Chapter 19

Numerical Methods for Differential Equations

From Chap. 1 we know that an ODE of the first order is of the form $F(x, y, y') = 0$ and can often be written in the explicit form $y' = f(x, y)$. An **initial value problem** for this equation is of the form

$$\textbf{(1)} \qquad y' = f(x, y), \qquad y(x_0) = y_0$$

where x_0 and y_0 are given and we assume that the problem has a unique solution on some open interval $a < x < b$ containing x_0.

In this section we shall discuss methods of computing approximate numeric values of the solution $y(x)$ of (1) at the equidistant points on the x-axis

$$x_1 = x_0 + h, \qquad x_2 = x_0 + 2h, \qquad x_3 = x_0 + 3h, \qquad \cdots$$

where the **step size** h is a fixed number, for instance, 0.2 or 0.1 or 0.01, whose choice we discuss later in this section. Those methods are **step-by-step methods,** using the same formula in each step. Such formulas are suggested by the Taylor series

$$(2) \qquad y(x + h) = y(x) + hy'(x) + \frac{h^2}{2} y''(x) + \cdots .$$

For a small h the higher powers h^2, h^3, $\cdots$ are very small. This suggests the crude approximation

$$\begin{aligned} y(x + h) &\approx y(x) + hy'(x) \\ &= y(x) + hf(x, y) \end{aligned}$$

(with the second line obtained from the given ODE) and the following iteration process. In the first step we compute

$$y_1 = y_0 + hf(x_0, y_0)$$

which approximates $y(x_1) = y(x_0 + h)$. In the second step we compute

$$y_2 = y_1 + hf(x_1, y_1)$$

which approximates $y(x_2) = y(x_0 + 2h)$, etc., and in general

$$\textbf{(3)} \qquad y_{n+1} = y_n + hf(x_n, y_n) \qquad (n = 0, 1, \cdots).$$

This is called the **Euler method** or the **Euler–Cauchy method.** Geometrically it is an approximation of the curve of $y(x)$ by a polygon whose first side is tangent to this curve at x_0 (See Fig. 417)

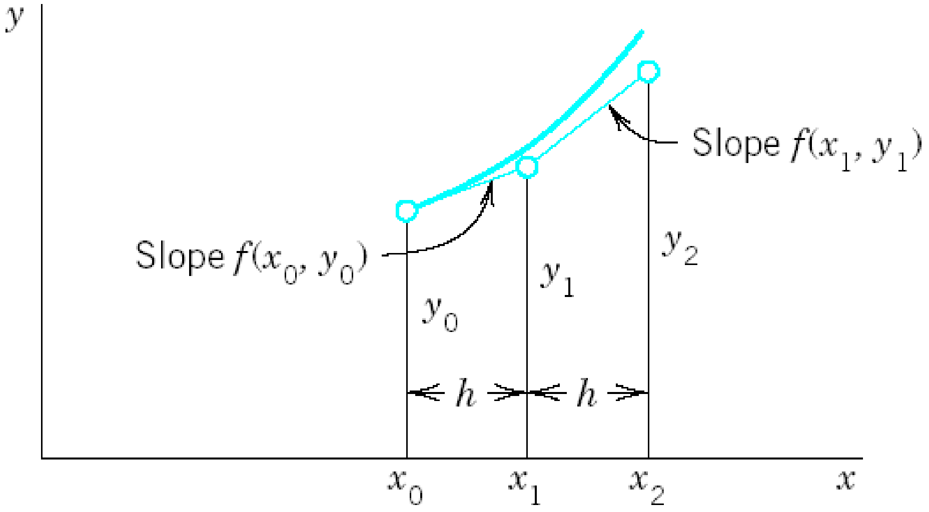


Fig. 417. Euler method

This crude method is hardly ever used in practice, but since it is simple, it nicely explains the principle of methods based on the Taylor series.

Taylor's formula with remainder has the form

$$y(x + h) = y(x) + hy'(x) + \tfrac{1}{2}h^2y''(\xi)$$

(where $x \leqq \xi \leqq x + h$). It shows that in the Euler method the *truncation error in each step* or **local truncation error** is proportional to h^2, written $O(h^2)$, where O suggests *order* (see also Sec. 20.1). Now over a fixed x-interval in which we want to solve an ODE the number of steps is proportional to $1/h$. Hence the *total error* or **global error** is proportional to $h^2(1/h) = h^1$. For this reason, the Euler method is called a **first-order method.** In addition, there are **roundoff errors** in this and other methods, which may affect the accuracy of the values $y_1, y_2, \cdots$ more and more as n increases, as we shall see.

Table 19.1 **Euler Method Applied to (4) in Example 1 and Error**

n	x_n	y_n	$0.2(x_n + y_n)$	Exact Values	Error ϵ_n
0	0.0	0.000	0.000	0.000	0.000
1	0.2	0.000	0.040	0.021	0.021
2	0.4	0.040	0.088	0.092	0.052
3	0.6	0.128	0.146	0.222	0.094
4	0.8	0.274	0.215	0.426	0.152
5	1.0	0.489		0.718	0.229

EXAMPLE 1

Euler Method

Apply the Euler method to the following initial value problem, choosing $h = 0.2$ and computing $y_1, \cdots, y_5$:

$$y' = x + y, \qquad y(0) = 0. \tag{4}$$

Solution. Here $f(x, y) = x + y$; hence $f(x_n, y_n) = x_n + y_n$, and we see that (3) becomes

$$y_{n+1} = y_n + 0.2(x_n + y_n).$$

Table 19.1 shows the computations, the values of the exact solution

$$y(x) = e^x - x - 1$$

obtained from (4) in Sec. 1.5, and the error. In practice the exact solution is unknown, but an indication of the accuracy of the values can be obtained by applying the Euler method once more with step $2h = 0.4$, letting $y_n{}^*$ denote the approximation now obtained, and comparing corresponding approximations. This computation is:

x_n	$y_n{}^*$	$0.4(x_n + y_n)$	y_n in Table 19.1	Difference $y_n - y_n{}^*$
0.0	0.000	0.000	0.000	0.000
0.4	0.000	0.160	0.040	0.040
0.8	0.160		0.274	0.114

Let ϵ_n and $\epsilon_n{}^*$ be the errors of the computations with h and $2h$, respectively. Since the error is of order h^2, in a switch from h to $2h$ it is multiplied by $2^2 = 4$, but since we need only half as many steps as before, it will be multiplied only by $4/2 = 2$. Hence $\epsilon_n{}^* \approx 2\epsilon_n$ so that the difference is $\epsilon_n{}^* - \epsilon_n \approx 2\epsilon_n - \epsilon_n = \epsilon_n$. Now $y = y_n + \epsilon_n = y_n{}^* + \epsilon_n{}^*$ by the definition of error; hence $\epsilon_n{}^* - \epsilon_n = y_n - y_n{}^*$ indicates ϵ_n qualitatively. In our computations, $y_2 - y_2{}^* = 0.04 - 0 = 0.04$ (actual error 0.052, see Table 19.1) and $y_4 - y_4{}^* = 0.274 - 0.160 = 0.114$ (actually 0.152).

In the **improved Euler method** or **improved Euler–Cauchy method** (sometimes also called **Heun method**), in each step we compute first the auxiliary value

(7a) $$y^*_{n+1} = y_n + hf(x_n, y_n)$$

and then the new value

(7b) $$y_{n+1} = y_n + \tfrac{1}{2}h\,[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})].$$

This method has a simple geometric interpretation. In fact, we may say that in the interval from x_n to $x_n + \frac{1}{2}h$ we approximate the solution y by the straight line through (x_n, y_n) with slope $f(x_n, y_n)$, and then we continue along the straight line with slope $f(x_{n+1}, y^*_{n+1})$ until x reaches x_{n+1}.

The improved Euler–Cauchy method is a **predictor–corrector method,** because in each step we first ***predict*** a value by (7a) and then ***correct*** it by (7b)

Error of the Improved Euler Method. *The local error is of order h^3 and the global error of order h^2, so that the method is a* **second-order method.**

In algorithmic form, using the notations $k_1 = hf(x_n, y_n)$ ir (7a) and $k_2 = hf(x_{n+1}, y^*_{n+1})$ in (7b) we can write this method as shown in Table 19.2

Table 19.2 Improved Euler Method (Heun's Method)

ALGORITHM EULER (f, x_0, y_0, h, N)

This algorithm computes the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ at equidistant points $x_1 = x_0 + h$, $x_2 = x_0 + 2h$, $\cdots$, $x_N = x_0 + Nh$; here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec 1.9 .

INPUT: Initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, \cdots, N - 1$

```
    For $n = 0, 1, \cdots, N - 1$ do:
        $x_{n+1} = x_n + h$
        $k_1 = hf(x_n, y_n)$
        $k_2 = hf(x_{n+1}, y_n + k_1)$
        $y_{n+1} = y_n + \frac{1}{2}(k_1 + k_2)$
        OUTPUT $x_{n+1}, y_{n+1}$
    End
    Stop
End EULER
```

Example 2

Improved Euler Method

Apply the improved Euler method to the initial value problem (4), choosing $h = 0.2$, as before.

Solution. For the present problem we have in Table 19.2

$$k_1 = 0.2(x_n + y_n)$$

$$k_2 = 0.2(x_n + 0.2 + y_n + 0.2(x_n + y_n))$$

$$y_{n+1} = y_n + \frac{0.2}{2}(2.2x_n + 2.2y_n + 0.2) = y_n + 0.22(x_n + y_n) + 0.02.$$

Table 19.3 shows that our present results are more accurate than those in Example 1; see also Table 19.6

Table 19.3 **Improved Euler Method Applied to (4) and Error**

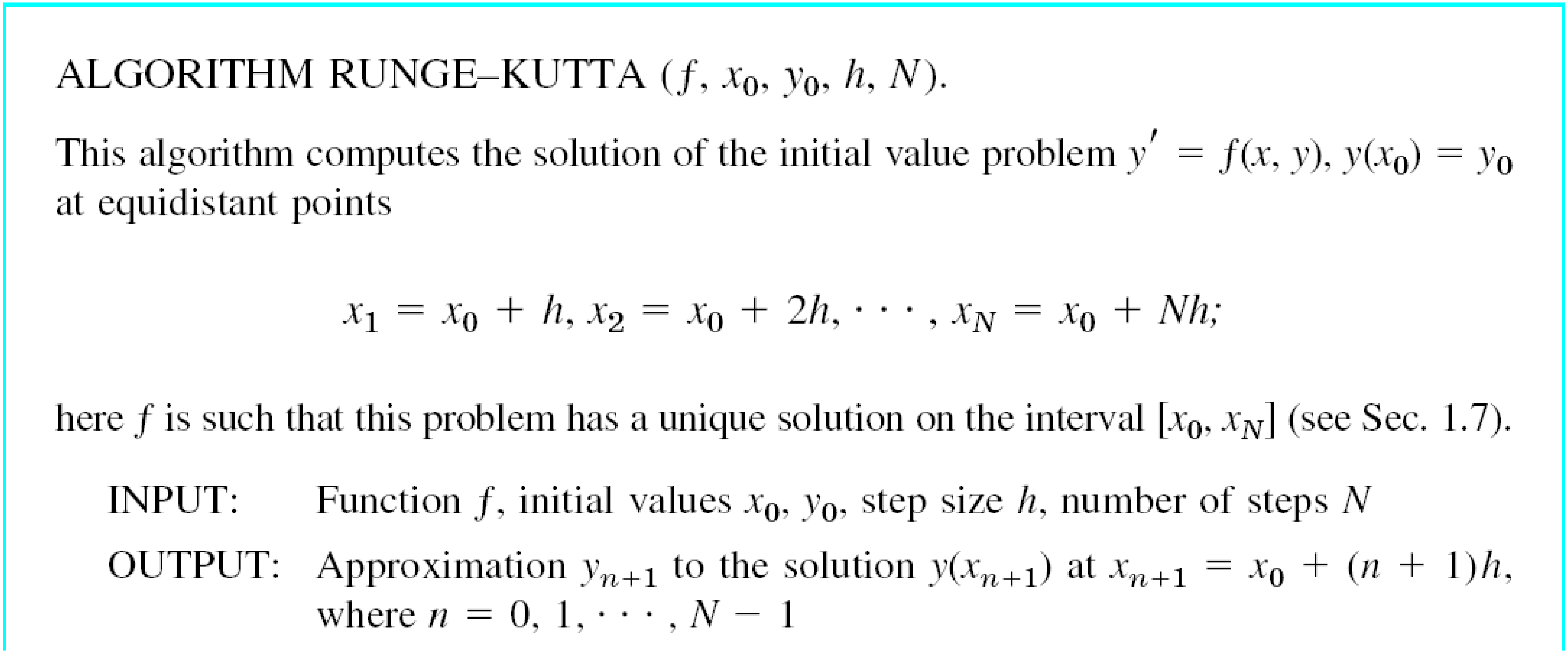
n	x_n	y_n	$0.22(x_n + y_n)$ $+ 0.02$	Exact Values (4D)	Error
0	0.0	0.0000	0.0200	0.0000	0.0000
1	0.2	0.0200	0.0684	0.0214	0.0014
2	0.4	0.0884	0.1274	0.0918	0.0034
3	0.6	0.2158	0.1995	0.2221	0.0063
4	0.8	0.4153	0.2874	0.4255	0.0102
5	1.0	0.7027		0.7183	0.0156

Runge–Kutta Methods (RK Methods)

A method of great practical importance and much greater accuracy than that of the improved Euler method is the *classical Runge–Kutta method of fourth order,* which we call briefly the **Runge–Kutta method.**[1] It is shown in Table 19.4 4. We see that in each step we first compute four auxiliary quantities k_1, k_2, k_3, k_4 and then the new value y_{n+1}. The method is well suited to the computer because it needs no special starting procedure, makes light demand on storage, and repeatedly uses the same straightforward computational procedure. It is numerically stable.

Note that if f depends only on x, this method reduces to Simpson's rule of integration Sec. 17.3 5). Note further that $k_1, \cdots, k_4$ depend on n and generally change from step to step.

Table 19.4 **Classical Runge–Kutta Method of Fourth Order**



ALGORITHM RUNGE–KUTTA (f, x_0, y_0, h, N).

This algorithm computes the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ at equidistant points

$$x_1 = x_0 + h, x_2 = x_0 + 2h, \cdots , x_N = x_0 + Nh;$$

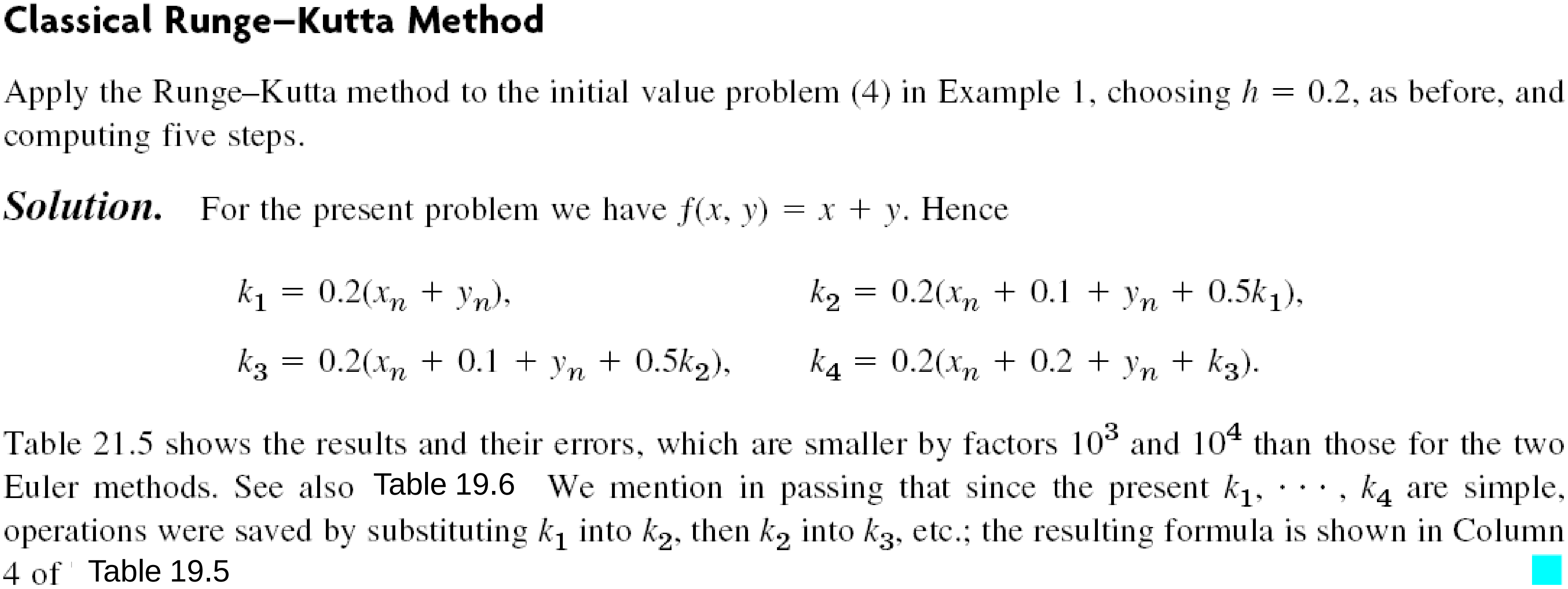
here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec. 1.7).

INPUT: Function f, initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, 1, \cdots , N - 1$

```
        For $n = 0, 1, \cdots, N - 1$ do:
            $k_1 = hf(x_n, y_n)$
            $k_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1)$
            $k_3 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2)$
            $k_4 = hf(x_n + h, y_n + k_3)$
            $x_{n+1} = x_n + h$
            $y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$
            OUTPUT $x_{n+1}, y_{n+1}$
        End
        Stop
End RUNGE–KUTTA
```

Example 3



Classical Runge–Kutta Method

Apply the Runge–Kutta method to the initial value problem (4) in Example 1, choosing $h = 0.2$, as before, and computing five steps.

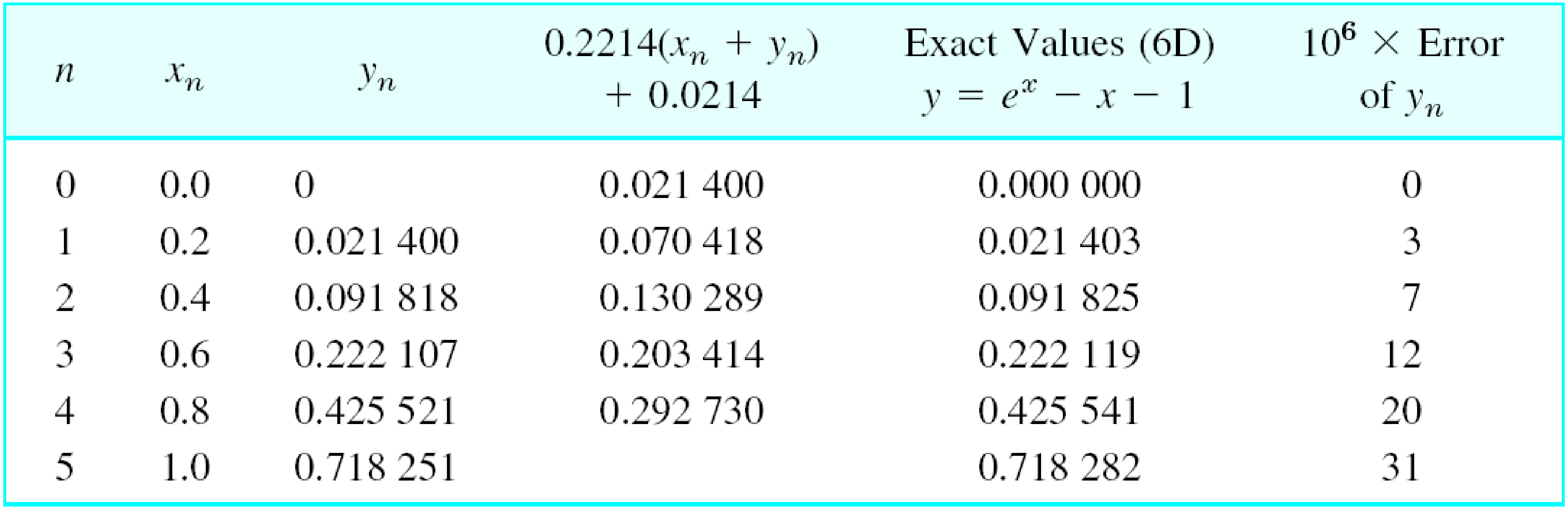
Solution. For the present problem we have $f(x, y) = x + y$. Hence

$$k_1 = 0.2(x_n + y_n), \qquad k_2 = 0.2(x_n + 0.1 + y_n + 0.5k_1),$$

$$k_3 = 0.2(x_n + 0.1 + y_n + 0.5k_2), \qquad k_4 = 0.2(x_n + 0.2 + y_n + k_3).$$

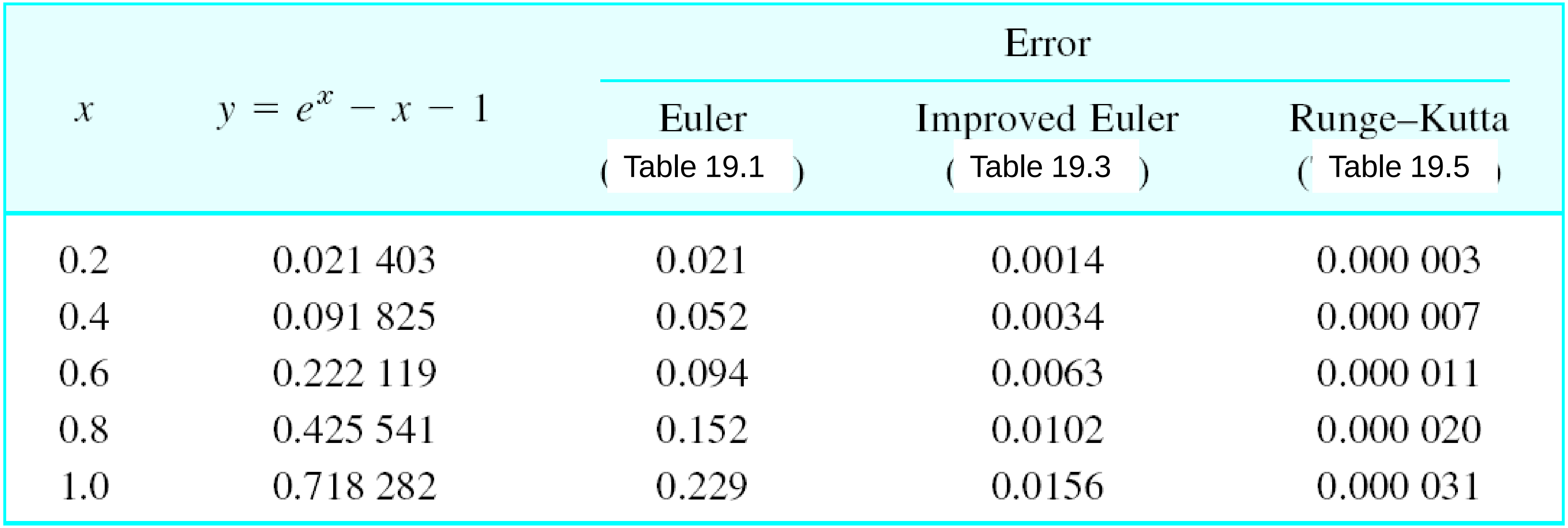
Table 21.5 shows the results and their errors, which are smaller by factors 10^3 and 10^4 than those for the two Euler methods. See also Table 19.6 We mention in passing that since the present $k_1, \cdots, k_4$ are simple, operations were saved by substituting k_1 into k_2, then k_2 into k_3, etc.; the resulting formula is shown in Column 4 of Table 19.5 ■

Table 19.5 **Runge–Kutta Method Applied to (4)**



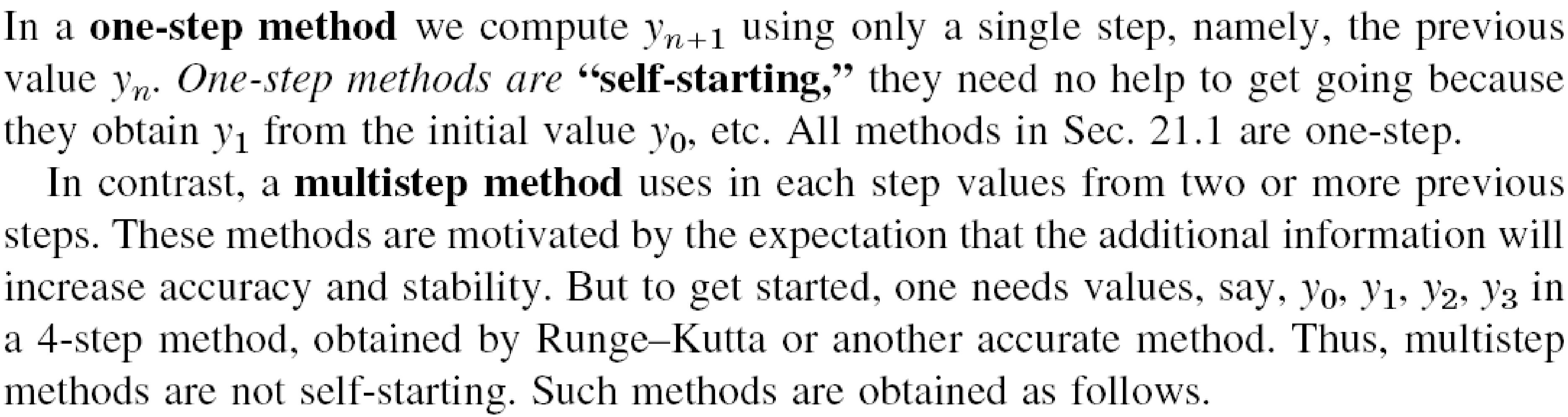
n	x_n	y_n	$0.2214(x_n + y_n)$ $+ 0.0214$	Exact Values (6D) $y = e^x - x - 1$	$10^6 \times$ Error of y_n
0	0.0	0	0.021 400	0.000 000	0
1	0.2	0.021 400	0.070 418	0.021 403	3
2	0.4	0.091 818	0.130 289	0.091 825	7
3	0.6	0.222 107	0.203 414	0.222 119	12
4	0.8	0.425 521	0.292 730	0.425 541	20
5	1.0	0.718 251		0.718 282	31

Table 19.6 **Comparison of the Accuracy of the Three Methods Under Consideration in the Case of the Initial Value Problem (4), with $h = 0.2$**



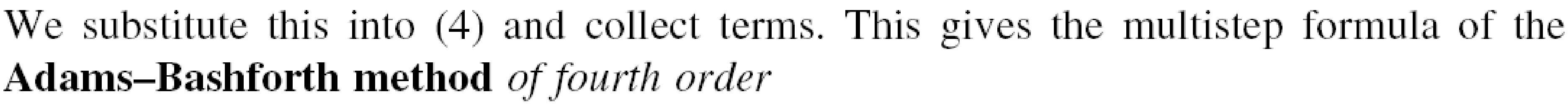
x	$y = e^x - x - 1$	Error		
		Euler (Table 19.1)	Improved Euler (Table 19.3)	Runge–Kutta (Table 19.5)
0.2	0.021 403	0.021	0.0014	0.000 003
0.4	0.091 825	0.052	0.0034	0.000 007
0.6	0.222 119	0.094	0.0063	0.000 011
0.8	0.425 541	0.152	0.0102	0.000 020
1.0	0.718 282	0.229	0.0156	0.000 031

19.2 Multistep methods

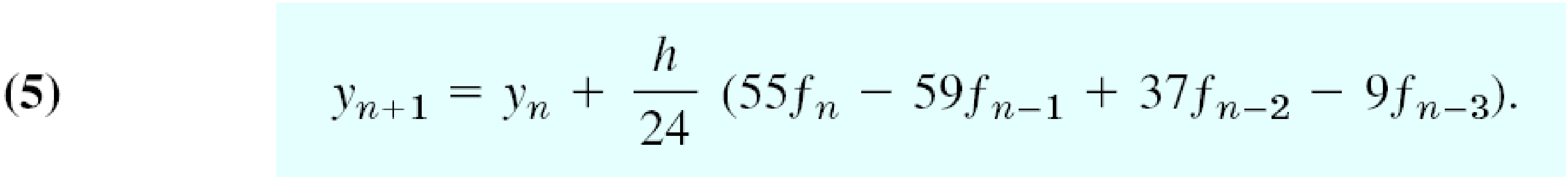


In a **one-step method** we compute y_{n+1} using only a single step, namely, the previous value y_n. *One-step methods are* **"self-starting,"** they need no help to get going because they obtain y_1 from the initial value y_0, etc. All methods in Sec. 21.1 are one-step.

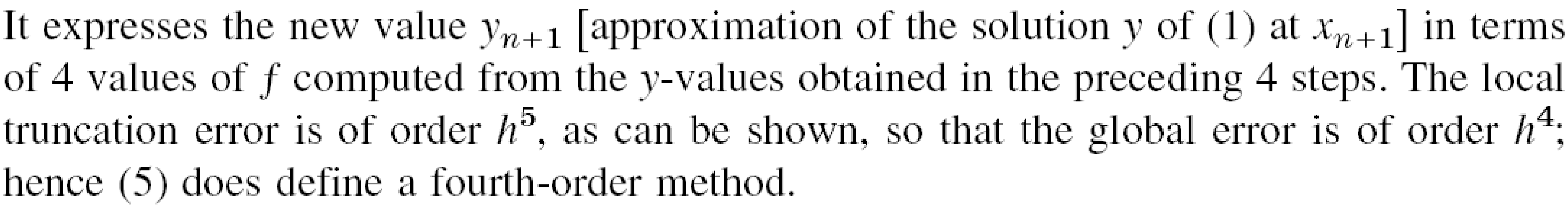
In contrast, a **multistep method** uses in each step values from two or more previous steps. These methods are motivated by the expectation that the additional information will increase accuracy and stability. But to get started, one needs values, say, y_0, y_1, y_2, y_3 in a 4-step method, obtained by Runge–Kutta or another accurate method. Thus, multistep methods are not self-starting. Such methods are obtained as follows.



We substitute this into (4) and collect terms. This gives the multistep formula of the **Adams–Bashforth method** *of fourth order*



$$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}). \tag{5}$$



It expresses the new value y_{n+1} [approximation of the solution y of (1) at x_{n+1}] in terms of 4 values of f computed from the y-values obtained in the preceding 4 steps. The local truncation error is of order h^5, as can be shown, so that the global error is of order h^4; hence (5) does define a fourth-order method.

$$(6) \qquad y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} \widetilde{p}_3(x)\,dx = y_n + \frac{h}{24}\,(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}).$$

This is usually called an **Adams–Moulton formula.** It is an **implicit formula** because $f_{n+1} = f(x_{n+1}, y_{n+1})$ appears on the right, so that it defines y_{n+1} only *implicitly,* in contrast to (5), which is an **explicit formula,** not involving y_{n+1} on the right. To use (6) we must ***predict*** a value y^*_{n+1}, for instance, by using (5), that is,

(7a)

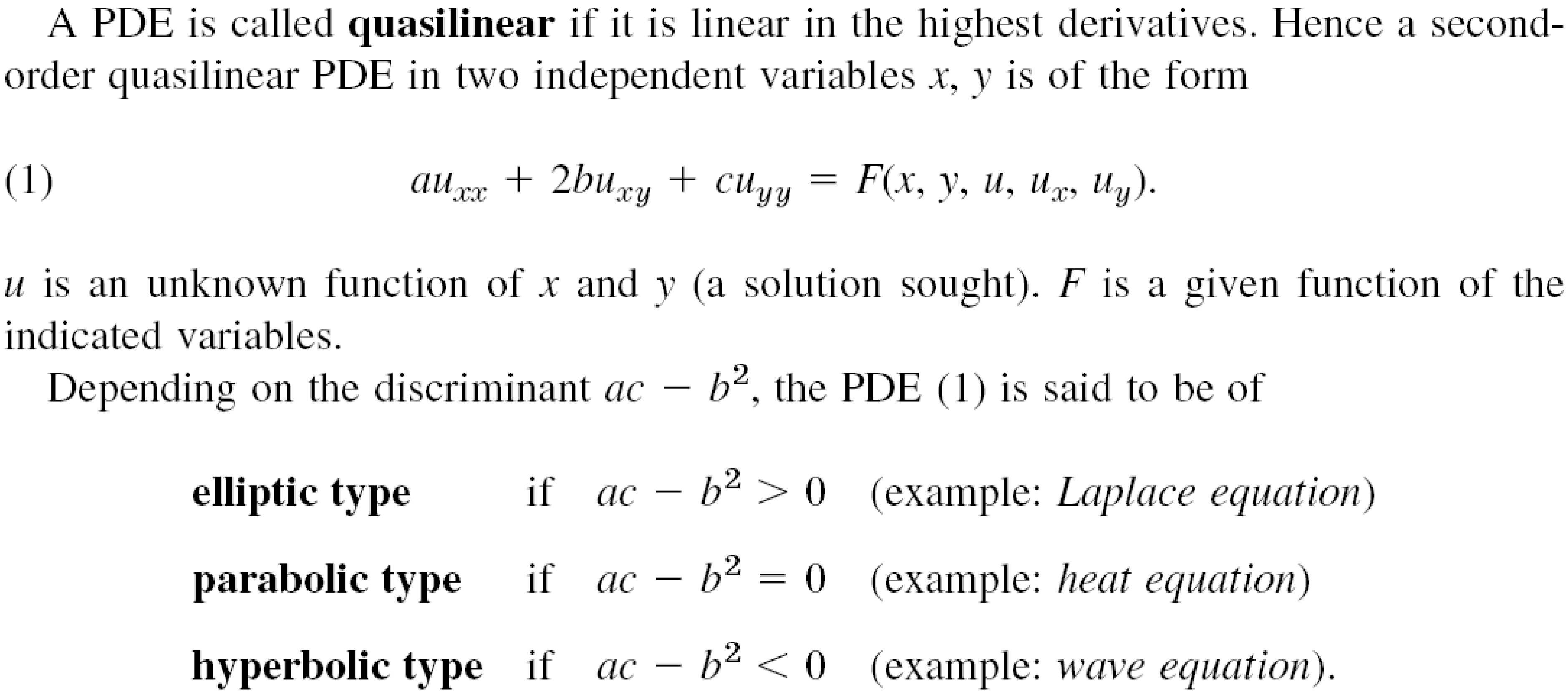
$$y^*_{n+1} = y_n + \frac{h}{24}\,(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}).$$

The ***corrected*** new value y_{n+1} is then obtained from (6) with f_{n+1} replaced by $f^*_{n+1} = f(x_{n+1}, y^*_{n+1})$ and the other f's as in (6); thus,

(7b)

$$y_{n+1} = y_n + \frac{h}{24}\,(9f^*_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}).$$

19.4 Methods for elliptic partial differential equations



A PDE is called **quasilinear** if it is linear in the highest derivatives. Hence a second-order quasilinear PDE in two independent variables x, y is of the form

$$au_{xx} + 2bu_{xy} + cu_{yy} = F(x, y, u, u_x, u_y). \tag{1}$$

u is an unknown function of x and y (a solution sought). F is a given function of the indicated variables.

Depending on the discriminant $ac - b^2$, the PDE (1) is said to be of

elliptic type	if	$ac - b^2 > 0$	(example: *Laplace equation*)
parabolic type	if	$ac - b^2 = 0$	(example: *heat equation*)
hyperbolic type	if	$ac - b^2 < 0$	(example: *wave equation*).

In this section we consider the **Laplace equation**

$$\nabla^2 u = u_{xx} + u_{yy} = 0 \tag{2}$$

and the **Poisson equation**

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y). \tag{3}$$

These are the most important elliptic PDEs in applications. To obtain methods of numeric solution, we replace the partial derivatives by corresponding **difference quotients,** as follows. By the Taylor formula,

$$(4)\quad \begin{aligned} &\text{(a)}\quad u(x+h, y) = u(x, y) + hu_x(x, y) + \tfrac{1}{2}h^2 u_{xx}(x, y) + \tfrac{1}{6}h^3 u_{xxx}(x, y) + \cdots \\ &\text{(b)}\quad u(x-h, y) = u(x, y) - hu_x(x, y) + \tfrac{1}{2}h^2 u_{xx}(x, y) - \tfrac{1}{6}h^3 u_{xxx}(x, y) + \cdots \end{aligned}$$

We subtract (4b) from (4a), neglect terms in $h^3, h^4, \cdots$, and solve for u_x. Then

$$u_x(x, y) \approx \frac{1}{2h}\,[u(x+h, y) - u(x-h, y)]. \tag{5a}$$

Similarly,

$$u(x, y+k) = u(x, y) + ku_y(x, y) + \tfrac{1}{2}k^2 u_{yy}(x, y) + \cdots$$

and

$$u(x, y-k) = u(x, y) - ku_y(x, y) + \tfrac{1}{2}k^2 u_{yy}(x, y) + \cdots .$$

By subtracting, neglecting terms in $k^3, k^4, \cdots$, and solving for u_y we obtain

$$u_y(x, y) \approx \frac{1}{2k}\,[u(x, y+k) - u(x, y-k)]. \tag{5b}$$

We now turn to second derivatives. Adding (4a) and (4b) and neglecting terms in $h^4, h^5, \cdots$, we obtain $u(x + h, y) + u(x - h, y) \approx 2u(x, y) + h^2 u_{xx}(x, y)$. Solving for u_{xx}, we have

(6a)
$$u_{xx}(x, y) \approx \frac{1}{h^2}\,[u(x + h, y) - 2u(x, y) + u(x - h, y)].$$

Similarly,

(6b)
$$u_{yy}(x, y) \approx \frac{1}{k^2}\,[u(x, y + k) - 2u(x, y) + u(x, y - k)].$$

We shall not need (see Prob. 1)

(6c)
$$\begin{aligned} u_{xy}(x, y) \approx \frac{1}{4hk}\,[&u(x + h, y + k) - u(x - h, y + k) \\ &- u(x + h, y - k) + u(x - h, y - k)]. \end{aligned}$$

Figure 452a shows the points $(x + h, y), (x - h, y), \cdots$ in (5) and (6).

We now substitute (6a) and (6b) into the ***Poisson equation*** (3), choosing $k = h$ to obtain a simple formula:

$$u(x + h, y) + u(x, y + h) + u(x - h, y) + u(x, y - h) - 4u(x, y) = h^2 f(x, y). \tag{7}$$

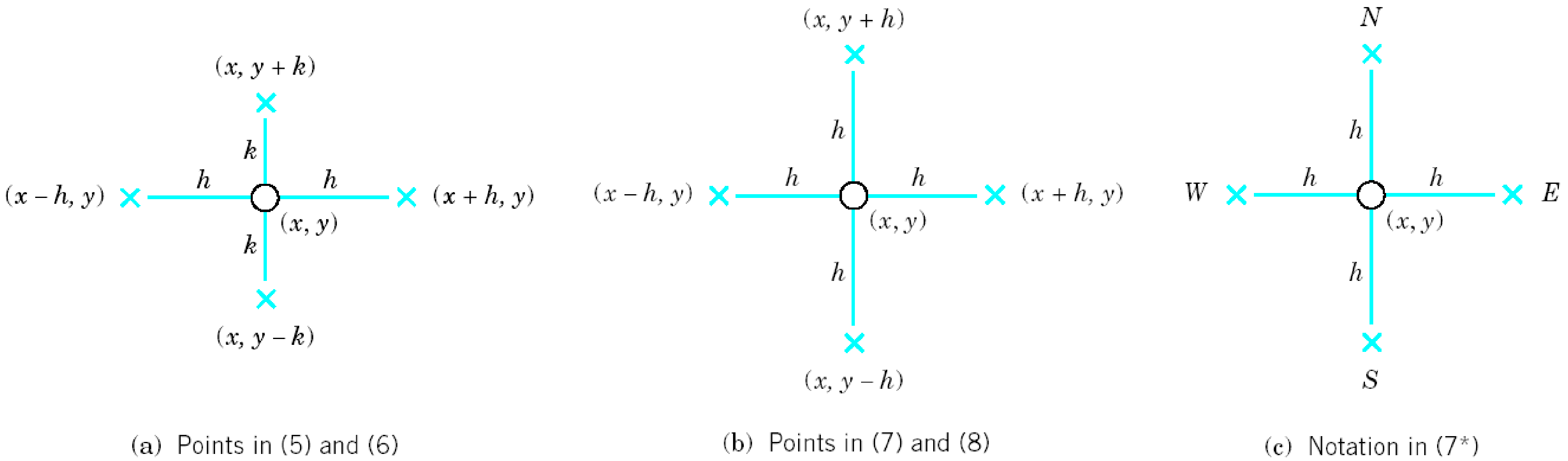
This is a **difference equation** corresponding to (3). Hence for the ***Laplace equation*** (2) the corresponding difference equation is

$$u(x + h, y) + u(x, y + h) + u(x - h, y) + u(x, y - h) - 4u(x, y) = 0. \tag{8}$$

h is called the **mesh size.** Equation (8) relates u at (x, y) to u at the four neighboring points shown in Fig. 452b. It has a remarkable interpretation: u at (x, y) equals the mean of the values of u at the four neighboring points. This is an analog of the mean value property of harmonic functions (Sec. 18.6).

Those neighbors are often called E (East), N (North), W (West), S (South). Then Fig. 452b becomes Fig. 452c and (7) is

$$u(E) + u(N) + u(W) + u(S) - 4u(x, y) = h^2 f(x, y). \tag{7*}$$



(a) Points in (5) and (6)

(b) Points in (7) and (8)

(c) Notation in (7*)

Fig. 452. Points and notation in (5)–(8) and (7*)

Dirichlet Problem

In numerics for the Dirichlet problem in a region R we choose an h and introduce a square grid of horizontal and vertical straight lines of distance h. Their intersections are called **mesh points** (or *lattice points* or *nodes*). See Fig. 453.

Then we approximate the given PDE by a difference equation [(8) for the Laplace equation], which relates the unknown values of u at the mesh points in R to each other and to the given boundary values (details on p. 913). This gives a linear system of *algebraic* equations. By solving it we get approximations of the unknown values of u at the mesh points in R.

We shall see that the number of equations equals the number of unknowns. Now comes an important point. If the number of internal mesh points, call it p, is small, say, $p < 100$, then a direct solution method may be applied to that linear system of $p < 100$ equations in p unknowns. However, if p is large, a storage problem will arise. Now since each unknown u is related to only 4 of its neighbors, the coefficient matrix of the system is a **sparse matrix,** that is, a matrix with relatively few nonzero entries (for instance, 500 of 10000 when $p = 100$). Hence for large p we may avoid storage difficulties by using an iteration method, notably the Gauss–Seidel method (Sec. 20.3), which in PDEs is also called **Liebmann's method.** Remember that in this method we have the storage convenience that we can overwrite any solution component (value of u) as soon as a "new" value is available.

(10) $$P_{ij} = (ih, jh), \qquad u_{ij} = u(ih, jh).$$

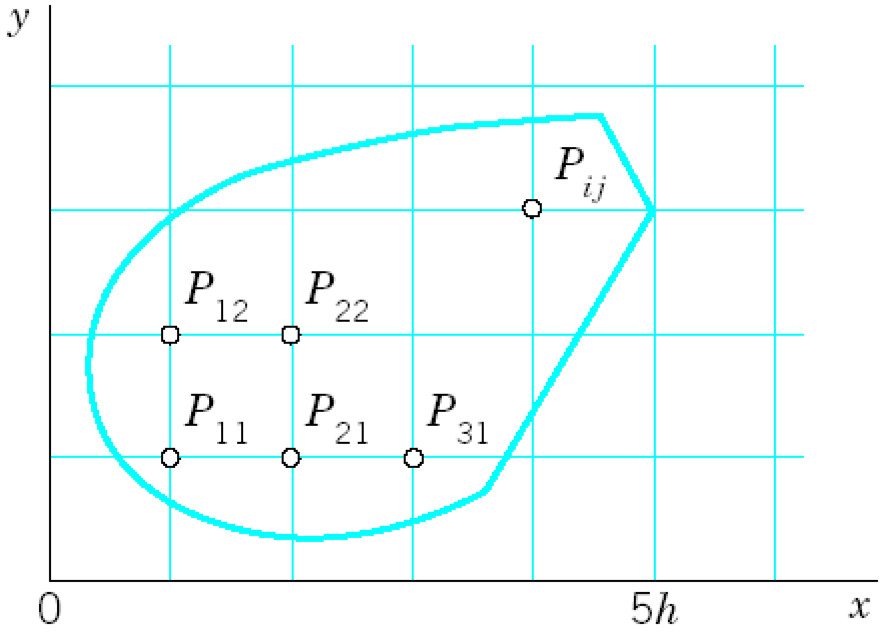


Fig. 453. Region in the *xy*-plane covered by a grid of mesh *h*, also showing mesh points $P_{11} = (h, h), \cdots, P_{ij} = (ih, jh), \cdots$

With this notation we can write (8) for any mesh point P_{ij} in the form

(11) $$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} = 0.$$

EXAMPLE 1

Laplace Equation. Liebmann's Method

The four sides of a square plate of side 12 cm made of homogeneous material are kept at constant temperature 0°C and 100°C as shown in Fig. 454a. Using a (very wide) grid of mesh 4 cm and applying Liebmann's method (that is, Gauss–Seidel iteration), find the (steady-state) temperature at the mesh points.

Solution. In the case of independence of time, the heat equation (see Sec. 10.8)

$$u_t = c^2(u_{xx} + u_{yy})$$

reduces to the Laplace equation. Hence our problem is a Dirichlet problem for the latter. We choose the grid shown in Fig. 454b and consider the mesh points in the order P_{11}, P_{21}, P_{12}, P_{22}. We use (11) and, in each equation, take to the right all the terms resulting from the given boundary values. Then we obtain the system

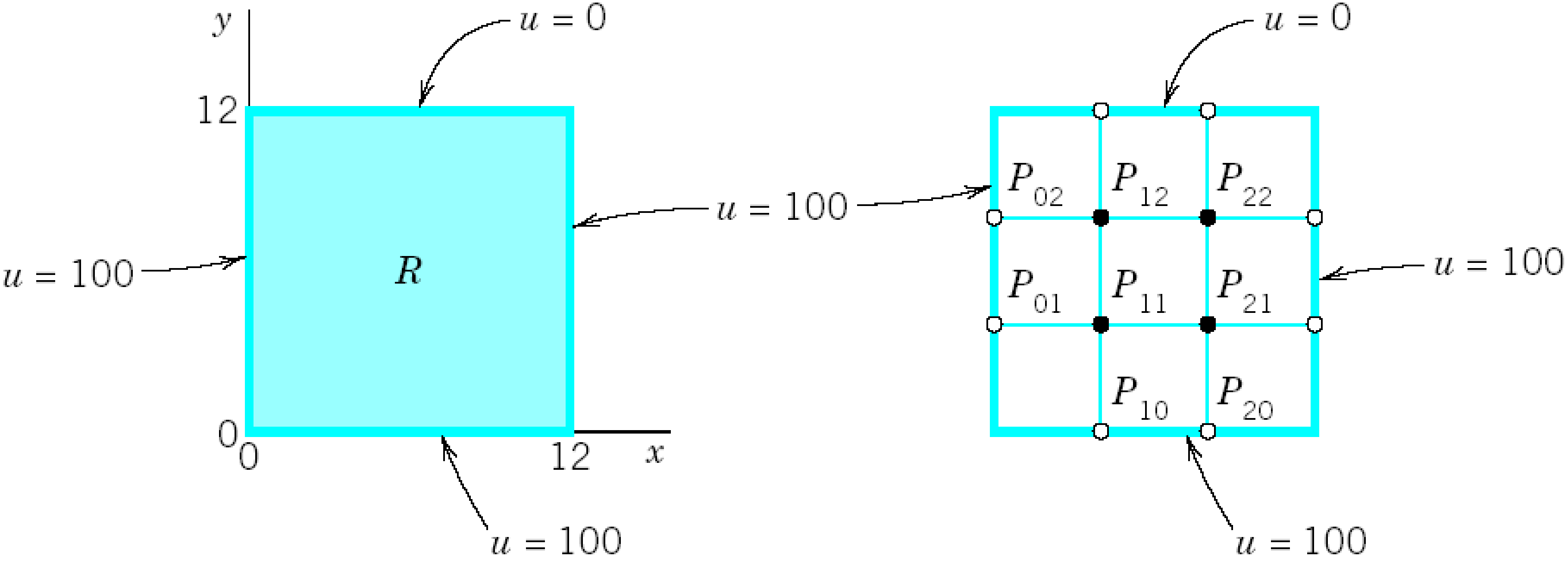
$$\begin{aligned} -4u_{11} + u_{21} + u_{12} \phantom{{}+ u_{22}} &= -200 \\ u_{11} - 4u_{21} \phantom{{}+u_{12}} + u_{22} &= -200 \\ u_{11} \phantom{{}+u_{21}} - 4u_{12} + u_{22} &= -100 \\ u_{21} + u_{12} - 4u_{22} &= -100. \end{aligned} \tag{12}$$

In practice, one would solve such a small system by the Gauss elimination, finding $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$.

More exact values (exact to 3S) of the solution of the actual problem [as opposed to its model (12)] are 88.1 and 61.9, respectively. (These were obtained by using Fourier series.) Hence the error is about 1%, which is surprisingly accurate for a grid of such a large mesh size h. If the system of equations were large, one would solve it by an indirect method, such as Liebmann's method. For (12) this is as follows. We write (12) in the form (divide by -4 and take terms to the right)

$$\begin{aligned}
u_{11} &= \phantom{0.25u_{11}} \quad 0.25u_{21} + 0.25u_{12} \phantom{+ 0.25u_{22}} + 50 \\
u_{21} &= 0.25u_{11} \phantom{+ 0.25u_{21} + 0.25u_{12}} + 0.25u_{22} + 50 \\
u_{12} &= 0.25u_{11} \phantom{+ 0.25u_{21} + 0.25u_{12}} + 0.25u_{22} + 25 \\
u_{22} &= \phantom{0.25u_{11}} \quad 0.25u_{21} + 0.25u_{12} \phantom{+ 0.25u_{22}} + 25.
\end{aligned}$$

These equations are now used for the Gauss–Seidel iteration. They are identical with (2) in Sec. 20.3, where $u_{11} = x_1$, $u_{21} = x_2$, $u_{12} = x_3$, $u_{22} = x_4$, and the iteration is explained there, with 100, 100, 100, 100 chosen as starting values. Some work can be saved by better starting values, usually by taking the average of the boundary values that enter into the linear system. The exact solution of the system is $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$, as you may verify.



(a) Given problem

(b) Grid and mesh points

Fig. 454. Example 1

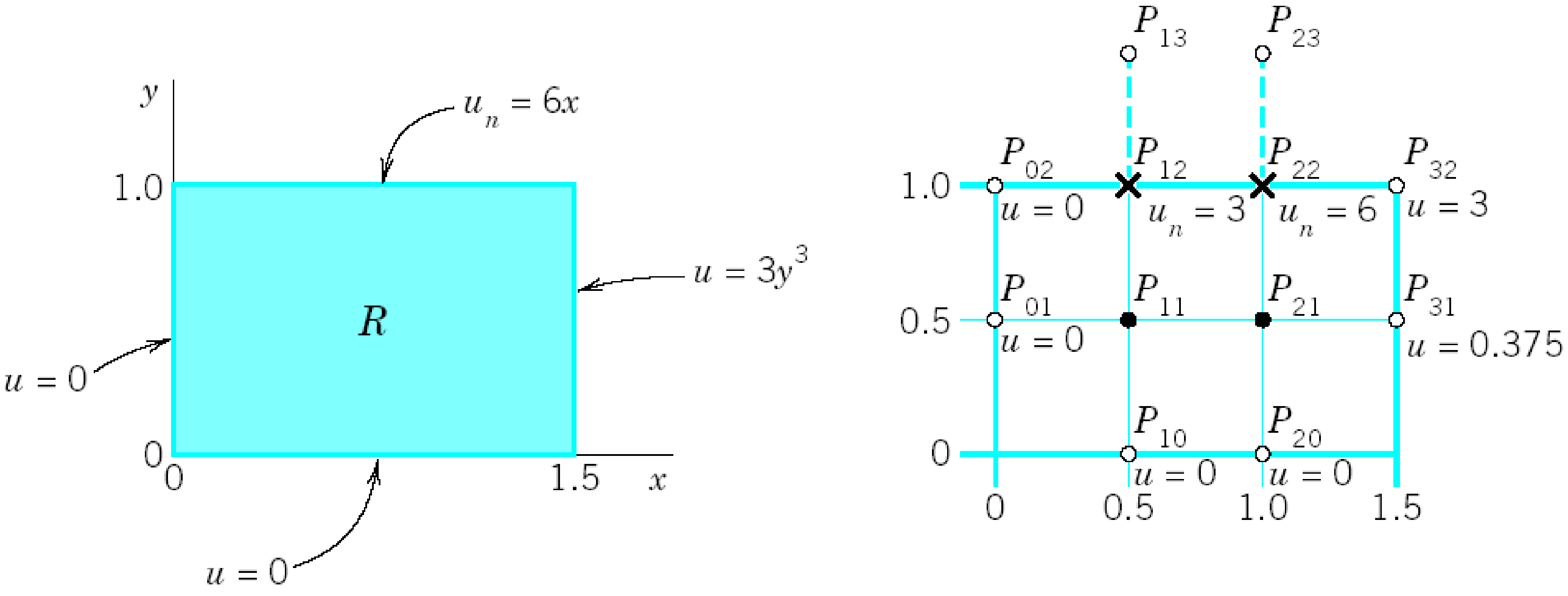
EXAMPLE 1

Mixed Boundary Value Problem for a Poisson Equation

Solve the mixed boundary value problem for the Poisson equation

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y) = 12xy$$

shown in Fig. 457a.



(a) Region R and boundary values

(b) Grid ($h = 0.5$)

Fig. 457. Mixed boundary value problem in Example 1

Solution. We use the grid shown in Fig. 457b, where $h = 0.5$. We recall that (7) in Sec. 21.4 has the right side $h^2 f(x, y) = 0.5^2 \cdot 12xy = 3xy$. From the formulas $u = 3y^3$ and $u_n = 6x$ given on the boundary we compute the boundary data

$$u_{31} = 0.375, \qquad u_{32} = 3, \qquad \frac{\partial u_{12}}{\partial n} = \frac{\partial u_{12}}{\partial y} = 6 \cdot 0.5 = 3, \qquad \frac{\partial u_{22}}{\partial n} = \frac{\partial u_{22}}{\partial y} = 6 \cdot 1 = 6. \tag{1}$$

P_{11} and P_{21} are internal mesh points and can be handled as in the last section. Indeed, from (7), Sec. 21.4, with $h^2 = 0.25$ and $h^2 f(x, y) = 3xy$ and from the given boundary values we obtain two equations corresponding to P_{11} and P_{21}, as follows (with -0 resulting from the left boundary).

$$\begin{aligned} -4u_{11} + u_{21} + u_{12} \phantom{+ u_{22}} &= 12(0.5 \cdot 0.5) \cdot \tfrac{1}{4} - 0 = 0.75 \\ u_{11} - 4u_{21} \phantom{+ u_{12}} + u_{22} &= 12(1 \cdot 0.5) \cdot \tfrac{1}{4} - 0.375 = 1.125 \end{aligned} \tag{2a}$$

The only difficulty with these equations seems to be that they involve the unknown values u_{12} and u_{22} of u at P_{12} and P_{22} on the boundary, where the normal derivative $u_n = \partial u/\partial n = \partial u/\partial y$ is given, instead of u; but we shall overcome this difficulty as follows.

We consider P_{12} and P_{22}. The idea that will help us here is this. We imagine the region R to be extended above to the first row of external mesh points (corresponding to $y = 1.5$), and we assume that the Poisson equation also holds in the extended region. Then we can write down two more equations as before (Fig. 457b)

$$
\text{(2b)}\qquad
\begin{aligned}
u_{11} \quad - 4u_{12} + \ u_{22} + u_{13} \qquad &= 1.5 - 0 = 1.5 \\
u_{21} + \ u_{12} - 4u_{22} \qquad + u_{23} &= 3 - 3 = 0.
\end{aligned}
$$

On the right, 1.5 is $12xyh^2$ at (0.5, 1) and 3 is $12xyh^2$ at (1, 1) and 0 (at P_{02}) and 3 (at P_{32}) are given boundary values. We remember that we have not yet used the boundary condition on the upper part of the boundary of R, and we also notice that in (2b) we have introduced two more unknowns u_{13}, u_{23}. But we can now use that condition and get rid of u_{13}, u_{23} by applying the central difference formula for du/dy. From (1) we then obtain (see Fig. 457b)

$$3 = \frac{\partial u_{12}}{\partial y} \approx \frac{u_{13} - u_{11}}{2h} = u_{13} - u_{11}, \qquad \text{hence} \qquad u_{13} = u_{11} + 3$$

$$6 = \frac{\partial u_{22}}{\partial y} \approx \frac{u_{23} - u_{21}}{2h} = u_{23} - u_{21}, \qquad \text{hence} \qquad u_{23} = u_{21} + 6.$$

Substituting these results into (2b) and simplifying, we have

$$
\begin{aligned}
2u_{11} \quad - 4u_{12} + \ u_{22} &= 1.5 - 3 = -1.5 \\
2u_{21} + \ u_{12} - 4u_{22} &= 3 - 3 - 6 = -6.
\end{aligned}
$$

Together with (2a) this yields, written in matrix form,

$$\begin{bmatrix} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ 2 & 0 & -4 & 1 \\ 0 & 2 & 1 & -4 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ 1.5 - 3 \\ 0 - 6 \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ -1.5 \\ -6 \end{bmatrix}. \tag{3}$$

(The entries 2 come from u_{13} and u_{23}, and so do -3 and -6 on the right). The solution of (3) (obtained by Gauss elimination) is as follows; the exact values of the problem are given in parentheses.

$$u_{12} = 0.866 \quad (\text{exact } 1) \qquad u_{22} = 1.812 \quad (\text{exact } 2)$$

$$u_{11} = 0.077 \quad (\text{exact } 0.125) \qquad u_{21} = 0.191 \quad (\text{exact } 0.25).$$

■

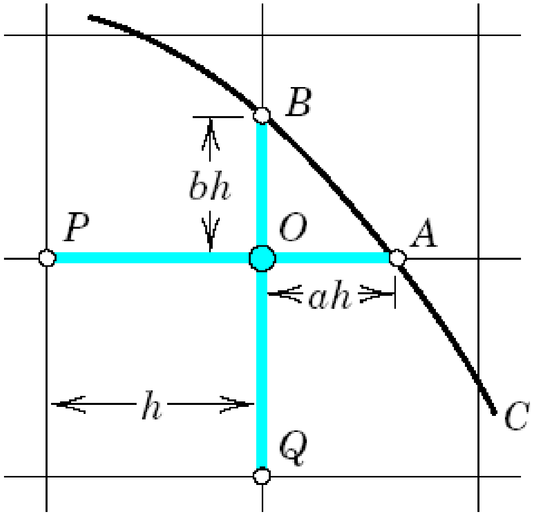


Fig. 458. Curved boundary C of a region R, a mesh point O near C, and neighbors A, B, P, Q

$$\textbf{(5)}\quad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(1+a)} + \frac{u_B}{b(1+b)} + \frac{u_P}{1+a} + \frac{u_Q}{1+b} - \frac{(a+b)u_O}{ab}\right].$$

For example, if $a = \frac{1}{2}$, $b = \frac{1}{2}$, instead of the stencil (see Sec. 21.4)

$$\left\{\begin{matrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{matrix}\right\} \quad \text{we now have} \quad \left\{\begin{matrix} & \frac{4}{3} & \\ \frac{2}{3} & -4 & \frac{4}{3} \\ & \frac{2}{3} & \end{matrix}\right\}.$$

because $1/[a(1+a)] = \frac{4}{3}$, etc. The sum of all five terms still being zero (which is useful for checking).

Using the same ideas, you may show that in the case of Fig. 459.

$$\textbf{(6)} \quad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(a+p)} + \frac{u_B}{b(b+q)} + \frac{u_P}{p(p+a)} + \frac{u_Q}{q(q+b)} - \frac{ap+bq}{abpq}\,u_O\right],$$

a formula that takes care of all conceivable cases.

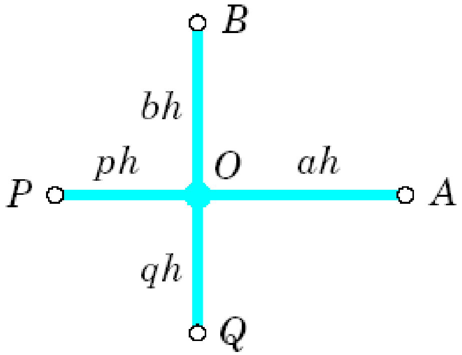


Fig. 459. Neighboring points *A*, *B*, *P*, *Q* of a mesh point *O* and notations in formula (6)

EXAMPLE 2

Dirichlet Problem for the Laplace Equation. Curved Boundary

Find the potential u in the region in Fig. 460 that has the boundary values given in that figure; here the curved portion of the boundary is an arc of the circle of radius 10 about (0, 0). Use the grid in the figure.

Solution. u is a solution of the Laplace equation. From the given formulas for the boundary values $u = x^3$, $u = 512 - 24y^2$, · · · we compute the values at the points where we need them; the result is shown in the figure. For P_{11} and P_{12} we have the usual regular stencil, and for P_{21} and P_{22} we use (6), obtaining

$$(7) \qquad P_{11}, P_{12}: \left\{\begin{matrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{matrix}\right\}, \qquad P_{21}: \left\{\begin{matrix} & 0.5 & \\ 0.6 & -2.5 & 0.9 \\ & 0.5 & \end{matrix}\right\}, \qquad P_{22}: \left\{\begin{matrix} & 0.9 & \\ 0.6 & -3 & 0.9 \\ & 0.6 & \end{matrix}\right\}.$$

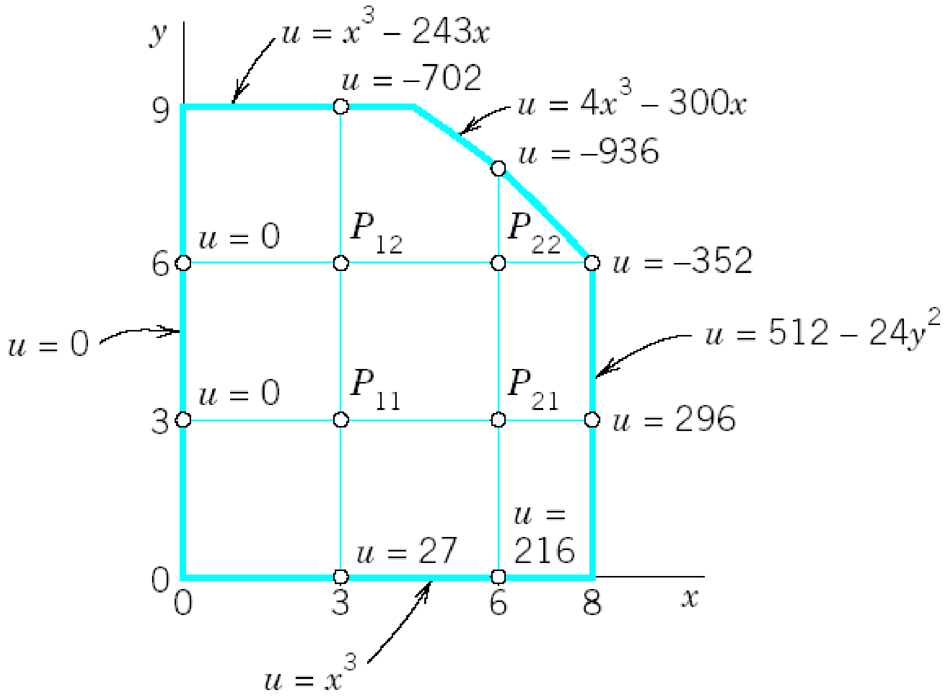


Fig. 460. Region, boundary values of the potential, and grid in Example 2

We use this and the boundary values and take the mesh points in the usual order $P_{11}, P_{21}, P_{12}, P_{22}$. Then we obtain the system

$$\begin{array}{rcl}
-4u_{11} + u_{21} + u_{12} \phantom{+ 0.5u_{22}} & = & 0 - 27 = -27 \\
0.6u_{11} - 2.5u_{21} \phantom{+ u_{12}} + 0.5u_{22} & = & -0.9 \cdot 296 - 0.5 \cdot 216 = -374.4 \\
u_{11} \phantom{+ u_{21}} - 4u_{12} + u_{22} & = & 702 + 0 = 702 \\
0.6u_{21} + 0.6u_{12} - 3u_{22} & = & 0.9 \cdot 352 + 0.9 \cdot 936 = 1159.2.
\end{array}$$

In matrix form,

$$\begin{bmatrix} -4 & 1 & 1 & 0 \\ 0.6 & -2.5 & 0 & 0.5 \\ 1 & 0 & -4 & 1 \\ 0 & 0.6 & 0.6 & -3 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -27 \\ -374.4 \\ 702 \\ 1159.2 \end{bmatrix}. \tag{8}$$

Gauss elimination yields the (rounded) values

$$u_{11} = -55.6, \qquad u_{21} = 49.2, \qquad u_{12} = -298.5, \qquad u_{22} = -436.3.$$

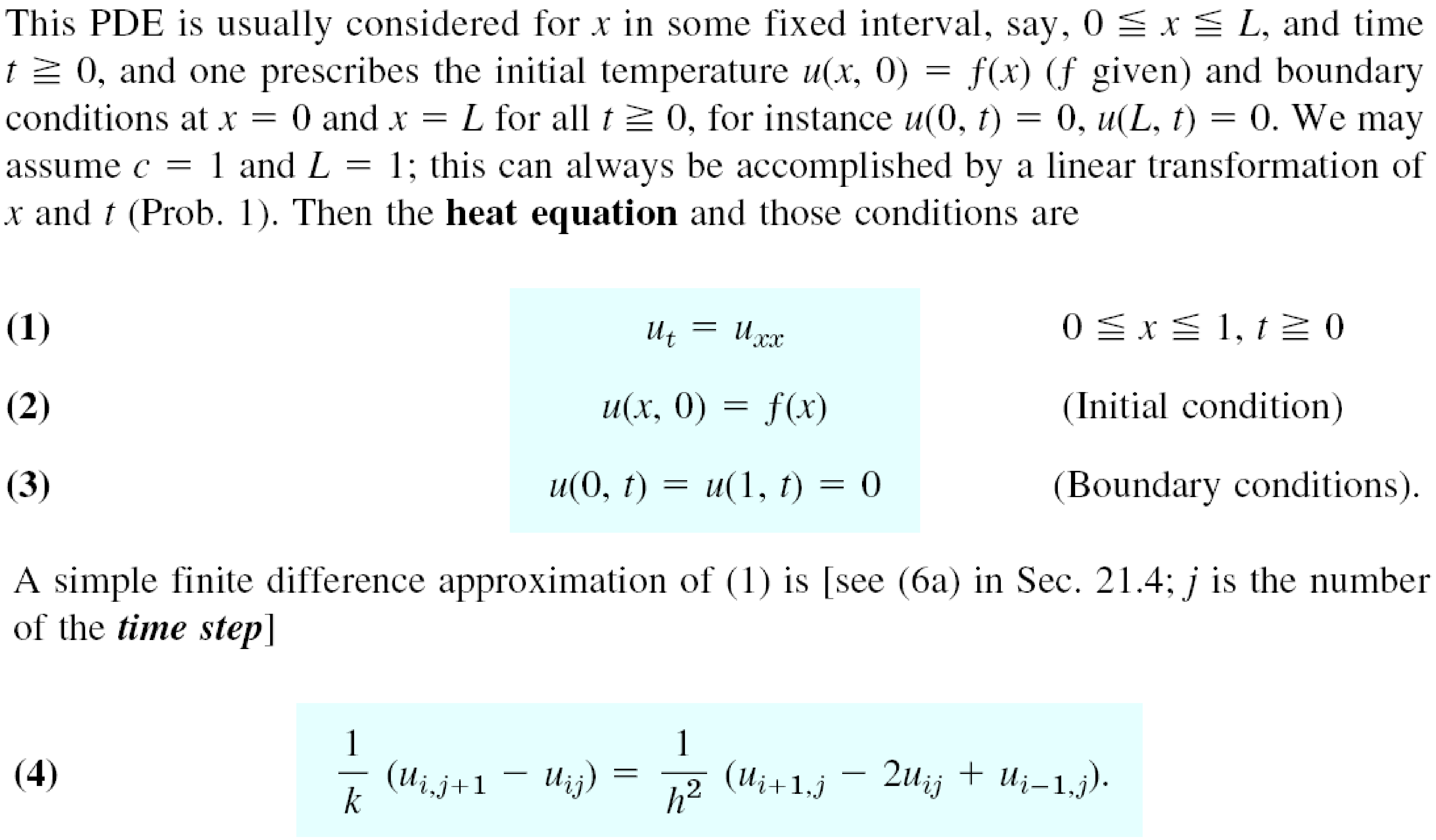
Clearly, from a grid with so few mesh points we cannot expect great accuracy. The exact solution of the PDE (not of the difference equation) having the given boundary values is $u = x^3 - 3xy^2$ and yields the values

$$u_{11} = -54, \qquad u_{21} = 54, \qquad u_{12} = -297, \qquad u_{22} = -432.$$

In practice one would use a much finer grid and solve the resulting large system by an indirect method. ■

Parabolic PDEs

$$u_t = c^2 u_{xx}$$



This PDE is usually considered for x in some fixed interval, say, $0 \leqq x \leqq L$, and time $t \geqq 0$, and one prescribes the initial temperature $u(x, 0) = f(x)$ (f given) and boundary conditions at $x = 0$ and $x = L$ for all $t \geqq 0$, for instance $u(0, t) = 0$, $u(L, t) = 0$. We may assume $c = 1$ and $L = 1$; this can always be accomplished by a linear transformation of x and t (Prob. 1). Then the **heat equation** and those conditions are

(1) $$u_t = u_{xx} \qquad 0 \leqq x \leqq 1,\ t \geqq 0$$

(2) $$u(x, 0) = f(x) \qquad \text{(Initial condition)}$$

(3) $$u(0, t) = u(1, t) = 0 \qquad \text{(Boundary conditions).}$$

A simple finite difference approximation of (1) is [see (6a) in Sec. 21.4; j is the number of the ***time step***]

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

Figure 464 shows a corresponding grid and mesh points. The mesh size is h in the x-direction and k in the t-direction. Formula (4) involves the four points shown in Fig. 465. On the left in (4) we have used a *forward* difference quotient since we have no information for negative t at the start. From (4) we calculate $u_{i,j+1}$, which corresponds to time row $j + 1$, in terms of the three other u that correspond to time row j. Solving (4) for $u_{i,j+1}$, we have

$$u_{i,j+1} = (1 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}), \qquad r = \frac{k}{h^2} \, . \tag{5}$$

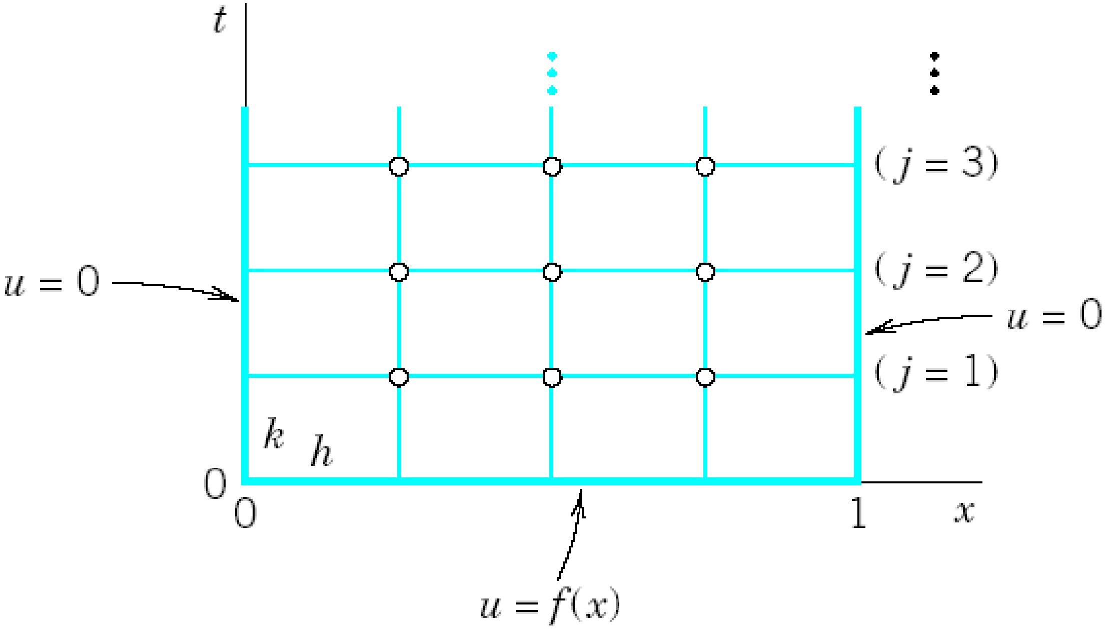


Fig. 464. Grid and mesh points corresponding to (4), (5)

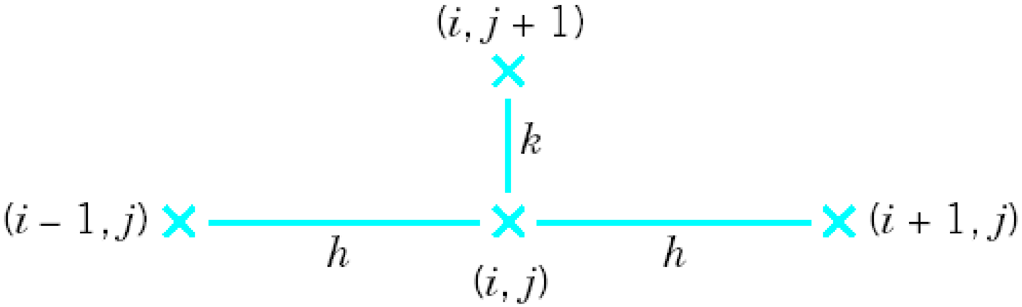


Fig. 465. The four points in (4) and (5)

Computations by this **explicit method** based on (5) are simple. However, it can be shown that crucial to the convergence of this method is the condition

$$r = \frac{k}{h^2} \leqq \frac{1}{2}\ . \tag{6}$$

That is, u_{ij} should have a positive coefficient in (5) or (for $r = \frac{1}{2}$) be absent from (5). Intuitively, (6) means that we should not move too fast in the t-direction. An example is given below.

Crank–Nicolson Method

Condition (6) is a handicap in practice. Indeed, to attain sufficient accuracy, we have to choose h small, which makes k very small by (6). For example, if $h = 0.1$, then $k \leqq 0.005$. Accordingly, we should look for a more satisfactory discretization of the heat equation.

A method that imposes no restriction on $r = k/h^2$ is the **Crank–Nicolson method,** which uses values of u at the six points in Fig. 466. The idea of the method is the replacement of the difference quotient on the right side of (4) by $\frac{1}{2}$ times the sum of two such difference quotients at two time rows (see Fig. 466). Instead of (4) we then have

$$\textbf{(7)} \qquad \begin{aligned} \frac{1}{k}(u_{i,j+1} - u_{ij}) &= \frac{1}{2h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \\ &+ \frac{1}{2h^2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}). \end{aligned}$$

Multiplying by $2k$ and writing $r = k/h^2$ as before, we collect the terms corresponding to time row $j + 1$ on the left and the terms corresponding to time row j on the right:

$$\textbf{(8)} \qquad (2 + 2r)u_{i,j+1} - r(u_{i+1,j+1} + u_{i-1,j+1}) = (2 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}).$$

How do we use (8)? In general, the three values on the left are unknown, whereas the three values on the right are known. If we divide the x-interval $0 \leqq x \leqq 1$ in (1) into n equal intervals, we have $n - 1$ internal mesh points per time row (see Fig. 464, where $n = 4$). Then for $j = 0$ and $i = 1, \cdots, n - 1$, formula (8) gives a linear system of $n - 1$ equations for the $n - 1$ unknown values $u_{11}, u_{21}, \cdots, u_{n-1,1}$ in the first time row in terms of the initial values $u_{00}, u_{10}, \cdots, u_{n0}$ and the boundary values u_{01} $(= 0)$, u_{n1} $(= 0)$. Similarly for $j = 1$, $j = 2$, and so on; that is, for each time row we have to solve such a linear system of $n - 1$ equations resulting from (8).

Although $r = k/h^2$ is no longer restricted, smaller r will still give better results. In practice, one chooses a k by which one can save a considerable amount of work, without making r too large. For instance, often a good choice is $r = 1$ (which would be impossible in the previous method). Then (8) becomes simply

(9) $$4u_{i,j+1} - u_{i+1,j+1} - u_{i-1,j+1} = u_{i+1,j} + u_{i-1,j} \,.$$

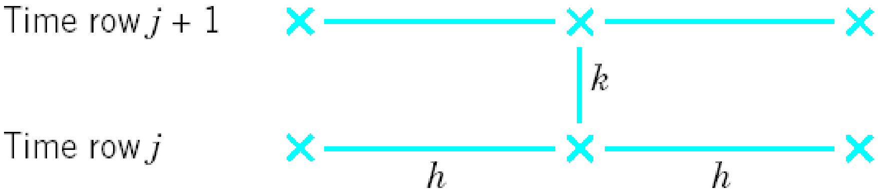


Fig. 466. The six points in the Crank–Nicolson formulas (7) and (8)

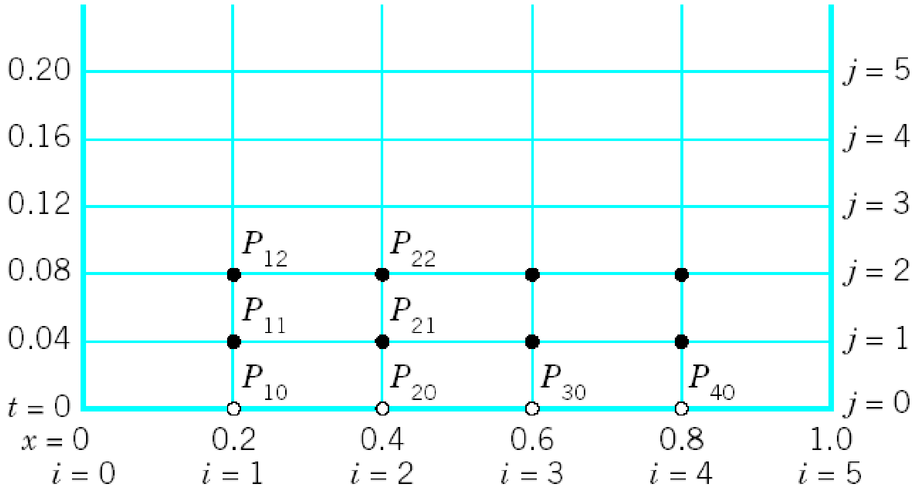


Fig. 467. Grid in Example 1

Temperature in a Metal Bar. Crank–Nicolson Method, Explicit Method

Consider a laterally insulated metal bar of length 1 and such that $c^2 = 1$ in the heat equation. Suppose that the ends of the bar are kept at temperature $u = 0$°C and the temperature in the bar at some instant—call it $t = 0$—is $f(x) = \sin \pi x$. Applying the Crank–Nicolson method with $h = 0.2$ and $r = 1$, find the temperature $u(x, t)$ in the bar for $0 \leqq t \leqq 0.2$. Compare the results with the exact solution. Also apply (5) with an r satisfying (6), say, $r = 0.25$, and with values not satisfying (6), say, $r = 1$ and $r = 2.5$.

Solution by Crank–Nicolson. Since $r = 1$, formula (8) takes the form (9). Since $h = 0.2$ and $r = k/h^2 = 1$, we have $k = h^2 = 0.04$. Hence we have to do 5 steps. Figure 467 shows the grid. We shall need the initial values

$$u_{10} = \sin 0.2\pi = 0.587\,785, \qquad u_{20} = \sin 0.4\pi = 0.951\,057.$$

EXAMPLE 1

Also, $u_{30} = u_{20}$ and $u_{40} = u_{10}$. (Recall that u_{10} means u at P_{10} in Fig. 467, etc.) In each time row in Fig. 467 there are 4 internal mesh points. Hence in each time step we would have to solve 4 equations in 4 unknowns. But since the initial temperature distribution is symmetric with respect to $x = 0.5$, and $u = 0$ at both ends for all t, we have $u_{31} = u_{21}$, $u_{41} = u_{11}$ in the first time row and similarly for the other rows. This reduces each system to 2 equations in 2 unknowns. By (9), since $u_{31} = u_{21}$ and $u_{01} = 0$, for $j = 0$ these equations are

$$\begin{aligned} &(i = 1) \qquad 4u_{11} - \quad u_{21} \qquad\quad = u_{00} + u_{20} = 0.951\,057 \\ &(i = 2) \qquad -u_{11} + 4u_{21} - u_{21} = u_{10} + u_{20} = 1.538\,842. \end{aligned}$$

The solution is $u_{11} = 0.399\,274$, $u_{21} = 0.646\,039$. Similarly, for time row $j = 1$ we have the system

$$\begin{aligned} &(i = 1) \qquad 4u_{12} - \quad u_{22} = u_{01} + u_{21} = 0.646\,039 \\ &(i = 2) \qquad -u_{12} + 3u_{22} = u_{11} + u_{21} = 1.045\,313. \end{aligned}$$

The solution is $u_{12} = 0.271\,221$, $u_{22} = 0.438\,844$, and so on. This gives the temperature distribution (Fig. 468):

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.00	0	0.588	0.951	0.951	0.588	0
0.04	0	0.399	0.646	0.646	0.399	0
0.08	0	0.271	0.439	0.439	0.271	0
0.12	0	0.184	0.298	0.298	0.184	0
0.16	0	0.125	0.202	0.202	0.125	0
0.20	0	0.085	0.138	0.138	0.085	0

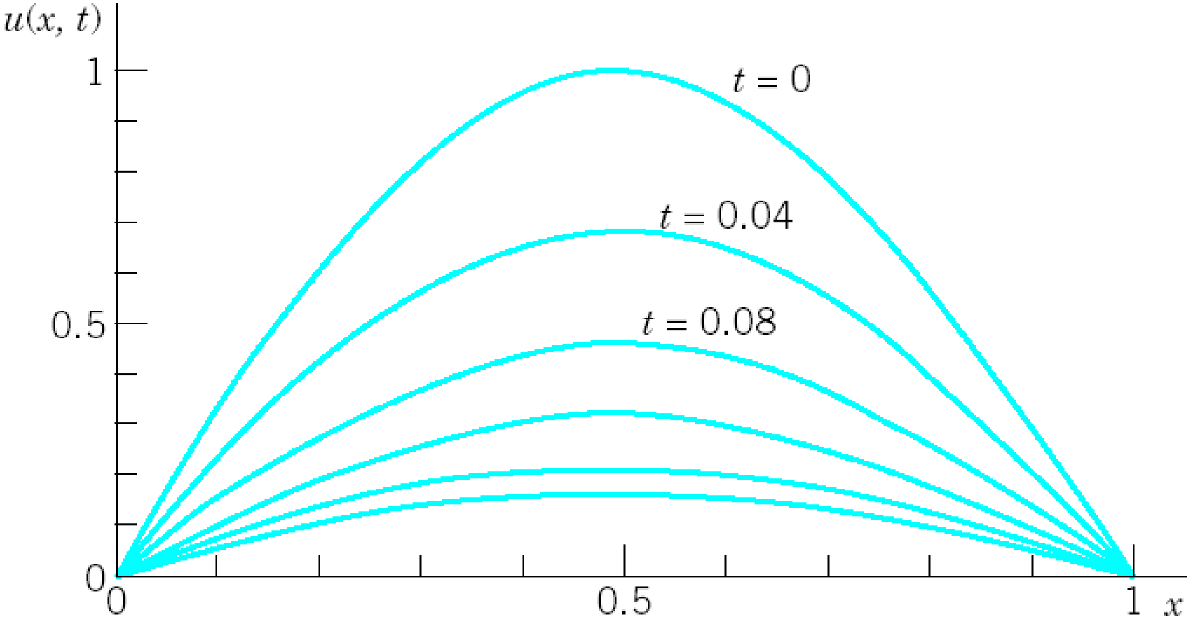


Fig. 468. Temperature distribution in the bar in Example 1

Comparison with the exact solution. The present problem can be solved exactly by separating variables (Sec. 12.5); the result is

$$u(x, t) = \sin \pi x \, e^{-\pi^2 t}. \tag{10}$$

In this section we consider the numeric solution of problems involving hyperbolic PDEs. We explain a standard method in terms of a typical setting for the prototype of a hyperbolic PDE, the **wave equation:**

(1) $$u_{tt} = u_{xx} \qquad 0 \leqq x \leqq 1,\ t \geqq 0$$

(2) $$u(x, 0) = f(x) \qquad \text{(Given initial displacement)}$$

(3) $$u_t(x, 0) = g(x) \qquad \text{(Given initial velocity)}$$

(4) $$u(0, t) = u(1, t) = 0 \qquad \text{(Boundary conditions).}$$

Note that an equation $u_{tt} = c^2 u_{xx}$ and another x-interval can be reduced to the form (1) by a linear transformation of x and t. This is similar to Sec. 21.6, Prob. 1.

Replacing the derivatives by difference quotients as before, we obtain from (1) [see (6) in Sec. 21.4 with $y = t$]

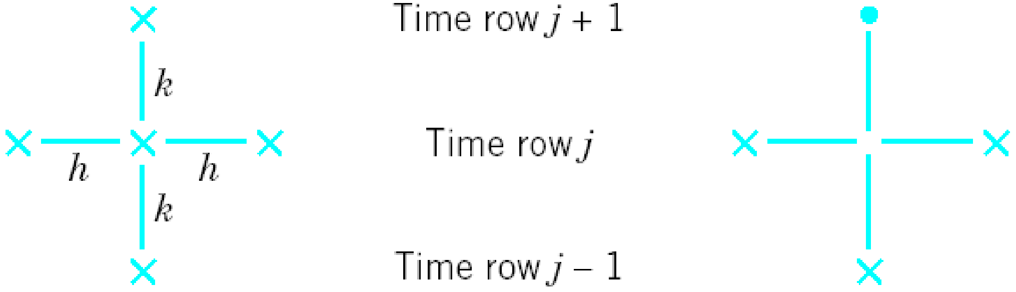
$$\frac{1}{k^2}\,(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) = \frac{1}{h^2}\,(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \tag{5}$$

where h is the mesh size in x, and k is the mesh size in t. This difference equation relates 5 points as shown in Fig. 469a. It suggests a rectangular grid similar to the grids for parabolic equations in the preceding section. We choose $r^* = k^2/h^2 = 1$. Then u_{ij} drops out and we have

(6)
$$u_{i,j+1} = u_{i-1,j} + u_{i+1,j} - u_{i,j-1}$$
(Fig.

It can be shown that for $0 < r^* \leqq 1$ the present **explicit method** is stable, so that from (6) we may expect reasonable results for initial data that have no discontinuities. (For a

Equation (6) still involves 3 time steps $j-1, j, j+1$, whereas the formulas in the parabolic case involved only 2 time steps. Furthermore, we now have 2 initial conditions.



(a) Formula (5)

(b) Formula (6)

Fig. 469. Mesh points used in (5) and (6)

So we ask how we get started and how we can use the initial condition (3). This can be done as follows.

From $u_t(x, 0) = g(x)$ we derive the difference formula

$$\frac{1}{2k}\,(u_{i1} - u_{i,-1}) = g_i, \qquad \text{hence} \qquad u_{i,-1} = u_{i1} - 2kg_i \tag{7}$$

where $g_i = g(ih)$. For $t = 0$, that is, $j = 0$, equation (6) is

$$u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i,-1}.$$

Into this we substitute $u_{i,-1}$ as given in (7). We obtain $u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i1} + 2kg_i$ and by simplification

$$u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}) + kg_i. \tag{8}$$

This expresses u_{i1} in terms of the initial data. It is for the beginning only. Then use (6).

EXAMPLE 1

Vibrating String, Wave Equation

Apply the present method with $h = k = 0.2$ to the problem (1)–(4), where

$$f(x) = \sin \pi x, \qquad g(x) = 0.$$

Solution. The grid is the same as in Fig. 467, Sec. 21.6, except for the values of t, which now are 0.2, 0.4, · · · (instead of 0.04, 0.08, · · ·). The initial values u_{00}, u_{10}, · · · are the same as in Example 1, Sec. 21.6. From (8) and $g(x) = 0$ we have

$$u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}).$$

From this we compute, using $u_{10} = u_{40} = \sin 0.2\pi = 0.587\,785$, $u_{20} = u_{30} = 0.951\,057$,

$$(i = 1) \quad u_{11} = \tfrac{1}{2}(u_{00} + u_{20}) = \tfrac{1}{2} \cdot 0.951\,057 = 0.475\,528$$

$$(i = 2) \quad u_{21} = \tfrac{1}{2}(u_{10} + u_{30}) = \tfrac{1}{2} \cdot 1.538\,842 = 0.769\,421$$

and $u_{31} = u_{21}$, $u_{41} = u_{11}$ by symmetry as in Sec. 21.6, Example 1. From (6) with $j = 1$ we now compute, using $u_{01} = u_{02} = \cdots = 0$,

$$(i = 1) \qquad u_{12} = u_{01} + u_{21} - u_{10} = 0.769\,421 - 0.587\,785 \qquad\qquad = 0.181\,636$$

$$(i = 2) \qquad u_{22} = u_{11} + u_{31} - u_{20} = 0.475\,528 + 0.769\,421 - 0.951\,057 = 0.293\,892,$$

and $u_{32} = u_{22}$, $u_{42} = u_{12}$ by symmetry; and so on. We thus obtain the following values of the displacement $u(x, t)$ of the string over the first half-cycle:

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.0	0	0.588	0.951	0.951	0.588	0
0.2	0	0.476	0.769	0.769	0.476	0
0.4	0	0.182	0.294	0.294	0.182	0
0.6	0	−0.182	−0.294	−0.294	−0.182	0
0.8	0	−0.476	−0.769	−0.769	−0.476	0
1.0	0	−0.588	−0.951	−0.951	−0.588	0

These values are exact to 3D (3 decimals), the exact solution of the problem being (see Sec. 12.3)

$$u(x, t) = \sin \pi x \cos \pi t.$$

The reason for the exactness follows from d'Alembert's solution (4), Sec. 12.4. (See Prob. 4, below.)