AN ABSTRACT OF THE THESIS OF

RALPH MARVIN TOMS for the degree DOCTOR OF PHILOSOPHY
(Name) (Degree

in APPLIED MATHEMATICS presented on August 9, 1973
(Major) (Date)

Title: GLOBALLY OPTIMAL RUNGE-KUTTA METHODS

Abstract approved: Redacted for Privacy
(J. Davis)

A Runge-Kutta method has been developed to minimize the global error in the numerical solution of certain classes of differential equations problems. The distinguishing feature of the method is that the coefficients of the numerical integration formula depend on the initial conditions present at the time of solution. The method is intended for use in situations where a set of differential equations is to be solved repeatedly for different initial conditions. The method is particularly applicable to real time control system applications.

Globally Optimal Runge-Kutta Methods

by

Ralph Marvin Toms

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1974

APPROVED:

Redacted for Privacy

Associate Professor of Mathematics

In Charge of Major

Redacted for Privacy

Chairman of Department of Mathematics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented August 9, 1973

Typed by Sharon Holfeltz for Ralph Marvin Toms

ACKNOWLEDGEMENT

I am grateful to my thesis advisor, Joel Davis, for his encouragement and guidance in the preparation of this thesis.

I would also like to acknowledge the partial financial support provided by the Boeing Company through their doctoral fellowship program.

I would like to extend my appreciation to Edward Costello of the Boeing Company for his support in this endeavor.

This thesis is dedicated to my wife Gwen.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

GLOBALLY OPTIMAL RUNGE-KUTTA METHODS

I INTRODUCTION

Considerable effort has been expended to obtain efficient numerical procedures for solving systems of ordinary differential equations. Much of this effort is directed at the development of methods suitable for use in computer library subroutines. A recent paper by Hull and Others [4] summarizes the current status of these efforts. The development of all purpose routines is complicated by the fact that they must work in as wide a class of situations as is possible. Thus, the methods employed must be relatively insensitive to the particular problem to be solved.

There is a contrasting situation in which maximal efficiency is required. This is in the area of the real time applications associated with digital process control. For any particular control system the associated system of differential equations is not a general system but is quite specific. The functional form of the equations remains fixed throughout the solution of the problem. A method which is efficient in a general setting may not be the most appropriate selection in a specific case.

In this thesis we develop a numerical integration procedure which can be optimized for specific problems. We have chosen a Runge-Kutta method for this purpose. This choice was motivated by the fact that such methods are frequently employed in real time digital control system applications. In such an application the choice of the numerical integration method is often critical. A method is generally needed which is stable, provides sufficiently accurate results and requires a minimal amount of computer time. Runge-Kutta methods are used for this purpose since they have good stability properties, provide flexibilty in the choice of integration step size, are easy to code and are efficient when accuracy requirements are modest. The method we shall propose produces a very small global error even when a coarse integration step is used. This has the effect of producing accurate results for a minimal amount of computation time and is therefore particularly applicable to real-time applications.

The usual approach to increasing the accuracy of a numerical integration procedure is to increase the order of the formula or to decrease the step size. We purposely avoid either of these approaches. Low order methods have several practical advantages. They allow the utmost flexibility in the selection of the integration step size. This is particularly apropos of the situation where discontinuities are present in the coefficients of the differential equations. The combination of a low order formula coupled with a coarse integration step has obvious computational advantages provided that the results are sufficiently accurate. As an alternative to reducing the step size or increasing the order of the method we develop a procedure that uses optimally weighted formulas to increase assuracy. This is accomplished by adjusting a free parameter appearing in the general second order

Runge-Kutta method so that the resultant formula is optimally weighted.

As is usually the case, the word optimal must be viewed with a degree of caution. What is optimal in one situation may be ineffective in another. There are several references in the literature to optimal Runge-Kutta methods [1, 5, 6, 7, 9]. In each of these the free parameter or parameters are selected to minimize certain terms appearing in the expression for the local truncation error. The development is in every case for a general first order differential equation. Once the free parameter has been selected it remains fixed not only throughout the solution of a problem but from one problem to another. Certain specific values of the free parameter yield the classical second order Runge-Kutta methods such as the Heun method and the modified Euler method (see Henrici [2], p 67).

The optimal methods developed herein differ considerably from those developed previously. The main differences are that:

1. the free parameter is selected to minimize the total (global) error rather than the local truncation error,
2. the optimal value of the free parameter depends strongly on the particular differential equation (or system) to be solved,
3. the optimal value of the parameter depends on the initial conditions of the problem and, possibly, on other parameters appearing in the problem.

It is the dependence of the Runge-Kutta formula on the initial conditions of the problem that is the distinguishing feature of our method. The basic idea is to first solve the differential equations for a sample set of initial conditions (usually taken from some compact set). These sample solutions are obtained using any convenient numerical procedure with as small an integration step as is needed. For each of the initial conditions of the selected sample, optimal weighting parameters are determined so that the total error is small (usually zero). The optimal parameter is then treated as a function of the initial condition. Standard approximation techniques are used to fit the optimal parameter as a function of the initial condition. Thus, in its real time application, the numerical integration procedure uses a dynamically computed weighting factor which depends on the initial conditions at the time of solution.

This method is clearly intended for use in very specific situations. These are characterized by the following conditions:

1. the system of differential equations is to be solved repeatedly for different initial conditions,
2. computational speed is at a premium,
3. the total error at the end of a specified interval is to be minimized,
4. it is economically feasible to perform the required optimization calculations.

II BACKGROUND AND PRELIMINARY ANALYSIS

II.1 <u>The Basic Problem</u>

In this chapter we will consider a single differential equation in one unknown function. Most of the analysis will be applicable, with certain modifications, to a system of differential equations. Systems will be discussed in a subsequent chapter.

It will be assumed that the differential equation has been transformed so that the independent variable lies in the unit interval. The problem under consideration is defined by,

$$\begin{aligned} Y' &= f(t , Y) \\ Y(o) &= Y_o \end{aligned} \qquad (2\text{-}1)$$

where f is a real valued function, $-\infty < a \leq Y_o \leq b < \infty$ and $t \in [0 , 1]$. It will be assumed throughout that for each $Y_o \in [a , b]$ there is a unique function $Y(t)$ which is continuous on $[0 , 1]$ and which satisfies the differential equation. We are therefore considering a family of continuous solutions to the problem (2-1) whose initial values lie in $[a , b]$. We will use the notation $Y(t , Y_o)$ to denote a specific member of the family. If Y_o is fixed for a particular argument we will normally denote the solution corresponding to Y_o as $Y(t)$. We will assume that $Y(t , Y_o)$ is a continuous function of Y_o. Derivatives of Y with respect to t will be denoted by the usual prime notation.

Taking the integration step as h we introduce a uniform grid

I_h on $[0 , 1]$ where,

$$I_h = \{t_i \mid t_o = 0,\ t_{i+1} = t_i + h, t_N = 1, \quad i=0,\ 1,\ \ldots,\ N-1\}$$

This definition implies that $h = 1/N$ where N is the number of integration steps. A uniform grid is used for algebraic convenience only; the procedure we shall develop can also be used with a non-uniform grid. The exact solution evaluated at a grid-point will be denoted by $Y(t_i)$ or Y_i. The numerical solution of (2-1) is a function $\{U_i\}$ mapping I_h into the reals. The value of the numerical solution at the ith grid-point is U_i. Naturally we would like the U_i to approximate Y_i for all grid-points.

Throughout the discussion we will make use of the definitions both to establish notation and concepts.

II.2 Review of the Method

In order to motivate the analysis which follows we shall briefly review the second order Runge-Kutta Method. The derivation of the method is given in several references, e.g., Hildebrand [3]. The following definition will be needed.

Definition 2.1

i. $f_t(t , y)$, $f_{ty} (t , y)$, ... denotes partial derivatives; we may omit the arguments and write f_t , f_{ty}, ...

ii. $f'(t ,y) = f_t (t , y) + f_y (t , y) f (t , y)$

The basic procedure consists of computing the numerical solution of (2-1) from the formula,

$$U_{i+1} = U_i + \frac{h}{2c} \Big[(2c - 1)\ f(t_i\ ,\ U_i) + \qquad (2\text{-}2)$$

$$f\big(t_i + ch,\ U_i + ch\ f(t_i\ ,\ U_i)\big)\Big]$$

$$i = 1,\ 2,\ \ldots,\ N\text{-}1$$

where c is a nonzero real parameter. This is the free parameter mentioned previously. The well known modified Euler method is obtained for $c = 1/2$ while for the method of Heun $c = 1$ (Henrici [2]). In the optimal method developed by Ralston [9] c turns out to be 2/3. Later we shall generalize the method to also include the case where $c = 0$.

The local truncation error of the method is given by

$$\frac{h^3}{12} \Big[2Y_i''' - 3c \big(Y_i''' - Y_i''\ f_y(t_i\ ,\ Y_i) \big) \Big] + 0\ (h^4) \qquad (2\text{-}3)$$

It should be noted that usually no choice of c will make (2-3) vanish at every grid-point for arbitrary f. Note also that the truncation error depends not only on drivatives of the solution Y but also on f_y. This is typical of Runge-Kutta methods; in many other numerical methods the error depends only on derivatives of Y.

It can be seen from (2-3) that if $Y''' - Y''\ f_y$ vanishes identically the principle part of the truncation error does not depend on c. There are cases where the entire truncation error of the method does not depend on c.

As an example of this type of behavior consider the differential equation,

$$Y' = a_1 + a_2\ t + a_3\ Y \qquad (2\text{-}4)$$

where the coefficients are constant. It is easy to see that $Y''' - Y''\ f_y$ is identically zero. In this case formula (2-2) reduces to,

$$\begin{aligned} U_{i+1} = U_i + h \left[a_1 + a_2\ t_i + a_3\ U_i \right] + \\ \frac{h^2}{2} \left[a_2 + a_2\ (a_1 + a_2\ t_i + a_3\ U_i) \right] \end{aligned} \qquad (2\text{-}5)$$

Using Definition 2.1,

$$f(t_i\ ,\ U_i) = a_1 + a_2\ t_i + a_3\ U_i$$

and

$$f'\ (t_i\ ,\ U_i) = a_2 + a_3\ (a_1 + a_2\ t_i + a_3\ U_i)$$

so that (2-5) can be written,

$$U_{i+1} = U_i + h\ f(t_i\ ,\ U_i) + \frac{h^2}{2}\ f'\ (t_i\ ,\ U_i)$$

This formula is recognizable as a three term power series approximation and the method **is approximate analytical continuation. Note** that the numerical solution is independent of c. It should be noted also that the reduction to a power series method can also occur for higher order Runge-Kutta methods.

In the subsequent discussion, sufficient conditions on the problem (2-1) will be given for the existence and uniqueness of a value of c which minimizes the total error of the method. In addition, simple transformations of (2-1) will be suggested which will ensure these conditions.

II.3 Preliminary Analysis

In the next sections sufficient conditions are given so that the method converges for all c in a prescribed compact interval. In order to include negative values for c in the analysis we will need to enlarge the domain of f to include values of t outside $[0 , 1]$.

Definition 2.2

Let S be the infinite strip defined by,

$$S = \{(t , y) \mid -\varepsilon \leq t \leq 1 + \varepsilon, y\}$$

where $\varepsilon > 0$ is a fixed constant.

Definition 2.3

The function f appearing in the differential equation (2-1) satisfies Property A if:

i. $f : S \subset R^2 \to R$ where R denotes the real numbers,

ii. f and its partial drivatives through order two exist and are continuous on S.

Definition 2.4

i. $G (t , y , c) = \frac{1}{c} \left[f \left(t + ch, y + ch\ f(t , y) \right) - f (t, y) \right]$

when $c \neq 0$

$G (t , y , o) = f' (t , y)$ when $c = o$

ii. $G_y(t, y, 0) = f_{ty}(t, y) + f_{yy}(t, y) f_y(t, y)$

If some of the arguments of G are understood from context we may omit them.

Definition 2.5

Let λ be a fixed positive constant and $c \in [-\lambda, \lambda]$. The general second order Runge-Kutta method is defined by

$$U_{i+1}(c) = U_i(c) + h f\left(t_i, U_i(c)\right) + \frac{h}{2} G\left(t_i, U_i(c), c\right) \qquad (2\text{-}6)$$

$$i = 0, 1, 2, \ldots, N-1$$

We shall often omit some of the arguments and write, for example,

$$U_{i+1} = U_i + h f(t_i, U_i) + \frac{h}{2} G(t_i, U_i)$$

We now establish a simple Lemma which guarantees that for small enough h the points $t_i + ch$ are in $[-\varepsilon, 1 + \varepsilon]$.

Lemma 2.1

Let $\varepsilon > 0$ and $\lambda > 0$ be fixed constants. Then there is an $h_\varepsilon \leq 1$ such that for all $h \leq h_\varepsilon$, $t + ch \in [-\varepsilon, 1 + \varepsilon]$ for every $c \in [-\lambda, \lambda]$ and $t \in [0, 1]$

Proof:

Take $h_\varepsilon = \min \{1, \varepsilon/\lambda\}$.

Lemma 2.2

Let f satisfy Property A and $h \leq h_\varepsilon$ then $G(t , y , c)$ is continuous for all points (t, y, c) in $S \times [-\lambda , \lambda]$.

Proof:

Let (t^* , y^*) be in S, $c \in [-\lambda , \lambda]$ and $c \neq 0$, then G is continuous at (t^*, y^*, c) since it is the finite composition of continuous functions. F or the case $c = 0$ we have, by Definition 2.4,

$$G(t , y, c) = \frac{1}{c}\left[f\left(t + ch, y + ch\, f(t , y)\right) - f\,(t , y)\right]$$

where $\left(t + ch, y + ch\, f\,(t , y)\right)$ is in S since $h < h_\varepsilon$ (Lemma 2.1) and $f{:}S \rightarrow R$.

Since f satisfies Property A we can apply Taylor's Theorem in R^2 to obtain

$$G\,(t , y , c) = f_t\left(t + \theta ch, y + \theta ch\, f(t , y)\right) + f_y\left(t + \theta ch, y + \theta ch\, f\,(t , y)\right) f(t , y)$$

where $0 < \theta < 1$. Since S is a convex set $\left(t + \theta ch, y + \theta chf(t,y)\right)$ is in S. Using the fact that f, f_t and f_y are continuous on S and Definition 2.4 we take limits on both sides to obtain,

$$\lim_{\substack{c \rightarrow o \\ t \rightarrow t^* \\ y \rightarrow y^*}} G(t , y , 0) = f_t\,(t^*, y^*) + f_y\,(t^*, y^*)\, f\,(t^*, y^*)$$

$$= f'\,(t^*, y^*) = G\,(t^* , y^* , 0)$$

and G is continuous at $(t^*, y^*, 0)$.

Lemma 2.3

If f satisfies Property A, $U_o = Y_o$ and $h \leq h_\varepsilon$ then $U_i(c)$ is a continuous function of c on $[-\lambda, \lambda]$, $i = 0, 1, \ldots, N$.

Proof:

The proof will be by induction. Since $U_o = Y_o$ by Definition 2.5

$$U_1(c) = Y_o + h\, f(t_o\ , Y_o) + \frac{h}{2} G\ (t_o\ , Y_o\ , c)$$

For $c \neq 0$, $U_1(c)$ is the composition of continuous functions and is therefore continuous. Taking the limit on both sides, using Lemma 2.2 and Definition 2.4 yields

$$\lim_{c \to o} U_1\ (c) = Y_o + h\ f\ (t_o, Y_o) + \frac{h^2}{2}\ f'\ (t_o, Y_o) = U_1\ (0)$$

Thus $U_1(c)$ is continuous on $[-\lambda, \lambda]$ for $h \leq h_\varepsilon$.

The induction hypothesis is that $U_i\ (c)$ is continuous. From Definition 2.5,

$$U_{i+1}\ (c) = U_i\ (c) + h\ f\ \left(t_i, U_i(c)\right) + \frac{h}{2} G\left(t_i, U_i\ (c), c\right)$$

From the induction hypothesis, the continuity of f and Lemma 2.2, $U_{i+1}(c)$ is continuous on $[-\lambda\ , \lambda]$.

Lemma 2.4

Let $\{\xi_i\}$ be a sequence of real numbers satisfying an inequality of the form,

$$|\xi_{i+1}| \leq A\ |\xi_i| + B \qquad i=0,\ 1,\ \ldots,\ N$$

where A and B are non-negative and independent of i. Then, if $A \neq 1$,

$$|\xi_i| \leq A^i\ |\xi_o| + \left[\frac{A^i - 1}{A - 1}\right] B$$

for $i = 0, 1, \ldots, N$.

Proof:

The proof is by induction and may be found in Henrici [2].

<u>Lemma 2.5</u>

For any real number δ, $1 + \delta \leq e^{\delta}$

Proof:

Since e^{δ} is a convex function, $1 + \delta \leq e^{\delta}$.

<u>Lemma 2.6</u>

If $\delta > 0$, B is a non-negative constant and $\{\xi_i\}$ is a sequence of real numbers satisfying $|\xi_i| \leq (1 + \delta)\ |\ \xi_{i-1}\ | + B$, i=0, 1,...,N then

$$|\xi_i| \leq e^{i\delta}\ |\xi_o| + B\ \frac{(e^{i\delta} - 1)}{\delta}$$

Proof:

Lemmas 2.4 and 2.5 with $A = 1 + \delta$

Definition 2.6

A function $f:S \rightarrow R^2$ satisfies a Lipschitz condition or is Lip K in y on S if there is a constant $K > 0$ such that,

$$| f(t, y_1) - f(t, y_2) | \leq K | y_1 - y_2 |$$

for all points (t, y_1), (t, y_2) in S.

Ordinarily the convergence of the numerical solution generated by (2-6) to the exact solution is established by assuming that f is Lip K on S, e.g., Henrici [2]. From this assumption it can be shown that the U_i lie in a compact set and from this fact the convergence is readily established. In general, the parameter c is not allowed to assume values outside the interval $(0, 1]$.

The functions f, which we shall be interested in, often fail to satisfy a Lipschitz condition on S. Because of this we attack the problem of convergence from a different direction.

We shall start by assuming that the points (t_i, U_i) and (t_i, Y_i) lie in some convex compact subset of S for all i. The differentiability properties of f are then used to obtain a Lipschitz condition for f and G on the compact set. The ensuing convergence proof is nearly the standard one, except that we allow c to assume values outside $(0, 1]$. Later, we will show that, for functions satisfying Property A, the points (t_i, U_i) and (t_i, Y_i) necessarily lie in a compact subset of S if h is sufficiently small.

Lemma 2.7

Let f satisfy Property A, then for each convex compact subset S_1 of S there is a $K > 0$ such that f is Lip K in y on S_1.

Proof:

Let (t, y_1) and (t , y_2) be arbitrary points in the convex compact set $S_1 \subset S$. By the Mean Value Theorem,

$$f(t , y_1) - f (t , y_2) = f_y (t, y_1 + \theta y_2) (y_1 - y_2)$$

where $0 < \theta < 1$. Since S_1 is convex $(t, Y_1 + \theta y_2) \in S_1$. Since S_1 is compact the continuous function f_y is bounded on S_1. We can take $K > 0$ to be this bound so that,

$$| f(t , y_1) - f (t , y_2) | \leq K | y_1 - y_2 |$$

and f is Lip K in y on S_1.

Lemma 2.8

If f satisfies Property A, h is sufficiently small and $c \in [-\lambda , \lambda]$ then $G_y(t , y , c)$ is continuous on $S \times [-\lambda , \lambda]$.

Proof:

Take $h \leq h_\varepsilon$ so that $(t + ch , y)$ is in S for all y.

For $c \neq 0$ we use the chain rule along with Definition 2.4 to obtain

$$G_y (t , y , c) = \frac{1}{c} \Big[f_y \big(t + ch, y + ch f(t , y)\big) - f_y (t , y) \Big]$$
$$+ ch f_y (t , y) f_y \big(t + ch, y + ch f (t , y)\big) \qquad (2\text{-}7)$$

Since f satisfies Property A we apply Taylor's Theorem in R^2 to obtain,

$$\begin{aligned} G_y(t, y, c) &= f_{ty}\big(t + \theta ch, y + \theta ch\, f(t, y)\big) + \\ &f_{yy}\big(t + \theta ch, y + \theta ch\, f(t, y)\big)\, f_y(t, y) + \\ &ch\, f_y(t, y)\, f_y\big(t + ch, y + ch\, f(t, y)\big) \end{aligned} \tag{2-8}$$

which is continuous at (t, y, c) for $c \neq 0$.

To show that G_y is continuous at a point $(t^*, y^*, 0)$ we compute from (2-8)

$$\lim_{\substack{c \to 0 \\ t \to t^* \\ y \to y^*}} G_y(t, y, c) = f_{ty}(t^*, y^*) + f_{yy}(t^*, y^*)\, f_y(t^*, y^*) \tag{2-9}$$

where we have again used the continuity properties of f and its partial derivatives. By Definition 2.4 the limit (2-9) is $G_y(t^*, y^*, 0)$ and the result follows.

Lemma 2.9

Let f satisfy Property A, $c \in [-\lambda, \lambda]$ and $h \leq h_\varepsilon$. If S_1 is a convex compact subset of S there is a constant $K > 0$ such that f and G are both Lip K on S_1.

Proof:

The proof is nearly identical to that of Lemma 2.7 with an appeal to Lemma 2.8. We take K to be an upper bound on $|f_y|$ and $|G_y|$ which are continuous on S_1.

Definition 2.7

For (t , y) and $(\bar{t} , \bar{y})$ in S,

i. $D f(\bar{t} , \bar{y} ; t , y) = f_t (\bar{t} , \bar{y}) + f_y (\bar{t} , \bar{y}) f (t , y)$

ii. $D^2 f (\bar{t} , \bar{y} ; t , y) = f_{tt} (\bar{t} , \bar{y}) + 2 f_{ty} (\bar{t} , \bar{y}) f (t ,y)$
$+ f_{yy} (\bar{t} , \bar{y}) \left[f (t , y)\right]^2$

Lemma 2.10

If f satisfies Property A, $h \leq h_\varepsilon$ and $c [-\lambda , \lambda]$ then

$$G(t, y, c) = h f' (t , y) + \frac{ch^2}{2} D^2 f (t + \theta ch, y + \theta ch f (t, y); t,y)$$

where $0 < \theta < 1$.

Proof:

By Taylor's Theorem and Definition 2.1,

$$f \left(t + ch, y + ch f(t , y)\right) = f (t , y) + ch f' (t, y)$$
$$+ (ch)^2 \left[f_{tt} + 2 f_{ty} f (t , y) + f_{yy} \left[f (t , y)\right]^2\right]$$

where the partial derivatives are evaluated at $\left(t + \theta ch, y + \theta ch f(t, y)\right)$ and $0 < \theta < 1$. For $c \neq 0$ we can divide by c and invoke the definitions of G and D^2 to attain the desired result. Taking the limit as $c \to 0$ and using the continuity of G (Lemma 2.2) the result holds for $c = 0$.

II.4 Convergence of the Method

In this section we will be dealing with two types of error; local (truncation) error and global (total) error. It is presumed throughout that all arithmetic is exact (no round-off error).

Definition 2.8

The truncation (or discretization) error induced by using (2-6) is defined to be the amount that Y (t) restricted to I_h fails to satisfy the difference equation (2-6). That is,

$$Y_{i+1} = Y_i + h\, f\, (t_i\, , Y_i) + \frac{h}{2}\, G\, (Y_i) + \tau_{i+1} \qquad (2\text{-}10)$$

Where τ_{i+1} is the truncation error. $i = 0, i, \ldots, N-1$

Definition 2.9

The total (or global) error at a point t_i of I_h is defined to be,

$$E_i = U_i - Y_i$$

The total error at $t_N = 1$ is then

$$E_N = U_N - Y_N = U_N - Y\ (1)$$

In addition to propagated truncation error the total error could include an error in the initial condition (i.e., $E_o = U_o - Y_o$).

Theorem 2.1

If f satisfies Property A, $h \leq h_\varepsilon$ and $c \in [-\lambda , \lambda]$ the truncation error (2-10) can be written in the form,

$$\tau_{i+1} = h^3 \left[\frac{Y'''(\xi_{i+1})}{6} - \frac{c}{4} D^2 f(t_i + \theta_i c h, Y_i + \theta_i c h f(t_i, Y_i) ; t_i , Y_i) \right] \qquad (2\text{-}11)$$

where $0 < \theta_i < 1$, $t_i < \xi_{i+1} < t_{i+1}$ and $i = 0, 1, 2, \ldots, N-1$

Proof:

By Taylor's Theorem,

$$Y_{i+1} = Y_i + h Y_i' + \frac{h^2}{2} Y_i'' + \frac{h^3}{6} Y''' (\xi_{i+1})$$

where $t_i < \xi_{i+1} < t_{i+1}$

The definition of truncation error coupled with Lemma 2.10 yields,

$$Y_{i+1} = Y_i + h f (t_i , Y_i) + \frac{h^2}{2} f' (t_i , Y_i) +$$

$$\frac{ch^3}{4} D^2 f \left(t_i + \theta_i ch, Y_i + \theta_i ch f(t_i , Y_i); t_i, Y_i\right) + \tau_{i+1}$$

Subtracting the two relations for Y_{i+1} and using the fact that $Y' = f$ and $Y'' = f'$ yields the desired result.

Lemma 2.11

Let f satisfy Property A and ξ be a constant such that $0 < \xi \leq 1$. If there is a Y^* and $h^* \leq h_\varepsilon$ such that for all $h \leq h^*$,

$$| Y_{i-1} + ch\ f(t_{i-1}, Y_{i-1}) - Y_o | \leq Y^*$$

for all $t_i \leq \xi$ and $c \in [-\lambda, \lambda]$ then there is a constant τ depending only on ξ and λ such that,

$$| \tau_i | \leq h^3 \tau$$

for all i such that $t_i \leq \xi$ and $h \leq h^*$

Proof:

Let R_ξ be the compact rectangle $[-\varepsilon, \xi] \times [Y_o - Y^*, Y_o + Y^*]$. From Definition 2.7,

$$D^2 f(\bar{t}, \bar{y} ; t, y) = f_{tt} (\bar{t}, \bar{y}) + 2 f_{ty} (\bar{t}, \bar{y}) f (t, y)$$
$$+ f_{yy} (\bar{t}, \bar{y}) \left[f(t, y) \right]^2$$

Since f satisfies Property A each term is continuous on R_ξ. Therefore each term is bounded on R_ξ.

Consider now the expression (2-11) of Theorem 2.1. Since Y''' is continuous on [0 , 1] there is a bound for $Y''' (\xi_{i+1})$ for all $i \leq N - 1$. The point $\left(t_{i-1} + ch, Y_{i-1} + ch\ f (t_{i-1}, Y_{i-1}\right)$ is in R_ξ by hypothesis for all $t_i \leq \xi$. Since $c = 0$ is in $[- \lambda, \lambda]$ it follows that (t_{i-1}, Y_{i-1}) is in R_ξ. Since R_ξ is convex and $0 < \theta_{i-1} < 1$ the point $\left(t_{i-1} + \theta_{i-1} ch, Y_{i-1} + \theta_{i-1} ch\ f(t_{i-1}, Y_{i-1})\right)$ is in R_ξ. Noting that $|c| \leq \lambda$ we see that the expression in square brackets in (2-11) is bounded for all $t_i \leq \xi$ whenever $h \leq h^*$. It is sufficient to take the bound to be τ.

Theorem 2.2

Let f satisfy Property A, $U_o = Y_o$ and ξ be a constant such that $0 < \xi \leq 1$. Suppose that there are constants Y* and h* such that for all $h \leq h^*$,

$$|U_{i-1} - Y_o| \leq Y^*$$

$$|U_{i-1} + c\ h\ f(t_{i-1}, U_{i-1}) - Y_0| \leq Y^*$$

$$|Y_{i-1} + c\ h\ f(t_{i-1}, Y_{i-1}) - Y_0| \leq Y^*$$

for all $t_i \leq \xi$ and $c \in [-\lambda\ , \lambda]$. Then the U_i converge, uniformly in i, to Y_i for all $t_i \leq \xi$.

Proof:

The proof is similar to that found in standard works, e.g., Henrici [2].

$$\text{Let } R_\xi = [-\varepsilon, \xi] \times [Y_o - Y^*\ , Y_o + Y^*].$$

Under the hypothesis, f and G are Lip K on the convex compact set R_ξ by Lemma 2.9. From Lemma 2.11 there is a τ such that $|\tau_i| \leq h^3\ \tau$ on R_ξ. Of course, both K and τ depend on ξ. Subtracting (2-10) from (2-6), using the definition of τ and Definition 2.9 yields,

$$|E_i| \leq |E_{i-1}| + h\ |\ f(t_{i-1}, U_{i-1}) - f(t_{i-1}, Y_{i-1})\ |$$

$$+ \frac{h}{2}\ |\ G\ (t_{i-1}, U_{i-1}) - G\ (t_{i-1}, Y_{i-1})\ | + |\ \tau_i\ |$$

which hold for $t_i \leq \xi$. Applying the Lipschitz conditions yields,

$$|E_i| \leq |1 + \frac{3}{2} h K| \quad |E_{i-1}| + h^3 \tau$$

Employing Lemma 2.6 with $\delta = \frac{3}{2} h K$ and $B = h^3 \tau$ we obtain,

$$|E_i| \leq \frac{2}{3} \frac{(e^{\frac{3}{2} i h K} -1)}{h K} h^3 \tau$$

for $t_i \leq \xi$. But $i h = t_i \leq \xi$ so that,

$$|E_i| \leq \frac{2}{3} \frac{(e^{\frac{3}{2} h \xi} -1)}{K} h^2 \tau$$

which holds for $t_i \leq \xi$ and $h \leq h_\varepsilon$. Thus, as $h \to 0$, $|E_i| \to 0$ for $t_i \leq \xi$ and the convergence is uniform in t_i for $t_i \leq \xi$.

Theorem 2.3

Let f satisfy Property A, $c \in [-\lambda , \lambda]$ and $U_o = Y_o$. Then the U_i converge, uniformly in i, to Y_i.

Proof:

We start by selecting $h \leq h_\varepsilon$. Let $\delta > 1$ be a fixed positive constant. Since $Y (t)$ is continuous on $[0 , 1]$ there is a constant $A > 0$ such that $|Y (t) - Y_o| \leq A$. Let $Y^* = A + 2\delta$ and form the rectangle $R^* = [-\varepsilon, 1 + \varepsilon] X [Y_o - Y^*, Y_o + Y^*]$. Since f and G are continuous there is an $M > 0$ such that,

$$|f(t,y) + \frac{1}{2} G(t, y, c) | \leq M \tag{2-12}$$

and

$$|ch\ f(t\ ,\ y)| \leq M \qquad (2\text{-}13)$$

on the compact set $R^* \times [-\lambda\ ,\ \lambda]$.

We form a sequence of positive numbers $\xi_i = \text{Min}\ (1,\ (y^* - \delta + i\ \delta)/M)$. The points of the sequence represent the abscissas of the intersection of the lines $y = Y_o \pm Y^*$ with a set of lines having slope M (see Figure 2-1). The sequence has the property that for some n_1, $\xi_n = 1$ for all $n \geq n_1$.

We assert that for each n there exists an h(n) such that for all $h \leq h(n)$,

$$|U_{i-1} - Y_o| \leq Y^* \qquad (2\text{-}14)$$

and

$$|U_{i-1} + ch\ f(t_{i-1}\ ,\ U_{i-1}) - Y_o| \leq Y^* \qquad (2\text{-}15)$$

and

$$|Y_{i-1} + ch\ f\ (t_{i-1}\ ,\ Y_{i-1}) - Y_o| \leq Y^* \qquad (2\text{-}16)$$

for all $t_i \leq \xi_n$. The proof will be by induction on n.

For n = 1, select $h \leq h(1) \leq \text{Min}\ \{h_\varepsilon\ ,\ \xi_1\}$. We shall prove, by induction on i, that (2-14), (2-15) and (2-16) hold for all $t_i \leq \xi_1$.

When i = 1, (2-14) follows immediately since $U_o = Y_o$. Since $(t_o\ ,\ U_o) \in R^*$ and $U_o = Y_o$ we apply (2-13) to obtain,

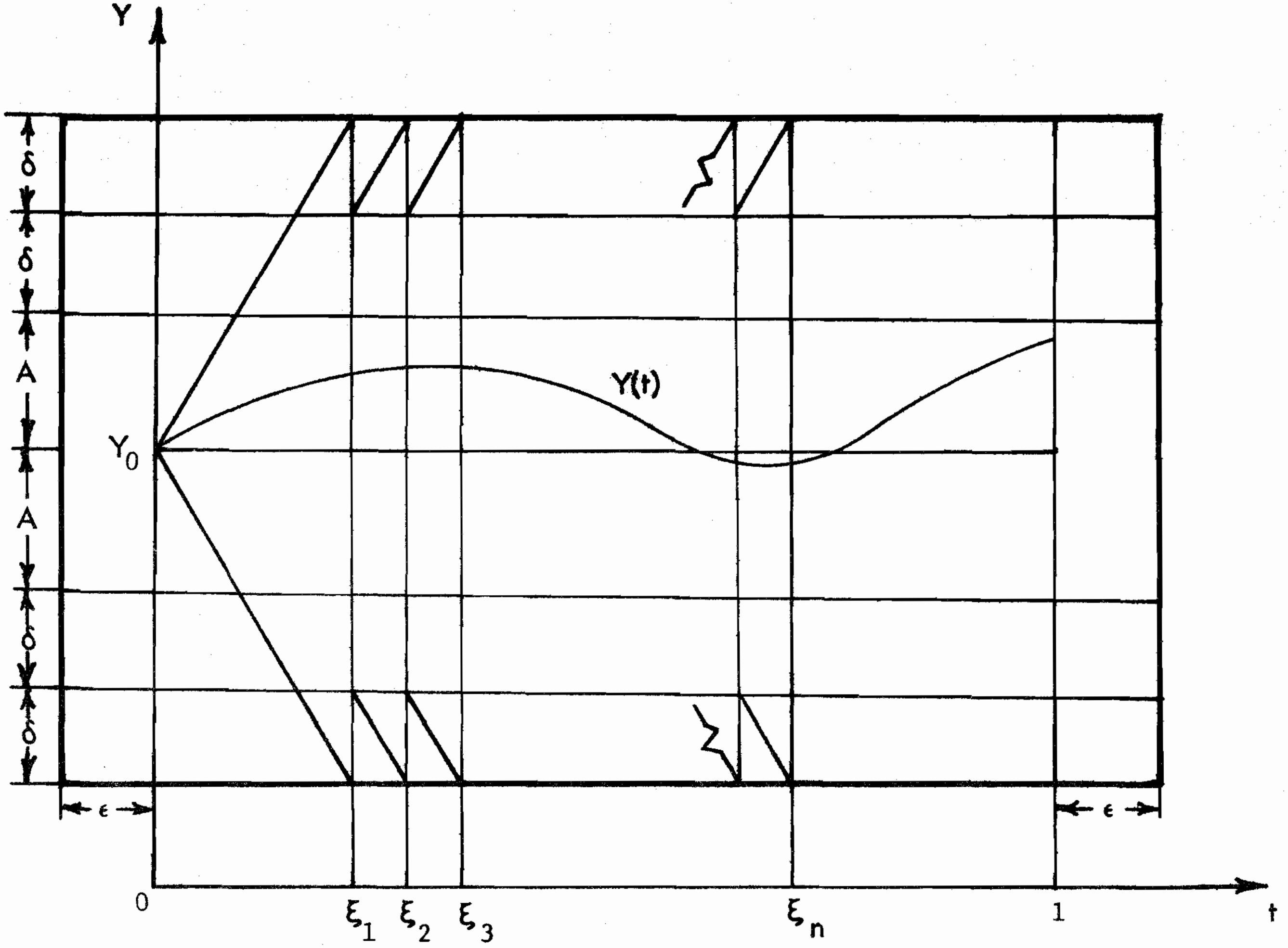


Diagram for Theorem 2-3
Figure 2-1

$$|ch\ f(t_o, Y_o)| \leq M\ h \leq M\ h(1) \leq Y^*$$

and this establishes both (2-15) and (2-16) for i = 1.

Our induction hypothesis on i is that the three inequalities hold for i = k. By (2-6),

$$U_k = Y_o + h\sum_{m=0}^{k-1} \left(f(t_m, U_m) + \frac{1}{2} G\ (t_m, U_m)\right) \qquad (2\text{-}17)$$

The induction hypothesis implies that, $(t_{i-1}, U_{i-1}) \in R^*$ for $t_i \leq t_k \leq \xi_1$. Therefore, by (2-12),

$$|U_k - Y_o| \leq k\ h\ M \leq k\ h(1)\ M < t_{k+1}\ M \leq \xi_1\ M \leq Y^*$$

which holds for $t_{k+1} \leq \xi_1$. This establishes (2-14). Similarly, from (2-12) and (2-17)

$$|U_k + ch\ f(t_k\ , U_k) - Y_o\ | \leq k\ h\ M + h\ M < t_{k+1}\ M \leq \xi_1\ M \leq Y^*$$

which holds for $t_{k+1} \leq \xi_1$. This establishes (2-15) for i=k + 1. The inequality (2-16) is treated in the same fashion. By induction the assertion is proved for all i for which $t_i \leq \xi_1$.

The induction hypothesis on n is, that there is an h (k) such that, for all h $\leq$ h(k), (2-14), (2-15) and (2-16) hold for all $t_i \leq \xi_k$.

The induction hypothesis is sufficient for Theorem 2.2 to be applied with h* = h(k). Since the convergence is uniform in i for $t_i \leq \xi_k$, there is an h(k+1) $\leq$ h (k) such that,

$$|U_i - Y_o| \leq Y^* - \delta \qquad (2\text{-}18)$$

for all $t_i \leq \xi_k$

Let $n = k + 1$. We shall show, by induction on i, that for all $h \leq h\ (k + 1)$, (2-14), (2-15) and (2-16) are satisfied when $t_i \leq \xi_{k+1}$. The case where $i = 1$ is the same as above when n was equal to 1.

The induction hypothesis on i is that (2-14), (2-15) and (2-16) hold for $i \leq p$ whenever $t_p \leq \xi_{k+1}$. Let i^* be the largest value of i for which $t_i \leq \xi_{k+1}$. Consider the case where $i = p + 1$. By (2-15),

$$U_{p+1} = U_{i^*} - h \sum_{m=i^*+1}^{p} \left| f(t_m, U_m) + \frac{1}{2}\ G(t_m, U_m) \right|$$

Subtracting Y_o from both sides and applying (2-18) and (2-12) yields,

$$|U_{p+1} - Y_o| \leq Y^* - \delta + (p - i^* - 1)\ h\ M$$

$$\leq Y^* - \delta + (p - i^* - 1)\ h\ (k + 1) \leq Y^* - \delta + (t_p - t_{i^*+1})\ M$$

$$\leq Y^* - \delta + (\xi_{k+1} - \xi_k)\ M \leq Y^* - \delta + \delta = Y^*$$

Thus, by induction, we have (2-14) satisfied for all i such that $t_i \leq \xi_{k+1}$. But this means that $h\ (k+1)$ satisfies our assertion on n for $n = k + 1$. Thus, by induction, (2-14) holds for all n. The inequalities (2-15) and (2-16) are established in a similar fashion.

We have proved that there is an $h(n)$ such that (2-14), (2-15) and (2-16) are satisfied for all $h \leq h\ (n)$ whenever $t_i \leq \xi_n$. Taking $h^* = h(n)$, we apply Theorem 2.2 to prove that the U_i converge uniformly in i for $t_i \leq \xi_n$ to Y_i. Since the ξ_n are all one for large enough n we have convergence on $[0\ ,\ 1]$.

Theorem 2.3 shows that the points $(t_i\ ,\ U_i)$, $\left(t_i + c\ h,\ U_i + ch\ f(t_i\ ,\ U_i)\right)$, $(t_i\ ,\ Y_i)$ and $\left(t_i + ch,\ Y_i + ch\ f(t_i\ ,\ Y_i)\right)$ all lie in a compact subset of S for sufficiently small h. Furthermore, since $Y\ (t,\ Y_o)$ is continuous there is a compact subset of S containing these points for all Y_o in $[a\ ,\ b]$. It is convenient to work with the compact rectangle defined below.

Definition 2.10

Let $R_{\overline{Y}} = [-\varepsilon,\ 1 + \varepsilon] \times [-\overline{Y}\ ,\ \overline{Y}]$ where $\overline{Y}$ is selected large enough so that,

$$| U_i - Y_o| \leq \overline{Y}$$

$$|U_i + ch\ f(t_i\ ,\ U_i) - Y_o| \leq \overline{Y}$$

$$|Y_i + ch\ f(t_i\ ,\ Y_i) - Y_o| \leq \overline{Y}$$

for all sufficiently small h, $c \in [-\lambda\ ,\lambda\ ]$ and $Y_o\ \varepsilon [a\ ,\ b]$.

III GLOBAL OPTIMIZATION

III.1 Introductory Statement

In this chapter we investigate the conditions under which;

A. there is a c for which $|E_N(c)|$ is minimized,

B. there is a unique c for which $|E_N(c)|$ is minimized,

C. there is a unique c for which $E_N(c) = 0$

From the preceding chapter we know that if f is sufficiently differentiable and h is small enough then $E_N(c)$ is a continuous function of c on $[-\lambda , \lambda]$ for fixed N. Since $[-\lambda , \lambda]$ is compact, $|E_N(c)|$ achieves a minimum for some c and the conditions for A are determined.

The conditions for B are not so easy. The example discussed in Chapter I shows that in some cases c may not be unique. From Theorem 2.1 it can be seen that if D^2f vanishes identically the local error does not depend on c. In this chapter we determine conditions on f under which both B and C are achieved. Furthermore, a simple transformation of variables will be proposed by which c can be achieved for a transformed form of the basic problem (2-1).

III.2 The Vanishing of D^2f

We are interested in classifying the types of functions f for which D^2f does not vanish at any points of I_h. From Definition 2.7,

$$D^2 f\ (t\ ,\ y\ ;\ t_i\ ,\ y_i) = f_{tt}\ (t\ ,\ y) + 2\ f_{ty}\ (t\ ,\ y)\ f(t_i\ ,\ y_i)$$

$$+\ f_{yy}\ (t\ ,\ y) \left[f(t_i\ ,\ y_i) \right]^2 \tag{3-1}$$

This expression can be developed from another point of view which will better illuminate the geometric properties of f.

<u>Definition 3.1</u>

For f satisfying Property A the <u>Hessian</u> of f is the matrix of partial derivatives,

$$H(t\ ,\ y) = \begin{pmatrix} f_{tt} & f_{ty} \\ f_{yt} & t_{yy} \end{pmatrix}$$

where it is understood that the derivatives are evaluated at (t , y). Since f is twice continuously differentiable H is symmetric.

With this definition (3-1) can be written as the real quadratic form,

$$z^T\ H\ z \tag{3-2}$$

where z is the column vector representation of the vector whose first component is 1 and whose second component is $f(t_i\ ,\ y_i)$.

Definition 3.2

A real quadratic form with symmetric matrix A of dimension n is positive definite if

$$x^T A x > 0$$

for all vectors $x \neq 0$ in R^n.

Definition 3.3

A real symmetric matrix is positive definite if its associated quadratic form is positive definite.

It is known that if H is positive definite at all points (t , y) in the domain of f then f is a strictly convex function of t and y (Rockafeller [10]). If f is strictly convex then (3-1) and (3-2) are positive on the strip S. In this case, we see by Theorem 2.1 that the global error depends on c.

Since the vector z of (3-2) has a first component which is never zero the class of functions for which (3-1) is positive is broader than the set of strictly convex functions. This is illustrated via the following definition.

Definition 3.4

Let $S_1 = \{(1 , x) \mid x \text{ real}\}$. Let f satisfy Property A and let the Hessian of f be positive definite on S_1. That is,

$$(1 \; x) \; H \; (t , y) \begin{pmatrix} 1 \\ x \end{pmatrix} > 0$$

for all $(t , y) \in S$ and for all $(1 , x) \in S_1$. We will denote the class of all such functions by $\mathcal{G}$.

We note that $\mathcal{G}$ contains the class of functions satisfying Property A which are strictly convex on S. In addition, $\mathcal{G}$ contains functions which are not strictly convex on S. For example, $f(t, y) = t^2 + y$ is not strictly convex but,

$$(1\ x)\ H \begin{pmatrix} 1 \\ x \end{pmatrix} = (1\ x) \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} \equiv 2 > 0$$

We will show that if $f \varepsilon \mathcal{G}$ then $E_N(c)$ is a strictly increasing function of c.

<u>Lemma 3.1</u>

For f satisfying Property A, $c \in [-\lambda, \lambda]$ and h sufficiently small,

$$\frac{d\, f\left(t_i\, , U_i\, (c)\right)}{d\, c} = f_y\left(t_i\, , U_i\, (c)\right) \frac{d\, U_i\, (c)}{d\, c}$$

Proof:

By the chain rule.

<u>Lemma 3.2</u>

For f satisfying Property A, $c \in [-\lambda\, , \lambda]$ and h sufficiently small,

$$\frac{d\ f\ (\bar{t}\ ,\ \bar{y})}{d\ c} = h\ Df\ (\bar{t}\ ,\ \bar{y}\ ;\ t_i\ ,\ U_i) +$$

$$f_y(\bar{t}\ ,\ \bar{y}) \left[1 + ch\ f_y(t_i\ ,\ U_i)\right] \frac{dU_i}{dc}$$

where $\bar{t} = t_i + c\ h$ and $\bar{y} = U_i + c\ h\ f\ (t_i\ ,\ U_i)$ $i = 0, 1, \ldots N\text{-}1$

Proof:

Taking the derivative yields,

$$\frac{d\ f\ (\bar{t}\ ,\ \bar{y})}{d\ c} = h\ f_t\ (\bar{t}\ ,\ \bar{y}) + f_y\ (\bar{t}\ ,\ \bar{y}) \left[\frac{dU_i}{dc} + h\ f\ (t_i\ ,\ U_i) + c\ h\ \frac{d\ f\ (t_i\ ,\ U_i)}{d\ c}\right]$$

Collecting terms, applying Definition 2.7 and Lemma 3.1 yields the desired result.

Lemma 3.3

For f satisfying Property A, $c \in [-\lambda\ ,\ \lambda]$, $c \neq 0$ and h sufficiently small,

$$\frac{d\ G\ (t_i\ ,\ U_i)}{d\ c} = -\ \frac{G\ (t_i\ ,\ U_i)}{c} + \frac{1}{c}\ Df(\bar{t}\ ,\ \bar{y}\ ;\ t_i\ ,\ U_i)$$

$$+ \left[hf_y\ (\bar{t}\ ,\ \bar{y})\ f_y(t_i\ ,\ U_i) + \frac{1}{c}\left(f_y\ (\bar{t}\ ,\ \bar{y}) - f_y\ (t_i\ ,\ U_i)\right)\right] \frac{d\ U_i}{d\ c}$$

where $\bar{t} = t_i + c\ h$, $\bar{y} = U_i + c\ h\ f\ (t_i\ ,\ U_i)$

Proof:

From the definition of G, (Definition 2.4) and Lemmas 3.1 , 3.2,

$$\frac{d\,G}{d\,c}(t_i, U_i) = -\frac{G(t_i , U_i)}{c} + \frac{1}{c}\left[\frac{d\,f\,(\bar{t} , \bar{y})}{d\,c} - \frac{d\,f\,(t_i , U_i)}{d\,c}\right]$$

$$= -\frac{G(t_i , U_i)}{c} + \frac{h}{c}\,Df\,(\bar{t} , \bar{y} ; t_i , U_i) + \left[h\,f_y(\bar{t} , \bar{y})\,f_y(t_i , U_i) + \frac{1}{c}\left(f_y(\bar{t} , \bar{y}) - f_y\,(t_i , U_i)\right)\right]\frac{d\,U_i}{d\,c}$$

with $\bar{t} = t_i + c\,h$, $\bar{y} = U_i + c\,h\,f\,(t_i , U_i)$.

Lemma 3.4

For f satisfying Property A, $c \in [-\lambda , \lambda]$, $C \neq 0$ and h sufficiently small

$$\frac{d\,U_{i+1}}{d\,c} = \left[1 + h\,f_y\,(t_i , U_i) + \frac{h^2}{2}\,f_y\,(\bar{t} , \bar{y})\,f_y(t_i, U_i) + \frac{h}{2c}\left(f_y\,(\bar{t} , \bar{y}) - f_y\,(t_i , U_i)\right)\right]\frac{d\,U_i}{d\,c} + \frac{h}{2}\left[\frac{Df\,(\bar{t} , \bar{y} ; t_i , U_i) - G\,(t_i , U_i)}{c}\right]$$

where $\bar{t} = t_i + c\,h$, $\bar{y} = U_i + c\,h\,f\,(t_i , U_i)$

Proof:

Differentiating (2-6), yields,

$$\frac{d\,U_{i+1}}{d\,c} = \frac{d\,U_i}{d\,c} + h\,\frac{d\,f\,(t_i\,,\,U_i)}{d\,c} + \frac{h}{2}\,\frac{d\,G(t_i,\,U_i)}{d\,c}$$

using Lemmas 3.1, 3.3 and collecting terms yields the result.

Lemma 3.5

For f satisfying Property A, $c \in [-\lambda\,,\,\lambda]$, $c \neq 0$ and h sufficiently small,

$$\frac{1}{c}\left[f_y\left(t_i + c\,h,\, U_i + c\,h\,f\,(t_i\,,\,U_i)\right) - f_y\,(t_i\,,\,U_i)\right]$$

$$= h\left[f_{ty} + f(t_i\,,\,U_i)f_{yy}\right]$$

where the partial derivatives are evaluated at

$$\left(t_i + \theta_i\,c\,h,\, U_i + \theta_i\,c\,h\,f\,(t_i\,,\,U_i)\right) \text{ and } 0 < \theta_i < 1$$

Proof:

By the Mean Value Theorem and the continuity of the partial derivatives.

Lemma 3.6

Let f satisfy Property A, $c \in [-\lambda\,,\,\lambda]$ and $c \neq 0$ then there is an h such that,

$$P_i = 1 + hf_y\,(t_i\,,\,U_i) + \frac{h^2}{2}\,f_y\left(t_i + c\,h,\, U_i + c\,h\,f(t_i\,,\,U_i)\right)f_y(t_i,U_i)$$

$$+ \frac{h}{2c}\left[f_y\left(t_i + c\,h,\, U_i + c\,h\,f\,(t_i\,,\,U_i)\right) - \right.$$

$$\left. f_y\,(t_i\,,\,U_i)\right] > 0$$

on the rectangle $R_{\overline{Y}}$ of Definition 2.10.

Proof:

We first take h small enough so that all the involved points are in S. Lemma 3.5 and the fact that all partial derivatives are bounded below on the compact set $R_{\overline{Y}}$ yield the result; that is, $P_i \approx 1$ for small enough h.

Lemma 3.7

If $f \in \mathcal{G}$ and h is sufficiently small then,

$$\frac{h}{c}\left[Df\ (t_i + c\,h,\ U_i + c\,h\,f\ (t_i\ ,\ U_i)\ ;\ t_i\ ,\ U_i) - G\ (t_i\ ,\ U_i)\right]$$

is positive for all non-zero $c \in [-\lambda\ ,\ \lambda]$

Proof:

By Taylor's Theorem and the definition of Df,

$$f\ (t_i\ ,\ U_i) = f\ \left(t_i + c\,h,\ U_i + c\,h\,f\ (t_i\ ,\ U_i)\right)$$
$$- c\,h\,Df\ \left(t_i + c\,h,\ U_i + c\,h\,f\ (t_i\ ,\ U_i)\ ;\ t_i\ ,\ U_i\right)$$
$$+ \frac{c^2\,h^2}{2}\ (z^T\,H\,z)$$

where $z^T = \left(1,\ f(t_i\ ,\ U_i)\right)$ and the partial derivatives in H are evaluated at $\left(t_i + \theta_i\,c\,h,\ U_i + \theta_i\,c\,h\,f\ (t_i\ ,\ U_i)\right)$. The latter point is in $R_{\overline{Y}}$ since $(t_i\ ,\ U_i)$ and $\left(t_i{+}c\,h,\ U_i + c\,h\,f\ (t_i\ ,\ U_i)\right)$ are in $R_{\overline{Y}}$ and it is a convex set. Since $f \in \mathcal{G}$ implies $z^T\,H\,z > 0$ it follows that,

$$-f\bigl(t_i + c\,h,\ U_i + c\,h\,f\,(t_i\ ,\ U_i)\bigr) + f(t_i\ ,\ U_i)$$

$$+\ c\,h\,Df\,(t_i + c\,h,\ U_i + c\,h\,f(t_i\ ,\ U_i)\ ;\ t_i\ ,\ U_i) > 0$$

Since $c \neq 0$ the desired result is obtained by dividing twice by c.

<u>Lemma 3.8</u>

If $f \in \mathcal{G}$ and h is sufficiently small then,

$$\frac{d\,U_1(c)}{d\,c} > 0$$

for all $c \neq 0$ in $[-\lambda\ ,\ \lambda]$.

Proof:

Since $\frac{d\,U_o}{d\,c} = 0$ and $\frac{d\,f\,(t_o\ ,\ U_o)}{d\,c} = 0$ by (2-6) and Lemma 3.3 we have,

$$\frac{d\,U_1}{d\,c} = \frac{h}{2}\ \frac{d\,G(t_o\ ,\ U_o)}{d\,c} = -\frac{h}{2c}\ G(t_o,\ U_o)$$

$$+\ \frac{h}{2c}\ Df\bigl(t_o + c\,h,\ U_o + c\,h\,f\,(t_o\ ,\ U_o)\ ;\ t_o\ ,\ U_o\bigr)$$

which is positive by Lemma 3.7.

<u>Theorem 3.1</u>

If $f \in \mathcal{G}$ and h is sufficiently small and fixed then $E_N(c)$ is a strictly increasing function of c on $[-\lambda\ ,\ \lambda]$.

Proof:

We shall show that $\frac{d}{d\,c}\,E_N\,(c) > 0$ for $c \neq 0$.

We take h small enough so that all points involved lie in $R_{\overline{Y}}$ and the expression of Lemma 3.6 is positive. Then fix h (or N). We have,

$$\frac{d\,E_N\,(c)}{d\,c} = \frac{d}{d\,c}\left[U_N\,(c) - Y_N\right] = \frac{d\,U_N\,(c)}{d\,c}$$

By Lemmas 3.4 and 3.6

$$\frac{d\,U_N\,(c)}{d\,c} = P_N\,\frac{d\,U_{N-1}(c)}{d\,c} +$$

$$\frac{h}{2\,c}\Big[Df(t_{N-1} + c\,h,\ U_{N-1} + c\,h\,f(t_{N-1}\ ,\ U_{N-1})\ ;\ t_{N-1},\ U_{N-1}) - G(t_{N-1},\ U_{N-1})\Big] \tag{3-5}$$

where P_N is the positive quantity (3-4) of Lemma 3.6. The last term in (3-5) is strictly positive by Lemma 3.7. Thus $\frac{d\,U_N}{d\,c}$ is positive whenever $\frac{d\,U_{N-1}}{d\,c}$ is positive. Since N is fixed it is sufficient to show that $\frac{d\,U_1(c)}{d\,c}$ is positive and the result follows by induction. But $\frac{dU_1(c)}{d\,c} > 0$ by Lemma 3.8.

Therefore, $\frac{d\,E_N(c)}{d\,c} > 0$ for all $c \in [-\lambda\ ,\ \lambda]$ except possibly $c = 0$. Since E_N is continuous and has a positive derivative on all but a finite point set it is strictly increasing on $[-\lambda\ ,\ \lambda]$.

Theorem 3.2

If $f \in \mathscr{G}$ and h is sufficiently small then there is a unique c $[-\lambda , \lambda]$ which minimizes $|E_N (c)|$.

Proof:

By Theorem 3.1, E_N is a strictly increasing function of c on $[-\lambda , \lambda]$. If $E_N > 0$ for all c $[-\lambda , \lambda]$ then $|E_N|$ has its minimum at $-\lambda$. If $E_N < 0$ then $|E_N|$ has its minimum at λ. If E_N has both positive and negative values there is a c^* such that $E_N(c^*) = 0$ by the Intermediate Value Theorem. In this case $|E_N(c^*)| = 0$.

This theorem establishes the conditions for B in the introduction of this chapter.

Theorem 3.3

For $f \in \mathscr{G}$, h small enough and λ large enough there is a μ such that $\tau_{i+1} (\mu) < 0$ and $\tau_{i+1} (-\mu) > 0$ for $i = 0, 1,\ldots,N-1$.

Proof:

Let $(\bar{t} , \bar{y})$ and (t , y) be any two points in $R_{\bar{Y}}$. For $z = \big(1, f\,(t , y)\big)$, $D^2 f\,(\bar{t} , \bar{y} ; t , y) \equiv z H z^T$ where the partial derivatives in H are evaluated at $(\bar{t} , \bar{y})$. Since $f \in \mathscr{G}$, $z H z^T > 0$ for sufficiently small h. In addition, f and its partial derivatives are continuous on $R_{\bar{Y}}$ so that there is a number $L > 0$ such that $z H z^T \geq L > 0$ for all pairs of points in $R_{\bar{Y}}$.

From Theorem 2.1,

$$\tau_{i+1}\ (c) = \frac{h^3}{4}\left[\frac{2}{3}\ Y'''\ (\xi_{i+1}) - c\ z_i\ H\ z_i^T\right]$$

where $t_i < \xi_{i+1} < t_{i+1}$, $z_i = \left(1,\ f\ (t_i\ ,\ Y_i)\right)$, H is evaluated at $\left(t_i + \theta_i\ c\ h,\ Y_i + \theta_i\ c\ h\ f\ (t_i\ ,\ Y_i)\right)$ and $0 < \theta_i < 1$.

Now, Y''' is continuous so that it achieves a maximum and minimum on [0 , 1]. Let Y_m''' denote the minimum and Y_M''' denote the maximum.

We select $\mu_1 > \frac{2}{3}\ \frac{Y_M'''}{L}$ and λ large enough so that $|\mu_1| \leq \lambda$. Since,

$$\mu_1\ (z_i\ H\ z_i^T) \geq \mu_1\ L > \frac{2}{3}\ Y_M''' \geq \frac{2}{3}\ Y'''\ (\xi_{i+1})$$

for $c = \mu_1$ we have $\tau_{i+1}\ (\mu_1) < 0$. In addition, if $c < \mu$, then $\tau_{i+1}\ (c) < \tau_{i+1} < (\mu)$.

Similarly, there is a $U > 0$ such that $z_i\ H\ z_i^T < U$ on $R_{\overline{Y}}$. For,

$$\mu_2 < \frac{2}{3}\ \frac{Y_m'''}{U} \text{ and } -\lambda \leq \mu_2 \leq \lambda$$

$$\mu_2\ (z_i\ H\ z_i^T) \leq \mu_2\ U < \frac{2}{3}\ Y_m''' < \frac{2}{3}\ Y'''\ (\xi_{i+1})$$

so that

$$\tau_{i+1}\ (\mu_2) > 0$$

Now let $\mu = \max\ (|\mu_1|\ ,\ |\mu_2|)$ then $\mu_1 < \mu$, and $-\mu < \mu_2$ so that,

$$\tau_{i+1}\ (\mu) < 0 \text{ and } \tau_{i+1}\ (-\mu) > 0$$

Lemma 3.9

Let $f \in \mathcal{F}$, h be sufficiently small and $|\mu| < \lambda$. Then $E_i(\mu) > 0$ and $E_1(-\mu) < 0$.

Proof:

By definition, $E_1 = U_1 - Y_1$. Applying the formula (2-6) along with Lemma 2.10 and subtracting the Taylor series expansion for Y_1 yields,

$$E_1(\mu) = U_o - Y_o + h\left[f(t_o, U_o) - f(t_o, Y_o)\right]$$

$$+ \frac{h^2}{2}\left[f'(t_o, U_o) - f'(t_o, Y_o)\right] +$$

$$\frac{h^3}{4}\left[\mu\, z_o H z_o^T - \frac{2}{3} Y'''(t_o + \xi_i t_i)\right]$$

where $0 < \xi_i < 1$, $z_o = \big(1, f(t_o, U_o)\big)$ and the partial derivatives in H are evaluated at the usual point. Since $U_o = Y_o$ and the fact that the last term is $-\tau_1(\mu)$ by Theorem 2.1

$$E_1(\mu) = -\tau_1(\mu)$$

By Theorem 3.3 $E_1(\mu) > 0$.

A similar argument establishes that $E_1(-\mu) < 0$.

Theorem 3.4

If $f \in \mathcal{F}$, h is sufficiently small and $|\mu| < \lambda$ there is a unique c such that $E_N(c) = 0$.

Proof:

Using (2-6), Lemma 2.10 and subtracting the power series expansion of Y_{i+1} yields

$$E_{i+1} = U_{i+1} - Y_{i+1} = U_i - Y_i + h \left[f(t_i\ , U_i) - f(t_i\ , Y_i)\right]$$

$$+ \frac{h^2}{2} \left[f'\ (t_i\ , U_i) - f'\ (t_i\ , Y_i)\right] - \tau_{i+1}(c)$$

Applying the Mean Value Theorem yields,

$$E_{i+1} = \left[1 + h\ f_y + \frac{h^2}{2}\ f'_y\right] E_i - \tau_{i+1}\ (c)$$

where f_y and f'_y are evaluated at points in $R_{\overline{Y}}$ and thus in view of their continuity properties are bounded on $R_{\overline{Y}}$. Therefore, there is an h for which

$$Q_i = \left[1 + h\ f_y + \frac{h^2}{2}\ f'_y\right] > 0 \text{ for all points in } R_{\overline{Y}}\ .$$

For h small enough and fixed consider

$$E_{i+1}\ (\mu) = Q_i\ E_i\ (\mu) - \tau_{i+1}\ (\mu)$$

where μ is defined in Theorem 3.3. Since $-\tau_{i+1}\ (\mu) > 0$ if $E_i\ (\mu)$ is > 0 then $E_{i+1}\ (\mu) > 0$.

From Lemma 3.9, $E_1\ (\mu) > 0$ so that $E_2\ (\lambda) > 0$. Proceeding recursively precisely N-1 times we conclude that

$$E_N\ (\mu) > 0$$

By a symmetrical argument $E_N\ (-\mu) < 0$. Since $E_N\ (c)$ is a continuous function of c on $[-\lambda ,\lambda]$ by the Intermediate Value Theorem there is a c for which $E_N(c) = 0$. Since E_N (c) is a strictly increasing function, the c so obtained, is unique.

At this point several observations are in order. If y''' is positive for all t then the μ_2 of Theorem 3.4 is positive and we can develop a theorem like Theorem 3.4 for $c \in [\mu_2 , \lambda]$. In this case we would not have to bother with the c = 0 case which has complicated the previous analysis.

Clearly we could have started with the hypothesis that H was negative definite in Definition 3.4. There would then be obvious duals to each of the theorems in this chapter. In particular, E_N would be strictly decreasing and again we would show the existence of a unique c for which $E_N(c) = 0$. At times it may be more convenient to use the dual theory.

III.3 Transformations

In many cases f will satisfy Property A but D^2f will not be one-signed on $R_{\overline{Y}}$. This problem can be circumvented by introducing a simple transformation of variables. We assume throughout that the derivatives involved are continuous. The arguments of functions will often be omitted unless they are particularly pertinent to the discussion. In view of the remarks following Theorem 2.3 it is sufficient to consider f defined on the compact rectangle $R_{\overline{Y}}$ of Definition 2.10.

Let

$$W(t) = Y(t) + g(t) \tag{3-6}$$

then

$$W' = Y' + g'$$

and from (2-1)

$$W' = f(t\ ,\ W - g) + g'$$

$$W(o) = Y(o) + g(o)$$

Letting $F(t\ ,\ W) = f(t\ ,\ W-g) + g'(t)$ the transformed problem becomes,

$$\begin{aligned} W' &= F\ (t\ ,\ W) \\ W_o &= Y_o + g_o \end{aligned} \qquad (3\text{-}7)$$

In this formulation F is defined on the compact rectangle $R_{Y*} = [-\varepsilon,\ 1 + \varepsilon]\ X\ [-Y*\ ,\ Y*]$ where $|Y*| = |\bar{Y}| + \underset{-\varepsilon \leq t \leq 1+\varepsilon}{\text{Max}} |g(t)|$. Thus, $F : R_{Y*} \to R$.

We shall begin by computing D^2F and from this determine the nature of $g(t)$ such that $D_F^2 > 0$ on R_{Y*}. Using the chain rule,

$$F_t = f_t - g'\ f_y + g''$$

$$F_{tt} = f_{tt} - g''\ f_y - 2\ g'\ f_{yt} + (g')^2\ f_{yy} + g''$$

$$F_W = f_y \qquad\qquad F_{WW} = f_{yy}$$

$$F_{tW} = f_{ty} - g'\ f_{yy}$$

The above relations, along with some fortuitous cancellations yield,

$$D^2F(\bar{t}, \bar{W}; t, W) = D^2f(\bar{t}, \bar{y}; t, y) + g'''(t) - g''(t)\, f_y(t, y) \tag{3-8}$$

We would like to choose g(t) so that D^2F is strictly positive on R_{Y*}. We assume that D^2f has both positive and negative values on $R_{\bar{Y}}$. Since D^2f is continuous on $R_{\bar{Y}}$ there is a constant $L > 0$ such that $D^2f > -L$ on $R_{\bar{Y}}$. We wish to choose g such that

$$g''' - g''\, f_y \geq L \tag{3-9}$$

on $R_{\bar{Y}}$. If this can be done, from (3-8)

$$D^2F = D^2f + g''' - g''\, f_y \geq D^2f + L > 0$$

We will arbitrarily choose $g''' > 0$ and $g'' \geq 0$ for $t \in [-\varepsilon, 1+\varepsilon]$. Since f_y is continuous there is an $M > 0$ such that $|f_y| \leq M$ on $R_{\bar{Y}}$. Under the assumptions on g the left side of (3-9) is the smallest when,

$$g''' - M\, g'' = L \tag{3-10}$$

This is a linear differential equation with the solution,

$$g'' = -L/M + Ae^{Mt} \tag{3-11}$$

where A is a constant of integration. We must select A so that $g''' > 0$ and $g'' \geq 0$. By selecting $A \geq \frac{L}{M}$ we have

$$g'' \geq \frac{L}{M}\left[-1 + e^{Mt}\right] \geq 0 \text{ as required} \tag{3-12}$$

Furthermore,

$$g''' = AMe^{Mt} = Le^{Mt} > 0$$

as required. By taking $A = L/M$ we determine that,

$$g(t) = \frac{-L}{2M} t^2 + \frac{L}{M^3} e^{Mt} \qquad (3\text{-}13)$$

is sufficient for (3-9).

We recall that (2-6) requires f' to be computed if c takes on the value zero. It would be desireable to avoid this computation. From the remarks following Theorem 3.4 we know that if $Y'''(t)$ is positive on $[0 , 1]$ then the optimal c is positive. A modification of (3-12) will ensure this.

Since $Y'''(t)$ is continuous on $[0 , 1]$ there is an $m > 0$ such that $-m < Y'''(t)$. In (3-11) choose $A \geq \text{Max} \{ \frac{m}{M} e , \frac{L}{M} \}$ then $g''' > 0$ and by (3-12) $g'' \geq 0$. In addition,

$$W''' = Y''' + g''' > - m + g'''$$

$$= - m + AMe^{Mt} \geq - m + me^{1+mt} > 0$$

as desired.

Thus the transformation,

$$g = - \frac{L}{2M} t^2 + \frac{A}{M^2} e^{Mt} \qquad (3\text{-}14)$$

with $A = \max \{ \frac{L}{M} , \frac{m}{M} e \}$ is sufficient to guarantee that $W''' > 0$ and $D^2F > 0$ on R_{Y*}.

For digital computation the exponential function is relatively expensive to compute. It would be desirable to have g be a simpler function, for instance, a low order polynomial in t. There are two special cases of (3-11) which are particularly useful. In both cases a preliminary transformation is used to make $W''' > 0$.

Using the lower bound $-m$ on Y''' let,

$$W = Y + \frac{mt^3}{6} \tag{3-15}$$

Then $W''' = Y''' + m > 0$ and the transformed problem (2-1) is,

$$W' = f(t\ ,\ W - \frac{mt^3}{6}) + \frac{m\ t^2}{2} \tag{3-16}$$

$$W_o = Y_o$$

In view of (3-16) we will assume below that $Y''' > 0$ for (2-1).

Case 1.

If there is a constant M_1 such that $0 < M_1 \leq f_y$ on R_Y then (3-9) is satisfied by,

$$g(t) = -\frac{L}{2M_1} t^2 \tag{3-17}$$

Since $Y''' > 0$ by assumption we have $W''' = Y''' > 0$ so that the optimal c is positive.

Case 2.

If there is a constant $M_2 > 0$ such that $f_y \leq -M_2 < 0$ then,

$$g = \frac{L}{2M_2}\ t^2 \tag{3-18}$$

satisfies (3-9) with $W''' > 0$.

While the above transformations provide insight into the type of transformation that is needed this is not the whole story. All of the suggested transformations are sufficient for our purpose and are by no means necessary conditions. For instance, for large enough n, $g(t) = (t + 1)^n$ will satisfy (3-9). We can also get another family of transformations by selecting g so that $g''' > 0$ while $g'' \leq 0$. Since the selection of an appropriate transformation is highly problem dependent we will not pursue the subject except to point out certain facets of the problem.

It should be noted that it may be more desirable to choose g so that $D_F^2 < 0$ and to employ the dual theory. The analysis is similar in this case.

In practice, we recommend trying the method directly without introducing the transformations. The conditions we have given for the existence of a unique optimal c are only sufficient conditions. Our experience with the method indicates that it applies to a much wider class of situations. If a transformation must be used the values of the precise bounds on f_y and Y''' may not be needed. In fact, we suggest estimating some values and trying the transformation. If optimal values of c are obtained which are sufficiently smooth functions of Y_o this will suffice.

It also should be noted that while a transformation might be theoretically usable it may be worthless in practice. In particular if M is very large in (3-13) both g and W will be large but opposite in sign. In order to recover Y we would have to add large numbers with opposite signs. The potential for losing significance is high in this case.

IV. APPLICATION TO SYSTEMS

The procedure developed in the previous chapters is also applicable to systems of differential equations. A theoretical development paralleling that of the preceding chapter is considerably more difficult in this case. However, some analogous results can be obtained. We shall indicate what is involved and leave a detailed analysis to later investigations.

Throughout this chapter, superscripts will generally represent indices used to denote vector components. Superscripts on scalars will represent powers. Subscripts will be used in the same manner as in previous chapters.

We shall denote a set of real valued functions, defined on the domain $[-\varepsilon, 1 + \varepsilon]$, by $Y^j(t)$. The index j will always range over $1, 2, \ldots, n$.

The argument t will frequently be omitted so that $y^j(t) \equiv y^j$. Similarly, let $f^j\ (t\ , y^1\ , y^2\ , \ldots, y^n)$ be a set of real valued functions defined for $t \in [-\varepsilon\ ,1 + \varepsilon]$, y^j real.

It will be convenient to adopt vector notation. Thus, $Y(t)$ is a vector with components $y^j(t)$ while the vector $f(t\ , y)$ has components $f^j\ (t\ , y)$. We may also write these, respectively, as $f(t\ , y^1\ , \ldots, y^n)$ and $f^j\ (t\ , y^1\ , \ldots, y^n)$.

Using vector initiation the problem we shall consider is,

$$Y'(t) = f\left(t , Y(t)\right)$$
$$Y(t_o) = Y_o \tag{4-1}$$

where $0 \leq t \leq 1$, $Y'(t)$ and $Y(o)$ are vectors with components $Y^{j'}(t)$ and $Y^j(o) = Y_o^j$ respectively.

It is assumed that a solution to (4-1) exists, is unique and continuous for $t \in [0 , 1]$. The grid, I_h , on $[0 , 1]$ is the same as before. The exact solution of (4-1) will be denoted by the vector Y while the vector U_i will represent the numerical solution at t_i. The subscript i, will range over 0, 1, ..., N. The components of U_i will be denoted by U_i^j where j ranges over 1, 2, ..., n. In addition, c is a vector with components c^j. The variables h and t are scalars.

In analogy with the preceding chapters we adopt the following definitions.

Definition 4.1

i. $R_{\overline{Y}} = [-\varepsilon , 1 + \varepsilon] \times [-\overline{Y}^1 , \overline{Y}^1] \times \ldots \times [-\overline{Y}^n , \overline{Y}^n]$

where $\varepsilon > 0$ and the $\overline{Y}^j$ are real

ii. $S = \{(t , y) \mid -\varepsilon \leq t \leq 1 + \varepsilon, y \in R^n\}$

Definition 4.2

The real functions f^j defined on S will satisfy Property A if f^j and all partial derivatives through order two exist and are continuous on S.

Definition 4.3

Let $\{\lambda^j\}$ be a set of fixed positive constants and let c^j be in $[-\lambda^j , \lambda^j]$ for each j. The general second order Runge-Kutta method for systems of equations is given by,

$$U_{i+1} = U_i + h\, f(t_i , U_i) + \frac{h}{2}\, G\left(t_i , U_i(c)\right) \tag{4-2}$$

where the components of the vector valued function G are given by,

$$G^j (t , y , c) = \frac{1}{c^j}\left[f^j\left(t + c^j h , y + c^j h\, f (t , y)\right) - f^j (t , y)\right] \quad \text{when } c^j \neq 0$$

and

$$G^j (t , y , 0) = h\, f^{j'} (t , y)$$

$$\equiv h\left[f^j_t (t , y) + f^j_y (t , y)\, f (t , y)\right]$$

when $c^j = 0$.

Note that in Definition 4-3 the variable c is a vector. This is a generalization of the conventional Runge-Kutta method for systems where c is treated as a scalar. Of course (4-2) includes the usual formulation as a special case.

Definition 4.4

The truncation (discretization) error induced by applying (4-2) to (4-1) is the vector τ_{i+1} defined by,

$$Y_{i+1} = Y_i + h\ f\ (t_i\ ,\ Y_i) + \frac{h}{2}\ G(t_i\ ,\ Y_i) + \tau_{i+1} \qquad (4-3)$$

Definition 4.5

The total (global) error at a point t_i of I_h is the vector,

$$E_i = U_i - Y_i$$

which has components

$$E_i^j = U_i^j - Y_i^j$$

For convenience, we assume that $Y_o = U_o$ so that $E_o = o$.

In view of the above discussion we can pose analogues of the propositions of section III.1. For any convenient norm for R^n we investigate the conditions under which;

A. there is a vector c for which $||E_n(c)||$ is minimized,

B. there is a unique c for which $||E_N(c)||$ is minimized,

C. there is a unique c for which $E_N(c) = 0$

We consider A first. If the f^j satisfy Property A it can be shown that the total error at $t = 1$ is a continuous function of c for fixed h. Since norms are continuous it follows that $||E_N(c)||$ is a continuous function of c. Because the c^j are in $[-\lambda^j\ ,\ \lambda^j]$ for each j the vector c lies in a compact subset of R^n. Thus, $||E_n(c)||$ attains a minimum for some value of c.

The other two propositions are more difficult to analyze. However, if each component f^j of the right hand side of (4-1) satisfies a certain convexity property, analogous to that of the previous chapter, some partial results can be obtained.

Definition 4.6

For f^j satisfying Property A the Hessian of f^j is the matrix of partial derivatives,

$$Hf\ (t, y) = \begin{pmatrix} f^j_{tt} & f^j_{ty^1} & \cdots & f^j_{t\ y^n} \\ f^j_{y^1t} & f^j_{y^1y2} & \cdots & f^j_{y^1y^n} \\ & \vdots & & \\ f^j_{y^nt} & f^j_{y^ny^1} & \cdots & f^j_{y^{nn}} \end{pmatrix}$$

where it is understood that the partial derivatives are to be evaluated at a point $(t\ ,\ y)$.

Definition 4.7

Let $S_1 = \{(1, x) \mid x \in R^n\}$. Let f^j satisfy Property A and let the Hessian of f^j be positive definite on S_1. That is,

$$(1\quad x)\ H\ f\ (t\ ,\ y) \begin{pmatrix} 1 \\ x \end{pmatrix} > 0$$

for all $(t\ ,\ y)$ in S and all $(1, x)$ in S_1. Then we shall denote the family of all such functions f^j by $\mathcal{F}$.

In what follows, we shall assume for f satisfying Property A , $c^j \in [-\lambda^j\ ,\ \lambda^j]$ and h sufficiently small that there is a compact set $R_{\overline{Y}}$ (see Definition 4.1) such that,

$$|U_i^j - U_o^j| \leq \overline{Y}^j$$

$$|U_i^j + c^j\ h\ f^j\ (t_i\ ,\ U_i) - Y_o^j| \leq \overline{Y}^j$$

$$|Y_i^j + c^j\ h\ f^j\ (t_i\ ,\ Y_j^i) - Y_o^j| \leq \overline{Y}^j$$

for all $i \leq N - 1$.

Lemma 4.1

If f satisfies Property A, h is sufficiently small and $c^j \in [-\lambda^j\ ,\ \lambda^j]$ then the jth components of the truncation error (4-3) can be written,

$$\tau_{i+1} = h^3 \left[\frac{Y^{j'''}(\xi_{i\ +1})}{6} - \frac{c^j}{4}\ z^T\ \overline{H}\ z \right]$$

where $\overline{H} = Hf^j$ evaluated at $(\overline{t}\ ,\ \overline{Y})$, $\overline{t}_i = t_i + \theta_i^j\ h$, $\overline{Y}_i^j = Y_i^j + \theta_i^j\ c^j\ h\ f^j\ (t_i\ ,\ Y_i)$, $0 < \theta_i^j < 1$, $t_i < \xi_{i+1} < t_{i+1}$ and $z^T = \left(1,\ f\ (t_i\ ,\ Y_i\ )\right)$.

Proof:

The proof is nearly identical to that of Theorem 2.1.

We consider first the case where c is treated as a scalar; more precisely, let ξ be real and set $c^j = \xi$ for all j. Then, for a fixed h, the total error E_N can be viewed as a vector valued function of the scalar ξ. In particular, the components E_N^j are functions of ξ. In this context the following theorem can be proved.

Theorem 4.1

Let $f^j \in \mathcal{G}$ for all j, h be sufficiently small and $\xi \in [-\lambda^j, \lambda^j]$ for all j. If, $\frac{\partial f^j}{\partial y^m} \geq 0$ when $j \neq m$, then $E_N^j(\xi)$ is strictly increasing function of ξ for all j.

The proof of this theorem follows along the lines of Theorem 3.1 of the previous chapter. Note, however, the additional condition on the partial derivatives of f^j.

What this Theorem says is that all the components of E_N are strictly increasing in ξ. This would imply that it may be possible to select a unique ξ so that one of the components of E_N vanishes. This can be established in a fashion analogous to that of Theorem 3.1. A practical example of such an application is included in the next chapter.

At first sight, the condition that the "off diagonal" partial derivatives be non-negative appears to be overly restrictive. In practice, this may not be important. First, systems of differential equations are often obtained by transforming a higher order equation to a first order system. The process by which this is achieved yields a system in which the off diagonal partial derivatives of the first n-1 equations are all zero. Secondly, practical experience with the method has revealed that the error components increase with ξ even when the partial derivative condition is violated.

Aside from the partial results we have obtained problems B and C remain open.

V. APPLICATIONS AND EXAMPLES

This chapter contains some experimental results obtained by applying our technique to a set of sample problems. The use of the transformations suggested in Chapter III will be illustrated. We also show some results associated with a successful application of the method to a real time control system problem.

Problem P1

$$z' = z \qquad 0 \le t \le 1$$
$$z_o = 1$$

This example was chosen because $D_f^2 \equiv 0$. Thus, the error of the method does not depend on c. The problem was solved for $N = 5$ and $N = 10$. The results are plotted in Figure 5-1 and will be used for comparison with the results obtained in the next problem.

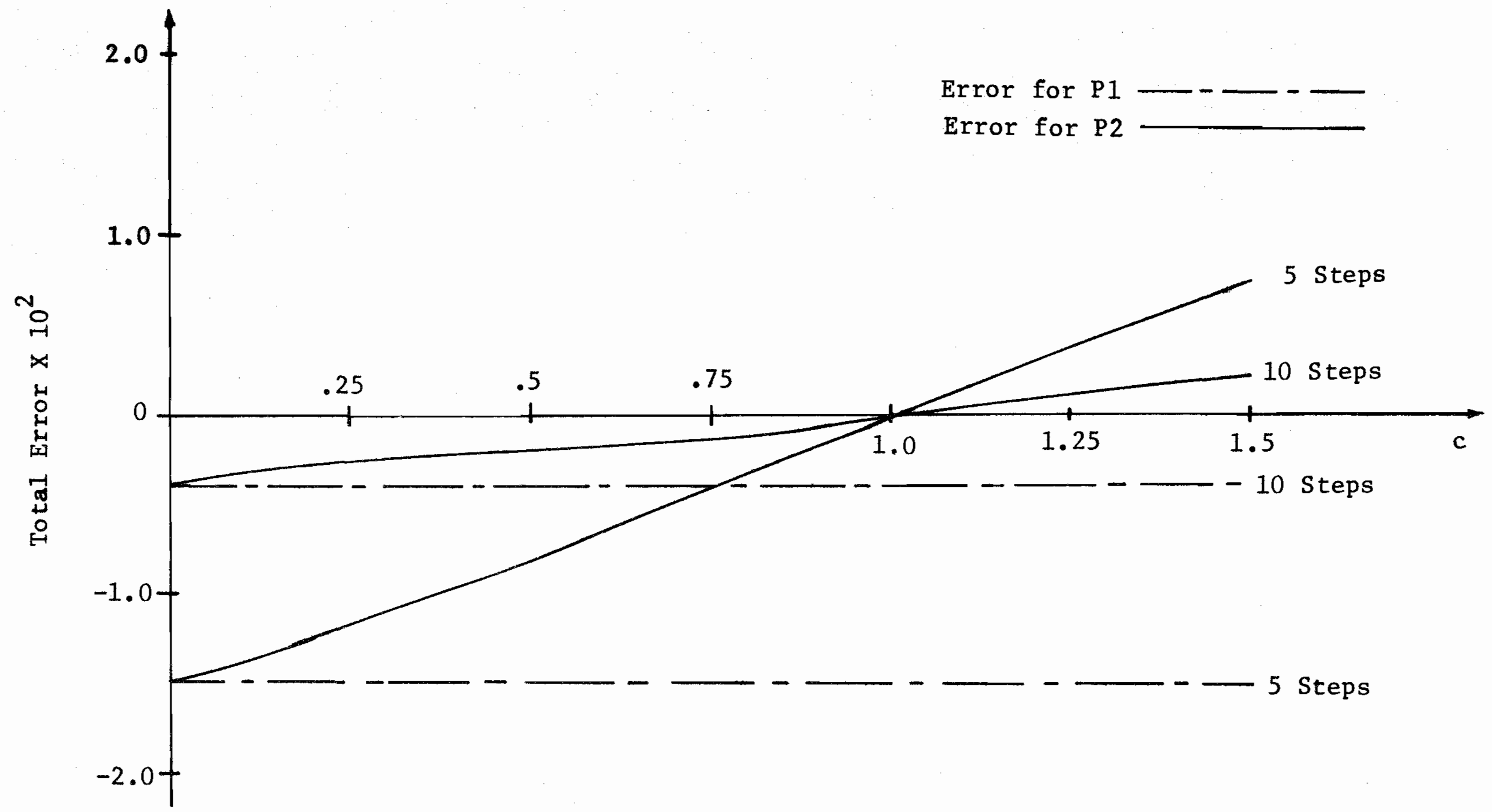
Problem P2

In this problem we introduce a transformation of the previous problem which produces a strictly increasing error function. We first note that in P1, $f_z > 0$ which is case 1 of the transformations of Chapter III. Therefore, we let,

$$g = -\frac{L}{2M}t^2$$

where L and M are to be specified. Since $D_f^2 \equiv 0$ and $f_z = 1$ we take $L = M = 1$.

Under the transformation $Y = z - \frac{t^2}{2}$ the problem of P1 becomes,



Error vs. **c** For P1 and P2
Figure 5-1

$$Y' = Y + \frac{t^2}{2} - t \qquad 0 \leq t \leq 1$$

$$Y_o = 1$$

which has the exact solution $Y = e^t - \frac{t^2}{2}$. The transformed problem was solved using the same grid as in P1 for a sample set of c values. The total error is plotted as a function of c in Figure 5-1. The total error is an increasing function of c as was desired. The value of c which gives zero error is slightly larger than one for both N = 5 and N = 10. We note that the optimal method of Ralston (c = 2/3) does not produce zero error.

In Figure 5-2 the total error is plotted for several values of c as a function of t_i. It is interesting to note that the error at all grid points tends to be small when the optimal value of c is used ($c \gtrsim 1$).

Problem P3

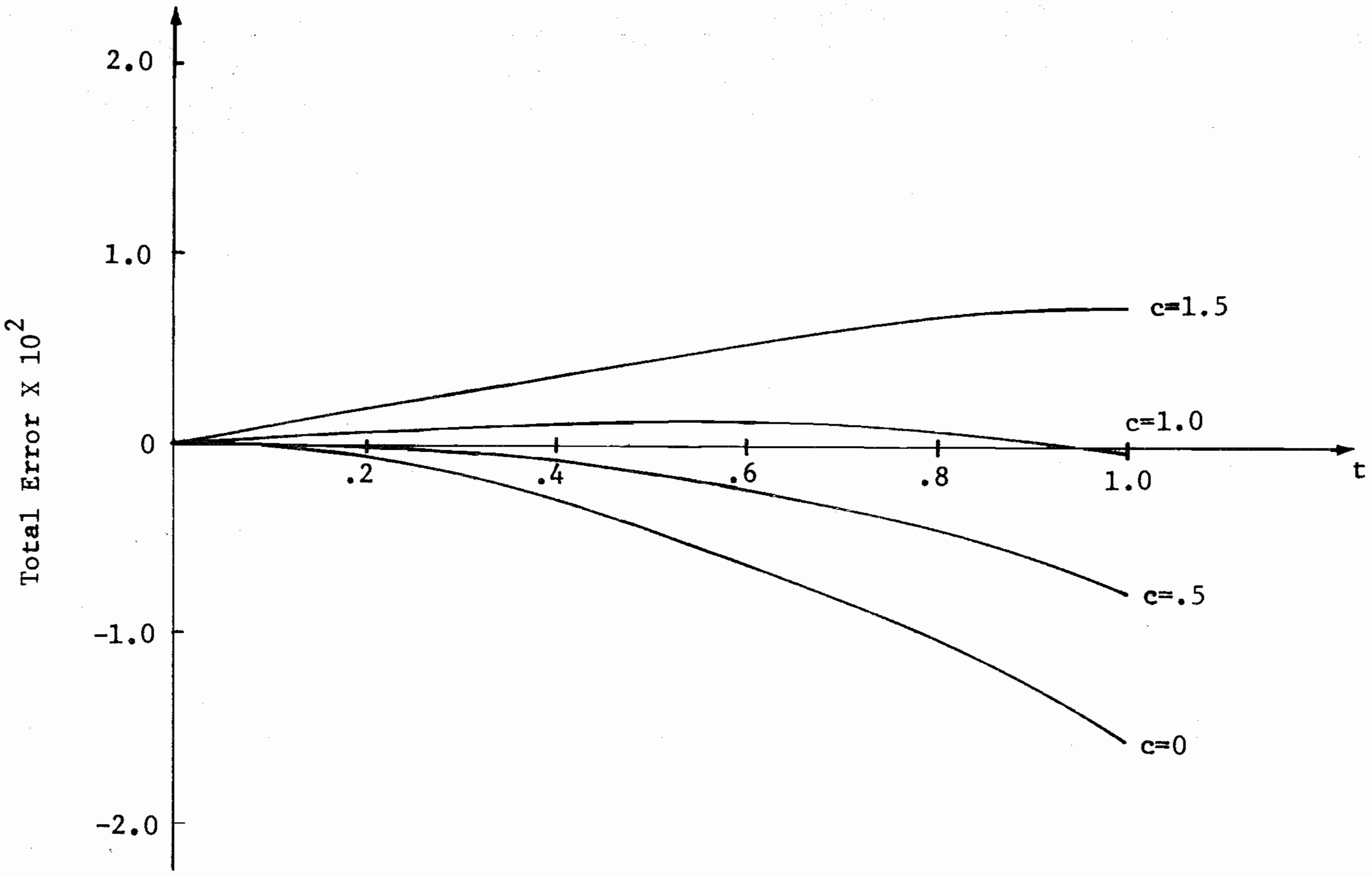
$$Y' = \frac{-Y^3}{2} \qquad 0 \leq t \leq 1$$

$$Y_o = 1$$

which has the exact solution $Y = (t + 1)^{-\frac{1}{2}}$. In this case,

$$D_f^2 (t , y ; t_i , U_i) = \left(1 - \frac{U_i^{\ 3}}{2}\right)\begin{pmatrix} 0 & 0 \\ 0 & -3Y \end{pmatrix} \begin{pmatrix} 1 \\ \frac{-U_i^{\ 3}}{2} \end{pmatrix}$$

$= - \frac{3}{4} Y U_i^6 < 0$ if $Y > 0$. Thus, the total error should be a strictly decreasing function of c for sufficiently small h. The



Error vs. t For P2

Figure 5-2

problem was solved for N = 5 and the results are shown in Figure 5-3. As expected, the error is decreasing, with the optimal $c \simeq 1.5$.

Problem P4

$$Y' = Y \text{ Cos } t \qquad 0 \leq t \leq 1$$

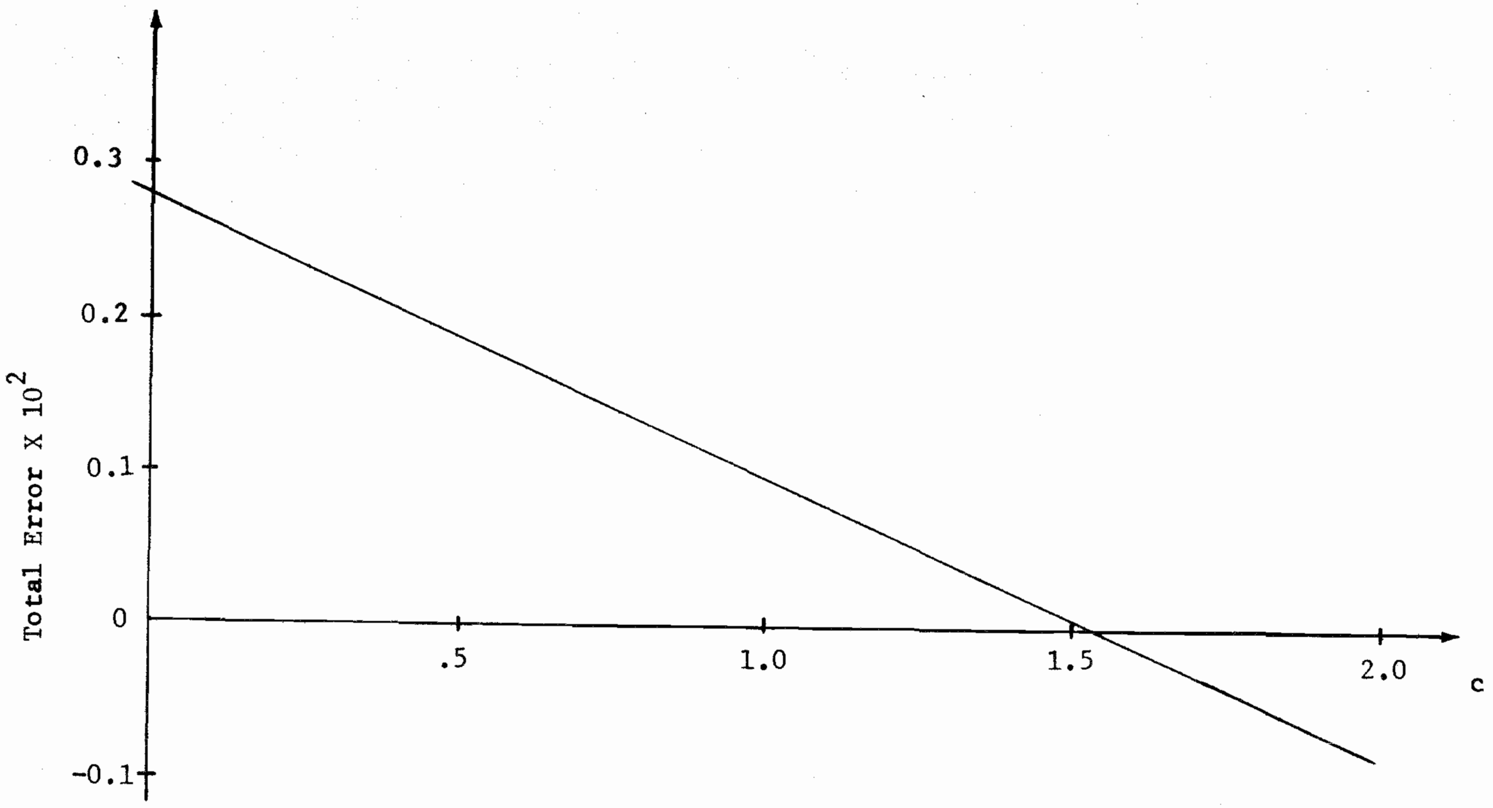
$$Y(o) = Y_o \qquad 1 \leq Y_o \leq 2$$

where the exact solution is $Y = Y_o \, e^{\text{Sin } t}$.

This problem was solved for both 5 and 10 integration steps for three initial conditions taken from the interval [1 , 2]. The total error is plotted as a function of c for each initial condition in Figure 5-4 (5 step case). These results are interesting because the optimal c is insensitive to the initial condition. Similar results were obtained for the 10 step case.

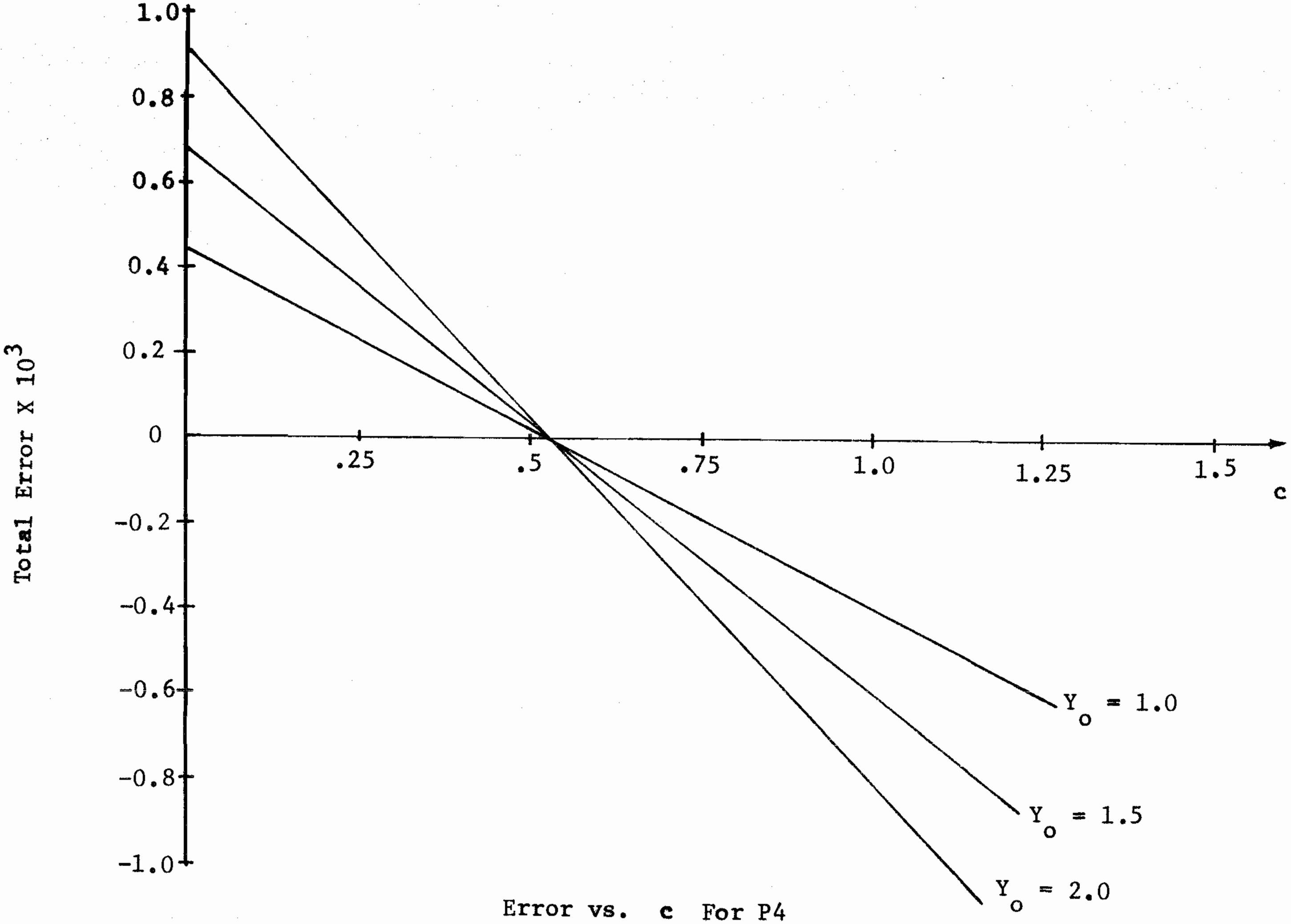
Problem P5

The method has been successfully applied to a real time control system (Toms et. al., [12, 13]). This application involves the development of a weapon delivery algorithm suitable for use in a small airborne digital computer. The primary function of this algorithm is to predict the impact point of projectiles for a given set of release conditions. In this system the impact point of the projectile is computed almost continuously. Deviations of the impact point from the target location are used to provide steering commands to the aircraft. With the high speeds involved there is an inherent requirement that the computations be done



Error vs. c For P3

Figure 5-3



Error vs. **c** For P4
Figure 5-4

in real time. In addition, the typical airborne computer has a basic computation rate which is much slower than large ground-based computers. Finally, the airborne computer is simultaneously performing other computational tasks related to flight control and navigation. The point is, that in a system like this, computation time is at a premium.

The primary computation involved in the trajectory algorithm is the numerical solution of a system of ordinary differential equations. These equations govern the flight path of ballistic objects released within the atmosphere of the earth. These equations need to be repeatedly solved for continuously changing initial conditions.

Using time as the independent variable and employing suitable simplifying assumptions the equations of motion for a projectile can be developed [8]. The principle assumptions are that the projectile is a point mass acted on only by the force of gravity and retardation forces due to air resistance. The Earth is assumed to be flat and non-rotating. The trajectory is restricted to a vertical plane by ignoring cross-track effects such as winds.

It has been noticed that if the equations of motion are re-formulated with range and altitude as independent variables their numerical solutions are dramatically better behaved [12]. This is the formulation used in the trajectory algorithm.

The following symbols are employed:

X = range

Z = altitude

V_X = the velocity component in the X direction

V_Z = the velocity component in the Y direction

g = the gravitational constant

H = the coefficient of drag

S = frontal area

P = air density (depends on Z)

V = projectile speed

M = Mach number

$C_D(M)$ = ballistic coefficient

The ballistic coefficient C_D is a known function of Mach number which in turn depends on V and certain atmospheric properties related to air temperature. The relation between speed and the velocity components is $V = (V_x^2 + V_z^2)^{\frac{1}{2}}$. In addition, we have,

$$V_x = V \text{ Cos } \theta$$

$$V_z = V \text{ Sin } \theta$$

where θ is the instantaneous angle between the velocity vector and the horizontal axis. These relations permit the computation of initial values for the velocity components.

With the above notation the equations of motion with range as the independent variable can be easily developed from those where time is the independent variable [12]. They are,

$$\frac{d\,Z}{d\,X} = \frac{V_z}{V_x}$$

$$\frac{d\,V_x}{d\,X} = -\,H \tag{5-1}$$

$$\frac{d\,V_z}{d\,X} = -\,\frac{H\,V_z - g}{V_x}$$

$$\frac{d\,t}{d\,X} = \frac{1}{V_x}$$

The corresponding equations with altitude as the independent variable are,

$$\frac{d\,X}{d\,Z} = \frac{V_x}{V_z}$$

$$\frac{d\,V_z}{d\,Z} = -H - \frac{g}{V_z} \tag{5-2}$$

$$\frac{d\,V_x}{d\,Z} = -\,H\,\frac{V_x}{V_z}$$

$$\frac{d\,t}{d\,Z} = \frac{1}{V_z}$$

The system (5-1) has a singular point when $V_x = 0$ while (5-2) is singular when $V_z = 0$. The trajectory algorithm uses the system (5-1) during the early portion of the trajectory when θ is not too large. As the projectile starts to fall steeply the algorithm automatically switches to the system (5-2). For most projectiles, including the one used in the example which follows, the impact angle is relatively small. This means that most of the time the system 5-1 is being used.

In the trajectory computation the important component of the solution vector is the impact range. For this reason we chose to select the vector c in (4-2) too minimize the error in range at impact. The resulting optimal vector has three equal scalar components which we shall denote by c.

The process for determining c so that the range error is small (zero) is best illustrated by example. For this purpose we selected a projectile which is retarded by a parachute device. For such projectiles the drag forces are substantial and the deceleration is very rapid. For such projectiles, a very fine grid is required if conventional Runge-Kutta methods are used. We shall show that our method yields a satisfactory solution for a very coarse grid.

For the initial condition sample, 27 data points were selected. They were all combinations that can be formed from,

$$V_o = 200,\ 400,\ 600 \text{ (Knots)}$$

$$Z_o = 100,\ 3000,\ 5000 \text{ (ft)}$$

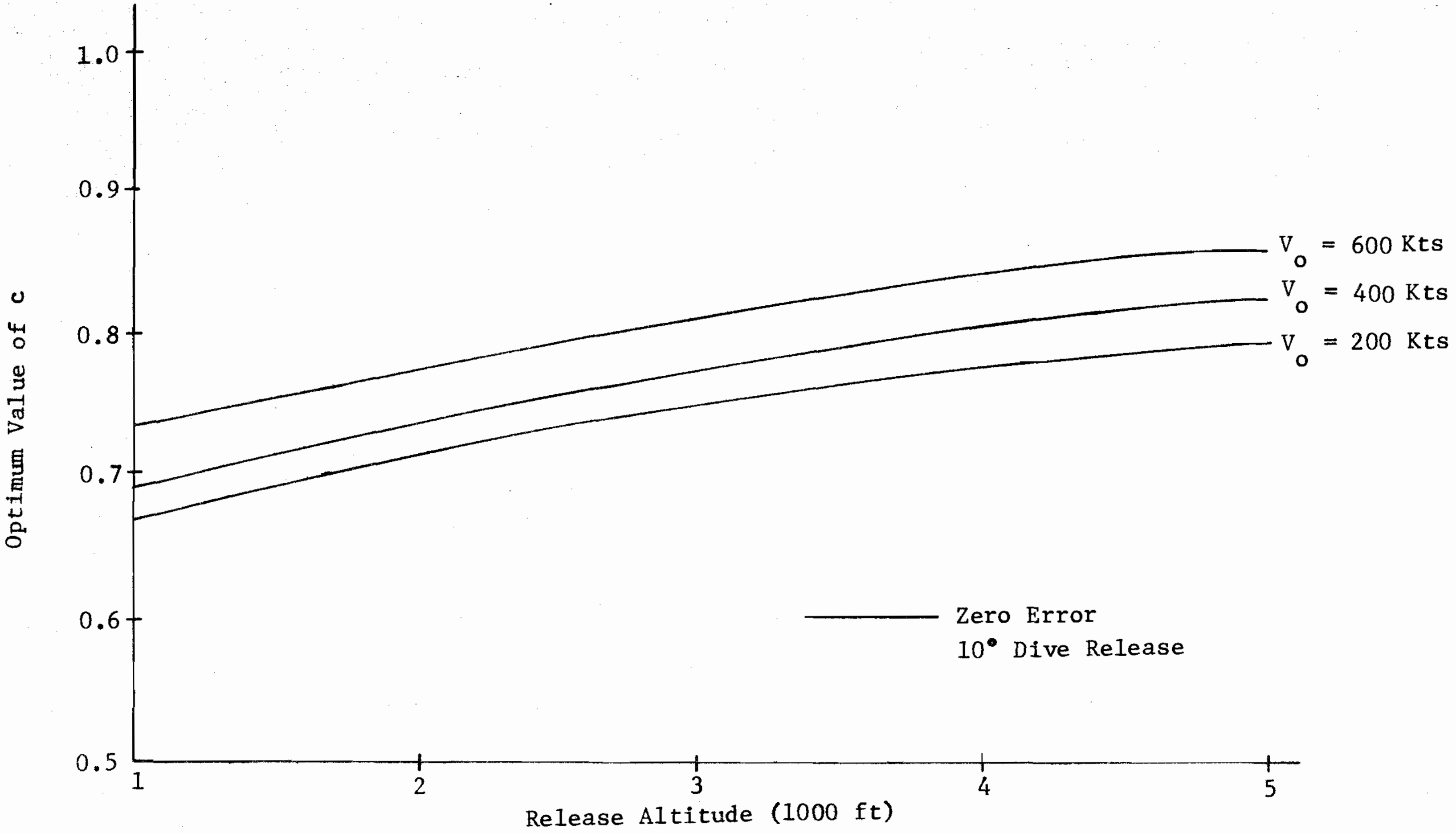
$$\theta_o = 0,\ -15,\ -30 \text{ (degrees)}$$

The resulting combinations of initial conditions span the entire delivery envelope for this projectile.

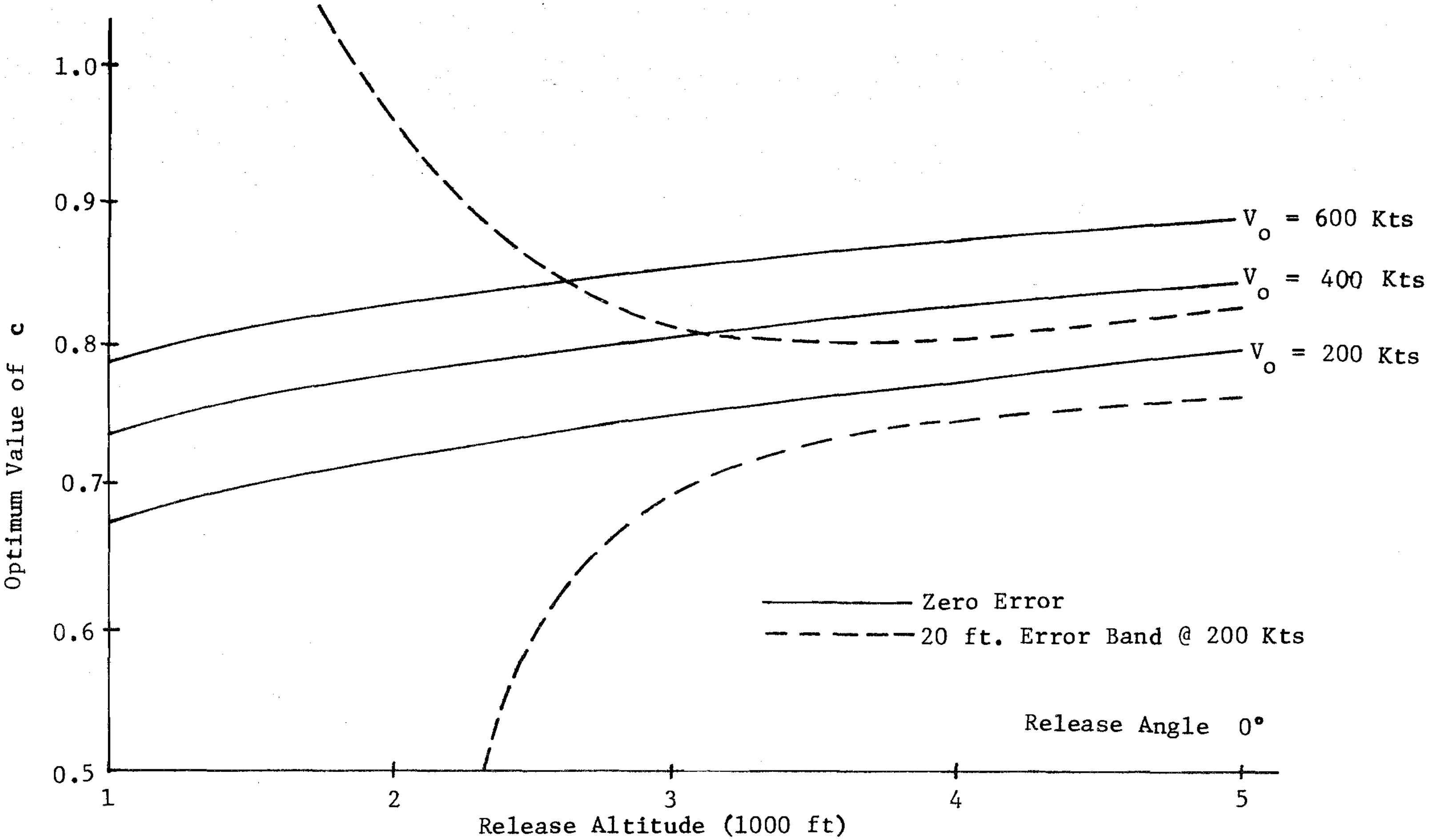
The ballistics equations we solved for each of the 27 sample initial conditions using a very small step size and the classical fourth order Runge-Kutta method. For our purpose this solution was treated as being the exact one. For each set of initial conditions the formula (4-2) was used for a selected sequence of c values. These solutions were obtained using only five integration steps. The range error at impact turned out to be monotonically increasing with c for each initial condition. This permitted the selection of a value of c which gave zero error for a particular initial condition. Some plots of these optimal c values versus the initial conditions are given in Figures 5-5 and 5-6. The solid lines in Figure 5-5 show the values of c for which a zero impact error was attained. The dashed lines enclosing the 200 knot curve show the values of c for which the error was ± 20 feet. This gives an indication of the sensitivity of the error to the value of c. In this instance the approximation of c as a function of initial conditions can be relatively gross.

The optimal c values were approximated by a linear function in V_o, θ_o and Z_o. This was done using a standard least square curve fit of the 27 data points. This approximation was then used in conjunction with (4-2) to again obtain solutions using five steps. Initial conditions were chosen both at and between the 27 sample points. The maximal error was found to be much less than ten feet for all cases.

Similar results have been obtained for other projectiles. Since many projectiles have similar drag characteristics the



Optimal c for Dive Release
Figure 5-6



Optimal c for Level Release

Figure 5-5

same approximating function for c is used for classes of projectiles. Eventually, it is expected that the drag coefficient will be included directly in the optimal c calculation. In this way a single approximating function for c will suffice for all projectiles processed by the system.

In an application of this type there is an obvious tradeoff between the allowable complexity in the approximating function for c and the number of steps employed. As the step size decreases the error becomes relatively insensitive to the value of c. On the other hand, if the steps are large a very precise fit of the optimal c data may be required.

VI RECOMMENDATIONS AND CONCLUSIONS

The preceding development generates several interesting topics for future investigations. For example, we have primarily used the global error in the process as a criteron for optimization. It was indicated in Chapter IV that sometimes a component of the error can be made zero by appropriate selection of c. The last example in Chapter V shows that this can be a desirable criterion in some applications. In other instances an alternate criterion may be more useful. That is, it may be more important to minimize some norm of the error instead of a component.

A similar study could be performed for Runge-Kutta methods of higher order where there would be more than one free parameter to adjust. A higher order method may prove to be worthwhile in some applications. The analytical treatment in this case appears to be considerably more complex. Since higher order methods involve more than one parameter the approximation of these as a function of initial conditions would likely be more difficult. One possible approach would be to reduce the set of free parameters to a single one by making some of them depend on the others. For example, in the case of a fourth order method there are two parameters say a and b. One could select $b = 2a$. In this way there is only one free parameter to adjust but the method is still fourth order.

One of the principle features of the method we have developed is that the numerical integration formula depends on the initial conditions present at the time of solution. In essence, the formulas used are weighted so that a small error is produced for a specific set of initial conditions. It is reasonable to postulate that this basic idea can be applied to other numerical integration methods. This would include other single step procedures such as implicit Runge-Kutta methods and also multi-step methods. In addition, it is probable that the same sort of thing could be done for boundary value problems. Thus, this general approach should be applicable in any situation where systems of ordinary or partial differential equations are to be repeatedly solved for similar boundary conditions. One example of this type is a digital simulation of electrical power systems networks. Another, might be, a simulation of transient heating effects in nuclear reactors.

Some of the recently developed numerical integration procedures utilize dynamically computed error estimates in order to change the order of the method as the integration proceeds [4]. This is akin to those methods in which the integration step size is modified in order to account for local variations in the estimated error. The variable order methods sometimes are much more efficient than fixed order methods [4]. However, they have the disadvantage of being difficult to code and do not seem to be effective when discontinuities are present in the differential equations [4]. One can postulate that the procedure developed in the previous chapters can be modified to provide a dynamically adjustable weighting factor. What we have in mind here is a method which uses a local error estimate in order to select an appropriate value of the free parameter. In such a method the order of the process would remain fixed. This would permit the use of a low order method which would simplify the coding problem and might yield good results in the neighborhood of discontinuities. This approach should be explored.

The principle unsolved theoretical problem associated with our method is to determine the conditions under which there is a vector c for which $E_N(c) = 0$ for a system of differential equations. A possible approach to the problem is to assume the existence and uniqueness of such a c on some compact set in R^n. Then one could apply such devices as the contractive mapping principle, the Newton-Kantorovich Theorem or the Implicit Function Theorem to deduce what properties the right hand side of the differential equation must have to guarantee a solution. Our attempts have so far not produced a complete answer.

BIBLIOGRAPHY

1. Ceschino, F. and Kuntzmann, J., "Numerical solution of initial value problems" Prentice-Hall, Inc., New Jersey, 1966, 318 p.

2. Henrici, P., "Discrete variable methods in ordinary differential equations," John Wiley and Sons, N.Y., fourth printing, 1964, p 407.

3. Hildebrand, F.B., "Introduction to numerical analysis", McGraw-Hill, New York, 1956, p 511.

4. Hull, T. E., Enright, W. H., Fellen, B. M. and Sedgwick, A. E., "Comparing Numerical methods for ordinary differential equations," SIAM J. Numer. Anal., Vol. 9, No. 4, Dec. 1972, p 603-637.

5. Hull, T. E. and Johnston, R. L., "Optimum Runge-Kutta methods", Mathematics of computation, Vol 18, No. 86, April 1964, p 306-310.

6. Johnston, R. L., "On optimum Runge-Kutta methods for the numerical solution of ordinary differential equations", MA Thesis, University of British Columbia, 1961.

7. Kuntzmann, J., "Deux formules optimales du type de Runge-Kutta," Chiffres, V. 2, 1959, p 21-26.

8. McShane, E. J., Kelley, J. L., and Reno, F. V., "Exterior ballistics," University of Denver Press, Denver, Colorado, 1953, pp. 239-240.

9. Ralston, A., "Runge-Kutta methods with minimum error bounds," Mathematics of Computation, Vol. 16, 1962, p 431-437; Corrigendum, V 17, 1963, p 488.

10. Rockafellar, B. T., "Convex analysis", Princeton University Press, Princeton, N.J. 1970, p 27.

11. Toms, R. M., Onyshko, S., Etter, R. F., and Hamilton F., "Algorithm for fire control," D162-10026-1, June 1969, The Boeing Company, Seattle, Washington.

12. Toms, R. M., Onyshko, S., Hamilton, F., and Hanvey, L.A., "Fire control algorithm applications", D162-10250-1, June 1970, The Boeing Company, Seattle, Washington.