ON A HIGH-ACCURACY DIFFERENCE SCHEME FOR AN ELLIPTIC EQUATION WITH SEVERAL SPACE VARIABLES*

A. A. SAMARSKII and V. B. ANDREYEV

(Moscow)

(*Received* 21 *June* 1963)

1. Suppose that in the region $D_p = \{0 < x_\alpha < 1, \alpha = 1, \ldots, p\}$ we are looking for a solution to the differential equation

$$Lu = \sum_{\alpha=1}^{p} L_\alpha u = -f(x), \qquad L_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2}, \tag{1}$$

which satisfies the condition

$$u|_\Gamma = g(x) \tag{2}$$

on the boundary Γ. Let $\bar{\omega}_h = \{x_i = (i_1 h, i_2 h, \ldots, i_p h) \in \bar{D}_p\}$ be a square net with step $h = 1/N$; and let γ be the boundary of the net $\bar{\omega}_h$. The numerical solution of the problem (1)-(2) is usually found with the use of the difference scheme

$$\Lambda y + f(x) = 0, \qquad y|_\gamma = g(x), \tag{3}$$

where

$$\Lambda = \sum_{\alpha=1}^{p} \Lambda_\alpha, \qquad \Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha} \tag{4}$$

(see [1] for the notation). This scheme gives second order accuracy. There are many iterative methods for solving the problem (3), and of these we have picked out those used in [2]-[8] which give the fastest rate of convergence. Without going into detail about any one method we

---

* *Zh. vych. mat.* 3, No. 6, 1006-1013, 1963.

note that the methods of [2]-[4] are applicable only for a parallelepiped and for $p = 2$ or $p = 3$, and [6]-[8] for a few more complicated regions and [6] for $p = 2$, [8] for arbitrary p. The paper [5] generalises [2], [3] for some wider problems.

2. To find the numerical solution of the problem (1)-(2) we use the scheme

$$\Lambda' y = \Lambda y + \frac{h^2}{6} \sum_{\alpha=1}^{p} \sum_{\beta>\alpha} \Lambda_\alpha \Lambda_\beta y = -\varphi(x), \qquad y|_\gamma = g, \tag{5}$$

where

$$\varphi(x) = f(x) + \frac{h^2}{12}. \tag{6}$$

This scheme has fourth order approximation in the class of sufficiently smooth solutions of (1), so that

$$\psi = \Lambda' u + \varphi = O(h^4). \tag{7}$$

It is not difficult to show that the scheme (5) has fourth order accuracy. Let us introduce the scalar products (see [1]):

$$(\eta, y) = \sum_{\omega_h} y_i \eta_i h^p, \qquad (y, \eta]_\alpha = \sum_{\omega_h + \alpha} y_i \eta_i h^p, \tag{8}$$

and the norms:

$$\|\eta\| = \sqrt{(\eta, \eta)}, \qquad \|\eta_{\bar{x}_\alpha}\| = \sqrt{(\eta_{\bar{x}_\alpha}, \eta_{\bar{x}_\alpha}]_\alpha}\,. \tag{9}$$

Let u be the solution of the problem (1)-(2), and y the solution of problem (5). For their difference $z = y - u$ we obtain

$$\Lambda' z = -\psi, \qquad z|_\gamma = 0. \tag{10}$$

Making a scalar multiplication of this equation by z we write down the energy identity (see [1]):

$$I = \frac{h^2}{6} \sum_{\alpha=1}^{p} \sum_{\beta>\alpha} \|z_{\bar{x}_\alpha \bar{x}_\beta}\|^2 + (\psi, z), \qquad I = \sum_{\alpha=1}^{p} \|z_{\bar{x}_\alpha}\|^2. \tag{11}$$

We use the obvious inequalities:

$$\begin{gathered} \|z\|^2 \leqslant \frac{1}{4p} I, \qquad \frac{h^2}{6} \sum_{\alpha=1}^{p} \sum_{\beta>\alpha} \|z_{\bar{x}_\alpha \bar{x}_\beta}\|^2 \leqslant \frac{p-1}{3} I, \\ (\psi, z) \leqslant \|z\| \|\psi\| \leqslant \frac{1}{\sqrt{4p}} I^{1/2} \|\psi\| \leqslant c_0 I + \frac{1}{16 c_0 p} \|\psi\|^2, \end{gathered} \tag{12}$$

where c_0 is an arbitrary positive constant. We insert these estimates in (11) and choose c_0 correspondingly. We then obtain

$$\|z\| \leqslant M_p \|\psi\|, \text{ where } M_p = \frac{3}{4p(4-p)}, \quad p \leqslant 3. \tag{13}$$

We have thus proved the following theorem.

Theorem 1. If the condition

$$\|\psi\| \leqslant Mh^4, \tag{14}$$

is satisfied then the difference scheme (5) for $p \leqslant 3$ converges in the mean at a rate $O(h^4)$ so that

$$\|y - u\| \leqslant M'h^4, \qquad M' = M \cdot M_p, \tag{15}$$

where M is a positive constant which does not depend on h.

Note 1. If instead of (1) we consider the equation

$$\bar{L}u = Lu - q(x)\, u = -f(x), \quad 0 < c_1 \leqslant q(x), \qquad u|_\Gamma = g(x), \tag{1'}$$

then it is easy to see that the solution of the problem

$$\Lambda' y - dy + \varphi(x) = 0, \qquad y|_\gamma = g(x), \tag{5'}$$

where

$$d(x) = q(x) + \frac{h^2}{12} \Lambda q(x),$$

converges in the mean at a rate $O(h^4)$ to the solution of the problem (1′) for $p = 4$ also.

3. Let us examine the following iterative scheme for the approximate solution of problem (5) for $p = 2, 3$:

$$v_{\bar{t}} = \Lambda v + \frac{h^2}{6} \sum_{\alpha=1}^{p} \sum_{\beta>\alpha} \Lambda_\alpha \Lambda_\beta \check{v} + \varphi, \quad v|_\gamma = g(x), \quad v(x, 0) = v^0(x), \tag{16}$$

where $v = v^{n+1}$ is the $(n + 1)$-th iteration, $\check{v} = v^{(n)}$, $v_{\bar{t}} = (v - \check{v})/\tau_n$; $\tau_n > 0$ is an iterative parameter to be chosen later. The initial value $v(x, 0) = v^{(0)}(x)$ is determined by the choice of the zero iteration. Let us construct two one-dimensional alternating direction algorithms for the numerical solution of problem (16).

A. We insert $\Lambda v = \Lambda \check{v} + \tau \Lambda v_{\bar{t}}$ in (16) and, following [6], replace the operator $(E - \tau\Lambda)$ where E is the unit operator by the operator A, where

$$A = \prod_{\alpha=1}^{p} A_\alpha, \quad A_\alpha = E - \tau\Lambda_\alpha.$$

Then instead of (16) we have the scheme

$$Av = [A + \tau\Lambda']\check{v} + \tau\varphi, \qquad v|_\gamma \tilde{=} g, \qquad v(x, 0) = v^{(0)}(x), \tag{17}$$

which we shall call the generating scheme. Introducing intermediate values $v_{(1)}, \ldots, v_{(p)} = v$ we reduce the solution of problem (5) to the solution of p one-dimensional problems:

$$A_1 v_{(1)} = [A + \tau\Lambda']\check{v} + \tau\varphi, \tag{18}$$

$$A_\alpha v_{(\alpha)} = v_{(\alpha-1)}, \quad \alpha = 2, \ldots, p; \qquad v_{(\alpha)} = A_{\alpha+1} \ldots A_p g \text{ for } x_\alpha = 0, 1.$$

B. Putting $w = v_{\bar{t}}$, we rewrite the generating scheme in the form

$$Aw = \Lambda'\check{v} + \varphi, \qquad w|_\gamma = 0. \tag{19}$$

From this we have the alternating direction algorithm

$$A_1 w_{(1)} = \Lambda'\check{v} + \varphi, \tag{20}$$

$$A_\alpha w_{(\alpha)} = w_{(\alpha-1)}, \quad \alpha = 2, \ldots, p; \quad w_{(\alpha)} = 0 \text{ for } x_\alpha = 0, 1,$$

$$v = \check{v} + \tau w_{(p)}.$$

To go from $\check{v}$ to v during the computation we must store the two layers: $\check{v}$ and $w_{(\alpha)}$, $\alpha = 1, 2, \ldots, p$. However this algorithm requires fewer operations than (18) (thus it is not necessary to calculate $A\check{v}$) and, furthermore, the functions $w_{(\alpha)}$ always satisfy zero boundary conditions. For $p = 2$ by analogy with [2] we can use an algorithm which does not contain the product $\Lambda_1\Lambda_2\check{v}$:

$$A_1 w_{(1)} = \Lambda_1\check{v} + \left(1 + \frac{h^2}{6\tau}\right)\Lambda_2\check{v} + \varphi, \tag{21}$$

$$A_2 w_{(2)} = w_{(1)} - \frac{h^2}{6\tau}\Lambda_2\check{v}, \quad v = \check{v} + \tau w_{(2)}; \qquad w_{(\alpha)} = 0, \quad x_\alpha = 0, 1.$$

Each of the equations $A_\alpha w_{(\alpha)} = \varphi_\alpha$ where φ_α is a given function can be solved using the formulae of one-dimensional successive substitution (see [9], pp. 283-309). All the computing algorithms which we have mentioned give the same generating scheme (17) as we can see by eliminating the intermediate values of $v_{(\alpha)}$ or $w_{(\alpha)}$, $\alpha = 1, \ldots, p-1$.

4. We show that the iterations defined on scheme (17) converge whatever the choice of the zero iteration $v^{(0)}(x)$ and of the sequence $\{\tau_n\}$ satisfying the condition $0 < c_1 \leqslant \tau_n \leqslant c_2$, where c_1 and c_2 are constants which do not depend on the iteration number n. Following [3] we give a method of choosing $\{\tau_n\}$ for which the rate of convergence of the iterations will be "sufficiently fast". We obtain the following conditions for the difference $z = v - y$, where y is the solution of the initial problem (5), $v = v^{(n)}$ is the solution of problem (17):

$$A z_{\bar{t}} = \Lambda' \check{z}, \qquad z|_\gamma = 0, \qquad z(x, 0) = z^{(0)}(x) = v^{(0)} - y(x). \tag{22}$$

Let us expand z and $\check{z}$ in terms of the eigenfunctions

$$\mu_k = \prod_{\alpha=1}^{p} \sin k_\alpha \pi x_\alpha, \quad k_\alpha = 1, \ldots, N-1, \quad k = \{k_1, \ldots, k_p\}, \quad x_\alpha = i_\alpha h, \tag{23}$$

of the operators Λ_α:

$$z = z^{(n+1)} = \sum_k a_k^{(n+1)} \mu_k, \qquad \check{z} = z^{(n)} = \sum_k a_k^{(n)} \mu_k. \tag{24}$$

Inserting (24) in (22) and using the linear independence of $\{\mu_k\}$ we obtain

$$a_k^{(n+1)} = \rho_k^{(n+1)} a_k^{(n)}, \tag{25}$$

$$\rho_k^{(n+1)} = 1 - \lambda \Big[\sum_{\alpha=1}^{p} \xi_\alpha - \frac{2}{3} \sum_{\alpha=1}^{p} \sum_{\beta>\alpha} \xi_\alpha \xi_\beta \Big] \prod_{\alpha=1}^{p} (1 + \lambda \xi_\alpha)^{-1},$$
$$\lambda = \lambda_{n+1} = \frac{4\tau_{n+1}}{h^2}, \qquad \xi_\alpha = \xi_{k_\alpha} = \sin^2 \frac{k_\alpha \pi h}{2}. \tag{26}$$

Theorem 2. The iterative method (17) for $p = 2, 3$ converges in the metric $L_2(\omega_h)$ whatever parameters τ_n satisfying the condition $0 < c_1 \leqslant \tau_n \leqslant c_2$ are chosen.

Thus using (25) we can write

$$a_k^{(n+1)} = \prod_{s=1}^{n+1} \rho_k^{(s)} a_k^{(0)}, \tag{27}$$

and it follows from (24) and (27) that

$$z_i^{(n+1)} = \sum_k a_k^{(0)} \prod_{s=1}^{n+1} \rho_k^s \mu_k(i). \tag{28}$$

Hence

$$\| z^{(n+1)} \| = \Big[\sum_{\omega_h} h^p \Big(\sum_k a_k^{(0)} \prod_{s=1}^{n+1} \rho_k^{(s)} \mu_k(i) \Big)^2 \Big]^{1/2} \leqslant \max_k \prod_{s=1}^{n+1} \rho_k^{(s)} \| z^0 \|, \tag{29}$$

where $z^{(0)} = v^{(0)} - y$ is the difference between the zero approximation and the exact solution of (5). We have to show that

$$\max_k \prod_{s=1}^{n+1} \rho_k^{(s)} \to 0 \quad \text{as } n \to \infty.$$

Let us first estimate $\rho_k^{(s)}$. Since $1 \leqslant k_\alpha \leqslant N-1$ we have

$$\sin^2 \frac{\pi h}{2} \leqslant \xi_\alpha < 1 \tag{30}$$

and therefore

$$2\xi_\alpha \xi_\beta \leqslant \xi_\alpha^2 + \xi_\beta^2 < \xi_\alpha + \xi_\beta. \tag{31}$$

Using (26) and (31) we obtain

$$0 < \rho_k^{(s)} \leqslant 1 - \frac{\lambda_s \left(1 - \frac{p-1}{3}\right) \sum_{\alpha=1}^{p} \xi_\alpha}{\prod_{\alpha=1}^{p} (1 + \lambda_s \xi_\alpha)}. \tag{32}$$

It follows from (32) that

$$\rho_k^{(s)} \leqslant \rho < 1 \tag{33}$$

for $0 < \overset{*}{c}_1 \leqslant \lambda \leqslant \overset{*}{c}_2$, $p = 2, 3$, where $\overset{*}{c}_1, \overset{*}{c}_2$ and ρ do not depend on the number of the iteration.

Using (29) and (33) we obtain

$$\| z^{(n+1)} \| \leqslant \rho^{n+1} \| z^{(0)} \| \to 0 \quad \text{as } n \to \infty.$$

Note 2. For equation (1′) and the corresponding difference scheme (5′) (see Note 1) the generating scheme is

$$Av = [A + \tau\Lambda' - \tau d]\, \check{v} + \varphi. \tag{17′}$$

It is not difficult to see that Theorem 2 remains valid for (17′) for $p = 4$ also.

5. *Lemma 1.* If $0 < m < 1 < M$ and

$$\rho(a, b) = 1 - \frac{2(a+b)}{3(1+a)(1+b)}, \tag{34}$$

then

$$\rho_2 = \max_{m\leqslant a,b\leqslant M} \rho(a, b) = \max\left[1 - \frac{4M}{3(1+m)^2},\quad 1 - \frac{4M}{3(1+M)^2}\right]. \tag{35}$$

In fact it is not difficult to see by a direct check that

$$\rho^2(a, b) \leqslant \rho(a, a)\,\rho(b, b). \tag{36}$$

Since the region of definition of $\rho(a, b)$ together with the point (a, b) also contains the points (a, a), (b, b) and conversely, on the basis of (36) we can state that $\max\limits_{m\leqslant a,b\leqslant M} \rho(a, b)$ is attained when $a = b$. Let us examine the behaviour of the function

$$\bar\rho(a) = 1 - \frac{4a}{3(1+a)^2}.$$

$$\frac{d\bar\rho}{da} = -\frac{4(1-a)}{3(1+a)^3} = \begin{cases} > 0 \text{ for } a > 1, \\ \leqslant 0 \text{ for } a \leqslant 1. \end{cases} \tag{37}$$

It follows from (37) that $\max \bar\rho(a)$ is attained either when $a = m$ or when $a = M$, which the following lemma proves.

Lemma 2. If $0 < m < ½ < M$ and

$$\rho(a, b, c) = 1 - \frac{1}{3}\theta(a, b, c), \tag{38}$$

where

$$\theta(a, b, c) = \frac{a+b+c}{(1+a)(1+b)(1+c)},$$

then

$$\rho_3 = \max_{m\leqslant a,b,c\leqslant M} \rho(a, b, c) = \max\left[1 - \frac{m}{(1+m)^3},\quad 1 - \frac{M}{(1+M)^3}\right]. \tag{39}$$

The lemma will be proved if we can show that the minimum of the function $\theta(a, b, c)$ is attained either when $a = b = c = m$, or when $a = b = c = M$. Let us fix a and examine the behaviour of $\theta(a, b, c)$ depending on the change in b and c. By a direct check we see that

$$\theta^2(a,b,c) \geqslant \theta(a,b,b)\,\theta(a,c,c). \tag{40}$$

Noting that

$$\frac{\partial\theta(a,e,c)}{\partial b} = 2\,\frac{1-a-b}{(1+a)(1+b)^2} = \begin{cases} \geqslant 0 & \text{for } b \leqslant 1-a, \\ \leqslant 0 & \text{for } b \geqslant 1-a, \end{cases} \tag{41}$$

and using (40) we obtain

$$\bar\theta(a) = \min_{m\leqslant b,c\leqslant M} \theta(a,b,c) = \min_{m\leqslant b\leqslant M} \theta(a,b,b) = \\ = \min\left[\frac{a+2m}{(1+a)(1+m)^2},\quad \frac{a+2M}{(1+a)(1+M)^2}\right]. \tag{42}$$

Further,

$$\frac{d\bar\theta(a)}{da} = \begin{cases} \text{either } \dfrac{1-2m}{(1+a)^2(1+m)^2} > 0, \\ \quad\text{or } \dfrac{1-2M}{(1+a)(1+M)^2} < 0. \end{cases} \tag{43}$$

On the basis of (42) and (43) we conclude:

$$\min_{m\leqslant a\leqslant M} \bar\theta(a) = \min_{m\leqslant a,b,c\leqslant M} \theta(a,b,c) = \min\left(\frac{3m}{(1+m)^2},\ \frac{3M}{(1+M)^2}\right). \tag{44}$$

6. Now let us choose the sequence $\{\lambda_n\}$ so that it satisfies the conditions

$$\lambda_n \xi^{(n)} = m, \qquad \lambda_n \xi^{(n+1)} = M, \qquad \xi^{(1)} = \sin^2\frac{\pi h}{2}, \tag{45}$$

and the number of iterations $n = n_0$ so that

$$\xi^{(n_0)} < 1, \qquad \xi^{(n_0+1)} \geqslant 1. \tag{46}$$

It follows that

$$\lambda_n = mq^{n-1}\sin^{-2}\frac{\pi h}{2}, \qquad \xi^{(n)} = q^{-n+1}\sin^2\frac{\pi h}{2}, \tag{47}$$

$$2\ln\sin\frac{\pi h}{2}\cdot\ln^{-1}q \leqslant n_0 \leqslant 2\ln\sin\frac{\pi h}{2}\ln^{-1}q + 1, \tag{48}$$

where $q = m/M$.

Lemma 3. If a cycle of n_0 iterations using the method (17) is carried out with the set of parameters $\{\lambda_n\}$ defined in (47) then

$$\|z^{(n_0)}\| \leqslant \rho_p \|z^{(0)}\|, \tag{49}$$

where ρ_p is given by formulae (35) and (39).

Thus when

$$\xi^{(n)} \leqslant \xi_\alpha \leqslant \xi^{(n+1)}, \tag{50}$$

$$m \leqslant \lambda_n \xi_\alpha \leqslant M, \tag{51}$$

and the intervals $[\xi^n, \xi^{n+1}]$ cover the whole region of values of ξ_α for each value of ξ_α, there exists in consequence at least one value of n for which

$$\rho_k^{(n)} < \rho_p. \tag{52}$$

The inequality (43) and Theorem 2 prove the lemma.

Note 3. Bearing (37) and (41) in mind it is easy to see that $\max(M/m)$ (or min n_0, which is equivalent) is attained for fixed ρ_p when the first and second terms on the right-hand side of (35) and (39) are equal:

$$1-\frac{4m}{3\,(1+m)^2}=1-\frac{4M}{3\,(1+M)^2}\,; \quad 1-\frac{m}{(1+m)^3}=1-\frac{M}{(1+M)^3}\,. \tag{53}$$

Then when $p = 2$

$$M=\frac{1}{m}, \tag{54}$$

when $p = 3$

$$M=\frac{\sqrt{(3+m)^2+4/m}-(3+m)}{2}\,. \tag{55}$$

Theorem 3. In order to reduce L_2, the norm of the error $\|z^0\|$, $1/\varepsilon$ times using method (17) it is sufficient to carry out k_0 cycles of n_0 iterations with the set of parameters $\{\lambda_n\}$ given by (47), where n_0 is defined by (48) and k_0 by (56):

$$k_0 > \frac{\ln(1/\varepsilon)}{\ln(1/\rho_p)}\,. \tag{56}$$

The proof of the theorem follows from Lemma 3.

Note 4. It follows from Theorem 3 that the total number of iterations required to reduce the error $\|z^0\|$ $1/\varepsilon$ times is

$$\nu \succeq \frac{2\ln\sin\frac{\pi h}{2}\ln\varepsilon}{\ln q\ln\rho_p}\,. \tag{57}$$

Using Note 3 and optimising ν w.r.t. m we obtain:

for $p = 2$

$$\nu_{opt} = 3.00 \ln \frac{1}{\sin(\pi h/2)} \ln \frac{1}{\varepsilon}, \tag{58}$$

$$m_{opt} = 0.24, \quad \bar{\rho}_{3opt} = 0.79, \quad q = \frac{m}{M} = 0.058; \tag{59}$$

for $p = 3$

$$\nu_{opt} = 8.40 \ln \frac{1}{\sin(\pi h/2)} \ln \frac{1}{\varepsilon}, \tag{60}$$

$$m_{opt} = 0.13, \quad \bar{\rho}_{3opt} = 0.91, \quad q = \frac{m}{M} = 0.080. \tag{61}$$

REFERENCES

1. SAMARSKII, A.A., *Zh. vychisl. Mat. mat. Fiz.*, 3, No.3, 431-466, 1963.

2. PEACEMAN, D.W. and RACHFORD, H.H., *J. Soc. Industr. and Appl. Math.*, 3, No. 1, 28-41, 1955.

3. DOUGLAS, J. and RACHFORD, H.H., *Trans. Amer. Math. Soc.*, 82, No. 2, 421-439, 1956.

4. DOUGLAS, J., *Numer. Math.*, 4, No. 1, 41-63, 1962.

5. BIRKHOFF, G. and VARGA, R.S., *Trans. Amer. Math. Soc.*, 92, No. 1, 13-24, 1959.

6. D'YAKONOV, Ye.G., *Dokl. Akad. Nauk SSSR*, 143, No. 1, 21-24, 1962.

7. D'YAKONOV, Ye.G., *Zh. vychisl. Mat. mat. Fiz.*, 2, No.4, 549-568, 1962.

8. D'YAKONOV, Ye.G., The Solution of Some Multi-dimensional Problems in Mathematical Physics using the Method of Nets (Reshenie nekotorykh mnogomernykh zadach matematicheskoi fiziki pri pomoshchi metoda setok). Candidate Dissertation in Physics and Mathematics, Matem. in-t Akad. Nauk SSSR, Moscow, 1962.

9. GODUNOV, S.K. and RYABEN'KII, V.S., Introduction to the Theorem of Difference Schemes (Vvedenie v teoriyu raznostnykh skhem). Fizmatgiz, Moscow, 1962.