1961-06-01

An analysis of the precepts available for synthesizing feedback control system when output characteristics are specified

Meyer, Wayne Eugene; Noble, Thomas Inglehart

Monterey, California. Naval Postgraduate School

AN ANALYSIS OF THE PRECEPTS AVAILABLE FOR SYNTHESIZING FEEDBACK CONTROL SYSTEMS WHEN OUTPUT CHARACTERISTICS ARE SPECIFIED

WAYNE E. MEYER
and
THOMAS I. NOBLE

AN ANALYSIS OF THE PRECEPTS AVAILABLE FOR SYNTHESIZING FEEDBACK CONTROL SYSTEMS WHEN OUTPUT CHARACTERISTICS ARE SPECIFIED

by

Wayne Eugene Meyer

Lieutenant Commander, United States Navy

B.S.E.E. University of Kansas, 1946
B.S.E.E. Massachusetts Institute of Technology, 1947
B.S.E.E. U. S. Naval Postgraduate School, 1960

and

Thomas Iglehart Noble

Lieutenant Commander, United States Navy

B.S. United States Naval Academy, 1950
B.S.E.E. U. S. Naval Postgraduate School, 1960

Submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE
IN
AERONAUTICS AND ASTRONAUTICS

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1961

ACKNOWLEDGMENTS

The authors wish to express their appreciation to the following persons who assisted them in the preparation of this thesis:

Professor Robert K. Mueller of the Department of Aeronautics and Astronautics for his unique and reassuring comments and encouragement as thesis supervisor during the course of this work.

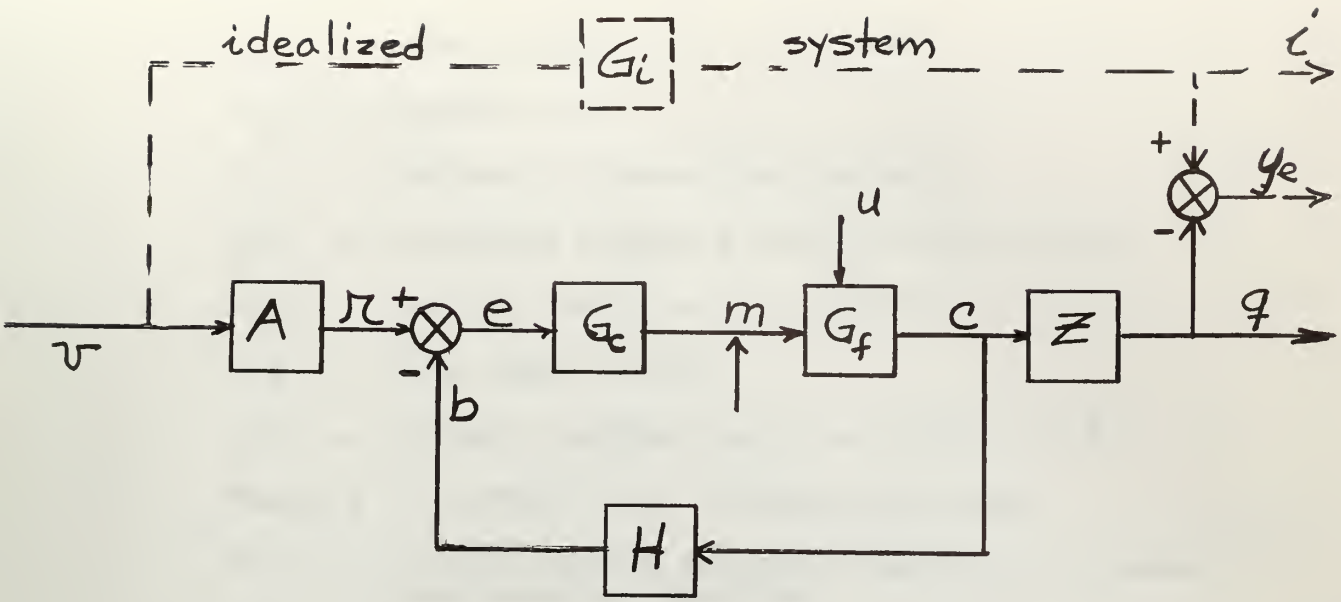
Dr. George J. Thaler and Dr. R.C.H. Wheeler of the United States Naval Postgraduate School, Monterey, California, who, as their teachers, equipped the authors with the concepts necessary for the execution of this thesis.

The graduate work for which this thesis is a partial requirement was performed while the authors were assigned by the Chief of Naval Personnel for graduate training at the Massachusetts Institute of Technology.

The authors also wish to thank Mrs. Virginia Worthington and Mrs. Eleanor Noble for their work in typing the manuscript.

TABLE OF CONTENTS

LIST OF SYMBOLS *



A = Reference input elements

G_c = Control elements (or compensating element)

G_f = Controlled system (or fixed elements or plant elements)

G_i = idealized system

Z = Indirectly controlled system

H = feedback elements

V, v = command

R, r = reference input

E, e = actuating signal

B, b = primary feedback

M, m = manipulated variable

U, u = disturbance

C, c = controlled variable

I, i = ideal value

Y_e, y_e = system error

Q, q = indirectly controlled variable

E/R = actuating signal ratio = $1/(1 + G\,H)$

C/R = control ratio = $G/(1 + G\,H)$

B/E = loop ratio = GH

B/R = primary feedback ratio = $GH/(1 + G\,H)$

Where G is forward loop transfer function.

W, w = overall system transfer function or closed loop transfer function

Q/V = W

RHP, LHP = Right or left half plane

Large letters represent frequency domain (or transforms) and small letters represent time domain.

* Above are standard symbols taken from reference 1.

AN ANALYSIS OF THE PRECEPTS AVAILABLE FOR SYNTHESIZING FEEDBACK CONTROL SYSTEMS WHEN OUTPUT CHARACTERISTICS ARE SPECIFIED

by

Wayne E. Meyer

and

Thomas I. Noble

Submitted to the Department of Aeronautics and Astronautics on 20 May 1961, in partial fulfillment of the requirements for the degree of Master of Science in Aeronautics and Astronautics.

ABSTRACT

This thesis contains a compilation and comparison of some of the techniques in use today for synthesizing and analyzing servomechanisms. The techniques are divided into the categories of Trial and Error Design Methods and Analytical Design Methods. Trial and Error Design Methods include such items as Root Locus techniques, Bode and Nichols plots, and other frequency plane plots. Considerable space is devoted to the Mitrovič method and to the Ross-Warren/Mariotti technique of compensation.

Analytical Design Methods are generally considered to be those which use some definition of error as a performance index, where the objective is to minimize (or maximize) the performance index. Considerable space is allocated to those methods published by Newton. Some effort is also devoted to the problem of obtaining a proper statement of specifications.

Thesis Supervisor: Robert K. Mueller

Title: Associate Professor of Aeronautics and Astronautics

Chapter 1

THE PROBLEM

1.1 General.

Feedback Control System synthesis means the determination of system and component specifications to meet the requirements of a specified job. The first phase of such a procedure involves the selection or design of a power element adequate to drive the load. This paper, however, is not concerned with the selection of components, and the power element is stipulated.

The type of Feedback Control System considered is a servomechanism. By definition here, a servomechanism is a particular type of feedback control system in which the controlled variable is a mechanical position. The output is the mechanical position of one object relative to another.

The words "synthesis" and "analysis" are frequently used interchangeably and loosely. Here, we will use J. R. Burnett's[1*] definition:

(1) The synthesis problem. Given the input to a system and the required output, determine the transfer function of the system.

(2) The analysis problem. Given the input to a system and the transfer function of the system, find the

* numerical superscripts refer to Bibliography

output of the system.

The type of system under consideration is lumped parameter, finite, time-invariant, and linear.

Stability is a problem to be reckoned with. We can say that stability is the primary consideration in all control systems. In that sense, it can be considered as a basic specification, always implied. However, in another sense, it is of only secondary consideration, because it is virtually always possible to render a system stable by some form of compensation. For linear systems, stability is a function of the system alone and is not dependent upon the input to the system. A system whose response will eventually become arbitrarily small, once the input is removed is "stable". Stated another way, a system is stable, if the impulse response of the system approaches a constant value and remains constant for large values of time after the impulse has occurred. If the closed-loop system function is stated as

$$\frac{C(s)}{R(s)} = K \frac{s^n + a_{n-1} s^{n-1} + \cdots}{s^m + b_{m-1} s^{m-1} + \cdots} \equiv K \frac{N(s)}{D(s)} \qquad (m > n)$$

then D(s) is a Hurwitz polynomial, all of whose roots must have a negative, non-zero, real part, for stability.

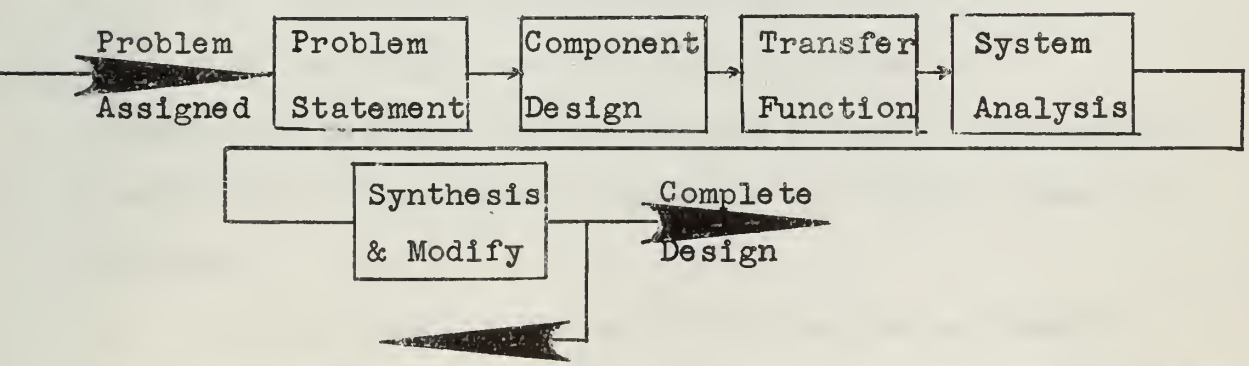
In the preparation of this paper we have researched a great many writings. We find all of them in agreement on one thing: In the design of Feedback Control Systems, arriving at an accurate and complete statement of the

problem is the most difficult phase. The difficulty of this phase is aggravated by the fact that Feedback Control Systems have permeated all walks of our modern life. The science of feedback control knows no bounds as it cuts across all historic lines of endeavor. The science has become so broad with so many people affected by it, that a statement of a particular design problem has little universal application. Each problem statement seems to be unique unto itself.

The vastness of this new science is evidenced by the fact that many engineers no longer call themselves "Electrical" or "Mechanical" or "Aeronautical" Engineers, but rather "Control" Engineers. Many books are being written and published as "Control Engineering" books. There is a great deal of discussion encouraging the divorce of Feedback Control Systems from all other historical lines of engineering, to let it grow as an independent science.

1.2 Flow Diagram of Design

A typical design flow diagram might be as follows: (See, for example, writings of R. B. Wilcox[1]).



Note that "feedback" continually exists through all the

design phases. This is what makes control system design so fascinating, but at the same time so maddening.

The problem statement may be given in any one of three forms:

(1) a literal description of what the system is supposed to do, or

(2) a numerical specification of the system performance, or

(3) a graphical representation of the system.

From the designer's viewpoint, numerical specification is preferred to eliminate ambiguity, but the price is loss of flexibility in meeting changing design considerations. If no graphical representation is provided the designer is immediately required to prepare a functional block diagram of some kind so that he may proceed with a component design analysis. Completion of this phase leads to the determination of suitable transfer functions for the components. Component transfer functions may be determined by experimental or analytical means. There are two basic experimental methods; determine the transfer function from:

(1) the component frequency response

(2) the component transient response.

Analytical methods require the use of basic mathematical and physical principles (for example, Newton's Laws of Motion).

In any event, a system block diagram may then be constructed, from which the "system" (or what we prefer to

here call "plant") transfer function is determined in either algebraic or graphical form. A pitfall comes to light here, because at this juncture, there is a tendency to forget the limitations imposed on the assumed mathematical model. These limitations might result from approximation, initial conditions, noise, drift or some other special input.

System analysis might then proceed using time domain or frequency domain methods with or without the aid of analog or digital computers. It is this analysis which then forms a basis for any synthesis or modification technique.

1.3 Problem Statement

The type or manner of statement of the problem requirements often determines the feasibility of any particular approach. Although ideally, the requirements are documented in the form of complete numerical specifications, more often than not, the only information is the general requirement that some given operation must be automatically controlled. The designer, therefore, must conduct a study to formalize his own detailed specifications.

1.4 Component Design and Analysis

This paper is not concerned with this phase of design. It will here be assumed that the components of the plant are specified. We are concerned with the

component design of the compensator only in the sense of whether the component is physically realizable.

1.5 Transfer Functions

Since we here assume linear time-invariant systems, transfer functions may be manipulated following algebraic rules; isolation between component transfer functions is implied. It follows that the concept of superposition is valid. This means the output of the system may be determined as the result of all the various inputs and disturbances, each applied separately. It is also possible to distinguish between dynamic and steady-state performance in each of these cases. The error may also be analyzed both dynamically and in steady state. However, a transfer function represents the assumption of some mathematical model, and as such, the limitations imposed by the assumptions must be kept constantly in mind. In the mechanics of attempting system optimization, for example, the limitation of saturation is always present. As the signal approaches saturation, the assumption of linearity becomes less and less valid.

1.6 System Analysis and Synthesis

The time and frequency analysis of the system characteristics is an analytical and/or graphical prediction that the system will perform as the dynamic specifications require. If it does not, then synthesis is required to determine necessary modifications and revisions to fulfill the response and accuracy

specifications. Synthesis and Analysis then go hand in hand; the final analysis is the one which defines all the characteristics of the completed design. We generally refer to system modifications and revisions as system "compensation". Compensation will mean the inclusion of that network which is necessary in order to: (1) make the system stable, and (2) meet the specifications.
A study of the literature reveals this phase to be the one where writers and designers reach a parting of the ways. They proceed down one of two broad paths. The best descriptive titles for these paths that we have found are those used by Newton[2]:
(1) Trial and Error Design Methods, and (2) Analytical Design Methods. Some of the more lucid writings in the current literature on these two philosophies are those of Thaler and Brown[3] for Trial and Error Design Methods and those of Newton[2] for Analytical Design Methods.

1.7 Objective of This Paper

In this paper, we attempt to bring under one cover a synopsis of the current methods suggested or in use for the design of the servomechanism. Our objective is to provide insight into that phase of design surrounding the problem statement where the designer makes his decision as to what approach to attempt first. To that end an understanding of the capabilities and limitations of the various techniques is required. The methods considered,

although not all inclusive, are those which, in the opinion of the authors, do have the greatest usefulness or offer the most potential.

In Chapter 2 we explore more thoroughly the statement of specifications. Chapters 3 and 4 contain a summation of the basic procedures and mechanics employed in the above two design methods. In Chapters 5 and 6, we have listed some of the modifications and refinements that are found in the current literature. The compilation is not all inclusive but the techniques listed attempt in some way to make the problem of compensation easier within their specified limitations.

CHAPTER 2

SPECIFICATIONS

2.1 General

It may often happen that the specifications given to the servo designer are either incomplete, incompatible, or incomprehensible. Specifications fall into two main categories: (1) The specification of control-system dynamics and performance, and (2) the general specifications, such as power-supply variations and environmental conditions, which influence the dynamic characteristics and performance. We mentioned earlier that the designer usually must conduct a study to formalize his own detailed specifications. This study may immediately reveal contradictory specifications. It may indicate that relaxation of a particular requirement will greatly simplify the control system. It may indicate that one specification is so domineering that satisfaction of that one is tantamount to overall success.

A servomechanism is expected to perform any or all of the following functions:

(1) Bring about a change in the actual value of the output so that it conforms to a desired value at all times;

(2) Minimize the effect of varying component performance on the output;

(3) Minimize the effect of disturbances.

The performance specifications determine the degree of

excellence with which the servo must carry out the above functions. This means that the performance specifications must be given in terms of the desired output for a given input. It is, therefore, necessary to explore the type of inputs that might arise.

2.2 Inputs

The input signal may be one of the following types:

(a) Aperiodic, noise free

(b) Aperiodic, with noise

(c) Periodic, noise free

(d) Periodic, with noise

(e) Stochastic, noise free

(f) Stochastic, with noise

(Noise is regarded as any input-signal variation that is not a measure of the information carried by the input.)

Some typical specifications for the six types of input signals listed are as follows (see chapter 16 of reference 4):

(a) Aperiodic, noise free

(1) The system dynamic error shall neither exceed a specified maximum value, nor shall the steady state error exceed a specified maximum value.

(2) The integral-square system error shall not exceed a specified value.

(3) The system error, in addition, shall not exceed a specified amount in the presence of a specified disturbance that occurs at some specified point in the system.

(4) See paragraph 4.2 for other possible error specifications.

(b) Aperiodic, with noise

(1) Given that the first component of the system error is that for zero noise step input; the second component is the value of the output from noise alone; then the square root of the sum of the squares of the two components shall not exceed a specified value.

(Aperiodic signals commonly considered are steps, ramps, impulses, pulses, or an input expressed as a power series in time.)

(c) Periodic, noise free (only fundamental frequency present)

(1) The frequency response shall be characterized by a specified peak magnitude ratio (output/input) occurring at a specified frequency.

(2) The magnitude ratio (output/input) shall be within a band of some specified number of decibels over a specified frequency range, and phase shift (output/input) shall not exceed a specified amount over this same range.

(d) Periodic, with noise

(1) Error expressed in a similar manner to that of aperiodic, with noise, above.

(e) Stochastic, noise free (see paragraph 2.4 for definition.)

(1) System error shall not exceed a specified rms value when the input autocorrelation function has a given value.

(f) Stochastic, with noise

(1) System error shall not exceed a specified rms value when the input signal autocorrelation function, the input-noise autocorrelation function, and the signal-to-noise cross-correlation function are given.

2.3 Disturbances

It is also necessary to specify performance in response to a given load and/or disturbance occurring at points different from the input. The load or disturbance can also be classed as aperiodic, periodic, or stochastic. Typical specifications take on the same form as those above for inputs. A load specification, however, usually prescribes the amount of time allowed for the output to recover to within a specified deviation.

2.4 Stochastic Signals

A stochastic process is one in which there is an element of chance. Sometimes the input to a system is not completely predictable and cannot be described by a

mathematical function of time. A typical example of a stochastic process is a radar signal mixed with noise. Since the value of a stochastic signal cannot be determined with certainty at a given instant of time, probability density functions and other statistical characterizations such as the average value, the rms value, and the correlation function are used to describe the signal (see Appendix B). However, it is necessary to think of a stochastic signal as a member of a family of signals,[2] each generated by an identical process. Such a family of signals is called an ensemble and the statistical characterization (such as a correlation function) of the stochastic process is related to the ensemble rather than to a particular member of the ensemble. It, therefore, follows that the determination of the response of a system to a stochastic input does not yield a function of time, but rather a statistical characterication of the output signal ensemble (see Chapter 4).

2.5 Philosophy on Choice of Test Input

It may be necessary for the designer to prescribe his own inputs when analyzing the effects of noise, or load disturbances, or the effects of environmental conditions such as temperature, humidity, corrosion, etc. Even though the time domain characteristics are frequently specified in terms of the response ratio of the output to a step-function input, many others may be specified or implied. Furthermore, they may be extremely complicated—requiring

graphical description either in the time domain or as a power density spectrum. It might be well here to quote the feelings of some writers in the field. It is to be noted that they are not all in agreement. For example, the following quotation (page 308, reference 5) succinctly expresses one viewpoint:

"The characteristics of a particular servo should be determined by the actual input, the actual uncontrolled disturbances acting on the system and the actual output requirements. It is clearly not sufficient to assume the input to be a step in displacement or velocity, nor is it sufficient to require only that the transient response be well damped and that the velocity lag be small .... In general, the actual input and noise as well as the output requirements need a statistical description .... One can, of course, conceive of specialized servo problems, in which the input is a displacement step and in which the requirements are based on the transient response; but they are the exception rather than the rule." (words of R. S. Phillips).

On the other hand, quoting from the same reference (page 18): "The performance of a servo can also be specified in terms of its response to a step function. The procedure of experimentally and theoretically studying a servo through its response to a step-function input is extremely useful and is widely used for a number of reasons. The experimental techniques used in such testing are simple and require a minimum of instrumentation.

The characteristics of any truly linear system are, of course, completely summarized by its response to a step-function input. That is, if the step-function response is known, the response to any other arbitrary input signal, can be determined. It would be expected, therefore, and it is true, that with proper interpretation the step-function response is a powerful and useful criterion of overall system quality." (words of I. A. Getting).

But, progress and advances in technology must be recognized. Less than ten years later, Dr. T. C. Fry, Bell Telephone Lab, Inc.[6] said: ".... in any actual guidance or fire-control problem, we are not really concerned with the response of the system to some particular, ideally defined tactical input. We are concerned with its response to the whole gamut of possible tactical situations, including all the possible variability in the enemy path and all the possible errors which may be produced by random perturbations in the input data or in the mechanism itself. Obviously, this adds elements of information theory, statistical theory or whatnot and greatly increases the level of (essentially mathematical) insight required for effective work."

The writers seem to be hitting at the same old saw of the tug-of-war between abstract mathematics and engineering approximations. To easily and practically accomplish a design by hand, on paper, requires simple inputs. The use of the step input, for example, does take

into consideration the the whole frequency range. But, as control systems become more and more sophisticated, it becomes necessary to consider more sophisticated inputs. This fortunately becomes feasible in the modern day with the use of computing machinery.

2.6 Static Characteristics

The static performance specifications describe the steady state value of the system output. Although the statement of them is a simple matter, their importance lies in the fact that they sometimes immediately establish the type[3] system which must be used. For example, if a system is to have no steady state error in response to a step input, it must be at least Type 1 (meaning one pure integration must exist in the open-loop transfer function); if no steady state error is to exist in response to a ramp input, the system must be at least Type 2. (Two pure integrations must exist in the open-loop transfer functions.)

Normally an acceptable following error must be stated for the response to a ramp input. Some maximum allowable output is usually stated for the response to a disturbing input.

2.7 Dynamic Characteristics

The desired dynamic characteristics may be specified in terms of transient or frequency response. (If defined in both domains, their compatibility must be determined.) The three main themes of engineering design might be tabulated as follows:

TABLE I

SPEED OF RESPONSE		RELATIVE STABILITY	
Transient Domain	Frequency Domain	Transient Domain	Frequency Domain
a) time constant	a) bandwidth	a) % of first overshoot	a) Peak value of closed loop output-input amplitude ratio
b) rise time	b) cutoff frequency	b) number of overshoots	b) gain and phase margin of open-loop frequency characteristic
c) settling time	c) frequency of peak overshoot		

ACCURACY
Allowable error in terms of % or per unit of controlled variable, stated as:
a) Maximum
b) Average
c) rms value

2.8 Required Specifications

The three main themes of design specifications are essentially those of (1) speed of response, (2) stability, and (3) accuracy. The Trial and Error Design Methods specification requirements are:

(1) Input Signal

(2) Desired Output

(3) Disturbances and special inputs

(4) Allowable error

(5) Plant elements (fixed)

(6) Relative stability

The Analytical Design Methods specification requirements are:

(1) Input signal

(2) Desired Output

(3) Disturbances and special inputs

(4) Performance index and required value of same

(5) Degree of Freedom allowed in compensation

Notice that the specifications for the two methods have two basic differences:

(a) The Analytical Design Method calls for a performance index vice allowable error and relative stability specifications. A performance index is simply a single number which is used as an indirect measure of system performance. Its use is an attempt to replace the functional description of the performance of a system through its response parameters (such as peak overshoot,

rise time, etc.) with a numerical description that rates the system performance with a single number.

(b) The Analytical Design Method calls for the degree of freedom allowed in compensation vice relative stability and plant elements specifications. This specification is not strictly necessary, but practically it is. With constraints imposed by degrees of freedom, it is possible to categorize large portions of the computational mathematics for all time, and make use of only the results of this categorization in a particular problem. More will be said about these differences in the next two chapters.

CHAPTER 3

BASIC TRIAL AND ERROR DESIGN METHODS

3.1 General

Almost all Trial and Error Design Methods depend upon the ability to express the output of a component or a servomechanism with respect to an input in terms of differential equations. These equations are almost universally transformed into algebraic equations by the Laplace transform. After this transformation is made, components can be collected into one overall mathematical model by well known, easily applied methods, and analysis and synthesis of the system is made directly in terms of these equations.

Most methods utilize and depend upon the open-loop transfer function or equation, the closed-loop transfer function, or a combination of the two functions. Test inputs with simple Laplace transforms are applied to these equations, the output is compared to the input, and various parameters are measured to determine the acceptability of the mathematical model of the servomechanism. Then the necessary hardware is determined to match the model, or the model is changed and another comparison or analysis is made.

Specifications and design parameters which must be satisfied by servomechanism adjustment are expressed in terms of either transient response to a nonperiodic test

input or frequency response to a periodic test input (almost universally a sinusoidal function).

Several design methods work with the transfer functions directly in the frequency domain, taking advantage of the fact that the analytical or graphical results will be directly related to the frequency response parameters such as bandwidth or amplitude of response at resonant frequency.

In the analysis of a servomechanism, most methods treat the servomechanism as a modification of a system with a second order differential equation. The output of any second order system is exactly known for any of the useful test inputs and the family of curves is not difficult to reproduce. This is done to simplify the description of the output with a given input to one that can be easily formulated. It is found in practice that the above treatment is almost always a useful approximation if it is but remembered that it is an approximation.

3.2 The Mathematical Model and Graphs of Response

An early task of the designer is to assemble his likeliest components, or the given components, into a representative mathematical model. The whole process of design depends upon the validity of the model, being only as accurate as the model which is used to describe the assembly of hardware. The transfer functions of the components can be gotten from experimental tests made upon them or from the differential equation of the physics of

the component. Even when the form of the differential equation is known, experimentation is often required to obtain the values of the parameters. In this paper, analysis is described as the determination of the output/input characteristics from the transfer function, but in the determination of the equations for a component all of the analysis theory and methods can be used in reverse.

There are graphical displays which are used to portray either the mathematical model or the response of the system to a given input. Some of the more useful ones in analysis and synthesis are:

(a) Frequency Response Graphs-- amplitude of the output/-input ratio and phase shift vs. frequency.

(b) Bode Diagram or Attenuation Diagram--log amplitude of the open loop output/input ratio and phase shift vs. log frequency.

(c) Nyquist Diagram- a polar plot of amplitude of the open loop output/input ratio and phase shift with frequency as a parameter.

(d) Inverse Complex Plane Diagram--a polar plot of the inverse of the amplitude of the open loop output/input ratio and phase shift with frequency as a parameter.

(e) Nichols Chart- a log amplitude of the open loop output/input ratio vs. phase angle with frequency as a parameter. Log amplitude of the closed loop (unity feedback) vs. phase shift is overlaid on

the chart.

(f) Root Locus Diagram-- the complex plane upon which the poles and zeros of the mathematical model are plotted.

3.3 Specifications

The operation of a servomechanism can be described by its specifications. These specifications are defined in such a way as to describe the speed of response, the stability, and the accuracy of the servomechanism. The specifications used in Trial and Error Design Methods all stem from the ratio of the output of the servomechanism to the input for a given type of test input.

There are three major types of specifications. There are those specifications which are determined from the open-loop frequency response; others which are determined from the closed-loop frequency response; and a third type which are determined from the transient response to a specified input, usually a unit step function.

Theoretically the closed-loop response to a step input and the closed-loop frequency response can be shown to be equivalent[4]. In practice, the correlation between the two is usually quite remote, and the conversion from one response to the other involves graphical methods of integration with many repeated calculations. There is direct correspondence between the open-loop and the

closed-loop frequency response and this relationship is not hard to determine.

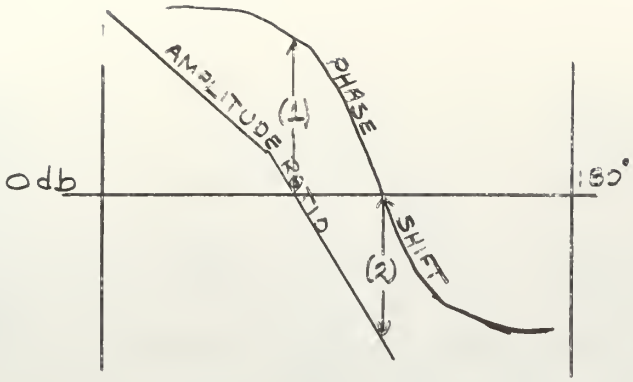
There are a great many different measurements used today in servomechanism design to specify the character of the servo. Most of these measurements are interrelated, and many of them are either synonymous or at least they describe the same type of output.

The following list of specifications is intended to be merely a sample of those more frequently used.[1-16] Many of them require amplifying modifiers not to be ambiguous.

(a) Open-Loop Frequency Response Specifications

(1) Phase Margin is 180° minus the angular difference in phase between the output and the input at the highest frequency where the output is of the same amplitude as the input. It is a measure of the relative stability of the system, being unstable at non-positive values.

(2) Gain Margin is the ratio of the amplitude of the input to the amplitude of the output at that frequency where the phase shift between the input and the output is 180°. It is usually expressed in decibels and is a measure of relative stability, being unstable for non-positive values of decibels.



(1) Phase Margin

(2) Gain Margin

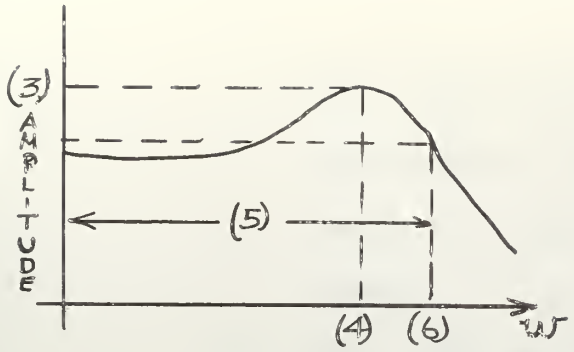
(b) Closed-Loop Frequency Response Specifications

(3) Peak Amplitude Ratio (Mp) is the maximum ratio of output to input. It is a measure of the relative stability of the system.

(4) Frequency of Peak Amplitude Ratio is a measure of the speed of response of the system.

(5) Bandwidth is one of the specification parameters which has no standardized definition. It is a measure of the frequency range at which the amplitude falls within specified limits. A popular limit is that the amplitude ratio be between + 3db and -3db. It is a measure of speed of response.

(6) Cutoff Frequency is much like bandwidth. It is the upper frequency at which the amplitude ratio reaches some specified value. Common ones are 0 db, -3db, -6db, or -20db.



(3) Peak Amplitude Ratio

(4) Frequency of Peak Amplitude Ratio

(5) Bandwidth

(6) Cutoff Frequency

(7&8) Damping Ratio (ζ) and Undamped Resonant Frequency (ω_n) are specification factors in that they completely specify the response of a linear second order servomechanism. All of the other specifications listed are fixed by (ζ) and (ω_n) for a second order system. The vast majority of servomechanisms have a dominant factor which is second order in character. Consequently (ζ) and (ω_n) specify the type of response of a higher order system to a more or less close degree.

(c) Transient Response Specifications to a Unit Step Input

(9) Peak Overshoot (M_{p_t}) is the maximum amplitude of the first overshoot measured from the final steady state output. It is a measure of relative stability; the more the overshoot, the less stable the system. A system with no overshoot is said to be "overdamped" while one with an overshoot is said to be "underdamped".

(10) Time of Peak Overshoot (t_p) is the time from

step input until the time of the maximum amplitude of the first overshoot. It is a measure of the speed of response of the system.

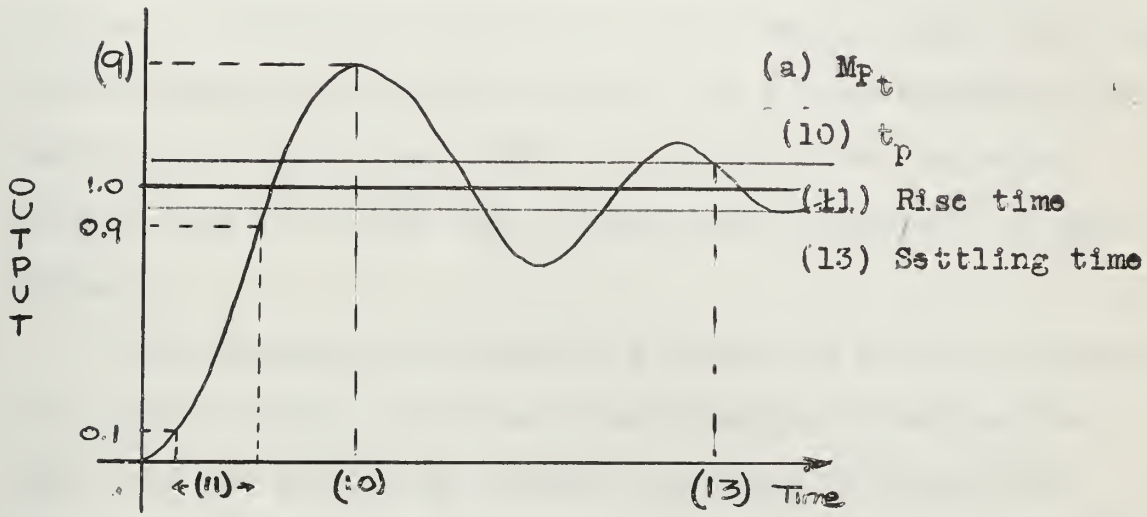
(11) Rise Time is the time for the output of the system to cancel a certain portion of the error. A commonly accepted Rise Time is the time for the system to move from 10% of the final value to 90% of the final value. It is a measure of the speed of response.

(12) Characteristic time (τ_c or $\zeta\omega_n$) is the logarithmic decrement of the dominant portion of the response. It is a measure of speed of response.

(13) Settling time is that time required for the servomechanism to reduce the error below and remain less than a certain percentage of the input step. Commonly accepted percentages are 5, 2, and 1 which make the settling time become approximately 3, 4, and 5 times the characteristic time. It is a measure of the speed of response of the system.

(14) Number of Overshoots is the number of overshoots of the output during the settling time. It is a measure of the relative stability of the system.

(15) The Error Coefficient is a measure of the steady state error of the system. It is dependent upon the type of the system and the gain of the system.



3.4 Analysis

A very important part of the synthesis of a servomechanism is the analysis of the system under test to see if it does or does not meet the required specifications. As we have seen, these specifications are invariably an expression of the comparison of one form of output-to-input relationship or another. The more quickly one is able to determine this relationship, the shorter the job of synthesis will be. Many of the analysis methods available are in reality shorthand techniques at approximating the output-to-input relationships.

Remember that linear servomechanisms have output-to-input relationships which are expressible as linear differential equations. The type of input to which these systems are subjected are the boundary conditions for the differential equations. Since it has been found that the equations can be handled or solved more readily in the Laplace transformed condition, mainly because the operators can be manipulated algebraically, the system equations are normally in this form. This is known as the frequency domain, and it is the one in which most synthesis is performed.

The frequency response of a system is easily obtained by simply solving the closed-loop transfer function for $j\omega$. If the open-loop transfer function is known and available, rather than the closed-loop transfer function, one can obtain the frequency response directly without first solving for the closed-loop transfer function. This is done by plotting the frequency response of the open-loop system and then transforming the coordinates. If there happens to be a feedback function other than unity, the transformation requires one more step, but it is still useful.

Unfortunately many of the specifications are prescribed in the time domain using the transient response to an aperiodic input. This situation has come about largely because it is easier for the specifications writer, and indeed for anyone, to visualize the effect upon, and the

action of, the servomechanism in that plane. So there is a requirement to know the response of the system to a transient input. There are several avenues of approach available to determine the transient response.

The most obvious approach, but unfortunately the one which usually entails the most labor, is actually to solve the differential equation with the prescribed boundary conditions. Normally this amounts to taking the Inverse Laplace transform of the closed-loop transfer function multiplied by the Laplace transform of the input. A modification of this technique is to get the approximate output response by solving only for the dominant features of response, ignoring the less important features. Note that some prior experience is helpful in determining what is a less important feature.

A second avenue of approach is to determine first the frequency response and then plot the time response from the frequency response. It has been shown that the two responses are uniquely equivalent. It is extremely unfortunate that the equivalency is too obscure to be seen immediately by the average eye--or for that matter by many a trained eye. The transformation from the frequency response to the transient response to an impulse or step involves repeated graphical integrations, and without a digital computer can be a tedious operation.

A third, and the most frequently used approach is to

assume that the system is like a second order system and that the relationship among its transfer function, its frequency response, and its transient response are closely related to a second order system. The great majority of servomechanisms are dominated by one pair of complex conjugate roots possibly with a single dipole near the origin. If the system is assumed to act like a second order one with only the complex conjugate roots, full advantage can be taken of the known relationships of second order systems with respect to transfer functions, frequency response characteristic, and transient response characteristics. The results can be appropriately modified if there is a dipole near the origin.

The frequency response can be determined from the transient response. This is of value when a physical component is available for transient response testing, and the transfer function is unknown.

In summary, it is seen that the loop transfer function, the frequency response, and the transient response of a servomechanism each uniquely defines its characteristics. Analysis is then the method of jumping from one of these four descriptions of the system to another.

3.5 Synthesis

The determination of a transfer function or a set of components which will comply to given specifications is synthesis. Often our choice of components is either limited or partially specified. This restricts our ability

to determine the transfer function or the remaining components.

In synthesis as well as in analysis, heavy reliance on second order (or possibly 3rd order) approximations is required in many of the present methods, expecially if the specifications are given in terms of the transient response.

An analysis of the specifications will quickly show what type of closed-loop transfer function is required in order to remain within the specifications. The real trick of synthesis is to translate this closed-loop transfer function, either real or implied, into an open-loop function which can then be stated in terms of components.

This transformation is not a unique one and, therefore, there is an infinity of combinations of components which will fall within the specifications. Unfortunately there is a much greater infinity of combinations which will not fall within the specifications.

The designer is allowed a considerable amount of leeway in the selection of a transfer function or of the components he must use. This leeway is allowed partly because the specifications usually require the system to fall within certain fairly broad limits and partly because the change in specifications due to a change in servomechanism parameters is usually not a sharp change. Since, to improve upon one specification, the servomechanism must often be allowed to relax another specification, and since it is unlikely that the specifications can be described as

optimum to begin with, the design of servomechanisms is as much an art as a science.

The most usual case is the one in which the whole system, except for a compensator or two, is already chosen, and it is only a matter of deciding what the compensator will be. Even the selection of the place to put the compensation may be restricted due to the limited physical accessability of the signal flow.

3.6 Compensation

There are two major methods of compensating a servomechanism. One is to place the compensator in the forward path (called cascade compensation), and the other is to place the compensator in the feedback path (called feedback compensation).

Cascade compensation is the most widely used of the two because it generally requires simpler and less expensive elements. It is easier to synthesize a system using cascade compensation because the relationship between open-loop and closed-loop frequency responses and transfer functions is more direct. Drift of the parameters of the active elements in the forward path disturbs the effect of cascade compensation more seriously than it would disturb feedback compensation. Cascade compensation elements are normally placed in the forward path at signal power levels rather than at output power levels. This allows the use of smaller network elements since little power must be

dissipated. The two types of cascade compensation that are generally used are lead compensators and lag compensators.

Lead compensators act as a derivative path in parallel with a direct path. The derivative path tends to increase the output to input ratio at high frequencies. At these same frequencies it decreases the amount of phase shift that would occur in the uncompensated system. The importance of the lead compensator is that it tends to cause the 180° phase shift to occur at a higher frequency providing a larger degree of phase margin than would otherwise be available. The main objection to lead compensation is that it causes a great attenuation of the forward signal, requiring a substantial increase in required gain in the power element in order to maintain the same steady state accuracy as before compensation. Also lead compensators tend to increase the bandwidth of the servomechanism making it more susceptible to high frequency noise. The transfer function of the compensator is of the form:

$$G_C(s) = \frac{1}{\alpha} \frac{\alpha\tau s + 1}{\tau s + 1} \qquad 1 < \alpha < 20$$

Lag compensators act as an integrating path in parallel with a direct path. The integral path tends to increase the output to input ratio at low frequencies while the phase shift is increased at the same frequencies. The lag compensator is used to increase the steady state accuracy of the uncompensated system. The time constant

of the pole of the lag compensator must be made large enough that the accompanying phase shift does not impair the stability of the system, and yet it cannot be too large or the transient error will tail-off too slowly. The transfer function of the compensator is of the form:

$$G_c(s) = \frac{\tau s + 1}{\beta \tau s + 1} \qquad 1 < \beta < 20$$

Lead-lag compensators complement the virtues of the lead compensator with those of the lag compensator. The relative stability is increased with the lead compensator while the steady state accuracy is maintained, or at least the attenuation of the lead compensator is offset, with the lag compensator.

Combinations of lead and lag compensators can be used to gain further benefits from these compensators. It must be remembered that isolation must exist between adjacent compensators for the transfer functions to be correct when multiplied. This isolation can be achieved with a buffer amplifier or cathode follower.

Feedback compensation, while more complex and expensive in components, allows a greater flexability than cascade compensation. At reasonably large gains, it tends to nullify the effect of the forward components around which it is placed, making the whole appear like the inverse of the feedback function. Thus it is an ideal method of replacing an undesirable function with a desirable one. A forward component with a shifting or uncertain gain or parameters in the transfer function can be made quite

rigid by the proper use of feedback compensation.

Tachometer feedback compensation is a very common method of dampening a servomechanism to make it more stable or to reduce the transient overshoot. The steady state accuracy suffers with tachometer feedback, but this may be corrected by placing a filter in cascade with the tachometer in the feedback loop.

3.7 Summary

The Trial and Error Design Method might be summarized as follows:

(a) The given specifications are: input signal, desired output, disturbances, allowable error, plant elements, and degree of stability required.

(b) On the basis of experience or preliminary approximations, select a form of compensation.

(c) Establish the parameters for the compensation exclusive of system gain.

(d) Adjust the system gain in accordance with the stability requirements.

(e) Analyze system to see if the error is satisfactory.

(f) If the error exceeds allowable limits, repeat the process, using different compensation; continue until error specifications are met.

Although the theoretical design objective is the attainment of the "best possible" servo, a more mundane design objective is the attainment of an "adequate" servo.

CHAPTER 4

BASIC ANALYTICAL DESIGN METHODS

4.1 General

The authors received almost all their training and schooling in servos using Trial and Error Design Methods. The particular problem that motivated them for this type of thesis was the fact that they were unable to recognize an inconsistent set of specifications. Background training was more than ample to select some method and to forge ahead to a design; if it failed to satisfy the specification, to start over; to keep starting over until a solution was found or patience was exhausted, never really knowing whether a solution existed, or in the case where a solution was found, whether it was the best solution.

Analytical Design Methods purport to solve this dilemma. And this is true in the sense that if it is assumed that a performance index is able to incorporate all the specifications into it, the method, being a pure mathematical minimizing or maximizing process, is able to immediately reveal, solely by the mathematics, whether a stated performance index can be achieved.

Recall that the three main themes in specifications are speed of response, stability and accuracy; it seems reasonable to believe that some performance index that includes error, time and an appropriate weighting function for the error, can represent the specifications, or at least most of them.

In selecting a performance index, we need an answer to the question of what kind of output is desirable for the servo. If there were no uncontrolled disturbances (or noise) our goal would be to make the output follow the input perfectly. But, generally, in the presence of disturbances, if the servo follows the input perfectly, it will also do a good job of following the noise. To establish a figure of merit or performance index, then, requires a compromise. The performance index must be practicable, not too difficult to apply, and of general applicability. It should also be a measure of the average behavior of the servo, rather than be affected by short-lived deviations from the mean, or shifts in the time axis. Quite a number of performance indices are in use or have been proposed, some of which are discussed below.

4.2 Performance Indices for Transient Signals

(a) Performance Indices which are not time weighted:

(1) $P_1 = \int_0^\infty e(t)\,dt$ (called Control Area)

(2) $P_2 = \int_0^\infty |e(t)|\,dt$ (called Integral Absolute Error - IAE)

(3) $P_3 = \int_0^\infty e^2(t)\,dt$ (called Integral-Square Error - ISE)

Two of the above indices (P_1 and P_3) can be used in purely analytical procedures whereas P_2 contains discontinuities. P_1 was proposed by T. M. Stout[7], P_2 by Fickeison & Stout[8], and P_3 by Hall[9] and Sartorius[10]. P_3 is sometimes called the Hall-Sartorius criterion. It can be seen that all of these indices heavily weight the error at the beginning of

the transient. Even though they favor rapid speed of response, they also tend to cause adjustments that are, in general, less damped than is usually required. From this viewpoint an IAE index is better than an ISE index. IAE is, however, normally treated with the use of analog computers. Note that none of these indices, though, place a penalty on errors occurring late in the transient, and thus, there is no good assurance that the servo does not have a long tail off. To provide for this, and at the same time, recognizing that no physical system can respond instantaneously, another set of performance indices are time weighted.

(b) Performance Indices which are time weighted:

(4) $P_4 = \int_0^\infty t\, e(t)\, dt$ (called Time Weighted Control Area)

(5) $P_5 = \int_0^\infty t\, |e(t)|\, dt$ (called Integral-Time-Multiplied Absolute Error- ITAE)

(6) $P_6 = \int_0^\infty t\, e^2(t)\, dt$ (called Time Weighted ISE)

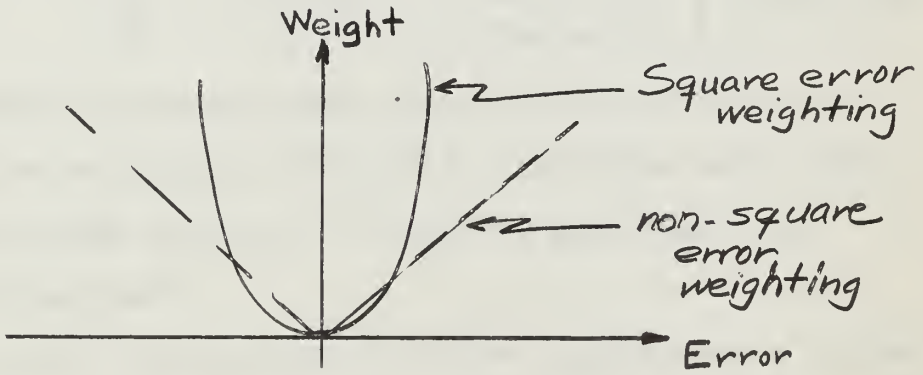
P_4 was proposed by Nims[11] and P_5 by Graham and Lathrop[12]. Again, P_4 and P_6 can be treated in purely analytical procedures, whereas P_5 is usually treated with an analog computer. P_5 can also be treated with a set of what are called standard forms of which the foremost proponents are Graham and Lathrop. It is discussed further in Chapter 6. The major objection to these indices is that they tend to tip the scales the other way: they put too

much weight on the latter portion of the transient. They might be more practical, if they ignored error after it reached a certain limit. Unfortunately such an expression would again not be analytical. However, P_5 is quite popular in the literature for computer use.

(c) Another index proposed by Nims[11] is:

$$(7)\quad P_7 = \left[\int_0^\infty t\, e(t)\, dt\right] \Big/ \left[\int_0^\infty e(t)\, dt\right]^2$$

However, we can find no other literature, for or against it. It can be used where the objective is to make P_1 and P_4 some value other than zero. P_3 (ISE) is by far the most popular index for purely analytical methods. One reason for this is that error becomes more undesirable as it increases in magnitude. As the below figure shows, weighting the square of the error rapidly penalizes large error.



However, probably the greatest reason for the wide usage of ISE (and Mean Square Error discussed in the next paragraph) is mathematical convenience. There is a highly polished body of mathematical knowledge that has been developed particularly around the idea of a mean square value.

4.3 Performance Indices for Stochastic Signals

It was indicated earlier that usual specifications concern the response of the system to typical inputs, whereas systems are actually subjected to random inputs. This is one of the basic reasons for interest in statistical criteria. By far the most commonly used and described index is the root-mean-square criterion, first proposed by A. C. Hall[9] and developed by James, Nichols and Phillips[5]. It is characterized by its mean square value, thusly:

$$P_s = \overline{e^2(t)} = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} e^2(t)\,dt \tag{8}$$

Another criterion suggested by Oldenburg and Sartorius[13] is characterized by the following integral;

$$P_s = \overline{|e(t)|} = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} |e(t)|\,dt \tag{9}$$

The use of the root-mean-square criterion was inspired by the writings of Wiener[14] and it has been highly developed by many writers, of whom two good ones are Newton[2] and Laning[15].

Recall that determination of the response of a system to a stochastic input does not yield a function of time, but rather a statistical characterization of the output signal ensemble. Stochastic signals consist of two classes:

(1) The process is stationary if the statistical behavior of the process that generates the ensemble is independent of time.

(2) The process is non-stationary if the statistical behavior varies with time.

Many books are written on the fact that if a stochastic process is stationary, it is possible to use a single member of the ensemble of the process to determine the process statistics. (Various books also justify the assumption that stationary signals are valid for use in servo work. No one has successfully "cracked the barrier" of non-stationary signals yet, except for special cases.) Furthermore, the average value of a signal from a stationary process can be found either by taking an ensemble average at a particular time or by taking the time average of a single member of the ensemble. Because of this (see Appendix B), the mean-square value of a stationary signal is precisely equal to its auto-correlation function, evaluated with the argument equal to zero. If P_8 is then selected as the performance index, an incredibly powerful tool results.

The rms value of a random quantity can often be calculated practically under conditions in which many of its other statistical properties cannot. For example, if a random quantity has a normal probability distribution, all statistical properties are determined from its rms and mean value.

It should be stressed, though, that the rms value of the error does not characterize the error completely, and use of this criteria may result in overlooking some important aspects of the problem. $\overline{e^2(t)}$ is independent of the

distribution of e(t) in the frequency spectrum. Since a system usually operates with other systems connected to it, the other systems will transmit e(t) in a manner that depends on their dominating frequency and on the spectrum of e(t). Gille[16] indicates as an example the stabilization of an airplane carrying passengers. Resonant frequencies of the order of magnitude of the duration of a human pace should be avoided because such frequencies are the most adverse to comfort. It is just as important to consider the frequency spectrum of the error as its rms value. Nevertheless, there is widespread use of P_8 in design analysis. This seems to be due less to its intrinsic worth as a criterion than to the convenience attached to its calculation.

4.3 Configurations

By establishing standard configurations, much of the mathematics can be worked out for one time and only the results are then necessary for design. The configurations used are:

(a) Fixed Configurations

In this configuration, all the physical elements, including the compensation, are essentially fixed. Optimization procedure then consists of the adjustment of a few free parameters (such as gain, or one time constant of a filter, or perhaps the ratio of a pair of time constants of a filter) so as to minimize the performance index. This technique is frequently used in conjunction with the Trial and Error Methods where the form of the system has been

fixed by other considerations and only the best numerical values of the free parameters are sought. On the surface it appears easy to use since the only requirement is to differentiate the performance index with respect to the free parameters and to set the partial derivatives equal to zero. This technique is also easy to apply in connection with the analog computer.

(b) Free Configuration

In this configuration, quite a high degree of mathematical sophistication is called for, because here the entire transfer function (or weighting function, if in the time domain) of the system is allowed to vary in minimizing the performance index. This can only be done by the use of the calculus of variations. This is one of Wiener's big contributions, producing an implicit transfer function in the form of an integral equation known as the Wiener-Hopf equation. Solution of this equation is extremely difficult; Wiener conceived a process for solution called "spectrum factorization". Although this kind of procedure is very satisfying intellectually, it, in general, leads to a transfer function which is physically unrealizable.

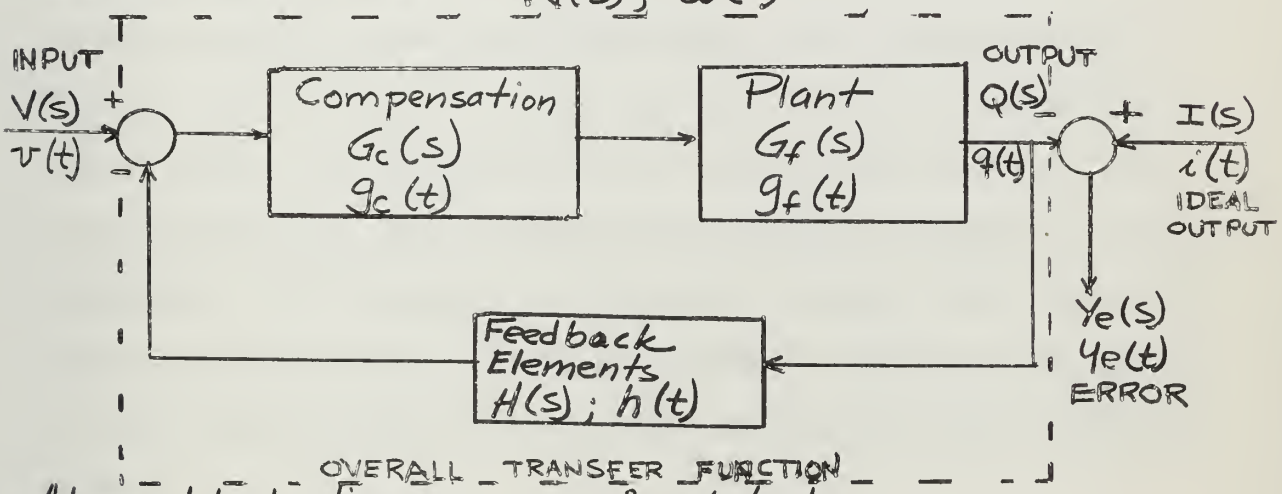
(c) Semi-Free Configuration

This can be thought of as the usual type design problem, and is the type that one first thinks of in the Trial and Error Design vernacular: All the plant elements are specified as fixed elements, but there is no constraint

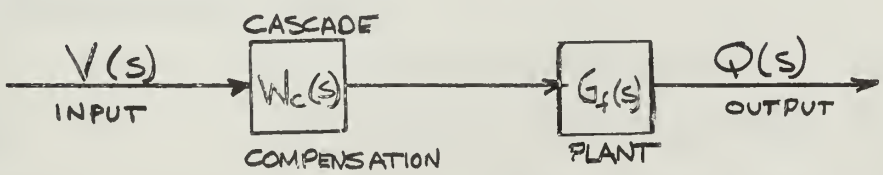
on the type of compensation allowed. This technique again frequently leads to a compensation which is physically unrealizable, but it has a very real practical advantage: it provides a goal to aim for in compensating by Trial and Error Methods. In providing the goal, it frequently indicates the direction one must take in the selection of a compensator type.

4.4 Rudiments of Analytical Design

All of the methods use the overall or closed-loop transfer function. Adopt the following symbology (taken from Newton[2]) (Same as that on symbols page, but repeated here for convenience):



Above block diagram manipulated:



$$W_c(s) = \frac{G_c(s)}{1 + G_c(s)\,G_f(s)\,H(s)} \; ; \quad G_c(s) = \frac{W_c(s)}{1 - G_f(s)\,W_c(s)\,H(s)}$$

Because the closed loop function is used, it sometimes is not too easy to manipulate back to the original block diagram. However, this disadvantage (if, in fact, it is one) is offset by the fact that these methods do not involve solving for the roots of the characteristic equation, unless they are desired in the final completed paper design. This is a very real advantage if the characteristic equation gets above an order of about five.

Practically speaking, the application of these methods can lead to failure if the limitations of the mathematical model are not continuously kept in mind. They almost inevitably lead to non-linear operation because the performance index demands minimum error. The minimization of dynamic error calls for higher gain, leading to saturation; and it calls for cancellation, or near cancellation of the plant elements by the compensator elements, leading to wide bandwidth. (For example, design of a second order system by these techniques will invariably lead to infinite gain which is not only physically impossible, but practically, undesirable.)

(a) Transient Input

Let us first consider procedures for a transient input. Use of the ISE criterion has been highly developed by Newton & Gould[2,4]; use of ITAE criterion has been developed by Graham & Lathrop[12] (discussed in Chapter 6). The ISE criterion is represented as:

$$I_y = \int_{-\infty}^{\infty} y e^2(t)\, dt$$

By Parseval's Theorem, the integral can be rewritten as:

$$I_y = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} Y_e(s)\, Y_e(-s)\, ds$$

where $Y_e(s)$ is the Fourier transform of $y_e(t)$. If $Y_e(s)$ is a rational function, I_y can be written in the form:

$$I_y = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{C(s)C(-s)}{D(s)\,D(-s)}\, ds$$

where C(s) and D(s) are polynomials in s. Definite integrals of this form have been evaluated and tabulated in terms of the coefficients of the polynomials (for example, see Appendix E of reference 2). Evaluation of the integral then gives:

$$I_y = I_y(P_1, P_2 \cdots P_k)$$

where P represents the free parameter.

I_y can be minimized by solving the set of K equations of the form:

$$\frac{\partial I_y}{\partial P_k} = 0$$

Except for the simplest forms, solution of this set of equations can be difficult. Many times, the easiest approach is simply to plot I_y versus P_k, holding other parameters constant; the minimum value can be determined closely enough for engineering purposes. An important condition for use of the developed integral tables is that D(s) must have all its zeros (roots) in the left half plane. This is also the condition for a stable system. Therefore, the parameters must not be allowed to take on such values as to cause D(s) to have zeros in the right hand plane. Otherwise the integral table is invalid and, besides, the system is unstable.

Note that another limitation exists above: Ye(s) must be a rational function. (It is not meant to cite this as a disadvantage compared to Trial and Error Methods, since those methods, in general, require rational functions also.) See Chapter 6, under Newton's Method, for a way around this limitation.

The above scheme can only be used with fixed configuration systems. For semi-free configurations with transient inputs, Newton uses translation functions. This is also postponed to Chapter 6.

(b) Stochastic Input

Let v(t) and q(t) represent stochastic input and output respectively of a linear system whose weighting function is w(t). From the convolution integral we know that

$$q(t) = \int_{-\infty}^{\infty} w(t_1)\, v(t - t_1)\, dt$$

$$q(t+\tau) = \int_{-\infty}^{\infty} w(t_2)\, v(t+\tau - t_2)\, dt_2$$

From the definition of autocorrelation functions (Appendix B), and from manipulating the above integrals, the following relation is obtained:

$$\varphi_{qq}(\tau) = \int_{-\infty}^{\infty} w(t_1)\, dt_1 \int_{-\infty}^{\infty} w(t_2)\, \varphi_{vv}(\tau + t_1 - t_2)\, dt_2$$

which expresses the autocorrelation function of the output in terms of the autocorrelation function of the input and the weighting function.

In a similar manner:

$$\varphi_{vq}(\tau) = \int_{-\infty}^{\infty} w(t_2)\, \varphi_{vv}(\tau - t_2)\, dt_2$$

which gives cross-correlation between the input and the output in terms of autocorrelation of the input and the weighting function. If the correlation functions and the weighting functions are Fourier transformable, then

$$\Phi_{qq}(s) = W(-s)\,W(s)\,\Phi_{vv}(s)$$

$$\Phi_{vq}(s) = W(s)\,\Phi_{vv}(s)$$

The latter expression is frequently used to evaluate W(s), since it is the ratio of the cross-power-density spectrum to input-power-density spectrum.

From the block diagram, $y_e(t) = i(t) - q(t)$. The mean-square-error is identically the value of the autocorrelation function of error, with $\tau = 0$:

$$\overline{y_e^2(t)} = \varphi_{yy}(0)$$

After suitable manipulation, the power density spectrum is obtained as:

$$\Phi_{yy}(s) = \Phi_{ii}(s) - W(s)\Phi_{vi}(-s) - W(-s)\Phi_{vi}(s) + W(-s)W(s)\Phi_{vv}(s)$$

This equation is suitable for many variations. For example, if $v(t) = v_d(t) + v_n(t)$ where $v_d(t)$ represents the data or signal component and $v_n(t)$ the noise component, and we assume them to be uncorrelated, then:

$$\Phi_{yy}(s) = \left[1 - W(s)\right]\left[1 - W(-s)\right]\Phi_{dd}(s) + W(s)W(-s)\Phi_{nn}(s)$$

(Obtained by letting $\Phi_{vv}(s) = \Phi_{dd}(s) + \Phi_{nn}(s)$ and $\Phi_{dd}(s) = \Phi_{ii}(s) = \Phi_{vi}(s)$ in the previous equations.)

Now, the problem, similar to that described for transient input is to find the area beneath the error-power-density

spectrum along the imaginary axis and then minimize this error. This can be done by evaluating and minimizing the following integral:

$$\overline{y_e^2(t)} = \varphi_{44}(0) = \frac{1}{j}\int_{-j\infty}^{j\infty} \varphi_{44}(s)\,ds$$

Of course, the complexity of the mathematics involved has increased greatly over that for the fixed configuration transient signal. Although the error may again be minimized by setting the partial derivatives equal to zero, it is usually best to proceed with a series of plots of error versus the free parameters.

(c) Stochastic Input, Free and Semi-free Configuration:

Newton shows that the actual output can be eliminated from the error expression, and the error then is a function only of the input and the desired output as follows:

$$\overline{y_e^2(t)} = \varphi_{ii}(0) - 2\int_{-\infty}^{\infty} w(t_1)\varphi_{vi}(t_1)\,dt_1 + \int_{-\infty}^{\infty} w(t_1)\,dt_1 \int_{-\infty}^{\infty} w(t_2)\varphi_{vv}(t_1 - t_2)\,dt_2$$

He then proceeds to prove that minimization of the error for a physically realizable weighting function requires satisfaction of this expression:

$$\int_{-\infty}^{\infty} w_m(t_2)\varphi_{vv}(t_1 - t_2)\,dt_2 - \varphi_{vi}(t_1) = 0 \text{ for } t_1 \geq 0$$

This is the Wiener-Hopf integral equation in the time domain, of which $w_m(t)$ is the implicit solution, where $w_m(t)$ represents that weighting function that minimizes $\overline{y_e^2(t)}$.

It is an understatement to say that solution of this equation is difficult. There are a few rare situations where its solution in the time domain is self-evident.

Newton (Chapter 5, reference 2) has a very good exposé on the explicit solution to this equation in the frequency domain. Suffice to say that its solution consists of Wiener's "spectrum factorization" technique whereby a function is split in such a manner that one portion consists of that component whose poles lie in the LHP and the other portion, that which lies in RHP. By discarding the RHP portions, the explicit solution of $W_m(s)$ gives a guaranteed stable transfer function. This solution in the frequency domain is much easier if the functions are Fourier transformable. Some authors have developed a solution whereby the whole derivation for optimum transfer function is conducted in the frequency domain.

Since solution for a free configuration is somewhat academic, the equation has been modified to provide for the semi-free configuration. We quote that solution here:

$$W_{cm}(s) = \frac{\left\{ \dfrac{G_f(-s)\,\Phi_{vi}(s)}{\left[G_f(-s)\,G_f(s)\right]^{-}\,\Phi_{vv}^{-}(s)} \right\}_{+}}{\left[G_f(-s)\,G_f(s)\right]^{+}\,\Phi_{vv}^{+}(s)}$$

Although this expression is quite ominous at first glance, a breakdown of the symbols helps:

$W_{cm}(s)$ = optimum cascade compensation for minimizing $\overline{\epsilon^2(t)}$

G_f = fixed or plant elements

$\Delta^+(s)$ = any factor of $\Delta(s)$ which includes all the poles and zeros of $\Delta(s)$ in the LHP.

$\gamma(s)_+$ = component of $\gamma(s)$ which has all its poles in LHP such that $\gamma(s) - \gamma(s)_+$ has all its poles in the RHP.

$\Delta^-(s) = \Delta(s)/\Delta^+(s)$ = remaining factor of $\Delta(s)$ which includes poles and zeros of $\Delta(s)$ in RHP.

So, if the correlation functions are known, then solution of the above equation for $W_{cm}(s)$ becomes a mathematical factoring problem, where stability is guaranteed.

If the fixed elements are minimum phase functions the above expression reduces to:

$$W_{cm}(s) = \left[\frac{1}{G_f(s)}\right] W_m(s)$$

$$\text{where } W_m(s) = \left\{\Phi_{vi}(s)/\Phi_{vv}(s)\right\}_+ \Big/ \Phi^+_{vv}(s)$$

For our purposes a minimum phase transfer function is one which has no zeros in the RHP. Now, $W_m(s)$ is the solution for the overall system transfer function for minimum $\overline{\epsilon^2(t)}$ where there are no fixed elements. Therefore, the overall system transfer function is independent of $G_f(s)$ if $G_f(s)$ is minimum phase. What this means is that when the desired ouptut of a system is equal to the input, the above equation will always call for a $W_{cm}(s)$ such that the overall system transfer function is 1 (thus giving a mean-square-error of zero) if the fixed elements are minimum phase. And this sounds reasonable, since the ultimate goal of any compensation is to cancel all attenuation and phase shift caused by

the plant over the whole frequency spectrum, in order that the output will be an exact replica of the input. Even though this is the ultimate goal, we, of course recognize that physically, the goal is unattainable. Hence, this method in general leads to a compensating function which is physically unattainable. The above methods rather nonchalantly assumed that the correlation functions were available, whereas actually, the obtaining of a correlation function in analytical form is another barrier to be surmounted.

4.4 Correlation Functions

(a) Derivation from Theoretical Considerations:

Appendix B lists some autocorrelation functions that can be derived. Because of this, they are popular for use in design. Furthermore, their use does not represent too great a departure from the world of practicality.

(b) Derivation from Experiment:

Any sophisticated or complicated design calls for the determination of the correlation function by experiment. Newton[2] presents an approximate numerical procedure for computing them from oscillograph traces. He also indicates an analog computer method whereby a stylus is manually moved over the oscillograph tracings. He states that currently the M.I.T. Servomechanisms Laboratory prefers a numerical procedure because

computers can be used.

After determination by experiment, the functions are frequently not analytical and some approximation scheme is necessary if they are to be used in analytical equations. For example, reference 17 describes a scheme for approximating a correlation function by a series of damped cosine functions. If a function is approximated by any such scheme, generally the whole design problem is of such size that computers are necessary for the entire design problem.

4.5 Summary

The Analytical Design Method might be summarized as follows:

(a) The given specifications are: input signal, desired output, disturbances, performance index and required value of same, plant elements, and degree of freedom allowed in compensation.

(b) Classify the problem according to free, semi-free, or fixed configuration.

(c) For free or semi-free configuration, use an appropriately derived formula. For fixed configuration, express the performance index as a function of the free parameters; minimize the performance index by adjusting the parameters.

(d) See if the compensation thus determined yields the required value of performance index. If it does not, the specifications cannot be met; if it does, practical realization may begin.

The design objective is the attainment of the "best possible" servo to meet the specifications, using the chosen performance index as the criterion.

CHAPTER 5

TRIAL AND ERROR DESIGN METHOD REFINEMENTS

5.1 Introduction

The purpose of Chapter 5 is to aid a servomechanisms designer to find an appropriate method of approaching the solution of his particular design problem. To this end, a summarization of the techniques found profitable by the authors is made along with references to more complete study of the technique involved.

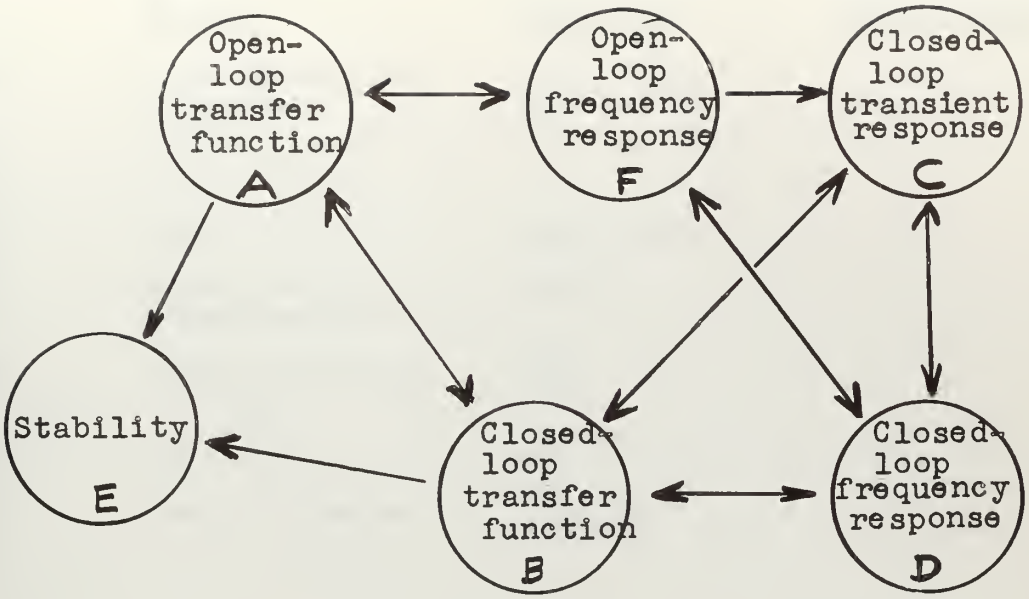
The design problem, when determined by the designer, will contain three parts: (1) the specifications which must be met, (2) fixed elements to be included, and (3) any limitations or preferences in the type of compensators to be used. The open-loop transfer function or frequency response of the simplest possible servo containing the required elements should be analyzed immediately in order to decide the amount and type of compensation required. This is done by synthesizing a system which will meet the specifications and then, using the compensators and variable gains, constrain the actual servo model to be similar to the system first synthesized. The final, or possibly feedback, step is to analyze the final servo model to insure that it does indeed meet the specifications. More care must be taken with approximations in this final analysis than in the preliminary analyses.

Following a guide to design methods available, some of the less well publicized methods are discussed. Some

root locus techniques, Mitrovic's method, and some transformation methods are explained to the extent that they can be used with the help of this paper. Theory and development of the methods is left to the references listed, with only a hint as to the validity of the methods. A knowledge of the basic frequency domain and time domain methods is assumed. The best general review of the methods available is either "Ordnance Engineering Design Handbook[4]" or "Handbook of Automation Computation and Control, Volume 1[19]"

5.2 Guide to Design

The various methods to design a servomechanism can be listed in any number of ways: type transfer function used, domain used, the graphical plane used, analysis or synthesis, etc. None of these groupings is an aid to the designer when he is attempting to make his choice for a design problem at hand. The breakdown of methods, shown on the following pages, while redundant, does allow the designer to concentrate his attention to the methods found fruitful by the authors. It is by no means an exhaustive collection.



	Method	Reference	Limitations
AB1	root locus	16,25	
AB2	solve equation $\frac{G}{1+G}$	3	
AB3	Mitrovic	par. 5.5	
AE1	Bode diagram	3	unity feedback
AE2	Nyquist criterion	20	unity feedback
AF1	Bode diagram	3	
AF2	solve equation G(jw)	20	
BA1	root locus	par. 5.4	unity feedback
BA2	solve equation $\frac{W}{1-W}$	3	unity feedback
BC1	root locus	par. 5.3	
BC2	second order approximation	3	
BC3	third order approximation	par. 5.3	

	Method	Reference	Limitations
BC4	inverse Laplace transform	20	
BD1	Bode diagram	3	
BD2	root locus	par. 5.3	
BD3	second order approximation	3,16	
BD4	solve equation $W(jw)$	20	
BE1	root locus	16	
BE2	Routh criteria	19, 20	
CB1	Bode diagram	3	
CB2	second order approximation	3, 4, 16, 19, 20	
CB3	curve fitting	20	
CD1	Nixon's method	par. 5.7(a)	
CD2	Chestnut & Mayer	par. 5.7(b)	
DB1	Bode diagram	3	
DB2	second order approximation	3, 19	
DC1	Floyd's method	par. 5.6(a)	
DC2	Guillemin's	par. 5.6(b)	
DC3	Stallard	4	
DF1	Nichol's chart	3	must know feedback
DF2	Nyquist diagram	20	must know feedback
DF3	Inverse polar plane plot	20	must know feedback

	Method	Reference	Limitations
FA1	Bode diagram	3	
FC1	Chestnut & Mayer diagrams	1, 4, 19	
FD1	Nichol's chart	3	
FD2	Nyquist diagram	20	
FD3	Inverse polar plane plot	20	

(Note: paragraphs indicated above refer to this paper.)

5.3 Root Locus Analysis

The root locus method was first developed by Walter Evans[18] in 1948. It has gained considerable popularity with usage. There have been many modifications of, and techniques developed about, the root locus.

Basic uses of the root locus are the plotting of the open-loop transfer function poles and zeros on the complex plane and the plotting of the locus of the roots of the closed-loop transfer function on the complex plane using the poles and zeros of the open-loop transfer function. The plotting of the root locus is basic to the general method and is assumed to be known by the designer[3,19,20].

The limits of gain which will allow a system to remain stable, and the second order approximation of the relative stability for a given gain, can be determined from the root locus.

Analysis of the root locus, from the position of the closed-loop poles and zeros, can be determined by the following methods which are listed approximately in order

of increasing difficulty, but also in order of increasing accuracy.

(a) Graphical Representation of Dominant Pole-Zero Locations Versus Selected Specifications.

Elgerd and Stephens[21] have graphed the step input response for almost every combination of roots and zeros up to fourth order for selected distances from the origin and for selected conjugate pole distances.

Abbott and Patton[22] have graphed the time for the output to first equal the input, the peak overshoot, the time of peak overshoot, and the settling time (all for a step input) for closed-loop systems containing a complex conjugate root pair (1) alone (second order) (2) with a zero (second order), or (3) a real root (third order). They further show that systems containing a complex conjugate root pair, a real root, and one or two zeros can be broken into partial fractions of the above three types and the transient responses then added.

Both of the above techniques require the system to be normalized to make one of the roots or zeros unity. Elgerd and Stephens consider more types of systems and their article (Trans AIEE, Vol 78, Pt II, 1959) is more apt to be available, at this time. The method of Abbott and Patton requires essentially no interpolation due to their mapping of the responses on the complex plane. Both methods show, along with Chu's (see below) that a real pole or zero loses most of its dominant influence when it is more than twice as far from the origin than one of

the complex conjugate roots. See Figure 1 at the end of this chapter.

(b) Chu's Approximation Equations[23]

This method computes, either graphically or analytically, the dominant terms of the transient response to a step input. From this he finds simple equations for the peak overshoot and the time of peak overshoot. Let

$$F_c(s) = \frac{kA(s)}{B(s)} = \frac{k\prod(s+z_i)}{\prod(s+q_i)}$$

$$T_p = \frac{1}{\omega_0}\left\{\frac{\pi}{2} - Ang\, A(s_0) + Ang\, B'(s_0)\right\}$$

$$M_p = \frac{2\omega_0}{\sigma_0+\omega_0}\left|\frac{kA(s_0)}{B'(s_0)}\right| \epsilon^{-\sigma_0 T_p}$$

$$C(t) \cong \frac{kA(0)}{B(0)} + 2\left|\frac{kA(s_0)}{B'(s_0)}\right| \epsilon^{-\sigma_0 t}\cos(\omega_0 t + \psi(s_0)) + \left\{\frac{kA(\sigma_1)}{B'(\sigma_1)}\epsilon^{-\sigma_1 t}\right\}$$

necessary only if $\sigma_1 \leq 3/T_p$

where: k = a factorable constant

$A_{(0)}$ = the product of the distances from all the zeros to the origin

$B_{(0)}$ = the product of the distance from all the roots to the origin

S_0 = $\sigma_0 + j\omega_0$ the location of one of the dominant complex roots

$A_{(s_o)}$ = the product of the distances from all of the zeros to s_o

$B'_{(s_o)}$ = the product of the distances from all the other roots to s_o

$\psi_{(s_o)} = Ang\, A_{(s_o)} - Ang\, B'_{(s_o)} - Ang\, s_o$

$Ang\, A_{(s_o)}$ = the sum of the angles from the real axis to the line from the zeros to s_o.

$Ang\, B'_{(s_o)}$ = the sum of the angles from the real axis to the line from the other roots to s_o.

$Ang\, s_o$ = the angle from the positive real axis to the line from the origin to s_o.

$A_{(\sigma_1)}$ = the product of the distances from all the zeros to σ_1

$B'_{(\sigma_1)}$ = the product of the distances from all the other roots to σ_1.

(c) Wheeler's Approximation Equations[24]

Wheeler's equations are similar to Chu's, except that he makes the substitution:

$$k = \frac{P_{(s_o)}}{A_{(s_o)}} = \frac{P_{(\sigma_1)}}{A_{(\sigma_1)}}$$

$$k\,\frac{A_{(0)}}{B_{(0)}} = C(\infty) = \text{Final value}$$

Thus, if one knows the final value of $C(t)$ or $F_c(s=0)$, he may plot the step response from the open-loop poles and the closed-loop roots. In any case, he can plot the error to a step.

$$T_P = \frac{1}{\omega_0}\left\{\frac{\pi}{2} - \text{Ang } A_{(s_0)} + \text{Ang } B'_{(s_0)}\right\}$$

$$M_P = C(\infty) + \frac{2}{\omega_n}\left[\frac{P_{(s_0)}}{B'_{(s_0)}}\right]\epsilon^{-\sigma_0 T_P}$$

$$\left[M_P - C(\infty)\right] = \frac{2}{\omega_n}\left[\frac{P_{(s_0)}}{B'_{(s_0)}}\right]\epsilon^{-\sigma_0 T_P}$$

$$C(t) = C(\infty) + \frac{2}{\omega_n}\left[\frac{P_{(s_0)}}{B'_{(s_0)}}\right]\epsilon^{-\sigma_0 t}\cos(\omega_0 t + \psi_{(s_0)}) + \left\{\frac{P_{(\sigma_1)}}{B'_{(\sigma_1)}}\epsilon^{-\sigma_1 t}\right\}$$

necessary only if $\sigma_1 \leq 3/T_P$

where: $P_{(s_0)}$ = the product of the distances for all the open-loop poles to s_0.

(d) Graphical Residue[25]

The graphical residue gives the true transient response to an input step. If there are repeated roots, it will not work in the form given below but the above reference should be consulted:

$$C(t) = \frac{k\,A_{(0)}}{B_{(0)}} + \sum_{\sigma_k \text{ real}} \frac{k\,A_{(\sigma_k)}}{B'_{(\sigma_k)}}\epsilon^{-\sigma_k t} + \sum_{(\sigma+j\omega)_k \text{ imag}} 2\left|\frac{k\,A_{(\sigma+j\omega)k}}{B'_{(\sigma+j\omega)k}}\right|\epsilon^{-\sigma_k t}\cos(\omega_k t + \text{Ang } A_\Delta - \text{Ang } B'_\Delta - \text{Ang } \Delta)^*$$

(e) Frequency Response from Root Locus[24]

The frequency response of the system can be gotten graphically for any given ω by measuring the distances from all the roots and zeros to jw on the imaginary axis.

* $\Delta \equiv (\sigma+j\omega)_k$

$$M(\omega) = \frac{k A(j\omega)}{B(j\omega)} = \text{magnitude of response}$$

$$N(\omega) = \sum \text{Ang } A(j\omega) - \sum \text{Ang } B(j\omega) = \text{phase shift of response}$$

5.4 Root Locus Synthesis

Having found where the desirable locations for the dominant roots are, the designer must now be able to "maneuver" the system's roots into these favorable locations. This can be attempted by varying the parameters of the given system or by adding a compensator in cascade.

(a) Effect of Varying One Parameter of Root Locus[24]

Dr. Wheeler demonstrated that any one parameter of the open-loop equation may be factored out of part of the characteristic equation and then be used as the gain of the final root locus, thus showing the effect on the system of varying this parameter. After factoring out the variable parameter, a root locus plotter would be a definite aid. This type of plotter is available in different degrees of sophistication[26]. The real use of this technique would be to see the true effect of one variable parameter.

(b) Placement of Lead or Lag Compensators[27]

The Ross-Warren technique is a natural outgrowth of the root locus equations and of earlier efforts by Walters[28] and Aseltine[29]. The locus is found to a point where the dominant complex conjugate roots are desired and the phase angle is noted. The placement of the lead or lag elements is partially fixed by this phase angle because they must be placed so as to change that angle to -180°. The compensator

is further, and completely, fixed by stipulating that the error coefficient (K_v) be unchanged by the compensation. This stipulation is unnecessarily restrictive and is one of the major drawbacks of the method.

Measure: θ_c, G_c, ω_{nc}, $\cos^{-1} \zeta_c$, K.

$$\varphi = (2n-1)\pi - \theta_c \text{ where } n \text{ is such that } |\varphi| < 180^\circ$$

$$\lambda = \cot^{-1}\left(\cot\varphi - \frac{K}{G_c}\csc\varphi\right)$$

$$z = \omega_n K \sin\lambda / G_c \sin\theta$$

$$p = \omega_n \sin\lambda / \sin(\theta - \varphi)$$

where:

θ_c = the phase angle measured to the desired root location.

G_c = the product of distances from the zeros to the desired root location divided by the product of the distances from the poles to the desired root location.

ω_{nc}, ζ_c = description of the desired root, in second order terminology.

K = the gain of the uncompensated equation.

z = zero location of the compensator

p = pole location of the compensator

z/p = attenuation of the compensator.

A modification of the above method allows the designer to approximate immediately, a lead filter with a

zero, and to use other than the uncompensated K_v. See Figure 2 at the end of this chapter with which one enters with the angle to be compensated and obtains the zero placement[30]. The effect of a real pole can be approximated from the same graph by subtracting the angle found at the pole distance from the angle found at the zero distance.

Mariotti has developed a similar method[30] for the determination of pole-zero location. The Ross-Warren technique, modified, requires successive approximations to place the compensator at the place with the best error coefficient, while Mariotti has graphed this placement so that it can be accomplished immediately. His work shows that the error coefficient increases as the compensator is moved further from the imaginary axis.

5.5 The Method of Mitrović[3,31]

(a) Theory

The method of synthesis presented by Dušan Mitrović is a different and potentially powerful approach to the solution of the synthesis problem. This method utilizes the characteristic equation in the unfactored polynomial form.

$$f_s = a_n s^n + a_{n-1} s^{n-1} + \cdots a_1 s + a_0$$

Separating the real and the imaginary parts of the characteristic equation and solving for a_0 and a_1 produces two parametric equations in terms of the higher order coefficients and with the variables ζ and ω. Inspection of these equations reveals that each power of ω contains a single coefficient of the characteristic equation and an

easily tabulated function of ζ.

$$a_0 = -\omega^2\left[a_2\varphi_1(\zeta) + a_3\varphi_2(\zeta)\omega + \cdots + a_n\varphi_{n-1}(\zeta)\omega^{n-2}\right]$$

$$a_1 = a_2\varphi_2(\zeta)\omega + a_3\varphi_3(\zeta)\omega^2 + \cdots + a_n\varphi_n(\zeta)\omega^{n-1}$$

$$\varphi_k(\zeta) = -\left[2\zeta\varphi_{k-1}(\zeta) + \varphi_{k-2}(\zeta)\right] \text{ except } \begin{cases}\varphi_0 = 0\\ \varphi_1 = -1\end{cases}$$

The basic consideration of the Mitrović method is to let a_0 and a_1 become variable and to observe the (a_1, a_0) plane.

(b) Properties

The locus $\Gamma(\zeta)$ of the parametric equations is plotted for a specific ζ and is found to have some remarkable properties:

(1) $\Gamma(0)$

The servomechanism is stable if certain conditions are met. This is because the total phase rotation as $\omega \to \infty$ is fixed for a stable system similar in nature to the Nyquist criteria. This curve immediately shows the range of a_1 and a_0 for which the system will be stable.

(2) $\Gamma(\zeta_k)$

The characteristic equation has no roots with ζ greater than ζ_k if certain conditions are met. This curve immediately shows the range of a_1 and a_0 for which the system will have a ζ less than ζ_k. More important, it will show the exact a_1 and a_0 necessary to make the system have roots

at any given ω_n and ζ_k.

(3) $\Gamma(1)$

The characteristic equation has all real roots if certain conditions are met. Here again, the range of a_1 and a_o are shown. The real roots are quickly determinable. The effect of changing a_1 or a_o is immediately evident. If there is but one complex pole pair, it can be determined easily once the real roots are known.

(4) Third order equations

If the system can be approximated by a third order equation (i.e. a dominant complex conjugate pole pair and a maximum of one dominant real root with no more than 3 dominant zeros) this method is considered the ultimate in flexibility by the authors.

A third order characteristic equation can be normalized so that a_3 and a_2 are both one. Now all third order equations can be solved by one set of $\Gamma(\zeta)$. This set is tabulated once and for all[3,31]. The a_1 and a_o required for all dominant root combinations is immediately discernable. Thus a very powerful tool is available to determine the parameters of any system with a closed-loop characteristic equation of third order or less.

(c) Plotting of $\Gamma(\zeta)$

The equations developed in the theory can be used to plot $\Gamma(\zeta)$ with a table of $\varphi_k(\zeta)$.

An easier method is available, though, using a table that has the $\varphi_k(\zeta)\,\omega^{k-1}$ already constructed for the ζ in question. Tables II, III, IV and V (at the end of this chapter) are constructed for values of ζ of 0, 0.5, 0.7 and 1.0.

$$a_0 = -\omega^2\left[a_2\psi_1(\zeta,\omega) + a_3\psi_2(\zeta,\omega) + \cdots + a_n\psi_{n-1}(\zeta,\omega)\right]$$

$$a_1 = a_2\psi_2(\zeta,\omega) + a_3\psi_3(\zeta,\omega) + \cdots + a_n\psi_n(\zeta,\omega)$$

$$\psi_k(\zeta,\omega) = \varphi_k(\zeta)\,\omega^{k-1}$$

These tables are constructed for values of ω from zero to one. To insure that the area of interest falls within the above ω, the characteristic equation must be normalized so that both a_n and a_{n-1} are unity. This is accomplished easily by the substitution $s = a_{n-1}\,P_T$, and it insures (for equations with all negative real parts in the roots) that all roots lie between 0 and -1. The transformed parameter is related to the actual parameter by:

$$\omega_T = \frac{\omega}{a_{n-1}}$$

$$a_{0T} = \frac{a_0}{(a_{n-1})^n} \quad ; \quad a_{1T} = \frac{a_1}{(a_{n-1})^{n-1}}$$

(d) Determining Stability

The $\Gamma(0)$ curve is plotted and point (a_1, a_0) is

noted.

The system is stable if:

(1) The point (a_1, a_0) lies in the first quadrant, and

(2) while ω varies between 0 and $+\infty$ the curve alternately cuts the lines a_0 and a_1 with the provision that a_0 is cut first, and

(3) that the total number of points of intersection is n, where n is the order of the characteristic equation.

(e) Determining ζ

The $\Gamma(\zeta_k)$ curve is plotted and the point (a_1, a_0) is noted. All roots have damping factor greater than ζ_k if:

(1) The point (a_1, a_0) lies in the first quadrant, and

(2) while curve $\Gamma(\zeta_k)$ alternately cuts the lines a_0 and a_1 with the provision that a_0 is cut first, and

(3) that the total number of points of intersection is m, where m is the largest integer which fulfills the inequality:

$$m + \left[1 - (-1)^m\right] \frac{\theta}{\pi} < n\left(1 + \frac{2\theta}{\pi}\right)$$

where n is the order of the characteristic equation and $\theta = \frac{\pi}{2} - \cos^{-1}\zeta_k$.

If the point (a_1, a_0) lies on the $\Gamma(\zeta_k)$ curve,

the damping factor of a complex conjugate root pair is ζ_k, and ω_n is the ω of that point on the curve.

(f) Determining the Real Roots

The $\Gamma(1)$ curve is plotted and the point (a_1, a_0) is noted.

All real roots lie on the $\Gamma(1)$ curve at a distance σ_i (when $\zeta = 1.0$, $\omega = \sigma$) from the origin such that a line tangent to σ_i will pass through (a_1, a_0). Conversely, at any place that a line from (a_1, a_0) is tangent to the curve, there is a root on the negative real axis at σ.

(g) Lead and Lag Compensation

The use of the Mitrović method is quite easy with lead or lag compensators. The characteristic equation of a lag compensated transfer function will probably have the lag pole in coefficients other than a_1 or a_0. It will be found that the effect of the lag pole on these coefficients is so small that it can be deleted.

The characteristic equation of a lead compensated system has the same problem of having the compensator pole in the higher order coefficients. Unfortunately, it is a larger term and cannot be ignored. Instead it must be fixed, in order to get numerical values for the coefficients. This would also have to be done to the zeros and the gain if they appeared in the higher order terms. Then any computations involving

variations of parameters which were fixed in the higher order terms would either be approximations or the curve would have to be continuously replotted.

5.6 Frequency Response to Transient Response

The several methods for obtaining the transient response to an impulse input from the frequency response curve make use of the following relationship:

$$f(t) = \frac{2}{\pi}\int_0^\infty \cos \omega t \; Re\left[F(j\omega)\right] d\omega$$

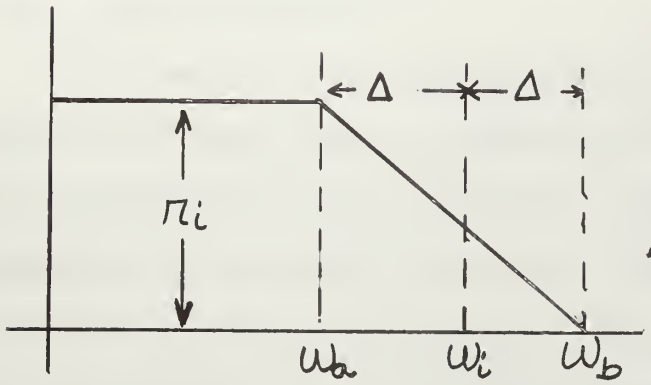
where $Re[F(j\omega)]$ is the real part of the frequency response.

The real part of the frequency response can be computed directly from the closed-loop frequency response curve. If the frequency response were plotted in polar coordinates, the real part could be read directly. Chen and Shen[32] suggest a chart which has the closed-loop frequency response plotted in polar coordinates and overlaid upon it is the open-loop, unity feedback frequency response. Gould[4] recommends a chart, similar to a Nichols chart, on which the open-loop frequency response curve is plotted as log-gain vs. phase angle (See Figure 3 at the end of this chapter.) Overlaid upon the chart is the magnitude of the real part of the closed-loop frequency response for a unity feedback system.

(a) Floyd's Method[4,19]

The real part of the closed-loop frequency response curve is approximated by the sum of a series of trapezoids:

$$f(t) = \sum_{i=1}^{k} \frac{2}{\pi} \pi_i \omega_i \left[\frac{\sin \omega_i t}{\omega_i t}\right] \left[\frac{\sin \Delta_i t}{\Delta_i t}\right]$$



$$\omega_i = \frac{\omega_a + \omega_b}{2}$$

$$\Delta_i = \frac{\omega_b - \omega_a}{2}$$

A check on the correctness of the trapezoids is:

$$f(0) = \frac{2}{\pi} \sum_{i=1}^{k} \pi_i \omega_i$$

(b) Guillemin's Method [4,25]

The straight line approximation of the real part of the closed-loop frequency response is differentiated twice, leaving a series of impulses of height a_i and at frequency ω_i:

$$f(t) = -\frac{2}{\pi t^2}\left[\frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos \omega_j t\right]$$

A check on the correctness of the impulses is:

$$\frac{a_0}{2} + \sum_{j=1}^{m} a_j = 0$$

$$f(0) = \frac{1}{\pi} \sum_{j=1}^{m} a_j \omega_j^2$$

(c) A Computer Method [33]

Levadi has combined Floyd's method and Guillemin's method in a way that is easy to program on a computer.

The output from the computer is the impulse response, the step response, and the response to a ramp input.

5.7 Transient Response to Frequency Response

(a) Nixon's Method[20]

A straight line approximation of the transient response to a step input is made and it is broken into the sum of a series of ramp functions. These functions could be expressed in the time domain as a magnitude multiplied by a unit ramp with a time delay. Instead, they are expressed directly in the frequency domain (by Laplace transform) and multiplied by s. This is the transfer function because, before multiplying by s, this sum is the Laplace equation for the output to a unit step input. If the transient response to an impulse were used, the multiplication by s would have been unnecessary. To find the frequency response, substitute $j\omega$ for s:

$$F(j\omega) = \sum_{i=0}^{n} \left[a_i \epsilon^{-t_i(j\omega)} \right] / (j\omega)$$

Since the equation is indeterminate at $\omega = 0$, L'Hospital's rule must be used.

(b) Chestnut and Mayer's Method[4]

The transient response to a step input is broken up into the sum of a series of step functions each occurring at an equal interval of time, but being of a variable height ΔC_i:

$$F(j\omega) = \sum_{i=1}^{n} \Delta C_i \epsilon^{-t_i(j\omega)}$$

The transient response to an impulse is broken up into the sum of a series of pulses of uniform width Δt but of variable height C_i:

$$F(j\omega) = \sum_{i=1}^{n} C_i \Delta t \epsilon^{-t_i(j\omega)}$$

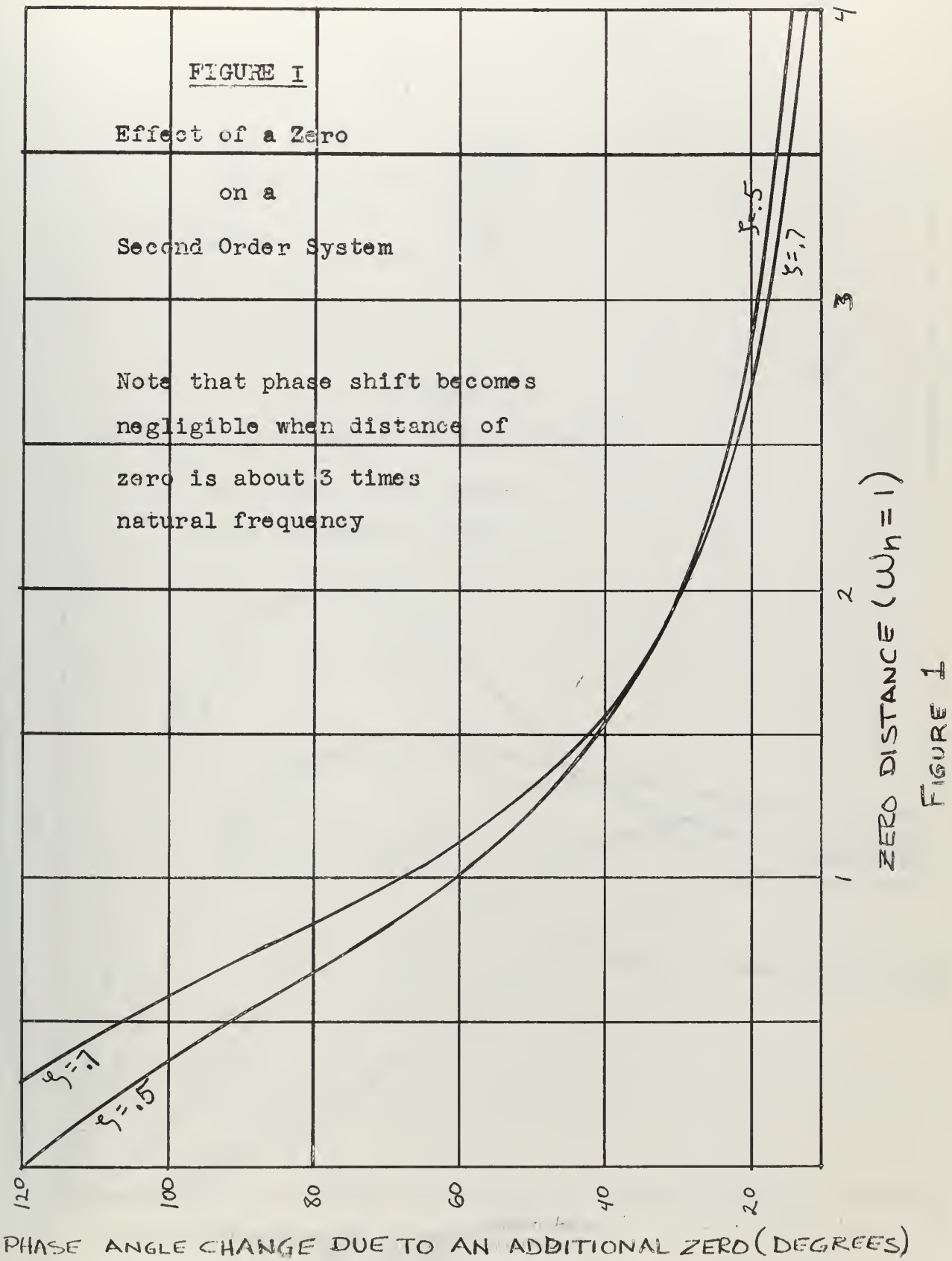


FIGURE I
Effect of a Zero
on a
Second Order System
Note that phase shift becomes
negligible when distance of
zero is about 3 times
natural frequency
ζ=.5
ζ=.7
ζ=.7
ζ=.5
4
3
2
1
ZERO DISTANCE (ωn=1)
120
100
80
60
40
20
PHASE ANGLE CHANGE DUE TO AN ADDITIONAL ZERO (DEGREES)

FIGURE 1

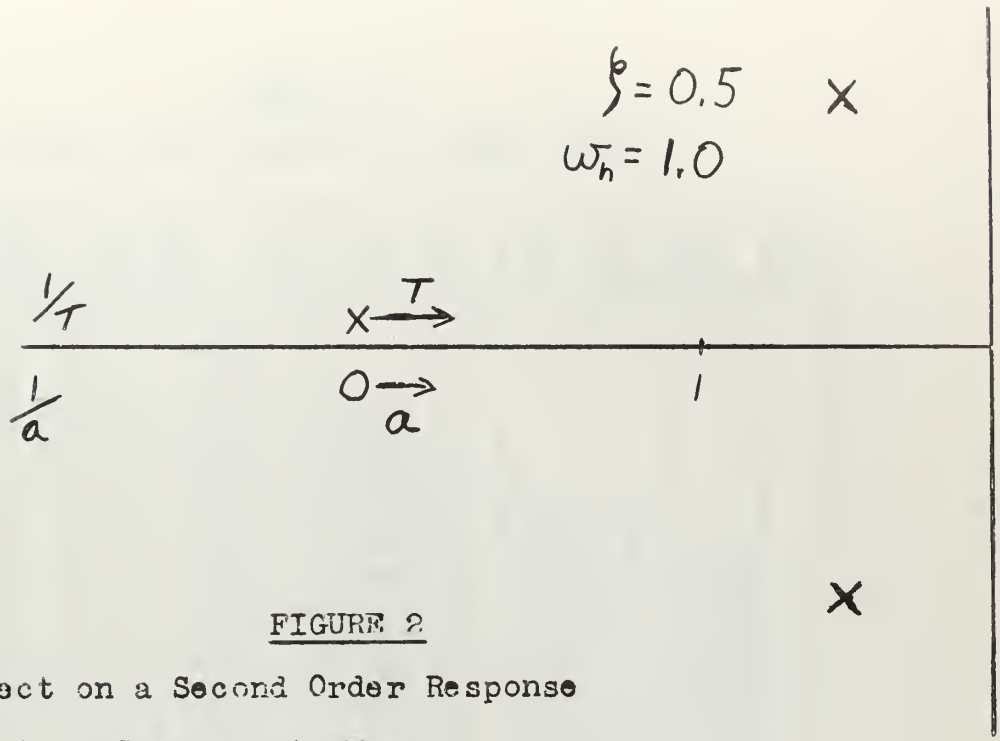


FIGURE 2

Effect on a Second Order Response to a Step Input When a Pole or Zero Is Added

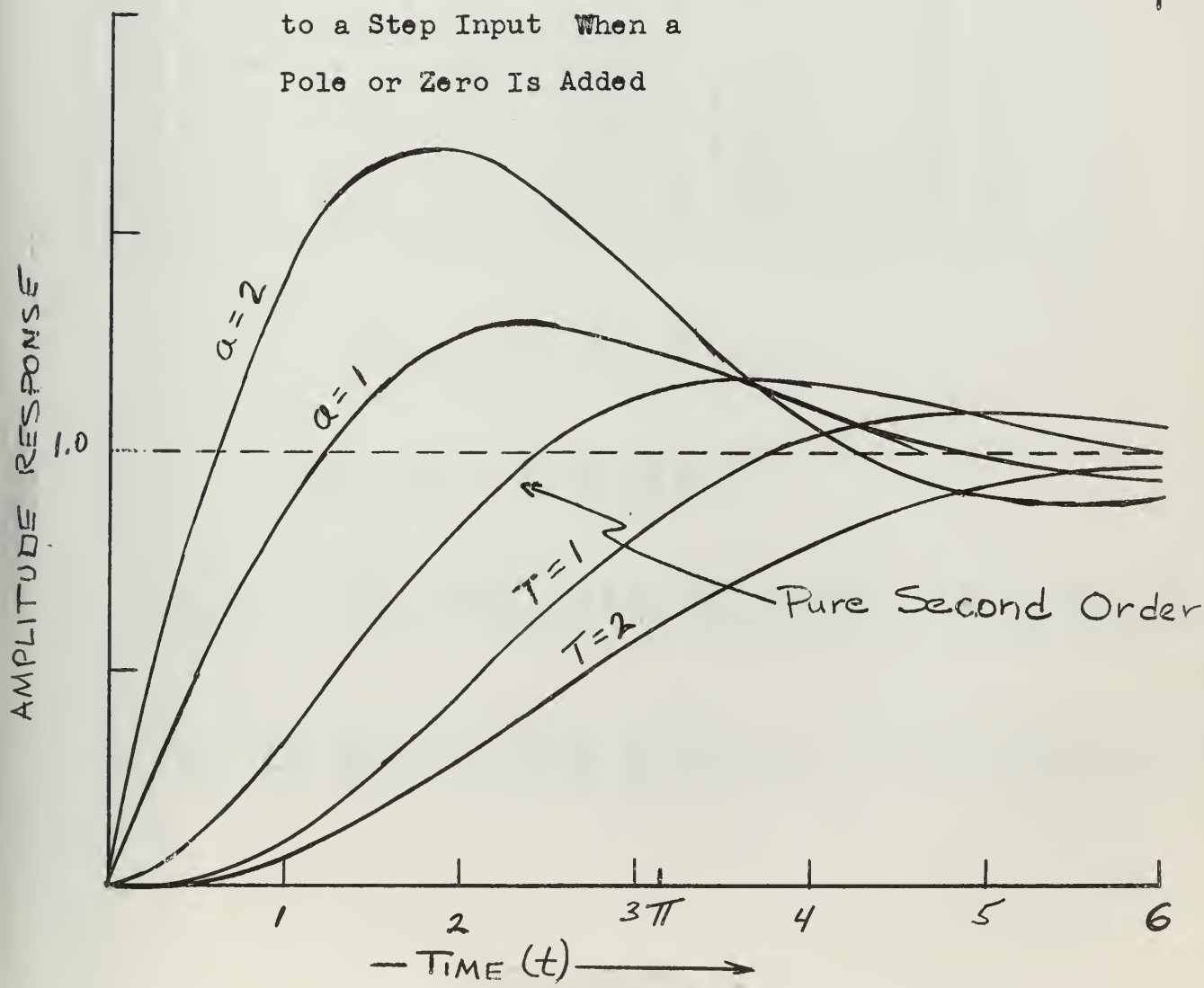


TABLE II

TABLE FOR COMPUTING MITROVIC'S $\Gamma(o)$ CURVE

ω	w^2	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9	ψ_{10}
0	0	-1	0	0	0						
.05	.003	-1	0	.003	0						
.10	.01	-1	0	.010	0	0					
.15	.023	-1	0	.023	0	.001					
.20	.04	-1	0	.040	0	.002	0				
.25	.063	-1	0	.063	0	.004	0	0			
.30	.09	-1	0	.090	0	.008	0	.001	0		
.35	.123	-1	0	.123	0	.015	0	.002	0	0	
.40	.16	-1	0	.160	0	.026	0	.004	0	.001	0
.45	.203	-1	0	.203	0	.041	0	.008	0	.002	0
.50	.25	-1	0	.250	0	.063	0	.016	0	.004	0
.55	.303	-1	0	.303	0	.092	0	.028	0	.008	0
.60	.36	-1	0	.360	0	.130	0	.047	0	.017	0
.65	.423	-1	0	.423	0	.179	0	.075	0	.032	0
.70	.49	-1	0	.490	0	.240	0	.118	0	.058	0
.75	.563	-1	0	.563	0	.316	0	.178	0	.100	0
.80	.64	-1	0	.640	0	.410	0	.262	0	.168	0
.85	.723	-1	0	.723	0	.522	0	.377	0	.273	0
.90	.81	-1	0	.810	0	.656	0	.531	0	.431	0
.95	.903	-1	0	.903	0	.815	0	.735	0	.663	0
1.0	1.0	-1	0	1.0	0	1.0	0	1.0	0	-1.0	0

ω vs. ψ for $\varphi = 0$

$$\psi_k(\varphi, \omega) = \varphi_k(\varphi)\omega^{k-1}$$

$$a_0 = -w^2\left[a_2\psi_1(\varphi,\omega) + a_3\psi_2(\varphi,\omega) + \cdots + a_n\psi_{n-1}(\varphi,\omega)\right]$$

$$a_1 = a_2\psi_2(\varphi,\omega) + a_3\psi_3(\varphi,\omega) + \cdots + a_n\psi_n(\varphi,\omega)$$

TABLE III

TABLE FOR COMPUTING MITROVIC'S $\Gamma(0.5)$ CURVE

ω	w^2	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9	ψ_{10}
0	0	-1	0								
.05	.003	-1	.050	0	0						
.10	.01	-1	.100	0	-.001	0					
.15	.023	-1	.150	0	-.003	.001					
.20	.04	-1	.200	0	-.008	.002					
.25	.063	-1	.250	0	-.016	.004	0	0			
.30	.09	-1	.300	0	-.027	.008	0	-.001			
.35	.123	-1	.350	0	-.043	.015	0	-.002	.001		
.40	.16	-1	.400	0	-.064	.026	0	-.004	.002	0	0
.45	.203	-1	.450	0	-.091	.041	0	-.008	.004	0	-.001
.50	.25	-1	.500	0	-.125	.063	0	-.016	.008	0	-.002
.55	.303	-1	.550	0	-.166	.092	0	-.028	.015	0	-.005
.60	.36	-1	.600	0	-.216	.130	0	-.047	.028	0	-.010
.65	.423	-1	.650	0	-.275	.179	0	-.075	.049	0	-.021
.70	.49	-1	.700	0	-.343	.240	0	-.118	.082	0	-.040
.75	.563	-1	.750	0	-.422	.316	0	-.178	.134	0	-.075
.80	.64	-1	.800	0	-.512	.410	0	-.262	.210	0	-.134
.85	.723	-1	.850	0	-.614	.522	0	-.377	.321	0	-.232
.90	.81	-1	.900	0	-.729	.656	0	-.531	.478	0	-.387
.95	.903	-1	.950	0	-.857	.815	0	-.735	.698	0	-.630
1.0	1.0	-1	1.00	0	1.0	1.0	0	1.0	1.0	0	-1.0

ω vs. ψ for $\zeta = 0.5$

$$\psi_k(\zeta, \omega) = \varphi_k(\zeta)\, \omega^{k-1}$$

$$a_0 = -\omega^2 \left[a_2 \psi_1(\zeta, \omega) + a_3 \psi_2(\zeta, \omega) + \cdots + a_n \psi_{n-1}(\zeta, \omega) \right]$$

$$a_1 = a_2 \psi_2(\zeta, \omega) + a_3 \psi_3(\zeta, \omega) + \cdots + a_n \psi_n(\zeta, \omega)$$

TABLE IV

TABLE FOR COMPUTING MITROVIC'S $\Gamma(0.7)$ CURVE

ω	w^2	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9
0.0	.000	-1	0	0						
.05	.003	-1	.070	-.048						
.10	.010	-1	.140	-.096		0				
.15	.023	-1	.210	-.144		.001				
.20	.040	-1	.280	-.192	0	.002	0			
.25	.063	-1	.350	-.240	-.001	.004	-.001	0		
.30	.090	-1	.420	-.288	-.002	.008	-.003	.001		
.35	.123	-1	.490	-.336	-.002	.016	-.007	.002		
.40	.160	-1	.560	-.384	-.004	.027	-.014	.004		0
.45	.203	-1	.630	-.432	-.005	.042	-.026	.008	0	-.002
.50	.250	-1	.700	-.480	-.007	.065	-.044	.014	.001	-.004
.55	.303	-1	.770	-.528	-.010	.095	-.070	.025	.002	-.009
.60	.360	-1	.840	-.576	-.012	.134	-.109	.043	.003	-.018
.65	.423	-1	.910	-.624	-.015	.185	-.162	.069	.005	-.034
.70	.490	-1	.980	-.672	-.019	.249	-.235	.108	.009	-.062
.75	.563	-1	1.050	-.720	-.024	.328	-.332	.163	.015	-.108
.80	.640	-1	1.120	-.768	-.029	.424	-.458	.241	.023	-.180
.85	.723	-1	1.190	-.816	-.034	.541	-.620	.346	.036	-.293
.90	.810	-1	1.260	-.864	-.041	.680	-.825	.488	.054	-.463
.95	.903	-1	1.330	-.912	-.048	.844	-1.082	.675	.078	-.713
1.0	1.0	-1	1.400	-.960	-.056	1.036	-1.398	.918	.112	-1.075

ω vs. ψ for $\gamma = 0.7$

$$\psi_k(\gamma, \omega) = \varphi_k(\gamma)\,\omega^{k-1}$$

$$a_0 = -\omega^2\left[a_2\psi_1(\gamma,\omega) + a_3\psi_2(\gamma,\omega) + \cdots + a_n\psi_{n-1}(\gamma,\omega)\right]$$

$$a_1 = a_2\psi_2(\gamma,\omega) + a_3\psi_3(\gamma,\omega) + \cdots + a_n\psi_n(\gamma,\omega)$$

TABLE V

TABLE FOR COMPUTING MITROVIC'S $\Gamma(1.0)$ CURVE

ω	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9
0	-1	0	0						
.05	-1	.100	-.008	.000	-.000				
.10	-1	.200	-.030	.004	-.001	.000			
.15	-1	.300	-.068	.014	-.003	.000			
.20	-1	.400	-.120	.032	-.001	.000	.000		
.25	-1	.500	-.188	.062	-.020	.006	-.001	.000	
.30	-1	.600	-.270	.108	-.041	.015	-.005	.002	.000
.35	-1	.700	-.368	.172	-.075	.032	-.013	.005	-.002
.40	-1	.800	-.480	.256	-.128	.061	-.028	.013	-.006
.45	-1	.900	-.608	.364	-.205	.111	-.058	.030	-.013
.50	-1	1.000	-.750	.500	-.312	.188	-.109	.062	-.035
.55	-1	1.100	-.908	.666	-.458	.302	-.194	.122	-.076
.60	-1	1.200	-1.080	.864	-.648	.467	-.327	.224	-.151
.65		1.300	-1.268	1.108	-.893	.696	-.528	.392	-.287
.70	-1	1.400	-1.470	1.372	-1.201	1.009	-.823	.659	-.518
.75		1.500	-1.688	1.688	-1.582	1.424	-1.246	1.068	-.901
.80	-1	1.600	-1.92	2.048	-2.048	1.966	-1.835	1.678	-1.510
.85		1.700	-2.168	2.456	-2.610	2.662	-2.640	2.565	-2.453
.90	-1	1.800	-2.430	2.916	-3.281	3.540	-3.720	3.826	-3.875
.95		1.900	-2.708	3.430	-4.073	4.643	-5.146	5.586	-5.971
1.0	-1	2.000	-3.0	4.0	-5.0	6.0	-7.0	8.0	-9.0

ω vs. ψ for $\mathcal{S} = 1.0$

$$\psi_k(\mathcal{S}, \omega) = \varphi_k(\mathcal{S})\,\omega^{k-1}$$

$$a_0 = -\omega^2\left[a_2\psi_1(\mathcal{S},\omega) + a_3\psi_2(\mathcal{S},\omega) + \cdots + a_n\psi_{n-1}(\mathcal{S},\omega)\right]$$

$$a_1 = a_2\psi_2(\mathcal{S},\omega) + a_3\psi_3(\mathcal{S},\omega) + \cdots + a_n\psi_n(\mathcal{S},\omega)$$

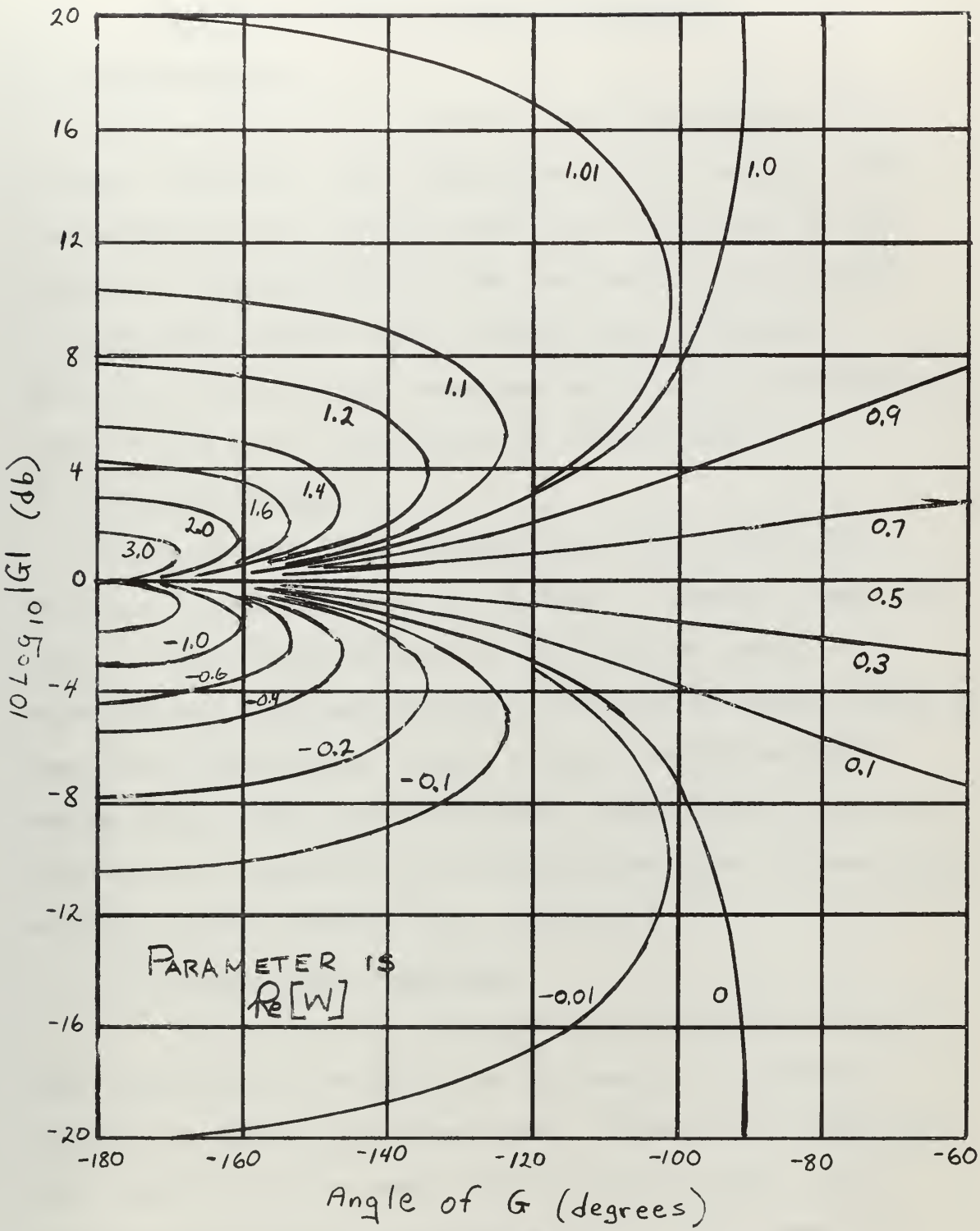


FIGURE 3

(GAIN-PHASE LOCI OF CONSTANT REAL PART OF $W = \frac{G}{1+G}$

CHAPTER 6

ANALYTICAL DESIGN METHOD REFINEMENTS

6.1 Introduction

In this chapter we summarize selected methods from current literature that concern Analytical Design. Some of the methods do not strictly meet the definitions of the Analytical Design Method at the beginning of this paper, but they are included here, because the literature generally considers any technique which uses a performance index as its main criterion to be "analytical".

6.2 Newton's Methods[2]

Chapter 4 was built almost exclusively around the writings of Newton, Gould and Kaiser[2]. However, they have done a great deal more than to develop some basic mathematics; they have expended much effort toward applying the basic mathematics. Their methods use ISE and MSE exclusively, since, as they state, these indices are the only ones of engineering usefulness that lead to reasonably straightforward mathematical analysis.

(a) Translation Functions

Newton proves, by means of variational calculus, that translation functions can be treated in a manner similar to correlation functions. Translation functions (see appendix B) are used with transient inputs. The formulas indicated in Chapter 4, then, can be used by merely substituting the correct translation function in the place of the correlation function. Most translation

functions can be found by Parseval's Theorem. Use of step and ramp inputs, even though they have no Fourier transform, impose no special limitation, since it is possible to introduce a convergence factor to find a Fourier transform. (The Fourier transform found this way is the same as the Laplace transform; however, special care is required in taking the inverse transform). This scheme is also much easier to use if the fixed elements are minimum phase. What appears to be a particular advantage of the method is its indifference to the kind of transfer function that exists in the plant. For example, it is able to handle a $G_f(s) = \epsilon^{-Ts}$ as readily as a $G_f(s) = \frac{1}{sT+1}$. Although this method normally employs the frequency domain for finding the best compensation, it frequently finds the minimum value of ISE by working exclusively in the time domain.

(b) Saturation

As indicated earlier, compensation by error minimization will almost invariably lead to saturation in a linear system, rendering the assumed mathematical model invalid. The reason for this is that the "called for" compensation transfer function is usually that which will cancel the effect of the plant transfer function. Since the plant normally consists of integrations and lags, this means that the compensation (or, at least, the equivalent cascade compensation) will call for several differentiations. Differentiations lead to large signal excursions in the system, with the result that saturation occurs somewhere. Differentiation also calls for more bandwidth, since the

signals have high frequency components. Many physical systems have the characteristic of linearity for large amplitudes at low frequencies, but can only tolerate small amplitudes at higher frequencies if linearity is to be preserved. To make the Analytical Design Method practical, this reference introduces constraints into the equations to insure that saturation does not occur.

(1) Transient Inputs - The most direct approach for avoiding saturation is to limit the peak value of that signal in the mathematical model that corresponds to the signal in the physical system that is likely to cause saturation. However, it is impractical to express the peak value of a signal as a function of the free parameters for systems above second order. The reason is that finding the roots of the characteristic equation above that order requires the use of numerical values so specific values of the parameters are necessary. But the integral-square value of a signal can frequently be expressed in terms of the free parameters; since the integral-square value gives large weight to large values, then some degree of control of the peak value is obtained. One procedure is to express the integral-square error as a function of the free parameters; also express the integral-square value of the saturaration signal as a function of the free parameters. Then

solve the equations by adjustment of the parameters. For fixed configurations this is best done by trial and error. For free and semi-free configurations, this reference introduces a more refined technique, employing a method developed by LaGrange. By use of a synthetic function and a constant called the LaGrange multiplier, the minimization of error problem is converted from one with a constraint to one without a constraint. Either of these methods works well in a fixed configuration problem, such as a positional servo. In this type of servo, saturation most frequently results from demanding an acceleration of the output member or load, which the motor is simply not capable of providing. It is to be noted that each introduction of a constraint effectively eliminates one free parameter in system design.

(2) Stochastic Inputs - Newton's method for handling saturation with stochastic inputs is quite impressive. The method involves the derivation of a modified Wiener-Hopf equation with the insertion of a LaGrange multiplier. Since explicit solution of the multiplier is difficult, graphical procedures are usually resorted to, for determination of the multiplier. The method seems to have no restrictions as long as the use of the rms value of the saturating signal is an acceptable criterion for constraining saturation. If this is unacceptable,

then there seems to be no choice, analytically speaking, but to separate the design problem into two parts: a linear and a non-linear mathematical model, employing non-linear techniques after saturation occurs.

(c) Bandwidth Minimization

The above paragraph indicated one reason why design by this method called for increased bandwidth. There is another reason: Since the input signals are represented by their power-density spectra, or energy-density spectra (for transient inputs), minimization of ISE or MSE will demand that the system transmit all those frequencies in the input with negligible amplitude and phase change. If the input signal is aperiodic, its spectrum includes all frequencies; if it is periodic, it includes discrete frequencies which are integer multiples of the fundamental out towards infinity. Hence, for minimum error, namely zero, an infinite bandwidth is called for.

Excessive bandwidth has several undesirable features:

(1) It allows the transmission of excessive noise;

(2) It causes unwanted saturation by raising the internal signal level; along with this it increases the requirements for components operating at high power levels, since output peak power is generally associated with the higher frequencies;

(3) It complicates the compensation, since the high frequency dynamics of the components cannot be

neglected in determining their transfer function, and the compensation must effectively offset these additional components.

This reference presents an analytical method for minimizing bandwidth by transposing the problem to one of minimizing the mean-square value of a transmitted signal. The method is good for any configuration and for both transient and stochastic signals. It uses a bandwidth test. For the test, a stationary stochastic noise signal (characterized by that to which the system is expected to be subjected) is used as the input to the system. The resultant output is passed through a filter whose rms output is measured. A standard system is defined with variable bandwidth. It is fed the same noise signal and its output passed through a filter. The rms value of the control system filtered output is correlated with the control system bandwidth by comparison with the standard system filtered output. The bandwidth of the standard system is adjusted until the two rms filtered outputs are equal; by definition, the two bandwidths are then equal. Only one restriction exists: The standard system, along with the noise source and filter, must produce a monotonically increasing rms output with increasing bandwidth; otherwise, minimizing the control system filtered output would not necessarily minimize its bandwidth. By this scheme the problem of minimizing bandwidth is equivalent to determining that overall system transfer (or weighting) function that

minimizes the mean square value of its filtered output subject to the constraints required by the performance index of the normal input signal.

It follows that additional specifications are required for establishing the noise source, the standard system and the filter. A frequently chosen noise source is white noise because it has components at all frequencies and is easily handled analytically. The filter is chosen with regard to the cutoff characteristics desired in the system; the filter must have a finite mean-squared output since this is the signal to be minimized. This reference recommends and uses a pure differentiator of (n-1)th order, requiring system cutoff at the rate of $1/w^n$ The standard systems recommended and used are the simple binomial filter or the Butterworth filter (see p.220, reference 2) chosen so that there is produced a finite value of mean-square filtered output. Since the problem is now transposed to that of minimizing the noise transmitted through the system, subject to the constraints imposed by the performance index, another problem in variational calculus results. The solution proceeds as indicated in the paragraph about saturation: The LaGrangian multiplier is introduced which eventually leads to a Wiener-Hopf equation.

(d) Stability

All of the above techniques have system stability inherent in them because of the "spectrum factorization" technique required to solve the Wiener-Hopf equation.

There is another matter to consider: These techniques use equivalent cascade compensation, requiring manipulation back to the feedback loop after $G_c(s)$ has been determined. The possibility exists that, if the plant elements are not minimum phase (zeros in RHP), $G_c(s)$ introduces a pole to cancel the RHP zero, and thus $G_c(s)$ is unstable in its own right. To prevent this the above techniques require that the plant elements be stable. This is not a restriction, since all that is necessary is to provide some feedback loop around the plant elements to make them stable, and use the equivalent transfer function as the fixed elements for the rest of the design.

(e) Disturbances

The above line of reasoning applies to disturbances also: The fact that they are applied at a place other than the input merely requires that they be manipulated out to the input. This modified input is then used in the design.

6.3 Graham & Lathrop Methods[12]

These methods are not Analytical Design Methods, but their objective is the same: The Synthesis of "Optimum" Transient Response. This reference, first of all, seems to have about the most complete exposè in print on the merits of the various performance indices. Because of the exposè it is concluded that the best index is Integral-of-Time-Multiplied-Absolute-Error (ITAE). Since this function is not analytic, strictly analytical methods cannot be applied.

An almost exhaustive study was made on the following criteria, as applied to a second order system with a step

input (this was a zero-displacement error system whose transfer function was of the form $\frac{C}{R}(s) = \frac{1}{s^2+2\zeta s+1}$):

$$P_1 = \int_0^\infty e(t)\,dt \;; \qquad P_5 = \int_0^\infty t\,|e(t)|\,dt$$

$$P_2 = \int_0^\infty |e(t)|\,dt \;; \qquad P_6 = \int_0^\infty t\,e^2(t)\,dt$$

$$P_3 = \int_0^\infty e^2(t)\,dt \;; \qquad P_{10} = \int_0^\infty t^2\,e^2(t)\,dt$$

$$P_4 = \int_0^\infty t\,e(t)\,dt \;; \qquad P_{11} = \int_0^\infty t^2\,|e(t)|\,dt$$

Plots of the various criteria versus damping ratio resulted in the following:

(1) P_1 and P_4 dictate a $\zeta = 0$ for minimum error and were rejected on this ground.

(2) P_3 called for $\zeta = .5$ but had very little selectivity; but it can be handled analytically and can be mechanized on a computer.

(3) P_2 and P_5 called for $\zeta \approx .5$; both are easily mechanized for the computer; both have better selectivity than P_3 with P_5 giving much sharper selectivity.

(4) P_6, P_{10}, P_{11}, called for $\zeta \approx .6 - .8$ and gave good selectivity. All were rejected because they are too difficult to handle either analytically or by computer.

P_2, P_3, P_5 were retained and applied to a zero velocity error system of form

$$\frac{C}{R}(s) = \frac{2\zeta s + 1}{s^2 + 2\zeta s + 1}$$ and thence

applied to higher order systems. It was found that P_5(ITAE) was the only criterion which retained any selectivity. It was, therefore, concluded that the ITAE criterion was best. (It is interesting to note that one of the advantages of ISE criteria offered by Newton is the fact that it is not very selective, thus allowing wide latitude practically in adjusting the system.)

Next in the study, a set of limitations is established: The servo must be a "duplicator"; namely, it will be subjected only to a step input and it will have zero-displacement error (meaning the closed loop transfer function has unity in the numerator). With these restrictions the servo is merely a low pass filter. The servo could, therefore, be represented by standard forms of either a binomial filter, or a Butterworth filter, or a "minimum ITAE" filter. It is a simple algebraic matter then, to force the coefficients of the servo characteristic equation to equal the coefficients of the standard forms. It was found that the minimum ITAE filter response combined the good response time of the Butterworth filter with the smaller oscillation (and overshoot) of the binomial filter. However, in the attempt to extend the method to zero-velocity and zero-acceleration-error systems, step-function responses to the minimum ITAE filter were no better than, and in some cases, worse than,

the binomial filter.

The use of standard forms does not involve solving for the roots of equations, plots or graphical constructions, integration, or inverse Laplace transformations. As the reference suggests, it is a true synthesis method, in that it leads directly to a description of the required system in terms of its design parameters. There are a large number of servos which fall within the imposed restrictions.

6.4 Methods of Zaborszky and Diesel[34,35]

Looking behind all the indirectness of feedback control system design, it develops that actual system design is based on the one ultimate measure of performance: How much error is there at the times when the system output is utilized? Zaborszky and Diesel attempt a mathematical formulation of such a measure of performance. In contrast to most standard techniques which concentrate on isolated phases of performance such as transient or steady state, this measure unites all of these, favoring none, in a single concept. This measure not only concerns itself with the amount of the error, but also the times when the output is being utilized.

This measure is formulated by the following line of reasoning:

(1) The specific environment and function of each control system will set a specific penalty valuation on errors of a given size. This penalty valuation can be expressed in the form of a penalty function F(e), a single-valued function of the error e. Error itself is a function of time, and the valuation may vary with extraneous parameters, or with time.

Therefore let $F(e) \triangleq F\left(e(t), t, v_1, v_2 \cdots v_n\right)$

(2) A second element of performance evaluation is the time when the output is utilized. For simple systems, all times of utilizing the output are generally equally probable, but this is generally not so in advanced or complicated systems. This means that the times elapsing from activation of the system to all times of utilization of its output can be arranged into a probability distribution p(t).

The measure can be mathematically expressed as:

$$S = \int_0^\infty F(e)\, p(t)\, dt$$

where the symbol S is the end sigma, the form of letter sigma used at the end of Greek words. For a deterministic input, the above gives the average value of the penalized error at all times when the output is utilized. If the environment consists of several inputs, or is stochastic, then

$$S = \int_0^\infty \overline{F(e)\, p(t)}\, dt$$

where the bar denotes the averaging over the ensemble. If p(t) is independent of the input (and it frequently is), then

$$S = \int_0^\infty \overline{F(e)}\, p(t) dt$$

Conceptually then, the above integrals are the average value of the penalized error at such times when the output is utilized, and are referred to as the "probabilistic error" or "end sigma error". As a performance index, the integral is called the "end sigma criterion". This measure unites in

a single concept the transient and steady states of operation as well as any intermediate states. None of these states is discriminated for or against, because p(t) determines the weight allotted to each state. The most fascinating thing about the above integral is its complete generality: All of the other performance indices mentioned in this paper are simply special cases of this integral.

It is possible to evaluate this integral in the s (frequency) domain if the functions involved are Laplace transformable. This means it can be used directly with a root locus plot without the necessity of going to the time domain. It can also be evaluated in the time domain in the same manner as indicated in Chapter 4. Use of this criterion for optimizing a system leads again to the Wiener-Hopf integral equation. The second reference indicated gives a complete design process to obtain the impulse response function for the system which has the smallest average square error for such times when its output is used. All of the necessary equations are presented in an organized (albeit complicated) form for the programming of a digital computer. As they indicate, the solution of problems of a complexity common in control engineering practice today, requires computers.

6.5 Methods of Schultz and Rideout[36]

These writers propose a general integral of the form

$$E = \int_0^{\infty} F\left(e(t), t\right) dt$$

which is the same as that of Zaborszky and Diesel

with no consideration to the times of utilization of the output. However, their attitude is the recognition of the fact that physical systems can never respond instantaneously, and therefore the immediate initial error should be ignored by a suitable delay. They express error as follows:

$$e(t,\tau) = r(t-\tau) - c(t)$$

where $r(t-\tau)$ represents the input delayed by an amount τ, and $c(t)$ is the output. This error is called "delay-error" and it is the one used in the indices as follows:

$$E(\tau) = \int_0^\infty F[e(t,\tau), t, \tau]\,dt$$

Their study shows that the integral can be used with transient or stochastic inputs, and that an optimum value of τ exists for a given set of parameters. They have also shown that IADE (Integral of Absolute value of Delay-Error), ISDE (Integral of Square of Delay-Error), and ITADE (Integral of Time-weighted Absolute value of Delay-Error) can all be easily instrumented on an analog computer.

CHAPTER 7

CONCLUSIONS

7.1 General

It is impossible to formulate in detail a universal approach to all servo design problems, but it is possible to list in proper sequence a number of design steps which can serve as a guide. As mentioned in the beginning, feedback control systems have penetrated all walks of life. All kinds of engineers and mathematicians are in the field, all developing methods to fit their own particular problems. There is another large group of persons in this field--this group never really designs servos per se--they look to the field for stimulation and challenge to their ingenuity: the mathematics and graphical and block diagram manipulations are "fun", and a major portion of them are not hard to learn. And so long as the design is a paper design, no "hardware" experience is really necessary. As automation makes further strides, less experience is needed, in one sense, because more and better pieces of hardware become available. Every day it is more nearly possible to realize a compensation which was not realizable a few years ago. But at the same time, two other things are happening: 1) cost control becomes more important, and 2) advanced systems used with refined components require a more sophisticated analysis.

Given a set of specifications, many solutions to the

design problem are possible. Owing to the great variety of techniques and problems, experience, in the final analysis, plays just as important a role as the use of methods. As Gille[16] puts it, the principle most often overlooked is that a feedback control system constitutes an entity, and each component of it must be considered as a part of the overall system. We would presume that this principle never really strikes home until one has attended the school of "hard knocks".

7.2 Analytical Design Methods

There can be no doubt that these techniques are rapidly coming into vogue. The general reason for this is that higher performance is continually being demanded: Feedback Control Systems accomplish all kinds of sophisticated jobs that were not done ten years ago or even conceived twenty years ago. The rapid development of improved components continues to make it easier for the designer to translate a complicated paper design into a useful physical system.

The first hurdle in the use of these techniques is to bring oneself to accept the idea that a performance index can properly express the specifications of the system. Once this has been accepted, application of the methods becomes much more palatable. The only two criteria in extended use are Mean Square Error (MSE) or Integral Square Error (ISE) and Integral Time Weighted Absolute Error (ITAE). For purely analytical techniques only MSE can be used.

Since the objective of the design is the minimization of the performance index, a particular advantage of these methods lies in their ability to recognize inconsistent specifications, in the sense that if the desired performance index is less than the theoretical minimum value, then the specification cannot be fulfilled.

Another advantage lies in the fact that the procedures are readily susceptible to modification to restrict the bandwidth or to limit saturation tendencies.

Another advantage lies in the methods' ability to handle virtually any kind of input or output. Whereas most Trial and Error procedures hinge on the desired output being equal to the input, these methods are generally indifferent as to what the desired output is. This means they are readily able to accommodate noisy signals. For that matter, they can handle any stationary stochastic signal as long as it is possible to obtain the correlation function, or correlation function transform.

Another advantage, albeit a philosophical one, is that the mathematics associated with the methods is the same type of mathematics used in information and statistical theory and in all advanced communication theory. Persons schooled and trained to think in those lines can readily adapt their thinking to the Analytical Design Methods.

Besides the indicated disadvantage of only one performance index, another disadvantage, that could be important sometimes, is the large number of numerical calculations associated with these methods. In one sense,

this is not serious since any method requires numerous calculations if the system is an advanced one; besides, computers are readily available today to do much of this detail.

From the viewpoint of minimizing MSE, these methods bring home very clearly three theoretical performance limitations on linear systems:

1) Noise and disturbances make it impossible for a system to establish equality between the desired and actual output--which means the minimum MSE is not zero under these conditions, nor as low as it could be in the absence of the noise or disturbance.

2) The best compensation cannot overcome the effect of a pure time delay in the fixed elements. This means that a feedback control system cannot predict the future value of a signal with zero error, since it inherently must operate on present or past information.

3) There is no way to eliminate completely the effect of a non-minimum-phase fixed element. Intuitively the cancellation of a zero in the RHP can be approximated but not completely accomplished because of the threat of instability. Thus zero MSE cannot be obtained.

We know that a practical performance limitation is that of saturation. In the Trial and Error Methods, there is a tendency to overlook this very important point until one attains a great deal of experience. In the Analytical Design Methods, saturation can be made to appear in the forefront from the beginning of the design.

7.3 Trial and Error Design Methods

Categorically the transfer function approach will most directly predict frequency response while the root locus approach will most directly predict transient response. Use of the system equation most directly predicts stability. However, all approaches are used in actuality.

One of the drawbacks of frequency response techniques is the difficulty of trying to visualize the transient response. The root locus method fills this gap because the motion of the closed-loop poles can be easily observed as the gain factor is varied. The root locus technique then serves as a fine educational tool also. Another advantage lies in the fact that the compensation employed or being investigated is readily evident and the problem of unrealizable compensation seldom presents itself. Direct determination of stability is also easy whereas the frequency response techniques can sometimes be misleading for indication of stability.

A peculiar advantage of frequency response techniques (or steady state analysis) is that knowledge of the mathematical model is not a requirement and it is particularly easy to select a compensator to cause the response curve to take on the right shape. In root locus techniques (and for that matter, in Analytical Design Methods), the mathematical models of different parts of the servo must be known; in practice it may be difficult to determine the constants required for such a representation.

The major drawback to the root locus is that accurate plotting of the locus is a time consuming task. As indicated earlier, a number of theorems for approximating the locus are available; a number of computing schemes have also been developed including one, by one of the authors, for use on the NCR 102 computer.

Translation from frequency response to transient response is also difficult at best. Here again, digital computers can play an important role in speeding up the arithmetic required in the various approximation methods.

Analog computers of course greatly facilitate analysis and design. In addition, good analog computers are extremely valuable as an aid to remaining in the linear zone, or observing the effects caused by moving out of the linear zone. To recognize their use in this respect is merely to recognize that nature just is not linear. Probably their strongest contribution is their application to solution of optimization problems, after the configuration is fixed.

All of the various aids and charts developed in these techniques come down to one point: they are a scheme for relocating the roots of the closed-loop equation by juggling the roots of the open-loop equation. In one case the given specifications define a closed-loop equation and the problem is to find an open-loop equation to fit it. In the second more advanced case the problem is to find an open-loop equation in which the plant elements fit also.

7.4 Epilogue

The servo designer has to answer three major questions:

a) What is the Frequency Response?

b) What is the Transient Response?

c) Is it stable?

No single method answers these questions. Careful examination of a proponent of a particular method reveals that the method works well for some particular type problem, but has limitations when extended to some other problem.

No matter what method is considered, the first step is the formulation of the problem by gathering the appropriate specifications and plant element descriptions. Perhaps the next best step is the application of Analytical Design Theory to the development of an appropriate formula for compensation. It is then necessary to make some reasonable approximations to reduce the complexity of this formula or the computations associated with it, after which the compensation is still unduly complex and generally unrealizable. At this point the Trial and Error Methods should be injected into the amalgam to ameliorate the situation. Simpler forms of compensation found by these methods will almost always yield performance close to that determined theoretically. How often this is true depends upon the broadness of the minimum of the chosen performance index. Furthermore several different forms of compensation can be found and compared to the theoretical one for their relative efficiencies. The reason that this comparison is possible

is that it is relatively easy to optimize the parameters after the selection of a compensator by Analytical Design Methods, since the system becomes one of a fixed configuration. This can also be done with an analog computer. However, the practical value of optimization can be overemphasized. When there are more than about two free parameters, optimization is "but a modern, more systematic variation of the very old engineering practice of compromise" (Quote from Gille[16]).

Very powerful tools are available today for synthesizing linear feedback control systems. All are of value, providing the designer is well trained in their use. Apparently no optimum set of methods exists, although there exists a broad enough family of methods to arrive at "almost the best possible" servo for a given set of conditions. It would seem that further refinements of the basic methods would have to await additional improvement in servo components.

APPENDIX A

SELECTED DEFINITIONS

1. Dynamic Response - the output response to an input that is a varying function of time.

2. Steady-State-Response - the output response to an input that is constant with time.

3. Transient Response - the time variation of one or more of the system outputs following a sudden change in one or more of the system inputs or the derivatives or integrals of the system inputs. A given transient response must be referred to the type of input that caused it.

4. Frequency Response - the variation of the output to an input which is a constant-amplitude variable-frequency sinusoid.

5. Forced Response - the time response of an output of the system to an arbitrary, but completely defined, variation of one of the system inputs. Forced response is distinguished from transient response in that the input variation associated with the forced response of a system is considered as a continuous time function with no discontinuities in any of its derivatives. (A sinusoidal input is a special case of a forcing input which is isolated for special attention because of its theoretical importance.)

6. Transient Test Inputs -

   a. Impulse - A unit impulse is a time function that is infinite at t=a and zero everywhere else. It is

defined as follows, where $\delta_0(t-a)$ is a unit impulse function occurring at t=a:

$$\int_{-\infty}^{\infty} \delta_0(t-a)\,dt = 1$$

$$\int_{-\infty}^{\infty} \delta_0(t-a)\,f(t)\,dt = f(a)$$

$$\delta_0(t-a) = 0, \quad a < t < a$$

b. Step - the unit step function $\delta_{-1}(t-a)$ is merely the integral of the unit impulse $\delta_0(t-a)$. It is defined as follows, where $\delta_{-1}(t-a)$ is a unit step occurring at t=a:

$$\delta_{-1}(t-a) = \int_{-\infty}^{t} \delta_0(x-a)\,dx$$

$$= \begin{cases} 0, & t < a \\ 1, & t > a \end{cases}$$

c. Ramp - The unit ramp function $\delta_{-2}(t-a)$ is the integral of the unit step $\delta_{-1}(t-a)$. The unit ramp is defined as follows, where $\delta_{-2}(t-a)$ is a unit ramp occurring at t=a:

$$\delta_{-2}(t-a) = \int_{-\infty}^{t} \delta_{-1}(x-a)\,dx$$

$$= \begin{cases} 0, & t < a \\ t, & t > a \end{cases}$$

d. Parabolic - The unit parabolic function $\delta_{-3}(t-a)$ is the integral of the unit ramp $\delta_{-2}(t-a)$. The unit parabolic is defined as follows, where $\delta_{-2}(t-a)$ is a unit parabolic occurring at t=a:

$$\delta_{-3}(t-a) = \int_{-\infty}^{t} \delta_{-2}(x-a)\,dx$$

$$= \begin{cases} 0, & t < a \\ \frac{1}{2}t^2, & t > a \end{cases}$$

e. Displaced cosine - the displaced cosine is one 360° segment of a cosine wave displaced so that it is always positive going:

$$f_{(\text{displaced cosine})} = \left[\delta_{-1}(t-a) - \delta_{1}\left(t-a+\frac{2\pi}{\omega}\right)\right]\left[1-\cos\omega(t-a)\right]$$

Note that the first derivative of the displaced cosine is zero for times t-a and t-a+ $\frac{2\pi}{\omega}$

Note that all of the above functions are equal to zero for all t<a, and that they are discontinuous, or one or more of their derivatives are discontinuous at the instant of occurrence.

7. The Convolution Integral

a. If y(t) is the input, x(t) the output, and w(t) the impulse response of the system, then the output x can be found by evaluating the convolution integral

$$x(t) = \int_{-\infty}^{\infty} w(t_1)\, y(t-t_1)\, dt_1$$

or

$$x(t) = \int_{-\infty}^{\infty} w(t-t_1)\, y(t_1)\, dt_1$$

b. If the system being studied is a physical system, then

$$w(t) = 0 \quad \text{for } t < 0$$

and the convolution integral reduces to:

$$x(t) = \int_{0}^{\infty} w(t_1)\, y(t-t_1)\, dt_1$$

or

$$x(t) = \int_{-\infty}^{t} w(t-t_1)\, y(t_1)\, dt_1$$

c. If y(t) and w(t) are both zero for t<0, then the convolution integral reduces to

$$x(t) = \int_0^t w(t_1)\, y(t - t_1)\, dt_1$$

or

$$x(t) = \int_0^t w(t - t_1)\, y(t_1)\, dt_1$$

8. The Fourier Transform - The Fourier transform of a function and its inverse are defined as follows:

$$\mathcal{F}[f(t)] \triangleq F(s) = \int_{-\infty}^{\infty} \epsilon^{-st} f(t)\, dt$$

$$\mathcal{F}^{-1}[F(s)] \triangleq f(t) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \epsilon^{st} F(s)\, ds$$

where s = the complex variable $\sigma + j\omega$

The Fourier transform is applicable to functions that exist for all time t. To insure the existence of the Fourier transform of a function, Dirichlet's conditions must be satisfied.

9. The Laplace Transform - The Laplace transform of a function and its inverse are defined as follows:

$$\mathcal{L}[f(t)] \triangleq F(s) \triangleq \int_0^{\infty} \epsilon^{-st} f(t)\, dt$$

$$\mathcal{L}^{-1}[F(s)] \triangleq f(t) \triangleq \frac{1}{2\pi j} \int_{c-j\omega}^{c+j\omega} \epsilon^{st} F(s)\, ds$$

where s = the complex variable $\sigma + j\omega$. Note that the Laplace transform is used for functions that are zero for $t < 0$. The constant c is used in the inverse as a convergence factor that enables one to apply the Laplace transform to functions whose Fourier transforms do not exist. A function must also satisfy the Dirichlet conditions to be Laplace transformable.

10. Gain - Gain of a system or element is the ratio of magnitude of the output with respect to the magnitude of

sinusoidal input. The frequency and conditions of operation and measurement must be specified.

11. Nyquist Diagram - The Nyquist Diagram is a closed polar plot of a loop transfer function from which stability may be determined. For a single-loop system, it is a map on the F(s) plane of an s-plane contour which encloses the entire right half of the s-plane, excluding poles for the loop transfer function which lie on the imaginary axis.

12. Response time - Response time is the time required for the output first to reach a specified value after the application of a step input or disturbance.

13. Rise time - Rise time is the time required for the output to increase from one specified percentage of the final value to another, following the application of a step input. Usually the specified percentages are 10% and 90%.

14. Settling Time - The settling time of a system or element is the time required for the absolute value of the difference between the output and its final value to become and remain less than a specified amount, following the application of a step input or disturbance. The specified amount is often expressed in terms of per cent of the final value.

15. Gain Margin - Gain margin is the amount by which the magnitude of the loop ratio of a stable system is different from unity at phase crossover; it is usually expressed in decibels.

16. Phase Crossover - Phase crossover is a point on the

plot of loop ratio at which its phase angle is 180°.

17. Phase Margin - Phase margin is the angle by which the phase of the loop ratio of a stable system differs from 180°.

18. Gain Crossover - Gain crossover is a point in the plot of loop ratio at which the magnitude of the loop ratio is unity.

19. Loop Ratio - Loop Ratio is the frequency response of the primary feedback to the actuating signal. Under linear conditions, the ratio is expressed as GH where G represents the forward elements and H the feedback elements.

APPENDIX B

THE FORMULATION OF THE TRANSLATION AND CORRELATION FUNCTIONS

B-1 Probability Density Functions

The analysis of stochastic signals requires the use of probability density functions and other statistical characterizations such as the average value, the root-mean-square value or mean-square value, and the correlation functions.

The probability density functions are direct measures of the chance of occurrence of certain events in a process. The first probability density function of a stochastic (or random) variable v(t) is denoted by

$$P_1(v_1,t_1) \triangleq \text{probability that the variable has a value } v_1 \text{ at time } t_1$$

The second probability density function is denoted by

$$P_2(v_1,t_1;v_2t_2) \triangleq \text{probability that the variable has a value } v_1 \text{ at time } t_1 \text{ and a value } v_2 \text{ at time } t_2 \text{ simultaneously.}$$

For a stationary stochastic process (one whose statistics are independent of time), $P_1(v_1,t_1)$ is independent of time t_1; $P_2(v_1,t_1;v_2,t_2)$ is a function only of the time difference (t_2-t_1). (Note then, that a process can be defined with one less variable if it is stationary).

In general, the average or mean value of a stochastic variable v(t) is given by

$$\overline{v(t)} \triangleq \int_{-\infty}^{\infty} v\, P(v,t)\, dv$$

The mean-square value is given by:

$$\overline{v^2(t)} \triangleq \int_{-\infty}^{\infty} v^2 P(v,t)\, dv$$

The root-mean-square (rms) value is given by the square root of the mean-square value. The variance of a stochastic process is given by

$$V \triangleq \overline{[v - \bar{v}]^2}$$

The standard deviation σ is the square root of the variance. It can be expressed as follows:

$$\sigma = \left[\overline{v^2} - \bar{v}^2\right]^{1/2}$$

Since the statistics of a stationary stochastic variable are independent of time, the mean value is:

$$\overline{v(t)} \triangleq \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} v(t)\, dt$$

and the mean-square value is given by

$$\overline{v^2(t)} \triangleq \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} v^2(t)\, dt$$

Two commonly used probability density functions are the normal distribution and the Poisson distribution. The normal distribution is given by:

$$P(v)\, dv = \frac{1}{\sigma\sqrt{2\pi}} \epsilon^{-\frac{1}{2}\left(\frac{v-\bar{v}}{\sigma}\right)^2} dv$$

where P(v)dv is the probability of finding v between v and v + dv.

The Poisson distribution is given by

$$P(n,\Delta t) = \frac{(\nu \Delta t)^n \epsilon^{-\nu \Delta t}}{n!}$$

where $P(n, \Delta t)$ is the probability of finding n events in a time interval Δt, and ν is the average frequency of occurrence of the events.

B-2 Correlation Functions for Stationary Stochastic Signals

The autocorrelation function $\varphi_{vv}(\tau)$ of a stationary stochastic process v(t) is defined as the mean value of the product of function v at time t by the function v at time $(t + \tau)$:

$$\varphi_{vv}(\tau) = \overline{v(t)\, v(t+\tau)} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} v(t)\, v(t+\tau)\, dt$$

A function analogous to the autocorrelation function for a single signal is the cross-correlation function for a pair of signals. The cross correlation function $\varphi_{vu}(\tau)$ between two stationary stochastic processes v(t) and u(t) is defined as the mean value of the product of the function v at time t by the function u at time $(t + \tau)$:

$$\varphi_{vu}(\tau) = \overline{v(t)\, u(t+\tau)} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} v(t)\, u(t+\tau)\, dt$$

Some of the useful properties of correlation functions are:

(1) $\varphi_{vv}(\tau) = \varphi_{vv}(-\tau)$ (even function)

(2) $|\varphi_{vv}(\tau)| \leq \varphi_{vv}(0)$

(3) $\lim_{\tau \to \infty} \varphi_{vv}(\tau) = \overline{v}^2$

(For a stationary signal this means the autocorrelation function approaches the square of the mean value of the signal as τ approaches infinity.)

(4) $\varphi_{vu}(\tau) = \varphi_{uv}(-\tau)$ (not an even function)

(5) $|\varphi_{vu}(\tau)| \leq \sqrt{\varphi_{vv}(0)\,\varphi_{uu}(0)}$

(6) $\lim_{\tau \to \infty} \varphi_{vu}(\tau) = \overline{v} \cdot \overline{u}$

We observe from the definition of the autocorrelation function, that the mean square value of the signal equals the value of the corresponding autocorrelation function with zero argument: $\varphi_{vv}(0) = \overline{v^2}$

B-3 Examples of Correlation Functions

(a) Example 1

A common type stochastic variable used is the case of v(t) as a rectangular wave with values $+\beta$ and $-\beta$ and with zero crossings located at event points that are Poisson-distributed in time. This autocorrelation function can be derived as:

$$\varphi_{vv}(\tau) = \beta^2 \epsilon^{-2\lambda|\tau|}$$

(b) Example 2

v(t) is a rectangular wave with amplitude values distributed in any fashion and with zero crossings Poisson-distributed in time:

$$\varphi_{vv}(\tau) = \sigma^2 \epsilon^{-\lambda|\tau|} + \overline{v(t)}^2$$

where σ is the standard deviation of the amplitude distribution and $\overline{v}$ is the mean value of the amplitude distribution.

(c) Example 3

v(t) is a train of identical finite pulses whose starting points are Poisson-distributed in time (known as "shot noise"):

$$\varphi_{vv}(\tau) = \nu \int_{-\infty}^{\infty} f(t)\, f(t+\tau)\, d\tau + \overline{v(t)}^2$$

where f(t) is the waveform of a single pulse and $\overline{v} = \nu \int_{-\infty}^{\infty} f(t)\, dt$. (This derivation is an extension of Campbell's Theorem- see p.102, reference 2)

(d) Example 4

v(t) is pure or white noise:

$$\varphi_{vv}(\tau) = \gamma\, \delta_0(\tau)$$

where γ is a constant that depends on how the process is generated, and $\delta_0(t)$ is an impulse. For example, if white noise is considered as a limiting case of shot noise generated by exponential pulses of amplitude A, time constant T, and area s under the pulse, then

$$\gamma = \frac{\nu s}{2} \quad \text{as} \quad \begin{cases} A \to \infty \\ T \to 0 \\ s = \text{constant} \end{cases}$$

B-4 Correlation Function Transforms

Because correlation functions are completely defined as functions of a time variable τ, they are Fourier transformable. By convention, $\frac{1}{2\pi}$ times the Fourier transform of a correlation function is called a power spectrum or a power-density spectrum. Since the

correlation functions involve convolution of one or more functions, this means that we may multiply the functions in the frequency plane and this is the significant advantage of taking their transforms.

The power density spectrum $\Phi_{vv}(s)$ of a stochastic process is defined as

$$\Phi_{vv}(s) \triangleq \frac{1}{2\pi}\int_{-\infty}^{\infty} \epsilon^{-s\tau}\varphi_{vv}(\tau)\,d\tau$$

The cross-power-density spectrum $\Phi_{vu}(s)$ between two stochastic processes v(t) and u(t) is defined as

$$\Phi_{vu}(s) \triangleq \frac{1}{2\pi}\int_{-\infty}^{\infty} \epsilon^{-s\tau}\varphi_{vu}(\tau)\,d\tau$$

The inverse transformations are

$$\varphi_{vv}(\tau) = \frac{1}{j}\int_{-j\infty}^{j\infty} \Phi_{vv}(s)\,\epsilon^{s\tau}\,ds$$

$$\varphi_{vu}(\tau) = \frac{1}{j}\int_{-j\infty}^{j\infty} \Phi_{vu}(s)\,\epsilon^{s\tau}\,ds$$

Since the above integrations are along the imaginary axis of the s-plane, the inverse transform may be rewritten in terms of the real frequency ω:

$$\varphi_{vv}(\tau) = \int_{-\infty}^{\infty} \Phi_{vv}(j\omega)\,\epsilon^{j\omega\tau}\,d\omega$$

Recalling that the mean-square value of the signal is the value of the autocorrelation function with $\tau = 0$, then

$$\overline{v^2} = \varphi_{vv}(0) = \int_{-\infty}^{\infty} \Phi_{vv}(j\omega)\,d\omega$$

Because the correlation function is an even function of τ,

$\Phi_{vv}(j\omega)$ can be written as $\Phi_{vv}(\omega)$, or $\varphi_{vv}(0) = \int_{-\infty}^{\infty} \Phi_{vv}(\omega) d\omega$

This means that the area underneath the frequency function $\Phi_{vv}(\omega)$ over the infinite frequency range (commonly accepted definition of power-density spectrum) is equal to the mean-square value of the signal.

Useful properties of the power spectra are

$$\Phi_{vv}(s) = \Phi_{vv}(-s) \qquad \text{(even function)}$$

$$\Phi_{vu}(s) = \Phi_{uv}(-s)$$

B-5 Translation Functions

Since transient type signals exhibit no statistical properties, they are not subject to description by correlation functions. Newton introduces the use of translation functions (see page 51, reference 2) for this type signal.

If $x_1(t)$ is an arbitrary transient signal, then the autotranslation function $I_{11}(\tau)$ is defined as:

$$I_{11}(\tau) \triangleq \int_{-\infty}^{\infty} x_1(t)\, x_1(t+\tau)\, dt$$

If $x_1(t)$ ans $x_2(t)$ are two arbitrary transient signals, then their cross-translation function $I_{12}(\tau)$ is defined as:

$$I_{12}(\tau) \triangleq \int_{-\infty}^{\infty} x_1(t)\, x_2(t+\tau)\, dt$$

Although these functions characterize the signal, this characterization is not unique; there are a number of different functions that can give rise to the same translation function. Useful properties of the translation function are

$$I_{11}(\tau) = I_{11}(-\tau) \qquad \text{(even function)}$$

$$I_{12}(\tau) = I_{21}(-\tau)$$

Note from the definition of the autotranslation function, that if $\tau = 0$,

$$I_1 \triangleq I_{11}(0) = \int_{-\infty}^{\infty} x_1^2(t)\,dt$$

where I_1 represents the integral-square value of the time function $x_1(t)$

B-6 Translation Function Transforms

Because of Parseval's Theorem (see p.44, reference 2), the integral I_1, stated above, can be expressed in terms of its transform as

$$I_1 = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} X_1(-s)\, X_1(s)\,ds$$

The Fourier transform of a translation function $I_{12}(\tau)$ is defined as follows:

$$I_{12}(s) \triangleq \int_{-\infty}^{\infty} \epsilon^{-s\tau} I_{12}(\tau)\,d\tau$$

The inverse transform is defined as

$$I_{12}(\tau) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} I_{12}(s)\, \epsilon^{s\tau}\,ds$$

Applying this definition to I_1:

$$I_1 \triangleq I_{11}(0) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} I_{11}(s)\,ds$$

It can be shown (see p. 57, reference 2) that

$$I_{11}(s) = X_1(-s)\, X_1(s).$$

The transforms of translation functions are sometimes called energy density spectra as contrasted to power density spectra in correlation functions. It is to be noted that the

translation functions for many commonly encountered transient signals are infinite since their defining integrals do not converge. Particular examples are the autotranslation functions for the step function and the ramp. Fourier transforms for these functions also do not exist. In such situations the introduction of a convergence factor will frequently permit a solution to be obtained. However, in general, the determination of the integral square value of a function is usually done in the time domain unless the Fourier transform is a rational function.

BIBLIOGRAPHY

1. Truxal, J. G., Editor-in-Chief, Control Engineers Handbook, McGraw-Hill Co., New York, 1958.
2. Newton, Gould, and Kaiser, Analytical Design of Linear Feedback Controls, Wiley and Sons, New York, 1957.
3. Thaler and Brown, Analysis and Design of Feedback Control Systems, McGraw-Hill Co., New York, 1960.
4. Army Ordnance Corps Pamphlet ORDP20-136, 137,138,139, Ordnance Engineering Design Handbook - Servomechanisms, August 1959.
5. James, Nichols and Phillips, Theory of Servomechanisms, McGraw-Hill Co., New York, 1947.
6. Report PEL216/1 Ad Hoc Working Group on Basic Research in Electronics, Technical Advisory Panel on Electronics, Office of the Assistant Secretary of Defense, Research and Development, 22 May 1956.
7. Stout, T. M., "A Note on Control Area", J. Appl. Physics, Vol. 21, pp 1129-1131, Nov. 1950.
8. Fickieson, F. C., and Stout, T. M., "Analogue Methods for Optimum Servomechanism Design", Trans, AIEE, Vol. 71, Part II, pp 244-250, 1952.
9. Hall, A. C., Analysis and Synthesis of Linear Servomechanisms, Technology Press, Cambridge, Mass., 1943.
10. Sartorius, H., "Die zweckmassige Festlegung der frei whalbaren Regelungskoustauten", dissertation at the Technische Hochschule, Stuttgart, Germany, 1944.
11. Nims, P. T., "Some Design Criteria for Automatic

Controls", Trans. AIEE, Vol. 70, Part I, pp 606-611, 1951.

12. Graham, D., and Lathrop, R. C., "The Synthesis of 'Optimum' Transient Response: Criteria and Standard Forms", Trans. AIEE, Vol. 72, Part II, pp 273-288, 1953.

13. Oldenbourg, R. and Sartorius, H., "A Uniform Approach to the Optimum Adjustment of Control Loops" ASME Paper 53-A-18, 1954

14. Wiener, N., "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series", NDRC Report, Cambridge, Mass., 1942. (since republished by Wiley)

15. Laning and Battin, Random Processes in Automatic Control, McGraw-Hill Co., New York, 1956.

16. Gille, Pelegrin and Decaulne, Feedback Control Systems, Analysis, Synthesis, and Design, McGraw-Hill Co., 1959.

17. "Approximation of $\varphi_{vv}(\tau)$ by Series of Damped Cosine Functions", National Electronics Conference Proceedings, Vol 14, p 709, 1958.

18. Evans, W. R., "Graphical Analysis of Control Systems", Trans. AIEE, Vol. 67, pp 547-551, 1948.

19. Grabbe, Ramo, Wooldridge, Editors, Handbook of Automation Computation and Control, New York, 1958.

20. Nixon, F. E., Principles of Automatic Control, Prentice-Hall, Englewood Cliffs, N. J., 1953.

21. Elgerd and Stephens, W., "Effect of Closed-Loop Transfer Function Pole and Zero Locations on the Transient

Response of Linear Systems", Trans. AIEE, Vol. 78, Pt II, pp. 121-127, 1959.

22. Abbott, W. B., and Patton, W. B., "Rapid Approximation of Servomechanism Transient Response", Masters Thesis, M.I.T., 1961.

23. Chu, Yaohan, "Synthesis of Feedback Control Systems by Phase-Angle Loci", Trans. AIEE, Vol. 71, Pt II, pp 330-339, 1952.

24. Wheeler, R. C. H., "Lecture Notes, USNPG School, Monterey, Calif., Course EE672", Unpublished, 1959.

25. Truxal, J. G., Automatic Feedback Control System Synthesis, McGraw-Hill Co., New York, 1955.

26. Liether, Houpis and D'Azzo, "An Automatic Root Locus Plotter Using an Analog Computer", Trans. AIEE, Applications and Industry, pp 523-527, January 1961.

27. Ross, Warren and Thaler, "Design of Servo Compensation Based on the Root Locus Approach", Trans. AIEE, Applications and Industry, pp 272-277, September 1960.

28. Walters, L. G., "Optimum Lead-Controller Synthesis in Feedback Control Systems", Trans. IRE-PGCT-1, pp 45-48, March 1954.

29. Aseltine, J. A., "On the Synthesis of Feedback Systems with Open-Loop Constraints", Trans. IRE-PGAC-4, no. 1, pp 31-36, May 1959.

30. Mariotti, Franco, "A Direct Method of Compensating Linear Feedback Systems", Trans. AIEE, Applications and Industry, pp 527-538, January 1961.

31. Mitrovič, Dušan, "Graphical Analysis and Synthesis of Feedback Control Systems", Trans. AIEE, Vol. 76, Pt II, pp 476-503, 1959.

32. Chen, C. F. and Shen, D. W. C., "A New Chart Relating Open-Loop and Closed-Loop Frequency Response of Linear Control Systems", Trans. AIEE, Vol. 78, Pt II, pp 252-255, 1959.

33. Levadi, V. S., "Simplified Method of Determining Transient Response from Frequency Response of Linear Networks and Systems", Trans. IRE-PGAC-4, no. 2, pp 55-66, November, 1959.

34. Zaborszky and Diesel, "Measure of Control-System Performance", Trans. AIEE, Vol. 78, Pt II, pp 163-168, 1959.

35. Zaborszky and Diesel, "Control Design for Minimum Probabilistic Error", Trans. AIEE, Applications and Industry, pp 44-54, May 1960.

36. Schultz, W. and Rideout, C, "A General Criterion for Servo Performance", National Electronics Conference Proceedings, vol. 13, pp 549-560, 1957.