TRANSLATION
MATHEMATICAL PROBLEMS IN CREATING A DIGITAL GRID ANALYZER
By
M. A. Aleksidze

FOREIGN TECHNOLOGY DIVISION

AIR FORCE SYSTEMS COMMAND

WRIGHT-PATTERSON AIR FORCE BASE
OHIO
MATHEMATICAL PROBLEMS IN CREATING A DIGITAL GRID ANALYZER

BY: M. A. Aleksidze

English Pages: 29

A basic obstacle to achieving great accuracy in solving boundary problems for equations in mathematical physics by the grid method is the comparatively small memory of present electronic digital computers. As a result, it is necessary to divide the grid into parts during solution of these problems, and to return to the external storage upon each iteration, or to use Schwartz's alternating method. As was shown in note [2], the latter method is preferable when the grid region is divided into two parts. If the number of subsets of mesh points to be added is increased, the rate of convergence of Schwartz's iteration process decreases rapidly, and this method too becomes rather inefficient. We should note that when the region is divided into parts, programming difficulties appear as well. In view of what we have said, it is a very attractive idea to construct a specialized digital computer whose working storage uses either a magnetic drum [14] or closed magnetic tape [16], which could be used for problems involving Laplace, Poisson, wave, or diffusion equations. The mathematical prerequisite for this idea is simplicity of carrying through the iteration process, both with the very simple and with the nine-point grid approximation of the Laplacian, which enters into all the equations mentioned above.

The aim of the present article is to investigate the desirability from the viewpoint of machine time, of using various iteration pro-
cesses in specialized digital computers. As we shall show below, the parallel processing proposed earlier for all columns (rows), while substantially enlarging the apparatus, yields almost no advantage with the most expedient iteration process, the superrelaxation process.

In order to investigate the expediency of using the various iteration processes, it is necessary to know the asymptotic relationships between the local and integral errors for the different iteration processes. These relationships are derived in §1.

In §2, we examine the various iteration processes from the viewpoint of convenience of the calculations set up and machine time.

In §3, we discuss one grid approximation for the Laplacian operator, and one effective iteration process for solution of the first boundary problem is proposed.

In §4, we derive relationships between the computer digit capacity and the size of the working storage that are optimum from the viewpoint of a difference solution to the boundary problems.

§1. Asymptotic Relationships Between Local and Integral Errors in Solving Boundary Problems

1. We know that the difference equations obtained by substituting difference relationships for elliptical differential operators are, as a rule, solved by an iteration method. Only residues can be used to find the end of the iteration process and the integral error in the approximate solution to the difference equations. Nonetheless, the literature contains no discussions of problems of the asymptotic relationship between the residue and the integral error in various iteration processes. Gershgorin [8] derived an accurate estimate of the error in terms of the residue for the solution of difference equations by the majorant method. For the simplest grid approximation of the La-
placian differential operator, the Gershgorin estimate takes the following form

$$|s| \leq \frac{M_u h^2}{4 h^2},$$

(1.1)

while for the nine-point approximation

$$|s| \leq \frac{M_u h^2}{24 h^2},$$

(1.2)

where $M_u$ is the maximum residue, $l$ is the radius of a circle intersecting the given region as a whole, $h$ is the mesh increment. It is clear from (1.1), for example, that where $l/h = 60$, the residue should be three orders of magnitude smaller than the anticipated accuracy $\varepsilon = 10^{-k}$, and this cannot be ensured if the machine contains fewer than $(k + 3)$ decimal digits, to say nothing of the fact that this leads to protraction of the iteration process. In this paragraph, we shall derive asymptotic estimates for the various iteration processes, allowing fairly large residues in order to provide the required accuracy.

2. We write the finite-difference equivalent of the Dirichlet problem for the Laplace equation in matrix form

$$A\varphi = \psi,$$

(1.3)

where $\psi \neq 0$ only at the boundary mesh points. For an error of $\varepsilon = u - \bar{u}$, the equation $A\varepsilon = A(u - \bar{u}) = a_{1,j}$ or $\varepsilon = A^{-1}a_{1,j}$ will hold, where $a_{1,j}$ is the residue at the $(1,j)$ mesh point and $\bar{u}$ is an approximate solution to Eq. (1.3). We write $A^{-1}$ in the following form [12]:

$$A^{-1} = -\sum_{l=0}^{\infty} D^r D^r A^{-1},$$

where $D = E + rA$, $E$ is the unit matrix, $r$ is a relaxation factor so chosen that the spectral norm $\delta$ of operator $E + rD$ will be a minimum.
Since \( r \) may be so chosen that all characteristic values of operator \( D \) will have moduli less than unity, when \( n \to \infty \), \( D^n \to 0 \), and for \( \epsilon \), we obtain

\[
\|s\| \leq \sum_{j=0}^{\infty} (D)^j r s_j = \frac{r M_o}{1 - D}.
\]

(1.4)

We let \( p_h \) and \( q_h \) be the height and width of the smallest rectangle completely enclosing the given region. For the simplest approximation of the two-dimensional Laplacian operator, \( r, D \), and the corresponding estimate \( \epsilon \) will take on the following values:

For Richardson's iteration process, \( r = \frac{1}{4} \), \( D = 1 - \frac{\pi^2}{4} (p^{-2} + q^{-2}) \) and

\[
|s| \leq \frac{M_o p^4 q^4}{\pi^2 (p^2 + q^2)}.
\]

(1.5)

For Liebman's iteration process, \( r = \frac{1}{4} \), \( D = \left[ 1 - \frac{\pi^2}{4} (p^{-2} + q^{-2}) \right] \) and

\[
|s| \leq \frac{M_o p^4 q^4}{2\pi^2 (p^2 + q^2)}.
\]

(1.6)

For a superrelaxation iteration process (Liebman's method extrapolated [13]) \( r < \frac{1}{2} \), \( D = 1 - \sqrt{\frac{\pi^2}{2}} \sqrt{p^{-2} + q^{-2}} \) and

\[
|s| \leq \frac{M_o p q}{\pi \sqrt{2} \sqrt{p^2 + q^2}}.
\]

(1.7)

For a second-order Richardson process [13], \( r \) must be taken as the sum of the optimum values of the two parameters that determine the rate of convergence for this process. Then \( r < 1 \), \( D = 1 - \frac{\pi}{2} \sqrt{p^{-2} + q^{-2}} \) and

\[
|s| \leq \frac{\sqrt{2} M_o p q}{\pi \sqrt{p^2 + q^2}}.
\]

(1.8)

Estimates (1.5) and (1.6) permit larger residue values for ensuring the required accuracy than does the Gershgorin estimate (1.1) for any region, while the linear dimensions, in units of \( h \), appear in estimates (1.7) and (1.8) in first degree, which substantially reduces

- 4 -
the number of iterations.

3. The characteristic values of the nine-point approximation \( M \) to the Laplacian operator that have the largest and smallest moduli are, for the Richardson iteration process, equal to, respectively [5],

\[
M_n = -323 + 2x^2 (p^2 + q^2),
\]

\[
M_0 = -6x^2 (p^2 + q^2)
\]

while we obtain

\[
r = \frac{2}{M_n + M_0} \frac{1}{16 + 2x^2 (p^2 + q^2)}
\]

for the optimum value of \( r \).

For this value of \( r \), the spectral norm of the operator \( \|E + rM\| \) equals \( D = 1 - \frac{3}{8} \pi^2 (p^2 + q^2) \), and for the estimate \( \varepsilon \) we obtain

\[
|\varepsilon| = \frac{M_n p^2 q^2}{\pi^2 (p^2 + q^2)}.
\]  \( \text{(1.9)} \)

A Richardson iteration process of second order is carried out for the operator \( M \) in accordance with the following formula

\[
d_{i+1,j} = d_{i,j} + r \left( M d_{i,j} + r [d_{i,j} - d_{i+1,j}] \right).
\]

If \( r_1 \) and \( r_2 \) are calculated from the equations

\[
1 - r_1 M_0 + r_2 - 2 \sqrt{r_1} = 0
\]

\[
1 - r_1 M_0 + r_2 + 2 \sqrt{r_1} = 0
\]

then we obtain \( D = 1 - \frac{3}{2} \pi \sqrt{p^2 + q^2} \), which leads to the following estimate for \( \varepsilon \)

\[
|\varepsilon| = \frac{2 M_n p q}{\pi \sqrt{p^2 + q^2}}.
\]  \( \text{(1.10)} \)

4. Estimates (1.5)-(1.10), just as in the case of the Gershgorin estimates (1.1) and (1.2), cannot be used if drift operators are used for the boundary values in the region. Let us examine the following difference operators\(^*\)
\[
\begin{align*}
M_{1_{i,j}} &= \frac{\delta_{i+1,j}}{(1+\delta_{i+1,j})} u_{i+1,j} + \frac{\delta_{i,j+1}}{(1+\delta_{i,j+1})} u_{i,j+1} \\
&\quad + \frac{1}{2(1+\delta_{i+1,j})} u_{i+1,j} + \frac{1}{2(1+\delta_{i,j+1})} u_{i,j+1} - u_{i,j}
\end{align*}
\]

\[
\begin{align*}
M_{2_{i,j}} &= \frac{\delta_{i+1,j}}{\delta_{i+1,j}+\delta_{i,j+1}+1} \left[ -\frac{1}{1+\delta_{i+1,j}} (\delta_{i+1,j} - u_{i+1,j}) \\
&\quad + \delta_{i,j+1} \right] + \frac{\delta_{i,j+1}}{\delta_{i+1,j}+\delta_{i,j+1}+1} \left[ -\frac{1}{1+\delta_{i,j+1}} (\delta_{i,j+1} - u_{i,j+1}) \\
&\quad + \delta_{i+1,j} \right] - u_{i,j}
\end{align*}
\]

where

\[p, q = 0, \pm 1; r^2 + s^2 > 0; r, s = 0, \pm 1; r^2 + s^2 > 0;\]

\[\delta_{i+1,j}\] is the distance from mesh point (1, j) to the boundary \(\Gamma\) in the positive direction with respect to the Ox axis (to point A), divided by \(h\) (Fig. 1); \(u_{i+1,j}\) is the value of function \(u\) at point A on boundary \(\Gamma\); \(\delta_{i+1,j+1}\) is the distance from mesh point (1, j) to the boundary \(\Gamma\) in the direction of the bisector of the angle formed by lines \(x = x_i\) and \(y = y_j\) (to point B), divided by \(\sqrt{2h}\); \(u_{i+1,j+1}\) is the value of function \(u\) at point B on the boundary \(\Gamma\), etc.

![Fig. 1](image-url)
The following two lemmas were proved in [5]:

**Lemma 1.** If $M_{\nu}$ is the maximum modulus of the solution to a difference equation

\[
DV_{ij} = e_{ij}, \quad r_{ij} \equiv 2h,
\]
\[
\nu_{ij} = 0, \quad 0 < r_{ij} < 2h,
\]
\[
U_{ij} = 0,
\]

then $M_{\nu}$ is the maximum modulus of the solution to the difference equation

\[
DV_{ij} = e_{ij}, \quad r_{ij} \equiv 2h,
\]
\[
\nu_{ij} = 0, \quad 0 < r_{ij} < 2h,
\]
\[
U_{ij} = 0,
\]

and satisfies the inequality

\[
M_{\nu} \equiv (1 + \delta)^{t} \left[ M_{\nu} + \max_{i,j} (e_{ij}) \right]. \tag{1.11}
\]

**Lemma 2.** If $M_{\nu}$ is the maximum modulus of the solution to the difference equation

\[
DV_{ij} = e_{ij}, \quad r_{ij} \equiv 2h,
\]
\[
\nu_{ij} = 0, \quad 0 < r_{ij} < 2h,
\]
\[
U_{ij} = 0,
\]

then $M_{\nu}$ is the maximum modulus to the solution of the difference equation

\[
DV_{ij} = e_{ij}, \quad r_{ij} \equiv 2h,
\]
\[
\nu_{ij} = 0, \quad 0 < r_{ij} < 2h,
\]
\[
U_{ij} = 0,
\]

and satisfies the inequality

\[
M_{\nu} \equiv \left\{ \frac{(1 + \gamma)(1 + \delta)}{(1 + \delta) + \delta(1 + \delta)} \left[ M_{\nu} + \max_{i,j} (e_{ij}) \right] \right\} \tag{1.12}
\]

where $r_{i,j}$ is the minimum distance from mesh point $(i,j)$ to the boundary of the region. $\delta = \max_{i,j} [\min \{\gamma_{i,j}, \nu_{i,j}\}]$, $\gamma = \max_{i,j} [\min \{\gamma_{i,j}, \nu_{i,j}\}]$. Estimates (1.11) and (1.12) are accurate inasmuch as they are also valid for the asymp-
totic estimates. Substituting estimates (1.5)-(1.10) into (1.11) and (1.12) in place of \( M_y \), we obtain an estimate for the error where the operators \( D \) and \( M_1 \) or \( D \) and \( M_2 \) are employed simultaneously, for the various iteration processes.

5. The boundary problem for the rectangle \( p = 30, q = 32 \) was solved by the following two methods

\[
(A) \quad w^N = \frac{1}{4} \left[ u^{(n)}_{i,j} + u^{(n)}_{i+1,j} + u^{(n)}_{i,j+1} + u^{(n)}_{i+1,j+1} \right]
\]

\[
(B) \quad w^N = u^{(n)}_{i,j} + r [u^{(n)}_{i+1,j} + u^{(n)}_{i,j+1} + u^{(n)}_{i+1,j+1} - 4 u^{(n)}_{i,j}] ,
\]

where \( r = 0.452 \) is the optimum value of the relaxation factor for the given rectangle. In both cases, the iteration processes were continued until the residues became smaller than \( 10^{-5} \). The first method required 986 iterations, the second method 86. The maximum integral error in the first case was reached at the center of the rectangle, and equaled \( 2 \cdot 10^{-4} \), and in the second case it was found at the mesh point \((i = 15, j = 8)\), and equaled \( 2.2 \cdot 10^{-5} \). The results are in good agreement with estimates (1.6) and (1.7). If the Gershgorin estimate (1) were used to obtain an integral error of \( 2.2 \cdot 10^{-5} \), the residue would have to have been reduced to less than \( 6.2 \cdot 10^{-7} \), which would have required \( \frac{\ln [6 \cdot 10^{-1}]}{1.572 + 1.572} = 248 \) iterations of type (A) and \( \frac{\ln [6 \cdot 10^{-1}]}{16 \sqrt{2} \cdot \sqrt{1.572 + 1.572}} = 16 \) iterations of type (B). Consequently, in comparing the convergences of various iteration processes, in general the residues should not be reduced to the same size for all the iteration processes, since the same \( M_y \) yields different integral errors for the various iteration processes. We should also note that in iteration process (B), the approximation already is not aperiodic; thus, for example, if the boundary problem is solved for \( u = 1 \), and \( u_0 = 0 \) is taken as a first approximation, while the iteration process (B) is progressing, values \( u > 1 \) may appear at several mesh points, which
must be taken into account in solving problems on a fixed-point computer.

§2. Iteration Processes Connected with a Difference Solution to the First Boundary Problem

In calculating machine time, we shall assume that the working storage uses a magnetic drum with a capacity of 25,000 numbers, and a speed of 50 rps.

The author of [14] deals basically with the Richardson iteration process

\[ u_{n+1}^{(2)} = \frac{1}{4} \left[ u_{n+1}^{(1)} + u_{n+1}^{(2)} + u_{n+1}^{(3)} + u_{n+1}^{(4)} \right]. \]

In order to examine the desirability of this process, we shall discuss the two-dimensional Dirichlet problem for a square with side 150 h (22,500 mesh points). We shall assume that the initial residues are \( |a_{i,j}^{(0)}| \leq 1 \), where

\[ a_{i,j}^{(n)} = u_{i,j+1}^{(n)} + u_{i-1,j}^{(n)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n)} - 4u_{i,j}^{(n)}. \]

For the solution it is necessary that the integral errors be less than \( 10^{-6} \). As was shown in §1, in order for the Richardson iteration process to provide an integral error of \( 10^{-6} \) in the square under consideration, it is necessary that the local errors, i.e., the residues, be brought to a magnitude of \( \frac{2 \cdot 10^{-4} \cdot \pi^2}{150^2} = 0.8 \cdot 10^{-6} \). The spectral norm \( \lambda \) of operator \( A \) equals \( \lambda = 1 - \frac{\pi^2}{2 \cdot 150^2} \), while the number of iterations required is \( N = \frac{\ln[0.8 \cdot 10^{-6}]}{\ln \lambda} \approx 153,500 \). If each column is processed individually, then one iteration will require 150 cycles, and the solution will take about 86 hours. If all of the columns have processed in parallel, about 150 sequential-type single-digit accumulators will be required, and the calculation time will be about 40 minutes. The Richardson method is clearly best suited to magnetic tape, since the writing of a new value into an old position creates additional diffi-
oulties. On magnetic tape, however, it is difficult to set up such a large number (150) of tracks. Thus, we shall not discuss magnetic tapes in the future.

For the Liebman iteration process

\[ u_j^{(n)} = \frac{1}{4} \left[ u_j^{(n-1)} + u_j^{(n-2)} + u_j^{(n-3)} + u_j^{(n-4)} \right] \]

the spectral norm of operator A for the problem under consideration will be \( \lambda = \sqrt{\frac{1}{2} \cdot 150^2} \), the local errors must be reduced to \( 1.7 \cdot 10^{-9} \), and we obtain \( N = \left\lfloor \frac{\ln(1.7 \cdot 10^{-9})}{\ln \lambda} \right\rfloor = 50,000 \) for the number of iterations. With sequential processing of the columns, this will take about 42 hours. With parallel processing of the columns, a Liebman iteration process cannot be carried out, since the \( (k+1) \)th approximation of the functions at mesh point \( (i,j-1) \) is still unknown when the \( (k+1) \)th approximation is being computed for the function at mesh point \( (i,j) \). In this case, as a result, only the following iteration process is possible

\[ u_j^{(n)} = \frac{1}{4} \left[ u_j^{(n-1)} + u_j^{(n-2)} + u_j^{(n-3)} + u_j^{(n-4)} \right]. \] (2.1)

In order to find the spectral norm for operator A with iteration process (2.1), we examine the following iteration process

\[ u_j^{(n)} = u_j^{(n-1)} + \alpha [u_j^{(n-2)} + u_j^{(n-3)} + u_j^{(n-4)} - u_j^{(n-5)}], \] (2.2)

which is the same as (2.1) when \( \alpha = 1/4 \).

Let us seek the characteristic functions in the form

\[ u_j^{(n)} = u_j \sin \frac{x_i}{150} \sin \frac{x_j}{150}. \] (2.3)

We write (2.2) as follows:

\[ u_j^{(n)} = u_j + \alpha [u_j^{(n-1)} + u_j^{(n-2)} + u_j^{(n-3)} - u_j^{(n-4)}]. \] (2.4)

Substituting (2.3) into (2.4), we obtain
We find the arbitrary constant $R$ from the condition $A = R^2$; we then obtain

$$R^2 = a t_2 R^{-1} + 4 a - 2 a t_2 - c,$$

where

$$t_2 = \cos \frac{\pi r}{150}, \quad t_s = \cos \frac{\pi s}{150};$$

Where $\alpha = 1/4$, the largest characteristic value corresponds to mesh point $r = s = 1$, and equals

$$\Delta = 1 - \frac{2}{3} \left( \frac{\pi}{150} \right)^3.$$

Iteration Process (2.1) will in this case converge $4/3$ times faster than Richardson's iteration process, and $2/3$ times more slowly than Liebman's iteration process. With parallel processing of all columns, and with iteration Process (2.1) carried out, the machine time will be about 25 minutes.

The Richardson and Liebman iteration processes have an important advantage over other iteration processes in that they do not require multiplication or division (division by 4 is accomplished by a shift during write-in).

The computer must be capable of multiplication, however, in order to process boundary mesh points for an arbitrary region, and in order to solve boundary problems in the case of selfadjoint differential op-
We shall now discuss the superrelaxation iteration process

\[ u^{(r,s)} = u^{(r-1,s-1)} + \frac{1}{4} \left[ a^{(r-1,s-1)} + a^{(r-1,s)} + a^{(r,s-1)} + a^{(r,s)} - 4 u^{(r,s)} \right] \quad (2.5) \]

for such a machine. The optimum value of the relaxation factor \( \alpha \) is found from the equations

\[ 4 \frac{a^2 \cos^2 \frac{x}{2}}{150 - 4a + 1} = 0 \]

and is approximately equal to

\[ \alpha_{\text{opt}} = \frac{1}{2} - \frac{\pi}{300} \]

As close as this value is to one-half, systematic use in (2.5) of the value \( \alpha = 1/2 \) leads to divergence of the iteration process. The spectral norm for the operator \( ||E - \alpha_{\text{opt}} A|| \) equals

\[ \lambda = 1 - \frac{\pi}{75} \]

If a superrelaxation iteration process is used to solve the problem under consideration, the residues must be reduced to values of 0.8 \( \cdot 10^{-7} \) and the number of iterations will be

\[ N = \frac{\ln (0.8 \cdot 10^{-7})}{\ln \lambda} = 350. \]

A solution will take about 18 minutes when the columns are processed sequentially.

A superrelaxation iteration process with parallel processing of the columns is impossible. Consequently, we shall examine the iteration process

\[ u^{(r,s)} = u^{(r-1,s-1)} + \frac{1}{4} \left[ a^{(r-1,s-1)} + a^{(r-1,s)} + a^{(r,s-1)} + a^{(r,s)} - 4 u^{(r,s)} \right] \quad (2.6) \]

with the optimum parameter \( \alpha \).

An increase in \( \alpha \) from 1/4 to 1/2 will decrease the characteristic value \( A \) \((r = 1, s = 1) \) at mesh point \((r = 1, s = 1)\), but the modulus
of the complex characteristic value at the mesh point \( r = 1, s = 149 \)
will rapidly begin to increase
\[
A(r=1; s=149) = (a_0 + i \sqrt{2a_1 + 4s - 3a_1^2 - 1})^l,
\]
where
\[
t_1 = \max t_1; t_4 = \max t_4.
\]

At \( \alpha = 1/3 \), the modulus of the characteristic value \( A(r = 1, s = 149) \) is already the spectral norm of the operator
\[
\left( \left[ \frac{1}{3} t_4 + i \sqrt{\frac{2}{3} t_4 + 1 - \frac{1}{9} a_1} \right] \right)^4 = 1 - \frac{a_1^2}{3.15a_1^2}.
\]
The optimum \( \alpha \) is calculated from the following equation:
\[
a_4^2 + \sqrt{a_4^2 + 2a_4^2 + 4a_4^4} - \sqrt{2a_4^4 + 4a_4^6} = 0.
\]

Squaring both sides of this equation, we obtain a third-degree equation in \( \alpha \)
\[
a^3(2a_4^2 + 4a_4^4 - a^2(16 + 8a_4^2 - 1) = 0.
\]
The discriminant \( D \) of this equation equals
\[
D = \left[ \frac{(a_4 + 16)^3}{2(2a_4^2 + 4a_4^4)} + \frac{4(2a_4^2 + 16)}{3(2a_4^2 + 4a_4^4) - 4a_4^2 + 8a_4^4} \right] +
\]
\[
+ \left[ \frac{24(2a_4^2 + 4a_4^4)(2a_4^2 + 16)^3}{9(2a_4^2 + 4a_4^4) - 4a_4^2 + 8a_4^4} \right] = 0, \quad \text{for} \quad (140 - b)
\]
and can be shown to be negative, which leads to inconveniences in the use of Cardan's formulas. In contrast to the superrelaxation method, on the other hand, an accurate optimum value for \( \alpha \) will in this case yield an insignificant increase in the rate of convergence. Since
where \( \alpha = 3/10 \) the spectral norm for the operator \( \|E - 0.3A\| = \Xi \)
will equal
\[
A = 1 - \frac{6}{7} \left( \frac{\pi}{150} \right)^3,
\]
and the iteration process converges 9/7 times more rapidly than when

- 13 -
\[ \alpha = 1/4. \]

Thus, when iteration Process (2.6) is executed with an optimum relaxation parameter, the machine time will be about 17 minutes.

A second-order Richardson iteration process has a convergence of about the same order as a superrelaxation process, but is inconvenient to carry out, since it requires the two preceding approximations to find a new approximation and, consequently, twice as large a memory.

Finally, we can say that the most expedient iteration process for a specialized digital computer is superrelaxation with sequential processing of columns. If the computer cannot execute the multiplication operation, then it is best to execute iteration Process (2.5) using \( \alpha = 1/4 \) and \( \alpha = 1/2 \) alternately.

§3. One Grid Approximation for the Laplacian Differential Operator

Let us consider the following finite-difference equivalent for the Laplacian differential operator \( L \)

Let us examine the various iteration processes using the operator \( s \).

Let us assume that the solution is sought in a rectangle measuring \( ph \) and \( qh \). We use the following formula for the Richardson iteration process:

The characteristic functions of the operator \( s \) will then be

\[ s_{i,j} = \sin \frac{\pi i}{p} \sin \frac{\pi j}{q}. \]

We now calculate the corresponding characteristic values \( s(i,j) \);
\[-4 \sin \frac{\pi r}{p} \sin \frac{\pi j}{q} = 4 \left( \cos \frac{\pi r}{p} \cos \frac{\pi j}{q} - 1 \right) \sin \frac{\pi r}{p} \sin \frac{\pi j}{q} \]

\[r^{(\pi)} = 4 \left( \cos \frac{\pi r}{p} \cos \frac{\pi j}{q} - 1 \right).\]

The spectral norm of this operator will be
\[\max_{r \in \mathbb{R}} |r^{(\pi)}| = \pi = 4 \left[ \cos \frac{\pi}{p} \cos \frac{\pi(q-1)}{q} - 1 \right] = -\pi^2 (p^{-1} + q^{-1}).\]

One of the minimum characteristic values is obtained for the point \( r = s = 1 \), and equals
\[\min_{r \in \mathbb{R}} |r^{(\pi)}| = \pi = 4 \left[ \left( 1 - \frac{\pi^2}{2p^2} \right) \left( 1 - \frac{\pi^2}{2q^2} \right) - 1 \right] = -\pi^2 (p^{-1} + q^{-1}).\]

The optimum value for \( \alpha \) is obtained from the condition \( 1 + \cos \theta = -(1 + \cos \theta) \) and equals \( 1/4 \). The spectral norm for operator \( E = (1/4)s \) for the Richardson iteration process is
\[\lambda = 1 - \frac{\pi^2}{2} (p^{-1} + q^{-1}).\]

For the Liebman iteration process, the working formula is written as
\[s \eta^2 = s [w \eta^2 + \sqrt{s \eta^2} + \sqrt{s \eta^2} - 4 \eta^2].\]

we look for the characteristic functions in the form
\[s \eta^2 = \lambda \sin \frac{\pi r}{p} \sin \frac{\pi j}{q},\]

and we then will have
\[\begin{align*}
(r+4s-1) s \eta^2 = & s \sin \frac{\pi r}{p} \frac{\pi r}{p} \left[ S \lambda^{j+1} \sin \frac{\pi j}{q} + \\
& + S \lambda^{j+1} \sin \frac{\pi j}{q} + S \lambda^{j+1} \sin \frac{\pi j}{q} + \\
& + S \lambda^{j+1} \sin \frac{\pi j}{q} \right].
\end{align*}\]

Selecting the coefficient \( \lambda \) as follows: \( \lambda^2 = s \), we obtain
\[\begin{align*}
\lambda^2 = & 4 \lambda \cos \frac{\pi r}{p} \cos \frac{\pi j}{q} \\
& + 4 \lambda \cos \frac{\pi r}{p} \cos \frac{\pi j}{q}.
\end{align*}\]
For the characteristic values of the operator \( \mathbf{a} \), we have
\[
\lambda^2 - 4 \lambda \mathbf{d} + 4 \mathbf{a} = 0,
\]
where
\[
\lambda = \cos \frac{\pi f}{p} \cos \frac{\pi f}{q}.
\]

If we again set \( \alpha = 1/4 \) (ordinary Liebman iteration process), we obtain
\[
\lambda = \cos \frac{\pi f}{p} \cos \frac{\pi f}{q},
\]
and the spectral norm \( \| \) for the operator \( \mathbf{B} - (1/4) \mathbf{s} \) will equal
\[
1 = 1 - 4^2 (p^{-2} + q^{-2}).
\]

If \( \alpha \) is so chosen that \( \mathbf{A} \) has multiple roots (Liebman method extrapolated), we obtain for \( \alpha \) the condition
\[
4 \alpha^3 \lambda^2 + 4 \alpha + 1 = 0,
\]
whence
\[
\alpha = \frac{1}{2} \sqrt[4]{\frac{1}{2}^2 + \frac{1}{2}^2 - 4 \lambda^2}.
\]

The second-order Richardson iteration process is based upon the analogy between the solution to the Dirichlet problem and the equation
\[
\frac{\partial^2 u}{(\partial + 2) + \partial u/\partial t} - \mathbf{L}u = 0,
\]
whose finite-difference equivalent may be written, for carrying out integration, as
\[
\frac{u(\mathbf{r} + 1) - u}{r_1} + r_1 \frac{u(\mathbf{r} - 1) - u}{r_2} + r_2 [u(\mathbf{r} + 1) - u]\frac{u(\mathbf{r} - 1) - u}{r_2},
\]
where \( r_1 \) and \( r_2 \) are relaxation factors whose optimum values are calculated from the equations [13]
\[
1 + r_1 = 2 \sqrt{r_2},
\]
\[
1 + r_2 = 2 \sqrt{r_1}.
\]

These values of the relaxation factors yield for the spectral norm of the corresponding operator \( \mathbf{a} \) a value
\[
\| \mathbf{a} \| = 16.
\]
Similar problems for the simplest grid approximation to the Laplacian operator $D$

$$D_{i,j} = u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}$$

were discussed in [13].

Comparison of the spectral norms for the operators shows that the Richardson and Liebman iteration processes converge twice as fast for the operator $s$ as they do for the operator $D$, while the extrapolated Liebman method and the Richardson iteration process of second order converge more rapidly by a factor of $\sqrt{2}$. It is not difficult to show that the operator $s$ yields two independent systems of equations

$$(A) u_{ij} = \phi_{ij} (i=1,2,\ldots,p-1; j=1,2,\ldots,q-1; i+j=2k;$$

$$r=0; p; \sigma=0, q; r+s=2h)$$

$$(B) u_{ij} = \phi_{ij} (i=1,2,\ldots,p-1; j=1,2,\ldots,q-1; i+j=2k+1; r=0; p; \sigma=0, q; r+s=2h+1),$$

where $k$ is an integer. We have used this property of the operator $s$ to reduce the number of arithmetic operations in the iteration process.

In order to check the results obtained, the Dirichlet problem was solved for the Laplace equation

$$\Delta u = 0, \quad u|_{\partial R} = f$$

in a rectangle with sides $30h$ and $32h$, using the following methods:

I. Liebman iteration process with operator $D$

II. Liebman iteration process with operator $s$

III. Superrelaxation iteration process with operator $D$ (the optimum relaxation factor equaled 0.452).

IV. Superrelaxation iteration process with operator $s$ (optimum relaxation factor equaled 0.461).

V. Superrelaxation iteration process with operator $s$ for system
(A) alone, then a single application of operator D to system (B) alone, and then method III.

The value of the function on the boundary \( f(s) \) varied within wide limits (four problems were solved in all), but the number of iterations for all of the problems nonetheless remained constant to within 4%. The calculations were halted when the local errors became less than \( 2^{-25} \). We should note, however, that as §1 implies, such a local error results in a different integral error for the various integration processes. Since the operator \( s \) is somewhat less accurate than the operator D, the solution obtained with method II was corrected by means of method I, and method IV by method III. The corresponding numbers of iterations required for such a correction are given in parentheses in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 число итераций</td>
<td>123</td>
<td>123 (9)</td>
<td>71 (3)</td>
<td>71 (3)</td>
<td></td>
</tr>
<tr>
<td>2 число арифметич. операций</td>
<td>( 163 \times 10^3 )</td>
<td>( 85 \times 10^2 )</td>
<td>( 163 \times 10^3 )</td>
<td>( 90 \times 10^2 )</td>
<td>( 46 \times 10^2 )</td>
</tr>
</tbody>
</table>

1) Number of iterations; 2) number of arithmetic operations.

The number of instructions necessary to obtain a new approximation for each mesh point is 12 with methods I and II, and 14 for methods III-IV-V. The fact that the difference in the number of instructions is so small is explained by the fact that the residue was calculated in all cases in order to determine the point at which the iteration process was to halt.

Comparison of these data with theoretical values indicated good agreement, although with methods III-IV, a somewhat smaller number of iterations were predicted. The calculations were carried out on the
§4. Electronic Digital Computer Parameters that Are Optimum from the Viewpoint of Boundary-Problem Difference Solutions

In order to design a specialized digital computer to solve boundary problems [14, 16] we may derive the most favorable relationship between the size of the computer operating storage and the number of digits in machine numbers. In this case, naturally, the indicators of the differential properties for that class of functions to which the solution for the problem under consideration belongs will enter into these relationships.

Let us examine the general $n$-dimensional second-order differential equation of elliptical type

$$
\frac{d^4 u}{dx_1^4} + \cdots + \frac{d^4 u}{dx_n^4} + \frac{d^4 u}{dx_a^4} + \cdots + \frac{d^4 u}{dx_b^4} + g(x_1, \ldots, x_n) u = f(x_1, \ldots, x_n)
$$

with the very general boundary conditions

$$
A_1(x_1, \ldots, x_n) \frac{du}{dx_1} + A_2(x_1, \ldots, x_n) \frac{du}{dx_2} + A_3(x_1, \ldots, x_n) u = \psi(x_1, \ldots, x_n)
$$

and let us write some finite-difference equivalent for the problem (4.1)-(4.2)

$$
P_{nm} = \sum_{j=1}^{g} \epsilon_{mj} - \left( \sum_{j=1}^{g} \epsilon_{mj} + G_j \right) u_j = F_j \quad (j = 1, 2, \ldots, N),
$$

where $g$ is the number of adjacent points, in the sense of the difference operator $p_n$ used; $N$ is the number of mesh points in the closed $n$-dimensional region $D$ for which the problem (4.1), (4.2) is stated. For $A_1 \neq 0$ and $A_2 \neq 0$, the points at which the lines generating the grid region intersect the boundary of the region are also regarded as mesh points.
As a rule, System (4.3) is solved by an iteration method. It can be shown that when the conditions \( g \leq 0, A_3 \leq 0, A_3 \) and \( g \) not simultaneously identically equal to 0 are fulfilled, the simple iteration process for System (4.3) converges. Let us write the general iteration formula for (4.3)

\[
\begin{align*}
\tilde{y}^q & = y^{q-1} + \sum_{i=1}^{j} C_{n,i} \tilde{y}^{q-1} \sum_{m=1}^{j} C_{m,i} \tilde{y}^{q-1} - \left( \sum_{p=1}^{j} C_{p,r} + G \right) y^{q-1} \\
& + \sum_{r=1}^{j} \alpha_{s,r} [\tilde{y}^{q-1} - y^{q-1}] + F_j (j=1, 2, \ldots, N),
\end{align*}
\]

where \( r_j \) and \( \alpha_{s,j} (s = 1, 2, \ldots, j) \) are the parameters of the iteration formula, and are so selected that the convergence will be as rapid as possible; \( j \) is the number of different approximations of the function that are used to reinforce convergence. It is assumed that the \((k+1)\)th approximation of the function is written at location \((k-1)\), and that all quantities entering into (4.4), with the exception of \( u_{1,j} \), are computed once at the beginning of the iteration process and then remain stored in the memory. It is not difficult to see that Iteration Formula (4.4) will then require \((q + e + 3)\) storage locations for each mesh point (each number occupies one location). If we let \( N_1 \) be the number of locations in the operating storage of the computer, and \( N_2 \) be the number of instructions in the iteration program [4], we obtain \((N_1 - N_2)/(q + 1 + 3)\) for the number of mesh points. With no loss of generality, we can consider \( D \) to be a unit \( n \)-dimensional cube; the mesh increment will then be \( \left( \frac{N_1 - N_2}{q + 1 + 3} \right)^{-1/n} \). We assume that \(|\tilde{u} - \tilde{u}| = \| M a^{1/n} = M \left( \frac{N_1 - N_2}{q + 1 + 3} \right)^{-1/n} \). According to the well-known Banach theorem on convergence for the method of successive approximations, the maximum residue in iteration Process (4.4) will reach a value 

\[
\left( 1 - M (1 - M^q \left( \frac{N_1 - N_2}{q + 1 + 3} \right)^{-1/n} \right)
\]

in order to provide an accuracy.
a, where \( \| p \| \) is the spectral norm for the corresponding difference operator.

We must also take into account rounding errors of two types: a) the error \( \varepsilon' \) due to rounding in the course of iteration Process (4.4); b) the error \( \varepsilon'' \) due to rounding in calculating the coefficients for Eq. (4.3).

The mathematical expectation for the mean-square error \( \varepsilon' \) in the case of the Dirichlet problem for the Laplace equation [1] after the \( k \)th iteration due to all of the preceding roundings will first go up as \( \ln k \), and will then approach asymptotically

\[
\frac{2}{\pi \ln k} = \frac{\ln \left( \frac{N_1 - N_2}{\varepsilon + \varepsilon + 1} \right)}{\varepsilon + \varepsilon + 1}.
\]

From this it follows that for the residues to be reduced to a value \( M \), rounding must be carried out in the right-hand digit of the number

\[
M \ln \left( \frac{N_1 - N_2}{\varepsilon + \varepsilon + 1} \right)^{2/\pi}.
\]

It is easy to verify that \( \varepsilon'' = \bar{u} - \bar{u} \) satisfy the following difference equation (\( \bar{u} \) and \( \bar{u} \) are solutions to (4.3) for the accurate and approximate values of the coefficients, respectively)

\[
\sum_{j=1}^{N} \xi(C_{\theta_j}) \varepsilon' = \left[ \sum_{j=1}^{N} \xi(C_{\theta_j}) + \xi(F_j) \right] \varepsilon'' \quad (j = 1, 2, \ldots, N),
\]

where \( \xi(\theta) \) is the rounding error in \( \theta \). The mathematical expectation for \( \xi(\theta) \) equals zero but, nonetheless, the probability estimate may yield a value that is too low for \( \varepsilon'' \) for the case in which \( g(x_1, \ldots, x_n) = \text{const.} \) and \( f(x_1, \ldots, x_n) = \text{const.} \). The error \( \varepsilon'' \) does not depend upon the method used to obtain a solution to (4.3).

We shall now show that for the two-dimensional Poisson equation, and for the Liebman and Richardson iteration processes,

\[
\left| \frac{\varepsilon'}{\varepsilon''} \right| \lesssim \left( \frac{\ln M}{M} \right)^{1/2}
\]

(the symbol \( \lesssim \) indicates that \( 0 < \left| \frac{\varepsilon'}{\varepsilon''} \right| \leq \infty \)).
Let us find Green's difference function $R(r,s)$ in the following manner: $R(r,s)$ for fixed indices $r$ and $s$ is a solution to the following difference equations:

$$D_{i,j}^g = \begin{cases} 1, & (i-r)+(j-s)=0, \quad r_{ij}=2h, \\ 0, & (i-r)+(j-s)=0, \quad r_{ij}=2h, \\ M, & 0<r_{ij}<2h, \\ R^{(m)}(r,s)=0. \end{cases}$$

The solution of the difference equations

$$Du_{i,j} = \xi(F_{i,j}), \quad r_{ij}=2h,$$

$$M, \quad 0<r_{ij}<2h,$$

$$u_{i,j}^{(m)}=0,$$  

(4.5)

can be represented in terms of the Green difference function $R(r,s)$ in the following manner:

$$u_{i,j} = \sum_r \sum_s R^{(r,s)}(F_{i,j}),$$  

(4.6)

For $R(r,s)$ in [5] we have the estimate

$$R^{(r,s)} = \frac{(r+s)^2}{2} \ln \left( \frac{\sqrt{2} \cdot s}{h} \right),$$  

(4.7)

where $a$ is the side of a square completely enclosed by the given region (the estimate given in [11] is used in deriving (4.7)). Substituting (4.7) into (4.6) and assuming that the number of mesh points is of the order of $h^{-2}$, we obtain

$$u_{i,j} = \frac{B}{h^2 \ln h},$$  

(4.8)

where $B$ is some constant that does not depend upon $h$. For the Liebman and Richardson iteration processes, $\|p\| = 1 - ch^2$ and, therefore,

$$\frac{\epsilon}{\|p\|} = \frac{c_1}{h^2 \ln h},$$  

(4.9)

where $c_1$ also does not depend upon $h$. We obtain (4.5) from (4.8) and (4.9).
Finally, we obtain for the optimum number of digits in the mantissa of the number for fixed $N$

$$
\gamma = \sum_{i=1}^{d} \ln \left( \frac{M(1-|\mu|)^{n^{i-1}} \left( \frac{N_{2}-N_{1}}{y+1} \right)^{i}} {\ln y} \right), \quad (4.10)
$$

(where $y$ is the radix of the number system used in the computer. We assume that by transforming Eqs. (4.1)-(4.2), the solution of the difference equations will be $|\bar{u}| \leq 1$ (for example, for the Dirichlet problem, we can use the principle of the maximum to do this). This imposes definite limitations upon the coefficients of Eqs. (4.1), (4.2).

Let us examine some concrete cases: $N_{1} - N_{2} = 10,000$, $y = 2$, a superrelaxation iteration process is used, $a_k = 1$, $b_k = 0$, ($k = 1, 2, \ldots, n$),

$$
f \equiv 0, A_{i} \equiv A_{k} \equiv 0, |M| < 1, \frac{1}{r_{j}} = \text{const} = \frac{1}{n^{i-1}+b}, |\mu| = 1 - 2^{1-x} [2]:
$$

then we obtain from (4.10)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$s=1$</th>
<th>$s=2$</th>
<th>$s=3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d=2$</td>
<td>$\gamma_{\text{err}}$</td>
<td>56</td>
<td>36</td>
</tr>
<tr>
<td>$d=4$</td>
<td>$\gamma_{\text{err}}$</td>
<td>83</td>
<td>53</td>
</tr>
</tbody>
</table>

In practice, evidently, we most frequently encounter problems in which the boundary function $f \in H(P, A, \lambda)$, where $P + \lambda \leq 2$. For the Dirichlet problem, with the Laplace equation, the best estimates are contained in [6, 7]. In particular, it is shown in [6] that the error is $O(h^{p+\lambda})$ for the simplest approximation to the Laplace operator, with $p + \lambda < 2$. It follows from [9] that asymptotically ($h \to 0$), the order of these estimates cannot be improved. It has been shown in [15] that if a discontinuity of the boundary function occurs at a mesh
point, the error will be $O(h^3)$. The fastest processes from the point of view of the number of iterations and simplicity of the apparatus needed for the calculation scheme is the superrelaxation iteration process. It follows from what we have said that when $|\lambda| \leq 1$, for a specialized digital computer of the types discussed in [14, 16], the relationship between $\gamma_{opt}$ and $N$ will be ($N_2 = 0$, since these computers use a schematic program)

$$
\gamma_{opt} = E\left[ \frac{3\ln N_1 + 3\ln^2 N_1 + 2(1+\ln n)}{\ln 2} \right].
$$

In this case, it is assumed that the algorithm given in [3] is used for the Poisson equation.

If no provision for rounding is made in the computer, then the mathematical expectation for the error $\varepsilon$ due to discarding of the right-hand digits will satisfy Eq. (4.5). If we derive the estimate $\varepsilon$ in a manner similar to that used for $\varepsilon''$, we obtain for such a computer

$$
\gamma_{opt} = E\left[ \ln \left( \frac{3\ln M - \ln(1+\varepsilon N^{1/3})}{\ln e} \right) \right].
$$

In conclusion, we wish to thank Ye.A. Volkov for his helpful comments and advice.

Submitted 29 November 1958
Inequality (1.4) also follows at once from the familiar Banach convergence theorem for the method of successive approximations, if $A$ is regarded as an operator that transforms the space $\mathbb{R}^n$ into itself. In the case of a space $m$, $\|A\| = 1$, but within this space the condition $\|A\| < 1$ is sufficient but not necessary for convergence of the successive approximations.

$M_1$ and $M_2$ are the drift operators for the function boundary values according to Mikeladze [10], where one or two neighboring points, respectively, are missing at the boundary point. In this notation, however, there is a possibility that the boundary values will drift not only along the lines forming the grid region, but also along lines parallel to the bisectrices of the coordinate angles (for $|p| = |q| = 1$). They are essentially non-symmetrical difference approximations of the Laplacian differential operator and have a local error of the order of $1(h^3)$ under certain conditions.

In Lemmas 1 and 2, the angles at which $r_{1,j} < 2h$ are regarded as boundaries. Such "expansion" of the boundary-point zone is carried out because the following premise can be demonstrated: If the boundary of a region is a closed Lyapunov curve, then there exists an $h_0 > 0$ such that for any $h \leq h_0$ at any mesh point where $r_{1,j} < 2h M_1$ or $M_2$ may be written in at least one notation so that $\delta_{i+p, j+q} \leq 2.5$ for $M_1$ and $\delta_{i+p, j+q} \leq 2.5 \geq \delta_{i+r, j+s}$ for $M_2$. The proofs of Lemmas 1 and 2 do not rely on this assumption and if more rigid requirements are imposed on the boundary of the region and the boundary belt is "narrowed" to $r_1 < h$, then $\delta \leq 1 \geq \delta$, which improves the evaluations (1.11) and (1.12).


Laplace Equation], Dissertation, Moscow State University, 1958.


14. Leondes, C. and Rubinoff, DJNA, a digital analyzer for Laplace, Poisson, diffusion, and wave equations. Trans. AIEE, F. 1,

### DISTRIBUTION LIST

<table>
<thead>
<tr>
<th>DEPARTMENT OF DEFENSE</th>
<th>MAJOR AIR COMMAND</th>
<th>Nr. Copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEADQUARTERS USAF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFCIN-3D2</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>ARL (ARB)</td>
<td>SCFDD</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ASTIA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TDBTL</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TDBDP</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ASD (ASYIM)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SSD (SSF)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>APGC (PGP)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ESD (EST)</td>
<td>1</td>
</tr>
<tr>
<td>OTHER AGENCIES</td>
<td>AEC</td>
<td>1</td>
</tr>
<tr>
<td>CIA</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>NSA</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>DIA</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>AID</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>OTS</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>AFSWC (SWP)</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>AFMT (MTW)</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>AEDC (AEC)</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>AFFTC (FAT)</td>
<td>AFSC</td>
<td>1</td>
</tr>
<tr>
<td>RDFA</td>
<td>AFSC</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total: 29 copies**