The Numerical Solution of Elliptic Partial Differential Equation

Takashi Yoshimura

1. Iterative Methods

In the numerical solution by differences of boundary value problems involving elliptic partial differential equations, one is led to consider linear systems of high order of the form

$$\sum_{j=1}^{N} a_{i,j}\, u_j + d_j = 0 \qquad (i=1,2,\ \ldots\ ,N) \tag{1}$$

where $u_1, u_2, \ldots, u_N$ are unknown and where the real numbers $a_{i,j}$ and d_i are known.

For linear systems of the size encountered in practise the Gauss elimination method are not practical. One is led, instead, to consider iterative methods.

The simplest of the iterative methods is the Gauss–Seidel method, where, starting with arbitrary initial approximation to the solution, one improves this approximation using improved values as soon as available. Thus the improvement formula is

$$u_i^{(m+1)} = \Big(-\sum_{j=1}^{i-1} a_{i,j}\, u_j^{(m+1)} - \sum_{j=i+1}^{N} a_{i,j}\, u_j^{(m)} - d_i\Big) / a_{i,i}\,. \tag{2}$$

A complete iteration consists of improving the approximate values for all unknowns. Having traversed all the unknowns, one starts over again at the "first" unknown and repeat the process until $d_m < \varepsilon$, where ε is a prescrived tolerance and where

$$d_m = \sum_{i=1}^{N} \left| u_i^{(m)} - u_i^{(m-1)} \right| . \tag{3}$$

In the Jacobi method one does not use improved values until after a complete iteration. The improvement formula for this method is

$$u_i^{(m+1)} = \sum_{\substack{j=1 \\ j\neq i}}^{N} b_{i,j}\, u_j^{(m)} + c_i \tag{4}$$

where

$$b_{i,j} = \begin{cases} -a_{i,j}/a_{i,i} & (i\neq j) \\ 0 & (i=j) \end{cases} \tag{5}$$

and

$$c_i = -d_i / a_{i,i} \qquad (i=1,\ 2,\ \ldots\ ,N)\,. \tag{6}$$

By a simple modification of (2) we can make a substantial improvement in the rate of convergence. We use the following formula :

$$u_i^{(m+1)} = \omega \left\{ \sum_{j=1}^{i-1} b_{i,j}\, u_j^{(m+1)} + \sum_{j=i+1}^{N} b_{i,j}\, u_j^{(m)} + c_i \right\} - (\omega-1)\, u_i^{(m)} . \tag{7}$$

Here ω is a parameter known as relaxation factor. This method is known as the successive overrelaxation method. Evidently, where $\omega=1$, the successive overrelaxa-

tion method reduces to the Gauss–Seidel method.

We may write (1) in matrix notation

$$Au + d = 0, \tag{8}$$

where A is a given $N \times N$ matrix and d is a given column matrix and where u is an unknown column matrix.

Jacobi method may be written in the form

$$u^{(m+1)} = Bu^{(m)} + c, \tag{9}$$

where B is defined in terms of A by (5), and where c is defined by (6).

For the successive overrelaxation method we have

$$u^{(m+1)} = \omega (Lu^{(m+1)} + Uu^{(m)} + c) - (\omega - 1) Iu^{(m)}$$

where $L + U = B$ where L is lower triangular, and U is upper triangular.

If we let

$$L\omega = (I - \omega L)^{-1} \{U - (\omega - 1) I\}$$

we have

$$u^{(m+1)} = L\omega u^{(m)} + (I - \omega L)^{-1} \omega c. \tag{10}$$

Equation (9) and (10) may be written both in the form

$$u^{(m+1)} = T u^{(m)} + f, \tag{11}$$

where $u^{(m)} = (u_1^{(m)}, u_2^{(m)}, \ldots, u_N^{(m)})$, $f = (f_1, f_2, \ldots, f_N)$,

f is fixed, and T denotes a linear operator.

In order to investigate the convergence of the sequence $u^{(m)}$ defined by (11) we study the behavior as $m \to \infty$ of the error $e^{(m)} = u^{(m)} - u$

where u is the true solution of (8).

Since $u = Tu + f$ we have, by linearity of T

$$e^{(m+1)} = T e^{(m)} = T^{m+1} e^{(0)}.$$

As a mesure of error $e^{(m)}$, we use the Euclidean norm. Let V_N denote the N-dimensional vector space of N-tuple of complex numbers, and let the Euclidean norm of an element $v = (v_1, v_2, \ldots, v_N)$ be defined by

$$\| v \| = \left[\sum_{i=1}^{N} | v_i |^2 \right]^{\frac{1}{2}} \tag{12}$$

Evidently, in order for $u^{(m)}$ to converge to u for all $u^{(0)}$, it is necessary and sufficient that for all $v \in V_N$, we have $\lim_{m \to \infty} \| T^m v \| = 0$.

A linear transformation T of V_N into itself is said to be convergent if for all $v \in V_N$ $\lim_{m \to \infty} \| T^m v \| = 0$

LEMMA. T is a convergent transformation if and only if all the eigenvalues of

T are less than one in absolute value.

Proof: Let λ be an eigenvalue of T, then there exists $v \neq 0$ such that $Tv=\lambda v$. Then $T^m v = \lambda^m v$. Since $\lim_{m\to\infty} \| T^m v \| = 0$, it follows that $\lim_{m\to\infty} \| \lambda^m v \| = 0$. By considering nonzero component of v, we have $\lambda^m \to 0$ $(m\to\infty)$, hence $|\lambda| < 1$.

THEOREM 1. If the matrix A is symmetric and positive definite, then the Gauss–Seidel method converges for every initial vector.

Proof: Let be decomposed by writing

$$A = L_e + D + R$$

where L_e is (left) lower triangular, D is diagonal, and R is (right) upper triangular. Then in the Gauss–Seidel method, we have after $m+1$ passes through all the equations:

$$u^{(m+1)} = D^{-1}(-L_e u^{(m+1)} - Ru^{(m)} - d)$$

or $$u^{(m+1)} = H u^{(m)} + M d$$

where $H = -(D + L_e)^{-1} R$, $M = -(D + L_e)^{-1}$.

We shall show that all the eigenvalues of H are less than one in absolute value. Since A is symmetric and positive definite,

$$L_e + D - R^* = D$$

is diagonal, and positive definite. For, $x^*D x = \sum_{i=1}^{N} a_{i,i} |x_i|^2 > 0$ for any vector. Now let λ be an eigenvalue of $-H$, then there exists $x \neq 0$ such that $-Hx = \lambda x$, hence $Rx = \lambda (D + L_e) x$ and hence

$$Ax = (D + L_e + R) x = (D + L_e) x + \lambda (D + L_e) x = (1 + \lambda) (D + L_e) x.$$

Here $\lambda \neq -1$, for if $\lambda = -1$ then $|A| = 0$ this is contradict.

Therefor $\dfrac{x^*Ax}{1+\lambda} = x^* (D + L_e) x.$

On the other hand,

$$x^*R^* = \bar{\lambda}\, x^* (D + L_e)^* = \bar{\lambda} x^* (A - R^*) = \bar{\lambda} x^* A - \bar{\lambda} x^* R^*$$

hence $(1 + \bar{\lambda})\, x^*R^* = \bar{\lambda}\, x^* A.$

Therefor $x^* R^* x = \dfrac{\bar{\lambda}}{1+\bar{\lambda}} x^* Ax.$

Consequently

$$x^* (D + L_e - R^*) x = \left(\frac{1}{1+\lambda} - \frac{\bar{\lambda}}{1+\bar{\lambda}}\right) x^*Ax = \frac{1 - \bar{\lambda}\lambda}{(1+\lambda)(1+\bar{\lambda})} x^*Ax = \frac{1-|\lambda|^2}{|1+\lambda|^2} x^*Ax.$$

Since $D + L_e - R^*$ and A are positive definite, the coefficient of x^*Ax is positive. Hence $1 - |\lambda|^2 > 0$, and then $|\lambda| < 1$.

THEOREM 2. If the matrix A is irreducible: given any two nonempty disjoint subsets S and T of the set W of the first N positive integers such that $S + T = W$,

there exists $a_{ij} \neq 0$ *such that* $i \in S$ *and* $j \in T$, *and if the matrix A is diagonl dominant:* $|a_{i,j}| \geqq \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}|$ *and for at least one value of i the strict inequality holds, Then the Jacobi method converges from any starting vector to the solution.*

Proof: It is easy to show that the matrix A is nonsingular, and hence that a unique solution exists. For the Jacobi method we have in terms of the matrix A,

$$u^{(m+1)} = -D^{-1}(L_e + R)\, u^{(m)} - D^{-1} d.$$

If we assme that there exists an eigen value λ of $-D^{-1}(L_e + R)$, i. e. a root of $|L_e + D + R| = 0$ such that $|\lambda| \geqq 1$, then evidently, the matrix $L_e + D + R$ is also diagonal dominant and irreducible, and hence nonsingular. Hence $|L_e + D + R| \neq 0$, this is contradict. Therefore all the eigenvalues of $D^{-1}(L_e + R)$ are less than one in absolute value. And hence, by lemma, iteration scheme is convergent.

We remark that if A is diagonal dominant and irreducible and if $A^* = (a^*_{i,j})$ is symmetric, where $a^*_{i,j} = a_{i,i}\, a_{i,j} / |a_{i,i}|$ $(i, j = 1, 2, \ldots, N)$, then A^* is positive definite, and hence the Gauss-Seidel method converges.

2. Partial difference equations of elliptic type

The results of preceding section can be applied to many systems of linear equations arising from elliptic boundary value problems.

Now let us consider the following problem: given a closed bounded region Ω in Euclidean n-space with interior R and boundary S, and a function g (x) defined on S, the problem is to find a function u (x) which is continuous in Ω, twice differentiable in R and which satisfies

$$H(u(x)) + G(x) = 0 \qquad \text{for } x \in R, \tag{13}$$

and

$$u(x) = g(x) \qquad \text{for } x \in S, \tag{14}$$

where the differential operator H (u) is defined by

$$H(u) = \sum_{k=1}^{N} \left(A_k \frac{\partial^2 u}{\partial x_k^2} + B_k \frac{\partial u}{\partial x_k} \right) + F u$$

It is assumed that the functions $F, G, A_1, \ldots, A_n, B_1, \ldots, B_n$ are given function of x which are continuous and twice differentiable in Ω and satisfy the conditions

$$A_k(x) > 0 \quad (k=1, 2, \ldots, n), \qquad F(x) \leqq 0.$$

We write H (u) in the form

$$H(u) = \sum_{k=1}^{N} \left\{ \frac{\partial}{\partial x_k} \left(A_k \frac{\partial u}{\partial x_k} \right) + C_k \frac{\partial u}{\partial x_k} \right\} + F u, \tag{15}$$

where

$$C_k = B_k - \frac{\partial A_k}{\partial x_k} \qquad (k=1, 2, \ldots, n).$$

If $C_k = 0$ $(k=1, 2, \ldots, n)$, then H is said to be self-adjoint.

To set up our finite difference analogue we construct a rectangular net whose nodes are points $x = (x_1, x_2, \ldots, x_n)$ such that

$$x_k = p_k h_k \qquad (k=1, 2, \ldots, n)$$

where the p_k are integers and for each k, h_k is the mesh size in the direction e_k. Two nodes with coordinates $p_k h_k$ and $p'_k h_k$ are adjacent if $\sum_{k=1}^{n} (p_k - p'_k) = 1$. We denote by Ω_h the set of all nodes contained in Ω. The set of nodes such that all adjacent nodes belong to Ω_h is called the interior of Ω_h and is denoted by R_h, all other nodes of Ω_h belong to the boundary of Ω_h, denoted by S_h.

The set R_h is connected if any two nodes of R_h can be connected by an unbroken chain of segments adjoining adjacent nodes of R_h. We assume that Ω has the property that there exists $\bar{h}$ such that if for all k, $h_k < \bar{h}$, then R_h is connected.

Let N and M denote respectively the number of nodes of R_h and S_h. To each node of Ω_h we assign an integer i such that $i \leqq N$ implies $x^{(i)} \in R_h$ and $N < i \leqq N+M$ implies $x^{(i)} \in S_h$. The coordinates of $x^{(i)}$ are $p_k^{(i)} h_k$ $(k=1, 2, \ldots, n)$.

In order to derive a difference equation analogue of (13) we replace

$$\frac{\partial}{\partial x_k}\left(A_k \frac{\partial u}{\partial x_k}\right) \quad \text{by} \quad h_k^{-2}(1-E_k^{-1}) \left\{ [E_k^{\frac{1}{2}} A_k(x)]\, [(E_k - 1)\, u(x)] \right\}$$

and

$$\frac{\partial u}{\partial x_k} \quad \text{by} \quad (2h_k)^{-1} (E_k - E_k^{-1})\, u(x),$$

where the difference operator E_k^a is defined by

$$E_k^a\, u(x) = u(x + a h_k e_k).$$

Substituting in (13) and (15) we get

$$\sum_{k=1}^{n} u(x + h_k e_k) \left\{ h_k^{-2} \left[A_k \left(x + \tfrac{1}{2} h_k e_k\right) + \tfrac{1}{2} h_k C_k \right] \right\} + \sum_{k=1}^{n} u(x - h_k e_k) \left\{ h_k^{-2} \left[A_k \left(x - \tfrac{1}{2} h_k e_k\right) - \tfrac{1}{2} h_k C_k \right] \right\} - u(x) \left\{ \sum_{k=1}^{n} h_k^{-2} \left[A_k \left(x + \tfrac{1}{2} h_k e_k\right) + A_k \left(x - \tfrac{1}{2} h_k e_k\right) \right] - F(x) \right\} + G(x) = 0$$

$$(x = x^{(1)}, x^{(2)}, \ldots, x^{(N)}) \tag{16}$$

and

$$u(x) = g^*(x) \qquad (x = x^{(N+1)}, \ldots, x^{(N+M)}) \tag{17}$$

Here $g^*(x) = g(x')$ where x' is some point of near to x, such as a nearest point.

If we replace $u(x^{(i)})$ by u_i for $i \leqq N$ we obtain a system of N linear algebraic equations and N unknowns of the form (1) where, for $i, j = 1, 2, \ldots, N$,

$$-a_{i,i} = \sum_{k=1}^{n} h_k^{-2} \left[A_k \left(x^{(i)} + \tfrac{1}{2} h_k e_k\right) + A_k \left(x^{(i)} - \tfrac{1}{2} h_k e_k\right) \right] - F\left(x^{(i)}\right),$$

$$a_{i,j} = h_k^{-2} \left[A_k \left(x^{(i)} + \tfrac{1}{2} h_k e_k\right) + \tfrac{1}{2} h_k C_k \right], \qquad \text{if } x^{(j)} = x^{(i)} + h_k e_k,$$

$$a_{i,j} = h_k^{-2} \left[A_k \left(x^{(i)} - \tfrac{1}{2} h_k e_k\right) - \tfrac{1}{2} h_k C_k \right], \qquad \text{if } x^{(j)} = x^{(i)} - h_k e_k,$$

$$a_{i,j} = 0, \qquad \text{if } x^{(i)} \text{ is not adjacent to } x^{(j)} \text{ and } i \neq j,$$

$$d_i = G\left(x^{(i)}\right) + \sum_{k=1}^{n}{}' h_k^{-2} \left\{ A_k \left(x^{(i)} + \tfrac{1}{2} h_k e_k\right) + \tfrac{1}{2} h_k C_k \right\} g^*\left(x^{(i)} + h_k e_k\right)$$

$$+ \sum_{k=1}^{n}{}'' h_k^{-2} \left\{ A_k \left(x^{(i)} - \tfrac{1}{2} h_k e_k\right) - \tfrac{1}{2} h_k C_k \right\} g^*\left(x^{(i)} - h_k e_k\right)$$

where $\sum_{k=1}^{n}{}'$ and $\sum_{k=1}^{n}{}''$ denote respectively summation over all k such that $x^{(i)} + h_k e_k$ and $x^{(i)} - h_k e_k$ are nodes of S_h.

Evidently $a_{i,i} < 0$ $(i = 1, 2, \ldots, N)$, and if the h_k are chosen so that R_h is connected and such that

$$h_k < 2 \left(\operatorname*{Min}_{x \in \Omega} A_k(x) \,/\, \operatorname*{Max}_{x \in \Omega} |C_k(x)| \right) \qquad (k = 1, 2, \ldots, n)$$

then $|a_{i,i}| \geqq \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}|$. Moreover since Ω is bounded there exists i such that $x^{(i)}$ is adjacent to some node of S_h; hence $|a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}|$, then the matrix $A = (a_{i,j})$ is diagonal dominant.

The matrix A is also irreducible, since R_h is connected and since if $x^{(i)}$ and $x^{(j)}$ are distinct nodes of R_h, then $a_{i,j} \neq 0$.

If H is self-adjoint then $C_k = 0$ $(k = 1, 2, \ldots n)$. If $a_{i,j} \neq 0$ and $i \neq j$, then for some k, $x^{(j)} = x^{(i)} + h_k e_k$ or $x^{(j)} = x^{(i)} - h_k e_k$. In the former case $a_{i,j} = h_k^{-2} A_k \left(x^{(i)} + \tfrac{1}{2} h_k\right)$. Moreover $x^{(i)} = x^{(j)} - h_k e_k$ and $a_{j,i} = h_k^{-2} A_k \left[\left(x^{(i)} + h_k e_k\right) - \tfrac{1}{2} h_k e_k \right] = a_{i,j}$. Similarly if $x^{(j)} = x^{(i)} - h_k e_k$ we have $a_{j,i} = a_{i,j}$. Thus when H is self-adjoint, the matrix A is symmetric.

For the Dirichlet problem we have

$$H(u) = \sum_{k=1}^{n} \frac{\partial^2 u}{\partial x_k^2} = 0.$$

Since $A_k = 1$, $B_k = 0$, $F = 0$, then $C_k = 0$, hence H is self-adjoint, the matrix A is symmetric.

For the difference analogue we have

$$b_{i,j} = \begin{cases} \dfrac{h_k^{-2}}{2\sum_{k=1}^{n} h_k^{-2}} & \text{if } x^{(i)} - x^{(j)} = \pm h_k e_k \\ 0 & \text{if } x^{(j)} \text{ is not adjacent to } x^{(i)}, \end{cases}$$

where the $b_{i,j}$ are defined by (5).

As an example we wish to solve the Dirichlet problem in a plane region Ω. We overlay Ω with a square net with mesh size h, and we assume that the boundary S is a closed polygon composed by mesh lines. The difference equation is

$$4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} = 0.$$

Here $u_{i,j} = u(ih, jh)$.

The Jacobi method is given by

$$u_{i,j}^{(m+1)} = \frac{1}{4}\left(u_{i+1,j}^{(m)} + u_{i-1,j}^{(m)} + u_{i,j+1}^{(m)} + u_{i,j-1}^{(m)}\right)$$

and successive overrelaxation method is given by

$$u_{i,j+1}^{(m+1)} = \omega\left[\frac{1}{4}\left(u_{i-1,j}^{(m+1)} + u_{i,j-1}^{(m+1)} + u_{i+1,j}^{(m)} + u_{i,j+1}^{(m)}\right)\right] - (\omega-1)\, u_{i,j}^{(m)} .$$

We assume that for the interior net point (i, j), i goes from 1 to IMAX, and j goes from JL (I) to JU (I) for every I ($1 \leqq I \leqq IMAX$) .

The following programme is written in HARP 103 language for the HIPAC 103 computing machine.

```
#  SUCCESSIVE OVERRELAXATION METHOD
   DIMENSION U (7,7) , JL(6), JU(6)
#  BOUNDARY VALUES
   READO, M
   DO 11 L=1, M
   READO, I, J
11 READ1, U (I, J)
#  RANGE
   READO, IMAX
   DO 12 I=2, IMAX
12 READO, JL (I) , JU (I)
#  CONSTANTS
   READ1, EPS, OMEGA
#  INITIAL VALUES
   DO 13 I=2, IMAX
   JMIN=JL (I)
   JMAX=JU (I)
```

```
      UIJ=U (I,JMIN-1)
      SLOPE= (U (I,JMAX+1) -UIJ) /FLOATF (JMAX-JMIN+2)
      DO 13 J=JMIN, JMAX
      UIJ=UIJ+SLOPE
   13 U (I,J) =UIJ
   #  ITERATION
      K=0
   14 K=K+1
      ENORM=0.
      DO 15 I=2, IMAX
      JMIN=JL (I)
      JMAX=JU (I)
      DO 15 J=JMIN,JMAX
      EIJ=OMEGA* ((U (I+1,J) +U (I,J+1) +U (I-1,J) +U(I,J-1)) /4. -U(I,J))
      U (I,J) =U (I,J) +EIJ
   15 ENORM=ENORM+ABSF (EIJ)
      PRINT 16, ENORM
   16 FORMAT (F 15. 10)
      IF (ENORM-EPS) 18, 18, 14
   #  OUTPUT; NUMBER OF ITERATION, RESULTS
   18 PRINTO, K
      DO 19 I=2, IMAX
      JMIN=JL (I)
      JMAX=JU (I)
      DO 19 J=JMIN, JMAX
   19 PRINT 20, I, J, U (I, J)
   20 FORMAT (2I5, F 15. 10)
      STOP
      END
```

References

〔1〕 D. Young, Iterative methods for solving partial difference equations of elliptic type, Trans. Math. Soc. 76 (1954) , pp. 92—111.

〔2〕 D. Young, The numerical solution of elliptic and parabolic partial differential equations, J. Todd, A survey of numerical analysis, Ch. 11.

〔3〕 A. Ralston, H. S. Wilf, Mathematical methods for digital computers, Wiley (1965) .