ON OPTIMUM RUNGE-KUTTA METHODS FOR THE NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

by

ROBERT LAURENCE JOHNSTON

B. Comm., University of British Columbia, 1958

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in the department of

MATHEMATICS

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April, 1961

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics

The University of British Columbia,
Vancouver 8, Canada.

Date April 12, 1961

ABSTRACT

After a brief discussion of numerical methods for the solution of the ordinary differential equation $x' = f(t, x)$ the problem of finding optimum methods is considered. The thesis then deals with this problem for the family of Runge-Kutta methods. Criteria for optimum methods are discussed and then the derivation of third-order methods is examined in detail.

The next part of the thesis deals with possible approaches to finding optimum methods. The first approach consists of finding some sort of estimate for f and its derivatives contained in the truncation error T_n. The resulting expression is then dependent on some free parameter or parameters (depending on the order of the method) which are chosen in order to minimize this expression. The second approach assumes the independence of terms in the truncation error and minimizes, in some sense, the coefficients of these terms. Several procedures based on these approaches, are used to predict optimum second-order and third-order methods and comparisons are made with experimental results. While no definite conclusions could be drawn it was seen that one particular procedure gave a good prediction. This result encourages further studies in this area.

I hereby certify that this abstract is satisfactory.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The author wishes to acknowledge the assistance of Dr. T. E. Hull and Dr. C. Froese in connection with this thesis. Special thanks are due Dr. Hull for his interest and encouragement.

Thanks are also due the Computing Centre of the University of British Columbia for the use of their computing facilities without which the evaluation of the various procedures would have been impossible. In this regard, the author wishes to express his appreciation of the generous financial support given by the Defence Research Board of Canada.

CHAPTER I Introduction

The discussions throughout this thesis will be concerned with the numerical solution of the first order ordinary differential equation

(1.1) $$x' = f(t, x(t))$$

with initial condition $x(t_0) = x_0$. The problem of solving a higher order equation such as

(1.2) $$x^{(m)} = f(t, x^{(m-1)}(t), \ldots\ldots, x'(t), x(t))$$

with $x(t_0) = x_0$ can be dealt with by reducing (1.2) to a system of m first-order equations and these can be solved by treating (1.1) as a vector equation and generalizing the numerical procedure to apply in this situation.

We shall define

$$t_n = t_0 + nh$$
$$x_n = x(t_n) .$$

We shall also introduce y_n which satisfies exactly an equation such as

$$y_{n+1} = y_n + hy_n + \frac{h^2}{2!}y_n ,$$

whereas x_n would satisfy

$$x_{n+1} = x_n + hx_n + \frac{h^2}{2!}x_n + T_{n+1}$$

where T_{n+1} is the truncation error. Hence, y_n is an approximation to the true value x_n. Note, however, that y_n does not contain round-off error.

There are two general families of methods for solving (1.1) numerically, namely, multi-step methods and Runge-

Kutta methods. The most general multi-step methods are based on formulae of the form -

$$(1.3) \qquad \sum_{i=0}^{k} \alpha_i \, y_{n-i} = \sum_{j=1}^{l} h^j \left(\sum_{i=0}^{k} \beta_{ij} \, y_{n-i}^{(j)} \right).$$

The most common multi-step methods in use are those with $l=1$ while a few with $l=2$ are also in use. The advantages of these methods over Runge-Kutta methods are (1) there are usually only two evaluations of the function $f(t, x)$ for each step as compared to several evaluations for Runge-Kutta methods which could mean a substantial saving in time if $f(t, x)$ is a complicated function; (2) they have a fairly simple expression for the truncation error, i.e., the truncation error for (1.3) is usually of the form -

$$T = h^s K \, x^{(s)}(\xi) \qquad , \; x_{n-i} < \xi < x_n$$

where $0 < K \ll 1$ and $s > kl$ whereas the truncation error for Runge-Kutta methods (cf. (2.16) below) is much more complicated.

The most general Runge-Kutta method is given by (2.1) and (2.2) below. The advantages of using methods of this type are (1) no special starting values are needed as is the case with multi-step methods when $k > 1$; (2) there is no instability; (3) it is much easier to alter the step size h ; (4) they are explicit methods and, hence, no iterations are necessary whereas multi-step methods with $\beta_{0j} \neq 0$ are implicit and require iterating; (5) they can follow curves with sharp bends better then multi-step methods.

The problem of choosing between multi-step and Runge-Kutta methods is of no concern here but what we do wish to consider is the problem of selecting a best or optimum method. Hull and Newbery (5) have dealt with the question of selecting a best multi-step method. The problem considered by them was to find, for $l=1$ and fixed k, a stable and consistent formula which was the most accurate in terms of size of the propagated error. The question of finding best Runge-Kutta methods is investigated in this thesis.

Preliminary to a determination of optimum methods, one must establish criteria by which different methods can be compared in order to decide which one is the best. One basis for comparison is time taken per step in the calculation. This factor depends, as we shall see, on the number of evaluations of $f(t, x)$ and, hence, we shall consider the class of methods for a given number of function evaluations and try to find the best method in each such class. Another basis for comparison is ease of calculation, e.g., it is easier to multiply by $\frac{1}{3}$ than by, say, .334782. However, this is of little concern with to-day's high speed computers. Similarly, the need for economy of storage as considered by Gill (3) is not important now that computers have large homogeneous memories. One could also require that the magnitude of the parameters for a particular method be restricted in order to keep round-off error down but multiple-precision arithmetic can be used at the expense of some speed in order to compensate for large parameters. The most important basis for comparison is size of the propagated error. With

Runge-Kutta methods the propagated error depends on the choice of formula only through the truncation error and not, for example, on any other factors such as stability as is the case with multi-step methods. Therefore, in using propagated error as a basis for comparing Runge-Kutta methods we can confine our attention to the truncation error.

In Chapter II, we shall investigate the derivation of a specific class of Runge-Kutta methods. Chapter III describes different approaches to finding optimum methods and assesses these approaches with the aid of experimental results for second-order methods. In Chapter IV, experiments with third-order methods are used to evaluate our ideas.

CHAPTER II Summary of Runge-Kutta Methods

The general Runge-Kutta method for solving (1.1) numerically is

$$(2.1) \qquad y_{n+1} = y_n + a_0k_0 + a_1k_1 + \cdots\cdots + a_rk_r,$$

where

$$(2.2) \quad \begin{cases} k_0 = hf(t_n, y_n) \\ k_1 = hf(t_n + m_1h, y_n + s_{10}k_0) \\ k_2 = hf(t_n + m_2h, y_n + s_{21}k_1 + s_{20}k_0) \\ \vdots \\ k_r = hf(t_n + m_rh, y_n + s_{r,r-1}k_{r-1} + \cdots\cdots + s_{r0}k_0). \end{cases}$$

and from the following it always turns out that $m_i = \sum_j s_{ij}$. The parameters a_i, m_j, s_{ij} are obtained by expanding the terms on the right side of (2.1) in Taylor series about the point (t_n, y_n), collecting terms according to powers of h, and equating coefficients to corresponding coefficients in

$$y_{n+1} = y_n + hy_n' + \frac{h^2}{2!}y_n'' + \frac{h^3}{3!}y_n''' + \cdots\cdots\cdots\cdots$$

$$= y_n + hf(t_n, y_n) + \frac{h^2}{2!}f'(t_n, y_n) + \frac{h^3}{3!}f''(t_n, y_n) + \cdots.$$

If we equate coefficients of h^i for $i=0, 1, 2, \ldots, m$ then the truncation error will be $O(h^{m+1})$ and, in fact, will have the general form

$$T_n = h^{m+1}\, G(a_i, m_i, s_{ij}) \qquad i,j = 0, 1, 2, \ldots, k \quad (k \leqslant m).$$

Geometrically, we see from (2.1) that y_{n+1} is obtained by

adding to y_n a weighted average of slopes in the neighbourhood of the point (t_n, x_n).

One of the principal disadvantages of Runge-Kutta methods as mentioned above is the number of evaluations of the function $f(t, x)$. Of course, one way to overcome this is to make the subscript r in (2.1) small and hence, when $f(t, x)$ is complicated, the time taken per step will be comparable to the time required for a multi-step method. However, as will be seen later, we will then be increasing the size of the truncation error by limiting the number of coefficients of h, h^2, ... that can be equated. If, for example, we make $r=2$ then we could only equate terms up to h^2 and the truncation error would be $O(h^3)$ whereas when $r=4$, the truncation term is $O(h^5)$. In general a method with truncation error $= O(h^{p+1})$ will be called a pth - order method. It is evident then that we want to restrict r (the number of function evaluations) and, at the same time, have p as large as possible. The following table shows the minimum r required for the corresponding p -

p	r	T
2	2	$O(h^3)$
3	3	$O(h^4)$
4	4	$O(h^5)$
5	6	$O(h^6)$.

We see that r increases at least as fast as p and, therefore, in order to reduce the truncation error by a factor of h we must pay the price of introducing at least one extra function evaluation. Hence for a given value of p we will deal with the problem of finding the method which has the smallest truncation error. This problem will be considered for the relatively simple second and third-order cases and on the basis of the information gained, we hope to be able to tell how to predict optimum methods of higher order.

Before attempting to derive optimum methods we should consider the derivation of conditions on the parameters a_i, m_j, s_{ij} (by equating powers of h). This is shown below for third-order methods. In order to avoid subscripts we shall change notation used in (2.1) and (2.2). The general third-order method is –

$$y_{n+1} = y_n + ak_0 + bk_1 + ck_2 , \tag{2.3}$$

where

$$\begin{cases} k_0 = hf(t_n, y_n) \\ k_1 = hf(t_n + mh, y_n + mk_0) \\ k_2 = hf(t_n + vh, y_n + (v-r)k_0 + rk_1), \end{cases} \tag{2.4}$$

and so we must determine the six parameters a, b, c, m, v, r. In addition to the parameters we want the truncation error T so we use the exact solution x_n in place of y_n and then (2.3) becomes

$$x_{n+1} = x_n + ak_0 + bk_1 + ck_2 + T_{n+1} , \tag{2.3$'$}$$

where

$$(2.4)^1 \quad \begin{cases} k_0 = hf(t_n, x_n) \\ k_1 = hf(t_n + mh, x_n + mk_0) \\ k_2 = hf(t_n + vh, x_n + (v-r)k_0 + rk_1) . \end{cases}$$

Expanding in Taylor series and equating coefficients of powers of h will give the equations for the parameters while the terms in h^4 and higher powers of h will be the truncation error. Now a Taylor series expansion in two variables is -

$$(2.5) \qquad f(t_n + A, x_n + B) = f + D^{*1}f + \frac{1}{2!}D^{*2}f + \frac{1}{3!}D^{*3}f + \ldots\ldots\ldots\ldots ,$$

where $\quad f = f(t_n, x_n)$

and $\quad D^{*p} = (A\frac{\partial}{\partial t} + B\frac{\partial}{\partial x})^p f$.

Then to expand k_1 put $A=mh$ and $B=mk_0$ and we get -

$$D^*f = mhf_t + mk_0f_x \qquad \text{where } f_x = \frac{\partial f}{\partial x}$$

$$= mhDf ,$$

where $\quad D = \frac{\partial}{\partial t} + f\frac{\partial}{\partial x}$

$$D^{*2}f = m^2h^2f_{tt} + 2m^2hk_0f_{tx} + m^2k_0^2f_{xx}$$

$$= m^2h^2D^2f$$

$$D^{*3}f = m^3h^3D^3f ,$$

and by (2.5) we have -

$$(2.6) \qquad k_1 = h(f + mhDf + \frac{m^2h^2}{2!}D^2f + \frac{m^3h^3}{3!}D^3f + \ldots\ldots\ldots).$$

To expand k_2 put $A=vh$ and $B=vk_0 + r(k_1 - k_0)$ and obtain

$$D^*f = vhDf + mrh^2f_xDf + \frac{m^2rh^3}{2!}f_yD^2f + \ldots\ldots$$

$$D^{*2}f = v^2h^2D^2f + 2mvrh^3DfDf_x + \ldots\ldots$$

$$D^{*3}f = v^3h^3D^3f + \ldots\ldots\ldots ,$$

so that

$$(2.7) \qquad k_2 = h(f + vhDf + \frac{v^2h^2}{2!}D^2f + \frac{v^3h^3}{3!}D^3f + \ldots\ldots\ldots$$

$$+ mvrh^3DfDf_x + \ldots\ldots\ldots).$$

Now we have

$$x_{n+1} = x_n + hx_n' + \frac{h^2}{2!}x_n'' + \frac{h^3}{3!}x_n''' + \frac{h^4}{4!}x_n^{iv} + \ldots\ldots\ldots$$

$$(2.8) \qquad = x_n + hf + \frac{h^2}{2!}Df + \frac{h^3}{3!}(D^2f + f_xDf)$$

$$+ \frac{h^4}{4!}(D^3f + f_xD^2f + f_x^2Df + 3DfDf_x) + \ldots\ldots.$$

Substituting (2.6) and (2.7) in (2.3) and equating coefficients of h^0, h^1, h^2, h^3 to corresponding coefficients in (2.8) we get

$$(2.9) \qquad \begin{cases} a + b + c = 1 \\ bm + cv = \frac{1}{2} \\ bm^2 + cv^2 = \frac{1}{3} \\ crm = \frac{1}{6} \end{cases} .$$

Equating coefficients of h^4 we get the truncation error -

$$(2.10) \qquad T_n = h^4\Big[A_1(D^3f) + A_2(DfDf_x) + A_3(f_xD^2f) + A_4(f_x^2Df)\Big] + O(h^5),$$

where
$$A_1 = \frac{1}{24} - \frac{1}{6}(bm^3 + cv^3)$$
$$A_2 = \frac{1}{8} - crmv$$
$$A_3 = \frac{1}{24} - \frac{1}{2}crm^2$$
$$A_4 = \frac{1}{24} .$$

From (2.9) we see that there are two degrees of freedom. We shall see in a later section that it is convenient to choose m and v as the free parameters and hence (2.10) can be expressed in the form (4.1).

The corresponding results for second-order methods are easily shown to be -

$$(2.11) \qquad \begin{cases} y_{n+1} = y_n + ak_0 + bk_1 \\ \text{where} \quad k_0 = hf(t_n, y_n) \\ k_1 = hf(t_n + mh, y_n + mk_0) . \end{cases}$$

The conditions on the three parameters are -

$$(2.12) \qquad \begin{cases} a + b = 1 \\ bm = \frac{1}{2} . \end{cases}$$

Picking m as the free parameter, the truncation error is -

$$T_n = h^3\left[\left(\tfrac{1}{4}m - \tfrac{1}{6}\right)D^2f + \tfrac{1}{6}f_xDf\right] + O(h^4). \tag{2.13}$$

The results for fourth-order methods are -

where

$$\begin{cases} y_{n+1} = y_n + ak_0 + bk_1 + ck_2 + dk_3 , \\ k_0 = hf(t_n, y_n) \\ k_1 = hf(t_n + mh, y_n + mk_0) \\ k_2 = hf(t_n + vh, y_n + (v-r)k_0 + rk_1) \\ k_3 = hf(t_n + ph, y_n + (p-s-t)k_0 + sk_1 + tk_2). \end{cases} \tag{2.14}$$

The conditions on the ten parameters are -

$$\begin{cases} a + b + c + d = 1 \\ bm + cv + dp = \frac{1}{2} \\ bm^2 + cv^2 + dp^2 = \frac{1}{3} \\ bm^3 + cv^3 + dp^3 = \frac{1}{4} \\ crm + d(sm+tv) = \frac{1}{6} \\ crm^2 + d(sm^2+tv^2) = \frac{1}{12} \\ crmv + dp(sm+tv) = \frac{1}{8} \\ drtm = \frac{1}{24} . \end{cases} \tag{2.15}$$

Some algebraic manipulation shows that $p=1$. The two free parameters can be chosen as m and v as in the third-order case. The truncation error is -

$$(2.16) \qquad T_n = h^5\Big[B_1(D^4f) + B_2(DfD^2f_x) + B_3(Df_xD^2f_x) + B_4(f_xDfDf_x) + B_5(f_xD^3f) + B_6(f_x^2D^2f_x) + B_7(f_x^3Df) + B_8(f_{xx}D^2f)\Big] + O(h^6) ,$$

where

$$B_1 = \frac{1}{120} - \frac{1}{24}\Big[bm^4 + cv^4 + dp^4\Big]$$

$$B_2 = \frac{1}{20} - \frac{1}{2}\Big[crmv^2 + d(sm + tv)p^2\Big]$$

$$B_3 = \frac{1}{30} - \frac{1}{2}\Big[crm^2v + d(sm^2 + tv^2)p\Big]$$

$$B_4 = \frac{7}{120} - drtm(v + p)$$

$$B_5 = \frac{1}{120} - \frac{1}{6}\Big[crm^3 + d(sm^3 + tv^3)\Big]$$

$$B_6 = \frac{1}{120} - drtm^2$$

$$B_7 = \frac{1}{120}$$

$$B_8 = \frac{1}{40} - \frac{1}{2}\Big[cr^2m^2 + d(sm + tv)^2\Big] .$$

CHAPTER III Procedures to Obtain Optimum Methods

We will now look at the problem of minimizing truncation error in order to find optimum methods. Two different approaches to this problem are described below and a comparison between them is made in the light of experimental results. No attempt has been made to prove which approach is correct although each one has some theoretical justification. Only the relatively simple second-order methods are considered in this chapter in an attempt to develop some ideas which can be projected to higher-order methods.

The first approach is concerned with finding some sort of estimate which describes the behaviour of T_n over the region R. This estimate should also reflect the dependence of T_n on the parameters involved in the expression for T_n. The second approach assumes the independence of terms involved in the expression for T_n and on this basis we minimize in some sense the coefficients of these terms.

To begin with, we can rewrite (2.13) so that the truncation error for second-order methods is -

$$(3.1) \qquad T_n = h^3\left[\left(\frac{1}{6} - \frac{m}{4}\right)x_n''' + \frac{m}{4}f_x x_n''\right] + O(h^4) ,$$

and in order to obtain an estimate for T_n as in the first approach we must obtain estimates for x''' and x'' in the region R. This, in turn, implies that we must find estimates for the function $f(t, x)$ and its partial derivatives of second and lower orders. The natural thing to do in this case

is to use upper bounds for $f(t, x)$ and its derivatives. The first type of bound tried was found in Bieberbach's book (1, p.55)

$$\left|f(t, x)\right| < M$$

$$\left|f_{t^i x^j}\right| < \frac{N}{M^{j-1}}$$

for the domain $\left\{\begin{array}{l} |t-t_0| < a \\ |x-x_0| < b \end{array}\right\}$ with $a > h$ and $b > Ma$.

However, as we shall see, it is convenient to be able to express a bound for (3.1) in the form

$$\left|T_n\right| \leq h^3 K \left|A(m)\right| , \tag{3.2}$$

where K is some positive constant not involving h or m and since these bounds are not suitable for doing this they were disregarded.

A type of bound that does give an expression in the form (3.2) was found in Lotkin's paper (8). He assumes the existence of an L and an M such that

$$\left|f(t, x)\right| \leq M$$

$$\left|f_{t^i x^j}\right| \leq \frac{L^{i+j}}{M^j} M . \tag{3.3}$$

The existence of such L and M can easily be shown by listing the bounds on $f(t, x)$ and its derivatives, namely,

$$|f| \leq S$$

$$\left|f_{t^i x^j}\right| \leq S_{i,j} \quad ,$$

and letting $M=S$ and $L = \max_{i,j}(L_{ij})$ where

$$L_{ij}^{i+j} = S_{ij}M^{j-1}$$

Applying (3.3) to (3.1) we have

$$\left|T_n\right| \leq h^3\left|(\tfrac{1}{6} - \tfrac{m}{4})x''' + \tfrac{m}{4}f_x x''\right|$$

$$\leq h^3L^2M(\left|1 - \tfrac{3m}{2}\right| + \left|\tfrac{m}{2}\right|) \ .$$

Now, keeping in mind the fact that we wish to apply these ideas to higher-order methods, it becomes apparent that even though we may have very good bounds for $f(t, x)$ and its derivatives, the process of taking absolute values could lead to a bound which is not realistic. Consequently, a different procedure for obtaining estimates for T_n will be used.

To obtain an alternate estimate for T_n we suppose R is divided into subregions R_k which are sufficiently small so that

$$f(t, x) \approx M_k \qquad \text{in } R_k \text{ for each } k.$$

We then find L_k so that

$$f(t, x) \approx e^{L_k t}\, e^{xL_k/M_k} \qquad \text{in } R_k \text{ for each } k.$$

We also assume that for each R_k

$$f_{t^i x^j} \approx \frac{L_k^{i+j}}{M_k^j} M_k .$$

Of course this does not follow from our approximation for $f(t, x)$ and perhaps this assumption weakens the succeeding arguments but let us proceed with the analysis to see where it leads us. For each subregion R_k we use these approximations for $f(t, x)$ and its derivatives to get -

$$T_{n,k} \approx L^3 L_k^2 M_k \left[(1 - \tfrac{3m}{2}) + \tfrac{m}{2}\right]$$

$$= h^3 L_k^2 M_k A(m) ,$$

where $A(m) = 1-m$ and $T_{n,k}$ is the estimate for the truncation error in the region R_k. The constants L_k and M_k will, of course, vary between regions accordingly as $f(t, x)$ and its derivatives vary and in this way we get an estimate of the behaviour of T_n and its dependence on the parameter m for the whole region R. Clearly, the dependence on m is reflected in the function $A(m)$ and, therefore, in order to minimize T_n we want to pick m so as to minimize $|A(m)|$. The obvious choice for m is seen to be $m=1$ and so the best second-order method would be

$$y_{n+1} = y_n + \tfrac{1}{2}k_0 + \tfrac{1}{2}k_1$$

where $\quad k_0 = hf(t_n, y_n)$

$$k_1 = hf(t_n + h, x_n + k_0) .$$

It is interesting to note that this choice of m corresponds to the method given in Hildebrand (4, p.235) which he has illustrated as a "convenient choice" for the free parameter. Before comparing this theoretically "best" method with experimental results we will now look at the second approach to obtaining optimum methods.

As was indicated before the basic idea behind the second approach is the assumption that the terms in the expression for T_n are independent. To explain this idea we rewrite equation (3.1) -

$$T_n = h^3\left[(\tfrac{1}{6} - \tfrac{m}{4})(D^2f + f_xDf) + \tfrac{m}{4}f_xDf\right] + O(h^4)$$

$$= h^3\left[(\tfrac{1}{6} - \tfrac{m}{4})D^2f + \tfrac{1}{6}f_xDf\right] + O(h^4)\,, \tag{3.4}$$

and hence we want to show that there is some basis to assume that D^2f and f_xDf are independent. Now the term D^2f contains second-order partial derivatives only whereas f_xDf contains first-order partial derivatives. It is this fact that motivates the idea that there is no reason to believe that D^2f and f_xDf are dependent, and hence, we make the assumption that they are independent.

The next problem is to choose m so as to minimize the coefficients of D^2f and f_xDf in some way. One way to do this is to minimize the sum of coefficients, i.e., minimize the function

$$B(m) = \left|\tfrac{1}{6} - \tfrac{m}{4}\right| + \left|\tfrac{1}{6}\right| .$$

The minimum for this function occurs at $m=\frac{2}{3}$ and the optimum

method in this case would be

$$y_{n+1} = y_n + \frac{1}{4}k_0 + \frac{3}{4}k_1$$

where $k_0 = hf(t_n, y_n)$

$$k_1 = hf(t_n + \frac{2}{3}h, y_n + \frac{2}{3}k).$$

Another way to minimize coefficients is to minimize the sum of squares of the coefficients -

$$C(m) = (\frac{1}{6} - \frac{m}{4})^2 + (\frac{1}{6})^2.$$

Once again the minimum is seen to occur at $m=\frac{2}{3}$ and the optimum method for this case is the same as above.

To derive another "optimum" method we use the expression (3.1) -

$$T_n = h^3\left[(\frac{1}{6} - \frac{m}{4})x_n''' + \frac{m}{4}f_x x_n''\right] + O(h^4).$$

We note that (3.4) was derived from (3.1) by a regrouping of terms. Assuming the independence of x''' and $f_x x''$ we want to minimize the sum of squares of coefficients -

$$D(m) = (\frac{1}{6} - \frac{m}{4})^2 + (\frac{m}{4})^2$$

to get $m = \frac{1}{3}$ for this case. This value of m gives the method

$$y_{n+1} = y_n - \frac{1}{2}k_0 + \frac{3}{2}k_1$$

where $k_0 = hf(t_n, y_n)$

$$k_1 = hf(t_n + \frac{1}{3}h, y_n + \frac{1}{3}k_0).$$

We should note that if we tried to minimize the sum of absolute values of coefficients in this case the best m occurs for any $m \geq \frac{1}{6}$.

The next stage in our attempt to derive a procedure for obtaining an optimum method is to obtain some experimental results. Table I shows the results of integrating several differential equations using methods obtained by varying the parameter m over the range $0 \leq m \leq 1$. The table shows the maximum error committed over the range of integration using the step size indicated. The error was obtained after each step by subtracting the numerical solution y_i from the true solution.* A record of the maximum error was kept and output at the end of the calculations. In order to avoid the necessity of dealing with round-off error the step size h was taken to be fairly large. Therefore, when analyzing Table I one should consider the relative size of the error for different values of m rather than the magnitude of the error. Several comments should be made with respect to the choice of equations for the experiments. A bound for the propagated error is of the form

$$|\epsilon| \leq Ke^{(t_n - t_0)G}$$

where K and G are constants and G is found by assuming that f_x satisfies the inequality $G \geq f_x$. From this it is seen that

---

* The term "true solution" means that t_i was substituted in the equation for the true solution of the differential equation. This figure will not be the exact solution since it will contain round-off error.

TABLE I

Errors Using Second-Order Methods

Equation	$x' = -2x + \sin t$	$x' = x - \frac{2t}{x}$	$x' = x \cos t$	$x' = t - \frac{2x}{t}$
Solution	$x = \frac{1}{5}(2\sin t - \cos t)$	$x = \sqrt{2t+1}$	$x = e^{\sin t}$	$x = \frac{1}{4}\left(t^2 - \frac{1}{t^2}\right)$
No. of steps	256	224	224	224
Step Size	$\frac{1}{64}$ Right angle	$\frac{5}{128}$	$\frac{3}{64}$	$\frac{3}{128}$

Errors

m				
0	∞	>1.600	>10	>10
.1	>3	.886	>10	5.86
.2	>3	.559	7.70	3.52
.3	2.21	.268	2.97	1.96
.4	.90	.004 ←	1.71	.85
.5	.51	.240	1.13	.02 ←
.6	.35	.466	.80	.63
.7	.26	.679	.72 ←	1.14
.8	.21	.880	.74	1.56
.9	.17	1.070	.75	1.91
1.0	.14 ←	1.252	.76	2.21

Note: Minimum error is indicated by ←

if $f_x > 0$, the propagated error increases exponentially whereas if $f_x < 0$ it is damped out. Equation 1 will, therefore, have damped out error, equations 2 and 4 will have exponentially increasing error, while equation 3 will have damped out error for part of the range of integration and exponentially increasing error for the balance of the range. Also note that equations 1, 3 and 4 are linear in x whereas equation 2 is non-linear. Therefore, it is seen that the choice of equations is representative of a reasonably large class of differential equations.

In analyzing Table I we see that no firm conclusions can be drawn as to which value of m is best, and, hence, there is no ground for rejecting any of the predicted values for best m derived above. The conclusion to be drawn from this is that we have obtained all the information possible from the second-order case and the ideas developed in this chapter and summarized in Table II should be tested using third-order methods. In Chapter IV we shall see the results of these tests.

TABLE II

Summary of Procedures for Finding Best Value of m

Procedure	Best m
Approximate $f(t, x)$ by $e^{L_k t} \quad e^{xL_k/M_k}$	
1. minimize $\lvert A(m) \rvert$	1
Independence of terms in T_n	
a) collect coeff. of D^2f, $f_x Df$	
2. minimize sum of abs. values	$\frac{2}{3}$
3. minimize sum of squares	$\frac{2}{3}$
b) collect coeff. of x''', x''	
4. minimize sum of abs. values	$\geq \frac{1}{6}$
5. minimize sum of squares	$\frac{1}{3}$

CHAPTER IV Optimum Third-Order Methods

In this chapter we consider the results of applying the procedures described in Chapter III to third-order methods. We shall consider each procedure in the same order as in Table II and then the various predicted optimum methods will be compared with experimental results. The truncation term that we must deal with involves two free parameters m and v, and hence we will be concerned with choosing values of these parameters such that

$$(4.1) \qquad T_n = h^4\Big[X_1(D^3f) + X_2(DfDf_x) + X_3(f_xD^2f) + X_4(f_x^2Df)\Big] + O(h^5),$$

where

$$X_1 = \frac{1}{72}(3 + 6mv - 4m - 4v)$$

$$X_2 = \frac{1}{24}(3 - 4v)$$

$$X_3 = \frac{1}{24}(1 - 2m)$$

$$X_4 = \frac{1}{24},$$

is as small as possible.

Using the first procedure we note that in each subregion R_k we have -

$$D^3f \approx 8L_k^3M_k$$

$$DfDf_x \approx 4L_k^3M_k$$

$$f_xD^2f \approx 4L_k^3M_k$$

$$f_x^2Df \approx 2L_k^3M_k ,$$

and hence we must minimize the function

$$(4.2) \qquad \left|A(m, v)\right| = \left|8X_1 + 4X_2 + 4X_3 + 2X_4\right|$$
$$= \left|39 + 24mv - 28m - 40v\right|$$

which is equivalent to solving the equation

$$39 + 24mv - 28m - 40v = 0 .$$

This defines a one-parameter family of best methods. We find later that the experimental evidence definitely indicates that the minimum error occurs at a specific point in the m-v plane and not along a curve. Therefore, we will not consider this procedure in any further detail. In any event the curve it produces does not even go through the minimum points.

The other procedures depend on the assumption that the terms D^3f, $DfDf_x$, f_xD^2f and f_x^2Df involved in third-order methods are independent. An argument similar to the one given in Chapter III will show that there is no reason to believe that they are dependent in any way and so we will assume they are pairwise independent. The reader may have noted in Chapter III that the essential difference between this approach and the first approach is that the coefficients are not weighted as they are in, say, equation (4.2).

Using the above assumption, the second procedure is concerned with minimizing the function -

$$B(m, v) = \left|X_1\right| + \left|X_2\right| + \left|X_3\right| + \left|X_4\right| .$$

This leads to the problem of solving the system

$$3 + 6mv - 4m - 4v = 0$$
$$3 - 4v = 0$$
$$1 - 2m = 0.$$

But this system is overdetermined so we search for an m and a v that will make

$$\left|3 + 6mv - 4m - 4v\right| + 3\left|3 - 4v\right| + 3\left|1 - 2m\right|$$

as small as possible. One pair of values which gives a relative minimum is $m=\frac{1}{2}$, $v=\frac{3}{4}$. If we restrict m and v to the interval $[0, 1]$ it is easy to show that $m=\frac{1}{2}$, $v=\frac{3}{4}$ gives the absolute minimum for $B(m, v)$ and so the optimum method will be -

$$y_{n+1} = y_n + \frac{1}{9}(2k_0 + 3k_1 + 4k_2),$$

where
$$k_0 = hf(t_n, y_n)$$
$$k_1 = hf(t_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_0)$$
$$k_2 = hf(t_n + \tfrac{3}{4}h, y_n + \tfrac{3}{4}k_1).$$

The third procedure involves minimizing the sum of squares of coefficients -

$$C(m, v) = X_1^2 + X_2^2 + X_3^2 + X_4^2.$$

Taking derivatives with respect to m and v and equating to zero we have -

$$(4.3) \qquad \frac{\partial}{\partial m}C = \frac{1}{18}\left[(3 + 6mv - 4m - 4v)(3v - 2) - 3(1 - 2m)\right] = 0$$

$$(4.4) \qquad \frac{\partial}{\partial v}C = \frac{1}{18}\Big[(3 + 6mv - 4m - 4v)(3m - 2) - 6(3 - 4v)\Big] = 0 .$$

We are now faced with the problem of solving these two equations simultaneously. Noting that (4.3) is linear in m and (4.4) is linear in v the graphs of the equations can easily be drawn. It was found that there was only one point of intersection of the graphs and hence there is only one real solution for the equations. This solution is approximately $m=\frac{1}{3}$, $v=\frac{3}{4}$. The optimum method given by this procedure is, therefore,

$$y_{n+1} = y_n + \frac{1}{30}(5k_0 + 9k_1 + 16k_2) ,$$

where

$$k_0 = hf(t_n, y_n)$$

$$k_1 = hf(t_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}k_0)$$

$$k_2 = hf(t_n + \tfrac{3}{4}h, y_n - \tfrac{3}{16}k_0 + \tfrac{15}{16}k_1) .$$

The last two procedures in Table II vary from the second and third in that T_n is regrouped giving different coefficients to minimize. The new coefficients are found by collecting terms involving x^{iv}, x''', x'' so that we would have -

$$T_n = h^3\Big[X_1x^{iv} + (X_3 - X_1)f_xx''' + (X_4 - X_3)f_x^2x'' + (X_2 - X_1)Df_xx''\Big] + O(h^5)$$

$$= h^3\Big[Y_1x^{iv} + Y_2f_xx''' + Y_3f_x^2x'' + Y_4Df_xx''\Big] + O(h^5) .$$

Now if we find m and v such that the function

$$Y = Y_1^2 + Y_2^2 + (Y_3 f_x^2 + Y_4 Df_x)^2$$

is a minimum it is clear that m and v would be dependent on f and its derivatives. Alternatively, we could minimize

$$Y^* = Y_1^2 + Y_2^2 + S_1 Y_3^2 + S_2 Y_4^2 ,$$

where S_1 and S_2 are weights based on estimates of f_x^2 and Df_x. If, for example, we use (3.3) to obtain these estimates we would have

$$S_1 = L^2 \quad \text{and} \quad S_2 = 2L^2$$

where the size of L depends on the function f. It is then clear that, as before, m and v will depend on the particular function f. Since we are looking for "universal" optimum methods, i.e., methods which are independent of f, it is clear that we are unable to use these procedures.

For easy reference, the methods derived above are summarized in Table III. Also included in Table III are "standard" third-order methods as given in Hildebrand (4, p.237).

We are now ready to examine some numerical results. From (2.9) we have, for $m \neq v$ and $m,v \neq 0$:

$$(4.5) \quad \begin{cases} a = 1 - b - c \\ b = \dfrac{3v - 2}{6m(v-m)} \\ c = \dfrac{2 - 3m}{6v(v-m)} \\ r = \dfrac{1}{6cm} \end{cases}$$

TABLE III

Summary of "Best" and "Standard" Third-Order Methods

Procedures	Best Methods	
	m	v
1. Approximate $f(t, x)$ by $e^{L_k t} e^{xL_k/M_k}$	$v = \frac{28m - 39}{8(3m-5)}$	
2. Minimize sum of absolute values of coefficients	$\frac{1}{2}$	$\frac{3}{4}$
3. Minimize sum of squares of coefficients	$\frac{1}{3}$	$\frac{3}{4}$
Standard Methods		
4. Kutta	$\frac{1}{2}$	1
5. Heun	$\frac{1}{3}$	$\frac{2}{3}$

We also give results for two situations which have not been mentioned so far. They arise from "singular "solutions of (2.9). If m=v, then we have the one-parameter family

$$(4.6)\quad \begin{cases} a = \frac{1}{4} & m = v = \frac{2}{3} \\ b = \frac{3}{4} - c & r = \frac{1}{4c} \end{cases} .$$

Secondly, if v=0, we have the single method

$$(4.7)\quad \begin{cases} a = -\frac{1}{2} & m = \frac{2}{3} \\ b = c = \frac{3}{4} & r = \frac{1}{3} \end{cases} .$$

For the experiments, values of m and v in the interval $[0,1]$ were used in (4.5), and for (4.6) c was given values in $[0,1]$. As before, h was made large in order to make round-off error negligible.

Table IV (a) and (b) shows the results (maximum magnitude of the error) of solving the problem

$$(4.8)\quad \begin{cases} x' = x - \frac{2t}{x} & , x(0) = 1 \\ \text{No. of steps} = 128 \quad \text{with} \quad h = \frac{3}{32} \end{cases} .$$

The main point about this table is that there is a wide range in the size of the errors. There is, however, a definite minimum and in this particular problem our third procedure (Table III) yields values of m and v closer to the minimum than either Heun's or Kutta's method, although they are quite good.

Table V (a) and (b) shows the results of solving the

TABLE IV (a)

Errors in Solution of Problem (4.8) Using Third-Order Methods

Units of 10^{-2} L: error ≥ 64

v \ m	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	—	L	L	L	L	L	36.11	36.34	36.44	L
.2	28.31	—	L	L	L	L	L	35.19	36.11	37.05
.3	20.96	23.97	—	L	L	L	32.34	34.15	35.88	L
.4	14.35	17.34	20.96	—	L	L	30.53	33.04	35.56	37.93
.5	7.83	11.52	15.04	19.89	—	L	28.68	32.04	35.30	38.55
.6	.98	5.54	10.03	14.41	18.70	—	26.96	31.11	35.15	39.06
.7	6.09	.59	4.75	10.08	15.18	20.31	—	30.06	34.84	39.55
.8	13.16	6.73	.38	5.74	11.82	12.53	L	—	34.70	40.10
.9	15.43	13.00	5.56	1.44	8.37	15.11	L	L	—	40.56
1.0	14.43	14.54	10.78	2.85	5.05	12.53	L	L	L	—

TABLE IV (b)

Methods Using "Singular" Values of m and v for Problem (4.8)

$v = m = 2/3$

c	Error (Units of 10^{-2})
.1	24.18
.2	24.24
.3	24.29
.4	24.27
.5	24.35
.6	24.27
.7	24.26
.8	24.26
.9	24.26
1.0	24.26

$v = 0$

Error $\geqslant 64$

TABLE V (a)

Errors in Solution of Problem (4.9) Using Third-Order Methods

v \ m	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0	1.1	1.2
.1	—	P	L	L	8.13	.528	-.257	-.164	-.115	-.085	-.066	-.053
.2	P	—	L	L	L	.575	-.253	-.155	-.106	-.077	-.060	-.047
.3	.204	.206	—	L	L	.668	-.246	-.143	-.094	-.068	-.051	-.041
.4	.185	.175	.188	—	L	.935	-.234	-.125	-.079	-.056	-.042	-.032
.5	.153	.132	.132	.156	—	7.94	-.212	-.097	-.059	-.040	-.029	-.022
.6	.089	.065	.059	.065	.090	—	-.151	-.052	-.029	-.019	-.013	-.010
.7	-.075	-.040	-.033	-.033	-.040	-.074	—	.037	.018	.011	.008	.006
.8	.607	-.195	-.139	-.127	-.138	-.192	.529	—	.098	.056	.037	.026
.9	L	.408	-.247	-.209	-.207	-.243	.376	8.24	—	.127	.079	.055
1.0	L	.817	.357	-.273	-.255	-.271	.337	.683	L	—	.144	.097

L: error ≥ 10

P: parameters were not calculated

TABLE V (b)

Methods Using "Singular" Values of m and v

for Problem (4.9)

v = m = 2/3

c	Error
.1	-.243
.2	-.188
.3	-.140
.4	-.099
.5	-.065
.6	-.036
.7	-.011
.8	.010
.9	.028
1.0	.044

v = 0

Error = -.303

problem

$$(4.9)\quad \begin{cases} x' = -2x + \sin t \quad , \; x(0) = -\frac{1}{5} \\ \text{No. of steps} = 128 \quad \text{with } h = \frac{3}{64} \quad \text{right-angle.} \end{cases}$$

The negative values in the table indicate that the numerical solution is larger than the "true" solution. This fact has little importance but it is clear that omission of the minus signs would be very misleading. The main feature of this table is the fact that there is again a minimum in the general area of $m=\frac{1}{3}$, $v=\frac{3}{4}$. A comparison with the standard methods shows that Heun's is again about the same as that predicted by procedure 3, but Kutta's is quite poor. The small errors in the right half of the table were an unexpected result and, as yet, we have no explanation for this phenomenon.

To summarize the results of this chapter, we were able to predict three possible optimum methods using procedures developed in Chapter III. Experimental results show that procedure 3 (Table III) gives the best predictions. We also saw that the method so predicted was better than Kutta's method and at least as good as Heun's method. Of course, it is wrong to draw any conclusions from the results of only two experiments. However, the equations used were quite different (f_x is positive for one and negative for the other; one is linear and the other is nonlinear). Even so, the minimum error occurred for approximately the same values of m and v. This gives a strong indication that the optimum

third-order method is given by $m=\frac{1}{3}$, $v=\frac{3}{4}$.

To conclude, it appears that we have good evidence as to which procedure should be used in obtaining an optimum Runge-Kutta method. This procedure is to minimize the sum of squares of coefficients of Df, D^2f, Df_x, etc. in the truncation error expression. Further tests should, of course, be made with third-order methods. Eventually, we hope to test this procedure with fourth-order methods.

BIBLIOGRAPHY

(1) BIEBERBACH, Ludwig, Theorie der Differentialgleichungen, New York, Dover, 1944.

(2) GALLER, B.A., "Notes on the Runge-Kutta Methods for Numerical Solution of Ordinary Differential Equations," in Bartels, R.C.F., and Carr, John W. III, ed., Notes for Advanced Numerical Analysis, Ann Arbor, University of Michigan Press, 1957.

(3) GILL, S., "A Process for the Step-by-Step Integration of Differential Equations in an Automatic Digital Computing Machine," Proceedings of the Cambridge Philosophical Society, vol. 47 (January, 1951), pp. 96-108.

(4) HILDEBRAND, F.B., Introduction to Numerical Analysis, Toronto, McGraw-Hill, 1956.

(5) HULL, T.E., and NEWBERY, A.C.R., "Integration Procedures which Minimize Propagated Errors," Journal of the Society for Industrial and Applied Mathematics, vol. 9 (1961), to appear.

(6) INCE, E.L., "Numerical Integration of Ordinary Differential Equations," Appendix B in Ordinary Differential Equations, New York, Dover, 1926.

(7) KOPAL, Z., Numerical Analysis, New York, Wiley, 1955.

(8) LOTKIN, M., "On the Accuracy of Runge-Kutta's Methods," Mathematical Tables and Aids to Computation, vol. 5, (1951), pp. 128-33.