JOURNAL OF COMPUTATIONAL PHYSICS 51, 357–363 (1983)

Note

Nonpolynomial Finite Difference Schemes and the Use of the Fast Fourier Transform

1. Introduction

Several authors have considered various forms of difference approximations to partial differential equations in which the coefficients in the finite difference equations involve the exponential function. Such schemes were called "unified" difference schemes by Roscoe [1] and probably the first method of this type was formulated by Allen and Southwell [2]. Their method was analysed by Dennis [3] and applications and extensions may be found in Allen [4], Dennis *et al.* [5], Dennis [6], Roscoe [1], and Dennis and Hudson [7, 8]. These authors have concentrated on the use of such schemes with iterative methods, particularly successive overrelaxation.

In this note we show that, for certain problems, the systems of linear equations generated from such difference approximations may be solved directly by FFT techniques. Computational results are presented for a model problem for various grid sizes and these results are compared with those obtained using standard difference formulae based on local polynomial approximations.

2. The Model Problem

The general approach is sufficiently illustrated by considering a function $\phi(x, y)$ which satisfies an equation of the form

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + 2p(y)\frac{\partial \phi}{\partial y} = q(x, y), \tag{1}$$

over a rectangular region R. An equation of a form similar to (1) with p constant and $q = 0$ was considered by Allen [4] in a determination of the temperature distribution near to a sliding contact.

The derivation of the difference scheme (Section 3) is valid for general p and q, but in order to solve the resulting equations by FFT methods p must be, at most, a function of y. (See, for example, Le Bail [9], who also gives a broader classification of equations which may be solved by FFT methods using standard polynomial based approximations.)

0021-9991/83 $3.00

We consider in detail the case

$$p(y) = -\frac{3}{1.1 - y}, \tag{2}$$

and

$$q(x, y) = \frac{3\sqrt{xy}}{(3 - x)(2 - y)}, \tag{3}$$

with R the unit square $0 \leqslant x, y \leqslant 1$. For simplicity we impose Dirichlet boundary conditions given by

$$\phi(0, y) = \phi(1, y) = \phi(x, 0) = 0, \qquad \text{and} \qquad \phi(x, 1) = 4x(1 - x). \tag{4}$$

As is well known, FFT methods may be used also with Neumann or periodic conditions.

Some comments on our choice of equation are relevant here. It is of course the case that, for a given equation, if one has some a priori knowledge of the expected form of its exact solution then this can help to determine the most suitable form of approximation to employ in the construction of a numerical solution. Thus, in this way the use of a particular type of approximation may yield good results with one equation but not with another. It is unlikely that Eq. (1) with the above choice of p, q, and boundary conditions has such an intrinsic bias in favour of either exponential or polynomial based approximations.

One feature of our chosen equation is that near to the boundary $y = 1$ relatively large values (compared with unity) of $p(y)$ are generated. It is well established in [1] that for elliptic equations with such relatively large first derivative coefficients, the particular type of exponentially based approximation employed there produces an improved form of matrix representation of the differential equation compared with that derived using polynomial approximations. For our example we have made use of finite difference equations of the form given in [1] since these are applicable to a general second-order elliptic equation and hence the results obtained are likely to be indicative of what might be achieved for a wider class of problems.

Numerical solutions to the above problem have been determined for five different grid sizes using the finite difference approximation derived in the following section and the standard five-point central difference approximation to (1). In view of the earlier remarks we anticipate that the results obtained using the exponential form of approximation will show some advantage over those obtained using polynomial approximations. This does indeed turn out to be the case and details are given in Section 4.

3. The Finite Difference Equations

Detailed arguments supporting the use of the form of approximation described below may be found in [1]. Essentially the motivation is based on the idea of looking for difference equations whose solutions are locally identical to those of differential equations. For ordinary differential equations the application of these ideas is straightforward but for elliptic problems Roscoe [1] has shown that a difference representation with the required property would contain an infinite number of terms. Therefore he advocates the use of an alternative approximate procedure which involves splitting the differential equation into two parts, one containing all the x derivatives and the second all the y derivatives. The methods developed for ordinary differential equations are then applied to each part separately. The idea of splitting the equation and treating each part in the above fashion was described earlier by Allen [4] and his form of approximation turns out to be identical with that derived and analysed in [1].

We rewrite Eq. (1) in the form

$$\frac{\partial^2 \phi}{\partial x^2} = \alpha(x, y), \tag{5}$$

so that

$$\frac{\partial^2 \phi}{\partial y^2} + 2p(y)\frac{\partial \phi}{\partial y} = q(x, y) - \alpha(x, y), \tag{6}$$

and at a mesh point (x_i, y_j) we solve Eq. (6) as if it were an ordinary differential equation to obtain the local approximation

$$\phi = Ae^{-2p_j y} + B + (q_{i,j} - \alpha_{i,j})\, y/2p_j, \tag{7}$$

where A and B are arbitrary constants. By setting $y = y_{j-1}$, y_j and y_{j+1} in turn, equations for $\phi_{i,j-1}$, $\phi_{i,j}$, and $\phi_{i,j+1}$ are obtained from (7) and A and B may thus be eliminated. From the resulting relation, $\alpha_{i,j}$ may be eliminated using (5) with a standard polynomial approximation for ϕ_{xx}. The difference representation of (6) based on (7) reduces to the standard polynomial approximation in the limit $p \to 0$. Thus at (x_i, y_j) for a square mesh of side h we obtain

$$\phi_{i-1,j} - (2 + a_j + b_j)\,\phi_{i,j} + \phi_{i+1,j} + a_j\phi_{i,j-1} + b_j\phi_{i,j+1} = h^2 q_{i,j}, \tag{8}$$

where

$$a_j = c_j/(e^{c_j} - 1), \qquad b_j = a_j\, e^{c_j}, \tag{9}$$

and $c_j = 2p_j h$. Applying Roscoe's general finite difference scheme [his Eq. (7.3)] to our particular problem produces an equation which is identical to (8).

The complete set of equations (8) for $i = 1, 2,..., N$, $j = 1, 2,..., N$ may be written in the matrix form as

$$\begin{aligned} A_1\boldsymbol{\phi}_1 + b_1\boldsymbol{\phi}_2 &= \mathbf{Q}_1 - a_1\boldsymbol{\phi}_0, \\ a_j\boldsymbol{\phi}_{j-1} + A_j\boldsymbol{\phi}_j + b_j\boldsymbol{\phi}_{j+1} &= \mathbf{Q}_j \qquad (j = 2, 3,..., N-1), \\ a_N\boldsymbol{\phi}_{N-1} + A_N\boldsymbol{\phi}_N &= \mathbf{Q}_N - b_N\boldsymbol{\phi}_{N+1}, \end{aligned} \tag{10}$$

where $\boldsymbol{\phi}_j = (\phi_{1,j}, \phi_{2,j},..., \phi_{N,j})^{\mathrm{T}}$, A_j is the $N \times N$ matrix

$$A_j = \begin{pmatrix} -(2 + a_j + b_j) & 1 & & & \\ 1 & -(2 + a_j + b_j) & 1 & & \\ & 1 & -(2 + a_j + b_j) & 1 & \\ & & 1 & -(2 + a_j + b_j) & \end{pmatrix}, \tag{11}$$

$\mathbf{Q}_j = h^2(q_{1,j}, q_{2,j},..., q_{N,j})^{\mathrm{T}}$, and $\boldsymbol{\phi}_0$, $\boldsymbol{\phi}_{N+1}$ are, respectively, known vectors of ϕ values on $j = 0, N + 1$.

By expanding $\boldsymbol{\phi}_j$ in terms of the known (orthogonal) eigenvectors $\mathbf{x}_s$ of (11), so that

$$\boldsymbol{\phi}_j = \sum_{s=1}^{N} \bar{\phi}_{s,j}\mathbf{x}_s, \tag{12}$$

we obtain systems of equations for the Fourier harmonics $\bar{\phi}_{s,j}$ in the usual form of N decoupled tridiagonal systems each of order N. A typical system is of the form

$$\begin{aligned} \lambda_{s,1}\bar{\phi}_{s,1} + b_1\bar{\phi}_{s,s} &= \bar{d}_{s,1}, \\ a_j\bar{\phi}_{s,j-1} + \lambda_{s,j}\bar{\phi}_{s,j} + b_j\bar{\phi}_{s,j+1} &= \bar{d}_{s,j} \qquad (j = 2, 3,..., N-1), \\ a_N\bar{\phi}_{s,N-1} + \lambda_{s,N}\bar{\phi}_{s,N} &= \bar{d}_{s,N}, \end{aligned} \tag{13}$$

where $\lambda_{s,j}$ denotes an eigenvalue of A_j and is given by

$$\lambda_{s,j} = -(a_j + b_j) - 4\sin^2\frac{s\pi}{2(N+1)} \qquad (s, j = 1, 2,..., N), \tag{14}$$

and $\bar{d}_{s,j} = \mathbf{x}_s^{\mathrm{T}}\mathbf{Q}_j^*/|\mathbf{x}_s|^2$, where $\mathbf{Q}_1^* = \mathbf{Q}_1 - a_1\boldsymbol{\phi}_0$, $\mathbf{Q}_j^* = \mathbf{Q}_j$, $j = 2, 3,..., N-1$, and $\mathbf{Q}_N^* = \mathbf{Q}_N - b_N\boldsymbol{\phi}_{N+1}$. Since $a_j, b_j > 0$ for all values of p_j, using (14) it is easy o deduce that the usual Gauss elimination (Thomas) algorithm applied to (13) will always be stable with respect to the growth of round-off errors. This is not the case if standard central difference approximations are used; the resulting systems equivalent to (13) are formally stable only for $|p_j| < 1/h$ although round-off propagation was not found to be a serious problem for the range of values of N used on an ICL 1906S machine which has 11 decimal digit accuracy. This point is considered further in the following section. Both the exponential-based and polynomial-based schemes have a leading truncation error $O(h^2)$.

FFT methods can be used to form the right-hand sides $\bar{d}_{s,j}$ and to reconstruct the vectors ϕ_j using (12). The basic form of cyclic reduction which is often used together with Fourier analysis for the soluion of Poisson's equation (Hockney [10], Swartztrauber [11], Temperton [12]) cannot be used on the more general systems (10) since the coefficients in these equations depend on j.

4. Numerical Results and Discussion

Equation (1) was solved over the unit square for five different values of N in the range 7 to 127 using both schemes. It was found that, as N was increased, considerably greater changes took place in the computed solutions based on standard polynomial approximations than the corresponding solutions obtained using the exponential scheme. Thus, for the purposes of comparison, the solution for $N = 127$ based on the exponential scheme was regarded as "accurate" and the values of RMS error and maximum modulus error given in Table I were calculated on this basis. It is clear that in all cases the RMS error for the exponantial scheme is less than the corresponding error for the standard scheme and additionally that the rate of increase in value of the error as N decreases is less for the exponential scheme than for the standard scheme.

Similar remarks apply to the values of maximum modulus error given in Table I. It is worth noting that the major feature of the solution is a relatively large positive peak concentrated in the region $0.85 \leqslant y \leqslant 1$, $0.25 \leqslant x \leqslant 0.75$ with maximum value 1 at the point $(\frac{1}{2}, 1)$. The maximum modulus error was always found to occur in this area. It is perhaps therefore unrealistic to expect worthwhile results for $N = 7$ ($h = 0.125$) and there appears to be no particularly clear explanation for the slightly anomalous value of maximum modulus error produced for this value of N using the exponential scheme. For $N = 15$ the maximum modulus errors of the standard scheme and exponential scheme represent, respectively, percentage errors of approx-

TABLE I

	RMS Error		Max. Mod. Error	
N	Standard	Exponential	Standard	Exponential
7	1.35,–2	9.25,–4	3.24,–1	2.18,–2
15	2.31,–3	3.22,–4	1.77,–1	2.39,–2
31	3.15,–4	6.42,–5	6.52,–2	1.28,–2
63	3.99,–5	7.43,–6	1.72,–2	3.12,–3
127	5.79,–6	—	4.94,–3	—

Note. Values of RMS error and maximum modulus error (derived as explained in the text) for various values of N using both the standard polynomial approximation and an approximation involving the exponential function where, for example, the notation 1.35,–2 means 1.35×10^{-2}.

imately 30% and 4%. The corresponding values for $N = 63$ are approximately 0.5% and 0.1%.

As a check on the accuracy of our computed solutions of the finite difference equations, these solutions were substituted back into the basic difference equations and the maximum modulus residual R_M (over all the grid points) was calculated for each case. For the exponential scheme values of R_M were found to decrease as N decreased, ranging from 1.5×10^{-10} for $N = 127$ to 1.8×10^{-12} for $N = 7$. For the polynomial-based scheme values of R_M were found to be very similar to those for the exponential scheme only as far as $N = 15$, where R_M had the value 2.0×10^{-11} (the value for the exponential scheme was 1.5×10^{-11}). For $N = 7$ the value of R_M was found to be 5.1×10^{-11} which is almost twice as large as its value for $N = 31$ and about thirty times the corresponding value for the exponential scheme. The most probable explanation for this result may be attributed to round-off error propagation effects in the solution of the tridiagonal systems. As was noted in Section 3 the Thomas algorithm as applied to (13) is always stable whereas for the polynomial-based scheme we require $|p_j| < 1/h$. This condition is satisfied for all j for $N = 127,63$, and 31 but for $N = 7$ $(h = \frac{1}{8})$ the maximum value of $|p_j|$ is approximately 13. For $N = 15$ $(h = \frac{1}{16})$ the corresponding value is approximately 18.

A disadvantage of methods based on schemes involving exponentials is that they are likely to involve more computational effort than methods based on standard schemes and this has been confirmed, for a given problem and method, in [8]. These authors were particularly concerned with producing diagonally dominant schemes in order to generate stable iterative methods of solution and they found that one iteration using an exponentially based scheme took approximately twice as long as an iteration using a standard scheme not involving exponentials. By running our programs many times we were able to conclude that typically the method based on the exponential scheme took about 10% more time than that using the standard scheme.

Thus, overall, it seems reasonable to state that, for certain problems, the use of exponential schemes together with FFT techniques can be worthwhile.

REFERENCES

1. D. F. ROSCOE, *J. Inst. Math. Its Appl.* **16** (1975), 291.
2. D. N. DE G. ALLEN AND R. V. SOUTHWELL, *Q. J. Mech. Appl. Math.* **8** (1955), 129.
3. S. C. R. DENNIS, *Q. J. Mech. Appl. Math.* **13** (1960), 487.
4. D. N. DE G. ALLEN, *Q. J. Mech. Appl. Math.* **15** (1962), 11.
5. S. C. R. DENNIS, J. D. HUDSON, AND N. SMITH, *Phys. Fluids* **11** (1968), 933.
6. S. C. R. DENNIS, *Lect. Notes Phys.* **19** (1973), 120.
7. S. C. R. DENNIS AND J. D. HUDSON, *J. Inst. Math. Its Appl.* **23** (1979), 43.
8. S. C. R. DENNIS AND J. D. HUDSON, *J. Inst. Math. Its Appl.* **26** (1980), 369.
9. R. C. LE BAIL, *J. Comput. Phys.* **9** (1972), 440.
10. R. W. HOCKNEY, "Methods in Computational Physics," Vol. 9, pp. 136–210, Academic Press, New York/London, 1970.

11. P. SWARTZTRAUBER, *SIAM Rev.* **19** (1977), 490.
12. C. TEMPERTON, *J. Comput. Phys.* **34** (1980), 314.

RECEIVED: May 12, 1982; REVISED: January 4, 1983

W. M. PICKERING

Department of Applied and Computational Mathematics, The University, Sheffield S10 2TN, England