INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 9, 583–598 (1989)

STABLE SECOND-ORDER ACCURATE ITERATIVE SOLUTIONS FOR SECOND-ORDER ELLIPTIC PROBLEMS

AVI LIN

Department of Mathematics, Temple University, Philadelphia, PA 19122, U.S.A.

SUMMARY

The paper describes a numerical scheme for solving a convection–diffusion elliptic system with very small diffusion coefficients. This iterative numerical procedure is unconditionally stable and converges very rapidly. Although only linear equations are considered here, this technique can be easily extended to non-linear equations, while keeping its main features as for the linear case. The numerical experiments presented are quite general and confirm most of these features. These examples also show a good way of implementing this scheme.

KEY WORDS Upwind Second order Stable schemes

1. INTRODUCTION

The present paper investigates the numerical solutions for an elliptic linear PDE of the convection–diffusion type. Without loss of generality the numerical scheme will be presented for the two-dimensional case, since the extension for higher-dimensional problems is quite straightforward. The following two-dimensional equation for ϕ will be considered in the present study:

$$L(\phi) = u\phi_x + v\phi_y - \varepsilon(\phi_{xx} + \phi_{yy}) - R = 0. \tag{1}$$

This is solved over the two-dimensional domain Ω with boundary $\partial\Omega$. L is a linear two-dimensional partial differential elliptic operator, $u = u(x, y)$ and $v = v(x, y)$ are the convection coefficients, ε is the diffusion coefficient and $R = R(x, y)$ is a known source term. The first and second derivatives of ϕ with respect to x are denoted by ϕ_x and ϕ_{xx} respectively. Similar notations are used for the y-derivatives of ϕ. Equation (1) is subject to certain boundary conditions for ϕ on $\partial\Omega$. When solving this equation by a finite difference (FD) technique, a FD grid is spread over the domain Ω. Then equation (1) is approximated at all the grid points of Ω, which leads to a linear system of equations for the discrete values of Φ, $\bar{\Phi}$ defined over this grid. Numerical treatment of this problem has a long history,[1,2] especially in solving the Navier–Stokes equations.[3] The main problem is how to treat numerically the convection terms in equation (1) in regions of Ω where the convection dominates the diffusion. Basically it is known that central differencing (CD) for these terms may lead to a non-physical oscillatory behaviour or even to a non-converged solution,[4] while upwind differencing is a non-diverging technique but introduces a false diffusion term into the original equation.[3] Some more accurate upwind schemes have been proposed in the past, especially for fluid flows,[5] like the KR method[1] and the Hibrid method,[4] among others. However,

0271–2091/89/050583–16$08.00

Received 8 March 1988
Revised 11 August 1988

their use was found to be very limited, as is discussed in the respective papers. The finite difference approximation of the convection terms is the main subject of the present paper.

2. THE NUMERICAL SET-UP

In order to simplify the presentation of the numerical scheme, let us assume that Ω is a rectangular in the (x, y) plane. The FD grid consists of M discrete points in the x-direction and N points in the y-direction, where the spacings are not necessarily equal to each other. Let us denote by x_i the x-value of the ith discrete point on the x-axis and by y_j the y-value of the jth discrete point on the y-axis. The local intervals in the x- and y-directions are defined as follows (see Figure 1):

$$h_i = x_i - x_{i-1}, \quad 1 < i \leqslant M, \tag{2}$$

$$k_j = y_j - y_{j-1}, \quad 1 < j \leqslant N. \tag{3}$$

The following definitions for the local mesh ratio and the local average interval length are also needed:

$$\sigma_i = \frac{h_{i+1}}{h_i}, \qquad \bar{h}_i = \tfrac{1}{2}(h_{i+1} + h_i) = \tfrac{1}{2}(x_{i+1} - x_{i-1}), \qquad 1 < i < M, \tag{4}$$

$$\tau_j = \frac{k_{j+1}}{k_j}, \qquad \bar{k}_j = \tfrac{1}{2}(k_{j+1} + k_j) = \tfrac{1}{2}(y_{j+1} - y_{j-1}), \qquad 1 < j < N. \tag{5}$$

The FD approximation of equation (1) at the grid point (i, j) is based on the truncated Taylor series for ϕ at this point.[6] Let us denote by $\bar{Q}$ the FD approximation of the quantity Q, where the bars are omitted for ϕ itself, understanding that it is defined at the grid points. We will define the truncation error $T(Q)$ for the approximation $\bar{Q}$ for the quantity Q as

$$\bar{Q} + T(Q) \equiv Q. \tag{6}$$

For the second derivatives for ϕ we will use the standard central difference approximation over a non-uniform grid:[6]

$$\overline{(\Phi_{xx} + \Phi_{yy})}_{i,j} = \frac{\Phi_{i+1,j} - (1+\sigma_i)\Phi_{i,j} + \sigma_i \Phi_{i-1,j}}{[2\sigma_i/(1+\sigma_i)]\bar{h}_i^2} + \frac{\Phi_{i,j+1} - (1+\tau_j)\Phi_{i,j} + \tau_j \Phi_{i,j-1}}{[2\tau_i/(1+\tau_i)]\bar{k}_i^2} + T_{i,j}. \tag{7}$$

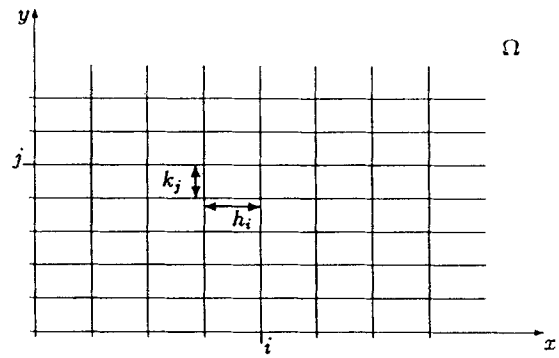

Figure 1. Notations for the FD grid

Here $T_{i,j}$ is the value of the truncation error function T of this approximation at the point (i,j), which is given by

$$T = -\frac{1}{3}\left[\frac{\sigma-1}{\sigma+1}\bar{h}\Phi_{xxx} + \frac{\sigma^3+1}{(\sigma+1)^3}\bar{h}^2\Phi_{xxxx} + \frac{\tau-1}{\tau+1}\bar{k}\Phi_{yyy} + \frac{\tau^3+1}{(\tau+1)^3}\bar{k}^2\Phi_{yyyy}\right]. \tag{8}$$

If the sequences of the FD intervals $\{h_i\}_{i=2}^{M}$ and $\{k_j\}_{j=2}^{N}$ are analytically continuous, i.e. $\sigma_i = 1 + O(h_i)$ and $\tau_j = 1 + O(k_j)$, then the approximation given by equation (7) is of the second order.[7] The FD approximation of the convection terms is the main subject of the present paper and is discussed in the next section.

3. FD MODELS FOR THE CONVECTION

The choice of the FD approximation for the first derivatives of ϕ appearing in equation (1) depends on the way that the final algebraic system of equations is solved. If this system is solved by some direct methods,[8] then second-order accurate solutions may be obtained (when standard second-order CD approximations are used) to any machine accuracy. However, if iterative procedures are used, then most of the solutions will diverge for large convection coefficients. Since direct methods are very limited in their use (mainly because of the special structure of the matrices and the usage of large computer resources), most of the time this system is solved by iterative techniques. The present paper considers mainly iterative techniques for solving the final elliptic FD algebraic equations.

A necessary convergence condition for most of the (explicit and implicit) iteration procedures used to solve this problem is that the algebraic system be diagonal dominant,[3] independent of the way the iterative procedure is introduced. If second-order CD is used, for instance the following first derivative at some inner point ξ:

$$\overline{(\Phi_x)|_\xi} = \frac{\Phi(x_{i+1}, y_j) - \Phi(x_{i-1}, y_j)}{x_{i+1} - x_{i-1}} + O(\bar{h}_i^2), \quad x_{i-1} < \xi < x_{i+1}, \tag{9}$$

then this condition ends up with certain limitations on the so-called local mesh Reynolds number.[3] The local directed mesh Reynolds numbers R^x and R^y are defined as follows:

$$R^x_{p,q} = \frac{|u_{p,q}|\bar{h}_p}{\varepsilon}, \qquad R^y_{p,q} = \frac{|v_{p,q}|\bar{k}_q}{\varepsilon}. \tag{10}$$

When equation (9) is used as the FD approximations for the first derivatives of the convection part of equation (1), together with equation (7) as the FD approximation of the diffusion part, then the convergence is assured if

$$R^x_{i,j},\ R^y_{i,j} \leqslant 2. \tag{11}$$

This way of solving equation (1) is the classical method (CM), and for very large convection coefficients its iteration procedure might not be stable. For these cases most of the numerical approaches suggest using the upwind differencing approach, which may be written at the grid point (i,j) as follows:

$$\overline{(\Phi_x)_{i,j}} = \begin{cases} \dfrac{\Phi_{i,j} - \Phi_{i-1,j}}{h_i}, & \text{if } u_{i,j} \geqslant 0, \\ \dfrac{\Phi_{i+1,j} - \Phi_{i,j}}{h_{i+1}}, & \text{if } u_{i,j} < 0, \end{cases} \tag{12}$$

$$\overline{(\Phi_y)_{i,j}} = \begin{cases} \dfrac{\Phi_{i,j} - \Phi_{i,j-1}}{k_j}, & \text{if } v_{i,j} \geqslant 0, \\ \dfrac{\Phi_{i,j-1} - \Phi_{i,j}}{k_{j+1}}, & \text{if } v_{i,j} < 0. \end{cases} \tag{13}$$

By substituting equation (12), (13) and (7) into equation (1), it can be verified that the diagonal dominance of the algebraic system is preserved and any standard explicit or implicit method will be stable. However, this approach has two basic limitations:

1. The rate of convergence is slowed down dramatically as any of the directed local mesh Reynolds numbers increases.[9] Since this approach creates a very non-symmetric coefficient matrix for the system of equations, the regular accelerating techniques, like the conjugate gradient technique,[10] will help little in increasing the convergence rate.
2. The second limitation, which is much more severe than the first, is the fact that equations (12) and (13) present FD approximations which are accurate only to order one. Because the truncation error here is proportional to $h\phi_{xx}$ and $k\phi_{yy}$, some artificial (numerical) diffusion is added to ε, and equation (1) is solved for a different (and in fact much larger) diffusion coefficient from that which appears in the original equation (1), producing, of course, wrong solutions for the problem.[3]

In the past, several second-order upwind schemes for the convection terms have been proposed, some of them being only marginally stable.[5] One of the most commonly used second-order methods is the KR scheme,[1] which may be written for the ϕ_x term at the grid point (i, j) as follows:

$$\overline{(\Phi_x)_{i,j}} = \begin{cases} \left(\dfrac{\Phi_{i,j} - \Phi_{i-1,j}}{h_i}\right)^{(n)} + C(\Phi_{i,j}^{(n-1)}), & \text{if } u_{i,j} \geqslant 0, \\ \left(\dfrac{\Phi_{i+1,j} - \Phi_{i,j}}{h_{i+1}}\right)^{(n)} - C\left(\dfrac{\Phi_{i,j}^{(n-1)}}{\sigma_i}\right), & \text{if } u_{i,j} < 0, \end{cases} \tag{14}$$

where C is the correction function, defined as:

$$C(\Phi_{i,j}) = \frac{\Phi_{i+1,j} - (1+\sigma_i)\Phi_{i,j} + \sigma_i \Phi_{i-1,j}}{2\bar{h}_i} . \tag{15}$$

Similar expressions may be obtained for $\overline{(\Phi_y)_{i,j}}$. The upper index in equation (14) indicates the iteration number and $C(\Phi)$ is a term which has to be added to the appropriate expressions in equations (12) and (13) in order to make them second-order accurate as given by equations (14) and (15). When using equations (14) and (15), the solutions obtained at each stage n of the iteration procedure are accurate only to the first order, while they are second-order accurate in the converged state. However, this approach suffers from several deficiencies:

1. It is not quite clear how to use this kind of iteration approach in the framework of the other iteration loops used to solve this problem, like the iteration scheme for overcoming the two- or three-dimensional nature of the field, or iterations due to the non-linearity of the system if it exists.
2. As in the first-order upwind case, the rate of convergence decreases dramatically as R^x and R^y increase. This slowing down of the convergence rate of equations (14) and (15) is much more pronounced than that of equations (12) and (13); in fact, the spectral radius of the

iteration procedure expressed by equation (14) is proportional to $1-h^2$, where h is the smallest grid spacing in Ω, when equation (1) is undimensionalized appropriately.[11]

In the present work a new upwind second-order accurate FD approximation for the convection terms will be presented in the spirit of the KR scheme[1] and which is based also on the one-sided extrapolation of the CD approximations for the first derivatives. This is discussed in the following section.

4. SECOND-ORDER UPWIND APPROXIMATION

Let us derive the FD approximation for $(u\Phi_x)_{i,j}$ assuming $u_{i,j} \geqslant 0$ with the following features:

(a) should be order-two accurate
(b) should be unconditionally stable when used with implicit methods.

The final formulation can be extended also for negative values of this coefficient. Let us denote by 1 the point $(i-1/2, j)$, by 2 the point $(i-3/2, j)$ and by 0 the point (i, j) as shown in Figure 2.

Lemma 1. For $f \in \mathbb{C}^2$, the following second-order FD approximation for f_0 exists:

$$f_0 = \left(1 + \frac{\sigma_{i-1}}{1+\sigma_{i-1}}\right) f_1 - \frac{\sigma_{i-1}}{1+\sigma_{i-1}} f_2 + T_r \tag{16}$$

with a truncation error

$$T_r = \frac{1}{4}\left(1 + \frac{1}{2\sigma_i}\right) h_i^2 (f_0)_{xx} + \ldots . \tag{17}$$

Lemma 2. If f_1 and f_2 in Lemma 1 are replaced by $\bar{f}_1$ and $\bar{f}_2$ respectively, then $\bar{f}_0$ given by this lemma is $O(h^2)$ accurate.

Lemma 3. The quantity $(\Phi_x)_{i,j}$ can be approximated to order two by

$$\overline{(\Phi_x)_{i,j}} = \left(1 + \frac{\sigma_{i-1}}{1+\sigma_{i-1}}\right) \frac{\overline{\Phi_{i,j}} - \overline{\Phi_{i-1,j}}}{h_i} - \frac{\sigma_{i-1}}{1+\sigma_{i-1}} \frac{\overline{\Phi_{i-1,j}} - \overline{\Phi_{i-2,j}}}{h_{i-1}}. \tag{18}$$

The above three lemmas can be proved easily. In order to formulate the present upwind technique, let us consider the time-dependent version of equation (1):

$$a\Phi_t + L(\Phi) = 0. \tag{19}$$

The sign of the artificial coefficient a is such that this equation is parabolic in t, where t is the time-like co-ordinate. It is desired to get the steady state solutions for this equation in a stable and rapid manner, beginning from some given initial conditions. Denoting by Δ the time step, and by

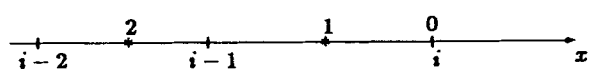

Figure 2. Location of the grid points for the present scheme

n the time step index, and using equations (7) and (16), the following FD approximations are suggested for $u_{i,j} > 0$:

$$\overline{(\Phi_t)}_{i,j} = \frac{\Phi_{i,j}^{(n)} - \Phi_{i,j}^{(n-1)}}{\Delta}, \tag{20}$$

$$\overline{(\Phi_x)}_{i,j} = \left(1 + \frac{\sigma_{i-1}}{1+\sigma_{i-1}}\right)\frac{\Phi_{i,j}^{(n)} - \Phi_{i-1,j}^{(n)}}{h_i} - \frac{\sigma_{i-1}}{1+\sigma_{i-1}}\frac{\Phi_{i-1,j}^{(n-1)} - \Phi_{i-2,j}^{(n-1)}}{h_{i-1}}, \tag{21}$$

$$\overline{(\Phi_{xx})}_{i,j} = \frac{\Phi_{i+1,j}^{(n)} - (1+\sigma_i)\Phi_{i,j}^{(n)} + \sigma_i \Phi_{i-1,j}^{(n)}}{[2\sigma_i/(1+\sigma_i)]\bar{h}_i^2}. \tag{22}$$

This second-order upwind scheme will be referred hereafter as the S method. It is worth noting that the truncation error of this approximation is

$$T_r = \frac{1}{3}\frac{\sigma}{\sigma+1}\bar{h}^2\left[\frac{1+2\sigma}{1+\sigma}\bar{h}\Phi_{xxxx} - 2\Phi_{xxx}\right]. \tag{23}$$

Theorem 1. The solutions for equation (19) using equations (20)–(22) are unconditionally stable for all $n \geqslant 1$.

Proof. Let us use here the Von Neumann analysis, applied for an equally spaced mesh:

$$\sigma_i = \tau_j = 1, \qquad h_i = h, \qquad k_i = k. \tag{24}$$

Denoting

$$\alpha = \frac{p\pi}{M}, \qquad \beta = \frac{q\pi}{N}, \qquad 1 \leqslant p \leqslant M-1, \qquad 1 \leqslant q \leqslant N-1, \tag{25}$$

it is assumed that

$$\Phi_{i,j}^{(n)} = \lambda^n \exp^{ii\alpha} \exp^{ij\beta}, \tag{26}$$

where $i^2 = -1$ and λ is a complex number defined as the amplification factor. Substitution of equation (26) into equations (20)–(22) and then into equation (12) gives the following equation for the absolute value of the amplification factor:

$$|\lambda|^2 = \frac{[1 + c_1(\cos\alpha - \cos 2\alpha) + c_2(\cos\beta - \cos 2\beta)]^2 + [c_1(\sin\alpha - \sin 2\alpha) + c_2(\sin\beta - \sin 2\beta)]^2}{[1 + (d_1 + 6c_1)\sin^2(\alpha/2) + (d_2 + 6c_2)\sin^2(\beta/2)]^2 + 9[c_1\sin\alpha + c_2\sin\beta]^2}, \tag{27}$$

where

$$c_1 = \frac{u\Delta/a}{2h}, \qquad c_2 = \frac{v\Delta/a}{2k}, \qquad d_1 = \frac{4\varepsilon\Delta/a}{h^2}, \qquad d_2 = \frac{4\varepsilon\Delta/a}{k^2} \tag{28}$$

are positive coefficients. The following can be proved:

1. From the fact that

$$\sin\gamma - \sin 2\gamma \leqslant 3\sin\gamma, \quad 0 \leqslant \gamma \leqslant \pi, \tag{29}$$

the second term in the numerator of equation (27) is smaller than the second term in the denominator of that equation.

2. From the fact that

$$|\cos\gamma - \cos 2\gamma| \leqslant 6\sin^2(\gamma/2), \tag{30}$$

the first term in the numerator is smaller than the first term of the denominator, independent of the value of α and β.

Therefore one can get for all p and q in equation (27)

$$|\lambda| \leqslant 1, \tag{31}$$

even for the case where $a/\Delta \to \infty$. □

In fact, the rate of convergence towards the steady state is determined mainly by the short waves of the error propagation, given approximately by

$$|\lambda| \simeq 1 - \tfrac{1}{2}[(d_1/2 + 7c_1)\alpha^2 + (d_2/2 + 7c_2)\beta^2]. \tag{32}$$

Thus if l and U are the respective length scale and convection coefficient scale used to normalized equation (1), then the choice of $\Delta = l/U$ gives the maximum rate of convergence, which is about $3{\cdot}5\pi^2$. In practice, one usually gets a smaller rate of convergence, as will be shown in the examples.

5. BOUNDARY CONDITIONS

Since the governing equation is elliptic, boundary conditions (BC) for Φ should be imposed at *all* the grid points of $\partial\Omega$. Generally, two types of BC may be imposed: the Dirichlet BC and the Neumann BC. In this section we will formulate numerical models for these two types of BC which are suitable and consistent with the suggested numerical FD approximation for the convection part of equation (1). Let us derive these formulations for a boundary located at $x =$ constant, which is a straight line parallel to the y-axis, and where the convection coefficient of the Φ_x term is positive near this boundary. A sketch of the present boundary is given in Figure 3. All other kinds of boundaries and convection coefficients may be treated in a very similar way.

5.1. Dirichlet boundary conditions

The Dirichlet BC are such that the values of Φ are given on the boundary, i.e. the value of Φ at the point 0, Φ_0, is given. The FD approximation for the x-convection term at the point 1 is derived from Lemma 3 to give in the case $u_1 \geqslant 0$

$$\overline{(\Phi_x)_1} = \left(1 + \frac{\sigma_0}{1+\sigma_0}\right)\frac{\overline{\Phi_1} - \overline{\Phi_0}}{h_1} - \frac{\sigma_0}{1+\sigma_0}\frac{\overline{\Phi_0} - \overline{\Phi_{-1}}}{h_0}, \tag{33}$$

where the point -1 is an artificial point located at a distance $h_0 = h_1$ behind the boundary (with $\sigma_0 = 1$). The value of Φ_{-1} is obtained from equation (1) by applying it at the boundary. In this situation it may be written as follows:

$$u_0(\Phi_x)_0 - \varepsilon(\Phi_{xx})_0 = S_0, \tag{34}$$

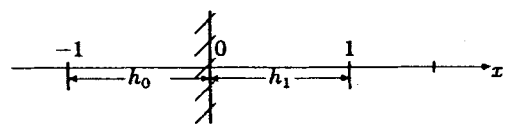

Figure 3. Location of points near the boundary

where

$$S = R - v\Phi_y + \varepsilon\Phi_{yy} \tag{35}$$

is a known quantity at the boundary points 0. The left-hand side of equation (34) will be approximated by second-order central differences for the first and second derivatives. By substituting equations (7) and (9) into equation (34), the following expression for Φ_{-1} is obtained:

$$\Phi_{-1} = \frac{R_0^x - 2}{R_0^x + 2}\Phi_1 + \frac{4}{R_0^x + 2}\Phi_0 - \frac{2S_0 h^2}{\varepsilon(R_0^x + 2)}. \tag{36}$$

After substituting this expression into equation (33), the following equation is obtained:

$$\overline{(\Phi_x)_1} = \frac{1 + R_0^x}{1 + R_0^x/2}\frac{\overline{\Phi_1} - \overline{\Phi_0}}{h_1} - \frac{R_0^x}{R_0^x + 2}\frac{S_0}{u_0}. \tag{37}$$

Thus this FD model for the Dirichlet BC may be implemented at the new time step n since it is always diagonal dominant without any delayed or lagged correction terms.

5.2. *Neumann boundary conditions*

This condition means that the values of the first derivative $(\Phi_x)_0$ are given at the boundary point 0. Denoting by K the quantity

$$\Phi_{xx} = \frac{u\Phi_x - S}{\varepsilon} \equiv K, \tag{38}$$

which may be known at the point 0 from the previous iteration, say, the following FD approximation can be proven to be of second order:

$$\overline{(\Phi_x)_1} = \frac{\overline{\Phi_1} - \overline{\Phi_0}}{h_1} + \frac{K_0 h_1}{2}. \tag{39}$$

This equation presents a second-order diagonal dominant formulation for the convection near the boundary, where it is considered to be applied in the new time step n without any lagged (correction) terms.

6. CONSIDERATION OF THE ITERATIVE SCHEME

The present scheme, which is given by equations (20)–(22), can be written for $u, v \geqslant 0$ at each stage n as follows:

$$\left(1 + \frac{\sigma_{i-1}}{1 + \sigma_{i-1}}\right)u_{i,j}\frac{\Phi_{i,j} - \Phi_{i-1,j}}{h_i} + \left(1 + \frac{\tau_{j-1}}{1 + \tau_{j-1}}\right)v_{i,j}\frac{\Phi_{i,j} - \Phi_{i,j-1}}{k_j}$$
$$- \varepsilon\left(\frac{\Phi_{i+1,j} - (1 + \sigma_i)\Phi_{i,j} + \sigma_i\Phi_{i-1,j}}{[2\sigma_i/(1 + \sigma_i)]\bar{h}_i^2} + \frac{\Phi_{i,j+1} - (1 + \tau_j)\Phi_{i,j} + \tau_j\Phi_{i,j-1}}{[2\tau_j/(1 + \tau_j)]\bar{k}_j^2}\right) - F_{i,j}^* = 0. \tag{40}$$

Here Φ is the discrete function's values at the iteration level n, and equation (33) is solved at each stage n with the following expression for $F_{i,j}$ (for the case $u, v \geqslant 0$):

$$F_{i,j} = \frac{\sigma_{i-1}}{1 + \sigma_{i-1}}u_{i,j}\frac{\Phi_{i-1,j} - \Phi_{i-2,j}}{h_i} + \frac{\tau_{j-1}}{1 + \tau_{j-1}}v_{i,j}\frac{\Phi_{i,j-1} - \Phi_{i,j-2}}{k_j} + R_{i,j}. \tag{41}$$

The convection parts in equation (40) are replaced by equation (33) or equation (39) at points on

the boundaries (with appropriate modifications for the directions and locations of the boundaries). These equations are also used at the grid points next to the boundaries. They help in replacing the values of Φ at the points $i \pm 2$ and $j \pm 2$ by again using the boundary conditions. Equation (41) may be treated at the n level ($* \equiv n$) for a direct solution or at the $n-1$ level ($* \equiv n-1$) for an iterative solution. Obviously, when using the second possibility in the iterative mode for equation (40), the solution at the stage n does not simulate the final solution, not even to first order as happens, for instance, with the KR method.[1] On the other hand, it converges very fast to the final solution, as expressed by equation (32). Since equation (40) presents a general two-dimensional FD equation for Φ, it will be solved in the present study in an iterative manner. The case where $F_{i,j}$ is treated at the n level is not the main subject of the present paper. However, it is not difficult to prove the following theorem:

Theorem 2. When $F_{i,j}^{(n)}$ is used in equation (40), and it is solved iteratively by the line relaxation or the alternating line relaxation technique, then the procedure is unconditionally stable.

Proof. This theorem assumes that the FD approximation for equation (1) is given by

$$\frac{u_{i,j}}{\bar{h}_i}\Bigg[\overbrace{\left(2-\frac{1}{1+\sigma_{i-1}}\right)\Phi_{i,j}}^{x,y} - \overbrace{\left(3-\frac{2}{1+\sigma_{i-1}}\right)\Phi_{i-1,j}}^{x} + \overbrace{\left(1-\frac{1}{1+\sigma_{i-1}}\right)\Phi_{i-2,j}}^{x}\Bigg]$$

$$+ \frac{v_{i,j}}{\bar{k}_j}\Bigg[\overbrace{\left(2-\frac{1}{1+\tau_{j-1}}\right)\Phi_{i,j}}^{x,y} - \overbrace{\left(3-\frac{2}{1+\tau_{j-1}}\right)\Phi_{i,j-1}}^{y} + \overbrace{\left(1-\frac{1}{1+\tau_{j-1}}\right)\Phi_{i,j-2}}^{y}\Bigg]$$

$$- \varepsilon\Bigg(\overbrace{\frac{\Phi_{i+1,j}-(1+\sigma_i)\Phi_{i,j}+\sigma_i\Phi_{i-1,j}}{[2\sigma_i/(1+\sigma_i)]\,\bar{h}_i^2}}^{x} + \overbrace{\frac{\Phi_{i,j+1}-(1+\tau_j)\Phi_{i,j}+\tau_j\Phi_{ij-1}}{[2\tau_j/(1+\tau_j)]\,\bar{k}_j^2}}^{y}\Bigg) = 0, \tag{42}$$

where elements that are denoted by $\overset{x}{\cdots}$ are treated implicitly in the first (x) sweep, and elements that are denoted by $\overset{y}{\cdots}$ are treated implicitly in the second (y) sweep of each iteration. Applying the Von Neumann analysis expressed in equation (26) for the case with $\sigma_i = 1$ and where h_i and k_j are constants, the amplification factor A of the error between two successive alternating line relaxation procedures as a function of the modes α and β is given by

$$A = A_1 * A_2, \tag{43}$$

where

$$A_1 = \left|\frac{R_v e^{-ij\beta}(4-e^{-ij\beta})+4\cos j\beta}{R_u(3-e^{-ii\alpha})(1-e^{-ii\alpha})+3R_v+4(3-2\cos i\alpha)}\right|, \tag{44}$$

$$A_2 = \left|\frac{R_u e^{-ii\alpha}(4-e^{-ii\alpha})+4\cos i\alpha}{R_v(3-e^{-ij\beta})(1-e^{-ij\beta})+3R_u+4(3-2\cos\beta)}\right|. \tag{45}$$

Here the specific local mesh Reynolds numbers are

$$R_u = \max_{i,j} R^x_{i,j}, \qquad R_v = \max_{i,j} R^y_{i,j}.$$

It can be shown that $\max_{i,j} A$ occurs

(a) for the first modes of the error, i.e. $\alpha = \pi/M$ and $\beta = \pi/N$
(b) for $\alpha R_u - \beta R_v = O(\alpha^2, \beta^2)$.

Using these results, the error amplification factor can be approximated by

$$A \simeq 1 - \frac{29\alpha\beta}{3(R_u + R_v) + 8}, \tag{46}$$

which is less than 1 for all practical M and N (such that $M \times N > 36$). □

Using this approach, the solution to the problem is obtained by applying two iteration loops: the outer loop which iterates on the lagged correction terms of the convection part, and the inner loop which solves the two-dimensional field, equation (40), for each state of these correction terms. Schematically in this S1 iteration scheme, equation (40) can be rewritten in the following general form:

$$\mathbf{A}\boldsymbol{\Phi}^{(n)} = \mathbf{F}^{(n-1)}, \tag{47}$$

where $\mathbf{A}$ is the coefficient matrix whose elements are given by equation (40), $\boldsymbol{\Phi}$ is the vector of variables, consisting of the Φ-values at all the grid points, and $\mathbf{F}$ is the source term vector whose elements are given by equation (41). If $MN = M \times N$, then the length of Φ and $\mathbf{F}$ is MN and the dimension of $\mathbf{A}$ is $MN \times MN$. $\mathbf{A}$ is a five-diagonal matrix: it has three main diagonals and the other two are located at a distance N (or M, depending on the ordering of the grid points) on both sides of the main diagonal. In a similar way we can write the definition of the source term $\mathbf{F}$ as follows:

$$\mathbf{F} = \mathbf{C}\boldsymbol{\Phi} + \mathbf{R}, \tag{48}$$

where $\mathbf{R}$ is a vector containing the original source term contributions of equation (1) at the various grid points, and the matrix $\mathbf{C}$ is in general a six-diagonal matrix. For example, when considering implicit solutions in the x-direction, the entries of $\mathbf{C}$ will come from the values of Φ at the points $j \pm 1, j \pm 2$ and $i \pm 2$. Generally, when solving equation (47), the following splitting of $\mathbf{A}$ is considered:

$$\mathbf{A} = \mathbf{P} - \mathbf{Q}, \tag{49}$$

where $\mathbf{P}$ is a matrix chosen such that its inverse can be found directly at the expense of a reasonable amount of the CPU time and computer memory allocation (for example, by some kind of factorization technique). In this case equation (47) is solved iteratively by

$$\boldsymbol{\Phi}^{(n)}_m = \mathbf{B}\boldsymbol{\Phi}^{(n)}_{m-1} + \mathbf{P}^{-1}\mathbf{F}^{(n-1)}, \tag{50}$$

where m is the iteration index and

$$\mathbf{B} = \mathbf{I} - \mathbf{P}^{-1}\mathbf{A} \tag{51}$$

is the iteration matrix. A necessary and sufficient condition for the iteration procedure, equation (50), to converge is that the spectral radius Λ of the iteration matrix will be bounded by 1:

$$\Lambda(\mathbf{B}) \leqslant 1. \tag{52}$$

The following Stein theorem is needed to prove Theorem 4 below:[12]

Theorem 3. Any real matrix $\mathbf{G}$ has $\Lambda(\mathbf{G}) \leqslant 1$ if and only if there exists a real positive definite matrix $\mathbf{E}$ such that the real matrix $\mathbf{D}$ given by

$$\mathbf{D} = \mathbf{E} - \mathbf{G}\mathbf{E}\mathbf{G}^{t} \tag{53}$$

is positive definite.

The following theorem is of some advantage in implementing the present numerical scheme.

Theorem 4. Any two-dimensional iteration procedure of the present second-order upwind scheme that fulfils the condition (52) will also converge if $m = n$ in equation (50).

Comment. This theorem means that after choosing a stable inner iteration procedure, one might collapse together the outer and inner iteration loops, producing one iteration loop in which the second-order upwind correction terms are also updated, and still get convergence. We will denote this kind of iteration scheme by S2.

Proof. We will outline only the main steps of the proof; it is easy to complete all the other details. It is convenient to consider the matrix $\mathbf{A}$ as consisting of two matrices:

$$\mathbf{A} = [\mathrm{CON}] + [\mathrm{DIF}], \tag{54}$$

where [CON] is the convection contribution matrix and [DIF] is the diffusion contribution matrix. It should be remembered that [DIF] is a positive definite matrix.[12] Now, a necessary condition for the iteration procedure defined by equation (50) to be acceptable is that it will converge for a pure diffusion-type problem. Thus we should also have $\Lambda(\mathbf{I} - \mathbf{P}^{-1}[\mathrm{DIF}]) \leqslant 1$. Yet, because of the existence of condition (52), the matrix $\mathbf{G} = \mathbf{P}^{-1}[\mathrm{CON}]$ has the property that $\Lambda(\mathbf{G}) \leqslant 1$ by using Theorem 3. The matrix $\mathbf{C}$ can be obtained from the matrix [CON] by using the scaling and shifting matrix $\mathbf{K}$:

$$\mathbf{C} = \mathbf{K} \times [\mathrm{CON}], \tag{55}$$

where $\mathbf{K} \geqslant \mathbf{0}$, and can be proven to be positive definite. Thus, by using the Stein theorem again, but applying it now to $\mathbf{P}^{-1}\mathbf{C}$, this convergence theorem is proved. □

In the examples that will be discussed later in the paper, it will be shown that the convergence when $m = n$ is in fact much faster than when the two iteration loops are considered separately.

7. IMPLEMENTATION OF THE PRESENT SCHEME

In order to make use of the scheme presented above while implementing the various theorems that have been discussed, the modified strongly implicit (MSI) method[13] has been chosen as the iterative procedure which can be presented also by equation (50). For two-dimensional fields the strongly implicit (SI) method[14] suggests the following iterative scheme for solving equations of the form of equation (40).

$$\Phi_{i,j} = a_{i,j}\Phi_{i-1,j} + b_{i,j}\Phi_{i,j-1} + c_{i,j}, \tag{56}$$

where the matrices $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are calculated using the elliptic FD equation (like equation (40)) based on the values of Φ from the last iteration. Examining the source term of the SI method c, it may be shown that it can be split into three terms, two of which contain elements of $\Phi_{i+1,j-1}$ and $\Phi_{i-1,j+1}$. Moving these terms to the left-hand side of equation (56), the MSI method[13] is recovered. This suggests solving the elements of Φ along *diagonal lines* of the field,

using the implicit tridiagonal inversion algorithm. It can also be shown[13] that there are several ways to implement the MSI procedure. Here we will use the simple form of the MSI technique, trying to parametrize it in a simple manner[15] by using the following form of the iterative technique:

$$\Phi^*_{i,j} = \omega_{i,j}(a_{i,j}\Phi_{i-1,j} + b_{i,j}\Phi_{i,j-1} + c_{i,j}) + (1-\omega_{i,j})\Phi_{i,j}, \tag{57}$$

where * denotes the values at the new iteration level and the rest of the terms are treated as known from the previous iteration level. It should be noted that the **c** source term in equations (56) and (60) should be split according to the MSI scheme, to achieve an implicit system along the diagonals of the field. ω is a matrix which maximizes the rate of convergence of the MSI technique; usually the elements of this matrix are $1 \leqslant \omega_{i,j} \leqslant 1$, where the standard SI procedure[14] is recovered for $\omega_{i,j} = 1$. The coefficient matrices **a**, **b** and **c** are the coefficient matrices given originally by the SI method.[14] Application of the appropriate recursion formula to the present case results in the following set of equations (given here for $u, v \geqslant 0$):

$$a_{i,j} = \frac{1}{D}\left[\left(2 - \frac{1}{1+\sigma_{i-1}}\right)\frac{u_{i,j}}{h_i} + \frac{\varepsilon(1+\sigma_i)}{2\bar{h}_i^2}\right], \tag{58}$$

$$b_{i,j} = \frac{1}{D}\left[\left(2 - \frac{1}{1+\tau_{j-1}}\right)\frac{v_{i,j}}{k_j} + \frac{\varepsilon(1+\tau_j)}{2\bar{k}_j^2}\right], \tag{59}$$

$$c_{i,j} = \frac{\varepsilon}{D}\left[\frac{1+\sigma_i}{2\sigma_i\bar{h}_i^2}(b_{i+1,j}\Phi_{i+1,j-1} + c_{i+1,j}) + \frac{1+\tau_j}{2\tau_j\bar{k}_j^2}(a_{i,j+1}\Phi_{i-1,j+1} + c_{i,j+1}) - R_{i,j}\right], \tag{60}$$

where

$$D = \left(2 - \frac{1}{1+\sigma_{i-1}}\right)\frac{u_{i,j}}{h_i} + \left(2 - \frac{1}{1+\tau_{j-1}}\right)\frac{v_{i,j}}{k_j} + \varepsilon\left[\frac{1+\sigma_i}{2\sigma_i\bar{h}_i^2}\left(1 + \sigma_i - \frac{\alpha_{i+1,j}}{\sigma_i}\right) + \frac{1+\tau_j}{2\tau_j\bar{k}_j^2}\left(1 + \tau_j - \frac{b_{i,j+1}}{\tau_j}\right)\right]. \tag{61}$$

Equations (57)–(60) imply that each iteration consists of the usual two sweeps of the MSI technique:

1. In the first sweep the matrices **a**, **b** and **c** are calculated using these equations.
2. In the second sweep $\boldsymbol{\Phi}$ is calculated using equation (56)

Here the evaluation of Φ is done by solving a tridiagonal system along the diagonal of Ω.[15] It should be noted that one has to be careful in incorporating the boundary conditions into these matrices, as is discussed in other papers.[15]

8. NUMERICAL EXPERIMENTS

For the numerical tests we have chosen the two-dimensional domain $\Omega = (0, 1) \times (0, 1)$. The tests are done by choosing

(a) an analytical solution for Φ
(b) analytical convection coefficient functions $u(x, y)$ and $v(x, y)$
(c) for a given diffusion coefficient ε, the source function $R(x, y)$ is calculated analytically.

By using the discrete values of u, v and R on a given FD grid, the values of Φ are calculated by the numerical approach that has been presented here and are compared with the exact values. In

what follows 'ERROR' means the maximum difference between the exact solution and the numerical solution in the domain:

$$\text{ERROR} = \max_{\Omega} |\Phi_{\text{exact}} - \Phi_{\text{numerical}}|. \tag{62}$$

Test 1. The following function has been chosen:

$$\Phi(x, y) = \sin \pi x + \cos \pi y. \tag{63}$$

The convection coefficients were chosen to be constant across the domain, with $\varepsilon = 0{\cdot}01$. The BC were chosen to be of the Dirichlet type. This problem was run first under the two different modes of iteration procedure:

(a) using two iteration loops (the S1 scheme)
(b) using only one iteration loop (the S2 scheme).

The test was executed with large convection coefficients: $u = 10$, $v = 10$. Table I gives the variation of the error for $M = N = 51$ and for $M = N = 101$ grids, where the grid points were evenly spread. It can be seen that the rate of convergence of the one-iteration-loop approach is much faster than the originally suggested two-iteration-loop approach. The local mesh Reynolds numbers for these FD grids are 20 and 10 respectively. When this problem was solved with the standard second-order FD approximation, the CM method, using the MSI iteration technique, the procedure was not stable and the solution diverged. Of course it is possible to find values for $\omega_{i,j}$ (in the interval [0, 1]) such that the solution will not diverge but will not converge either. Table II summarizes the errors for this problem for different grids. The second-order accuracy of the solution can be readily verified. It is also interesting to compare the converged solutions of the present approach with those of the CM scheme. To do this we have chosen $u = v = 0{\cdot}1$ so that for $M, N \geqslant \sim 20$ the MSI procedure will also be stable for the classical scheme. The comparison is presented in Table III for different FD grids. The main conclusion from these results is that although both

Table I. Comparison of the number of iterations to reduce the error to a prespecified ERROR level

	51 × 51 grid		101 × 101 grid	
ERROR	S1 method	S2 method	S1 method	S2 method
0·1	10	4	22	6
0·05	21	6	30	8
0·01	40	8	59	11
0·001	66	9	125	15
0·0001	84	11	17	179

Table II. Errors produced by the present scheme for different FD grids for $u = v = 10$ and a convergence criterion of 10^{-13}

$M = N$	20	40	80	160
Dirichlet BC	1·4E + 1	1·8E + 0	2·4E − 1	3·6E − 2
Neumann BC	3·2E − 1	2·7E − 2	3·1E − 3	5·7E − 4

Table III. Comparison between the CD method and the present second-order upwind method for $u = v = 0{\cdot}1$ and a convergence criterion of 10^{-13} for the Dirichlet BC

	Central differences (CM)		Present method (S)	
$M = N$	Number of iterations	$\max\lvert\Phi - \Phi_{exact}\rvert$	Number of iterations	$\max\lvert\Phi - \Phi_{exact}\rvert$
20	74	4·81E−2	102	8·47E−2
40	266	5·20E−3	305	1·20E−2
80	980	1·01E−3	1002	2·06E−3
160	3650	1·23E−4	3600	3·93E−4

schemes, the CM scheme and the S2 scheme, are at least second-order accurate, the results of the CM method are closer to the exact solutions than those of the S2 method. This difference in the errors cannot come from the approximations at the inner points of Ω, since the truncation error of the S method is only about twice as large as that of the CM method. Rather, the difference arises mainly because of the FD approximations at the points on $\partial\Omega$, as stated in the following theorem:

Theorem 5. If $\Phi \in \mathbb{C}^4$, then the ratio between the truncation errors of the S method and the CM method when using Dirichlet BC reaches a maximum near the boundary $\partial\Omega$ and is around 2–3 plus a linear function of the local specific mesh Reynolds number.

Proof. For small diffusion coefficients, the main contribution to the truncation error comes from the convection terms. Let us denote by T_r the truncation error of the x-convection term when the CM procedure is used. Using equation (34), it can be shown that the truncation error of the appropriate term in the S method (equation (37)) near the boundary, after substituting it into the governing equation to be solved (equation (1)) is

$$\left(3 - \frac{2}{R^x + 2}\right) T_r + \frac{1}{4} h^3 \Phi_{xxxx}. \tag{64}$$

Thus the ratio between the two truncation errors is

$$3 - \frac{2}{R^x + 2} + \frac{3}{2} h \frac{\Phi_{xxxx}}{\Phi_{xxx}}. \tag{65}$$

This approximation happens to be larger than that in the inner field. Under mild assumptions, the ratio $\frac{3}{2} h \Phi_{xxxx}/\Phi_{xxx}$ can be approximated by taking the governing equation (1) to be approximately R^x. Thus the above ratio can be written as $1{\cdot}5R^x + 3 - 2/(R^x + 2)$, so that for small values of R^x this ratio is ~ 2, and for large values of R^x it is $\sim 3(1 + R^x/2)$. This result can be roughly verified from the results given in Table III. Moreover, as M and N become smaller, the value of this last term may become larger than 1·5. This fact is also verified by this table. □

Test 2. In this we have considered the same problem as in Test 1, but the equation is subject to the following Neumann BC:

$$\frac{\partial \Phi}{\partial x} = 1 \quad \text{at } x = 0, \qquad \frac{\partial \Phi}{\partial x} = -1 \quad \text{at } x = 1, \tag{66}$$

$$\frac{\partial \Phi}{\partial y} = 0 \quad \text{at } y = 0 \quad \text{and} \quad y = 1, \tag{67}$$

with $\Phi(x = 0, y = 0{\cdot}5) = 0$. Table II summarizes the errors for different FD grids. The second-order accuracy of the solutions can be verified. It can be seen very clearly from this table that the results of the Neumann BC problem are much more accurate than those of the Dirichlet problem. The reason for this is that by using equation (39) it can be proven that the maximum value of the ratio of the truncation errors for this case is 2.

9. CONCLUSIONS

The paper presents a second-order upwind numerical scheme for solving a linear convection–diffusion elliptic equation. The scheme is of the iterative type, where in each iteration a similar problem to the original one is solved, using a standard upwind scheme with larger convection coefficients and appropriate correction terms. This method was proven and found to be unconditionally stable, its main feature being its rapid rate of convergence. This feature does not change much even when the above iteration procedure is combined with the iteration procedure due to the sparsity of the system (or to the multidimensionality nature of the governing equation). Although only a two-dimensional formulation has been presented, the scheme can be extended very easily to any number of elliptic dimensions.

It is possible to extend this method to the solution of numerically non-linear elliptic PDEs. Three iterative loops can be generated for this case:

(a) the inner loop due to the second-order upwind correction term
(b) the middle loop due to the multidimensionality of the problem
(c) the outer loop due to the non-linearity of the problem.

A number of questions remain which require further investigation:

1. How does this kind of iterative strategy converge?
2. Does the order in which the three loops are executed affect the rate of convergence?
3. Is it possible to get a higher rate of convergence by combining two of the loops or even collapsing all three loops into one loop?

REFERENCES

1. P. K. Khosla and S. G. Rubin, 'A diagonally dominant second-order accurate implicit scheme', *Comput.Fluids*, **2**, 207–209 (1974).
2. A. Lin and S. Rubin, 'A conditionally stable symmetric numerical scheme for solving convection–diffusion equations', in W. Pilkey *et al.* (eds), *Innovative Numerical Analysis in Applied Engineering*, University Press of Virginia, 1980.
3. P. Roache, Computational Fluid Dynamics, Hermosa, Albuquerque, NM, 1976.
4. B. Leonard, 'A stable and accurate convective modelling procedure based on quadratic upstream interpolation', *Comput. Methods Appl. Mech. Eng.*, **19**, 59–98 (1979).
5. M. Atias, M. Wolfshtein and M. Israeli, 'A study of the efficiency of various Navier–Stokes solvers', *AIAA 2nd Computational Fluid Dynamics Conf. Proc.*, 1975, pp. 81–90.
6. G. E. Forsythe and W. G. Warsow, *Finite-Difference Methods for Partial Differential Equations*, Wiley, 1960.
7. S. G. Rubin and P. K. Khosla, 'Navier–Stokes calculations with a coupled strongly-implicit method, Part 1: Finite-difference solutions', *Comput. Fluids*, **9**, 163–180 (1981).

8. P. N. Schwarztrauber and R. A. Sweet, 'The direct solution of the discrete Poisson equation on a disk', *SIAM J. Numer. Anal.*, **5**, 950–970 (1977).
9. A. Lin, 'High order three points schemes for boundary value problems. II: Non-linear problems', *J. Comput. Appl. Math.*, **15**, 269–282 (1986).
10. D. S. Kershaw, 'The ICCG method for the iterative solution of systems of linear equations', *J. Comput. Phys.*, **26**, 43–65 (1978).
11. A. Lin, 'High order three points schemes for boundary value problems. I: Linear problems', *SIAM J. Stat. Sci. Comput.* **7** (Part 3), 940–958 (1986).
12. D. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.
13. A. Lin, 'The modified strongly implicit method', Submitted.
14. H. L. Stone, 'Iterative solution of implicit approximations of multidimensional partial diffferential equations', *SIAM J. Numer. Anal.*, **5**, 530–559 (1968).
15. A. Lin, 'The parametrized strongly implicit method', *Int. j. numer. methods fluids*, **5**, 381–391 (1984).