The Discretization Error in Boundary Value Problems for the Second Order Differential Equations (I).

Shoichi SEINO

1. Introduction

Although a number of papers guarantee that a consistent method is convergent, it is hardly useful for purposes of estimating the discretization error. In particular it fails to indicate the order of the discretization error. In [1], P. Henrici described the discretization error in boundary value problems of class M which will be said if it is of the form

$$y''=f(x, y), \qquad y(a)=\alpha, \qquad y(b)=\beta,$$

where $a<b$ and α and β are given constants, and suggested that a similar result could be true for the generalized boundary value problems.

In this paper, we consider the boundary value problem (BVP for short)

$$(1) \quad y'' = g(x)y' + f(x, y), \qquad y(a) = \alpha, \qquad y(b) = \beta,$$

where $-\infty<a<b<+\infty$, α and β are arbitrary constants.

We assume that the function $g(x)$

(a) is defined and continuous in the interval $[a, b]$, where a and b are finite,

(b) is monotone decreasing,

(c) is nonnegative in $[a, b]$,

also that the function $f(x, y)$

(d) is defined and continuous in the strip $a\leqq x\leqq b$, $-\infty<y<+\infty$,

(e) satisfies a Lipschitz condition with respect to y,

(f) $f_y(x, y)$ is continuous and satisfies $f_y(x, y) \geqq 0$ in $a\leqq x\leqq b$, $-\infty<y<+\infty$.

We shall establish the error bound and asymptotic formula for the discretization error. These results are stated as Theorem 3 and Theorem 4.

2. Preliminary Results

We shall first list some definitions.

Let W be the set of the first n integers, $W=\{1, 2, \cdots\cdots\cdots, n\}$. A matrix $\mathbf{A}=(a_{ij})$ is called reducible if it is possible to decompose W into two nonempty, disjoint subsets S and T, such that $a_{ij} = 0$ for $i \in S$ and $j \in T$. A matrix which is not reducible is called irreducible.

[Lemma 1]

A tridiagonal matrix $\mathbf{A}= (a_{ij})$ *is irreducible if and only if*

$$a_{i,i-1} \neq 0 \quad (i=2, 3, \cdots\cdots\cdots\cdots, n) \text{ and}$$

$$a_{i,i+1} \neq 0 \quad (i=1, 2, \cdots\cdots\cdots\cdots, n-1).$$

A matrix $\mathbf{A}$ with real elements is called monotone if $\mathbf{Az}\geqq\mathbf{O}$ implies $\mathbf{z}\geqq\mathbf{O}$, where by the notation $\mathbf{z} \geqq \mathbf{O}$ we mean that all components z_i of the vector $\mathbf{z}$ satisfy $z_i \geqq 0$.

As a consequence of this definition we obtain the following lemmas.

〔Lemma 2〕

A matrix $\mathbf{A}$ *is monotone if and only if the elements of the inverse matrix* $\mathbf{A}^{-1}$ *are nonnegative.*

〔Lemma 3〕

Let the matrix $\mathbf{A} = (a_{ij})$ *be irreducible and satisfy the conditions*

(i) $a_{ij} \leqq 0, \quad i \neq j\,;\ i,\ j = 1, 2, \cdots\cdots\cdots, n,$

(ii) $\displaystyle\sum_{j=1}^{n} a_{ij} \begin{cases} \geqq 0 & \text{for } i = 1, 2, \cdots\cdots\cdots, n, \\ > 0 & \text{for at least one } i. \end{cases}$

Then $\mathbf{A}$ *is monotone.*

〔Lemma 4〕

Let the matrices $\mathbf{A}$ *and* $\mathbf{B}$ *be monotone, and assume that* $\mathbf{A} - \mathbf{B} \geqq \mathbf{O}$.

Then $\mathbf{B}^{-1} - \mathbf{A}^{-1} \geqq \mathbf{O}$.

Now we introduce a few notations.

Let vectors $\mathbf{y}$, $\mathbf{f}(\mathbf{y})$ and $\mathbf{a}$ be

$$\mathbf{y} = [y_1, \cdots\cdots\cdots, y_{N-1}]^T, \qquad \mathbf{f}(\mathbf{y}) = [f(x_1, y_1), \cdots\cdots\cdots\cdots, f(x_{N-1}, y_{N-1})]^T,$$

$$\mathbf{a} = [\alpha - \beta_0 h^2 f(x_0, \alpha),\ 0, \cdots\cdots\cdots, 0,\ \beta - \beta_2 h^2 f(x_N, \beta)]^T,$$

matrices $\mathbf{J}$ and $\mathbf{B}$ be

$$\mathbf{J} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \beta_1 & \beta_2 & & & \\ \beta_0 & \beta_1 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_0 & \beta_1 & \beta_2 \\ & & & \beta_0 & \beta_1 \end{pmatrix}.$$

Let the system of finite difference equations

(2) $\quad \mathbf{J}\mathbf{y} + h^2 \mathbf{B}\mathbf{f}(\mathbf{y}) - \mathbf{a} = \mathbf{O}$

have arisen from BVP of class M. If p denotes the order of the finite difference operator, assume that the exact solution $y(x)$ of BVP of class M has a continuous $(p+2)$nd derivative in $[a, b]$ and let

$$Z = \max_{a \leqq x \leqq b} \left| y^{(p+2)}(x) \right|.$$

Assume that G is a positive constant which depends only on the difference operator and that

(3) $\quad \beta_i \geqq 0 \quad (i=0,\ 1,\ 2), \qquad \beta_0 + \beta_1 + \beta_2 = 1,$

and

(4) $\quad h^2 L < 1,$

where L denotes the Lipschitz constant of $f(x,\ y)$.

〔Theorem 1〕

Let the values y_n *satisfy the equations*

(5) $\quad -y_{n-1} + 2y_n - y_{n+1} + h^2\{\beta_0 f_{n-1} + \beta_1 f_n + \beta_2 f_{n+1}\} = \theta_n K h^{q+2},$

where the θ_n *are arbitrary numbers satisfying* $|\theta_n| \leqq 1$, *K and p are arbitrary nonnegaitve* *constants, and where* $f_n=f(x_n, y_n)$.

If (3) *and* (4) *hold, then the discretization error* $e_n=y_n-y(x_n)$ *satisfies*

$$|e_n| \leqq \frac{(x_n-a)(b-x_n)}{2} \cdot (GZh^p+Kh^q), \qquad n=1,\ 2, \cdots\cdots\cdots, N-1.$$

〔Theorem 2〕

In addition to the hypotheses of Theorem 1, we shall now that the order $p \geqq 2$ *and* $K=0$, *and that the difference operator satisfies* $\beta_0=\beta_2$, *and that exact solution* $y(x)$ *is* $(p+4)$ *times continuously differentiable. Then the error* e_n *satisfies*

$$e_n = e(x_n)h^p + 0(h^{p+2}),$$

where $e(x)$ *denotes the solution of BVP*

$$e''(x) = g(x)e(x) - Cy^{(p+2)}(x), \qquad e(a)=e(b)=0,$$

where C is called an error constant defined by Henrici.

Proofs of these lemmas and theorems are given in [1].

3. A priori Bound and Asymptotic Behavior of the Discretization Error of BVP (1)

We extend the result of BVP of class M to BVP (1).

For the direct numerical solution of BVP (1), we introduce the points $x_n = a + nh$ $(n=1,\ 2, \cdots\cdots, N)$, where $h=(b-a)h^{N-1}$ and N is an appropriate integer. The difference equation which determinate numbers y_n approximating the values $y(x_n)$ of the true solution at the points x_n is

$$(6) \qquad -y_{n-1}+2y_n-y_{n+1}+h^2\{\beta_0(g_{n-1}y'_{n-1}+f_{n-1})+\beta_1(g_ny'_n+f_n)+\beta_2(g_{n+1}y'_{n+1}+f_{n+1})\}=0,$$

where $g_n=g(x_n)$, $\quad f_n=f(x_n, y_n)$.

We shall assume that β_i (i =0, 1, 2) satisfy (3) and $\beta_0=\beta_2$.

Furthermore we approximate y' by

$$y'_{n-1}=h^{-1}(y_n-y_{n-1}), \qquad y'_n=(2h)^{-1}(y_{n+1}-y_{n-1}), \qquad y'_{n+1}=h^{-1}(y_{n+1}-y_n).$$

Then (6) is as follows ;

$$(7) \qquad -y_{n-1}+2y_n-y_{n+1}+h\{A_{n-1}y_{n-1}+B_ny_n+C_{n+1}y_{n+1}\}+h^2\{\beta_0f_{n-1}+\beta_1f_n+\beta_2f_{n+1}\}=0,$$

where $A_{n-1}=-(\beta_0g_{n-1}+\frac{1}{2}\beta_1g_n)$, $\quad B_n=\beta_0g_{n-1}-\beta_2g_{n+1}$, $\quad C_{n+1}=\frac{1}{2}\beta_1g_n+\beta_2g_{n+1}$.

The further discussion is simplified by the use of vector and matrix notation.

Defining the vector

$$\mathbf{b} = [\alpha+hA_1\alpha-h^2\beta_0f(x_0, \alpha), 0, \cdots\cdots\cdots 0,\ \ \beta-hC_N\beta-h^2\beta_2f(x_N, \beta)]^T,$$

and matrix

$$\mathbf{C}=\begin{pmatrix} B_1 & C_2 & & & \\ A_1 & B_2 & C_3 & & \\ & \ddots & \ddots & \ddots & \\ & & A_{N-3} & B_{N-2} & C_{N-1} \\ & & & A_{N-2} & B_{N-1} \end{pmatrix},$$

the system of equations arising from demanding that (7) hold for $n=1,\ 2, \cdots\cdots, N-1$ can be written in the form

$$(8) \qquad \mathbf{Jy}+h\mathbf{Cy}+h^2\mathbf{Bf}(\mathbf{y})=\mathbf{b},$$

where vectors $\mathbf{y}$, $\mathbf{f}(\mathbf{y})$ and matrices $\mathbf{J}$, $\mathbf{B}$ are defined in § 2.

As usual, we mean by the discretization error the quantity $e_n = y_n - y(x_n)$, where y_n is the exact solution of the difference equation (8), and $y(x_n)$ is the exact solution of BVP (1).

We denote the order of the finite difference operator by p. We assume that the exact solution y(x) has a continuous (p+2)nd derivative in [a, b], and that Z and G are defined in § 2.

Let us require that the stepsize h be so small that

$$(9)\qquad X<h<\min\{X_1,\ X_2,\ X_3\},$$

where

$$X_1=\frac{-C_k+\sqrt{C^2_k+4\beta_2h_k}}{2\beta_2h_k},\quad (k=2,\ 3,\cdots\cdots\cdots,N-1),$$

$$X_2=\frac{A_l+\sqrt{A_l^2+4\beta_0h_l}}{2\beta_0h_l},\quad (l=1,\ 2,\cdots\cdots\cdots,N-2),$$

$$X_3=\frac{C_N-\sqrt{C_N^2-4(\beta_0h_{N-2}+\beta_1h_{N-1})}}{2(\beta_0h_{N-2}+\beta_1h_{N-1})},$$

and

$$X=\frac{A_l}{\beta_0h_l},\quad (l=1,\ 2,\cdots\cdots\cdots,N-2),$$

where $h_n=h(x_n,\ y_n)=f_y(x_n,\ y_n)$.

[Theorem 3]

Instead of (8), let the values y_n satisfy the equations

$$(10)\qquad -y_{n-1}+2y_n-y_{n+1}+h\{A_{n-1}y_{n-1}+B_ny_n+C_{n+1}y_{+1}\}+h^2\{\beta_0f_{n-1}+\beta_1f_n+\beta_2f_{n+1}\}=\theta_nKh^{q+2},$$

where the θ_n are arbitrary numbers satisfying $|\theta_n|\leqq 1$ and where K and q are arbitrary nonnegative constants. If (9) holds, then the discretization error satisfies

$$(11)\qquad |e_n|\leqq\frac{(x_n-a)(b-x_n)}{2}(GZh^p+Kh^q),\qquad n=1,\ 2,\cdots\cdots\cdots,N-1.$$

Proof. The exact solution y(x) satisfies

$$(12)\qquad -y(x_{n-1})+2y(x_n)-y(x_{n+1})+h\{A_{n-1}y(x_{n-1})+B_ny(x_n)+C_{n+1}y(x_{n+1})\}$$
$$+h^2\{\beta_0f(x_{n-1},\ y(x_{n-1}))+\beta_1f(x_n,\ y(x_n))+\beta_2f(x_{n+1},\ y(x_{n+1}))\}=\theta'_nGZh^{p+2},$$

where $|\theta'_n|\leqq 1$.

Subtracting this relation from (10), we get

$$-e_{n-1}+2e_n-e_{n+1}+h\{A_{n-1}e_{n-1}+B_ne_n+C_{n+1}e_{+1}\}+h^2\{\beta_0(f_{n-1}-f(x_{n-1},\ y(x_{n-1})))$$
$$+\beta_1(f_n-f(x_n,\ y(x_n)))+\beta_2(f_{n+1}-f(x_{n+1},\ y(x_{n+1})))\}=\theta''(GZh^{p+2}+Kh^{q+2}).$$

Applying the mean value theorem to the difference $f(x_n,\ y_n)-f(x_n,\ y(x_n))$, we have

$$(13)\qquad -e_{n-1}+2e_n-e_{n+1}+h\{A_{n-1}e_{n-1}+B_ne_n+C_{n+1}e_{n+1}\}+h^2\{\beta_0h_{n-1}e_{n-1}+\beta_1h_ne_n+\beta_2h_{n+1}e_{n+1}\}=\theta_n''(GZh^{p+2}+Kh^{q+2}).$$

Denoting by $\mathbf{H}$ the diagonal matrix with elements h_n $(n=1,\ 2,\cdots\cdots,N-1)$, we obtain

$$(14)\qquad (\mathbf{J}+h\mathbf{C}+h^2\mathbf{BH})\,\mathbf{e}=(GZh^{p+2}+Kh^{q+2})\,\Theta,$$

where Θ is a vector whose components numerically do not exceed 1. As the stepsize h satisfies (9), the matrix $\mathbf{J}+h\mathbf{C}+h^2\mathbf{BH}$ satisfies the conditions of Lemma 3, and furthermore we have $\mathbf{J}+h\mathbf{C}+h^2\mathbf{BH}\geqq\mathbf{J}$. From Lemma 4, it follows that

$$\mathbf{O} \leqq (\mathbf{J}+h\mathbf{C}+h^2\mathbf{BH})^{-1} \leqq \mathbf{J}^{-1}.$$

If we put $\mathbf{J}^{-1}=(j_{mn})$, we obtain

$$|e_n| \leqq (GZh^{p+2}+Kh^{q+2}) \sum_{n=1}^{N-1} j_{mn}.$$

Through the fact that

$$\sum_{n=1}^{N-1} j_{mn} = \frac{(x_m-a)(b-x_m)}{2h^2},$$

we get

$$|e_m| \leqq \frac{(x_m-a)(b-x_m)}{2}(GZh^p+Kh^q).$$

In addition to the hypotheses of Theorem 3, we shall now that the order $p \geqq 2$, and that $K=0$. Then it follows that $|\mathbf{e}| = 0(h^2)$.

We shall assume that the exact solution $y(x)$ of BVP (1) is $(p+4)$ times continuously differentiable.

【Theorem 4】

Let $e(x)$ be the solution of BVP

(15) $\quad e''(x) = g(x)e'(x)+h(x)\,e(x)-D_{p+2}y^{(p+2)}(x), \quad e(a)=e(b)=0,$

where $h(x)=f_y(x, y(x))$, and D_{p+2} is a nonzero constant.

Under the above hypotheses, the discretization error e_n satisfies

$$e_n=e(x_n)h^p+0(h^{p+2}), \qquad n=1, 2, \cdots\cdots\cdots, N-1.$$

Proof. As the order of the finite difference operator is p, we have

(16) $\quad -y(x_{n-1})+2y(x_n)-y(x_{n+1})+h\{A_{n-1}y(x_{n-1})+B_ny(x_n)+C_{n+1}y(x_{n+1})\}$
$\quad +h^2\{\beta_0f(x_{n-1}, y(x_{n-1}))+\beta_1f(x_n, y(x_n))+\beta_2f(x_{n+1}, y(x_{n+1}))\}$
$\quad =-D_{p+2}y^{(p+2)}(x_n)h^{p+2}+0(h^{p+3}).$

Since $e_n = 0(h^2)$,

(17) $\quad f(x_n, y_n)-f(x_n, y(x_n))=h_ne_n+0(h^4).$

Furthermore, by (3) and $\beta_0=\beta_2$

(18) $\quad y^{(p+2)}(x_n)=\beta_0y^{(p+2)}(x_{n-1})+\beta_1y^{(p+2)}(x_n)+\beta_2y^{(p+2)}(x_{n+1}).$

On the other hand, values y_n satisfy

(19) $\quad -y_{n-1}+2y_n-y_{n+1}+h\{A_{n-1}y_{n-1}+B_ny_n+C_{n+1}y_{n+1}\}+h^2\{\beta_0f_{n-1}+\beta_1f_n+\beta_2f_{n+1}\}=0.$

We subtract (16) from (19), applying (17) and (18),

(20) $\quad -e_{n-1}+2e_n-e_{n-1}+h\{A_{n-1}e_{n-1}+B_ne_n+C_{n+1}e_{n+1}\}+h^2\{\beta_0h_{n-1}e_{n-1}+\beta_1h_ne_n$
$\quad +\beta_2h_{n+1}e_{n+1}\}=D_{p+2}\{\beta_0y^{(p+2)}(x_{n-1})+\beta_1y^{(p+2)}(x_n)+\beta_2y^{(p+2)}(x_{n+1})\}+0(h^{p+3}).$

We now introduce new quantity, to be called magnified errors, by

$$\bar{e}_n=e_n\cdot h^{-p}, \qquad n=1, 2, \cdots\cdots\cdots, N-1.$$

Then (20) becomes

(21) $\quad -\bar{e}_{n-1}+2\bar{e}_n-\bar{e}_{n+1}+h\{A_{n-1}\bar{e}_{n-1}+B_n\bar{e}_n+C_{n+1}\bar{e}_{n+1}\}+h^2\{\beta_0(h_{n-1}\bar{e}_{n-1}$
$\quad -D_{p+2}y^{(p+2)}(x_{n-1}))+\beta_1(h_n\bar{e}_n-D_{p+2}y^{(p+2)}(x_n))+\beta_2(h_{n+1}e_{n+1}$
$\quad -D_{p+2}y^{(p+2)}(x_{n+1}))\}\}=0\,(h^3).$

By the finite difference method the same relation would result from solving BVP (15)

denoting by e_n the approximation to $e(x_n)$.
We may appeal to Theorem 3 to conclude that

$$\bar{e} = e(x_n) + 0(h^2),$$

i.e., $$e_n = e(x_n) \cdot h^p + 0(h^{p+2}), \qquad n=1, 2, \cdots\cdots\cdots, N-1.$$

and the theorem follows.

References

〔1〕. P. Henrici : Discrete Variable Methods in Ordinary Differential Equations, John Wiley & Sons, Inc. (1962).

〔2〕. I. Babuška, M. Prāger and E. Vitásek : Numerical Processes in Differential Equations, John Wiley & Sons, Inc. (1966).

〔3〕. Symposium on the Numerical Treatment of Ordinary Differential Equations, Integral and Integro-Differential Equations, Birkhauser Verlag (1960).

〔4〕. L. F. Shampine : Boundary Value Problems for Ordinary Differential Equations. SIAM J. Numer. Anal., Vol. 5, No. 2. (1968).

〔5〕. L. F. Shampine : Some nonlinear boundary value problems, Arch. Rational Mech. Anal., ,25 (1967).