RUNGE-KUTTA FORMULAS OF OPTIMUM STABILITY

A Thesis

by

HECTOR GONZALES SIERRA

Submitted to the Graduate College of
Texas A&M University in
partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

May 1969

Major Subject Mathematics

RUNGE-KUTTA FORMULAS OF OPTIMUM STABILITY

A Thesis

by

HECTOR GONZALES SIERRA

Approved as to style and content by:

H. A. Luther
(Chairman of Committee)

H. A. Luther
(Head of Department)

(Member)

Edgar J. Boedeker
(Member)

(Member)

Bill C. Mecca
(Member)

(Member)

May 1969

ABSTRACT

RUNGE-KUTTA FORMULAS OF OPTIMUM STABILITY. (May 1969)

Hector G. Sierra, B.A., University of Texas;

Directed by: Dr. H. A. Luther

This study presents a derivation of a fourth-order Runge-Kutta formula used in the numerical method solution of a single ordinary differential equation.

Three definitions of stability of the Runge-Kutta single-step process are given. Also two theorems showing that the single-step method is stable are presented.

In this thesis, two of the stability definitions were studied and it was found that for the first stability definition (H-Stability), optimum stability will be obtained by fourth-order Runge-Kutta formulas with parameters, $R_i \geq 0$, $(i = 1, \ldots, 4)$. Optimum stability for the second stability definition is given by Runge-Kutta formulas with parameters $R_i \geq 0$ and $b_{kj} \geq 0$, $(k = 2, \ldots, 4 ; j = 1, \ldots, 3)$. In particular, a formula due to Runge is the only formula satisfying this criteria when $0 < a_2$, $a_3 < 1$.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to:

Dr. H. A. Luther for his guidance and assistance in developing this thesis;

The United States Air Force for the opportunity to undertake this study;

My wife, Carlota, and our boys for their patience and understanding.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

With the advent of the high-speed computer, numerical methods for solving ordinary differential equations have gained considerable importance in applied mathematics.

A numerical procedure which approximates solutions to a differential equation is that referred by Henrici [1] as a discrete variable method. Simply stated, this method consists of replacing a problem involving continuous variables by one involving discrete variables.

The discrete variable methods are normally classed into one-step methods and multi-step methods. In a one-step method only the value of y_n, an approximate solution to the differential equation, is required in order to determine y_{n+1}; while the multi-step method requires knowledge of more than just the previous point.

Among the one-step methods, one of the most widely used is the Runge-Kutta method.

The citations on the following pages follow the style of the SIAM Journal On Numerical Analysis.

Development of the Method

The development of the Runge-Kutta method started with the work of Runge [2] in 1895. Runge's method was improved by Heun [3] in 1900 and in 1901 by Kutta [4] who generalized the method, thus giving the method the name Runge-Kutta.

Up until recent years most of the investigation done on this method involved Runge-Kutta formulas of order four or less. Some of the recent works of higher processes are those of Butcher [5], Luther [6] and Cassity [8].

The introduction of the high-speed computer also prompted investigations into this method. A fourth-order process to minimize storage requirement was developed by Gill [9] in 1951.

Within the last eighteen years much of the work has dealt with the truncation error and error bounds of this method. Studies concerning this type of investigation have been presented by Lotkin [10], and Ralston [11].

Another area of interest is that of stability of the method. Carr [12] in 1958 presented a paper which gave a bound on the propagated error to indicate stability of the Runge-Kutta method. In

recent years Karim [13] and Lawson [14] have written papers on the region of stability for the Runge-Kutta method.

Although the stability of the Runge-Kutta method is established, at least for certain regions, the literature fails to give a Runge-Kutta formula which will minimize the bound required in the definition of stability.

It is thus the purpose of this study to determine Runge-Kutta formulas which will give optimum stability of the Runge-Kutta method. The study will be restricted to the fourth-order Runge-Kutta process.

CHAPTER II

DERIVATION OF FOURTH ORDER RUNGE-KUTTA FORMULA

As previously mentioned, the fourth-order Runge-Kutta method will be the basis for this study. Thus, in this chapter a fourth-order formula will be developed both for familiarization of the reader and for the derivation of relationships which will be used in subsequent chapters. It is noted that the relationships developed also apply to a system of differential equations when the fourth-order formula is used. The formula will be developed in the same manner as that of Ince [15].

We consider a first-order differential equation of the form

$$\frac{dy}{dx} = f(x,y) \tag{2.1}$$

with the initial condition $y(x_o) = y_o$.

Now by a Taylor's expansion of (2.1) about $x = x_o$ we have

$$y(x_o+h) = y_o + hf(x_o,y_o) + \frac{h^2}{2!} f'(x_o,y_o) + \frac{h^3}{3!} f''(x_o,y_o) + \cdots \tag{2.2}$$

The derivatives may be expressed as partial derivatives of f by first defining the operator

(2.3) $$D \triangleq \frac{\partial}{\partial x} + f \frac{\partial}{\partial y}$$

such that

$$Du = \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} f$$

where

$$f = f(x,y).$$

Now

$$\frac{dy}{dx} = f$$

and

$$\frac{d^2y}{dx^2} = \frac{df}{dx} = f_x + ff_y = Df.$$

Also,

$$\frac{d^3y}{dx^3} = \frac{d^2f}{dx^2} = D^2f + f_y\ Df$$

where

$$D^2f = \left(\frac{\partial^2}{\partial x^2} + 2f \frac{\partial^2}{\partial x \partial y} + f^2 \frac{\partial^2}{\partial y^2}\right) f$$

Likewise,

$$\frac{d^4y}{dx^4} = \frac{d^3f}{dx^3} = D^3f + f_y\ D^2f + f^2{}_y\ Df + 3(Df_y)(Df)$$

where

$$D^3 = \frac{\partial^3}{\partial x^3} + 3f \frac{\partial^3}{\partial x^2 \partial y} + 3f^2 \frac{d^3}{\partial x \partial y^2} + f^3 \frac{\partial^3}{\partial y^3}, \text{ etc.}$$

Now we rewrite (2.2) as

$$(2.4)\qquad y(x_o+h) = y(x_o) + \{hf + \frac{h^2}{2!}\ Df + \frac{h^3}{3!}\ (D^2f + f_y\ Df) + \frac{h^4}{4!}\ [D^3f + f_y\ D^2f + f^2_y\ Df + 3(Df_y)\ (Df)\ ] + \cdots\ \}_o .$$

Next we seek to replace (2.4) by an approximation of the form

$$(2.5)\qquad y(x_o+h) = y(x_o) + R_1k_1 + R_2k_2 + R_3k_3 + R_4k_4 + \cdots$$

where

$$k_1 = h\ f(x_o,y_o)$$

$$k_2 = hf\ (x_o+a_2h,\ y_o+b_{21}k_1)$$

$$k_3 = hf\ (x_o+a_3h,\ y_o+b_{31}k_1+b_{32}k_2)$$

$$k_4 = hf\ (x_o+a_4h,\ y_o+b_{41}k_1+b_{42}k_2+b_{43}k_3).$$

Here the constants R_i, a_i, and b_{ij} are to be determined such that equation (2.5) will agree with equation (2.4) up to and including the term of order h^4. Hence, we expand k_2, k_3 and k_4 by using Taylor's expansion for two variables.

Now to expand k_2, let

$$D_1 = (a_2 \frac{\partial}{\partial x} + b_{21} f_o \frac{\partial}{\partial y}). \tag{2.6}$$

Then

$$hD_1 = (a_2 h \frac{\partial}{\partial x} + b_{21} k_1 \frac{\partial}{\partial y}).$$

Thus,

$$k_2 = h[f + hD_1 f + \frac{h^2}{2!} D_1^2 f + \frac{h^3}{3!} D_1^3 f + \ldots]_o. \tag{2.7}$$

To expand k_3, let

$$D_2 = a_3 \frac{\partial}{\partial x} + (b_{31} f_o + b_{32} f_o) \frac{\partial}{\partial y} . \tag{2.8}$$

Then,

$$ha_3 \frac{\partial}{\partial x} + [b_{31} k_1 + b_{32} k_2] \frac{\partial}{\partial y}$$

$$= hD_2 + b_{32} (k_2 - f_o h) \frac{\partial}{\partial y}$$

$$= hD_2 + b_{32} h^2 [D_1 f + \frac{h}{2!} D_1^2 f + \ldots]_o \frac{\partial}{\partial y} .$$

Therefore,

$$k_3 = h[f + hD_2 f + \frac{h^2}{2!} D_2^2 f + \frac{h^3}{3!} D_2^3 f + \ldots + h^2 b_{32} (f_y D_1 f + \frac{h}{2!} f_y D_1^2 f + hD_1 f D_2 f_y + \ldots)]_o . \tag{2.9}$$

To expand k_4, let

$$D_3 = a_4 \frac{\partial}{\partial x} + (b_{41} + b_{42} + b_{43}) f_o \frac{\partial}{\partial y} . \tag{2.10}$$

And in the same manner as before,

$$
\begin{aligned}
(2.11)\qquad k_4 = h[&f + hD_3 f + \frac{h^2}{2!} D_3{}^2 f + \frac{h^3}{3!} D_3{}^3 f + \cdots \\
&+ h^2(b_{42} D_1 f + b_{43} D_2 f) f_y \\
&+ h^3(b_{42} D_1 f + b_{43} D_2 f) D_3 f_y \\
&+ \frac{h^3}{2!}(b_{42} D_1{}^2 f + b_{43} D_2{}^2 f + 2 b_{32} b_{43} f_y D_1 f) f_y \\
&+ \cdots]_o \, .
\end{aligned}
$$

We now substitute equations (2.7), (2.9) and (2.11) in (2.5). Next we equate terms of like powers of h of equations (2.4) and (2.5) obtaining the following:

$$
\begin{aligned}
&R_1 + R_2 + R_3 + R_4 = 1 \\
&a_2 R_2 + a_3 R_3 + a_4 R_4 = \tfrac{1}{2} \\
&a_2{}^2 R_2 + a_3{}^2 R_3 + a_4{}^2 R_4 = \tfrac{1}{3} \\
(2.12)\qquad &a_2{}^3 R_2 + a_3{}^3 R_3 + a_4{}^3 R_4 = \tfrac{1}{4} \\
&a_2 b_{32} R_3 + (a_2 b_{42} + a_3 b_{43}) R_4 = \tfrac{1}{6} \\
&a_2{}^2 b_{32} R_3 + (a_2{}^2 b_{42} + a_3{}^2 b_{43}) R_4 = \tfrac{1}{12} \\
&a_2 a_3 b_{32} R_3 + (a_2 b_{42} + a_3 b_{43}) a_4 R_4 = \tfrac{1}{8} \\
&a_2 b_{32} b_{43} R_4 = \tfrac{1}{24}
\end{aligned}
$$

where

$$
\begin{aligned}
&a_2 = b_{21} \\
(2.13)\qquad &a_3 = b_{31} + b_{32} \\
&a_4 = b_{41} + b_{42} + b_{43}
\end{aligned}
$$

Since the eleven equations in (2.12) and (2.13) contain thirteen unknowns, we assume two of the unknowns to be arbitrary and solve for the remaining unknowns in terms of them. We do this by the following procedure.

From the equations of (2.12) add the second equation multiplied by a_2a_4 and the third multiplied by $-(a_2 + a_4)$ and add them to the fourth, obtaining

$$(2.14) \qquad R_3a_3(a_2-a_3)(a_4-a_3) = \frac{a_2a_4}{2} - \frac{a_2+a_4}{3} + \frac{1}{4} .$$

From the fifth and seventh equations it follows that

$$(2.15) \qquad R_3(a_2b_{32})(a_4-a_3) = \frac{a_4}{6} - \frac{1}{8} ,$$

while from the fifth and sixth we have

$$(2.16) \qquad R_4(a_3b_{43})(a_3-a_2) = \frac{1}{12} - \frac{a_2}{6} .$$

By eliminating R_4 from (2.16) above and equation eight of (2.12) we find that

$$(2.17) \qquad a_2b_{32} = \frac{a_3(a_2-a_3)}{2(2a_2-1)} , \text{ if } a_2 \neq \frac{1}{2} .$$

Now substitute (2.17) in (2.15) obtaining

$$(2.18) \qquad R_3a_3(a_2-a_3)(a_4-a_3) = (\frac{a_4}{3} - \frac{1}{4})(2a_2-1)$$

Comparison of (2.18) with (2.14) yields

$$(2.19) \qquad \frac{a_2a_4}{2} - \frac{a_2+a_4}{3} + \frac{1}{4} = (2a_2-1)(\frac{a_4}{3} - \frac{1}{4})$$

$$= \frac{2a_2a_4}{3} - \frac{a_2}{2} - \frac{a_4}{3} + \frac{1}{4} .$$

And hence,

$$a_2 a_4 = a_2 \quad (2.20)$$

But from the last equation of (2.12) it is clear that $a_2 \neq 0$, thus,

$$a_4 = 1 \quad (2.21)$$

Also from equation eight of (2.12) $R_4 \neq 0$ and therefore R_3 from equation (2.15) is not equal to zero.

Now R_1, R_2, R_3, and R_4 and be determined uniquely in terms of a_2 and a_3 from the first four equations of (2.12) if their determinant which has the value

$$a_2 a_3 (a_2 - a_3)(a_3 - 1)(1 - a_2) \quad (2.22)$$

is non-singular. The values for this non-singular case are:

$$\begin{aligned} R_1 &= \frac{1}{2} + \frac{1 - 2(a_2 + a_3)}{12 a_2 a_3} \\ R_2 &= \frac{2a_3 - 1}{12 a_2 (a_3 - a_2)(1 - a_2)} \\ R_3 &= \frac{1 - 2a_2}{12 a_3 (a_3 - a_2)(1 - a_3)} \\ R_4 &= \frac{1}{2} + \frac{2(a_2 + a_3) - 3}{12 (1 - a_2)(1 - a_3)} \end{aligned} \quad (2.23)$$

From the fifth, sixth, and seventh equations of (2.12) we determine b_{32}, b_{42}, and b_{43} in terms of a_2 and a_3 provided their determinant whose value is

(2.24) $R_3 R_4^2 a_2^2 a_3 (a_3-a_2)(a_3-1)$

is non-singular. The values are found to be:

$$b_{32} = \frac{a_3(a_3-a_2)}{2a_2(1-2a_2)}$$

(2.25)
$$b_{42} = \frac{(1-a_2)\,[a_2+a_3-1-(2a_3-1)^2\,]}{2a_2(a_3-a_2)\,[6a_2a_3-4(a_2+a_3)+3\,]}$$

$$b_{43} = \frac{(1-2a_2)\,(1-a_2)\,(1-a_3)}{a_3(a_3-a_2)\,[6a_2a_3-4(a_2+a_3)+3\,]}\,.$$

Now any two conditions consistent with the foregoing equations may be imposed. If we impose a condition of symmetry such that

(2.26) $R_1 = R_4$ and $R_2 = R_3$,

and a second condition requiring that the range from x_o to $x_1 = x_o + h$ be divided into three equal parts so that

(2.27) $a_2 = \frac{1}{3}$ and $a_3 = \frac{2}{3}$,

we find the values

(2.28)
$$\begin{array}{llll} R_1 = \frac{1}{8} & & & \\ R_2 = \frac{3}{8} & a_2 = \frac{1}{3} & b_{21} = \frac{1}{3} & \\ R_3 = \frac{3}{8} & a_3 = \frac{2}{3} & b_{31} = -\frac{1}{3} & b_{32} = 1 \\ R_4 = \frac{1}{8} & a_4 = 1 & b_{41} = 1 & b_{42} = -1 \\ & & & b_{43} = 1\,. \end{array}$$

Finally we arrive at the formula due to Kutta:

(2.29) $$y_{n+1} = y_n + \frac{1}{8}[k_1+3k_2+3k_3+k_4] ,$$

where

$$k_1 = hf(x_n, y_n)$$

$$k_2 = hf(x_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}k_1)$$

$$k_3 = hf(x_n + \tfrac{2}{3}h, y_n - \tfrac{1}{3}k_1 + k_2)$$

$$k_4 = hf(x_n + h, y_n + k_1-k_2+k_3) .$$

Lastly we consider the possibilities when the determinants of (2.22) and (2.24) are singular. It is found that the only cases possible are the following:

(2.30) Case 1:

$$a_2 = a_3 = \frac{1}{2} \quad \text{and} \quad a_4 = 1$$

with

$$R_1 = \frac{1}{6} \qquad b_{32} = \frac{1}{6R_3}$$

$$R_2 = \frac{2}{3} - R_3 \qquad b_{42} = 1 - 3R_3$$

$$R_4 = \frac{1}{6} \qquad b_{43} = 3R_3 .$$

(2.31) Case 2:

$$a_2 = 1 \quad ; \quad a_3 = \frac{1}{2} \quad \text{and} \quad a_4 = 1$$

with

$$R_1 = \frac{1}{6} \qquad b_{32} = \frac{1}{8}$$

$$R_2 = \frac{1}{6} - R_4 \qquad b_{42} = -\frac{1}{12R_4}$$

$$R_3 = \frac{2}{3} \qquad b_{43} = \frac{1}{3R_4} \ .$$

(2.32) Case 3:

$$a_2 = \frac{1}{2} \ ; \quad a_3 = 0 \quad \text{and} \quad a_4 = 1$$

with

$$R_1 = \frac{1}{6} - R_3 \qquad b_{32} = \frac{1}{12R_3}$$

$$R_2 = \frac{2}{3} \qquad b_{42} = \frac{3}{2}$$

$$R_4 = \frac{1}{6} \qquad b_{43} = 6R_3 \ .$$

CHAPTER III

SOLUTION APPROXIMATION OF THE ORDINARY DIFFERENTION EQUATION

As stated in the introduction, the solution of a given ordinary differential equation subject to given initial conditions can be found by numerical methods or, more exactly by discrete variable methods. Hence, we will find approximate solutions to the ordinary differential equation by finding solutions to certain equations called difference equations which approximate the differential equation. Therefore, it seems appropriate at this point to present a few important aspects of difference equations before proceding to discuss the solutions of differential equations.

Difference Equations

The theory of difference equations is very similar to the theory of differential equations. The main difference between the two theories is that the difference equation theory seeks as a solution a sequence instead of a function. Normally the sequence with the general element u_k is denoted as $\{u_o, u_1, \ldots\}$

or more commonly as $\{u_k\}$.

The general difference equation with constant coefficients can be written in the form

$$(3.1) \qquad a_o u_j + a_1 u_{j+1} + \ldots + a_n u_{j+n} = C_{j+n}, \quad j = 0, 1, \ldots ;$$

where the C_{j+n} are the non-homogeneous terms. The difference equation in (3.1) is of order n and generally, a solution $\{u_j\}$ is determined by specifying n initial conditions.

If $C_{j+n} = 0$ in (3.1), we can then have nth order homogeneous difference equations expressed as

$$(3.2) \qquad a_o u_j + a_1 u_{j+1} + \ldots + a_n u_{j+n} = 0, \quad j = 0, 1, \ldots .$$

For the homogeneous difference equations the set of r solutions, $\{u_i^{(1)}\}$, $\{u_i^{(2)}\}$, ... , $\{u_i^{(r)}\}$ are lineraly independent iff $\alpha u_i^{(1)} + \ldots + \beta u_i^{(r)} = 0$, i = 0, 1, ...; implies $\alpha = \ldots = \beta = 0$.

A set of n independent solutions of the homogeneous difference equations of order n is called a fundamental set of solutions.

Any solutions, say $\{v_i\}$ of the homogeneous difference equations (3.2) can always be expressed uniquely in terms of the fundamental set of solutions.

A fundamental set of solutions for (3.2) can be found by trying as a solution the powers of some scalar,

say $u_i = \alpha x^i$, $i = 0, 1, \ldots$.

Then (3.2) becomes

$$(3.3) \qquad (a_n x^n + a_{n-1} x^{n-1} + \ldots a_0)(\alpha x^i) = 0.$$

If $\alpha = 0$, the solution is trivial. Thus we consider only the roots of

$$(3.4) \qquad P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0 = 0 .$$

The polynomial $P_n(x)$ is called the characteristic polynomial of (3.2). If the roots of the characteristic polynomial are distinct, then a fundamental set of solutions is given by $\{u_i^k\} = \{x_k^i\}$, $k = 1, 2, \ldots, n$. If the roots are not distinct, we can still obtain a fundamental set of solutions, W_i^k by using derivatives of the powers of the root.

Lastly we introduce a linear difference equations theorem which will later be used.

Theorem 3.1. *Let* $\{u_j^{(\nu)}\}$ *be the fundamental set of solutions of the* n^{th} *order homogeneous difference equation which satisfy the initial conditions*

$$(3.5) \qquad u_i^{(\nu)} = \delta_{i\nu} \; ; \; i = 0, 1, \ldots, n-1 \; ; \quad \nu = 0, 1, \ldots, n-1 .$$

Then the solution of the non-homogeneous equation subject to initial conditions is

$$(3.6) \qquad \vec{u}_j = \sum_{\nu=0}^{n-1} \vec{u}_\nu u_j^{(\nu)} + \frac{1}{a_n} \sum_{k=0}^{j-n} \vec{C}_{k+n} u_{j-k-1}^{(n-1)},$$

We also define

$$(3.7) \qquad u_i^{(n-1)} \equiv 0, \text{ for all } i < 0$$

and

$$(3.8) \qquad \vec{C}_j \equiv \vec{0}, \text{ for all } j < n.$$

The proof for the above theorem is given by Isaacson and Keller [16]. For present purposes, $n=1$, $a_n=1$, $a_o=-1$.

Next, since we seek numerical methods solutions to the differential equation, there are certain errors which must be taken into consideration.

Numerical Methods Errors

When using a numerical method, one must take into account the error of approximation. Actually there are errors to consider. The first, called the discretization error, is due to the fact that the number $\vec{u}_k$ given by the theoretical method will not agree with $\vec{y}_k = \vec{y}(x_k)$, the true solution to the differential equation. The discretization error is denoted by

$$\vec{e}_k = \vec{u}_k - \vec{y}_k. \tag{3.9}$$

A second error is due to the limitations of any computing machinery. Thus instead of the number $\vec{u}_k$, the number actaully obtained by the computing equipment is $\vec{u}^*_k$. The difference between the number we should have gotten by the method being used and the number actually obtained is called the round-off error and is written as

$$h\vec{\rho}_{k+1} = \vec{u}^*_{k+1} - \vec{u}^*_k - h\vec{F}(x, h, \vec{u}^*_k, \vec{f}). \tag{3.10}$$

In this definition, $\vec{F}(x, h, \vec{u}, \vec{f})$ can be of rather general character. For the purpose of this thesis, $\vec{F}$ is as defined by (3.21) and (3.22). This round-off error is dependent on such things as the precision used in the computer (single or double precision) and the type of operation used (fixed or floating).

Then for the numerical method total error denoted by $\vec{u}^*_k - \vec{y}_k$, we find that

$$|\vec{u}^*_k - \vec{y}_k| \leq |(\vec{u}^*_k - \vec{u}_k)| + |(\vec{u}_k - \vec{y}_k)|. \tag{3.11}$$

The local truncation error, denoted by τ_{j+k} measures the difference between the differential equation and the difference equation and is normally defined in terms of

(3.12) $h\vec{\tau}(x,h) = \vec{y}(x+h)-\vec{y}(x)-h\vec{F}(h,x,\vec{y}(x),\vec{f})$.

Here $\vec{F}$ is the functional in (3.10); once again for our purpose use (3.21).

Approximate Solutions

Now we are ready to discuss the approximate solutions of a system of ordinary differential equations expressed in vector form as

(3.13) $$\frac{d\vec{y}}{dx} = \vec{f}(x,\vec{y})$$

and having an exact solution

(3.14) $\vec{y} = \vec{\bar{y}}(x)$

in some interval

(3.15) $a \leq x \leq b$

and subject to the initial conditon

(3.16) $\vec{y}(a) = \vec{y}_o$.

By use of a numerical procedure, in particular a single-step method, we seek a value $\vec{u}_j$ which approximates $\vec{y}(x_j) = \vec{y}_j$, the exact solution. Here we have that

(3.17) $x_j = a+jh, \quad j=0, 1, \ldots, N$

and

(3.18) $h = \frac{(b-a)}{N}$

where N is a positive integer.

We will assume that $\vec{f}$ belongs to a class F as defined below.

Definition 3.1. F is defined as a class of real vector-valued functions

$$\vec{f} = ({}^{1}f(x,\vec{y}),\ {}^{2}f(x,\vec{y}),\ \ldots,\ {}^{p}f(x,\vec{y}))\ ,$$

where $\vec{y} = ({}^{1}y, {}^{2}y, \ldots, {}^{p}y)$ and such that $\vec{f}$, $\vec{f}_x$, $\vec{f}_{1_y}$, $\vec{f}_{2_y}$, $\ldots$, f_{p_y} and all partial derivatives of the first four orders are continuous and uniformly bounded in S_B : $\{(x,\vec{y}) \mid a \leq x \leq b,\ \|\vec{y}\| < \infty\}$

Next, let $\vec{u}_o$ be defined by

(3.19) $\vec{u}_o = \vec{y}_o + \vec{e}_o$

where $\vec{e}_o$ is the initial discretization error and is a function of h only.

For $1 \leq j \leq n$ we let $\vec{u}_j$ be uniquely defined (assuming h is sufficiently small) by the difference equation

(3.20) $\vec{u}_{j+1} - \vec{u}_j = h\,\vec{F}\ \{h, x_j, \vec{u}_j, \vec{f}\}$

where $\vec{F}\ \{h, x_j, \vec{u}_j, \vec{f}\}$ is a fourth order Runge-Kutta process; that is,

$$(3.21)\qquad \vec{F}\{h, x, \vec{u}, \vec{f}\} = R_1\vec{k}_1 + R_2\vec{k}_2 + R_3\vec{k}_3 + R_4\vec{k}_4 .$$

Here R_1, R_2, R_3, R_4 are constants and

$$\vec{k}_1 = \vec{k}_1(x, h, \vec{u}) = \vec{f}(x, \vec{u})$$

$$\vec{k}_2 = \vec{k}_2(x, h, \vec{u}) = \vec{f}(x + a_2h, \vec{u} + hb_{21}\vec{k}_1)$$

$$(3.22)\qquad \vec{k}_3 = \vec{k}_3(x, h, \vec{u})$$

$$= \vec{f}(x + a_3h, \vec{u} + hb_{31}\vec{k}_1 + hb_{32}\vec{k}_2)$$

$$\vec{k}_4 = \vec{k}_4(x, h, \vec{u})$$

$$= \vec{f}(x + a_4h, \vec{u} + hb_{41}\vec{k}_1 + hb_{42}\vec{k}_2 + hb_{43}\vec{k}_3)$$

where the a_i, b_{im} and R_i are real and must satisfy the relationships of (2.12).

We shall call the numerical method (3.20) the theoretical numerical approximation to (3.13).

Next let $\vec{\rho}_0$ be a function of h only and define

$$(3.23)\qquad \vec{u}^*_0 = \vec{y}_0 + \vec{\rho}_0 .$$

For the interval $1 \leq j \leq n$ and for h sufficiently small, $\vec{u}^*_j$ is uniquely determined by

$$(3.24)\qquad \vec{u}^*_{j+1} - \vec{u}^*_j = h\vec{F}\{h, x_j, \vec{u}^*_j, \vec{f}\} + h\vec{\rho}_{j+1} .$$

Of course the function $\vec{F}$ is defined as in (3.21) with $u_j = u^*_j$ and $\vec{\rho}_{j+1}$ is the local rounding error. We shall call method (3.24) the computed approximation to (3.13).

CHAPTER IV

STABILITY OF SINGLE-STEP METHOD

In this chapter we will show that the Runge-Kutta method is stable by presenting theorems concerning three different types of stability. We begin by stating some definitions which we will utilize in proving stability. The following three types of stability will be defined as by Luther [7].

Definition 4.1. Stability: Let the sequences $\{\vec{u}^*_j\}$ and $\{\vec{u}'^*_j\}$ be solutions of method (3.24), both for the same $\vec{F}$, $\vec{f}$, and h but perhaps with different round-off errors.

Then method (3.24) is stable iff, for $\vec{f}$ belonging to F, there is an h_o and M, such that for all $0 \leq h \leq h_o$ we have $||\vec{u}^*_i - \vec{u}'^*_i|| \leq M\varepsilon$, $0 \leq i \leq N$, provided $||\vec{\rho}_i - \vec{\rho}'_i|| \leq \varepsilon$, $0 \leq i \leq N$.

Now let the sequences $\{\vec{u}^*_j\}$ and $\{\vec{u}_j\}$ be solutions of method (3.24) and method (3.20) respectively.

Then method (3.20) is stable iff, for $\vec{f}$ belonging to F, there is an h_o and M such that for $0 \leq h \leq h_o$ we have $||\vec{u}_j - \vec{u}^*_j|| \leq M\varepsilon$, $0 \leq j \leq N$, provided $||\vec{\rho}_o - \vec{e}_o|| \leq \varepsilon$ and $||\vec{\rho}_j|| \leq \varepsilon$, $j \geq 1$.

Definition 4.2. L-Stability: Let the sequences $\{\vec{u}_j\}$ and $\{\vec{u}'_j\}$ be solutions of method (3.20). Then method (3.20) is L-stable iff, for $\vec{f}$ belonging to F, there is an h_o and M such that for all $0 \leq h \leq h_o$ we have $||\vec{u}_j - \vec{u}'_j|| \leq M\varepsilon$, $0 \leq j \leq N$, provided $||\vec{e}_o - \vec{e}'_o|| \leq \varepsilon$.

Definition 4.3. H-Stability: Method (3.24) is said to be H-stable iff, for $\vec{f}$ belonging to F, there is an h_o and M (ε) such that for all $0 \leq h \leq h_o$ we have $\max_{0 \leq j \leq N} ||\vec{u}^*_j|| \leq M(\varepsilon)$, provided $||\vec{u}^*_o|| \leq \varepsilon$ and $||\vec{\rho}_j|| \leq \varepsilon$, $1 \leq j \leq N$.

Method (3.20) is said to be H-stable iff, for $\vec{f}$ belonging to F, there is an h_o and M (ε) such that for all $0 \leq h \leq h_o$ we have $\max_{0 \leq j \leq N} ||\vec{u}_j|| \leq M(\varepsilon)$ provided $||\vec{u}_o|| \leq \varepsilon$.

Now we state two Lemmas which follow from the stability definitions.

Lemma A. For the given class F, stability of method (3.24) implies stability of method (3.20) and stability of method (3.20) implies L-stability of method (3.20).

Lemma B. For the given class F, H-stability of method (3.24) implies H-stability of method (3.20).

Proof for both lemmas: Note that method (3.20) is a special case of method (3.24). Also note that

for $j \geq 1$, $\vec{\rho}_j$ and $\vec{\rho}'_j$ can be chosen.

Since it will later be required, it seems convenient at this time to introduce another definition involving stability.

Definition 4.4. Root Condition: Let the process have the character

$$\sum_{s=0}^{n} a_s u_{j+s} = h\,\vec{G}\,(x_j, h, \vec{u}_{j-m}, \ldots, \vec{u}_{j+m}, \vec{f})$$

where $\vec{G}$ is determined uniquely when the function $\vec{f}$ is known, as well as h, x_j, $\vec{u}_{j-m}$, ... , $\vec{u}_{j+m}$; m a non-negative integer and a_n, $a_o \neq 0$.

Now let the polynomial $P(\zeta) = \sum_{s=0}^{n} a_s \zeta^s$ be associated with the LHS of the process formula. Then $P(\zeta)$ is said to satisfy the root condition iff all zeros of $P(\zeta)$ are one or less than one in modulus and any zero of modulus one is simple.

For our single step process, $P(\zeta) = \zeta - 1$ is the polynomial associated with the LHS of the difference equation (3.24) or (3.20). Since the only root of $P(\zeta)$ is one, $P(\zeta)$ satisfies the root condition.

To show stability of method (3.20) and method (3.24) we will have to establish that $\vec{F}$ satisfies the following properties (see [7]):

$$(4.1) \qquad \| \vec{F}\{h, x_j, \vec{u}_j, \vec{f}\} \| \leq K$$

$$(4.2) \qquad \| \vec{F}\{h, x_j, \vec{u}_j, \vec{f}\} - \vec{F}\{h, x_j, \vec{v}_j, \vec{f}\} \| \leq C \| \vec{u}_j - \vec{v}_j \| .$$

where K and C are constants independent of x_j, $\vec{u}_j$, and $\vec{v}_j$, but may depend on the upper bounds of $\vec{f}$ and on a finite number of its partial derivatives.

We now introduce some properties of vector norms which will be used to prove (4.1) and (4.2).

Vector norm. For every vector $\vec{x}$ in a linear space S, there corresponds a unique real number $\| \vec{x} \|$. This number is called the norm of $\vec{x}$ iff:

$$(4.3) \qquad \| \vec{x} \| \geq 0 \text{ , for all } \vec{x} \text{ belonging to S.}$$

$$(4.4) \qquad \| \vec{x} \| = 0, \text{ iff } \vec{x} = 0.$$

$$(4.5) \qquad \| c\vec{x} \| = |c| \cdot \| \vec{x} \| \text{ , for all scalars c and } \vec{x} \text{ belonging to S.}$$

$$(4.6) \qquad \| \vec{x} + \vec{y} \| \leq \| \vec{x} \| + \| \vec{y} \| .$$

Although there are several examples of norms we shall make use only of the maximum norm defined as

$$(4.7) \qquad \| \vec{x} \|_\infty = \max_j |x_j| .$$

Here $\vec{x} = (x_1, x_2, \ldots, x_p)$.

Establishing Property (4.1). Now we proceed to show that (4.1) holds for the fourth-order Runge-Kutta method. From (3.21) and the use of norms we have

$$(4.8) \qquad \|\vec{F}\{h, x_j, \vec{u}_j, \vec{f}\}\| \leq |R_1|\|\vec{k}_1\| + |R_2|\|\vec{k}_2\| + |R_3|\|\vec{k}_3\| + |R_4|\|\vec{k}_4\| .$$

Applying the definition of the maximum norm yields

$$(4.9) \qquad \|\vec{F}\{h,x_j,\vec{u}_j,\vec{f}\}\|_\infty \leq |R_1|\,|^{\alpha}f(x_j,\vec{u}_j) + |R_2|\,|^{\beta}f(x_j+a_2h,\vec{u}_j+hb_{21}\vec{k}_1) + |R_3|\,|^{\gamma}f(x_j+a_3h,\vec{u}_j+hb_{31}\vec{k}_1+hb_{32}\vec{k}_2)| + |R_4|\,|^{\delta}f(x_j+a_4h,\vec{u}_j+hb_{41}\vec{k}_1+hb_{42}\vec{k}_2+hb_{43}\vec{k}_3)|$$

where α, β, γ, δ denote the component yielding the maximum valued element of the vector functions $\vec{f}$ of $\vec{k}_1$, $\vec{k}_2$, $\vec{k}_3$ and $\vec{k}_4$ respectively.

But by definition 3.1, each has an upper bound say, M_1. Therefore we have

$$(4.10) \qquad \|\vec{F}\{h, x_j, \vec{u}_j, \vec{f}\}\| \leq K$$

where

$$(4.11) \qquad K = M_1\{|R_1| + |R_2| + |R_3| + |R_4|\} .$$

Hence property (4.1) is established.

Establishing Property (4.2). For ease of presentation we define

$$(4.12)\qquad \vec{F}\{h,x_j,\vec{u}_j,\vec{f}\}=R_1\vec{k}_{1(u_j)} + R_2\vec{k}_{2(u_j)} + R_3\vec{k}_{3(u_j)} + R_4\vec{k}_{4(u_j)}$$

where

$$(4.13)\qquad \begin{aligned}
\vec{k}_{1(u_j)} &= \vec{f}(x_j,\vec{u}_j)\\
\vec{k}_{2(u_j)} &= \vec{f}(x_j+a_2h,\ \vec{u}_j+hb_{21}\vec{k}_{1(u_j)})\\
\vec{k}_{3(u_j)} &= \vec{f}(x_j+a_3h,\vec{u}_j+hb_{31}\vec{k}_{1(u_j)}+hb_{32}\vec{k}_{2(u_j)})\\
\vec{k}_{4(u_j)} &= \vec{f}(x_j+a_4h,\vec{u}_j+hb_{41}\vec{k}_{1(u_j)}+hb_{42}\vec{k}_{2(u_j)}\\
&\quad + hb_{43}\vec{k}_{3(u_j)}).
\end{aligned}$$

Now by the properties of norms and employing the definition of the maximum norm we have

$$(4.14)\qquad \begin{aligned}
&||\vec{F}\{h,x_j,\vec{u}_j,\vec{f}\}-\vec{F}\{h,x_j,\vec{v}_j,\vec{f}\}||_\infty\\
&\quad\le |R_1|\,|{}^{\alpha}f(x_j,\vec{u}_j) - {}^{\alpha}f(x_j,\vec{v}_j)|\\
&\quad + |R_2|\,|{}^{\beta}f(x_j+a_2h,\vec{u}_j+hb_{21}\vec{k}_{1(u_j)})\\
&\qquad - {}^{\beta}f(x_j+a_2h,\vec{v}_j+hb_{21}\vec{k}_{1(v_j)})|\\
&\quad + |R_3|\,|{}^{\gamma}f(x_j+a_3h,\vec{u}_j+hb_{31}\vec{k}_{1(u_j)}+hb_{32}\vec{k}_{2(u_j)})\\
&\qquad - {}^{\gamma}f(x_j+a_3h,\vec{v}_j+hb_{31}\vec{k}_{1(v_j)}+hb_{32}\vec{k}_{2(v_j)})|\\
&\quad + |R_4|\,|{}^{\delta}f(x_j+a_4h,\vec{u}_j+hb_{41}\vec{k}_{1(u_j)}+hb_{42}\vec{k}_{2(u_j)})\\
&\qquad + hb_{43}\vec{k}_{3(u_j)}) - {}^{\delta}f(x_j+a_4h,\vec{v}_j+hb_{41}\vec{k}_{1(v_j)})\\
&\qquad + hb_{42}\vec{k}_{2(v_j)} + hb_{43}\vec{k}_{3(v_j)})|
\end{aligned}$$

where α, β, γ, δ denote the components yielding the maximum value element of the vector differences of $\vec{k}_1$, $\vec{k}_2$, $\vec{k}_3$, and $\vec{k}_4$ respectively.

Now applying Taylor's formula for functions of several variables to the second factor of the first term of the RHS of (4.14) we find it is equal to

$$(4.15) \quad \left| ({}^{1}u_j - {}^{1}v_j) \frac{\partial}{\partial {}^{1}u} {}^{\alpha}f(x_j, \vec{v}_j + \theta_1 \{\vec{u}_j - \vec{v}_j\}) \right.$$

$$\left. + ({}^{2}u_j - {}^{2}v_j) \frac{\partial}{\partial {}^{2}u} {}^{\alpha}f(x_j, \vec{v}_j + \theta_1 \{\vec{u}_j - \vec{v}_j\}) + \ldots \right|$$

where $0 < \theta_1 < 1$ and ${}^{i}u_j$, ${}^{i}v_j$, $(i=1, \ldots, P)$ are elements of the vectors $\vec{u}_j$ and $\vec{v}_j$. But by definition 3.1, the partial derivatives have a common upper bond, say M_2; where $M_2 > 0$. Thus (4.15) is less or equal to

$$(4.16) \quad M_2 \left| {}^{1}u_j - {}^{1}v_j \right| + M_2 \left| {}^{2}u_j - {}^{2}v_j \right| + \ldots .$$

Replacing each term by the maximum term denoted by $M_2({}^{L}u_j - {}^{L}v_j)$ yields

$$(4.17) \quad M_2 P \left| {}^{L}u_j - {}^{L}v_j \right| \geq \left| {}^{\alpha}f(x_j, \vec{u}_j) - {}^{\alpha}f(x_j, \vec{v}_j) \right|$$

where P is the number of terms due to the number of partial derivatives with respect to the vector $\vec{u}_j$ of length P and where $P \geq 1$.

Next we apply Taylor's formula to the second factor of the second term of the RHS of (4.14) and

obtain

$$(4.18) \quad \left| \left(\{ {}^1u_j + hb_{21} {}^1k_{1(u_j)} \} - \{ {}^1v_j + hb_{21} {}^1k_{1(v_j)} \} \right) \frac{\partial^\beta f}{\partial {}^1u} \right.$$

$$+ \left(\{ {}^2u_j + hb_{21} {}^2k_{1(u_j)} \} \right.$$

$$\left. \left. - \{ {}^2v_j + hb_{21} {}^2k_{1(v_j)} \} \right) \frac{\partial^\beta f}{\partial {}^2u} + \ldots \right|$$

where $0 < h \leq 1$ and

$$\frac{\partial^\beta f}{\partial {}^i u} = \frac{\partial}{\partial {}^i u} {}^\beta f(x_j + a_2 h, \vec{v}_j + hb_{21} \vec{k}_{1(v_j)}$$

$$+ \theta_2 (\vec{u}_j + hb_{21} \vec{k}_{1(u_j)} - \vec{v}_j - hb_{21} \vec{k}_{1(v_j)} \quad)$$

with $0 < \theta_2 < 1$.

But as before, the partial derivatives are bounded by M_2 and hence (4.18) is less or equal to

$$(4.19) \quad M_2 \{ |{}^1u_j - {}^1v_j| + |{}^2u_j - {}^2v_j| + \ldots \} + M_2 h |b_{21}| \{$$

$$|{}^1k_{1(u_j)} - {}^1k_{1(v_j)}| + |{}^2k_{1(u_j)} - {}^2k_{1(v_j)}|$$

$$+ \ldots \} .$$

Now consider the first bracket term of the above equation and replace each of its terms by the maximum term denoted by $M_2 ({}^Iu_j - {}^Iv_j)$, thus obtaining

$$(4.20)\qquad M_2P\,|\,{}^{L}u_j-{}^{L}v_j\,| + M_2h\,|b_{21}|\,\{\,|{}^{1}k_{1(u_j)}-{}^{1}k_{1(v_j)}\,| + |{}^{2}k_{1(u_j)}-{}^{2}k_{1(v_j)}\,| + \dots\,\}\,.$$

Then recalling that $k_1(u_j) = f(x_j, \vec{u}_j)$ and $k_1(v_j) = f(x_j, \vec{v}_j)$ and using (4.17) we finally obtain

$$(4.21)\qquad M_2P\,|\,{}^{L}u_j-{}^{L}v_j\,| + M_2h\,|b_{21}\,|\{M_2P\,|\,{}^{L}u_j-{}^{L}v_j\,| + M_2P\,|{}^{L}u_j-{}^{L}v_j\,| + \dots\,\} = (M_2P+hM_2{}^2P^2\,|\,b_{21}|\,)\,|\,{}^{L}u_j-{}^{L}v_j\,|\,.$$

Following the same procedure we find that the second factor of the third term is less or equal to

$$(4.22)\qquad (M_2P + hM_2{}^2P^2\{\,|b_{31}| + |b_{32}\,|\,\} + h^2M_2{}^3P^3\,|b_{32}|\,|b_{21}\,|\,)\,|\,{}^{L}u_j-{}^{L}v_j\,|\,.$$

Likewise the second factor of the fourth term is less or equal to

$$(4.23)\qquad (M_2P + hM_2{}^2P^2\,\{|b_{41}| + |\,b_{42}| + |\,b_{43}\,|\,\} + h^2M_2{}^3P^3\,\{|\,b_{42}\,||\,b_{21}\,|+|\,b_{43}\,|\,(|b_{31}\,|+|\,b_{32}|\,)\,\} + h^3M_2{}^4P^4\,|\,b_{43}\,||\,b_{32}||\,b_{21}\,|\,)\,|\,{}^{L}u_j-{}^{L}v_j\,|\,.$$

Now using (4.17), (4.21), (4.22), and (4.23) in (4.14) and letting $h = 1$, we get

$$(4.24)\quad \|\vec{F}\{h,x_j,\vec{u}_j,\vec{f}\} - \vec{F}\{h,x_j\vec{v}_j,\vec{f}\}\| \le \|\vec{u}_j-\vec{v}_j\| \{$$
$$|R_1|C_1 + |R_2|C_1 + |R_2|C_2|b_{21}| + |R_3|C_1$$
$$+ |R_3|C_2(|b_{31}| + |b_{32}|) + |R_3|C_3|b_{32}||b_{21}|$$
$$+ |R_4|C_1 + |R_4|C_2(|b_{41}| + |b_{42}| + |b_{43}|)$$
$$+ |R_4|C_3(|b_{42}||b_{21}| + |b_{43}|\{|b_{31}| + |b_{32}|\})$$
$$+ |R_4|C_4|b_{43}||b_{32}||b_{21}|\}$$

where $C_1 = M_2P$, $C_2 = (M_2P)^2$, $C_3 = (M_2P)^3$ and $C_4 = (M_2P)^4$. Letting C_{max} be the largest valued C_i, $(i = 1, \ldots, 4)$, we finally have

$$(4.25)\quad \|\vec{F}\{h,x_j,\vec{u}_j,\vec{f}\} - \vec{F}\{h,x_j,\vec{v}_j,\vec{f}\}\| \le C\|\vec{u}_j-\vec{v}_j\|$$

where

$$(4.26)\quad C = C_{max}[(|R_1| + |R_2| + |R_3| + |R_4|)$$
$$+ (|R_2||b_{21}| + |R_3|\{|b_{31}| + |b_{32}|\}$$
$$+ |R_4|\{|b_{41}| + |b_{42}| + |b_{43}|\})$$
$$+ (|R_3||b_{32}||b_{21}|$$
$$+ |R_4|(|b_{42}||b_{21}| + |b_{43}|\{|b_{31}| + |b_{32}|\}))$$
$$+ |R_4||b_{43}||b_{32}||b_{21}|] .$$

Thus property (4.2) is established.

Runge-Kutta Process Stability

We now show stability of the Runge-Kutta single-step method by stating and proving the following theorems.

Theorem 4.1. Method (3.24) is stable, method (3.20) is stable, and method (3.20) is L-stable if property (4.2) and the root condition are satisfied.

Proof: The proof will be presented in a manner similar to that of Luther [7]. Also because of lemma A we need only to prove stability of method (3.24). It has been shown that property (4.2) is satisfied by the Runge-Kutta method and also that the root condition is satisfied by definition 4.4. Further by stability definition 4.1 we have $||\vec{\rho}_k - \vec{\rho}'_k|| \leq \varepsilon$, $0 \leq k \leq N$.

We now seek for $\vec{f}$ belonging to F an h_o and M, independent of h, such that for $0 \leq h \leq h_o$, $||\vec{u}^*_k - \vec{u}'^*_k|| \leq M\varepsilon$, $0 \leq k \leq N$.

Hence let $\vec{E}_k = \vec{u}^*_k - \vec{u}'^*_k$, $0 \leq k \leq N$, and $\omega_k = \max_{0 \leq j \leq k} ||\vec{E}_j||$, $0 \leq k \leq N$.

Then,

$$\vec{E}_0 = \vec{u}^*_0 - \vec{y}_0 - [\vec{u}'^*_0 - \vec{y}_0] = \vec{\rho}_0 - \vec{\rho}'_0 \quad .$$

Thus,

$$\omega_0 = \| \vec{E}_0 \| = \| \vec{\rho}_0 - \vec{\rho}'_0 \| \leq \varepsilon .$$

Then for $0 \leq k \leq N - 1$, we have

$$(4.41) \qquad \vec{E}_{k+1} - \vec{E}_k = \vec{C}_{k+1} = h[\vec{F}\{h,x_k,\vec{u}^*_k,\vec{f}\} - \vec{F}\{h,x_k,\vec{u}'^*_k,\vec{f}\}] + h\{\vec{\rho}_{k+1} - \vec{\rho}'_{k+1}\} .$$

Now by theorem 3.1, for $1 \leq k \leq N$, we have

$$(4.42) \qquad \vec{E}_k = \sum_{\nu=0}^{n-1} \vec{E}_\nu W_k^{(\nu)} + \sum_{j=0}^{k-n} \vec{C}_{j+n} W_{k-j-1}^{(n-1)} .$$

But $n = 1$ in the single-step method, hence we have

$$(4.43) \qquad \vec{E}_k = \vec{E}_0 W_k^{(0)} + \sum_{j=0}^{k-1} \vec{C}_{j+1} W_{k-j-1}^{(0)}$$

Using in (4.43) $\vec{C}_{j+1}$ as defined by (4.41) and applying norms yields

$$(4.44) \qquad \| \vec{E}_k \| \leq |W_k^{(0)}| \, \| \vec{E}_0 \| + kh \, |W_{k-j-1}^{(0)}| \max_{0 \leq j \leq k-1} \| \vec{F}(h,x_j,\vec{u}^*_j,\vec{f}) - \vec{F}(h,x_j,\vec{u}'^*_j,\vec{f}) \| + kh \, |W_{k-j-1}^{(0)}| \, \| \vec{\rho}_{j+1} - \vec{\rho}'_{j+1} \| .$$

For the single-step method, $|W_k^{(0)}| = 1$. Thus,

$$(4.45)\qquad \|\vec{E}_k\| \leq \varepsilon + khC \max_{0\leq j\leq k-1} \|\vec{u}^*_j - \vec{u}'^*_j\| + kh\varepsilon$$
$$= \varepsilon + khC\,\omega_k + kh\varepsilon \ .$$

But we note now that $\|\vec{E}_j\| = \omega_k$ for some $j \leq k$. Also note that the C above is the constant in (4.2). Hence (4.45) becomes

$$(4.46)\qquad \omega_k \leq \varepsilon + kh\varepsilon + khC\omega_k \ .$$

Now limit values of k such that $khC \leq \frac{1}{2}$ or

$$(4.47)\qquad k \leq [\tfrac{1}{2hC}] = t,$$

where $[r]$ denotes largest integer not exceeding r. Now we find from (4.46) and (4.47) that

$$(4.48)\qquad \tfrac{1}{2}\omega_k \leq \varepsilon + kh\varepsilon \leq \varepsilon + \tfrac{1}{2C}\,\varepsilon$$

or

$$(4.49)\qquad \omega_k \leq 2[1+\psi]\,\varepsilon = M_1\varepsilon \ , \ 0 \leq k \leq t_1$$

where $\psi = \frac{1}{2C}$.

Of course (4.49) implies that $0 \leq t_1 \leq N$. If not, then (4.49) holds for $0 \leq k \leq N$.

By continuing to repeat this procedure, we will eventually bound ω_k for $0 \leq k \leq N$.
Thus letting $M_o = 1$, we have,

$$(4.50)\qquad \omega_k \leq 2[M_0+\psi]\varepsilon = M_1\varepsilon \quad , \ 0 \leq k \leq t_1$$

$$\omega_k \leq 2[M_1+\psi]\varepsilon = M_2\varepsilon \quad , \ t_1 \leq k \leq 2t_1$$

$$\vdots$$

$$\omega_k \leq 2[M_q+\psi]\varepsilon = M_{q+1}\varepsilon \quad , \ qt_1 \leq k \leq N \leq (q+1)t_1$$

and in general,

$$(4.51)\qquad \|\vec{E}_k\| \leq \omega_k \leq [2^{(q+1)} + 2\psi(2^{(q+1)}-1)]\varepsilon = M\varepsilon$$

for $0 \leq k \leq N$.

We now proceed to prove (4.51) by induction.

(a) We know (4.51) is true for $q = 0$; that is,

$$M_1\varepsilon = [2+2\psi]\varepsilon = 2[1+\psi]\varepsilon .$$

And this is identical to the first equation of (4.50).

(b) Next we assume (4.51) holds for $q = s$. Hence,

$$M_s\varepsilon = [2^{s+1} + 2\psi(2^{(s+1)}-1)]\varepsilon .$$

(c) Now we let $q = s+1$. From (4.50) we have

$$M_{s+1}\varepsilon = 2[M_s+\psi]\varepsilon .$$

Substituting the value M_s from (b) we find

$$M_{s+1}\varepsilon = [2^{(s+2)} + 2\psi(2^{(s-2)}-1)]\varepsilon .$$

Since (4.51) holds for s+1, the proof is complete. What remains to be shown in (4.51) is that M is independent of h. This is done by showing that q is

independent of h and hence, M is independent of h.

Thus we define

$$p = [2C(b-a)] \leq 2C(b-a) \tag{4.52}$$

Also let

$$h_o \leq \frac{1}{2C(p+1)} \tag{4.53}$$

Then

$$\frac{p}{2C} \leq b-a \leq \frac{p+1}{2C} \quad . \tag{4.54}$$

Note that from (4.53) and with $h \leq h_o$, we obtain

$$\frac{1}{2hC} \geq \frac{1}{2h_oC} \geq p+1 \tag{4.55}$$

or

$$t_1 \geq p+1 \quad . \tag{4.56}$$

Now since $t_1 = [\frac{1}{2hC}]$ we have,

$$pt_1 \leq [\frac{p}{2hC}] \leq \frac{b-a}{h} = N \leq [\frac{p+1}{2hC}] < (p+1)t_1 + (p+1) \leq (p+2)t_1 \quad . \tag{4.57}$$

Hence we get

$$pt_1 \leq N \leq (p+2)t_1 \quad . \tag{4.58}$$

But from the last equation of (4.50) we have

$$qt_1 \leq N \leq (q+1)t_1 \quad . \tag{4.59}$$

And from (4.58) and (4.59) we see that

(4.60) $\quad p \leq q \leq p+1 \quad .$

Thus q is independent of h. As for h_o we require only that h_o make method (3.24) unique.

Theorem 4.2. If property (4.2) and the root condition are satisfied, then method (3.24) is H-stable and method (3.20) is H-stable.

Proof: It has been shown that property (4.2) is satisfied and also that the root condition is satisfied.

From the definition of H-stability we have that $\| \vec{u}^*_o \| \leq \varepsilon$ and $\| \vec{\rho} \| \leq \varepsilon$, $1 \leq k \leq N$.

We now seek an h_o and $M(\varepsilon)$ such that for $0 \leq h \leq h_o$ we obtain $\max_{0 \leq k \leq N} \| \vec{u}^*_k \| \leq M(\varepsilon)$.

Hence let

(4.61) $\quad \omega_j = \max_{0 \leq k \leq j} \| \vec{u}^*_k \| \quad , 0 \leq k \leq N.$

Then $\omega_o \leq \varepsilon$.

Also for $0 \leq k \leq N-1$, we have

(4.62) $\quad \vec{u}^*_{k+1} - \vec{u}^*_k = h\vec{F}(h, x_k, \vec{u}^*_k, \vec{f}) + h\vec{\rho}_{k+1}$.

Now in the same manner as in the previous theorem we have

(4.63) $\quad \| \vec{u}^*_j \| \leq \omega_j \leq \varepsilon + (b-a)(\varepsilon+K) = M(\varepsilon).$

Again we require an h_o that makes method (3.24) unique.

CHAPTER V

STABILITY OPTIMIZATION

In this chapter we discuss the main purpose of this study. Briefly restated, we seek to find Runge-Kutta formulas which will yield optimum stability when the Runge-Kutta method is used. We are interested in the two types of stability defined by definition 4.1 and definition 4.3. From (4.51) it is easily seen that optimum stability will be achieved by finding the smallest possible q. By (4.60) we see that if p is minimized q will be minimized. And from (4.52) we have that p is dependent on the Lipschitz constant C of property (4.2). Hence our problem becomes one of minimizing C. For optimum H-stability we have from equation (4.63) that we need to minimize the constant K of property (4.1)

Minimizing the K Constant

From equation (4.11) of Chapter IV we found that for the fourth order Runge-Kutta process we have

$$K = M_1 \{ |R_1| + |R_2| + |R_3| + |R_4| \} \tag{5.1}$$

where M is a constant.

In this study we will require $0 < a_2, a_3 < 1$. Since we seek to minimize K, we see that our task is to minimize $\sum_{i=1}^{4} |R_i|$. Hence let us now define

$$R_{min} = \min [\max |R_i|] . \tag{5.2}$$

In Chapter II it was stated that for the singular case there are only three possibilities which give finite solutions of R_i. From these singular cases, we see that for $a_2=a_3=\frac{1}{2}$, we have $R_{min}=R_2=R_3=\frac{1}{3}$ and $R_1=R_4=\frac{1}{6}$. These parameters yield the formula due to Runge:

$$\vec{y}_{j+1} = \vec{y}_j + \frac{1}{6}(\vec{k}_1+2\vec{k}_2+2\vec{k}_3+\vec{k}_4) \tag{5.3}$$

where

$$\begin{aligned} \vec{k}_1 &= h\vec{f}(x_j,\vec{y}_j) \\ \vec{k}_2 &= h\vec{f}(x_j+\tfrac{1}{2}h,\ \vec{y}_j+\tfrac{1}{2}\vec{k}_1) \\ \vec{k}_3 &= h\vec{f}(x_j+\tfrac{1}{2}h,\ \vec{y}_j+\tfrac{1}{2}\vec{k}_2) \\ \vec{k}_4 &= h\vec{f}(x_j+h,\ \vec{y}_j+\vec{k}_3) . \end{aligned} \tag{5.4}$$

Next we recall the first equation of (2.12) which is

$$R_1 + R_2 + R_3 + R_4 = 1 \tag{5.5}$$

Since we are interested in R_{min} for any set a_2 and a_3, it is easily seen that if one or more R_i's are permitted to be negative, the sum of the remaining R_i will increase in order to satisfy (5.5). For example, if we allow R_1 to be negative we find

(5.6) $\quad R_2 + R_3 + R_4 > 1 .$

Then R_{min} will be greater than $\frac{1}{3}$. Thus we conclude that for R_{min} we must have $R_i \geq 0$. Hence we have

(5.7) $\quad |R_1| + |R_2| + |R_3| + |R_4| = 1 .$

Using (5.7) in (5.1) we find the minimum K to be

(5.8) $\quad K = M_1 .$

Therefore, for optimum H-stability we only require that the fourth order formulas have positive R_i . Well known formulas which meet this requirement is the one due to Runge, (5.3) and the one due to Kutta, (2.29). There are of course other formulas also satisfying this criteria.

It is now interesting to note that $R_{min} = \frac{1}{3}$ for all $0 < a_2, a_3 < 1$.

Minimum $[\max|R_i|]$

We know from above that R_{min} requires $R_i \geq 0$.

We also know that for the singular case $R_{min} = \frac{1}{3}$.

For the non-singular case ($a_2 \neq a_3$) in the region where $0 < a_2, a_3 < 1$, we now show that $R_{min} = \frac{1}{3}$.

Using (2.23) we have $R_1 = 0$ iff

(5.9) $\quad 6a_2a_3 - 2a_3 - 2a_2 + 1 = 0.$

This is the equation of a hyperbola for which the asympotes are $a_2 = \frac{1}{3}$ and $a_3 = \frac{1}{3}$. Its intercepts occur at ($a_2 = \frac{1}{2}$, $a_3 = 0$) and ($a_2 = 0$, $a_3 = \frac{1}{2}$) . Also this hyperbola intercepts the region boundary at ($a_2 = 1$, $a_3 = \frac{1}{4}$) and ($a_3 = 1$, $a_2 = \frac{1}{4}$) . Thus R_1 is zero for all points on the hyperbola and found to be positive for all values of a_2 and a_3 in the region designated by the plus sign in figure 5.1. That is, $R_1 > 0$ iff

(5.10) $\quad 6a_2a_3 - 2a_3 - 2a_2 + 1 > 0$.

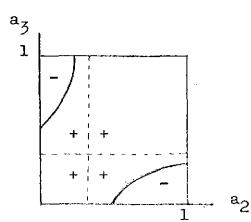

Fig. 5.1. Region for positive R_1

Next, from (2.23) we find $R_2 = 0$ iff $a_3 = \frac{1}{2}$ and $R_3 = 0$ iff $a_2 = \frac{1}{2}$. Also $R_2, R_3 \geq 0$ iff $0 < a_2 < \frac{1}{2} \leq a_3 < 1$ or $0 < a_3 < \frac{1}{2} < a_2 < 1$.

This is illustrated in figure 5.2 by the cross hatched region.

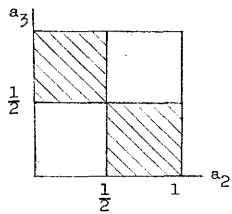

Fig. 5.2. Region for positive R_2 and positive R_3

We next see that $R_4 = 0$ iff

(5.11) $\quad 6a_2a_3 - 4a_3 - 4a_2 + 3 = 0$.

This is the equation of a hyperbola for which the asympotes are $a_2 = \frac{2}{3}$ and $a_3 = \frac{2}{3}$. Its intercepts are found to be $(a_2 = \frac{3}{4}, a_3 = 0)$ and $(a_2 = 0, a_3 = \frac{3}{4})$. Further, this hyperbola intercepts the region boundary at $(a_2 = 1, a_3 = \frac{1}{2})$ and $(a_3 = 1, a_2 = \frac{1}{2})$. From figure 5.3 we see that

R_4 will be positive in the region denoted by the plus sign. That is $R_4 > 0$ iff

(5.12) $\quad 6a_2a_3 - 4a_3 - 4a_2 + 3 > 0$.

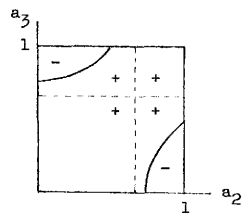

Fig. 5.3. Region for positive R_4

Now by considering figures 5.1, 5.2, and 5.3 we obtain figure 5.4 whose cross hatched region (boundaries included) represents the region we seek such that $R_i \geq 0$ for $0 < a_2, a_3 < 1$.

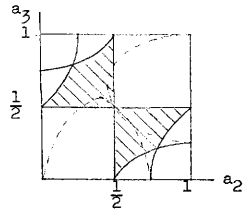

Fig. 5.4. Region for positive R_i

From the above discussion it is clear that $R_{min} = \frac{1}{3}$ on the boundary of the region described by figure 5.4.

Finally we examine the interior of the region in figure 5.4. Consider first the part of the region where $0 < a_2 \leq \frac{1}{2}$ and $\frac{1}{2} \leq a_3 < 1$, $(a_2 \neq a_3)$. Recall

(5.13) $\quad R_1 = \frac{1}{2} + [1-2(a_2+a_3)] \,/\, 12a_2a_3$.

Now when $a_2 = \frac{1}{2}$ or $a_3 = \frac{1}{2}$, (5.13) yields $R_1 = \frac{1}{6}$.

Suppose that $R_1 > \frac{1}{6}$. Then we have

(5.14) $\quad \frac{1}{2} + [1-2(a_2+a_3)\ ] \,/\, 12a_2a_3 > \frac{1}{6}$

which gives $a_3 < \frac{1}{2}$, a value of a_3 outside our region of interest and hence it becomes obvious that $R_1 \leq \frac{1}{6}$. In the same manner we find $R_4 \leq \frac{1}{6}$.

Consider next the part of the region where $\frac{1}{2} \leq a_2 < 1$ and $0 < a_3 \leq \frac{1}{2}$, $(a_2 \neq a_3)$. Using the same procedure as above we again obtain $R_1 \leq \frac{1}{6}$ and $R_4 \leq \frac{1}{6}$. It then follows that for our entire region

(5.15) $\quad R_1 + R_4 \leq \frac{1}{3}$.

Whence

(5.16) $\quad R_2 + R_3 \geq \frac{2}{3}$.

It is now easily seen that from (5.16) $R_{min} = \frac{1}{3}$, which is what we set out to show.

Next we proceed to show that for our region of interest, (5.3) is the only formula for which $b_{ij} \geq 0$ as well as being the formula for which $R_i \geq 0$. From the above discussion it is clear (5.3) satisfies $R_i \geq 0$. We now show it satisfies $b_{ij} \geq 0$.

Using (2.25), we find (see [7])

(5.16') $\quad b_{31} = a_3(3a_2 - 4a_2^2 - a_3) \,/\, [2a_2(1-a_2)\,]$

$\qquad\quad b_{41} = Q \,/\, [2a_2a_3(6a_2a_3 - 4(a_2+a_3) - 3\,)\,]$

where

$$Q = 12a_2^2a_3^2 - 12a_2^2a_3 - 12a_2a_3^2 + 15a_2a_3 + 4a_3^2 - 6a_2 - 5a_3 + 2 - 4a_2^2 .$$

Now using $b_{31} \geq 0$ and $b_{32} \geq 0$, we find

$a_2 < \frac{1}{2}$ and $3a_2 - 4a_2^2 > a_3 > a_2$ or $a_2 > \frac{1}{2}$ and $3a_2 - 4a_2^2 < a_3 < a_2$ or a_2. From $6a_3a_2 - 4(a_2+a_3) + 3 > 0$ and $b_{42} \geq 0$ we have $a_3 > a_2$ and $a_2 > 2 - 5a_3 + 4a_3^2$ or $a_3 < a_2$ and $a_2 < 2 - 5a_3 + 4a_3^2$. From $b_{43} \geq 0$ we have $a_2 < \frac{1}{2}$ and $a_3 > a_2$ or $a_2 > \frac{1}{2}$ and $a_3 < a_2$. Note that $a_3 = 3a_2 - 4a_2^2$ and $a_2 = 2 - 5a_3 + 4a_3^2$ are tangent at $(\frac{1}{2}, \frac{1}{2})$. When we include $6a_3a_2 - 2(a_2+a_3) + 1 \geq 0$, we find that unless $b_{41} \geq 0$ changes the result, we must have a_2 and a_3 in the region bounded by the two hyperbolas and the two parabolas (see fig. 5.4).

Turn now to $b_{41} \geq 0$. If we let $a_2 = \frac{1}{2} + u$, $a_3 = \frac{1}{2} - v$, (see (5.16')) we find

$$Q = 12u^2v^2 + v^2 - uv - \frac{u}{2} - \frac{v}{2}$$

where $0 \leq u$, $v \leq \frac{1}{2}$. Using

$$v = \left[u + \frac{1}{2} + (24u^3 + u^2 + 3u + \frac{1}{4})^{\frac{1}{2}}\right] / \left[24u^2 + 2\right]$$

we readily find $Q < 0$ for the region, except at $u = v = 0$. The region found just above for the other parameters to be nonzero is seen to be within

this last; so that $u = v = 0$ is the only choice for all parameters to be non-negative. This of course means $a_2 = a_3 = \frac{1}{2}$.

Minimizing the C Constant

While in the preceding section we found a requirement for Runge-Kutta formulas which give optimum H-stability, here we seek a formula which will optimize the stability as given by definition 4.1. To accomplish this, it has been shown at the beginning of this chapter that we desire to minimize the Lipschitz constant C of property (4.2), previously determined for the fourth order Runge-Kutta to be

$$
\begin{aligned}
(5.17)\qquad C = C_{max}\,[\,&(|R_1| + |R_2| + |R_3| + |R_4|\,) \\
&+ (\,|R_2||b_{21}| \\
&+ |R_3|\{|b_{31}| + |b_{32}|\} + |R_4|\{|b_{41}| + |b_{42}| \\
&\quad + |b_{43}|\,)) \\
&+ (|R_3||b_{32}||b_{21}| + |R_4|\{|b_{42}||b_{21}| \\
&\quad + |b_{43}|\{|b_{31}| \\
&+ |b_{32}|\}\,)\,) + |R_4||b_{43}||b_{32}||b_{21}|\,]\ .
\end{aligned}
$$

where C_{max} is a constant. Using (2.12) and (2.13)

we get

$$(5.18)\quad |R_1| + |R_2| + |R_3| + |R_4| \geq$$

$$|R_2|\{|b_{21}| + |R_3|\{|b_{31}| + |b_{32}|\}$$

$$+ |R_4|\{|b_{41}| + |b_{42}| + |b_{43}|\} \geq \frac{1}{2}$$

$$|R_3||b_{32}||b_{21}| + |R_4|(|b_{42}||b_{21}|$$

$$+ |b_{43}|\{|b_{31}| + |b_{32}|\}) \geq \frac{1}{6}$$

$$|R_4||b_{43}||b_{32}||b_{21}| = \frac{1}{24}$$

It seems clear that to minimize (5.17) we have to minimize the relationships of (5.18). This will be accomplished easily when $R_i \geq 0$ and $b_{ij} \geq 0$. One formula which satisfies this requirement is Runge's formula given in (5.3). Then the minimum C is given as

$$(5.19)\quad C = C_{max}\left[1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24}\right] = \frac{41}{24}\, C_{max}\,.$$

It is important to note that for our region of interest, $0 < a_2, a_3 < 1$, Runge's formula is the only formula which meets the criteria. This was shown in the previous section.

CHAPTER VI

CONCLUSION

In many engineering and scientific problems it is necessary to consider numerical procedures for obtaining an approximate solution to an ordinary differential equation. One numerical method useful for this purpose is the Runge-Kutta process. However numerical methods raise questions of their own.

One important question is that of stability of the method. The Runge-Kutta method is indeed stable as we have shown in this study. Karim and Lawson have found regions of stability for the Runge-Kutta method of order four and higher. However, neither indicates that their choice of Runge-Kutta formulas is the one which gives optimum stability in our sense of having a minimum bound. It might prove worth-while to investigate, for example, Karim's work to see if indeed his fourth order Runge-Kutta formula is the one for optimum stability according to his stability definition.

In conclusion we again emphasize the main purpose of this study. This is that when the fourth-

order Runge-Kutta method is used to approximate a solution we are interested in having optimum stability. Hence that method which has the minimum bound in our definition of stability is in this sense best. Thus we conclude that any fourth-order Runge-Kutta formula having $R_i \geq 0$ will give optimum H-stability. For optimum stability of our second definition, we need fourth-order Runge-Kutta formulas having $R_i \geq 0$ and $b_{ij} \geq 0$. In particular, Runge's formula meets this criteria. Moreover, this is the only formula meeting their criteria when $0 < a_2, a_3 < 1$.

REFERENCES

[1] P. Henrici, Discrete Variable Methods in Ordinary Differential Equations, Wiley, New York, 1962.

[2] C. Runge, Über die numerische Auflösung von Differentialgleichungen, Math. Ann., 46 (1895), 167-178.

[3] K. Heun, Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen, Z. Math. Phys., 45 (1900), 23-38.

[4] W. Kutta, Beitrag zur näherungsweisen Integration totaler Differentialgleichungen, Z. Math. Phys., 46 (1901), 435-453.

[5] J. C. Butcher, Coefficients for the study of Runge-Kutta integration processes, J. Austral. Math. Soc., 3 (1963), 185-201.

[6] H. A. Luther, Further Explicit Runge-Kutta Formulas, SIAM Rev., 8 (1966), pp. 374-380.

[7] _____, Unpublished class notes, Texas A&M University, 1968.

[8] C. R. Cassity, Solutions of the fifth order Runge-Kutta equations, SIAM J. Numer. Anal., 3 (1966), 598-606.

[9] S. Gill, A process for the step-by-step integration of differential equations in an automatic digital computing machine, Proc. Cambridge Philos. Soc., 47 (1951), 96-108.

[10] M. Lotkin, On the accuracy of Runge-Kutta's method, Math. Tables Aids Comput., 5 (1951), 128-132.

[11] A. Ralston, Runge-Kutta methods with minimum error bounds, Math. Comp., 16 (1962), 431-437.

[12] J. W. Carr, Error bounds for the Runge-Kutta single step integration process, J. Assoc. Comp. Mach., 5 (1958), 39-44.

[13] Abbas I. Abdel Karim, The stability of the fourth order Runge-Kutta method for the solution of systems of Differential equations, Comm. ACM, 9 (1966), pp. 113-116.

[14] J. D. Lawson, An order five Runge-Kutta process with extended region of stability, SIAM J. Numer. Anal., 3 (1966), pp. 593-597.

[15] E. I. Ince, Ordinary Differential Equation, Dover, New York, 1956.

[16] E. Isaacson and H. B. Keller, Analysis of Numerical Methods, Wiley, New York, 1966.

VITA

Hector G. Sierra, son of Edmundo H. and Frances G. Sierra, was born on October 3, 1937, in San Antonio, Texas.

He attended San Antonio public schools and graduated from Thomas Jefferson High School in 1955. In June 1960 he received his Baccalaureate Degree, Bachelor of Arts in Mathematics, from the University of Texas. At graduation exercises he was commissioned a second Lieutanent in the United States Air Force and presently holds the rank of Captain.

He entered active duty in August 1960, and received his navigator's wings in 1961. Upon completion of USAF Electronic Warfare school, he was assigned to the Strategic Air Command at Castle AFB, California. While at Castle he completed Squadron Officer's School in residence during the Fall of 1964. This year he is being reassigned to an overseas tour in Thailand.

Capt. Sierra is married. He and his wife, Carlota, have three boys, Charles, Ricardo, and Michael. His permanent address is 327 Matthews St., San Antonio, Texas.

The typist for this thesis was Mrs. Jan Want.