0, \quad m = 0, 1, 2, \ldots,$$

has the solution $v_m = q_0^m$, $|q_0| < 1$, $m = 0, 1, \ldots$

We will consider this solution only for $m = 0, 1, \ldots, M$, setting

$$u = (u_0, u_1, \ldots, u_M) = (1, q_0, \ldots, q_0^M),$$

and will calculate, for this net function u, whose graph in the case $q = 1/2$ is shown in Fig. 56, the norm of the vector $w \equiv R_h u - \lambda_0 u$. From the equation

$$|w_M| = |(1 - r - \lambda_0)q_0^m + rq_0^{m+1}| = 0, \quad m = 0, 1, \ldots, M-1,$$

$$|w_M| = |q_0|^M$$

it follows that $||w|| = |q_0|^M = |q_0|^{1/h}$. If h is so small that $q_0^{1/h} < \varepsilon$, then $||w|| = ||R_h u - \lambda_0 u|| < \varepsilon = \varepsilon||u||$, since $||u|| = 1$.

Thus it has been shown that, in our example, all the points of the sets $\overleftrightarrow{\Lambda}$, $\overrightarrow{\Lambda}$ and $\overleftarrow{\Lambda}$ belong to the spectrum of the family of difference operators.

Let us show now that any point λ_0, not belonging to the sets $\overleftrightarrow{\Lambda}$, $\overrightarrow{\Lambda}$ or $\overleftarrow{\Lambda}$ does not belong to the spectrum of the family $\{R_h\}$. Specifically, we will show that there exists a number $A > 0$, not depending on h, such that, for any function $u = (u_0, u_1, \ldots, u_M)$, we may write the inequality

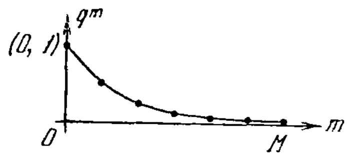


Fig. 56.

$$||R_h u - \lambda_0 u|| \geq A||u||. \tag{8}$$

Then for $\varepsilon < A$ the inequality $||R_h u - \lambda_0 u|| < \varepsilon||u||$ has no solution, and the point λ_0 does not lie in the spectrum. If we define $f \equiv R_h u - \lambda_0 u$, inequality (8) takes the form

$$||f|| \geq A||u||. \tag{9}$$

It is this inequality which we will derive. The equation $R_h u - \lambda_0 u = f$ will first be written in the expanded form

$$\left.\begin{aligned} (1 - r - \lambda_0)u_m + ru_{m+1} &= f_m, \quad m = 0, 1, \ldots, M-1, \\ -\lambda_0 u_M &= f_M. \end{aligned}\right\} \tag{10}$$

We will regard these relations as equations for u, and think of f as a given right-hand side. Next we write the solution $u = \{u_m\}$ in the form of a sum, setting

$$u_m = \alpha_m + \beta_m, \quad m = 0, 1, \ldots, M, \tag{11}$$

where the α_m are the components of the bounded solution $\alpha = \{\alpha_m\}$ of the following equation:

$$(1 - r - \lambda_0)\alpha_m + r\alpha_{m+1} = F_m =$$

$$= \begin{cases} 0, & \text{if } m < 0 \\ f_M, & \text{if } m = 0, 1, \ldots, M-1, \\ 0, & \text{if } m \geq M. \end{cases} \tag{12}$$

Then, by virtue of linearity, the vector $\beta = \{\beta_m\}$, whose components enter into Eq. (11), is the solution of the equation

$$\left.\begin{aligned} (1 - r - \lambda_0)\beta_m + r\beta_{m+1} = 0, \quad m = 0, 1, \ldots, M-1, \\ -\lambda_0\beta_M = f_M + \lambda_0\alpha_M. \end{aligned}\right\} \tag{13}$$

To prove bound (9), which for the given choice of norm can be written in the form $u_m < (\max_m |f_m|)/A$ it is sufficient (since $u_m = \alpha_m + \beta_m$) to establish bounds of the form

$$|\alpha_m| \leq A_1 \max |f_m|, \tag{14}$$

$$|\beta_m| \leq A_2 \max |f_m|, \tag{15}$$

where A_1 and A_2 are constants. Let us begin with bound (14). Note that Eq. (12) is an equation of first order of the form

$$a\alpha_m + b\alpha_{m+1} = F_m, \quad m = 0, \pm 1, \ldots,$$

where $a = 1 - r - \lambda_0$, $b = r$. An equation of this form was discussed in §2, where we arrived at the bound

$$|\alpha_m| \leq \frac{\max_m |F_m|}{|a| - |b|} = \frac{\max_m |f_m|}{|a| - |b|}. \tag{16}$$

In the example considered here $|a| - |b| > \delta_0/2$, where δ_0 is the distance from the point λ_0 to the set $\overset{\leftarrow}{\Lambda} + \overset{\rightarrow}{\Lambda} + \overset{\leftrightarrow}{\Lambda}$. Inequality (16), therefore, implies (14), the inequality we set out to prove. Bound (15) follows from (13), written in the following form

$$\beta_m = -\frac{f_M + \lambda_0\alpha_M}{\lambda_0} q_0^{m-M}, \tag{17}$$

where q_0 is defined by the relation $(1 - r - \lambda_0) + rq_0$. By assumption, the point λ_0 does not belong to the set Λ, and therefore lies outside the circle with radius r and center at $1 - r$. But in this case $|q_0| > 1$. Further $|\lambda_0| = \delta_1 > 0$, since if it were true that $\lambda_0 = 0$, then λ_0 would belong to the set $\ddot{\Lambda} + \dot{\Lambda} + \hat{\Lambda}$. Thus, using Eq. (17), and taking account of the already-demonstrated bound (14), we get inequality (15):

$$|\beta_m| = \left|\frac{f_M + \lambda_0\alpha_M}{\lambda_0}\right| \cdot \left|q_0^{m-M}\right| \leq \frac{|f_M|}{|\lambda_0|} + |\alpha_M| \leq$$

$$\leq \frac{\max|f_m|}{\delta_1} + A_1\max|f_m| = A_2\max|f_m|.$$

And thus it has been shown that the spectrum of a family of operators $\{R_h\}$, defined by Eq. (1), coincides with the union of the sets $\ddot{\Lambda} + \dot{\Lambda} + \hat{\Lambda}$ in the complex plane.

* * * * * *

2. Algorithm for computing the spectrum in the general case.

Theorem. *Let the operator* R_h, $b = R_h a$, a,b *in* U_h, *be given by the equation* $B_h b = A_h a$, *where* A_h *and* B_h *are linear operators, defined on a finite-dimensional linear normed space* U_h, *with values in some linear normed space* F_h. *Suppose, further, that the operators* A_h *and* B_h *are bounded uniformly in* h, *and that the operator* B_h *has a uniformly bounded inverse* B_h^{-1}: $||A_h||$, $||B_h||$, $||B_h^{-1}|| < C$.

In this case the spectrum of the family $\{R_h\}$ *excludes those and only those* λ, *in the complex plane, for which the operator* $A_h - \lambda B_h$ *has, for all sufficiently small h, an inverse operator uniformly bounded in h.*

The proof is obvious, and we will not present it here.

Suppose, now, the operator R_h, $v = R_h u$, is given by the difference relations

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} B_{km} v_{m+k} = \sum_{k=-k_0}^{k_0} A_{km} u_{k+m}, \quad k_0 \leq m \leq M - k_0, \\ &\sum_{i=0}^{2k_0} b_i v_i = \sum_{i=0}^{2k_0} a_i u_i, \quad \sum_{i=0}^{2k_0} \beta_i v_{N-i} = \sum_{i=0}^{2k_0} \alpha_i u_{N-i}, \end{aligned}\right\} \tag{18}$$

and, moreover, that the problem

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} B_{km} v_{m+k} = \phi_m, \quad k_0 \le m \le M - k_0, \\ &\sum_{i=0}^{2k_0} b_i v_i = \psi_1, \qquad \sum_{i=0}^{2k_0} \beta_i v_{M-i} = \psi_2 \end{aligned}\right\} \tag{19}$$

is well conditioned.

It will be assumed that

$$A_{km} = A_k\left(\frac{m}{M}\right), \qquad B_{km} = B_k\left(\frac{m}{M}\right),$$

where $A_k(x)$ and $B_k(x)$ are square matrices, defined on the interval $0 \le x \le 1$, and satisfying, on this interval, the smoothness conditions (14) §4: a_i, b_i, α_i and β_i are rectangular numerical matrices not depending on M. In this case the theorem is applicable, and the spectrum of the family of operators $\{R_h\}$ consists of all those λ's for which the difference boundary-value problem

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} (\lambda B_{km} - A_{km}) u_{m+k} = \phi_m, \quad k_0 \le m \le M - k_0, \\ &\sum_{i=0}^{2k_0} (\lambda b_i - a_i) u_i = \psi_1. \qquad \sum_{i=0}^{2k_o} (\lambda \beta_i - \alpha_i) u_{M-1} = \psi_2 \end{aligned}\right\} \tag{20}$$

is ill-conditioned. To determine whether problem (20) is well-conditioned, for each λ one can use the criterion of Sect. 7§4.

PROBLEMS

1. Prove that for the family of difference operators $\{R_h\}$, $v = R_h u$, given by the equations

$$\left.\begin{aligned} v_m &= (1 - r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ v_M &= 0, \qquad\qquad\qquad Mh = 1, \end{aligned}\right\}$$

and considered in this section, the spectrum does not change if the norm is defined, not by the equation $||u|| = \max_m |u_m|$, but by the equation

$$||u|| = \left(h \sum_m |u_m|^2\right)^{1/2}.$$

2. Prove that the spectrum of the family of difference operators $\{R_h\}$, $v = R_h u$, given by the equations

$$\left.\begin{aligned} v_m &= (1 - r + \gamma h)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ v_M &= 0, \qquad\qquad\qquad\qquad Mh = 1, \end{aligned}\right\}$$

does not depend on the value of the constant γ, and coincides with the spectrum constructed in this section for $\gamma = 0$.

3. Compute the spectrum of the family of operators $\{R_h\}$, $v = R_h u$, given by the equations

$$\left.\begin{aligned} v_m &= (1 - r)u_m + r(u_{m-1} + u_{m+1}), \quad m = 1, 2, \ldots, M-1, \\ au_0 &+ bu_1 = 0, \quad u_M = 0, \quad Mh = 1, m \quad r = \text{const}, \end{aligned}\right\}$$

where a and b are given numbers. Consider the case $|a| > |b|$ and $|a| < |b|$.

§46. The kernels of the spectra of families of operators

* * * * * *

Suppose that R_h reflects the linear normed space U_h', of dimension N, $N = N(h)$, into itself. We will write, in place of R_h and U_h', respectively, R_N and U_N, so that the notation will indicate the dimensionality explicitly. It is to be assumed that $N \to \infty$ as $h \to 0$.

Here we consider to what extent the spectrum of a family of operators $\{R_h\}$ depends on the choice of a sequence of norms $||\cdot||_N$ in the spaces U_N and, thus, to what extent the spectral criterion for the boundedness of the norms of powers of the operator R_N (Theorem 1 of §44) is invariant with respect to the choice of norm.

As regards the family of operators $\{R_N\}$ we will postulate that the eigenvalues of all the operators R_N are bounded in totality, i.e. lie in some circle

$$|\lambda| \leq c = \text{const}. \tag{1}$$

Clearly, for the validity of condition (1) it is sufficient that there exist at least one sequence of norms, $||\cdot||_N$, such that the inequality $||R_N|| < c' = \text{const}$ will be satisfied. Thus it is clear that bound (1) is natural: it is satisfied for the families of operators $\{R_N\}$, effecting transitions from level to level, and arising in the course of the consideration of evolutional difference boundary-value problems. Let us now go on to the definition of the concept of the kernel of a spectrum, which we will use to formulate the results of this section.

Suppose we are given: some sequence of norms $||\cdot||_N$, a number a in [0,1], and an integer $k \geq 0$. Denote by $\Lambda(a,k,N)$ the set of points, λ, for which the inequality $||R_N u - \lambda u|| < a^{N-k}||u||$ has a solution u in U_N. In the notation for the norm, the subscript N will be omitted.

Definition. The kernel, of index a, of a family of operators $\{R_N\}$ (with a in the interval [0,1]) we define to be the following set, $\Lambda(a)$, in the complex plane:

$$\Lambda(a) \equiv \bigcap_{k\geq 0} \bigcap_{s>0} \left(\overline{\bigcup_{N>s} \Lambda(a, k, N)} \right).$$

Here

$$\overline{\bigcup_{N>s} \Lambda(a, k, N)} \equiv \Lambda_s(a, k)$$

is the set-theoretical closure of the union of sets $\Lambda(a, k, N)$ for all $N > s$; further

$$\bigcap_{s>0} \Lambda_s(a, k) \equiv \Lambda(a, k)$$

is the intersection of all sets $\Lambda_s(a, k)$ while

$$\bigcap_{k\geq 0} \Lambda(a, k) = \Lambda(a)$$

is the intersection of all sets $\Lambda(a, k)$.

Theorem 1. *The kernel $\Lambda(a)$, a in [0,1], is completely contained in the spectrum of the family of operators $\{R_N\}$, and is closed.*

Proof. We will prove that, if λ_0 does not belong to the spectrum of the family of operators $\{R_N\}$, then neither does it belong to the kernel. In fact there is an $\varepsilon > 0$ and an N_0 such that, for all $N > N_0$ and any u in U_N, the inequality $||R_N u - \lambda_0 u|| \geq \varepsilon||u||$ is satisfied. But then for all λ in the circle $|\lambda - \lambda_0| < \frac{1}{2}\varepsilon$ the inequality $||R_N u - \lambda u|| \geq \frac{1}{2}\varepsilon||u||$ is also satisfied. Therefore for $N > N_0$ not one set $\Lambda(a, k, N)$ contains a point of the circle $|\lambda - \lambda_0| < \frac{1}{2}\varepsilon$. But this implies the validity of the first assertion of the theorem. To prove that the kernel $\Lambda(a)$ is closed we note that the $\Lambda_s(a,k)$ are closed by construction, and the set $\Lambda(a)$, as the intersection of closed sets, is also closed.

Example. Let us compute the kernel $\Lambda(a)$, a in [0,1], of the family of operators $\{R_N\}$, if the operator R_{N+1}, $v = R_{N+1}u$, is given by the equations

$$\left.\begin{aligned} v_n &= (1 - r)u_n + ru_{n+1}, \quad n = 0, 1, \ldots, N-1, \\ v_N &= 0, \end{aligned}\right\} \tag{2}$$

and the norm by the equation $||u|| = ||(u_0, u_1, \ldots, u_N)|| = \max|u_n|$. We will show that $\Lambda(a)$ consists of the point $\lambda = 0$, along with the closed circle of radius ar, with center at 1-r:

$$|\lambda - (1 - r)| \leq ar. \tag{3}$$

In fact $\lambda = 0$, as we saw in 1§42, is an eigenvalue for all operators R_N, so that it belongs to all the sets $\Lambda(a, k, N)$, and therefore to the kernel. Further, for any λ_0 lying strictly inside circle (3) there is a real $\alpha > 0$ and real $b > 1$ for which we may write

$$\lambda_0 = 1 - r + \frac{ar}{b} e^{i\alpha}.$$

The inequality $||R_N u - \lambda_0 u|| < a^N N^{-k} ||u||$, for any fixed k and all large enough N, has the solution

$$u_n = \begin{cases} (a/b)^n e^{i\alpha n}, & n = 0, 1, \ldots, N-1, \\ 0, & n = N. \end{cases}$$

It follows that, for all large enough N, the set $\Lambda(a, k, N)$ contains the point λ_0 and, therefore, this point is also contained in $\Lambda(a)$. Thus the interior points of circle (3) belong to the kernel $\Lambda(a)$, and in view of the fact that the kernel is closed it must also contain the boundary of circle (3).

If the point $\lambda_0 \neq 0$ does not belong to circle (3), i.e.

$$\lambda_0 = 1 - r + \frac{ar}{b} e^{i\alpha}, \quad a \geq 0, \quad b = 1 - 2\delta, \quad \delta > 0,$$

then, writing out the Green's function for the first-order difference equation (§2), it is possible to establish that, for any λ in the circle $|\lambda - \lambda_0| < \min[|\lambda_0|, ar/(1 - \delta)]$, for all large enough N and all u in U_N, we have the inequality $||R_N u - \lambda u|| > a^N ||u||$. It follows that the points of this circle do not belong to the sets $\Lambda(a, k, N)$ if N is large enough and, therefore, neither can they belong to the union of their closures $\Lambda_s(a, k)$, nor to the kernel $\Lambda(a)$.

Note that the kernel $\Lambda(0)$ of index $a = 0$ in the above example consists of the two points $\lambda = 0$ and $\lambda = 1 - r$, and the kernel $\Lambda(1)$ coincides with the whole spectrum of the family of operators $\{R_N\}$, computed in §45.

At this point we conclude our discussion of the example, and return to general considerations.

Definition. The kernel $\Lambda(0)$ will be called the "absolute kernel".

Theorem 2. *The absolute kernel of the family of operators* $\{R_N\}$ *does not depend on the choice of the sequence of norms* $||\cdot||_N$.

The proof follows from the fact that, for $a = 0$, the set $\Lambda(a, k, N)$ coincides for each N with the set of eigenvalues of the operator R_N, which does not depend on the norm in the space U_N.

Theorem 3. *Under condition (1) the sequence of norms,* $||\cdot||_N$, *can always be so chosen that the spectrum of the family of operators* $\{R_N\}$ *will coincide with its absolute kernel.*

Proof. We will demonstrate the construction of norms whose existence is asserted by the theorem. Choose a basis in the space U_N in such a way that the transformation matrix R_N, in this basis, will be in Jordan form, with the absolute values of all off-diagonal elements smaller than $1/N$. Introduce a scalar product, and the associated norm, stipulating that the basis is orthonormal. If λ_0 is an arbitrary point not belonging to $\Lambda(0)$, and $\varepsilon > 0$ is the distance from this point to the set $\Lambda(0)$ (closed by virtue of theorem 1) then one can verify that $||Ru - \lambda_0 u|| \geq \frac{\varepsilon}{4} ||u||$ for all $N > 8/\varepsilon$ and all u in U_N, so that λ_0 does not belong to the spectrum of the family of operators $\{R_N\}$.

Thus, if the spectrum of the family of operators $\{R_N\}$ does not coincide with its kernel $\Lambda(0)$ of index $a = 0$ for the given choice of norm, as in the above example (2) with norm $||u|| = \max|u_n|$, then by choosing another sequence of norms one can get as a spectrum the narrower set $\Lambda(0)$.

However, in the theory of difference schemes one uses norms which are not completely arbitrary.

We designate by $||\cdot||_{c_N}$ the norm equal to the maximum of the absolute values of all components which constitute a net function (or vector function) in U_N. Now we single out a class of sequences of norms $||\cdot||_N$ for which there exists a positive integer, s, depending on the sequence, and such that for all large enough N

$$\sup_{||u||_{c_N}=1} ||u||_N \leq N^s \inf_{||u||_{c_N}=1} ||u||_N. \tag{4}$$

Clearly the norm $||\cdot||_{c_N}$ itself, and all the norms we have encountered in dealing with difference equations will, as N increases, form sequences belonging to class (4).

Theorem 4. *The kernel* $\Lambda(a)$ *of index* a *in* $[0,1]$ *does not depend on the choice of norm sequences from among those satisfying requirement (4).*

The proof follows immediately from the definitions.

Let us now consider the family of operators, $\{R_h\}$, defined by Eqs. (18) and (19) of §45, making the supplementary assumption that the matrix coefficients A_k and B_k are constant: $A_k(x) \equiv A_k(0)$, $B_k(x) \equiv B_k(0)$. For this family of operators we may state the following important

Theorem 5. (A. V. Sokolov). *If in the spaces* $U_h' = U_N$, *in which the operators* $R_h = R_N$ *act, one introduces the norms* $||\cdot||_{c_N}$, *then the kernel* $\Lambda(1)$, *with index* $a = 1$, *of the spectrum of operators* $\{R_N\}$, *coincides with the whole spectrum of this family.*

From this theorem and theorem 4 it follows that, for any sequence of norms satisfying (4), the spectrum of the family of operators $\{R_N\}$ contains the spectrum of the family of operators $\{R_N\}$ obtained through use of the

norm $||\cdot||_{C_N}$; which spectrum in turn may be computed by the methods used in 2§45. Therefore, if the spectral condition for the boundedness of norms of powers of the operators R_N (i.e. theorem 1 of §44) is not satisfied for the norms $||\cdot||_{C_N}$, then it will not be satisfied for any other choice of a sequence of norms from among those subject to condition (4).

The proof of this theorem of A. V. Sokolov involves a very complicated line of reasoning, and we will not present it here.

§47. On the stability of iterative algorithms for the solution of nonselfadjoint difference equations

The solution of stationary problems via the time-evolution of a steady state may be regarded as a sort of iterative process, and the results obtained on successive time levels as successive approximations. In §35 we considered, as an example, the Dirichlet difference problem for the Poisson equation.

For a vanishing solution on the boundary this is a selfadjoint difference problem. Correspondingly, in the approach to steady state it was possible to expand the error in a complete orthogonal system of eigenfunctions. Via arguments based on the eigenvalues one could draw conclusions, simultaneously, about the rate of error reduction and, as well, about the influence of roundoff errors committed at intermediate time-levels.

It turns out that, in solving nonselfadjoint difference equations by the time-evolution method the situation, generally speaking, is different. An instability may develop, in this case, in spite of the convergence of the iterative process, as a result of a strong sensitivity to roundoff errors. Here this phenomenon will first be defined more precisely, and then discussed. In our discussion the concept of the spectrum, and the kernels of the spectrum, of a family of difference operators will turn out to be useful.

Let

$$u = R_N u + f_N \tag{1}$$

be a family of linear equations (a "difference equation") in some unknown element u of an N-dimensional linear normed space U_N, a family depending on the positive integer parameter N. We will consider the iterative process

$$u^{m+1} = R_N u^m + f_N, \quad m = 0, 1, \ldots, \tag{2}$$

for computing the solution u. It will be assumed that all the eigenvalues $\lambda_k = \lambda_k(N)$ of the operator R_N are smaller then 1 in absolute magnitude,

$$|\lambda_k| \leq \rho_N < 1, \tag{3}$$

i.e. that the well-known criterion for the convergence of process (2) is satisfied, with

$$||u - u^m|| = O(\rho_N^m). \tag{4}$$

Suppose, now, that computation (2) is carried out approximately, with some number, $p = q + \alpha$, of significant digits, i.e. via the equation

$$\tilde{u}^{m+1} = R_N\tilde{u}^m + f_N + 10^{-p}||\tilde{u}^m||\delta_m, \tag{5}$$

where δ_m is an arbitrary element of U_N with $||\delta_m|| \leq 1$.

We now choose a positive integer q and require that, for arbitrary δ_k, $||\delta_k|| \leq 1$, $k = 0, 1, \ldots$,

$$\overline{\lim_{m\to\infty}}\ ||u - \tilde{u}^m|| \leq 10^{-q}||u||. \tag{6}$$

Inequality (6) guarantees that one can calculate the solution u, using Eq. (5), with an error not exceeding one in the q'th decimal place (in the sense of the norm in U_N).

Lemma. *To satisfy condition (6) it is necessary that the number* α *of "extra decimal digits" in Eq. (5) satisfy the inequality*

$$(1 - 10^{-q})\phi \leq 10^{\alpha},$$

and sufficient that it satisfy the inequality

$$(1 + 10^{-q})\phi \leq 10^{\alpha},$$

where

$$\phi = \lim_{m\to\infty} \max_{||\delta_k||=1} \left|\left| \sum_{k=0}^{m} R_N^{m-k}\delta_k \right|\right| .$$

We leave the proof to the reader.

Note that the existence of $\phi = \phi(N)$ follows from condition (3). Below we will mean, by the symbol $\alpha = \alpha(N)$, the smallest integer which guarantees that requirement (6) is satisfied. From the lemma it is clear that such a number exists, is non-negative, and depends on q either weakly or not at all.

Definition. A convergent iteration algorithm (2) will be called "stable" if there exists a constant C, independent of N, for which

$$\alpha(N) < C; \tag{7}$$

a convergent iterative algorithm will be called "weakly stable" if there exists a constant C, independent of N, for which

$$\alpha(N) \leq C \ln N, \tag{8}$$

but the algorithm is not stable. Finally, a convergent iterative process will be called "unstable" if it is neither stable nor weakly stable.

Example. We write the equation

$$\left.\begin{aligned} &-2u_n + u_{n+1} - f_n = 0, \quad n = 0, 1, \ldots, N-1, \\ &u_N = 0 \end{aligned}\right\} \tag{9}$$

in the form

$$\left.\begin{aligned} &u_n = (1 - 2r)u_n + ru_{n+1} + rf_n, \quad n = 0, 1, \ldots, N-1, \\ &u_N = 0. \end{aligned}\right\} \tag{10}$$

treating r as a parameter. One iteration algorithm (2) for Eq. (10) proceeds as follows:

$$u_n^{m+1} = (1 - 2r)u_n^m + ru_{n+1}^m + rf_n, \quad n = 0, 1, \ldots, N-1, \tag{11}$$

so that the operator R_N, $v = R_N u$ is defined by the equations

$$\begin{aligned} &v_n = (1 - 2r)u_n + ru_{n+1}, \quad n = 0, 1, \ldots, N-1, \\ &v_N = 0. \end{aligned}$$

This operator has, as can easily be seen, only the two eigenvalues $\lambda_1(N) = 1 - 2r$ and $\lambda_2(N) = 0$.

Inequality (3) is satisfied and iterative process (11) converges for $r < 1$. We will take, as a norm, $||u|| = \max_n |u_n|$, and show that for $r < 2/3$ the algorithm is stable, while for $r > 2/3$ it is unstable. In fact if $r < 2/3$, then

$$\max_n |v_n| \leq \max(|1 - 3r|, |1 - r|) \max_n |u_n|,$$

so that $||R_N|| \leq \max(|1 - 3r|, |1 - r|) = \rho < 1$. Therefore $\phi(N) \leq 1/(1 - \rho)$, and by virtue of the lemma bound (7) holds for $C = -2\ln(1 - \rho)$.

Now suppose that $r > 2/3$. In (11) let $f_n = 0$, $u_n^0 = (-1)^n$. It is easy to see that in this case $u_n^m = (1 - 3r)^m(-1)^n$, $n = 0, 1, \ldots, N-m$. It follows, then, that $||R_N^m|| \geq \rho^m$, $m = 1, 2, \ldots, N-1$, where $\rho = |1 - 3r| > 1$. Therefore $\phi(N) > \rho^N$, and we find from the lemma that $\alpha = N \lg \rho$, which proves instability. One can show that for $r = 2/3$ the iteration algorithm (11) is weakly stable.

Thus the spectral convergence criterion (3) for the iteration algorithm does not determine its stability. The spectral criterion and stability conditions are properly formulated, not in terms of the disposition of the spectra of each of the operators R, but in terms of the location of the spectrum and kernels of the spectrum of the family of operators $\{R_N\}$. In fact under the assumption that the family of operators $\{R_N\}$ is uniformly bounded, $||R_N|| < C$, it is easy to verify the following assertion.

Lemma. *In order that, for all large enough N, the iteration algorithm (2) be convergent, it is sufficient that the radius, ρ, of any kernel of the spectrum of the family of operators $\{R_N\}$ be strictly less than unity.*

Stability criterion. *In order that the iteration algorithm (2) be stable it is necessary and sufficient that the spectrum of the family of operators $\{R_N\}$ lie strictly inside the unit circle.*

Theorem. *In order that the iteration algorithm (2) be convergent, and either stable or weakly stable, it is sufficient that the radius, ρ, of the kernel $\Lambda(1)$ of the spectrum of the family of operators $\{R_N\}$ be strictly less than unity; in order for a convergent iteration algorithm (2) to be unstable it is sufficient that the radius, ρ, of this kernel of the family of operators $\{R_N\}$ be strictly greater than unity.*

In §46 it has been shown that the kernel $\Lambda(1)$ of the spectrum of the family of operators $\{R_N\}$ does not depend on the choice of norms from among those of class (4)§46, a natural class of norms for difference equations. From this it follows, in particular, that if the operators R_N are uniformly-in-N contracting, $||R_N|| \leq \rho < 1$, so that the spectrum (and thus also the kernel $\Lambda(1)$ of the spectrum) of the family of operators $\{R_N\}$ lies in the circle $|\lambda| \leq \rho \leq 1$, then iteration algorithm (2) is stable and remains stable (strongly or weakly) in any other norm (4)§46, in which the operators R_N may no longer be contracting.

In the above example the spectrum of the family of operators consists of the circle $|\lambda - (1 - 2r)| \leq r$ and the point $\lambda = 0$, coinciding with its kernel $\Lambda(1)$. The assertion regarding the stability of algorithm (11) for $r < 2/3$, and its instability for $r > 2/3$ can, therefore be based on spectral criteria as well as on the theorem.

To compute the solution of a (nonselfadjoint) equation

$$A_N u + f_N = 0 \tag{12}$$

one may try to construct an iteration algorithm of the form

$$B_N u^{m+1} = B_N u^m + \left(A_N u^m + f_N\right). \tag{13}$$

Here the operator B_N must be so chosen as to be easy to invert numerically, and so that the spectrum of the family of operators $\{B_N^{-1}A_B\}$ will have a radius, ρ, smaller than one, and as small as possible. By virtue of the bound $||R_N^m|| \leq C(\varepsilon)\cdot(\rho + \varepsilon)^m$, (Eq. (13)§44) where $\varepsilon > 0$ is arbitrary and $C(\varepsilon)$ does not depend on N, this last condition guarantees rapid convergence; and by virtue of the stability criterion, formulated above it also guarantees the stability of iteration algorithm (13).

This Page Intentionally Left Blank

APPENDIX
METHOD OF INTERNAL BOUNDARY CONDITIONS

* * * * * *

In the theory of boundary-value problems for analytic functions (i.e. for the solution of systems of Cauchy-Riemann equations) and also for the solution of more general systems of partial differential equations, one sometimes applies the method of singular integral equations. This method consists in the reduction of the boundary-value problem to an integral equation with the integral taken over the boundary of the region under consideration. In addition to the given boundary condition, one also makes use of consequences of the system of differential equations itself, of the relations which must be satisfied by functions (and their normal derivatives) on the region boundary so that it will be possible to construct a solution of the equation by extending the domain of definition of the functions into the region's interior. In the case of analytic functions the necessary relation is the classical Sokhotski-Plemelj condition, which may be developed from the Cauchy integral formula

$$\phi(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} \, d\zeta$$

by going to the limit where z tends to the boundary γ. In the case of a differential equation of second order the corresponding condition falls out of Green's formula, expressing the solution at each point of the region in terms of values of this solution, and of its normal derivative, on the boundary. To obtain this condition one must also go to the limit where a point inside the region tends to its boundary, making use of the properties of potentials of simple and double layers.

The method of internal boundary conditions is, in concept, analogous to the above-described method, which reduces boundary-value problems for partial differential equations to integral equations at the boundary. The role of auxiliary boundary conditions, analogous to the Sokhotski-Plemelj condition, is taken over by internal boundary conditions evolving from the difference analogue of the integral formula of Cauchy (or the difference analogue of Green's formula).

1. Class of systems of difference equations. We will be concerned, here, with boundary-value problems for general systems of difference equations with constant coefficients which, in vector notation, take the form

$$Lu \equiv \sum_{k \text{ in } K} A_k u_{n+k} - f_n = 0, \tag{1}$$

where $n = (n_1, n_2, \ldots, n_s)$ and $k = (k_1, k_2, \ldots, k_s)$ are multi-indices, the A_k are quadratic matrices, f_n is given and u_n is the desired vector-function, while K is a finite set (a "stencil"). We will suppose that system (1) satisfies the following algebraic condition: the characteristic matrix

$$A(\xi) \equiv \sum_{k \text{ in } K} A_k \xi^k, \tag{2}$$

(where $\xi^k \equiv (\xi_1^{k_1}, \ldots, \xi_s^{k_s})$, and $\xi_1, \ldots, \xi_s$ are complex parameters), is not degenerate identically in ξ:

$$\det A(\xi) \not\equiv 0. \tag{3}$$

This restriction is a natural one: one can show that in the case $\det A(\xi) \equiv 0$ Eq. (1) has no solution for any finite (in n) right-hand side f_n.

2. Fundamental solution. The matrix function G_n will be called a "fundamental solution of system (1)" if it simultaneously satisfies the following two equations:

$$\sum_{k \text{ in } K} A_k G_{n-k} = \delta_n^0 E, \tag{4}$$

$$\sum_{k \text{ in } K} G_{n-k} A_k = \delta_n^0 E. \tag{4'}$$

Lemma. *Let* $Q(\xi_1, \ldots, \xi_t)$ *be an arbitrary polynomial in an arbitrary number t of complex arguments, a polynomial not identically equal to zero. Then it is possible to choose radii,* r_j, *of circles* $|\xi_j| = r_j$, *so that* $Q(\xi_1, \ldots, \xi_t) \neq 0$ *if* $|\xi_1| = r_1, \ldots, |\xi_t| = r_t$.

We carry out the proof by induction on the number of arguments, t. For $t = 1$ the number of roots of $Q(\xi_1) = 0$ is finite and the assertion is obviously true. Assuming that it has been proven for $t = p$ we now establish that the assertion of the lemma is also true for $t = p + 1$. Expand the polynomial $Q(\xi_1, \ldots, \xi_{p+1})$ in powers of ξ_{p+1}:

$$Q(\xi_1, \ldots, \xi_{p+1}) = Q_0(\xi_1, \ldots, \xi_p)\xi_{p+1}^M + \ldots + Q_M(\xi_1, \ldots, \xi_p),$$

where M is some positive integer, and $Q_0(\xi_1, \ldots, \xi_p)$ does not vanish identically. Choose $r_1, \ldots, r_p$ such, that $Q_0(\xi_1, \ldots, \xi_p) \neq 0$ for

$|\xi_1| = r_1, \ldots |\xi_p| = r_p$. This is possible by our induction assumption. Now by taking r_{p+1} great enough one can arrive at a situation such that for $|\xi_j| = r_j$, $j = 1, \ldots, p+1$ we have $Q(\xi_1, \ldots, \xi_{p+1}) \neq 0$.

Theorem 1. *The matrix* G_n, *defined by the equation*

$$G_n = \frac{1}{(2\pi i)^s} \oint \ldots \oint_{|\xi_j|=r_j} \ldots \oint \frac{A^{-1}(\xi)}{\xi_1^{n_1+1} \ldots \xi_s^{n_s+1}} d\xi_1 \ldots d\xi_s \qquad (5)$$

is a fundamental solution.

Here the r_j are chosen, in accordance with the lemma, so that $\det A(\xi) \neq 0$ if $|\xi_j| = r_j$.

This theorem may be proven by direct substitution. Taking account of the properties of residues we find that

$$\sum_{k \text{ in } K} A_k G_{n-k} = \frac{1}{(2\pi i)^s} \oint \ldots \oint \frac{A(\xi)A^{-1}(\xi)}{\xi_1^{n_1+1} \ldots \xi_s^{n_s+1}} d\xi_1 \ldots d\xi_s = \delta_n^0 E,$$

$$\sum_{k \text{ in } K} G_{n-k} A_k = \frac{1}{(2\pi i)^s} \oint \ldots \oint \frac{A^{-1}(\xi)A(\xi)}{\xi_1^{n_1+1} \ldots \xi_s^{n_s+1}} d\xi_1 \ldots d\xi_s = \delta_n^0 E.$$

3. Boundary of net-region. Consider Eq. (1) on some bounded set

$$Lu \equiv \sum_{k \text{ in } K} A_k u_{n+k} = f_n, \quad n \text{ in } D_0, \qquad (6)$$

where D_0 is an arbitrary net-domain of definition of the right-hand side f_n. Then the region of definition of the solution u_n is the set D, generated by the point n+k when n and k independently run through the points of D_0 and K respectively. With each r in D we associate a subset K_r of the set K, a subset consisting of all those k in K for which r-k is not in D_0. We designate as the "boundary" Γ the set of all those points r in D for which K_r is non-empty. For example for the simplest difference analog of the Poisson equation

$$Lu \equiv u_{n_1-1,n_2} + u_{n_1,n_2+1} + u_{n_1+1,n_2} + u_{n_1,n_2-1} - 4u_{n_1n_2} = h^2 F_{n_1n_2},$$

$$|n_1| < N, \quad |n_2| < N; \quad Nh = 1,$$

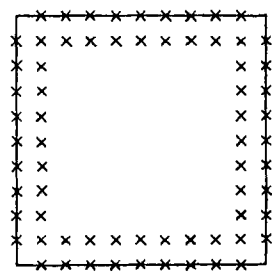


Fig. 57.

the set D_0 consists of those points (n,h, n_2h) which fall inside the square $|x_1| \leq 1$, $|x_2| \leq 1$; the set K of the five vectors (1,0), (0,1),

(-1,0), (0,-1), (0,0); the set D of the totality of all integer-points of the square $|n_1| < N$, $|n_2| \leq N$, except for the four corner-points $|n_1| = |n_2| = N$. The boundary Γ consists of two layers of points, marked by crosses in Fig. 57.

4. Difference analogs of Cauchy and Cauchy-type integral formulas.

Lemma. *Let B_n be an arbitrary matrix-function such that right-hand multiplication of this function is meaningful for any n'th order square matrix defined at all points of an integer-net. Then one may write the following identity:*

$$\sum_{n \text{ in } D_0} B_{-n} \sum_{k \text{ in } K} A_k u_{n+k} \equiv$$

$$\equiv \sum_{n \text{ in } D} \left(\sum_{k \text{ in } K} B_{-n+k} A_k \right) u_n - \sum_{r \text{ in } \Gamma} \left(\sum_{k \text{ in } K_r} B_{-r+k} A_k \right) u_r. \qquad (7)$$

Proof. The vector-function u_n, n in D, may be written in the form

$$u_n = \sum_{t \text{ in } D} \delta_n^t u_n.$$

The left- and right-hand sides of identity (7) depend linearly on u. Therefore to prove this identity it is sufficient to verify its validity for the vector-function

$$v_n \equiv v_n^t \equiv \delta_n^t u_n = \begin{cases} 0, & \text{if } n \neq t, \\ u_t, & \text{if } n = t, \end{cases}$$

for each fixed t in D:

$$\sum_{n \text{ in } D_0} B_{-n} \sum_{k \text{ in } K} A_k v_{n+k} \equiv \sum_{k \text{ in } K} \sum_{n \text{ in } D_0} B_{-n} A_k v_{n+k} \equiv$$

$$\equiv \sum_{k \text{ in } K} \delta_{D_0}^{t-k} B_{-t+k} A_k v_t \equiv \sum_{k \text{ in } K} B_{-t+k} A_k v_t - \sum_{k \text{ in } K_t} B_{-t+k} A_k v_t \equiv$$

$$\equiv \sum_{n \text{ in } D} \left(\sum_{k \text{ in } K} B_{-n+k} A_k \right) v_n - \sum_{r \text{ in } \Gamma} \left(\sum_{k \text{ in } K_r} B_{-r+K} A_k \right) v_r,$$

where

$$\delta_D^{t-k} = \begin{cases} 1 \text{ for } t-k \text{ in } D_0, \\ 0 \text{ for } t-k \text{ not in } D_0. \end{cases}$$

Theorem 2. *Let* $\{u_n\}$, *n in D, be an arbitrary solution of Eq. (6), while* G_n *is any arbitrary fundamental solution. Then*

$$\sum_{r \text{ in } \Gamma} \sum_{k \text{ in } K_r} \left(G_{n-r+k} A_k\right) u_r + \sum_{m \text{ in } D_0} G_{n-m} f_m = \begin{cases} u_n, & \text{for } n \text{ in } D, \\ 0, & \text{for } n \text{ not in } D. \end{cases} \tag{8}$$

Proof. We multiply both sides of Eq. (6) on the left by the matrix G_{t-n} and sum over all n *in* D_0. Using identity (7), and then Eq. (4′), we get Eq. (8).

Consequences. *Every solution* $\{u_n\}$ *of Eq. (6) is completely determined by its values on* Γ *and may be constructed from these values via Eq. 8.*

Theorem 3. *Let* $\{v_r\}$ *be an arbitrary vector-function, of dimension m, defined on* Γ, and let G_n be any fundamental solution. Then the equation

$$u_n = \sum_{r \text{ in } \Gamma} \left(\sum_{k \text{ in } K_r} G_{n-r+k} A_k \right) v_r + \sum_{m \text{ in } D_0} G_{n-m} f_m, \quad n \text{ in } D, \tag{9}$$

gives a solution of Eq. (6).

Proof. Applying the operator L to the vector-function $\{u_n\}$ defined by Eq. (9), we find that

$$Lu_n = \sum_{r \text{ in } \Gamma} \left[\sum_{k \text{ in } K_r} (LG_{n-r+k}) A_k \right] v_r + \sum_{m \text{ in } D_0} (LG_{n-m}) f_m, \quad n \text{ in } D_0. \tag{10}$$

Let us now calculate the right-hand side. By virtue of (4) we have

$$LG_{n-r+k} = \begin{cases} E, & \text{for } n = r - k, \\ 0, & \text{for } n \neq r - k. \end{cases}$$

But by the definition of the set K_r the point n = r-k does not belong to D_0, so that the first term on the right-hand side of Eq. (10) is the null-vector. The second term is, clearly, equal to f_n, so that $Lu_n = f_n$, and the theorem is proven.

Equation (8) is analogous to the Cauchy integral formula for analytic functions $\phi(z)$ in the bounded region d with boundary γ:

$$\frac{1}{2\pi i} \oint_{\zeta \text{ in } \gamma} \frac{\phi(\zeta)}{\zeta - z} d\zeta = \begin{cases} \phi(z), & \text{for } z \text{ in } d, \\ 0, & \text{for } z \text{ not in } d \cup \gamma. \end{cases} \tag{11}$$

Here the roles of analytic functions, of the boundary region and the Cauchy integrand $1/[2\pi i(\zeta - z)]$ are taken over, respectively, by the solutions of problem (6), the boundary Γ of the net-region D and the expression

$\left(\sum_{k \text{ in } K_r} G_{n-r+k}A_k\right)$: this expression takes into account, via set K_r within which the summation is carried out, the structure of the boundary near the point r in Γ.

It is natural, in this case, to compare Eq. (9) with integral formulas of the Cauchy type. Equation (8) is analogous to Green's formula for the Laplace equation.

We underscore, however, the following essential difference between Eqs. (11) and (8): the Cauchy integral formula is valid only strictly inside region d, while (8) is valid at all points of D, including its boundary points. There is an analogous difference between Eq. (9) and Green's formula.

5. Internal boundary conditions.

Theorem 4. *Suppose* G_n *is some fundamental solution of Eq. (1). If a given vector function* $\{u_r\}$, *given on* Γ, *(i.e with r in* Γ*) is to be extendable over the whole bounded net-region D so as to constitute a solution of Eq. (6), it is necessary and sufficient that, for all n in* Γ

$$\sum_{r \text{ in } \Gamma}\left(\sum_{k \text{ in } K_r} G_{n-r+k}\right)u_r + \sum_{m \text{ in } D_0} G_{n-m}f_m = u_n, \quad n \text{ in } \Gamma. \tag{12}$$

Proof. If $\{u_r\}$, with r in Γ, is to be extendable over D so as to form a solution $\{u_n\}$, n in D then, applying Eq. (8) to this solution, and then considering the resulting equation only for n in Γ, we verify that (12) is satisfied. Conversely, if $\{u_r\}$, for r in Γ, satisfies Eq. (12), then we take $v_r \equiv u_r$ and construct a solution $\{u_n\}$, n in D, via Eq. (9). By virtue of (12) the boundary values of this solution $\{u_r\}$, r in Γ, coincide with the given boundary values.

Theorem 4, just demonstrated above, gives us justification to call Eqs. (12) "internal boundary conditions": these conditions are not imposed externally, but are a consequence of the differential equation itself.

If Eqs. (8) and (9) are regarded as analogs of Cauchy and Cauchy-type integral formulas, then the internal boundary conditions can be thought of as analogous to the Sokhotski-Plemelj conditions, by which a function $\phi(z)$, given on the boundary, γ, of a region d in the complex plane, may be extended over the whole region d to form an analytic function.

Equation (8) may be considered as a difference-Green's-formula for system (6) which implicitly takes into account the "potential jump" on the boundary Γ, and tends to internal boundary conditions (12).

6. Boundary projection operator. It is possible to write the internal boundary conditions in a form different from (12). We will designate by U_Γ the linear space of all net vector-functions $u_r = \{u_r\}$, r in Γ, and by U_Γ° the subspace of those among them which may be extended over all of D to form a solution $\{u_n\}$, n in D, of the homogeneous equation corresponding to Eq. (6).

Define a linear mapping P, $u_\Gamma = Pv_\Gamma$, of space U_Γ into itself, by the following equation:

$$u_n = \sum_{r \text{ in } \Gamma} \left(\sum_{k \text{ in } K_r} G_{n-r+k} A_k \right) v_r, \quad n \text{ in } \Gamma. \tag{13}$$

Theorem 5. Operator P is a projection operator, projecting U_Γ onto U'_Γ.

Proof. In fact, for any v_Γ in U_Γ, by theorem 3 the element $u_\Gamma = Pv_\Gamma$ belongs to U'_Γ. If v_Γ is in U'_Γ then, by theorem 2, we get $Pu_\Gamma = u_\Gamma$. The theorem is proven.

The operator P, defined by Eq. (13), we will call the "boundary projection operator". With its help internal boundary conditions (12), in the case $f_n \equiv 0$, may be written in the form

$$u_\Gamma - Pu_\Gamma = 0. \tag{14}$$

It should be stressed that the boundary projection operator depends on the choice of a fundamental solution G_n.

7. General boundary-value problem. Given the stated consequence of theorem 2, we see that each solution of Eq. (6) may be reconstructed from its values on the boundary Γ. This fact gives us the justification to define the general linear boundary-value problem for Eq. (6) as a boundary-value problem of the form

$$\left.\begin{aligned} Lu \equiv \sum_{k \text{ in } K} A_k u_{n+k} &= f_n, \quad n \text{ in } D_0, \\ \ell u_\Gamma &= \phi, \quad \phi \text{ in } \Phi, \end{aligned}\right\} \tag{15}$$

where ℓ is some linear operator mapping the space U_Γ onto a linear space Φ.

Natural difference schemes approximating the first, second or third boundary-value problems for the Poisson equation, for example, may easily be written in form (15).

The name "general boundary-value problem" is somewhat arbitrary: one can find difference boundary-value problems which are not of form (15). This is true, for example, of natural difference schemes for differential boundary-value problems in which the order of the differential equation is lower than the order of the differential boundary conditions.

8. Basic idea of the method of internal boundary conditions. Suppose, for simplicity, that $f_n \equiv 0$. Between the difference boundary-value problem

$$Lu = 0, \quad \ell u_\Gamma = \phi \tag{16}$$

and the problem

$$u_\Gamma - Pu_\Gamma = 0, \quad \ell u_\Gamma = \phi \tag{17}$$

there is a very close connection. Specifically, by Theorem 2 the boundary values $u_\Gamma = \{u_r\}$ (with r in Γ), for each solution $\{u_n\}$ (n in D) of problem (16) must satisfy Eq. (17). Conversely, each solution $u_\Gamma = \{u_r\}$ (r in Γ) of problem (17) must, by theorem 4 and the stated consequence of theorem 2, be extendable uniquely, over all of D, to form a solution of problem (16). The basic idea of the method of internal boundary conditions consists in the passage, from the original boundary-value problem (16), to the system of equations (17) on the boundary Γ. Progress is achieved in this way for two reasons. First, because the number of unknowns in problem (17) is small compared with that in (16). Second, because of the special form of system (7), whose structure contains the boundary projection operator as a built-in integral component.

9. Stability of internal boundary conditions. One may, perhaps, fear that the internal boundary condition $u_\Gamma - Pu_\Gamma = 0$ is "almost degenerate", and that therefore problem (17) is ill-conditioned regardless of the form of operator ℓ, so that the passage from problem (16) to problem (17) is connected with the loss of computational stability.

Assuming that space Φ is contained in space U_Γ, we will introduce in space U_Γ (and therefore also in space Φ within U_Γ) the norm $||\cdot||$, then prove a theorem showing that, in the transition from problem (16) to problem (17), there is no loss of computational stability.

Theorem 6. Suppose that problem (17) has a solution u_Γ for an ϕ in Φ and, moreover, that

$$||u_\Gamma|| \leq c||\phi||, \tag{18}$$

where c does not depend on ϕ. Further let v_Γ be any arbitrary element of U_Γ. Introduce the notation

$$v_\Gamma - Pv_\Gamma = \tilde{\psi}, \quad \ell v_\Gamma = \tilde{\phi}. \tag{19}$$

Then v_Γ is subject to the bound

$$||v_\Gamma|| \leq c(||\tilde{\phi}|| + ||\ell||\,||\tilde{\psi}||) + ||\tilde{\psi}|| \tag{20}$$

If we regard (19) as an equation which determines v_Γ, then bound (20) signifies that the sensitivity of the solution of problem (19) to perturbations $\tilde{\psi}$ of the right-hand side of the internal boundary conditions is characterized by the constant c of bound (18), i.e. by the sensitivity of the solution to perturbations in the right-hand side of the given boundary condition $\ell u_\Gamma = \phi$.

Proof. Define

$$z_\Gamma \equiv v_\Gamma - \tilde{\psi} = Pv_\Gamma.$$

By theorem 5, $z_\Gamma = Pv_\Gamma$, with Pv_Γ in U_Γ'. Therefore

$$z_\Gamma - Pz_\Gamma = 0, \quad \ell z_\Gamma = \ell(u_\Gamma - \tilde{\psi}) = \tilde{\phi} - \ell\tilde{\psi},$$

i.e. z_Γ satisfies an equation-system of form (17) and, by virtue of (18), is subject to the bound

$$||z_\Gamma|| \leq c||\tilde{\phi} - \ell\tilde{\psi}|| \leq c(||\tilde{\phi}|| + ||\ell||\,||\tilde{\psi}||).$$

From this bound, taking account of the identity $z_\Gamma = v_\Gamma - \tilde{\psi}$, we get Eq. (20).

10. Supplementary idea. We now develop an idea which is useful for the computational solution of boundary-value problems for partial differential equations, an idea applicable to the following problem.

Suppose the function $u(x,y)$ is defined in some domain, d, with a sufficiently smooth boundary γ, as the solution of the Dirichlet problem

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (x,y) \text{ in } d, \\ u|_\gamma = a(s), \end{aligned}\right\}$$

and one is required to determine the derivative

$$\left.\frac{\partial u}{\partial n}\right|_\gamma = b(s)$$

directed towards the inward normal. Such a problem arises if, for a temperature $u|_\gamma = a(s)$ on the boundary, γ, one wants to find the steady-state heat-flux through the boundary.

Let s be the arc-length along the boundary γ and assume, for the sake of definiteness, that the whole length of the boundary γ is 2π. We set out to determine the function

$$\left.\frac{\partial u}{\partial n}\right|_\gamma = b(s)$$

approximately, in the form of a partial sum

$$b(s) = \sum_{j=0}^{k} (\alpha_j \cos js + \beta_j \sin js)$$

of its Fourier series. To determine the coefficients α_j and β_j we will use the method of internal boundary conditions.

Given $h > 0$, we construct the net

$$(x_{n_1},\ y_{n_2}) = (n_1 h,\ n_2 h)$$

and the difference equation

$$u_{n_1+1,n_2} + u_{n_1-1,n_2} + u_{n_1,n_2+1} + u_{n_1 n_2-1} - 4u_{n_1 n_2} = 0.$$

Assign to $D_0 = D_0^h$ all points of the net which, along with their four neighboring points, belong to $d \cup \gamma$. One can then define the net region $D = D^h$, its boundary $\Gamma = \Gamma^h$ and the internal boundary condition $u_\Gamma - Pu_\Gamma = 0$.

The idea proposed here is that the function $u|_\gamma = a(s)$ and the function

$$\left.\frac{\partial u}{\partial n}\right|_\gamma = b(s),$$

written in the form of a series with undetermined coefficients, be extended by Taylor's formula from the boundary, γ, into the adjacent band containing the boundary, Γ^h, of the net-region; the undetermined coefficients

$$\alpha_j = \alpha_j^h, \quad \beta_j = \beta_j^h$$

would then be chosen so as to minimize the residual which develops when the extension of the function, $u(x,y)$, from the boundary into the near-boundary region, is substituted into the boundary conditions.

11. Comparison of the method of internal boundary conditions with the method of singular integral equations. At the beginning of this Appendix we pointed out the analogy between the method of internal boundary conditions and the method of singular integral equations, an analogy which is not quite complete. Here we compare these methods, refining the analogy and bringing out explicitly the essential differences.

For purposes of comparison we first describe the idea of the method of singular integral equations for boundary-value problems, for example for the problem

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} - \mu u = 0, \quad x = (x_1,\ x_2) \text{ in } d, \tag{21}$$

$$a_0 u_0 + a_1 u_1 = \phi(x), \quad x = (x_1,\ x_2) \text{ in } \gamma, \tag{22}$$

where $\mu = \text{const} > 0$ and d is a bounded region with γ as its boundary. Boundary condition (22) connects the solution $u = u_0(x)$ on the region-boundary with its derivative along the inward-pointing normal $\partial u/\partial \nu = u_1(x)$. The coefficients a_0 and a_1 are given operators.

Let us first write out the classical Green's formula for Eq. (21):

$$u(x) = \int_{y \text{ in } \gamma} \left[g(x-y)\,\frac{\partial u}{\partial \nu} - u\,\frac{\partial g}{\partial \nu}\right] d\sigma_y, \tag{23}$$

where g(x) is the fundamental solution of Eq. (21) tending to zero at infinity. Now let x tend towards the boundary γ. Using the properties of potentials of single and double layers we get, on the boundary γ, a relation of the form

$$u_0 = b_0 u_0 + b_1 u_1, \tag{24}$$

connecting the solution u(x) with its normal derivative $\partial u/\partial \nu = u_1(x)$ on the region-boundary; here b_0 and b_1 are known integral operators. The transition from problem (21), (22), to the equivalent system of equations (23), (24), for the functions $u_0(x)$ and $u_1(x)$ defined on the boundary γ, is precisely the essential feature of the method of singular integral equations.

For comparison let us now consider the method of internal boundary conditions as applied to the following general boundary value problem for the difference analog of Eq. (21) in a square net-region

$$u_{n_1-1,n_2} + u_{n_1,n_2+1} + u_{n_1+1,n_2} + u_{n_1,n_2-1} - (4+\mu)u_{n_1,n_2} = 0, \tag{25}$$

$$-N < n_1, n_2 < N,$$

$$\ell u_\Gamma = \phi. \tag{26}$$

We will write the internal boundary condition $u_\Gamma - Pu_\Gamma = 0$ in a form which will be useful below. It is easy to verify that Eq. (9) in this case may be rewritten in the form

$$u_n \equiv \sum_{r \text{ in } Q_0} \left[G_{n-r}(\Delta_\nu u_r) - u_r(\Delta_\nu G_{n-r})\right] + \sum_{r \text{ in } Q_0} u_r \delta^r_n, \quad n \text{ in } D, \tag{27}$$

where Q_0 is the set of all points, Γ, lying on the sides of the square $|n_1| = N$, $|n_2| = N$, i.e on the outer layer of the two-layer boundary Γ of the net-square region (Fig. 57), and Δ_ν is the difference analog of the derivative along the inward-directed normal.

We note that Eq. (27) would be completed analogous to the classical Green's formula (23) in the absence, on the right-hand side, of the "singular term" $\sum \delta^r_n u_r$. However in this case Eq. (27) would be valid, not for all n in D, but only for n in D_0. It would then be impossible to arrive at the internal boundary conditions $u_\Gamma - Pu_\Gamma = 0$. These conditions are obtained from (27) if n runs over, not the whole region D, but only the points on the boundary Γ; they may be written as two systems of equations

$$u_n = \sum_{r \text{ in } Q_0} \left[G_{n-r}(\Delta_\nu u_r) - u_r(\Delta_\nu G_{n-r})\right] + u_n, \quad n \text{ in } Q_0, \tag{28}$$

$$u_n = \sum_{r \text{ in } Q_0} [G_{n-r}(\Delta_\nu u_r) - u_r(\Delta_\nu G_{n-r})], \quad n \text{ in } \Gamma \setminus Q_0, \tag{29}$$

corresponding, respectively, to the points n, n in Q_0, of the outer layer, and the points n in $\Gamma \setminus Q_0$ of the inner layer of the double boundary Γ. As the function G_n, in Eqs. (28) and (29), we take a bounded fundamental solution. One can show that the internal boundary condition $u_\Gamma - Pu_\Gamma = 0$, i.e. the system of equations (28), (29), is algebraically equivalent to each of the subsystems, (28) or (29), taken separately.

Subsystem (28) is analogous to integral relation (24), so that the difference analog of problem (22), (24) is problem (27), (28), but not the problem

$$\ell u_\Gamma = \phi, \quad u_\Gamma - Pu_\Gamma = 0,$$

specified by Eqs. (27)-(29), which is considered in the method of internal boundary conditions.

There is an obvious difference between the internal boundary conditions $u_\Gamma - Pu_\Gamma = 0$, i.e system (28), (29), and subsystem (28) alone. The internal boundary conditions

$$u_\Gamma - Pu_\Gamma = 0$$

contain the extra equations (29). In this sense the difference-internal-boundary-conditions

$$u_\Gamma - Pu_\Gamma = 0$$

are similar, not to integral relation (24), but to the Sokhotski-Plemelj conditions for analytic functions. These latter take the form of two real relations connecting two real functions, but they are not independent, and the manifold of pairs of functions satisfying the Sokhotski-Plemelj conditions depends on one real arbitrary function.

We note that the internal boundary conditions

$$u_\Gamma - Pu_\Gamma = 0$$

have an advantage over the equivalent subsystem (28) in that within their structure they contain the boundary projection operator. Thanks to this circumstance the problem

$$\ell u_\Gamma = \phi, \quad u_\Gamma - Pu_\Gamma = \psi$$

is stable, in the sense of theorem 6, with respect to perturbations of the right-hand side ψ. In the general case one can, by striking some of the equations from among those constituting the system $u_\Gamma - Pu_\Gamma = 0$, produce a

subsystem algebraically equivalent to the original equation-set, but no longer having this stability property.

One can show that, in our example (25), (26), instead of

$$u_\Gamma - Pu_\Gamma = 0$$

it is more convenient to use, not subsystem (28), but subsystem (29) which is stable and, in contrast to subsystem (28), consists of independent equations; its rank is equal to the number of equations it contains.

Thus in the above example the analogy between the method of internal boundary conditions, and the method of singular integral equations like the Sokhotski-Plemelj conditions, is not complete.

Moreover there is not a complete analogy with the classical method of integral equations in which the function sought is not itself the solution of the original problem (21), (22) on the boundary, but is some auxiliary density of the potential of a single or double layer.

In conclusion we note that we use the expression "method of singular integral equations" because the Sokhotski-Plemelj condition contains a singular integral. In the example of this section condition (26) contains a convergent improper integral.

* * *

This Page Intentionally Left Blank

BIBLIOGRAPHICAL COMMENTARIES

Ch. 1, §§1,2. One can become acquainted with the general theory of linear difference equations, for example, by reading Ch. 5 of Ref. 7.

Ch. 2, §5. The author first was introduced to the FEBS method and its underlying theory, as applied to several classes of difference boundary-value problems, in 1953 via the manuscript of an article by I. M. Gelfand and O. V. Lokytsievskii, entitled "The forward-elimination, back substitution method for the solution of difference equations". (See, for example, Ref. 10.) There exist variants of the FEBS method, designed for the computational solution of difference boundary-value problems not considered in this book. Various results along with a bibliography, will be found in Refs. 4, 15, 23, and others.

Ch. 3. The idea of using directly, as a basis for forward-elimination, back-substitution, the good-conditioning property of difference boundary-value problems was proposed by N. S. Bakhvalov. Some steps toward realization of this idea were taken in the presentation of FEBS in Ref. 10, and then by V. V. Ogneva (U.S.S.R. Comp. Math. and Math. Phys. 7, #4 (1967)) who is responsible for the idea of considering the truncated systems. A modified presentation of this work will be found in Ref. 8.

The mathematical theory of good-conditioning of difference boundary-value problems, as presented in §6, uses the thesis work of Bagisbaev, a student at Novosibirsk University who is responsible in particular for the example showing that the coefficient-smoothness conditions cannot be ignored.

Ch. 6, §§19,20. One can acquire a more detailed familiarity with methods for the numerical solution of ordinary differential equations through Ref. 4, and through the literature cited there.

Difference schemes for some important classes of differential equations with discontinuous coefficients are constructed by A. N. Tikhonov and A. A. Samarskii in their theory of homogeneous difference schemes, and are presented in one of the chapters of Ref. 23.

Ch. 7 §21. The concept of stability of difference schemes with respect to rounding errors, for given initial conditions, was first described

by J. Von Neumann and R. D. Richtmeyer in 1950* in work devoted to the computation of gas-dynamics discontinuities. The first overall system for determining stability and approximation, in which convergence is a consequence of approximation and stability, was proposed by V. S. Ryabenkii, Soviet Math., Doklady, **86**, #6 (1952), in the case of the difference analog of Cauchy's problem for partial differential equations.

The system of basic definitions adopted in this book, and the theorem stating that approximation and stability imply convergence, are close to those proposed by A. F. Filippov, Soviet Math., Doklady, **100**, #6 (1955). See also Ref. 22 or Ref. 10. The main difference consists in that we use a more universal definition of approximation.

There exist other natural systems of definition of the basic concepts, for which approximation and stability guarantee convergence. Among these the best known is the system of definitions of P. D. Lax, proposed in 1956 (see, for example, Ref. 20). In Lax's theory he considers difference schemes for nonstationary problems but postulates that the difference schemes act, not in the space of net functions, but in the same function space as the differential equations. With this (supplementary) assumption it is demonstrated that, for an approximating difference scheme, stability and convergence take place simultaneously. This equivalence theorem of Lax is one of the concrete forms of the more general construct of L. V. Kantorovich, Russian Math. Surveys **3**, issue 6 (1948).

In the last few years A. A. Samarskii, jointly with A. V. Gulin, has proposed and developed a stability theory applicable to a whole wide class of difference schemes (see Refs. 23 and 24, and §43 of this book).

New results, along with a bibliography and surveys of work on the stability of difference schemes, may be found in Refs. 10, 15 and 20-28.

It should be said that in the 1928 work of R. Courant, K. Friedrichs and G. Levy (see Russian Math. Surveys **8** (1940)) and in much other work where the method of finite differences is used to prove the existence of solutions of differential equations, the authors proved inequalities which, in modern terminology, could be interpreted as stability in one norm or another. However the concept of stability developed in connection with the use of difference schemes for the computation of approximate solutions assuming that these solutions exist. Therefore stability is usually studied in weaker norms than those used for proof of existence. Note that the method of finite differences was first used to prove the existence of solutions of partial differential equations in 1924 by L. A. Lyusternik (see Russian Math. Surveys **8** (1940)), in work which dealt with Laplace's equation.

---

*J. von Neumann and R. D. Richtmeyer, "A Method for the Numerical Calculation of Hydrodynamical Shocks, J. Appl. Phys. **21** #3 (1950).

Ch. 7, 3§22. The method presented here for the construction of difference schemes, was proposed in the works of: P. L. I. Brian, A.I.Ch.E. J. 7 (1961); J. Douglas, Num. Math. 4 (1962); J. Douglas, Trans. Amer. Math. Soc. 89 (1958); and S. K. Godunov, Difference Methods for the Solution of the Equations of Gas Dynamics, Novosibirsk (1962) (in Russian). The two-dimensional variant of the predictor-corrector Lax-Wendroff scheme (Ref. 20), considered in this section, was proposed for gas dynamics problems by L. A. Chudov. (See the review article by G. S. Roslyakov and G. F. Telenin in the collection "Computational Methods of Gas Dynamics" Moscow, Moscow University Press, issue 2 (1963) (in Russian)). The idea of the Runga-Kutta method was used by V. V. Rusanov (preprint, In-t prikl. matematika AN SSSR(1967, in Russian)) for the construction of difference schemes of third-order accuracy for gas-dynamics calculations.

L. A. Chudov (article in the collection "Some applications of net methods in Gas Dynamics", Vol. 1 "Flow in the boundary layer", Moscow State University Press, (1971) (in Russian)), has, for equations of parabolic type, constructed a difference scheme of Runge-Kutta type with second-order accuracy and good smoothing properties. Predictor-corrector schemes are used in many gas-dynamics calculations. See, for example, Ref. 1. There are, in addition, other methods for the construction of difference schemes (see Refs. 4, 13, 19-28).

Ch. 8, 5§25. So fas as is known by the authors, the possibility of using differential approximations for the study of difference equations was first noted in the 1950's by A. I. Zhukov (communications of a seminar of the Institute for Applied Math), who proposed the example used here. The theory of differential approximations, in which one studies the asymptotic and group properties of interesting classes of difference equations, was constructed by N. N. Yanenko and Yu. I. Shokin, Sib. Matem. Zh. **10**, #5 (1969); Chislennie Metodi Mekh. Sploshnoy Sredi 2, #2 (1971) (in Russian). The same class of problems was addressed by N. N. Kuznetsov, Soviet Math., Doklady **200**, #5 (1971); Soviet Math., Doklady **204**, #2 (1972); U.S.S.R. Comp. Math. and Math. Physics 12, #12 (1972).

Ch. 8, 1§26. The idea of freezing coefficients at interior points was proposed in the above-cited article of Von Neumann and Richtmeyer (see comments on §21).

Ch. 8, 2§26. The criterion of K. I. Babenko and I. M. Gelfand was reported in their paper, authored jointly with O. V. Lokutszevskii, and presented at a conference on functional analysis in 1956 in Moscow. See also Ref. 2 and the comments, below, on Ch. 14.

Ch. 8, §27. There exists an algorithm for the calculation of coefficients in finite Fourier series, wich is very economical in the number of

arithmetic operations, and is commonly called the "Fast Fourier Transform." See, for example, Refs. 4 or 5.*

Finite Fourier series were, apparently, first used for the analysis of nonstationary difference equations by O. A. Ladizhenska. With the aid of this apparatus she found a convergent implicit difference scheme for equation-systems hyperbolic in the sense of Petrovski. Apparently this was the first example of a convergent implicit difference scheme (O. A. Ladizhenska, author's summary of dissertation, Leningrad State University, March 1949, in Russian. See also Ref. 13.).

Ch. 9. See Refs. 1-3, 9, 13, 14, 21, 26 and their bibliographies. In journals and collections of papers one constantly finds new work on computational methods applied to the mechanics of continuous media.

Ch. 10. The alternating direction scheme (12) §32 was constructed by D. Peaceman and G. Rachford in 1956 (see, for example, Refs. 5 or 28). Splitting scheme (7) §31 was proposed by N. N. Yanenko, Soviet Mathematics, Doklady **125**, #6 (1959). At this point splitting schemes have been constructed for many of the basic problems of mathematical physics. See, for example, Refs. 5, 15, 23, 27 and 28; also the monograph by E. G. Dyakonov entitled "Difference Methods for the Solution of Boundary-Value Problems, Part 1 (1971) and Part 2 (1972)", (Moscow State University Press, in Russian) and its bibliography.

A variant of the alternating direction method, obtained via combination of this method with the Ritz variational method, has been proposed and used for the computation of eigenvalues of strongly-elliptic operators, and for the solution of Laplace difference equation in: G. P. Prokopov, U.S.S.R. Comp. Math. and Math. Phys. **8**, #1 (1968); S. K. Godunov and G. P. Prokopov, U.S.S.R. Comp. Math. and Math. Phys. **9**, #2 (1969); S. K. Godunov, V. V. Ogneva and G. P. Prokopov, in "Partial Differential Equations", a collection of papers, proceedings of a symposium dedicated to the 60th birthday of academician C. L. Sobolev, 1970 (in Russian). The original locally one-dimensional scheme was proposed by I. V. Fryazinov, U.S.S.R. Comp. Math. and Math. Phys. **13**, #3 (1973).

Ch. 10, §33. Relating to the method of super-particles of O. M. Belotserkovski and Yu. M. Davidov, and to its applications, aside from the work cited in §33 see Ref. 3; also the text of the review paper by O. M. Belotserkovski and V. E. Yanitsko given at the Fourth U.S.S.R. Conference on the Dynamics of Rarified Gases in 1975 at Zvenigorod (in Russian); and the text of the lecture given by O. M. Belotserkovski at the Von Karman lectures in Brussels, 1976.

Ch. 11, §34. For the Poisson difference equation in a rectangle the most economical computational solution method is the fast Fourier transform

---

*Also, for example, "The Fast Fourier Transform," E. O. Brigham, Prentice Hall (1974). (Translator's note.)

(see the comments on §27). Many authors, starting with L. A. Lyusternik in 1924, have worked on difference schemes for the Laplace and Poisson equations in regions with curvilinear boundaries. See, for example, Refs. 4, 16, 23 and their bibliographies.

Error estimates, expressed directly in terms of the initial conditions, have been obtained for a series of schemes approximating the Dirichlet and von Neumann problems, and the mixed boundary-value problem for the Laplace and Poisson equations in a rectangle, a rectangular parallelopiped and certain triangles. See E. A. Volkov, Tr. Matem. in-ta im. V. A. Steklova, 74 (1966) 105 (1969) (in Russian), and I. A. Sultanova, U.S.S.R. Comp. Math. and Math. Phys. 11, #5 (1971), with bibliography. E. A. Volkov also established (Tr. Matem. in-ta im. V. A. Steklova, 117 (1972) (in Russian) that, if the difference operator at the boundary net-points satisfies a certain adequacy condition with respect to the standard five-point Laplace difference-operator, then the solution of the Poisson difference equation extended from the net onto a closed region with curvilinear boundaries will, for smooth enough initial data, approximate to second order in the net step-width the solution itself, and all its derivatives up to and including the n'th, for arbitrary n.

We also mention, in particular, an error bound for a difference solution of the Poisson equation obtained by E. A. Volkov (Tr. Matem. in-ta im. V. A. Steklova, 117 (1972) (in Russian) in a situation where the Laplace operator is not approximated to second order in the number of net-levels, a number which grows without bound as the net is refined. This bound is, at the same time, stronger than a uniform second-order bound since it implies an additional falling-off of the error near the boundary.

Ch. 11, §35. The idea of considering solutions of stationary problems as limits of solutions of nonstationary problems as $t \to \infty$ was first used, in the 1930's, by A. N. Tikhonov.

One of the approach-to-steady-state difference schemes for the treatment of supersonic gas flow around immersed bodies was proposed by S. K. Godunov, A. V. Zabrodin and G. P. Prokopov, U.S.S.R. Comp. Math. and Math. Phys. 1, #6 (1961), (see Ref. 9). It is interesting to note that the arguments relating to the stability of this scheme, described in the work of K. A. Bagrinovski and S. K. Godunov (Soviet Mathematics, Doklady, 115 #3 (1957)) make use of the splitting of the difference operator. There now exists a whole series of works, by many authors, directed towards the computational treatment of stationary problems via the establishment of a steady state.

One of the first effective methods for accelerating the solution of the Poisson difference equation was indicated by Lyusternik (Tr. Matem. in-ta im. V. A. Steklova 20 (1947) (in Russian)).

Ch. 11, §36. Chebyshev polynomials have been used for optimizing sets of iteration parameters in various problems, starting with the works of A. A. Abramov, M. K. Gaburin and Flanders and Shortley, all appearing in about 1950.

New results, bibliographies, and review papers written from various points of view, relating to iterative methods for solving elliptic difference-boundary value problems, can be found in Refs. 5, 16, 23 and 28. Also in the monographs "Iterative Methods for the Solution of Difference Analogs of Boundary-Value Problems of Elliptic Type", E. G. Dyakonov, Kiev (1970): "Iterative Methods and Quadratic Functionals," G. I. Marchuk and Yu. A. Kuznetsov, Novosibirsk, Nauka, Siberian Section, 1972, both in Russian; and in the review paper by R. P. Fedorenko, Russian Mathematical Surveys **28**, #2 (1973), as well as other works.

Ch. 12. The basic idea underlying the construction of variational-difference schemes is contained in the work of R. Courant (Courant R., Bull. Amer. Math. Soc. **49**, #1 (1943)). Independently, in engineering calculations of structural strength, various realizations of variational-difference schemes were often used without theoretical justification, under the name "finite-element methods".

The monograph by L. A. Oganesyan, V. Ya. Rivkind and L. A. Rykhobetz, entitled "Variational-Difference Methods for the Solution of Elliptic Equations" (in Parts 1 and 2 of Tr. seminara po differents. uravneniam, In-t fiziki i matemateki AN Litovskoy SSR, issue 5, Vilna, 1973 and issue 8, Vilna 1974, (in Russian)) is devoted to a systematic presentation of the foundations of the theory of variational-difference schemes, and of some of its applications. This monograph was used in the preparation of Ch. 12. See also, for example, Refs. 12, 18 and 25.

At the present time variational-difference schemes have been implemented in the form of well-developed programs on fast computers, for a whole series of problems in the theory of elasticity. See, for example, Ref. 12. There are also numerical implementations of the projection-difference method for some other (not only elliptic) problems. A series of recent works has been collected in "Variational-Difference Methods in Mathematical Physics, Novosibirsk, 1974 and Novosibirsk, 1976 (in Russian).

Ch. 13, §42. Stationary solutions are often used to elucidate the character of convergence close to boundaries. See, for example, S. K. Godunov, Matem. Sb. **47** (89), 3 (1957, in Russian).

Ch. 13, 4§43. Here we have used Sect. 4§6, of Ref. 22, written by A. F. Fillipov.

Ch. 13, 5§43. The choice of scalar product $(u,v)_{B_h}$ via Eq. (21), apparently, was first proposed by N. Min'o in 1953, for the special case of the difference analog of the heat-equation with variable coefficients, and then presented in more general form in §15 of Ref. 22, which also contains a modified presentation of the above-cited work of N. Min'o.

Ch. 13, 6§43. The first of the Samarski stability criteria introduced in this section is obtained from theorem 5, section 6§1, Chapter VI of Ref. 23, if, instead of Hilbert space, one considers Euclidean space, and sets $\rho = 1$. See also Sec. 7§1, Chapter VI of Ref. 23.

Ch. 14, §44. The concept of the spectrum of a family of operators was introduced in Ref. 10 where in particular the authors, with the help of this concept, derived the criterion of K. I. Babenko and I. M. Gelfand for stability of nonstationary problems on line-segements. There it was also shown that disposition of the spectrum of a family of operators in the unit circle is necessary for stability.

Theorem 2 was derived by V. C. Ryaben'kii, Soviet Math. Doklady, **185**, #2 (1969).

Ch. 14, §46. The concept of the kernel of the spectrum of a family of operators was introduced by Ryaben'kii, Soviet Math. Doklady, **185**, #2 (1969). There, also, the author formulated theorems 1-4.

The theorem of A. V. Sokolov for the case of scalar coefficients A_k and B_k was published in Soviet Math., Doklady, **208**, #2 (1973). A proof in the general cases of matrix coefficients is contained in his article, Tr. Mosk. matem. obsch. **35**, Moscow State University Press, 1976 (in Russian).

Ch. 14, §47. Here we present a paper of V. S. Ryaben'kii, Soviet Math., Doklady **193**, #3 (1970).

Appendix. The method of internal boundary conditions (MIBC) was proposed by V. S. Ryaben'kii, Doctoral Dissertation, In-t prikl. matematiki AN SSSR (1969) (in Russian). In Sects. 1-9 and 11 we present part of a paper by V. S. Ryaben'kii, Math. Surveys **26**, #3 (1971). This paper also describes some applications of MIBC to the study and computational solution of boundary-value problems in simple and compound regions.

The content of Sect 10 was published in a report presented by V. S. Ryaben'kii at a conference honoring the 70'th birthday of academician I. G. Petrovski, held at Moscow State University (January 1976).

Appendix, Sect. 2. A. Ya. Belyankov, Matem. Zametki **18**, #5 (1975, in Russian), proved the existence of a fundamental solution which grows, for $||n||^2 = n_1^2 + \ldots + n_s^2 \to \infty$, no faster than some power of $||n||$.

He also constructed the so-called cyclic fundamental solution, which allows one to construct internal boundary conditions and makes possible an effective construction technique based on the fast Fourier transform. See his article in the collection "Problems of Mechanics and Mathematical Physics" (in Russian) dedicated to the memory of academician I. G. Petrovski, Nauka, 1976.

A. V. Zabrodin and V. V. Ogneva (preprint, In-t prikl. matematiki AN SSSR, 1973, in Russian) used their own variant of MIBC for the computational treatment of a nonlinear heat conduction problem.

This Page Intentionally Left Blank

BIBLIOGRAPHY

1. Alalikin, G. B., Godunov, S. K., Kireeva, I. L., Pliner, L. A., "Reshenie odnomernikh zadach gazovoi dinamiki v podvizhnikh setkakh," M., "Nauka" (1970).

2. Babenko, K. I. Voskresenskii, G. P., Lyubimov, A. N., Rusanov, V. V., "Prostranstvennoie obtekanie gladkikh tel idealnim gazom," M., "Nauka" (1964): translated as "Three-Dimensional Flow of Ideal Gases Around Smooth Bodies," Jerusalem, Israel Program for Scientific Translation (1968).

3. Belotserkovskii, O. M. et al., "Chislennoie issledovanie sovremennikh zadach gazovoi dinamiki," M., "Nauka" (1974).

4. Bakhvalov, N. S., "Chislennie metodi," M., "Nauka" (1975): translated as "Numerical Methods," Mir.

5. Forsythe, G. E. and Wasow, W. R., "Finite Difference Methods for Partial Differential Equations," Wiley (1960).

6. Gavurin, M. K., "Lektsii po metodam vichislenii," M., "Nauka" (1971).

7. Gelfand, A. O., "Ischislenie konechnikh raznostei," M., "Nauka" (1967).

8. Godonuv, S. K., "Uravnenia matematicheskoi fiziki," M., "Nauka" (1971).

9. Godonuv. S. K., Zabrodin, A. V., Ivanov, M. Ya., Kraiko, A. N., Prokopov, G. P., "Chislennoie mnogomernikh zadach gazovoi dinamiki," M., "Nauka", 1976.

10. Godynov, S. K. and Ryabenkii, V. S., "Vvednie v teoriyu raznostnikh skhem," M., Fizmatgiz (1962): translated as "Theory of Difference Schemes, an Introduction," Interscience (1964).

11. Dyachenko, V. F., "Osnovnie ponyatia vichislitelnoi matematiki," M., "Nauka" (1972).

12. Zienkiewicz, O. C., "The Finite Element Method," 3rd ed., McGraw-Hill (1977).

13. Ladizhenskaya, O. A., "Kraevie zadachi matematicheskoi fiziki," M., "Nauka" (1973): translated as "The Boundary Value Problems of Mathematical Physics," Springer (1985).

14. Lyubimov, A. N. and Rusanov, V. V., "Techenie gaza okolo tupikh tel," Ch. 1, M. (1970).

15. Marchuk, G. I. "Metodi vichislitelnoi matematiki," Novosibirsk, "Nauka" (1978): translated as "Methods of Numerical Mathematics," Springer (1985).

16. Marchuk, G. I. and Lebedev, V. I., "Chislennie metodi v teorie perenosa neitronov," M., Atomizdat (1971): translated as "Numerical Methods in the Theory of Neutron Transport," Harwood (1986).

17. Mikhlin, S. G., "Chislennaya realizatsiya variatsionnikh metodov," M. "Nauka" (1968): translated as "The Numerical Performance of Variational Methods" Groningen, Volters-Noordhoff (1971).

18. Oden, J. T., "Finite Elements of Nonlinear Continua," McGraw-Hill (1972).

19. Petrovskii, I. G., "Lektsii ob uravneniakh s chastnimi proizvodnimi," M., Fizmatgiz (1961): translated as "Lectures on Partial Differential Equations," Interscience (1954).

20. Richtmeyer, R. D. and Morton, K. W., "Difference Methods for Initial Value Problems," 2'nd ed., Interscience (1967).

21. Rozhdestvenskii, B. L. and Yanenko, N. N., "Sistemi kvazilineinikh uravnenii i ikh prilozheniya v gazovoi dinamike," M. "Nauka" (1968): translated as "Systems of Quazilinear Equations and Their Applications to Gas Dynamics," Am. Math. Soc. (1983).

22. Ryabenkii, V. S. and Fillipov, A. F. "Ob ustoichivosti raznostnikh uravnenii," M. Gostekhizdat (1956).

23. Samarskii, A. A., "Vvedenie v teoriyu raznostnikh skhem," M. "Nauka" (1971).

24. Samarskii, A. A. and Gulin, A. V., "Ustoichivost raznostnikh skhem," M. "Nauka" (1973).

25. Samarskii, A. A. and Andreev, V. B. "Raznostnie metodi dlya ellipticheskikh uravnenii," M., "Nauka" (1975).

26. Samarskii, A. A. and Popov, Yu. P., "Raznostnie skhemi gazovoi dinamiki," M., "Nauka" (1975).

27. Tikhonov, A. N. and Samarskii, A. A. "Uravneniya matematicheskoi fiziki," M. "Nauka" (1972): translated as "Equations of Mathematical Physics," 2nd ed., Pergamon Press (1963).

28. Yanenko, N. N., "Metod drobnikh shagov resheniya mnogomernikh zadach matematicheskoi fiziki," Novosibirsk, "Nauka" (1967): translated as "The Method of Fractional Steps: the Solution of Problems of Mathematical Physics in Several Variables," Springer (1971).

29. Strang, W. G. and Fix, G. J., "An Analysis of the Finite Element Method," Prentice Hall (1973).

INDEX

This Page Intentionally Left Blank