THE DIFFERENCE CORRECTION METHOD FOR NON-LINEAR TWO-POINT BOUNDARY VALUE PROBLEMS

BY

VICTOR PEREYRA

TECHNICAL REPORT CS18

FEBRUARY 25, 1965

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

THE DIFFERENCE CORRECTION METHOD FOR NON-LINEAR, TWO-POINT BOUNDARY VALUE PROBLEMS.

by

Victor Pereyra[†]

1. Introduction

This paper will deal with the theory and application of the difference correction method to two-points boundary value problems of monotonic type, i.e.:

$$y'' = f(x,y) \tag{1.1}$$

$$\alpha\, y(a) - \beta\, y'(a) = A \tag{1.1'}$$

$$\gamma\, y(b) + \delta\, y'(b) = B \tag{1.1"}$$

with several conditions on $f(x,y)$ and the constants $\alpha, \beta, \gamma, \delta$.

A thorough discussion on the practical aspects of the difference correction method can be found in Fox [1957] and Fox [1961] where the method is applied to a wide variety of problems. Considering boundary value problems for the Poisson equation in two dimensions Bickley, Michaelson and Osborne [1961] have pointed out some theoretical aspects of the difference correction when applied to that problem.

In Henrici's book, "Discrete variable methods in ordinary differential equations" [1962] p.377, it is indicated that, if a difference correction

---

† Departamento de Matematicas e Instituto de Calculo, F. C. E. y N., Universidad de Bs.As. Argentina. Present address Stanford Computation Center. Stanford, California, U.S.A.

 This report was supported in part by Office of Naval Research Contract Nonr-225(37) (NR-044-211) at Stanford University.

is added to an approximate solution of 1.1 then the order of the discretization error is increased in at least two units. After giving some notation in Section 2 a discussion of (1.1) with $\alpha = y = 1$, $\beta = \delta = 0$ is given in detail in the following sections. The asymptotic behavior of the discretization error is discussed in Section 3, following the lines of Henrici with certain changes which make it more general and allow us to introduce several ways of performing the difference correction.

In Section 4 the h^2 improvement property of a generalized difference correction is proved.

In Section 5 two possibilities (different from the classical) are investigated for the case $p = 2$, and in Section 6 numerical results and comparisons with other methods are presented, showing that it is faster and more accurate to use correction differences than a direct method of equivalent order.

In Section 7, the results of Sections 3 and 4 are extended to the general problem (1.1) and in Section 8 a numerical example is presented.

2. Notation and known results

As we want to use several results by Henrici [1962] 'Chapter 7, and we prefer to avoid repeated references, we will adopt its notation and we will give a summary of these results.

A non linear boundary value problem will be called of class M, if it is of the form (1.1) and, a) the initial value problem $y'' = f(x,y)$ $y(a) = \alpha$, $y'(a) = A$ with A arbitrary, has a unique solution. b) $f_y(x,y)$ is continuous and

$$f_y(x,y) \geq 0 \quad \text{for } a \leq x \leq b, \ -\infty < y < \infty \ . \tag{2.1}$$

c) the boundary conditions are,

$$y(a) = \alpha \quad , \; y(b) = \beta \; .$$

It is proved then that a problem of class M always has a unique solution. The finite difference approximations that we will discuss are of the form,

$$(2.2) \qquad -y_{n-1} + 2y_n - y_{n+1} + h^2 \{\beta_0 f_{n-1} + \beta_1 f_n + \beta_2 f_{n+1}\} = 0$$
$$n = 1,2, \ldots . N-1$$

where $\beta_0 + \beta_1 + \beta_2 = 1$, $\beta_0 = \beta_2$, $h = (b-a)/N$ (N integer), $y_0 = \alpha$, $y_N = \beta$ and the rest is standard notation. The limitation of taking this kind of equations appears naturally if we do not want to consider grid points outside of the interval [a,b]. By introducing some special matrices and vectors, part of the following discussion can be simplified. We will use no special notation for matrices or vectors, but we hope that their meanings will be clear in each context. Let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \end{pmatrix} \quad f(y) = \begin{pmatrix} f(x_1, y_1) \\ f(x_2, y_2) \\ \vdots \\ f(x_{N-1}, y_{N-1}) \end{pmatrix} \quad a = \begin{pmatrix} \alpha - \beta_0 h^2 f(x_0, \alpha) \\ 0 \\ \vdots \\ 0 \\ \beta - \beta_0 h^2 f(x_N, \beta) \end{pmatrix}$$

$$J = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & -1 \\ & & & & -1 & 2 \end{pmatrix} \qquad B = \begin{pmatrix} \beta_1 & \beta_2 & & & & \\ \beta_o & \beta_1 & \beta_2 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \beta_2 \\ & & & & \beta_o & \beta_1 \end{pmatrix}$$

(2.3)

$$F(y) = \begin{pmatrix} f_y(1) & 0 & & & \\ 0 & f_y(2) & \cdot & & \\ & & \cdot & \cdot & \\ & & \cdot & \cdot & \\ & & & & 0 \\ & & & 0 & f_y(N-1) \end{pmatrix}$$

where $f_y(j) = f_y(x_j, y_j)$.

For instance, formula (2.2) can now be written,

$$Jy + h^2 Bf(y) - a = 0 \tag{2.4}$$

where the vector a takes care of the boundary values.

A Newton type iteration used to solve the system of non-linear equations (2.4) is insured to be convergent under certain restrictions, mainly on the first approximation and on the step length h (Th. 7.7, p. 373, l.c.) . If the first approximation is called $y^{(0)}$, then the formulas for Newton method are in this case,

$$r(y^{(i)}) = Jy^{(i)} + h^2 Bf(y^{(i)}) - a, \tag{2.5}$$

$$\Delta y^{(i)} = -(J + h^2 BF(y^{(i)}))^{-1} r(y^{(i)}) \tag{2.6}$$

and finally,

$$Y^{(i+1)} = y^{(i)} + \Delta y^{(i)} \tag{2.7}$$

If the computed approximation is called y^* and the exact solution of (1.1) is called y, then theorem 7.8, p. 374 gives for the components of the discretization error, $e = y^* - y$, the following bound,

$$|e_n| \leq \frac{(x_n - a)(b - x_n)}{2} (C h^p + K h^q) \tag{2.8}$$

where C is a constant which depends on the method and on the problem itself, and p is the order of the method. K and q are arbitrary non negative constants which stem from the assumption that the Newton iteration is stopped when the components of the residual vector satisfy

$$|r_n| \leq K h^{q+2} \tag{2.9}$$

This is a very important practical fact, because it permits us to perform an incomplete iteration (the only possible kind in actual computation) before applying the difference correction technique. We will assume that $q > p + 4$ in order to avoid interference of this term in the discussion of the discretization error.

A difference operator $L[y(x);h]$ is naturally associated to the difference scheme (2.2),

$$L[y(x);h] = -y(x_{n-1}) + 2y(x_n) - Y(x_{n+1}) \tag{2.10}$$

$$+ h^2\{\beta_o\, y''(x_{n-1}) + \beta_1\, y''(x_n + \beta_2\, y''(x_{n+1})\} \quad .$$

$L[y(x);h]$ operates on all functions $y(x)$ sufficiently differentiable. By expanding in Taylor series all the terms of (2.10) it is possible to find,

$$L[y(x);h] = h^{p+2}C_{p+2}\, y^{(p+2)}(x) + h^{p+4}C_{p+4}\, y^{(p+4)}(x) + O(h^{p+6}) \tag{2.11}$$

where p is called the <u>order</u> of the method.

We will also need some notions about monotone matrices.

A matrix A is said to be <u>reducible</u> if and only if it is similar to a block matrix of the form,

$$P^T AP = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11}, A_{22} are square and P is a permutation matrix. In particular, a tridiagonal matrix $A = (a_{ij})$ is irreducible if and only if,

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

and

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

If by the notation $v > 0$ (either for vectors or matrices) we mean that all the elements are non-negative, then we can define: a matrix A is monotone if $A z \geq 0$ implies $z > 0$. A direct consequence of the definition is that every monotone matrix is not singular.

A fundamental result of this theory is: a matrix A is monotone iff $A^{-1} \geq 0$.

Another important fact is the following: if A is irreducibly diagonally dominant and has non-positive off-diagonal elements then A is monotone.

Finally we quote for further use, if A and B are monotone and $-B \leq A$ then $A^{-1} \leq B^{-1}$

3. Asymptotic behavior of the discretization error

Following the lines of Henrici, pp. 375-377, we will now derive an expression for the discretization error which will be useful in the discussion of the difference correction method.

We recall that, if formula (2.2) is used as a finite difference approximation to problem (1.1), and y_n is the approximate solution of the system of equations, then the discretization error, $e_n = y_n - y(x_n)$ $(n = 0,1,\ldots,N)$ satisfies (2.8). We will assume that $p \geq 2$ and that the exact solution $y(x)$ is $(p+6)$ times continuously differentiable.

Therefore,

$$f(x_n,y_n) - f(x_n,y(x_n)) = f_y(x_n,y(x_n))(y_n-y(x_n)) + O(h^{2p})$$

or, by calling $g_n = f_y(x_n,y(x_n))$,

$$f(x_n,y_n) - f(x_n,y(x_n)) = g_n \cdot e_n = O(h^{2p}) \tag{3.1}$$

As $\beta_o + \beta_1 + \beta_2 = 1$ and $\beta_o = \beta_2$ we get,

$$y^{(p+2)}(x_n) = \beta_0\, y^{(p+2)}(x_{n-1}) + \beta_1\, y^{(p+2)}(x_n) + \beta_2\, y^{(p+2)}(x_{n+1}) - \beta_0\, h^2\, y^{(p+4)}(x_n) + O(h^4) \tag{3.2}$$

Now we will construct a difference equation for the discretization error, by subtracting (2.11) from (2.2)

$$-y_{n-1} + 2\,y_n - Y_{n+1} + h^2(\beta_o\, f_{n-1} + \beta_1\, f_n + \beta_2\, f_{n+1}) - L[y(x_n),h] =$$

$$= -\,h^{p+2}\, C_{p+2}\, y^{(p+2)}(x_n) - h^{p+4}\, C_{p+4}\, y^{(p+4)}(x_n) + O(h^{p+6}) \quad .$$

Or, letting $d_n = f(x_n,y_n) - f(x_n,y(x_n))$

$$-e_{n-1} + 2\,e_n - e_{n+1} + h^2(\beta_0\, d_{n-1} + \beta_1\, d_n + \beta_2\, d_{n+1}) =$$

$$= - h^{p+2}\, C_{p+2}\, y^{(p+2)}(x_n) - h^{p+4} C_{p+4}\, y^{(p+4)}(x_n) + O(h^{p+6}) .$$

Using now the relation (3.1), dividing through by h^p and defining the magnified error $\bar{e}_n = h^{-p} e_n$ we get,

$$(3.3) \qquad -\bar{e}_{n-1} + 2.\,\bar{e}_n - \bar{e}_{n+1} + h^2\{\beta_0\, g_{n-1}\, \bar{e}_{n-1} + \beta_1\, g_n\, \bar{e}_n +$$

$$+\beta_2\, g_{n+1}\, \bar{e}_{n+1}\} = - h^2 C_{p+2}\; Y^{(p+2)}(x_n) - h^4\, C_{p+4}\, y^{(p+4)}(x_n)$$

$$+ O(h^6) .$$

Introducing now (3.2) and defining,

$$(3.4) \qquad \Phi_n = g_n\, \bar{e}_n + Y^{(p+2)}(x_n)\, C_{p+2}$$

(3.3) is transformed in,

$$(3.5) \qquad -\bar{e}_{n-1} + 2\,\bar{e}_n - \bar{e}_{n+1} + h^2(\beta_0\, \Phi_{n-1} + \beta_1\, \Phi_n + \beta_2\, \Phi_{n+1}) =$$

$$= h^4\,(C_{p+2}\,\beta_0 - C_{p+4})\, y^{(p+4)}(x_n) + O(h^6) .$$

If we solve the boundary value problem of class M,

$$e''(x) = g(x)\, e(x) + C_{p+2} y^{(p+2)}(x) \tag{3.6}$$

$$e(a) = e(b) = 0$$

by the method (2.2), we will obtain equations (3.5) with zeros in the right hand sides. Then, by (2.8) we get,

$$e_n = h^p e(x_n) + \Theta_n \frac{(b-a)^2}{8} \{ h^{p+2} \quad _{p+2}\beta_o - C_{p+4}) y^{(p+4)}(x_n) + \tag{3.7}$$

$$+ h^{2p} G Z \} + O(h^{p+4})$$

where $|\Theta_n| < 1$.

In other words,

$$y(x_n) = y_n - h^p e(x_n) + O(h^{p+2}) \tag{3.8}$$

with the error leading term given in (3.7).

4. The difference correction.

The last formula of Section 3 indicates a way of improving the approximate solution y_n by at least two orders in h. To do so, we have to know how to compute $e(x_n)$. Actually, it is enough to know how to compute a quantity e_n^* which satisfies,

$$e_n^* = e(x_n) + O(h^2) ,$$

and that is the one which we will be able to obtain. The difficulty in solving (3.6) is that we do not know $y^{(p+2)}(x_n)$. Consequently, a reasonable step is to replace $y^{(p+2)}(x_n)$ by a known appropriate value, $r(x_n)$. The only thing we will require from $r(x_n)$ is that it fulfills

$$y^{(p+2)}(x_n) = r(x_n) + s(x_n)\, h^2 + O(h^3) \tag{4.1}$$

where $s(x)$ is a sufficiently differentiable function, and $|s(x)| < K \quad a \leq x \leq b$.

Let us now define the following differende problem,

$$-e^*_{n-1} + 2\, e^*_n - e^*_{n+1} + h^2 (\beta_0\, \Phi^*_{n-1} + \beta_1\, \Phi^*_n + \beta_2\, \Phi^*_{n+1}) = 0 \tag{4.2}$$

with

$$\Phi^*_n = g_n\, e^*_n + C_{p+2} \quad r(x_n) .$$

The problem has a solution since the matrix $\Omega = (J + h^2\, B\, F)$ being monotone (for h sufficiently small) has an inverse. Moreover, $0 < \Omega^{-1} \leq J^{-1}$.

On the other hand, the exact solution of (3.6) satisfies,

$$(4.3)\quad -e(x_{n-1}) + 2\,e(x_n) - e(x_{n+1}) + h^2\,(\beta_0\,\Phi_{n-1} + \beta_1\,\Phi_n + \beta_2\,\Phi_{n+1})$$

$$= h^{p+2}\,C_{p+2}\,e^{(p+2)}(\tilde{x}_n)$$

$$\Phi_n = g_n\,e(x_n) + C_{p+2}\,y^{(p+2)}(x_n) \quad .$$

The difference of (4.2) and (4.3) gives us an equation for the error $\eta_n = e^*_n - e(x_n)$,

$$(4.4)\ -\eta_{n-1} + 2\,\eta_n - \eta_{n+1} + h^2\,(\beta_0\,(g_{n-1}\,\eta_{n-1} - C_{p+2}\,s(\tilde{\tilde{x}}_{n-1})\,h^2)$$

$$+ \ldots\ldots) = -h^{p+2}\,C_{p+2}\,e^{(p+2)}(\tilde{\tilde{x}}_n)$$

or

$$(J + h^2\,B\,F)\,\eta = C_{p+2}\,s(\tilde{\tilde{x}})\,h^4 - h^{p+2}\,C_{p+2}\,e^{(p+2)}(\tilde{x})$$

$$= v \quad .$$

Using now the fact that $\Omega^{-1} = (J + h^2\,B\,F)^{-1}$ is a positive matrix we get,

$$(4.5)\qquad |\eta| = |\Omega^{-1}\,v| \le \Omega^{-1}|v| \le J^{-1}|v| \quad .$$

It is clear that,

$$|v| \le C_{p+2}\,(K\,h^4 + h^{p+2}\,E_{p+2})\xi$$

where ξ is a vector with all components equal to one, and

$|e^{(p+2)}(x)| \leq E_{p+2} \qquad x \in [a, b]$.

Then

$$|e_n^* - e(x_n)| \leq C_{p+2} (K h^4 + h^{p+2} E_{p+2})(J^{-1} \xi)_n$$

with

$$(J^{-1} \xi)_n = \frac{(x_n - a)(b - x_n)}{2h^2}$$

or by using an uniform bound,

$$\|e_n^* - e(x_n)\|_\infty \leq C_{p+2} (K h^2 + h^p E_{p+2}) \frac{(b - a)^2}{8} \tag{4.6}$$

which finally gives the desired result,

$$e_n^* - e(x_n) = O(h^2) \ . \tag{4.7}$$

Summarizing, the complete procedure to obtain an h^{p+2} order in the discretization error is,

1) Compute $y_n (n = 0,1,...,N)$ by the method of order p given by formula (2.2). The iteration in New-ton method can be stopped when the residuals are less than $K h^{p+2}$.
2) Compute $-h^p e_n^*$ by using (4.2), and add this quantity to y_n. The new approximation will hold (3.8).

The remaining discussion will deal with some possible choices for the approximation (4.1).

The classical choice is,

$$r(x_n) = h^{-p-2}\Delta^{p+2}y_{n+q+1} \qquad (P = 2q)\,. \tag{4.8}$$

By (3.8)

$$\Delta^{p+2}y_{n+q+1} = h^{p+2}\,y^{(p+2)}(x_n) + h^{p+4}\,s(x_n) + O(h^{p+6})$$

where we have assumed enough differentiability on y(x).

In this case the quantity $h^p\,e^*_n$ is called the difference correction by Fox [1957].

By extension we will keep calling difference correction to any quantity computed in this way, whatever the approximation r(x) be.

In the next Section we will give two more expressions for r(x) in the case p = 2. We will also show there, that the use of the difference correction instead of a direct formula with the same order, results in less computational work for the same accuracy. There are two reasons for this saving; on one side the formula used in the Newton iteration is much simpler and on the other side, the number of iterations needed is smaller. That is explained since, when the difference correction is used, the q of (2.9) has only to be equal to p + 2, while in the other case it has to be at least p + 4.

5. Two expressions for the correction term,

As we are considering the equation,

$$y'' = f(x,y)$$

and we want to approximate $y^{(4)}(x)$ $(p = 2)$, a natural idea is to consider,

$$y^{(4)}(x) = \frac{d^2 f(x,y(x))}{dx^2} \tag{5.1}$$

which immediately gives place to two new forms for r(x). We will prove they are valid expressions, in the sense that they satisfy (4.1).

i)

$$r(x) = h^{-2} \delta^2 f(x_n,y_n) \ . \tag{5.2}$$

We want to prove that, if

$$y_n = y(x_n) + h^2 e(x_n) + O(h^4) \tag{5.3}$$

then,

$$\frac{y^{(4)}(x)}{4} \quad \frac{x}{dx} = \frac{d^2 f(x,y(x))}{dx^2} = \frac{\delta^2 f(x_n,y_n)}{h^2} + O(h^2) \ . \tag{5.4}$$

If we were using $y(x_n)$ instead of y_n then (5.4) would be trivially true, but as y_n only satisfies (5.3), some manipulations are needed.

$$\frac{d^2 f(x,y(x))}{dx^2} = f_{xx} + 2 f_{xy} y' + f_{YY}(y')^2 + f_y y'' \ . \tag{5.5}$$

On the other hand,

$$(5.6) \qquad \delta^2 f(x_n,y_n) = f(x_{n-1},y_{n-1}) - 2 f(x_n,y_n) + f(x_{n+1},y_{n+1})$$

and by developing in Taylor series we get,

$$(5.7) \quad \delta^2 f(x_n,y_n) = (\delta^2 y_n) f_y(x_n,y(x_n)) + h (y_{n+1} - y_{n-1})$$

$$f_{xy}(x_n, y(x_n)) + [(y_{n-1} - y(x_n))^2 + (y_{n+1} - y(x_n))^2$$

$$- 2(y_n - y(x_n))^2] + \frac{1}{2} f_{yy}(x_n,y(x_n)) + h^2 f_{xx}(x_n,y(x_n)$$

$$+ O(h^4) .$$

The coefficient of f_{yy} can be expressed in a more convenient way. By using(5.3),

$$(y_{n-1} - y(x_n))^2 + (y_{n+1} - y(x_n))^2 - 2(y_n - y(x_n))^2 =$$

$$= (y(x_{n-1}) - y(x_n) + h^2 e(x_{n-1}))^2 + (y(x_{n+1}) - y(x_n) + h^2 e(x_{n+1}))^2 +$$

$$+ O(h^4) = (-y'(x_n)h + [\frac{1}{2} y''(x_n) + e(x_{n-1})] h^2)^2 +$$

$$+ (y'(x_n)h + [\frac{1}{2} y''(x_n) + e(x_{n+1})] h^2)^2 + O(h^4)$$

and the final expression is,

$$(5.8) \qquad (y_{n-1} - y(x_n))^2 + (y_{n+1} - y(x_n))^2 - 2(y_n - y(x_n))^2 =$$

$$= 2\,(y'(x_n))^2\,h^2 + O(h^4)\ .$$

Then (5.7) and (5.8) imply,

$$\frac{\delta^2 f(x_n,y_n)}{h^2} = f_{xx}(x_n,y(x_n)) + \frac{\delta^2 y_n}{h^2} f_y(x_n,y(x_n)) + 2\frac{Y_{n+1} - y_{n-1}}{2h} \cdot$$

$$\cdot\ f_{xy}(x_n,y(x_n)) + (y'(x_n))^2 f_{yy}(x_n,y(x_n)) + O(h^2)\ .$$

$$= \frac{d^2 f}{dx^2}(x_n,y(x_n)) + O(h^2)$$

which proves (5.4).

An immediate advantage of using $\delta^2 f$ instead of $\delta^4 y$ is that no external values are required to compute the differences at points close to the boundary, avoiding the use of special formulas and information unrelated with the problem.

Since the values $f(x_n,y_n)$ are already computed (from the last iteration in the solution of (2.2)) no extra work is necessary and there is always less computation in carrying 2nd differences compared with the 4th.

ii) In cases in which $f(x,y)$ is easily differentiated, it would be worth to use the approximation,

$$(5.9) \qquad r(x_n) = f_{xx}(x_n,y_n) + f_{xy}(x_n,y_n)\,\frac{y_{n+1} - y_{n-1}}{h} +$$

$$+ f_{yy}(x_n,y_n)\,\frac{(y_{n+1} - y_{n-1})^2}{4h^2} \quad f_y(x_n,y_n)\,f(x_n,y_n) \ .$$

For instance, if $f(x,y)$ is independent of x, (5.9) becomes,

$$r(x_n) = f_{yy}(x_n, y_n)\,\frac{(y_{n-1} - y_{n+1})^2}{4h^2} + f_y(x_n, y_n)\,f(x_n,y_n) \ .$$

If $f(x,y) = g(x)\,y + h(x)$ then,

$$r(x_n) = g''(x_n)\,y_n + g'(x_n)\,\frac{y_{n+1} - y_{n-1}}{h} + g^2(x_n)\,y_n + h''(x_n) +$$
$$+ g(x_n)\,h(x_n)$$

and so on.

The proof that (5.9) is an approximation to $y^{(4)}(x)$ of order at least h^2 goes in the same fashion than the proof for (5.2).

6. Numerical results and comparison of different methods.

We will now state two other finite difference procedures, the Numerov-Milne fourth order approximation, and a truncated version of the Fox difference correction. After that, we will compare them with the two methods described in the previous section and with a shooting type technique.

The Numerov-Milne fourth order method is,

$$(6.1) \qquad Jy = -h^2\,Bf(x,y) + a$$

with $\beta_0 = \beta_2 = 1/12$, $\beta_1 = 10/12$.

B can also be written as,

$$B = I - \frac{1}{12} J \quad .$$

The Fox difference correction with fixed fourth order length uses first, a second order approximation given by the solution of,

$$J\bar{y} = -h^2 f(x,\bar{y}) + a \tag{6.2}$$

then one difference correction in the form,

$$Je = -h^2 F(x,\bar{y})\, e - \frac{1}{12} J^2 \bar{y} \tag{6.3}$$

and finally

$$Y = \bar{y} + h^2 e \quad . \tag{6.4}$$

Thus, the use of fourth differences makes it necessary to compute external values for $\bar{y}$. Fox suggests the use of equation (6.2) to extrapolate values out of the interval of integration, giving the two special formulas,

$$\begin{aligned} y_{-1} &= 2\alpha - y_1 + h^2 f(a,\alpha) \\ y_{N+1} &= 2\beta - y_{N-1} + h^2 f(b,\beta) \quad . \end{aligned} \tag{6.5}$$

Equatians (6.1), and (6.2) through (6.5) will be referred to as Methods I and II, respectively. Methods III and IV will be the ones which stem from formulas (5.2) and (5.9).

The procedure used for these methods is similar to the one used for Method II, the change appearing in equation (6.3).

For Method III we get instead of (6.3),

$$Je = -h^2 F(x,\bar{y})\, e + \frac{1}{12} Jf(x,\bar{y}) \,. \tag{6.6}$$

Method IV expressed in components is,

$$\begin{aligned} -e_{n-1} + 2e_n - e_{n+1} = h^2 f_y(x_n,\bar{y}_n)\, e_n - \frac{1}{12}\Big[& h^2 f_{xx}(x_n,\bar{y}_n) + \\ & + h\, f_{xy}(x_n,\bar{y}_n)(\bar{y}_{n+1} - \bar{y}_{n-1}) + \frac{1}{4} f_{yy}(x_n,\bar{y}_n)(\bar{y}_{n+1} - \bar{y}_{n-1})^2 \\ & + h^2 f_y(x_n,\bar{y}_n)\, f(x_n,\bar{y}_n)\Big] \,. \end{aligned} \tag{6.7}$$

In spite of its complicated aspect, method IV turns out to be the fastest and the most accurate whenever the partial derivatives of f-(x,y) are simple and can be calculated easily.

Now we want to point out a common feature of the methods using the correction difference. We recall that if Newton's method is used to solve (6.2) the formulas are (care has to be taken on the boundary points),

(6.8) $$r(y^{(i)}) = Jy^{(i)} + h^2 f(x, y^{(i)})$$

(6.9) $$\Delta y^{(i)} = -(J + h^2 F(y^{(i)}))^{-1} r(y^{(i)})$$

and

(6.10) $$y^{(i+1)} = y^{(i)} + \Delta y^{(i)} .$$

In solving either the linear systems (6.3), (6.6) or (6.7) we get equations which resemble very much those above. In fact, the changes are: in the expressions for $r(y^{(i)})$; (6.10) becomes $y^{(i+1)} = y^{(i)} - h^2 \Delta y^{(i)}$ and only one iteration is required.

The r(y) corresponding to (6.3), (6.6) and (6.7) are respectively,

(6.11) $$r(y) = -\frac{1}{12h^2} J^2 y$$

(6.12) $$r(y) = \frac{1}{12} Jf(x, \bar{y})$$

(6.13) $$r(y) = -\frac{1}{12} v .$$

In (6.13), v stands for the vector obtained from the second term in the right-hand side of (6.7).

Thus, if the difference correction is combined with Newton's method in the earlier stages, practically the same code can be used in both parts. We have written an Extended Algol program for the B5000 at Stanford which took advantage of this situation. The program modifications for the

different methods were very slight, and the procedure followed in the numerical comparisons has been to introduce these modifications in the most direct fashion.

Another important observation, from the time consuming point of view, is that the quantities $f(x,y)$ and $F(x,y)$ do not have to be computed again in order to perform the difference correction since the values calculated for the last iteration of the Newton method are in general good enough, and no noticeable improvement is observed when these values are recomputed.

We have chosen as our first example a problem which has a known analytical solution and is completely worked out in Collatz [1960] pp.145-147. The method used there is a combination of shooting and interpolation.

By using the same step length, $h = 1/5$, we have computed approximate solutions with the four methods described above.

The problem is,

$$y'' = \frac{3}{2} y^2; \qquad y(0) = 4, \qquad y(1) = 1 \tag{6.14}$$

with one solution equal to

$$y(x) = \frac{4}{(1+x)^2} \tag{6.15}$$

In all the methods the first guess $y^{(0)}$ was constructed from a linear interpolation of the given data

$$y^{(0)}(x) = -3x + 4 \quad .$$

In Table I the values of the five approximate solutions are given; and in Table II information about number of iterations, computing time, and deviation from the true solution is recorded. The subscripts stand for the numbering we have given to the different methods. Method V is the one used in Collatz and y(x) is the exact solution (6.15).

TABLE I

x	y(x)	y_I	y_{II}	y_{III}	y_{IV}	y_V
0	4.00000	4.00000	4.00000	4.00000	4.00000	4.00000
0.2	2.77778	2.77680	2.77718	2.77719	2.77757	2.79464
0.4	2.04082	2.03995	2.04019	2.04019	2.04054	2.05787
0.6	1.56250	1.56191	1.56202	1.56202	1.56226	1.57519
0.8	1.23457	1.23427	1.23431	1.23431	1.23443	1.24138
1.0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00003

TABLE II

	y_I	y_{II}	y_{III}	y_{IV}	y_V
Number of Iterations in Newton Part.	4	3	3	3	-
$\|y(x) - y_{APR.}\|_2$	9.75×10^{-4}	6.29×10^{-4}	6.27×10^{-4}	2.78×10^{-4}	293×10^{-4}
Computation time in seconds 1/	1.70	1.63	1.62	1.63	

1/ In the Burroughs B5000 at Stanford Computation Center.

It is observed that this is a problem in which method IV is fairly convenient. In fact, (6.13) becomes

$$r_n(\bar{y}) = -\frac{1}{12}\left(0.75\,(\bar{y}_{n-1} - \bar{y}_{n+1})^2 + h^2\,4.5\,\bar{y}_n^3\right).$$

Method V is included as a matter of reference, but no attempt is made in comparing it with the finite differences type procedures since they are completely different in principle.

Methods I through IV have been numbered-in order of increasing speed and accuracy. There is no discussion about the accuracy in this example. One word has to be said about the speed. The figures in the third row of Table II show that the computation time was practically the same in all four methods with a tiny seven hundreth of a second in favor of the difference correction. This situation will also be noted in the second problem presented at the end of this section. However, we can mention some reasons which lead us to believe that the ordering is also meaningful in so far as computational speed is concerned.

The solution by Newton's method of the system (6.1) is more complicated than the solution of (6.2) which is basic for all the methods using the difference correction. Moreover, as was mentioned in Section 4, the requirements of precision in these latter methods are less than for the Numerov-Milne method. That implies, that in general less iterations can be expected for methods II, III, and IV than for method I. That is shown in the first row of Tables II and IV. Of course, one more iteration (the difference correction) has to be counted, but in general, as can be seen

in formulas (6.11), (6.12) and (6.13), this iteration involves less computation than the one corresponding to the regular Newton formulas. That is more noticeable after recalling that f and F do not have to be recomputed for this correction.

A last remark is that all the linear systems involved in this discussion are tridiagonal, and a simplified Gauss-type elimination procedure can be used, saving both computation and storage (see, for instance, Henrici [1962] pp. 351-354, or D. H. Thurnau [1963]).

To finish with this section, we present another example which behaves in the same fashion as the first one.

$$y'' = -e^{-2y} \quad ; \quad y(1) = 0 \ , \ y(2) = \ln(2) \ .$$

The exact solution is $y(x) = \ln(x)$.

The step length used was $h = 1/16$, and in Tables III and IV we give the numerical results corresponding to the nodal points $x = 1, 1.25, 1.5, 1.75$. Since

$$f(x,y) = -e^{-2y} \quad ; \quad f_y(x,y) = 2\, e^{-2y} \quad ; \quad f_{yy}(x,y) = -2f_y(x,y)$$

(6.13) becomes

$$r_n(\bar{y}) = -\frac{1}{12} f_y(x_n,\bar{y}_n)[h^2 f(x_n,\bar{y}_n) - 0.5 (\bar{y}_{n+1} - \bar{y}_{n-1})^2] \ .$$

TABLE III

x	y(x)	y_I	y_{II}	y_{III}	y_{IV}
1	0	0	0	0	0
1.25	0.223143551	0.223143676	0.223143656	0.223143656	0.223143525
1.50	0.405465108	0.405465223	0.405465209	0.405465209	0.405465088
1.75	0.559615788	0.559615853	0.559615847	0.559615847	0.559615778

TABLE IV

	y_I	y_{II}	y_{III}	y_{IV}
Number of Iterations in Newton Part.	4	3	3	3
$\|y(x) - y_{APR.}\|_2$	12.9×10^{-8}	10.9×10^{-8}	10.9×10^{-8}	2.7×10^{-8}
Computation time in seconds	4.24	4.17	4.20	4.13

We note again that methods II, III, and IV are about the same in speed and somehow faster than method I. Methods II and III gave practically the same results when h was fairly small. There is a noticeable increase in accuracy when passing from method I to IV.

7. Boundary conditions of Sturm-Liouville type.

In the more general problem (1.1) the conditions on f(x,y) are the same as in the problems of class M discussed before. For the constants a, β, 7, δ we have the following requirements,

$$(7.1) \qquad 0 \leq \alpha, \beta, \gamma, \delta \; ; \quad \alpha\gamma + \alpha\delta + \beta\gamma > 0 \quad .$$

- Under these conditions this problem is also of monotonic type and it has an unique solution (Schröder [1956]).

Now the finite difference procedure has to be modified in points close to the boundary.

To clarify the ideas we will only consider in detail the case p = 2, and the corresponding difference correction. For n = 1 . . . N - 1 the approximation is the same as described in (2.2) (with $\beta_0 = \beta_2 = 0$, $\beta_1 = 1$). Observe that now y_0 and y_n are also unknown. To handle these two new unknowns we need two more equations.

By using the formula,

$$(7.2) \qquad y'(x) = \frac{y(x+h) - y(x-h)}{2h} - \frac{1}{6} h^2 y'''(\bar{x})$$

(without $-\frac{1}{6} h^2 y'''(\bar{x})$) combined with the first boundary condition (1.1') we get at x = a,

(7.3) $$-y_{-1} = \frac{2h}{\beta}(\alpha y_0 - A) - y_1 .$$

By applying (2.2) at x = a (n = 0) we get,

$$-y_{-1} + 2y_0 - y_1 = -f_0 h^2$$

and by using (7.3) and multiplying through by β,

(7.4) $$(2\beta + 2ah) y_0 - 2\beta y_1 = -\beta f_0 h^2 + 2h A$$

Similarly, at x = b (n = N),

(7.5) $$2\delta y_{N-1} + (2\delta + 2\gamma h) y_N = 2B h - \delta f_N h^2$$

With (7.4), (7.5) and the N - 1 equations (2.2) we have now as many equations as unknowns. On the other hand the exact solution satisfies,

(7.6) $$(2\beta + 2\alpha h) y(x_0) - 2\beta y(x_1) =$$

$$-\beta f(x_0, y(x_0)) h^2 + 2h A - \frac{\beta}{12} h^4 y^{IV}(x_0)$$

$$-\frac{h^3}{3}\beta y'''(x_0) + O(h^5)$$

and,

$$(7.7) \qquad (2\delta + 2h\ 7)\ y(x_N) - 2\delta\ y(x_{N\,1}) =$$

$$= -\ h^2\ \delta\ f(x_N,\ y\ (x_N)\ ) + 2h\ B + \delta\ \frac{h^3}{3}\ y'''\ \ (x_N) - \frac{1}{12}\ h^4\ \delta\ y^{IV}(x_N) + O(h^5)$$

Now, by taking the corresponding differences we get for the error of discretization at the boundary points,

$$(7.8) \qquad (2 + \frac{2\alpha h}{\beta} + g_o\ h^2)\ e_o - 2\ e_1 = \Theta_o''\ (h^3\ M_1 + h^{q+2} K)$$

$$(2 + \frac{2h\gamma}{\delta} + g_N\ h^2)\ \ e_N - 2e_{N-1} = \ \Theta_N''\ (h^3\ M_2 + h^{q+2} K)\ .$$

Together with (7.8) we have the equations for the inner points,

$$(7.9) \quad -\ e_{n\,1} + 2\ e_n - e_{n+1} + h^2\ g_n\ e_n = \Theta_n''\ (h^{q+2}\ K + h^4\ G\ Z) \quad \text{(Henrici (l.c.) p.375)}$$

or in matrix form,

$$(7.10) \qquad s\ e = b$$

where,

$$(7.11) \qquad s_{ii} = 2 + h^2\ g_i \qquad i = 1 \ldots N\text{-}l$$

$$S_{00} = 1 + \frac{\alpha h}{\beta} + \frac{g_o h^2}{2}$$

$$S_{NN} = 1 + \frac{\gamma h}{\delta} + \frac{g_N h^2}{2}$$

$$S_{i,i+1} = S_{i-1,i} = -1 \qquad i = 0, \ldots, N$$

and

$$b_i = \Theta_i'' \left(h^{q+2} K + h^4 G Z\right) \qquad 7 = 1 \ldots N-1$$

(7.12)
$$b_0 = \Theta_0'' \left(h^{q+2} \frac{K}{2} + h^3 \frac{M}{2}\right)$$

$$b_N = \Theta_N'' \left(h^{q+2} \frac{K}{2} + h^3 \frac{M}{2}\right)$$

If β and δ are different from zero then S is irreducible, otherwise we can skip the corresponding equation and the resulting matrix will be irreducible. Moreover,

i) $$S_{ij} \leq 0 \qquad i \neq j \qquad i,j = 0 \ldots N$$

ii) $$\sum_{j=0}^{N} S_{0j} = \frac{a h}{\beta} + g_0 \frac{h^2}{2} > 0$$

$$\sum_{j=0}^{N} S_{ij} = h^2 g_i \geq 0$$

$$\sum_{j=0}^{N} S_{Nj} = \frac{\gamma h}{\delta} + g_N \frac{h^2}{2} > 0 \ .$$

Consequently S is monotone.

Now we will try to find a bound for the discretization error.

(7.13)

$$S = \begin{pmatrix} 1 + P_0 & -1 & 0 & & & \\ -1 & 2 + P_1 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 + P_{N-1} & -1 \\ & & & 0 & -1 & 1 + P_N \end{pmatrix} > \geq \begin{pmatrix} A & v \\ v^T & 1 + mh \end{pmatrix} = \tilde{S}$$

-where

$$v^T = (0, 0, \ldots, 0, -1)$$

$$m = \gamma/\delta \qquad (\alpha, \gamma > 0)$$

and

$$A = \begin{pmatrix} 1 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 2 \end{pmatrix}$$

We know that

$$0 < S^{-1} \leq \tilde{S}^{-1} \quad .$$

Let's compute a bound for $\tilde{S}^{-1}$.

$$\tilde{S}^{-1} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

where the C_{ij} are blocks with the same sizes as the corresponding ones in $\tilde{S}$.

From Householder [1953], p. 78, we obtain for $\tilde{S}^{-1}$,

$$C_{22} = \frac{1}{1 + mh - a_{NN}^{-1}}$$

$$C_{21} = C_{22} \, (a_{N1}^{-1}, \ldots, a_{NN}^{-1})$$

(7.14)

$$C_{12} = C_{21}^{I}$$

$$C_{11} = A^{-1} \left(I + \begin{pmatrix} \bigcirc \\ C_{21} \end{pmatrix}\right)$$

with a_{ij}^{-1} being the elements of A^{-1} .

It is easy to show that,

$$A^{-1} = \begin{pmatrix} N & N-1 & N-2 & \dots & 1 \\ N-1 & N-1 & N-2 & \dots & 1 \\ \cdot & & & \cdots & \\ \cdot & & \cdot & \cdots & \\ \cdot & \cdot & & \cdots & \\ 1 & 1 & & \dots & 1 \end{pmatrix}$$

Using this result in (7.14) we obtain,

$$C_{22} = \frac{1}{1 + mh - 1} = \frac{1}{mh}$$

$$C_{21} = \frac{1}{mh} (1, \dots 1)$$

$$(C_{11})_{ij} = a_{ij}^{-1} + \frac{a_{iN}}{mh} \leq (N-i) + \frac{1}{mh}$$

Then, $$|e_i| < S^{-1} \|b\|$$

$$|e|_i \leq |b_0| [N-i + \frac{1}{mh}] + (N-i + \frac{1}{mh}) \sum_{j=1}^{N-1} |b_j| + \frac{1}{mh} |b_N|$$

$$i = 0, \dots N-1$$

$$|e|_N \leq \frac{1}{mh} [|b_0| + |b_n| + \sum_{j=1}^{N-1} |b_j|]$$

Since $$(N-i + \frac{1}{mh}) \sum_{j=1}^{N-1} |b_j| \leq (N-1)^2 [h^{q+2} K + h^4 G Z]$$

$$\leq h^q K' + h^2 G' Z (b-a)$$

and $$|b_0| [N-i + \frac{1}{mh}] < h^{q+1} K' + h^2 M'$$

we finally obtain,

(7.15) $$\|e\|_\infty \leq 2h^2 [G' Z (b-a) + M'] + 2K'h^q[1+h]$$

which is the result (2.8) corresponding to this problem. For the interior points the treatment of Section 4 is carried over in this case, the only changes appearing at the boundary points. Consequently (4.2) is used for n=1 . . N-1, and also at 0, N in order to construct the corresponding modified equations. We also need $y'''(x) = t(x) + h^2 u(x) + 0 (h^4)$ with t, u smooth. Now we are able to define the two boundary equations to be added to (4.2),

$$(2h\alpha+2 \beta + h^2 \beta g(0))e_0^* - 2 \beta e_1^* = \frac{\beta h^2}{12} r(0) + \beta \frac{h}{3} t(0)$$

(7.16)

$$(2h\gamma + 26 + h^2\delta g(N)) e_N^* - 26 e_{N-1}^* = \frac{\delta}{12} h^2 r(N) - \frac{\delta h}{3} t(N) .$$

The equivalent of (3.6) is,

$$e''(x) = g(x)\, e(x) + C_4 y^{(4)}(x)$$

(7.17)
$$\alpha\, e(a) - \beta\, e'(a) = \frac{\beta}{6} y'''(a)$$

$$y\, e(b) + \delta\, e'(b) = \frac{\delta}{6} y'''(b)$$

whose solution satisfies (4.3) at the interior mesh points and

$$(2h\alpha + 2\beta + h^2 \beta\, g(0))\, e(x_0) - 2\beta\, e(x_1) =$$

$$= \beta \frac{h^2}{12} y^{(4)}(x_0) + \frac{h\beta}{3} y'''(x_0) + \frac{h^4}{12} e^{(4)}(\bar{\bar{x}}_0)$$

$$(2\gamma h + 2\delta + h^2 \delta\, g(N))\, e(x_N) - 2\, 6\, e(x_{N\,1})$$

$$= \frac{\delta}{12} h^2 y^{(4)}(x_N) - \frac{h\delta}{3} y'''(x_N) + \frac{h^4}{12} e^{(4)}(\bar{\bar{x}}_N)$$

at the boundary.

From here we obtain the η equations,

$$(1 + \frac{\alpha}{\beta} h + \frac{h^2}{2} g(0))\, \eta_0 - \eta_1 =$$

(7.19)
$$= -\frac{h^4}{12} s(\tilde{x}_0) - \frac{h^3}{3} u(\tilde{\tilde{x}}_0) - \frac{h^4}{12} e^{(4)}(\bar{\bar{x}}_0)$$

$$(1 + \frac{\gamma}{\delta} h + \frac{h^2}{2} g_N)\, \eta_N - \eta_{N-1} = -\frac{h^4}{12} s(\tilde{x}_N) + \frac{h^3}{3} u(\dot{\tilde{\tilde{x}}}_N) - \frac{h^4}{12} e^{(4)}(\bar{\bar{x}}_N)$$

But now the η system we have obtained is of the form,

$$S\,\eta = v \tag{7.20}$$

where v has components with the same orders in h as b in (7.10). Hence the same result (7.15) is obtained for $\|\eta\|_\infty$ (with different constants),

$$\|\eta\|_\infty = O(h^2) \tag{7.21}$$

which finally implies,

$$e_n^* - e(x_n) = O(h^2)\ . \tag{7.22}$$

Consequently, in this more general problem, the difference correction, being applied not only to the differential equation, but also to the boundary conditions, improves the solution in two orders in h, as in the simpler case.

8. Numerical example for the Sturm-Liouville case.

Equation (6.14) with the boundary conditions,

$$\begin{cases} y(0) - 2\,y'(0) = 20 \\ \\ 2\,y(1) + 3\,y'(1) = -1 \end{cases} \tag{8.1}$$

was integrated by using a suitable modification of method III. Since this

problem has the same solution (6.15) as before we only list the new results. This time, the first guess (a linear function) turned out to be fairly bad,. forcing several Newton iterations before reaching the required precision.

Step (h)	Number of Newton iter.	comput. Time (Sec)	$\epsilon = \| y(x) - y_{APR} \|_2$ (before diff. correc.)	ϵ after diff. correc.
1/5	5	2.6	1.1 x 10 -1	9.6 x 10-3
1/20	6	8.1	7.8 x 10-3	4 x 10 -5

Acknowledgments

I want to express my gratitude to Professor Gene Golub, of Stanford University, who called my attention to this problem and who has guided. this research always with patience and kindness.

I also give my sincere thanks to the authorities and staff of the Stanford Computation Center who have made it possible for me to complete this work.

REFERENCES

Bickley W. G. Michaelson S., and Osborne M. R. [1960] "On finite-difference methods for the numerical solution of boundary-value problems" Proc. Roy. Soc. London, A262, pp. 219-236.

-Brown, Robert R.[1962] "Numerical solution of boundary value problems using nonuniform grids" J. Soc. Indust. Appl. Math. 10 pp. 475-495.

Collatz L.[1960] The Numerical Treatment of Differential Equations, 3rd. Edition, Springer, Berlin.

Fox, L. [1957] The Numerical Solution of Two Points Boundary Value Problems in Ordinary Differential Equations, Oxford Univ. Press.

Fox, L. [1962] Numerical Solution of Ordinary and Partial Differential Equations, Pergamon Press; Oxford.

Fox, L. and Goodwin, E. T. [1949] "Some new methods for the numerical integration of ordinary differential equations' Proc. Camb. Phil. Soc. 45 pp. 373-388.

Haselgrove, C. B. [1961] "The solution of non-linear equations and of differential equations with two-point boundary conditions" Comp. J., 4 pp. 255-259.

Henrici, Peter [1962] Discrete Variable Methods in Ordinary Differential Equations, Wiley, New York.

Householder, Alston S. [1953] Principles of Numerical Analysis, McGraw-Hill Book Co., New York.

Mayers, D. F. [1964] "The deferred approach to the limit in ordinary differential equations" Comp. J., 7 pp. 54-57.

Milne, William E. [1953] Numerical Solution of Differential Equations, Wiley, New York.

P.I.C.C. Symposium on the Numerical Treatment of Ordinary Differential Equations, Integral and Integro Differential Equations [1960]. Birkhauser, Stuttgart.

Schröder, Johann [1956] "Über das Differenzenverfahren bei nichtlinearen Randwertaufgaben" I, Z. angew. Math. Mech., 36, pp. 319-331; II, pp. 443-455.

Thurnau, Donald H. [1963] "Algorithm 195, Bandsolve" CACM, 6 p. 441.

Todd, John [1950] "Solution of differential equations by recurrence relations" MTAC, 4 pp. 39-44.

Varga, Richard S. [1962] Matrix Iterative Analysis, Prentice Hall, New Jersey.