MATHEMATICS OF COMPUTATION, VOLUME 24, NUMBER 111, JULY, 1970

On the Numerical Solution of the Diffusion Equation

By Øystein Tødenes

Abstract. A proof given by C. E. Pearson [1] for the asymptotic convergence of the numerical solution of the diffusion equation is discussed, and found insufficient. A new, direct proof is given. A method given by Pearson, for improving the numerical solution when a discontinuity is present in the initial-boundary conditions, is considered in more detail.

1. Introduction. C. E. Pearson [1] has studied the numerical solution of the system:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \tag{1.a}$$

$$u(x,0) = 0, \qquad 0 \leqq x \leqq 1, \tag{1.b}$$

$$\left.\begin{aligned} u(0,t) &= 1 \\ u(1,t) &= 0 \end{aligned}\right\} \quad t > 0, \tag{1.c, 1.d}$$

with special emphasis on the effect of the discontinuity at the point (0, 0).

The system is approximated with the well-known formula

$$u_{j,k+1} - u_{j,k} = \rho\{\theta(u_{j+1,k+1} - 2u_{j,k+1} + u_{j-1,k+1}) + (1-\theta)(u_{j+1,k} - 2u_{j,k} + u_{j-1,k})\}, \tag{2.a}$$

where $\rho = \Delta t/\Delta x^2$, $\Delta x = 1/M$, and $\theta = \frac{1}{2}$, and

$$u_{j,0} = 0, \qquad j = 1, 2, \cdots, M-1, \tag{2.b}$$

$$u_{0,0} = P, \qquad u_{0,k} = 1, \qquad k = 1, 2, \cdots, \tag{2.c}$$

$$u_{M,k} = 0, \qquad k = 0, 1, 2, \cdots. \tag{2.d}$$

The determination of the optimum value of P is Pearson's main problem. The problem is divided into two parts:

(a) Showing that the effect of P diminishes as k increases, this leads to a consideration of the asymptotic convergence.

(b) The value of P is then determined to satisfy an accuracy criterion for "small" k.

2. Asymptotic Convergence. This problem is treated by Pearson as follows: An exact solution of Eq. (1.a) with initial condition

$$u(x,0) = 0, \qquad x > 0, \tag{3}$$

Received August 29, 1969.

AMS Subject Classifications. Primary 6568, 3507; Secondary 3516, 3538.

Key Words and Phrases. Asymptotic convergence, stationary solution, singular boundary conditions.

and boundary condition (1.c) (semi-infinite rod), is found to be

$$u(x, t) = \operatorname{erfc}(x/2\sqrt{t}).$$

The first two terms of the asymptotic expansion of erfc $(x/2\sqrt{t})$, when $t \to \infty$, is

$$\operatorname{erfc}\left(\frac{x}{2\sqrt{t}}\right) \underset{t\to\infty}{\sim} 1 - \frac{x}{(\pi t)^{1/2}},$$

i.e..

$$u(x, t) \underset{t\to\infty}{\sim} 1 - \frac{x}{(\pi t)^{1/2}}. \tag{4}$$

Secondly, a solution of the discrete analogue of the system (1.a), (3), (1.c), i.e., of (2.a), with the initial condition

$$u_{j,0} = 0, \qquad j = 1, 2, \cdots, \tag{5}$$

and boundary condition (2.c), is found by a sine transform. The first two terms of the asymptotic expansion of this solution, when $k \to \infty$, is found (by a steepest descents technique) to be:

$$u_{j,k} \underset{k\to\infty}{\sim} 1 - \frac{j\Delta x}{(\pi k \Delta t)^{1/2}}. \tag{6}$$

From (4) and (6) Pearson concludes "$\cdots$ that the numerical solution given by Eq. (2) (here Eq. (2.a)) will normally become asymptotically correct, as n (here k) grows."

From Pearson's paper, it is evident that the problem he wants to solve is whether the solution of the *system* (2) converges asymptotically to the solution of the *system* (1). The changes of the initial and boundary conditions, represented by (3) and omission of (1.d), and by (5) and omission of (2.d), done for the sake of computational convenience, also results in changes of the respective solutions.

The boundary conditions have a decisive effect on the asymptotic convergence, see, e.g., Parker and Crank [3], therefore, one cannot from

$$\lim_{k\to\infty} u_{j,k} = \lim_{k\to\infty} u(j\Delta x, k\Delta t),$$

where $u_{j,k}$ is the solution of the system (2.a), (5), (2.c), and $u(x, t)$ is the solution of the system (1.a), (3), (1.c), conclude that a similar relation holds with an arbitrary other set of boundary conditions. Whether such a conclusion holds for *special* changes on the boundary conditions is an open, unanswered question. Therefore, the proof for the asymptotic convergence of the solution of the system (2) to the solution of the system (1) is not sufficient.

The problem can be settled in a direct way, as follows: From Carslaw and Jaeger [2, p. 99], we find the solution of the system (1)

$$u(x, t) = 1 - x - \frac{2}{\pi} \sum_{i=1}^{\infty} \frac{1}{i} \exp(-i^2\pi^2 t) \sin i\pi x \tag{7}$$

and the stationary solution, $U(x)$, is

$$U(x) = \lim_{t\to\infty} u(x, t) = 1 - x. \tag{7.a}$$

The system (2) may be written in matrix vector notation as follows:

$$A\mathbf{u}_1 = B\mathbf{u}_0 + \mathbf{b}' + \mathbf{b},$$

$$A\mathbf{u}_{k+1} = B\mathbf{u}_k + \mathbf{b}, \qquad k = 1, 2, \cdots,$$

where

$$A = I + \rho\theta Q_{M-1}, \qquad B = I - \rho(1-\theta)Q_{M-1},$$

$$Q_M = \begin{Bmatrix} -2 & 1 & 0 & 0 & \cdots\cdots\cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots\cdots\cdots & \\ 0 & & & & & 0 \\ \vdots & & & & 1 \quad -2 & 1 \\ 0 & \cdots\cdots\cdots\cdots & & & 0 \quad 1 & -2 \end{Bmatrix} \quad (M \times M),$$

$$\mathbf{b}' = \rho\theta(P - 1, 0, \cdots, 0)^T,$$

$$\mathbf{b} = \rho(1, 0, \cdots, 0)^T,$$

$$\mathbf{u}_k = (u_{1,k}, u_{2,k}, \cdots, u_{M-1,k})^T.$$

It should be pointed out that P is introduced in the system by $\mathbf{b}'$.

The eigenvalues of Q_{M-1}:

$$\lambda_n(Q_{M-1}) = 4\sin^2(n\pi/2M), \qquad n = 1, 2, \cdots, M-1.$$

The eigenvalues of A:

$$\lambda_n(A) = 1 + \rho\theta 4\sin^2(n\pi/2M) \neq 0, \qquad n = 1, 2, \cdots, M-1,$$

i.e., A is nonsingular and A^{-1} exists. By repeated substitution, we get

$$\mathbf{u}_{k+1} = (A^{-1}B)^{k+1}\mathbf{u}_0 + (A^{-1}B)^k A^{-1}\mathbf{b}' + \sum_{i=0}^{k} (A^{-1}B)^i A^{-1}\mathbf{b}. \tag{8}$$

By consideration of eigenvalues, the matrix $(A^{-1}B - I)$ is easily found to be nonsingular, and so $(A^{-1}B - I)^{-1}$ exists. From (8):

$$\mathbf{u}_{k+1} = (A^{-1}B)^{k+1}\mathbf{u}_0 + (A^{-1}B)^k A^{-1}\mathbf{b}' + \frac{(A^{-1}B)^{k+1} - I}{(A^{-1}B - I)} A^{-1}\mathbf{b}, \tag{9}$$

$(A^{-1}B)^T = (A^{-1}B)$, thus, the spectral radius

$$\sigma(A^{-1}B) = \max_n |\lambda_n(A^{-1}B)|,$$

$$\sigma(A^{-1}B) = \max_n \left| \frac{1 - \rho(1-\theta)4\sin^2 \dfrac{n\pi}{2M}}{1 + \rho\theta 4\sin^2 \dfrac{n\pi}{2M}} \right| < 1$$

when $\frac{1}{2} \leqq \theta \leqq 1$.

Thus, when $\frac{1}{2} \leqq \theta \leqq 1$, the solution of (8) converges to the stationary solution, $\mathbf{U}$,

$$\mathbf{U} = \lim_{k\to\infty} \mathbf{u}_k = -(A^{-1}B - I)A^{-1}\mathbf{b} = \frac{1}{\rho} Q_{M-1}^{-1}\mathbf{b}.$$

Since Q_{M-1} is nonsingular, there is a one-one correspondence between $\mathbf{U}$ and $\mathbf{b}$, and the equation above is written:

$$\rho Q_{M-1}\mathbf{U} = \mathbf{b}.$$

This constitutes a set of difference equations that is easily solved with respect to the components of $\mathbf{U}$, giving

$$\mathbf{U} = (1 - \Delta x, 1 - 2\Delta x, \cdots, 1 - (M - 1)\Delta x)^T.$$

This is seen to be the discrete analogue of the stationary solution of the system (1), given by (7.a), and so, the asymptotic convergence is proved.

3. Solution Near the Singularity. The solution of the difference system (2), given by (9), shall now be developed in more detail by an eigenvector expansion, and is, at first, written in the following form:

$$\mathbf{u}_k = (A^{-1}B)^k(\mathbf{u}_0 - \mathbf{U}) + (A^{-1}B)^{k-1}A^{-1}\mathbf{b}' + \mathbf{U}.$$

The eigenvectors of Q_{M-1}, $\mathbf{y}_n$,

$$\mathbf{y}_n = \left(\sin\frac{n\pi}{M}, \sin 2\frac{n\pi}{M}, \cdots, \sin(M-1)\frac{n\pi}{M}\right)^T,$$

constitute a basis.

The expansion involves the summation of the series $\sum_{n=1}^{M-1} n \sin n\beta$, which is performed by integration with respect to β, complex representation of $\cos n\beta$, summation of geometric series, and differentiation with respect to β.

$$u_{j,k}(P) = 1 - j\Delta x - \Delta x \sum_{i=1}^{M-1} \left[\frac{1 - \rho(1-\theta)4\sin^2\frac{i\pi}{2M}}{1 + \rho\theta 4\sin^2\frac{i\pi}{2M}}\right]^k \cot g\frac{i\pi}{2M}\sin j\frac{i\pi}{M}$$

(10)

$$+ 2\rho\theta(P-1)\Delta x \sum_{i=1}^{M-1} \frac{\left(1 - \rho(1-\theta)4\sin^2\frac{i\pi}{2M}\right)^{k-1}}{\left(1 + \rho\theta 4\sin^2\frac{i\pi}{2M}\right)^k} \sin\frac{i\pi}{M}\sin j\frac{i\pi}{M},$$

where $u_{j,k}$ is written as a function of P. We note the analogy between this formula, when $P = 1$, and the formula (7). The optimum value, P_1, of the parameter P, is determined by Pearson by minimizing the error vector

$$||\tilde{\mathbf{u}}_1 - \mathbf{u}_1(P_1)|| = \min_P ||\tilde{\mathbf{u}}_1 - \mathbf{u}_1(P)||, \tag{11}$$

where the Euclidean norm is taken, and

$$\tilde{\mathbf{u}}_k = (u(\Delta x, k\Delta t), u(2\Delta x, k\Delta t), \cdots, u((M-1)\Delta x, k\Delta t))^T.$$

Actually, what is done is introducing a parameter P into the numerical solution and choosing some optimum value for it. The parameter is given a strong connection to the singular point (0, 0). Thinking of the many other ways a correcting parameter could be introduced, it is not a priori evident that the connection between P and the point (0, 0) is specially attractive.

Formula (10) reveals that the parameter results in an additive "correction term". To discuss the features of this term in more detail, define

$$P_{j,k}, \qquad j = 1, 2, \cdots, M-1, \qquad k = 1, 2, \cdots,$$

by

$$u_{j,k}(P_{j,k}) = u(j\Delta x, k\Delta t).$$

(The "best way" of introducing a correcting parameter, P, would be one for which $P_{j,k}$ is constant.)

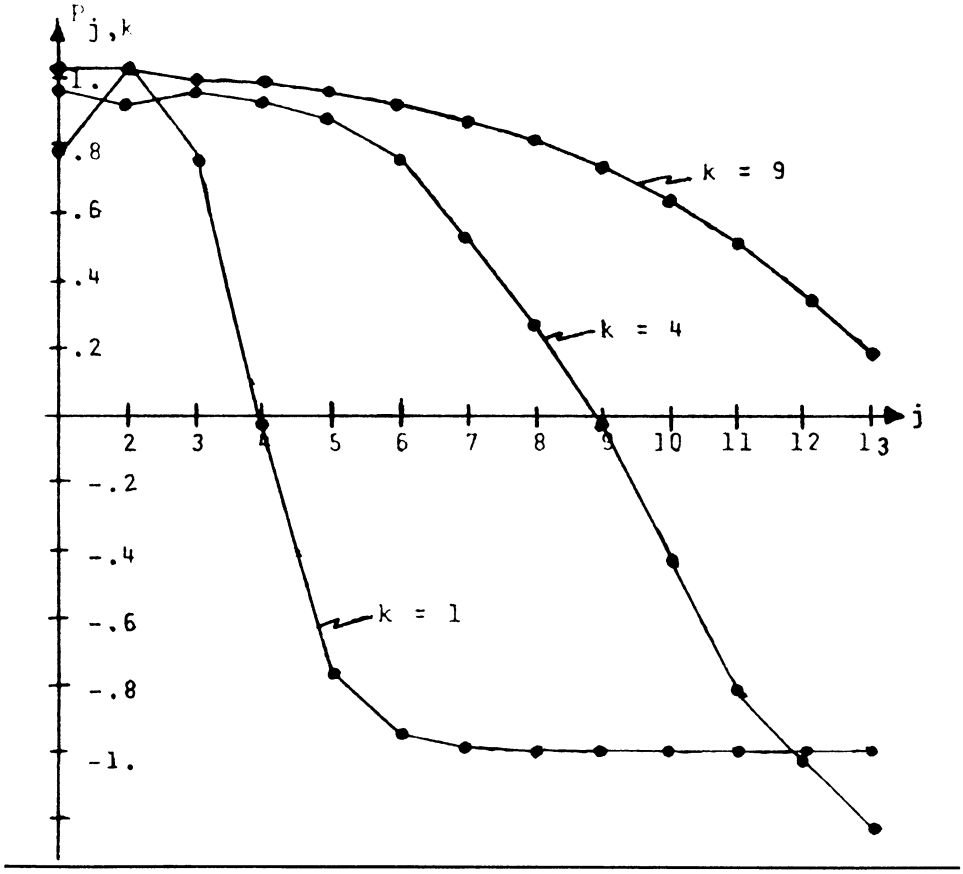
In analogy to (11), we defined P_k, $k = 1, 2, \cdots$, by

$$||\tilde{u}_k - u_k(P_k)|| = \min_P ||\tilde{u}_k - u_k(P)||.$$

The distribution of $P_{j,k}$, $k = 1, 4, 9$, $j = 1, 2, \cdots, 13$,

$$\Delta x = \tfrac{1}{100}, \qquad \Delta t = \tfrac{1}{10000}, \qquad \theta = \tfrac{1}{2},$$

is reported roughly on Fig. 1.



$\Delta x = \frac{1}{100}$, $\Delta t = \frac{1}{10000}$, $\theta = \frac{1}{2}$

FIGURE 1

From Fig. 1, it is seen that the inequality, $0 \leqq P_{i,k} \leqq 1$, is not fulfilled (contrary to Pearson's supposition), and that $P_{i,k}$ is far from being constant.

k	1	2	3	4	5	$\cdots$	9
P_k	0.816	0.889	0.917	0.928	0.935	$\cdots$	0.945

TABLE 1

$\Delta x = \frac{1}{100}, \quad \Delta t = \frac{1}{10000}, \quad \theta = \frac{1}{2}$

From the table above, it is seen that P_k is not constant, i.e., the value of P that minimizes the Euclidean norm of the error vector for $k = 1$, does not minimize the norms of the error vectors for $k = 2, 3, \cdots$. (This is, of course, not surprising.) However, noting that the multiplicative term, $g_{i,k}$, in (10),

$$g_{j,k} = \Delta x \sum_{i=1}^{M-1} \frac{\left(1 - \rho(1-\theta)4 \sin^2 \frac{i\pi}{2M}\right)^{k-1}}{\left(1 + p\theta 4 \sin^2 \frac{i\pi}{2M}\right)^{k}} \sin \frac{i\pi}{2M} \sin j \frac{i\pi}{2M},$$

diminishes as k and/or j increase, the value of P should become less important as k and/or j grow. To investigate this more closely, the effect of choosing $P = 1$ (the a priori most "natural" choice) is compared to the effect of choosing $P = P_1$. Define $\epsilon_{j,k}(P)$,

$$\epsilon_{j,k}(P) = u(j\Delta x, k\Delta t) - u_{j,k}(P) = 2\rho\theta g_{j,k}(P_{j,k} - P).$$

The results for $k = 4$ are reported in the following table.

j	$\epsilon_{j,4}(P_1)$	$\epsilon_{j,4}(1)$
1	0.0027	−0.0008
2	0.0040	−0.0020
3	0.0043	−0.0017
4	0.0031	−0.0023
5	0.0011	−0.0022
6	−0.0004	−0.0022
7	−0.0013	−0.0022
8	−0.0011	−0.0016
9	−0.0007	−0.0009
10	−0.0004	−0.0005
11	−0.0002	−0.0002

TABLE 2

$\Delta x = \frac{1}{100}, \quad \Delta t = \frac{1}{10000}, \quad \theta = \frac{1}{2}$

In the case above, it would be rather hard to decide what value of P should be chosen. For $k > 4$, $P = 1$ would, in a way, be better than $P = P_1$, whereas, for $k = 1, 2, 3$, $P = P_1$ would, perhaps, be better than $P = 1$. Similar results hold for other values of Δx and Δt.

In conclusion, it should be difficult to find an adequate criterion under which $P = P_1$, *generally*, is a better value than, for example, $P = 1$.

The calculation of $P_{j,k}$ and P_k were carried out on the IBM/360 of the University of Bergen, using double-precision FORTRAN programs.

Universitetet i Bergen
Avd. for elektronisk databehandling
Lars Hillesgt. 19
5000 Bergen, Norway

1. C. E. PEARSON, "Impulsive end condition for diffusion equation," *Math. Comp.*, v. 19, 1965, pp. 570–576. MR **33** #1980.
2. H. S. CARSLAW & J. C. JAEGER, *Conduction of Heat in Solids*, Oxford Univ. Press, London, 1959.
3. I. B. PARKER & J. CRANK, "Persistent discretization errors in partial differential equations of parabolic type," *Comput. J.*, v. 7, 1964, pp. 163–167. MR **32** #608.