ON SOME NUMERICAL INTEGRATION METHODS FOR SOLVING ALGEBRAIC EQUATIONS

By

Kai-Wen Tu

DCS-TR-#29

Department of Computer Science
Rutgers University
The State University of New Jersey
New Brunswick, New Jersey 08903

May 1974

1. Introduction:

This paper examines the iterative schemes arise from the numerical methods for the integration approach [3,5,7] to nonlinear algebraic equations. Convergence of some multistep methods and single-step methods has been demonstrated in [1,6]. While these methods do not require as good an initial approximation as Newton-Raphson method, they have not been widely implemented due to (a) unavailability of convergence rate, (b) a lack of rapid convergence close to the solution. In this paper we shall show that the problems of (a) and (b) are closely related to the stepsize for numerical integration. Specifically, rapid convergence can be attained with an judicious choice of stepsize and the resulting rate of convergence is either quadratic or weaker depending on the numerical method. It follows that Euler's method, third order Runge-Kutta method and the two-correction trapezoidal rule are quadratically convergent with stepsizes 1, 1.596071638 and 1.295597743 approximately. The first result is obvious since the Euler's method with stepsize 1 reduces to the Newton-Raphson method. The findings on non-Euler methods provide significant insight into the numerical integration approach to algebraic problems.

2. The difference equations

Let

$$F(X) = 0 \qquad (2.1)$$

be a system of nonlinear algebraic equations. Following [2,10]

we apply numerical integration techniques to the differential equation

$$\frac{dX}{dt} = - X + G(X) \tag{2.2}$$

where

$$G(X) = J^{-1}(F + JX) \tag{2.3}$$

J = the Jacobian of F

to find a solution X^* of (2.1). Since we are mainly interested in the convergence at a small region containing X^*, we will be concerned only with fixed mesh integration processes here. For a linear k-step method we have

$$\sum_{i=0}^{k} (\alpha_i - h\beta_i)X_{n-1} + h\sum_{i=0}^{k} \beta_i G_{n-i} = 0 \tag{2.4}$$

Let

$$e_{n-i} = X_{n-i} - X^* \tag{2.5}$$

With this definition e_{n-i} is not the usual numerical error between the solutions of (2.2) and (2.4). Combining (2.4) and (2.5) we have

$$\sum_{i=0}^{k} (\alpha_i - h\beta_i)e_{n-i} = -h\sum_{i=0}^{k} \beta_i (G_{n-i} - X^*) \tag{2.6}$$

Similarly, with the application of an explicit Runge-Kutta type method [6] to (2.2) we obtain the difference equation

$$e_n - r(h)e_{n-1} = s_n \tag{2.7}$$

Where r(h) is a polynomial of h and s_n depends on h and G values in a neighborhood of X_{n-1}.

Both (2.6) and (2.7) can be treated in a general setting. Consider the linear difference equation

$$c_k z_n + c_{k-1} z_{n-1} + \cdots + c_o z_{n-k} = d_n, \quad n = k, k+1, \ldots \tag{2.8}$$

where $c_o, c_1, \ldots, c_k$ are fixed constants, while d_n is a given sequence with

$$|d_n| \leq \hat{\xi} |z_{n-1}| \qquad \text{for some } \hat{\xi} > 0 \tag{2.9}$$

We shall show that the solution growth of (2.8) is directly related to the roots of the characteristic polynomial

$$p(z) = c_k z^k + c_{k-1} z^{k-1} + \cdots + c_o \tag{2.10}$$

The solution of (2.8) can be obtained [8] from the power series development of

$$z(x) = \sum_{\tau=1}^{t} \frac{P_\tau(x) D(x)}{(1 - w_\tau x)^{m_\tau}} \qquad \text{for some } t \leq k \tag{2.11}$$

where

$$D(x) = \sum_{i=0}^{k-1} d_i x^i + \sum_{i=k}^{\infty} d_i x^i \tag{2.12}$$

and $\sum_{i=0}^{k-1} d_i x^i$ can be chosen arbitrarily, w_j's are the roots of (2.10) with multiplicity m_τ, and $P_\tau(x)$ is a polynomial of degree $m_\tau - 1$. We write

$$P_\tau(x) D(x) = \sum_{i=0}^{\infty} d_i^{(\tau)} x^i \tag{2.13}$$

Assume $m_\tau = 1$ for all τ and choose $w = \max |w_\tau|$. The coefficient of x^n in (2.11) is

$$z_n = \sum_{\tau=1}^{t} \sum_{j=0}^{n} w_\tau^{(n-j)} d_j^{(\tau)} \tag{2.14}$$

Thus

$$|z_n| \leq \xi \left(w^n + |z_o| w^{n-1} + \cdots + |z_{n-1}|\right) \tag{2.15}$$

where ε depends on $\hat{\varepsilon}$ and is of $O(\hat{\varepsilon})$. By choosing the initial points $z_0, z_1, \ldots, z_{k-1}$ conveniently and using an induction argument it follows that

Theorem 1 The solution z_n of (2.8) satisfies

$$|z_n| \leq \varepsilon (w + \varepsilon)^{n-1}(w + |z_0|) \tag{2.16}$$

provided (2.9) holds and the roots of (2.10) are simple.

It is clear that the preceding results extend to vector-valued difference equation when the scalar norm $|.|$ is replaced by the uniform norm $\|.\|$. Thus, according to (2.16), we seek to minimize w to accelerate the reduction of $\|e_n\|$ in (2.6) and (2.7).

3. Optimum step size

The coefficients $c_0, c_1, \ldots, c_k$ in (2.8) are dependent on h. Thus our main objective here is to find for each method a stepsize h* which yields the smallest $w = w^*$. Such an h* exists but the corresponding w* is not necessarily 0. In fact this is the source of shortcomings for k-step methods when $k \geq 2$. We elaborate with the following discussions.

(a). Adams type methods

It is easy to see that we have $h^* = 1$, $w^* = 0$ for Euler's method. For an Euler-trapezoidal rule predictor-corrector method the underlying difference equations are prediction:

$$X_n^{(p)} - (1 - h)X_{n-1} = hG_{n-1} \tag{3.1}$$

first correction:

$$X_n^{(1)} - (1 - h + h^2/2)X_{n-1} = h/2\,(G_n^{(p)} + (1 - h)G_{n-1}) \tag{3.2}$$

second correction:

$$X_n^{(2)} - (1 - h + h^2/2 - h^3/4)X_{n-1}$$

$$= h/2\ (G_n^{(1)} - h/2\ G_n^{(p)} - (1 + h/2 - h^2/2)G_{n-1}) \tag{3.3}$$

where $G_n^{(p)} = G(X_n^{(p)})$ and $G_n^{(1)} = G(X_n^{(1)})$

Transforming (3.1) and (3.2) into the same form as (2.6) and subsequently (2.8) we find that

$$w^{(1)} = |1 - h + h^2/2| \tag{3.4}$$

$$w^{(2)} = |1 - h + h^2/2 - h^3/4| \tag{3.5}$$

where $w^{(i)}$ is the w for the ith corrector. It is obvious that

$$1 - h + h^2/2 > 0 \qquad \text{for all } h \tag{3.6}$$

and

$$\min_h |1 - h + h^2/2| = |1 - h + h^2/2|_{h=1} = 1/2 \tag{3.7}$$

However, $w^{(2)}$ vanishes when h is approximately 1.295597743. Therefore the two-correction scheme is preferred here.

Let $\tilde{G} = G - X^*$ and $e = X - X^*$. When $w^{(2)} = 0$ we transform (3.3) into an equation correspond to (2.6):

$$e_n^{(2)} = h/2\ (\tilde{G}_n^{(1)} - h/2\ \tilde{G}_n^{(p)} - (1 - h/2 + h^2/2)\tilde{G}_{n-1}) \tag{3.8}$$

According to Ortega and Rheinboldt [7] we have

$$\|\tilde{G}\| \leq A\|e\|^2 \tag{3.9}$$

for an appropriate G and sufficiently small $\|e\|$, where A is a constant. Now the right side of (3.8) can be bounded in a manner similar to Euler's method with h = 1. Namely, there exists

$A_t \geq 0$ such that

$$\| h/2\ (\tilde{G}_n^{(1)} - h/2\ \tilde{G}_n^{(p)} - (1 + h/2 - h^2/2)\tilde{G}_{n-1})\| \leq A_t \|e_{n-1}\|^2 \quad (3.10)$$

In general, when $k \geq 2$ the condition $w^* = 0$ implies that $c_1, \ldots, c_k$ are to be annihilated simultaneously with $h = h^*$ for some h^*. According to (2.6) we then have

$$\alpha_i - h\beta_i = 0, \qquad i = 1, 2, \ldots, k \quad (3.11)$$

Consequently, the ratio α_i/β_i should be a constant equal to h^* for all i. For Adams type methods this is clearly unattainable since $\alpha_1, \beta_1, \beta_2, \ldots, \beta_k$ are nonvanishing whereas $\alpha_2 = \alpha_3 = \ldots = \alpha_k = 0$. Hence $w^* > 0$ when $k \geq 2$. This points out the shortcomings of k-step Adams methods ($k \geq 2$) in the sufficiently small regions containing the solution X^*.

(b). Runge-Kutta type methods

With an explicit Runge-Kutta method the underlying characteristic polynomial (2.8) is linear. Hence we only have to deal with the norm reduction of a single root. Let r_i be such a root for the ith order Runge-Kutta method. According to (2.7) we have

$$r_1(h) = 1 - h \quad (3.12)$$

$$r_2(h) = 1 - h + h^2/2 \quad (3.13)$$

$$r_3(h) = 1 - h + h^2/2 - h^3/6 \quad (3.14)$$

$$r_4(h) = 1 - h + h^2/2 - h^3/6 + h^4/24 \quad (3.15)$$

These are the truncated exponential series of exp(-h). When i is even and $h \geq 0$

$$r_i(h) = \exp(-h) + (h^{i+1}/(i+1)!)\exp(-\tilde{h}) > 0 \quad (3.16)$$

where $\tilde{h}$ is between 0 and h.

Thus $w^* > 0$ for the even order Runge-Kutta methods.

When i = 1 we again have Euler's method. It is found that r_3 vanishes when h is approximately 1.596071638. This again leads to quadratic convergence here since for some suitable $A_r \geq 0$ it can be shown that

$$\| S_n \| \leq A_r \| e_{n-1} \|^2 \tag{3.17}$$

The preceding results can be extended to a general ith order Runge-Kutta method. In view of the alternating sign of the coefficients in $r_i(h)$ a real root of $r_i(h)$, whenever it exists, is always positive. Thus we can always find an optimum stepsize h* for an ith order Runge-Kutta method when i is odd.

We have established that

Theorem 2 Euler's method, third order Runge-Kutta method and the two correction trapezoidal rule are quadratically convergent, ie, w* = 0, with the integration stepsizes h* = 1, 1.596071638, 1.295597743 approximately.

4. Numerical results and discussions

We have carried out numerical computation of the following four problems on IBM 360/67:

1. $f_1 = 10 (x_2 - x_1^2)$

   $f_2 = 1 - x_1$

   Initial value: (0.8,0.4), Solution: (1,1)

2. $f_1 = 1/2 \sin (x_1 x_2) - x_2/4\pi - x_1/2$

   $f_2 = (1 - 1/4\pi) (\exp(2x_1) - e) + ex_2/\pi - 2ex_1$

   Initial value: (0.55,3.0), Approximate solution: (0.3,2.8)

3. $f_1 = 4 + x_1 + x_2 - x_1^2 + 2x_1 x_2 + 3x_2^2$

   $f_2 = 1 + 2x_1 - 3x_2 + x_1^2 + x_1 x_2 - 2x_2^2$

   Initial value: (1.0,-4.0), Approximate solution: (3.339,-2.984)

4. $f_1 = x_1^2 - x_2 + 1$

$f_2 = x_1 - \cos(\pi x_2/2)$

Initial value: (0.2,0.8), Solution: (0,1)

The convergence criteria is $\|F_n\| < 10^{-12}$. The numerical results, as summarized in Table 1, indicate that

(a) All methods with $w^* = 0$ and $h = h^*$ require comparably the same number of integration steps for convergence, and

(b) A method (the classical Runge-Kutta method) with $w^* \neq 0$ and $h = h^*$ requires substantially more integration steps for convergence.

In a sufficiently small region U containing the solution X^* the underlying trajectory $X(t)$ of (2.2) is not necessarily t-optimal. In other words, assuming the numerical integration is exact the relation between $\|X(t_n) - X^*\|$ and $\|X(t_{n-1}) - X^*\|$ may be weaker than quadratic. This has been observed by previous researchers (see for example Meyer [11]). Thus in the region U it is more desirable to attain quadratic convergence than to maintain high accruacy in numerical integration. The general problem of (2.2) in a broader region containing U can be treated by variable mesh integration methods [1,6,10]. In a previous work [10] we have established necessary step changing conditions for convergence. Interestingly, the use of h^* as the maximum stepsize for a variable mesh scheme is consistent with these conditions and leads to a more efficient stepsize control algorithm.

Table 1 Number of steps required to achieve $\| F \| < 10^{-12}$

		Problem Number			
Method	Stepsize	(1)	(2)	(3)	(4)
Euler	0.5	43	38	47	40
	1.0	3	5	7	6
	1.5	42	39	46	39
RK3	1.0	28	25	30	26
	1.596071638	4	5	7	6
	2.0	27	26	29	26
RK4	1.0	31	28	34	28
	1.596071638	23	21	25	22
	2.0	27	25	31	25
TR	1.0	22	20	24	21
	1.295597743	4	5	7	6
	1.5	19	19	22	19

RK3 -- third order Runge-Kutta method

RK4 -- fourth order Runge-Kutta method

TR -- two correction trapezoidal rule

References

(1) P.T. Boggs, "The solution of nonlinear systems of equations by A-stable integration techniques", SIAM J. Numer. Anal. Vol. 8, No. 4, 767-785 (1971)

(2) C.G. Broyden, "A class of methods for solving nonlinear simultaneous equations", Math. Comp. 19, 577-593 (1965)

(3) D.F. Davidenko, "On a new method of numerical solution of system of nonlinear equations", Dokl, Akad. Nauk. SSSR 88, 601-602 (1953)

(4) P. Henrici, "ERROR PROPAGATION FOR DIFFERENCE METHODS", John Wiley, New York (1962)

(5) M.N. Jacovlev, "On the solution of systems of nonlinear equations by differentiation with respect to a parameter", USSR Comput. Math. and Math Phys. 4, 146-149 (1964)

(6) W. Kizner, "A numerical method for finding solutions of nonlinear equations", SIAM J. Appl. Math., 12, 424-428 (1964)

(7) J.M. Ortega and W.C. Rheinboldt, "ITERATIVE SOLUTION OF NONLINEAR EQUATIONS IN SEVERAL VARIABLES", Academic Press, New York, (1970)

(8) A.M. Ostrowski, "SOLUTIONS OF EQUATIONS AND SYSTEMS OF EQUATIONS", Academic Press, N. Y. (1966)

(9) P. Rabinowitz, "NUMERICAL METHODS FOR NONLINEAR ALGEBRAIC EQUATIONS", Gordon and Breach, New York (1970)

(10) K.W. Tu, "A class of variable mesh multistep methods for solving simultaneous nonlinear equations", Department of

Computer Science Report DCS-TR-#27, Rutgers University (1974)

(11) G.H. Meyer, "On solving nonlinear equations with a one-parameter operator embedding", SIAM J. Numer. Anal. 5, 739-752 (1968)