7

Parabolic Equati

7.1 INTRODUCTION

The prototype parabolic differential equation is the heat equation

$$u_{xx} - u_y = 0 ,$$

and we will examine it first. Because, physically, the variable y represents ***time*** problem to be studied, we will set y=t and examine the heat equation in its more custo form

(7.1) $$u_t = u_{xx} .$$

Two kinds of problems are of fundamental interest both mathematically physically with regard to (7.1). These are the ***initial value*** problem and the ***in boundary*** problem, which are defined as follows. In an initial value problem for one is given a function f(x) which is continuous for all values of x and one is asked t a function u(x,t) which is

(a) defined and continuous for $-\infty<x<\infty$, $0\le t$;

(b) satisfies (7.1) for $-\infty<x<\infty$, $0<t$; and

(c) satisfies the initial conditions $u(x,0)=f(x)$ at time $t=0$ for $-\infty<x<\infty$.

As shown geometrically in Figure 7.1, initial value problems are defined upper-half plane.

In an initial-boundary problem, one is given a constant a>0 and three conti functions

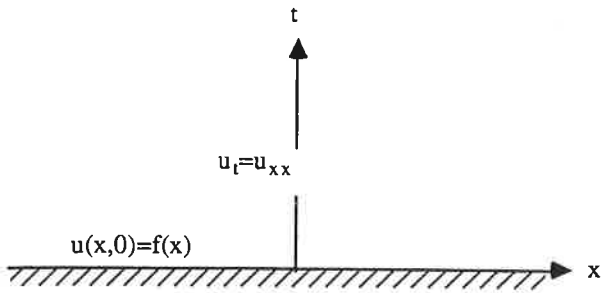

Figure 7.1

$g_1(t)$, $t\geq 0$; $g_2(t)$, $t\geq 0$; $f(x)$, $0\leq x\leq a$,

which satisfy $g_1(0)=f(0)$, $g_2(0)=f(a)$, and one is asked to find a function $u(x,t)$ which is

(a) defined and continuous for $t\geq 0$, $0\leq x\leq a$;

(b) satisfies (7.1) on $0<x<a$, $t>0$; and

(c) satisfies the initial and boundary conditions

(7.2) $u(x,0) = f(x)$, $0\leq x\leq a$, (initial condition)

(7.3) $u(0,t) = g_1(t)$, $u(a,t) = g_2(t)$, $t\geq 0$, (boundary conditions).

As shown geometrically in Figure 7.2, initial-boundary problems are defined on a semi-infinite strip.

The unique solution of an initial value problem can be given in terms of the Fourier integral, while that for an initial-boundary problem can be given in terms of series. However, again, because analytical solutions so given are not evaluated easily at particular points of interest, and, because methods for generating such solutions do not extend to nonlinear problems, we turn to numerical methods. For clarity, we will concentrate on initial-boundary problems, though most of the ideas extend to the initial value problem as well.

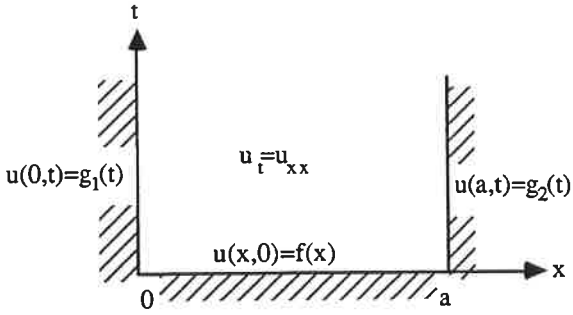

Figure 7.2

7.2 AN EXPLICIT NUMERICAL METHOD FOR THE HEAT EQUATION

Let us first develop a simple numerical method to solve the initial-boundary problem fo heat equation. Because in the initial-boundary problem the time variable t can vary i unbounded range $0 \le t < \infty$, and since the computer cannot calculate forever, we will replace $t \ge 0$ by $0 \le t \le T$. The constant value T is usually determined by the physics o phenomenon under study. Thus, in fast reaction type problems, like those related example, to the release of nuclear energy, it is often sufficient to choose T relatively s On the other hand, slow reaction type problems, like those related, for exampl radioactive decay, may require the choice of a relatively large T. In any case, T is constant, positive value.

To begin with, it is important to know that, like harmonic functions, solutio the heat equation possess the ***max-min*** property. Specifically [Forsythe and W (1960)], in the rectangle $0 \le x \le a$, $0 \le t \le T$, ***any solution of*** (7.1) ***is bounded above by the maximum and from below by the minimum of its initial boundary values.*** (The line t=T is not considered part of the boundary.)

With the above considerations in mind, let us begin by subdividing the interval into n equal parts, each of length h=a/n, by the R_{n+1} set $x_i=ih$, i=0,1,2,...,n. Nex interval [0,T] is divided into m equal parts, each of length k=T/m, by the R_{m+1} set j=0,1,2,...,m. Thus, as indicated in Figure 7.3, on the rectangle with vertices (0,0), (a,T) and (0,T), we have defined a set of planar grid points $(x_i,t_j) \in R_{n+1,m+1}$. The with coordinates $(x_i,0)$, i=0,1,2,...,n, are called the ***initial*** grid points, the points coordinates (x_0,t_j) and (x_n,t_j), j=1,2,...,m, are called the ***boundary*** grid points, a points with coordinates (x_i,t_j), i=1,2,...,n–1, j=1,2,...,m, are called the ***interio*** points. Those grid points whose coordinates are (x_i,t_j), i=0,1,2,...,n, are called the j of grid points, or, the grid points at ***time level*** t_j.

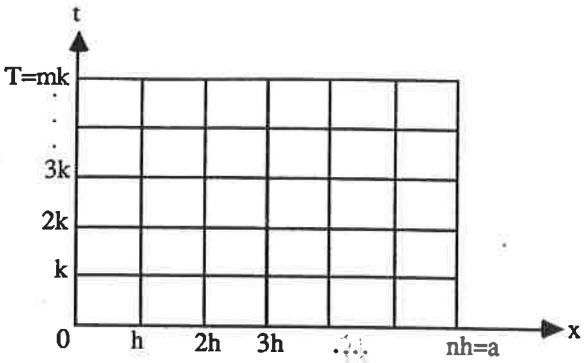

Figure 7.3

A simple numerical method for approximating the solution of initial-boundary problem (7.1)-(7.3) can be formulated now as follows. At the 0^{th} row of grid points the initial condition (7.2) implies

(7.4) $$u_{i,0} = f(x_i), \qquad i=0,1,2,\ldots,n.$$

At x=0 and at x=a, the boundary conditions (7.3) yield

(7.5) $$u_{0,j} = g_1(t_j), \qquad u_{n,j} = g_2(t_j), \qquad j=1,2,\ldots,m.$$

At each interior grid point, for the point arrangement shown in Figure 7.4, consider from Section 3.6 the approximations

(7.6) $$u_t(x_i,t_j) \approx \frac{u_{i,j+1}-u_{i,j}}{k}$$

(7.7) $$u_{xx}(x_i,t_j) \approx \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2},$$

substitution of which into (7.1) yields the approximation

(7.8) $$\frac{u_{i,j+1}-u_{i,j}}{k} = \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2},$$

or, equivalently,

(7.9) $$u_{i,j+1} = u_{i,j} + \frac{k}{h^2}(u_{i+1,j}-2u_{i,j}+u_{i-1,j}).$$

In (7.9), setting

$$\alpha = \frac{k}{h^2}$$

yields finally

(7.10) $$u_{i,j+1} = \alpha u_{i+1,j} + (1-2\alpha)u_{i,j} + \alpha u_{i-1,j}.$$

Now, on the first row of grid points, apply (7.10) with j=0 to approximate $u_{i,1}$ explicitly for each of i=1,2,...,n–1. Then, using the numerical results generated for row 1, set j= and approximate u explicitly on the second row by means of (7.5) and (7.10). Continue i the indicated fashion to approximate u explicitly at each grid point of row j+ j=2,3,...,m–1, by applying (7.10), and by making use of (7.5) and the numeric approximation generated on row j. The values u_{ij}, i=0,1,2,...,n, j=0,1,2,...,m, are calle the numerical solution on the $R_{n+1,m+1}$ set.

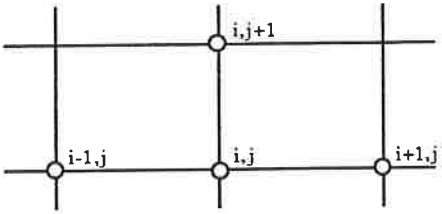

Figure 7.4

EXAMPLE 1. Consider the initial-boundary problem defined by (7.1), a=3 and

$$u(x,0) = f(x) = x^2, \qquad 0 \le x \le 3 \tag{7.11}$$

$$\begin{aligned} u(0,t) &= g_1(t) = 0 , \\ u(3,t) &= g_2(t) = 9 , \end{aligned} \qquad t>0 . \tag{7.12}$$

Set T=3, n=3, m=6 so that h=1, k=0.5, and the $R_{4,7}$ set is (i,0.5j), i=0,1,2,3, j=0,1,2,3,4,5,6. Then, at the initial time t_0=0 the initial condition (7.11) implies

$$u_{00} = 0 , \quad u_{10} = 1 , \quad u_{20} = 4 , \quad u_{30} = 9 . \tag{7.13}$$

At the boundary grid points the boundary conditions (7.12) yield

$$\begin{aligned} &u_{01} = u_{02} = u_{03} = u_{04} = u_{05} = u_{06} = 0 , \\ &u_{31} = u_{32} = u_{33} = u_{34} = u_{35} = u_{36} = 9 . \end{aligned} \tag{7.14}$$

At the interior grid points, since α=0.5, equation (7.10) reduces to

$$u_{i,j+1} = 0.5(u_{i-1,j}+u_{i+1,j}) , \qquad i=1,2. \tag{7.15}$$

Thus, for j=0 and by use of (7.13), equation (7.15) yields u_{11}=2 and u_{21}=5. Hence, the numerical solution at time level t_1=0.5 is

$$u_{01} = 0 , \quad u_{11} = 2 , \quad u_{21} = 5 , \quad u_{31} = 9 . \tag{7.16}$$

Next, for j=1 and by use of (7.16), equation (7.15) yields u_{12}=2.5 and u_{22}=5.5. Hence the numerical solution at time level t_2=1 is

$$u_{02} = 0 , \quad u_{12} = 2.5 , \quad u_{22} = 5.5 , \quad u_{32} = 9 . \tag{7.17}$$

Continuing in the indicated fashion for j=2,3,4,5 one finds

$$
\begin{array}{llll}
u_{03}=0\,, & u_{13}=2.75\,, & u_{23}=5.75\,, & u_{33}=9\,,\\
u_{04}=0\,, & u_{14}=2.875\,, & u_{24}=5.875\,, & u_{34}=9\,,\\
u_{05}=0\,, & u_{15}=2.9375\,, & u_{25}=5.9375\,, & u_{35}=9\,,\\
u_{06}=0\,, & u_{16}=2.96875\,, & u_{26}=5.96875\,, & u_{36}=9\,.
\end{array}
\tag{7.18}
$$

The numerical solution on the $R_{4,7}$ set is given finally by (7.13), (7.16), (7.17) and (7.18)

We say that the numerical solution possesses the max-min property when th discrete function u_{ij}, i=0,1,2,...,n, j=0,1,2,...,m, takes on its maximum and its minimu on the union of of the initial and the boundary grid points. Note that the discrete function u in the above example possesses the max-min property.

EXAMPLE 2. Let us reconsider the initial-boundary problem defined by (7.1), (7.11) an (7.12), and fix T=30. Set h=1. Since k=0.5 will then result in 60 rows of grid points, le us set k=5 so that only 6 rows will result. Then, α=5, and equation (7.10) reduces to

$$
u_{i,j+1} = 5u_{i-1,j} - 9u_{i,j} + 5u_{i+1,j}\,, \qquad i=1,2. \tag{7.19}
$$

Thus, for j=0 and by use of (7.13), equation (7.19) yields u_{11}=11 and u_{21}=14. Hence, th numerical solution at time level t_1=5 is

$$
u_{01}=0\,, \quad u_{11}=11\,, \quad u_{21}=14\,, \quad u_{31}=9\,. \tag{7.20}
$$

Next, for j=1,2,3,4,5 and by use of the numerical solution generated for row j, equatio (7.19) yields

$$
\begin{array}{llll}
u_{02}=0\,, & u_{12}=-29\,, & u_{22}=-26\,, & u_{32}=9\,,\\
u_{03}=0\,, & u_{13}=131\,, & u_{23}=134\,, & u_{33}=9\,,\\
u_{04}=0\,, & u_{14}=-509\,, & u_{24}=-506\,, & u_{34}=9\,,\\
u_{05}=0\,, & u_{15}=2051\,, & u_{25}=1964\,, & u_{35}=9\,,\\
u_{06}=0\,, & u_{16}=-8639\,, & u_{26}=-7376\,, & u_{36}=9\,.
\end{array}
\tag{7.21}
$$

Thus, the numerical solution on the $R_{4,7}$ set is now given by (7.13), (7.20) and (7.21). Note, however, that this solution has a nonphysical behavior, since it oscillates with

positive and negative values. Indeed it does not possess the max-min property, and, if one proceeds further with j=6,7,8,..., overflow will occur readily.

To develop a condition for (7.10) to yield a numerical solution which possesses the max-min property, consider the following simple argument. For a given initial-boundary problem, let h=a/2, k=T/m, so that the points of $R_{3,m+1}$ are shown in Figure 7.5. At the point (a/2,0), set $u_{10}=\varepsilon>0$. At the remaining initial and boundary grid points, set $u_{0j}=u_{2j}=0$, j=0,1,2,...,m. Then, application of (7.10) at the interior grid points yields

$$u_{11} = (1-2\alpha)\varepsilon\,, \qquad u_{12} = (1-2\alpha)u_{11} = (1-2\alpha)^2\varepsilon\,, \qquad u_{13} = (1-2\alpha)^3\varepsilon\,, \qquad \ldots\,,$$

and, in general,

$$u_{1j} = (1-2\alpha)^j\varepsilon\,, \qquad j=0,1,2,\ldots,m. \tag{7.22}$$

Now, to have the max-min property, we require from (7.22) that $0\le u_{1j}\le\varepsilon$, j=1,2,...,m, that is,

$$0 \le (1-2\alpha)^j\varepsilon \le \varepsilon\,.$$

or

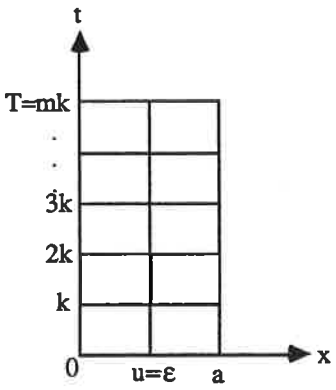

Figure 7.5

$$0 \le 1-2\alpha \le 1 .$$

Thus, since $\alpha>0$, this implies

$$\alpha \le \frac{1}{2}, \tag{7.23}$$

which is equivalent to

$$k \le \frac{h^2}{2}. \tag{7.24}$$

Inequality (7.24) is, indeed, the correct mathematical condition for the finite difference approximation (7.8) to possess the max-min property. This condition was violated in Example 2 where we selected $\alpha=5>1/2$. A rigorous development will be given later.

7.3 THE GENERAL LINEAR PARABOLIC EQUATION

The ideas developed thus far will be extended next to general linear parabolic equations as follows.

Let $a>0$ be a fixed constant and let $P(x,t)$, $Q(x,t)$, $R(x,t)$ and $S(x,t)$ be bounded and continuous for $0\le x\le a$, $t\ge 0$. Then consider the initial-boundary problem

$$u_t = P(x,t)u_{xx}+Q(x,t)u_x+R(x,t)u+S(x,t) , \qquad 0<x<a , \qquad t>0 \tag{7.25}$$

$$u(x,0) = f(x) , \qquad 0\le x\le a \tag{7.26}$$

$$u(0,t) = g_1(t) , \qquad u(a,t) = g_2(t) , \qquad t\ge 0, \tag{7.27}$$

in which $g_1(0)=f(0)$ and $g_2(0)=f(a)$. We will assume that $P(x,t)\ge v>0$ and $R(x,t)\le 0$, in order to be assured that the solution of the initial-boundary problem (7.25)-(7.27) exists, is unique, and has a general ***max-min*** property, which can be stated as follows. If $S(x,t)\equiv 0$, then in the rectangle $0\le x\le a$, $0\le t\le T$, any positive maximum of a solution of (7.25) is bounded above by the maximum of its initial and boundary values. Similarly, any negative minimum is bounded below by the minimum of its initial and boundary values. If, in

addition, $R(x,t)\equiv 0$, then any solution of (7.25) is bounded above by the maximum and below by the minimum of its initial and boundary values. (The line $t=T$ is not considered part of the boundary.)

A numerical method for (7.25)-(7.27) can be formulated as follows. Select a final time $T>0$. Subdivide the interval $[0,a]$ into n equal parts, each of length $h=a/n$, by an R_{n+1} set $x_i=ih$, $i=0,1,2,\ldots,n$. Next, subdivide the interval $[0,T]$ into m equal parts, each of length $k=T/m$, by an R_{m+1} set $t_j=jk$, $j=0,1,2,\ldots,m$. At the initial grid points, from (7.26), set

$$u_{i,0} = f(x_i), \qquad i=0,1,2,\ldots,n. \tag{7.28}$$

At the boundary grid points, from (7.27), set

$$u_{0,j} = g_1(t_j), \qquad u_{n,j} = g_2(t_j), \qquad j=1,2,\ldots,m. \tag{7.29}$$

In order to compute u_{ij} at the interior grid points, consider the approximations

$$u_t(x_i,t_j) \approx \frac{u_{i,j+1}-u_{i,j}}{k}, \tag{7.30}$$

$$u_x(x_i,t_j) \approx \frac{u_{i+1,j}-u_{i-1,j}}{2h}, \tag{7.31}$$

$$u_{xx}(x_i,t_j) \approx \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}. \tag{7.32}$$

Substitution of (7.30)-(7.32) into (7.25) then yields

$$\frac{u_{i,j+1}-u_{i,j}}{k} = P_{ij}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} + Q_{ij}\frac{u_{i+1,j}-u_{i-1,j}}{2h} + R_{ij}u_{i,j} + S_{ij}. \tag{7.33}$$

It follows that (7.33) can be rewritten, equivalently, as

$$u_{i,j+1} = (\alpha P_{ij}-\beta Q_{ij})u_{i-1,j}+(1+kR_{ij}-2\alpha P_{ij})u_{i,j}+(\alpha P_{ij}+\beta Q_{ij})u_{i+1,j}+kS_{ij}, \tag{7.34}$$

where

$$\alpha = \frac{k}{h^2}, \qquad \beta = \frac{k}{2h}.$$

Equation (7.34) can now be used to generate explicitly the numerical solution at each interior grid point at time t_{j+1}. The numerical method so obtained is called the ***explicit central difference*** method.

EXAMPLE. Consider the initial-boundary problem defined by

$$u_t = u_{xx} + (x-2)u_x - 3u \tag{7.35}$$

$$u(x,0) = x^2 - 4x + 5, \qquad 0 \le x \le 4 \tag{7.36}$$

$$u(0,t) = u(4,t) = 5e^{-t}, \qquad t > 0 \tag{7.37}$$

the exact solution of which is $(x^2-4x+5)e^{-t}$. To generate the numerical solution, let T=1, n=4, m=10 so that h=1, k=0.1, and the $R_{5,11}$ set is (i,0.1j), i=0,1,2,3,4, j=0,1,2,...,10. Then, at the initial time t_0=0 the initial condition (7.36) implies

$$u_{0,0} = 5, \quad u_{1,0} = 2, \quad u_{2,0} = 1, \quad u_{3,0} = 2, \quad u_{4,0} = 5. \tag{7.38}$$

At the boundary grid points the boundary conditions (7.37), to three decimal places, yield

$$\begin{aligned}
u_{0,0} &= 5.000 & u_{4,0} &= 5.000 \\
u_{0,1} &= 4.524 & u_{4,1} &= 4.524 \\
u_{0,2} &= 4.094 & u_{4,2} &= 4.094 \\
u_{0,3} &= 3.704 & u_{4,3} &= 3.704 \\
u_{0,4} &= 3.352 & u_{4,4} &= 3.352 \\
u_{0,5} &= 3.033 & u_{4,5} &= 3.033 \\
u_{0,6} &= 2.744 & u_{4,6} &= 2.744 \\
u_{0,7} &= 2.483 & u_{4,7} &= 2.483 \\
u_{0,8} &= 2.247 & u_{4,8} &= 2.247 \\
u_{0,9} &= 2.033 & u_{4,9} &= 2.033 \\
u_{0,10} &= 1.839 & u_{4,10} &= 1.839 .
\end{aligned} \tag{7.39}$$

At the interior grid points, since $\alpha=0.1$ and $\beta=0.05$, equation (7.34) becomes

$$u_{i,j+1} = 0.1[1-0.5(x_i-2)]u_{i-1,j}+0.5u_{i,j}+0.1[1+0.5(x_i-2)]u_{i+1,j}. \tag{7.40}$$

Thus, by using (7.38), (7.39), and calculating to three decimal places, equation (7.40), for $j=0,1,2,\ldots,9$, yields

$$\begin{aligned}
u_{1,1} &= 1.800 & u_{2,1} &= 0.900 & u_{3,1} &= 1.800\\
u_{1,2} &= 1.624 & u_{2,2} &= 0.810 & u_{3,2} &= 1.624\\
u_{1,3} &= 1.466 & u_{2,3} &= 0.730 & u_{3,3} &= 1.466\\
u_{1,4} &= 1.325 & u_{2,4} &= 0.658 & u_{3,4} &= 1.325\\
u_{1,5} &= 1.198 & u_{2,5} &= 0.594 & u_{3,5} &= 1.198\\
u_{1,6} &= 1.084 & u_{2,6} &= 0.537 & u_{3,6} &= 1.084\\
u_{1,7} &= 0.980 & u_{2,7} &= 0.485 & u_{3,7} &= 0.980\\
u_{1,8} &= 0.887 & u_{2,8} &= 0.439 & u_{3,8} &= 0.887\\
u_{1,9} &= 0.802 & u_{2,9} &= 0.397 & u_{3,9} &= 0.802\\
u_{1,10} &= 0.726 & u_{2,10} &= 0.359 & u_{3,10} &= 0.726,
\end{aligned}$$

which is the numerical solution at the interior grid points on the given $R_{5,11}$ set.

With regard to the max-min property of the numerical solution of parabolic equation (7.25), we proceed as follows.

DEFINITION 7.1. For $S(x,t)\equiv 0$, a numerical solution u_{ij} of (7.25) is said to possess the weak max-min property if u_{ij} satisfies the following inequalities for $j=1,2,\ldots,m$:

$$u_{ij} \geq \min[0, \min_{1\leq q\leq J}(u_{0q}), \min_{0\leq p\leq n}(u_{p0}), \min_{1\leq q\leq J}(u_{nq})] \tag{7.41}$$

$$u_{ij} \leq \max[0, \max_{1\leq q\leq J}(u_{0q}), \max_{0\leq p\leq n}(u_{p0}), \max_{1\leq q\leq J}(u_{nq})]. \tag{7.42}$$

Note that the numerical solution obtained from the example above possesses the weak max-min property since inequalities (7.41) and (7.42) are satisfied.

DEFINITION 7.2. For $S(x,t)\equiv R(x,t)\equiv 0$, a numerical solution u_{ij} of (7.25) is said to possess the strong max-min property if u_{ij} satisfies the following inequalities:

$$u_{ij} \geq \min\left[\min_{1\leq q\leq j}(u_{0q}), \min_{0\leq p\leq n}(u_{p0}), \min_{1\leq q\leq j}(u_{nq})\right] \tag{7.43}$$

$$u_{ij} \leq \max\left[\max_{1\leq q\leq j}(u_{0q}), \max_{0\leq p\leq n}(u_{p0}), \max_{1\leq q\leq j}(u_{nq})\right] . \tag{7.44}$$

If either of Definition 7.1 or 7.2 is applicable, one says that the numerical solution has the max-min property. Note also that only when $S(x,t)\equiv 0$ is the max-min property defined.

Before continuing, it is important to make the following observation. The explicit finite difference scheme (7.34) may yield a numerical solution which does not possess the max-min property. In fact, for $S(x,t)\equiv 0$, equation (7.34) reduces to

$$u_{i,j+1} = (\alpha P_{ij}-\beta Q_{ij})u_{i-1,j}+(1+kR_{ij}-2\alpha P_{ij})u_{i,j}+(\alpha P_{ij}+\beta Q_{ij})u_{i+1,j} . \tag{7.45}$$

Now, if any one of the coefficients of $u_{i-1,j}$, $u_{i,j}$ $u_{i+1,j}$ in the right-hand side of (7.45) is negative, then it is easy to construct an example of an initial-boundary problem whose numerical solution does not possess the max-min property. Assume, for instance, that for $j=0$ and $0<i<n$ one has $(\alpha P_{i0}-\beta Q_{i0})<0$. At the point (x_{i-1},t_0), set $u_{i-1,0}=\varepsilon>0$. At the remaining initial and boundary grid points of $R_{n+1,m+1}$, set $u=0$ so that the max-min property implies $0\leq u\leq\varepsilon$ at each interior grid point. However, application of (7.45) at (x_i,t_0) yields

$$u_{i,1} = (\alpha P_{i0}-\beta Q_{i0})\varepsilon < 0 .$$

In order to be assured, ***a priori***, that the finite difference scheme (7.45) does imply the max-min property, it is necessary and sufficient that the space step h and the time step k satisfy appropriate conditions. The sufficiency is proved in the following theorem. The necessity appears later as an exercise.

THEOREM 7.1. For $S(x,t)\equiv 0$, let the initial-boundary problem (7.25)-(7.27) be defined on $[0,a]$, $0\leq t\leq T$, where $P(x,t)$, $Q(x,t)$ and $R(x,t)$ are continuous, and hence bounded. Assume that $V\geq P(x,t)\geq v>0$, $|Q(x,t)|\leq M$ and $-N\leq R(x,t)\leq 0$ on $0\leq x\leq a$, $0\leq t\leq T$. If

(7.46) $\mathbf{Mh \le 2\nu}$

and

(7.47) $$k \le \frac{h^2}{Nh^2+2V},$$

then the numerical solution of (7.25)-(7.27), obtained with the central difference formula (7.34), possesses the max-min property.

PROOF. Note first that inequality (7.46) implies

$$(\alpha P_{ij}-\beta Q_{ij}) \ge (\alpha P_{ij}-\beta|Q_{ij}|) \ge (\alpha\nu-\beta M) \ge 0,$$

and

$$(\alpha P_{ij}+\beta Q_{ij}) \ge (\alpha P_{ij}-\beta|Q_{ij}|) \ge (\alpha\nu-\beta M) \ge 0,$$

while inequality (7.47) implies

$$(1+kR_{ij}-2\alpha P_{ij}) \ge (1-kN-2\alpha V) \ge 0.$$

Moreover,

(7.48) $$(\alpha P_{ij}-\beta Q_{ij})+(1+kR_{ij}-2\alpha P_{ij})+(\alpha P_{ij}+\beta Q_{ij}) = 1+kR_{ij} \le 1.$$

Thus, (7.45) implies

(7.49) $$u_{i,j+1} \le (\alpha P_{ij}-\beta Q_{ij}) \max_{0\le p\le n}(u_{pj})+(1+kR_{ij}-2\alpha P_{ij}) \max_{0\le p\le n}(u_{pj})$$

$$+(\alpha P_{ij}+\beta Q_{ij}) \max_{0\le p\le n}(u_{pj}) = (1+kR_{ij}) \max_{0\le p\le n}(u_{pj}),$$

which, since $0\le(1+kR_{ij})\le 1$, implies

(7.50) $$u_{i,j+1} \le \max[0, \max_{0\le p\le n}(u_{pj})] = \max[0, u_{0,j}, \max_{0<p<n}(u_{p,j}), u_{n,j}] .$$

Now, because (7.50) is valid for i=1,2,...,n–1, it follows that

(7.51) $$\begin{aligned} u_{i,j} &\le \max[0, u_{0,j-1}, \max_{0<p<n}(u_{p,j-1}), u_{n,j-1}] \\ &\le \max[0, u_{0,j-1}, u_{0,j-2}, \max_{0<p<n}(u_{p,j-2}), u_{n,j-2}, u_{n,j-1}] \\ &\le \max[0, u_{0,j-1}, u_{0,j-2}, u_{0,j-3}, \max_{0<p<n}(u_{p,j-3}), u_{n,j-3}, u_{n,j-2}, u_{n,j-1}] \\ &\ \ \vdots \\ &\le \max[0, \max_{0\le q<j}(u_{0,q}), \max_{0<p<n}(u_{p,0}), \max_{0\le q<j}(u_{n,q})] , \end{aligned}$$

which implies (7.42). Inequality (7.41) can be proved in an entirely similar way.

When R_{ij}=0, i=0,1,2,...,n, j=0,1,2,...,m, inequality (7.49) reduces to

$$u_{i,j+1} \le \max_{0\le p\le n}(u_{pj}) = \max[u_{0,j}, \max_{0<p<n}(u_{p,j}), u_{n,j}] ,$$

which can be used to yield

(7.52) $$\begin{aligned} u_{i,j} &\le \max[u_{0,j-1}, \max_{0<p<n}(u_{p,j-1}), u_{n,j-1}] \\ &\le \max[u_{0,j-1}, u_{0,j-2}, \max_{0<p<n}(u_{p,j-2}), u_{n,j-2}, u_{n,j-1}] \\ &\le \max[u_{0,j-1}, u_{0,j-2}, u_{0,j-3}, \max_{0<p<n}(u_{p,j-3}), u_{n,j-3}, u_{n,j-2}, u_{n,j-1}] \\ &\ \ \vdots \\ &\le \max[\max_{0\le q<j}(u_{0,q}), \max_{0<p<n}(u_{p,0}), \max_{0\le q<j}(u_{n,q})] , \end{aligned}$$

which implies (7.44). Inequality (7.43) can be proved in a similar fashion.

Note that under the hypotheses of Theorem 7.1, if f(x), $g_1(t)$ and $g_2(t)$ are bounded,

the numerical solution of initial-boundary problem (7.25)-(7.27), obtained with the explicit central difference scheme (7.41), remains bounded for any T. Hence, inequalities (7.46), (7.47) can also be regarded as the ***sufficient*** (not necessary) ***stability conditions*** for the method. Note also that (7.46) is a restriction on the space step h. Since h=a/n, it implies that the number of parts n in which the interval [0,a] is to be subdivided must satisfy the following inequality:

$$n \geq \frac{aM}{2v}. \tag{7.53}$$

Once h has been fixed, one then should choose a time step k in such a way that inequality (7.47) is also satisfied. Since h=a/n and k=T/m, inequality (7.47) is equivalent to

$$m \geq T\left[N+\frac{2V}{h^2}\right] = T\left[N+\frac{2Vn^2}{a^2}\right]. \tag{7.54}$$

From (7.53), (7.54) implies that if the space subdivision n is to be taken large, the corresponding time subdivision m will be required to be much larger, in which case the numerical method may no longer be convenient.

7.4 AN EXPLICIT UPWIND METHOD

If one wishes to eliminate the restriction (7.46) on the space step h, one can use an upwind, rather than a central, difference approximation to discretize the term u_x in (7.25). Specifically, one proceeds as follows. The terms u_t and u_{xx} are approximated by (7.30) and (7.32), respectively, while the term u_x in (7.25) is approximated by

$$u_x \approx \frac{u_{i+1,j}-u_{i,j}}{h}, \quad \text{if} \quad Q_{ij} \geq 0; \qquad u_x \approx \frac{u_{i,j}-u_{i-1,j}}{h}, \quad \text{if} \quad Q_{ij} < 0; \tag{7.55}$$

Substitution of (7.30), (7.32) and (7.55) into (7.25) then yields

$$\frac{u_{i,j+1}-u_{i,j}}{k} = P_{ij}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} + Q_{ij}\frac{u_{i+1,j}-u_{i,j}}{h} + R_{ij}u_{i,j} + S_{ij}, \quad \text{if} \quad Q_{ij} \geq 0;$$

$$\frac{u_{i,j+1}-u_{i,j}}{k} = P_{ij}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} + Q_{ij}\frac{u_{i,j}-u_{i-1,j}}{h} + R_{ij}u_{i,j} + S_{ij}, \quad \text{if} \quad Q_{ij} < 0,$$

or, equivalently,

$$(7.56) \qquad u_{i,j+1} = [\alpha P_{ij}+\beta(|Q_{ij}|-Q_{ij})]u_{i-1,j} + [1+kR_{ij}-2(\alpha P_{ij}+\beta|Q_{ij}|)]u_{i,j}$$

$$+ [\alpha P_{ij}+\beta(|Q_{ij}|+Q_{ij})]u_{i+1,j}+kS_{ij} ,$$

where, of course, $\alpha=k/(h^2)$ and $\beta=k/(2h)$. The algorithm to be used to generate the numerical solution at each interior grid point is the same as described in the previous section but with (7.56) replacing (7.34). The method so obtained is called the ***explicit upwind*** method.

EXAMPLE. Let us reconsider the initial-boundary problem (7.35)-(7.37). Let T=1, n=4, m=10 so that h=1, k=0.1, and the $R_{5,11}$ set is (i,0.1j), i=0,1,2,3,4, j=0,1,2,...,10. Then, by using (7.38), equation (7.56) yields for j=0

$$u_{1,1} = 1.900 \qquad u_{2,1} = 0.900 \qquad u_{3,1} = 1.900 .$$

Next, by use of (7.38) and the numerical solution just generated, for j=1, equation (7.56) yields

$$u_{1,2} = 1.758 \qquad u_{2,2} = 0.830 \qquad u_{3,2} = 1.758 .$$

Continuing in the indicated fashion, for j=2,3,...,9, one finds, to three decimal places,

$$\begin{array}{lll} u_{1,3} = 1.604 & u_{2,3} = 0.766 & u_{3,3} = 1.604 \\ u_{1,4} = 1.459 & u_{2,4} = 0.704 & u_{3,4} = 1.459 \\ u_{1,5} = 1.324 & u_{2,5} = 0.647 & u_{3,5} = 1.324 \\ u_{1,6} = 1.201 & u_{2,6} = 0.587 & u_{3,6} = 1.201 \\ u_{1,7} = 1.088 & u_{2,7} = 0.533 & u_{3,7} = 1.088 \\ u_{1,8} = 0.985 & u_{2,8} = 0.484 & u_{3,8} = 0.985 \\ u_{1,9} = 0.892 & u_{2,9} = 0.439 & u_{3,9} = 0.892 \\ u_{1,10} = 0.807 & u_{2,10} = 0.398 & u_{3,10} = 0.807 , \end{array}$$

Table 7.1 - Explicit methods.

exact	$u_{1,10}=0.736$	$u_{2,10}=0.368$	$u_{3,10}=0.736$
central	$u_{1,10}=0.726$	$u_{2,10}=0.359$	$u_{3,10}=0.726$
upwind	$u_{1,10}=0.807$	$u_{2,10}=0.398$	$u_{3,10}=0.807$

which is another approximation to the exact solution $u=(x^2-4x+5)e^{-t}$. Note that the present numerical solution possesses the discrete max-min property. It, however, is not as accurate as the one which was obtained in the previous section by the central difference method. For clarity in this comparison, we have listed in Table 7.1 the values of the exact solution at time T=1, the results by the central difference method from the example of the previous section, and the results just generated.

Although the explicit upwind method is not very accurate, it yields a numerical solution which possesses the max-min property for any choice of the space step h. This is proved in the next theorem.

THEOREM 7.2. For $S(x,t)\equiv 0$, let the initial-boundary problem (7.25)-(7.27) be defined on $0\le x\le a$, $0\le t\le T$, where $P(x,t)$, $Q(x,t)$ and $R(x,t)$ are continuous, and hence bounded. Assume that $0<\nu\le P(x,t)\le V$, $|Q(x,t)|\le M$ and $-N\le R(x,t)\le 0$ on $0\le x\le a$, $0\le t\le T$. If

$$k \le \frac{h^2}{Nh^2+Mh+2V}, \tag{7.57}$$

then the numerical solution of (7.25)-(7.27), obtained with the upwind difference formula (7.56), possesses the max-min property.

PROOF. Note first that, for $S(x,t)\equiv 0$, equation (7.56) reduces to

$$u_{i,j+1} = [\alpha P_{ij}+\beta(|Q_{ij}|-Q_{ij})]u_{i-1,j} + [1+kR_{ij}-2(\alpha P_{ij}+\beta|Q_{ij}|)]u_{i,j} + [\alpha P_{ij}+\beta(|Q_{ij}|+Q_{ij})]u_{i+1,j}, \tag{7.58}$$

where

$$\alpha P_{ij}+\beta(|Q_{ij}|-Q_{ij}) \ge 0, \qquad \alpha P_{ij}+\beta(|Q_{ij}|+Q_{ij}) \ge 0,$$

and, by inequality (7.57),

$$1+kR_{ij}-2(\alpha P_{ij}+\beta|Q_{ij}|) \geq 1-kN-2(\alpha V+\beta M) \geq 0 .$$

Moreover,

$$(7.59) \qquad [\alpha P_{ij}+\beta(|Q_{ij}|-Q_{ij})]+[1+kR_{ij}-2(\alpha P_{ij}+\beta|Q_{ij}|)]+[\alpha P_{ij}+\beta(|Q_{ij}|+Q_{ij})]=1+kR_{ij} \leq 1 .$$

The proof now follows directly as in Theorem 7.1.

Note that under the hypotheses of Theorem 7.2, if $f(x)$, $g_1(t)$ and $g_2(t)$ are bounded, the numerical solution of initial-boundary problem (7.25)-(7.27) remains bounded for any fixed T, and hence, inequality (7.57) also constitutes a ***stability condition*** for the method.

7.5 NUMERICAL SOLUTION OF MILDLY NONLINEAR PROBLEMS

The explicit central difference and upwind method extend in a natural way to mildly nonlinear initial-boundary problems defined by (7.26), (7.27) and

$$(7.60) \qquad u_t = P(x,t)u_{xx}+Q(x,t)u_x+F(x,t,u) , \qquad P(x,t) > 0 .$$

To assure that the solution of (7.26), (7.27) and (7.60) exists, is unique, and has properties in common with the solution of the corresponding linear equation (7.25), it will be assumed that F is bounded and F_u exists and satisfies

$$(7.61) \qquad -N \leq F_u \leq 0 , \qquad 0 \leq x \leq a, \qquad 0 \leq t \leq T, \qquad -\infty < u < \infty .$$

The only modifications necessary in the algorithms for the central difference and upwind method when (7.60) replaces (7.25) are that the difference equations must be modified appropriately. For the ***central difference*** method, one need only replace (7.34) with

$$(7.62) \qquad u_{i,j+1} = (\alpha P_{ij}-\beta Q_{ij})u_{i-1,j}+(1-2\alpha P_{ij})u_{i,j}+(\alpha P_{ij}+\beta Q_{ij})u_{i+1,j}+kF(x_i,t_j,u_{i,j}) .$$

For the ***upwind*** method, one need only replace (7.56) with

$$u_{i,j+1} = [\alpha P_{ij}+\beta(|Q_{ij}|-Q_{ij})]u_{i-1,j} + [1-2(\alpha P_{ij}+\beta|Q_{ij}|)]u_{i,j} \tag{7.63}$$

$$+ [\alpha P_{ij}+\beta(|Q_{ij}|+Q_{ij})]u_{i+1,j}+kF(x_i,t_j,u_{i,j}) .$$

With regard to stability, the central difference and the upwind methods continue to remain stable under the constraints of Theorems 7.1 and 7.2, respectively, with F_u replacing R.

EXAMPLE. Consider the initial-boundary problem defined by

$$u_t = u_{xx}-[u+\cos(u)] , \qquad 0<x<5, \qquad 0<t\le 1 \tag{7.64}$$

$$u(x,0) = 1 , \qquad 0\le x\le 5 \tag{7.65}$$

$$u(0,t) = u(5,t) = 1 , \qquad 0<t\le 1 . \tag{7.66}$$

Note first that $F(x,t,u)=-[u+\cos(u)]$. Thus, $F_u=-1+\sin(u)$, which implies $-2\le F_u\le 0$. Moreover, since in this example $Q(x,t)\equiv 0$, one has $M=0$ and the central difference method and the upwind method are equivalent. Now, since $M=0$, h, and hence n, can be chosen arbitrarily. Let us then fix $n=5$, so that $h=1$. The stability restriction (7.47) implies $k\le 1/4$. Hence, fix $m=5$, so that $k=1/5=0.2$, and the finite difference formula (7.62), in this example, reduces to

$$u_{i,j+1} = \alpha u_{i-1,j}+(1-2\alpha)u_{i,j}+\alpha u_{i+1,j}-k[u_{i,j}+\cos(u_{i,j})] ,$$

that is,

$$u_{i,j+1} = (0.2)u_{i-1,j}+(0.4)u_{i,j}+(0.2)u_{i+1,j}-(0.2)\cos(u_{i,j}) . \tag{7.67}$$

Now, at $t_0=0$, (7.65) yields

$$u_{00} = u_{10} = u_{20} = u_{30} = u_{40} = u_{50} = 1 .$$

At x=0 and at x=5, (7.66) yields

$$u_{01} = u_{02} = u_{03} = u_{04} = u_{05} = 1$$
$$u_{51} = u_{52} = u_{53} = u_{54} = u_{55} = 1 .$$

By setting j=0, i=1,2,3,4, and calculating to four decimal places, (7.67) yields, explicitly, the following numerical solution at time level t_1=k:

$$u_{11} = 0.6919 \qquad u_{21} = 0.6919 \qquad u_{31} = 0.6919 \qquad u_{41} = 0.6919 .$$

Next, by setting j=1 and i=1,2,3,4, the numerical solution that one obtains from (7.67) at t_2=2k is

$$u_{12} = 0.4611 \qquad u_{22} = 0.3995 \qquad u_{32} = 0.3995 \qquad u_{42} = 0.4611 .$$

Continuing in the indicated fashion for j=2,3,4, (7.67) yields

$$u_{13} = 0.2852 \qquad u_{23} = 0.1477 \qquad u_{33} = 0.1477 \qquad u_{43} = 0.2852$$
$$u_{14} = 0.1517 \qquad u_{24} = -0.0521 \qquad u_{34} = -0.0521 \qquad u_{44} = 0.1517$$
$$u_{15} = 0.0525 \qquad u_{25} = -0.2006 \qquad u_{35} = -0.2006 \qquad u_{45} = 0.0525 ,$$

which gives the numerical solution.

*7.6. CONVERGENCE OF EXPLICIT FINITE DIFFERENCE METHODS

Throughout this section, we will consider mildly nonlinear boundary value problems of the form (7.26), (7.27), (7.60) under the assumptions that F is bounded, $-N \le F_u \le 0$, $0 < v \le P(x,t) \le V$, $|Q(x,t)| \le M$, so that the analytical solution exists and is unique.

In developing convergence analyses, let us assume that the numerical solution u_{ij} is free from roundoff errors, so that the only difference between u_{ij} and the exact solution $U(x_i,t_j)$ is the error made by replacing (7.60) by the difference equation, that is, the local truncation error.

In solving (7.26), (7.27), (7.60) numerically, consider, first, the central difference approximation (7.62), for which we will prove the following theorem.

THEOREM 7.3. Let U(x,t) be the analytical solution of the mildly nonlinear initial-boundary problem (7.26), (7.27), (7.60). Let u_{ij}, i=0,1,2,...,n, j=0,1,2,...,m be the numerical solution obtained on $0\le t\le T$ with the central difference approximation (7.62), and let $E(x_i,t_j)=E_{ij}$, be the error, defined by $E_{ij}=U(x_i,t_j)-u_{ij}$. Assume that U(x,t) has continuous partial derivatives up to and including order two in t and order four in x. If $Mh\le 2\nu$ and $k\le h^2/(Nh^2+2V)$, then

$$|E_{ij}| \le T\left[\frac{h^2}{12}\left(VM_4+2MM_3\right)+\frac{k}{2}N_2\right], \tag{7.68}$$

where $N_2=\max(|U_{tt}|)$, $M_3=\max(|U_{xxx}|)$ and $M_4=\max(|U_{xxxx}|)$.

PROOF. For j=0, for i=0 and for i=n, one has, respectively,

$$E_{i0} = |U(x_i,0)-u_{i0}| = |f(x_i)-f(x_i)| = 0, \qquad i=0,1,2,\ldots,n,$$

$$E_{0j} = |U(0,t_j)-u_{0j}| = |g_1(t_j)-g_1(t_j)| = 0, \qquad j=0,1,2,\ldots,m,$$

$$E_{nj} = |U(a,t_j)-u_{nj}| = |g_2(t_j)-g_2(t_j)| = 0, \qquad j=0,1,2,\ldots,m,$$

so that (7.68) is valid. One must then prove (7.68) for j=1,2,...,m, and i=1,2,...,n–1.

Since U(x,t) is the analytical solution of (7.60), it follows that

$$U_t(x_i,t_j) = P_{ij}U_{xx}(x_i,t_j)+Q_{ij}U_x(x_i,t_j)+F(x_i,t_j,U(x_i,t_j)),$$

which, by (3.30), (3.31), (3.37), (3.38) and (3.40), (3.41), is equivalent to

$$\left[\frac{U(x_i,t_{j+1})-U(x_i,t_j)}{k}+\frac{k}{2}U_{tt}(x_i,\tau)\right] = P_{ij}\left[\frac{U(x_{i+1},t_j)-2U(x_i,t_j)+U(x_{i-1},t_j)}{h^2}\right.$$

$$\left.+\frac{h^2}{24}\left[U_{xxxx}(\xi_1,t_j)+U_{xxxx}(\xi_2,t_j)\right]\right]+Q_{ij}\left[\frac{U(x_{i+1},t_j)-U(x_{i-1},t_j)}{2h}\right.$$

$$\left.+\frac{h^2}{12}\left[U_{xxx}(\xi_3,t_j)+U_{xxx}(\xi_4,t_j)\right]\right]+F(x_i,t_j,U(x_i,t_j)),$$

that is,

(7.69) $$U(x_i,t_{j+1}) = (\alpha P_{ij}-\beta Q_{ij})U(x_{i-1},t_j)+(1-2\alpha P_{ij})U(x_i,t_j)+(\alpha P_{ij}+\beta Q_{ij})U(x_{i+1},t_j)$$
$$+kF(x_i,t_j,U(x_i,t_j))+k\frac{h^2}{24}\{P_{ij}[U_{xxxx}(\xi_1,t_j)+U_{xxxx}(\xi_2,t_j)]$$
$$+2Q_{ij}[U_{xxx}(\xi_3,t_j)+U_{xxx}(\xi_4,t_j)]\}-\frac{k^2}{2}U_{tt}(x_i,\tau)\,.$$

Now, by subtracting (7.62) from (7.69), one finds

$$E_{i,j+1} = (\alpha P_{ij}-\beta Q_{ij})E_{i-1,j}+(1-2\alpha P_{ij})E_{i,j}+(\alpha P_{ij}+\beta Q_{ij})E_{i+1,j}+k\frac{\partial F(x_i,t_j,\mu)}{\partial u}E_{i,j}$$
$$+k\frac{h^2}{24}\{P_{ij}[U_{xxxx}(\xi_1,t_j)+U_{xxxx}(\xi_2,t_j)]+2Q_{ij}[U_{xxx}(\xi_3,t_j)+U_{xxx}(\xi_4,t_j)]\}$$
$$-\frac{k^2}{2}U_{tt}(x_i,\tau)\,.$$

By replacing $\partial F(x_i,t_j,\mu)/\partial u$ by R_{ij}, which is an extension of the notation R_{ij}, one has $-N\le R_{ij}\le 0$ and the last equation can be written as

$$E_{i,j+1} = (\alpha P_{ij}-\beta Q_{ij})E_{i-1,j}+(1+kR_{ij}-2\alpha P_{ij})E_{i,j}+(\alpha P_{ij}+\beta Q_{ij})E_{i+1,j}$$
$$+k\frac{h^2}{24}\{P_{ij}[U_{xxxx}(\xi_1,t_j)+U_{xxxx}(\xi_2,t_j)]+2Q_{ij}[U_{xxx}(\xi_3,t_j)+U_{xxx}(\xi_4,t_j)]\}$$
$$-\frac{k^2}{2}U_{tt}(x_i,\tau)\,.$$

From the hypotheses, $(\alpha P_{ij}-\beta Q_{ij}) \ge 0$, $(1+kR_{ij}-2\alpha P_{ij}) \ge (1-kN-2\alpha P_{ij}) \ge 0$ and $(\alpha P_{ij}+\beta Q_{ij}) \ge 0$. Thus,

$$|E_{i,j+1}| \le (\alpha P_{ij}-\beta Q_{ij})|E_{i-1,j}|+(1+kR_{ij}-2\alpha P_{ij})|E_{i,j}|+(\alpha P_{ij}+\beta Q_{ij})|E_{i+1,j}|$$
$$+k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2]\,,$$

which implies

$$\max_i(|E_{i,j+1}|) \le [(\alpha P_{ij}-\beta Q_{ij})+(1+kR_{ij}-2\alpha P_{ij})+(\alpha P_{ij}+\beta Q_{ij})]\max_i(|E_{i,j}|)$$

$$+k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2]$$

$$=(1+kR_{ij})\max_i(|E_{i,j}|)+k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2]$$

$$\le \max_i(|E_{i,j}|)+k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2] .$$

By using this recurrence relation one obtains

$$\max_i(|E_{i,j+1}|) \le \max_i(|E_{i,j}|)+k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2]$$

$$\le \max_i(|E_{i,j-1}|)+2k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2]$$

$$\le \max_i(|E_{i,j-2}|)+3k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2]$$

$$\vdots$$

$$\le \max_i(|E_{i,0}|)+(j+1)k[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2] .$$

But, $(j+1)k=t_{j+1}\le T$ and $\max_i(|E_{i,0}|)=0$. Thus,

$$\max_i(|E_{i,j+1}|) \le T[\frac{h^2}{12}(VM_4+2MM_3)+\frac{k}{2}N_2] ,$$

which implies (7.68) and the Theorem is proved.

Theorem 7.3 allows us to give a convergence theorem which is as follows.

THEOREM 7.4. Under the assumptions of Theorem 7.3, the numerical solution of a mildly nonlinear initial-boundary problem obtained with the central difference method converges to the analytical solution as $h \to 0$, that is,

$$\lim_{h \to 0} |U(x_i,t_j) - u_{ij}| = 0 , \qquad i=0,1,2,\ldots,n, \qquad j=0,1,2,\ldots,m. \tag{7.70}$$

PROOF. Since, by hypotheses, k is required to satisfy the inequality

$$k \le \frac{h^2}{Nh^2+2V} ,$$

error bound (7.68) implies (7.70).

Analogous results to those of Theorems 7.3 and 7.4 are given, next, for the explicit upwind method (7.63).

THEOREM 7.5. Let $U(x,t)$ be the analytical solution of the mildly nonlinear initial-boundary problem (7.26), (7.27), (7.60). Let u_{ij} be the numerical solution obtained on $0 \le t \le T$ with the upwind difference formula (7.63), and let $E_{ij}=E(x_i,t_j)$, $i=0,1,2,\ldots,n$, $j=0,1,2,\ldots,m$ be the error. Assume that $U(x,t)$ has continuous partial derivatives up to and including order two in t and order four in x. If $k \le h^2/(Nh^2+Mh+2V)$, then

$$|E_{ij}| \le T\left[\frac{h}{12}\left(hVM_4+6MM_2\right)+\frac{k}{2}N_2\right] , \tag{7.71}$$

where $N_2=\max(|U_{tt}|)$, $M_2=\max(|U_{xx}|)$ and $M_4=\max(|U_{xxxx}|)$.

The proof of Theorem 7.5 is similar to that of Theorem 7.3.

From error bound (7.71) a convergence theorem follows readily.

THEOREM 7.6. Under the assumptions of Theorem 7.5, the numerical solution of a mildly nonlinear initial-boundary problem obtained with the upwind difference method converges to the analytical solution as $h \to 0$, that is,

$$\lim_{h \to 0} |U(x_i,t_j) - u_{ij}| = 0 , \qquad i=0,1,2,\ldots,n, \qquad j=0,1,2,\ldots,m.$$

In general, the reason why the central difference method, when applicable, is to be preferred to the upwind method is explained by the respective error bounds, for, if h is

small, the error bound of the central difference method is, in general, smaller than the error bound of the upwind method. In the latter case, for small h, the error bound is dominated by the term $(h/2)TMM_2$, which is called the ***artificial*** diffusion term.

Note, finally, that the results given in this section also apply to linear initial-boundary problems as a particular case when $F(x,t,u)=R(x,t)u+S(x,t)$.

7.7 IMPLICIT CENTRAL DIFFERENCE METHOD

Suppose now that one wishes to construct a method which possesses the max-min property for ***all*** choices of time step k. Such a method is desirable, for example, if one has to calculate for long periods of time. If, say, one wishes to have a numerical approximation of the simple heat equation up to T=100 and one has to choose h=0.01, then (7.24) implies that one must choose k to satisfy

$$k \le \frac{1}{2}\left(\frac{1}{100}\right)^2 = 0.00005 .$$

To generate a numerical solution at t=100 would therefore require, using the explicit method, computation on a minimum of two million rows of points of $R_{n+1,m+1}$.

Interestingly enough, conditions (7.43) and (7.57) for the central difference method and for the upwind method, respectively, can be eliminated simply by replacing the point pattern shown in Figure 7.4 by the one shown in Figure 7.6, or, more precisely, by approximating u_t with a backward rather than a forward finite difference, that is,

$$u_t \approx \frac{u_{i,j}-u_{i,j-1}}{k} . \tag{7.72}$$

To begin with, let us consider the central difference method for the mildly nonlinear initial-boundary problem. Substitution of (7.31), (7.32), (7.72) into (7.60) yields

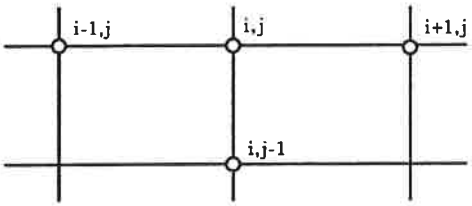

Figure 7.6

$$\frac{u_{i,j}-u_{i,j-1}}{k} = P_{ij}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} + Q_{ij}\frac{u_{i+1,j}-u_{i-1,j}}{2h} + F(x_i,t_j,u_{i,j}) \, ,$$

or, equivalently,

$$(\alpha P_{ij}-\beta Q_{ij})u_{i-1,j}-(1+2\alpha P_{ij})u_{i,j}+(\alpha P_{ij}+\beta Q_{ij})u_{i+1,j}+kF(x_i,t_j,u_{i,j})+u_{i,j-1}=0 \tag{7.73}$$

Equation (7.73) can now be used to generate the numerical solution at each interior grid point at time t_j. In fact, by setting

$$\begin{aligned} f_1(u_{1,j}) &= u_{1,j-1}+kF(x_1,t_j,u_{1,j})+(\alpha P_{1j}-\beta Q_{1j})u_{0,j}, \\ f_i(u_{i,j}) &= u_{i,j-1}+kF(x_i,t_j,u_{i,j}) \, , \qquad i=2,3,\ldots,n-2 \\ f_{n-1}(u_{n-1,j}) &= u_{n-1,j-1}+kF(x_{n-1},t_j,u_{n-1,j})+(\alpha P_{n-1,j}+\beta Q_{n-1,j})u_{n,j} \, , \end{aligned} \tag{7.74}$$

equation (7.73), written consecutively for i=1,2,…,n−1, yields

$$\begin{aligned} &-(1+2\alpha P_{1j})u_{1,j}+(\alpha P_{1j}+\beta Q_{1j})u_{2,j} && +f_1(u_{1,j}) = 0 \\ &(\alpha P_{2j}-\beta Q_{2j})u_{1,j}-(1+2\alpha P_{2j})u_{2,j}+(\alpha P_{2j}+\beta Q_{2j})u_{3,j} && +f_2(u_{2,j}) = 0 \\ &\quad\cdot \qquad \cdot \qquad \cdot && \cdot \quad \cdot \\ &\quad\cdot \qquad \cdot \qquad \cdot && \cdot \quad \cdot \\ &\quad\cdot \qquad \cdot \qquad \cdot && \cdot \quad \cdot \\ &(\alpha P_{n-1,j}-\beta Q_{n-1,j})u_{n-2,j}-(1+2\alpha P_{n-1,j})u_{n-1,j}+f_{n-1}(u_{n-1,j}) = 0 \end{aligned} \tag{7.75}$$

Starting with j=1, on the first row of grid points, if one inserts the known initial and boundary values in (7.74), equations (7.75) result in a mildly nonlinear linear system of n−1 equations in the n−1 unknowns $u_{1,1}, u_{2,1}, u_{3,1}, \ldots, u_{n-1,1}$. A solution of this system, together with the boundary values, is called the numerical solution at time t_1. Once a numerical result for row 1 has been generated, set j=2, and solve a new mildly nonlinear system (7.75) of n−1 equations in the n−1 unknowns $u_{1,2}, u_{2,2}, u_{3,2}, \ldots, u_{n-1,2}$. Continue in the indicated fashion for each subsequent row j, j=3,4,…,m. The numerical method so

obtained is called the ***implicit central difference*** method.

In general, the mildly nonlinear system (7.75) may not have a solution, or may have more than one solution. Note, however, that the linear part of system (7.75) is tridiagonal with $-(1+2\alpha P_{ij})$ on the ***main*** diagonal, $(\alpha P_{ij}-\beta Q_{ij})$ on the ***subdiagonal***, and $(\alpha P_{ij}+\beta Q_{ij})$ on the ***superdiagonal***. Thus, if h satisfies the inequality $hM<2\nu$, then system (7.75) satisfies all the assumptions of Theorem 1.4, and in fact,

$$-(1+2\alpha P_{ij}) \le -(1+2\alpha\nu) < 0$$

$$(\alpha P_{ij}-\beta Q_{ij}) \ge (\alpha\nu-\beta M) > 0$$

$$(\alpha P_{ij}+\beta Q_{ij}) \ge (\alpha\nu-\beta M) > 0 ,$$

and

$$|\alpha P_{ij}-\beta Q_{ij}|+|\alpha P_{ij}+\beta Q_{ij}| = |2\alpha P_{ij}| < |1+2\alpha P_{ij}| = |-(1+2\alpha P_{ij})| .$$

Hence the numerical solution u_{ij}, $i=0,1,2,\ldots,n$, ***exists*** and is ***unique*** for each $j=0,1,2,\ldots,m$. Moreover, the solution of these systems can be found by the generalized Newton's method which, in this case, will converge for all ω in a subrange of $0<\omega<2$ and for all initial guesses. For efficiency, however, since the numerical solution at time level t_j, in general, is expected to be close to the solution at previous time level t_{j-1}, it is convenient to use $u_{i,j-1}$ as an initial guess for $u_{i,j}$, $i=1,2,\ldots,n-1$ in the Newtonian iteration formulas.

The implicit central difference method for linear initial-boundary problems can be derived from (7.73) as a particular case. In fact, for $F(x,t,u)=R(x,t)u+S(x,t)$, (7.73) becomes

$$(\alpha P_{ij}-\beta Q_{ij})u_{i-1,j}-(1-kR_{ij}+2\alpha P_{ij})u_{i,j}+(\alpha P_{ij}+\beta Q_{ij})u_{i+1,j}+u_{i,j-1}+kS_{ij} = 0 . \tag{7.76}$$

Equations (7.76), for each $j=1,2,\ldots,m$, constitute a linear tridiagonal system of $n-1$ equations in the $n-1$ unknowns $u_{1,j}, u_{2,j}, \ldots, u_{n-1,j}$. Each of these systems, when $Mh<2\nu$, is diagonally dominant, with negative elements on the main diagonal and positive ones on the subdiagonal and on the superdiagonal. Thus, by Theorem 1.2, the numerical solution ***exists*** and is ***unique*** on the entire $R_{n+1,m+1}$ set. It is interesting to note that the condition $Mh<2\nu$ can be extended to $Mh\le 2\nu$. However, the proof must then invoke Theorem 1.3 rather than Theorem 1.2. Moreover, when $Mh\le 2\nu$, the implicit finite difference method (7.76) also possesses the ***max-min*** property.

THEOREM 7.7. Let the initial-boundary problem (7.25)-(7.27) be defined on $0\le x\le a$, $0\le t\le T$, with $S(x,t)\equiv 0$, and with continuous $P(x,t)$, $Q(x,t)$ and $R(x,t)$. Assume that $P(x,t)\ge\nu>0$, $|Q(x,t)|\le M$ and $R(x,t)\le 0$ on $0\le x\le a$, $0\le t\le T$. If $Mh\le 2\nu$ then the numerical solution of (7.25)-(7.27), obtained with the implicit central difference method, possesses the max-min property.

PROOF. Let $u_{pj}=\max(u_{ij})$. If p=0, p=n, or j=0, the theorem is valid. Hence assume $j>0$ and $0<p<n$. Then, since $S_{ij}=0$, (7.76) yields

$$(7.77)\qquad (1-kR_{pj})u_{p,j} = u_{p,j-1}+(\alpha P_{pj}-\beta Q_{pj})u_{p-1,j}-2\alpha P_{pj}u_{p,j}+(\alpha P_{pj}+\beta Q_{pj})u_{p+1,j}$$

$$\le u_{p,j-1}+(\alpha P_{pj}-\beta Q_{pj})u_{p,j}-2\alpha P_{pj}u_{p,j}+(\alpha P_{pj}+\beta Q_{pj})u_{p,j} = u_{p,j-1},$$

from which follows, independently of the sign of u_{pj}, since $R_{pj}\le 0$,

$$(7.78)\qquad \max_i(u_{i,j}) \le \max[0,u_{0,j},\max_p(u_{p,j-1}),u_{n,j}] .$$

Now, by repeating the above argument, one finds

$$\max_i(u_{i,j}) \le \max[0,u_{0,j},\max_{0<p<n}(u_{p,j-1}),u_{n,j}]$$

$$\le \max[0,u_{0,j},u_{0,j-1},\max_{0<p<n}(u_{p,j-2}),u_{n,j-1},u_{n,j}]$$

$$\le \max[0,u_{0,j},u_{0,j-1},u_{0,j-2},\max_{0<p<n}(u_{p,j-3}),u_{n,j-2},u_{n,j-1},u_{n,j}]$$

$$\vdots$$

$$\le \max[0,\max_{1\le q\le j}(u_{0,q}),\max_{0<p<n}(u_{p,0}),\max_{1\le q\le j}(u_{n,q})]$$

which implies (7.42). Inequality (7.41) can be proved in an entirely similar way.

If $R_{ij}=0$, i=0,1,2,...,n, j=0,1,2,...,m, inequality (7.77) implies

$$(7.79)\qquad \max_i(u_{i,j}) \le \max[u_{0,j},\max_p(u_{p,j-1}),u_{n,j}] ,$$

which implies, as above,

$$\max_i(u_{i,j}) \le \max[u_{0,j}, \max_{0<p<n}(u_{p,j-1}), u_{n,j}]$$

$$\le \max[u_{0,j}, u_{0,j-1}, \max_{0<p<n}(u_{p,j-2}), u_{n,j-1}, u_{n,j}]$$

$$\le \max[u_{0,j}, u_{0,j-1}, u_{0,j-2}, \max_{0<p<n}(u_{p,j-3}), u_{n,j-2}, u_{n,j-1}, u_{n,j}]$$

$$\vdots$$

$$\le \max[\max_{1\le q\le j}(u_{0,q}), \max_{0<p<n}(u_{p,0}), \max_{1\le q\le j}(u_{n,q})] .$$

Hence, (7.44) is established. Inequality (7.43) can be proved in a similar fashion.

Note that under the hypotheses of Theorem 7.7, if f(x), $g_1(t)$ and $g_2(t)$ are bounded, the numerical solution of initial-boundary problem (7.25)-(7.27), obtained with the implicit central difference method remains bounded for any fixed T. Hence, inequality (7.46) can also be regarded as a ***sufficient stability condition*** for the method.

EXAMPLE. Let us reconsider the initial-boundary problem (7.35)-(7.37), namely

$$u_t = u_{xx} + (x-2)u_x - 3u$$

$$u(x,0) = x^2 - 4x + 5 , \qquad 0 \le x \le 4$$

$$u(0,t) = u(4,t) = 5e^{-t}, \qquad t > 0 .$$

Let T=1, n=4, and, in order to take advantage of the less restrictive condition imposed by the implicit method, fix m=2, so that h=1 and k=0.5. Now, at the initial time t_0=0, the given initial condition implies

$$u_{00} = 5 , \qquad u_{10} = 2 , \qquad u_{20} = 1 , \qquad u_{30} = 2 , \qquad u_{40} = 5 ,$$

while, at the boundary grid points the given boundary conditions, to three decimal places, yield

$$u_{0,1} = 3.033 \qquad u_{4,1} = 3.033$$

$$u_{0,2} = 1.839 \qquad u_{4,2} = 1.839 .$$

At the interior grid points, since $\alpha=1/2$ and $\beta=1/4$, equation (7.76) becomes

$$\left[\frac{1}{2}-\frac{1}{4}(x_i-2)\right]u_{i-1,j}-\left[1+\frac{3}{2}+1\right]u_{i,j}+\left[\frac{1}{2}+\frac{1}{4}(x_i-2)\right]u_{i+1,j}+u_{i,j-1}=0\,.$$

Now, for j=1 and i=1,2,3, one has

$$\frac{3}{4}u_{01}-\frac{7}{2}u_{11}+\frac{1}{4}u_{21}+u_{10}=0$$

$$\frac{1}{2}u_{11}-\frac{7}{2}u_{21}+\frac{1}{2}u_{31}+u_{20}=0$$

$$\frac{1}{4}u_{21}-\frac{7}{2}u_{31}+\frac{3}{4}u_{41}+u_{30}=0$$

which, by using the known initial and boundary values yields

$$\begin{aligned} -14u_{11}+u_{21} &= -17.099 \\ 2u_{11}-14u_{21}+2u_{31} &= -4 \\ u_{21}-14u_{31} &= -17.099\,, \end{aligned}$$

the unique solution of which, to three decimal places, is

$$u_{11}=1.268, \qquad u_{21}=0.648\,, \qquad u_{31}=1.268\,.$$

Next, for j=2, by using the above numerical results, and the required boundary conditions, one obtains

$$\begin{aligned} -14u_{12}+u_{22} &= -10.589 \\ 2u_{12}-14u_{22}+2u_{32} &= -2.592 \\ u_{22}-14u_{32} &= -10.589\,, \end{aligned}$$

which, to three decimal places, yields the following results at $t_2=1$:

$$u_{12} = 0.786\,, \qquad u_{22} = 0.410\,, \qquad u_{32} = 0.786\,,$$

and the example is complete.

7.8 IMPLICIT UPWIND METHOD

If one wishes to eliminate also the restriction (7.42) on the space step h, one can use an implicit upwind difference approximation to discretize (7.60). Specifically, the terms u_{xx}, u_x and u_t are approximated by (7.32), (7.55) and (7.72), respectively. The resulting approximation for (7.60) is then given by

$$\frac{u_{i,j}-u_{i,j-1}}{k} = P_{ij}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} + Q_{ij}\frac{u_{i+1,j}-u_{i,j}}{h} + F(x_i,t_j,u_{i,j})\,, \quad \text{if } Q_{ij} \geq 0\,,$$

$$\frac{u_{i,j}-u_{i,j-1}}{k} = P_{ij}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} + Q_{ij}\frac{u_{i,j}-u_{i-1,j}}{h} + F(x_i,t_j,u_{i,j})\,, \quad \text{if } Q_{ij} < 0\,,$$

or, equivalently, by

$$(7.80) \qquad [\alpha P_{ij}+\beta(|Q_{ij}|-Q_{ij})]u_{i-1,j}-[1+2(\alpha P_{ij}+\beta|Q_{ij}|)]u_{i,j}$$
$$+[\alpha P_{ij}+\beta(|Q_{ij}|+Q_{ij})]u_{i+1,j}+kF(x_i,t_j,u_{i,j})+u_{i,j-1} = 0\,.$$

The algorithm to be used to generate the numerical solution at each time level t_j, $j=1,2,\ldots,m$, is the same as described above, but with (7.80) replacing (7.73). Specifically, starting with $j=1$, if one writes (7.80) consecutively for fixed j and $i=1,2,\ldots,n-1$, ***and*** one also inserts the known numerical solution at time level t_{j-1} and the known boundary values, there results a mildly nonlinear linear system of $n-1$ equations in the $n-1$ unknowns $u_{1,j}$, $u_{2,j}$, …, $u_{n-1,j}$. This system satisfies all the assumptions of Theorem 1.4 for ***any*** given h ***and*** k. Thus, the numerical solution u_{ij} ***exists*** and is ***unique*** on the entire $R_{n+1,m+1}$ set. The method so obtained is called the ***implicit upwind*** method.

The implicit upwind method for ***linear*** initial-boundary problems can be derived directly from (7.80). By setting $F(x,t,u)=R(x,t)u+S(x,t)$, (7.80) becomes

(7.81) $$[\alpha P_{ij}+\beta(|Q_{ij}|-Q_{ij})]u_{i-1,j}-[1-kR_{ij}+2(\alpha P_{ij}+\beta|Q_{ij}|)]u_{i,j}$$

$$+[\alpha P_{ij}+\beta(|Q_{ij}|+Q_{ij})]u_{i+1,j}+u_{i,j-1}+kS_{ij}=0\ .$$

Equations (7.81), for each j=1,2,...,m, constitute a linear tridiagonal system of n–1 equations in the n–1 unknowns $u_{1,j}, u_{2,j}, \ldots, u_{n-1,j}$. For ***all*** h and k, these systems are diagonally dominant, with negative elements on the main diagonal and positive ones on the subdiagonal and the superdiagonal. Thus, by Theorem 1.2, the numerical solution *exists* and is ***unique*** on the entire $R_{n+1,m+1}$ set.

Although the implicit upwind method is relatively less accurate, it yields a numerical solution which possesses the discrete max-min property for any choice of h and k. This is stated in the next theorem.

THEOREM 7.8. Let the initial-boundary problem (7.25)-(7.27) be defined on $0\le x\le a$, $0\le t\le T$, with $S(x,t)\equiv 0$, and with continuous $P(x,t)$, $Q(x,t)$ and $R(x,t)$. Assume that $R(x,t)\le 0$, $|Q(x,t)|\le M$, $P(x,t)\ge \nu>0$ on $0\le x\le a$, $0\le t\le T$. Then the numerical solution of (7.25)-(7.27), obtained with the implicit upwind method, possesses the max-min property.

The proof of this theorem is entirely similar to that of Theorem 7.7.

Note that under the hypotheses of Theorem 7.8, if f(x), $g_1(t)$ and $g_2(t)$ are bounded, the numerical solution of initial-boundary problem (7.25)-(7.27), obtained with the implicit upwind method remains bounded for any T. Hence, since no restrictions are imposed on h and k, this method is ***unconditionally*** stable.

EXAMPLE. Let us consider, once again, the initial-boundary problem (7.35)-(7.37), that is,

$$u_t = u_{xx}+(x-2)u_x-3u$$

$$u(x,0) = x^2-4x+5\ , \qquad 0\le x\le 4$$

$$u(0,t) = u(4,t) = 5e^{-t}, \qquad t>0\ .$$

For the implicit upwind method, set T=1, n=4 and m=2, so that h=1 and k=0.5, α=1/2, β=1/4. The given initial and boundary conditions, to three decimal places, are

$$u_{02} = 1.839 \qquad\qquad\qquad\qquad\qquad\qquad u_{42} = 1.839$$

$$u_{01} = 3.033 \qquad\qquad\qquad\qquad\qquad\qquad u_{41} = 3.033$$

$$u_{00} = 5.000 \qquad u_{10} = 2.000 \qquad u_{20} = 1.000 \qquad u_{30} = 2.000 \qquad u_{40} = 5.000\ .$$

Thus, for j=1 and i=1,2,3, and by use of the required initial and boundary conditions, equation (7.81) implies

$$\begin{aligned} -8u_{11} + u_{21} &= -10.066 \\ u_{11} - 7u_{21} + u_{31} &= -2 \\ u_{21} - 8u_{31} &= -10.066 , \end{aligned}$$

the unique solution of which, to three decimal places, is

$$u_{11} = 1.342 , \qquad u_{21} = 0.669 , \qquad u_{31} = 1.342 .$$

Next, for j=2, and i=1,2,3, equation (7.81) implies

$$\begin{aligned} -8u_{12} + u_{22} &= -6.362 \\ u_{12} - 7u_{22} + u_{32} &= -1.338 \\ u_{22} - 8u_{32} &= -6.362 , \end{aligned}$$

which, at t_2=1, yields

$$u_{12} = 0.849 , \qquad u_{22} = 0.434 , \qquad u_{32} = 0.849 .$$

Table 7.2 Implicit methods.

exact	$u_{1,2}$=0.736	$u_{2,2}$=0.368	$u_{3,2}$=0.736
central	$u_{1,2}$=0.786	$u_{2,2}$=0.410	$u_{3,2}$=0.786
upwind	$u_{1,2}$=0.849	$u_{2,2}$=0.434	$u_{3,2}$=0.849

Table 7.3 Grid size restrictions.

	Central difference	Upwind
Explicit	$hM \le 2\nu$, $k \le \frac{h^2}{Nh^2+2V}$	$k \le \frac{h^2}{Nh^2+Mh+2V}$
Implicit	$hM \le 2\nu$	*No restrictions*

For comparison, Table 7.2 shows the values of the exact solution at time T=1, the results from the implicit central difference method from the example in the previous section and the results just generated. As expected, both numerical approximations possess the discrete max-min property, and better accuracy is obtained when central differences are used. Finally, a comparison between Table 7.1 and Table 7.2 indicates that the numerical results listed in Table 7.1, obtained for the very same initial-boundary problem using the explicit upwind and the central difference method, have higher accuracy because a smaller time step has been used.

In summary, we observe that the simplest method for a general initial-boundary problem is the explicit central difference method. This method, however, as indicated in Table 7.3, requires limitations on h an k. If the limitation on h is too restrictive, one can use the explicit upwind method. When the limitation on k is too restrictive, one can use an implicit central difference method. If both limitations on h and k are too restrictive, then the implicit upwind method has to be used. Of course, upwind methods are only first order accurate in space, hence their use should be limited only to the cases when the limitation on h is relatively severe. On the other hand, implicit methods require one to solve a system of n–1 equations at each time step, so that the limitation on k is eliminated at the price of greater computational complexity. Note, finally, that while explicit methods apply to ***both*** initial value problems and initial-boundary problems, the implicit methods apply ***only*** to initial-boundary problems.

7.9 THE CRANK-NICOLSON METHOD

Since central difference methods have relatively high accuracy in space, the next problem is to improve on the time accuracy. For this purpose, note that the use of symmetry in the construction of difference equations can lead to better accuracy in the sense that the truncation error is of a higher order of magnitude than when symmetry is not used. Thus, for example, the error in approximation (3.30) is O(h), while that of (3.37), which uses symmetry, is $O(h^2)$. With this observation as an intuitive guide, let us modify the previous methods to yield greater time accuracy in the following simple way. In place of the point patterns shown in Figures 7.4 and 7.6, consider the expanded point pattern shown in Figure 7.7. The center of symmetry of the six points shown there is the point $(x_i,t_{j-1/2})=[ih,(j-1/2)k]$ which is ***not*** a grid point. If one wishes, now, to develop formulas

ymmetrically about $(x_i,t_{j-1/2})$, then note first that

$$u_t(x_i,t_{j-1/2}) \approx \frac{u_{i,j}-u_{i,j-1}}{k}, \tag{7.82}$$

oes use points located symmetrically about $(x_i,t_{j-1/2})$. Further, since

$$u_{xx}(x_i,t_j) \approx \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}$$

nd

$$u_{xx}(x_i,t_{j-1}) \approx \frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{h^2},$$

is reasonable to set

$$u_{xx}(x_i,t_{j-1/2}) \approx \frac{1}{2}\left[u_{xx}(x_i,t_j)+u_{xx}(x_i,t_{j-1})\right],$$

that

$$u_{xx}(x_i,t_{j-1/2}) \approx \frac{1}{2}\left[\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}+\frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{h^2}\right]. \tag{7.83}$$

milarly, when centered finite differences are used to approximate u_x, we set

$$u_x(x_i,t_{j-1/2}) \approx \frac{1}{2}\left[\frac{u_{i+1,j}-u_{i-1,j}}{2h}+\frac{u_{i+1,j-1}-u_{i-1,j-1}}{2h}\right]. \tag{7.84}$$

sing (7.82)-(7.84) in (7.60) yields

$$\frac{u_{i,j}-u_{i,j-1}}{k} = P_{i,j-1/2}\left[\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{2h^2}+\frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{2h^2}\right] \tag{7.85}$$

$$+ Q_{i,j-1/2}\left[\frac{u_{i+1,j}-u_{i-1,j}}{4h}+\frac{u_{i+1,j-1}-u_{i-1,j-1}}{4h}\right]+F\left(x_i,t_{j-1/2},\frac{u_{i,j}+u_{i,j-1}}{2}\right),$$

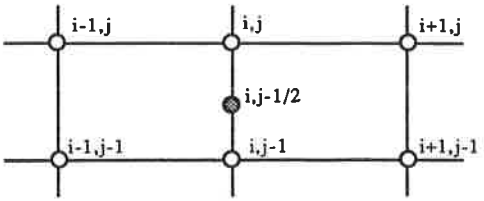

Figure 7.7

which, in each of its parts, is symmetrical about $(x_i,t_{j-1/2})$. Formula (7.85) is called the ***Crank-Nicolson*** formula, or the central difference Crank-Nicolson formula. It has greater accuracy than both the fully explicit and the fully implicit formulas. When this formula is used in place of (7.73) in the implicit method, the resulting method is called the Crank-Nicolson method, or the central difference Crank-Nicolson method. It does lead, at each time step, to a mildly nonlinear system which, for $hM<2v$, satisfies all the assumptions of Theorem 1.4. Moreover, for linear initial-boundary problems, when $hM\le 2v$, the Crank-Nicolson method at each time step yields a system which satisfies all the assumptions of Theorem 1.3. Thus, it has a ***unique*** solution for all k, but, when $S(x,t)\equiv 0$, in order to be assured that this method possesses the ***max-min*** property, the time step k must satisfy the condition $k\le 2h^2/(Nh^2+2V)$.

EXAMPLE. For completness, calculating to three decimal places, let us solve the initial boundary problem (7.35)-(7.37) with a Crank-Nicolson method. In this case, for T=1, n=4 and m=10, one has h=1, k=0.1, and (7.85) reduces to

$$\frac{u_{i,j}-u_{i,j-1}}{0.1} = \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{2} + \frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{2}$$

$$+ (x_i-2)\left[\frac{u_{i+1,j}-u_{i-1,j}}{4} + \frac{u_{i+1,j-1}-u_{i-1,j-1}}{4}\right] - 3\frac{u_{i,j}+u_{i,j-1}}{2},$$

or, equivalently,

$$0.1(4-x_i)u_{i-1,j}-5u_{i,j}+0.1x_iu_{i+1,j} = -0.1(4-x_i)u_{i-1,j-1}-3u_{i,j-1}-0.1x_iu_{i+1,j-1},$$

which, by using the initial conditions (7.38) and the boundary conditions (7.39), for j=1, and i=1,2,3 yields the system

$$\begin{aligned} -5u_{11}+0.1u_{21} &= -8.957 \\ 0.2u_{11} - 5u_{21}+0.2u_{31} &= -3.800 \\ 0.1u_{21} - 5u_{31} &= -8.957 . \end{aligned}$$

The solution of this system is $u_{11}=1.810$, $u_{21}=0.905$, $u_{31}=1.810$. Next, by using the boundary conditions (3.39) and the results just obtained, one has for $j=2$

$$\begin{aligned} -5u_{12}+0.1u_{22} &= -8.104 \\ 0.2u_{12} - 5u_{22}+0.2u_{32} &= -3.438 \\ 0.1u_{22} - 5u_{32} &= -8.104 \end{aligned}$$

whose solution is $u_{12}=1.637$, $u_{22}=0.819$, $u_{32}=1.637$. Continuing in the indicated fashion for $j=3,4,\ldots,10$, one finds at $t_{10}=1$

$$u_{1,10} = 0.736 , \qquad u_{2,10} = 0.368 , \qquad u_{3,10} = 0.736 ,$$

which, by comparison with the results shown in Tables 7.1 and 7.2, indicates clearly the superiority in accuracy of the Crank-Nicolson method.

As usual, if one is willing to sacrifice a degree of accuracy in space, the restriction on h can be eliminated by using, in place of (7.85), an ***upwind*** Crank-Nicolson formula which is given by

$$\begin{aligned} (7.86) \qquad \frac{u_{i,j}-u_{i,j-1}}{k} &= P_{i,j-1/2}\left[\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{2h^2}+\frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{2h^2}\right] \\ &+\frac{(|Q_{i,j-1/2}|+Q_{i,j-1/2})u_{i+1,j}-2|Q_{i,j-1/2}|u_{i,j}+(|Q_{i,j-1/2}|-Q_{i,j-1/2})u_{i-1,j}}{4h} \\ &+\frac{(|Q_{i,j-1/2}|+Q_{i,j-1/2})u_{i+1,j-1}-2|Q_{i,j-1/2}|u_{i,j-1}+(|Q_{i,j-1/2}|-Q_{i,j-1/2})u_{i-1,j-1}}{4h} \\ &+ F(x_i,t_{j-1/2},\frac{u_{i,j}+u_{i,j-1}}{2}) . \end{aligned}$$

The mildly nonlinear system that one obtains at each time step from (7.86) satisfies the conditions of Theorem 1.4 for any h and k. Thus the numerical solution always exists, is unique, and also possesses the max-min property for linear equations with $S(x,t)\equiv 0$ provided k satisfies the following inequality:

(7.87) $$k \le \frac{2h^2}{Nh^2+Mh+2V}.$$

For programming purposes, the above results are often unified as follows. The fully explicit, fully implicit and Crank-Nicolson space centered formulas can be written compactly as

(7.88) $$\frac{u_{i,j}-u_{i,j-1}}{k} = P_{i,j-\theta}\left[(1-\theta)\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}+\theta\frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{h^2}\right]$$

$$+ Q_{i,j-\theta}\left[(1-\theta)\frac{u_{i+1,j}-u_{i-1,j}}{2h}+\theta\frac{u_{i+1,j-1}-u_{i-1,j-1}}{2h}\right]+F(x_i,t_{j-\theta},(1-\theta)u_{i,j}+\theta u_{i,j-1}),$$

which, for $\theta=1$, is equivalent to the explicit formula (7.62), for $\theta=0$ yields the implicit formula (7.73), and for $\theta=1/2$ yields the Crank-Nicolson formula (7.85). In general, a method that results for $0\le\theta\le1$ is called a θ-method. The general conditions for a θ-method to possess the max-min property for linear equations with $S(x,t)\equiv0$ are as follows:

(7.89) $$Mh \le 2v$$

(7.90) $$\theta k \le \frac{h^2}{Nh^2+2V}.$$

The fully explicit, fully implicit and Crank-Nicolson upwind formulas can be written, more compactly, as

(7.91) $$\frac{u_{i,j}-u_{i,j-1}}{k} = P_{i,j-\theta}\left[(1-\theta)\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}+\theta\frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{h^2}\right]$$

$$+ (1-\theta)\frac{(|Q_{i,j-\theta}|+Q_{i,j-\theta})u_{i+1,j}-2|Q_{i,j-\theta}|u_{i,j}+(|Q_{i,j-\theta}|-Q_{i,j-\theta})u_{i-1,j}}{2h}$$

$$+ \theta\frac{(|Q_{i,j-\theta}|+Q_{i,j-\theta})u_{i+1,j-1}-2|Q_{i,j-\theta}|u_{i,j-1}+(|Q_{i,j-\theta}|-Q_{i,j-\theta})u_{i-1,j-1}}{2h}$$

$$+ F(x_i,t_{j-\theta},(1-\theta)u_{i,j}+\theta u_{i,j-1}),$$

which, for $\theta=1$, is equivalent to the explicit upwind formula (7.63), for $\theta=0$ yields the implicit upwind formula (7.80), and for $\theta=1/2$ yields the upwind Crank-Nicolson formula (7.86). In general, an upwind θ-method that one obtains from (7.91) for $0\le\theta\le1$ yields a unique solution which possesses the max-min property for linear equations with $S(x,t)\equiv0$ under the following condition:

$$\theta k \le \frac{h^2}{Nh^2+Mh+2V}. \tag{7.92}$$

It should be emphasized, finally, that the conditions (7.89), (7.90), (7.92) yield numerical solutions in every case which are unique and possess the max-min property for appropriate linear equations, just as their analytical counterparts do. The ***stability*** of the methods can be established, usually, with less restrictive assumptions [Casulli (1987)]. Thus, for example, the Crank-Nicolson method for the heat equation is stable for all h and k, but need not yield the physically significant max-min property.

EXERCISES

Basic Exercises

1. Given the initial-boundary problem for $u_t=u_{xx}$ with $a=1$, $g_1(t)=e^{-t}$, $g_2(t)=2e^{-t}$, and $f(x)=x^2+1$, find the numerical solution by the explicit central difference method up to t_{10} for each of the following choices:

   (a) $h = 1/4$, $k = 1/10$
   (b) $h = 1/4$, $k = 1/20$
   (c) $h = 1/4$, $k = 1/40$
   (d) $h = 1/4$, $k = 1/80$.

   Which of the above calculations are stable? Which will lead eventually to overflow? Which possess the max-min property?

2. Given the initial-boundary problem for $u_t=u_{xx}+xu_x-3u$ with $a=1$, $g_1(t)=e^{-t}$, $g_2(t)=2e^{-t}$, and $f(x)=x^2+1$, find the numerical solution by the explicit central difference method at $T=1$ for each of the following choices:

(a) $h = 1/4$, $k = 1/10$

(b) $h = 1/4$, $k = 1/20$

(c) $h = 1/4$, $k = 1/40$

(d) $h = 1/4$ $k = 1/80$.

Which of the above calculations are stable? Which will lead eventually to overflow? Which possess the max-min property? Compare your results with the exact solution $u=(x^2+1)e^{-t}$.

3. Repeat Exercise 2 by using the explicit upwind method. Then compare the results of the two methods with the exact solution.

4. Complete the proof of Theorem 7.2.

5. By using an explicit method, find the numerical solution at T=3 for the initial-boundary problem defined by

$$u_t = u_{xx} - \arctan(u), \quad 0<x<1, \quad t>0$$
$$f(x) = x, \quad 0\le x\le 1,$$
$$g_1(t) = 0, \quad g_2(t) = e^{-t}, \quad t\ge 0.$$

6. Give a proof to Theorem 7.5.

7. Repeat Exercise 1 by using the implicit method with h=1/5 and k=1.

8. Given the initial-boundary problem for $u_t=u_{xx}+xu_x-3u$ with $a=1$, $g_1(t)=e^{-t}$, $g_2(t)=2e^{-t}$, and $f(x)=x^2+1$, find the numerical solution by the implicit central difference method at T=1 for each of the following choices:

(a) $h = 1/4$, $k = 1/10$

(b) $h = 1/4$, $k = 1/5$

(c) $h = 1/4$, $k = 1/2$

(d) $h = 1/4$, $k = 1$.

Which of the above calculations are stable? Which possess the max-min property? Compare your results with the exact solution $u=(x^2+1)e^{-t}$.

9. Give an example of an initial-boundary problem for which, when $Mh>2\nu$, the implicit central difference method, when applied to linear equations with $S(x,t)\equiv 0$, does not possess the max-min property.

10. Repeat Exercise 8 but use the implicit upwind method. Then compare the numerical results with those of Exercise 8.

11. Give a proof to Theorem 7.8.

12. Repeat Exercise 1 but use the Crank-Nicolson method with $h=1/5$ and $k=1$.

13. Using the Crank-Nicolson method, repeat Exercise 8 and compare the numerical results obtained.

14. Prove that when $Mh\leq 2\nu$ and $k\leq h^2/(Nh^2+V)$, the central difference Crank-Nicolson method possesses the max-min property when applied to linear equations with $S(x,t)\equiv 0$.

15. Give an example of an initial-boundary problem for which, when $Mh>2\nu$, the central difference Crank-Nicolson method does not possess the max-min property.

16. Give an example of an initial-boundary problem for which, when $k>h^2/(Nh^2+V)$, the central difference Crank-Nicolson method does not possess the max-min property.

17. Prove that when $k\leq h^2/(Nh^2+Mh+V)$, the upwind Crank-Nicolson method, when applied to linear equations with $S(x,t)\equiv 0$, possesses the max-min property.

18. Give an example of an initial-boundary problem for which, when $k>h^2/(Nh^2+Mh+V)$, the upwind Crank-Nicolson method, when applied to linear equations with $S(x,t)\equiv 0$, does not possess the max-min property.

Supplementary Exercises

19. Extend the explicit methods developed for initial-boundary problems to initial value problems.

20. Consider the initial and boundary conditions

$$u(x,0) = x, \qquad 0 \le x \le 1$$
$$u(0,t) = 0, \qquad t \ge 0$$
$$u(1,t) = e^{-t}, \qquad t \ge 0.$$

Find an approximate solution of the associated initial-boundary problem for

$$u_t = u_{xx} - xu_x$$

at (0.5,5) and compare your result with the exact solution $u=xe^{-t}$.

21. Repeat Exercise 20 but replace the parabolic equation by

$$u_t = u_{xx} - u - u^3 - x^3 e^{-3t}.$$

22. Consider the initial and boundary conditions

$$u(x,0) = 0, \qquad 0 \le x \le \pi$$
$$u(0,t) = 0, \qquad t \ge 0$$
$$u(\pi,t) = 0, \qquad t \ge 0.$$

Find an approximate solution of the associated initial-boundary problem for

$$u_t = u_{xx} + \sin(x)[\sin(t) + \cos(t)]$$

when $T=4\pi$. Compare your result with the exact solution

$$u = \sin(x)\sin(t).$$