The numerical solution of implicit first order ordinary differential equations with initial conditions

M. A. Wolfe

Department of Applied Mathematics, University of St. Andrews

Predictor-corrector methods are used in conjunction with iteration to obtain numerical solutions of a class of first order ordinary differential equations with an initial condition in which the derivative cannot be expressed explicitly as a function of the independent and dependent variables.

(Received February 1970)

1. Introduction

Many methods are available for the numerical solution of first order ordinary differential equations with an initial condition of the form

$$y^{(1)}(x) = f(x, y(x)); y(x_0) = y_0; x_0, x \in [a,b] \ . \quad (1.1)$$

In order to apply these methods it is assumed that, as in (1.1) it is possible to express the first derivative $y^{(1)}$ of the unknown function y explicitly as a function of x and y. If this is not possible then a problem of the form

$$y^{(1)}(x) = f(x, y(x), y^{(1)}(x)); y(x_0) = y_0; x_0. x \in [a,b] \quad (1.2)$$

is obtained to which the methods previously mentioned are not immediately applicable. The equation (1.2) will be called an implicit equation. Implicit initial value problems of the form

$$f(x, y(x), y^{(1)}(x)) = 0; y(x_0) = y_0; x, x_0 \in [a,b] \quad (1.3)$$

have been considered by Altman (1960) and by Verner (1969).

For existence theorems corresponding to (1.1), (1.2) and (1.3) the reader is referred to Murray and Miller (1954).

It is the purpose of this paper to explain how well-known predictor-corrector methods for the solution of (1.1) may be applied to (1.2).

2. The general method

Suppose it is required to obtain the numerical solution of (1.2) on $[a, b]$, assuming that such a solution exists. Let $x_0 = a$ and let y_n and $y_n^{(1)}$ denote the numerical estimates of the values $y(x_n)$ and $y^{(1)}(x_n)$ of y and $y^{(1)}$ respectively at $x = x_n$, where

$$x_n = x_0 + nh \ (n = 0, 1, 2, \ldots, N) \ , \quad (2.1)$$

and $x_N = b$.

Choose $\{\alpha_k; k = 0, \ldots, 6\}$ so that

$$y(x_{n+1}) = \sum_{k=0}^{2} \alpha_k y(x_{n-k}) + h \sum_{k=3}^{6} \alpha_k y^{(1)}(x_{n-k+3}) + A_n h^p \ , \quad (2.2)$$

where p is a positive integer and A_n is a bounded function of n; this is possible provided that a solution of the differential equation exists having a derivative of order p which is continuous on $[a, b]$.

Choose $\{\beta_k; k = 0, \ldots, 5\}$ so that

$$y^{(1)}(x_{n+1}) = \sum_{k=0}^{5} \beta_k y^{(1)}(x_{n-k}) + B_n h^{p-1} \quad (2.3)$$

where B_n is a bounded function of n.

Finally, choose $\{\gamma_k; k = 0, \ldots, 6\}$ so that

$$y(x_{n+1}) = \sum_{k=0}^{2} \gamma_k y(x_{n-k}) + h \sum_{k=3}^{6} \gamma_k y^{(1)}(x_{n-k+4}) + C_n h^p \ , \quad (2.4)$$

where $|C_n| < |A_n|$ for all values of n.

Clearly $p \leqslant 7$ but in most practical applications of (2.2), (2.3) and (2.4) as a predictor-corrector system $p \leqslant 5$ and the values of the coefficients are chosen to ensure stability as well as the required accuracy.

The formulae (2.2), (2.3), and (2.4) suggest the following algorithm for the numerical solution of (1.2) with errors $0(h^{p-1})$:

Algorithm 1:

Given y_j and $y_j^{(1)}$ $(j = 0, \ldots, s)$ with errors $0(h^{p-1})$, where $2(s + 1)$ is the number of starting values required, compute $y_{n+1}^{(1)}$ and y_{n+1} for $n = s, \ldots, N-1$ as follows:

Compute
$$u_{1,0} = \sum_{k=0}^{2} \alpha_k y_{n-k} + h \sum_{k=3}^{6} \alpha_k y_{n-k+3}^{(1)} \quad (2.5)$$

and
$$u_{2,0} = \sum_{k=0}^{5} \beta_k y_{n-k}^{(1)} \ . \quad (2.6)$$

Then for $i = 0, 1, 2, \ldots$

compute
$$u_{1,i+1} = h\gamma_3 f(x_{n+1}, u_{1,i}, u_{2,i}) + \delta_n \ , \quad (2.7)$$

and
$$u_{2,i+1} = f(x_{n+1}, u_{1,i}, u_{2,i}) \ , \quad (2.8)$$

where
$$\delta_n = \sum_{k=0}^{2} \gamma_k y_{n-k} + h \sum_{k=4}^{6} \gamma_k y_{n-k+4}^{(1)} \ . \quad (2.9)$$

Using (2.7) and (2.8) iterate until

$$\text{Max } \{E_1, E_2\} \leqslant \varepsilon h^p \quad (2.10)$$

where
$$E_1 = |u_{1,i} - h\gamma_3 f(x_{n+1}, u_{1,i}, u_{2,i}) - \delta_n| \ , \quad (2.11)$$

and
$$E_2 = |u_{2,i} - f(x_{n+1}, u_{1,i}, u_{2,i})| \ . \quad (2.12)$$

In (2.10) ε is a given parameter.

If the value of i for which (2.10) is satisfied is I, then take

$$y_{n+1} = u_{1,I} \ ; \ y_{n+1}^{(1)} = u_{2,I} \ . \quad (2.13)$$

In this algorithm the number of starting values required depends upon the values of the coefficients in (2.2), (2.3) and (2.4). It remains to ascertain under what conditions Algorithm 1 will provide a numerical solution of (1.2) with error $0(h^{p-1})$.

3. Convergence of the method

The following notation to denote partial differentiation will be used:

$$f_{i,j}^{(m,n)}(v_1,v_2,v_3) = \frac{\partial^{m+n} f}{\partial v_i^m \partial v_j^n} f(v_1,v_2,v_3) \ . \quad (3.1)$$

The following result is easily established:

Lemma:

If (i) $f_i^{(1)}(x, y, z)$ $(i = 2, 3)$ exist and are continuous in a closed interval $S \subset R^3$ where R denotes the set of real numbers;

(ii) the equations

$$u_1 = h\gamma_3 f(x,u_1,u_2)+\delta$$
$$u_2 = f(x,u_1,u_2)$$

have a solution (u_1^*, u_2^*) such that $(x, u_1^*, u_2^*) \in S$, where h and γ_3 have been defined and δ is a constant;

(iii) $f_3^{(1)}(x,u_1,u_2)+h\gamma_3 f_2^{(1)}(x,u_1,u_2) \neq 1$,
$f_3^{(1)}(x, u_1, u_2) \neq 1$,
$h\gamma_3 f_2^{(1)}(x, u_1, u_2) \neq 1$
$\forall (x, u_1, u_2) \in S$;

(iv) for u_1' and u_2' given, and $\varepsilon > 0$ given

$$|u_1' - h\gamma_3 f(x, u_1', u_2') - \delta| \leqslant \varepsilon ,$$

and

$$|u_2' - f(x, u_1', u_2')| \leqslant \varepsilon ,$$

then there exist finite numbers K_1 and K_2 such that

$$|u_i' - u_i^*| \leqslant K_i \varepsilon \ (i = 1,2) .$$

This lemma will be used in establishing the validity of the general method.

Let y_r^* and $y_r^{(1)*}$ be the exact solutions of the system of equations

$$u_1 = h\gamma_3 f(x_r, u_1, u_2)+\delta_{r-1}^* , \quad (3.2)$$
$$u_2 = f(x_r, u_1, u_2)$$

where

$$\delta_r^* = \sum_{k=0}^{2} \gamma_k y_{r-k}^* + h \sum_{k=4}^{6} \gamma_k y_{r-k+4}^{(1)*} \quad (3.3)$$

Let y_r^{**} and $y_r^{(1)**}$ be the exact solutions of the system

$$u_1 = h\gamma_3 f(x_r, u_1, u_2)+\delta_{r-1} , \quad (3.4)$$
$$u_2 = f(x_r, u_1, u_2)$$

where δ_r is defined by (2.9) with n replaced by r.

Define e_r^*, $e_r^{(1)*} e_r^{**}$, $e_r^{(1)**}$, by

$$e_r^* = y_r^* - y(x_r) , \quad (3.5)$$
$$e_r^{(1)*} = y_r^{(1)*} - y^{(1)}(x_r) , \quad (3.6)$$
$$e_r^{**} = y_r^* - y_r^{**} , \quad (3.7)$$
$$e_r^{(1)**} = y_r^{(1)*} - y_r^{(1)**} . \quad (3.8)$$

Define the total cumulative errors e_r and $e_r^{(1)}$ by

$$e_r = y_r - y(x_r) , \quad (3.9)$$
$$e_r^{(1)} = y_r^{(1)} - y^{(1)}(x_r) . \quad (3.10)$$

Then

$$|e_r| \leqslant |y_r - y_r^{**}| + |e_r^{**}| + |e_r^*| , \quad (3.11)$$
$$|e_r^{(1)}| \leqslant |y_r^{(1)} - y_r^{(1)**}| + |e_r^{(1)**}| + |e_r^{(1)*}| . \quad (3.12)$$

Using the lemma and the convergence criterion (2.10)

$$|y_r - y_r^{**}| \leqslant K_1 \varepsilon h^p , \quad (3.13)$$
$$|y_r^{(1)} - y_r^{(1)**}| \leqslant K_2 \varepsilon h^p , \quad (3.14)$$

where $r = s + 1, s + 2, \ldots$.

Putting $r = n + 1$ and using (3.2), (3.4) and the mean value theorem,

$$e_{n+1}^{**} = h\gamma_3 \{ f_2^{(1)}(x_{n+1}, \xi_{n+1}, \eta_{n+1}) e_{n+1}^{**}$$
$$+ f_3^{(1)}(x_{n+1}, \xi_{n+1}, \eta_{n+1}) e_{n+1}^{(1)**} \} + (\delta_n^* - \delta_n^{**}) + (\delta_n^{**} - \delta_n) \quad (3.15)$$

where $n = s, s + 1, \ldots$, and

$$\delta_n^{**} = \sum_{k=0}^{2} \gamma_k y_{n-k}^{**} + h \sum_{k=4}^{6} \gamma_k y_{n-k+4}^{(1)**} . \quad (3.16)$$

Also

$$e_{n+1}^{(1)**} = f_2^{(1)}(x_{n+1}, \xi_{n+1}, \eta_{n+1}) e_{n+1}^{**}$$
$$+ f_3^{(1)}(x_{n+1}, \xi_{n+1}, \eta_{n+1}) e_{n+1}^{(1)**} , \quad (3.17)$$

where $n = s, s + 1, \ldots$.

In (3.15) and (3.17) ξ_{n+1} and η_{n+1} lie between y_{n+1}^*, y_{n+1}^{**} and $y_{n+1}^{(1)*}$, $y_{n+1}^{(1)**}$ respectively.

With M defined by

$$M = \operatorname*{Max}_{(xyz) \in S} \left\{ \left| \frac{f_2^{(1)}}{1 - f_3^{(1)}} \right| \right\} \quad (3.18)$$

where $f_2^{(1)}$ and $f_3^{(1)}$ are evaluated at (x,y,z), and provided

$$h|\gamma_3| M < 1 , \quad (3.19)$$

$$0 < (1 - h|\gamma_3|M)|e_{n+1}^{**}| \leqslant \sum_{k=0}^{2} \{|\gamma_k| + hM|\gamma_{k+4}|\} |e_{n-k}^{**}|$$
$$+ K_1 \varepsilon h^p \sum_{k=0}^{2} |\gamma_k| + K_2 \varepsilon h^{p+1} \sum_{k=4}^{6} |\gamma_k| , \quad (3.20)$$

where $n = s + 3, \ldots$.

Consider now a class of formulae (2.4) in which

$$\sum_{k=0}^{2} |\gamma_k| = 1 . \quad (3.21)$$

This permits the commonly used predictor-corrector formulae, in particular those of Adams and Moulton to be included in Algorithm 1.

Define $v_{n+1} (n = s + 3, \ldots)$ by

$$(1 - h|\gamma_3|M) v_{n+1} = \sum_{k=0}^{2} \{|\gamma_k| + hM|\gamma_{k+4}|\} v_{n-k}$$
$$+ \varepsilon h^p (K_1 + hK_2 \sum_{k=4}^{6} |\gamma_k|) \quad (3.22)$$

with

$$v_r \geqslant \operatorname*{Max}_{s+1 \leqslant i \leqslant s+3} \{|e_i^{**}|\} = \delta'' , \quad (3.23)$$

for $r = s + 1, s + 2, s + 3$.

Then it is easily shown by induction that

$$v_r \geqslant |e_r^{**}| \quad (r = s+1, s+2, \ldots) . \quad (3.24)$$

The characteristic equation corresponding to (3.22) is

$$P(z) = (1 - h|\gamma_3|M) z^3 - \sum_{k=0}^{2} \{|\gamma_k| + hM|\gamma_{k+4}|\} z^{2-k} = 0 \quad (3.25)$$

which has a root z^* given by

$$z^* = 1 + \mu h \ (\mu > 0) , \quad (3.26)$$

when (3.19) and (3.21) are satisfied.

A solution of (3.22) which satisfies (3.23) is

$$v_n = \delta'' z^{*n-s-1} + |e^{**}| . (z^{*n-s-1} - 1) , \quad (3.27)$$

where $n = s + 1, s + 2, \ldots$, and

$$e^{**} = \frac{-\varepsilon h^{p-1} (K_1 + K_2 h \sum_{k=4}^{6} |\gamma_k|)}{M \sum_{k=3}^{6} |\gamma_k|} . \quad (3.28)$$

Hence

$$|e_n^{**}| \leqslant \delta'' \exp (x_{n-s-1} - x_0)\mu + |e^{**}| [\exp (x_{n-s-1} - x_0)\mu - 1] \quad (3.29)$$

where $n = s + 1, s + 2, \ldots$.

So provided δ'' is $0(h^{p-1})$, e_n^{**} is $0(h^{p-1})$ and will tend to zero as $n \to \infty$ and $h \to 0$ in such a way that nh remains constant.

Consider now e_r^*. An argument similar to the preceding one shows that provided (3.19) is satisfied,

$$0 < (1 - h|\gamma_3|M)|e_{n+1}^*| \leqslant \sum_{k=0}^{2} |\gamma_k| . |e_{n-k}^*|$$
$$+ h \sum_{k=4}^{6} |\gamma_k| . M|e_{n-k+4}^*| + Ch^p \quad (3.30)$$

where $n = s + 3, \ldots$, and $|C_n| \leqslant C \ (\forall n)$.

Define w_{n+1} $(n = s+3, \ldots)$ by

$$(1-h|\gamma_3|M)w_{n+1} = \sum_{k=0}^{2}\{|\gamma_k|+hM|\gamma_{k+4}|\}w_{n-k}+Ch^p \qquad (3.31)$$

with $$w_r \geqslant \operatorname*{Max}_{s+1\leqslant i\leqslant s+3}\{|e_i^*|\} = \delta' \ . \qquad (3.32)$$

The characteristic equation corresponding to (3.31) is just (3.25).

A solution of (3.31) which satisfies (3.32) is

$$w_n = \delta' z^{*n-s-1}+|e^*| \ . \ (z^{*n-s-1}-1) \ , \qquad (3.33)$$

where $n = s+1, \ldots,$ and

$$e^* = -Ch^{p-1}/M\sum_{k=3}^{6}|\gamma_k| \ . \qquad (3.34)$$

Hence

$$|e_n^*| \leqslant \delta' \exp(x_{n-s-1}-x_0)\mu+|e^*|\,[\exp(x_{n-s-1}-x_0)\mu-1] \qquad (3.35)$$

where $n = s+1, \ldots$.

So provided δ' is $0(h^{p-1})$, e_n^* is $0(h^{p-1})$ and will tend to zero as $n \to \infty$ and $h \to 0$ in such a way that nh remains constant.

From (3.11), (3.13), (3.29) and (3.35) it is established that Algorithm 1 is valid provided that the conditions of the lemma are satisfied, (3.19) is satisfied, δ' and δ'' are both $0(h^{p-1})$ and the iterative procedure used to solve the system (2.7) and (2.8) is convergent. That δ' and δ'' are at least $0(h^{p-1})$ can be established easily, albeit tediously, by constructing expressions for e_r^* and $e_r^{**}(r = s+1, s+2, s+3)$ in terms of the starting values. The convergence of the iterative procedure is examined in Section 4.

4. Convergence of the iterative procedure

Let $\mathbf{u}$ be the column vector defined by

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \qquad (4.1)$$

and let $\mathbf{F}(\mathbf{u})$ be the vector-valued function defined by

$$\mathbf{F}(\mathbf{u}) = \begin{pmatrix} h\gamma_3 f(x_{n+1}, u_1, u_2)+\delta_n \\ f(x_{n+1}, u_1, u_2) \end{pmatrix} \equiv \begin{pmatrix} f_1(\mathbf{u}) \\ f_2(\mathbf{u}) \end{pmatrix} . \qquad (4.2)$$

Then if the system

$$\mathbf{u} = \mathbf{F}(\mathbf{u}) \qquad (4.3)$$

has a solution $\mathbf{u}^*$, and in the ball $U(\mathbf{u}^*, \rho) = \{\mathbf{u}; \|\mathbf{u}^*-\mathbf{u}\| < \rho\}$ where

$$\|\mathbf{u}^*-\mathbf{u}\| = \operatorname*{Max}_{1\leqslant j\leqslant 2}\{|u_j^*-u_j|\} \ , \qquad (4.4)$$

and

$$\|\mathbf{F}(\mathbf{u})-\mathbf{F}(\mathbf{u}^*)\| \leqslant \lambda\|\mathbf{u}-\mathbf{u}^*\| \qquad (4.5)$$

with $0 < \lambda < 1$, then for all $\mathbf{u}_0$ in $U(\mathbf{u}^*, \rho)$ where $\mathbf{u}_0$ is the initial iterate in a sequence $\{\mathbf{u}_i\}$ generated from

$$\mathbf{u}_{i+1} = \mathbf{F}(\mathbf{u}_i) \qquad (4.6)$$

(i) all the iterates $\mathbf{u}_i$ lie in $U(\mathbf{u}^*, \rho)$;
(ii) $\mathbf{u}_i \to \mathbf{u}^*$ $(i \to \infty)$;
(iii) the solution $\mathbf{u}^*$ is unique in $U(\mathbf{u}^*, \rho)$.

From this it follows that if f_1 and f_2 have continuous first partial derivatives such that

$$\left|\frac{\partial f_j}{\partial u_k}(\mathbf{u})\right| \leqslant \frac{\lambda}{2} \ (j,k = 1,2) \ (\mathbf{u}\in U(\mathbf{u}^*,\rho)) \qquad (4.7)$$

then for any $\mathbf{u}_0 \in U(\mathbf{u}^*, \rho)$

(iv) all the iterates $\mathbf{u}_i$ lie in $U(\mathbf{u}^*, \rho)$;
(v) $\mathbf{u}_i \to \mathbf{u}^*$ $(i \to \infty)$.

For a proof of these results see, for example Isaacson and Keller (1966).

From (4.2) this means that sufficient conditions for the convergence of the iterative procedure of Algorithm 1 are

$$|f_2^{(1)}(x_{n+1}, u_1, u_2)| < \tfrac{1}{2}; \ |f_3^{(1)}(x_{n+1}, u_1, u_2)| < \tfrac{1}{2} \qquad (4.8)$$

for all $\mathbf{u}$ in a ball $U(\mathbf{u}^*, \rho)$ containing $\mathbf{u}_0 = (u_{1,0}, u_{2,0})$.

Since conditions (4.8) may not be satisfied by f it seems desirable to obtain other iterative procedures which will converge for a wider class of function f. One possibility is to modify the iterative procedure (4.6) and another is to use Newton-Raphson iteration. These possibilities will now be discussed.

It can be shown as in Isaacson and Keller (1966) that if f_1 and f_2 have continuous first partial derivatives which satisfy

$$\left|1-\frac{\partial f_1}{\partial u_1}(\mathbf{u})\right| > \left|\frac{\partial f_1}{\partial u_2}(\mathbf{u})\right| \ , \qquad (4.9)$$

and

$$\left|1-\frac{\partial f_2}{\partial u_2}(\mathbf{u})\right| > \left|\frac{\partial f_2}{\partial u_1}(\mathbf{u})\right| \qquad (4.10)$$

in a ball $\bar{U}(\mathbf{u}^*, \rho) = \{\mathbf{u}; \|\mathbf{u}-\mathbf{u}^*\| \leqslant \rho\}$ then the sequence $\{\mathbf{u}_i\}$ generated from

$$\mathbf{u}_0 \in \bar{U}(\mathbf{u}^*,\rho)$$
$$\mathbf{u}_{i+1} = \Theta_i\mathbf{F}(\mathbf{u}_i)+(\mathrm{I}-\Theta_i)\mathbf{u}_i \ (i = 0, 1, 2, \ldots) \qquad (4.11)$$
$$\Theta_i = (\theta_{ij}\delta_{jk}) \quad (j,k = 1, 2)$$

Θ_i being a nonsingular diagonal matrix with

$$\theta_{ij} = 1/\{1-\frac{\partial f_j}{\partial u_j}(\mathbf{u}_i)\} \ (j = 1, 2) \ (i = 0, 1, 2, \ldots) \qquad (4.12)$$

will converge to the solution $\mathbf{u}^*$ of (4.3).

Applying (4.9) and (4.10) to (4.2) the conditions to be satisfied for the modified iterative procedure (4.11) to be applicable to (4.2) are

$$|1-h\gamma_3 f_2^{(1)}(x_{n+1}, u_1, u_2)| > |h\gamma_3 f_3^{(1)}(x_{n+1}, u_1, u_2)| \ , \qquad (4.13)$$

and

$$|1-f_3^{(1)}(x_{n+1}, u_1, u_2)| > |f_2^{(1)}(x_{n+1}, u_1, u_2)| \qquad (4.14)$$

for all $\mathbf{u}$ in $\bar{U}(\mathbf{u}^*, \rho)$.

Clearly (4.13) is easily satisfied by taking h sufficiently small which leaves (4.14) as the decisive criterion. Referring to (3.18), (4.14) is equivalent to the requirement

$$M < 1 \ . \qquad (4.15)$$

The following algorithm is then obtained for the numerical solution of (1.2):

Algorithm 2:
Given y_j and $y_j^{(1)}$ $(j = 0, \ldots, s)$ with errors $0(h^{p-1})$ compute y_{n+1} and $y_n^{(1)}$ for $n = s, \ldots, N-1$ as follows:

Compute $u_{1,0}$ and $u_{2,0}$ from (2.5) and (2.6) respectively. Then for $i = 0, 1, 2, \ldots$ compute θ_{1i}, θ_{2i}, $u_{1,i+1}$, $u_{2,i+1}$ from

$$\theta_{1,i} = 1/\{1-h\gamma_3 f_2^{(1)}(x_{n+1}, u_{1,i}, u_{2,i})\}$$
$$\theta_{2,i} = 1/\{1-f_3^{(1)}(x_{n+1}, u_{1,i}, u_{2,i})\}$$
$$u_{1,i+1} = \theta_{1,i}\{h\gamma_3 f(x_{n+1}, u_{1,i}, u_{2,i})+\delta_n\}+(1-\theta_{1i})u_{1,i}$$
$$u_{2,i+1} = \theta_{2,i}f(x_{n+1}, u_{1,i}, u_{2,i})+(1-\theta_{2,i})u_{2,i}$$

where δ_n is defined by (2.9).

Iterate until criterion (2.10) is satisfied and then obtain y_{n+1} and $y_{n+1}^{(1)}$ from (2.13).

If it is difficult to calculate or evaluate $f_2^{(1)}$ and $f_3^{(1)}$ they could be approximated by a simple difference replacement such as

$$\left.\begin{aligned} f_2^{(1)}(x,y,z) &\simeq \{f(x,y+\varepsilon_1,z)-f(x,y,z)\}/\varepsilon_1 \\ f_3^{(1)}(x,y,z) &\simeq \{f(x,y,z+\varepsilon_2)-f(x,y,z)\}/\varepsilon_2 \end{aligned}\right\} \qquad (4.16)$$

where ε_1 and ε_2 are conveniently chosen parameters.

Since the iterative procedure described in Algorithm 2 requires the evaluation of the partial derivatives $f_2^{(1)}$ and $f_3^{(1)}$ one could with little additional effort use Newton-Raphson iteration for the solution of (4.3) provided that the convergence conditions are not too stringent.

Consider the system

$$\mathbf{G}(\mathbf{u}) = \begin{pmatrix} g_1(\mathbf{u}) \\ g_2(\mathbf{u}) \end{pmatrix} = \mathbf{0}. \tag{4.17}$$

Then it can be shown that if:

(i) g_1 and g_2 are defined and twice continuously differentiable on a subset T of R^2;

(ii) $\mathbf{G}(\mathbf{u}) = \mathbf{0}$ has a solution $\mathbf{u}^* \in T$;

(iii) Det $(H(\mathbf{u}^*)) \neq 0$,

where the matrix $H(\mathbf{u})$ has elements $\dfrac{\partial g_i}{\partial u_j}(\mathbf{u})$ $(i, j = 1, 2)$ $(\mathbf{u} \in T)$, then there is a number $\rho > 0$ such that the sequence $\{\mathbf{u}_i\}$ generated by Newton iteration starting from an initial iterate $\mathbf{u}_0$ converges quadratically to $\mathbf{u}^*$ for all $\mathbf{u}_0 \in \bar{U}(\mathbf{u}^*, \rho) \subset T$. For discussion of this result see Henrici (1964).

With

$$\left.\begin{aligned} g_1(\mathbf{u}) &= u_1 - h\gamma_3 f(x_{n+1}, u_1, u_2) - \delta_n \\ g_2\ \mathbf{u}) &= u_2 - f(x_{n+1}, u_1, u_2) \end{aligned}\right\} \tag{4.18}$$

Newton iteration will converge provided that:

(i) $'f_{ij}^{(11)}(x_{n+1}, u_1, u_2)$ is continuous in T $(i, j = 2, 3)$;

(ii) $'\ \mathbf{u}^* \varepsilon T$;

(iii) $'f_3^{(1)}(x_{n+1}, u_1^*, u_2^*) + h\gamma_3 f_2^{(1)}(x_{n+1}, u_1^*, u_2^*) \neq 1.$ (4.19)

Hence it would appear that Newton iteration will converge for a wide class of function f. This gives rise to the following algorithm:

Algorithm 3:

Given y_j and $y_j^{(1)}$ $(j = 0, \ldots, s)$ with errors $0(h^{p-1})$ compute y_{n+1} and $y_{n+1}^{(1)}$ $(n = s, \ldots, N-1)$ as follows:

Compute $u_{1,0}$, $u_{2,0}$ and δ_n from (2.5), (2.6) and (2.9) respectively. Then for $i = 0, 1, 2, \ldots$ compute $u_{1,i+1}$ and $u_{2,i+1}$ from

$$u_{1,i+1} = u_{1,i} + \Delta_{1,i}$$
$$u_{2,i+1} = u_{2,i} + \Delta_{2,i}$$

where

$$\Delta_{1,i} = \frac{\{(\delta_n - u_{1,i})(1 - f_3^{(1)}(i)) + h\gamma_3(f(i) - u_{2,i} f_3^{(1)}(i))\}}{H}$$

$$\Delta_{2,i} = \frac{\{f(i) - (u_{1,i} - \delta_n) f_2^{(1)}(i) - u_{2,i}(1 - \gamma_3 h f_2^{(1)}(i))\}}{H}$$

in which $f_j^{(1)}(i) = f_j^{(1)}(x_{n+1}, u_{1,i}, u_{2,i})$ $(j = 2, 3)$,

$$f(i) = f(x_{n+1}, u_{1,i}, u_{2,i}),$$

and $$H = 1 - f_3^{(1)}(i) - h\gamma_3 f_2^{(1)}(i).$$

Iterate until criterion (2.10) is satisfied and obtain y_{n+1} and $y_{n+1}^{(1)}$ from (2.13).

The analysis of Section 3 will remain valid for Algorithm 3, as indeed it will for Algorithm 2.

5. Second order method

In this section a second order method for the numerical solution of (1.2) is given in which the values of α and γ correspond to the second order Adams-Moulton method and the values of β have been chosen to give a sufficiently accurate initial estimate of $y_{n+1}^{(1)}$. For a discussion of the Adams-Moulton formulae see Henrici (1962). In this section a starting procedure is also given for the second order method.

In (2.2) take $\alpha_0 = 1$, $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = 3/2$, $\alpha_4 = -1/2$, $\alpha_5 = \alpha_6 = 0$, so that $p = 3$; in (2.3) take $\beta_0 = 2$, $\beta_1 = -1$, $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$; in (2.4) take $\gamma_0 = 1$, $\gamma_1 = \gamma_2 = 0$, $\gamma_3 = \gamma_4 = 1/2$, $\gamma_5 = \gamma_6 = 0$. With these values of the coefficients, all three algorithms may be used to give second order accuracy provided that a suitable starting method is available and that the function f has the appropriate properties for convergence of the iterative procedure.

To provide starting values for the second order procedure, denote the numerical estimates of $y(x_n)$ and $y^{(1)}(x_n)$ with errors $0(h^p)$ by $y_{p,n}$ and $y_{p,n}^{(1)}$ respectively. The if f is such that the iterative procedure of Algorithm 2 is likely to converge, compute $y_{2,0}^{(1)}$ from

$$\theta_i = 1/\{1 - f_3^{(1)}(x_0, y_0, u_i)\} \tag{5.1}$$

and $$u_{i+1} = \theta_i f(x_0, y_0, u_i) + (1 - \theta_i) u_i \ (i = 0, 1, 2, \ldots)\ . \tag{5.2}$$

Iterate until $$|u_i - f(x_0, y_0, u_i)| \leqslant \varepsilon h^2\ . \tag{5.3}$$

If criterion (5.3) is satisfied when $i = I$ take

$$y_{2,0}^{(1)} = u_I\ . \tag{5.4}$$

In this iteration x_0 and y_0 are known and u_0 can either be found directly from the equation in fortuitous cases* or estimated graphically or by the method of bisection, the latter being more suitable for machine use.

Next compute $y_{2,1}$ from

$$y_{2,1} = y_0 + h y_{2,0}^{(1)}\ . \tag{5.5}$$

Finally, compute $y_{2,1}^{(1)}$ from

$$\theta_i = 1/\{1 - f_3^{(1)}(x_1, y_{2,1}, u_i)\} \tag{5.6}$$

$$u_{i+1} = \theta_i f x_1, y_{2,1}, u_i) + (1 - \theta_i) u_i \ (i = 0, 1, 2, \ldots) \tag{5.7}$$

with u_0 given by

$$u_0 = (y_{2,1} - y_0)/h\ , \tag{5.8}$$

iterating until (5.3) is satisfied when, say, $i = J$ and taking

$$y_{2,1}^{(1)} = u_J\ . \tag{5.9}$$

Alternatively it might be possible to use Newton-Raphson iteration in the same way to obtain $y_{2,0}^{(1)}$, $y_{2,1}$, and $y_{2,1}^{(1)}$.

The validity of the starting procedures given in this section is established using the following lemma which is easily proved:

Lemma:

If

(i) the function f of a single variable x has a first derivative which exists and is continuous on a closed interval I;

(ii) the equation $x = f(x)$ has a root $x^* \in I$;

(iii) $f^{(1)}(x) \neq 1$ $(\forall x \in I)$;

(iv) $|x' - f(x')| \leqslant \varepsilon$ where $x' \in I$ is given and $\varepsilon > 0$ is given;

then $|x' - x^*| \leqslant L\varepsilon$ where

$$L = \operatorname*{Max}_{x \in I} \left\{ \frac{1}{|1 - f^{(1)}(x)|} \right\}.$$

This lemma ensures that (5.4) gives $y^{(1)}$ with error $0(h^2)$. Indeed

$$|y_{2,0}^{(1)} - y^{(1)}(x_0)| \leqslant L_1 \varepsilon h^2\ , \tag{5.10}$$

where $$L_1 = \operatorname*{Max}_{(xyz) \in S} \left\{ \frac{1}{|1 - f_3^{(1)}(x,y,z)|} \right\}. \tag{5.11}$$

Using (5.5), (5.10), and

$$y(x_1) = y_0 + h y^{(1)}(x_0) + \frac{h^2}{2} y^{(2)}(\xi_0) \tag{5.12}$$

*In which case this iteration is unnecessary.

Table 1 Solution of (7.1) using Algorithm 2

x	SECOND ORDER METHOD ($e_r \times 10^{15}$) $h = 0{\cdot}1$	$h = 0{\cdot}05$	FOURTH ORDER METHOD ($e_r \times 10^{15}$) $h = 0{\cdot}1$	$h = 0{\cdot}05$
0·0	0·0	0·0	0·0	0·0
0·2	−0·028	−0·069	−0·014	−0·042
0·4	−0·069	−0·17	−0·014	−0·19
0·6	−0·097	−0·26	−0·083	−0·32
0·8	−0·14	−0·35	−0·18	−0·46
1·0	−0·12	−0·42	−0·35	−0·64
1·2	−1·1	−2·0	−1·3	−2·7
1·4	−2·2	−3·6	−2·4	−5·1
1·6	−2·9	−4·9	−3·8	−7·5
1·8	−4·0	−6·2	−4·9	−10·0
2·0	−4·7	−7·5	−6·0	−12·0

where $x_0 < \xi_0 < x_1$,

$$|y_{2,1} - y(x_1)| \leqslant L_1 \varepsilon h^3 + \frac{h^2}{2} |y^{(2)}(\xi_0)| \ . \tag{5.13}$$

If $y^{(2)}(x)$ exists and is continuous on $[a, b]$, so that

$$|y^{(2)}(x)| \leqslant L_2 \tag{5.14}$$

for some L_2, (5.13) implies that (5.5) gives $y(x_1)$ with error $0(h^2)$.

Finally,

$$y_{2,1}^{(1)} - y^{(1)}(x_1) = y_{2,1}^{(1)} - f(x_1, y_{2,1}, y_{2,1}^{(1)}) + f(x_1, y_{2,1}, y_{2,1}^{(1)}) - f(x_1, y(x_1), y^{(1)}(x_1))$$

implies that

$$y_{2,1}^{(1)} - y^{(1)}(x_1) = \frac{[y_{2,1}^{(1)} - f(x_1, y_{2,1}, y_{2,1}^{(1)})]}{[1 - f_3^{(1)}(x_1, \xi_1, \eta_1)]} + \frac{f_2^{(1)}(x_1, \xi_1, \eta_1)}{[1 - f_3^{(1)}(x_1, \xi_1, \eta_1)]} (y_{2,1} - y(x_1)) \tag{5.15}$$

where ξ_1 and η_1 lie between $y_{2,1}$, $y(x_1)$ and $y_{2,1}^{(1)}$, $y^{(1)}(x_1)$ respectively.

Then (5.15) implies that

$$|y_{2,1}^{(1)} - y^{(1)}(x_1)| \leqslant L_1 \varepsilon h^2 + M(L_1 \varepsilon h^3 + h^2 L_2/2)$$

where M is defined by (3.18).

Hence (5.9) gives $y^{(1)}(x_1)$ with error $0(h^2)$.

6. Fourth order method

To obtain a fourth order method the values of α and γ corresponding to the fourth order Adams-Moulton method may be used. In (2.2) take $\alpha_0 = 1$, $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = 55/24$, $\alpha_4 = -59/24$, $\alpha_5 = 37/24$, $\alpha_6 = -9/24$ so that $p = 5$; in (2.3) take $\beta_0 = 4$, $\beta_1 = -6$, $\beta_2 = 4$, $\beta_3 = -1$, $\beta_4 = \beta_5 = 0$; in (2.4) take $\gamma_0 = 1$, $\gamma_1 = \gamma_2 = 0$, $\gamma_3 = 9/24$, $\gamma_4 = 19/24$, $\gamma_5 = -5/24$, $\gamma_6 = 1/24$.

A fourth order starting procedure has been constructed but is extremely complicated and it would seem preferable to start the fourth order method using the second order method with a shorter step length.

Table 2 Solution of (7.3) using Algorithm 2

x	SECOND ORDER METHOD ($e_r \times 10^2$) $h = 0{\cdot}1$	$h = 0{\cdot}05$	FOURTH ORDER METHOD ($e_r \times 10^5$) $h = 0{\cdot}1$	$h = 0{\cdot}05$
0·0	0·0	0·0	0·0	0·0
0·2	−0·47	−0·12	0·0	0·0040
0·4	−0·45	−0·091	0·51	0·031
0·6	−0·44	−0·085	0·51	0·059
0·8	−0·46	−0·079	0·39	0·062
1·0	−0·41	−0·078	−0·073	0·061

Table 3 Solution of (7.1) using Algorithm 3

x	SECOND ORDER METHOD ($e_r \times 10^{15}$) $h = 0{\cdot}1$	$h = 0{\cdot}05$	FOURTH ORDER METHOD ($e_r \times 10^{15}$) $h = 0{\cdot}1$	$h = 0{\cdot}05$
0·0	0·0	0·0	0·0	0·0
0·2	−0·014	−0·028	−0·014	−0·04
0·4	−0·042	−0·069	−0·069	0·0
0·6	−0·069	−0·11	0·069	0·069
0·8	−0·097	−0·15	0·21	0·17
1·0	−0·12	−0·19	0·31	0·24
1·2	−0·67	−0·89	0·44	0·0
1·4	−1·3	−1·5	0·0	−0·67
1·6	−2·2	−2·2	−0·67	−1·3
1·8	−2·9	−2·9	−1·1	−2·0
2·0	−3·5	−3·5	−1·6	−2·7

7. Numerical results

As an illustration of the use of the algorithms presented in Sections 5 and 6 two equations are solved.

Firstly, the equation

$$x^2 \left(\frac{dy}{dx}\right)^5 + \frac{dy}{dx} - xy = 1; \ y(0) = 0 \tag{7.1}$$

is solved on [0, 2]. The analytical solution of (7.1) is $y(x) = x$. For (7.1),

$$f = xy - x^2 (y^{(1)})^5 + 1 \ , \tag{7.2}$$

so that (4.8) is not satisfied. However $M < 1$ for f given by (7.2) so that Algorithm 2 should be applicable. Also for f given by (7.2) (4.19) is satisfied for h sufficiently small so Algorithm 3 should be applicable.

For (7.1) the exact value of $y^{(1)}(x_0)$ is known from the equation itself. The numerical solution of (7.1) would be expected to have high accuracy since $y^{(n)}(x) \equiv 0$ for all $n \geqslant 2$.

A more testing equation is

$$\left(\frac{dy}{dx}\right)^5 - \frac{dy}{dx} + y = e^{5x} \ ; \ y(0) = 1 \tag{7.3}$$

which is solved on [0, 1]. The analytical solution of (7.3) is $y(x) = e^x$. For (7.3),

Table 4 Solution of (7.3) using Algorithm 3

x	SECOND ORDER METHOD ($e_r \times 10^2$) $h = 0{\cdot}1$	$h = 0{\cdot}05$	FOURTH ORDER METHOD ($e_r \times 10^6$) $h = 0{\cdot}1$	$h = 0{\cdot}05$
0·0	0·0	0·0	0·0	0·0
0·2	−1·0	−0·25	0·0	0·0094
0·4	−0·99	−0·24	0·35	0·052
0·6	−0·96	−0·23	1·2	0·10
0·8	−0·92	−0·22	2·1	0·17
1·0	−0·88	−0·21	3·3	0·25

$$f = (y^{(1)})^5 + y - e^{5x}, \tag{7.4}$$

so that again (4.8) is not satisfied. However, $M < 1$ for f given by (7.4) and (4.19) is certainly satisfied on [0, 1] so both Algorithm 2 and 3 should be applicable.

The numerical results for the solution of (7.1) and (7.3) using Algorithm 2 are shown in **Tables 1 and 2** respectively, which give the cumulative errors e_r, defined by (3.9). The numbers shown in Table 1 should be multiplied by 10^{-15}, and the errors are due in this case to round off owing to the fortuitous behaviour of the higher derivatives of y. The numbers for the second order algorithm in Table 2 should be multiplied by 10^{-2}, and those for the fourth order algorithm in Table 2, by 10^{-15} The errors given in Table 2 are primarily due to truncation owing to the fact that for (7.3), y has non-zero derivatives of all orders, and round off is small due to the use of double precision.

The numerical results for the solution of (7.1) and (7.3) using Algorithm 3 are shown in **Tables 3 and 4** respectively.

The numbers shown in Table 3 should be multiplied by 10^{-15}, and those in Table 4 should be multiplied by 10^{-2} for the errors of the second order method and by 10^{-6} for the errors of the fourth order method.

Finally it should be noted that the methods given in this paper could be applied to systems of implicit equations.

References

ALTMAN, M. (1960). Iterative methods for the numerical solution of ordinary differential equations, *Symposium on the numerical treatment of ordinary differential equations, integral and integro-differential equations*, pp. 174-176, Birkhauser Verlag.

HENRICI, P. (1964). *Elements of Numerical Analysis*, pp. 97-107, John Wiley and Sons Inc.

HENRICI, P. (1962). *Discrete Variable Methods in Ordinary Differential Equations*, pp. 194-199, 225-228, 247, John Wiley and Sons Inc.

ISAACSON, E., and KELLER, H. (1966). *Analysis of Numerical Methods*, pp. 109-122, John Wiley and Sons Inc.

MURRAY, F. J., and MILLER, K. S. (1954). *Existence Theorems for Ordinary Differential Equations*, pp. 28, 44, New York University Press.

VERNER, J. H. (1969). Implicit methods for differential equations, *Conference on the Numerical Solution of Differential Equations* (Dundee), pp. 261-266, Springer-Verlag.