ON THE 9-POINT DIFFERENCE FORMULA FOR LAPLACE'S EQUATION

by

A. I. van de Vooren* and A. C. Vliegenthart**

1. *Introduction.*

One method of solving boundary value problems for elliptic differential equations is to replace the differential equation by a difference equation. In case of the two-dimensional Laplace equation this is most frequently performed by aid of the so-called 5-point formula, which for a square net of mesh length h involves a truncation error of order h^4. A number of techniques for the rapid solution of the resulting system of algebraic equations has been developed, a. o. by Young (overrelaxation method based upon the property A of the matrix) [1, 2] and by Rachford (alternating-direction methods) [3].

The present paper is concerned with the investigation of the 9-point formula and its comparison with the 5-point formula. The truncation error of the 9-point formula is of order h^8 for a square net. It has already been shown by Gerschgorin in 1930 [4] that the discretization error, by which is meant the difference between the exact solutions of the differential equation and of the system of difference equations, is of order h^2 for a rectangular region if the 5-point formula is applied and if the solution is four times differentiable, including at the contour. Analogously it can be shown that for the 9-point formula the discretization error in a rectangular region is of order h^6, provided the eighth order derivatives exist everywhere. Thus the 9-point formula yields results of the same accuracy with a much smaller mesh length than the 5-point formula. This means both a smaller system of equations and a larger convergence rate which outweighs the disadvantage of the more complicated form of the 9-point formula. In this paper convergence rates are compared for point iteration methods (Jacobi, Gauss-Seidel and overrelaxation). It is shown that although the matrix of the 9-point formula does not possess property A, the most favourable relaxation factor and the convergence rate for small h and for a rectangular region can be calculated by an analytic method, based on separation of variables. Our result, which is confirmed by numerical calculations, differs from an earlier result obtained by Garabedian [5].

The authors wish to acknowledge useful discussions with Mr. H. J. Burema.

2. *The 5-point and 9-point formulas and their truncation errors.*

In order to replace the two-dimensional Laplace equation by a difference equation, we introduce a rectangular net with mesh lengths h and k in x- and y-direction. We present the difference equations for regular points of the region, that are points of which all neighbouring points, whose function values also appear in the difference equation, lie within the region or on its boundary. The derivation, which makes use of Taylor series expansions for the exact solution u(x, y) of Laplace's equation, has been given in detail in the authors' report [6].

When U(x, y) denotes the solution of the system of difference equations, then the 5-points formula is

* Dept. of Mathematics, University of Groningen, the Netherlands.

** Technological University, Delft, the Netherlands.

$$U(x,y) = \frac{k^2}{2(h^2+k^2)}\left\{U(x+h,y)+U(x-h,y)\right\} + \frac{h^2}{2(h^2+k^2)}\left\{U(x,y+k)+U(x,y-k)\right\} \tag{2.1}$$

with the truncation error $-\frac{1}{24}h^2k^2\frac{\partial^4 u(x,y)}{\partial x^4}$.

The 9-point formula is

$$U(x,y) = \frac{1}{20}\left\{U(x+h,\ y+k) + U(x-h,\ y+k) + U(x-h,\ y-k) + U(x+h,\ y-k)\right\}$$
$$-\frac{1}{10}\frac{h^2-5k^2}{h^2+k^2}\left\{U(x+h,y)+U(x-h,y)\right\} + \frac{1}{10}\frac{5h^2-k^2}{h^2+k^2}\left\{U(x,y+k)+U(x,y-k)\right\} \tag{2.2}$$

with the truncation error $\frac{1}{400}h^2k^2(h^2-k^2)\frac{\partial^6 u(x,y)}{\partial x^6}$.

In the case h = k the 9-point formula becomes

$$U(x,y) = \frac{1}{20}\left\{U(x+h,\ y+k) + U(x-h,\ y+k) + U(x-h,\ y-k) + U(x+h,\ y-k)\right\}$$
$$+\frac{1}{5}\left\{U(x+h,y) + U(x-h,y) + U(x,y+k) + U(x,y-k)\right\} \tag{2.3}$$

with the truncation error $-\frac{1}{10080}h^8\frac{\partial^8 u(x,y)}{\partial x^8}$.

In the last case the truncation error is of smaller order than if $h \neq k$.

3. *General remarks on the iterative method.*

We now shall consider the Dirichlet problem for a rectangular region with sides a and b in x and y-direction respectively. The net points have coordinates

$$x = ph\ (p = 1,2\ldots,\ m-1), \qquad y = qk\ (q = 1,2\ldots,\ n-1)$$
$$\text{while } mh = a \text{ and } nk = b.$$

Since all net points are regular points, the difference equation, given by one of the formulas (2.1) through (2.3), completely replaces the differential equation. The difference equation is solved by an iterative process which generally can be written as

$$\mathbf{U}^{(n+1)} = \mathbf{H}\mathbf{U}^{(n)} + \mathbf{k}. \tag{3.1}$$

$\mathbf{U}^{(n)}$ denotes the vector of which the elements are the values of U(x,y) in the net points, obtained after n iterations. **H** is the square matrix of order (m - 1)(n - 1) determined by the difference equation and the iterative method chosen. Solution by iteration is employed since the matrix **H** is

sparse. The vector $\mathbf{k}$ is due to the given boundary values. It enters into the difference equation if this is applied to a point of which one or more of the neighbouring points are on the boundary.

If $\mathbf{U}$ is the exact solution of the difference equation, then

$$\mathbf{U} = \mathbf{H}\mathbf{U} + \mathbf{k} \tag{3.2}$$

which shows that the error $\mathbf{v}^{(n)} = \mathbf{U} - \mathbf{U}^{(n)}$ satisfies

$$\mathbf{v}^{(n+1)} = \mathbf{H}\mathbf{v}^{(n)} \text{ or } \mathbf{v}^{(n)} = \mathbf{H}^n \mathbf{v}^{(0)}.$$

A necessary and sufficient condition for convergence (i.e. $\lim_{n\to\infty} \mathbf{v}^{(n)} = 0$) with an arbitrary starting vector $\mathbf{v}^{(o)}$ is that $\lim_{n\to\infty} \mathbf{H}^n = 0$. According to a well-known theorem of matrix theory [7], this is the case if the spectral radius $\lambda(\mathbf{H})$ of $\mathbf{H}$ is smaller than 1. Since the elements of $\mathbf{v}^{(n)}$ will vanish asymptotically as λ^n, they will become smaller than ϵ as soon as $c\lambda^n < \epsilon$, i.e. if $n \log \lambda$ becomes smaller than a certain value.

The number of steps n is inversely proportional to $\log \lambda$. Therefore the rate of convergence is defined by

$$R(\mathbf{H}) = -\log \lambda(\mathbf{H}) \tag{3.3}$$

In the following sections we shall compare the convergence rates for the 5- and 9-point formulas if the Jacobi, Gauss-Seidel and overrelaxation methods are applied.

4. *The Jacobi method.*

This method, also known as the method of simultaneous displacements, consists of substituting the n^{-th} approximation at the right hand side of the difference equation given in Sec. 2. The $n+1^{-th}$ approximation is obtained by using only values of the n^{-th} approximation for all points of the net.

The iteration formula is in the case of the 5-point formula, according to eq. (2.1)

$$U^{(n+1)}(x,y) = \frac{k^2}{2(h^2+k^2)}\left\{U^{(n)}(x+h,y) + U^{(n)}(x-h,y)\right\} + \frac{h^2}{2(h^2+k^2)}\left\{U^{(n)}(x,y+k) + U^{(n)}(x,y-k)\right\} \tag{4.1}$$

or, written in matrix form,

$$\mathbf{U}^{(n+1)} = \mathbf{B}^{(5)}\,\mathbf{U}^{(n)} + \mathbf{k},$$

$\mathbf{B}^{(5)}$ being a symmetric matrix.

The eigenvalues of $\mathbf{B}^{(5)}$ satisfy the equation

$$\mu\,\mathbf{U} = \mathbf{B}^{(5)}\,\mathbf{U},$$

which agrees with the difference equation

$$\mu\,U(x,y) = \frac{k^2}{2(h^2+k^2)}\left\{U(x+h,y)+U(x-h,y)\right\} + \frac{h^2}{2(h^2+k^2)}\left\{U(x,y+k) + U(x,y-k)\right\} \tag{4.2}$$

provided the values of U for points on the boundary are taken equal to 0.

For the 9-point formula we can derive from eq. (2.2) a Jacobi matrix $\mathbf{B}^{(9)}$ and a difference equation which is comparable to eq. (4.2), viz.

$$\mu U(x,y) = \frac{1}{20}\Big\{U(x+h,\ y+k) + U(x-h,\ y+k) + U(x-h,\ y-k) + U(x+h,\ y-k)\Big\}$$
$$- \frac{1}{10}\frac{h^2-5k^2}{h^2+k^2}\Big\{U(x+h,y)+U(x-h,y)\Big\} + \frac{1}{10}\frac{5h^2-k^2}{h^2+k^2}\Big\{U(x,y+k)+U(x,y-k)\Big\} \tag{4.3}$$

Eqs. (4.2) and (4.3) can be solved by aid of separation of variables, that is by putting

$$U(x,y) = X(x)\,Y(y).$$

The result (see [6] for details of derivation) is that the eigenvalues of eq. (4.2) are given by

$$\mu = \frac{k^2}{h^2+k^2}\cos\frac{p\pi}{m} + \frac{h^2}{h^2+k^2}\cos\frac{q\pi}{n} \tag{4.4}$$

and those of eq. (4.3) by

$$\mu = \frac{1}{5}\left(\cos\frac{p\pi}{m}\cos\frac{q\pi}{n} + \frac{5k^2-h^2}{h^2+k^2}\cos\frac{p\pi}{m} + \frac{5h^2-k^2}{h^2+k^2}\cos\frac{q\pi}{n}\right). \tag{4.5}$$

In both cases $p = 1,2\ldots, m-1$ and $q = 1,2\ldots, n-1$.

It follows from eq. (4.4) that the spectral radius of $\mathbf{B}^{(5)}$ is obtained by putting $p = q = 1$

$$\lambda(\mathbf{B}^{(5)}) = \frac{k^2}{h^2+k^2}\cos\frac{\pi h}{a} + \frac{h^2}{h^2+k^2}\cos\frac{\pi k}{b} \tag{4.6}$$

which is smaller than 1. For large values of m and n the convergence rate becomes

$$R(\mathbf{B}^{(5)}) = \frac{1}{2}\frac{h^2k^2}{h^2+k^2}\pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right). \tag{4.7}$$

The spectral radius of $\mathbf{B}^{(9)}$ is also obtained for $p = q = 1$, provided $1/\sqrt{5} < h/k < \sqrt{5}$. In that case the convergence rate is

$$R(\mathbf{B}^{(9)}) = \frac{3}{5}\frac{h^2k^2}{h^2+k^2}\pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right). \tag{4.8}$$

If $h > k\sqrt{5}$, all terms in eq. (4.5) become negative if $\cos q\pi/n$ is taken negative. The eigenvalue with the largest absolute value occurs for $p = 1$, $q = n-1$ and is equal to

$$\frac{1}{5}\left\{-1 - \frac{6h^2-6k^2}{h^2+k^2} + \left(1+\frac{h^2-5k^2}{h^2+k^2}\right)\frac{\pi^2}{2m^2} + \left(1+\frac{5h^2-k^2}{h^2+k^2}\right)\frac{\pi^2}{2n^2}\right\}.$$

The terms independent of m and n are

$$\frac{1}{5}\left\{-1-\frac{6h^2-6k^2}{h^2+k^2}\right\}=\frac{1}{5}\left\{-5-\frac{2h^2-10k^2}{h^2+k^2}\right\},$$

which is smaller than -1.

Convergence is here only possible if m and n are not too large.

The same reasoning holds for $k > h\sqrt{5}$.

The conclusion is drawn that if the ratio of the mesh sizes is larger than $\sqrt{5}$, convergence with the 9-point formula is not guaranteed and is certainly absent if h and k are sufficiently small. In that case the operator (2.2) is no longer positive, which means that if it is applied to positive U-values, the left hand side may become negative. This is inadmissible for a solution of Laplace's equation.

If the ratio of the mesh sizes is smaller than $\sqrt{5}$, the Jacobi procedure is convergent and gives even faster convergence in the case of the 9-point formula than for the 5-point formula, since $R(\mathbf{B}^{(9)}) = 1.2\ R(\mathbf{B}^{(5)})$.

5. *The Gauss-Seidel method.*

In this method, which is also known as the method of successive displacements, the $(n+1)^{th}$ approximation for an element in the vector **U** is used in the further calculations as soon as it is known. The properties of the method then depend upon the order of the elements in the vector. We shall take these elements in "reading" order with x increasing to the right and y increasing in downward direction.

If we divide the Jacobi matrix **B**, of which the diagonal entries are all zero, in the lower and upper triangular matrices **L** and **R**, the equation of the Gauss-Seidel method becomes in matrix form

$$\mathbf{U}^{(n+1)} = \mathbf{L}\,\mathbf{U}^{(n+1)} + \mathbf{R}\,\mathbf{U}^{(n)} + \mathbf{k}$$

$$\text{or}\qquad \mathbf{U}^{(n+1)} = (\mathbf{I}-\mathbf{L})^{-1}\,\mathbf{R}\,\mathbf{U}^{(n)} + (\mathbf{I}-\mathbf{L})^{-1}\,\mathbf{k}\,. \tag{5.1}$$

Convergence is determined by the spectral radius of

$$\mathbf{C} = (\mathbf{I}-\mathbf{L})^{-1}\,\mathbf{R}$$

which should be smaller than 1.

5-point formula.

The eigenvalues of $\mathbf{C}^{(5)} = (\mathbf{I}-\mathbf{L})^{-1}\,\mathbf{R}$ satisfy the equation

$$\mu\mathbf{U} = \mathbf{C}^{(5)}\,\mathbf{U} \text{ or } \mu(\mathbf{I}-\mathbf{L})\mathbf{U} = \mathbf{R}\,\mathbf{U}.$$

The last form corresponds to the difference equation (see eq. (2.1))

$$\mu\left\{U(x,y)-\frac{k^2}{2(h^2+k^2)}U(x-h,y)-\frac{h^2}{2(h^2+k^2)}U(x,y-k)\right\}=$$

$$\frac{k^2}{2(h^2+k^2)}U(x+h,y)+\frac{h^2}{2(h^2+k^2)}U(x,y+k). \tag{5.2}$$

The method of seperation of variables is again applied for the solution of eq. (5.2), see [6]. The result is, as is well-known [1,2], that the eigenvalues are the squares of the eigenvalues of the matrix $\mathbf{B}^{(5)}$. The convergence rate for large values of m and n is given by

$$R(\mathbf{C}^{(5)}) \approx \frac{h^2k^2}{h^2+k^2}\pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right) \tag{5.3}$$

and hence is twice that of the Jacobi method with the 5-point formula.

9-point formula.

The difference equation for the iteration is

$$U^{(n+1)}(x,y) = \frac{1}{20}\left\{U^{(n)}(x+h,\, y+k)+U^{(n)}(x-h,\, y+k)+U^{(n+1)}(x+h,\, y-k)+\right.$$

$$\left.+\, U^{(n+1)}(x-h,\, y-k)\right\} - \frac{1}{10}\frac{h^2-5k^2}{h^2+k^2}\left\{U^{(n)}(x+h,y)+U^{(n+1)}(x-h,y)\right\}+$$

$$+\frac{1}{10}\frac{5h^2-k^2}{h^2+k^2}\left\{U^{(n)}(x,y+k)+U^{(n+1)}(x,y-k)\right\}.$$

This fixes the matrices $\mathbf{L}$, $\mathbf{R}$ and $\mathbf{C}^{(9)} = (\mathbf{I}-\mathbf{L})^{-1}\mathbf{R}$ and hence, the equation for the eigenvalues and eigenvectors of $\mathbf{C}^{(9)}$ becomes

$$\mu\left\{U(x,y)-\frac{1}{20}U(x+h,y-k)-\frac{1}{20}U(x-h,\,y-k)+\frac{1}{10}\frac{h^2-5k^2}{h^2+k^2}U(x-h,y)-\frac{1}{10}\frac{5h^2-k^2}{h^2+k^2}\right.$$

(5.4)

$$\left.U(x,y-k)\right\} = \frac{1}{20}U(x+h,\,y+k)+\frac{1}{20}U(x-h,\,y+k)-\frac{1}{10}\frac{h^2-5k^2}{h^2+k^2}U(x+h,y)+\frac{5h^2-k^2}{h^2+k^2}$$

$U(x,y+k)$.

Separation of variables, see [6] for details, leads to the following cubic equation for $z = \sqrt{\mu}$

$$25z^3 - e_2(e_1^2f-10f+40)z^2 - (e_1^2e_2^2+e_1^2f^2-e_2^2f^2+8e_2^2f-16e_2^2)z - e_1^2e_2f = 0 \tag{5.5}$$

where $e_1 = \cos p\pi/m$, $e_2 = \cos q\pi/n$, $f = (5k^2-h^2)/(h^2+k^2)$. (5.6)

The roots of this equation have been investigated numerically by aid of the Telefunken TR4 computer of Groningen University for various values of the parameters f, e_1 and e_2. The parameter f is restricted to the interval $(-1,5)$ since h and k are real. It was found that $|z| = 1$ only occurs if at least one of the quantities e_1 and e_2 is equal to ± 1. Since, however, e_1 and e_2 are in absolute value always smaller than 1, the same holds for $|z|$, which means that there is always convergence.

The spectral radius of $\mathbf{C}^{(9)}$ again occurs for $p = q = 1$. For large values of m and n, we can obtain the deviation of z from 1 by substituting in eq. (5.5)

$$z = 1 + \delta z, \quad e_1 = 1 - \pi^2/2m^2, \quad e_2 = 1 - \pi^2/2n^2.$$

Neglecting higher powers of δz, $1/m^2$ and $1/n^2$, we find

$$\delta z = -\frac{3}{5}\frac{h^2k^2}{h^2+k^2}\pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right).$$

Hence, the convergence rate is

$$R(\mathbf{C}^{(9)}) = \frac{6}{5}\frac{h^2k^2}{h^2+k^2}\pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right) \tag{5.7}$$

which again is twice that of the Jacobi method in those cases where the latter method converges.

6. *The overrelaxation method.*

The overrelaxation method consists, in fact, of an extrapolation at each step of the result obtained by the Gauss-Seidel method. In matrixform

$$\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} + \omega\left\{\tilde{\mathbf{U}}^{(n+1)} - \mathbf{U}^{(n)}\right\},$$

where $\tilde{\mathbf{U}}^{(n+1)}$ is the Gauss-Seidel approximation (5.1)

$$\tilde{\mathbf{U}}^{(n+1)} = \mathbf{L}\,\mathbf{U}^{(n+1)} + \mathbf{R}\,\mathbf{U}^{(n)} + \mathbf{k}.$$

$\mathbf{U}^{(n+1)}$ in the right hand side of the latter equation is the vector of values obtained from the overrelaxation method. Elimination of $\tilde{\mathbf{U}}^{(n+1)}$ yields

$$\mathbf{U}^{(n+1)} = \mathbf{C}_\omega\,\mathbf{U}^{(n)} + \mathbf{k}_1 \tag{6.1}$$

where $\mathbf{C}_\omega = (\mathbf{I}-\omega\mathbf{L})^{-1}\left\{(1-\omega)\mathbf{I}+\omega\mathbf{R}\right\}$ and $\mathbf{k}_1 = (\mathbf{I}-\omega\mathbf{L})^{-1}\,\omega\mathbf{k}$. (6.2)

For convergence of the iteration the spectral radius of $\mathbf{C}_\omega$ should again be smaller than 1.

5-point formula.

The eigenvalues of $\mathbf{C}_\omega$ satisfy the equation

$$\mu(\mathbf{I}-\omega\mathbf{L})\mathbf{U} = \left\{(1-\omega)\mathbf{I}+\omega\mathbf{R}\right\}\mathbf{U}$$

or, written as difference equation for the 5-point formula (see eq. (2.1)):

$$\mu\left[U(x,y) - \omega\left\{\frac{k^2}{2(h^2+k^2)}U(x-h,y)+\frac{h^2}{2(h^2+k^2)}U(x,y-k)\right\}\right] =$$

$$(1-\omega)\,U(x,y)+\omega\left\{\frac{k^2}{2(h^2+k^2)}U(x+h,y)+\frac{h^2}{2(h^2+k^2)}U(x,y+k)\right\}. \tag{6.3}$$

Separation of variables leads to a quadratic equation in $z = \sqrt{\mu}$, viz

$$z^2 - \omega g z + \omega - 1 = 0 \tag{6.4}$$

where $g = \frac{k^2}{h^2+k^2}\cos\frac{p\pi}{m}+\frac{h^2}{h^2+k^2}\cos\frac{q\pi}{n}$. (6.5)

Further investigation of this case yields the well-known [1, 2] results

$$\omega_{opt} = \frac{2(1-\sqrt{1-g_1^2})}{g_1^2} \quad\text{and}\quad \lambda(\mathbf{C}_\omega^{(5)}) = \omega_{opt} - 1, \tag{6.6}$$

where g_1 is the value of g when $p = q = 1$ is substituted in eq. (6.5). For large values of m and n the results become

$$\left.\begin{aligned} &\lambda(\mathbf{C}_\omega^{(5)}) = 1 - t, \quad \omega_{opt} = 2 - t, \quad R(\mathbf{C}_\omega^{(5)}) = t \\ &\text{where } t = \frac{2\pi hk}{\sqrt{h^2+k^2}}\sqrt{\frac{1}{a^2}+\frac{1}{b^2}} \end{aligned}\right\} \tag{6.7}$$

Hence, the convergence speed is much better than in the case of the Gauss-Seidel method. If one takes m equal to n, $R(\mathbf{C}_\omega^{(5)}) = 2\pi/n$, while $R(\mathbf{C}^{(5)}) = \pi^2/n^2$.

9-point formula.

The equivalent to eq. (6.3) becomes

$$20(\mu+\omega-1)\,U(x,y) -$$

$$-\mu\omega\left\{U(x+h,\,y-k)+U(x-h,\,y-k)+2\frac{5h^2-k^2}{h^2+k^2}U(x,y-k)-2\frac{h^2-5k^2}{h^2+k^2}U(x-h,y)\right\} =$$

$$\omega\left\{U(x+h,\,y+k)+U(x-h,\,y+k)+2\frac{5h^2-k^2}{h^2+k^2}U(x,y+k)-2\frac{h^2-5k^2}{h^2+k^2}U(x+h,y)\right\}.$$

Separation of variables now leads to a quartic equation in $z = \sqrt{\mu}$, see [6], which is

$$25z^4-\omega e_2(\omega e_1^2 f-10f+40)z^3-(\omega^2e_1^2e_2^2+\omega^2e_1^2f^2-\omega^2e_2^2f^2+8\omega^2e_2^2f-16\omega^2e_2^2-50\omega+50)z^2 -$$

$$-\omega e_2(\omega e_1^2 f-10\omega f+10f+40\omega-40)z+25(\omega-1)^2=0. \tag{6.8}$$

The roots of this quartic have been investigated numerically for f = 2 (h = k) on the Telefunken TR 4 computer. The results are shown in fig. 1.

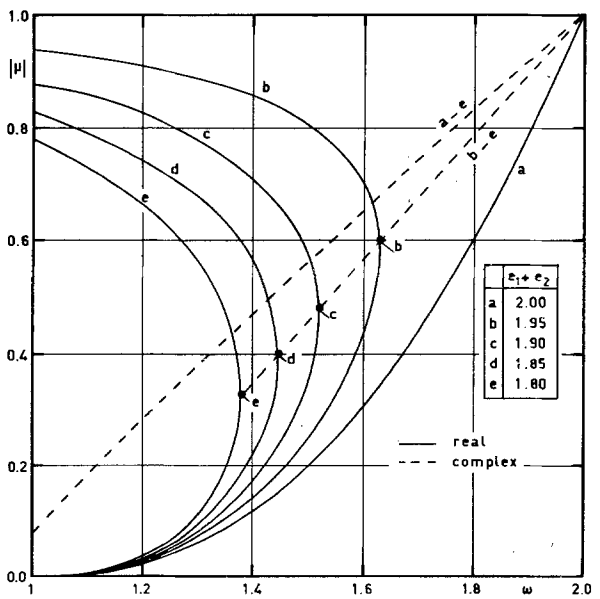

Fig. 1. Eigenvalue of $C^{(9)}$ for various values of $e_1 + e_2$. (h = k)

For e_1 and e_2 near 1, the roots appear to depend approximately only on the sum $e_1 + e_2$. In fig. 1 $|\mu|$ is given as a function of $\omega(1 \leqq \omega \leqq 2)$ for the values $e_1 + e_2 = 2,\ 1.95,\ 1.90,\ 1.85$ and 1.80. The drawn lines give real roots, the dotted lines give the moduli of complex roots. It is seen that there are either 2 or 4 complex roots.

The values of $|\mu|$ for the complex roots are for values of $e_1 + e_2$ between 1.8 and 2 nearly independent of $e_1 + e_2$. For smaller values of $e_1 + e_2$ the larger of the two $|\mu|$ - values corresponding to complex roots, decreases.

Hence, it can be concluded that $\omega = \omega_{opt}$ is equal to the ω-value of the point of intersection of the upper dotted line with the drawn line corresponding to the value of $e_1 + e_2$, with e_1 and e_2 obtained by taking $p = 1$, $q = 1$. For that value of $e_1 + e_2$ one pair of complex roots and one real root have the same modulus, while the other real root has a smaller modulus. Smaller values of $e_1 + e_2$ lead to smaller values of $|\mu|$.

We now shall investigate ω_{opt} for large values of m and n. At $\omega = \omega_{opt}$ the roots corresponding to the largest values of e_1 and e_2 that can occur, are $z = r$, $z = s$, $z = re^{+i\varphi}$ and $z = re^{-i\varphi}$ with $r > s$. The quartic (6.8) then should be of the form

$$(z - r)(z - s)(z - re^{i\varphi})(z - re^{-i\varphi}) = 0$$

or, putting $r(1 + 2\cos\varphi) = t$,

$$z^4 - (s + t)z^3 + t(r + s)z^2 - r(r^2 + st)z + r^3 s = 0. \tag{6.9}$$

We identify corresponding coefficients in (6.8) and (6.9) which yields

$$\left.\begin{aligned} \omega e_2(\omega e_1^2 f - 10f + 40) &= 25(s + t) \\ \omega^2 e_1^2 e_2^2 + \omega^2 e_1^2 f^2 - \omega^2 e_2^2 f^2 + 8\omega^2 e_2^2 f - 16\omega^2 e_2^2 - 50\omega + 50 &= -25t(r + s) \\ \omega e_2(\omega e_1^2 f - 10\omega f + 10f + 40\omega - 40) &= 25r(r^2 + st) \\ (\omega - 1)^2 &= r^3 s \end{aligned}\right\} \tag{6.10}$$

These are 4 equations for the four unknowns r, s, t and $\omega(= \omega_{opt})$.

It is seen from fig. 1 that for $e_1 = 1$, $e_2 = 1$, the following solution holds

$$\omega_{opt} = 2,\quad r = 1,\quad s = 1,$$

while the system of equations yields $t = \dfrac{55 - 16f}{25}$.

We now try to find the solution of the system for values of e_1 and e_2 which are slightly smaller than 1. This appears possible by making the following expansions

$$\left.\begin{aligned} e_1 &= 1 - \delta_1 \\ e_2 &= 1 - \delta_2 \\ \omega &= 2 - A_1\sqrt{\alpha\delta_1 + \beta\delta_2} + A_2\delta_1 + A_3\delta_2 + \ldots\ldots \\ r &= 1 - B_1\sqrt{\alpha\delta_1 + \beta\delta_2} + B_2\delta_1 + B_3\delta_2 + \ldots\ldots \\ s &= 1 - C_1\sqrt{\alpha\delta_1 + \beta\delta_2} + C_2\delta_1 + C_3\delta_2 + \ldots\ldots \\ t &= \frac{55 - 16f}{25} - D_1\sqrt{\alpha\delta_1 + \beta\delta_2} + D_2\delta_1 + D_3\delta_2 + \ldots\ldots, \end{aligned}\right\} \tag{6.11}$$

whence it will be clear that only the ratio α/β (and not α and β themselves) is of importance.

These expansions are substituted in eqs. (6.10). Identification of the

coefficients of $\sqrt{\alpha\delta_1 + \beta\delta_2}$ yields the system

$$\left.\begin{aligned} 2A_1(20 - 3f) &= 25C_1 + 25D_1 \\ 2A_1(55 - 16f) &= B_1(55 - 16f) + C_1(55 - 16f) + 50D_1 \\ 2A_1(60 - 13f) &= 2B_1(65 - 8f) + C_1(55 - 16f) + 25D_1 \\ 2A_1 &= 3B_1 + C_1 \end{aligned}\right\} \tag{6.12}$$

This homogeneous system has a matrix of rank 3 and the ratio of A_1, B_1, C_1 and D_1 can be determined. The result is

$$B_1 = \frac{1}{2}\cdot\frac{5 + 3f}{5 + 4f}A_1, \quad C_1 = \frac{1}{2}\cdot\frac{5 + 7f}{5 + 4f}A_1, \quad D_1 = \frac{275 + 85f - 48f^2}{50(5 + 4f)}A_1. \tag{6.13}$$

That A_1 is an independent parameter is due to the fact that the terms with $\sqrt{\alpha\delta_1 + \beta\delta_2}$ in ω, r, s and t are not yet affected by the change in e_1 and e_2. This means that one of these terms may be chosen with an arbitrary magnitude.

Identification of the coefficients of δ_1 gives the system

$$\left.\begin{aligned} &A_1^2 f\alpha + 2A_2(20 - 3f) - 8f = 25C_2 + 25D_2 \\ &A_1^2\alpha(15 - 8f) + 2A_2(55 - 16f) + 8f^2 + 8 = 25B_1D_1\alpha + 25C_1D_1\alpha + B_2(55 - 16f) + \\ &\qquad C_2(55 - 16f) + 50D_2 \\ &A_1^2\alpha(40 - 9f) + 2A_2(60 - 13f) - 8f = 75B_1^2\alpha + B_1C_1\alpha(55 - 16f) + 25B_1D_1\alpha + \\ &\qquad + 25C_1D_1\alpha + 2B_2(65 - 8f) + C_2(55 - 16f) + 25D_2 \\ &A_1^2\alpha + 2A_2 = 3B_1^2\alpha + 3B_1C_1\alpha + 3B_2 + C_2 \end{aligned}\right\} \tag{6.14}$$

Multiplication of these equations by -1, 1, -1 and 25 followed by addition yields

$$8f^2 + 16f + 8 = B_1C_1\alpha(20 + 16f). \tag{6.15}$$

Identification of the coefficients of δ_2 yields

$$\left.\begin{aligned} &A_1^2 f\beta + 2A_3(20 - 3f) + 16f - 80 = 25C_3 + 25D_3 \\ &A_1^2\beta(15 - 8f) + 2A_3(55 - 16f) - 8f^2 + 64f - 120 = 25B_1D_1\beta + 25C_1D_1\beta + B_3(55 - 16f) + \\ &\qquad + C_3(55 - 16f) + 50D_3 \\ &A_1^2\beta(40 - 9f) + 2A_3(60 - 13f) + 16f - 80 = 75B_1^2\beta + B_1C_1\beta(55 - 16f) + 25B_1D_1\beta + \\ &\qquad + 25C_1D_1\beta + 2B_3(65 - 8f) + C_3(55 - 16f) + 25D_3 \\ &A_1^2\beta + 2A_3 = 3B_1^2\beta + 3B_1C_1\beta + 3B_3 + C_3. \end{aligned}\right\} \tag{6.16}$$

The same linear combination as calculated for the system (6.14) now gives

$$-8f^2 + 32f + 40 = B_1C_1\beta(20 + 16f). \tag{6.17}$$

Comparison of eqs. (6.15) and (6.17) yields for the ratio α/β

$$\frac{\alpha}{\beta} = \frac{1 + f}{5 - f}. \tag{6.18}$$

Since $-1 < f < 5$ it is clear that this ratio is always positive as required in (6.11).

Eq. (6.15) gives

$$B_1 C_1 \alpha = 2\frac{(1+f)^2}{5+4f}.$$

Using the results obtained in (6.12) we may write also

$$A_1^2 \alpha = 8\frac{(1+f)^2(5+4f)}{(5+3f)(5+7f)}.$$

Substitution in (6.11) yields

$$\omega_{opt} = 2 - 2\sqrt{\frac{2(1+f)(5+4f)}{(5+3f)(5+7f)}}\cdot\sqrt{(1+f)\delta_1 + (5-f)\delta_2 + 0(\delta_1) + 0(\delta_2)}$$

$$\lambda(\mathbf{C}_\omega^{(9)}) = |\mu|_{max} = r^2 = 1 - 2\sqrt{\frac{2(1+f)(5+3f)}{(5+4f)(5+7f)}}\cdot\sqrt{(1+f)\delta_1 + (5-f)\delta_2 + 0(\delta_1) + 0(\delta_2)}$$

$$R(\mathbf{C}_\omega^{(9)}) = 2\sqrt{\frac{2(1+f)(5+3f)}{(5+4f)(5+7f)}}\cdot\sqrt{(1+f)\delta_1 + (5-f)\delta_2 + 0(\delta_1) + 0(\delta_2)}$$

With regard to the factor $5 + 7f$ which becomes negative for $f < -\frac{5}{7}$, we may remind the reader that it was assumed that $r \geq s$ since otherwise the foregoing considerations do not hold. From (6.11) it is clear that the condition $r \leq s$ is equivalent to $B_1 \leq C_1$ if terms of order δ_1 and δ_2 are neglected. It follows from (6.13) that $B_1 \leq C_1$ again is equivalent to $f \geq 0$.

Hence, the results obtained are valid only for $f \geq 0$ or $h \leq k\sqrt{5}$. Finally, substituting

$$f = -\frac{h^2 - 5k^2}{h^2 + k^2}, \qquad \delta_1 = 1 - \cos\frac{\pi}{m} \text{ and } \delta_2 = 1 - \cos\frac{\pi}{n}$$

the results can also be written as

$$\left.\begin{aligned} \omega_{opt} &= 2 - 6\pi hk^2\sqrt{\frac{h^2+25k^2}{(h^2+10k^2)(h^2+k^2)(-h^2+20k^2)}\cdot\left(\frac{1}{a^2}+\frac{1}{b^2}\right)} \\ R(\mathbf{C}_\omega^{(9)}) &= 12\pi hk^2\sqrt{\frac{h^2+10k^2}{(h^2+25k^2)(h^2+k^2)(-h^2+20k^2)}\cdot\left(\frac{1}{a^2}+\frac{1}{b^2}\right)} \end{aligned}\right\} \quad (6.19)$$

For a square ($a = b$) with equal meshsizes ($h = k$) we now derive results for ω_{opt} and R containing errors $0(h^3)$ instead of $0(h^2)$ as would be the case when using eq. (6.19).

Since then $e_1 = e_2 = e$, we make the expansions

$$\left.\begin{aligned} e &= 1 - \delta \\ \omega &= 2 - A\delta^{\frac{1}{2}} + A'\delta + A''\delta^{3/2} \\ r &= 1 - B\delta^{\frac{1}{2}} + B'\delta + B''\delta^{3/2} \\ s &= 1 - C\delta^{\frac{1}{2}} + C'\delta + C''\delta^{3/2} \\ t &= \frac{23}{25} - D\delta^{\frac{1}{2}} + D'\delta + D''\delta^{3/2} \end{aligned}\right\} \quad (6.20)$$

The relations between A, B, C, and D follow from eq. (6.13) by taking f equal 2

$$B = \frac{11}{26} A, \qquad C = \frac{19}{26} A \qquad \text{and} \qquad D = \frac{253}{650} A. \tag{6.21}$$

Identification of the coefficients of δ in the equations (6.10) leads to the system (compare (6.14) and (6.16))

$$\left.\begin{aligned} 2A^2 + 28A' - 64 &= 25C' + 25D' \\ -A^2 + 46A' + 16 &= 23B' + 23C' + 50D' + 25BD + 25CD \\ 22A^2 + 68A' - 64 &= 98B' + 23C' + 25D' + 75B^2 + 23BC + 25BD + 25CD \\ A^2 + 2A' &= 3B' + C' + 3B^2 + 3BC \end{aligned}\right\} \tag{6.22}$$

Multiplication of these equations by -1, 1, -1 and 25 followed by addition yields 13BC = 36. Hence A = 2.9928, B = 1.2662, C = 2.1871, D = 1.1649.

Three other linear combinations of eqs. (6.28) yield three relations between the four unknowns A', B', C', D'. We eliminate D' and obtain two relations for A', B' and C'. Multiplication of the equations by -2, +1, 0 and 27 and addition gives after substitution of the values obtained for A, B, C and D

$$11A' - 26B' = 28.4211.$$

The last equation of (6.22) becomes

$$2A' - 3B' - C' = 4.1605.$$

The third equation between A', B' and C' is obtained from identification of the coefficients of $\delta^{\frac{5}{2}}$ in eqs. (6.10). This gives the system

$$\begin{aligned} 44A - 4AA' + 28A'' &= 25C'' + 25D'' \\ -16A + 2AA' + 46A'' &= 23B'' + 23C'' + 50D'' - 25BD' - 25B'D - 25CD' - 25C'D \\ 84A - 44AA' + 68A'' &= 98B'' + 23C'' + 25D'' - 25B^3 - 150BB' - 25BCD \\ &\quad - 23BC' - 25BD' - 23B'C - 25B'D - 25CD' - 25C'D \\ - 2AA' + 2A'' &= 3B'' + C'' - B^3 - 3B^2C - 6BB' - 3BC' - 3B'C. \end{aligned}$$

Multiplication of the equations by 1, -1, 1 and -25 followed by addition leads to

$$144 A = 75B^2C - 25BCD + 52BC' + 52B'C$$

or, after substitution of the values for A, B, C and D

$$19B' + 11C' = 41.5385.$$

Finally we find

$$A' = 4.478, \qquad B' = 0.802, \qquad C' = 2.392.$$

With $\delta = \pi^2 h^2/2a^2$ this gives the results

$$\left.\begin{aligned} \omega_{opt} &= 2 - 2.116\ \pi h/a + 2.24\ (\pi h/a)^2 + 0(h^3) \\ \lambda(\mathbf{C}_\omega^{(9)}) &= 1 - 1.791\ \pi h/a + 1.60\ (\pi h/a)^2 + 0(h^3) \\ R(\mathbf{C}_\omega^{(9)}) &= 1.791\ \pi h/a + 0(h^3). \end{aligned}\right\} \tag{6.23}$$

These results can be compared with the corresponding results for the 5-point formula as given by eq. (6.6)

$$\left.\begin{aligned} \omega_{opt} &= 2 - 2\,\pi h/a + 2(\pi h/a)^2 + 0(h^3) \\ \lambda(\mathbf{C}_\omega^{(5)}) &= 1 - 2\,\pi h/a + 2(\pi h/a)^2 + 0(h^3) \\ R(\mathbf{C}_\omega^{(5)}) &= \quad 2\,\pi h/a + 0(h^3). \end{aligned}\right\} \qquad (6.24)$$

It may be noted that eq. (6.23) differs from a result obtained by Garabedian, viz.

$$\omega_{opt} = 2 - 2.04\,\pi h/a + 0(h^2).$$

There are a number of assumptions in Garabedian's work which are difficult to assess, but which apparently are not justified. It will be shown in the next section that the experimental results are in agreement with our results.

7. *Numerical results.*

First experiment.

The Dirichlet problem has been numerically solved for the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$ with boundary values taken from the function

$$u(x,y) = \log\left\{(x+1)^2 + y^2\right\},$$

which is a solution of Laplace's equation. The solution was obtained by aid of the overrelaxation method with both the 5-point and the 9-point formula. As soon as the solution obtained in two consecutive cycles differed less than 10^{-9} in all calculated points, the iteration was stopped. The values taken initially at all regular points of the mesh were equal to 0. In both directions the mesh lengths were taken equal to 1/30. For the 9-point formula the discretization error was smaller than 10^{-9} since in this accuracy the solution obtained, agreed with the values of the function $u(x,y)$.

The number N of cycles, which were performed before the iteration stopped, was investigated as a function of the relaxation factor ω and was compared between the 5-point and the 9-point formulae. The results are presented in fig. 2. For the 5-point formula N is minimal for $\omega = 1.810$ and at this value there is clearly a kink in the curve. Formula (6.24) gives for this case $\omega = 1.812$ with an error of order h^3. The exact formula for ω is given by eq. (6.6) and this yields $\omega = 1.811$. For the 9-point formula the minimal value of N occurs for $\omega = 1.807$, while the value given by eq. (6.23) is $\omega = 1.803$ with an error $0(h^3)$. Garabedian's formula yields $\omega = 1.786$, while eq. (6.23) yields $\omega = 1.778$ if in this formula the term $0(h^2)$ is also neglected. The agreement between the formulae (6.23) and (6.24) and the experimental values of ω is satisfactory. The exact value according to eq. (6.8) is $\omega = 1.801$.

The agreement with the convergence speeds $R(\mathbf{C}_\omega)$ is less good. According to eqs. (6.23) and (6.24) the convergence speed of the 5-point formula is larger than that of the 9-point formula, thus leading to a smaller value of N for the 5-point formula. This is not confirmed by the experiment. We therefore calculate the N values which correspond to the convergence speeds of eqs. (6.23) and (6.24). Assuming that the values of the solution are of the order 1, it follows that the number N of steps is given roughly by

$$\lambda^N \sim 10^{-9} \text{ or } NR(\mathbf{C}_\omega) \sim 9 \ln 10$$

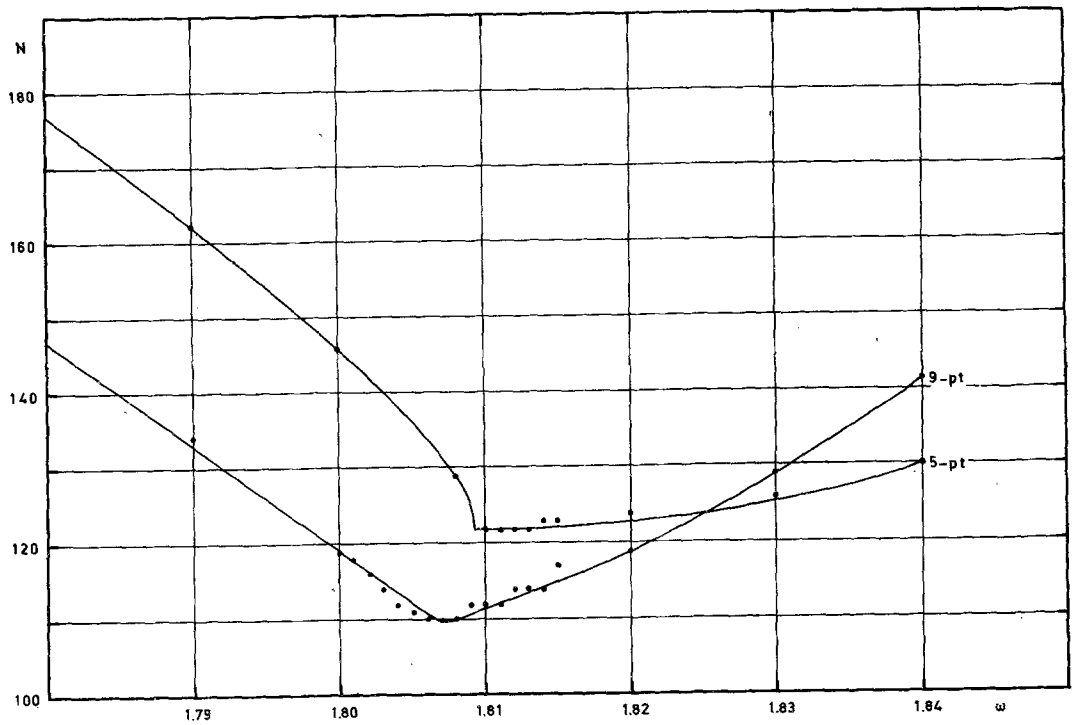

Fig. 2. Number of iterationsteps to obtain accuracy of 10^{-9}. Overrelaxation method. First experiment.

Using for $R(\mathbf{C}_\omega)$ the values of eqs. (6.24) and (6.23) we find

5-point formula	$N \sim 99.$	(experimentally 122)
9-point "	$N \sim 115.$	(" 110)

The reason for the discrepancy is that the results of eqs. (6.23) and (6.24) are asymptotic results, that is they only hold after a large number of steps. The experimental results give an average convergence rate during N steps and this differs from the asymptotic value. This conclusion is confirmed in the second experiment and there it is also shown that asymptotically the experimental and the theoretical convergence rate completely agree.

Second experiment [1].

The Dirichlet problem was numerically solved for the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$ with boundary values equal to 0. Since, in this experiment, we are only interested in the rate of convergence, which is independent of the boundary values, this choice is appropriate. The initial values at the regular points of the mesh were all taken equal to 1. Mesh lengths in both directions were 1/30. The solution approaches 0 and since the calculations were performed in floating decimal point allowing a smallest number of 10^{-153}, they could be continued almost arbitrarily long.

The spectral radius of the matrix $\mathbf{C}_\omega$ was calculated from the formula

$$\log \lambda(\mathbf{C}_\omega) \sim \frac{\log \| \mathbf{C}_\omega^{n+p} \mathbf{x} \| - \log \| \mathbf{C}_\omega^{n} \mathbf{x} \|}{p}, \qquad (7.1)$$

1) This experiment has been suggested by Mr. H.J. Burema, who also gave the derivation of formula (7.1).

where $\mathbf{x}$ denotes the vector of the initial values at the regular points. The vectornorm used is the norm, which is equal to the sum of the absolute values of all components. In fact, formula (7.1) expresses that after sufficient iterationsteps the norm of the vector decreases in each step by a factor λ. A derivation of the formula is given in [6].

Formula (7.1) has been evaluated for various values of n and p. For increasing n and p the formula will give a better approximation for the spectral radius λ. This experimental value of λ has been compared with the theoretical value obtained from the equation for $z = \sqrt{\mu}$ (eq. (6.4) for the 5-point formula and eq. (6.8) for the 9-point formula). The root of this equation which is largest in modulus is the spectral radius λ.

For $\omega < \omega_{opt}$ the final experimental value of $\lambda(\mathbf{C}_\omega)$ was usually reached if n+p was larger than 150. This value agreed with the theoretical value. For $\omega > \omega_{opt}$ the experimental value of $\omega(\mathbf{C}_\omega)$ oscillated around the theoretical value if n+p was taken sufficiently large, say 500. This is due to the fact that then there are many complex eigenvalues of the matrix $\mathbf{C}_\omega$ which all have nearly largest modulus. These are not only the two complex roots of largest modulus given by eq. (6.8), but also complex roots of the same equation for smaller e_1 and e_2, which correspond to other p and q (see eq. (5.6)) and which roots have also the same modulus according to fig. 1.

In all cases the experimental value of $\lambda(\mathbf{C}_\omega)$ obtained for small n and p is larger than the final value. This means that the convergence rate at the beginning of the iteration is smaller than its asymptotic value. In particular this is true for the 5-point formula, which agrees with the fact that in the first experiment the number of iterations necessary for obtaining a certain accuracy was larger than would be expected from the eqs. (6.23) and (6.24).

Fig. 3 gives the spectral radius as function of the overrelaxationfactor ω for both 5-point and 9-point formula (h = 1/30).

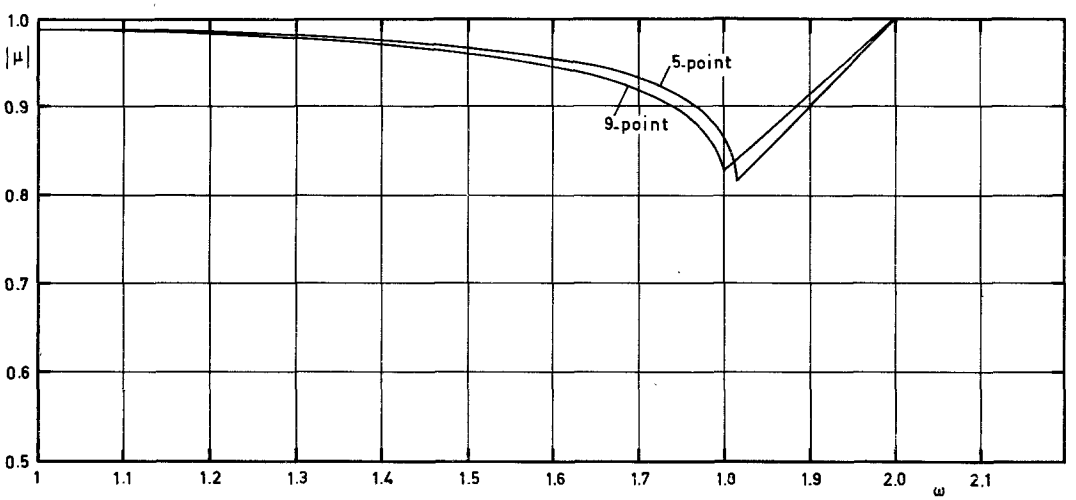

Fig. 3. Spectral radius of C_ω for unit square and h = k = 1/30.

Third experiment.

A final experiment has been conducted with regard to the maximal discretization error in the field. It has been shown by Gerschgorin [4] that if the fourth partial derivatives are bounded, the discretization error for the 5-point formula decreases as h^2 if h is the mesh size in both directions. In the same way it follows that, with partial derivatives up to the eighth order bounded, the discretization error decreases as h^6.

Since the discretization errors were very small for the harmonic function

taken in the first example, we now took as harmonic function

$$u(x,y) = e^{3x} \cos 3y.$$

Table 1 shows the results for the maximal discretization error in the field

Table 1

The maximal discretization error in the field as function of h for the 9-point formula

h	1/4	1/6	1/8	1/10	1/16
error	6400.10^{-8}	578.10^{-8}	99.10^{-8}	28.10^{-8}	2.10^{-8}

The error is proportional to h^6, which is in agreement with the theory.

as a function of h for the 9-point formula. This error is defined as the largest difference in absolute value between the exact harmonic function and the exact solution of the difference equations occurring somewhere in the field. It is clear from these results that the discretization error indeed decreases as h^6.

8. *References.*

1. G.E. Forsythe and W.R. Wasow: "Finite-difference methods for partial differential equations", Wiley and Sons, New York, 1960.

2. D.M. Young: "Iterative methods for solving partial difference equations of elliptic type", Trans. Amer. Math. Soc. Vol. 76, p. 92-111, 1954.

3. G. Birkhoff, R.S. Varga and D.M. Young: "Alternating direction implicit methods", Advances in Computers. Vol. 3, p. 190-273, 1962.

4. S. Gerschgorin: "Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen", Z.A.M.M. Vol. 10, p. 373-382, 1930.

5. P.R. Garabedian: "Estimation of the relaxation factor for small mesh size", Math. Tables Aids Comp. Vol. 10, p. 183-185, 1956.

6. A.I. van de Vooren and A.C. Vliegenthart: "On the 9-point difference formula for Laplace's equation". Report TW-42, Math. Institute, University of Groningen, 1967.

7. J.H. Wilkinson: "The algebraic eigenvalue problem", Clarendon Press, Oxford, 1965.

[Received May 25, 1967]