ON ACCURATE SOLUTION OF BOUNDARY VALUE ORDINARY DIFFERENTIAL EQUATIONS*

R. P. Tewarson† and S. Gupta

Applied Mathematics and Statistics Department
State University of New York
Stony Brook, N.Y. 11794

Abstract -- Zusammenfassung

On Accurate Solution of Boundary Value Ordinary Differential Equations. A method for solving two-point boundary value ordinary differential equations is given. It requires six function evaluations in each subinterval for an $O(h^7)$ local error, where h is the size of each subinterval. No information outside the subinterval is used. Results of computational experiments are given showing that the method compares quite favorably with classical Runge-Kutta methods.

Ueber eine präzise Lösung von Randwert Gewöhnlichen Differenzialgleichungen. Es wird eine Methode zur Lösungen von Zwei-Punkt Randwert Gewöhnlichen Differenzialgleichungen gegeben. Sie fordert sechs Funktionalbewertungen in jedem Sub-Interval für einen Lokalfehler von $O(h^7)$, wobei h die Grösse jedes Sub-Intervals ist. Dabei wird keine Auskunft ausserhalb des Sub-Intervals verwendet. Die Resultate der Rechnungsexperiemnte werden angegeben und zeigen, dass diese Methode Vorteile gegenüber der klassischen Runge-Kutta Methoden bietet.

†Please send all correspondence to Prof. Dr. R. P. Tewarson, Applied Math. Dept., Math Tower, SUNY at Stony Brook, N.Y. 11794.

*Research supported by NIH Grant No. AM17593.

1. Introduction

We consider the numerical solution of the system of differential equations

$$y'(x) - f(x,y(x)) = 0, \tag{1.1}$$

with the two-point boundary conditions

$$g(y(0),y(1)) = 0, \tag{1.2}$$

where y,f and g are functions with values in R^m, and $0 \leq x \leq 1$.

Let us subdivide the x range [0,1] into n equal parts, such that $h = 1/n$ and $x_i = ih$, $i = 0,1,\ldots,n$. If we integrate the p-th equation of the system (1.1) in the interval $[x_{i-1},x_i]$ and denote the resulting quantity by $\phi_{pi}(y)$ then we have

$$\phi_{pi}(y) = y_{pi} - y_{p,i-1} - \int_{x_{i-1}}^{x_i} f_p(x,y(x))dx = 0, \tag{1.3}$$

where $y_{pi} = y_p(x_i)$ and $p = 1,2,\ldots,m$; $i = 1,2,\ldots,n$. The integral in (1.3) can be evaluated by the various numerical schemes. For example, the well known trapezoidal rule leads to

$$\phi_{pi}(y) = y_{pi} - y_{p,i-1} - \frac{h}{2}\left[f_{pi}+f_{p,i-1}\right] + O(h^3) = 0, \tag{1.4}$$

where $f_{pi} = f_p(x_i,y(x_i))$, y is now a vector with the components y_{pi}. On the other hand, the use of Simpson's rule [1] yields

$$\phi_{pi}(y) = y_{pi} - y_{p,i-1} - \frac{h}{6}\left[f_{pi}+4f_{p,i-\frac{1}{2}}+f_{p,i-1}\right] + O(h^5). \tag{1.5}$$

In order to compute $f_{p,i-\frac{1}{2}}$ we require the $y_{p,i-\frac{1}{2}}$ values. It is easy to show

by Taylor's theorem that

$$y_{p,i-\frac{1}{2}} = \frac{y_{pi}+y_{p,i-1}}{2} - \frac{h}{8}\left\{y'_{p,i-1}-y'_{p,i}\right\} + O(h^4).$$

In view of (1.1) and the above equation we get

$$y_{p,i-\frac{1}{2}} = \frac{y_{pi}+y_{p,i-1}}{2} + \frac{h}{8}\left\{f_{pi}-f_{p,i-1}\right\} + O(h^4), \tag{1.6}$$

and it follows from (1.5) that the error will remain $O(h^5)$ when (1.6) is used to evaluate $f_{p,i-\frac{1}{2}}$.

We have shown in [2] how the cubic spline on spline and the quintic splines can be used to get $O(h^6)$ and $O(h^7)$ formulas (See equations (2.21) and (2.28) in [2]).

Let us denote by ϕ and g the vectors with the components ϕ_{pi} and g_p respectively. Then (1.4), (1.5) or equations (2.21) and (2.28) in [2] can be written as

$$\phi(y) = 0, \tag{1.7}$$

and (1.2) as

$$g(y) = 0. \tag{1.8}$$

Clearly, $\phi \in R^{mn}$, $y \in R^{m(n+1)}$, and $g \in R^m$. It is worth noting that in (1.7), m components of y can be eliminated directly by expressing them as linear functions of the rest of the components by using (1.8), provided that g(y) is a linear function of y. We will assume that g(y) is linear and the necessary elimination of y components has been done so that we only have to solve (1.7). Newton's method is used to solve (1.7) as follows: Given $y^{(0)}$, for k = 0,1,2,... the following steps are done until $\| \phi \ (y^{(k)}) \|$ is less than a given

quantitity. The system of linear equations

$$J(\phi)\delta y^{(k)} = \phi(y^{k}) \tag{1.9}$$

is solved for $\delta y^{(k)}$, where $y^{(k)}$ is the k-th approximation to a root of (1.7) and $J(\phi)$ denotes the Jacobian of ϕ with respect to y evaluated at $y^{(k)}$. Then the next approximation to the root is given by

$$y^{(k+1)} = y^{(k)} - \delta y^{(k)}. \tag{1.10}$$

The Jacobian matrices corresponding to the trapezoidal and Simpson rules are sparse but the cubic spline on spline or quintic spline methods lead to significantly less sparse Jacobians. Therefore, a modified Newton method, which does not have the quadratic convergence of the usual Newton method, must be used to handle this situation [2]. Also, the cubic spline on spline and the quintic spline methods require, respectively, the solution of tri-diagonal and penta-diagonal system of linear equations in addition to the solution of the linear system (1.9).

In the next section, we give a method which has an $O(h^7)$ local error. This method is, in a sense, an extension of our implementation of Simpson's rule [1]. As in the case of the Trapezoidal and Simpson rules, the Jacobian of the present method is sparse since no information outside the interval $[x_{i-1},x_i]$ is used. In the last section of this paper, we give some results of our computational experiments showing that the present method is significantly better than other methods in terms of accuracy and overall computational cost.

2. Main Results

For the sake of clarity of presentation, let us drop the subscript p on y and f in this section.

THEOREM: If $f \in C^6$ in $[x_{i-1}, x_i]$, y_i and y_{i-1} are given, and

$$\bar{y}_{i-3/4} = \frac{1}{64}\left[54y_{i-1}+10y_i+h(9f_{i-1}-3f_i)\right], \tag{2.1}$$

$$\bar{y}_{i-1/4} = \frac{1}{64}\left[10y_{i-1}+54y_i+h(3f_{i-1}-9f_i)\right], \tag{2.2}$$

$$\hat{y}_{i-1/2} = \frac{1}{2}(y_{i-1}+y_i)+h\left[\frac{1}{24}(f_{i-1}-f_i)+\frac{1}{6}(\bar{f}_{i-3/4}-\bar{f}_{i-1/4})\right], \tag{2.3}$$

$$\hat{y}_{i-3/4} = \frac{1}{256}\left[90y_{i-1}+22y_i+144y_{i-1/2}+h(9f_{i-1}-3f_i-36\hat{f}_{i-1/2})\right], \tag{2.4}$$

$$\hat{y}_{i-1/4} = \frac{1}{256}\left[22y_{i-1}+90y_i+144y_{i-1/2}+h(3f_{i-1}-9f_i+36\hat{f}_{i-1/2})\right], \tag{2.5}$$

then

$$\phi(y) = y_i - y_{i-1} - \frac{h}{90}\left[7(f_{i-1}+f_i)+32(\hat{f}_{i-3/4}+\hat{f}_{i-1/4})+12\hat{f}_{i-1/2}\right]+O(h^7) = 0, \tag{2.6}$$

where

$$\bar{f}_j = f(x_j,\bar{y}_j),\ \hat{f}_j = f(x_j,\hat{y}_j),\ j = i-\frac{3}{4},\ i-\frac{1}{4}.$$

PROOF: Using Taylor's theorem to expand all the quantities on the right-hand sides of (2.1) and (2.2) about the node $i-\frac{1}{2}$ and using (1.1) we have

$$y_{i-3/4} = \bar{y}_{i-3/4} - Ch^4 + O(h^5),$$

and

$$y_{i-1/4} = \bar{y}_{i-1/2} - Ch^4 + O(h^5),$$

where

$$C = -\frac{3}{2048}\,y^{(4)}_{i-1/2}.$$

($y_j^{(s)}$ denotes the s-th derivative of y at x_j).

Now,

$$f\left(x_{i-1/4},\bar{y}_{i-1/4}\right) = f\left(x_{i-1/4},y_{i-1/4}+Ch^4+O(h^5)\right)$$

$$= f_{i-1/2} + \frac{h}{4} f_x + \left(y_{i-1/4}+Ch^4-y_{i-1/2}\right)f_y$$

$$+ \frac{1}{2}\left(\frac{h}{4}\right)^2 f_{xx} + \left(\frac{h}{4}\right)\left(y_{i-1/4}+Ch^4-y_{i-1/2}\right)f_{xy}$$

$$+ \frac{1}{2}\left(y_{i-1/4}+Ch^4-y_{i-1/2}\right)^2 f_{yy} + O(h^5)$$

$$= f_{i-1/4} + Ch^4\left(f_y+O(y_{i-1/4}-y_{i-1/2})\right) + O(h^5)$$

$$= f_{i-1/4} + Ch^4\left(f_y+O(h)\right) + O(h^5),$$

since $y_{i-1/4} - y_{i-1/2} = O(h)$,

or

$$\bar{f}_{i-1/4} = f_{i-1/4} + Ch^4 f_y + O(h^5). \tag{2.7}$$

Similarily,

$$\bar{f}_{i-3/4} = f_{i-3/4} + Ch^4 f_y + O(h^5). \tag{2.8}$$

Therefore from (2.7) and (2.8), we have

$$f_{i-3/2} - f_{i-1/4} = \bar{f}_{i-3/4} - \bar{f}_{i-1/2} + O(h^5), \tag{2.9}$$

and from (2.3) it follows that

$$\hat{y}_{i-1/2} = \frac{1}{2}\left(y_{i-1}+y_i\right) + h\left[\frac{1}{24}\left(f_{i-1}-f_i\right) + \frac{1}{6}\left(f_{i-3/2}-f_{i-1/2}\right)\right] + O(h^6).$$

If the quantities on the right-hand side of the above equation are expanded by Taylor's theorem about $x_{i-1/2}$, then we have

$$\hat{y}_{i-1/2} = y_{i-1/2} + O(h^6).$$

because we have to compute

$$f_{i-1},\ \bar{f}_{i-3/4},\ \bar{f}_{i-1/4},\ \hat{f}_{i-1/2},\ \hat{f}_{i-3/4} \text{ and } \hat{f}_{i-1/2}\ .$$

This is in contrast with a sixth-order Runge-Kutta method (local error $O(h^7)$) that must be used for multiple shooting and requires seven function evaluations [4]. Multiple shooting from both ends of the interval $[x_{i-1},x_i]$ would require even more function evaluations for $O(h^7)$ local truncation error.

Note that we only require that $f \in C^6$ inside the interval $[x_{i-1},x_i]$, and therefore jump discontinuities in f at the node points create no problems. The cubic spline and spline and quintic spline [2] as well as other methods, e.g. those using numerical derivatives (deferred corrections [5]) may require a very fine mesh in order that numerical derivatives of sufficient accuracy be computed in case of jump discontinuities.

3. Computational Results

The results of our computational experiments comparing the present method with various other methods are given in Table 1. We have exhibited under each method the absolute value of the maximum error between the exact solution and the computed solution for two problems (these were labelled as problems 2 and 3 in [2]). The methods used were TR (Trapezoidal rule), SR (Simpson's rule), RK5 and RK6 (Runge-Kutta methods with local errors $O(h^6)$ and $O(h^7)$ respectively [3], p. 424 and [4], p. 193), CSS (cubic spline on spline), and QS (quintic spline). Mesh sizes 10, 20, 40 and 80 were used. The CSS and QS results are from [2] and therefore n = 80 errors are not available. As we mentioned in [2], the first problem in Table 1 has large values for the higher

derivatives of f_{2i}. For example, f_{2i}^6 have the factor $100\Pi^6 \approx 3.9 \times 10^5$. It is pointed out in [5] that for this problem the usual trapezpidal rule ($O(h^3)$ local error) required 1024 mesh points for 10^{-6} accuracy, and the deferred correction method needed 65 mesh points with 7 corrections.

The present method and RK5 both require six function evaluations. It is evident from Table 1 that in all cases the present method gives significantly more accurate results than RK5. It even beats RK6 which requires seven function evaluations. As the mesh size increases, the RK5 is still three orders of magnitude worse than the present method, even RK6 is two orders worse for problem 1!

Another interesting fact that emerges from Table 1 is that for even small values of the grid size n, the present method yields usable results for problem 1 which is not the case for RK5 or even RK6.

References

[1] Tewarson, R.P., On the use of Simpson's rule in renal models, Math. Biosciences 55, 1-5 (1981).

[2] Tewarson, R.P., On the use of splines for the numerical solution of non-linear two-point boundary value problems. BIT 20, 223-232 (1980).

[3] Fehlberg, E., Eine methode zur fehlerverkleinerung beim Runge-Kutta verfahren, Z. angew. Math. Mech. 38, 421-426 (1958).

[4] Butcher, J.C., On Runge-Kutta processes of high order, J. Austral. Math. Soc. 4, 179-194 (1964).

[5] Lentini, M. & Pereyra, V., A variable order finite-difference method for nonlinear multipoint boundary value problems, Math. Comp. 28, 981-1003 (1974).

Table 1. Maximum errors in the solutions

PROBLEM	METHOD	TR	SR [1]	RK5 [3]	CSS [2]	QS [2]	RK6 [4]	Present Method
	Local error / Mesh size	$O(h^3)$	$O(h^5)$	$O(h^6)$	$O(h^6)$	$O(h^7)$	$O(h^7)$	$O(h^7)$
1	10	.276(+1)	.150(+0)	.140(+1)	.42(-2)	.28(-2)	.158(+1)	.399(-2)
	20	.703(+0)	.109(-1)	.248(-1)	.40(-3)	.20(-3)	.127(-1)	.757(-4)
	40	.161(+0)	.650(-3)	.509(-3)	.18(-4)	.56(-5)	.122(-3)	.115(-5)
	80	.393(-1)	.402(-4)	.129(-4)	-	-	.149(-5)	.217(-7)
2	10	.336(-3)	.552(-7)	.642(-8)	.11(-8)	.70(-9)	.166(-9)	.247(-10)
	20	.839(-4)	.345(-8)	.197(-9)	.25(-10)	.12(-10)	.295(-11)	.387(-12)
	40	.210(-4)	.216(-9)	.615(-11)	.48(-12)	.20(-12)	.408(-13)	.616(-14)
	80	.524(-5)	.135(-10)	.192(-12)	-	-	.863(-15)	.258(-15)