ELECTRICAL ENGINEERING

ENGINEERING EXPERIMENT STATION

AUBURN UNIVERSITY

AUBURN, ALABAMA

QUANTIZATION ERROR-BOUNDS FOR HYBRID CONTROL SYSTEMS

PREPARED BY

SAMPLED-DATA CONTROL SYSTEMS GROUP

AUBURN UNIVERSITY

C. L. Phillips, Project Leader

Thirteenth Technical Report

28 May, 1968 to 28 September, 1968

CONTRACT NAS8-11274
GEORGE C. MARSHALL SPACE FLIGHT CENTER
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
HUNTSVILLE, ALABAMA

Approved By:

C. H. Holmes
Head Professor
Electrical Engineering

Submitted By:

C. L. Phillips
Professor
Electrical Engineering

FOREWORD

This document is a technical summary of the progress made since May 28, 1968, by the Auburn University Electrical Engineering Department toward fulfillment of phase B of Contract No. NAS8-11274. This contract was awarded to Engineering Experiment Station, Auburn, Alabama, May 28, 1964, and was extended September 28, 1966 by the George C. Marshall Space Flight Center, National Aeronautics and Space Administration, Huntsville, Alabama.

SUMMARY

The process of approximately representing more precise variables with variables of decreased word length is called "quantization." Several methods are given for determining bounds on system errors, which result from quantization in the digital components of a hybrid control system.

The concept of "Uniform Boundedness of Solutions," which arises in connection with Liapunov stability theory, is extended to discrete-time systems. This is accomplished by establishing the properties of a positive-definite function and its first forward-difference which are requisite to the definition of two sets M and M_r, in the state space, such that, if the system state enters M, it cannot leave M_r. The boundedness of motions idea is used to compute a system error-bound for each quantizer acting singly. A composite error bound is computed by combining these bounds. It is further established that an error-bound may be computed from the boundedness of motions theory by determining simultaneously the contributions of all quantizers.

The problem of establishing a least upper-bound on quantization error is formulated as a discrete-time, optimal-control theory problem. A closed form solution is given for the least upper-bound on quantization error at any sampling instant NT by the application of the "Discrete Maximum Principle." This solution is verified and generalized with another independent development. The optimal approach is used to obtain a quantization error-bound for a standard system. The resulting error-bound is significantly stronger than bounds which have been obtained

for this system by the application of other methods presented in the literature.

Finally, the z-transform theory is used to develop a second method for generating a least upper-bound on quantization error.

LIST OF PERSONNEL

The following named staff members of Auburn University have actively participated on this project:

C. L. Phillips--Professor of Electrical Engineering

R. K. Cavin III--Graduate Assistant in Electrical Engineering

D. L. Chenoweth--Graduate Assistant in Electrical Engineering

J. C. Johnson--Graduate Assistant in Electrical Engineering

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

A, B, C, D, E	symbols denoting matrices
$\underline{b}$, $\underline{c}$	symbols denoting column vectors
E^n	symbol representing n-dimensional Euclidean space
[f(k)]	the number sequence f(0), f(1),...
F(z)	the z-transform of [f(k)]
iM	means "the interior of set M"
M, $\hat{M}$, M_r, R, S,	symbols denoting point sets in E^n
P,Q	positive definite symmetric matrices
$\underline{p}(k)$	vector representing the state of the adjoint system at the k^{th} sampling instant
$\underline{q}(k)$	vector whose components are the difference between quantizer input and output at the k^{th} sampling instant
$\hat{\underline{q}}(k)$	an optimal quantizer input at the k^{th} sampling instant
Q(·)	means "the quantized value of (·)"
s	complex variable associated with the Laplace transform
$V[\underline{\nu}(k)]$	a positive definite, quadratic function of $\underline{\nu}(k)$
$\underline{x}(k)$	vector denoting system state at the k^{th} sampling instant
$\overline{\underline{x}}(k)$	vector denoting state of system containing quantizers at k^{th} instant
y(k)	scalar representing a system output

$\bar{y}(k)$	scalar representing the output of a system containing quantizers
$y_v(k)$	scalar representing difference between $y(k)$ and $\bar{y}(k)$
$y_v[\hat{\underline{v}}(k)]$	output resulting when optimal input sequence $[\hat{\underline{q}}(j);\ j=0,1,\ldots,k-1]$ is applied to a system
z	complex variable associated with the z-transform
δM	the boundary of set M
$\Delta V[\underline{v}(k)]$	the first forward difference for $V[\underline{v}(k)]$
ϵ	means "is a member of"
$\lambda(A)$	symbol denoting an eigenvalue of the matrix A
$\underline{v}(k)$	vector representing difference between $\bar{\underline{x}}(k)$ and $\underline{x}(k)$
$\hat{\underline{v}}(k)$	state resulting from application of optimal input sequence $[\hat{\underline{q}}(j),\ j=0,\ldots,k-1]$
ϕ, Ψ	symbols representing functions from E^1 into E^1

I. INTRODUCTION

The advances which have occurred in the past decade in the area of integrated electronics have many far-reaching implications for the design of automatic control systems. One of these implications is that economical, special purpose digital computers, which are capable of on-line process control, can be readily constructed from commercially available solid-state components. As a result of the increased availability and the flexibility of these digital elements, the designer of an automatic control system is presented with an attractive alternative to the conventional analog compensation of a system. The problem of specifying the difference equations which must be implemented with the digital device in order to obtain satisfactory system performance can be readily solved, for systems whose compensators use fixed interval sampling, by application of the z-transform theory in conjunction with classical synthesis methods.

A property of all physically realizable digital devices is that all variables processed by these devices are represented by binary words of finite length. Because of this characteristic, an analog signal which is applied to a digital device must be represented over its range by a finite number of binary words. It follows that errors of approximation are unavoidable in the process of analog to digital conversion. Approximations of the type that occur in analog to digital conversion belong to a more general class of "quantization approximations" which are defined as follows:

Let f be a function from the interval [a,b] into R, the set of all real numbers, and let h represent a number greater than zero. Then "g is a quantization approximation of f" if g is a function from [a,b] onto S, a finite subset of R, and if $|f(x) - g(x)| \leq h$ for all $x \in [a,b]$.

As an example of quantization in an analog to digital converter, consider the case where an analog signal whose maximum magnitude is 10 volts is to be represented by a binary word containing three bits. If one of these bits is used to denote the sign of the analog signal and if the least significant bit is chosen to represent 3.333 volts, then the input-output characteristic of the analog to digital converter is represented by Figure I-1. Devices exhibiting this type of characteristic are commonly called "truncators" in the literature [1], [2]. The motivation for this title is evident from Figure I-1 since

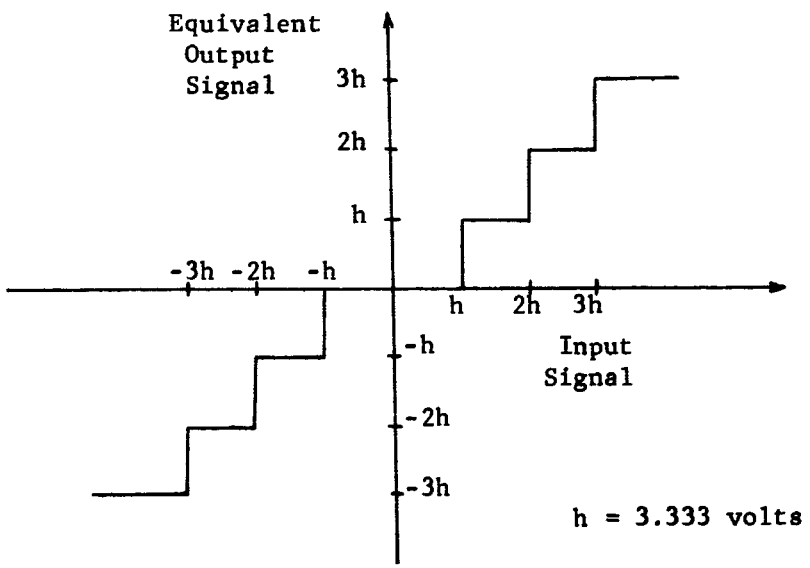

Figure 1. Input-output characteristic of a truncator.

the output signal is simply a truncation of the input signal. Note that a maximum error of approximation of h can be associated with the truncation characteristic of Figure I-1.

By far the most common quantization characteristic is the familiar "round-off" input-output characteristic of Figure I-2.

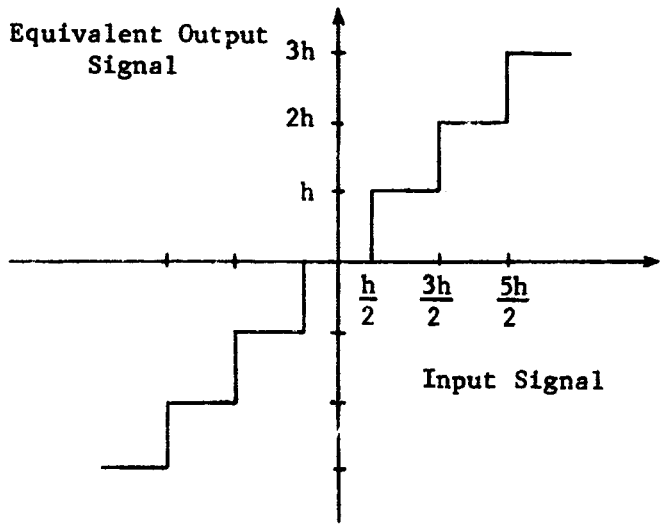

Figure 2. Input-output characteristic for round-off quantizer.

The round-off quantizer has a maximum approximation error of h/2.

There are many different types of quantization characteristics which can be realized. For example, one might utilize a non-uniform quantization characteristic to reduce approximation errors which occur at small input levels while accepting somewhat poorer approximations at larger input levels. In any event, all quantization characteristics exhibit the common property that the maximum magnitude of approximation error is bounded for all input levels below a specified limit.

There are two functions associated with a digital device, other than analog to digital conversion, that also involve the quantization phenomena. When an N bit digital variable is multiplied by an M bit constant, then a word of maximum length of M+N bits can result. However, it is quite often desirable to retain fewer than M+N bits in order to avoid an excessively complex machine or an unacceptable increase in computational time. Suppose, for example, that the binary variable 10.1 is to be multiplied by the coefficient, .11. The product of these two binary numbers is 01.111. If, however, the product is to be represented with only three bits, two of which are to the left of the binary point, one might choose the number 10.0 as the approximate product. If the least significant bit of the input signal represents 1/2 volt analog, then it follows that an analog equivalent of the digital multiplication is represented by Figure I-3, where Q has the same input-output characteristic as Figure I-2 with h=1/2 volt. The signal $e_s(t)$, which has a maximum range of 3 1/2 volts, can only attain a finite set of levels since it represents a digital variable.

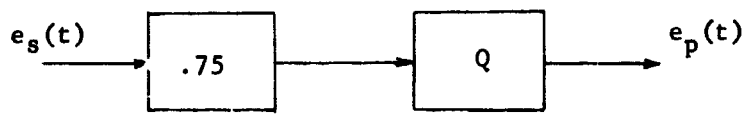

Figure 3. An analog representation of digital multiplication.

The third source of quantization in a digital device is the digital to analog converter. It is quite often the case that because of hardware limitations, it is not meaningful to convert more than a given number of bits of digital information into an analog signal even though a greater word length is carried in the machine. Consequently, the analog signal is a quantization of the more precise digital word.

The approximations due to quantization that occur in a realizable digital device are a potential source of system error when a system is digitally controlled. The primary objective of the work described in this dissertation is to develop methods for giving a quantitative measure of the system errors which result from quantization.

A survey of the literature pertaining to the effects of quantization on system performance is given in Chapter II. Chapter III contains a brief outline of the mathematical concepts that are used in the developments of Chapters IV, V, and VI. In Chapter IV, several related methods for determining bound on system quantization errors are given. All of the methods described in this chapter are based on the Direct Method of Liapunov and the related concept of "Uniform Boundedness of Solutions." The Discrete Maximum Principle is used in Chapter V to determine a least upper bound on system quantization error. Finally, in Chapter VI, results completely analogous to those given in Chapter V are obtained by utilization of transfer function methods. The conclusions which result from the work given in Chapters IV, V, and VI are contained in Chapter VII.

II. A SURVEY OF THE LITERATURE

The existing literature on the topic of amplitude quantization can be divided into essentially two categories. The study of the questions of system stability, of the existence of periodic motions of system variables, and of the character of these periodic motions if they exist comprises one of the major areas of endeavor as related by the literature. Significant among these studies are those of Tsypkin [3], Korshunov [4], Chow [5], and Biondi [6]. The second category of effort, and the one to which much of the published work belongs, pertains to the analytical determination of the errors which arise in a system containing quantizing nonlinearities. The idea of "quantization errors" implies the existence of a reference system with which one can compare the performance of the system containing quantizers. This reference system is almost always taken to be the system under investigation with all quantizing nonlinearities removed.

The available methods in the category of "quantization errors" can be subdivided into techniques having a statistical basis and techniques having a deterministic basis. The statistical techniques which have been presented in the literature can be used under certain conditions to obtain an estimate of the root-mean-square error due to quantization. In contrast, the published methods which use the deterministic approach invariably provide a means for computing an upper-bound on quantization error.

A. The Statistical Methods

W. R. Bennett [7] pointed out that, in many systems, a quantizer with its nonlinear staircase characteristic as shown in Figure II-1 can be replaced by a summing junction to which an independent source of noise is connected as shown in Figure II-2.

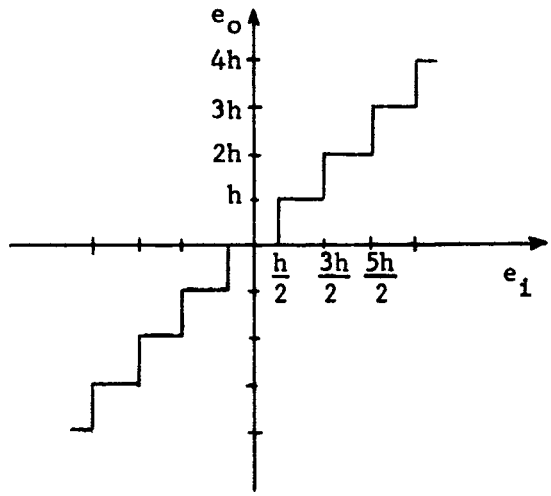

Figure 4. The staircase characteristic of a quantizer.

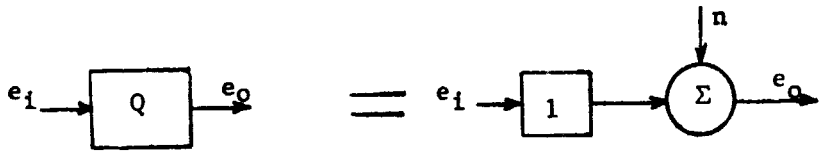

Figure 5. A statistical model of the quantizer.

Bennett theorized that the quantization noise can be considered to be independent of the input signal because the distortion spectrum of the quantizer output signal is practically independent of the input signal when it varies over several quantizer steps. He further asserted that even if the input signal does not extend over but a few quantizer steps, there is usually enough residual noise in actual systems to mask the relation between the quantization noise and the input signal.

B. Widrow [8,9] conducted extensive studies which gave mathematical foundation to the conclusions of Bennett. Widrow developed a "Quantizing Theorem" that is similar in structure to the "Sampling Theorem" of sampled-data theory. The Sampling Theorem states that in order to reconstruct the input signal from its sampled counterpart, the sampling frequency must be greater than or equal to twice the highest frequency present in the spectrum of the input signal. Widrow advanced the concept that quantization is a sampling process that acts not on the input function itself but on the probability density distribution of the input function. Widrow's observations led to the following statement of the Quantizing Theorem:

Let $\omega(e_i)$ denote the distribution density of the amplitude of the input signal, and let its characteristic function be given by

$$W_{e_i}(u) = \int_{-\infty}^{\infty} \omega(e_i) \epsilon^{-je_i u} de_i \quad . \tag{II-1}$$

Then when the quantization frequency, $2\pi/h$, is twice as high as the highest frequency component in $W_{e_i}(u)$, it is possible to recover $\omega(e_i)$

from the distribution of the output signal of the quantizer.

Widrow then established that if the Quantizing Theorem is satisfied, then the physical quantizer can be modeled as indicated by Figure II-2 where the distribution density of the noise is described as shown in Figure II-3.

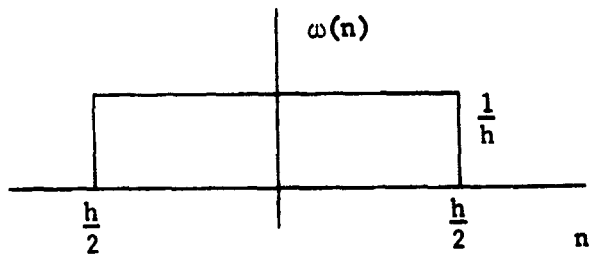

Figure 6. Noise distribution density.

B. The Deterministic Methods

The paper which has served as a basis for much of the deterministically oriented research on quantization errors in control systems was published by Bertram. [10] Because of the fundamental nature of this reference, and because of its direct bearing on this dissertation, the contents of this reference will be discussed in some detail.

Bertram considered the closed loop, digitally controlled, system shown in Figure II-4, in which the controller difference equation is implemented with a digital device. In order for the controller to be physically realizable, both the coefficients of the difference equation and the variables which are processed by the controller must be approximated by binary words of finite length.

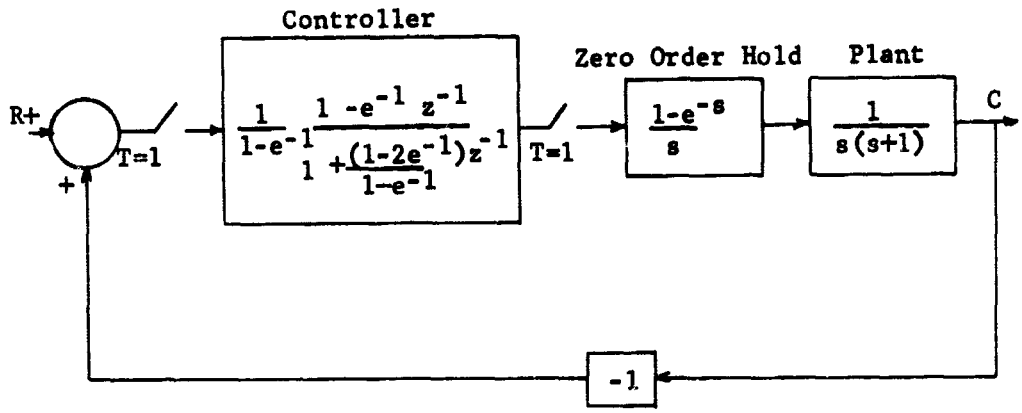

Figure 7. Block diagram of the system under consideration.

Bertram's paper provides a means of establishing an upper bound on system errors which arise due to quantization of process variables. The interesting problem of coefficient quantization is not considered.

Bertram used the phase variable form of the z-transform of $(1-e^{-s})/s^2(s+1)$ in order to obtain a discrete-time model for the Hold-Plant combination shown in Figure II-4. The resulting discrete-time representation of the system is shown in Figure II-5 with quantizers located at the input and the output of the controller and associated with each multiplication that is performed by the digital unit.

The symbols $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$ denote the state of the system shown by Figure II-5. Let x_1, x_2, and x_3 represent the state of a system

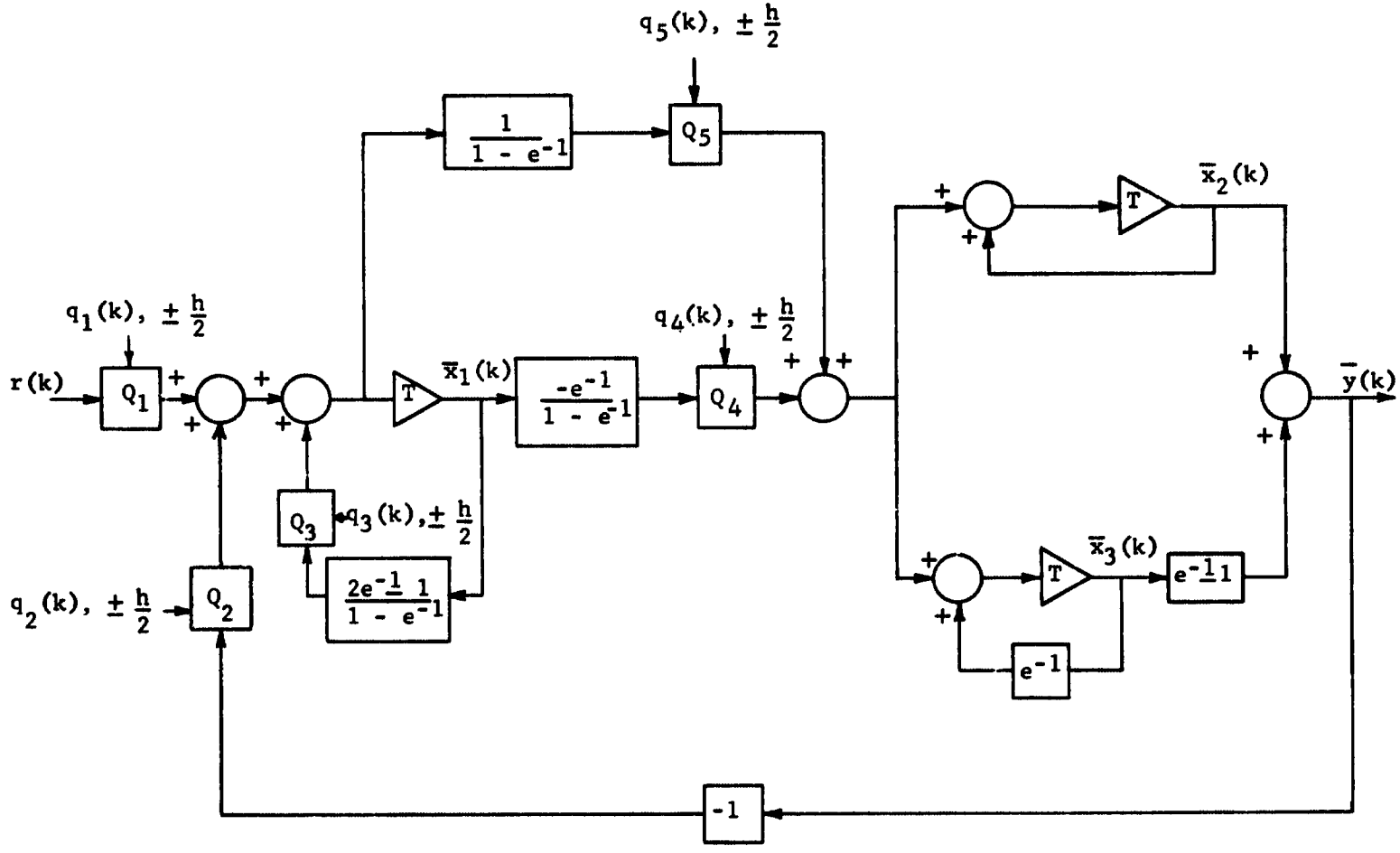

Figure 8.--A discrete-time model for the system of Figure II-4 showing quantization.

which is identical to that of Figure II-5 except that each of the quantizers is replaced by a unity gain element. It is evident then that this new system can be described by a set of first order linear equations of the form

$$\underline{x}(k+1) = A\underline{x}(k) + \underline{d}r(k) \tag{II-2}$$

$$y(k) = \underline{c}^T\underline{x}(k) \tag{II-3}$$

where

$$\underline{x}(k) = \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix}$$

$$A = \begin{bmatrix} \alpha_1 , & -1 , & 1-e^{-1} \\ \alpha_2 , & \frac{-e^{-1}}{1-e^{-1}} , & 1 \\ \alpha_2 , & \frac{-1}{1+e^{-1}} , & 1+e^{-1} \end{bmatrix} , \qquad \underline{d} = \begin{bmatrix} 1 \\ \frac{1}{1-e^{-1}} \\ \frac{1}{1-e^{-1}} \end{bmatrix} , \tag{II-4}$$

$$\underline{c} = \begin{bmatrix} 0 \\ 1 \\ e^{-1}-1 \end{bmatrix} , \qquad \alpha_1 = \frac{2e^{-1} - 1}{1 - e^{-1}} \text{ , and}$$

$$\alpha_2 = \frac{e^{-1} + e^{-2} - 1}{(1-e^{-1})^2} .$$

Now the recursion relations for the nonlinear system of Figure II-5 are

$$\bar{x}_1(k+1) = Q_3[\alpha_1\bar{x}_1(k)] + Q_2[-\bar{x}_2(k) + (1-e^{-1})\bar{x}_3(k)] + Q_1[r(k)] ,$$

$$\bar{x}_2(k+1) = Q_4\left[\left(\frac{-e^{-1}}{1-e^{-1}}\right)\bar{x}_1(k)\right] + \bar{x}_2(k) + Q_5\left\{\left(\frac{1}{1-e^{-1}}\right)\left[Q_3(\alpha_1\bar{x}_1(k)) + \right.\right.$$

$$Q_2\left(-\bar{x}_2(k) + (1-e^{-1})\bar{x}_3(k)\right) + Q_1(r(k))\Big]\Big\}, \qquad \text{(II-5)}$$

$$\bar{x}_3(k+1) = Q_4\left[\left(\frac{-e^{-1}}{1-e^{-1}}\right)\bar{x}_1(k)\right] + e^{-1}\bar{x}_3(k) + Q_5\left(\frac{1}{1-e^{-1}}\right)\Big[Q_3(\alpha_1\bar{x}_1(k))$$

$$+ Q_2(-\bar{x}_2(k) + (1-e^{-1})\,\bar{x}_3(k)) + Q_1(r(k))\Big]\Big\},$$

and

$$\bar{y}(k) = \bar{x}_2(k) - (1-e^{-1})\,\bar{x}_3(k)\,.$$

In the referenced paper, the same input-output quantization characteristic was used for each of the five quantizers in the system of Figure II-5. However, if a system is to be studied which employs different degrees of quantization, a set of equations entirely analogous to (II-5) can be written. The symbolism $Q_i(\cdot)$ which appears in (II-5) means "the quantized value of $(\cdot)$."

Equations (II-5) are in a relatively intractable form and do not readily convey some of the interesting properties of quantizing systems. In order to transform the system of equations (II-5) into a more useful form, a more detailed analysis of these quantized difference equations is required. For example, the first term on the right of the equation for $\bar{x}_3(k+1)$ can be rewritten as:

$$Q_4\left\{\left(\frac{-e^{-1}}{1-e^{-1}}\right)\bar{x}_1(k)\right\} = \left(\frac{-e^{-1}}{1-e^{-1}}\right)\bar{x}_1(k) + q_4(k) \qquad \text{(II-6)}$$

where $q_4(k)$ represents the roundoff error resulting from the quantization operation. Note that $q_4(k)$ is bounded in magnitude, i.e., $|q_4(k)| \leq h/2$, where h is the quantizer granularity. Another

of the terms of $\bar{x}_3(k+1)$ is

$$Q_5\left\{\left(\frac{1}{1-e^{-1}}\right)\left[Q_3(\alpha_1\bar{x}_1(k)) + Q_2\left(-\bar{x}_2(k) + (1-e^{-1})\bar{x}_3(k)\right) + Q_1(r(k))\right]\right\} =$$

$$Q_5\left\{\left(\frac{1}{1-e^{-1}}\right)\left[\alpha_1\bar{x}_1(k) + q_3(k) - \bar{x}_2(k) + (1-e^{-1})\bar{x}_3(k) + q_2(k) + r(k) + q_1(k)\right]\right\} = \tag{II-7}$$

$$\frac{\alpha_1}{1-e^{-1}}\,\bar{x}_1(k) + \frac{1}{1-e^{-1}}\,q_3(k) - \frac{1}{1-e^{-1}}\,\bar{x}_2(k) + \bar{x}_3(k) + \frac{1}{1-e^{-1}}\,q_2(k)$$

$$+\frac{r(k)}{1-e^{-1}} + \frac{1}{1-e^{-1}}\,q_1(k) + q_5(k)\,,$$

where $q_1(k)$, $q_2(k)$, $q_3(k)$, and $q_5(k)$ denote quantizer round-off at the k^{th} sampling instant. Further, each of $q_1(k)$, $q_2(k)$, $q_3(k)$, and $q_5(k)$ cannot exceed h/2 in magnitude. Therefore, $\bar{x}_3(k+1)$ may be written as

$$\begin{aligned}\bar{x}_3(k+1) = \alpha_2\bar{x}_1(k) &- \frac{1}{1-e^{-1}}\,\bar{x}_2(k) + (1+e^{-1})\bar{x}_3(k) \\ &+ \frac{1}{1-e^{-1}}\,r(k) + q_4(k) + \frac{1}{1-e^{-1}}\,q_3(k) + \frac{1}{1-e^{-1}}\,q_2(k) \\ &+ \frac{1}{1-e^{-1}}\,q_1(k) + q_5(k)\,.\end{aligned} \tag{II-8}$$

By a procedure entirely analogous to the one given for the $\bar{x}_3(k+1)$ relation, equations for $\bar{x}_1(k+1)$ and $\bar{x}_2(k+1)$ can be rewritten. The resulting system of equations is

$$\bar{\underline{x}}(k+1) = A\bar{\underline{x}}(k) + \underline{d}r(k) + B\underline{q}(k)\,, \tag{II-9}$$

$$\bar{y}(k) = \underline{c}^T\bar{\underline{x}}(k)\,, \tag{II-10}$$

where

$$\bar{\underline{x}}(k) = \begin{bmatrix} \bar{x}_1(k) \\ \bar{x}_2(k) \\ \bar{x}_3(k) \end{bmatrix} \quad , \qquad \underline{q}(k) = \begin{bmatrix} q_1(k) \\ q_2(k) \\ q_3(k) \\ q_4(k) \\ q_5(k) \end{bmatrix} \tag{II-11}$$

$$B = \begin{bmatrix} 1 & , & 1 & , & 1 & , & 0 & , & 0 \\ \frac{1}{1-e^{-1}} & , & \frac{1}{1-e^{-1}} & , & \frac{1}{1-e^{-1}} & , & 1 & , & 1 \\ \frac{1}{1-e^{-1}} & , & \frac{1}{1-e^{-1}} & , & \frac{1}{1-e^{-1}} & , & 1 & , & 1 \end{bmatrix} ,$$

and the A, $\underline{c}$ and $\underline{d}$ matrices are identical to those described in equations (II-4).

Note that equations (II-9) and (II-10) can be written directly from the system diagram of Figure II-5 if each quantizer, Q_i, is replaced by a summing junction at which a round-off signal, q_i, is added.

Define

$$\underline{\nu}(k) = \bar{\underline{x}}(k) - \underline{x}(k) \ . \tag{II-12}$$

If equations (II-2) are subtracted from (II-9), the resulting expression is

$$\underline{\nu}(k+1) = A\underline{\nu}(k) + B\underline{q}(k) \ , \tag{II-13}$$

where

$$\left|q_i(k)\right| \leq h_i/2 \text{ for all } k = 0, 1, 2, \ldots, \text{ and } i = 1,2,3,4,5.$$

Similarly,

$$y_\nu(k) = \bar{y}(k) - y(k) = \underline{c}^T\underline{\nu}(k) \ . \tag{II-14}$$

Equations (II-13) and (II-14) imply that if all of the eigenvalues of A lie within the unit circle in the complex plane, then, since all of the $q_i(k)$ are bounded, the difference between the output of the system with quantization and the output of the system without quantization is bounded. Thus, if a system is stable without quantization of variables, then it will be stable if some or all of its variables are quantized.

Equations (II-13) and (II-14) are not of much practical value for the determination of the precise quantization error at a particular time, N, because the values of $q_i(k)$,--for i=1,2,3,4,5, and k=0,1,..,N-1--are complicated functions of $\underline{x}(k)$, k=0,1,...N-1, and r(k), k=0,1,..., N-1. These equations are, however, of significant value in the computation of a bound on the difference between $\overline{y}(k)$ and y(k).

Bertram used the triangular inequality to determine a bound on $y_v(k)$ for each sampling instant, k. However, the resulting bound cannot be written in closed form and its determination requires a significant amount of matrix multiplication.

J. B. Slaughter [11] developed a set of equations similar to those given in (II-13) and (II-14) by a process which is essentially the same as the procedure described above. However, in place of the term $B\underline{q}(k)$ in (II-13), Slaughter used $\underline{u}(k)$, where the components of $\underline{u}(k)$ are

$$u_i(k) = \underline{s}_i^T \underline{q}(k), \qquad i = 1,2,3,4,5, \tag{II-15}$$

where the symbol $\underline{s}_i^T$ denotes the i^{th} row of the matrix B. Then, the

maximum value of each of the u_i is computed using (II-15) and the bound $h_i/2$ on the magnitude of each of the q_i, $i = 1,2,3,4,5$. Thus Slaughter's equations are of the form

$$\underline{v}(k+1) = A\underline{v}(k) + \underline{u}(k) \ , \tag{II-16}$$

and

$$y_v(k) = \underline{c}^T\underline{v}(k) \ .$$

If the z-transform is evaluated for equations (II-16), the following expressions result:

$$\underline{v}(z) = z^{-1}A\underline{v}(z) + z^{-1}\,U(z) \tag{II-17}$$

$$Y_v(z) = \underline{c}^T\underline{v}(z) \tag{II-18}$$

Let $\underline{u}_M$ denote a vector whose components are equal to the maximum magnitude of u_i, $i = 1,2,3,4,5$. Slaughter assumed that $\underline{u}_M$ is the forcing function for (II-15) for all k. Thus (II-17) becomes

$$\underline{v}(z) = z^{-1}A\underline{v}(z) + \underline{u}_M \frac{z^{-1}}{1-z^{-1}} \ . \tag{II-19}$$

Slaughter did not attempt to prove that the selection of $\underline{u}_M$ as a forcing function will result in a sequence $\hat{y}_v(k)$ which is an upper-bound for $y_v(k)$ for all $k = 0,1,\ldots$. Although it seems reasonable to assert that the components of the worst case excitation vector would assume their maximum magnitudes, it is by no means evident that these components would not change sign as a function of system state or of the discrete-time argument, k.

From (II-19), it is evident that

$$\underline{V}(z) = \frac{1}{z-1}\,[I-Az^{-1}]^{-1}\,\underline{u}_M \,. \qquad \text{(II-20)}$$

The substitution of (II-20) into (II-18) gives

$$Y_V(z) = \frac{\underline{c}^T[I-Az^{-1}]^{-1}\,\underline{u}_M}{z-1} \,. \qquad \text{(II-21)}$$

The limit toward which the sequence $y_V(k)$, $k = 0,1,\ldots$ tends may be ascertained by the application of the Final Value Theorem from the z-transform theory:

$$\lim_{k\to\infty} y_V(k) = \lim_{z\to 1} (z-1)\,Y_V(z) = \lim_{z\to 1} \left[\underline{c}^T[I-Az^{-1}]^{-1}\underline{u}_M\right] . \qquad \text{(II-22)}$$

Since all of the eigenvalues of the A matrix are less than unity in magnitude, the limit indicated in (II-22) is always finite.

Monroe [12] presented a technique for the computation of a steady state bound on quantization error which is very similar in structure to the method of Slaughter discussed above.

Recently, Johnson [13,14] applied the concept of "Uniform Boundedness of Solutions" to the problem of determining a bound for $y_V(k)$ that is valid for all k. The "Uniform Boundedness of Solutions" concept is based on the application of the Liapunov stability theory to systems characterized by difference equations. Johnson considered the case where there is only one quantizer input in (II-13). By the manipulation of norm inequalities, he established a bound on the $\|\underline{V}\|$ and consequently on $|y_V(k)|$. Further, if several quantization error inputs must be considered in (II-13), then a bound for $|y_V(k)|$ can be computed by application of the triangular rule for the summation

of the absolute values of scalars. Since Chapter IV is devoted to the development of extensions of this method, a detailed discussion of this method is deferred until that chapter.

Kiovuniemi [15] recognized that the problem of bounding the response of a system characterized by (II-13), and (II-14) can be considered as an optimal-control problem. He then applied the discrete-time optimal-control theory to the problem of determining an optimal sequence $[\hat{\underline{q}}(k),\ k = 0,1,\ldots,\ N-1]$ such that the quadratic functional

$$J(\hat{\underline{q}},N) = (1/2)\underline{v}^T(N)S\underline{v}(N) \tag{II-23}$$

has the property

$$J(\hat{\underline{q}},N) \geq J(\underline{q},N) \tag{II-24}$$

for all admissible sequences $[\underline{q}(k),\ k=0,1,\ldots,N-1]$. If S is chosen as the matrix $\underline{cc}^T$, then (II-23) will provide the least upper bound on $y_v(N)$. However, a digital computer search routine is usually required in order to determine the optimal quantization error sequence $[\underline{q}(k),\ k = 0,1,\ldots,N]$.

In a significant correspondence, Dejka [16] set forth a method for computing a bound on quantization error by use of transfer functions. In order to apply the method, a transfer function is required from each quantizer input error source to the output variable. As an example, consider the simple system shown below where $Q_1(z)$ and $Q_2(z)$ denote the z-transform of quantization error sequences $q_1(k)$ and $q_2(k)$ respectively.

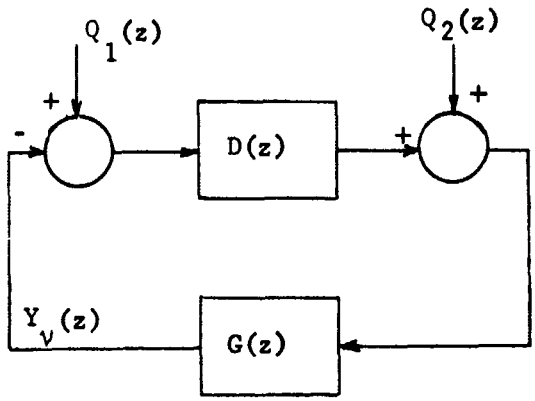

Figure 9. Diagram of a system used to illustrate Dejka's method.

From Figure II-6, it follows that

$$Y_v(z) = \frac{D(z)G(z)}{1+G(z)D(z)} Q_1(z) + \frac{G(z)}{1+G(z)D(z)} Q_2(z) \quad . \tag{II-26}$$

Let

$$G_1(z) = \frac{D(z)G(z)}{1+G(z)D(z)} = \sum_{n=0}^{\infty} g_1(nT)z^{-n} \tag{II-27}$$

and

$$G_2(z) = \frac{G(z)}{1+G(z)D(z)} = \sum_{n=0}^{\infty} g_2(nT)z^{-n} \quad . \tag{II-28}$$

It follows that

$$y_v(nT) = \sum_{k=0}^{n} q_1(kT)g_1[(n-k)T] + \sum_{k=0}^{n} q_2(kT)g_2[(n-k)T] \quad . \tag{II-29}$$

However, since

$$|q_1(kT)| \leq h_1/2, \text{ for all } k = 0,1,2,\ldots, \tag{II-30}$$

and

$$|q_2(kT)| \leq h_2/2, \text{ for all } k = 0,1,2,\ldots, \tag{II-31}$$

then

$$|y_v(nT)| \leq \frac{h_1}{2} \sum_{k=0}^{n} |g_1(kT)| + \frac{h_2}{2} \sum_{k=0}^{n} |g_2(kT)| \quad . \tag{II-32}$$

In his correspondence, Dejka fails to note several useful properties of the above procedure which are of significant practical interest. The above method is considered in depth in Chapter VI where it is shown that it may be to determine a least upper-bound on $|y_v(k)|$ for all $k = 0,1,2\ldots$.

III. MATHEMATICAL BACKGROUND

The purpose of this chapter is to present a brief summary of the mathematical concepts which underlie the developments presented in the following chapters.

A. Linear Spaces

Definition III-1: [17] A set E is called a "linear space" if addition and scalar multiplication are defined on E and the following rules apply:

a) $x + y = y + x$ for all $x,y \in E$.

b) $(x + y) + z = x + (y + z)$ for all $x,y,z \in E$.

c) There exists an element $0 \in E$ such that $x + 0 = x$ for all $x \in E$.

d) For any $x \in E$, there exists an element $y \in E$ such that $x + y = 0$.

e) $1x = x$ for all $x \in E$. (III-1)

f) $\alpha(\beta x) = (\alpha\beta)x$ for all $x \in E$ and any numbers α and β.

g) $(\alpha + \beta)x = \alpha x + \beta x$ for all $x \in E$ and any numbers α and β.

h) $\alpha(x + y) = \alpha x + \alpha y$ for any $x,y \in E$ and any number α.

Let E^n be a set in which the point $x \in E^n$ is defined to be a sequence of n real numbers, i. e.,

$$x = (x_1, x_2, \cdots, x_n) . \tag{III-2}$$

Then E^n is a real linear space if the following operations are defined:

a) $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$, (III-3)

b) $\alpha x = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n)$, (III-4)

for all $x, y \in E^n$ and any number α.

Definition III-2: The statement that "P(x,y) is an inner product" means that P is a mapping from E^n X E^n onto E^1 such that if $x \in E^n$, $y \in E^n$, $z \in E^n$ and k is a nonzero constant, then

a) $P(x,y) = P(y,x)$, (III-5)

b) $P(kx,y) = kP(x,y)$, (III-6)

c) $P(x,x) \neq 0$ if $x \neq (0,0,\ldots,0)$, and (III-7)

d) $P(x+y,z) = P(x,z) + P(y,z)$. (III-8)

A lemma of Schwarz may be proved as a consequence of the above properties.

Theorem III-1: [18] Suppose $x \in E^n$ and $y \in E^n$, then

$$\left[P(x,y)\right]^2 \leq P(x,x)\, P(y,y). \tag{III-9}$$

Theorem III-2: [19] If $f = \sum_{i=1}^{n} x_i y_i$, then f is an inner product for $x \in E^n$ and $y \in E^n$.

Denote

$$\langle x,y \rangle = \sum_{i=1}^{n} x_i y_i \ . \tag{III-10}$$

The concept of a norm can be introduced in terms of its properties in much the same manner as the idea of an inner product was introduced above. However, in this dissertation, the conventional Euclidean norm

is used exclusively as a measure of the distance between $x \epsilon E^n$ and $y \epsilon E^n$. The Euclidean norm is computed as follows:

$$\|x - y\| = \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{1/2} = \left[\langle x-y, x-y\rangle\right]^{1/2} \qquad \text{(III-11)}$$

Definition III-3: The statement that "f is a linear transformation from E^n into E^m" means that if $x \epsilon E^n$, $y \epsilon E^n$ and k is a constant different from zero, then

$$\text{a)} \quad f(x + y) = f(x) + f(y) \text{ and} \qquad \text{(III-12)}$$

$$\text{b)} \quad f(kx) = kf(x) \qquad \text{(III-13)}$$

In order to facilitate the utilization of the rules of matrix manipulation, each point $x \epsilon E^n$ will henceforth be written as the column vector

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \qquad \text{(III-14)}$$

If A is an m x n matrix, $\underline{x} \epsilon E^n$ and

$$f(\underline{x}) = A\underline{x}, \qquad \text{(III-15)}$$

then f is a linear transformation from E^n into E^m.

Theorem III-3: [20] If f is a linear transformation from E^n into E^m, there exists a number $\|f\|$ such that $\|f(\underline{x})\| \leq \|f\| \cdot \|\underline{x}\|$ for all $\underline{x} \epsilon E^n$. The number $\|f\|$ is positive if the Null transformation which maps each point of E^n onto the origin of E^m, is excluded from consideration.

In the case where the transformation is effected by a matrix, A, the number $\|f\|$ is written $\|A\|$ and is called "the norm of matrix A." The $\|A\|$ may be computed with the aid of the following theorem which is proved in Appendix B.

Theorem III-4: If A is a m x n matrix, then

$$\|A\| = [\text{Maximum Eigenvalue of } A^T A]^{1/2} \tag{III-16}$$

B. Some Properties of Quadratic Forms

Definition III-4: The statement that "the function f from E^n into E^1 is a quadratic form" means that the function is of the form

$$f(\underline{x}) = \underline{x}^T A \underline{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \tag{III-17}$$

where A is an n x n matrix and a_{ij} denotes the ij^{th} element of A.

Note from (III-17) that the coefficient of $x_i x_j$ is $(a_{ij} + a_{ji})$. Thus the quadratic form can always be written in terms of a symmetric matrix B by simply choosing the element $b_{ij} = b_{ji} = (a_{ij} + a_{ji})/2$ for i = 1, ..., n, and j = 1, ..., n.

Definition III-5: The quadratic form $f(\underline{x}) = \underline{x}^T A \underline{x}$ is said to be positive (negative) definite if $f(\underline{x})$ is positive (negative) for every $\underline{x} \epsilon E^n$ except $\underline{x} = \underline{0}$. Further, the quadratic form is said to be positive

(negative) semidefinite if it is non-negative (non-positive) for every $\underline{x} \epsilon E^n$ and there exist points $\underline{x} \neq \underline{0}$ for which $f(\underline{x})$ is zero.

Theorem III-6: [21] If $f(\underline{x}) = \underline{x}^T A \underline{x}$ is positive definite, then all of the eigenvalues of A are positive and real.

The statement that "A is positive (negative) definite" means that the function $f = \underline{x}^T A \underline{x}$ is positive (negative) definite.

Theorem III-7. [21] The matrix A is positive definite if and only if each of the members of the number sequence $[\Delta_1, \Delta_2, \ldots, \Delta_n]$ is positive, where

$$\Delta_1 = a_{11}, \quad \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{21}, & a_{22}, & a_{23} \\ a_{31}, & a_{32}, & a_{33} \end{vmatrix} \qquad \text{(III-18)}$$

$\ldots\ldots, \quad \Delta_n = |A|$ = the determinant of A.

Theorem III-8: [21] The matrix A is negative definite if and only if each of the members of the number sequence $[-\Delta_1, \Delta_2, -\Delta_3, \ldots, (-1)^n \Delta_n]$ is positive, where Δ_i, $i = 1,2,\ldots,n$ is defined by (III-18).

Let the symbol $\lambda_i(A)$ denote the i^{th} eigenvalue of the matrix A. The following theorem is of particular significance in the developments of Chapter IV.

Theorem III-9: [19] If A is a positive definite symmetric matrix, then

$$\lambda_{min}(A) \, \|\underline{x}\|^2 \leq \underline{x}^T A \underline{x} \leq \lambda_{max}(A) \, \| \underline{x} \|^2. \qquad \text{(III-19)}$$

C. Liapunov Stability Theory and Discrete-Time Systems

Consider the homogeneous first-order difference equation

$$\underline{x}(t_{k+1}) = \underline{F}(\underline{x}(t_k);\ t_k)\ , \tag{III-20}$$

where t_k is the independent discrete-time variable, $t_{k+1} > t_k$ for all integers k and $t_k \to \infty$ as $k \to \infty$.[22] Denote the solution to (III-20) for an initial state $\underline{x}^0$ and an initial time t_0 by

$$\underline{x}(t_k) = \underline{P}(t_k;\ \underline{x}^0,\ t_0) \tag{III-21}$$

for all $t_k \geq t_0$. It will be assumed that $\underline{P}$ is always continuous in all of its arguments and that $\underline{F}(\underline{0},\ t_k) = \underline{0}$ for all t_k.

Definition III-6: An equilibrium state is any state $\underline{x}^e$ with the property that $\underline{F}(\underline{x}^e,\ t_k) = \underline{x}^e$ for all $t_k \geq t_0$.

Definition III-7: The statement that "the equilibrium state $\underline{x} = \underline{0}$ is stable in the sense of Liapunov (L-Stable)" means that for any t_0 and any $\epsilon > 0$, there exists a number, $\delta(\epsilon, t_o) > 0$, such that if $\|\underline{x}^0\| < \delta(\epsilon, t_o)$, then $\|\underline{P}(t_k,\ \underline{x}^0,\ t_0)\| < \epsilon$ for all $t_k \geq t_0$.

The equilibrium state, $\underline{x} = \underline{0}$ is said to be "uniformly L-Stable" if the number δ in Definition III-7 does not depend on t_0.

Definition III-8: The statement that "the equilibrium state $\underline{x} = \underline{0}$ is asymptotically stable" means that it is L-Stable and that there exists a number $\eta(t_o) > 0$ such that the number sequence $[\|\underline{x}(t_k)\|,\ k = 0,1,\ldots]$ converges to $\underline{0}$ for all $\|\underline{x}^0\| < \eta(t_o)$.

Further, if η does not depend on t_0, the equilibrium state $\underline{x} = \underline{0}$ is "uniformly asymptotically stable." If the equilibrium state $\underline{x} = \underline{0}$ is asymptotically stable for any initial state $\underline{x}^0$, the equilibrium is said to be "asymptotically stable in the large."

In order to introduce Liapunov's theorems, the meaning of a "positive definite" function and a "decrescent" function must be given.

Definition III-9: The statement that "the scalar function, $V(\underline{x};t_k)$ is positive definite in the neighborhood, N_0, of the point $\underline{x} = \underline{0}$" means that $V(\underline{0}; t_k) = 0$ and that there exists a continuous, monotonically increasing function, ϕ, of the real variable, $\|\underline{x}\|$, such that

$$V(\underline{x}; t_k) \geq \phi(\|\underline{x}\|) \text{ for all } t_k \text{ and all } \underline{x} \epsilon N_0.$$

Definition III-10: The statement that "a positive definite function $V(\underline{x}; t_k)$ is decrescent in a neighborhood, N_0, of $\underline{x} = \underline{0}$" means that there exists a continuous, monotonically increasing function, Ψ, of the real variable, $\|\underline{x}\|$ such that

$$\Psi(0) = 0 \text{ and}$$

$$V(\underline{x};t_k) \leq \Psi(\|\underline{x}\|) \text{ for all } t_k \text{ and all } \underline{x} \epsilon N_0.$$

It is interesting to note from Definition III-5 and Theorem III-9 that a positive definite quadratic form is always a positive definite decrescent function in E^n.

Definition III-11: The statement that "$\Delta V(\underline{x}; t_k)$ is the first forward difference of $V(\underline{x}; t_k)$" means that

$$\Delta V(\underline{x};\ t_k) = \frac{V(\underline{x}(t_{k+1});\ t_{k+1}) - V(\underline{x}(t_k);\ t_k)}{t_{k+1} - t_k} \qquad \text{(III-22)}$$

Theorem III-10: [22] (Liapunov's Stability Theorem) The equilibrium state, $\underline{x} = \underline{0}$, is L-Stable if there exists a positive definite function $V(\underline{x};\ t_k)$ possessing a non-positive forward difference, $\Delta V(\underline{x};\ t_k)$.

Theorem III-11: [22] (Liapunov's Asymptotic Stability Theorem) The equilibrium state, $\underline{x} = \underline{0}$, is asymptotically stable if there exists a decrescent, positive definite, function $V(\underline{x};\ t_k)$ possessing a negative definite forward difference $\Delta V(\underline{x};\ t_k)$.

Theorem III-12: [22] The equilibrium state, $\underline{x} = \underline{0}$, is asymptotically stable in the large if it is asymptotically stable and if $V(\underline{x};\ t_k)$ has the property that $\phi(\|\underline{x}\|) \to \infty$ as $\|\underline{x}\| \to \infty$.

Theorem III-13: [22] The discrete-time system characterized by

$$\underline{x}(k + 1) = A\underline{x}(k)\ , \qquad \text{(III-23)}$$

where $\underline{x}$ is an n-component vector and A is an n x n matrix, is asymptotically stable if and only if all of the eigenvalues of A are of magnitude less than unity.

Finally, a theorem of considerable importance to the developments of Chapter IV may be stated for the discrete-time system described by (III-23).

Theorem III-14: [23] If all of the eigenvalues of A are of magnitude less than unity and if Q is an arbitrary, symmetric, positive definite, n x n matrix, then there exists a unique, symmetric, positive

definite, matrix, P, which is the solution of the matrix equation

$$A^T P A - P = -Q \ . \tag{III-24}$$

D. Topics in Optimization Theory

One of the most widely used techniques for determining the extremum of a function subject to a set of constraint equations is the Lagrange Multiplier Technique. Lagrange's method provides a necessary condition for an extremum and can be described as follows. Let $f(\underline{x})$ be an expression whose extreme values are sought when the variables are restricted by a certain number of side conditions, say $g_1(\underline{x}) = 0, \ \ldots, \ g_m(\underline{x}) = 0$. Define the function

$$F(\underline{x},\underline{\lambda}) = f(\underline{x}) + \lambda_1 g_1(\underline{x}) + \lambda_2 g_2(\underline{x}) + \ \ldots \ + \lambda_m g_m(\underline{x}), \tag{III-25}$$

where $\lambda_1, \ \lambda_2, \ \ldots, \ \lambda_m$ are m constants. The function F is next differentiated with respect to each coordinate, x_i, and set equal to zero. The resulting set of n equations are then adjoined to the set of m constraint equations forming a set of n + m equations in n + m unknowns, i. e.,

$$\begin{aligned} &\frac{\partial F(\underline{x},\underline{\lambda})}{\partial x_j} = 0, \ j = 1,2, \ \ldots, \ n \\ &g_i(\underline{x}) = 0, \ i = 1,2, \ \ldots, \ m. \end{aligned} \tag{III-26}$$

Lagrange showed that if the point, $\underline{x}$, is a solution to the extremum problem, then $\underline{x}$ must satisfy the n + m equations of (III-26). In order

to put the above method on a somewhat more rigorous footing, the theorem which follows may be proved.

Definition III-12: Let S be a set of points in E^n and assume $\underline{x} \epsilon S$. The statement that "$\underline{x}$ is an interior point of S" means that there exists a neighborhood of $\underline{x}$, $N(\underline{x}) \subset S$. The statement that "S is open" means that each of its points is an interior point.

Definition III-13: The statement that "$f \epsilon C'$" means that the components of f have continuous first-order partial derivatives.

Theorem III-15: [24] Let f be a real valued function such that $f \epsilon C'$ on an open set S in E^n. Let $g_1, \ldots, g_m$ represent m real valued functions such that $\underline{g} = (g_1, g_2, \ldots, g_m)^T \epsilon C'$ on S and assume $m < n$. Let X_0 be that subset of S on which $\underline{g}$ vanishes, i. e.,

$$X_o = [\underline{x};\ \underline{x} \epsilon S,\ \underline{g}(\underline{x}) = 0] \tag{III-27}$$

Assume that $\underline{x}_0 \epsilon X_0$ and assume that there exists a neighborhood, $N(\underline{x}_0)$ such that $f(\underline{x}) \geq f(\underline{x}_0)$ for all $\underline{x} \epsilon X_0 \cap N(\underline{x}_0)$ or such that $f(\underline{x}) \leq f(\underline{x}_0)$ for all $\underline{x} \epsilon X_0 \cap N(\underline{x}_0)$. Assume also that the rank of G, where $G = \left[\frac{\partial g_i}{\partial x_j}\right]$ is m. Then there exist m real numbers $\lambda_1, \lambda_2, \ldots, \lambda_m$, such that the following n equations are satisfied:

$$\frac{\partial}{\partial x_j} F(\underline{x}, \underline{\lambda}) = 0,\ j = 1, 2, \ldots, n, \tag{III-28}$$

where F is defined by (III-25).

Definition III-14: The statement that "$\underline{x}$ is on the straight line between $\underline{x}_1$ and $\underline{x}_2$" means that there exists a number $t \epsilon [0,1]$ such that

$$\underline{x} = t\underline{x}_1 + (1 - t)\underline{x}_2 \qquad \text{(III-29)}$$

Definition III-15: The statement that "set S is convex" means that if $\underline{x}_1 \epsilon S$ and $\underline{x}_2 \epsilon S$, then every point $\underline{x}$ on the straight line between $\underline{x}_1$ and $\underline{x}_2$ is a member of S.

Definition III-16: The statement that "$\underline{x}$ is a limit point of the set S" means that every neighborhood about $\underline{x}$ contains a point of S distinct from $\underline{x}$.

Definition III-17: The statement that "set S is closed" means that S contains all of its limit points.

The definitions III-14 through III-17 provide sufficient terminology for the definition of the absolute maximum of a function and for the statements which follow this definition.

Definition III-18: Let the function, $f(\underline{x})$, be defined over a closed set X in E^n. The statement that "$f(\underline{x})$ takes on its absolute maximum over X at $\underline{x}_o$" means that $f(\underline{x}) \leq f(\underline{x}_o)$ for all $\underline{x} \epsilon X$. The Definition III-18 can be altered in an obvious way to define the absolute minimum of f over the closed set S.

Definition III-19: The statement that "the function f is convex over a convex set X in E^n" means that for any two points $\underline{x}_1$ and $\underline{x}_2$ in X and for all $t \epsilon [0,1]$,

$$f[t\underline{x}_2 + (1 - t)\underline{x}_1] \leq tf(\underline{x}_2) + (1 - t)f(\underline{x}_1) \qquad \text{(III-30)}$$

Definition III-20: The statement that "the function $f(\underline{x})$ is strictly convex over a convex set X in E^n" means that for any two points $\underline{x}_1$ and $\underline{x}_2$ in X and for all $t \epsilon (0,1)$,

$$f[t\underline{x}_2 + (1 - t)\underline{x}_1] < tf(\underline{x}_2) + (1 - t)\, f(\underline{x}_1) \qquad \text{(III-31)}$$

Theorem III-16: [25] A linear function $f(\underline{x}) = \underline{c}^T\underline{x}$ is a convex function over all of E^n.

Theorem III-17: [25] A positive semidefinite quadratic form $f(\underline{x}) = \underline{x}^T A\underline{x}$ is a convex function over all of E^n.

Definition III-21: The statement that "$\underline{x}\epsilon X$ is an extreme point of the convex set X" means that there does not exist two distinct points $\underline{x}_1$ and $\underline{x}_2\epsilon X$ such that $\underline{x}$ lies on the straight line between $\underline{x}_1$ and $\underline{x}_2$.

An extreme point can be envisioned geometrically as a corner point of a convex set.

Theorem III-19: [25] Let X be a closed and bounded convex set. If the absolute maximum of the convex function $f(\underline{x})$ over X is finite, then the absolute maximum of $f(\underline{x})$ will be taken on at one or more extreme points of X.

IV. UNIFORM BOUNDEDNESS OF SOLUTIONS AND QUANTIZATION ERROR-BOUNDS

In this chapter, solutions are given for an upper-bound on the output of a stable, linear, time invariant, discrete-time system whose input signals are limited only in magnitude. Since the developments of this chapter rely heavily upon the idea of uniform boundedness as it relates to discrete-time systems, a section is devoted to a description of the properties of the motions of uniformly bounded systems. These basic concepts are then used to compute an error-bound for systems with one quantizer. It is established in the last section of the chapter that an upper-bound may be computed for multiple quantizer systems by application of the uniform boundedness concept.

A. Uniform Boundedness of Solutions

There are several excellent treatments in the western literature which describe the theory of "Uniformly Bounded Solutions" as it relates to Liapunov theory for systems governed by differential equations [26], [27], [28]. However, the definitions and theorems which have been established for the continuous-time systems have not, in general, been reformulated to consider the case of discrete-time systems. Therefore, those definitions and theorems which are pertinent to the developments which follow in this chapter will be stated and proved.

Consider the discrete-time system which is described by (III-20) and rewritten below:

$$\underline{x}(t_{k+1}) = \underline{F}(\underline{x}(t_k);\ t_k)\ , \qquad \text{(IV-1)}$$

where $\underline{F}$ is defined for all $t_o \leq t_k < \infty$ and for all $\underline{x}$. Let $\underline{x}(t_o) = \underline{x}_o$.

Definition (IV-1): The solutions of (IV-1) are said to be uniformly bounded if for any $\alpha > 0$ there exists a positive number, β, such that if $\| \underline{x}_o \| < \alpha$, then $\|\underline{x}(t_k, \underline{x}_o, t_o)\| < \beta$ for $t_k \geq t_o$, where β depends only on α and is independent of t_o.

Yoshizawa [28] also defines the properties of several stronger types of boundedness motions such as "equi-ultimately bounded" and "uniform-ultimately bounded." However, the developments of this chapter do not require the utilization of these boundedness characteristics and the associated theorems.

Let $V(t_k, \underline{x})$ be a decrescent and positive definite function in E^n with continuous first partial derivatives for all $\underline{x}$ and all $t_k \geq t_o$. Let L denote a bounded, open set in E^n and let M be the union of L with the boundary of L; i.e., $M=L \cup \delta L$, with the property that if $\underline{x}\epsilon M^c$ (the complement of M), then $\Delta V(\underline{x}, t_k) \leq 0$ for all $t_k \geq t_o$. Let M_r be a closed set with the property that $M \subseteq M_r$.

Theorem IV-1: If $\Delta V(\underline{x}(t_k), t_k) \leq 0$ for all $\underline{x}\epsilon M^c$ and if $V(\underline{x}_1(t_{k_1+1}), t_{k_1+1}) < V(\underline{x}_2(t_{k_2}), t_{k_2})$ for all $t_{k_2} > t_{k_1+1}$, $t_{k_1} \geq t_o$, all $\underline{x}_1(t_{k_1})$ in M and all $\underline{x}_2(t_{k_2})$ in M_r^c ; then each solution of (IV-1) which at some time $t_{k_o} \geq t_o$ is in M can never thereafter leave M_r.

Proof:

Assume that for some integer, k_o, $\underline{x}(t_{k_o})\epsilon M$. Further assume that there exists an integer $j > k_o$ such that $\underline{x}(t_j)\epsilon M_r^c$. Therefore, there exists a least integer, n, with the property that $\underline{x}(t_{n-1})\epsilon M$ and

$\underline{x}(t_k) \epsilon M^c$ for all $n \leq k \leq j$, $n \neq j$. Now, by the theorem hypothesis:

$$V(\underline{x}(t_j),\ t_j) > V(\underline{x}(t_n),\ t_n)\ . \qquad \text{(IV-2)}$$

However this cannot be since $\Delta V(\underline{x}(t_k), t_k) \leq 0$ for all $t_n \leq t_k \leq t_j$.

B. The Development of An Error Bound for Systems Containing a Single Quantizer

In this section, the problem of determining an upper bound on quantization error in a system with a single quantizer by application of the concept of uniform boundedness is considered. A set of first order difference equations can be written for a time-invariant, discrete-time system that is linear with the exception of one quantizer by application of the method of Bertram described in Chapter II. The equations which result are

$$\underline{\nu}(k + 1) = A\underline{\nu}(k) + \underline{b}q(k), \text{ and} \qquad \text{(IV-3)}$$

$$y_\nu(k) = \underline{c}^T\underline{\nu}(k), \qquad \text{(IV-4)}$$

where

A is an $(n \times n)$ matrix with the property that each of its eigenvalues is less than unity in magnitude,

$\underline{\nu}$ is an n-vector representing the difference between the state of the system with quantization and the state of the system without quantization,

$\underline{b}$ and $\underline{c}$ are constant n-vectors,

y_ν is the difference between the system output with quantization and the system output without quantization, and

q denotes the difference between the quantizer input variable and the quantizer output variable.

The magnitude of q must be less than or equal to one-half of the quantizer granularity, h, i.e.,

$$|q(k)| \leq \frac{h}{2} \text{ for all } k \geq k_o \; ; \; h > 0 \; . \tag{IV-5}$$

Suppose now that an arbitrary n x n, positive definite, symmetric matrix, Q, is chosen. Then from Theorem III-14, there exists a unique, symmetric, positive definite matrix, P, which is a solution of

$$A^T P A - P = -Q \; . \tag{IV-6}$$

Let

$$V[\underline{\nu}(k)] = \underline{\nu}^T(k) P \underline{\nu}(k) \; . \tag{IV-7}$$

Then the first forward difference of $V[\underline{\nu}(k)]$ is given by

$$\Delta V[\underline{\nu}(k)] = \frac{1}{T}\left\{-\underline{\nu}^T(k) Q \underline{\nu}(k) + 2q(k)\underline{b}^T P A \underline{\nu}(k) + \underline{b}^T P \underline{b} q^2(k)\right\} \; , \tag{IV-8}$$

where T can be set equal one without loss of generality. The problem of determining a bound on $|y_\nu(k)|$ that is valid for all k may now be considered in the context of the statements of Theorem IV-1. In order to use the uniform boundedness properties of this theorem, the set M^c must be defined such that if $\underline{\nu} \epsilon M^c$, $\Delta V[\underline{\nu}(k)] \leq 0$. Note that because of the second order term $-\underline{\nu}^T(k) Q \underline{\nu}(k)$ in (IV-8) and because of the boundedness of q(k), $\Delta V[\underline{\nu}(k)]$ must become negative as the $\|\underline{\nu}(k)\|$ increases along any ray from the origin. The forcing function, $\hat{q}(k)$, which maximizes $\Delta V(\underline{\nu}(k))$ may be determined by inspection of (IV-8) to be

$$\hat{q}(k) = \frac{h}{2}\,\mathrm{sgn}[\underline{b}^T P A \underline{v}(k)] \tag{IV-9}$$

where sgn is defined for this case by

$$\mathrm{sgn}[f] = +1 \text{ if } f \geq 0 \text{ and} \tag{IV-10}$$

$$\mathrm{sgn}[f] = -1 \text{ if } f < 0 .$$

Let $\Delta\hat{V}[\underline{v}(k)]$ represent the first forward difference of $V[\underline{v}(k)]$ when $\hat{q}(k)$ is applied and consider the set of points to which $\underline{v}(k)$ belongs only in case

$$\Delta\hat{V}(\underline{v}(k)) = 0. \tag{IV-11}$$

Suppose that $\underline{v}_1(k)$ satisfies (IV-11). Then $\Delta\hat{V}[t\underline{v}_1(k)] < 0$ for every $t > 1$ and for all $k \geq k_o$. This assertion may be easily verified as follows:

$$\underline{v}_1^T(k) Q \underline{v}_1(k) = h\underline{b}^T P A \underline{v}_1(k)\,\mathrm{sgn}[\underline{b}^T P A \underline{v}_1(k)] + \underline{b}^T P \underline{b}\,\frac{h^2}{4} . \tag{IV-12}$$

$$\Delta\hat{V}[t\underline{v}_1(k)] = -t^2 \underline{v}_1^T(k) Q \underline{v}_1(k) + th\underline{b}^T P A \underline{v}_1(k)\,\mathrm{sgn}[t\underline{b}^T P A \underline{v}_1(k)] + t^2 \underline{b}^T P \underline{b}\,\frac{h^2}{4} \tag{IV-13}$$

The substitution of (IV-12) into (IV-13) yields the desired result:

$$\Delta V[t\underline{v}_1(k)] = h\underline{b}^T P A \underline{v}_1(k)\,\mathrm{sgn}[\underline{b}^T P A \underline{v}_1(k)] \left\{t - t^2\right\} + \underline{b}^T P \underline{b}\,\frac{h^2}{4}\left\{1 - t^2\right\} < 0. \tag{IV-14}$$

Let the symbol $\delta\hat{M}$ denote the set of points to which $\underline{v}(k)$ belongs only in case $\Delta\hat{V}[\underline{v}(k)] = 0$. Then $\hat{M}$ is the set to which $\underline{v}(k)$ belongs only in case there exists a point $\underline{v}_1(k) \in \delta\hat{M}$ and a number $t \in [0,1]$ such that

$\underline{\nu}(k)=t\underline{\nu}_1(k)$. The set $\hat{M}$ is the least set with the property that if $\underline{\nu}(k)\epsilon\hat{M}^c$, then $\Delta V[\underline{\nu}(k)] < 0$ regardless of the selection $q(k)$ or of time, k.

By application of matrix and norm inequalities, it can be shown [14] that $\Delta V[\underline{\nu}(k)] < 0$ for all $\underline{\nu}(k)$ with the property that

$$\|\underline{\nu}(k)\| > \rho \quad , \tag{IV-15}$$

where

$$\rho = \left\{\frac{\|\underline{b}^T P A\| + \sqrt{\|\underline{b}^T P A\|^2 + \lambda_{min}(Q)|\underline{b}^T P \underline{b}|}}{\lambda_{min}(Q)}\right\} \frac{h}{2} \tag{IV-16}$$

Let $M_1 = \left\{\underline{\nu}(k); \|\underline{\nu}(k)\| \leq \rho\right\}$. Now, because of the approximations involved in the development of ρ, the set $\hat{M}$ described above is a subset of M_1. In order to describe $\hat{M}$ more precisely than the spherical approximation M_1, it is necessary to consider (IV-11), the equation of the boundary of $\hat{M}$, in more detail. If Q is chosen as the identity matrix, then the set $\hat{M}$ can be described in terms of simple geometrical concepts. Denote

$$\underline{\alpha}^T = [\alpha_1, \alpha_2, \ \ldots \ , \alpha_n] = [\underline{b}^T P A] \text{ and}$$

$$\beta = \underline{b}^T P \underline{b} \ .$$

If Q is chosen as the identity matrix and the above symbolism is used, (IV-11) may be written as

$$[\nu_1 - \alpha_1 \tfrac{h}{2} \operatorname{sgn}(\underline{\alpha}^T \underline{\nu}(k))]^2 + \ldots + [\nu_n - \alpha_n \tfrac{h}{2} \operatorname{sgn}(\underline{\alpha}^T \underline{\nu}(k))]^2 = \left(\frac{h}{2}\right)^2 \cdot \left\{\beta + \sum_{k=1}^{n} \alpha_k^2\right\} . \tag{IV-17}$$

The surface defined by (IV-17) is that generated by two hyperspheroids, S_1 and S_2, each of which intersects the hyperplane $H = \{\underline{\nu}(k);\ \underline{\alpha}^T\underline{\nu}(k) = 0\}$. S_1 is centered at $\frac{h}{2}\underline{\alpha}$, S_2 is centered at $-\frac{h}{2}\underline{\alpha}$ and each of these spheroids is of radius $\frac{h}{2}\left\{\beta + \sum_{k=1}^{n}\alpha_k^2\right\}^{1/2}$. A two dimensional geometric interpretation of (IV-17) is given by Figure IV-1.

Now, in order to utilize the boundedness properties of Theorem IV-1, it is necessary to develop a characterization of the set M_r. Positive definite quadratic forms exhibit several useful properties which can be employed in the generation of a suitable M_r.

Property IV-1: The set $U = \{\underline{\nu}(k);\ \underline{\nu}^T(k)P\underline{\nu}(k) \leq \xi\}$ where ξ is a positive constant is closed, bounded and convex.

Property IV-2: If $\xi_1 > \xi_2$, $\xi_2 > 0$ and $U_1 = \{\underline{\nu}(k);\underline{\nu}^T(k)P\underline{\nu}(k) \leq \xi_1\}$ and $U_2 = \{\underline{\nu}(k);\underline{\nu}^T(k)P\underline{\nu}(k) \leq \xi_2\}$, then $U_2 \subset U_1$.

Property IV-3: If $\underline{\nu}_1(k) \neq \underline{0}$ and $\underline{\nu}(k) = t\underline{\nu}_1(k)$ where $t\epsilon[0, \infty]$, then the function $V[t\underline{\nu}_1(k)] = t^2\underline{\nu}_1(k)P\underline{\nu}_1(k)$ is an increasing function of t.

Now, since $V[\underline{\nu}(k+1)] = V[\underline{\nu}(k)] + \Delta V[\underline{\nu}(k)]$, the set M_r described in Theorem IV-1 is simply $M_r = \{\underline{\nu}(k);\ V[\underline{\nu}(k)] \leq \xi_o\}$ where $\xi_o > 0$ is given by

$$\xi_o = \underset{\underline{\nu}(k)\epsilon\hat{M}}{\text{Max}} \left\{V[\underline{\nu}(k)] + \hat{\Delta V}[\underline{\nu}(k)]\right\} . \qquad \text{(IV-18)}$$

The argument of (IV-18) may be written

$$V[\underline{\nu}(k)] + \hat{\Delta V}[\underline{\nu}(k)] = \underline{\nu}^T(k)P\underline{\nu}(k) - \underline{\nu}^T(k)Q\underline{\nu}(k) + h\left|\underline{b}^TPA\underline{\nu}(k)\right| + \underline{b}^TP\underline{b}\,\frac{h^2}{4}$$

$$= \underline{\nu}^T(k)A^TPA\underline{\nu}(k) + h\left|\underline{b}^TPA\underline{\nu}(k)\right| + \underline{b}^TP\underline{b}\,\frac{h^2}{4} \qquad \text{(IV-19)}$$

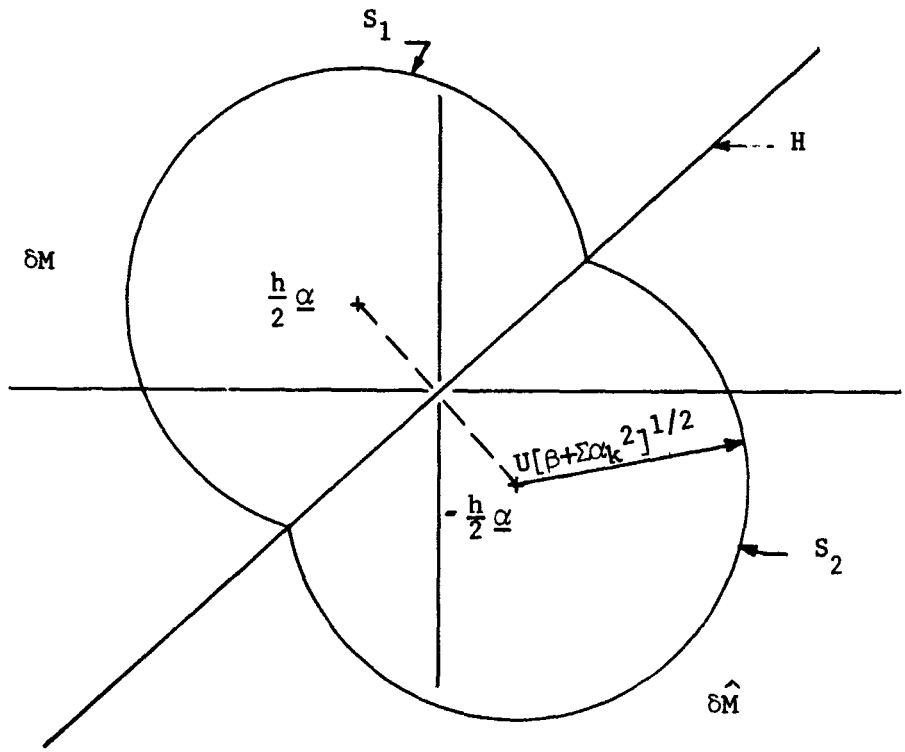

Figure 10.--A two dimensional representation of $\delta\hat{M}$

Inspection of (IV-19) shows that each term is non-negative, and non-decreasing along any ray from the origin of E^n. Consequently, it is a simple matter to show that if a $\underline{\nu}(k)$ satisfying (IV-18) exists, there is at least one $\underline{\nu}(k) \epsilon \delta M$ which satisfies (IV-18). The problem of defining M_r is equivalent to specifying the maximum value of V on the contour $\Delta\hat{V}(k) = 0$.

1. Methods of solution for the maximum V on $\delta\hat{M}$.

One can approach the problem of obtaining a solution for the maximum value of V on $\delta\hat{M}$ from either an analytical or a computational standpoint. In either event, it is quite probable that the gradient of $\Delta\hat{V}$ with respect to $\underline{\nu}$ will be required. It is apparent from (IV-8) using excitation (IV-9) that if $\underline{\nu} \epsilon H$, then the gradient of $\Delta\hat{V}$ is not defined. Thus, in order to apply most of the standard extremization techniques, it is necessary to provide a means of treating points which belong to $H \cap \hat{M}$. Fortunately, it turns out that the maximum value of V on $\delta\hat{M}$ cannot occur on $H \cap \delta\hat{M}$. This will next be shown using semi-geometric arguments.

It is not difficult to show that if $\underline{\nu}_1 \epsilon \delta\hat{M} \cap H$ and $\underline{\nu}_2 \epsilon \delta\hat{M} - \delta\hat{M} \cap H$, then

$$\|\underline{\nu}_2\| > \|\underline{\nu}_1\| \ . \qquad \text{(IV-20)}$$

(The argument, k, of $\underline{\nu}_1$ and $\underline{\nu}_2$ has been omitted for notational convenience). Assume that $V(\underline{\nu}_1) \geq V(\underline{\nu}_2)$. Now the $\underline{\nu}_2$ points in a neighborhood of $\underline{\nu}_1$ can be represented by $k\underline{\nu}_1 + \underline{\gamma}_1$ where $k > 1$ and $\langle \underline{\nu}_1, \underline{\gamma}_1 \rangle = 0$. Specifically, let $\underline{\nu}_2 = k\underline{\nu}_1 + \underline{\gamma}_1$ and $\underline{\nu}_2' = k\underline{\nu}_1 - \underline{\gamma}_1$. Consider now points $\underline{\gamma}$, on a line connecting $\underline{\nu}_2$ and $\underline{\nu}_2'$ given by

$$\underline{y} = \gamma\underline{\nu}_2 + (1-\gamma)\underline{\nu}_2' \qquad \text{where } 0 \leq \gamma \leq 1 \qquad \text{(IV-21)}$$

It is apparent that if $\gamma = 1/2$, $\underline{y} = k\underline{\nu}_1$. This point is evidently outside of the $V(\underline{\nu}_1)$ contour. Now, from the hypothesis, $V(\underline{\nu}_1) \geq V(\underline{\nu}_2)$, and $V(\underline{\nu}_1) \geq V(\underline{\nu}_2')$. Therefore it follows that the $V(\underline{\nu}_1)$ contour is not convex. However, this contradicts Property II. Thus it may be concluded that $V(\underline{\nu}) < V_{max}$ for all $\underline{\nu} \epsilon \delta M \cap H$, where V_{max} corresponds to the maximum value of V on $\delta\hat{M}$. (This conclusion is geometrically evident from Figure IV-1).

As a consequence of the above arguments, and because of the symmetry of the V and $\delta\hat{M}$ contours about H, the search for the maximum V on δM may be confined to either the positive input portion of E^n neglecting points on H or to the negative input portion of E^n.

One of the most practical methods for computing the maximum V on $\delta\hat{M}$ is by application of a computer search technique due to Fletcher and Powell [29]. The objective function which is to be minimized by this technique is $-V + \lambda(\Delta\hat{V})^2$, where λ is a suitable positive number chosen to force ΔV to approximately zero.

2. The development of an output-bound

After the set M_r has been defined, there are several approaches which may be taken to the problem of generating a bound for $y_\nu(k)$ that is valid for all $k \geq 0$. The first method is due to Johnson [14] and is completely analytic. Recall that $M_1 = \left\{\underline{\nu}(k);\ \|\underline{\nu}(k)\| \leq \rho\right\}$ where ρ is described by (IV-16). Thus from Theorem III-9, the maximum value of V on δM_1, is

$$V_{max}(M_1) = \rho^2 \lambda_{max}(P) \tag{IV-22}$$

Now according to Theorem IV-1, the system state must always remain in $M_{r1} = \left\{ \underline{v}(k);\ \underline{v}^T(k) P \underline{v}(k) \leq V_{max}(M_1) \right\}$. The boundary of M_{r1}, δM_{r1}, can itself be enclosed in a spherical ball. This can be easily shown by returning to Theorem IV-9.

$$\lambda_{min}(P)\ \|\underline{v}(k)\|^2 \leq V_{max}(M_1) \tag{IV-23}$$

for all $\underline{v}(k) \in M_{r1}$. If a spherical ball of radius

$$\rho' = \left\{ \frac{\lambda_{max}(P)}{\lambda_{min}(P)} \right\}^{1/2} \rho \tag{IV-24}$$

is constructed, then from (IV-23), M_{r1} will be contained within this spherical ball. Now, from Theorem IV-1, $\|\underline{v}(k)\| \leq \rho'$ for all $k \geq 0$. By application of (IV-4) and Theorem III-1, an output-bound for $y_v(k)$ can be computed as

$$|y_v(k)| \leq \|\underline{c}\| \cdot \|\underline{v}(k)\| \leq \|\underline{c}\|\ \rho' \text{ , for all } k=0,1,2,\ldots . \tag{IV-25}$$

Suppose that, instead of using the completely analytical procedure described above, the maximum value of V, V_{max}, on $\delta\hat{M}$ is determined by the application of a digital computer search technique. Then a somewhat less conservative bound for $\|\underline{v}(k)\|$ than that given by (IV-24) can be computed by essentially the same arguments given above, i.e.,

$$\|\underline{v}(k)\| \leq \left\{ \frac{V_{max}}{\lambda_{min}(P)} \right\}^{1/2} = r \ , \tag{IV-26}$$

and

$$|y_v(k)| \leq \|c\| \cdot \left\{ \frac{V_{max}}{\lambda_{min}(P)} \right\}^{1/2} = \|c\| \cdot r \tag{IV-27}$$

Still another method for obtaining a bound on $y_\nu(k)$ is to establish a bound on each of the components, ν_i, i=1,2,..., n of $\underline{\nu}\epsilon M_r$. This is equivalent to determining the minimum box which will contain $M_r = \left\{\underline{\nu}(k);\ \underline{\nu}^T(k)P\underline{\nu}(k) \leq V_{max}\right\}$. [30] Since,

$$\sum_{i=1}^{n} |c_i||\nu_i|_{max} \geq \|\underline{c}\| \cdot r \ , \qquad \text{(IV-28)}$$

the bound predicted by this method may be expected to be somewhat more conservative than the bound given by (IV-27).

None of the above methods will give the least upper-bound on $y_\nu(k)$ for $\underline{\nu}(k)\epsilon M_r$ for all systems. However, it turns out that a least upper-bound on $y_\nu(k)$ for $\underline{\nu}(k)\epsilon M_r$ can be obtained in closed form. Since M_r is convex and $y_\nu(k)$ is a linear combination of state variables, it may be argued that $y_{\nu max}$ occurs on δM_r. Consequently, it is desired to

$$\text{maximize} \quad y_\nu(k) = \underline{c}^T\underline{\nu}(k) \ , \qquad \text{(IV-29)}$$

$$\text{subject to} \quad \underline{\nu}^T(k)P\underline{\nu}(k) = V_{max}$$

This extremization problem will be solved by utilization of the Lagrange multiplier method outlined in Chapter III. Let

$$F = \underline{c}^T\underline{\nu}(k) + \lambda_1\left\{\underline{\nu}^T(k)P\underline{\nu}(k) - V_{max}\right\} \qquad \text{(IV-30)}$$

Then,

$$\text{grad}_\nu F = \underline{c} + 2\lambda_1 P\underline{\nu}(k) = 0 \qquad \text{(IV-31)}$$

Equations (IV-31) and the constraint equation of (IV-29) must next be

solved for $\underline{\nu}(k)$. First, λ_1 will be eliminated from these relations. From (IV-31)

$$\underline{\nu}^T(k) \text{ grad } F = \underline{\nu}^T(k)\underline{c} + 2\lambda_1\underline{\nu}^T(k)P\underline{\nu}(k) = 0 \qquad \text{(IV-32)}$$

However, from the constraint equation (IV-29), $\underline{\nu}^T(k)P\underline{\nu}(k) = V_{max}$. The substitution of this relation into (IV-32) and the solution for λ_1 yields

$$\lambda_1 = \frac{-\underline{\nu}^T(k)\underline{c}}{2V_{max}} \quad . \qquad \text{(IV-33)}$$

Consequently (IV-31) becomes

$$\underline{c} - \frac{(\underline{\nu}^T(k)\underline{c})}{V_{max}} P\underline{\nu}(k) = 0 \quad . \qquad \text{(IV-34)}$$

Equation (IV-34) may be rewritten as

$$\underline{\nu}(k) = \frac{V_{max}}{\underline{\nu}^T(k)\underline{c}} P^{-1}\underline{c} \quad . \qquad \text{(IV-35)}$$

The inner product of (IV-35) with $\underline{c}$ yields

$$\underline{c}^T\underline{\nu}(k) = \frac{V_{max}}{\underline{\nu}^T(k)\underline{c}} \underline{c}^T P^{-1}\underline{c} \qquad \text{(IV-36)}$$

It follows from (IV-36) that

$$|y_{\nu max}| = \sqrt{V_{max}\underline{c}^T P^{-1}\underline{c}} \qquad \text{(IV-37)}$$

4. An example of the procedures for bounding $y_\nu(k)$

Consider a system described by (IV-3) and (IV-4) with

$$A = \begin{bmatrix} 0 & , & 1.0 \\ -0.53, & & 1.135 \end{bmatrix} \quad , \qquad \underline{b} = \begin{bmatrix} 0. \\ 1. \end{bmatrix} \quad , \qquad \text{(IV-38)}$$

$$\underline{c} = \begin{bmatrix} 0.163 \\ -0.232 \end{bmatrix} \quad , \qquad d = 0, \text{ and } \underline{h}_2 = 1.0 \ .$$

Let Q = I. Then, the application of Theorem (III-14) leads to

$$P = \begin{bmatrix} 2.73 & -2.43 \\ -2.43 & 6.18 \end{bmatrix} \quad , \qquad \text{where} \qquad \text{(IV-39)}$$

$\lambda_1(P) = 7.44$, and $\lambda_2(P) = 1.48$.

The following relation, which results from (IV-24), may be used to bound the $\|\underline{v}(k)\|$.

$$\|\underline{v}(k)\| \leq \left[\frac{\lambda_{max}(P)}{\lambda_{min}(P)}\right]^{1/2} \rho \ , \qquad \text{for all } k \geq 0 \qquad \text{(IV-40)}$$

where ρ is given by (IV-16). For this example, inequality (IV-40) becomes

$$\|\underline{v}(k)\| \leq 26.6 \quad , \qquad \text{for all } k \geq 0 \ . \qquad \text{(IV-41)}$$

From (IV-25) the output is bounded by

$$|y_v(k)| \leq 7.55 \quad , \qquad \text{for all } k \geq 0 \ . \qquad \text{(IV-42)}$$

A somewhat tighter bound may be developed for $|y_v(k)|$ by application of the search technique described in [29] to the problem of defining the maximum V on $\delta\hat{M}$. The maximum V on $\delta\hat{M}$ is approximately

$$V_{max} = 1027 \qquad \text{(IV-43)}$$

The application of (IV-26) leads to

$$||\underline{y}(k)|| \leq \left\{ \frac{V_{max}}{\lambda_{min}(P)} \right\}^{1/2} = 26.3 \quad , \quad \text{for all } k \geq 0 \quad . \qquad \text{(IV-44)}$$

Then, from (IV-27), the output is bounded by

$$|y_v(k)| \leq 7.45 \quad , \qquad \text{for all } k \geq 0 \quad . \qquad \text{(IV-45)}$$

It is interesting to note that the application of (IV-26) in conjunction with the computed V_{max} gives almost no advantage, in this case, over the results obtained using strictly analytical techniques.

A third method by which a bound on $|y_v(k)|$ may be developed is by application of (IV-37). In this problem, the numerical results are

$$|y_v(k)| \leq 3.47 \quad , \qquad \text{for all } k \geq 0 \quad . \qquad \text{(IV-46)}$$

This third technique gives a substantially tighter bound for y for this example than the first two techniques provide. It has been noted that if the $\underline{c}$ vector is not approximately colinear with the eigenvector associated with the minimum eigenvalue of the P matrix, then the application of (IV-37) will provide a considerably tighter bound for $|y_v(k)|$, than can be obtained with the first two techniques discussed above.

C. The Development of an Output-Bound for a System Containing Several Quantizers

If several quantization nonlinearities are present in a control system, the computation of a bound on the system error becomes somewhat

more difficult. One method for obtaining an output-bound is to simply compute an output-bound for $|y_\nu^i(k)|$, i = 1, 2, ..., n, where the i superscript denotes the bound on error due to the $i^{\underline{th}}$ quantizer acting alone. Each of these n bounds can be determined by the methods of the preceding section. Since, by superposition,

$$y_\nu(k) = \sum_{i=1}^{n} y_\nu^i(k) , \qquad \text{(IV-47)}$$

then

$$|y_\nu(k)| \leq \sum_{i=1}^{n} |y_\nu^i(k)|_{max} \qquad \text{for all } k \geq 0 \qquad \text{(IV-48)}$$

An error-bound which is computed by (IV-48) is likely to be quite conservative since it is based on the triangular inequality for absolute values. In order to improve the quality of the estimate, a method is given in this section for computing a bound on multiple quantizer system errors based on the contribution of all quantizers acting simultaneously.

Consider the controllable system

$$\underline{\nu}(k + 1) = A\underline{\nu}(k) + B\,\underline{q}(k) , \qquad \text{(IV-49)}$$

and

$$y_\nu(k) = \underline{c}^T\underline{\nu}(k) , \qquad \text{(IV-50)}$$

where,

A is an n x n matrix whose eigenvalues are less than unity in magnitude,

B is an n x m matrix,

$\underline{c}$ is an n-vector,

$\underline{\nu}(k)$ is an n-component state vector, and

$\underline{q}(k)$ is an m-component excitation vector.

It will be assumed that each of the components of $\underline{q}(k)$, $q_i(k)$, $i = 1, 2, \ldots, n$, is bounded in magnitude by $\frac{h_i}{2}$. Let $\Omega = \left\{\underline{q}(k);\ |q_i| \leq \frac{h_i}{2},\ i = 1, 2, \ldots, n,\ k \geq 0\right\}$. Obviously the set Ω of admissible controls is closed, bounded and convex.

Just as was done in the preceding section, let P be a positive definite matrix which results from the solution of $A^TPA - P = -Q$ where Q is any positive definite symmetric matrix. Then, if $V[\underline{\nu}(k)] = \underline{\nu}^T(k)P\underline{\nu}(k)$, the first forward difference of $V[\underline{\nu}(k)]$ is

$$\Delta V[\underline{\nu}(k)] = -\underline{\nu}^T(k)Q\underline{\nu}(k) + 2\underline{q}^T(k)B^TPA\underline{\nu}(k) + \underline{q}^T(k)B^TPB\underline{q}(k) \quad . \qquad \text{(IV-51)}$$

Let

$$\hat{\Delta V}[\underline{\nu}(k)] = \max_{\underline{q}(k)\epsilon\Omega} \left\{\Delta V[\underline{\nu}(k)]\right\} \qquad \text{(IV-52)}$$

Because of the second order term in (IV-51), $\hat{\Delta V}[\underline{\nu}(k)]$ must eventually become negative as $\underline{\nu}(k)$ increases along any ray from $\underline{0}\epsilon E^n$. In fact, it is shown in Appendix A that $\hat{\Delta V}[\nu(k)]$ is equal to zero at no more than one point along each ray. This means that the surface described by

$$\hat{\Delta V}[\underline{\nu}(k)] = 0 \qquad \text{(IV-53)}$$

represents the boundary, $\delta\hat{M}$ of the set $\hat{M}$, where $\hat{M}$ is the set to which $\underline{\nu}(k)$ belongs only in case $\hat{\Delta V}[\underline{\nu}(k)] \geq 0$. Note that if $\underline{\nu}(k)\epsilon\hat{M}^c$, then $\Delta V[\underline{\nu}(k)] < 0$.

Let M_r be the set to which $\underline{\nu}(k)$ belongs only in case $\underline{\nu}^T(k)P\underline{\nu}(k) < \xi_o$ where

$$\xi_o = \max_{\underline{\nu}\epsilon\hat{M}} \left\{\hat{\Delta V}[\underline{\nu}(k)] + V[\underline{\nu}(k)]\right\} \quad . \qquad \text{(IV-54)}$$

It will be shown that (IV-54) is satisfied by at least one point, $\underline{v}(k) \epsilon \partial M$. Suppose that $\underline{q}_1(k)$ maximizes $\Delta V[\underline{v}_1(k)]$ for $\underline{v}_1(k) \epsilon iM$, and that $\underline{q}_2(k)$ maximizes $\Delta V[t\underline{v}_1(k)]$ for $t\underline{v}_1(k)$ where $t > 1$.

$$\Delta\hat{V}[t\underline{v}_1(k)] = -t^2\underline{v}_1^T(k)Q\underline{v}_1(k) + 2t\underline{q}_2^T(k)B^TPA\underline{v}_1(k) + \underline{q}_2^T(k)B^TPB\underline{q}_2(k) \quad . \tag{IV-55}$$

Since $\underline{q}_2(k)$ is assumed to be a maximizing input at $t\underline{v}_1(k)$,

$$2t\underline{q}_2^T(k)B^TPA\underline{v}_1(k) + \underline{q}_2^T(k)B^TPB\underline{q}_2(k) \geq 2t\underline{q}_1^T(k)B^TPA\underline{v}_1(k) + \underline{q}_1^T(k)B^TPB\underline{q}_1(k) \quad . \tag{IV-56}$$

Now since $\underline{q}_1(k)$ is a maximizing input at $\underline{v}_1(k)$, each of $2\underline{q}_1^T(k)B^TPA\underline{v}_1(k)$ and $\underline{q}_1^T(k)B^TPB\underline{q}_1(k)$ is nonnegative. Therefore (IV-56) can be written

$$2t\underline{q}_2^T(k)B^TPA\underline{v}_1(k) + \underline{q}_2^T(k)B^TPB\underline{q}_2(k) \geq 2\underline{q}_1^T(k)B^TPA\underline{v}_1(k) + \underline{q}_1^T(k)B^TPB\underline{q}_1(k) \quad . \tag{IV-57}$$

Further,

$$t^2\underline{v}_1^T(k)A^TPA\underline{v}_1(k) + 2t\underline{q}_2^T(k)B^TPA\underline{v}_1(k) + \underline{q}_2^T(k)B^TPB\underline{q}_2(k) \geq$$

$$\underline{v}_1^T(k)A^TPA\underline{v}_1(k) + 2t\underline{q}_1^T(k)B^TPA\underline{v}_1(k) + \underline{q}_1^T(k)B^TPB\underline{q}_1(k) \quad . \tag{IV-58}$$

This means that

$$V[t\underline{v}_1(k)] + \Delta V[t\underline{v}_1(k)] \geq V[\underline{v}_1(k)] + \Delta V[\underline{v}_1(k)] \tag{IV-59}$$

for all $t > 1$ and all $k \geq 0$. As a result of the above arguments, it may be asserted that

$$\max_{\underline{y}\in\hat{M}}\left\{\Delta\hat{V}[\underline{y}(k)] + V[\underline{y}(k)]\right\} = \max_{\underline{y}\in\delta\hat{M}}\left\{\Delta\hat{V}[\underline{y}(k)] + V[\underline{y}(k)]\right\} \qquad \text{(IV-60)}$$

Thus, just as in the single quantizer case, the problem of defining M_r is reduced to finding the maximum V on $\delta\hat{M}$.

1. The selection of the maximizing excitation

The problem of choosing that $\underline{q}(k)\in\Omega$ which maximizes $\Delta V[\underline{y}(k)]$ for a given $\underline{y}(k)\in E^n$ is not a simple one since $\Delta V[\underline{y}(k)]$ is a complex function of $\underline{y}(k)$ and $\underline{q}(k)$. The first term on the right hand side of (IV-51) is independent of $\underline{q}(k)$ and hence does not influence the selection of $\underline{q}(k)$. As a result, the problem of choosing the maximizing excitation function may be stated formally as

$$\text{maximize} \quad F = 2\underline{q}^T(k)B^TPA\underline{y}(k) + \underline{q}^T(k)B^TPB\underline{q}(k) \qquad \text{(IV-61)}$$

$$\text{subject to} \quad \underline{q}(k)\in\Omega \quad \text{for all } k \geq 0 .$$

Note that the function

$$F_1 = \underline{q}^T(k)B^TPB\underline{q}(k) = \underline{q}^T(k)\beta\underline{q}(k) \qquad \text{(IV-62)}$$

where $\beta = B^TPB$,

is a positive semidefinite form since P is a positive definite matrix. Therefore $F_1[\underline{q}(k)]$ is a convex function over Ω by Theorem III-17. Let

$$F_2 = 2\underline{q}^T(k)\ B^TPA\underline{y}(k) = \underline{q}^T(k)\underline{\alpha}(k) \text{ where} \qquad \text{(IV-63)}$$

$$\underline{\alpha}(k) = 2B^TPA\underline{y}(k) .$$

Then from Theorem III-16, F_2 is a convex function over Ω. Since F_1 and F_2 are convex functions over Ω, then

$$[\lambda \underline{q}_2(k) + (1-\lambda)\underline{q}_1(k)]^T \underline{\alpha}(k) = \lambda \underline{q}_2^T(k)\underline{\alpha}(k) + (1-\lambda)\underline{q}_1^T(k)\underline{\alpha}(k) \quad , \tag{IV-64}$$

and

$$[\lambda \underline{q}_2(k) + (1-\lambda)\underline{q}_1(k)]^T \beta [\lambda \underline{q}_2(k) + (1-\lambda)\underline{q}_1(k)] \quad \leq$$

$$\lambda \underline{q}_2^T(k)\beta \underline{q}_2(k) + (1-\lambda)\underline{q}_1^T(k)\beta \underline{q}_1(k) \tag{IV-65}$$

where

$$\lambda \epsilon [0,1] \ .$$

Now, if (IV-64) and (IV-65) are added, it is seen that

$$F[\lambda \underline{q}_2(k) + (1-\lambda)\underline{q}_1(k)] \leq \lambda F[\underline{q}_2(k)] + (1-\lambda)F[\underline{q}_1(k)] \ . \tag{IV-66}$$

Hence, F is a convex function over Ω. Now from Theorem III-19, it follows that the absolute maximum of F over Ω is taken on at one or more of the extreme points of Ω. Since Ω is a rectangular region in E^n, each of its corner points is an extreme point and there are no other extreme points in Ω. The significance of the above is that each component of the maximizing excitation at a point $\underline{v}(k) \epsilon E^n$ is always equal to its maximum magnitude. The problem of choosing the maximizing excitation for a given $\underline{v}(k)$ is reduced to testing 2^n candidate excitation vectors.

The problem of selecting functions $f_i[\underline{v}(k)]$, i=1,2,...,n such that the components of a maximizing excitation function are given by

$$u_i = \operatorname{sgn}[f_i[\underline{v}(k)]] \ , \ \begin{matrix} i = 1,2, \ \ldots, \ n \\ k \geq 0 \end{matrix} \tag{IV-67}$$

is quite formidable and no method of generating these functions is known. As a result, it is not possible to write a closed form expression for the surface $\delta\hat{M}$ described by $\Delta\hat{V}[\underline{v}(k)] = 0$. Therefore, the search

techniques which were described for finding the maximum V on δM for the single quantizer case are not as readily applicable for the multiple quantizer case since an expression for the gradient of $\Delta\hat{V}[\underline{v}(k)]$ with respect to $\underline{v}(k)$ cannot be written. Further, because of the possibility of 2^n maximizing excitations for $\Delta V[\underline{v}(k)]$, the problem of deciding in which region of E^n to search for the maximum V on δM is decidedly less well defined than for the single quantizer case.

3. An approximate definition for $\hat{M}$ and M_r

In this section, a set $R \supseteq \hat{M}$ will be defined such that if $\underline{v}(k) \epsilon R^c$, $\Delta V[\underline{v}(k)] < 0$. Let $\underline{q}_w$ denote an excitation vector which maximizes $\Delta V[\underline{v}(k)]$. Then,

$$\Delta\hat{V}[\underline{v}(k)] = \underline{v}^T(k)Q\underline{v}(k) + 2\underline{q}_w^T(k)B^TPA\underline{v}(k) + \underline{q}_w^T(k)\beta\underline{q}_w(k) = 0 \qquad \text{(IV-68)}$$

is the equation satisfied by $\underline{v}(k) \epsilon \delta\hat{M}$. From Theorem III-9, it is known that

$$\underline{v}^T(k)Q\underline{v}^T(k) \geq \|\underline{v}(k)\|^2 \lambda_{min}(Q) , \qquad \text{(IV-69)}$$

where $\lambda_{min}(Q)$ is the smallest eigenvalue of Q. Therefore, the set of points which satisfy the inequality

$$\lambda_{min}(Q)\|\underline{v}(k)\|^2 \geq 2\underline{q}_w^TB^TPA\underline{v}(k) + \underline{q}_w^T(k)\beta\underline{q}_w(k) \qquad \text{(IV-70)}$$

must be in the region in which $\Delta\hat{V}[\underline{v}(k)] \leq 0$. This line of reasoning may be extended by asserting that if

$$\lambda_{min}(Q)\|\underline{v}(k)\|^2 \geq 2\left|\underline{q}_w^TB^TPA\underline{v}(k)\right| + \underline{q}_w^T(k)\beta\underline{q}_w(k), \qquad \text{(IV-71)}$$

then

$\Delta\hat{V}[\underline{v}(k)] \leq 0.$

As a consequence of Theorem III-19, the term $\underline{q}_w^T(k)\beta\underline{q}_w(k)$ is bounded above and the bound may be determined by evaluating this term at each of the extreme points (corner points) of Ω. Let K_o denote the maximum value of $\underline{q}_w^T(k)\beta\underline{q}_w(k)$, for $\underline{q}_w(k)\epsilon\Omega$. K_o can also be computed by the formula $K_o = \lambda_{max}(\beta) \cdot \|\underline{q}_w(k)\|^2$ as shown in Appendix B. Consider next the term $2|\underline{q}_w^T(k)B^TPA\underline{v}(k)|$ and let

$$E = B^TPA \ . \tag{IV-72}$$

By application of the Schwarz inequality (Theorem III-1)

$$|\underline{q}_w^T(k)E\underline{v}(k)| \leq \|\underline{q}_w(k)\| \cdot \|E\underline{v}(k)\| \tag{IV-73}$$

However, from Theorem III-3, there exists a number, $\|E\|$, such that

$$\|E\underline{v}(k)\| \leq \|E\| \ \|\underline{v}(k)\| \ . \tag{IV-74}$$

A method for computing $\|E\|$ is developed in Appendix B.

As a result of the above arguments, all points $\underline{v}(k)\epsilon E^n$ which satisfy the inequality

$$\lambda_{min}(Q)\|\underline{v}(k)\|^2 \geq 2\|E\| \cdot \|\underline{q}_w(k)\| \cdot \|\underline{v}(k)\| + K_o \tag{IV-75}$$

also satisfy $\Delta\hat{V}[\underline{v}(k)] \leq 0$. Further, since each of the components of $\underline{q}_w(k)$ is always at its maximum magnitude,

$$\|\underline{q}_w(k)\| = \frac{1}{2}\sqrt{\sum_{i=1}^{n} h_i^2} \ . \qquad \text{for all } k \geq 0 \tag{IV-76}$$

It is therefore readily recognized that (IV-75) is a quadratic in

$\|\underline{y}(k)\|$. The largest value of $\|\underline{y}(k)\|$, ρ, which satisfies (IV-75) occurs when the equality is used. Consequently, it follows that the smallest number ρ with the property that if $\|\underline{y}(k)\| > \rho$, $\Delta\hat{V}[\underline{y}(k)] < 0$ is

$$\rho = \frac{\|E\| \cdot \|\underline{g}_w\| + \sqrt{\|E\|^2 \cdot \|\underline{g}_w\|^2 + \lambda_{min}(Q) \cdot K_o}}{\lambda_{min}(Q)} \quad . \tag{IV-77}$$

Note that if $K_o = \lambda_{max}(\beta)\|\underline{g}_w\|^2$ is used in (IV-77), the $\|\underline{g}_w\|$ appears as a factorable multiplier. The set $R = \left\{\underline{y}(k);\ \|\underline{y}(k)\| \leq \rho\right\}$ satisfies the properties of the set M given in Theorem IV-1. The maximum value of $V[\underline{y}(k)]$ on $\delta R = \left\{\underline{y}(k);\ \|\underline{y}(k)\| = \rho\right\}$ can be computed from Theorem III-9 to be

$$V_{max} = \lambda_{max}(P) \cdot \rho^2 \quad . \tag{IV-78}$$

Therefore the set $S = \left\{\underline{y}(k);\ \underline{y}^T(k)P\underline{y}(k) \leq V_{max}\right\}$ is analogous to the set M_r of Theorem IV-1. The maximum value of $y_\nu(k)$ for $\underline{y}(k) \in \delta S$ was established by equation (IV-37) in a preceding section of this chapter. The substitution of V_{max} from (IV-78) into (IV-37) gives

$$|y_{\nu max}| = \rho \sqrt{\lambda_{max}(P)\underline{c}^T P^{-1}\underline{c}} \tag{IV-79}$$

As was pointed out in (IV-77), ρ is proportional to $\|\underline{g}_w\|$, and as is evident from (IV-79), $|y_{\nu max}|$ is proportional to ρ; therefore the maximum magnitude of quantization error is proportional to $\|\underline{g}_w\|$.

3. An example

The system shown in Figure II-5, which was originally studied by Bertram [10], will be used in an effort to establish a bound on the error which arises in a digital control system due to quantization in

the digital device. The technique which has been developed for bounding $y_\nu(k)$ in the preceding sections of this chapter will now be applied to this system. As indicated by (II-9), (II-10) and (II-11), the system equations are,

$$\begin{bmatrix} \nu_1(k+1) \\ \nu_2(k+1) \\ \nu_3(k+1) \end{bmatrix} = \begin{bmatrix} \alpha_1 & -1 & \alpha_3^{-1} \\ \alpha_2 & -\alpha_3 e^{-1} & 1 \\ \alpha_3 & -\alpha_3 & 1 + e^{-1} \end{bmatrix} \begin{bmatrix} \nu_1(k) \\ \nu_2(k) \\ \nu_3(k) \end{bmatrix} +$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ \alpha_3 & \alpha_3 & \alpha_3 & 1 & 1 \\ \alpha_3 & \alpha_3 & \alpha_3 & 1 & 1 \end{bmatrix} \begin{bmatrix} q_1(k) \\ q_2(k) \\ q_3(k) \\ q_4(k) \\ q_5(k) \end{bmatrix}, \qquad \text{(IV-80)}$$

$$y_\nu(k) = [0,\ 1,\ -\alpha_3^{-1}]\ \underline{\nu}(k), \qquad \text{(IV-81)}$$

$$\alpha_3 = [1-e^{-1}]^{-1},\ \alpha_1 = [2e^{-1}-1]\alpha_3 \qquad \text{(IV-82)}$$

$$\alpha_2 = [e^{-1} - \alpha_3^{-1}]\ \alpha_3^2 \ . \qquad \text{(IV-83)}$$

Further, it will be assumed that the system employs uniform quantization, i.e.,

$$|h_i| \leq 1/2,\ i = 1, 2, \ldots, 5. \qquad \text{(IV-84)}$$

If it is assumed that Q = I, the solution of (IV-6) for P yields

$$P = \begin{bmatrix} 3.144 & 1.108 & -1.763 \\ 1.108 & 3.472 & -2.353 \\ -1.763 & -2.353 & 2.706 \end{bmatrix}$$

The eigenvalues of this matrix are approximately:

$\lambda = 0.49$, $\lambda = 6.62$ and $\lambda = 2.21$.

It was shown in Appendix B that the number $\|E\|$ is equal to the square root of the maximum eigenvalue of the matrix E^TE. The result of performing the indicated matrix multiplications and evaluating the eigenvalues of the resultant matrix is:

$$\|E\| = \left[\lambda_{max}(E^TE)\right]^{1/2} = 8.04 \ . \qquad \text{(IV-86)}$$

The number K_o was determined by evaluating the function $\underline{q}_w^T(k)B^TPB\underline{q}_w(k)$ for each extreme point of Ω. It was found that

$$K_o = 17.2 \ . \qquad \text{(IV-87)}$$

The number ρ can be calculated by utilization of the defining relation (IV-77).

$$\rho = \frac{5}{4}(8.04) + \frac{5}{4}(64.65) + 17.2 = 18.85 \qquad \text{(IV-88)}$$

The maximum value of $V[\underline{v}(k)]$ on δR is given by

$$V_{max} = \lambda_{max}(P)\ \rho^2 = (18.85)^2\ (6.62) \ . \qquad \text{(IV-89)}$$

The components of the $\underline{c}$ vector are

$$\underline{c}^T = [0,\ 1.0,\ -.632] \ . \qquad \text{(IV-90)}$$

Now from (IV-79), the output bound is given by

$$|y_{\nu max}| = V_{max}\, \underline{c}^T P^{-1} \underline{c} = 34.6 \qquad \text{(IV-91)}$$

Thus the output will always be less than 34.6 in magnitude provided that $\|\underline{\nu}(0)\| \leq 18.85$.

V. A LEAST UPPER-BOUND ON QUANTIZATION ERROR IN DIGITAL CONTROL SYSTEMS VIA THE DISCRETE MAXIMUM PRINCIPLE

The work presented in this chapter is based on the analysis of the quantization state-error equations (II-13) and (II-14) that were developed in Chapter II. These equations are repeated here for convenience:

$$\underline{\nu}(k+1) = A\underline{\nu}(k) + B\underline{q}(k), \tag{II-13}$$

$$y_{\nu}(k) = \underline{c}^{T}\underline{\nu}(k), \tag{II-14}$$

and

$$q_i(k) \leq \frac{h_i}{2} \quad \text{for all } k = 0,1,2,\ldots; \text{ and all } i = 1,2,\ldots,m. \tag{V-1}$$

Matrices A and B together with vectors $\underline{q}(k)$, $y_{\nu}(k)$ and $\underline{c}^{T}$ have the same properties as described in Chapter IV following (IV-50).

In this chapter, a method will be given for the computation of a least upper-bound for $y_{\nu}(k)$ that is valid for all integers $k \geq 0$. Let Ω be the set to which $\underline{q}(k)$ belongs only in case each of the components of $\underline{q}(k)$ satisfy (V-1). The problem can now be stated mathematically as: Determine a number U such that if $S = [x;\ x > y_{\nu}(k)]$ for all $k=0,1,\ldots,N$; all sequences $[\underline{q}(k)] \epsilon \Omega, k=0,1,2,\ldots N-1$, then $U \epsilon S$ and $U \leq x$ for all $x \epsilon S$. A method for computing U will be derived by application of the Discrete Maximum Principle which is discussed in the following section.

A. The Discrete Maximum Principle

The following statement of the Discrete Maximum Principle is a modification of a somewhat more general formulation given by Halkin in reference [31].

The following conditions must be fulfilled in order to ensure the applicability of the maximum principle.

[1] A^{-1} must exist

[2] The set $\{A\underline{v} + B\underline{q} \; ; \; \underline{q} \in \Omega\}$ must be convex for every $\underline{v}$ which is a member of E^n.

Comment: Condition 2 is always satisfied since Ω is a closed, bounded and convex set in E^m and these properties are preserved under linear transformation, B, into E^n.

It will be assumed that the system initial state is known and thus the initial transversality conditions will not be involved in the development which follows. Further, the terminal manifold for $\underline{v}$ is completely unspecified. In addition, a "sufficiently large" integer, N, will be selected at the outset to specify the terminal time of the process, NT.

Two sequences, $\hat{\underline{q}}(0)$, $\hat{\underline{q}}(1)$, ..., $\hat{\underline{q}}(N-1)$ and $\hat{\underline{v}}(0)$, $\hat{\underline{v}}(1)$, ..., $\hat{\underline{v}}(N)$ are said to be optimal if they satisfy recurrence relation (II-13) if $\underline{q}(k)\epsilon\Omega$ $k = 0,1,\ldots,N-1$ and if $y_\nu(\hat{\underline{v}}(N))$ is the maximum value of $y_\nu(\underline{v}(N))$ subject to these constraints.

The maximum principle may now be stated for the problem under consideration.

If the sequences $\hat{\underline{q}}(0)$, $\hat{\underline{q}}(1)$, ..., $\hat{\underline{q}}(N-1)$ and $\hat{\underline{v}}(0)$, $\hat{\underline{v}}(1)$, ..., $\hat{\underline{v}}(N)$ are optimal, then there exists a sequence of non-zero vectors $\hat{\underline{p}}(0)$, $\hat{\underline{p}}(1)$, ..., $\hat{\underline{p}}(N)$ such that

[1] $$<A\hat{\underline{v}}(k) + B\hat{\underline{q}}(k), \hat{\underline{p}}(k+1)> \geq <A\hat{\underline{v}}(k) + B\underline{q}(k), \hat{\underline{p}}(k+1)>$$

$$\underline{q}(k) \in \Omega, \text{ and } k = 0,1,2,\ldots,N-1, \tag{V-2}$$

[2] $$\hat{\underline{p}}(k) = A^T \hat{\underline{p}}(k+1), \tag{V-3}$$

[3] $$\hat{\underline{p}}(N) = \beta_o \operatorname{grad}_{\underline{v}} y_{\underline{v}}(k)\Big|_{\underline{v} = \hat{\underline{v}}(N)}, \text{ and} \tag{V-4}$$

$$\beta_o \geq 0. \tag{V-5}$$

B. Computation of The Error Bound

The application of equation (V-4) yields

$$\hat{\underline{p}}(N) = \beta_o \underline{c}. \tag{V-6}$$

Thus the value of the adjoint variable, $\underline{p}(k)$, $k \leq N$, may be shown from equation (V-3) to be

$$\hat{\underline{p}}(k) = [A^T]^{N-k} \hat{\underline{p}}(N) = \beta_o [A^T]^{N-k} \underline{c} \tag{V-7}$$

Now from (V-2), it is apparent that the terms $< A\hat{\underline{v}}(k), \hat{\underline{p}}(k+1) >$ may be neglected in maximizing the Hamiltonian with respect to $\underline{q}(k)$. By application of a well-known property of inner products that

$$<B\underline{q}(k), \hat{\underline{p}}(k+1)> = <\underline{q}(k), B^T\hat{\underline{p}}(k+1)>, \tag{V-8}$$

and letting the vector $\underline{b}_i$, i=1,2,...,m, denote the i^{th} column of B, it is obvious that the Hamiltonian, (V-2), is maximized if

$$q_i(k) = \frac{h_i}{2} \operatorname{sgn} < \underline{b}_i, \underline{p}(k+1) > , \qquad \text{(V-9)}$$

$$i=1,2,\ldots,m \text{ and}$$

$$k=0,1,\ldots,N-1 .$$

Substituting (V-7) into (V-9),

$$\hat{q}_i(k) = \frac{h_i}{2} \operatorname{sgn} \left\{ \beta_o < \underline{b}_i, [A^T]^{N-(k+1)} \underline{c} > \right\} \qquad \text{(V-10)}$$

$$i=1,2,\ldots,m, \text{ and}$$

$$k=0,\ldots,N-1 .$$

Several observations can be advanced from (V-10). Note that the magnitude of β_o is not significant in (V-10) and it will be subsequently assumed that β_o=1. It is possible that some of the inner products indicated in (V-10) will equal zero. In this case, the corresponding component of $\hat{\underline{q}}(k)$ has no bearing on the optimal solution. It is shown in Appendix C that the same solution for the components, $\hat{q}_i(k)$, can be obtained without introducing the Discrete Maximum Principle. In fact, as a result of this proof, if A is singular, the selection of $\hat{q}_i(k)$ given by (V-10) is also correct. For notational convenience let,

$$\hat{\underline{q}}(k) = \begin{bmatrix} \frac{h_1}{2} \operatorname{sgn} \langle \underline{b}_1, [A^T]^{N-(k+1)} \underline{c} \rangle \\ \vdots \\ \frac{h_m}{2} \operatorname{sgn} \langle \underline{b}_m, [A^T]^{N-(k+1)} \underline{c} \rangle \end{bmatrix} = \frac{1}{2} \underline{h} \operatorname{SGN} \left\{ B^T [A^T]^{N-(k+1)} \underline{c} \right\} \tag{V-11}$$

The substitution of (V-11) into (III-13) yields,

$$\hat{\underline{v}}(k+1) = A\hat{\underline{v}}(k) + \frac{1}{2} B \underline{h} \operatorname{SGN} \left\{ B^T [A^T]^{N-(k+1)} \underline{c} \right\} \tag{V-12}$$

for an N-stage process. Equation (V-12) may be solved explicitly for $\hat{\underline{v}}(N)$.

$$\hat{\underline{v}}(N) = A^N \underline{v}(0) + \frac{1}{2} \sum_{k=0}^{N-1} A^{N-(k+1)} B \underline{h} \operatorname{SGN} \left\{ B^T [A^T]^{N-(k+1)} \underline{c} \right\} \tag{V-13}$$

Thus from (III-14), the bound on system output is

$$y_v(\hat{\underline{v}}(N)) = \underline{c}^T \hat{\underline{v}}(N). \tag{V-14}$$

A useful property of the bound, $y_v(\hat{\underline{v}}(N))$ becomes evident upon inspection of (V-13) and (V-14) with $\underline{v}(0)$ set equal to zero. For this case, $y_v(\hat{\underline{v}}(N))$ is a non-decreasing function of N which converges to a number, Y_{max}, with increasing N. This convergence is a consequence of the initial assumption that the system is stable.

C. The Influence of Initial Conditions

It is quite often the case in control theory applications that the control system is comprised of a digital controller for a continuous

time process. Thus those state variables which appear within the digital device can assume only discrete levels while those which represent continuous time variables can take on a continuum of values. Physically this means that there may initially exist a "misalignment error" in the digital portion of the system. The magnitude of this initial error is dependent on the quantization granularity in the system. For convenience, let the first $t \leq n$ state variables of (II-2) be the digital variables and let the granularity for these state variables be given by

$$|\nu_i(0)| = |x_i(0) - x_i^q(0)| \leq \frac{m_i}{2}, \qquad i = 1,\ldots,t, \quad \text{(V-15)}$$

where $x_i(0)$ denotes the actual value of the i^{th} state in the digital device. Since by assumption, the remaining n-t variables are representative of the continuous portion of the system it is reasonable to assert that

$$\nu_i(0) = x_i(0) - x_i^q(0) = 0, \qquad i = t+1,\ldots,n. \quad \text{(V-16)}$$

It is desirable to determine the maximum error which can arise due to quantization when this initial error is included. Note from (V-13) that the optimal quantization input sequence is independent of $\underline{\nu}(0)$. Therefore, a worst-case initial error may be defined independently of the quantization sequence. The contribution of the $\underline{\nu}(0)$ term to $y_\nu(N)$ is

$$\begin{aligned} y_\nu'(N) &= \langle \underline{c}\ ,\ A^N\underline{\nu}(0) \rangle \\ &= \langle \underline{\nu}(0),\ [A^N]^T\underline{c} \rangle . \end{aligned} \quad \text{(V-17)}$$

Let $\underline{a}_i^N$ denote the i^{th} column of A^N. Then, the worst case initial error is

$$\nu_i(0) = \frac{m_i}{2} \operatorname{sgn} \langle \underline{a}_i^N, \underline{c} \rangle , \qquad i = 1,2,\ldots,t. \tag{V-18}$$

It should be pointed out that when (V-18) is used in (V-13) and (V-14), the resulting sequence $y_\nu(N)$ is not necessarily non-decreasing with increasing N.

D. Bertram's Example

The system shown in Figure (II-5), which was originally studied by Bertram [10], has served as a standard example for the application of various techniques which have been advanced for the establishment of a bound on quantization error. As indicated by (IV-80), the recurrence relation for this system is

$$\begin{bmatrix} \nu_1(k+1) \\ \nu_2(k+1) \\ \nu_3(k+1) \end{bmatrix} = \begin{bmatrix} \alpha_1 & -1 & \alpha_3^{-1} \\ \alpha_2 & -\alpha_3 e^{-1} & 1 \\ \alpha_2 & -\alpha_3 & 1+e^{-1} \end{bmatrix} \begin{bmatrix} \nu_1(k) \\ \nu_2(k) \\ \nu_3(k) \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ \alpha_3 & \alpha_3 & \alpha_3 & 1 & 1 \\ \alpha_3 & \alpha_3 & \alpha_3 & 1 & 1 \end{bmatrix} \begin{bmatrix} q_1(k) \\ q_2(k) \\ q_3(k) \\ q_4(k) \\ q_5(k) \end{bmatrix} \tag{V-19}$$

and the output is given by,

$$y_\nu(k)=[0,\ 1,\ -\alpha_3^{-1}]\ \underline{\nu}(k), \tag{V-20}$$

where $k = 0,1,\ldots,N-1$; $\alpha_3 = [1-e^{-1}]^{-1}$, $\alpha_1 = [2e^{-1}-1]\ \alpha_3$, and $\alpha_2 = [e^{-1} - \alpha_3^{-1}]\ \alpha_3^2$. Further, assume that uniform quantization is used, i.e.,

$$h_i = 1, \qquad i = 1,\ 2,\ 3,\ 4,\ 5\ . \tag{V-21}$$

The application of (V-13) and (V-14) with $\underline{\nu}(0)$ set to $\underline{0}$ yields

TABLE 1

OUTPUT BOUNDS FOR BERTRAM'S EXAMPLE WITH ZERO INITIAL CONDITIONS

N	4	6	8	10
$y_\nu(\hat{\underline{\nu}}(N))$	2.8507907	2.9089243	2.9167918	2.918565

It may be concluded from the data in Table V-1 that the system under consideration rapidly approaches its output bound with increasing N. The above results provide a much stronger bound than the bound of y_ν=7.9 which was obtained by Bertram in his original paper [10]. Koivuniemi [15] established that the maximum value of $\frac{1}{2}\ \|\hat{\underline{\nu}}(N)\|^2$ for this system in 38.078. The utilization of this result in conjunction with the norm inequality

$$\hat{y}_\nu \leq \|\ \hat{\underline{\nu}}(N)\ \| \cdot \|\ \underline{c}\ \|\ , \tag{V-22}$$

yields y_ν = 10.2 which is also a relatively weak bound. It is interesting to note that if Koivuniemi had used $[\underline{c}^T\underline{\nu}(N)]^2$ as a performance index, he

would have obtained results identical to those tabulated above. However, in order to obtain these results by Koivuniemi's methods, some sort of search method would be employed, whereas the proposed method provides a closed form solution for the output bound.

Next, the effect of worst case initial conditions will be considered. In Bertram's original formulation of this problem, x_1 represented the state of the discrete compensator while x_2 and x_3 reflect the dynamics of the continuous time segment of the system; thus $v_2(0) = 0$, $v_3(0) = 0$ and

$$|v_1(0)| \leq 1/2. \tag{V-23}$$

The application of (V-13), (V-14) and (V-18) yields the following results:

TABLE 2

OUTPUT BOUNDS FOR BERTRAM'S EXAMPLE USING WORST CASE INITIAL CONDITIONS

N	4	8	12	14	16
$y_v(\hat{\underline{v}}(N))$	2.8844069	2.9174075	2.9180119	2.9180216	2.9180176

The worst case error occurs for N=14. As expected, the output bound approaches the bound obtained in the previous case with $\underline{v}(0) = \underline{0}$ as N increases above 16.

The practical problem of initially choosing the value of N is worthy of discussion. Note from (V-13) that as N is increased, all of the original terms are retained and the additional terms which are summed provide a non-negative contribution to the output. Because of this property and the amenability of (V-13) to digital computer implementation, it has been found that the simplest solution to the problem of choosing N is to

compute the output bound for each integer from N=0 upward until changes in $y_\nu(\hat{\underline{\nu}}(N))$ are acceptably small for the system under study.

E. The Synthesis Problem

The problem of determining acceptable quantizer granularity for a system whose maximum allowable error due to quantization has been specified will now be discussed. It will be assumed that the system configuration is fixed and the only control which the designer can exercise over quantization error is by the adjustment of quantization granularity. Equation (V-13) may be rewritten in a more convenient form by utilization of equation (V-10), with $\underline{\nu}(0) = \underline{0}$,

$$\hat{\underline{\nu}}(N) = \frac{h_1}{2} \sum_{\ell=0}^{N-1} A^\ell \underline{b}_1 \operatorname{sgn} \langle \underline{b}_1, [A^T]^\ell \underline{c} \rangle + \ldots + \frac{h_m}{2} \sum_{\ell=0}^{N-1} A^\ell \underline{b}_m \operatorname{sgn} \langle \underline{b}_m, [A^T]^\ell \underline{c} \rangle , \tag{V-24}$$

and $y_\nu(\underline{\nu}(N))$ is

$$y_\nu(\underline{\nu}(N)) = \frac{h_1}{2} \sum_{\ell=0}^{N-1} \underline{c}^T A^\ell \underline{b}_1 \operatorname{sgn} \langle \underline{b}_1, [A^T]^\ell \underline{c} \rangle + \ldots + \frac{h_m}{2} \sum_{\ell=0}^{N-1} \underline{c}^T A^\ell \underline{b}_m \operatorname{sgn} \langle \underline{b}_m, [A^T]^\ell \underline{c} \rangle . \tag{V-25}$$

It can be easily seen that each of the summation terms shown in (V-25) is actually

$$\frac{h_i}{2} \sum_{\ell=0}^{N-1} \underline{c}^T A^\ell \underline{b}_i \operatorname{sgn} \langle \underline{b}_i, [A^T]^\ell \underline{c} \rangle = \frac{h_i}{2} \sum_{\ell=0}^{N-1} \left| \underline{c}^T A^\ell \underline{b}_i \right| , \qquad \text{(V-26)}$$

$$i = 1,2,\ldots m.$$

Since the system is assumed to be stable, each of these summations must converge with increasing N to a non-negative number, M_i, i.e.,

$$\lim_{N \to \infty} \sum_{\ell=0}^{N-1} \left| \underline{c}^T A^\ell \underline{b}_i \right| = M_i, \qquad i = 1,2,\ldots,m \ . \qquad \text{(V-27)}$$

Therefore the least upper bound for y_ν is given by

$$\hat{y}_\nu \ (\max) = \frac{1}{2} \sum_{i=1}^{m} M_i h_i \ . \qquad \text{(V-28)}$$

Equation (V-28) can be used to check a proposed quantizer granularity selection. Equation (V-26) is a particularly useful form in that it clearly demonstrates the conditions under which a quantization input sequence, $q_i(k)$, will have no effect on the system quantization error. Assuming that the eigenvalues of the system A matrix are distinct, then if $\underline{c}$ is an eigenvector of A^T and $\langle \underline{b}_i, \underline{c} \rangle = 0$, the i^{th} quantization input will not contribute to y_ν.

F. An Application

In order to demonstrate the applicability of the results obtained above, the problem of determining an upper-bound for quantization errors generated by a digital attitude-controller for Saturn V class space vehicles will be considered. A block diagram for the attitude control system is shown by Figure 11.

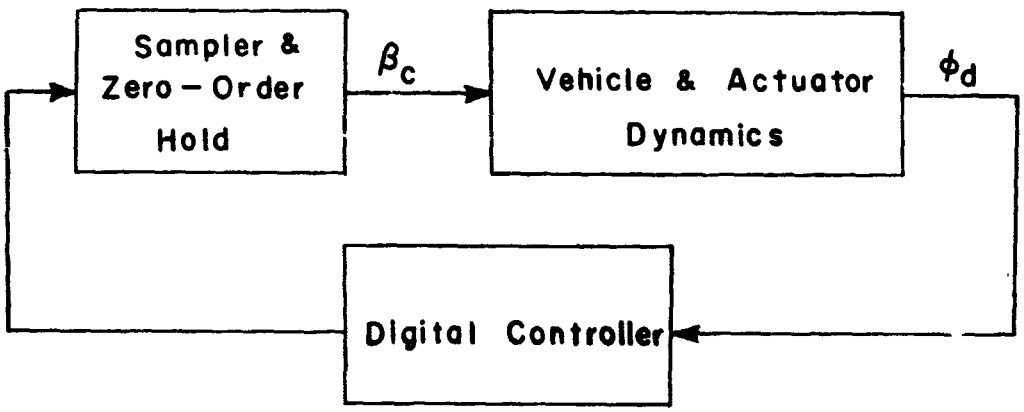

Figure 11.--Saturn V digitized attitude control system.

The differential equations which characterize the vehicle and actuator dynamics can be written in the vector-matrix notation,

$$\dot{\underline{x}} = A\underline{x} + \underline{b}\beta_c \qquad \text{(V-29)}$$

$$\phi_d = \underline{c}^T\underline{x} \qquad \text{(V-30)}$$

where $\underline{x}$ is a 14 x 1 vector, β_c and ϕ_d are scalars and the elements of A, $\underline{b}$ and $\underline{c}$ are defined in terms of vehicle parameters in Appendix D. Since the digital device accepts inputs and generates outputs every T seconds (T is a constant), it is necessary to determine the inputs to the controller at each sampling instant. This motivates the discretization of the continuous-time portion of the system which is represented by (V-29) and (V-30). The discrete-time model can be easily determined since $\beta_c(t)$ is held fixed between sampling instants by the zero-order hold element. A detailed description of a method for discretizing the continuous-time portion of a linear, stationary system is given in [32] . The discrete-time equations for the actuator and vehicle dynamics may be written as

$$\underline{x}(k + 1)T) = A_1 \underline{x}(kT) + \underline{b}_1 \beta_c(kT) \tag{V-31}$$

$$\phi_d(kT) = \underline{c}^T \underline{x}(kT) \tag{V-32}$$

Classical frequency domain techniques can be used to show that the transfer function for the digital controller,

$$D(z) = \left\{\frac{z^2 + a_1 z + a_2}{z^2 + b_1 z + b_2}\right\} \left\{\frac{z^2 + a_1' z + a_2'}{z^2 + b_1' z + b_2'}\right\} \left\{\frac{z^2 + a_1'' z + a_2''}{z^2 + b_1'' z + b_2''}\right\} \tag{V-33}$$

where,

$a_1 = -3530./2048.$ $\quad b_1 = -90289./65536.$

$a_2 = 1877./2048.$ $\quad b_2 = 34839./65536.$

$$a_1' = -128502./65536. \quad b_1' = -127717./65536 .$$
$$a_2' = 63017./65536 . \quad b_2' = 62341./65536.$$

and

$$a_1'' = -65274./65536. \quad b_1'' = -65405./65536. ,$$
$$a_2'' = 0.0 \quad b_2'' = 0.0$$

will satisfactorily stabilize the system.

It was decided to investigate the effects of quantization in two different programming forms which might be used to physically implement the transfer function of (V-33). These implementations, which are depicted in Figure 12 and Figure 13, differ only in the form of programming used to realize the second-order cascaded modules. The cascade-direct method is used in Figure 12 and the cascade-canonical method is shown in Figure 13.

1. The Cascade-Direct Configuration

Note from Figure 12 that there are seven sources of quantization associated with the cascade-direct form. The block labelled 255/15 which appears at the input of the digital device is an artifice which is useful in modelling controller operation. A maximum magnitude for ϕ_d of 15 degrees is converted by the analog-to-digital (A/D) converter to the binary word 1111111., which has the decimal equivalent of 255. Obviously then, any value of ϕ_d less than 15 degrees in magnitude will be converted by the A/D converter into an integer less than 255 in magnitude. Suppose, for example, that $\phi_d(kT)$ = 1.5 degrees. Then $\phi_d'(kT) = 25.5$ and $Q_1[\phi_d'(kT)] = 25$, since Q_1,

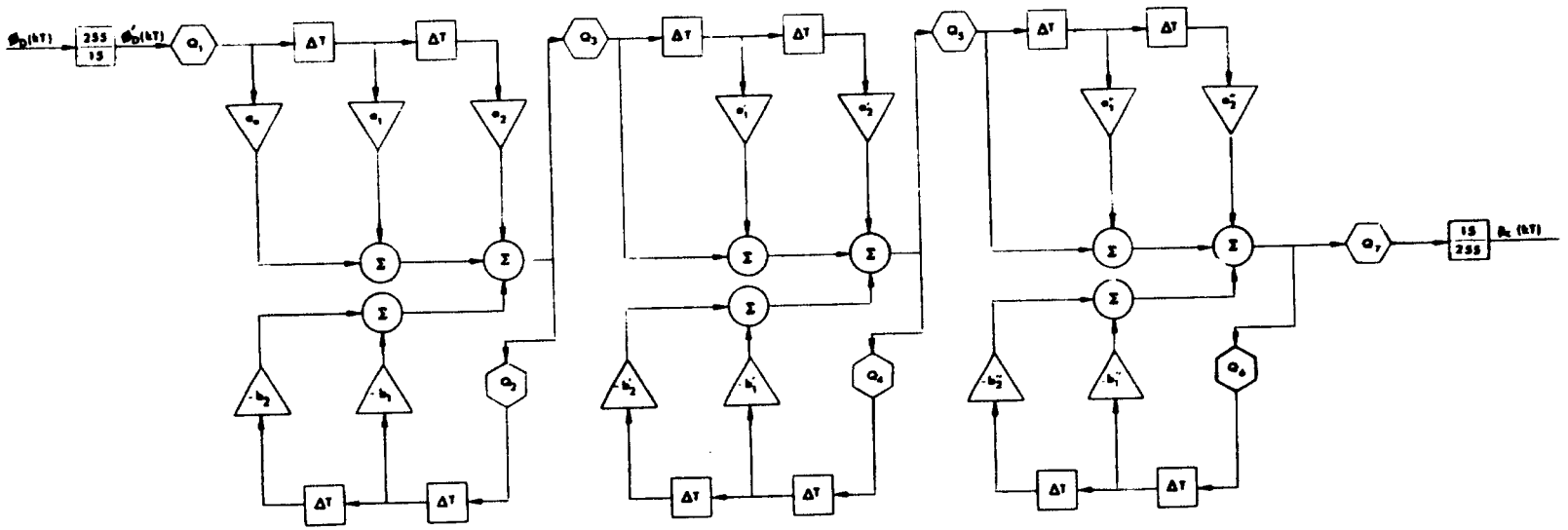

Fig. 12--Mathematical Model of the Cascade-Direct Realization

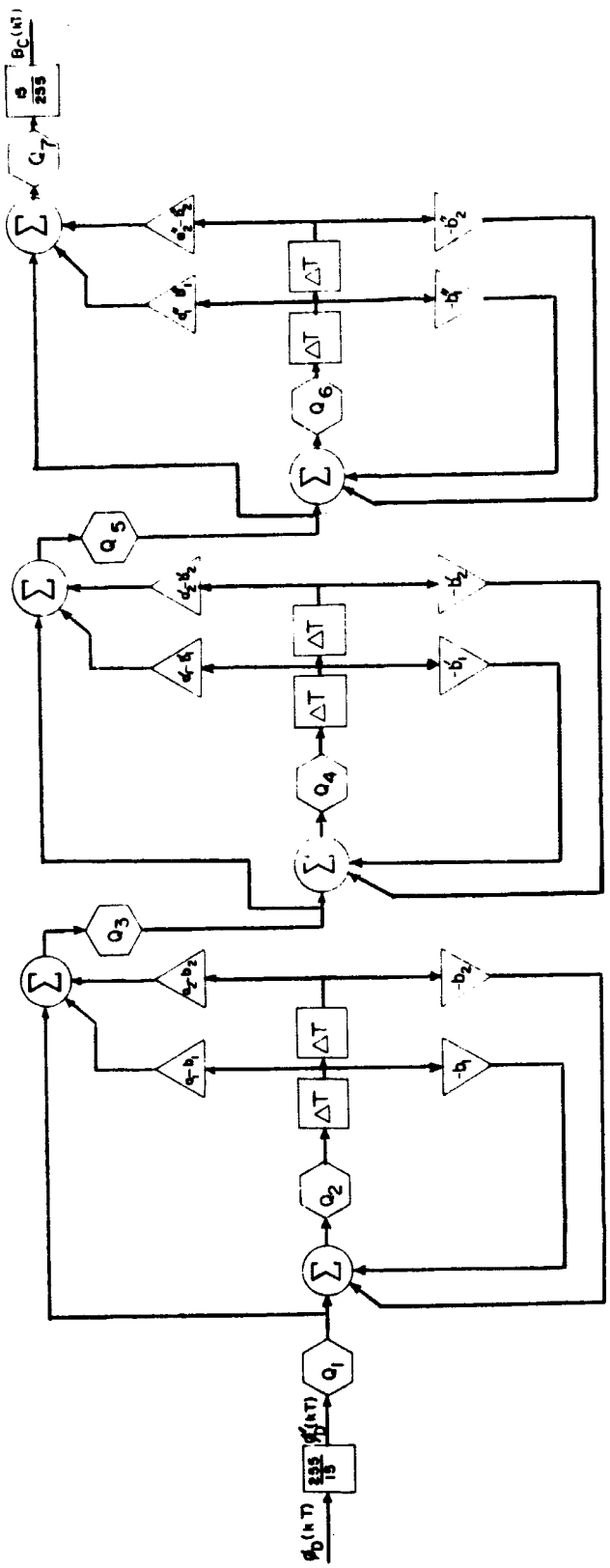

Fig 13 -- Mathematical Model of the Cascade-Canonical Realization

as well as the other 6 quantizers actually truncate in the physical system. As a consequence of the 255./15. scale change, the evaluation of quantizer granularity, h, is simplified. In general the granularity of any of the seven quantizers can be easily determined from the number of bits to the right of the binary point. If there are N bits to the right of the binary point, then $h = 2^{-N}$.

It is evident from Figure 12 that the difference equations for the ideal digital controller are of the form:

$$\underline{y}(k + 1)T) = D\underline{y}(kT) + \underline{f}\ \phi_d(kT) \tag{V-34}$$

$$\beta_c(kT) = \underline{g}^T\underline{y}(kT) + \alpha\ \phi_d(kT) \tag{V-35}$$

where $\underline{y}(kT)$ is a 10 x 1 vector,

$\phi_d(kT)$ is a scalar,

$\beta_c(kT)$ is a scalar, and

D, $\underline{f}$, $\underline{g}$, and α are of appropriate dimension.

Equations (V-31), (V-32), (V-34) and (V-35) can be combined to form a composite transition relation for the control system.

$$\begin{bmatrix} \underline{x}(\ (k + 1)T) \\ \underline{y}(\ (k + 1)T) \end{bmatrix} = \begin{bmatrix} A_1 + \alpha\underline{b}_1\underline{c}^T & \underline{b}_1\underline{g}^T \\ \underline{f}\underline{c}^T & D \end{bmatrix} \begin{bmatrix} \underline{x}(kT) \\ \underline{y}(kT) \end{bmatrix} \tag{V-36}$$

$$\phi_D(kT) = \underline{c}^T\underline{x}(kT) + 0^T\ \underline{y}(kT) = \underline{c}_1^T \begin{bmatrix} \underline{x}(kT) \\ \underline{y}(kT) \end{bmatrix} \tag{V-37}$$

Now each of the seven quantizers may be considered as an input summing point for a quantization error signal, and the quantization error transition equations may be written. These equations are:

$$\upsilon((k+1)T) = \begin{bmatrix} A_1 + \underline{\alpha} b_1 \underline{c}^T & \underline{b}_1 \underline{g}^T \\ \underline{f}\,\underline{c}^T & D \end{bmatrix} \upsilon(kT) + B\,\underline{q}(k) \quad , \tag{V-38}$$

$$\varepsilon_{\phi_d}(kT) = \underline{c}_1^T \; \underline{\upsilon}(kT) \tag{V-39}$$

where $\underline{q}(k)$ represents the quantization error input terms, and the elements of B are: (see Figure 12.)

$$b_i = b_{i3} = b_{i5} = b_{i7} = (\hat{b}_i)\ (15)/255,\ i = 1, \ldots, 14, \tag{V-40}$$

and $\hat{b}_i$ denotes the $i^{\underline{th}}$ component of vector $\underline{b}_1$. Further

$$\begin{aligned} &b_{15,1} = 1.,\ b_{17,1} = 1.,\ b_{17,2} = 1.,\ b_{19,1} = 1.,\ b_{19,3} = 1. \\ &b_{21,1} = 1.\ \ b_{21,3} = 1.\ \ b_{21,4} = 1.\ \ b_{23,1} = 1.\ \ b_{23,3} = 1. \\ &b_{23,5} = 1.\ \ b_{24,1} = 1.\ \ b_{24,3} = 1.\ \ b_{24,5} = 1.,\ b_{24,6} = 1. \end{aligned} \tag{V-41}$$

All other elements of the 24 x 7 B matrix are zero.

A digital computer program was written to perform the operations required by (V-27) in order to determine a bound for quantization error. Approximately 25 minutes of execution time on the IBM 7040 system at Auburn was required to determine the M_i numbers associated with each quantizer. This corresponded to the termination of the infinite summation of (IV-27) with N = 296. The change in each of

the M_i was less than .106% for this number of terms. The results are:

$$M_1 = 0.15779 \qquad M_2 = 0.76551 \qquad M_3 = 0.14172$$
$$M_4 = 109.22579 \qquad M_5 = 0.26475 \qquad M_6 = 26.49572 \qquad \text{(V-42)}$$
$$M_7 = 0.26430$$

Note the sensitivity of the system to the granularity of quantizers 4 and 6 as indicated by the large magnitude of these numbers relative to the other M_i. This implies that in the specification of machine word lengths, a large number of bits should be carried to the right of the binary point for quantizers 4 and 6. As an example, suppose that $h_1 = 1.$, $h_2 = 2^{-5}$, $h_3 = 1.$ $h_4 = 2^{-5}$, $h_5 = 2^{-5}$, $h_6 = 2^{-5}$ and $h_7 = 2^{-3}$. The quantization error-bound for this system is

$$|\epsilon_{\phi_d}|_{max} = \sum_{i=1}^{7} M_i h_i = 4.6060 \text{ degrees} \qquad \text{(V-43)}$$

(The 1/2 factor that appears in (V-28) was omitted since signals processed by the digital device are truncated rather than rounded to the nearest integer.) It has been found by simulation that the quantization errors which arise in the physical system are normally less than one-tenth of the magnitude of the bound given in (V-40). It is interesting to note that if the digital device was ideal, i.e., $h_2 = h_3 = h_4 = h_5 = h_6 = h_7 = 0$, then the maximum system error due to quantizer 1 is about .158 degrees. This .158 degrees represents an upper-bound on the inherent error due to quantization

in the flight computer which provides $\phi_d(k)$ to the digital controller in the physical system.

2. The Cascade-Canonical Configuration

Consider now the controller configuration shown in Figure 13 in which seven sources of quantization errors are present. Note that Q_1, Q_3, Q_5, and Q_7 are identical to their counterparts of Figure 12 in their effects upon the system dynamics. Thus, it is not unreasonable to conclude that M_1, M_3, M_5, and M_7 should be the same as the corresponding M_i numbers given by (V-42). These and the remaining M_i numbers will now be evaluated for the cascade-canonical configuration .

The procedure outlined in the foregoing discussion, (V-29) thru (V-41), can be directly applied to the configuration of Figure 13 with the following minor modifications. Instead of (V-41), (see Figure 13),

$$b_{15,1} = 1.,\ b_{15,2} = 1.,\ b_{17,1} = 1.,\ b_{17,3} = 1.,\ b_{17,4} = 1.,\ b_{19,1} = 1.,\ b_{19,3} = 1.,\ b_{19,5} = 1.,\ b_{19,6} = 1. \tag{V-44}$$

In addition, since the cascade-canonical form contains only five delay element, $y(kT)$ is a 5 x 1 vector, and the orders of the composite state transition matrix, $\upsilon(kT)$, B, and $\underline{c}_1$ in (V-38) and (V-39) are reduced by five.

The digital computer program mentioned in the previous section

was modified to accomodate the cascade-canonical form and the M_i numbers were generated by (IV-27) with termination at N = 296. The results are:

$$M_1 = 0.15779 \qquad M_2 = 0.06585 \qquad M_3 = 0.14172$$
$$M_4 = 0.18353 \qquad M_5 = 0.26475 \qquad M_6 = 0.05306 \qquad \text{(V-45)}$$
$$M_7 = 0.26430$$

It is interesting to make a comparison in this example of the error bounds resulting from the two different realizations of the controller described by (V-33). Suppose, for comparison purposes, that the same granularities exist in the quantizers of the cascade-canonical form as were selected in computing the error bound for the cascade-direct form of (V-43), i.e., $h_1 = 1.$, $h_2 = 2^{-5}$, $h_3 = 1.$, $h_4 = 2^{-5}$, $h_5 = 2^{-5}$, $h_6 = 2^{-5}$, and $h_7 = 2^{-3}$. Then the system output quantization-error bound is

$$|\epsilon_{\phi_d}|_{max} = \sum_{i=1}^{7} M_i h_i = 0.3502 \text{ degrees}, \qquad \text{(V-45)}$$

which is less than one tenth of that resulting in the system with the cascade-direct realization of the digital controller.

VI. A LEAST UPPER-BOUND ON QUANTIZATION ERROR IN DIGITAL CONTROL SYSTEMS BY THE TRANSFER FUNCTION METHOD

As indicated in Chapter II, Dejka [16] proposed a transfer function method for the computation of a bound on quantization error, but he did not point out that his method can be used for the analysis of multiple quantizer systems and that the application of this method results in the least upper bound on quantization error. Recently, Curry [2] stated without proof that the method proposed by Dejka provides a least upper bound on quantization errors. The purpose of this Chapter is to provide a mathematical development which verifies these assertions.

A. Mathematical Developments

In Chapter II, it was shown that the problem of determining a bound on quantization errors is equivalent to finding a bound for

$$y_v(k) = \underline{c}^T \underline{v}(k) \ , \tag{II-14}$$

where

$$\underline{v}(k+1) = A\underline{v}(k) + B\underline{q}(k) \ . \tag{II-13}$$

Each of the components $q_i(k)$, $i = 1,2,3, \ldots, m$, of $\underline{q}(k)$ are bounded in magnitude, i. e.,

$$\left|q_i(k)\right| \leq \frac{h_i}{2} \ , \quad i = 1, 2, \ldots, m, \text{ and } k = 0,1,2, \ldots \ . \tag{VI-1}$$

Since the system described by (II-13) and (II-14) is linear and

stationary, the z-transform theory is readily applicable. Let $G_i(z)$ represent the transfer function from the i^{th} quantizer input to the output, $y_\nu(k)$. Further, let

$$Y_\nu(z) = \mathcal{Z}[y_\nu(k)], \text{ and} \tag{VI-2}$$

$$Q_i(z) = \mathcal{Z}[q_i(k)]. \tag{VI-3}$$

Then by superposition,

$$Y_\nu(z) = \sum_{i=1}^{m} Q_i(z)\, G_i(z) + \sum_{j=1}^{n} \nu_j(0)\, H_j(z)\ , \tag{VI-4}$$

where $H_i(z)$ is the transfer function from the i^{th} state to $y_\nu(k)$. Consider next a term of (VI-4), $Q_i(z)\, G_i(z)$.

$$Q_i(z)\, G_i(z) = \left[\sum_{k=0}^{\infty} q_i(k) z^{-k}\right]\left[\sum_{n=0}^{\infty} g_i(k) z^{-k}\right] \tag{VI-5}$$

Now, the inverse z-transform of $Q_i(z)\, G_i(z)$ evaluated at $k = N$ is

$$\mathcal{Z}^{-1}\left[Q_i(z)\, G_i(z)\right]_{k=N} = \sum_{\ell=0}^{N} q_i(N-\ell)\, g_i(\ell). \tag{VI-6}$$

It is easy to see that if

$$\hat{q}_i(N-\ell) = \frac{h_i}{2} \operatorname{sgn} g_i(\ell)\ , \tag{VI-7}$$

for $\ell = 0, 1, 2, \ldots, N$,

then the summation in (VI-6) is maximized with respect to the sequence $q_i(0), q_i(1), \ldots, q_i(N)$. If the input sequence is chosen in accordance with (VI-7), then the summation of (VI-6) becomes

$$\sum_{\ell=0}^{N} \hat{q}_i(N-\ell)\, g_i(\ell) = \frac{h_i}{2} \sum_{\ell=0}^{N} |g_i(\ell)| \qquad \text{(VI-8)}$$

The number M_i defined by (V-27) can be written in terms of $q_i(\ell)$ as follows:

$$M_i = \lim_{N \to \infty} \sum_{\ell=0}^{N} |g_i(\ell)|, \qquad \text{(VI-9)}$$

$i = 1,2,\ldots,m.$

If $\underline{\nu}(0) = \underline{0}$, then it is apparent that

$$\hat{y}_\nu(N) = \frac{1}{2} \sum_{i=1}^{m} h_i \sum_{\ell=0}^{N} |g_i(\ell)|, \qquad \text{(VI-10)}$$

where, as in Chapter V, $\hat{y}_\nu(N)$ represents the maximum value of $y_\nu(N)$ for $\underline{q}(k) \epsilon \Omega$ for all $k = 0,1,\ldots,N$.

Suppose as in Chapter V that $\underline{\nu}(0)$ is nonzero, i.e.,

$$|\nu_j(0)| \leq \frac{m_j}{2} \quad j = 1,2,\ldots,t \qquad \text{(VI-11)}$$

and

$$|\nu_j(0)| = 0, \quad j = t+1,\ldots,n \quad . \qquad \text{(VI-12)}$$

Consider the inverse transform of the term $\nu_j(0)\, H_j(z)$, i.e.,

$$\mathcal{Z}^{-1}\left[\nu_j(0)\, H_j(z)\right]_{k=N} = \mathcal{Z}^{-1}\left[\nu_j(0) \sum_{n=0}^{\infty} h_j(n) z^{-n}\right] = \nu_j(0)\, h_j(N) \qquad \text{(VI-13)}$$

Thus it is apparent that the selection of $\nu_j(0)$ which maximizes $y_\nu(N)$ is

$$\nu_j(0) = \frac{m_j}{2} \operatorname{sgn} h_j(N) \ , \ j = 1,2,\ldots, t \ . \qquad \text{(VI-14)}$$

When initial conditions are included, the worst case value of $y_\nu(N)$ becomes

$$\hat{y}_\nu(N) = \frac{1}{2} \sum_{i=1}^{m} h_i \sum_{\ell=0}^{N} |g_i(\ell)| + \frac{1}{2} \sum_{j=0}^{t} m_j \, |h_j(N)| . \qquad \text{(VI-15)}$$

The method of computing $\hat{y}_\nu(N)$ as given by (IV-10) and (IV-15) was applied to the problem of Bertram described by (V-19) and (V-20). As expected, precisely the same results as those given in Tables V-1 and V-2 were obtained.

VII. CONCLUSIONS

The problem of determining an upper bound for the magnitude of system errors that may arise in a discrete-time, hybrid control system due to quantization of some of the system variables was considered in detail. Three distinct methods for computing a bound on quantization errors were advanced.

The first method, which was discussed in Chapter IV, is based on the concept of "Uniform Boundedness of Solutions" as it applies to discrete-time systems. This theoretical basis was used in essentially two different ways to establish quantization error-bounds. In one case, an output-bound was determined by computing a bound for each quantizer input and then determining an overall error-bound for the system by application of the principle of superposition and the triangular inequality. However, in order to obtain the bound on each input it is necessary to solve a constrained maximization problem by a computer search technique. This requirement has the following implications: (a) The output-bound cannot be found in closed form in terms of system parameters. (b) It is necessary to implement a generally complex computer search routine. The second application of the concept of "Uniform Boundedness of Solutions" that was set forth in Chapter IV results in a solution for a quantization error-bound which can be written in closed form in terms of system parameters and which simultaneously accounts for the error contributions of each of the system quantizers. Both of these approaches provide an upper-bound on

quantization error which is often quite conservative. For instance, the second of the above procedures was used to determine a bound on quantization error in the system originally studied by Bertram. It was determined that an error-bound for this system is 34.6. The methods developed in other chapters of this dissertation were used to show that the least upper-bound on quantization error is only 2.918 for this system.

The optimal control theory for discrete-time systems was used to develop a method for computing a least upper-bound on quantization error. This was accomplished by recognizing that the problem of determining a least upper-bound on quantization error can be cast into the form of an optimal control problem. A closed form solution was obtained for the least upper-bound on quantization error at any sampling instant, NT. Another result of considerable significance when the synthesis of a digital controller is considered was the development of a method for computing a non-negative number for each quantizer such that the summation of the products of each of these numbers with its respective maximum quantization error yields the least upper-bound on system quantization error. The magnitude of these non-negative numbers allows the designer to determine the degree of sensitivity of system quantization error to the granularity of each quantizer and to systematically choose levels of quantization which are needed in order to meet any specification on maximum system quantization error.

Another benefit accruing from the optimal-control theory viewpoint is that a method was developed for choosing the worst possible initial

errors which result from quantization. Then, these worst possible initial errors were used in conjunction with the results discussed above to give a least upper-bound on system error for any time NT where N is a positive integer and T is the sampling interval.

Finally, it was established that the optimal approaches taken above can be used to compute error-bounds on quantization for linear systems using deterministic sampling schemes.

In Chapter VI, it was established that the transfer function can be utilized to obtain precisely the same quantization error-bounds as were determined by the optimal control methods if a constant sampling rate is assumed. It appears that the transfer function method is preferable if the digital computer is controlling a plant of low order. However, as plant order increases, the problem of obtaining transfer functions with sufficient numerical accuracy become increasingly difficult and it becomes advantageous to utilize the vector-matrix notation associated with the optimal-control method for bounding quantization errors.

REFERENCES

[1] E. I. Jury, Theory and Application of the Z-Transform Methods. New York: John Wiley and Sons, 1964.

[2] E. E. Curry, "The Analysis of Round-Off and Truncation Errors in a Hybrid Control System," IEEE Transactions on Automatic Control, Volume AC-12, No. 5, October 1967, pp. 601-604.

[3] Ya. Z. Tsypkin, "Elements of Theory of Numerical Automatic Systems," Automatic and Remote Control, Proceedings First Int. Congress IFAC, Butterworths, England, 1961, pp. 286-294.

[4] Yu. M. Korshunov, "The Analysis of Periodic States Due to Level Quantization of Signal in Automatic Digital Systems," Automation and Remote Control, Volume 22, July 1961, pp. 778-787.

[5] C. K. Chow, "Contactor Servomechanisms Employing Sampled Data," AIEE Transactions on Applications and Industry, Volume 73, Part II, 1954, pp. 51-64.

[6] E. Biondi, A. Debenedetti, and P. Rotoloni, "Error Determination in Quantized Sampled-Data Systems," Proceedings of the Third Congress of the International Federation of Automatic Control, London, 1966, pp. 34D.1-34D.7.

[7] W. R. Bennet, "Spectra of Quantized Signals," Bell System Technical Journal, Volume 27, July 1948, pp. 446-472.

[8] B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," IRE Transactions on Circuit Theory, December, 1956, pp. 266-276.

[9] B. Widrow, "Statistical Analysis of Amplitude Quantized Sampled-Data Systems," AIEE Transactions on Applications and Industry, No. 52, January 1961, pp. 555-568.

[10] J. E. Bertram, "The Effect of Quantization in Sampled Feedback Systems," AIEE Transactions on Applications and Industry, Volume 77, September 1958, pp. 177-182.

[11] J. B. Slaughter, "Quantization Errors in Digital Control Systems," IEEE Transactions on Automatic Control, Volume AC-9, January 1964, pp. 70-74.

[12] A. J. Monroe, Digital Processes for Sampled Data Systems. New York: John Wiley and Sons, 1962.

[13] G. W. Johnson, "Upper Bound on Dynamic Quantization Error in Digital Control Systems via the Direct Method of Liapunov," IEEE Transactions on Automatic Control, Volume AC-10, No. 4, October 1965, pp. 439-448.

[14] G. W. Johnson and G. N. T. Lack, "Comments on Upper Bound on Dynamic Quantization Error in Digital Control Systems via the Direct Method of Liapunov," IEEE Transactions on Automatic Control, Volume AC-11, No. 2, April 1966, pp. 331-334.

[15] A. J. Kiovuniemi, "Quantization Error in the Design of Digital Control Systems," 1967 Joint Automatic Control Conference Preprints of Papers, June 1967, pp. 32-39.

[16] W. J. Dejka, "Quantization Error from Z-Transform of Transfer Function," IEEE Transactions on Automatic Control, Volume 8, No. 1, January 1963, pp. 72-73.

[17] W. A. Porter, Modern Foundations of Systems Engineering. New York: The Macmillan Company, 1966.

[18] M. Athans and P. L. Falb, Optimal Control. New York: McGraw Hill, 1966.

[19] L. A. Zadeh and C. A. Desoer, Linear System Theory. New York: McGraw Hill, 1963.

[20] P. R. Halmos, Finite Dimensional Vector Spaces. Princeton, New Jersey: D. Van Nostrand, 1958.

[21] P. M. Derusso, R. J. Roy and C. M. Close, State Variables for Engineers. New York: John Wiley and Sons, 1966.

[22] H. Freeman, Discrete-Time Systems. New York: John Wiley and Sons, 1965.

[23] W. Hahn, "Uber die Anwendung der Methode von Ljapunov auf Differenzengleichungen," Math Annalen, Volume 136, 1958, pp. 430-441.

[24] T. M. Apostol, Mathematical Analysis. Reading, Mass.: Addison-Wesley, 1957.

[25] G. Hadley, Nonlinear and Dynamic Programming. Reading, Mass.: Addison-Wesley, 1964.

[26] W. Hahn, Theory and Application of Liapunov's Direct Method. Englewood Cliffs, N. J.: Prentice-Hall, 1963.

[27] J. Lasalle and S. Lefschetz, Stability by Liapunov's Direct Method with Applications. New York: Academic Press, 1961.

[28] T. Yoshizawa, "Liapunov's Functions and Boundedness of Solutions," RIAS Tech. Rep. TN-59-1218, December 1959, 10 pages.

[29] R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," The Computer Journal, Volume 6, 1963, pp. 163.

[30] J. H. George, "A Method for Interpretation of Results Obtained by Liapunov's Method,' 1966 Joint Automatic Control Conference Preprints of Papers, Seattle, pp. 841-847.

[31] H. Halkin, "A Maximum Principle of the Pontryagin Type for Systems Described by Nonlinear Difference Equations," J. SIAM Control, Volume 4, No. 1, 1966, pp. 90-111.

[32] Phillips, C. L., et al., "Simulation of High-Order Hybrid Control System," Technical Report Number 12, NAS8-11274, Engineering Experiment Station, Auburn, Alabama, July, 1968.

APPENDIX A

In this appendix, it will be established that $\Delta\hat{V}[\underline{\nu}(k)]$ for the multiple quantizer system cannot equal zero at more than one point along each ray from the origin. Consider a ray through the point $\underline{\nu}_1 \neq 0$ and further assume that, (all arguments, k, are omitted for convenience)

$$\Delta\hat{V}[\underline{\nu}_1] = 0 = -\underline{\nu}_1^T Q \underline{\nu}_1 + 2\underline{q}_1^T B^T P A \underline{\nu}_1 + \underline{q}_1^T B^T P B \underline{q}_1 \ , \qquad \text{(A-1)}$$

where $\underline{q}_1 \epsilon \Omega$ is the input which maximizes $\Delta V[\underline{\nu}_1]$. It is easy to show that if $\underline{q}_1$ is a maximizing input, then terms two and three on the right hand side of (A-1) are non-negative.

Let us assume that $\underline{\nu}_1$ is the "first" point on the ray at which $\Delta\hat{V}$ is zero and that there exists at least one other point, $t\underline{\nu}_1$, $t > 1$ at which $\Delta\hat{V}[\underline{\nu}]$ is zero for maximizing input $\underline{q}_2$. It is not difficult to show that if $\underline{q}_2 = \underline{q}_1$, then $\Delta\hat{V}$ is zero only at $\underline{\nu}_1$ on the ray. Consider the following proof: Let

$$a = \underline{\nu}_1^T Q \underline{\nu}_1 \ , \ b = 2\underline{q}_1^T B^T P A \underline{\nu}_1 \ , \text{ and } c = \underline{q}_1^T B^T P B \underline{q}_1 \ . \qquad \text{(A-2)}$$

Then from (A-1)

$$-a + b + c = 0. \qquad \text{(A-3)}$$

Assume that at some other point $t\underline{\nu}_1$, $\Delta V[t\underline{\nu}_1] = 0$. Then if $\underline{q}_2 = \underline{q}_1$

$$-t^2 a + tb + c = 0 \ . \qquad \text{(A-4)}$$

It follows from (A-3) and (A-4) that

$$t = 1; \; -\frac{2c}{2a} < 0. \tag{A-5}$$

Consequently the only point on the ray at which $\Delta\hat{V}[\underline{v}] = 0$ occurs at $\underline{v}_1$ if $\underline{q}_2 = \underline{q}_1$.

Next the case where $\underline{q}_2 \neq \underline{q}_1$ will be considered. Equation (A-1) becomes

$$\Delta\hat{V}[t\underline{v}_1] = - t^2\underline{v}_1^T Q\underline{v}_1 + 2t\underline{q}_2^T B^T PA\underline{v}_1 + \underline{q}_2^T B^T PB\underline{q}_2 = 0. \tag{A-6}$$

Let

$$b_1 = 2\underline{q}_2^T B^T PA\underline{v}_1 \quad \text{and} \quad c_1 = \underline{q}_2^T B^T PB\underline{q}_2 \; . \tag{A-7}$$

Then (A-6) becomes

$$-t^2 a + tb_1 + c_1 = 0. \tag{A-8}$$

Equations (A-8) and (A-3) may be combined to eliminate a. The result is

$$t^2(b+c) = tb_1 + c_1 \; . \tag{A-9}$$

Now since $\Delta\hat{V}[\underline{v}_1]$ is maximized by $\underline{q}_1 \neq \underline{q}_2$, it follows that

$$b + c \geq b_1 + c_1 \quad . \tag{A-10}$$

Consequently, if inequality (A-10) is substituted into (A-9),

$$t^2 \, [b_1+c_1] \leq tb_1 + c_1 \quad . \tag{A-11}$$

However, since $t > 1$, $b_1 \geq 0$ and $c_1 \geq 0$ and b_1 and c_1 cannot be simultaneously zero, then (A-11) is a contradiction.

Therefore it has been shown that $\Delta\hat{V}[\underline{v}]$ cannot be zero more than once along any ray from $\underline{0} \epsilon E^n$.

APPENDIX B

Let the set of points, H, described by

$$\underline{\nu}^T\underline{\nu} = \|\underline{\nu}\|^2 = \psi^2 \ , \ \psi = \text{a constant} \tag{B-1}$$

represent the initial set in E^n for transformation E. Therefore, in order to compute $\|E\|$ the vector $\underline{\nu}_0 \epsilon H$ must be selected with the property that

$$\|E\underline{\nu}_0\|^2 = \underline{\nu}_0^T E^T E\underline{\nu}_0 \geq \underline{\nu}^T E^T E\underline{\nu} \ , \quad \forall\, \underline{\nu}\epsilon H. \tag{B-2}$$

Formally the problem becomes

Maximize: $\underline{\nu}^T E^T E\underline{\nu}$ (B-3)

Subject to $\underline{\nu}^T\underline{\nu} = \psi^2$ (B-4)

The application of Lagrange multiplier techniques yields,

$$F = \underline{\nu}^T E^T E\underline{\nu} + \lambda(\underline{\nu}^T\underline{\nu}), \text{ and} \tag{B-5}$$

$$\text{grad}_{\nu}\ F = 2E^T E\underline{\nu} + 2\lambda\,\underline{\nu} = 0 \tag{B-6}$$

The inner product of (B-6) with $\underline{\nu}$ gives

$$\underline{\nu}^T E^T E\underline{\nu} + \lambda\underline{\nu}^T\underline{\nu} = 0. \tag{B-7}$$

The utilization of (B-4) with (B-7) gives

$$\lambda = (-\underline{\nu}^T E^T E\underline{\nu}) / \psi^2 \tag{B-8}$$

Substituting this value of λ into (B-6) yields

$$E^T E\underline{\nu} = \quad {}^{T}\,{}^{T}E\underline{\nu}\,,\ \underline{\nu} \tag{B-9}$$

However, since $\underline{\nu}^T E^T E\underline{\nu}$ is a scalar, (B-9) can only be satisfied by the eigenvectors of E^TE. Consequently the term $\frac{\underline{\nu}^T E^T E\underline{\nu}}{\psi^2}$ is an eigenvalue of E^TE. Therefore, in order for $\underline{\nu}_0^T E^T E\underline{\nu}_0$ to be a maximum over H, $\underline{\nu}_0$ must be the eigenvector corresponding to the maximum eigenvalue of E^TE, i.e.,

$$\|E\underline{\nu}_0\|^2 = \lambda_{max}(E^TE)\psi^2 = \lambda_{max}(E^TE)\ \|\underline{\nu}_0\|^2 \tag{B-10}$$

$$\text{for } \underline{\nu}_0 \epsilon H.$$

From (B-10) it is evident that

$$\|E\underline{\nu}\| \leq \left[\lambda_{max}\ (E^TE)\right]^{1/2} \|\underline{\nu}\| \ . \tag{B-11}$$

and

$$\|E\| = \left[\lambda_{max}(E^TE)\right]^{1/2} . \tag{B-12}$$

APPENDIX C

The solution of $\underline{\hat{v}}(N)$, which is given in (V-13) and which was developed with the aid of the Discrete Maximum Principle, can also be obtained from the general solution of the III-13 with $q_i(k) \in \Omega$, $i = 1,2,\ldots,m$ and $k = 0,1,\ldots,N-1$. The general solution is

$$\underline{v}(N) = A^N \underline{v}(0) + \sum_{k=0}^{N-1} A^{N-(k+1)} B \, \underline{q}(k) \; . \tag{C-1}$$

Further,

$$y_v(N) = \underline{c}^T A^N \underline{v}(0) + \underline{c}^T \sum_{k=0}^{N-1} A^{N-(k+1)} B \, \underline{q}(k) \tag{C-2}$$

$$= \langle \underline{c}, \, A^N \underline{v}(0) \rangle + \sum_{k=0}^{N-1} \langle \underline{c}, \, A^{N-(k+1)} B \, \underline{q}(k) \rangle$$

or

$$y_v(N) = \langle \underline{v}(0), \, [A^N]^T \, \underline{c} \rangle + \sum_{k=0}^{N-1} \langle \underline{q}(k), \, B^T [A^{N-(k+1)}]^T \, \underline{c} \rangle \; . \tag{C-3}$$

However, since $[A^\ell]^T = [A^T]^\ell$ where ℓ is an integer, (C-3) may be written as

$$y_v(N) = \langle \underline{v}(0), \, [A^T]^N \, \underline{c} \rangle + \sum_{k=0}^{N-1} \langle \underline{q}(k), \, B^T [A^T]^{N-(k+1)} \, \underline{c} \rangle \; . \tag{C-4}$$

We can immediately conclude from (C-4) that

$$\hat{q}_i(k) = \frac{h_i}{2} \operatorname{sgn} \langle \underline{b}_i, [A^T]^{N-(k+1)} \underline{c} \rangle \qquad \text{(C-5)}$$

$$\forall\ i=1,2,\ldots,m \text{ and}$$

$$\forall\ k=0,1,\ldots,N-1\ .$$

This is precisely the same result expressed by (V-10) in Chapter V.

The method of proof used above suggests an extension of the results of this paper to quantizing systems with a non-uniform sampling period. As an example consider a system which alternately samples with a period of T_1 and T_2 seconds. Let the system equations be:

$$\begin{aligned} \underline{v}(k+1) &= A_1\underline{v}(k) + B_1\underline{q}(k) && \text{for } T = T_1 \text{ and} \\ \underline{v}(k+1) &= A_2\underline{v}(k) + B_2\underline{q}(k) && \text{for } T = T_2\ . \end{aligned} \qquad \text{(C-6)}$$

In order to avoid notational difficulties, consider the case where N=4. Then,

$$\begin{aligned} \underline{v}(4) &= A_2A_1A_2A_1\,\underline{v}(0) + A_2A_1A_2\,B_1\underline{q}(0) + A_2A_1\,B_2\underline{q}(1) \\ &\quad + A_2B_1\underline{q}(2) + B_2\underline{q}(3)\ . \end{aligned} \qquad \text{(C-7)}$$

Continuing as before,

$$\begin{aligned} y_v(4) &= \langle \underline{v}(0),\ A_1^TA_2^TA_1^TA_2^T\,\underline{c} \rangle + \langle \underline{q}(0),\ B_1^TA_2^TA_1^TA_2^T\,\underline{c} \rangle \\ &\quad + \langle \underline{q}(1),\ B_2^TA_1^TA_2^T\,\underline{c} \rangle + \langle \underline{q}(2),\ B_1^TA_2^T\,\underline{c} \rangle \\ &\quad + \langle \underline{q}(3),\ B_2^T\underline{c} \rangle\ . \end{aligned} \qquad \text{(C-8)}$$

The method of choosing the optimal quantization sequence is similar to that shown in (C-5) for the system with constant A and B matrices. It is evident that this technique can be easily extended to treat more complicated sampling schemes.

APPENDEX D

ELEMENTS OF THE A, B, AND C MATRICES

```
A IS A 14 X 14 MATRIX

B IS A 14 X 1 MATRIX

C IS A 14 X 1 MATRIX

A(1,1)=-2.*ZETA1*W1
A(1,2)=-W1**2
A(2,1)=1.0
A(3,2)=W1**2
A(3,3)=-2.*ZETA2*W2
A(3,4)=-W2**2
A(4,3)=1.0
A(5,2)=(ALCG*ASE+AIE)*W2**2*(-W1**2)/AIXX
A(5,3)=(ALCG*ASE+AIE)*W2**2*2.*ZETA2*W2/AIXX
A(5,4)=(ALCG*ASE+AIE)*W2**2*W2**2/AIXX + W2**2*
1 (-C2-AK3*ASE/AIXX)
A(5,6)=-C1
A(5, 8)= -(FC*(ALCG*YPB(1)+YB(1)))/AIXX
A(5,10)= -(FC*(ALCG*YPB(2)+YB(2)))/AIXX
A(5,12)= -(FC*(ALCG*YPB(3)+YB(3)))/AIXX
A(5,14)= -(FC*(ALCG*YPB(4)+YB(4)))/AIXX
A(6,5)=1.0
A(7,2)=(ASE*YB(1)-AIE*YPB(1))*W2**2*W1**2/GM(1)
A(7,3)=(ASE*YB(1)-AIE*YPB(1))*W2**2*(-2.*ZETA2*W2)/GM(1)
A(7,4)=(ASE*YB(1)-AIE*YPB(1))*W2**2*(-W2**2)/GM(1)+
1 W2**2*RP*YB(1)/GM(1)
A(7,7)=-2.*ZETA(1)*WB(1)
A(7,8)=-WB(1)**2
A(8,7)=1.0
A(9,2)=W2**2*(ASE*YB(2)-AIE*YPB(2))*W1**2/GM(2)
A(9,3)=W2**2*(ASE*YB(2)-AIE*YPB(2))*(-2.*ZETA2*W2)/GM(2)
A(9,4)=W2**2*(ASE*YB(2)-AIE*YPB(2))*(-W2**2)/GM(2) +
1 W2**2*RP*YB(2)/GM(2)
A(9,9)=-2.*ZETA(2)*WB(2)
A(9,10)=-WB(2)**2
A(10,9)=1.0
A(11,2)=W2**2*(ASE*YB(3)-AIE*YPB(3))/GM(3)*W1**2
A(11,3)=W2**2*(ASE*YB(3)-AIE*YPB(3))/GM(3)*(-2.*ZETA2*W2)
A(11,4)=W2**2*(ASE*YB(3)-AIE*YPB(3))/GM(3)*(-W2**2) +
1 W2**2*RP*YB(3)/GM(3)
A(11,11)=-2.*ZETA(3)*WB(3)
A(11,12)=-WB(3)**2
A(12,11)=1.0
A(13,2)=W2**2*(ASE*YB(4)-AIE*YPB(4))/GM(4)*W1**2
A(13,3)=W2**2*(ASE*YB(4)-AIE*YPB(4))/GM(4)*(-2.*ZETA2*W2)
```

```
 A(13,4)=W2**2*(ASE*YB(4)-AIE*YPB(4))/GM(4)*(-W2**2) +
1 W2**2*RP*YB(4)/GM(4)
 A(13,13)=-2.*ZETA(4)*WB(4)
 A(13,14)=-WB(4)**2
 A(14,13)=1.0

 ALL THE OTHER ELEMENTS ARE 0.0

 B(1)=1.0

 ALL THE OTHER ELEMENTS ARE 0.0

 C(6)=1.0
 C(8)=YPD(1)
 C(10)=YPD(2)
 C(12)=YPD(3)
 C(14)=YPD(4)

 ALL THE OTHER ELEMENTS ARE 0.0
```