# SURVEY OF DIGITAL SOLUTION
# OF DIFFERENTIAL EQUATIONS

by

ING-WEN HWANG

B.S.E.E., National Taiwan University, 1963

---

A MASTER'S REPORT

submitted in partial fulfillment of the

requirement for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1966

Approved by:

Charles A. Halijak

Major Professor

# TABLE OF CONTENT

# INTRODUCTION

Developments of modern science have confronted the scientist and the engineer with a variety of problems which cannot be solved formally. Hence the widenning interest in numerical analysis, a branch of mathematics which leads to approximate solutions by repeated application of the four basic operations of algebra. Numerical analysis has been applied to scientific and technological problems from the very beginning of applied science, but has been given new impetus by development of the electronic digital computer. Many classical numerical analysis techniques are available for engineering applications to solve differential equations. There is another group of procedures unique to the engineering literature which generates approximating recurrence relations from approximate Z-transforms. No matter how different all these techniques are in terms of their mathematical approaches and the algebra involved, they all have two things in common; the calculations are performed with discrete values and on a step-by-step basis. Consequently the time interval between steps is an important factor which affects the accuracy as well as the speed of the computation.

Up to a point, the smaller the step size used, the longer is the solution time required and the more accurate is the solution obtained. Since significant digits are limited in computation, the increased solution accuracy produced by the reduced step size is lost. This effect is termed, "round-off error" Therefore there exists an optional step size to be used in producing the numerical solution of optimal accuracy. In general no single technique is best in all cases. In fact, the effectiveness of approximate methods hinges on the type of function in question and the goodness of each method is measured by a number of factors, namely, the program set-up time, the solution speed, and solution accuracy and stability. In other words, it is a consideration of economy and accuracy and these two quantities are contradictory in nature. Therefore a brief examination of strong points and weaknesses of various types of digital techniques and comparisons among them would provide a guidepost for selection of an optimal technique.

The objective of this report is to review previous methods for digital solution of differential equations, and to illustrate their applications for approximating the solutions of ordinary differential equations.

## II. Survey of Methods for Digital Solution of Differential Equations.

Methods for approximating solution of ordinary differential equations are based on the principle of discretization. These methods have the common feature that no attempt is made to approximate the exact solution y(t) over a continuous range of the independent variable. Approximating values are sought only on a set of discrete points $t_o$, $t_1$, $t_2$, $\cdots t_n$. Generally speaking, a discrete variable method for solving a differential equation consists of an algorithm which furnishes a number $y_n$ corresponding to each lattice point $t_n$. Which is to be regarded as an approximation to the value $y(t_n)$ of the exact solution at the point $t_n$.

Discrete variable methods fall into two classes: classical numerical analysis techniques; and numerical transform techniques. Classical numerical analysis techniques yield one-step methods and multiple step methods. In a one-step method the value of $y_n$ can be found if only one initial value is known. In a multiple step method the calculation of $y_{n+1}$ requires explicit knowledge of more than one starting value. Oftentimes, a one-step method is used to start a multiple step method.

The numerical transform techniques were first introduced by Tustin and were further expanded by Madwed, Boxer-Thaler, and Halijak. These techniques generate approximating recurrence relations from approximate z-transform of $1/s^n$ and $\bar{f}/s^n$.

## III. Classical Numerical Analysis Techniques.

### 1). One-Step Methods

Let a typical first-order differential equation be given by

$$\frac{dy}{dt} = f(t,y)$$

$$y = y_o \qquad \text{at } t = t_o \tag{3.1}$$

and let it satisfy Lipschitz conditions in some closed region D. There exists a single-valued function y(t) continuous in D such that it satisfies Eq. (3.1), then it can be solved approximately by these methods.

One-step methods are usually divided into two classes; The first class includes Taylor series method and Picard's method. In these methods, y in Eq. (3.1) is approximated by a truncation series, the individual terms of which are functions of the independent variable t. The second class is represented by the methods of Euler and Runge-Kutta methods.

A. Taylor Series Method.

Consider the differential equation (3.1) with the initial condition $y=y_o$ at $t=t_o$. Let the required solution be

$$y = y(t) \tag{3.2}$$

If $t=t_o$ is not a singular point of the function, y(t) can be expanded in Taylor series about this point. Thus with

$$y_o^{(m)} = \left[\frac{d^m y}{dt^m}\right]_{t=t_o} \tag{3.3}$$

$$y = y_o + (t-t_o)y_o^{(1)} + \frac{1}{2!}(t-t_o)^2 y_o^{(2)} + \frac{1}{3!}(t-t_o)y_o^{(3)} + \cdots \tag{3.4}$$

a power series in $t$ that converges over some range $t_o \le t \le t_b$.

The value of $y^{(i)}$ is evaluated by making use of the differential equation (3.1) which can be written as

$$y^{(1)} = f(t,y) = g_1(t,y), \tag{3.5}$$

differentiating Eq. (3.5) yields

$$y^{(2)} = \frac{\partial g_1}{\partial t} + \frac{\partial g_1}{\partial y} y^{(1)} = g_2(t,y,y^{(1)}), \tag{3.6}$$

and by repeated differentiation

$$y^{(m)} = g_m(t,y,y^{(1)},y^{(2)},\ldots y^{(m-1)}) \qquad m = 1,2\cdots \tag{3.7}$$

where

$$g_m = \frac{\partial g_{m-1}}{\partial t} + \sum_{i=0}^{m-2} \frac{\partial g_{m-1}}{\partial y^{(i)}} y^{(i+1)}, \tag{3.8}$$

Setting $t=t_o$ and $y=y_o$, Eq. (3.7) becomes

$$y_o^{(m)} = g_m(t_o, y_o, y_o^{(1)}, \ldots y_o^{(m-1)}), \qquad m = 1,2\cdots \tag{3.9}$$

The truncation error after the mth term is given by

$$R_m = \frac{f^m(\xi)}{m!}(t - t_o)^m, \qquad \text{where} \qquad t_o \leq \xi \leq t \tag{3.10}$$

Replacing $f(\xi)$ with an upper bound $M^{(m)}$ to its value in the interval $(t, t_o)$. The truncation error is then bounded by

$$R_m \leq \left| \frac{M^{(m)}}{m!}(t - t_o)^m \right|, \tag{3.11}$$

In particular, for a convergent Taylor series with alternating sign, the truncation error after mth term cannot exceed the (m+1)th term Eq.(3.11) becomes

$$|R_m| \leq \left| \frac{f^{(m)}(t_o)}{m}(t - t_o)^m \right|, \tag{3.12}$$

B. 1. Euler's Method.

This method is of very little practical importance, but it illustrates in simple form the basic idea of those numerical methods which seek to determine the change of $\Delta y$ in y corresponding to a small increment of the argument.

Consider Eq.(3.1), the left-hand side of the first part is, by definition

$$\frac{dy}{dt} = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} \tag{3.13}$$

therefore, for small value of t

$$\Delta y \cong \frac{dy}{dt} \Delta t$$

Thus, the increase in y(t) when t increases to $t+\Delta t$ is approximately

$$y_{m+1} - y_m = f(t_m, y_m)T \tag{3.14}$$

where

$$T = \Delta t$$

However, the exact expression for y at $t=t_{m+1}$, using Taylor series formula is

$$z_{m+1} = z_m + Tf(t\ ,\ z\ ) + \frac{1}{2} T^2 y^{(2)}(\xi)_{m+1} \tag{3.15}$$

where

$$t_m < \xi_{m+1} < t_{m+1} \ .$$

Subtracting Eq.(3.14) from Eq.(3.15) results in

$$z_{m+1} - y_{m+1} = (z_m - y_m) + T(f(t_m, z_m) - f(t_m, y_m))$$

$$+ \frac{1}{2} T^2 y^{(2)}(\xi_{m+1}). \qquad (3.16)$$

The left-hand side of Eq.(3.16) is by definition the truncation error of y at $t=t_{m+1}$,

Let $\epsilon_{m+1} = z_{m+1} - y_{m+1}$ (3.17)

and $$k_m = \frac{f(t_m, z_m) - f(t_m, y_m)}{z_m - y_m} \cong \frac{\partial f(t_m, z_m)}{\partial y} \qquad (3.18)$$

Eq.(3.16) becomes

$$\epsilon_{m+1} = \epsilon_m + Tk_m\epsilon_m + \frac{1}{2} T^2 y^{(2)}(\xi_{m+1}). \qquad (3.19)$$

For $m = 0$

$$\epsilon_o = 0 \qquad (3.20)$$

$m = 1$

$$\epsilon_1 = \frac{1}{2} R^2 y^{(2)}(\xi_1), \qquad t_o < \xi_1 < t_1 \qquad (3.21)$$

$m = 2$

$$\epsilon_2 = (t + TK_1)\frac{1}{2} T^2 y^{(2)} (\xi_1) + \frac{1}{2} T^2 y^{(2)} (\xi_2) \qquad (3.22)$$

where

$t_1 < \xi_2 < t_2$ ;

proceeding in this manner, the truncation error at nth step is

$$\epsilon_n = T^2 \left[ \frac{1}{2} y^{(2)} (\xi_1) \prod_{j=1}^{n-1} (1 + TK_j) + \frac{1}{2} y^{(2)} (\xi_2) \prod_{j=2}^{n-1} (1 + TK_j) + \cdots + \frac{1}{2} y^{(2)} (\xi_n) \right] \qquad (3.23)$$

or $$\epsilon_n = T^2 \sum_{i=0}^{n-1} A_i a_i \qquad (3.24)$$

where

$$\begin{cases} A_i = \prod_{j=1}^{n-1} (1 + TK_j) & i \neq n \\ A_i = 1 & i = n \end{cases}$$

and $$a_i = \frac{1}{2} y^{(2)} (\xi_i) \qquad (3.25)$$

Thus errors $\frac{1}{2} y^{n}(\xi_i)$ introduced at each step because of the inaccuracy of formula (3.14) are to be multiplied by amplification factors $A_i$ before being summed.

C. Runge-Kutta Method.

This method is an algorithm designed to approximate the Taylor's series solution.

Consider the Taylor's Series

$$y_{n+1} = y_n + T\, y_n + \frac{T^2}{2}\, y_n + \cdots, \tag{3.26}$$

where

$$y_{n+1} = y((n+1)T) \qquad \text{and} \qquad y_n = y(nT).$$

If the Runge-Kutta formula is derived by retainning terms in the Taylor's series expression up to mth power of T, this formula is called the Runge-Kutta method of mth order accuracy or the Runge-Kutta mth order method.

C. 1. Second Order Runge-Kutta Formula.

Consider increments in y defined by the equations

$$\Delta y = R_1 \Delta' y + R_2 \Delta'' y \tag{3.27}$$

where

$$\Delta' y = f(t_n, y_n)\, T,$$

$$\Delta''y = f(t_n + aT,\ y_n + bT)\ T,$$

and $R_1$, $R_2$, a and b are constants.

Expanding Eq.(3.27) with respect to Taylor's series of two variables, results in

$$\Delta y = T(R_1 + R_2)\ f_n + (a(f_t)_n + b(f_y)_n)\ T^2\ R_2 + \cdots \tag{3.28}$$

where

$$f_n = f(t_n, y_n) = y_n\ , \qquad (f_t)_n = \left(\frac{\partial f}{\partial t}\right)\Bigg|_{t=t_n,\ y=y_n},$$

and

$$(f_y)_n = \left(\frac{\partial f}{\partial y}\right)\Bigg|_{t=t_n,\ y=y_n}$$

Eq.(3.28) will be equal to Eq.(3.26) up to second order of T if its parameters take the values

$$R_1 + R_2 = 1 \tag{3.29}$$

$$R_2 a = 1/2$$

$$R_2 b = 1/2$$

There are four constants to be determined and only three equations. Choose $R_2$ as parameter, the constants a, b, and $R_1$ can be determined in terms of $R_2$. Thus

$$a = b = \frac{1}{2R_2} \tag{3.30}$$

$$R_1 = 1 - R_2$$

Consequently, an infinite number of forms of Eq. (3.28) can be established. Formulas of higher order can be obtained in the similar manner although no formula beyond fifth order has ever been developed.

C. II. Third Order Runge-Kutta Formula.

Again take the differential equation

$$\frac{dy}{dt} = f(t, y). \tag{3.31}$$

and expand it with respect to $(t_n, y_n)$ by Taylor series for two variables to obtain

$$\Delta y = f_n T + \frac{1}{2} (f_t + f_y f)_n T^2 + \frac{1}{6} \left\{ f_{tt} + 2f_{ty} f + f_{yy} f^2 + (f_t + f_y f) f_y \right\}_n T^3 + \cdots \tag{3.32}$$

Define the increment of y by

$$\Delta y = R_1 \Delta' y + R_2 \Delta'' y + R_3 \Delta''' y \tag{3.33}$$

where

$$\Delta' y = f(t_n, y_n)\ T$$

$$\Delta'' y = f(t_n + mT,\ y_n + m\Delta' y)\ T$$

$$\Delta''' y = f(t_n + \lambda T,\ y_o + p\Delta'' y + (\lambda - p)\Delta' y)\ T$$

The symbols m, $\lambda$ , and p are constants. Expanding Eq.(3.33) with respect to Taylor's series at $(t_n, y_n)$, results in

$$\Delta' y = Tf_n$$

$$\Delta'' y = T\Big\{ f_n + mT(f_t)_n + m\Delta' y(f_y)_n + \frac{1}{2!}\Big[(mT)^2(f_{tt})_n + 2m^2T\Delta' y(f_{ty})_n + (m\Delta' y)^2(f_{yy})_n\Big] + \cdots\Big\}$$

$$\Delta''' y = T\Big\{ f_n + \lambda T(f_t)_n + \Big[\rho\Delta'' y + (\lambda - \rho)\Delta' y\Big](f_y)_n + \frac{1}{2!}\Big\{(\lambda T)^2(f_{tt})_n + 2\lambda T\Big[\rho\ \Delta'' y + (\lambda - \rho)\Delta' y\Big](f_{ty})_n + \Big[\rho\Delta'' y + (\lambda - \rho)\Delta' y\Big]^2(f_{yy})_n\Big\} + \cdots\Big\} \qquad (3.34)$$

where

$$(f_{yy})_n = \left(\frac{\partial^2 f}{\partial y^2}\right)_{t=t_n,\ y=y_n}. \qquad (f_y)_n = \left(\frac{\partial f}{\partial y}\right)_{t=t_n,\ y=y_n}$$

$$(f_t)_n = \left(\frac{\partial f}{\partial t}\right)_{t=t_n,\ y=y_n}. \qquad f_n = f(t_n, y_n)$$

Eliminating $\Delta' y$ and $\Delta'' y$ from the right hand side of the second and third equations of Eq.(3.34) and substiting into Eq.(3.33), the result coincides with Taylor's series expansion up to the third power of T provided the following relations hold among the constants

$$R_1 + R_2 + R_3 = 1 \tag{3.35}$$

$$R_2 m + R_3 \lambda = \frac{1}{2}$$

$$R_2 m^2 + R_3 \lambda^2 = \frac{1}{3}$$

$$R_3 \rho m = \frac{1}{6}$$

Since there are six constants to be determined, namely, $R_1$, $R_2$, $R_3$, m, $\lambda$, and $\rho$, and only four equations, the constants m and $\lambda$ can be chosen as parameters. The constants $R_1$, $R_2$, $R_3$ and $\rho$ as calculated from Eq.(3.35) are

$$R_1 = \frac{6m\lambda - 3(m + \lambda) + 2}{6m\lambda} \tag{3.36}$$

$$R_2 = \frac{2 - 3\lambda}{6m(m - \lambda)}$$

$$R_3 = \frac{2 - 3m}{6\lambda(\lambda - m)}$$

$$\rho = \frac{\lambda(\lambda - m)}{m(2 - m)}$$

From Eq.(3.35), it is clear that $R_3$ cannot be zero and $R_1$ and $R_2$ cannot both zero. If $R_2=0$, then by Eq.(3.35)

$$\lambda = \frac{2}{3} \qquad R_3 = \frac{3}{4} \qquad R_1 = \frac{1}{4} \qquad \text{and} \quad \rho m = \frac{2}{3} \tag{3.37}$$

Moreover, assume $\rho$ to be 1/3, then m is equal to 2/3. Therefore

$$\Delta y = \frac{1}{4} (\Delta' y + 3 \Delta''' y) \tag{3.38}$$

$$\Delta' y = f(t_n, y_n)\, T$$

$$\Delta'' y = f(t_n + \frac{2}{3} T,\ y_n + \frac{2}{3} \Delta' y)\, T$$

$$\Delta''' y = f\left[t_n + \frac{2}{3} T,\ y_n + \frac{2}{3} (\Delta' y + \Delta'' y)\right] T$$

On the other hand, if $\lambda=0$, then

$$m = \frac{2}{3} \qquad R_2 = \frac{3}{4} \qquad R_3 = \frac{1}{4\rho} \qquad \text{and} \qquad R_1 = \frac{\rho - 1}{4\rho} \tag{3.39}$$

If it is assumed that $\rho=1$, then another form can be constructed from Eq.(3.39) and Eq.(3.33); it is

$$\Delta y = \frac{1}{4} (3 \Delta'' y + \Delta''' y) \tag{3.40}$$

where

$$\Delta' y = f(t_n, y_n)\, T$$

$$\Delta''y = f(t_n + \frac{2}{3} T, y_n + \frac{2}{3} \Delta' y) T$$

$$\Delta'''y = f(t_n, y_n + \Delta'' y - \Delta' y) T$$

C. IV. Fourth Order Runge-Kutta Formula.

The Runge-Kutta fourth order formula has the form

$$\Delta y = R_1 \Delta' y + R_2 \Delta'' y + R_3 \Delta''' y + R_4 \Delta'''' y \tag{3.41}$$

where

$$\Delta' y = f(t_n, y_n) T$$

$$\Delta'' y = f(t_n + mT, y_n + m \Delta' y) T$$

$$\Delta''' y = f(t_n + \lambda T, y_n + \rho \Delta'' y_n + (\lambda - \rho) \Delta' y) T$$

$$\Delta'''' y = f(t_n + \mu T, y_n + \sigma \Delta''' y + \tau \Delta'' y + (\mu - \sigma - \tau) \Delta' y) T.$$

Expand second, third, fourth and fifth equations of Eq.(3.41) to obtain

$$\Delta' y = f(t_n, y_n) T \tag{3.42}$$

$$\begin{aligned}
\Delta''y = T\Big\{ & f(t_n, y_n) + mT(f_t)_n + m\Delta'y(f_y)_n \\
& + \frac{1}{2!}\left((mT)^2 (f_{tt})_n + 2(mT)(m\Delta'y)(f_{ty})_n + (m\Delta'y)^2(f_{yy})_n\right) \\
& + \frac{1}{3!}\Big[(mT)^3 (f_{ttt})_n + 3(mT)^2(m\Delta'y)(f_{tty})_n \\
& + 3(mT)(m\Delta'y)^2(f_{tyy})_n + (m\Delta'y)^3(f_{yyy})_n\Big] + \cdots \Big\}
\end{aligned}$$

$$\begin{aligned}
\Delta'''y = T\Big\{ & f(t_n, y_n) + \lambda T(f_t)_n + (\rho\Delta''y + (\lambda - \rho)\Delta'y)(f_y)_n \\
& + \frac{1}{2!}\Big[(\lambda T)^2(f_{tt})_n + 2(\lambda T)(\rho\Delta''y + (\lambda - \rho)\Delta'y)(f_{ty})_n \\
& + (\rho\Delta''y + (\lambda - \rho)\Delta'y)^2(f_{yy})_n\Big] \\
& + \frac{1}{3!}\Big[(\lambda T)^3(f_{ttt})_n + 3(\lambda T)^2(\rho\Delta''y + (\lambda - \rho)\Delta'y)(f_{tty})_n \\
& + 3(\lambda T)(\rho\Delta''y + (\lambda - \rho)\Delta'y)^2(f_{tyy})_n + (\rho\Delta''y + (\lambda - \rho)\Delta'y)^3 \\
& + (\rho\Delta''y + (\lambda - \rho)\Delta'y)^3(f_{yyy})_n\Big] + \cdots \Big\}
\end{aligned}$$

$$\begin{aligned}
\Delta^{(4)}y = T\Big\{ & f_n + \mu T(f_t)_n + (\sigma\Delta^{(3)}y + \tau\Delta^{(2)}y + (\mu - \sigma - \tau)\Delta'y)(f_y)_n \\
& + \frac{1}{2!}\Big[(\mu T)^2(f_{tt})_n + 2(\mu T)(\sigma\Delta^{(3)}y + \tau\Delta^{(2)}y \\
& + (\mu - \sigma - \tau)\Delta'y)(f_{ty})_n + ((\mu - \sigma - \tau)\Delta'y + \sigma\Delta^{(3)}y \\
& + \tau\Delta^{(2)}y)^2(f_{yy})_n\Big] + \cdots \Big\}
\end{aligned}$$

Eliminating $\Delta' y$, $\Delta'' y$ and $\Delta''' y$ from the right-hand side of Eq. (3.42) by successive substitution, putting these results into Eq. (3.41) and comparing with Taylor's series yield

$$R_1 + R_2 + R_3 + R_4 = 1 \tag{3.43}$$

$$R_2 m + R_3 \lambda + R_4 \mu = \frac{1}{2}$$

$$R_2 m^2 + R_3 \lambda^2 + R_4 \mu^2 = \frac{1}{3}$$

$$R_3 m \rho + R_4 (m \sigma + \lambda\ ) = \frac{1}{6}$$

$$R_2 m^3 + R_3 \lambda^3 + R_4 \mu^3 = \frac{1}{4}$$

$$R_3 m \lambda \rho + R_4 \mu (m \sigma + \lambda \tau) = \frac{1}{8}$$

$$R_3 m^3 \rho + R_4 (m_1^2 \sigma + \lambda^2 \tau) = \frac{1}{12}$$

$$R_4 m \lambda \tau = \frac{1}{24}$$

There is a one-parameter family of solutions derived by Kutta as follows:

$$R_1 = \frac{1}{6} \qquad R_2 = \frac{2 - t}{3} \qquad R_3 = \frac{t}{3} \qquad R_4 = \frac{1}{6} \tag{3.44}$$

$$m = 1/2, \quad \lambda = 1/2, \quad \rho = 1/2t, \quad \mu = 1, \quad \sigma = 1 - t, \quad \tau = t$$

If $t=1$, a Runge-Kutta fourth order formula can be constructed as follows:

$$\Delta y = \frac{1}{6} (\Delta' y + 2\Delta'' y + 2\,\Delta^{(3)} y + \Delta^{(4)} y) \tag{3.45}$$

where

$$\Delta y = y_{n+1} - y_n,$$

$$\Delta' y = T\, f(t_n, y_n),$$

$$\Delta^{(2)} y = T\, f(t_n + \frac{1}{2} T,\, y + \frac{1}{2} \Delta' y)$$

$$\Delta^{(3)} y = T\, f(t_n + \frac{1}{2} T,\, y_n + \frac{1}{2} \Delta'' y)$$

$$\Delta^{(4)} y = T\, f(t_n + T,\, y_n + \Delta'' y)$$

4). Fifth Order Runge-Kutta Formula

The 5th order Runge-Kutta formulas were derived recently by H. A. Luther and H. P. Konen. The derivations are as follows:

Let $dy/dt = f(t, y)$ together with $y(t_o) = y_o$ (3.46)

The fifth order Runge-Kutta formulas are phrased as

$$\Delta y = + \sum_{1}^{6} R_i\, \Delta^{(i)} y$$

$$\Delta' y = T\, f(t_n, y_n) \tag{3.47}$$

$$\Delta^{(s)} y = T\, f(t_n + a_s T,\ y + \sum_{j=1}^{s-1} b_{sj}\, \Delta^{(j)} y).$$

where $2 \leqslant s \leq 6$. and the coefficients $R_i$ are the constants to be determined. The usual procedure yields 44 equations involving the various parameters. Assume that

$$a_s = \sum_{j=1}^{s-1} b_{sj}, \qquad 2 \leqslant s \leqslant 6 \tag{3.48}$$

and eliminate easily identifiable combinations to obtain the following 16 relations

$$\left\{\begin{aligned}
&R_1 + R_2 + R_3 + R_4 + R_5 + R_6 = 1,\\
&a_2R_2 + a_3R_3 + a_4R_4 + a_5R_5 + a_6R_6 = \frac{1}{2},\\
&a_2^2R_2 + a_3^2R_3 + a_4^2R_4 + a_5^2R_5 + a_6^2R_6 = \frac{1}{3}\\
&a_2^3R_2 + a_3^3R_3 + a_4^3R_4 + a_5^3R_5 + a_6^3R_6 = \frac{1}{4}\\
&a_2^4R_2 + a_3^4R_3 + a_4^4R_4 + a_5^4R_5 + a_6^4R_6 = \frac{1}{5}
\end{aligned}\right\} \tag{3.49a}$$

$$\left\{\begin{array}{l}
c_1R_3 + c_2R_4 + c_3R_5 + c_4R_6 = \dfrac{1}{6} \; . \\[2ex]
a_3c_1R_3 + a_4c_2R_4 + a_5c_3R_5 + a_6c_4R_6 = 1/8, \\[2ex]
a_3^2c_1R_3 + a_4^2c_2R_4 + a_5^2c_3R_5 + a_6^2c_4R_6 = 1/10, \\[2ex]
d_1R_3 + d_2R_4 + d_3R_5 + d_4R_6 = 1/12, \\[2ex]
a_3d_1R_3 + a_4d_2R_4 + a_5d_3R_5 + a_6d_4R_6 = 1/15, \\[2ex]
c_1^2R_3 + c_2^2R_4 + c_3^2R_5 + c_4^2R_6 = 1/20,
\end{array}\right\} \qquad (3.49a)$$

$$\left\{\begin{array}{l}
a_2^3b_{32}R_3 + (a_2^3b_{42} + a_3^3b_{43})R_4 + (a_2^3b_{52} + a_3^3b_{53} + a_4^3b_{54})\, R_5 \\
\qquad + (a_2^3b_{62} + a_3^3b_{63} + a_4^3b_{64} + a_5^3b_{65})\, R = 1/20 \\[2ex]
c_1b_{32}R_3 + (c_1b_{53} + c_2b_{54})\, R_5 + (c_1b_{63} + c_2b_{64} + c_3b_{65})\, R_6 = 1/24, \\[2ex]
d_1b_{43}R_4 + (d_1b_{53} + d_2b_{54})\, R_5 + (d_1b_{63} + d_2b_{64} + d_3b_{65})\, R_6 = 1/60, \\[2ex]
(a_3 + a_4)c_1b_{43}R_4 + \left[(a_3 + a_5)c_1b_{53} + (a_4 + a_5)c_2b_{54}\right] R_{54} \\[2ex]
\quad + \left[(a_3 + a_6)c_1b_{63} + (a_4 + a_6)c_2b_{64} + (a_5 + a_6)c_3b_{65}\right] R_6 = 7/120 \\[2ex]
c_1b_{43}b_{54}R_5 + \left[\, c_1b_{43}b_{64} + (c_1b_{53} + c_2b_{54})b_{65}\right] R_6 = 1/120,
\end{array}\right\}$$

(3.49a)

where

$$c_i = \sum_{j=2}^{i+1} a_j b_{i+2,\, j} \,, \qquad d_i = \sum_{j=2}^{i+1} a_j^2 b_{i+2,\, j} \,. \tag{3.49b}$$

simplification requires that

$$R_2 = 0 \tag{3.50}$$

and

$$\frac{a_i^2}{2} = c_{i-2} \,, \qquad 3 \leqslant i \leqslant 6 \tag{3.51}$$

then Eq.(3.49) can be simplified considerably. After eliminating duplicates and combining some equations solve, in addition to (3.51),

$$R_1 + R_3 + R_4 + R_5 + R_6 = 1, \tag{3.52a}$$

and

$$a_3R_3 + a_4R_4 + a_5R_5 + a_6R_6 = 1/2,$$

$$a_3^2R_3 + a_4^2R_4 + a_5^2R_5 + a_6^2R_6 = 1/3,$$

$$a_3^3R_3 + a_4^3R_4 + a_5^3R_5 + a_6^3R_6 = 1/4, \tag{3.57b}$$

$$a_3^4R_3 + a_4^4R_4 + a_5^4R_5 + a_6^4R_6 = 1/5,$$

and

$$a_3b_{43}R_4 + (a_3b_{53} + a_4b_{54})\ R_5 + (a_3b_{63} + a_4b_{64} + a_5b_{65})\ R_6 = 1/6$$

$$a_3^2b_{43}R_4 + (a_3^2b_{53} + a_4^2b_{54})\ R_5 + (a_3^2b_{63} + a_4^2b_{64} + a_5^2b_{65})\ R_6 = 1/12$$

$$a_3^3b_{43}R_4 + (a_3^3b_{53} + a_4^3b_{54})\ R_5 + (a_3^3b_{63} + a_4^3b_{64} + a_5^3b_{65})\ R_6 = 1/20,$$

$$a_4a_3b_{43}R_4 + a_5(a_3b_{53}a_4b_{54})\ R_5 + a_6(a_3b_{63} + a_4b_{64} + a_5b_{65})\ R_6$$
$$= 1/8,$$

$$a_4a_3^2b_{43}R_4 + a_5(a_4^2b_{53} + a_4^2b_{54})\ R_5 + a_6(a_3^2b_{63} + a_4^2b_{64} + a_5^2b_{65})\ R_6$$
$$= 1/15, \tag{3.52c}$$

and

$$a_3b_{43}b_{54}R_5 + \left[\ a_3b_{43}b_{64} + (a_3b_{53} + a_4b_{54})\ b_{65}\right] R_6 = 1/24,$$

$$a_3^2b_{43}b_{54}R_5 + \left[\ a_3^2b_{43}b_{64} + (a_3^2b_{53} + a_4^2b_{54})\ b_{65}\right] R_6 = 1/60, \tag{3.52d}$$

The situation is now as follows. Equations (3.52b), (3.52c), and (3.52d) are to be solved independently. Then (3.52a) yields $R_1$. Equation (3.51) are used to find $b_{32}$, $b_{42}$, $b_{52}$, $b_{62}$. Then equations (3.48) determine $b_{21}$, $b_{31}$, $b_{41}$, $b_{51}$ and $b_{61}$. This, with $R_2$=0, completes the solution. The family of solutions due to Kutta may now be found by taking $b_{65}$=0. From (3.52d), $a_3$=2/5. It then develops that the third equation in (3.52c) has the left member equal to $-a_3a_4$ times the left member of the first of this group, plus $a_3 + a_4$ times the left member of the second of this group. This forces $a_4$ to be 1. Equations (3.52c) may now be solved for $b_{64}R_6$ and $b_{54}R_5$. When the results are substituted in $a_3 b_{43} (b_{54}R_5 + b_{64}R_6) = 1/24$ (the consequence of (3.52d) and $b_{65}=0$) it is found that $b_{43} = 15/4$. In summary, there results the following description of a family of solutions:

$$R_2 = 0, \quad b_{65} = 0, \quad b_{43} = 15/4, \quad a_3 = 2/5, \quad a_4 = 1, \tag{3.53}$$

with

$$\begin{aligned}
R_3 &= 125\left[10a_5a_6 - 5(a_5 + a_6) + 3\right] / \left[72(2 - 5a_5)(2 - 5a_6)\right], \\
R_4 &= \left[8a_5a_6 - 7(a_5 + a_6) + 6\right] / \left[36(1 - a_5)(1 - a_6)\right], \\
R_5 &= \left[1 - a_6\right] / \left[12a_5(1 - a_5)(5a_5 - 2)(a_5 - a_6)\right], \qquad (3.53) \\
R_6 &= \left[1 - a_5\right] / \left[12a_6(1 - a_6)(5a_6 - 2)(a_6 - a_5)\right], \\
R_1 &= 1 - R_3 - R_4 - R_5 - R_6,
\end{aligned}$$

and

$$\begin{aligned}
b_{53} &= 5\left[7 - 10a_6 - 108R_4(1 - a_6)\right] / \left[144R_5(a_5 - a_6)\right], \\
b_{54} &= \left[1 - a_6\right] / \left[36R_5(a_5 - a_6)\right], \\
b_{63} &= 5\left[7 - 10a_5 - 108R_4(1 - a_5)\right] / \left[144R_6(a_6 - a_5)\right], \qquad (3.53c) \\
b_{64} &= \left[1 - a_5\right] / \left[36R_6(a_6 - a_5)\right],
\end{aligned}$$

and

$$b_{32} = 2/\left| 25a_2 \right|,$$

$$b_{42} = 1/a_2,$$

$$b_{52} = \left[ 5a_5^2 - 4b_{53} - 10b_{54} \right] / \left[ 10a_2 \right] \qquad (3.53d)$$

$$b_{62} = \left[ 5a_6^2 - 4b_{63} - 10b_{64} \right] / \left[ 10a_2 \right],$$

and

$$b_{j1} = a_j - \sum_{j=2}^{j-1} b_{ji}, \qquad 2 \leqslant j \leqslant 6. \qquad (3.53e)$$

It becomes very simple to construct a fifth order Runge-Kutta formula based on the coefficients derived from equation (3.53). The result is

$$y_{n+1} = y_n + \left\{ 4\Delta' y + (16 + \sqrt{6})\,\Delta^{(5)}y + (16 - \sqrt{6})\,\Delta^{(6)}y \right\}/36,$$

$$\Delta' y = Tf(t_n, y_n),$$

$$\Delta'' y = Tf(t_n + 4T/11,\; y_n + 4\Delta' y/11), \qquad (3.54)$$

$$\Delta''' y = Tf(t_n + 2T/5,\; y_n + \left\{ 9\Delta' y + 11\Delta'' y \right\}/50),$$

$$\Delta^{(4)} y = Tf(t_n + T,\; y_n + \left\{ -11\Delta'' y + 15\Delta''' y \right\}/4).$$

$$\Delta^{(5)} y = Tf(t_n + (6 - \sqrt{6})T/10,\; y_n + \left\{ (81 + 9\sqrt{6})\Delta' y + (255 - 55\sqrt{6})\Delta''' y + (24 - 14\sqrt{6})\,\Delta^{(4)}y \right\}/600),$$

$$\Delta^{(6)} y = Tf(t_n + (6 + \sqrt{6})T/10,\; y_n + \left\{ (81 - 9\sqrt{6})\Delta' y + (255 + 55\sqrt{6})\Delta''' y + (24 + 14\sqrt{6})\,\Delta^{(4)}y \right\}/6000),$$

Let $b_{43} = 0$ instead of letting $b_{65} = 0$, then Eqs. (3.52d) yields

$$(a_3 b_{53} + a_4 b_{54})\, b_{65} R_6 = 1/24,$$

$$(a_3 b_{53} + a_4 b_{54})\, b_{65} R_6 = 1/60,$$

Using these equations in conjunction with the first, second, fourth, and fifth equations of (3.52c), the terms in $b_{63}$, $b_{64}$, and $b_{65}$ can be eliminated and there result two equations:

$$R\,(a_6 - a_5) = (4a_6 - 3)\, b_{65} R_6,$$

$$R\,(a_6 - a_5) = (5a_6 - 4)\, b_{65} R_6.$$

These lead to $a_6 = 1$. Another family of solution occurs with

$$R_2 = 0, \qquad b_{43} = 0, \qquad a_6 = 1, \tag{3.55a}$$

and

$$R_3 = \left[\, 3 - 5(a_4 + a_5) + 10a_4a_5 \,\right] / \left[ 60a_3(a_4 - a_3)(a_3 - a_5)(a_3 - 1) \right],$$

$$R_4 = \left[\, 3 - 5(a_3 + a_5) + 10a_3a_5 \right] / \left[\, 60a_4(a_3 - a_4)(a_4 - a_5)(a_4 - 1) \,\right],$$

$$R_5 = \left[\, 3 - 5(a_3 + a_4) + 10a_3a_4 \,\right] / \left[ 60a_5(a_3 - a_5)(a_5 - a_4)(a_5 - 1) \,\right],$$

$$R_6 = \left[\, 12 - 15(a_3 + a_4 + a_5) + 20(a_3a_4 + a_4a_5 + a_5a_3) - 30a_3a_4a_5 \,\right] / \left[\, 60(1 - a_3)(1 - a_4)(1 - a_5) \,\right],$$

$$R_1 = 1 - R_3 - R_4 - R_5 - R_6, \tag{3.55b}$$

and

$$b_{53} = \left[\, 2 - 5a_4 \,\right] / \left[\, 120R_5a_3(1 - a_5)(a_3 - a_4) \,\right].$$

$$b_{54} = \left[\, 2 - 5a_3 \,\right] / \left[\, 120R_5a_4(1 - a_5)(a_4 - a_3) \,\right],$$

$$b_{63} = \left[\, 6 - 2a_3 - 10a_4 - 14a_5 + 5a_3a_4 + 25a_4a_5 + 10a_5^2 - 20a_5^2a_4 \,\right] / \left[\, 120R_6a_3(1 - a_5)(a_3 - a_4)(a_3 - a_5)\right], \tag{3.55c}$$

$$b_{64} = \left[\, 6 - 2a_4 - 10a_3 - 14a_5 + 5a_3a_4 + 25a_3a_5 + 10a_5^2 - 20a_5^2a_3 \,\right] / \left[\, 120R_6a_4(1 - a_5)(a_4 - a_3)(a_4 - a_5)\right],$$

$$b_{65} = \left[\, 3 - 5a_3 - 5a_4 + 10a_3a_4 \,\right] / \left[\, 60R_6a_5(a_5 - a_3)(a_5 - a_4)\right],$$

and

$$b_{32} = a_3^2 \;/ \left[\, 2a_2 \,\right],$$

$$b_{42} = a_4^2 \;/ \left[\, 2a_2 \,\right], \tag{3.55d}$$

$$b_{52} = \left[\, a_5^2 - 2a_3b_{35} - 2a_4b_{54} \right] / \left[\, 2a_2 \right],$$

$$b_{62} = \left[\, 1 - 2a_3b_{63} - 2a_4b_{64} - 2a_5b_{65} \,\right] / \left[\, 2a_2 \,\right],$$

and

$$b_{j1} = a_j - \sum_{i=2}^{j-1} b_{ji}, \qquad 2 \le j \le 6 \tag{3.55e}$$

Thus a formula can be constructed as follows:

$$y_{n+1} = y_n + \left\{ \Delta' y + 5\Delta''' y + 5\Delta^{(5)} y + \Delta^{(6)} y \right\} / \ 12, \tag{3.56}$$

$$k_1 = Tf(t_n, y_n),$$

$$k_2 = Tf(t_n + T/2, y_n + k_1/2),$$

$$k_3 = Tf(t_n + (5 - 5)T/10, y_n + \left\{ 2k_1 + (3 - 5)k_2 \right\}/10),$$

$$k_4 = Tf(t_n + T/2, y_n + k_1 + k_2/4),$$

$$k_5 = Tf(t_n + (5 + \sqrt{5})T/10, y_n + \left\{ (1 - \sqrt{5})k_1 - 4k_2(5 + 3\sqrt{5})k_3 \right.$$

$$\left. + 8k_4 \right\} /20,$$

$$k_6 = Tf(t_n + T, y_n + ( 5 - 1)k_1 + (2\sqrt{5} - 2)k_2 + (5 - \sqrt{5})k_3$$

$$- 8k_4 + (10 - 2\sqrt{5})k_5/4.$$

This fifth order formula is not in Kutta's family. It belongs to Lobatto quadrature formulas which have errors of order $T^7$ rather than $T^6$.

The accumulated truncation error of Runge-Kutta's method can be calculated as follows:

Let

$$z_{n+1} = z_n + T\bar{\Delta} z(t_n, z_n ; T) \tag{3.57}$$

be the exact solution of Eq. (3.46),

and $$y_{n+1} = y_n + T\Delta y(t_n, y_n\ ; T) \tag{3.58}$$

be the calculated value of the solution

Substituting Equation (3.57) from Equation (3.58) yields

$$e_{n+1} = e_n + T\left[\Delta y(t_n, y_n\ ; T) - \bar{\Delta} z(t_n, z_n\ ; T)\right] \tag{3.59}$$

by application of the triangle inequality, Equation (3.59) becomes

$$\begin{aligned} \| e_{n+1} | &\leq | e_n | + T \| \Delta y(t_n, y_n\ ; T) - \Delta z(t_n, z_n\ ; T) \| \\ &+ T \| \Delta z(t_n, z_n\ ; T) - \bar{\Delta} z(t_n, z_n\ ; T) \| \end{aligned} \tag{3.60}$$

The Lipschitz condition yields

$$\begin{aligned} \| \Delta y(t_n, y_n\ ; T) - \Delta z(t_n, z_n\ ; T) \| &\leq L \| y_n - z_n \| \\ &= L \| e_n \| \end{aligned} \tag{3.61}$$

and $$\| \Delta z(t_n, z_n\ ; T) - \bar{\Delta} z(t_n, z_n\ ; T) \| \leq N(T^P) \tag{3.62}$$

where

$$N = \frac{1}{(p + 1)!} \quad \text{Max} \, \| f^{(p)}(y) \|$$

and P is the order of the Runge-Kutta formula

Hence from Equation (3.60) the required truncation error is

$$\| e_{n+1} \| \leq (1 + LT) \| e_n \| + T^{(p+1)} L \tag{3.63}$$

The value of L can be determined from a Runge-Kutta formula and different formulas correspond to different values of L. All L's have upper bounds.

The Runge-Kutta method seems to be tedious because values of f(t,y) have to be calculated a number of times per time increment. However, the formulas are systematic and hence can be easily programmed on an automatic machine. No special starting procedure is required and calculations can often be checked by repetition using a different step size. Furthermore, such a method is particularly useful if certain coefficients in the differential equation are empirical formulas for which analytical expressions are not known. The step size can be altered as desired. It should be understood that the derivatives as evaluated by Runge-Kutta process are not the actual derivatives at the various points within the step as commonly assumed.

2). Multiple step Methods.

The one-step methods are necessary for obtainning initial values in the solution of a differential equation. However, they involve too much labor to be used for obtaining a numerical solution over an extended range. This can be offset by using multiple step methods.

Multiple step methods are always expressed in the difference-differential equation form

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_o y_n = T\Big\{ \beta_k f_{n+k} + \beta_{k-1} f_{n+k-1}$$
$$+ \cdots + \beta_o f_n \Big\}$$

$$n = 0,\ 1,\ 2,\ \cdots \qquad (3.64)$$

where k is a fixed integer, $f_m = f(t_m, y_m)(m = 0,1,3,\cdots)$, and where $\alpha_\mu$ and $\beta_\mu (\mu = 0,1,2,\cdots)$ are real constants which do not depend on n. Any $\alpha_k$ is always assumed unequal to zero. Equation (3.64) defines the general linear k-step method. If $\beta_k$=0, the formula (3.64) is called "closed"; otherwise it is called open.

Unlike one-step methods, multiple step methods are not self-starting; if some values $y_{n+k-1}$, $y_{n+k-2}$, $\cdots$ $y_n$ are not known, these methods break down. Such is the case at the beginning of the computation, where the initial condition furnishes only one of the required k+1 values, or at places where the step T is changed.

Stability and convergence are two important factors that affect the availability of a multiple step method. To insure its stable and convergence, two rules must be fulfilled;

1. The characteristic polynomial

$$\alpha_k S^k + \alpha_{k-1} S^{k-1} + \cdots + \alpha_o = 0$$

of Equation (3.75) has no root with modulus exceeding 1. and the root of modulus 1 must be simple.

II. The order of the associated difference operator be at least 1.

A. Adams-Bashforth Method

Consider the initial value problem

$$y' = f(t, y) \qquad\qquad y(t_o) = y_o \tag{3.65}$$

An exact solution of the differential equation (3.65) by definition satisfies the identity

$$y(t + k) - y(t) = \int_t^{t+k} f(t, y)\, dt \tag{3.66}$$

for any two values of t in the interval [a,b]. Replace f(t,y) in the right-hand side of Equation (3.66) by an interpolating polynomial on a set of points $t_n$ where $y_n$ has already been computed or is just about to be computed. Equate the intergral and accept its value as the increment of the approximate values $y_n$ between t and t+k. If it is assumed that the interpolating points are $t_p$, $t_{p-1}$, $t_{p-2}$, ··· $t_{p-q}$, then the polynomial replacing f(t,y) is given by

$$p(t) = \sum_{m=0}^{q} (-1)^m \binom{-s}{m} \nabla^m f_p \tag{3.67}$$

$$s = \frac{t - t_p}{T}$$

where q is an arbitrary integer. This formula is developed in Appendix.

Equation (3.67) can be substituted into Equation (3.66) to yield

$$y_{p+1} - y_p \cong \int_{t_p}^{t_{p+1}} p(t)dt = T \sum_{m=0}^{q} r_m \nabla^m f_p \tag{3.68}$$

where $$r_m = (-1)^m \frac{1}{T} \int_{t_p}^{t_{p+1}} \binom{-s}{m} ds \tag{3.69}$$

and $$y_p = y(t_p)$$

Construct a generating function G(k) as follows

$$G(k) = \sum_{m=0}^{\infty} r^m k^m = \sum_{m=0}^{\infty} (-1)^m \int_{t_p}^{t_{p+1}} \binom{-s}{m} ds$$

$$= \int_{t_p}^{t_{p+1}} \Sigma(-1)^m \binom{-s}{m} ds = \int_{t_p}^{t_{p+1}} (1 - k)^{-s} \, ds \tag{3.70}$$

The identity $(1-k)^{-s} = \exp(-s\log(1-k))$ causes Eq.(3.70) to become

$$G(k) = - \frac{k}{(1 - k) \log(1 - k)}$$

this may be written as

$$- G(k) \frac{\log(1 - k)}{k} = \frac{1}{1 - k} .$$

Since $$\frac{1}{1 - k} = 1 + k + k^2 + \cdots$$

and $$- \frac{\log(1 - k)}{k} = 1 + \frac{1}{2} k + \frac{1}{3} k^2 + \cdots$$

one can conclude that

$$(r_o + r_1 k + r_2 k^2 + \cdots)(1 + \frac{1}{2} k + \frac{1}{3} k^2 + \cdots) = 1 + k + k^2 + \cdots$$

By comparing the coefficients of corresponding powers of k, a relation can be found

$$r_m + \frac{1}{2} r_{m-1} + \frac{1}{3} r_{m-2} + \cdots + \frac{1}{m+1} r = 1 \qquad (3.71)$$

thus it is possible to calculate $r_m$ recursively. Some values of $r_m$ calculated from Eq.(3.71) are

| m | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $r_m$ | 1 | $\frac{1}{2}$ | $\frac{5}{12}$ | $\frac{3}{8}$ | $\frac{251}{720}$ | $\frac{95}{288}$ | $\frac{19087}{60480}$ |

If $y^{(q+2)}(t)$ is continuous in $[a,b]$, then $y'(t)$ can be expressed as

$$y' = \sum_{m=0}^{q} (-1)^m \binom{-s}{m} \nabla^m y'(t_p) + (-1)^{(q+1)} \binom{-s}{q+1} T^{q+1} y^{(q+2)}(\xi)$$

where $\xi$ is a point between the largest and the smallest of the values $t, t_p$, and $t_{p-q}$. Integrating $y'$ between $t_p$ and $t_{p+1}$, we get

$$y(t_{p+1}) - y(t_p) = T \sum_{m=0}^{q} r_m \nabla^m f_p + R_q^{AB} \tag{3.72}$$

where

$$R_q^{AB} = (-1)^{(q+1)} T^{(q+1)} \int_{t_p}^{t_{p+1}} \binom{-s}{q+1} y^{(q+2)}(\xi)\, dt$$

since $\binom{-s}{q+1}$ is a constant sign in the interval $t_p < t < t_{p+1}$ and $y^{(q+2)}(\xi)$ is a continuous function of t, apply the second mean value theorem of the integral calculus

$$R_q^{AB} = (-1)^{(q+1)} T^{(q+1)} y^{(q+2)}(\xi') \int_{t_p}^{t_{p+1}} \binom{-s}{q+1} dt$$

where $t_{p-q} < \xi' < t_{p+1}$. By definition of $r_{q+1}$, this may be written

$$R_q^{AB} = T^{(q+2)}\, y^{(q+2)}(\xi')\, r_{q+1} \tag{3.73}$$

This is the desired expression for the remainder of the Adams-Bashforth formula.

B. Adams-Moulton Method.

This method uses the form

$$y_p - y_{p-1} = \int_{t_{p-1}}^{t_p} p(t)\, dt = T \sum_{m=0}^{q} r_m^* \nabla^m f \tag{3.74}$$

where

$$r_m^* = (-1)^m \frac{1}{T} \int_{t_{p-1}}^{t_p} \binom{-s}{m} dt = (-1)^m \frac{1}{T} \int_{-1}^{0} \binom{-s}{m} ds \tag{3.75}$$

the generating function of the coefficients is determined as follows:

$$G^*(k) = \sum_{m=0} r_m^* k^m = -\frac{k}{\log(1-k)} \tag{3.76}$$

or

$$-\frac{\log(1-k)}{k}\, G^*(k) = 1 \tag{3.77}$$

Expand the left-hand side of Eq. (3.77) to power series

$$\left(1 + \frac{1}{2} k + \frac{1}{3} k + \cdots\right)\left(r_o^* + r_1^* k + r_2^* k + \cdots\right) = 1 \qquad (3.78)$$

It follows that

$$r_m^* + \frac{1}{2} r_{m-1}^* + \frac{1}{3} r_{m-2}^* + \cdots + \frac{1}{m+1} r_o^* = \begin{cases} 1, & m=0 \\ 0, & m=1,2,3,\cdots \end{cases} \qquad (3.79)$$

The numerical values of $r_m^*$ are easily found from this recurrence relation. Some values of $r_m$ calculated from Eq. (3.79) are

| m | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $r_m^*$ | 1 | $-\frac{1}{2}$ | $-\frac{1}{12}$ | $-\frac{1}{24}$ | $-\frac{19}{720}$ | $-\frac{3}{160}$ | $-\frac{863}{60480}$ |

Since $y_p$ occurs as an argument in $f_p = f(t_p, y_p)$in the right hand side term of Eq.(3.74), it will not be possible to solve this equation explicitly. A better approach to this solution is by means of an iterative procedure.

Assuming an approximation of a solution of Eq.(3.74) has been obtained to be $y_p^{(o)}$, calculate $f^{(o)} = f(t_p, y_p^{(o)})$ and form the difference

$\nabla f_p^{(o)} = f_p^{(o)} - f_{p-1}$, $\nabla^2 f_p^{(o)} = \nabla f_p^{(o)} - \nabla f_{p-1}$, $\cdots$ . A better approximation is then obtained from

$$y_p^{(1)} = y_{p-1} + T \sum_{m=0}^{q} r_m \nabla^m f_p^{(o)} \tag{3.80}$$

Calculating $f_p^{(1)} = f(t_p, y_p^{(1)})$ and re-evaluating the differences, a still better value $y_p^{(2)}$ is

$$y_p^{(2)} = y_{p-1} + T \sum_{m=0}^{q} r_m^* \nabla^m f_p^{(1)} \tag{3.81}$$

Generally a sequence $y^{(r)}$ $(r = 0,1,2,3\cdots)$ of approximations is obtained recursively from the relation

$$y_p^{(r)} = y_{p-1} + T \sum_{m=0}^{q} r_m^* \nabla^m f_p^{(r-1)} \tag{3.82}$$

Since

$$\nabla^m f_p^{(i)} = \nabla^{m-1} f_p^{(i)} - \nabla^{m-1} f_p , \qquad \begin{array}{l}(m = 1,2\cdots)\\(i = 1,2\cdots)\end{array} \tag{3.83}$$

it follows that

$$\nabla^m f_p^{(r)} - \nabla^m f_p^{(r-1)} = f_p^{(r)} - f_p^{(r-1)} \tag{3.84}$$

Thus

$$y_p^{(r+1)} - y_p^{(r)} = T \sum_{m=0}^{q} r_m^* \left( \nabla^m f_p^{(r)} - \nabla^m f_p^{(r-1)} \right) \tag{3.85}$$

$$= T \sum_{m=0}^{q} r_m^* \left( f_p^{(r)} - f_p^{(r-1)} \right)$$

The Lipschitz condition yields

$$\left| f_p^{(r)} - f_p^{(r-1)} \right| \leqslant L \left| y_p^r - y_p^{r-1} \right| \tag{3.86}$$

Equation (3.85) becomes

$$y_p^{(r+1)} - y_p^{(r)} \leqslant T \sum_{m=0}^{q} r_m^* L \left| y_p^{(r)} - y_p^{(r-1)} \right| \tag{3.87}$$

or

$$y_p^{(r+1)} - y_p^{(r)} \leqslant (TL\ A)^r \left| y_p^{(1)} - y_p^{(o)} \right| \tag{3.88}$$

where

$$A = \sum_{m=0}^{q} r_m^*$$

The solution of $y_p$ of Eq.(3.74) is now obtained by summing terms of the series

$$y_p^* = y_p^{(o)} + (y_p^{(1)} - y_p^{(o)}) + (y_p^{(o)} - y_p^{(1)}) + \cdots$$

$$+ (y_p^{(n)} - y_p^{(n-1)}) + \cdots \qquad (3.89)$$

The series on the right-hand side of Eq. (3.89) will converge absolutely provided $0 \leq |TLA| < 1$. In this case the solution $y_p^*$ will exist and be unique.

The local truncation error is

$$R_q^{AM} = T^{(q+2)} \, y^{(q+2)} \, (\xi) \, r_{q+1}^* \qquad (3.90)$$

where

$$t_{p-q} < \xi < t_p.$$

C. Milne's Method

Milne's method requires predictor and corrector formulas. The predictor formula is of the form

$$y_{n+1} - y_{n-3} = \int_{t_{n-3}}^{t_{n+1}} f(t, y)\, dt \doteq \frac{4T}{3} (2f_{n-2} - f_{n-1} + 2f_n) \quad (3.91)$$

and the corrector is of the form

$$y_{n+1} - y_{n-1} = \int_{t_{n-1}}^{t_{n+1}} f(t, y)\, dt \doteq \frac{T}{3} (f_{n-1} + 4f_n + f_{n+1}) \qquad (3.92)$$

The predictor formula has truncation error

$$\text{Truncation error} = \frac{14}{45} T^5 y^{(5)} (\xi_1) \qquad (3.93)$$

where

$$t_n < \xi_1 < t_{n+1}$$

The corrector formula has truncation error

$$\text{Truncation error} = \frac{T^5}{90} y^{(5)} (\xi_2) \qquad (3.94)$$

where

$$t_{n-1} < \xi_2 < t_{n+1}$$

The procedure is as follows:

Step 1; Takes T small enough to insure that the remainder term ininvolving is small in the predictor formula, then find out $y_{n+1}$ from this formula.

Step 2; Obtain a first approximation to $y'_{n+1}$ by substituting the value of $y_{n+1}$ obtained from step 1 into the following equation

$$y' = f(t, y) \qquad\qquad y = y_0 \ \text{ at } \ t = t_0 .$$

Step 3; Obtain a better approximation of $y_{n+1}$ by means of the corrector formula Eq. (3.94).

After repeating step 2 and step 3, the value of $y_{n+1}$ becomes very accurate. When values of $y_{n+1}$ and $f_{n+1}$ have negligible error, the next pair of values $y_{n+2}$ and $f_{n+2}$ may be obtained by a repetition of the process.

Let $y_{n+1}^{(o)}$ be the value of $y_{n+1}$ obtained from step 1, $y_{n+1}^{(1)}$ be the value of $y_{n+1}$ introduced by the corrector at first time, and $y_{n+1}^{(m)}$ be the value of $y_{n+1}$ after introducing the corrector m times, then

$$f_{n+1}^{(1)} - f_{n+1}^{(o)} = k(y_{n+1}^{(1)} - y_{n+1}^{(o)}) \tag{3.95}$$

where

$$k = \frac{f_{n+1}^{(1)} - f_{n+1}^{(o)}}{y_{n+1}^{(1)} - y_{n+1}^{(o)}} \tag{3.96}$$

and

$$f_{n+1}^{(o)} = f(t_{n+1}, y_{n+1}^{(o)}), \qquad f_{n+1}^{(1)} = f(t_{n+1}, y_{n+1}^{(1)}) \tag{3.97}$$

If $f(t, y)$ possesses a continuous first derivative $f_y(t, y)$ with respect to y, then by the mean-value theorem

$$k = f(t_{n+1}, \eta_{n+1}) \tag{3.98}$$

where

$\eta_{n+1}$ lies between $y_{n+1}^{(o)}$ and $y_{n+1}^{(1)}$

Equation (3.92) together with Equation (3.95) yield

$$y_{n+1}^{(2)} - y_{n+1}^{(1)} = \frac{T}{3}\left[f_{n+1}^{(1)} - f_{n+1}^{(o)}\right] = \left[k(y_{n+1}^{(1)} - y_{n+1}^{(o)})\right] \qquad (3.99)$$

By the same method

$$y_{n+1}^{(3)} - y_{n+1}^{(2)} = (\frac{T}{3} k)(\frac{T}{3} k)(y_{n+1}^{(1)} - y_{n+1}^{(o)})$$

where the change of k is negligible.

Proceeding in this fashion a sequence

$$y_{n+1}^{(r)} - y_{n+1}^{(r-1)} = (\frac{T}{3} k)^{(r-1)} (y_{n+1}^{(1)} - y_{n+1}^{(o)})$$

is obtained.

Thus

$$y_{n+1}^{*} = y_{n+1}^{(o)} + (y_{n+1}^{(1)} - y_{n+1}^{(o)}) + (y_{n+1}^{(2)} - y_{n+1}^{(1)}) + \cdots$$

$$= y_{n+1}^{(o)} + (y_{n+1}^{(1)} - y_{n+1}^{(o)})\left[1 + \frac{T}{3} k + (\frac{T}{3} k)^2 + \cdots\right] \qquad (3.100)$$

Provided that T is sufficiently small, the value of $\frac{T}{3}k$ can be chosen so that $0 \leq |\frac{T}{3} k| < 1$ to insure convergence of the power series on the right-hand side of Eq.(3.100).

If iterations are finite, the value of $\tilde{y}_{n+1}$ obtained will differ from $y^*_{n+1}$ by

$$y^*_{n+1} - \tilde{y}_{n+1} = (y^{(1)}_{n+1} - y^{(o)}_{n+1})(\frac{T}{3} k)^{n+1} \left( \frac{3}{3 - Tk} \right) \tag{3.101}$$

where n is the number of steps in the finite step process.

Let the true values of the $\tilde{y}_i$ and $\tilde{y}'_i$ be $z_i$ and $z'_i$ respectively. Define errors in the $\tilde{y}_i$ and $\tilde{y}'_i$ by the equations

$$\begin{aligned} z_i &= \tilde{y}_i + e_i \\ z'_i &= \tilde{y}'_i + e'_i \end{aligned} \tag{3.102}$$

From the differential equation (3.64)

$$e'_i = z'_i - \tilde{y}'_i = f(t_i, z_i) - f(t_i, \tilde{y}_i) = k_i (z_i - \tilde{y}_i) \tag{3.103}$$

where $\quad k_i = f(t_i, \eta_i) \quad$ and $\quad y_i < \eta_i < z_i$

The $z_i$ and $z'_i$ satisfy the equation

$$z_{n+1} - z_{n-1} = \frac{T}{3}(z'_{n-1} + 4z'_n + z'_{n+1}) - T^5 z^{(5)}(\xi_{n+1})/90$$

where

$$t_n < \xi_{n+1} < t_{n+1} \tag{3.104}$$

Subtracting Equation (3.94) from Equation (3.104) results in

$$e_{n+1} - e_{n-1} = \frac{T}{3}(e'_{n-1} + 4e'_n + e'_{n+1}) - T^5 A_{n+1} \tag{3.105}$$

where

$$z^{(5)}(\xi_{n+1})/90 = A_{n+1} \qquad t_n < \xi < t_{n+1}$$

On making use of Equation (3.103), this difference equation may be written

$$(1 - \frac{1}{3}Tk_{n+1})\, e_{n+1} = (\frac{4}{3}Tk_n)\, e_n + (1 + \frac{1}{3}Tk_{n-1})\, e_{n-1} - T^5 A_{n+1} \tag{3.106}$$

Assume $\begin{cases} k_i = k, \\ A_i = A \end{cases}$ for $i = 1, 2, \cdots n$, over a sufficiently restricted range.

Solving the difference equation yields

$$e_n = C_1\lambda_1^n + C_2\lambda_2^n + \frac{T^5 A}{(6\frac{1}{3}Tk)} \tag{3.107}$$

where

$$\lambda_1 = \frac{\frac{2}{3}\,Tk + \sqrt{1 + \frac{1}{3}\,T^2k^2}}{1 - \frac{1}{3}\,Tk} \cong 1 + Tk \cong e^{3\theta} \tag{3.108}$$

$$\lambda_2 = \frac{\frac{2}{3}\,Tk - \sqrt{1 + \frac{1}{3}\,T^2k^2}}{1 - \frac{1}{3}\,Tk} \cong -\left(1 - \frac{Tk}{3}\right) \cong -\,e^{\theta} \tag{3.109}$$

$$\theta = \frac{1}{3}\,Tk \,.$$

Expressing the constants $c_1$ and $c_2$ in terms of $e_1$ and $e_o$ and setting $e_o$ to be zero, Equation (3.107) becomes

$$e_n \cong \frac{e_1}{2}\left(\lambda_1^n - \lambda_2^n\right) + \frac{T^5A}{6\theta}\left(1 - \lambda_1^n - \frac{3}{2}\,\theta\lambda_2^n\right) \tag{3.110}$$

As an approximation setting $\epsilon_1=0$, Eq.(3.110) can be replaced by

$$e_n \cong \frac{T^5A}{2k}\left[1 - e^{k(t_n - t_o)} + (-1)^{n+1}\,\frac{1}{2}\,Tke^{-\frac{2}{3}k(t_n - t_o)}\right] \tag{3.111}$$

If k is positive, the last term decreases as $t_n$ increases. In this case the error is always of the same sign and is multiplied by 10 each time $t_n$ increases by 2.3/k. This case corresponds to a differential equation in which the plots of the solutions for various boundary conditions, in the neighborhood of the desired solution, diverge to the right, and the accumulated relative error may decrease even though the accumulated error increases exponentially. Milne's method is thus applicable.

If k is negative, the second term vanishes, and the error is alternately positive and negative. This case corresponds to a differential equation in which the plots of the solutions for various boundary conditions converge to the right. The accumulated relative error increases rapidly without bound. This shows that Milne's method is unstable, and cannot be used.

Milne's method has two virtues: It supplies a running check that the method and interval size are suitable and that the computation is locally accurate enough to warrant going on. Moreover, only two evaluations of the derivatives per step forward are required.

D. Hamming Method

Milne's method is always unstable, because it has two roots with modulus one in the characteristic equation of its corrector foumula. Hamming has removed unstability by eliminating one of these roots and thus modified the corrector formula without changing the predictor formula.

D.I. Third Order formula

The third order formula is of the form

$$y_{n+1} = ay_n + by_{n-1} + T(cy'_{n+1} + dy'_n + ey'_{n-1}) \qquad (3.112)$$

Expanding both sides with respect to $y_n$ by Taylor series expansion and requiring exact fit for $1, T, T^2, T^3$, the coefficients a,b,c,d, and e can be determined to be

$$\begin{aligned} a &= -4 + 12c \qquad & d &= 4 - 8c \\ b &= 5 - 12c \qquad & e &= 2 - 5c \\ c &= c \end{aligned} \tag{3.113}$$

The truncation error has the value

$$\text{Truncation error} = \frac{1 - 3c}{6} T^4 y^{(4)}$$

The characteristic equation is

$$\rho^2 + (4 - 12c)\rho - (5 + 12c) = 0 \tag{3.114}$$

Solving this equation yields

$$\rho_1 = 1$$
$$\rho_2 = -5 + 12c$$

Stability requires

$$0 \leq |\rho_2| < 1$$

or $$\frac{1}{2} > C > \frac{1}{3}$$

Setting $c = 5/12$, Equation (3.112) becomes

$$y_{n+1} = y_n + \frac{T}{12}(5y'_{n+1} + 8y'_n - y'_{n+1}) \qquad (3.115)$$

D. II. Fourth Order Formulas

This method generalizes Milne's corrector formula to the form

$$y_{n+1} = ay_n + by_{n-1} + cy_{n-2} + T\left[dy'_{n+1} + ey'_n + fy'_{n-1}\right] \qquad (3.116)$$

and then stabilizes this formula by choosing suitable values of the coefficients to minimize one of the characteristic roots with modulus one in Milne's formula.

Expanding Equation (3.116) with respect $y_n$ by Taylor series expansion, and requiring exact fit for 1, $T$, $T^2$, $T^3$, $T^4$, yield

$$\begin{aligned} a &= \frac{27(1-b)}{24} \qquad & d &= \frac{9-b}{24} \\ b &= b & e &= \frac{18+14b}{24} \\ c &= \frac{-3(1-b)}{24} & f &= \frac{-9+17b}{24} \end{aligned} \qquad (3.117)$$

In this case the truncation error is found to be

$$kT^5 y^{(5)} = \frac{-9 + 5b}{360} T^5 y^{(5)}. \tag{3.118}$$

The characteristic equation of Equation (3.116), using Equation (3.117), is

$$8\rho^3 - 9(1 - b)\rho^2 - 8b\rho + (1 - b) = 0. \tag{3.119}$$

The root loci of $\rho$ with respect to b are shown in Fig. 2. For stable operation b should lie in the range $-0.6 < b < 1$ to insure that no characteristic root exceeds one and that the root with modulus one is simple. Milne's formula is a special case of Equation (3.116) with b=1. It is easily seen from Fig. 2 that in this case there are two characteristic roots with modulus one.

For the case b=0, a stable predictor-corrector formula can be constructed as follows:

predictor $\quad p_{n+1} = y_{n-3} + \frac{4T}{3}(2y'_n - y'_{n-1} + 2y'_{n-2})$

modify $\quad m_{n+1} = p_{n+1} - \frac{112}{121}(p_n - c_n)$

corrector $\quad c_{n+1} = \frac{1}{8}(9y_n - y_{n-2}) + 3T(f'_{n+1} - 2y'_n - y'_{n-1}) \qquad (3.120)$

final value $\quad y_{n+1} = c_{n+1} + \frac{9}{121}(p_{n+1} - c_{n+1}).$

Truncation error of corrector formula in this case is

$$\text{truncation error} = -\frac{1}{40} T^5 y^{(5)} .$$

as compared with that of Milne's corrector formula, there is a 125% increment.

In Hamming methods predictor formulas are always the same as Milne's formula because if the predictor is generalized by the same method used above, it would be unstable.

From the above discussion it is easily seen that to gain stability in a predictor-corrector method one must lose some accuracy. This loss in accuracy can be compensated by shortening the interval of integration.

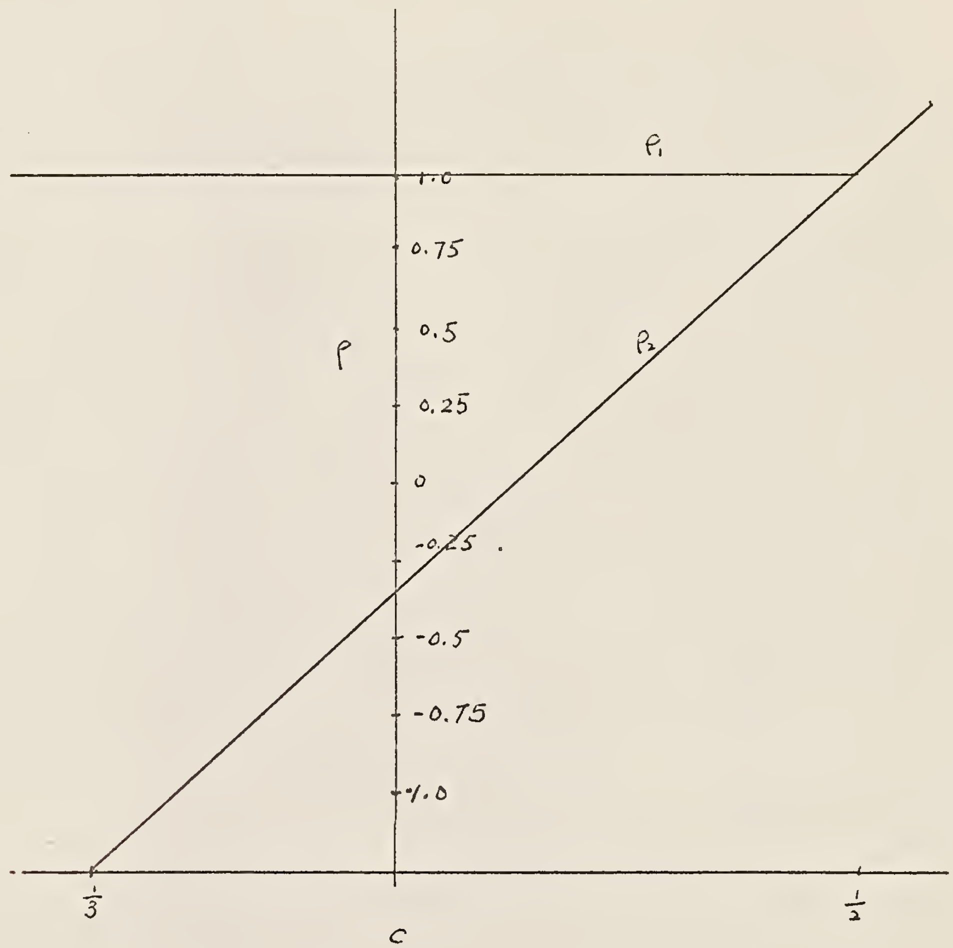


Fig. 1. Root Loci of Eq. (3.114)

$$\rho^2 + (4 - 12c)\rho - (5 + 12c) = 0$$

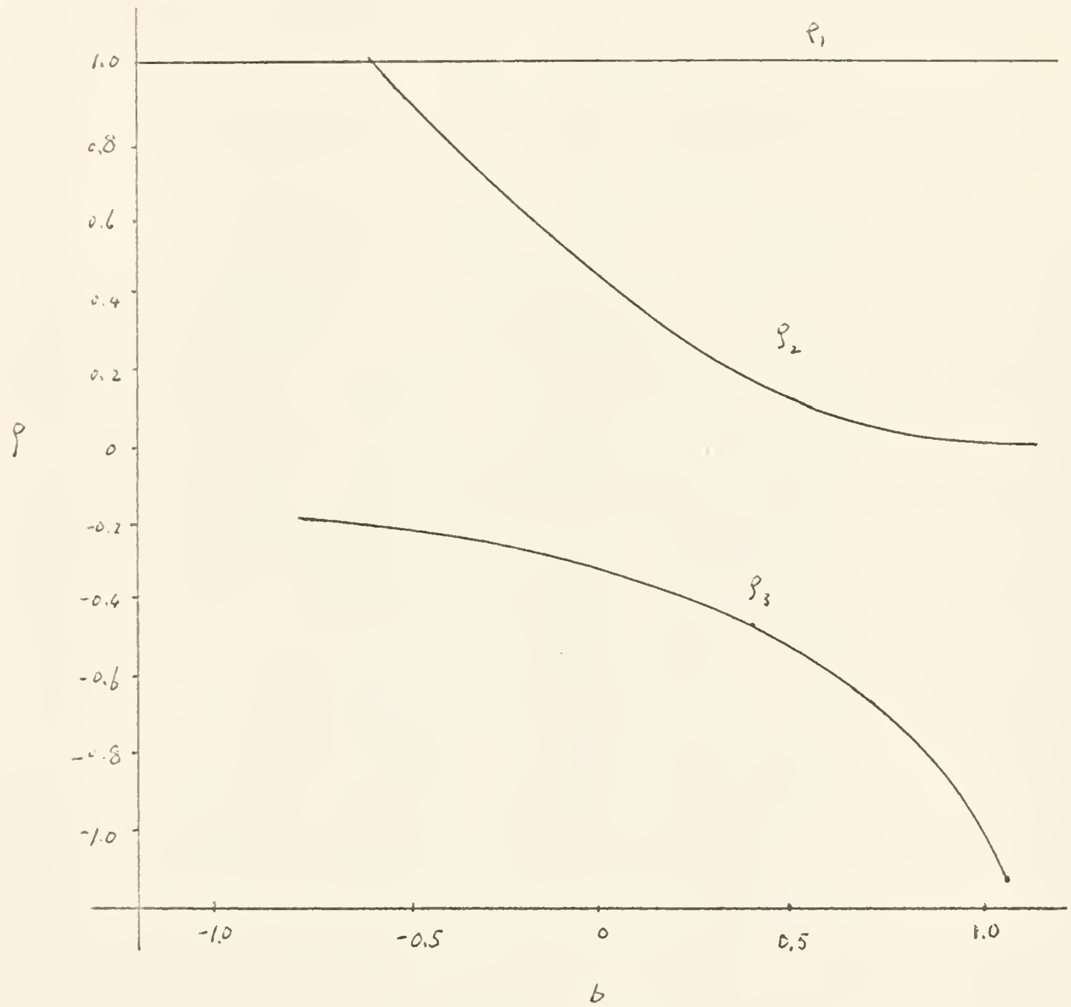


Fig. 2. Root Loci of Eq. (3.119)

$$8\rho^3 - 9(1 - b)\rho^2 - 8b\rho + (1 - b) = 0$$

C. Comparison Between One-Step Methods and Multiple Step Methods

One-step methods are useful for obtaining the first few of the solution of a differential equation, but involve too much labor to be used for obtaining a numerical solution over an extended range. However, multiple step methods require special starting procedures which are furnished by one-step methods.

IV. Numerical Transform Techniques.

Part of the solution of a linear differential equation with constant coefficients has the form of a convolution. The digital computer must approximate convolution on a denumerable set of equally spaced points. The tranpezoidal approximation is the commonly used quadrature method.

Trapezoidal approximation proceeds as follows. Let $g(x)$ be a function defined and continuous in a closed interval containning $x=0$, and let its first four derivatives be continuous in the same interval, the Taylor series representation of the function with remainder is

$$g(x) = g(0) + xg^{(1)}(0) + \frac{x^2}{2!} g^{(2)}(0) + \frac{x^3}{3!} g^{(3)}(0)$$

$$+ \int_o^x \frac{(x - \mu)^3}{3!} g^{(o)}(\mu)\, d\mu \qquad (4.1)$$

Integrating the given function results in

$$\frac{1}{h}\int_0^h g(x)\,dx = \frac{(g_o + g_1)}{2} - \frac{h^2}{12}\left(\frac{g_o^{(2)}}{2} + \frac{g_1^{(2)}}{2}\right) + R(h) \tag{4.2}$$

The remainder R(h) is given by

$$R(h) = \frac{h^4}{4!}\int_0^1 (\lambda - 2\lambda^3 + \lambda^4)\, g^{(4)}(\lambda h)\, d\lambda \tag{4.3}$$

$$\leq \frac{h^4}{120}\, g^{(4)}(\theta h) \qquad 0 < \theta < 1$$

A convolution can be approximated by dividing its interval into a finite number of subintervals each with same width T as follow:

$$\int_0^t f(\tau)\, g(t-\tau)\, d\tau = \sum_{k=0}^{n-1} \int_{kT}^{(k+1)T} f(\tau)\, g(t-\tau)\, d\tau \tag{4.4}$$

Trapezoidal approximation yields

$$\int_0^t f(\tau)\, g(t-\tau)\, d\tau = \frac{T}{2}\left(\sum_{k=0}^{n-1} f_k\, g_{n-k} + f_{k+1}\, g_{n-k-1}\right) \tag{4.5}$$

Taking z-transform of both sides of Equation (4.4) results in

$$Z[\bar{f}\bar{g}] = T\,[Zf][Z\bar{g}] - \frac{T}{2}[f_o Z\bar{g} + g_o Z\bar{f}] \tag{4.6}$$

There are many ways of utilizing trapezoidal convolution to solve a given differential equation. Each of these is called a program. Three classes of programs are discernible; The multiple integration substitution program, the Tustin integrator program, and the single integrator program. Differences among them are at the transitions from integration to multiple integration.

## A. Tustin Program

This method solves an n-th order differential equation by the state-space spproach. A Tustin-like program is demonstrated as follows:

Let an n-th order differential equation of the form

$$D^n y = x(t) \tag{4.7}$$

Define the row vectors

$$w' = (y, y', \cdots y^{(n-1)})$$

$$v' = (D, 0, \cdots x(t)) \tag{4.8}$$

Then

$$Dw + \begin{pmatrix} 0 & -I \\ 0 & 0 \end{pmatrix} w = v \tag{4.9}$$

is equivalent to an nth order differential equation. Take the Laplace transform and divided by s to obtain

$$\bar{w} + \frac{1}{s} A\bar{w} = \frac{1}{s} v + \frac{1}{s} w_o \tag{4.10}$$

Employing the sampling operation and trapezoidal integration yields

$$Z\bar{w} + AZ\bar{w} - Aw_o = \alpha Z\bar{v} - \beta v_o + \frac{2}{T} \beta w_o \tag{4.11}$$

where

$$\alpha = \frac{T}{2} \frac{1 + z}{1 - z}, \qquad \beta = \frac{T}{2(1 - z)}, \qquad \underset{(nxn)}{A} = \begin{pmatrix} 0 & -I \\ 0 & 0 \end{pmatrix}$$

Since $A^n = 0$, it follows that

$$(I + \alpha A^n) = I \qquad \text{or} \qquad (I + \alpha A) \left[ I - (\alpha A) + (\alpha A)^2 + \cdots (-\alpha A)^{n-1} \right] = I \tag{4.12}$$

Then the desired solution can be written as

$$Z\bar{y} = \alpha^n Z\bar{x} - \beta\alpha^{n-1} x_o + \frac{1}{1 - z} y_o^{(n-1)} + \frac{T^2 z}{(1 - z)^2} \sum_{k=2}^{n} \alpha^{k-2} y_o^{(n-k)} \qquad \text{for } n = 1,2,3\cdots \tag{4.13}$$

The above equation is the Tustin program, if all initial conditions are zero.

This program's great advantage is that it finds the solution and its first (n-1) derivatives in one sweep.

B. Single-integrator program.

Halijak has derived a single-integrator program which uses a sequence of ascending order differential equations derived from the given differential equation to solve the same equation. Consider the differential equation.

$$\left[D^{n} + a_1 D^{n-1} + \cdots + a_{n-1} D + a_n\right] y(t) = x(t), \tag{4.14}$$

First, set up the differential equation

$$(D + a_1)\, y_1 (t) = 1 \qquad\qquad y_1 (0) = 0 \tag{4.15}$$

The Laplace transform of this differential equation yields after division by s

$$\left(1 + \frac{a_1}{s}\right) \bar{y}_1 (s) = \frac{1}{s^2} \tag{4.16}$$

Taking Z-transform of both sides and employing approximate trapezoidal convolution, yields

$$Z\bar{y}_1 = \frac{2Tz}{\left[1 - z\right]\left[(2 + aT) - (2 - aT)\, z\right]} \doteq Z\left(\frac{1}{s(s + a)}\right) \tag{4.17}$$

Set up another differential equation of the form;

$$(D^2 + a_1 D + a_2)\, y_2\,(t) = 1, \qquad y_2(0) = y_2(0) = 0 \tag{4.18}$$

taking Laplace transform of both sides and dividing by s(s+a) yield

$$(1 + \frac{a^2}{s(s+a)})\, \bar{y}(s) = \frac{1}{s^2(s+a)} \tag{4.19}$$

taking Z-transform yields

$$Z\bar{y}(s) + Ta_2 \left[ Z(\frac{1}{s(s+a_1)}) \right]\left[ Z\bar{y}(s) \right] = \left[ Z(\frac{1}{s}) \right]\left[ Z(\frac{1}{s(s+a)}) \right] \tag{4.20}$$

Substituting equation (4.17) into equation (4.20) yields

$$Zy_2(s) = \frac{T}{2}\,\frac{(1+z)}{(1-z)} \cdot \frac{2Tz}{(2+a_1T) - (4 - 2a_2T^2)z + (2 - a_1T)z^2}$$

$$\doteq Z(\frac{1}{s(s^2 + a_1 s + a_2)}) \tag{4.21}$$

Proceeding in this manner, let a nth order differential equation be of the form

$$(D^{n-1} + a_1 D^{n-2} + \cdots + a_{n-1})\, y_{n-1}\,(t) = 1$$

$$y_{n-1}(0) = y'_{n-1}(0) = \cdots = y_{n-1}^{(n-2)}(0) = 0 \tag{4.22}$$

A recurrence relation can be constructed to be

$$Z\bar{y}_i(s) = \frac{T}{2}\left(\frac{1+z}{1-z}\right) Z\bar{y}_{i-1}(s) \Big/ \left[1 + a_i T Z\bar{y}_{i-1}(s)\right] \tag{4.23}$$

$$2 \leqslant i < n$$

for $i=n$,

$$Z\bar{y}_n(s)\left[1 + a_n T(Z(\bar{y}_{n-1}(s))\right] = T\left[Zx(s)\right]\left[Zy_{n-1}(s)\right] - \frac{T}{2}\left\{x_o Z\bar{y}_{n-1}(s)\right\} \tag{4.24}$$

Initial conditions other than zero can be introduced at the last step.

C. Multiple-Integrator Substitution Program.

The multiple integrator substitution program casts the Laplace transform of the given differential equation into inverse powers of s by a suitable division and substitutes for them definite functions of z defined to be $e^{-Ts}$.

C. I. Halijak's Integrator Substitution Program.

This method uses trapezoidal convolution to generate z-transform of

$(1/s^n)$ and $(1/s^n)f$. The procedure is as follows:

for $n = 1$, $\quad Z(1/s) = 1/(1 - z)$,

for $n = 2$, $\quad Z(1/s^2) = Z(1/s)Z(1/s)T - TZ(1/s)$,

$$= Tz/(1 - z)^2, \tag{4.25}$$

for $n \geqq 2$, $\quad Z(1/s^n) = Z(1/s^{n-1})Z(1/s)T - Z(1/s^n)T/2$,

$$= Z\left(\frac{1}{s^{n-1}}\right) \frac{T}{2} \frac{(1 + z)}{(1 - z)},$$

$$= Z\left(\frac{1}{s^2}\right) \frac{T}{2} \left[\frac{(1 + z)}{(1 - z)}\right]^{n-2}$$

Substituting Equation (4.25) into Equation (4.6) yields

$$Z\left(\frac{1}{s^n} f\right) \doteq \left[T^2 z\Big/(1 - z)^2\right]\left[\frac{T(1 + z)}{2(1 - z)}\right]^{n-2} (Z\bar{f} - 0.5f_o) \tag{4.26}$$

This method yields moderate accuracy approximations and is the basis of physically small computers.

# SUMMARY

The result of an approximate computation of the solution of differential equation is affected by errors. These errors arise from different causes and affect final result in different ways.

Three types of errors that are most important are truncation error, round-off error, and accumulated error. Round-off error usually affects the last retained digit of the decimal representation; its effect can be minimized by retaining additional digits. Truncation error is due to discarded terms in an infinite series. Sometimes the remainder exceeds the sum of terms retained, thus making the calculated result meaningless. Therefore, an estimate of truncation error is essential. The importance of accumulated errors depends on rate of accumulation. If the accumulation error is unbounded, the solution becomes meaningless.

The search for Runge-Kutta type formulas is important. It seems that the method given in this report in deriving coefficients of Runge-Kutta formulas can be applied to investigate six and higher order formulas.

Trapezoidal convolution using the integral of the first two terms of Taylor series coincides with the average of right and left Riemann sum approximations. The truncation error is of order $T^2$. Improved trapezoidal convolution using higher order Taylor Series terms seems worthy of further investigation.

Accuracy is not the only consideration for evaluating a computation process. The step size affects the solution accuracy as well as the cost; the more accurate solution is generally the more expensive. Frequently it is desired to have a solution within a certain accuracy with most economical computation. The selection of optimal size depends on the method used and the problem solved.

A number of numerical transform techniques have been developed in the Z-transform language. Many of these exhibit difficulties for the solution of differential equation with non-zero initial conditions. The programs generated by Halijak have shown that proper reintroduction of initial conditions is required for better approximation.

# APPENDIX 1

## Error Formulas for an Interpolating Polynomial

Let the function $z(x)$ be defined on an interval containing the $q+1$ distinct points $x_0, x_1, \cdots x_q$. It is well known that among all polynomials in x of degree not exceeding q there exists exactly one polynomial $P(x)$ which satisfies the relations

$$P(x_i) = z(x_i) \qquad i = 0,1,2,\cdots q \qquad (I - 1)$$

The uniqueness of this interpolating polynomial follows from the fact that the difference of any two such polynomials is a polynomial of degree $\leq q$ which has $q+1$ zeros and therefore vanishes identically. Existence can be proved by exhibiting the polynomial explicitly, in the form

$$P(x) = L(x) \sum_{i=0}^{q} \frac{z(x_i)}{L'(x_i)\ (x - x_i)} \qquad (I - 2)$$

where $\quad L(x) = (x - x_0)(x - x_1) \cdots (x - x_q)$

The error committed in this approximation can be estimated from the following Lemmas.

Lemma 1. Let $z(x)$ have a continuous derivative of order $q+1$ in J. Then for every point x in J there exists a point $\xi$ in the smallest interval I containing both x and the points $x_i$ $(i=1,2,\cdots q)$ such that

$$z(x) - P(x) = \frac{1}{(q + 1)!} L(x) z^{(q+1)}(\xi) \qquad (I - 3)$$

Lemma II. Let $z(x)$ satisfies the same hypothesis as in Lemma 1. Then for every $x_k (k=0,1,2,\cdots q)$ there exists a number $\xi$ such that

$$z'(x_k) - p'(x_k) = \frac{1}{(q + 1)!} z^{(q+1)}(\xi) L'(x_k) \qquad (I - 4)$$

REFERENCES


1. Tustin A.
   A method of analysing the behavior of linear systems in terms of time series, Journal IEE, proceeding at the Convention on Automatic Regulators and Servomechanisms, vol, 94, part II-A, pp 130ff, May, 1947.

2. Madwed A.
   Number Series Method of solving linear and non-linear Differential Equations, Rep. No. 6445-T-26, Instrumentation Lab., MIT; April, 1950.

3. Boxer R. and Thaler S.
   A Simplified Method of Solving Linear and Nonlinear Systems, Proceedings IRE, vol. 44, no. 1, pp. 84-101, Jan. 1956.

4. Armarel S., Lambert L. and Millman J.
   A Transformation Calculus for Obtaining Approximation Solutions to Linear Integro-Differential Equations with Constant Coefficients, Technical Report T-7/c, Columbia University Electronics Research Laboratory, August, 1955.

5. Halijak C. A.
   Digital Approximation of the Solutions of Differential Equations Using trapexoidal Convolution, ITM-64, Bendix Systems Division, The Bendix Corporation, Ann Arbor, Michigan, August 26, 1960.

6. Halijak C. A.
   Algebraic Methods in Numerical Analysis of Ordinary Differential Equations, Special Report No. 37, Kansas State University Bulletin.

7. Henrici P.
   Discrete Variable Methods in Ordinary Differential Equations, John Wiley & Sons, Inc., New York. 1962

8. Kunz K. S.
   Numerical analysis, McGraw-Hill Book Company, Inc., New York. 1942

9. Boxer R.
   A Note on Numerical Transform Calculus, Proceeding IRE, vol. 45, pp. 1401-1406; October, 1957.

10. Kwan R. K.
    On the Analysis of Closed-Loop systems by Means of Digital Computing Techniques. Master Thesis, McGill University, Montreal, Canada. 1963.

11. Hamming R. W.
Numerical methods for scientists and engineers, McGraw-Hill book Company, 1962.

# ACKNOWLEDGMENT

The author wishes to express his appreciation to Dr. Charles A. Halijak and Dr. Floyd W. Harris of the Department of Electrical Engineering for their tutelage and guidance which have implicitly contributed to this report.

# SURVEY OF DIGITAL SOLUTION OF DIFFERENTIAL EQUATIONS

by

ING-WEN HWANG

B.S.E.E., National Taiwan University, 1963

---

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1966

There are a number of numerical methods for approximating solution of a differential equation. These methods have two things in common; the calculations are performed with discrete values and on a step-by-step basis. The purpose of this report is to review those methods that are frequently used on digital computers. All these methods are divided into two classes; the classical numerical techniques and the numerical transform techniques. Classical numerical techniques yield two types; one-step methods and multiple step methods. In one-step methods the value of $y_n$ is solved from its previous step. In multiple step methods the value of $y_n$ is solved from its several previous steps $y_{n-1}$, $y_{n-2}$, ... $y_{n-q}$. The virtue of one-step methods is that they are self starting whereas multiple step methods require more than one starting point that they are not self starting. However, multiple step methods are more easier in continuing a solution than one-step methods. The numerical transform techniques were developed recently by Tustin, Madwed, Boxer-Thaler, and Halijak. These methods use z-transform as a language. Three classes are discernable; Tustin program, single integration program, and multiple integratinn substitution program. Differences among them are the transition from integration to multiple integration.

Truncation error and accumulated relative error are important factors that affect availability of a method. It is necessary to estimate these errors so as not to make computation meaningless.