

PRINCIPLES OF ANALYSIS OF PHYSICAL SYSTEMS

Notes from a series of lectures presented by Professor Elias P. Gyftopoulos, Associate Professor of Nuclear Engineering, Massachusetts Institute of Technology.

Compiled by

R. W. Garner

Special Power Excursion Reactor Test Project

PHILLIPS
PETROLEUM
COMPANY



ATOMIC ENERGY DIVISION

NATIONAL REACTOR TESTING STATION
US ATOMIC ENERGY COMMISSION

Garn-4-62A
May 14, 1962

PRINCIPLES OF ANALYSIS OF PHYSICAL SYSTEMS

Notes from a series of lectures presented by Professor Elias P. Gyftopoulos, Associate Professor of Nuclear Engineering, Massachusetts Institute of Technology.

Compiled by

R. W. Garner

Special Power Excursion Reactor Test Project

INTRODUCTORY REMARKS

These notes were compiled from hand-written notes and tape recordings of a series of lectures presented by Dr. Elias P. Gyftopoulos, Associate Professor of Nuclear Engineering, Massachusetts Institute of Technology. The lectures were conducted at the SPERT facilities at the National Reactor Testing Station during the period from July 17, to July 28, 1961. Professor Gyftopoulos also conducts a one year course in Control Theory (based on similar material) at MIT.

The purpose of these lectures was to describe, from the standpoint of control theory, various methods of interpreting experimental results. If one desired to attach a title to the subject matter presented herein, a most appropriate one would be "Principles of Analysis of Physical Systems".

In transcribing the notes, emphasis was placed on preserving the mood and spontaneity of presentation. This was made possible, in large part, by a great deal of effort by Professor Gyftopoulos in editing the entire series of notes.

TABLE OF CONTENTS

	<u>Page No.</u>
LECTURE NO. I	
Fourier Series	1
Aperiodic Functions (Fourier and Laplace Transforms)	8
Theory of Functions of Complex Variables	15
Properties of Laplace Transforms	24
LECTURE NO. II	
Analysis of Linear Systems	36
System Function - Transfer Function	48
Block Diagrams	52
Feedback Systems	54
Flow Graphs	56
LECTURES NO. III and IV	
The Inverse Laplace Transform	61
Complete Response of a Linear System	71
Steady State Response	76
Bode Diagrams	79
Stability of Linear Systems - Transient Response	85
Nyquist Stability Criterion	86
Examples	97
Root-Locus Diagrams	103
LECTURE NO. V	
Statistics	128
Correlation Functions	132
Autocorrelation Function	132

	<u>Page No.</u>
Crosscorrelation Function	139
Autocorrelation Function in the Frequency Domain (Power-Density Spectra)	143

LECTURES NO. VI and VII

Applications of Geometric Theory to Nonlinear Reactor Dynamics . .	148
Geometric Theory of Autonomous Differential Equations	150
Dynamics of Xenon Controlled Reactors	158
Dynamics of Reactors with Two Temperature Coefficients . . .	175

LECTURES NO. VIII and IX

Analysis of Nonlinear Systems	190
---	-----

Lecture No. I

FOURIER SERIES

Let's assume that someone has taken a series of measurements of some (at the present, undefined) physical quantity. The measurements were taken at different times and the magnitudes of the values obtained plotted as a function of time. The resulting graph is shown in Figure 1 where M denotes the magnitude in arbitrary units, and t refers to real time.

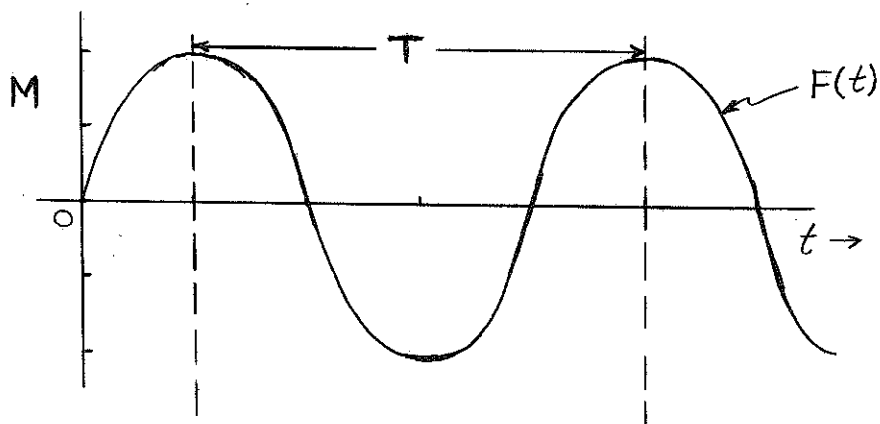


Fig. 1

Suppose that close examination of the graph shows that for every interval of time T, the magnitude repeats itself; i.e., the function obtained is periodic in time. Let's call this function $F(t)$.

Next, let us try to determine an analytical expression to describe the function $F(t)$ which will enable us to predict its value at any time. This means that, for a given value of the function at some time t_1 , we want to be able to say what value it will have at a time t_2 . There are several methods of approaching this problem, but let's assume that the function could be approximated by the Fourier series,

$$f(t) = \frac{a_0}{2} + a_1 \cos \omega t + a_2 \cos 2\omega t + \dots + a_n \cos n\omega t + b_1 \sin \omega t + b_2 \sin 2\omega t + \dots + b_n \sin n\omega t. \quad (1)$$

One might ask, why were sines and cosines chosen to represent the function? First, since the Fourier series contains both sines and cosines it is representative of both even and odd functions, and thus the origin of time is not important. Second, sines and cosines are orthogonal functions in the range $(0, T)$. This means that the integral of the products of any two sines or cosines satisfies the following relationships:

$$\left. \begin{aligned} \int_0^T \sin n\omega t \sin m\omega t dt &= 0 ; (m \neq n) , \\ \int_0^T \cos n\omega t \cos m\omega t dt &= 0 ; (m \neq n) , \\ \int_0^T \sin n\omega t \cos m\omega t dt &= 0 . \end{aligned} \right\} (2)$$

Another reason for using a Fourier series is that if the coefficients are evaluated by

$$\left. \begin{aligned} a_0 &= \frac{1}{T} \int_z^{z+T} F(t) dt , \\ a_n &= \frac{2}{T} \int_z^{z+T} F(t) \cos n\omega t dt ; n > 0 , \\ b_n &= \frac{2}{T} \int_z^{z+T} F(t) \sin n\omega t dt ; n > 0 , \end{aligned} \right\} (3)$$

Then we have the very important result that the error involved between $F(t)$ and $f(t)$ is a minimum in the least mean square sense. This means that

$$\frac{1}{T} \int_0^T (F(t) - f(t))^2 dt \quad \text{is a minimum}$$

and the error is given by

$$\begin{aligned} \overline{\varepsilon^2} &= \overline{F^2(t)} - \sum_i \frac{(a_i^2 + b_i^2)}{2} \\ &= \int_0^T F^2(t) dt - \sum_i \frac{(a_i^2 + b_i^2)}{2}, \quad (4) \end{aligned}$$

where the bar denotes average values. In other words, the approximation is "good" in the sense that $f(t)$ is a least square fit. From Equation (4) it is seen that the more terms involved in $F(t)$ the smaller the error.

A final reason for choosing the Fourier series is that each term is independent of all the others. Thus, adding more terms to decrease the error does not change the values already obtained for the previous terms.

The advantages of an orthogonal set of functions such as the sines and cosines just described may be better appreciated in terms of a geometrical picture. In general, a series in terms of orthogonal functions such as

$$F(t) \approx f_a^N(t) = a_1 \phi_1(t) + a_2 \phi_2(t) + \dots + a_N \phi_N(t) + \dots,$$

where $\phi_2(t)$ is a member of a set of orthogonal functions over the range τ and $\tau+T$, and

$$a_i = \frac{1}{T} \int_{\tau}^{\tau+T} F(t) \phi_i(t) dt \quad ,$$

is a least mean square error fit to $F(t)$. Consider $\phi_i(t)$ to be associated with one and only one direction in a multidimensional space. The directions of this space are perpendicular to each other because the $\phi_i(t)$'s are orthogonal. In fact, if the $\phi_i(t)$'s are normalized:

$$\frac{1}{T} \int_{\tau}^{\tau+T} \phi_n(t) \phi_m(t) dt = \delta_{nm}$$

where the Kronecker delta is defined by

$$\begin{aligned} \delta_{nm} &= 1 \quad ; \quad n = m \\ &= 0 \quad ; \quad n \neq m \end{aligned}$$

Then the terms $a_i \phi_i(t)$ are nothing else than the projections of $F(t)$ along the different axes. These projections are independent of each other and the more projections taken into account, the better the representation of $F(t)$. We will make further use of this geometrical representation of a function in later sections.

It should be emphasized at this point that the Fourier series does not give an exact representation for just any function $F(t)$, regardless of how many terms are used. As an example, Figure 2 shows a comparison of a Fourier series to a square wave.

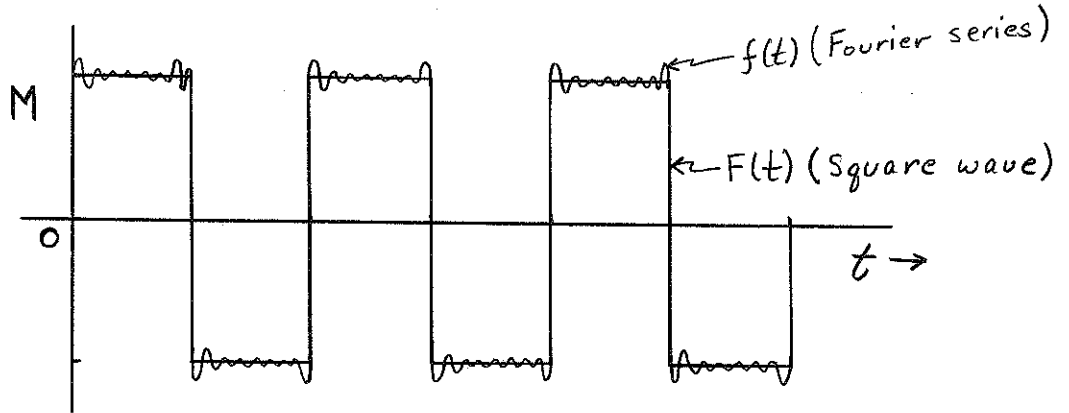


Fig. 2

By the Fourier series approximation there is always an appreciable overshoot (Gibbs' phenomenon). Another comparison is shown in Figure 3 where Tchebycheff polynomials are used to approximate a square wave.

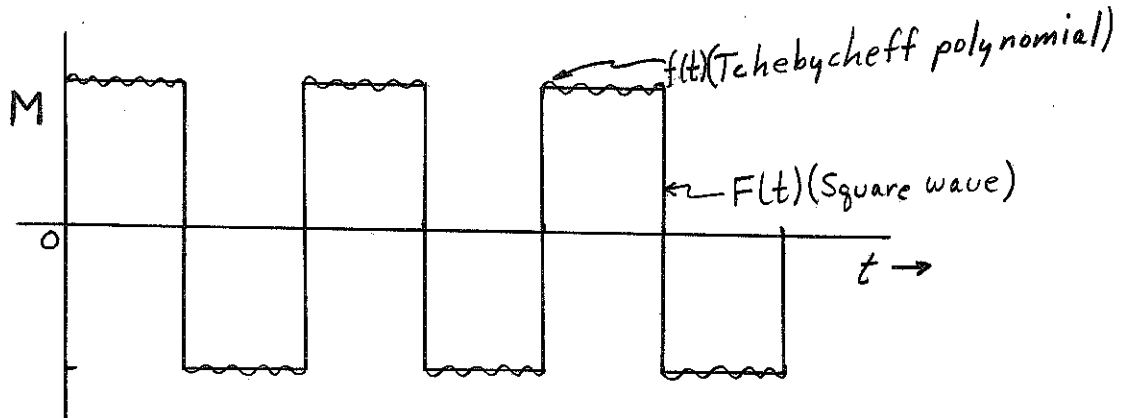


Fig. 3

A first glance at Figure 3 might indicate that Tchebycheff polynomials are a better approximation to a square wave than a Fourier series. However, it can be shown that the error involved in Figure 3 is greater than that of Figure 2. The point is, before attempting to approximate a function $F(t)$ by another function $f(t)$, be it an infinite series or what have you, considerable thought should be given to finding an orthogonal set which will introduce the least possible error. In the case of the square wave,

one should use a square wave, or even perhaps a combination of Walsh functions, which are shown in Figure 4, in order to make the error zero or small.

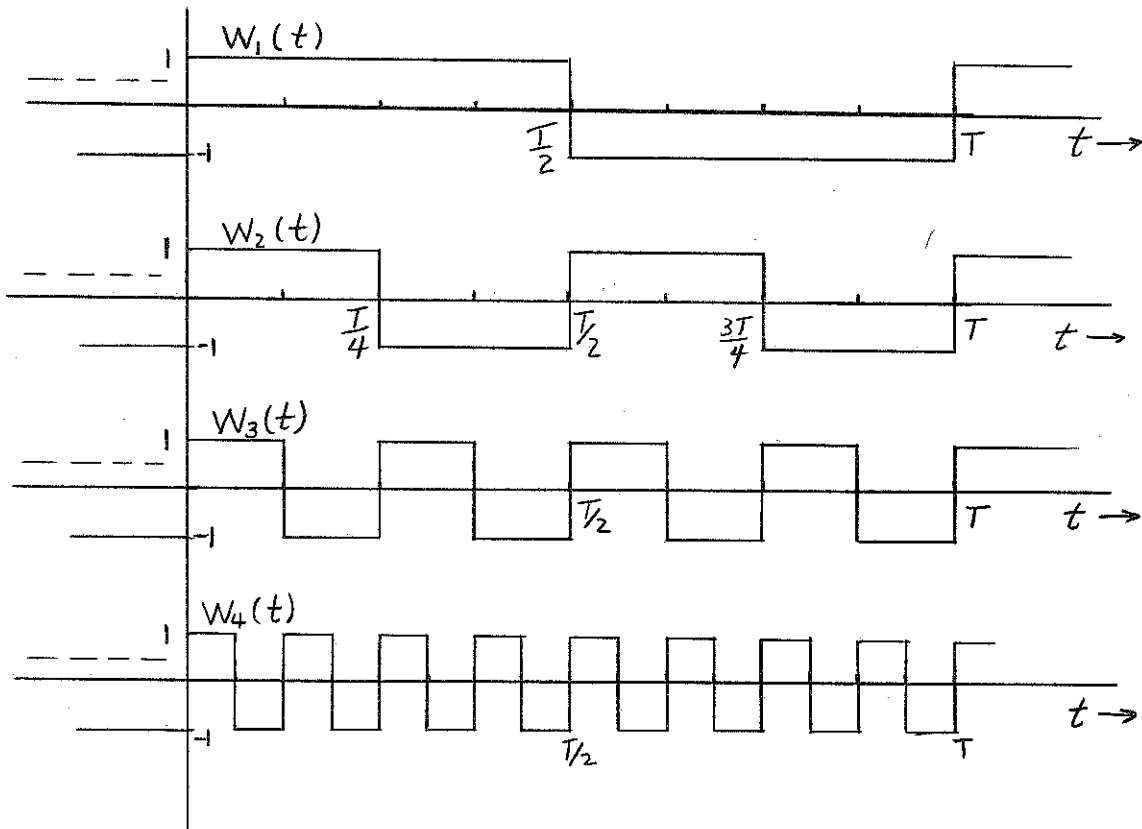


Fig. 4

The Fourier series

$$f(t) = \frac{a_0}{2} + \sum_{i=1}^{\infty} a_i \cos i\omega t + \sum_{i=1}^{\infty} b_i \sin i\omega t \quad (5)$$

may also be written in a more useful form by utilizing the relationship between the rectangular and exponential expressions for complex numbers.

Any complex number

$$Z = x + jy$$

can be written as

$$Z = |Z| (\cos \theta + j \sin \theta) \quad , \quad (6)$$

which is known as the polar form. If $\cos \theta + j \sin \theta$ is expanded in a Taylor series, then

$$\cos \theta + j \sin \theta = 1 + \frac{j\theta}{1!} + \frac{(j\theta)^2}{2!} + \dots + \frac{(j\theta)^n}{n!} + \dots \quad (7)$$

The series expansion of the exponential $e^{j\theta}$ gives

$$e^{j\theta} = 1 + \frac{j\theta}{1!} + \frac{(j\theta)^2}{2!} + \dots + \frac{(j\theta)^n}{n!} + \dots \quad (8)$$

Since Equations (7 and 8) are identical, we have

$$|Z| e^{j\theta} = |Z| (\cos \theta + j \sin \theta) \quad . \quad (9)$$

Similarly,

$$|Z| e^{-j\theta} = |Z| (\cos \theta - j \sin \theta) \quad . \quad (10)$$

Equations (9 and 10) are called the polar forms of complex numbers. Therefore, Equation (5) can be written as

$$f(t) = \frac{1}{T} \sum_{n=-\infty}^{+\infty} P_n e^{jn\omega t} \quad , \quad (11)$$

where

$$P_n = \int_{\tau}^{\tau+T} F(t) e^{-jn\omega t} dt \quad (12)$$

$$T = \frac{2\pi}{\omega}$$

and

$$\left. \begin{aligned} P_n + P_{-n} &= a_n T \\ P_n - P_{-n} &= j b_n T \end{aligned} \right\} (13)$$

The conclusions to be drawn from this form of the Fourier series are the following:

1. By determining the values of the P_n 's we can describe a function $F(t)$. Thus P_n is a transform.
2. If we plot $|P_n|$ and the angle $\angle P_n$ for various values of n , we determine a set of frequencies necessary to represent a periodic function, and the function may be represented by these discrete frequencies only. Thus P_n is an equivalent representation of $F(t)$ in the frequency domain.

At this point one might say, "O.K., so we know how to represent a periodic function, what do we do if the function to be approximated is not periodic?"

APERIODIC FUNCTIONS (FOURIER AND LAPLACE TRANSFORMS)

Since, by definition, a periodic function must necessarily repeat itself with a period T over the time scale from $-\infty$ to $+\infty$, in any practical situation we can never have truly periodic functions. Still

we see that we can approximate what we call periodic functions. However, many times the function we want to represent does not repeat itself consistently after any reasonable interval of time; i.e., one cannot assign a period T to the function. We would like to know then if we can represent an aperiodic (non-periodic) function in a manner similar to that for the periodic case.

Suppose we have a function $F(t)$ which looks like the one shown in Figure 5. Assume the scale has been expanded for purposes of illustration.

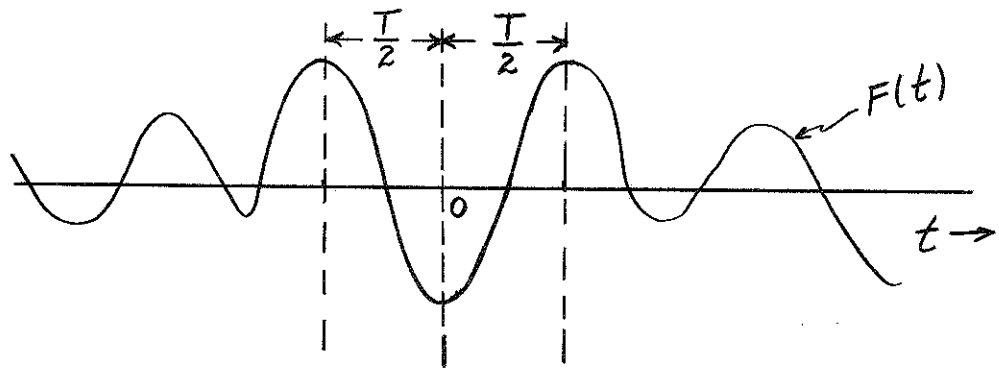


Fig. 5

Consider only a small section of the curve in Figure 5 and assume that this part of the function is a member of a periodic function and represents one period (as shown). Now make a Fourier series expansion over this small interval. Thus,

$$P_n = \int_{-T/2}^{T/2} F(t) e^{-jn\omega t} dt \quad , \quad (14)$$

where $\omega = \frac{2\pi}{T}$. Now let the period T increase gradually. We may then write

$$(n+1)\omega - n\omega = \omega = \frac{2\pi}{T} \rightarrow \Delta\omega$$

where $\Delta\omega \rightarrow 0$ as $T \rightarrow \infty$. Then, we can write

$$P(\omega) = \int_{-\infty}^{\infty} F(t) e^{-j\omega t} dt, \quad (15)$$

so that ω changes continuously rather than discretely. Equation (15) is called the "Fourier transform of $F(t)$ ".

Then from Equation (11) for periodic functions,

$$F(t) = \frac{1}{2\pi} \sum_n \omega P_n e^{jn\omega t}$$

and, similarly, for an aperiodic function

$$F(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) e^{j\omega t} d\omega, \quad (16)$$

which is the "inverse Fourier transform".

Thus, from Equations (15 and 16) we see that if we know either $P(\omega)$ or $F(t)$ we can find the other. Also, we see that if a function is aperiodic, a continuous spectrum of frequencies is required to represent it whereas for a periodic function, only a discrete spectrum of frequencies was necessary.

The bilateral integral (15) creates problems of existence. Since many functions of interest start at $t = 0$ and are zero before $t = 0$ we restrict the definition of (15) over the range $0-\infty$. Then

$$P(\omega) = \int_0^{\infty} F(t) e^{-j\omega t} dt. \quad (17)$$

However, Equation (17) does not necessarily converge.

A sufficient condition to guarantee the existence of Equation (17) is the requirement that $F(t)$ belong to a class of functions known as "functions of exponential order". A function $F(t)$ is said to be of exponential order if, for some constant c ,

$$\int_0^{\infty} |F(t)| e^{-ct} dt < M \quad (18)$$

Therefore, even though $F(t)$ may become infinitely large as $t \rightarrow \infty$, we see that $F(t)$ must not grow more rapidly than a multiple of some exponential function of t if Equation (17) is to exist. Then, if we write Equation (17) as

$$P(c, \omega) = \int_0^{\infty} F(t) e^{-j\omega t} e^{-ct} dt \quad (19)$$

and if the restriction, Equation (18), is satisfied, we may say that Equation (19) exists. Thus a new transform is defined.

One might ask here, why not rewrite Equation (19) to give

$$P(c + j\omega) = \int_0^{\infty} F(t) e^{-(c + j\omega)t} dt$$

However, c is a constant and we would be trying to work with a complex function which is not a function of a complex variable. Therefore, instead of making c a constant let it be a variable (not of time) so that c corresponds to the real part of a complex variable,

$$s = \sigma + j\omega$$

Then we have

$$P(\sigma+j\omega) = \int_0^{\infty} F(t) e^{-(\sigma+j\omega)t} dt \quad ; \sigma > \sigma_a$$

which may be written

$$F(s) = \int_0^{\infty} F(t) e^{-st} dt \quad ; \sigma > \sigma_a \quad (20)$$

As we said before, P is a transform and we call F(s) the "Laplace transform of F(t)".

Let's stop for a moment and see what we now have. First, we have a function, F(s), defined by Equation (20) which is a function of a complex variable, $s = \sigma + j\omega$. Second, the integration has been simplified by changing from $-\infty \rightarrow +\infty$ to $0 \rightarrow +\infty$ on the validity of Equation (20) is that $\text{Re}(s) \equiv \sigma > \sigma_a$, which defines the abscissa of absolute convergence, so that the integral will converge. However, as we will see later, it will be possible to remove even this restriction. Thus, we see that for functions F(t) for which F(t) vanishes for $t < 0$, the Fourier transform of Equation (15) becomes formally identical with the Laplace transform of Equation (20) if we replace $j\omega$ by $\sigma + j\omega$. Similarly, the "inverse Laplace transform", which is analogous to Equation (16), is given by

$$F(t) = \frac{1}{2\pi j} \int_{c-j\omega}^{c+j\omega} F(s) e^{st} ds \quad ; c > \sigma_a \quad (21)$$

We can summarize by noting that:

- a. F(s) is an equivalent representation of F(t) in the sense that if we have F(s) we can find F(t). The correspondence is one to one and

works both ways.

b. The transform $F(s)$ is a function of a complex variable s . Since there are many powerful theorems associated with functions of complex variables, we are at liberty to use these to our benefit.

c. The limitation $\text{Re}(s) > \sigma_a$ is only mathematical and, as we will see later, can easily be disregarded.

d. As we will also see later, the representation of $F(t)$ in the complex (not imaginary) frequency domain does not require an infinite number of frequencies.

Now let's consider a few examples of Laplace transforms to convince ourselves that they are analytical functions of complex variables.

Examples of Laplace Transforms

Unit Step Function $U(t)$

The unit step function is defined such that

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

and

$$U(t) = 1 \quad \text{for } t > 0$$

Then

$$F(s) = \int_0^{\infty} U(t) e^{-st} dt = \int_0^{\infty} e^{-st} dt = -\frac{1}{s} e^{-st} \Big|_0^{\infty} = \frac{1}{s}, \quad (22)$$

provided $\text{Re}(s) > 0$ or $\sigma_a = 0$. The necessity that for this case $\text{Re}(s) > 0$ can be seen from the evaluation of the integral.

$$\begin{aligned} -\frac{1}{s} e^{-st} \Big|_0^{\infty} &= -\frac{1}{s} e^{-s\infty} + \frac{1}{s} e^{-s0} \\ &= \frac{1}{s} - \frac{1}{s} e^{-s\infty} \end{aligned}$$

Now

$$|e^{-j\omega\infty}| = 1, \text{ and if } \sigma > 0, e^{-\sigma\infty} \rightarrow 0.$$

Then $\frac{1}{s} e^{-\sigma\infty} \rightarrow 0$ and Equation (22) is valid.

Exponential e^{-at}

For $F(t) = e^{-at}$ with $t > 0$

$$\begin{aligned} F(s) &= \int_0^{\infty} e^{-at} e^{-st} dt = -\frac{1}{s+a} e^{-(s+a)t} \Big|_0^{\infty} \\ &= \frac{1}{s+a} - \frac{1}{s+a} e^{-(s+a)\infty} \\ &= \frac{1}{s+a}, \quad (23) \end{aligned}$$

provided $\text{Re}(s) = \sigma > -a$. Thus for $F(t) = e^{-at}$, the abscissa of absolute convergence is $\sigma_a = -a$.

Sinusoidal Function $\sin \beta t$

For $F(t) = \sin \beta t$

$$\begin{aligned} F(s) &= \int_0^{\infty} \sin \beta t e^{-st} dt = \int_0^{\infty} \frac{e^{j\beta t} - e^{-j\beta t}}{2j} e^{-st} dt \\ &= \frac{\beta}{s^2 + \beta^2}, \quad (24) \end{aligned}$$

provided $\text{Re}(s) > 0$.

Table I lists the abscissas of absolute convergence for these and some other simple functions.

As has already been mentioned, since the Laplace transforms are functions of complex variables, we may be able to make use of the theorems concerning analytic functions of complex variables. It will be advantageous at this time to review briefly some theories of functions of complex variables.

THEORY OF FUNCTIONS OF COMPLEX VARIABLES

A complex number is one with a real part and an imaginary part. If both the real and imaginary parts are variables, the complex number can be represented on what we will call the s-plane as shown in Figure 6.

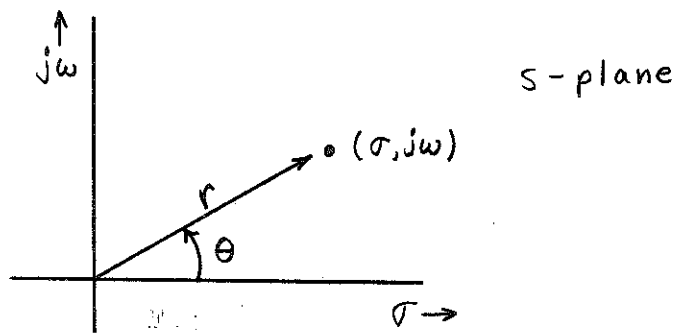


Fig. 6

In vector (polar) notation,

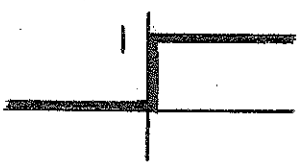

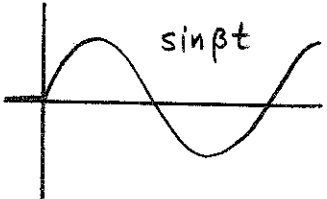
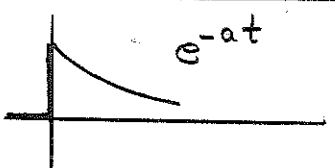
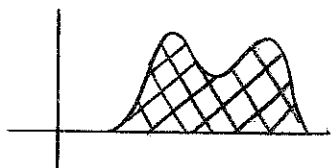
$$s = r e^{j\theta} \quad ; \quad \text{where } r = \sqrt{\sigma^2 + \omega^2} \quad , \quad \theta = \tan^{-1} \frac{\omega}{\sigma}$$

Function of a Complex Variable

Let $s = \sigma + j\omega$ be a complex variable. If there is another number Z , which is also a complex variable and which is so related to s that to each

TABLE I

Abscissas of Absolute Convergence

$f(t)$	$F(s)$	σ_a
	$\frac{1}{s}$	0
	$\frac{1}{s}$	0
	$\frac{\beta}{s^2 + \beta^2}$	0
	$\frac{1}{s+a}$	-a
$e^{-(a+j\beta)t} + e^{-(c+jd)t}$ $a > c > 0$	$\frac{1}{s+a+j\beta} + \frac{1}{s+c+jd}$	-c
		$-\infty$
e^{t^2}		No abscissa

value of s there corresponds a definite value, or set of values of Z , then Z is called a "function" of the complex variable s ; i.e.,

$$Z = G(s) \quad (25)$$

Then if Z is a function of s , separating Z into real and imaginary parts gives

$$Z = u + jv = G(\sigma + j\omega) \quad ,$$

and each of the real functions u and v are determined by the pair of real variables σ and ω ; i.e.,

$$u = u(\sigma, \omega) \quad \text{and} \quad v = v(\sigma, \omega) \quad .$$

Consequently,

$$Z = G(s) = u(\sigma, \omega) + jv(\sigma, \omega) \quad . \quad (26)$$

To plot the function G we also need a plane which we will call the G -plane, as shown in Figure 7.

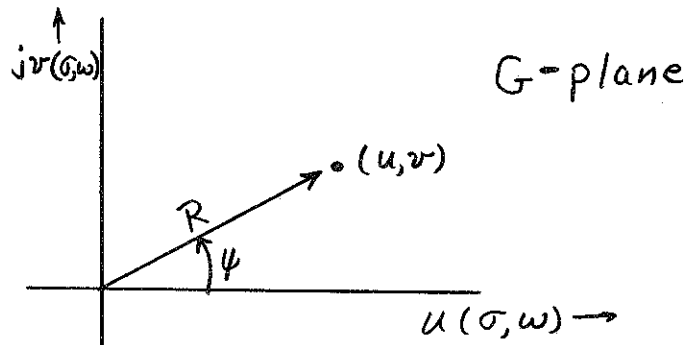


Fig. 7

In vector notation,

$$G(s) = R e^{j\psi} \quad ; \quad \text{where} \quad R = \sqrt{u^2 + v^2} \quad , \quad \psi = \tan^{-1} \frac{v}{u}$$

Analytic Functions of a Complex Variable (Differentiation)

A function $G(s)$ is said to be "analytic" in a region R of the complex s -plane if $G(s)$ has a unique derivative at each point of R .

To require that a function $Z = G(s)$ have a unique derivative at a point s is equivalent to requiring that the limit

$$\frac{dz}{ds} = \lim_{\Delta s \rightarrow 0} \left[\frac{G(s+\Delta s) - G(s)}{\Delta s} \right] = \lim_{\Delta s \rightarrow 0} \frac{\Delta Z}{\Delta s} \quad (27)$$

exist uniquely as $\Delta s \rightarrow 0$ from any direction in the complex plane. That is, the value approached by the ratio $\frac{\Delta Z}{\Delta s}$ must exist for any direction of approach and must not depend upon the direction.

As an example of a simple function which is not analytic anywhere, consider the relation

$$Z = \sigma - j\omega = \bar{s} \quad (28)$$

Here Z can be considered as a function of $s = \sigma + j\omega$, since if s is given, the real and imaginary parts of s are determined and hence Z is determined.

However, if we examine the ratio

$$\begin{aligned} \frac{\Delta Z}{\Delta s} &= \frac{G(s+\Delta s) - G(s)}{\Delta s} \\ &= \frac{\sigma - j\omega + \Delta\sigma - j\Delta\omega - \sigma + j\omega}{\Delta\sigma + j\Delta\omega} \\ &= \frac{\Delta\sigma - j\Delta\omega}{\Delta\sigma + j\Delta\omega} \end{aligned} \quad (29)$$

we see that if $\Delta s \rightarrow 0$ along a line parallel to the real (σ) axis, so that $\Delta\omega = 0$ throughout the limiting process, the limit of Equation (29) is $+1$,

whereas if $\Delta s \rightarrow 0$ along a line parallel to the imaginary ($j\omega$) axis, so that $\Delta \sigma = 0$ throughout the limiting process, the limit is -1. In general, if Δs approaches zero along a curve with slope $d\omega/d\sigma = m$ at the point considered in the complex plane, then we have

$$\lim_{\substack{\Delta s \rightarrow 0 \\ \Delta \sigma \rightarrow 0 \\ \Delta \omega \rightarrow 0}} \left[\frac{\Delta z}{\Delta s} \right] = \lim_{\substack{\Delta \sigma \rightarrow 0 \\ \Delta \omega \rightarrow 0}} \left[\frac{1 - j \frac{\Delta \omega}{\Delta \sigma}}{1 + j \frac{\Delta \omega}{\Delta \sigma}} \right] = \frac{1 - mj}{1 + mj} = e^{-2j\theta} \quad (30)$$

and hence, different values of the limit are approached for each value of m .

The whole purpose of this example has been to emphasize that not just any function can have a unique derivative at a point in the complex plane, and that for those functions which do, we reserve the name "analytic functions" of complex variables.

Suppose that we are given a function of a complex variable which is defined over some region. Obviously it is practically impossible to determine by trial if the function has a unique derivative at every point in the region. However, a simple set of conditions which are both necessary and sufficient for analyticity have been determined and are called the "Cauchy-Riemann conditions". These are derived in practically any book which discusses complex variables and are simply stated here as

$$\left. \begin{aligned} \frac{\partial u}{\partial \sigma} &= \frac{\partial v}{\partial \omega} \\ \frac{\partial u}{\partial \omega} &= -\frac{\partial v}{\partial \sigma} \end{aligned} \right\} \quad (31)$$

The fact that both the real and imaginary parts of a function of a complex variable satisfy the Cauchy-Riemann conditions is sufficient that the function be analytic.

PROPERTIES OF ANALYTIC FUNCTIONS OF COMPLEX VARIABLES

A. Theory of Conformal Mapping

For functions of real variables, usually denoted by $y = f(x)$, we can easily visualize the behavior of the function by plotting it in the xy -plane. However, in order to interpret the case of a function $Z = G(s)$ of a complex argument geometrically, we must use two planes, an s -plane and a G -plane as was previously described in Figure 7. This is because both $G(s)$ and s have two coordinates; σ and ω corresponding to the s -plane, and u and v corresponding to the G -plane.

Suppose that on the s -plane we choose several values of s and draw the curves shown in Figure 8 as AB and BC .

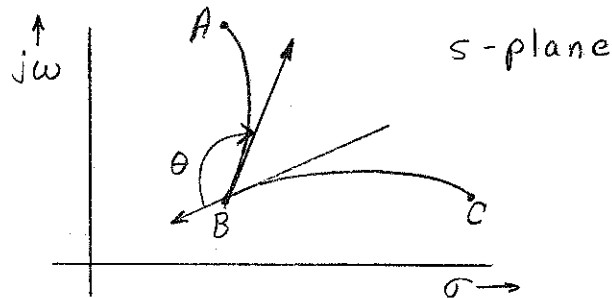


Fig. 8

Now draw tangents to the curves at the point B and direct them away from B in the manner shown. Measure the angle θ in a clockwise direction between the tangents. Suppose we have a relationship $Z = G(s)$ and wish to evaluate and plot this function for every value of s on the curves AB and BC of Figure 8. The result of plotting in the G -plane might look as shown in Figure 9. The shape and location of the curves in the G -plane would of course depend upon the function used, but the useful result is that the angle between the tangents (providing the directions are chosen the same as in Figure 8) is preserved if $Z = G(s)$ is an "analytic function". The

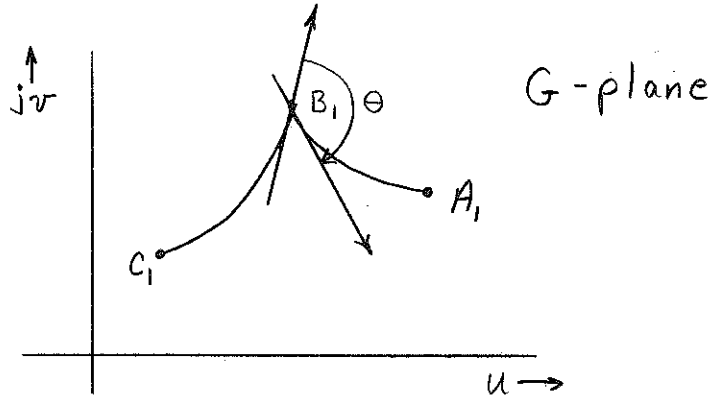


Fig. 9

importance of this result will become apparent when we later discuss the Nyquist theory.

B. Cauchy's Integral Theorem

Suppose we have a function $G(s)$ which is analytic for all values of s on the boundary of and within the arbitrary contour C shown in Figure 10. Then Cauchy's integral theorem states that integration counterclockwise around the contour C gives

$$\oint_C G(s) ds = 0 \quad (32)$$

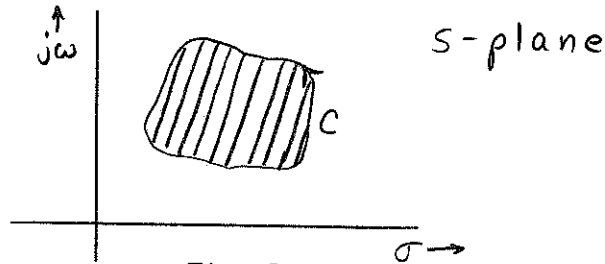


Fig. 10

This says that if the function $G(s)$ is defined (continuous) and differentiable (analytic) on and within the region C , the integral of the function between any two points of C is independent of the path of integration between them.

C. Cauchy's Integral Formula

Let's again consider that we have a $G(s)$ which is analytic for all values of s on the boundary of and within the contour C shown in Figure 11.

Then Cauchy's integral formula is

$$G(s_0) = \frac{1}{2\pi j} \oint_C \frac{G(s)}{s-s_0} ds \quad (33)$$

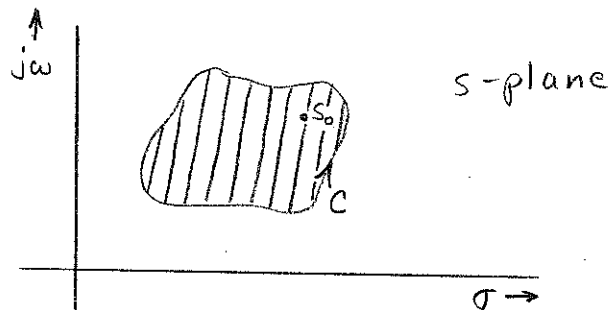


Fig. 11

This theorem states that if a function is known to be analytic, as prescribed above, on and within the contour C and if its values are known along the path C then the values of the function along the boundary completely determine the values of the function in the interior of C .

A very useful extension of Cauchy's integral formula is given by the following theorem:

The function $G(s_0)$ defined by

$$G(s_0) = \frac{1}{2\pi j} \oint_k \frac{G(s)}{s-s_0} ds,$$

where k is an arbitrary path along which $G(s)$ is defined and continuous, possesses derivatives in the region of definition of every order and these are given by the following formulas:

$$G'(s_0) = \frac{1}{2\pi j} \oint_k \frac{G(s)}{(s-s_0)^2} ds \quad ; \quad (34)$$

$$G''(s_0) = \frac{2!}{2\pi j} \oint_k \frac{G(s)}{(s-s_0)^3} ds \quad ; \quad (35)$$

and in general

$$G^{(n)}(s_0) = \frac{n!}{2\pi j} \oint_k \frac{G(s)}{(s-s_0)^{n+1}} ds \quad ; \quad (36)$$

for $n = 1, 2, 3, \dots$

D. The Principle of Analytic Continuation

Suppose we have two functions $G_1(s)$ and $G_2(s)$. Let $G_1(s)$ be defined and analytic in a region C_1 and also let $G_2(s)$ be defined and analytic in a region C_2 , as shown in Figure 12. Let the regions C_1 and C_2 overlap; i.e., there is a region g such that the values of s in g are common to both C_1 and C_2 . In the region g , let $G_1(s) = G_2(s)$. Under these conditions the functions $G_1(s)$ and $G_2(s)$ are then analytically continued from the one region to the other.

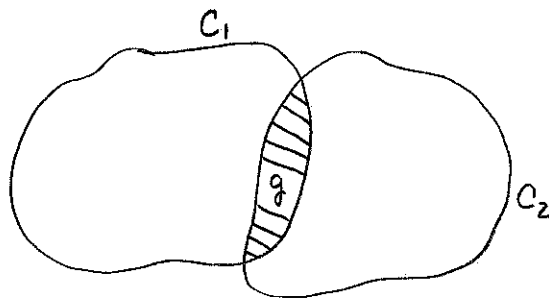


Fig. 12

We can say, therefore, that if the two regions C_1 and C_2 are in the position just described, and if an analytic function is defined in C_1 , then either there is no function at all or precisely one function which is analytic in C_2 and coincides with $G_1(s)$ in g . If such a function $G_2(s)$ exists, then the function $G_1(s)$ defined in C_1 is said to be "continuable" beyond C_1 into the region C_2 . The principle of analytic continuation will be used later to remove the restriction on the Laplace transform that the real part of s be greater than the abscissa of absolute convergence.

The above formulas and theorems from the theory of complex variables have been presented without proofs, however the proofs are given in most books which treat complex variables. As it will be seen later, these few statements about functions of a complex variable represent essentially all there is in the mathematical background of stationary linear systems.

Now that we are a little more familiar with complex variables, let's review some of the fundamental properties of Laplace transforms.

PROPERTIES OF LAPLACE TRANSFORMS

A. Poles and Zeros of Complex Functions

For our purpose, the functions of complex variables which are of interest are ratios of polynomials. Thus, we will have functions of the form

$$\begin{aligned} G(s) &= \frac{a_n s^n + a_{n-1} s^{n-1} + \dots + a_0}{b_m s^m + b_{m-1} s^{m-1} + \dots + b_0} \\ &= A \frac{(s-z_1)(s-z_2) \dots (s-z_n)}{(s-p_1)(s-p_2) \dots (s-p_n)} \quad , \quad (37) \end{aligned}$$

with

$$A = \frac{a_n}{b_m} \quad ; \quad n < m$$

In Equation (37) the Z_i 's (roots of the numerator) are zeros of $G(s)$ and the p_i 's (roots of the denominator) are poles of $G(s)$. By depicting all values obtained for the Z_i 's and p_i 's from the two polynomials on the s -plane, we have a graphical representation of the function $G(s)$ which is called a pole-zero plot of the function. A representative pole-zero plot is shown in Figure 13, where the number of zeros has been chosen to be four, and the number of poles to be five. In Figure 13, the poles are indicated by X's

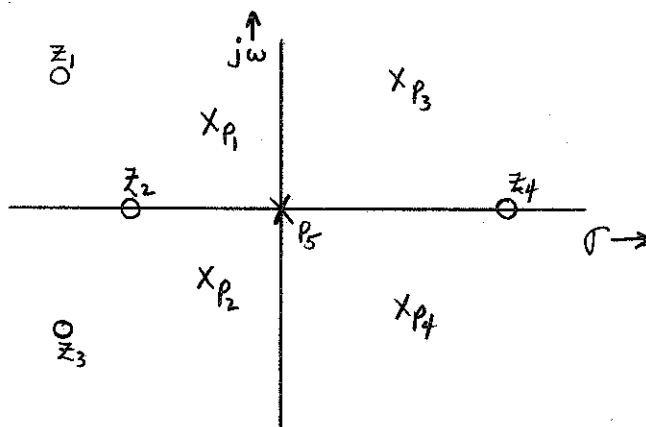


Fig. 13

and the zeros are indicated by circles, a notation which will be used consistently throughout this series of lectures. Notice that, apart from the constant A , all the information about a Laplace transform which is a ratio of two polynomials is given by the poles and zeros. In other words, only a few complex frequencies are required to describe the function.

Assume that we are given an $f(t)$ and from this have determined $F(s)$, the Laplace transform of $f(t)$. We would like to know if there is any relationship between this $F(s)$ and the Laplace transform of other functional relations related to $f(t)$, such as the derivative df/dt , etc. The answer is yes. There are some fundamental rules which determine the new Laplace transforms, and these are presented below for several fundamental operations.

1. Multiplication by e^{-at}

Here the problem is: given $f(t)$ and $F(s)$, find the Laplace transform of $e^{-at}f(t)$. Applying the basic formula,

$$\begin{aligned} \text{Laplace transform of } e^{-at}f(t) &\equiv \mathcal{L}\{e^{-at}f(t)\} = \int_0^{\infty} e^{-at}f(t)e^{-st}dt \\ &= \int_0^{\infty} f(t)e^{-(s+a)t}dt \\ &= F(s+a) \end{aligned} \quad (38)$$

Thus, the Laplace transform of $e^{-at}f(t)$ is determined by replacing s in $F(s)$ by $s + a$.

Example

$$\mathcal{L}\{\cos\beta t\} = \frac{s}{s^2 + \beta^2} \quad ;$$

$$\mathcal{L}\{e^{-at}\cos\beta t\} = \frac{s+a}{(s+a)^2 + \beta^2} \quad \cdot$$

2. Multiplication by t

Given $f(t)$ and $F(s)$, find the Laplace transform of $t f(t)$. By definition

$$F(s) = \int_0^{\infty} f(t)e^{-st}dt \quad \cdot$$

Then

$$-\frac{dF(s)}{ds} = \int_0^{\infty} t f(t)e^{-st}dt \quad \cdot \quad (39)$$

Thus, the Laplace transform of $t f(t)$ is determined by taking the negative of the derivative with respect to s of the Laplace transform of $f(t)$. In general,

$$\int_0^{\infty} t^n f(t) e^{-st} dt = (-1)^n \frac{d^n F(s)}{ds^n} \quad (40)$$

Example

For $f(t) = \sin \beta t$; $F(s) = \frac{\beta}{s^2 + \beta^2}$

Then

$$\begin{aligned} \mathcal{L}\{t \sin \beta t\} &= - \frac{dF(s)}{ds} \\ &= \frac{2\beta s}{(s^2 + \beta^2)^2} \end{aligned}$$

3. Some General Remarks

There is a unique correspondence between the location of the poles of a function of a complex variable and the behavior of the time function for which the function of a complex variable is the Laplace transform. This correspondence is described graphically in Table II. As an example, consider the time function to be $e^{-at} \sin \beta t$. Table II shows the graph of this function in the t -plane as a damped oscillation as a function of time. In the s -plane this function is represented by two poles in the left-half-plane; i.e., to the left of the $j\omega$ -axis; its Laplace transform is $\frac{\beta}{(s+a)^2 + \beta^2}$, and the abscissa of absolute convergence is $-a$. As another example, consider the time function $t \sin \beta t$. Table II shows the graph of this function to be divergent as a function of time. The Laplace transform of this function is $\frac{2\beta s}{(s^2 + \beta^2)^2}$ with an abscissa of absolute convergence of zero. The

TABLE II

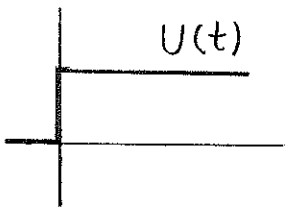
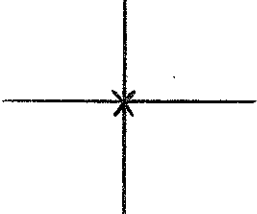
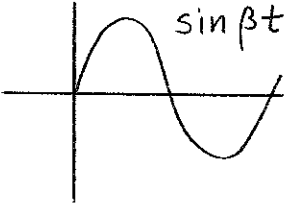
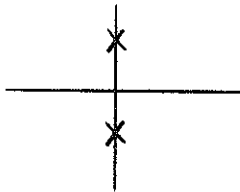
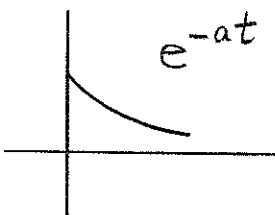
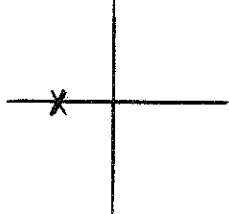
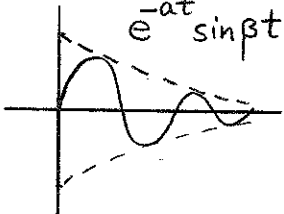
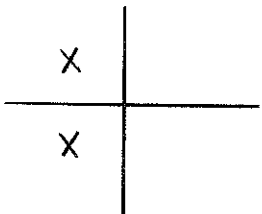
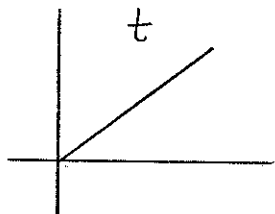
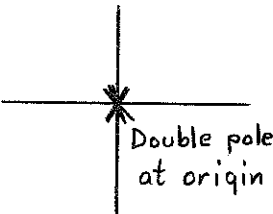
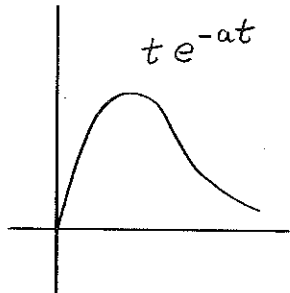
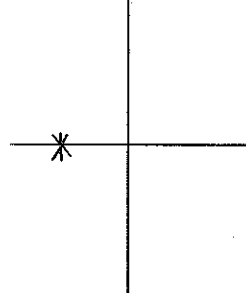
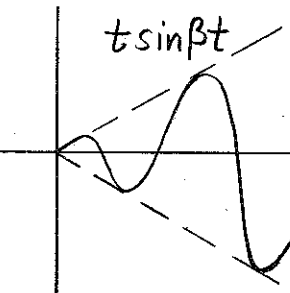
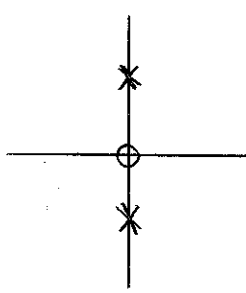
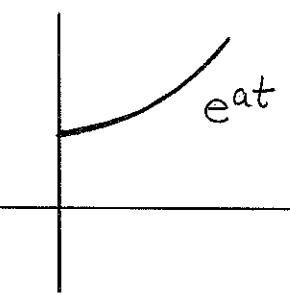
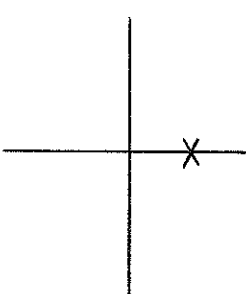
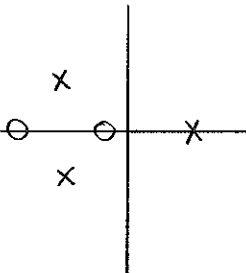
t-plane	F(s)	s-plane	abscissa of absolute convergence
	$\frac{1}{s}$		0
	$\frac{\beta}{s^2 + \beta^2}$		0
	$\frac{1}{s+a}$		$-a$
	$\frac{\beta}{(s+a)^2 + \beta^2}$		$-a$
	$\frac{1}{s^2}$	 <p>Double pole at origin</p>	0

TABLE II
(continued)

t-plane	F(s)	s-plane	abscissa of absolute convergence
	$\frac{1}{(s+a)^2}$		$-a$
	$\frac{2\beta s}{(s^2 + \beta^2)^2}$		0
	$\frac{1}{s-a}$		a
$e^{-at} \sin \beta t + e^{at}$			a

function $t \sin \beta t$ is represented in the s-plane by a zero at the origin and two double poles on the $j\omega$ -axis.

By considering the s-plane representation of several functions of time, some general rules can be determined for predicting the behavior of these functions in the time domain. These general rules apply to increasing, decreasing, and constant functions of time and are listed below.

<u>Increasing functions of time</u>	<u>Constant functions</u>	<u>Decreasing functions of time</u>
Poles in the right-half-plane of any order. Multiple poles on the j -axis.	Single order poles on the j -axis.	Poles in the left-half-plane of any order.

Now we will consider some more properties of Laplace transforms.

4. Differentiation in the Time Domain

The problem is, if we are given a $f(t)$ and the corresponding $F(s)$, find the Laplace transform of $\frac{df(t)}{dt}$. Applying the general formula

$$\mathcal{L}\left\{\frac{df(t)}{dt}\right\} = \int_0^{\infty} \frac{df(t)}{dt} e^{-st} dt \quad ,$$

by partial integration

$$\mathcal{L}\left\{\frac{df(t)}{dt}\right\} = f(t) e^{-st} \Big|_0^{\infty} + s \int_0^{\infty} f(t) e^{-st} dt \quad . \quad (41)$$

But the integral on the right hand side is just s times the Laplace transform of $f(t)$. Therefore

where $f(0)$ is the function $f(t)$ evaluated at time $t = 0$.

In arriving at Equation (41) we have assumed no mathematical restrictions on the function $f(t)$. However, in order to evaluate $e^{-st}f(t)\Big|_0^\infty$, $f(t)$ must be a "function of exponential order", as previously described. From Equation (41) we see that differentiation in the time domain corresponds to multiplying by s in the frequency domain.

By repeated application of this procedure, in general

$$\begin{aligned} \mathcal{L}\left\{\frac{d^n f(t)}{dt^n}\right\} &= \int_0^\infty \frac{d^n f(t)}{dt^n} e^{-st} dt \\ &= s^n F(s) - s^{n-1} f(0) - s^{n-2} \left. \frac{df(t)}{dt} \right|_{t=0} - \\ &\quad \dots - \left. \frac{d^{n-1} f(t)}{dt^{n-1}} \right|_{t=0} \end{aligned} \quad (42)$$

5. Integration in the Time Domain

Again, if we are given a $f(t)$ and $F(s)$, is there a relationship between the Laplace transform $F(s)$ and the Laplace transform of $\int f(t)dt$?
From

$$F(s) = \int_0^\infty f(t) e^{-st} dt,$$

with partial integration

$$\begin{aligned} F(s) &= \int_0^\infty f(t) e^{-st} dt = e^{-st} \int_0^t f(t) dt + s \int_0^\infty \left[\int_0^t f(t) dt \right] e^{-st} dt \\ &= - \int_0^\infty f(t) dt + s \int_0^\infty \left[\int_0^t f(t) dt \right] e^{-st} dt, \end{aligned}$$

where it has been assumed that $e^{-s\infty}f(\infty) \rightarrow 0$, i.e., $\int f(t)dt$ is of exponential order. Then $\int_0^{\infty} f(t)dt = f^-(0)$ (the initial value of the integral at time $t = 0$), and

$$\begin{aligned} \int_0^{\infty} \left[\int_0^t f(t)dt \right] e^{-st} dt &= \frac{F(s)}{s} + \frac{f^-(0)}{s} \\ &= \frac{F(s)}{s} \end{aligned} \quad (43)$$

since $\int_0^0 f(t)dt = 0$. Thus, integration in the time domain corresponds to division by s in the frequency domain. It will be noticed that the initial values of the function $f(t)$ have been included in this formulation.

Examples

Let $f(t) = \cos \beta t = \frac{1}{\beta} \frac{d}{dt} \sin \beta t$. Then

$$F(s) = \int_0^{\infty} \cos \beta t e^{-st} dt = \frac{1}{\beta} s \frac{\beta}{s^2 + \beta^2} = \frac{s}{s^2 + \beta^2} .$$

Let $f(t) = \int_0^t U(t)dt$. Then

$$F(s) = \int_0^{\infty} t e^{-st} dt = \frac{1}{s} \cdot \frac{1}{s} = \frac{1}{s^2} .$$

Another question is whether there is a correspondence between the magnitude of the function in the time domain and the magnitude of the function in the frequency domain. With two exceptions, the answer is that there is no correspondence. These two exceptions are known as the initial and final value theorems and correspond to the conditions $t = 0$; $s \rightarrow \infty$ and $t = \infty$; $s \rightarrow 0$, respectively.

6. Initial Value Theorem

The theorem states that

$$\lim_{s \rightarrow \infty} [s F(s)] = f(0) \quad (44)$$

The proof is as follows:

We know that

$$\int_0^{\infty} \frac{df(t)}{dt} e^{-st} dt = s F(s) - f(0)$$

By letting $s \rightarrow \infty$, the integral approaches zero and in the limit

$$\lim_{s \rightarrow \infty} [s F(s)] = f(0)$$

Example

$$\text{If } F(s) = \frac{a_1 s + a_0}{s^2 + b_1 s + b_0} \quad ; \quad \text{then } f(\infty) = \lim_{s \rightarrow \infty} s F(s) = a_1$$

7. Final Value Theorem

The theorem states that

$$\lim_{s \rightarrow 0} [s F(s)] = f(\infty) \quad , \quad (45)$$

provided $f(\infty)$ is finite; i.e., $F(s)$ has no poles in the right half plane.

The proof is as follows:

$$\int_0^{\infty} \frac{df(t)}{dt} e^{-st} dt = s F(s) - f(0)$$

If, before we integrate we let $s = 0$, then

$$\lim_{s \rightarrow 0} \int_0^{\infty} \frac{df(t)}{dt} e^{-st} dt = f(\infty) - f(0) = \lim_{s \rightarrow 0} [s F(s) - f(0)]$$

Thus

$$\lim_{s \rightarrow 0} [s F(s)] = f(\infty)$$

Examples

Suppose $f(t)$ is a unit step function. Then $F(s) = 1/s$ and $f(\infty) =$
 $sF(s) = s(1/s) = 1$.
 $s \rightarrow 0$

However, suppose $f(t) = e^{at}$. Then $F(s) = \frac{1}{s-a}$ and $\lim_{s \rightarrow 0} [sF(s)] = \lim_{s \rightarrow 0} \left[\frac{s}{s-a} \right] = 0$.
This is inconsistent with $f(\infty) = e^{a\infty} \rightarrow \infty$, and the final value of the func-
tion e^{at} cannot be determined by knowing only $F(s)$. This is in agreement with
the restriction that $f(\infty)$ be finite or $F(s)$ have no poles in the right half
plane.

8. Translation in Time

$$\text{Suppose } f(t) = \begin{cases} 0 & ; t \leq 0 \\ f(t) & ; t \geq 0 \end{cases} .$$

Then

$$\int_{\tau}^{\infty} f(t-\tau) e^{-st} dt = e^{-s\tau} F(s) \quad \bullet \quad (46)$$

Proof:

If $f(t)$ is translated through τ units of time in the positive direction,
 $f(t)$ becomes $f(t - \tau)$ when $t \geq \tau$ and zero otherwise. Then since the trans-
lated function vanishes when $0 \leq t < \tau$, its transform is defined by

$$\int_{\tau}^{\infty} f(t-\tau) e^{-st} dt \quad \bullet$$

If t is replaced by $t' + \tau$, the lower limit becomes zero and

$$\begin{aligned} \int_0^{\infty} f(t') e^{-s(t'+\tau)} dt' &= e^{-s\tau} \int_0^{\infty} f(t') e^{-st'} dt' \\ &= e^{-s\tau} F(s) \quad \bullet \end{aligned}$$

9. Linearity

The following relationships are satisfied by Laplace transforms

$\mathcal{L}f$

$$\int_0^{\infty} f(t) e^{-st} dt = F(s)$$

then

$$a. \int_0^{\infty} a f(t) e^{-st} dt = a F(s) \quad ; \quad a = \text{constant} \quad (47)$$

$$b. \int_0^{\infty} [a_1 f_1(t) + a_2 f_2(t)] e^{-st} dt = a_1 F_1(s) + a_2 F_2(s) \quad . \quad (48)$$

We will now proceed to the discussion of the analysis of linear systems.

Suggested References

1. M. F. Gardner and J. L. Barnes, Transients in Linear Systems. New York: John Wiley & Sons, 1956.
2. K. Knopp, Theory of Functions, Part One. New York: Dover Publications.
3. R. V. Churchill, Functions of Complex Variables.
4. H. Chestnut and R. W. Mayer, Servomechanisms and Regulating System Design, Volume 1. New York: John Wiley & Sons, 1953.

Lecture No. II

ANALYSIS OF LINEAR SYSTEMS

From our experiences with different physical processes we have come to classify these systems under different names; such as thermal, electrical, mechanical, hydraulic, etc. However, in attempting to describe the behavior of any of these systems mathematically as a function of time, the same principles are involved for one as for any other; i.e., we ultimately are faced with having to solve differential equations.

Let's examine what these system equations really say. We see that we can visualize physical quantities by certain laws. We have imposed upon nature some fundamental quantities which we give names, such as "energy". We also know that we can postulate that certain of these quantities are conserved; i.e., in the case of energy there is a balance between the total energy of a system and the energies associated with different aspects of the system. Just what energy is referred to here depends on what system one is talking about. If we propose that the system is electrical, we are talking about electrical energies; if it is a mechanical system, there are mechanical (i.e., kinetic, potential, etc.) energies involved.

We also define another quantity which is conserved and call it "momentum". Some systems are described in terms of "particles", and, as we well know, there is conservation of particles. Thus, when we attempt to describe a physical system we are really thinking in terms of the laws of conservation of the physical quantities involved. Therefore, the differential equations which one writes are really means of implementing the conservation laws for

the physical quantities involved in a given system. We do this by writing the equations in terms of the characteristic parameters of the system.

In order to define the characteristic parameters, let's consider the quantity "energy", and ask ourselves, "what can we do to energy"? One thing we can do is to store energy. In an electrical system, for instance, where current may be passed through an inductor there is a magnetic field associated with the inductor and we visualize this by saying there is energy stored in the inductor. We also know that we can dissipate energy. This is not to say that the energy is lost but, whereas in the case of the inductor we can get the energy back, there are many instances where energy is not recoverable. As an example of energy being dissipated, if current is passed through a resistor energy is used to heat the resistor but we cannot recover this energy.

Thus, from a knowledge of the characteristic properties of the system we can define the characteristic parameters of energy storage and dissipation. Before discussing these parameters of storage and dissipation further, let's first restrict ourselves to a particular system.

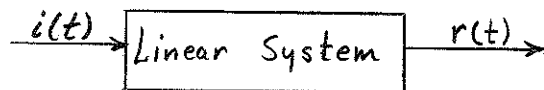
We want to consider for the present, only linear systems. By this we mean that if a system is characterized by two conjugate quantities, say current I and voltage V , then we can describe energy in terms of these quantities, and if we put a value of I into the system and measure a value of V , then if the system is linear, when we put in $2 I$ we will measure $2 V$.

If the system we wish to describe is an electrical system then we can use the quantities called resistance, capacitance and inductance. If the system is a thermal system we speak of thermal resistance, thermal capacitance, etc. For nuclear reactors the delayed neutron precursor concentration corresponds to storage and the decay constant λ corresponds to a loss.

In a broad sense, energy and particle processes can be visualized as equivalent.

Thus, we see that for different systems we are compelled to characterize them by similar parameters associated with potential energy, kinetic energy and dissipation of energy.

Having established these parameters on the basis of energy, we could look at the problem in another way. Let's forget about the conservation laws and invoke an observation of nature which we will call "laziness". Now we know that no physical system is capable of changing its status instantaneously. This is easily seen since the rate of change of a physical quantity cannot be infinite. Let's characterize a linear system by a black box where we have an input $i(t)$ and an output $r(t)$. For the present we are talking about linear systems in general, not just an electrical system.



If we talk about storage only, the response of the system must be given by one of two relationships;

$$r(t) = a \frac{di(t)}{dt} \quad , \quad (49)$$

or

$$r(t) = b \int i(t) dt \quad \cdot \quad (50)$$

In these equations a and b are constants and represent inductive or capacitive effects. In terms of dissipation or loss of energy, the relationship would be

$$r(t) = c i(t) \quad (51)$$

where the response is directly proportional to the input.

Now, looking at the problem from the viewpoint of storage and dissipation of energy, we are in a position to describe the system in terms of equations. We must do this systematically by taking the various parts of the system and describing them individually. The equations which we will arrive at for linear systems will be integrodifferential equations. In addition to requiring that the system is linear, we also assume that interactions within the system propagate infinitely fast. By this we mean that there is no delay time involved between a measurement in one part of the system and another measurement somewhere else in the system. This assumption therefore implies spatial independence and we call the system a "lumped-parameter" system. If spatial dependence were to be considered, the equations describing the system would be partial differential equations rather than ordinary differential equations.

Now let's take a specific system, formulate the equations describing the system, and by some method, solve the equations. Let's consider the electrical network shown in Figure 14. Although the diagram has been referred to as an electrical network, we could conceivably represent any physical system in the same manner since any system can be described in terms of the characteristic coefficients of inductance, capacitance and resistance which effectively represent the possibilities of storage and dissipation of energy.

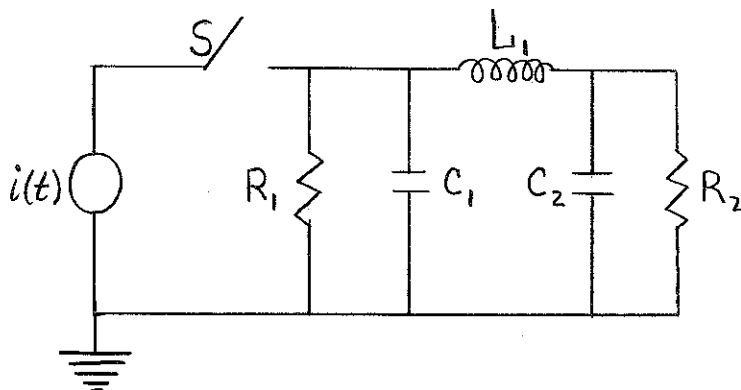


Fig. 14

We interpret the diagram as saying that when the switch S is closed, the excitation $i(t)$ is applied to the system. In this case we will call $i(t)$ current and when current is introduced to the different elements voltages are formed throughout the system. In this particular case, the construction of the diagram implies that in the first part of the system there is a dissipation of energy by passing current through the resistor R_1 . This energy is not recoverable. In the second part, we can think of the capacitor C_1 as a small tank in which we can store energy for later use. When current is passed through the inductance L_1 we are storing kinetic energy because of the magnetic field created. The same reasoning applies to the other portions of the network. We see then that we can interpret the physical system as a model and for our purposes we can think of it only as a mathematical model.

One method of attacking the problem of writing the equations for the model is to consider that energy is conserved and write the balance equations for each portion of the diagram. However, an equivalent statement of the conservation of energy is obtained from the laws relating the voltages and currents of the system. These are known as Kirchhoff's laws. There are two ways to express Kirchhoff's laws; first, by considering any closed loop, the algebraic sum of the voltages around the loop is zero; i.e., for a loop

$$\sum V_i = 0 \quad , \quad (52)$$

and second, by considering a node, where currents enter and leave, the algebraic sum of the currents into and out of the node is equal to zero; i.e., for a node

$$\sum i(t) = 0 \quad . \quad (53)$$

These two statements are equivalent to the proposition of conservation of energy.

In the case of Figure 14 it is convenient to use the second expression of Kirchhoff's laws to derive the equations of the system. Thus we will use the nodal approach and have that for each node, the algebraic sum of the currents into and out of that node is zero. The diagram of Figure 14 is redrawn in Figure 15 to show the two nodes to be considered, which are denoted by the voltages V_1 and V_2 , and to show the direction of current through each element. The voltages V_1 and V_2 are determined with respect to ground. The switch is now closed.

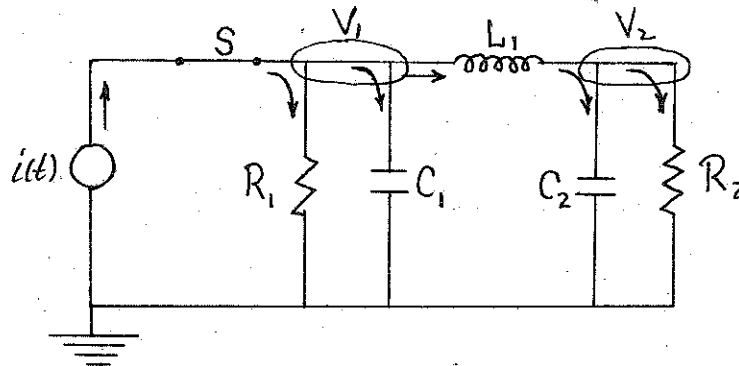


Fig. 15

We can write the equations for each node. For the first node:

$$i(t) = \frac{V_1}{R_1} + C_1 \frac{dV_1}{dt} + \frac{1}{L_1} \int (V_1 - V_2) dt \quad (54)$$

For the second node:

$$\frac{1}{L_1} \int (V_1 - V_2) dt = C_2 \frac{dV_2}{dt} + \frac{V_2}{R_2} \quad (55)$$

Thus we have a set of ordinary integrodifferential equations.

Now that we have the equations for the system we need to solve them to determine the voltages V_1 and V_2 since we can assume that we know the excitation signal $i(t)$ and all of the coefficients R , L , and C . For our purposes, we will assume that the coefficients are constants; i.e., the values of R , L , and C do not vary with time. We will also need to specify the initial values of the voltages at the time the switch is closed. It is possible that the capacitors have charge stored in them prior to closing the switch, thus giving rise to an initial value of the voltage across the capacitors different from zero. However, we will assume that there is no initial charge on the capacitors and that all voltages are zero prior to throwing the switch. Similarly we assume that the initial current through the inductance is zero.

Even though for the simple example we have chosen here there are other methods of solving the equations which are just as easy, for the reasons explained in the introduction we choose to solve them by using Laplace transforms. The entire first lecture was concerned with describing the Laplace transform as a tool, and now we will learn how to use this tool.

We start by taking the Laplace transform of all the terms in each equation. As a matter of notation the Laplace transform of a quantity will be denoted by a bar over the symbol, such as, the Laplace transform of the current i will be denoted by \bar{i} . Then, in transform language Equation (54) becomes

$$\bar{I} = \frac{\bar{V}_1}{R_1} + C_1 s \bar{V}_1 + \frac{1}{L_1} \left(\frac{\bar{V}_1}{s} - \frac{\bar{V}_2}{s} \right), \quad (56)$$

and for Equation (55)

$$\frac{1}{L_1} \left(\frac{\bar{V}_1}{s} - \frac{\bar{V}_2}{s} \right) = C_2 s \bar{V}_2 + \frac{\bar{V}_2}{R_2}. \quad (57)$$

By using Laplace transforms we see that the integrodifferential equations we started with have been transformed into algebraic equations. We have already decided that the initial voltages are zero, and by rewriting the equations we get

$$\bar{I} = \left[\frac{1}{R_1} + C_1 s + \frac{1}{L_1 s} \right] \bar{V}_1 - \frac{1}{L_1 s} \bar{V}_2 \quad , \quad (58)$$

and

$$0 = \left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right] \bar{V}_2 - \frac{1}{L_1 s} \bar{V}_1 \quad . \quad (59)$$

At this point let's examine the obvious properties of these equations.

1. They are algebraic.
2. The variables \bar{V}_1 and \bar{V}_2 are related in terms of coefficients involving the characteristic coefficients of the system.
3. The coefficients of the variables are functions of the complex variable s .
4. If $i(t)$ is the only input to the system, none of the coefficients of the variables are dependent on $i(t)$.

Solving Equation (59) for \bar{V}_1 , in terms of \bar{V}_2 we get

$$\bar{V}_1 = L_1 s \left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right] \bar{V}_2 \quad . \quad (60)$$

Substituting Equation (60) into Equation (58) gives

$$\bar{I} = \left[\frac{1}{R_1} + C_1 s + \frac{1}{L_1 s} \right] \left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right] L_1 s \bar{V}_2 - \frac{1}{L_1 s} \bar{V}_2 \quad ,$$

and solving for \bar{V}_2

$$\bar{V}_2 = \frac{\frac{1}{L_1 s}}{\left[\frac{1}{R_1} + C_1 s + \frac{1}{L_1 s} \right] \left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right] - \frac{1}{L_1^2 s^2}} \cdot \bar{I} \quad .$$

Then

$$\bar{V}_2 = Z(s) \cdot \bar{I} \quad , \quad (61)$$

where

$$Z(s) = \frac{\frac{1}{L_1 s}}{\left[\frac{1}{R_1} + C_1 s + \frac{1}{L_1 s} \right] \left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right] - \frac{1}{L_1^2 s^2}}$$

Solving Equation (60) for \bar{V}_1 we get

$$\begin{aligned} \bar{V}_1 &= \frac{\left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right]}{\left[\frac{1}{R_1} + C_1 s + \frac{1}{L_1 s} \right] \left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right] - \frac{1}{L_1^2 s^2}} \cdot \bar{I} \\ &= Y(s) \cdot \bar{I} \quad , \quad (62) \end{aligned}$$

where

$$Y(s) = \frac{\left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right]}{\left[\frac{1}{R_1} + C_1 s + \frac{1}{L_1 s} \right] \left[\frac{1}{R_2} + C_2 s + \frac{1}{L_1 s} \right] - \frac{1}{L_1^2 s^2}}$$

We see then that it is possible to describe the system in terms of quantities $Z(s)$ and $Y(s)$, which depend only on the characteristic coefficients (R , L , and C) of the system, which are known, and a term \bar{I} which depends only on the input of the system, which is also known. The quantities $Z(s)$ and $Y(s)$ are rightfully called the "system functions" and the term \bar{I} is called the "excitation function". \bar{V}_1 and \bar{V}_2 are called the Laplace transforms of the responses of the system.

Notice that regardless of which variable \bar{V}_1 or \bar{V}_2 we solve for, the denominators (i.e., the poles) of the terms $Z(s)$ and $Y(s)$ are identical.

This will always be the case for linear systems and we call these the eigenvalues of the system. We may now write

$$\left. \begin{aligned} V_1 &= \mathcal{L}^{-1} [Y(s) \cdot \bar{I}] \\ V_2 &= \mathcal{L}^{-1} [Z(s) \cdot \bar{I}] \end{aligned} \right\} ,$$

which we will leave in this form for the present time. \mathcal{L}^{-1} denotes the inverse Laplace transform; i.e., the operation required to transform back to the time domain.

To illustrate the applicability of this method of analysis to another physical system, let's consider an example of a mechanical system. A diagram of the physical system is shown in Figure 16.

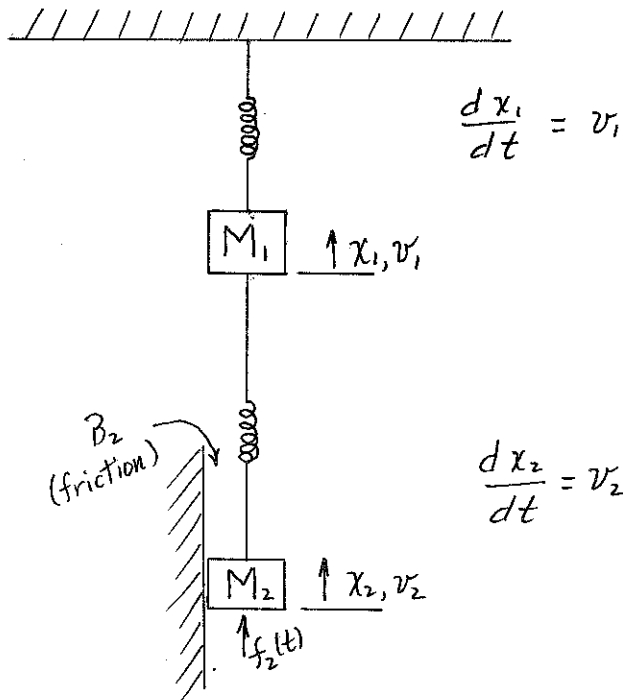


Fig. 16

For a mechanical system, Newton's second law of motion, or a slightly different expression of it known as D'Alembert's principle, corresponds to Kirchhoff's electric-network current law which was used in the first example. Assuming the motion of a body to be in the X-direction, this can be expressed by

$$\sum f(t) \text{ acting in the } x\text{-direction} = M \frac{d^2x}{dt^2} \quad (63)$$

Writing Equation (63) in the form

$$\sum f(t) \text{ acting in the } x\text{-direction} - M \frac{d^2x}{dt^2} = 0 \quad (64)$$

expresses D'Alembert's principle, namely:

The sum of the instantaneous external forces acting on a body in a given direction and the body's reaction force in that direction due to inertia is zero.

The equations for the system of Figure 16 are as follows: For M_2 ,

$$f_2(t) = M_2 \frac{dv_2(t)}{dt} + k_2 \int_0^t (v_2(t) - v_1(t)) dt + B_2 v_2(t), \quad (65)$$

and for M_1 ,

$$0 = M_1 \frac{dv_1(t)}{dt} + k_1 \int_0^t v_1(t) dt + k_2 \int_0^t (v_1(t) - v_2(t)) dt. \quad (66)$$

If we assume that all initial conditions are equal to zero, the Laplace transform of Equations (65) and (66) are:

$$\bar{F} = \left[M_2 s + B_2 + \frac{k_2}{s} \right] \bar{V}_2 - \frac{k_2}{s} \bar{V}_1, \quad (67)$$

and

$$0 = \left[M_1 s + \frac{k_1 + k_2}{s} \right] \bar{V}_1 - \frac{k_2}{s} \bar{V}_2 \quad (68)$$

Solving for \bar{V}_1

$$\bar{V}_1 = Y(s) \cdot \bar{F} \quad (69)$$

where

$$Y(s) = \frac{\frac{k_2}{s}}{\left[M_1 s + \frac{k_1 + k_2}{s} \right] \left[M_2 s + B_2 + \frac{k_2}{s} \right] - \frac{k_2^2}{s^2}}$$

Solving for \bar{V}_2

$$\bar{V}_2 = Z(s) \cdot \bar{F} \quad (70)$$

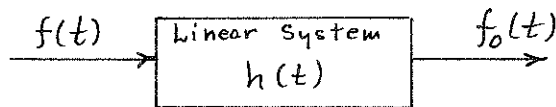
where

$$Z(s) = \frac{\left[M_1 s + \frac{k_1 + k_2}{s} \right]}{\left[M_1 s + \frac{k_1 + k_2}{s} \right] \left[M_2 s + B_2 + \frac{k_2}{s} \right] - \frac{k_2^2}{s^2}}$$

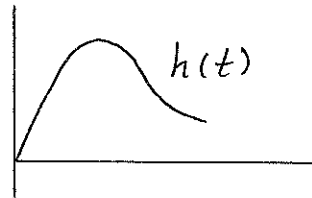
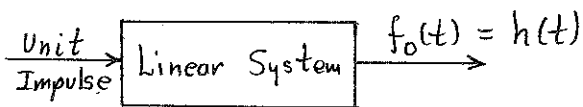
Thus we see that an identical procedure can be used to analyze the two linear systems, the electrical and the mechanical, and that for both cases we end up with the result that we can characterize any output by the product of two terms; one which is a function only of the system and another which depends only on the input signal to the system. The fact that we can do this is a direct property of the Laplace transform method of analysis, and to be even more general it will now be shown that for any linear system we will always have these two terms.

System Function - Transfer Function

Let's visualize a linear system by means of a diagram where there is an input $f(t)$ to the system, an output $f_0(t)$, and the system is characterized by the system function $h(t)$.

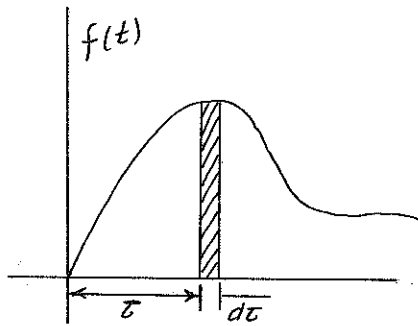


We can describe the function $h(t)$ by considering that the input to the system is an impulse. Mathematically an impulse is a function having zero width and infinite height, but the product of the width and the height is a finite number. Physically we can never achieve an impulse, but by making the width very close to zero we can approximate it. The response of the system to a unit impulse input will have some particular shape, and this output will necessarily be the system function itself.



Assuming then that we know the system function, that is, the response of the system to a unit impulse input, we would like to know what the response of the system will be for an arbitrary input. The arbitrary input may have any shape. Since the system is linear we know that we could calculate the output for any input by breaking the input up into several inputs, the sum of which gives the original input, measuring the output of each individual input and then adding them up. However, since presumably we already know the system response to an impulse input, we may mathematically break any input into a number of impulses.

Graphically then the input $f(t)$ can be characterized by many small impulses with an area given by

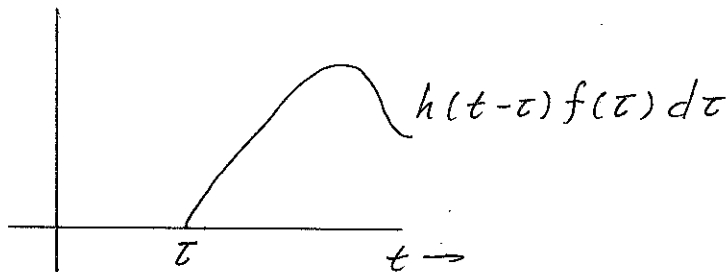


$$\text{Area of impulse} = f(\tau) d\tau$$

Then for any impulse $f(\tau)d\tau$, the output is

$$df_o(t) = h(t-\tau) f(\tau) d\tau \quad (71)$$

This is shown graphically as

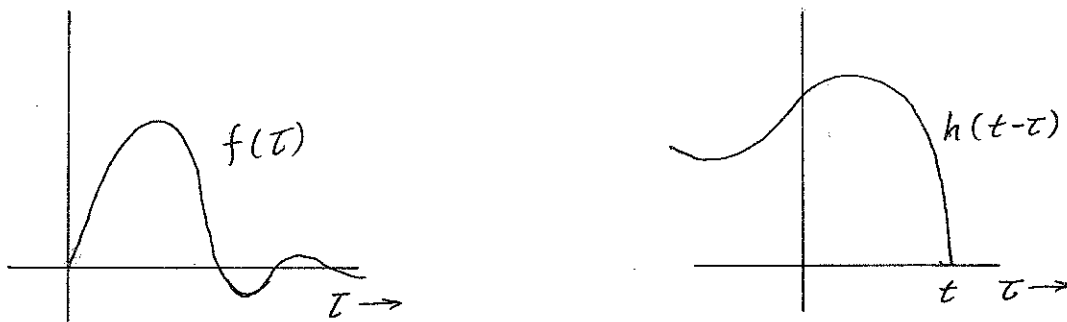


By considering all the possible impulses in the input, the output is given by

$$f_o(t) = \int_0^t h(t-\tau) f(\tau) d\tau \quad (72)$$

Equation (72) then describes the response of the system to an arbitrary driving force (input) in terms of the response of the system to a unit impulse, and the integral is called the real convolution integral,

describing convolution in the real domain. We can extend the limits of integration to $+\infty$ since the product of $h(t - \tau)$ and $f(\tau)$ is zero after $\tau = t$ anyway. This can be seen from the graphs of $h(t - \tau)$ and $f(\tau)$.



Then Equation (72) can be written as

$$f_0(t) = \int_0^{\infty} h(t-\tau) f(\tau) d\tau \quad (73)$$

Equation (73) can also be written in the form,

$$f_0(t) = \int_0^{\infty} h(\tau) f(t-\tau) d\tau \quad (74)$$

where the same reasoning applies.

These equations have all been written for the real time domain. In terms of Laplace transforms, from Equation (73)

$$\int_0^{\infty} f_0(t) e^{-st} dt = F_0(s) = \int_0^{\infty} e^{-st} dt \int_0^{\infty} h(t-\tau) f(\tau) d\tau \quad (75)$$

By changing the order of integration (assumed allowable for types of functions considered)

$$F_0(s) = \int_0^{\infty} f(\tau) d\tau \int_0^{\infty} h(t-\tau) e^{-st} dt \quad .$$

Multiplying and dividing by $e^{s\tau}$,

$$F_o(s) = \int_0^{\infty} e^{-s\tau} f(\tau) d\tau \int_0^{\infty} h(t-\tau) e^{-s(t-\tau)} dt.$$

Letting $t - \tau = t'$

$$F_o(s) = \int_0^{\infty} f(\tau) e^{-s\tau} d\tau \int_{-\tau}^{\infty} h(t') e^{-st'} dt'.$$

But since $h(t') = 0$ for $t' < 0$, the limits on the last integral can be from 0 to ∞ , and

$$F_o(s) = \int_0^{\infty} f(\tau) e^{-s\tau} d\tau \int_0^{\infty} h(t') e^{-st'} dt' \quad (76)$$

We can immediately recognize the first term on the right as the Laplace transform of the input function, and the second integral as the Laplace transform of the system function. Then

$$F_o(s) = H(s) \cdot F(s) \quad , \quad (77)$$

where $H(s)$ represents the system function and $F(s)$ corresponds to the input. We also call $H(s)$ the "transfer function" of the system.

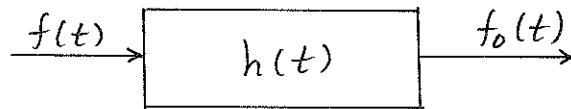
Equation (77) is of the same form as was derived for the electrical and mechanical systems used as examples. Since Equation (77) was obtained in general, the only restriction being that the system is linear, we see then that the response of any linear physical system can be described in terms of a factor involving only the system parameters, and a factor concerned only with the input. We can say then that for any linear system,

in the time domain the input and output are related by the convolution integral, and in the frequency domain they are related by the product of the transfer function of the system and the Laplace transform of the input.

A pictorial way of representing these results is through "block diagrams".

Block Diagrams

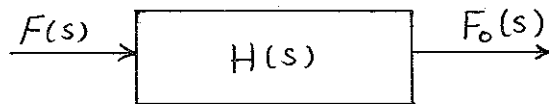
A block diagram is merely a shorthand notation of the relationship between the input and output of a linear system. As we have just shown, if we characterize our system by a black box with a system function $h(t)$, an input $f(t)$ and an output $f_0(t)$, then the diagram



corresponds to the equation

$$f_0(t) = \int_0^{\infty} h(\tau) f(t-\tau) d\tau = \int_0^{\infty} h(t-\tau) f(\tau) d\tau ,$$

whereas if our diagram is in terms of Laplace transforms;



the corresponding equation is

$$F_0(s) = H(s) \cdot F(s)$$

The reason we use block diagrams is because, for more involved systems the block diagram is a convenient means of visualizing what physical

processes we think are going on within the system. To illustrate this, let's consider the system shown in Figure 17.

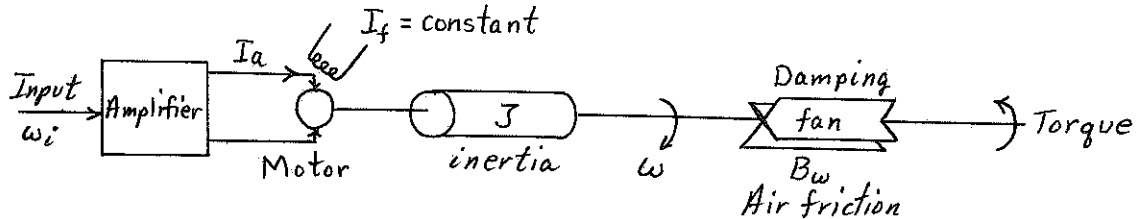


Fig. 17

From the diagram, a voltage ω_i is applied to the amplifier. The output of the amplifier is a current I_a which goes to the rotor of a motor. The shaft of the motor rotates with angular velocity ω and also has inertia. The shaft turns a fan which experiences a torque and involves friction, denoted by the friction coefficient for air, B_ω . The block diagram for this system is shown in Figure 18.

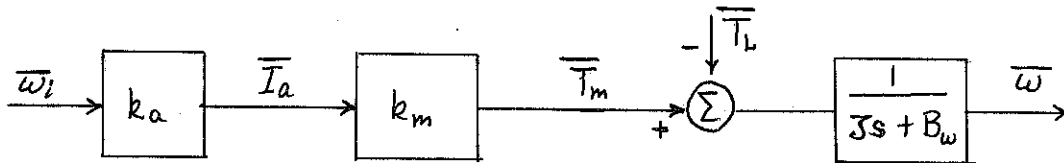


Fig. 18

We interpret Figure 18 by saying;

1. There is an input voltage ω_i to the amplifier.
2. The amplifier produces an output which is the current I_a .
3. The output from the amplifier goes to the motor which produces an output in the form of a torque T_m .

4. The torque T_m is what makes the system run because it overcomes the torque T_L at the other end, and at the same time overcomes the torques produced by inertia and friction.

5. The resultant torque then makes the shaft turn and give the speed ω . In terms of equations, we interpret Figure 18 as representing

$$\bar{I}_a = k_a \bar{\omega}_i \quad (\text{amplifier output}) , \quad (78)$$

$$\bar{T}_m = k_m \bar{I}_a \quad (\text{motor output, torque}) , \quad (79)$$

and

$$\bar{T}_m = k_m \bar{I}_a = J \frac{d\omega}{dt} + B_\omega \omega + \bar{T}_L \quad . \quad (80)$$

Block diagrams can become very involved. As an example, we will consider this same system when a feedback loop is added.

Feedback Systems

Basically a feedback system merely means that some output signal is fed back and compared with the input signal to determine if the system is operating as desired. In the case of Figure 17, say it is desired that for a certain input voltage the fan will rotate at a definite angular velocity. By measuring the output ω with a tachometer and comparing this measurement (now a voltage) with the input we have a feedback loop as shown in Figure 19.

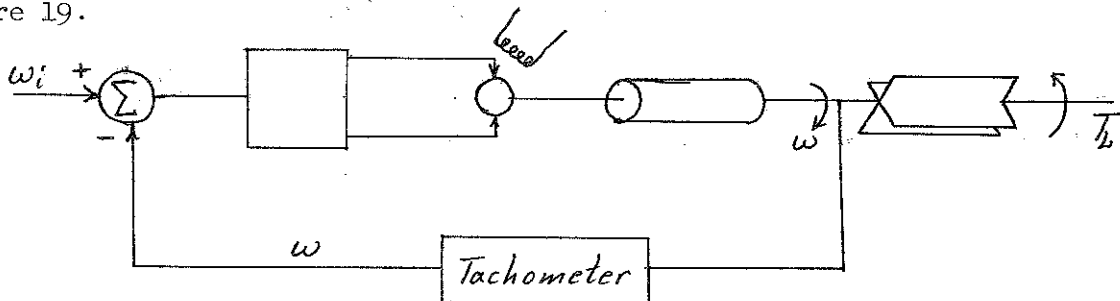


Fig. 19

Assume the system has been built such that if the voltage from the output ω is equal to the input voltage ω_i then the angular velocity is the desired value and nothing happens. However, if there is a difference in these two quantities, a signal is fed to the amplifier which essentially tells it whether to produce more current or to decrease the current output. In this manner then the feedback serves the purpose of controlling the response of the system. This is what might be called the "first kind" or "control feedback" because this was introduced purposely in the design of the system. However, there is another kind of feedback which is considered fictional because it is mentally created. That is to say, when we try to analyze a system we visualize that a particular feedback process is occurring as a result of our approach to the problem of analysis of the system.

A nuclear reactor is an example of a feedback system of this second kind. We have observed that in a nuclear reactor we have to deal with what we call power or neutron level, a quantity we call reactivity, which is associated with the position of the control rods, and the temperatures of different parts of the reactor. If we want to represent the behavior of the reactor by a block diagram we first have to decide which variable to consider as the input. If it is assumed that the reactivity is known then the block diagram might look as shown in Figure 20.

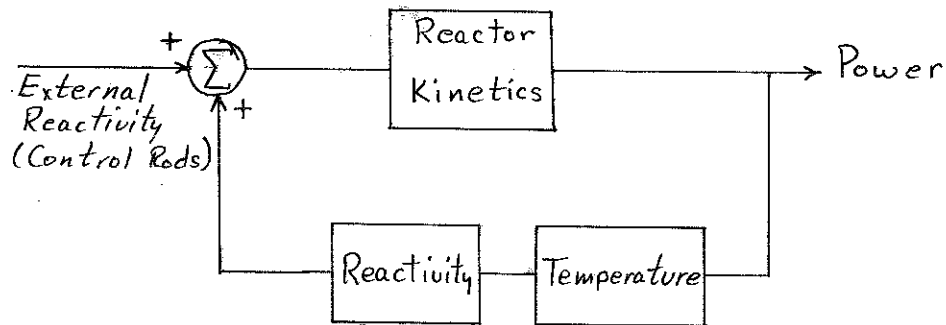


Fig. 20

We interpret Figure 20 as saying that the reactivity seen by the reactor is not really the same as what we put in because the power level affects the temperature, which in turn has an effect on the reactivity. This is a logical interpretation of the effects of the different variables on the behavior of the system, however, it is not the only interpretation.

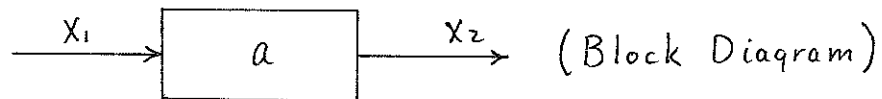
At this point Professor Gyftopoulos posed the problem that, since reactivity is not a directly measurable quantity because it is calculated from the measured period and the inhour formula (which in turn depends upon the control rod calibration, etc.) perhaps another variable, which is measurable, should be used as the input and the response of the system determined for this interpretation.

Now that we have established what block diagrams and feedback systems are, there is a very useful rule for handling feedback systems. This rule results from the method using flow graphs.

Flow Graphs

For other than the most simple systems, the block diagram for a system quickly becomes quite complicated. Since the desired results anyway are to determine the relationship between input and output, an easy method of getting this result would be most welcome. A method of doing just this will now be described briefly. The details of the method have been worked out some time ago, thus the method is by no means new.

Suppose we have a linear physical system described by the block diagram.



where X_1 is an input and X_2 the output. We have already said that a block diagram represents one or more equations; thus we could represent the block

diagram shown by the equation

$$X_2 = a X_1$$

where a is not necessarily a constant. Another way of representing this equation is by the flow graph



The implication of this graph is that X_1 is the cause and X_2 the effect. Several block diagrams, representative equations, and corresponding flow graphs are given in Table III.

TABLE III

<u>Block Diagram</u>	<u>Equation</u>	<u>Flow Graph</u>
	$X_2 = a X_1$	
	$X_2 = a X_1$ $X_3 = b X_2$	
	$X_3 = a X_1 + b X_2$	
	$X_3 = a X_1 + b X_2$ $X_4 = c X_3$	
	$X_2 = a X_1 + b X_3$ $X_3 = c X_2$	

Some rules and basic definitions to follow in using this method are:

1. Although there is no unique way of drawing a flow graph for a particular system, once you have chosen a direction for the arrows, use this convention throughout.

2. The value of an unknown is not changed by adding a branch out from the unknown. In other words, a variable can be a cause several times, but it can be an effect only once.

Definitions

Path: A path is a succession of branches from one node to another, all in the same direction and traced so that no node is encountered twice.

Loop: A continuous succession of branches traced in the same direction and forming a closed cycle, no node being encountered twice.

Let's assume a physical system for which the equations

$$\begin{aligned}x_2 &= f x_6 & x_6 &= e x_5 + l x_8 \\x_3 &= a x_1 + h x_7 & x_7 &= g x_3 + k x_8 \\x_4 &= b x_3 + d x_5 & x_8 &= m x_6 + i x_7 \\x_5 &= c x_4\end{aligned}$$

have been obtained. Assuming we know an input, say X_1 , we then have 7 equations and 7 unknowns. Solving this set of equations for all unknowns might prove to be very time consuming. However, we can construct the following flow graph shown in Figure 21.

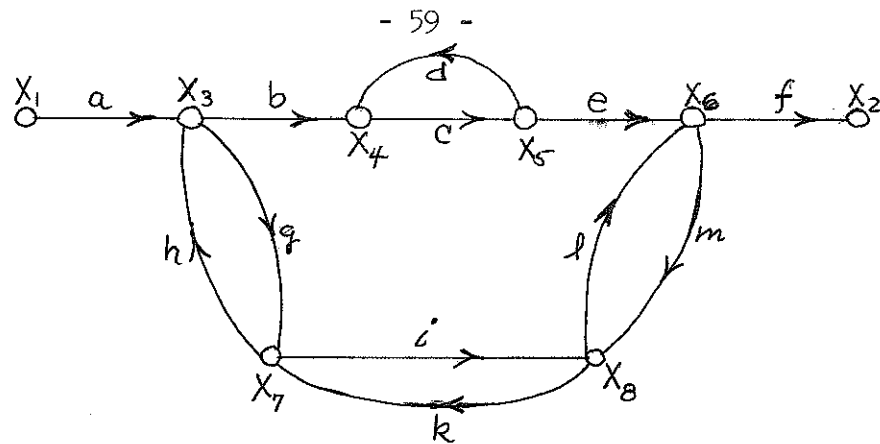


Fig. 21

This flow graph is merely a representation of the equations. Let's see what we have;

There are two paths from X_1 to X_2 . These are denoted by

$$G_1 = abcef \text{ and } G_2 = aqilf$$

There are five loops (or "loop gains"), denoted by

$$T_1 = qh$$

$$T_2 = ik$$

These are also called "loop gains" $T_3 = lm$

$$T_4 = cd$$

$$T_5 = bcemkh$$

The gain X_2/X_1 is given by

$$\frac{X_2}{X_1} = \frac{G_1 [1 - () + () - () + \dots] + G_2 [1 - () + () - \dots] + G_3 [\dots]}{1 - () + () - () + \dots} \quad (81)$$

where the G_i 's are defined by the paths. The general rules for determining the quantities in the parentheses are as follows:

For the denominator we have

For the denominator we have

$$1 - \left[\begin{array}{l} \text{Sum of loop gains;} \\ \text{i.e., } \sum T_i \end{array} \right] + \left[\begin{array}{l} \text{Sums of products} \\ \text{of 2 loop gains} \\ \text{which do not touch;} \\ \text{i.e., } \sum T_i T_j \text{ which} \\ \text{do not touch} \end{array} \right] - \left[\begin{array}{l} \text{Sums of products} \\ \text{of 3 loop gains} \\ \text{which do not touch;} \\ \text{i.e., } \sum T_i T_j T_k \text{ which} \\ \text{do not touch} \end{array} \right] + \text{etc.}$$

For the numerator we have

$$G_1 \left[1 - \left[\begin{array}{l} \text{Sum of loop gains} \\ \text{which do not touch} \\ \text{path } G_1 \text{ and do not} \\ \text{touch each other} \end{array} \right] + \left[\begin{array}{l} \text{Sum of products} \\ \text{of 2 loop gains} \\ \text{which do not touch} \\ \text{path } G_1 \text{ and do not} \\ \text{touch each other} \end{array} \right] - \dots + G_2 \left[\dots \right] + \text{etc.} \right]$$

For our example (Figure 20)

$$\frac{X_2}{X_1} = \frac{G_1(1 - T_2) + G_2(1 - T_4)}{1 - (T_1 + T_2 + \dots + T_5) + (T_1 T_3 + T_1 T_4 + T_2 T_4 + T_3 T_4) - T_1 T_3 T_4} \quad (82)$$

Continuing this process, we can solve for each variable in terms of X_1 if so desired, or in terms of any other variable. For each equation, such as Equation (82), we must reconstruct the flow graph with the independent variable as the input and the dependent variable as the output.

By this method then we can solve directly for the relationship between the input and the output.

Suggested References

1. M. F. Gardner and J. L. Barnes, Transients in Linear Systems. New York: John Wiley & Sons, 1956.
2. H. Chestnut and R. W. Mayer, Servomechanisms and Regulating System Design, Volume 1. New York: John Wiley & Sons, 1953.
3. John G. Truxal, Automatic Feedback Control System Synthesis. New York: McGraw-Hill Book Co., 1955.

Suggested References

4. S. J. Mason, Feedback Theory - Some Properties of Signal Flow Graphs, Proceedings IRE, Vol. 41, No. 9, pp 1144 - 1156, September, 1953.
5. G. C. Newton, L. A. Gould, and J. F. Kaiser, Analytic Design of Linear Feedback Controls. New York: John Wiley & Sons, 1957.



LECTURE NO. III

Let's review for a moment to see what has been accomplished up to this point. In our discussion of linear systems we showed how to describe the system by integrodifferential equations which were formulated strictly on the basis of conservation of energy or particles within the system. By using the Laplace transform method for representing the system in the frequency domain, the integrodifferential equations were transformed into algebraic equations and the transform of the response of any linear system was expressed in terms of the Laplace transform of the input function multiplied by the transfer function of the system. In the time domain the input and output are related by the convolution integral.

We have also discussed how to represent a physical system by block diagrams which correspond to the equations describing the system, and how to solve the equations for different response functions in terms of the input function by means of flow graphs. For our purposes we use flow diagrams only as a method of solving a set of linear, independent equations.

We still don't know the response of a system in the time domain because the relationships determined using flow graphs are actually relationships between the transfer function of the system and the Laplace transforms of the input and output. We will now discuss the method of transforming from the frequency domain back into the time domain.

THE INVERSE LAPLACE TRANSFORM

In the examples considered in the second lecture it was shown that the response function for a linear system is represented by a relationship of the form

$$F_o(s) = H(s) \cdot F(s)$$

where $H(s)$ (the system function or transfer function) is an algebraic function of the complex variable s . This is also true of the excitation function $F(s)$ provided the driving function is a constant, an exponential, sinusoid, etc., which will be the case for problems of interest to us. Since the product of two rational algebraic functions is also a rational algebraic function, the response function $F_0(s)$ is a function of the rational algebraic type. Then the final step in the solution of a set of linear, constant-coefficient integrodifferential equations with given boundary conditions reduces ultimately to obtaining the inverse Laplace transform of a ratio of rational polynomials of s .

In the first lecture, the inverse Laplace transform was defined as

$$f(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(s) e^{st} ds, \quad (83)$$

where $F(s)$ is analytic everywhere except at the poles, at which points it blows up. In general, we can express $F(s)$ as a partial fraction expansion

$$F(s) = \sum \frac{k_i}{s-p_i} + \sum \frac{k_j}{(s-p_j)^2} + \dots \quad (84)$$

Therefore, apart from the exact values of the k_i 's, the important aspect of $F(s)$ is the poles; i.e., the p_i 's. Consider a simple term $\frac{1}{s-p_1}$ and let p_1 be on the complex plane as shown in Figure 22.

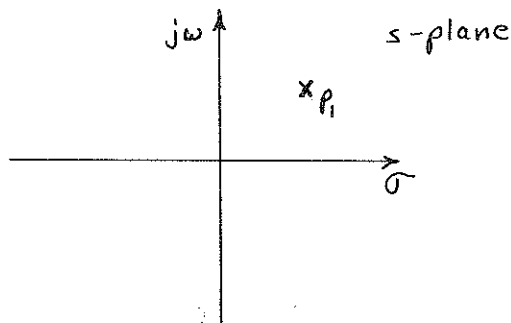


Fig. 22

Now draw a semicircle contour, C with radius R as shown in Figure 23.

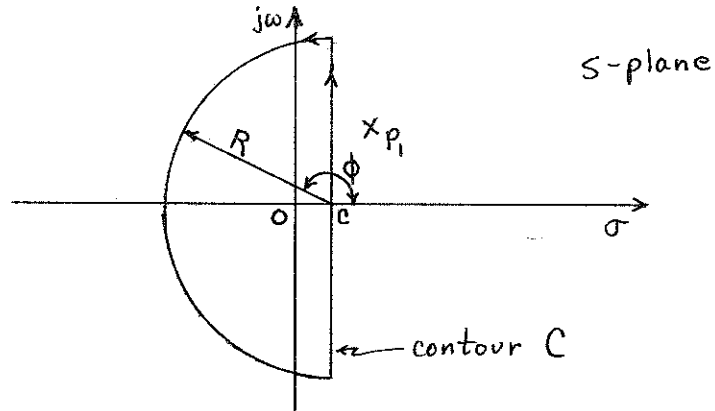


Fig. 23

Notice that C has been drawn to exclude the point p_1 . Then on and within the contour C the function $\frac{e^{st}}{s-p_1}$ has the following properties:

1. It is analytic within and on C. This is true since it is analytic everywhere except where it blows up and this point has been excluded from C

2. The function vanishes on the semicircle for $t > 0$. To visualize this, let R become very large. Then from

$$s = R e^{j\phi} = R [\cos\phi + j \sin\phi]$$

where $\phi > \pi/2$

$$e^{st} = e^{tR\cos\phi} e^{jtR\sin\phi}$$

Since $\cos\phi$ is always negative and R is large,

$$e^{st} \rightarrow 0 \text{ as } R \rightarrow \infty$$

Thus, according to Cauchy's integral theorem

$$\int_{c-j\infty}^{c+j\infty} \frac{e^{st}}{s-p_1} ds = \oint_C \frac{e^{st}}{s-p_1} ds = 0$$

Now let's redraw the contour C to include the point $s = p_1$, as shown in Figure 24.

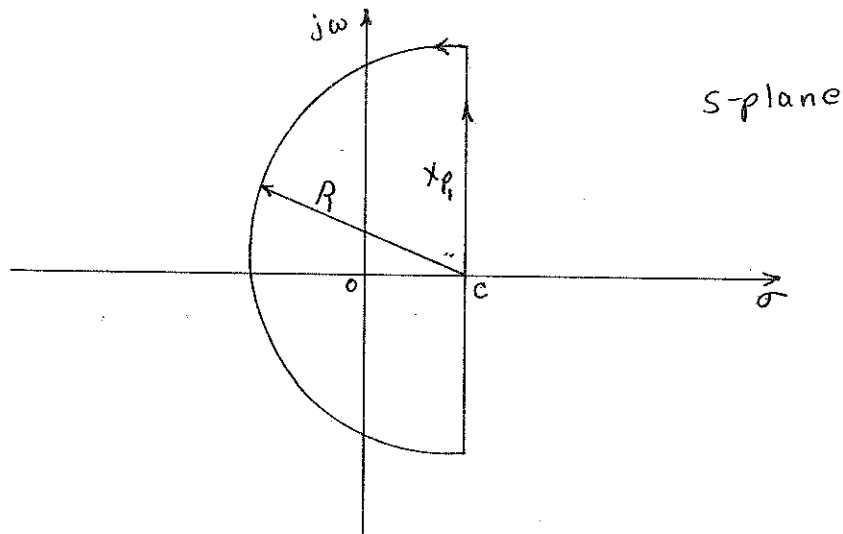


Fig. 24

Then from Cauchy's integral formula,

$$f(t) = \frac{1}{2\pi j} \oint_C \frac{e^{st}}{s - p_1} ds = e^{p_1 t} \quad (85')$$

Thus, when the contour C excludes the pole ($C < \text{Re } p_1$) the inverse Laplace transform vanishes and when the contour includes the pole ($C > \text{Re } p_1$), the function $f(t)$ is defined for $t > 0$. Thus the restriction on $c > \sigma_a$, namely to the right of the real part of the singularity, stems from the fact that when we take the inverse transform we desire to pick up all the poles.

However, by the method of analytic continuation we can remove this restriction as shown in Figure 25.

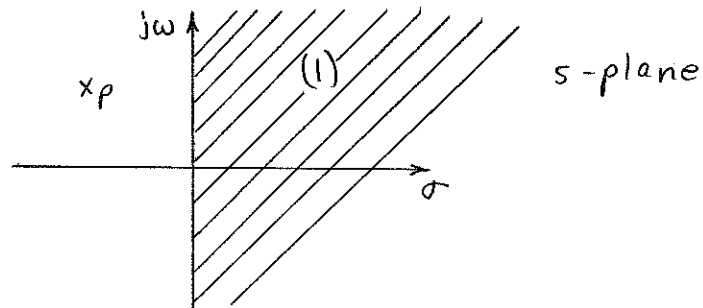


Fig. 25

If the Laplace transform is defined for region (1), then by drawing a small exclusion circle around the pole, p , the function is analytic everywhere except in the small circle.

Now we will see how all this is involved in determining the inverse Laplace transform.

Inverse Transforms of Ratios of Polynomials

As was stated in the first lecture, the functions of complex variables which are of interest to us are ratios of polynomials. We will be concerned then with functions of the form

$$F(s) = F_1(s) + \frac{P_1(s)}{Q(s)} = \frac{a_n s^n + \dots + a_0}{b_m s^m + \dots + b_0}$$
$$= k_{n-m} s^{n-m} + \dots + k_0 + \frac{c_{m-1} s^{m-1} + \dots + c_0}{b_m s^m + \dots + b_0} \quad (86)$$

We can consider $F_1(s)$ as being only mathematical fiction since it corresponds to multiple impulses which do not arise in physical problems. Therefore we concentrate on ratios of polynomials whose numerator is at least one degree lower than the denominator. We have then

$$F(s) = \frac{P_1(s)}{Q(s)} = \frac{k_1^{(n)}}{(s-p_1)^n} + \frac{k_1^{(n-1)}}{(s-p_1)^{n-1}} + \dots + \frac{k_1^{(1)}}{s-p_1} + \dots + \frac{k_2}{s-p_2} + \dots, \quad (87)$$

and we have to find inverse transforms of terms of the form $k/(s-p)$, $k/(s-p)^n$, etc. In order to determine the inverse transform of $F(s)$ then we need only to find the inverse transform of each term in the expansion and add them up.

Let's consider the term $\frac{k}{s-p}$. Then

$$\mathcal{L}^{-1}\left\{\frac{k}{s-p}\right\} = \frac{1}{2\pi j} \int_C \frac{k e^{st}}{s-p} ds \quad (88)$$

Applying Cauchy's integral formula where for the contour C the radius $R \rightarrow \infty$, then the limits of integration are from $\sigma - j\infty$ to $\sigma + j\infty$, and

$$\frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{k e^{st}}{s-p} ds = \frac{k}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{e^{st}}{s-p} ds = k f(\rho) = k e^{\rho t} \quad (89)$$

For the term $\frac{k}{(s-p)^2}$,

$$\mathcal{L}^{-1}\left\{\frac{k}{(s-p)^2}\right\} = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{k e^{st}}{(s-p)^2} ds \quad (90)$$

Equation (90) is of the same form as Equation (34) of the first lecture

$$G'(s_0) = \frac{1}{2\pi j} \int_C \frac{G(s)}{(s-s_0)^2} ds \quad (34)$$

Then

$$\mathcal{L}^{-1}\left\{\frac{k}{(s-p)^2}\right\} = k f'(\rho) = k t e^{\rho t} \quad (91)$$

By Equation (36) this reasoning can be extended such that in general,

$$\mathcal{L}^{-1}\left\{\frac{k}{(s-p)^n}\right\} = \frac{1}{2\pi j} \int_C \frac{k e^{st}}{(s-p)^n} ds = \frac{k}{(n-1)!} t^{n-1} e^{\rho t} \quad (92)$$

We see therefore that implementation of the inverse Laplace transform requires means for finding the k's in the partial fraction expansion and means for finding the roots of the polynomials. Let's assume that we know how to find the roots of a polynomial (a suggested reference is "Highschool Algebra", Ginn Publishing Co., 1910). Our problem then will be concerned with finding values for the k's which are called "residues" of the function $F(s)$ at the poles.

Let's consider the function

$$F(s) = \frac{k_1^n}{(s-p_1)^n} + \frac{k^{(n-1)}}{(s-p_1)^{n-1}} + \dots + \frac{k_1^{(1)}}{s-p_1} + \frac{k_2}{s-p_2} \quad .$$

For the multiple pole p_1 , which is of order n , we have the following procedure.

Multiply $F(s)$ by $(s-p_1)^n$ and find

$$G(s) = (s-p_1)^n F(s) = k_1^n + (s-p_1) \left[\begin{array}{l} \text{Terms} \\ \text{which are} \\ \text{finite for } s=p_1 \end{array} \right] \quad . \quad (93)$$

Therefore

$$k_1^n = \left[(s-p_1)^n F(s) \right]_{s=p_1} \quad . \quad (94)$$

Similarly

$$k_1^{(n-1)} = \frac{d}{ds} \left[(s-p_1)^n F(s) \right]_{s=p_1} \quad , \quad (95)$$

$$k_1^{(n-2)} = \frac{1}{2!} \frac{d^2}{ds^2} \left[(s-p_1)^n F(s) \right]_{s=p_1} \quad , \quad (96)$$

$$k_1 = \frac{1}{(n-1)!} \frac{d^{n-1}}{ds^{n-1}} \left[(s-p_1)^n F(s) \right]_{s=p_1} \quad . \quad (97)$$

This is a useful formula for the residue of $F(s)$ at an n^{th} order pole.

For a simple pole (1^{st} order) $n = 1$ and Equation (93) gives

$$G(p_i) = k_{-1} = \left[(s - p_i) F(s) \right]_{s=p_i} \quad (98)$$

Since we can write the function as

$$F(s) = \frac{P_1(s)}{Q(s)} \quad ,$$

Equation (98) can also be written as

$$k_{-1} = \left[\frac{P_1(s)}{\frac{dQ(s)}{ds}} \right]_{s=p_i} \quad (99)$$

We see then that if $F_0(s)$ has only simple poles

$$\mathcal{L}^{-1}\{F_0(s)\} = f(t) = \sum_i k_i e^{p_i t} \quad , (100)$$

where

$$k_i = \left[(s - p_i) F_0(s) \right]_{s=p_i} \quad .$$

Example

Suppose

$$F_0(s) = \frac{a_1 s + a_0}{(s + \alpha_1)(s + \alpha_2)(s + \alpha_3)} \quad ,$$

where $\alpha_1, \alpha_2,$ and α_3 are real numbers, all different. Then

$$\mathcal{L}^{-1}\{F_0(s)\} = k_1 e^{-\alpha_1 t} + k_2 e^{-\alpha_2 t} + k_3 e^{-\alpha_3 t}; \quad 0 \leq t \quad (101)$$

where

$$k_1 = \left[(s+d_1) \frac{a_1 s + a_0}{(s+d_1)(s+d_2)(s+d_3)} \right]_{s=-d_1}$$

$$= \frac{a_0 - d_1 a_1}{(d_2 - d_1)(d_3 - d_1)}$$

$$k_2 = \left[(s+d_2) \frac{a_1 s + a_0}{(s+d_1)(s+d_2)(s+d_3)} \right]_{s=-d_2}$$

$$= \frac{a_0 - d_2 a_1}{(d_1 - d_2)(d_3 - d_2)}$$

$$k_3 = \left[(s+d_3) \frac{a_1 s + a_0}{(s+d_1)(s+d_2)(s+d_3)} \right]_{s=-d_3}$$

$$= \frac{a_0 - d_3 a_1}{(d_1 - d_3)(d_2 - d_3)}$$

Note that for complex roots the poles appear in pairs which are conjugates. In this case the k_1 's for each pair of poles are also conjugates.

Then

$$\mathcal{L}^{-1}\{F_0(s)\} = f(t) = k_1 e^{p_1 t} + \bar{k}_1 e^{\bar{p}_1 t} + \dots$$

where the bar denotes a conjugate quantity and $p_1 = \sigma + j\omega$. Then

$$f(t) = |k_1| e^{j\phi_1} e^{\sigma t} e^{j\omega t} + |k_1| e^{-j\phi_1} e^{\sigma t} e^{-j\omega t} + \dots$$

$$= 2|k_1| e^{\sigma t} \cos(\omega t + \phi_1) + \dots$$

If $F_0(s)$ has n^{th} order poles,

$$\mathcal{L}^{-1}\{F_0(s)\} = f(t) = \sum_{l=1}^i \sum_{j=1}^n \frac{k_{lj}}{(n-j)!} t^{n-j} e^{p_l t}, \quad (102)$$

where

$$k_{ej} = \frac{1}{(n-j)!} \left[\frac{d^{n-1}}{ds^{n-1}} (s-p_i)^n F_0(s) \right]_{s=p_i}$$

Example

Suppose $F_0(s) = \frac{a_2 s^2 + a_1 s + a_0}{(s+d)^3 s^2}$ where d is a real number.

Then

$$\begin{aligned} \mathcal{L}^{-1}\{F_0(s)\} &= \mathcal{L}^{-1}\left\{ \frac{k_{11}}{(s+d)^3} + \frac{k_{12}}{(s+d)^2} + \frac{k_{13}}{(s+d)} + \frac{k_{21}}{s^2} + \frac{k_{22}}{s} \right\} \\ &= \left(\frac{k_{11}}{2!} t^2 + k_{12} t + k_{13} \right) e^{-dt} + k_{21} t + k_{22} \quad (103) \end{aligned}$$

where

$$k_{11} = \left[\frac{a_2 s^2 + a_1 s + a_0}{s^2} \right]_{s=-d} = \frac{a_2 d^2 - a_1 d + a_0}{d^2},$$

$$k_{12} = \left[\frac{d}{ds} \left(\frac{a_2 s^2 + a_1 s + a_0}{s^2} \right) \right]_{s=-d} = \frac{2a_0 - a_1 d}{d^3},$$

$$k_{13} = \frac{1}{2!} \left[\frac{d^2}{ds^2} \left(\frac{a_2 s^2 + a_1 s + a_0}{s^2} \right) \right]_{s=-d} = \frac{3a_0 - a_1 d}{d^4},$$

$$k_{21} = \left[\frac{a_2 s^2 + a_1 s + a_0}{(s+d)^3} \right]_{s=0} = \frac{a_0}{d^3}$$

and

$$k_{zz} = \left[\frac{d}{ds} \left(\frac{a_2 s^2 + a_1 s + a_0}{(s+d)^3} \right) \right]_{s=0} = \frac{a_1 d - 3a_0}{d^4} .$$

Assuming we now know how to determine the inverse Laplace transform of a given function, let's discuss the complete response of a linear system.

COMPLETE RESPONSE OF A LINEAR SYSTEM

From the general expression

$$F_o(s) = H(s)F(s) ,$$

where $H(s)$ is the transfer function of the system and $F(s)$ is the Laplace transform of the input signal, we know that, in the time domain

$$\mathcal{L}^{-1}\{F_o(s)\} = f_o(t) = \underbrace{\left[\begin{array}{l} \text{terms involving} \\ \text{poles of } H(s) \end{array} \right]}_{\text{Transient term}} + \underbrace{\left[\begin{array}{l} \text{terms involving} \\ \text{poles of } F(s) \end{array} \right]}_{\text{Steady-State term}}$$

and that the time response is a function of both the system and the input signal. Usually, it is desirable that the time response $f_o(t)$ follow directly the input function $f(t)$. However, in most systems this is not the case for all time because distortion is introduced through the system function $H(s)$ which gives rise to a transient term in the time response. After the transient term has died out, the system settles down to the steady-state response. We see then that the time response of a linear system is composed of two parts; a steady-state response and a transient response. Before discussing these components further, let's examine the solution of a general linear second-order differential equation with

constant coefficients. By doing this, we can explicitly point out the salient features of the solution which are common to the solutions of problems in which we will be interested.

Consider the equation

$$A \frac{d^2 y}{dt^2} + B \frac{dy}{dt} + C y = f(t) \quad (104)$$

The solution of this equation in terms of Laplace transforms is

$$\mathcal{L}\{y(t)\} = \bar{Y} = \frac{1}{As^2 + Bs + C} \left[\bar{F} + y(0)(As+B) + y'(0)A \right], \quad (105)$$

where $\bar{F} = \mathcal{L}\{f(t)\}$ and $y(0), y'(0)$ are the initial values of y and its derivative respectively. The time dependent solution is:

$$y(t) = \mathcal{L}^{-1}\{\bar{Y}\} = \mathcal{L}^{-1}\left\{ \frac{F(s) + y(0)(As+B) + y'(0)A}{As^2 + Bs + C} \right\} \quad (106)$$

Assume that the driving function $f(t)$ is a sinusoid. More specifically, $f(t) = F_m \cos(\omega_1 t + \psi)$, in which F_m, ω_1 , and ψ are real constants. Then

$$F(s) = \frac{g s + h \omega_1}{s^2 + \omega_1^2} \quad (107)$$

where

$$g = F_m \cos \psi \quad \text{and} \quad h = -F_m \sin \psi \quad .$$

Substituting Equation (107) into $\dot{Y}(s)$ gives

$$Y(s) = \frac{g s + h \omega_1 + [y(0)(As+B) + y'(0)A](s^2 + \omega_1^2)}{(s^2 + \omega_1^2)(As^2 + Bs + C)} \quad (108)$$

Note that the numerator has been cleared of fractions so that $Y(s)$ is a ratio of polynomials.

In order to determine $y(t)$ we must specify the poles and the corresponding residues of $Y(s)$. To this effect, note that the factor $s^2 + \omega_1^2$ in the denominator is introduced by the driving function and can be written

$$s^2 + \omega_1^2 = (s - s_1)(s - s_2) \quad ; \quad s_1, s_2 = \pm j\omega_1 \quad .$$

The other factor $(As^2 + Bs + C)$ is introduced by the system described by the second order differential Equation (104) and depends solely on the parameters of the system. This factor is usually called the characteristic function of the system and can be written as:

$$As^2 + Bs + C = A(s - s_3)(s - s_4) = A[(s + \alpha)^2 + \omega_d^2] \quad , \quad (109)$$

where

$$\alpha = \frac{B}{2A} \quad ; \quad \omega_d^2 = \omega_0^2 - \alpha^2 \quad ; \quad \omega_0^2 = \frac{C}{A}$$

$$s_3, s_4 = -\alpha \pm j\omega_d = \text{characteristic roots} \quad .$$

The form of $y(t)$ depends upon whether ω_d^2 is positive, zero, or negative.

Let us assume for this example that $A, B, C > 0$ and $\omega_d^2 > 0$ so that the characteristic roots (roots of the characteristic function) are complex and

in the left-half-plane. Thus:

$$y(t) = 2 \operatorname{Re} [k_1 e^{j\omega_1 t}] + 2 \operatorname{Re} [k_2 e^{(-\alpha + j\omega_d)t}] \quad (110)$$

where

$$\begin{aligned} k_1 &= \left[(s - j\omega_1) Y(s) \right]_{s=j\omega_1} = \frac{g - jk}{2A [(\omega_0^2 - \omega_1^2) + 2j\alpha\omega_1]} \\ &= \frac{F_m}{2A [(\omega_0^2 - \omega_1^2)^2 + 4\alpha^2\omega_1^2]^{1/2}} e^{-j(\psi - \theta)} \\ &= C_1 e^{-j(\psi - \theta)} \quad ; \quad (-\theta) = \tan^{-1} \frac{2\alpha\omega_1}{\omega_0^2 - \omega_1^2} \quad (111) \end{aligned}$$

$$\begin{aligned} k_2 &= \left[(s + \alpha - j\omega_d) Y(s) \right]_{s = -\alpha + j\omega_d} \\ &= \frac{1}{2A\omega_d} \frac{m - jn}{d^2 + \omega_1^2 - \omega_d^2 - 2j\alpha\omega_d} \\ &= \frac{1}{2A\omega_d} \left[\frac{m^2 + n^2}{(\omega_0^2 - \omega_1^2)^2 + 4\alpha^2\omega_1^2} \right]^{1/2} e^{j\lambda} \\ &= C_2 e^{j\lambda} \quad ; \quad \lambda = \tan^{-1} \left(-\frac{n}{m} \right) - \tan^{-1} \left(-\frac{2\alpha\omega_d}{d^2 + \omega_1^2 - \omega_d^2} \right) \quad (112) \end{aligned}$$

$$m = \omega_d \left[q + Ay(0)(3\alpha^2 + \omega_1^2 - \omega_d^2) - 2\alpha(By(0) + Ay'(0)) \right],$$

$$n = h\omega_1 - q\alpha - A\alpha y(0)(\alpha^2 + \omega_1^2 - 3\omega_d^2) + (By(0) + Ay'(0))(\alpha^2 + \omega_1^2 - \omega_d^2).$$

In terms of these shorthand notations the final result is:

$$y(t) = 2C_1 \cos(\omega_1 t + \psi - \theta) + 2C_2 e^{-\alpha t} \cos(\omega_d t + \lambda). \quad (113)$$

for $t \geq 0$.

This example brings out several interesting points which are more or less applicable to all linear systems no matter how complicated they are. The first term of $y(t)$ has the same sinusoidal form as the driving function, but for a difference in magnitude and phase. This term is called the steady-state portion of the response. Notice that the magnitude differs from that of the forcing function by a factor equal to the magnitude of the transfer function $1/(As^2 + Bs + C)$ evaluated at $s = j\omega_1$. In addition, the phase differs from that of the forcing function by an angle equal to the phase of the transfer function for $s = j\omega_1$.

The second term in Equation (113) depends on the characteristics of the system and the initial conditions, and is called the transient portion of the response. It arises from the two terms of the partial fraction expansion, corresponding to the two complex roots of the characteristic function. The shape of the transient depends only on the characteristic roots, namely the damping constant α and the angular frequency ω_d (ω_0 is the resonant frequency of the system and is the limit value of ω_d as

the damping goes to zero). The initial values affect only the amplitude and phase of the transient and therefore do not play any important role in the behavior of the transient.

The transient in this example is an exponentially decaying sinusoidal function. The time constant is the inverse of the damping constant and the frequency is equal to ω_d .

Of course if ω_d^2 were negative, the two characteristic roots of the characteristic function would be real and the transient would consist of two decaying exponentials. If $\omega_d^2 = 0$ the transient would consist of an exponential multiplied by the time variable.

Now that we have some idea of how a linear system could behave, we realize that there are three main questions to ask ourselves when analyzing the system:

1. Is the transient response bounded? By asking this question we are referring to the stability of the system.
2. If the answer to question (1) is yes, then we ask, "Is the transient well behaved?" We need to know what maximum values the transient term may obtain and the time constant involved in the disappearance of the transient response.
3. If the answers to questions (1) and (2) are yes, then we want to examine the steady-state portion of the response to establish how well the system follows the input commands.

STEADY STATE RESPONSE

For our discussion let's consider the last question first; i.e., "What is the steady-state response?" Let's specify the excitation

function to be a sinusoid of unit amplitude

$$f(t) = \sin \omega t \quad (114)$$

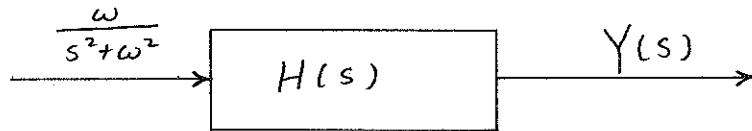
Then

$$F(s) = \frac{\omega}{s^2 + \omega^2} \quad (115)$$

There are two main reasons for choosing the input function to be a sinusoid.

First, if we determine the response of the system for one sinusoid we can do it for many sinusoids by superposition. Second, we can always expand any function in terms of sinusoids and therefore we can determine the response of a linear system for any input function.

With the sinusoidal excitation we want to examine the steady-state response of a linear system described by the following diagram;



Denoting the steady-state portion of $Y(s)$ by $Y(s)_{ss}$, then according to the previous discussion:

$$\mathcal{L}^{-1}\{Y(s)_{ss}\} = \frac{k_1}{s-j\omega} + \frac{k_2}{s+j\omega} \quad (116)$$

where

$$k_1 = \left[(s-j\omega) H(s) \frac{\omega}{(s-j\omega)(s+j\omega)} \right]_{s=j\omega}$$
$$= \frac{H(j\omega)}{2j} \quad (117)$$

and

$$\begin{aligned} k_2 &= \left[(s + j\omega) H(s) \frac{\omega}{(s - j\omega)(s + j\omega)} \right]_{s = -j\omega} \\ &= - \frac{H(-j\omega)}{2j} \\ &= \overline{k_1} \end{aligned} \quad (118)$$

Thus:

$$\begin{aligned} y_{ss}(t) &= k_1 e^{j\omega t} + k_2 e^{-j\omega t} \\ &= H(j\omega) \frac{1}{2j} e^{j\omega t} - H(-j\omega) \frac{1}{2j} e^{-j\omega t} \\ &= |H(j\omega)| \sin(\omega t + \angle H(j\omega)) \end{aligned} \quad (119)$$

Therefore, as was pointed out in the previous example, the steady-state response is a sinusoid of the same frequency although it has different amplitude, $|H(j\omega)|$ and phase $\angle H(j\omega)$ than the input function.

The quantity $H(j\omega)$ is the transfer function of the system for $s = j\omega$. The interesting implication of Equation (119) is that we can measure this transfer function experimentally. Indeed if we excite a linear system by a sinusoid and measure the amplitude and phase of the output for different frequencies we have all the information needed for $H(j\omega)$. In fact, we also have all the necessary information about the steady-state response of the system to any input which we visualize as a sum of sinusoids. Usually this information is presented in terms of magnitude and phase plots of $H(j\omega)$. These are called Bode diagrams.

Later we will see that Bode diagrams are also useful in determining stability characteristics of the system.

Let's now consider a method of plotting the magnitude and phase of the transfer function, namely the "Bode diagrams".

BODE DIAGRAMS

Assume the transfer function of the general form

$$H(s) = k \frac{(1 + \tau_1 s) \left(1 + 2\zeta_1 \frac{s}{\omega_1} + \frac{s^2}{\omega_1^2}\right) (\dots)}{s (1 + \tau_2 s) \left(1 + \zeta_2 \frac{s}{\omega_2} + \frac{s^2}{\omega_2^2}\right) (\dots)} \quad (120)$$

Let $s = j\omega$ and determine the operator

$$D \equiv 20 \log_{10} \quad (121)$$

Taking $20 \log_{10}$ of Equation (120) results in:

$$DH = \frac{Dk + D(1 + \tau_1 s) + D\left(1 + 2\zeta_1 \frac{s}{\omega_1} + \frac{s^2}{\omega_1^2}\right) + \dots}{-Ds - D(1 + \tau_2 s) - D\left(1 + \zeta_2 \frac{s}{\omega_2} + \frac{s^2}{\omega_2^2}\right) + \dots} ; s = j\omega \quad (122)$$

In other words, this operation results in an algebraic sum of similar logarithmic terms, a fact that greatly facilitates the calculation.

To see this clearly, let us first define the unit of decibel. If the magnitude of a quantity is 10 then

$$20 \log_{10}(10) = 1$$

This unit we call the decibel and denote by db. Next consider a typical term of Equation (122) such as $D(1 + \tau_1 s)$, $s = j\omega$ and make a plot of its magnitude as a function of ω_1 . The plot is shown in Figure 26.

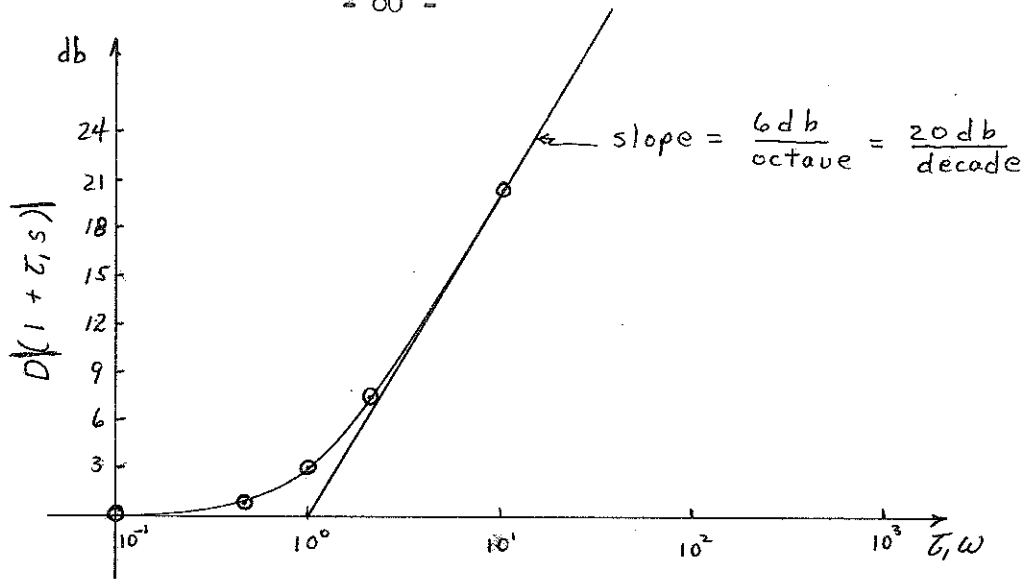


Fig. 26

Notice that for ω very small $1 + \tau_1 s \approx 1$. Therefore $D(1) = 0$ db. For ω very large $1 + \tau_1 s \approx \tau_1 s$. Therefore

$$D|j\tau_1 \omega| = D|\tau_1 \omega|$$

implying that the magnitude plot is linear with a slope of 20 db/decade or 6 db/octave on seimlog paper.

For intermediate frequencies;

If $\tau_1 \omega = 1$ then $D|(1 + j1)| = 3$ db ;

If $\tau_1 \omega = 2$ then $D|(1 + j2)| = 7$ db ;

If $\tau_1 \omega = 0.5$ then $D|(1 + j0.5)| = 1$ db ;

Thus from these few frequencies we can describe a curve which indicates the behavior of $20 \log_{10}$ of $(1 + \tau_1 s)$ as a function of frequency over a large frequency range.

If it is necessary to consider terms such as $-D|(1 + \tau_2 s)|$, then this is merely an inversion on the db axis as shown in Figure 27.

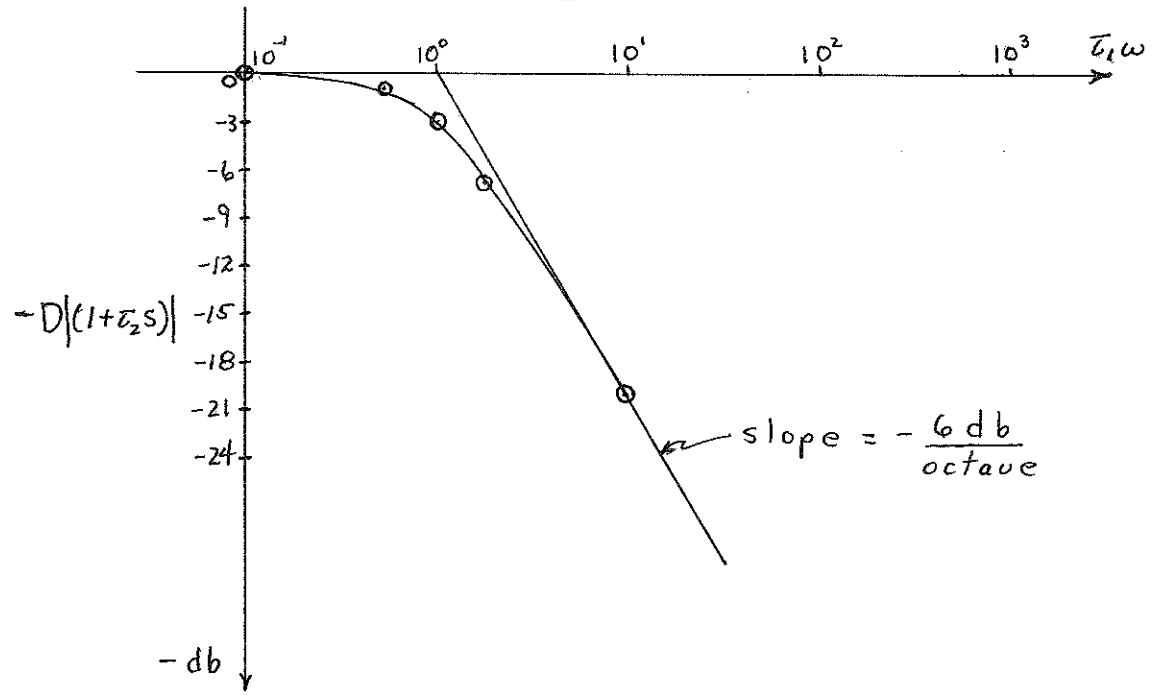


Fig. 27

For the phase of the transfer function we note that

$$\angle H(j\omega) = \angle k + \angle(1+z_1s) + \dots - \angle s - \angle(1+z_2s) - \dots \quad (123)$$

Then for a typical first order term

if $s = j\omega$ and $1 + z_1s = 1 + jz_1\omega$,

for small ω , phase of $(1 + z_1s) = \tan^{-1}(\frac{\omega}{T})$

$$\approx 0^\circ$$

for large ω , say $\omega = 100$, phase = $\tan^{-1}(\frac{\omega}{T})$

$$\approx 90^\circ$$

for $\omega = 1$

$$\begin{aligned} \text{phase} &= \tan^{-1}(1) \\ &= 45^\circ \end{aligned}$$

for $\omega = 0.5$

$$\begin{aligned} \text{phase} &= \tan^{-1}(0.5) \\ &= 26^\circ \end{aligned}$$

for $\omega = 2$

$$\begin{aligned} \text{phase} &= \tan^{-1}(2) \\ &\approx 64^\circ \end{aligned}$$

The plot of the phase of $(1 + \tau_1 s)$ for $s = j\omega$ as a function of $\tau_1 \omega$ is shown in Figure 28.

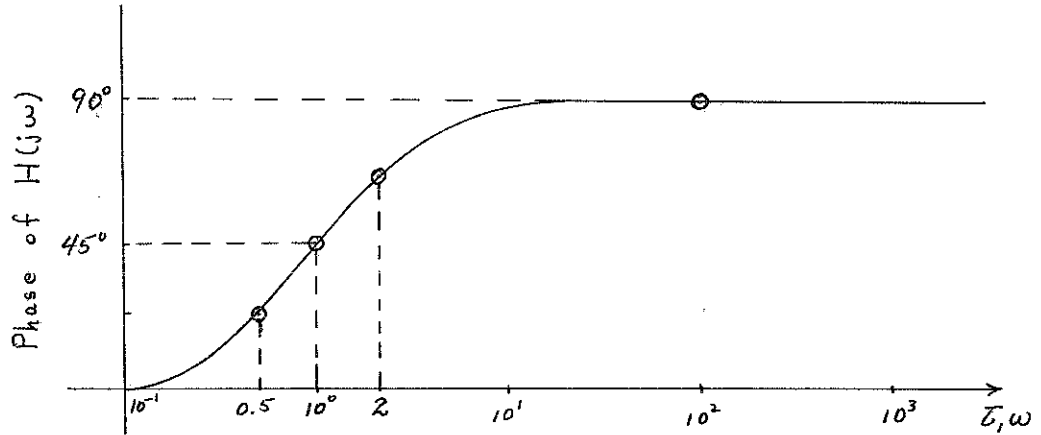


Fig. 28

A similar procedure can be applied to second order terms,

$$\left[\frac{1}{1 + 2\zeta_1 s/\omega_1 + s^2/\omega_1^2} \right] \quad \text{for } s = j\omega$$

The results of the plot of the magnitude of this kind of function versus ω/ω_1 with ζ_1 as a parameter is shown in Figure 29.

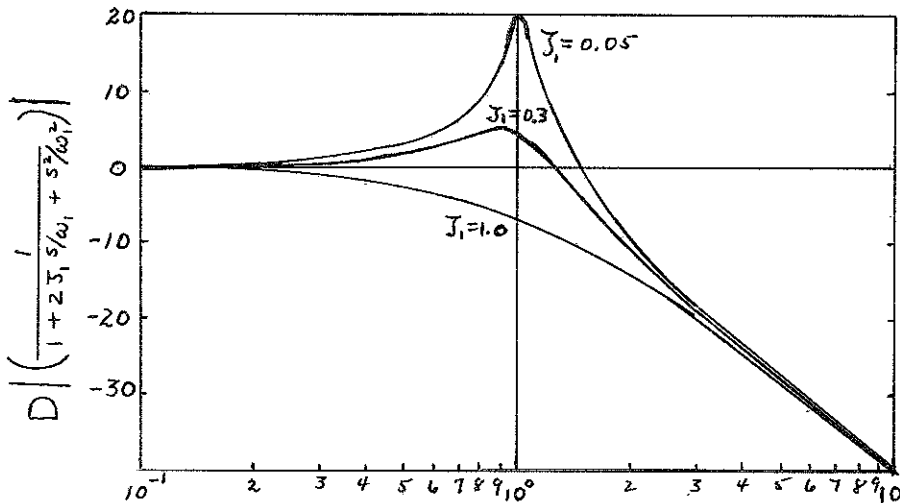


Fig. 29

For very small ω the function is asymptotic with zero slope and for large ω the slope is $-\frac{12 \text{ db}}{\text{Octave}} = -\frac{40 \text{ db}}{\text{Decade}}$. Incidentally the parameter ζ_1 is called the damping factor.

The phase shift for the function $\frac{1}{1 + 2\zeta_1 \frac{s}{\omega_1} + \frac{s^2}{\omega_1^2}}$ is shown in Figure 30.

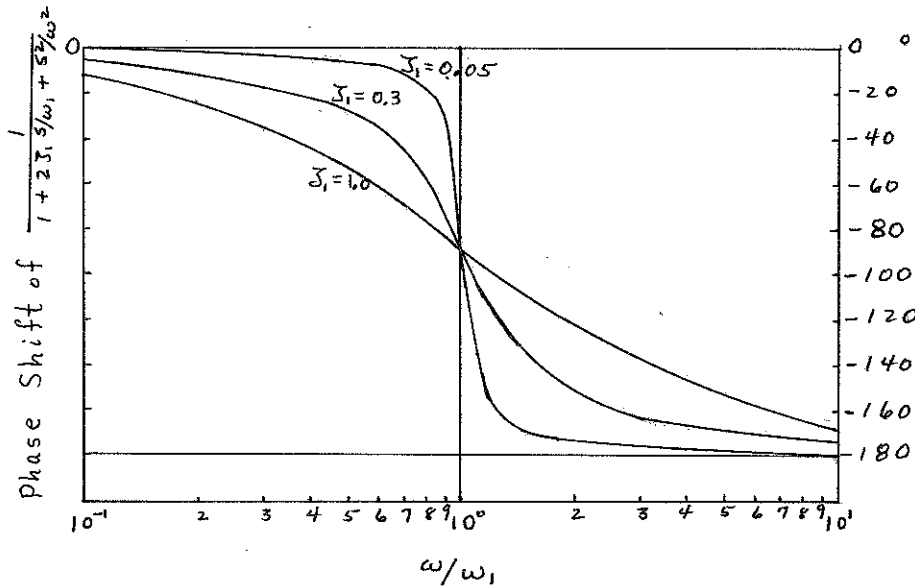


Fig. 30

Similar plots can be derived for terms of the form $\frac{1}{1 + 2\zeta_2 \frac{s}{\omega_2} + \frac{s^2}{\omega_2^2}}$.

Well, since any ratio of polynomials can be written in the form of Equation (120), it is evident how the universal normalized plots 26 - 30 can be used to plot the magnitude and phase of any complicated transfer function. Specifically, one can un-normalize the plots of magnitude and phase for each first and second order term and then add all terms algebraically. (Un-normalize of course means to reduce the plots to plots vs ω rather than $\tau_i \omega$ or ω/ω_i .)

Quite often plotting is greatly facilitated by using the asymptotes of the magnitude plots only. The asymptotes meet (for first order terms $(1 + \tau_1 s)$) at the point $\omega = 1/\tau_1$. For second order terms they meet at $\omega = \omega_c$. This is called the break point.

STABILITY

Let us now examine the first question: "Is the transient response bounded?" Before considering some specific techniques which are extremely helpful in answering this question, let's look at the problem in some generality.

Assume that the Laplace transform of a time dependent function $f(t)$ is

$$F(s) = \frac{a_n s^n + a_{n-1} s^{n-1} + \dots + a_0}{b_m s^m + b_{m-1} s^{m-1} + \dots + b_0} \quad (124)$$

The question is whether $f(t)$ is bounded. To this effect, given $F(s)$ we know that

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \sum_i k_i^{(n)} t^n e^{p_i t} \quad (125)$$

where $k_i^{(n)}$ is the residue at the n^{th} order pole p_i of $F(s)$. Of course $k_i^{(n)}$ is a constant, independent of time. Then the question that we have may be phrased as, "What would make $f(t)$ unbounded?" Evidently, if the $\text{Re } p_i$ of any pole is positive or if $\text{Re } p_i = 0$ but $n \neq 1$ then $f(t)$ is unbounded. The implication of this assertion is that the boundedness of any time function depends only on the sign of the real part of the poles of its Laplace transform. Consequently, for stability considerations it is ^{not} necessary to actually find the time dependent function. All that is needed is to examine the location of the poles of the Laplace transform in question. Specifically, if the poles are in the left-half-plane or they are single and on the j -axis the time function is bounded. If the poles are on the j -axis but are multiple or they are in the right-half-plane, the function is unbounded.

Returning again to the question of boundedness of the transient response, it is obvious that since the transient response is solely dependent on the poles of the transfer function of the system all that is needed to answer question number one is to examine the location of the poles of the transfer function. All stability criteria do exactly that by different but entirely equivalent means.

For example, if the transfer function is a ratio of two polynomials with unspecified poles there are Hurwitz's and Routh's criteria which establish the location of the poles without actually finding their actual value. These criteria will not be discussed here.

Instead we will start the discussion of stability with Nyquist's Criterion of Stability.

NYQUIST STABILITY CRITERION

From the second lecture we have some idea of what a feedback system is. The Nyquist criterion will be applied to feedback systems only. Consider the system shown in Figure 31.

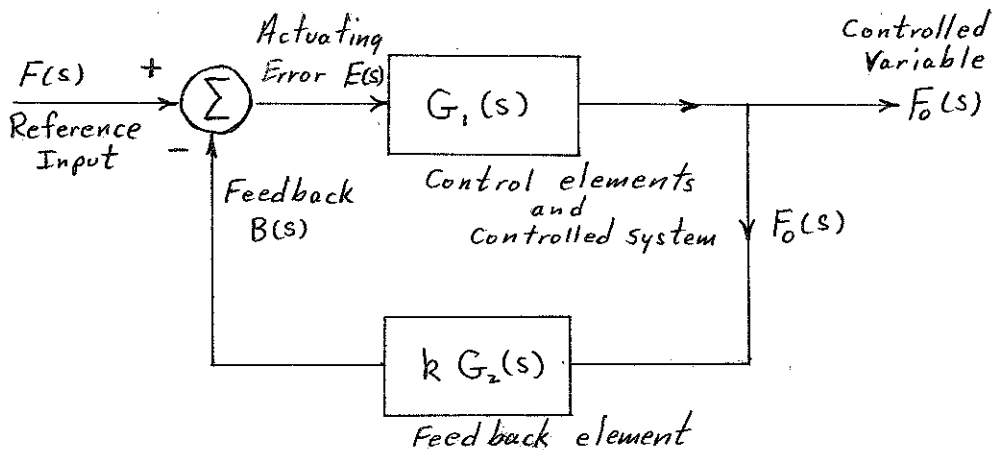


Fig. 31

We want to determine the stability of this system.

Considering the individual gains of the system we have;

$$F_0(s) = G_1(s)E(s) \quad , \quad (126)$$

where $G_1(s)$ is called the "forward gain" ;

$$F(s) - B(s) = E(s) \quad (127)$$

and

$$B(s) = kG_2(s)F_0(s) \quad , \quad (128)$$

where $kG_2(s)$ is the "feedback gain", k being an adjustable constant. Combining these three equations gives

$$F_0(s) = \frac{G_1(s) F(s)}{1 + k G_1(s) G_2(s)}$$

or

$$H(s) = \frac{F_0(s)}{F(s)} = \frac{G_1(s)}{1 + k G_1(s) G_2(s)} \quad , \quad (129)$$

where $H(s)$ is the system transfer function and $kG_1(s)G_2(s)$ is called the "loop gain". Assuming that the input function itself does not blow up, we need be concerned, as already indicated, only with the system function $H(s)$ to determine if the system output is stable. Thus, we want to investigate whether $\frac{G_1(s)}{1 + kG_1(s)G_2(s)}$ has poles in the right or left-half-planes or both.

To this effect, notice that both $G_1(s)$ and $G_2(s)$ may, in general, be ratios of polynomials.

Let us write these functions as

$$G_1(s) = \frac{N_1(s)}{D_1(s)} \quad \text{and} \quad G_2(s) = \frac{N_2(s)}{D_2(s)} \quad , \quad (130)$$

where $N_1(s)$ and $D_1(s)$ are polynomials of s . Thus the transfer function Equation (129) becomes

$$H(s) = \frac{N_1(s) / D_1(s)}{1 + k \frac{N_1(s) N_2(s)}{D_1(s) D_2(s)}} \quad . \quad (131)$$

The problem then of finding the location of the poles of $H(s)$ is identical with the problem of finding the zeros of the denominator of $H(s)$, namely the zeros of the function

$$1 + k \frac{N_1(s) N_2(s)}{D_1(s) D_2(s)} = \frac{D_1(s) D_2(s) + k N_1(s) N_2(s)}{D_1(s) D_2(s)} \quad , \quad (132)$$

or more simply, the zeros of the function

$$D_1(s) D_2(s) + k N_1(s) N_2(s) = 0 \quad . \quad (133)$$

One way of answering this problem is to ask whether knowledge of the location of the roots of the function $D_1(s)D_2(s)$ as well as knowledge of the functional relationship of the loop gain $kG_1(s)G_2(s)$ for different values of s can help in establishing the location of the poles of $H(s)$ or the zeros of $D_1(s)D_2(s) + kN_1(s)N_2(s)$. The answer is yes and the Nyquist criterion does just that. In order to prove this assertion let us divert for a while and consider some simple examples illustrative of this point.

Suppose that we have a simple function

$$G(s) = s - s_1 \quad , \quad (134)$$

where s_1 is given. Consider the contour C (Figure 32a) on the s -plane

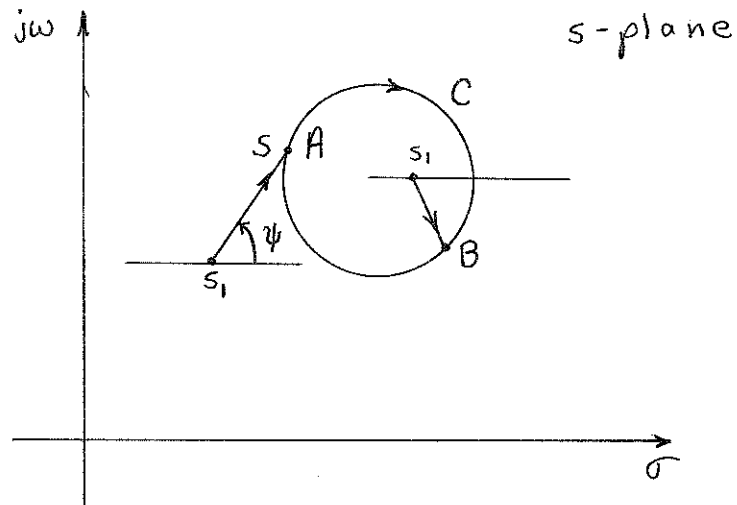


Fig. 32a

and point s_1 on the same plane. Assume that s_1 is outside the arbitrary contour C , and suppose that $G(s)$ is evaluated for all values of s on the contour. It is evident that, as s varies clockwise on the contour, the vector $G(s) = s - s_1$, changes its phase angle by 0° . In fact, if we plot $G(s)$ on the $G(s)$ -plane for the values of s along the contour C , the plot will look as in Figure 32b.

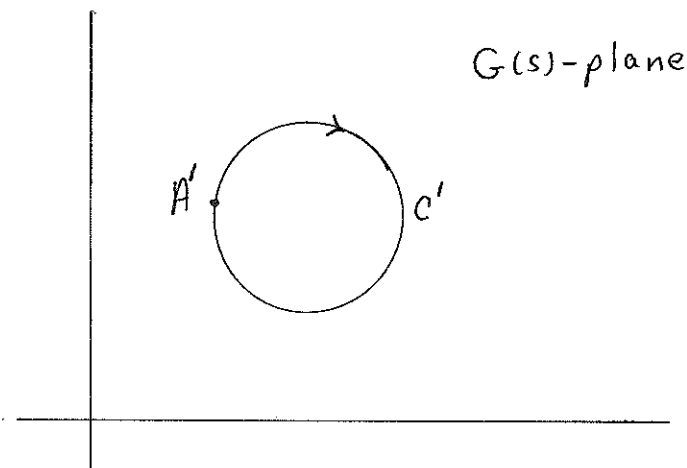


Fig. 32b

The way Figure 32f is drawn, the plot in the $G(s)$ -plane does not encircle the origin because there is one zero and one pole inside the contour C .

This is the end of our diversion. Let us now examine our original problem of finding the zeros of the function $1 + kG_1G_2$ that lie in the right-half-plane using the method of Nyquist.

Suppose that we know the zeros of $1 + kG_1G_2$ as well as its poles. Furthermore, suppose that these poles, namely the roots of $D_1(s)D_2(s)$, are all in the left-half-plane. Then make a pole-zero plot of $1 + kG_1G_2$ (Figure 33). According to our previous discussion, if we choose a contour C

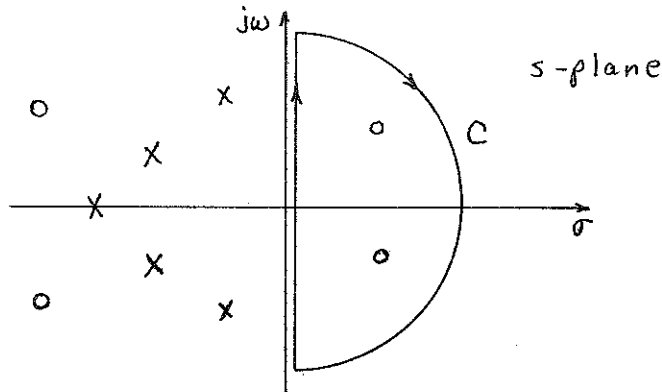


Fig. 33

such that it covers the entire right-half-plane (i.e., a contour defined by the $j\omega$ -axis and a semicircle of infinite radius) and make a plot of $1 + kG_1G_2$ for all values of s along this contour, then this plot will encircle the origin clockwise as many times as there are zeros of $1 + kG_1G_2$ inside the contour C . If the number of encirclements is zero, then there are no zeros in the right-half-plane; that is, there are no poles of $H(s)$ in the right-half-plane and the system is stable. If the number of encirclements is different than zero, then the system is unstable. At long last this is the Nyquist Criterion Stability.

Of course if the roots of $D_1(s)D_2(s)$ do lie in the right-half-plane then the number of clockwise encirclements of the origin of the $(1+kG_1G_2)$ -plane is equal to the zeros and the poles. If the number of the roots of $D_1(s)D_2(s)$ in the right-half-plane is known then Nyquist's criterion is fairly easy to implement. However, if this number is not known, the criterion is ambiguous.

In actual fact, the infinite semicircle of the C-contour necessary for Nyquist's criterion of stability is not needed. The reason is that for all physical systems the loop gain kG_1G_2 approaches zero or a constant for $s \rightarrow \infty$. Therefore, knowledge of the values of the loop gain for $s = j\omega$ for all values of ω is adequate to examine the stability of the feedback system.

There is even another way of presenting the Nyquist criterion of stability. Notice that in plotting the function $1+kG_1G_2$ for values of s along the contour C of Figure 33, we add the constant unity vector to the complex number kG_1G_2 . Thus we might as well plot kG_1G_2 directly and, instead of considering encirclements of the origin of the $(1+kG_1G_2)$ -plane we consider encirclements of the point $(-1,0)$ of the kG_1G_2 -plane and interpret the encirclements in an identical fashion. Actually this is the way that Nyquist's criterion is used in practice.

In summary then, in order to apply the Nyquist criterion of stability do the following:

First, consider the loop gain kG_1G_2 .

Second, make a plot of this gain for $s = j\omega$ (in other words, plot on a plane the values of kG_1G_2 for all $s = j\omega$).

In making this plot it is convenient to assume $k = 1$ and consider encirclements of the point $(-\frac{1}{k}, 0)$.

Third, examine the number of clockwise encirclements of the point $(-\frac{1}{k}, 0)$ by the contour G_1G_2 . If the number of poles of G_1G_2 in the right-half-plane is zero and the number of clockwise encirclements is zero then the system is stable. If the number of poles of G_1G_2 in the right-half-plane is different than zero the system is stable only when the number of encirclements counterclockwise is equal to the number of poles of G_1G_2 in the right-half-plane.

Frequently we need to consider systems where the term $kG_1(s)G_2(s)$ has $s = 0$ as a root of the denominator. As an example, consider the loop gain

$$k G_1(s) G_2(s) = \frac{k(1 + \tau_2 s)}{s(1 + \tau_1 s)(1 + \tau_3 s)} \quad (135)$$

For s equal to $j\omega$, as ω approaches zero from both the positive and negative values of ω we would expect to get infinite values of $kG_1(s)G_2(s)$. Plotting $kG_1(j\omega)G_2(j\omega)$ on the complex plane we would get a plot as shown in Figure 34. It is important to determine how this plot is joined

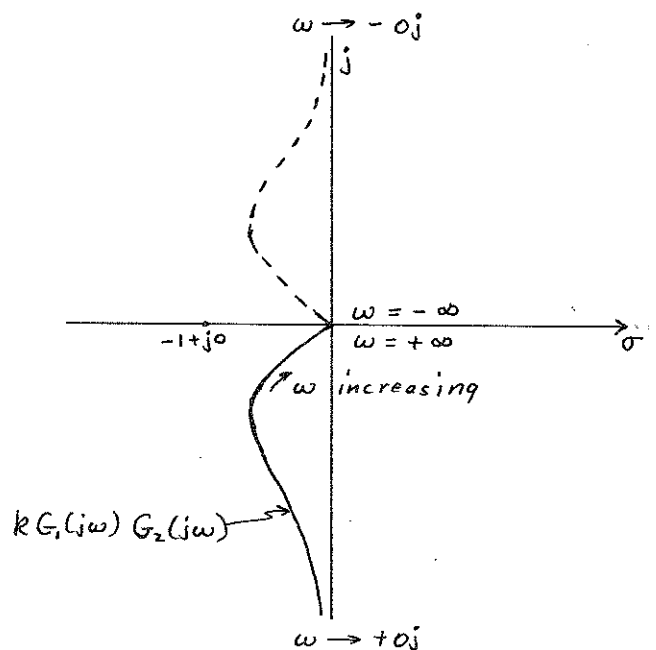


Fig. 34

from $\omega = -0j$ to $\omega = +0j$ since, if the $kG_1(j\omega)G_2(j\omega)$ contour is closed around the point $-1 + j0$, the system is unstable. This follows from the fact that a counterclockwise rotation of the -1 point would be realized as ω changed from $-\infty$ to $+\infty$, whereas Equation (135) shows that there are no poles with positive real parts.

The nature of the plot in the neighborhood of $\omega = 0$ may be resolved by considering the contour C to be along the negative imaginary axis from $s = -j\infty$ until $s = -j0$ gets very close to zero. Then let the path be a semicircle in the positive-half-plane of a very small radius until it reaches the positive imaginary axis at a very small value of $s = +j0$; after which it continues along the positive imaginary axis until $s = +j\infty$.

For the semicircular portion of the path

$$s = \delta e^{j\theta} \quad , \quad (136)$$

where $\delta \rightarrow 0$ and $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$. An expanded plot of this portion of the contour C is shown in Figure 35.

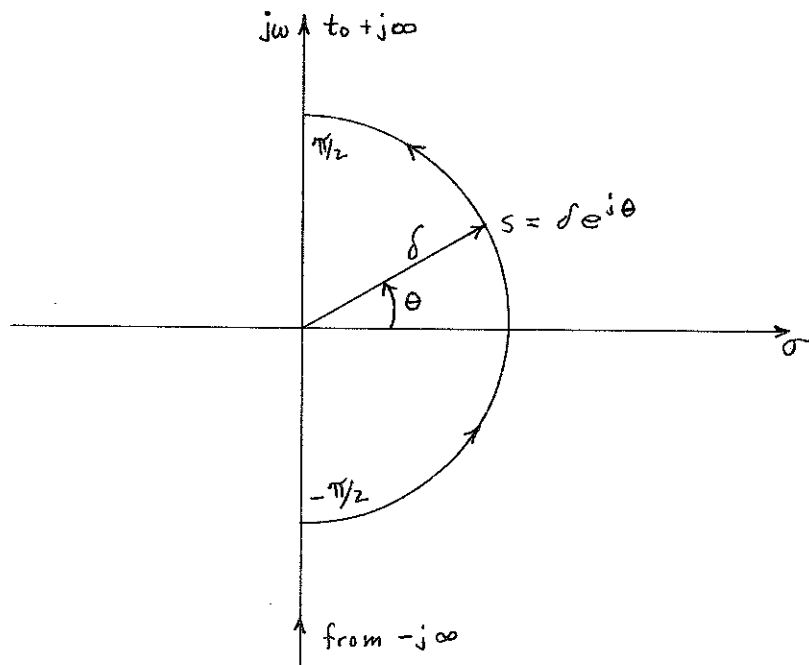


Fig. 35

Now consider the plot of $kG_1(s)G_2(s)$ for $s \rightarrow 0$.

$$k G_1(s) G_2(s) \approx \frac{k}{s} \quad (137)$$

Substituting Equation (136) in Equation (137) gives

$$k G_1(s) G_2(s) = \frac{k}{s} e^{-j\theta} \quad ; \quad \begin{matrix} -\frac{\pi}{2} < \theta < \frac{\pi}{2} \\ s \rightarrow 0 \end{matrix} \quad (138)$$

From this then, the magnitude of $kG_1(s)G_2(s) \rightarrow \infty$ as $s \rightarrow 0$, and the angle of Equation (138) goes from $\pi/2$ to $-\pi/2$ as θ goes through values of $-\pi/2$ to $\pi/2$. In Figure 34 this means that the points for $\omega \rightarrow -0j$ and $\omega \rightarrow +0j$ are joined by means of a semicircle of infinite radius in the first and fourth quadrants. The plot of $kG_1(s)G_2(s)$ would then finally look as shown in Figure 36 which indicates the system to be stable.

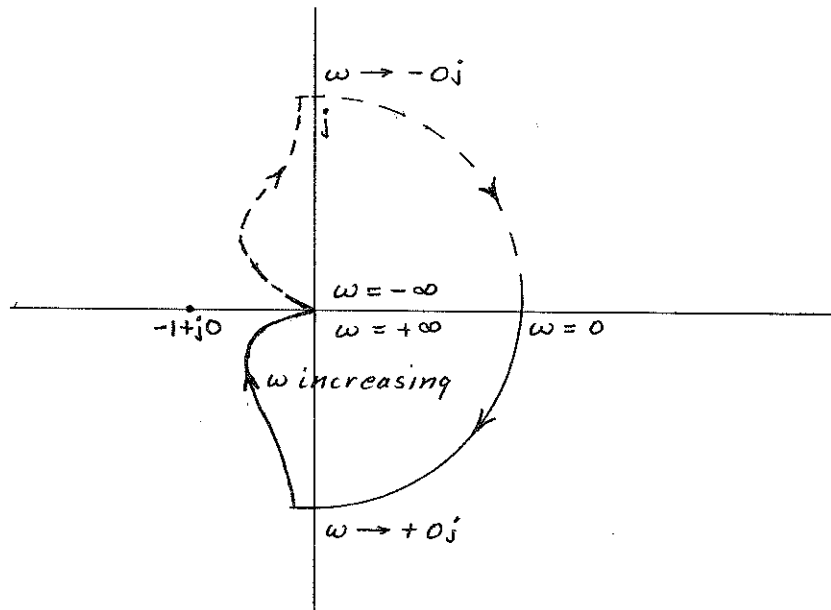


Fig. 36

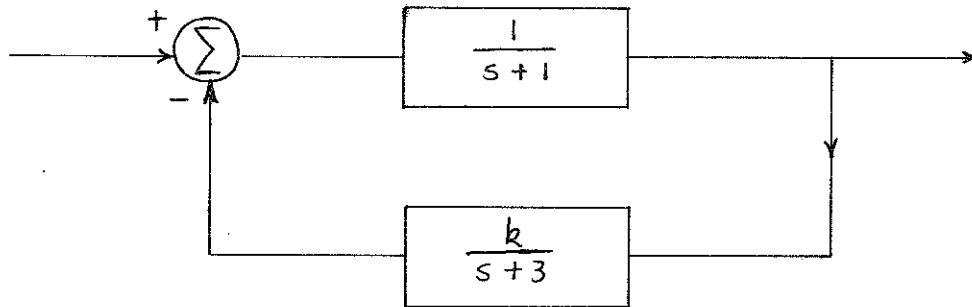
By similar reasoning, the complex plane plot for other cases where s^n occurs in the denominator of $kG_1(s)G_2(s)$ can be determined to show that when ω passes through zero, the $kG_1(s)G_2(s)$ plot makes n clockwise semi-

circles of infinite radius about the origin.

The conjugate nature of the $G_1(+j\omega)G_2(+j\omega)$ and $G_1(-j\omega)G_2(-j\omega)$ means that the plot of $G_1(s)G_2(s)$ for values of $-\infty < \omega < 0$ and $+\infty > \omega > 0$ is symmetrical about the real axis. Hence, if the shape of the plot is known for the range of values of $0 < \omega < +\infty$, it is not necessary to calculate the data for the range $-\infty < \omega < 0$.

EXAMPLES

Example 1



For this system

$$H(s) = \frac{\frac{1}{s+1}}{1 + k \left(\frac{1}{s+1}\right)\left(\frac{1}{s+3}\right)}$$

and

$$k G_1(s) G_2(s) = \frac{k}{(s+1)(s+3)}$$

A pole-zero plot of the loop gain on the complex s -plane is shown in Figure 37. Shown in Figure 38 is the plot on the complex plane of the function $G_1(s)G_2(s) = \frac{1}{(s+1)(s+3)}$ as $s = j\omega$ goes clockwise from $-j\infty$ to $+j\infty$.

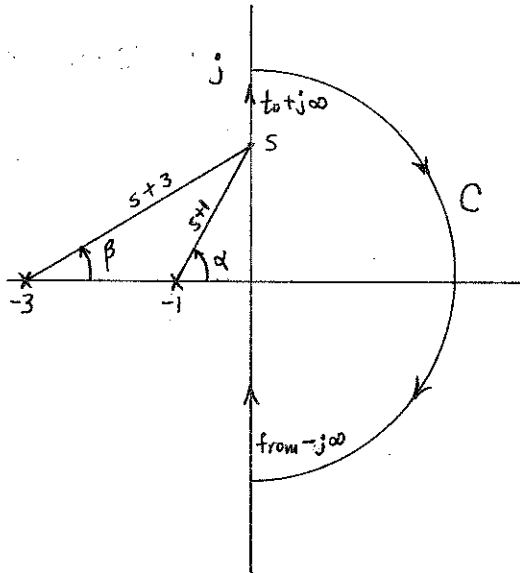


Fig. 37

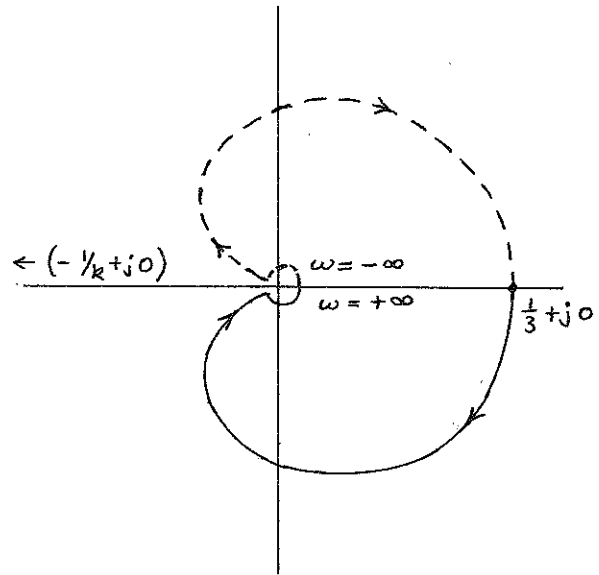


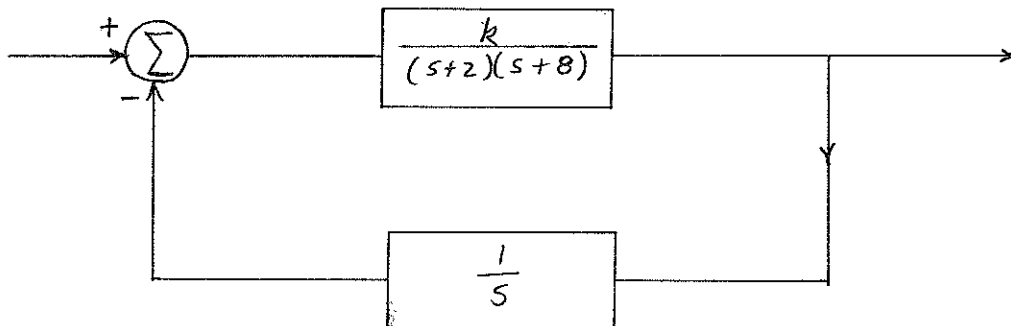
Fig. 38

Figure 38 is obtained from a few frequencies in the following manner:

At $s = j\omega = 0$, $G_1(j\omega)G_2(j\omega) = 1/3$ with zero angle.

As s approaches $+j\omega = +j\infty$, the angles α and β associated with $s + 1$ and $s + 3$, respectively, are always positive, making the angle associated with $G_1(s)G_2(s)$ negative. Also as $s = j\omega$ approaches $s = +j\infty$, $G_1(s)G_2(s)$ approaches zero and the angle approaches -180° . Actually, if the contour C in the s -plane included the infinite semicircle in the positive plane, the plot of $G_1(s)G_2(s)$ does not pass through zero but rather encircles the origin by a small circle of infinitely small radius. Since the point $(-1/k, j0)$ is never encircled, the system of this example is always stable, regardless of the value of k .

Example 2



For this system

$$H(s) = \frac{\frac{k}{(s+2)(s+8)}}{1 + \frac{k}{s(s+2)(s+8)}}$$

and

$$k G_1(s) G_2(s) = \frac{k}{s(s+2)(s+8)}$$

Since there is a pole at the origin $s = 0$, we can resolve the difficulties involved by letting the contour C have a small semicircle around $s = j\omega = 0$, as explained previously. The pole-zero plot then is as shown in Figure 39, and the complex plot of $kG_1(s)G_2(s)$ is shown in Figure 40. Depending on the value of k , the system may or may not be stable.

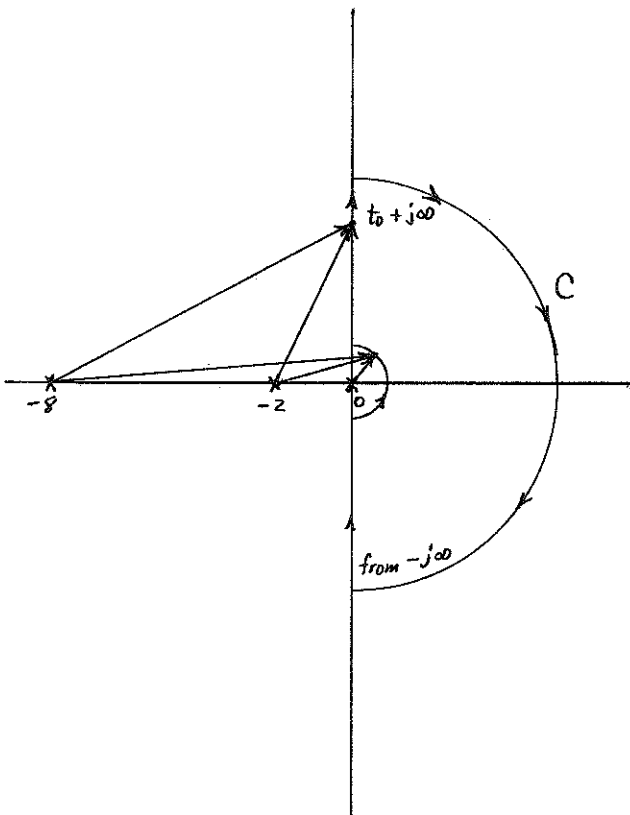


Fig. 39

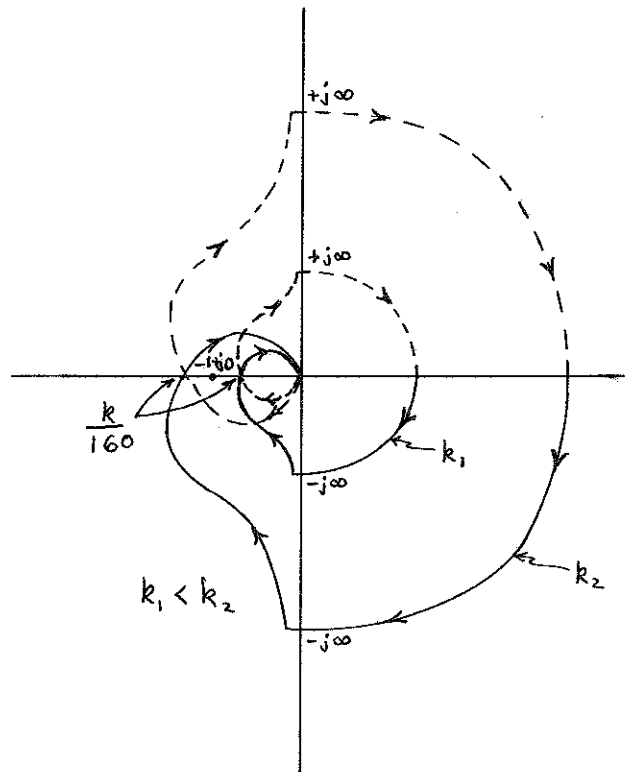


Fig. 40

As shown in Figure 40, for one value of k the system is stable and for another it is unstable. For this example, if $0 < k < 160$ the system is stable. If $k < 0$ the system is unstable since for negative values of k the plot is the reflection in the left-half-plane and the point $-1 + j0$ is always encircled. If $k > 160$, there is also instability.

Suggested References

1. M. F. Gardner and J. L. Barnes, Transients in Linear Systems. New York: John Wiley & Sons, 1956.
2. H. Chestnut and R. W. Mayer, Servomechanisms and Regulating System Design, Volume 1. New York: John Wiley & Sons, 1953.
3. John G. Truxal, Automatic Feedback Control System Synthesis. New York: McGraw-Hill Book Company, 1955.

LECTURE NO. IV

As a means of orientation, let's review what was covered in the third lecture.

Having described a physical system in terms of physical processes by writing integrodifferential equations, the Laplace transform method was used to obtain algebraic equations in the frequency domain. The behavior of a linear system was then described in the time domain by taking the inverse Laplace transform of the response transform in the frequency domain. From the inverse Laplace transform we could identify a steady-state term and a transient term.

In order to analyze a linear system then, we determined that we need to be able to answer the following three questions:

1. Is the system stable?
2. If the system is stable, what is the form of the transient response term?
3. What is the steady-state response term; i.e., how does the response of the system compare with the input function after the transient term has died out?

We discussed the question of the steady-state behavior first and found that it can be described by means of the transfer function of the system, $H(s)$, for values of $s = j\omega$ along the $j\omega$ -axis over the range $-\infty < \omega < +\infty$. Thus, by determining the magnitude and the phase of the transfer function for a sinusoidal input, we can analyze the system for any input since any input can be described by sinusoids. Bode diagrams were developed [plots of $20 \log_{10} |H(j\omega)|$ and $\angle H(j\omega)$ versus ω] for determining the steady-state response and some simple examples of Bode diagrams were described. The steady-state study (Bode diagrams) led us to a simple means of determining

the transfer function of a linear system experimentally since all that we have to do is excite the system with sinusoids of various frequencies and plot the amplitude and phase of the output of the system as a function of these frequencies.

We next discussed the stability of a linear system and determined that it is not necessary to know the time behavior to determine stability but that we need only to investigate the sign of the real part of the poles of the transfer function. If the real part of any pole is positive the system is unstable and if the real parts of all the poles are negative or zero (but then the poles are single) the system is stable.

There are several methods available for investigating the problem of stability and some may be more suitable for a given system than others. The methods of Hurwitz and Routh are particularly applicable when it is desired only to determine the sign of the real part of the roots of a function given in the form of a high order polynomial. The Nyquist criterion is useful both analytically and experimentally, provided certain conditions are fulfilled.

Analytically the system must be described as a feedback system consisting of a forward gain, feedback gain and an adjustable parameter k . Then the Nyquist criterion is applicable as described in Lecture III.

Experimentally one needs to be able to break the feedback loop in order to determine the open-loop gain since now the gain is just kG_1G_2 . We can measure this gain by exciting the system with sinusoids of different frequencies.

An example of an experiment of this sort would be possible with a nuclear reactor. For a reactor we can measure the forward gain by operating at low power where the feedback mechanisms do not interact. We can also determine the overall system transfer function by operating at high power. From these two conditions then we can determine the feedback gain, and therefore deduce

the loop gain. To the derived loop gain we could then apply the Nyquist criterion and establish the range of stable operation of the reactor.

We will now discuss a method of determining the behavior of the poles of the transfer function known as the "Root-Locus" method.

ROOT-LOCUS

The problem we are concerned with is that for the equation $1 + kG_1(s)G_2(s) = 0$; we want to determine the sign of the real parts of its roots. Let's rewrite the equation in the form

$$G_1(s)G_2(s) = -1/k \quad . \quad (139)$$

Since $G_1(s)G_2(s)$ is a complex number, for Equation (139) to be valid the following two conditions must be satisfied.

$$1. \quad |G_1(s)G_2(s)| = |1/k| \quad ; \quad (k > 0) \quad (140a)$$

$$2. \quad \angle G_1(s)G_2(s) = \angle -1/k = 180^\circ + 2\pi n \quad (140b)$$

where n is an integer.

For different values of k , there will be different values of s which satisfy conditions (140a, b). In fact, as k varies from $0 \rightarrow \infty$, the values of s that satisfy Equations (140a, b) move continuously along certain paths which constitute the root-locus of the equation

$$1 + kG_1G_2 = 0 \quad ; \quad 0 < k < \infty \quad .$$

The continuity of the paths of the root-locus stems from the fact that G_1G_2 is a continuous function.

Regardless of the particular value of k , it turns out that the phase equality (140b) is adequate to plot the locus. To see this point in a simple way, consider first a simple example such as:

$$s(s+2) + k = 0 \quad .$$

For a given value of k we can easily solve for the values of s . However, we will approach the problem differently.

The angle associated with a complex number is called the argument of the complex number; i.e., $\angle s = \arg s$, etc. We have then

$$\arg s + \arg(s+2) = \arg(-k) = 180^\circ + 2\pi n$$

For $k = 0$ we notice that the solutions of the previous equation are $s = 0$, $s = -2$ (Figure 41). For other values of k the solutions will be somewhere on the s -plane. Where they are exactly can be determined with the help of the phase condition. Indeed, let us first ask, "Is any value of s between $s = 0$ and $s = -2$ a solution for some undefined values of k ?" In Figure 41

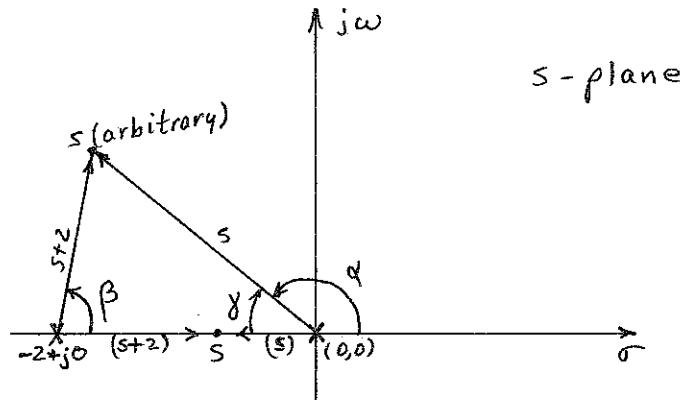


Fig. 41

let s be any value on the real axis between $s = 0$ and $s = -2$, then the vector s is in the direction from $s = 0$ to s with an angle of 180° as shown. Also, the vector $s + 2$ is in the direction from $s = -2$ to s with zero angle as shown. Then the sum of the angles for these two vectors gives

$$\arg s + \arg(s+2) = 180^\circ + 0 = 180^\circ$$

and satisfies the phase condition. Thus, values of s on the real axis between $s = 0$ and $s = -2$ are possible roots of the equation $s(s+2)+k = 0$.

Let s be an arbitrary complex number. Then the vectors s and $s+2$ are those drawn from $s = 0$ to s and from $s = -2$ to s , respectively, as shown in Figure 41. The angles involved are α for the vector s and β for the vector $(s+2)$. Now:

$$\arg s + \arg(s+2) = \alpha + \beta \quad ,$$

and only the values of s which give $\alpha + \beta = 180 + n2\pi$ are possible solutions. But $\alpha = 180 - \gamma$. Consequently only points s for which $\beta = \gamma$ are possible solutions. This implies that only points s which lie on the vertical drawn through the mid-point of the interval $(0, -2)$ are possible solutions of the given equation for positive values of k .

Thus the root-locus is as shown in Figure 42 where it is seen that the locus goes in from $s = 0$ and $s = -2$ to $s = -1$ and then splits up and down perpendicular to the real axis. From Figure 42 we see that the roots have always negative real parts. If this equation were the denominator of some transfer function, the system would be stable for all positive values of k .

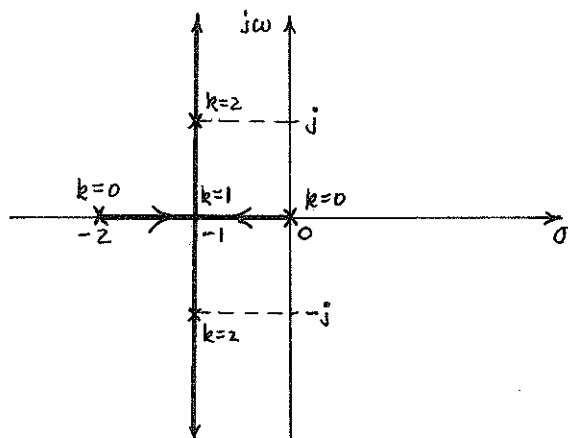


Fig. 42

Now, since we know what values of s are possible solutions (determined entirely from the phase requirement) we can determine what values of k correspond to each point of the locus s from the magnitude condition,

$$|s(s+2)| = |k|$$

For example; for the double root $s = -1$, $k = 1$. For $s = -1 \pm j$, $k = 2$, etc.

For another example, let us find the root-locus of:

$$1 + \frac{k}{s(s+1)(s+4)} = 0 ; k > 0$$

which may be visualized as the denominator of the transfer function of a feed-

back system. Accordingly, for $k = 0$, the poles of the loop gain are the solutions of the equation. These are $s = 0, -1, -4$. Now for the values of s on the real axis for $s < -4$, from Figure 43, which is a pole-zero plot of the loop gain, we have

$$\begin{aligned} \arg s + \arg(s+1) + \arg(s+4) &= 180 + 180 + 180 \\ &= 180 + 2\pi \end{aligned}$$

and these values of s are possible solutions. For values of

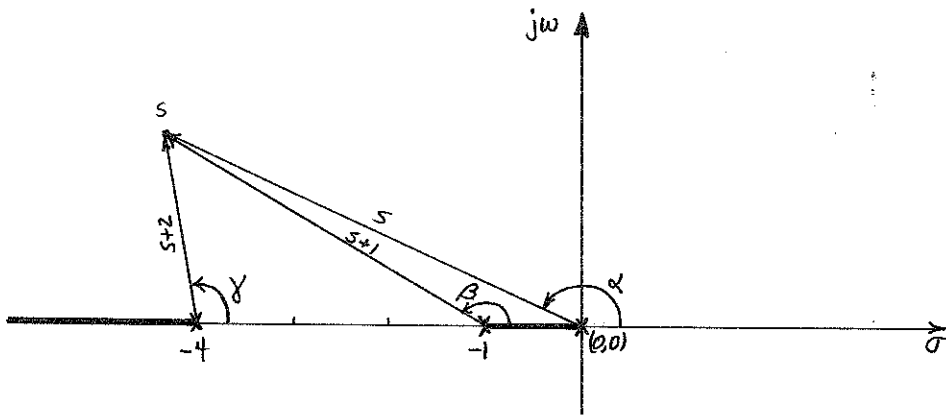


Fig. 43

s between $s = 0$ and $s = -1$ we have

$$\arg s + \arg(s+1) + \arg(s+4) = 180 + 0 + 0 = 180^\circ,$$

and these values of s are also possible solutions. The other segments of the real axis do not satisfy the phase condition and therefore cannot possibly contain values of s which satisfy the given equation for any value of k . In Figure 43 the regions of the real axis which contain possible solutions are indicated by the heavier lines.

If s is an arbitrary complex number, in order for it to be a solution we must again have (Figure 43)

$$\arg s + \arg(s+1) + \arg(s+4) = \alpha + \beta + \gamma = 180^\circ + 2\pi n.$$

In order to see which complex numbers satisfy the phase condition it is helpful to consider the following steps. First, let's consider what happens

when $k \rightarrow \infty$. From the equation

$$1 + \frac{k}{s(s+1)(s+4)} = 0$$

we get

$$s(s+1)(s+4) + k = 0$$

and, if s is large compared to 1 and 4,

$$s^3 \simeq -k \quad ,$$

$$s \simeq \pm \sqrt[3]{-k} \quad .$$

Thus, if $k \rightarrow \infty$, $\sqrt[3]{-k} \rightarrow \infty$ and s is also infinite. From the phase condition though, we see that

$$\angle s^3 = 3 \angle s = \angle -k = 180 + 2n\pi \quad .$$

If $s \rightarrow \infty$ the vectors s , $s+1$ and $s+4$ appear as if they all originate at the origin and the angles of α , β and γ are all equal. Thus, the phase requirement imposes that s must have associated with it an angle of 60° , 180° , or 300° . This follows from the fact that we can write

$$s = Ae^{j\phi}$$

$$s^3 = A^3 e^{3j\phi}$$

and for the phase condition to be satisfied 3ϕ must equal $180^\circ + 2n\pi$. This will be true only if $\phi = 60^\circ$, 180° or 300° .

Thus we see that for large k and large complex s , the values of s approach those which lie on asymptotes at 60° and 300° . These asymptotes are shown in Figure 44.

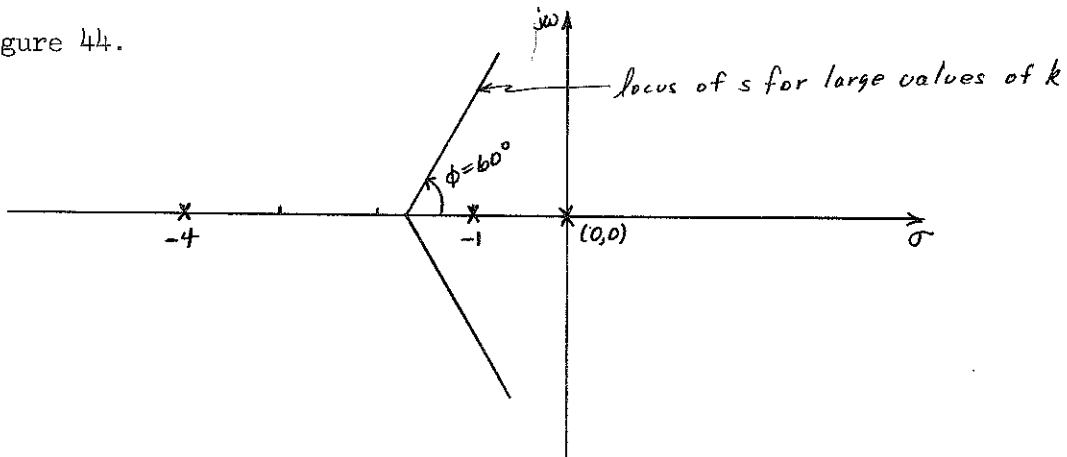


Fig. 44

The point of intersection of the asymptotes with the real axis can be found in the following manner. From the equation $s(s+1)(s+4) + k = 0$, it is seen that the sum of the roots is always -5 , regardless of the value of k . The sum of the roots is the negative of the coefficient of the second highest order term in s ; i.e., the negative of the coefficient of the s^2 term above. If there is no second highest order term the sum is zero. This is true for any polynomial. Since the sum of the roots is always the same, the asymptotes must meet at the point determined by the sum of the roots divided by the number of roots. In this case, the point is $-5/3$. This is the "center of gravity" of the roots.

From continuity considerations, we can show that on the real axis as k increases from zero to infinity, the root $s = -4$ moves to infinity while the roots $s = -1$ and $s = 0$ move toward each other, meet somewhere between $(0, -1)$ and then split toward the two asymptotes at 60° and 300° . We can find the break point analytically from geometric considerations and from the phase requirement. Figure 45 is an expanded pole-zero plot of the poles of the loop gain.

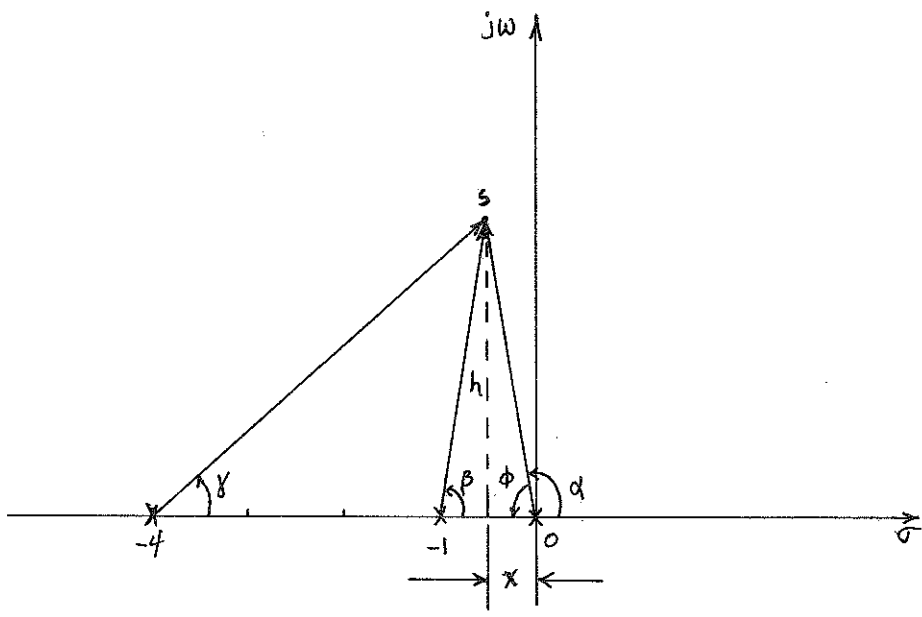


Fig. 45

Let the point s be complex but very close to the real axis so that we know

it is a possible root. Then we know that

$$\alpha + \beta + \gamma = 180^\circ \quad . \quad (141)$$

Also we know that

$$\alpha = 180 - \phi \quad . \quad (142)$$

Substituting this into Equation (141)

$$180^\circ - \phi + \beta + \gamma = 180$$

or

$$\phi = \beta + \gamma \quad . \quad (143)$$

Taking the tangent of both sides, since γ and β are very small angles, we have that $\tan(\beta + \gamma) = \tan \beta + \tan \gamma$, and

$$\tan \phi = \tan(\beta + \gamma) = \tan \beta + \tan \gamma \quad .$$

Thus, for this example,

$$\frac{h}{x} = \frac{h}{1-x} + \frac{h}{4-x} \quad .$$

Solving for x we get

$$x \simeq 0.5 \quad .$$

Thus, the poles at $s = -1$ and $s = 0$ move toward each other, meet at the point $s = -0.5$ and split perpendicular to the real axis. This is called the break-point. The fact that the roots split perpendicularly to the real axis can be seen as follows: Rewrite the given equation so that

$$s(s+1)(s+4) + k_1 + k_2 = 0$$

where k_1 is such that the roots of the equation

$$s(s+1)(s+4) + k_1 = 0$$

are: a double root at the break-point, one root on the real axis (smaller than -4). Thus, the given equation can be written as

$$(s+0.5)^2(s+q) + k_2 = 0$$

where $q < -4$. Then considering the root-locus of this equation it is

immediately seen that the escape angle from $s = -0.5$ for $k_2 \sim 0$, is $90^\circ, 270^\circ$.

We also know that eventually the values of s must approach the asymptotes, but what happens between the real axis and the asymptotes? This can be determined by picking values of s , applying the phase criteria and thus obtaining values of s between the two extremes. It may also be done with a protractor if it is not convenient analytically. At any rate, the locus of the values of s will be a half-hyperbola approaching the values of large s asymptotically. The complete root-locus for $1 + \frac{k}{s(s+1)(s+4)} = 0$ is shown in Figure 46.

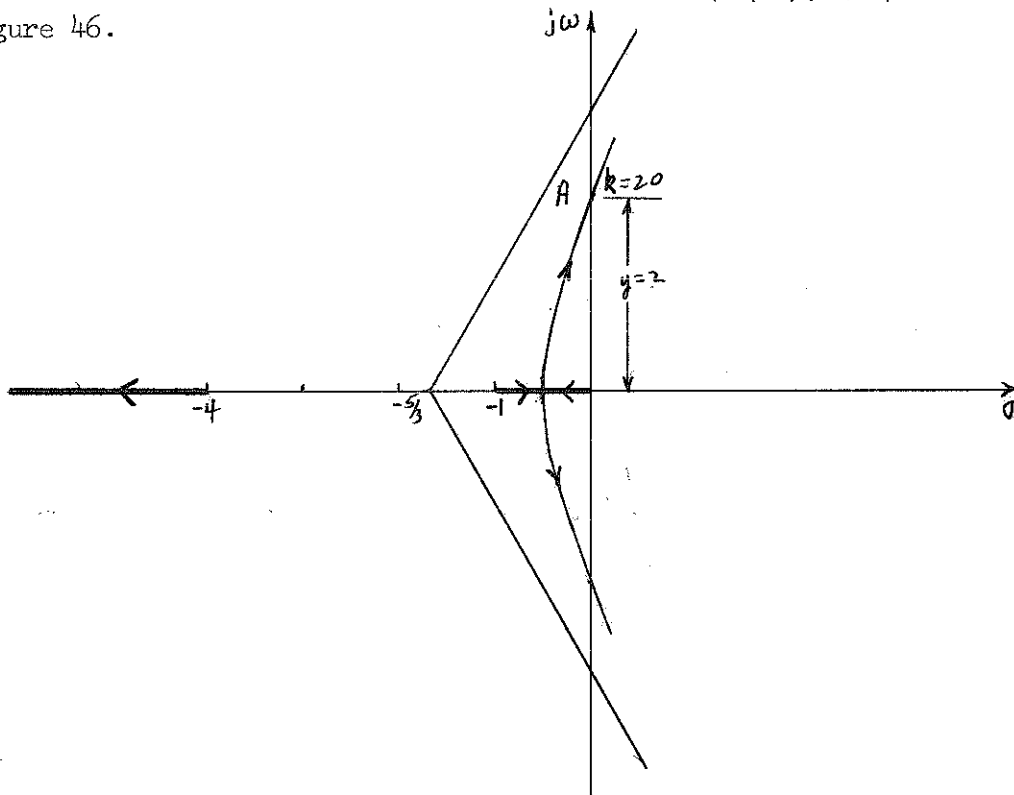


Fig. 46

(Note: See Insert A, Page 110a)

Now we can generalize the root-locus method to apply to any linear system in general. As has been previously pointed out, we can write the loop gain as $kG_1G_2 = k \frac{(s-z_1)(s-z_2)(s-z_3) \cdots (s-z_m)}{(s-p_1)(s-p_2)(s-p_3) \cdots (s-p_n)}$, and the problem is to determine how the zeros of the function one plus the loop gain vary as k varies. The steps to follow are:

I. For $k > 0$

A. Make a pole-zero plot of the loop gain.

Insert A

An important point which can be easily established is the point (A) at which the locus crosses the imaginary axis. This is called the cross-over point. From the angle condition we have

$$\alpha + \beta + \gamma = 180^\circ = 90^\circ + \beta + \gamma$$
$$\beta + \gamma = 90^\circ$$

Therefore (Figure 46)

$$\frac{\bar{x}}{1} = \frac{4}{x} ; x = 2$$

This point corresponds to a value of k given by

$$k = 2(\sqrt{4+1})(\sqrt{16+4}) = 20.$$

We can find other values of k for different points of the locus by a procedure similar to the one used in the previous example. In addition, it is evident that if the previous equation were the denominator of a transfer function, the system would become unstable for $k > 20$, because then the roots move in the right-half-plane.

B. Consider extreme values of k ;

1. For $k = 0$, the p_i 's are the solutions of $1 + kG_1G_2 = 0$

2. For $k = \infty$, the z_i 's are the solutions of $1 + kG_1G_2 = 0$

Usually there are more poles than zeros, thus there are $n-m$ values of s at infinity for $k = \infty$.

3. For values of k between these extremes, the solutions move continuously from the poles to the zeros of the loop gain.

C. Determine the slopes of the asymptotes. To this effect, note that:

For $s \rightarrow \infty$, the equation reduces to 1

$$1 + k \frac{1}{s^{n-m}} = 0;$$

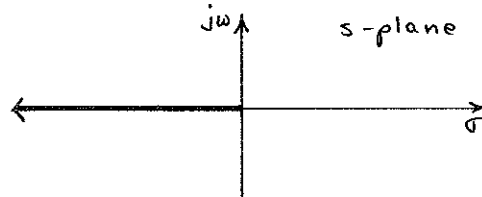
therefore

$$s^{n-m} = -k,$$

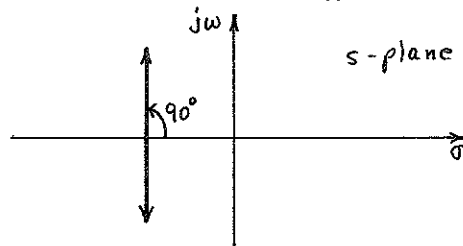
and

$$\angle s = \frac{180 + 2n\pi}{n-m}.$$

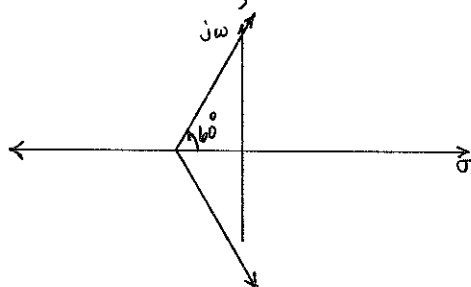
If $n-m = 1$ there is only one direction (180°) by which s can approach infinity.



If $n-m = 2$, the possible directions are two (90° and 270°) as shown.



If $n-m = 3$, there are three directions,



and so on.

D. Determine the intersection of the asymptotes. The equation can be written as:

$$1 + k \frac{s^m + a_1 s^{m-1} + \dots}{s^n + b_1 s^{n-1} + \dots} = 1 + \frac{k}{s^{n-m} + (b_1 - a_1) s^{n-m-1} + \dots}$$

$$\approx 1 + \frac{k}{s^{n-m} + (b_1 - a_1) s^{n-m-1}} ;$$

where $a_1 = -\sum$ zeros and $b_1 = -\sum$ poles. For s large, we may retain only the first two terms in the denominator and have that

$$s^{n-m} + (b_1 - a_1) s^{n-m-1} = -k. \quad (146)$$

Thus, the sum of the large solutions is $a_1 - b_1 = \sum$ poles - \sum zeros.

Then the asymptotes will meet at the point defined by

$$s = \frac{-(b_1 - a_1)}{n-m} = \frac{\sum \text{poles} - \sum \text{zeros}}{n-m}. \quad (147)$$

II. For $k < 0$, the statements are similar but the directions of the asymptotes are given by

$$\angle s = \frac{2\pi j}{n-m} \quad , \quad (148)$$

where j is an integer.

Now let's examine a couple more examples.

Assume we have

$$1 + k \left(\frac{s+2}{s-1} \right)^3 = 0 ,$$

as the denominator of a transfer function and wish to find the values of k for which the system is stable. The pole-zero plot of the loop gain, Figure 47, shows there is a triple pole at $s = 1 + j0$ corresponding to $k = 0$ and a triple zero at $s = -2$ corresponding to $k = \infty$. There are no solutions at infinity; in other words, the root-locus has no asymptotes.

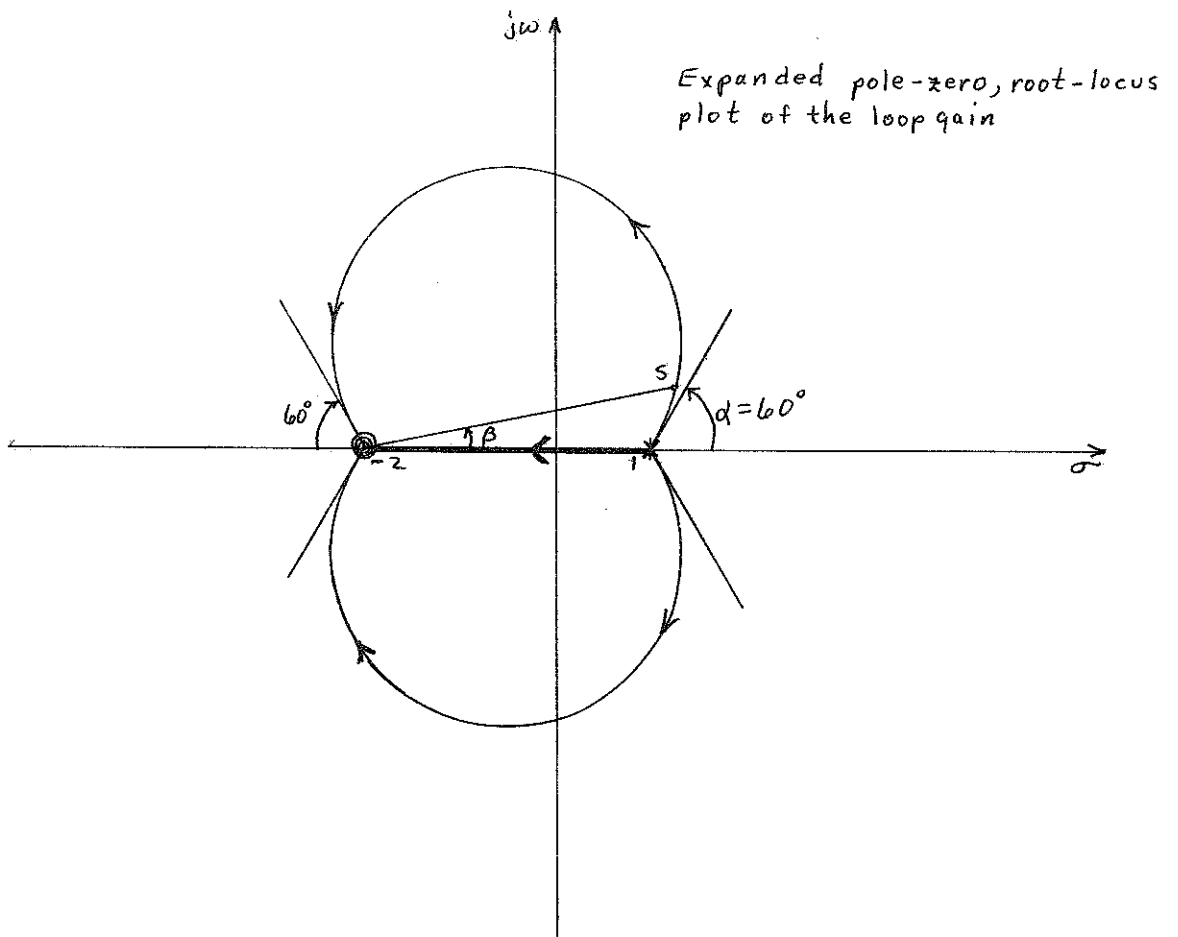


Fig. 47

For $k > 0$ let us define the escape angles from the triple root at $s = 1$ as k increases from zero. We consider the phase requirement for a point of the locus which is very close to $s = 1$. The phase angle of the vector $(s+2)$ is β (Figure 47) and is extremely small so that it can be assumed equal to zero. On the other hand, let us call the angle of the vector $(s - 1)$ α . Thus

$$3\alpha = 180 + 2n$$

$$\alpha = 60^\circ, 180^\circ, 300^\circ.$$

Consequently, one of the poles moves toward one of the zeros along the real axis and the other two escape from the real axis at angles of 60° and 300° .

The actual shape of the root-locus is the straight line between $s = 1$ and $s = -2$ and two circular arcs as shown in Figure 47. The reason for the circular arcs is the following. Consider an arbitrary point s (Figure 48). The phase requirement is

$$3\beta - 3\alpha = -180^\circ$$

$$\alpha - \beta = 60^\circ$$

But $\alpha - \beta = \gamma$. Therefore, all points of the locus see the segment between $s = 1$ and $s = -2$ with a constant angle of 60° which is possible only if the locus is a circular arc.

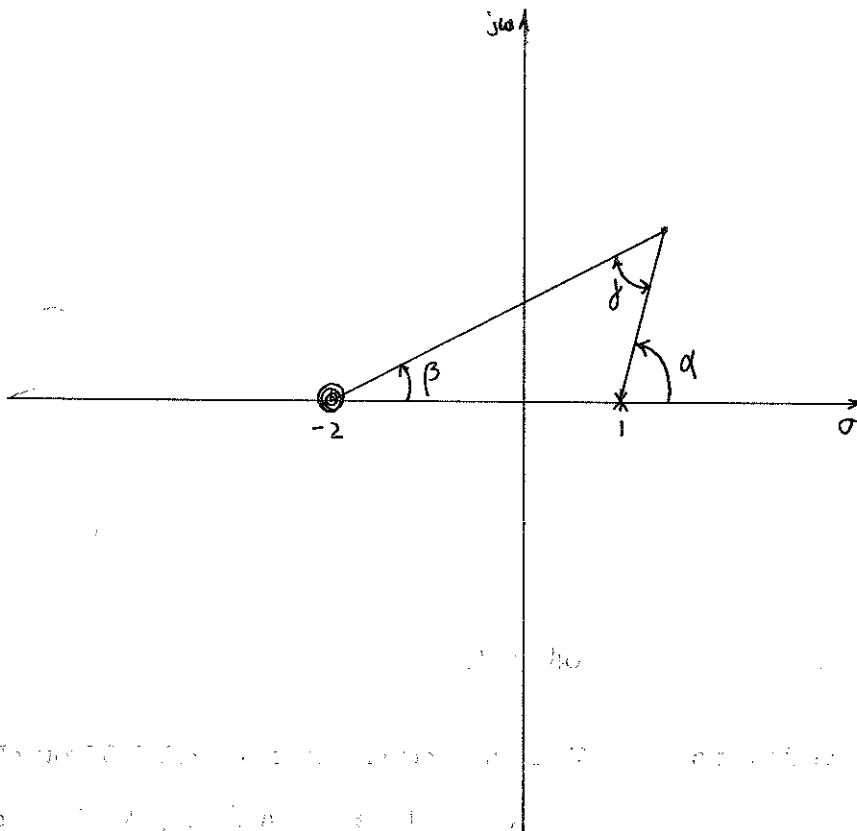


Fig. 48

We would like to know the points on the $j\omega$ -axis where the locus crosses. In Figure 48a the angle γ is

$$\gamma = 60^\circ .$$

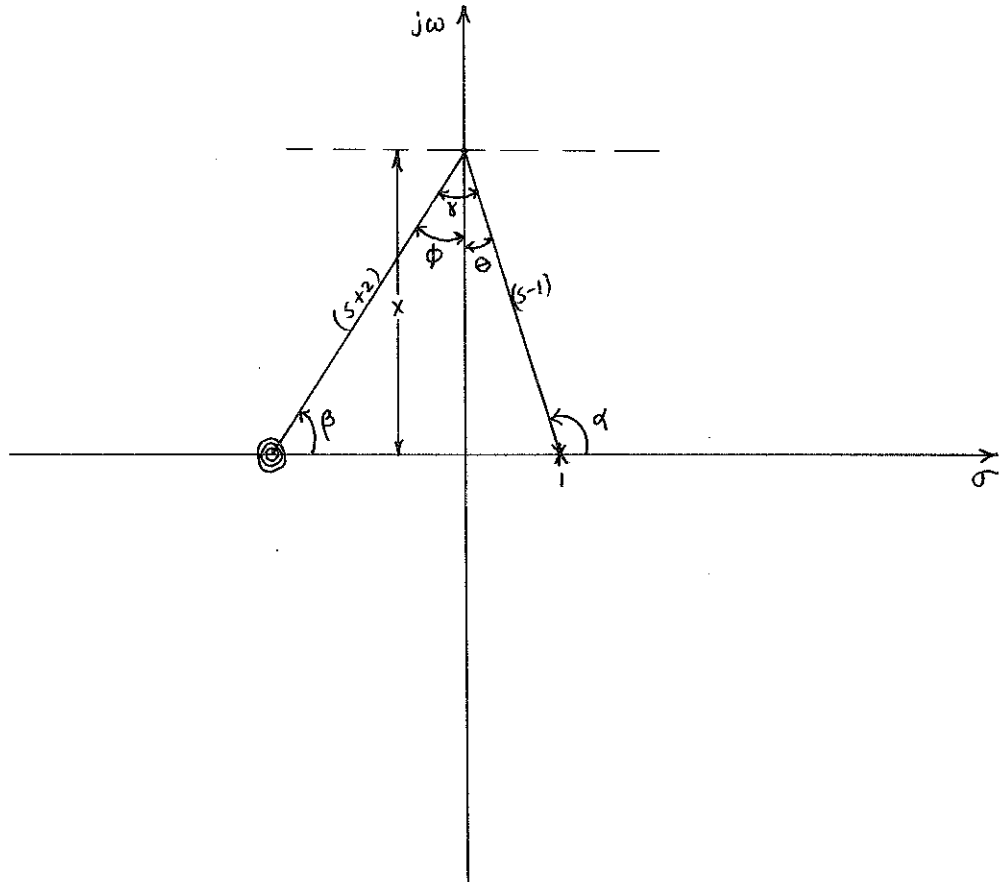


Fig. 48a

Knowing that $\gamma = 60^\circ$, we can determine the value of ω for $s = j\omega$ where the locus crosses the $j\omega$ -axis in the following manner:

From Figure 48a we have

$$\gamma = \phi + \theta = 60^\circ ,$$

and

$$\tan \theta = \frac{1}{x}; \quad \tan \phi = \frac{2}{x} ,$$

$$\tan 60^\circ = \frac{\tan \theta + \tan \phi}{1 - \tan \theta \tan \phi} = \frac{\frac{1}{x} + \frac{2}{x}}{1 - \frac{2}{x^2}} = \frac{3x}{x^2 - 2} = \sqrt{3} .$$

Thus we have

$$x^2 - \frac{3x}{\sqrt{3}} - 2 = 0$$

and

$$x = 2.53 \quad .$$

Therefore the locus crosses the $j\omega$ -axis at $s = +j2.53$ and $-j2.53$, and also at $s = 0$.

Since we know the values of s at which the poles cross the $j\omega$ -axis, we can determine the values of k for which the system is stable from the relationship

$$\left| \left(\frac{s-1}{s+2} \right)^3 \right| = \left| -k \right| .$$

For the cross-over point at the origin

$$\left| -k \right| = \left(\frac{1}{2} \right)^3 = \frac{1}{8} .$$

For the cross-over point at $s = \pm j\omega = +j2.53$,

$$\left| -k \right| = \left| \left(\frac{-1+j2.53}{2+j2.53} \right)^3 \right| = 0.6 \quad .$$

Thus for $k > 0.6$ the system is stable. In a similar manner it can be shown that for $k < 0$ the system is always unstable.

We will consider one last example which will illustrate better the applicability of the root-locus method in determining stability. We consider a system with a loop gain given as

$$\frac{k}{s} \left[\frac{r_1 a_1}{s+g_1} + \frac{r_2 a_2}{s+g_2} \right]$$

so that the denominator of the overall transfer function is

$$1 + \frac{k}{s} \left[\frac{r_1 a_1}{s+g_1} + \frac{r_2 a_2}{s+g_2} \right] = 1 + k(r_1 a_1 + r_2 a_2) \frac{1}{s} \frac{s + \frac{r_1 a_1 g_2 + r_2 a_2 g_1}{r_1 a_1 + r_2 a_2}}{(s+g_1)(s+g_2)} ,$$

where g_1 , g_2 , a_1 , and a_2 are positive constants and r_1 and r_2 are variables which can be either positive or negative. We want to find the possible values of k for which the system is stable.

Since r_1 and r_2 can be positive or negative, there are four possible locations of the zero of the loop gain which we must consider. The values of g_1 and g_2 are positive, therefore there are three poles, $s = 0$, $s = -g_1$, and $s = -g_2$. Figure 49 is a pole-zero plot showing the location of the poles and the relative possible locations of the zero of the loop gain.

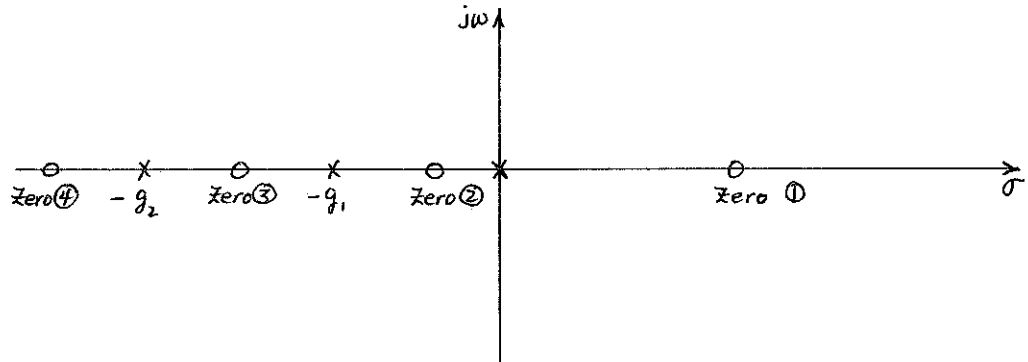


Fig. 49

For convenience let's define

$$\begin{aligned} r_1 a_1 + r_2 a_2 &= R_0 \\ r_1 a_1 g_1 + r_2 a_2 g_2 &= R_1 \\ r_1 a_1 g_2 + r_2 a_2 g_1 &= R_2 \end{aligned}$$

because, as it will be shown shortly, the different possibilities for the location of the zero of the loop gain depend on the sign of these quantities. Indeed, consider the following:

I. $R_2 < 0$. Then if $R_0 > 0$, the zero $(-R_2/R_0)$ is in the right-half-plane (position ①). The constant multiplier of the loop gain $kR_0 > 0$ (for $k > 0$).

Since there are three poles and one zero, there are two asymptotes for the locus of the poles of the transfer function and they are perpendicular to the real axis. Also, from $\frac{\sum \text{poles} - \sum \text{zeros}}{n-m}$ we see that the intersection

with the real axis must be to the left of the poles at the point

$$s = \frac{\sum \text{poles} - \sum \text{zeros}}{n-m} = \frac{-(g_1+g_2) - \left(-\frac{R_2}{R_0}\right)}{2}$$

which is a negative number with magnitude greater than the magnitude of any of the poles (Figure 50).

To determine the possible locus of the poles on the real axis, the phase condition is applied. To this effect, for the region between the origin and the zero we have $\arg(\text{Zero } \textcircled{1}) - \arg(s+g_2) - \arg s = 180^\circ - 0 - 0 - 0 = 180^\circ$. Therefore this is part of the locus.

For the interval between $s = 0$ and $s = -g_1$

$$\arg(\text{Zero } \textcircled{1}) - \arg s - \arg(s+g_1) - \arg(s+g_2) = 180 - 180 - 0 - 0 = 0,$$

and this interval is not a possible part of the locus. Finally, by the same procedure we determine that the region $-g_2 < s < -g_1$ is part of the locus while the region $s < -g_2$ is not. As k increases from zero the roots $(-g_1)$ and $(-g_2)$ move toward each other, meet and then split continuing their motion toward the asymptotes. The root-locus for this case is as shown in Figure 50.

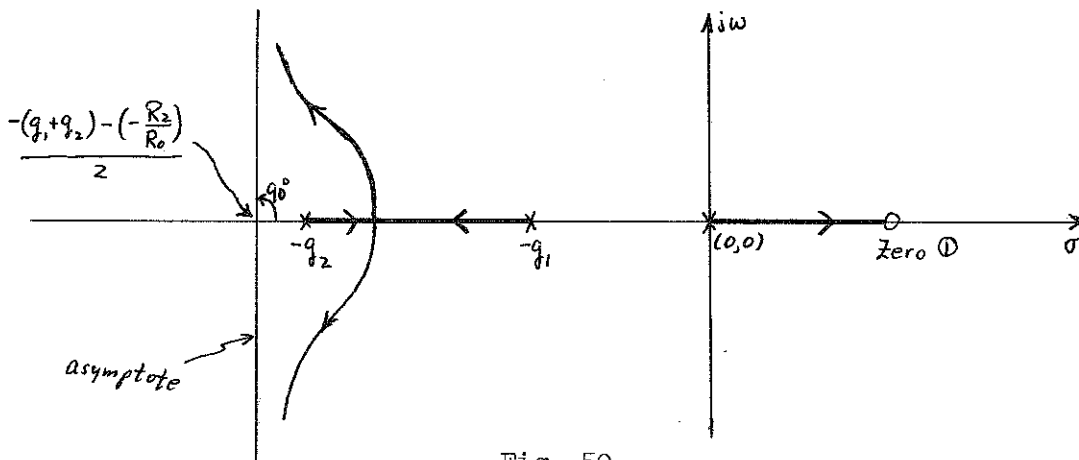


Fig. 50

If so desired, the point on the real axis at which the two roots meet and split can be found from a relationship similar to Equation (144). At any rate, for $R_2 < 0$, $R_0 > 0$, and $k > 0$ the system is always unstable because

the transfer function has always a pole in the right-half-plane, regardless of the magnitude of k .

II. Another case for which the zero can be in the right-half-plane is when $R_2 > 0$ and $R_0 < 0$. Then $kR_0 < 0$ for $k > 0$ and the phase condition now is that

$$\angle G_1(s)G_2(s) = 2\pi\ell \quad , \quad (149)$$

and for the asymptotes

$$\angle s = \frac{2\pi\ell}{n-m} \quad , \quad (150)$$

where ℓ is an integer. The asymptotes are at 0° and 180° .

To determine the possible locus of the poles on the real axis we now apply the new phase condition. For the region to the right of the zero

$$\arg(\text{Zero } \textcircled{1}) - \arg s - \arg(s+g_1) - \arg(s+g_2) = 0 - 0 - 0 - 0 = 0$$

Thus $\ell = 0$ in Equation (149) and this region is a possible locus of the roots.

For the interval between $s = 0$ and the Zero $\textcircled{1}$:

$$\arg(\text{Zero } \textcircled{1}) - \arg s - \arg(s+g_1) - \arg(s+g_2) = 180 - 0 - 0 - 0 = 180$$

which cannot be satisfied for any interger value of ℓ in Equation (149). Thus the interval is not a possible locus of the roots.

For the interval between $s = -g_1$, and $s = 0$,

$$\arg(\text{Zero } \textcircled{1}) - \arg s - \arg(s+g_1) - \arg(s+g_2) = 180 - 180 - 0 - 0 = 0,$$

and this interval is a possible locus of roots. In a similar manner it can be shown that the only other region of possible locus is for $s < -g_2$.

Following the procedure through as for the first case, we would find that the root-locus would finally look as shown in Figure 51. For this case then the system is stable until the value of $k = k_0$ where the locus intersects the $j\omega$ -axis. For values of $k > k_0$ the system becomes unstable.

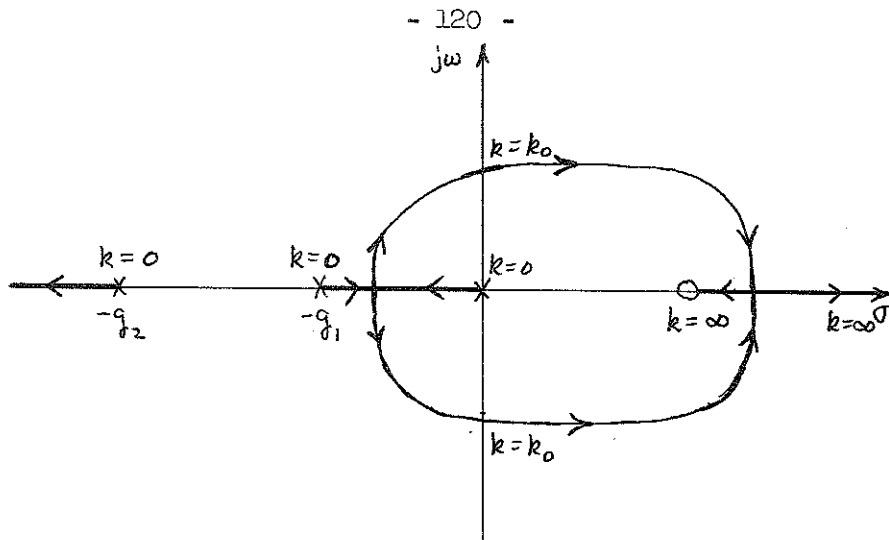


Fig. 51

III. For the conditions $R_2 > 0$ and $R_0 > 0$ the zero is in the left-half-plane in one of the three possible positions.

A. Let's assume first that the zero is in the interval from $s = -g_1$ to $s = 0$ as shown in Figure 52.

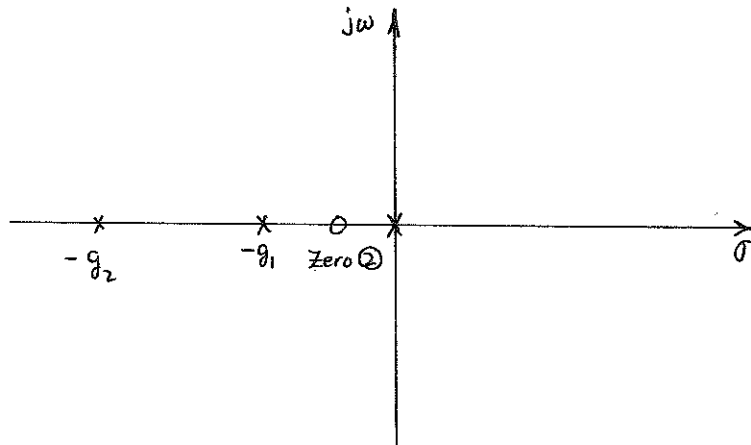


Fig. 52

For this case $k(r_1 a_1 + r_2 a_2) > 0$ and the phase condition

$$\angle G_1(s)G_2(s) = 180^\circ + 2\pi n$$

applies. Going through the standard procedure then it can be shown that the root-locus would be that of Figure 53 and the system is always stable.

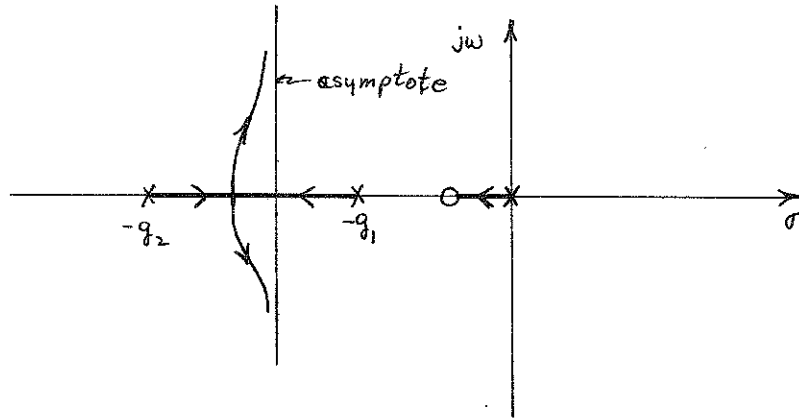


Fig. 53

B. Assume the zero to be in the interval from $s = -g_2$ to $s = -g_1$. Then Figure 54 shows the root-locus for this case and the system is again always stable ($kR_0 > 0$).

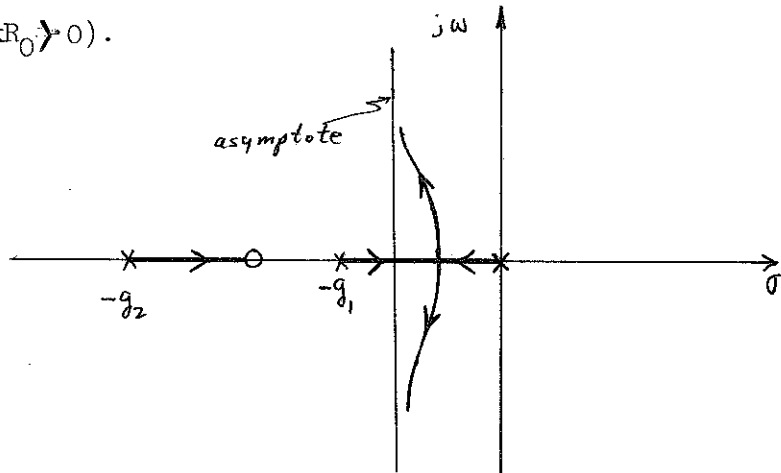


Fig. 54

C. Assume the zero is to the left of all the poles, but to the right of the point $s = -(g_1 + g_2)$. Then the root-locus is as shown in Figure 55 and the system is again always stable.

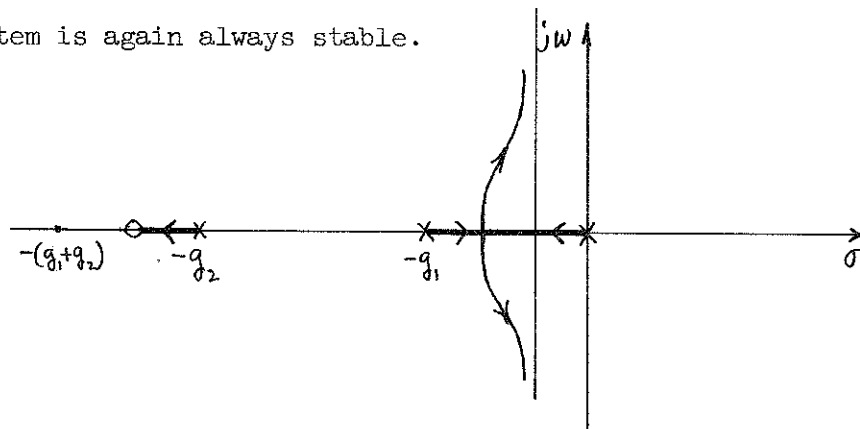


Fig. 55

Notice that as the location of the zero moves to the left the asymptotes move to the right and approach the $j\omega$ -axis. The significance of the point $-(g_1 + g_2)$ is that when the zero is at this point the asymptotes coincide with the $j\omega$ -axis.

D. One last possibility for the conditions of $R_2 > 0$ and $R_0 > 0$ is that the zero lies to the left of the point $-(g_1 + g_2)$. If this occurs, the asymptotes will be to the right of the $j\omega$ -axis and, for certain values of $k(r_1 a_1 + r_2 a_2)$, the system will become unstable. For this case the root-locus is as shown in Figure 56.

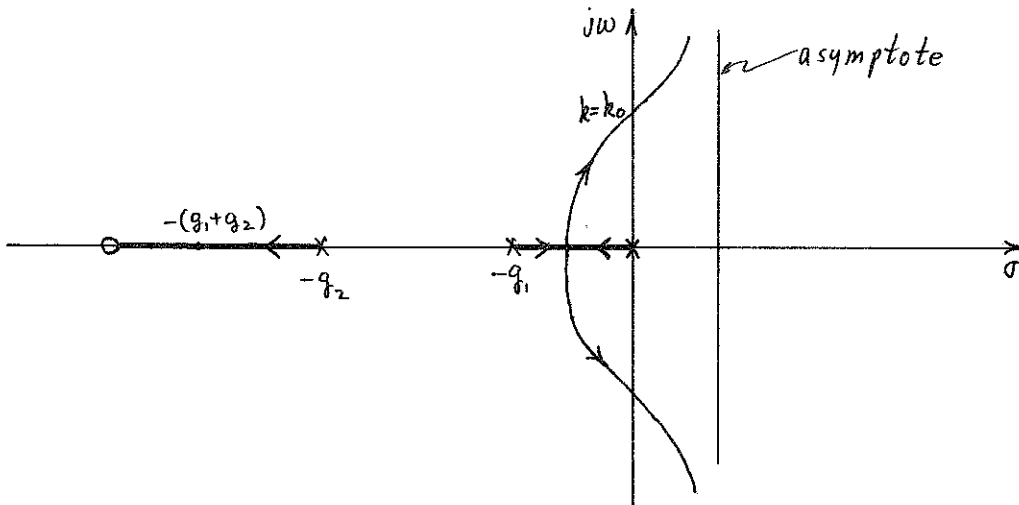


Fig. 56

Thus the system can be unstable when $\frac{R_2}{R_0} > (g_1 + g_2)$. This means that

$$r_1 a_1 g_2 + r_2 a_2 g_1 > r_1 a_1 g_1 + r_2 a_2 g_2 + r_1 a_1 g_2 + r_2 a_2 g_1$$

or

$$R_1 \equiv r_1 a_1 g_1 + r_2 a_2 g_2 < 0, \text{ and } k > k_0 .$$

IV. The last possible case to consider is when $R_2 < 0$ and $R_0 < 0$. Again the zero is in the left-half-plane in one of three possible locations, but $k(r_1 a_1 + r_2 a_2)$ is negative for $k > 0$. Thus we have to apply the phase conditions for negative k .

A. If the zero is in the interval between $s = 0$ and $s = -g_1$, the root-locus will be as in Figure 57.

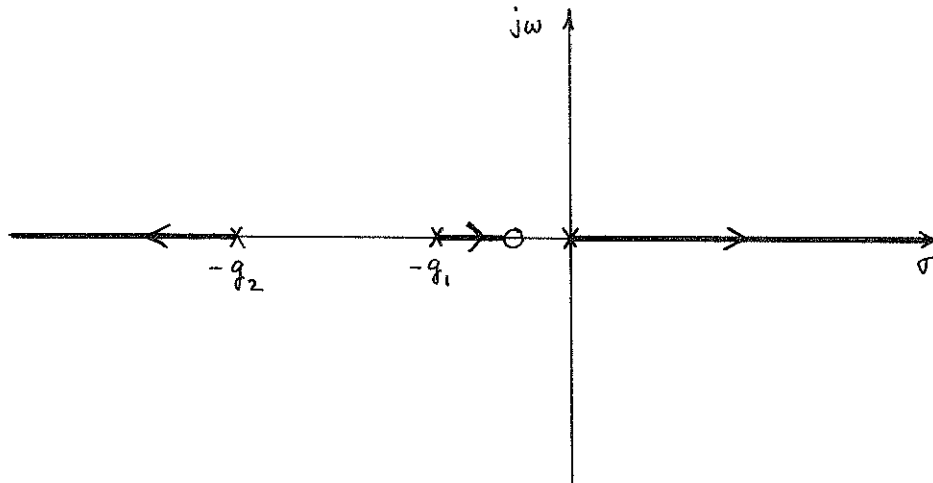


Fig. 57

The system is always unstable since the pole at $s = 0$ moves to the right half-plane.

B. If the zero is in the interval between $s = -g_1$ and $s = -g_2$ the root-locus is as shown in Figure 58 and the system is again always unstable.

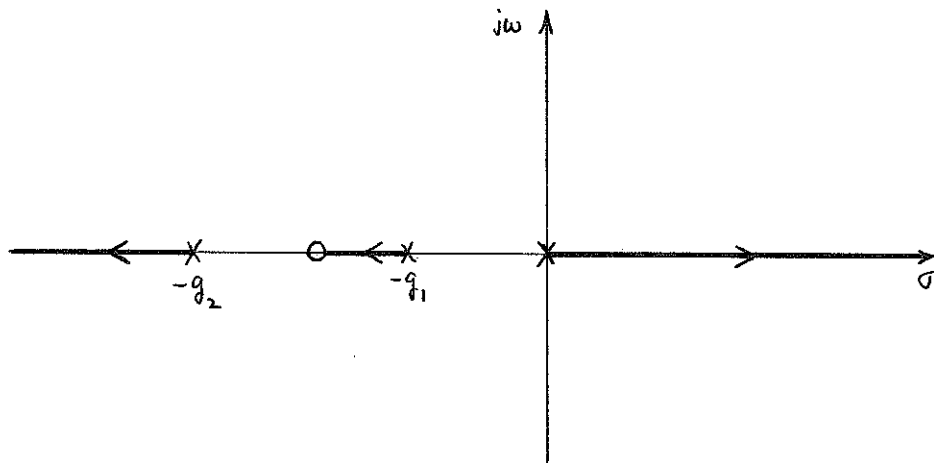


Fig. 58

C. For the position of the zero to the left of the poles, the root-locus is as shown in Figure 59 and again the system is always unstable.

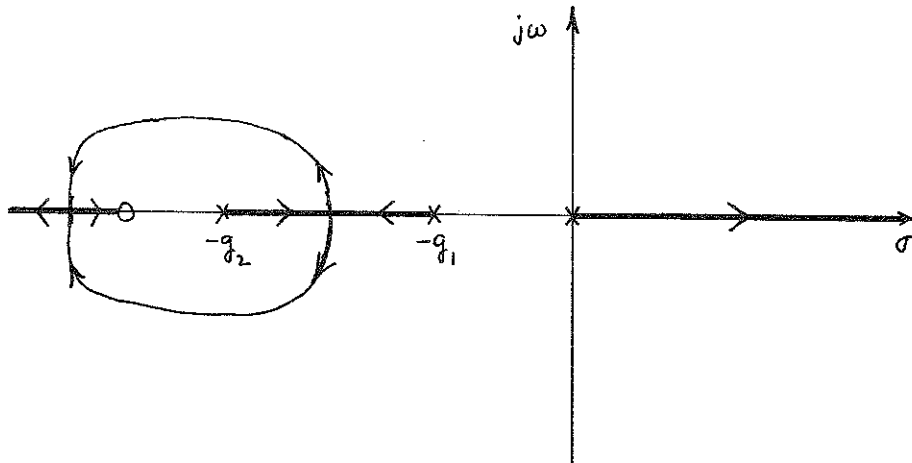


Fig. 59

Thus, for the conditions $R_2 < 0$ and $R_0 < 0$, the system is always unstable.

The four cases just investigated exhaust the possibilities for all variations of R_0 , R_1 , and R_2 . Since r_1 and r_2 are the only variables, we can consider looking at the possibility of stability or instability by identifying these regions on a plane of r_1 and r_2 .

I. Assume $g_1 > g_2 > 0$.

Consider the (r_1, r_2) -plane diagram of Figure 60. In this diagram the lines in the fourth quadrant emanating from the origin represent the values of r_1 and r_2 for which the R_i 's are zero. The regions above and below a particular R_i -line indicate whether R_i is positive or negative for different combinations of values of r_1 and r_2 in that region.

A. Now assume that $r_1, r_2 > 0$. Then $R_0, R_1, R_2 > 0$, and

$$(g_1 + g_2) > \frac{R_2}{R_0} \quad ,$$

(this inequality is equivalent to $R_1 > 0$). According to the preceding discussion, the system is always stable. That is, the system is stable in the first quadrant of the (r_1, r_2) -plane.

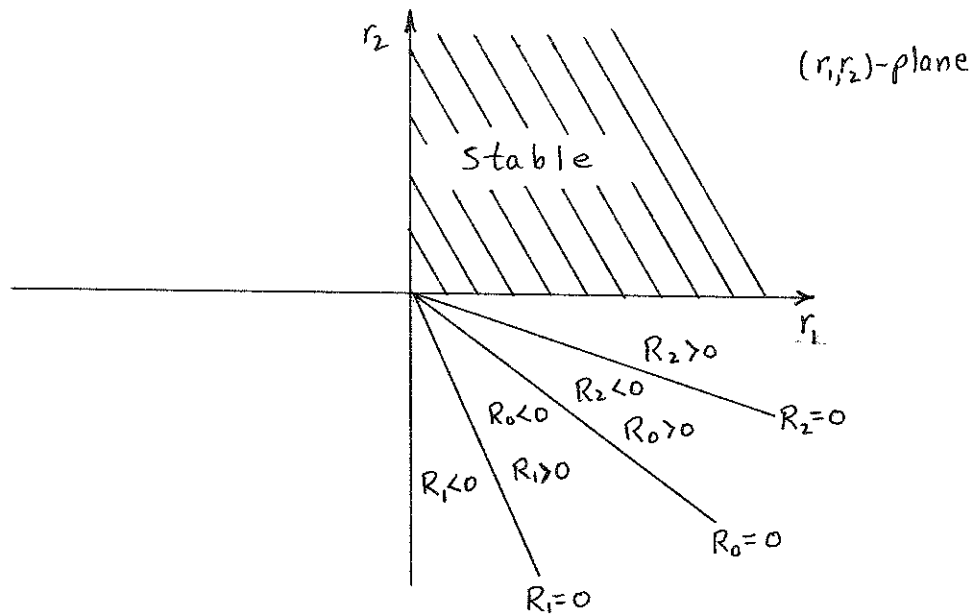


Fig. 60

B. Assume $r_1 \rightarrow 0$ and $r_2 \leftarrow 0$.

1. If $R_2 \rightarrow 0$; $R_0 \rightarrow 0$, and $R_1 \rightarrow 0$, thus the condition $\frac{R_2}{R_0} \ll (g_1 + g_2)$ is satisfied and the system is stable (everywhere above the line $R_2 = 0$; Figure 60).

2. If $R_0 \rightarrow 0$ but $R_2 \ll 0$ then the system is unstable as described for Case I above.

3. If $R_1 \rightarrow 0$ and $R_0 \ll 0$, this corresponds to Case IV above and the system is always unstable ($R_2 \ll 0$).

4. If R_0, R_1 , and $R_2 \ll 0$ the system is also always unstable (case IV).

From these considerations then we can label the regions as shown in Figure 61.

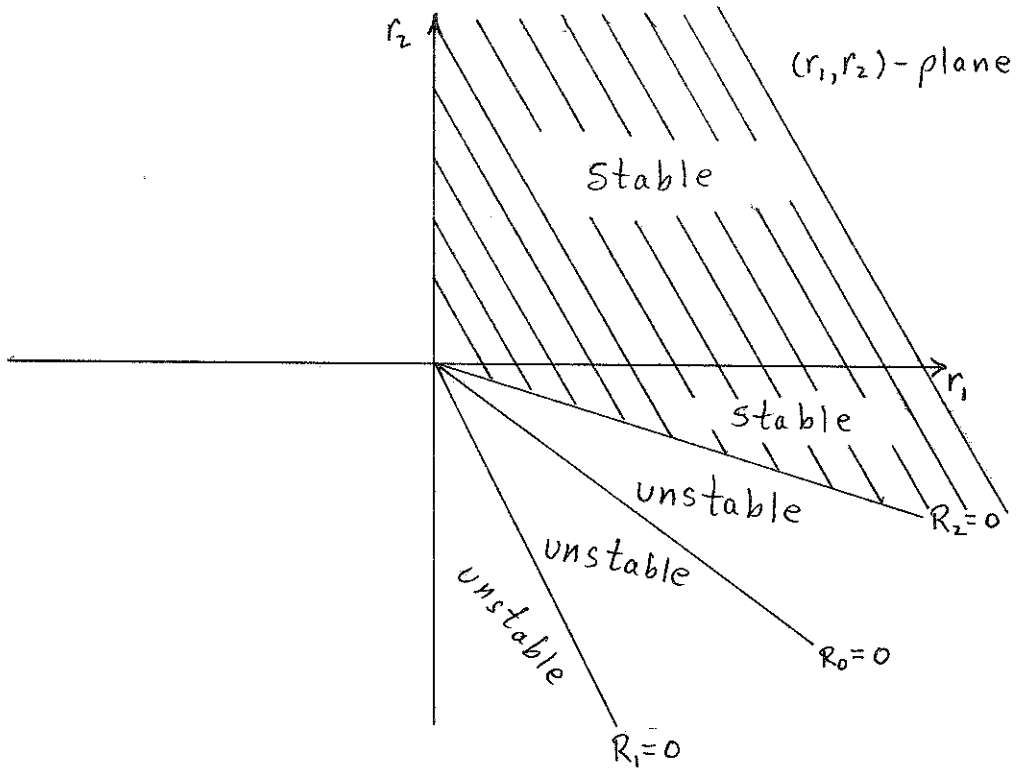


Fig. 61

II. Assume $0 < g_1 < g_2$. Following the same procedure we would obtain the diagram shown in Figure 62.

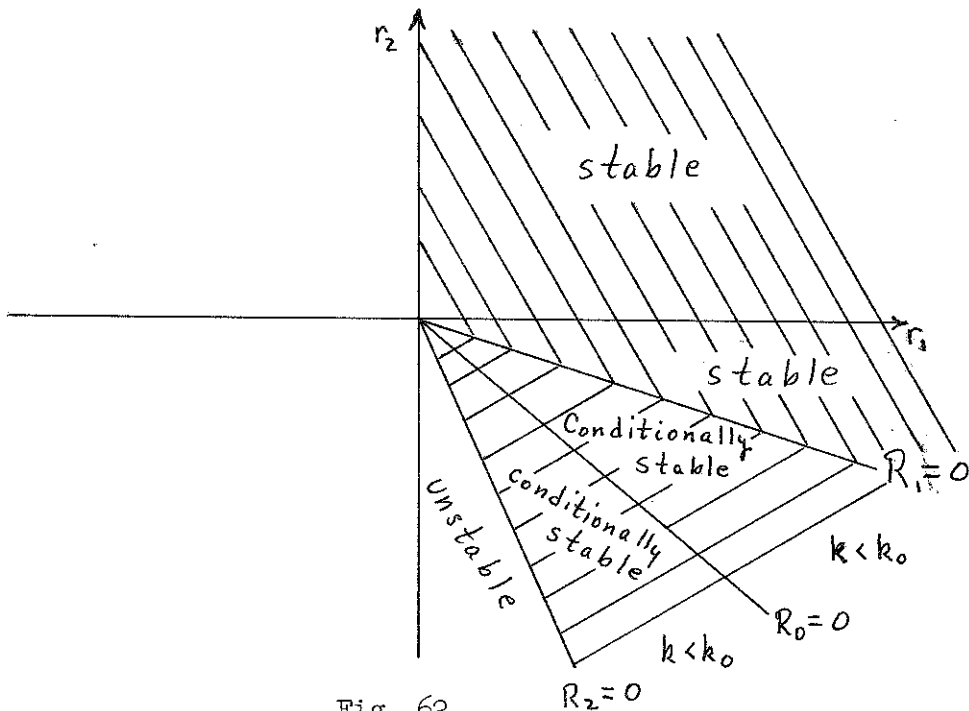


Fig. 62

Thus we have considered all possible variations. The value of k beyond which the conditionally stable system becomes unstable can be easily determined.

It turns out that in all cases

$$k_0 = - \frac{\varepsilon_1 \varepsilon_2 (\varepsilon_1 + \varepsilon_2)}{R_1}$$

(Note that $R_1 < 0$ in these cases.)

The preceding example is pertinent to a nuclear reactor with two temperature coefficients of reactivity r_1, r_2 .

Suggested References

1. H. Chestnut and R. W. Mayer, Servomechanisms and Regulating System Design, Volume 1. New York: John Wiley & Sons, 1953.
2. John G. Truxal, Automatic Feedback Control System Synthesis. New York: McGraw-Hill Book Company, 1955.

LECTURE NO. V

STATISTICS

In the previous lectures we have considered systems with deterministic inputs and outputs. The value of the input was always known; i.e., the probability of knowing the value of the input at any given time was always 100 percent. For most practical systems however, this is not the case. Rather than being an exact known function of time the input may be random in nature. In the case of a nuclear reactor the fission process itself is a random variable. Other examples of processes involving random variables would be boiling; in communication systems wave propagation through the atmosphere is random because of interference by cloud formations and other obstacles.

One way to handle random processes is to sample a large number of identical processes at the same time. In this way one could determine the probability that the variable will have a given value at that time. This method of analysis results in what might be called "ensemble statistics". The word "ensemble" is associated with the identity of the processes, and the word "statistics" means that the number of identical processes considered is large enough such that considering more identical processes would not change the values of the results. Analytically we determine ensemble statistics in the following manner:

Consider Figure 63 which is an ensemble of the outputs of three identical systems for the same input. Since the input is random the outputs will also be random. At the present there is no correlation between the outputs of the three systems; i.e., there is no way of determining the value of the output for one system, knowing the value of the output for another system. However, assume that at time t_1 we arbitrarily choose a value from $f(t) = f_1$ to $f(t) = f_1 + \Delta f_1$, as shown on Figure 63. Then the number of ensemble members that

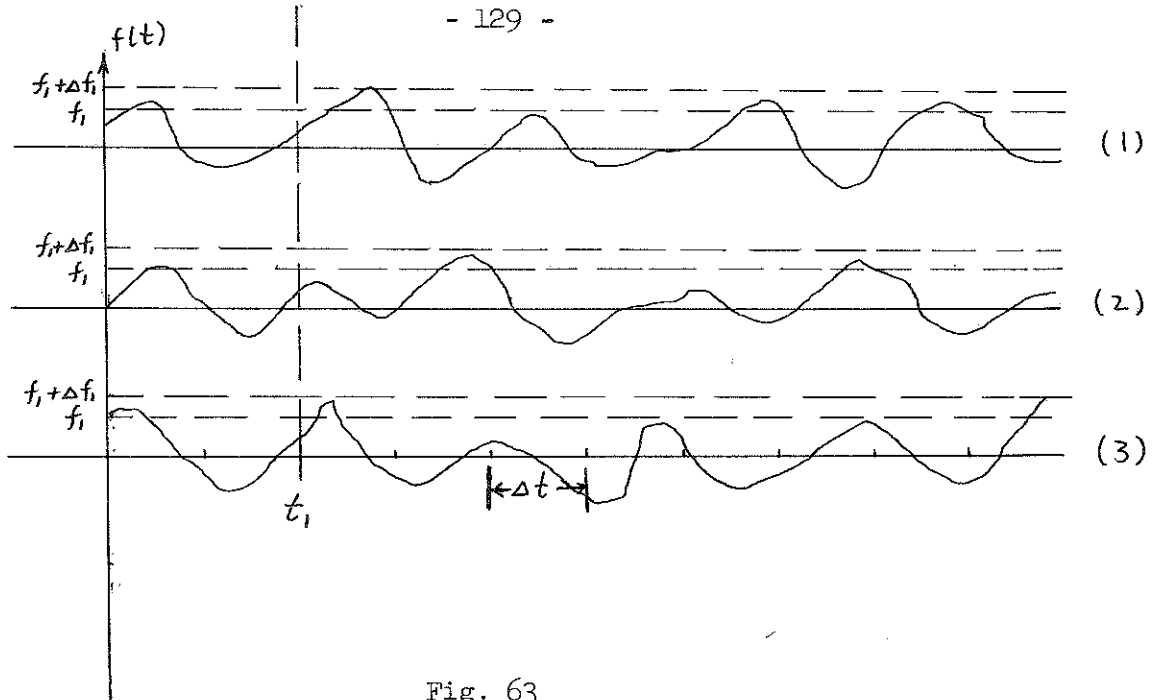


Fig. 63

have values between f_1 and $f_1 + \Delta f_1$ at time t_1 is given by $\Delta N_1(f_1, \Delta f_1, t_1, N)$ where N is the total number of the members in the ensemble.

If N is a large number then we can define the probability of the output having a value of f_1 at time t_1 by

$$P_1(f_1, t_1) = \frac{\Delta N_1}{N \Delta f_1} = \text{first probability density} \quad (151)$$

as $N \rightarrow \infty$ and $\Delta f_1 \rightarrow 0$.

Now let's assume that we perform another sampling at two values of time, t_1 and t_2 and wish to determine the probability of a member of the ensemble having a value between f_1 and $f_1 + \Delta f_1$ at time t_1 and also having a value between f_2 and $f_2 + \Delta f_2$ at time t_2 . This is the second probability density and, in a manner analogous to defining the first probability density, is given by

$$P_2(f_1, f_2, t_1, t_2) = \frac{\Delta N_2(f_1, \Delta f_1, t_1, f_2, \Delta f_2, t_2, N)}{N \Delta f_1 \Delta f_2} \quad (152)$$

as $N \rightarrow \infty, \Delta f_1, \Delta f_2 \rightarrow 0$.

We could continue this process of defining higher-order probability distributions and would finally be able to assign the probability that the function would have a given value of any time t . All probability density functions must exist if the ensemble of functions is to constitute a random process.

The first and second probability density functions are related by the equation

$$P_1(f_1, t_1) = \int_{-\infty}^{\infty} P_2(f_1, t_1, f_2, t_2) df_2 . \quad (153)$$

So far, we have defined P_1 and P_2 as functions of time. However, if the statistics measured are the same at different measuring times, the statistics are independent of time and the processes are called "stationary". Thus, if in determining the probability density P_1 we get the same value regardless of whether we take the time to be t_1 , t_2 , or t_3 say, then the process is said to be stationary in time. We can redefine the above probability density for stationary processes to obtain

$$p_1(f_1) = \frac{\Delta N_1(f_1, \Delta f_1, N)}{N \Delta f_1} , \quad (154)$$

For the second probability density and stationary systems the time dependence is not on t_1 and t_2 separately, but rather on the difference of t_1 and t_2 .

Then

$$p_2(f_1, f_2, \tau) = \frac{\Delta N_2(f_1, f_2, \tau, N)}{N \Delta f_1 \Delta f_2} , \quad (155)$$

where $\tau = t_1 - t_2$. In this respect, the second probability density is a function only of the time increment τ , not of time itself.

Probability densities are useful in establishing average values for the ensemble. For example the ensemble average is

$$\bar{f} = \frac{f_{11} + f_{12} + f_{13} + \dots + f_{1n}}{N} = \int_{-\infty}^{\infty} f_1 p_1(f_1) df_1, \quad (156)$$

and the second average

$$\bar{f}_{11} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1 f_2 p_2(f_1, f_2, \tau) df_1 df_2, \quad (157)$$

where the \sim means average over all members of the ensemble.

Another method of dealing with random processes is to consider only one member of the ensemble over a long period of time. Then the probability that the member has a value between f_1 and $f_1 + \Delta f_1$ over a given time T , is

$$P_{1T}(f_1) = \frac{\Delta T_1(f_1, \Delta f_1, T)}{T \Delta f_1} \quad (158)$$

as $T \rightarrow \infty$, $\Delta f_1 \rightarrow 0$, where ΔT is the time over which the function has values between f_1 and $f_1 + \Delta f_1$.

The average value estimated from one member of the ensemble is:

$$\bar{f} = \frac{1}{2T} \int_{-T}^T f(t) dt \quad ; T \rightarrow \infty, \quad (159)$$

where the bar denotes the average of one member rather than an ensemble.

Most physical stationary processes provide identical results under these two methods of measurement. Such processes, in which the statistics of one system over a long period of time are the same as the statistics of an ensemble of systems at one instant of time, are defined to be "ergodic" processes. An ergodic process will always be stationary, but a stationary process is not necessarily ergodic. We will be concerned only with ergodic processes.

CORRELATION FUNCTIONS

To make use of the above principles we will now consider correlation functions. These functions are related to the second probability density as we will soon see.

Autocorrelation Function

The autocorrelation function is concerned with the joint statistics of two successive values of a given signal spaced τ seconds apart in time. For an ensemble, consider the products

$$\begin{aligned} f_1(t_1) f_1(t_1 + \tau) \\ f_2(t_1) f_2(t_1 + \tau) \\ f_n(t_1) f_n(t_1 + \tau) \end{aligned}$$

where $f_i(t_1)$ is the value of each member at the time t_1 and $f_i(t_1 + \tau)$ is the value at a time $t_1 + \tau$. This τ interval can be taken anywhere on the records since the processes are assumed stationary. The average over the entire ensemble is then the autocorrelation function and is given by

$$\bar{\phi}_{11}(\tau) = \iint f_1 f_2 p_2(f_1, f_2, \tau) df_1 df_2. \quad (160)$$

Equation (160) then is a measure of the dependence of the value of the function in the future on the value of the function at the present.

For a single member over a long period of time

$$\bar{\phi}_{11}(\tau) = \frac{1}{2T} \int_{-T}^T f_1(t) f_1(t + \tau) dt, \quad (161)$$

as $T \rightarrow \infty$, and for ergodic processes

$$\tilde{\phi}_{11}(\tau) = \bar{\phi}_{11}(\tau) .$$

Properties of Autocorrelation Functions

Certain properties of the autocorrelation function are readily deduced from the definition

$$\phi_{11}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_1(t) f_1(t+\tau) dt .$$

(1) The autocorrelation function is an even function of τ ; that is, $\phi_{11}(\tau) = \phi_{11}(-\tau)$. $\phi_{11}(\tau)$ is measured by shifting the function τ seconds ahead and averaging the product of the original and shifted functions; $\phi_{11}(-\tau)$ is measured by shifting the function backward by τ seconds and averaging in the same way. Since the functions are averaged over a doubly infinite interval, the time origin is inconsequential and the averaged product is independent of the direction of the shift.

(2) The autocorrelation function of zero argument, $\phi_{11}(0)$, is the average of the square of the time function, since $\phi_{11}(0)$ is given by

$$\phi_{11}(0) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_1(t) f_1(t) dt = \overline{f^2(t)} > 0 .$$

(3) The value of the autocorrelation function never exceeds the value for zero argument; that is

$$|\phi_{11}(\tau)| \leq \phi_{11}(0) .$$

Again, this is apparent from the definition of autocorrelation since the maximum value inevitably occurs when the function is multiplied by itself

without shifting. Mathematically, since $f_1(t) \neq f_1(t + \tau)$ in general, we can write

$$\overline{[f_1(t) \pm f_1(t + \tau)]^2} \geq 0.$$

Then,

$$\overline{f_1^2(t)} + \overline{f_1^2(t + \tau)} \pm 2\overline{f_1(t) f_1(t + \tau)} \geq 0.$$

But $\overline{f_1^2(t)} = \phi_{11}(0)$ and $\overline{f_1^2(t + \tau)} = \phi_{11}(0)$ since the time involved is inconsequential. Therefore,

$$2\phi_{11}(0) \pm 2\phi_{11}(\tau) \geq 0.$$

Thus

$$\phi_{11}(0) \geq |\phi_{11}(\tau)|.$$

(4) For large τ the value of the autocorrelation function becomes $(\overline{f_1(t)})^2$.

This is apparent from the fact that for large τ the probability functions $p(f_1)$ and $p(f_2)$ are independent and

$$\begin{aligned} \phi_{11}(\tau) &= \int p(f_1) f_1 df_1 \int p(f_2) f_2 df_2 \\ &= [\overline{f_1(t)}] [\overline{f_1(t)}]. \end{aligned}$$

(5) If the signal contains periodic components (or a d-c value), the autocorrelation function contains components of the same periods (or a d-c component).

This property can be proved as follows: Assume

$$f_1(t) = r(t) + \sin \omega t$$

where $r(t)$ is random.

Then:

$$\begin{aligned}
 \phi_{rr}(\tau) &= \overline{(r(t) + \sin \omega t)(r(t+\tau) + \sin \omega(t+\tau))} \\
 &= \phi_{rr}(\tau) + \underbrace{\overline{r(t) \sin \omega(t+\tau)}}_0 + \\
 &\quad + \underbrace{\overline{r(t+\tau) \sin \omega t}}_0 + \overline{\sin \omega t \sin \omega(t+\tau)} \\
 &= \phi_{rr}(\tau) + \frac{1}{2} \cos \omega \tau
 \end{aligned}$$

One significant fact concerning periodic components should be noted. Since the origin of time is irrelevant in autocorrelation functions, the latter do not contain any information about the phase of the periodic component.

We will now consider some examples of autocorrelation functions.

Example I

Consider the random wave form shown in Figure 64. For this function, an event is equally likely to be positive or negative, and on the average over

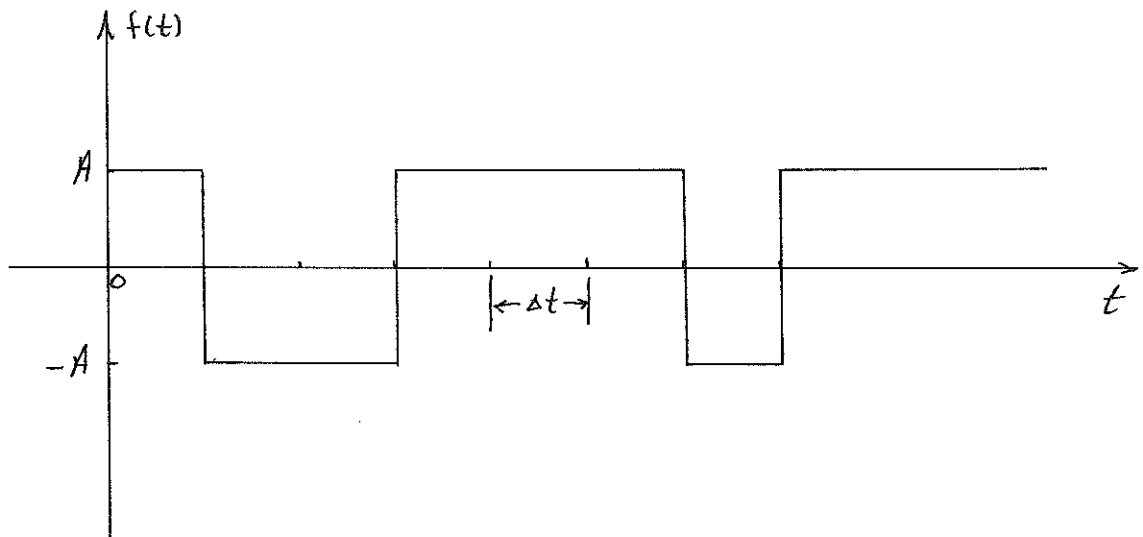


Fig. 64

a long period of time there are just as many positive as negative values. The function can, but does not necessarily, change sign, discontinuously, every Δt seconds.

Experimentally, the autocorrelation function can be determined by performing three operations:

(1) Replot and shift the waveform by an interval τ . This is shown in waveform (2) of Figure 65.

(2) Multiply the original, unshifted waveform by the shifted waveform. The product of the two is shown in waveform (3) of Figure 65.

(3) Integrate the area under the waveform (3) and divide by the time over which the integral is taken.

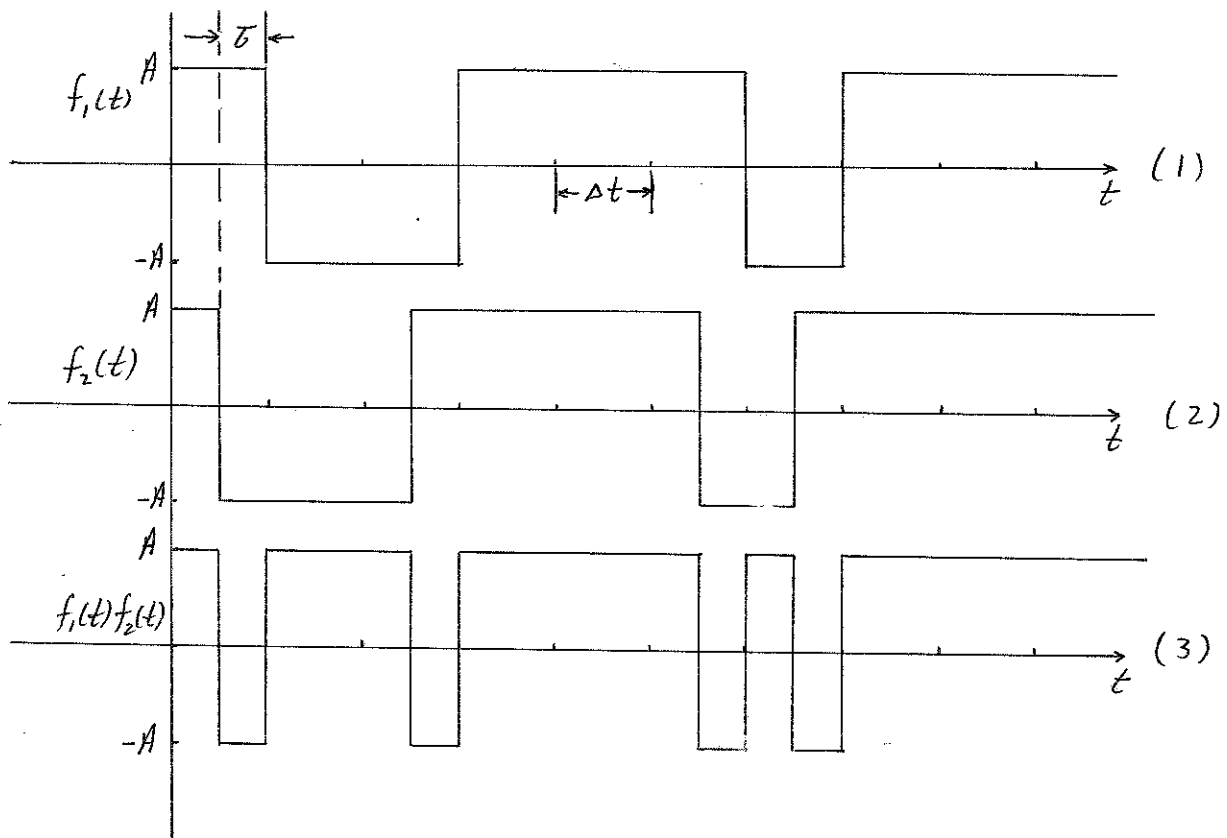


Fig. 65

If the shift $|\tau|$ is greater than the time interval Δt , then the probability that f_2 has the same sign as f_1 is one-half. This is independent of the sign of f_1 , thus there is no correlation between present and future values of $f(t)$. Since there is no correlation, the autocorrelation function is always zero for $|\tau| > \Delta t$.

If the shift $|\tau|$ is less than Δt , there is at most one positive or negative value in a Δt interval. The probability of getting a change in the sign of $f(t)$ in one τ interval is $\frac{\tau}{\Delta t}$. Then the autocorrelation function is given as

$$\phi_{11}(\tau) = \begin{cases} A^2(1 - \tau/\Delta t) & ; |\tau| < \Delta t \\ 0 & ; |\tau| > \Delta t \end{cases}$$

A plot of the autocorrelation function for Example 1 is shown in Figure 66. Thus, the values of $f_1(t)$ are correlated for $|\tau| < \Delta t$,

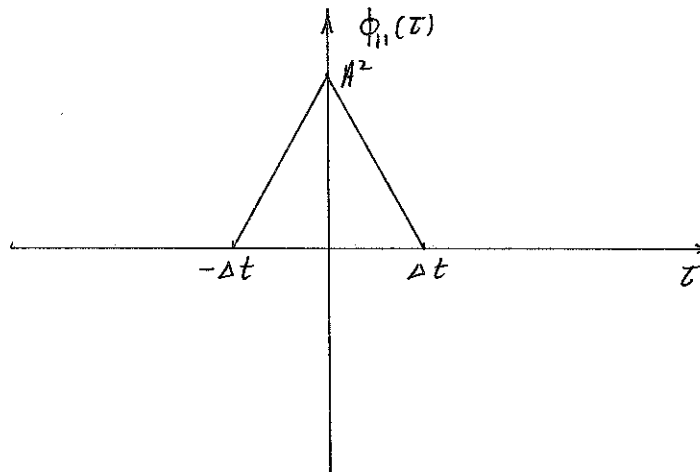


Fig. 66

but there is no correlation if $|\tau| > \Delta t$.

Analytically, if the distribution function of the magnitudes of $f(t)$ is known, the probability that $f(t)$ will have a certain value at a given time is known. Hence, the probability that $f(t)$ will have a value between f_1 and $f_1 + \Delta f_1$ at a given time and then, τ seconds later, having a value between f_2 and $f_2 + \Delta f_2$ is also known and is given as $p_2(f_1, f_2, \tau) df_1 df_2$. The autocorrelation function is then determined from

$$\phi_{11}(\tau) = \iint_{f_1, f_2} p_2(f_1, f_2, \tau) df_1 df_2,$$

as previously defined.

Example 2

Suppose the function $f(t)$ is given by

$$f(t) = E \sin(\omega_0 t + \psi)$$

Then the autocorrelation function is given by

$$\phi_{11}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T E \sin(\omega_0 t + \psi) E \sin[\omega_0(t + \tau) + \psi] dt$$

Since the integrand is periodic, the integration and limiting procedure can be replaced by the integral over one period and dividing by the period. Then

$$\phi_{11}(\tau) = \frac{\omega_0}{2\pi} E^2 \int_0^{2\pi/\omega_0} \sin(\omega_0 t + \psi) \sin[\omega_0(t + \tau) + \psi] dt$$

The integration is simplified by the change of variable

$$\omega_0 t + \psi = u$$

After substitution of the new variable we get

$$\phi_{11}(\tau) = \frac{E^2}{2\pi} \int_{\psi}^{2\pi + \psi} \sin u \sin(u + \omega_0 \tau) du$$

By expanding the integrand and performing the integration, the final result is

$$\phi_{11}(\tau) = \frac{E^2}{2} \cos \omega_0 \tau$$

The form of $\phi_{11}(\tau)$ is shown in Figure 67.

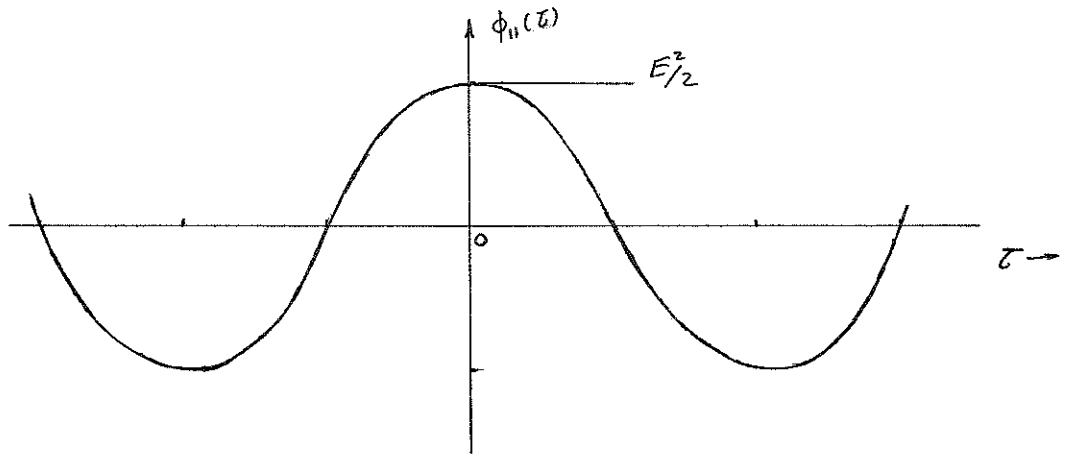


Fig. 67

Crosscorrelation Function

The autocorrelation function, as we have discussed it, has been concerned with a single kind of signal. However, many times we would like to consider two different signals. This method of correlation is called "crosscorrelation".

Consider a system with a statistical input $f_1(t)$ and an output time function $f_2(t)$, as shown in Figure 68. The system does not necessarily have to be linear. If no perturbing influences appear in the system, $f_2(t)$ will be

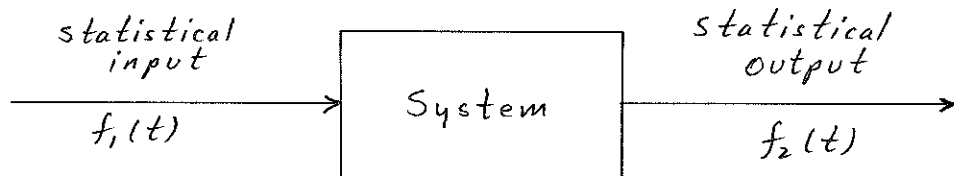


Fig. 68

uniquely related to $f_1(t)$ by the system function. If noise and other random disturbances are introduced by the system, however, $f_2(t)$ and $f_1(t)$ will be only partially related. The statistical relationship between $f_1(t)$ and $f_2(t)$ could be given by the joint probability distribution function $p(f_1, f_2, \tau)$, if it were known. On an ensemble basis then the crosscorrelation function

for the two signals would be

$$\phi_{12}(\tau) = \iint f_1(t) f_2(t) p(f_1, f_2, \tau) df_1 df_2, \quad (162)$$

where f_1 and f_2 are stationary processes. On a single member basis, the time average is

$$\phi_{12}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_1(t) f_2(t+\tau) dt. \quad (163)$$

For ergodic processes, the ensemble average is equal to the time average.

The following properties of the crosscorrelation function are determined from the definition.

(1) The crosscorrelation function $\phi_{ab}(\tau)$ is not an even function. In general, shifting $f_b(t)$ ahead by τ seconds yields a different result than retarding $f_b(t)$ by $-\tau$ seconds. $[\phi_{ab}(\tau) \neq \phi_{ab}(-\tau)]$

(2) $\phi_{ab}(\tau) \neq \phi_{ba}(\tau)$. It is not immaterial which variable is shifted ahead, as it was for the autocorrelation function.

(3) $\phi_{ab}(\tau) = \phi_{ba}(-\tau)$. A shift in $f_b(t)$ must yield the same result as a shift in $f_a(t)$ by the same amount in the opposite direction.

(4) Since the origin of time is important, the crosscorrelation function yields some information about the phase shift involved with periodic functions.

As an application of the crosscorrelation function to random inputs consider the following procedure:

The crosscorrelation between input and output for a system is given as

$$\phi_{i0}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_i(t) f_o(t+\tau) dt.$$

If the system is linear the output function $f_o(t+\tau)$ can be expressed as the convolution integral

$$f_o(t+\tau) = \int_0^{\infty} h(\lambda) f_i(t+\tau-\lambda) d\lambda \quad , \quad (164)$$

where $h(\lambda)$ is the system function. Substituting this into the equation for $\phi_{i_o}(\tau)$ gives

$$\begin{aligned} \phi_{i_o}(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_i(t) dt \int_0^{\infty} h(\lambda) f_i(t+\tau-\lambda) d\lambda \\ &= \int_0^{\infty} h(\lambda) d\lambda \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_i(t) f_i(t+\tau-\lambda) dt. \end{aligned} \quad (165)$$

The limiting function on the right is just the autocorrelation function of the input signal with a shift of $\tau - \lambda$. Thus

$$\phi_{i_o}(\tau) = \int_0^{\infty} h(\lambda) \phi_{i_i}(\tau-\lambda) d\lambda \quad , \quad (166)$$

and the crosscorrelation function between the input and output for a linear system with a statistical input is equal to the convolution integral of the autocorrelation function of the input and the system function.

Assume that the function $h(\lambda)$ is as shown in Figure 69. And that the

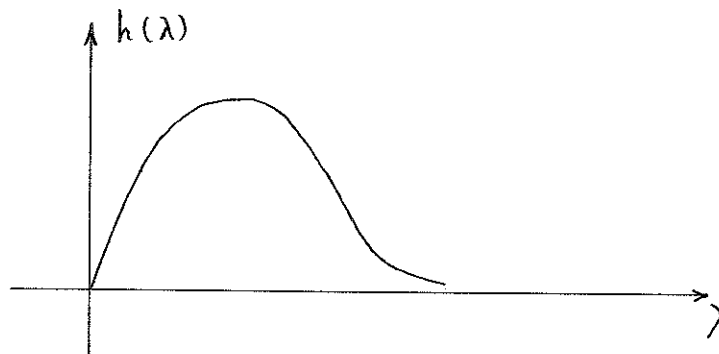


Fig. 69

autocorrelation function of the input is as shown in Figure 70, which may be considered as a good approximation to a unit impulse, (if the width Δt is very small).

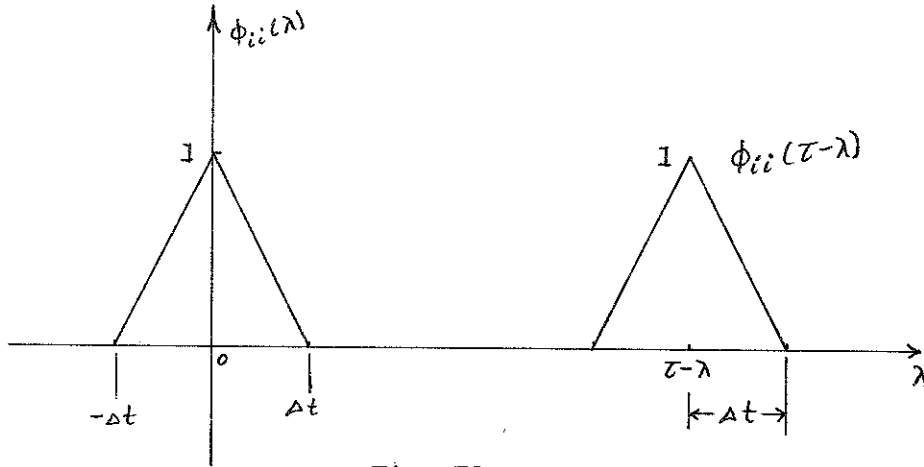


Fig. 70

Then multiplying the two values at each value of λ and integrating gives

$$\phi_{i0}(\tau) = \int_0^{\infty} h(\lambda) \phi_{ii}(\tau - \lambda) d\lambda \approx h(\tau) \Delta t \quad (167)$$

This says then that if the autocorrelation function of the input is a unit impulse, the crosscorrelation function of the input and output is approximately $h(\tau)$; i.e., the system function.

This suggests an experimental method of obtaining the linear portion of the time response of a linear system. Consider that we have a nuclear reactor and oscillate a control rod in and out in a statistical manner with Δt say 20 msec. If we delay the input by various times τ_1 , we can multiply the output and the delayed input and integrate, obtaining values of $h(\tau_1)$. Doing this for many values of τ_1 would give the crosscorrelation function of the input and output. We would thus obtain the system function.

The block diagram for such an experiment is shown in Figure 71.

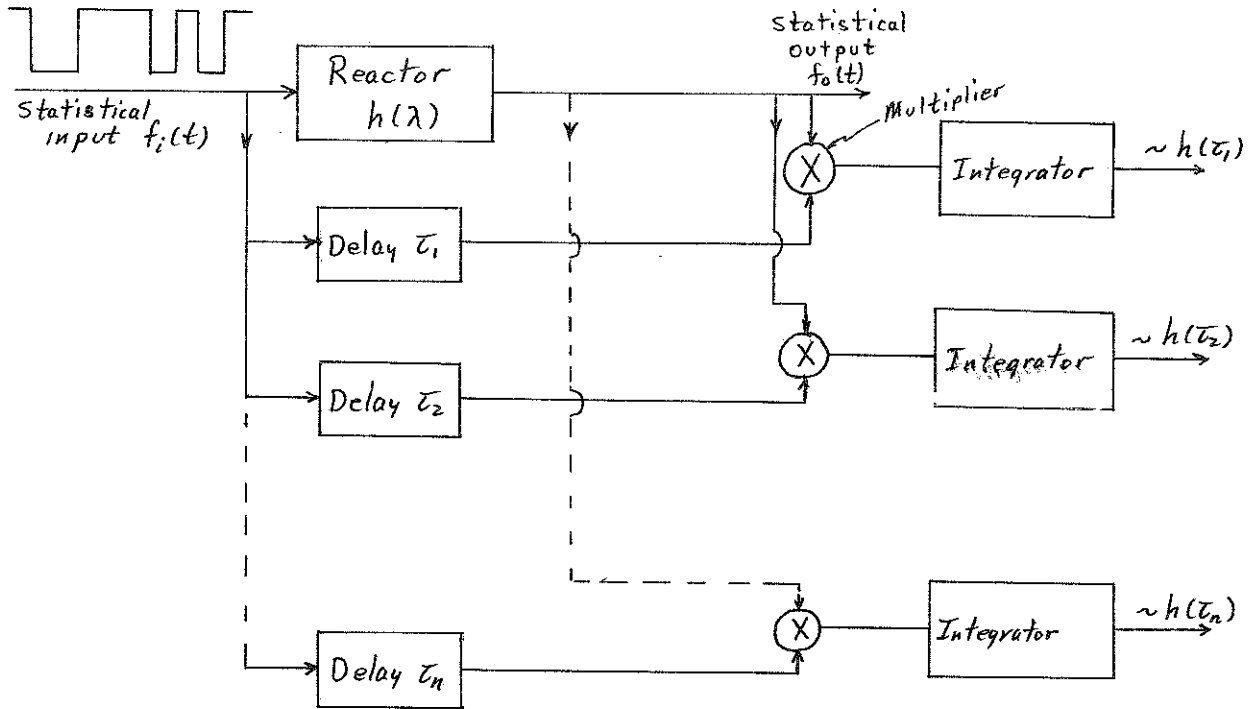


Fig. 71

Autocorrelation Function in the Frequency Domain (Power-Density Spectra)

In essence then, correlation functions describe the signals in terms of time-domain characteristics. For many purposes it is convenient, on the other hand, to describe certain signals in terms of frequency-domain characteristics.

Consider a linear system with a statistical input $f_i(t)$ and an output $f_o(t)$. Then

$$f_o(t) = \int_0^{\infty} h(\lambda) f_i(t-\lambda) d\lambda \quad (168)$$

The autocorrelation function of the output is

$$\phi_{oo}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_o(t) f_o(t+\tau) dt \quad , \quad (169)$$

as previously defined. Substituting Equation (168) into (169) for $f_o(t)$ and the similar expression

$$f_o(t+\tau) = \int_0^{\infty} h(\sigma) f_i(t+\tau-\sigma) d\sigma$$

for $f_o(t+\tau)$, we get

$$\phi_{oo}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T dt \int_0^{\infty} h(\lambda) f_i(t-\lambda) d\lambda \int_0^{\infty} h(\sigma) f_i(t+\tau-\sigma) d\sigma \quad (170)$$

Changing the order of integration

$$\phi_{oo}(\tau) = \int_0^{\infty} h(\lambda) d\lambda \int_0^{\infty} h(\sigma) d\sigma \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f_i(t-\lambda) f_i(t+\tau-\sigma) dt \quad (171)$$

But the limit function is just the autocorrelation function, $\phi_{ii}(\tau+\lambda-\sigma)$, of the input where the shift is $\tau+\lambda-\sigma$.

Therefore

$$\phi_{oo}(\tau) = \int_0^{\infty} h(\lambda) d\lambda \int_0^{\infty} h(\sigma) d\sigma \phi_{ii}(\tau+\lambda-\sigma) \quad . \quad (172)$$

Taking the Fourier transform of both sides of Equation (172) gives

$$\phi_{oo}(j\omega) = \int e^{-j\omega\tau} d\tau \int h(\lambda) d\lambda \int h(\sigma) d\sigma \phi_{ii}(\tau+\lambda-\sigma) \quad . \quad (173)$$

Changing the order of integration and multiplying by $e^{-j\omega(\lambda+\sigma)}$ and $e^{+j\omega(\lambda+\sigma)}$ gives

$$\phi_{oo}(j\omega) = \int h(\lambda) e^{j\omega\lambda} d\lambda \int h(\sigma) e^{-j\omega\sigma} d\sigma \int \phi_{ii}(\tau+\lambda-\sigma) e^{-j\omega(\tau+\lambda-\sigma)} d\tau. \quad (174)$$

In Equation (174) the first integral on the right-hand side is the complex conjugate of the transfer function of the system, $H^*(j\omega)$; the second integral is the transfer function of the system, $H(j\omega)$; and the third integral is the Fourier transform of the autocorrelation function of the input. Thus,

$$\begin{aligned} \phi_{oo}(j\omega) &= H^*(j\omega) \cdot H(j\omega) \cdot \phi_{ii}(j\omega) \\ &= |H(j\omega)|^2 \phi_{ii}(j\omega) \end{aligned} \quad (175)$$

Equation (175) indicates then that the Fourier transform of the autocorrelation function of the output is related to the Fourier transform of the autocorrelation function of the input by the square of the absolute value of the system transfer function.

The Fourier transform of the autocorrelation function of a random signal $f(t)$ is referred to as the power-density spectrum of $f(t)$.

Characteristics of Power-Density Spectra

There are several important characteristics associated with power-density-spectra functions. These are:

(1) $\phi_{aa}(j\omega)$ measures the power-density spectrum rather than the amplitude or phase spectra of the signal $f_a(t)$. Consequently, the relative phase of the various frequency components is lost when the signal is described by means of a power-density spectrum.

(2) As a result of discarding the phase information, a given power-density spectrum may correspond to a large number of different time functions.

(3) $\phi_{aa}(j\omega)$ is an even function of frequency,

$$\phi_{aa}(j\omega) = \phi_{aa}(-j\omega) \quad . \quad (176)$$

This characteristic follows directly from the original definition of $\phi_{aa}(j\omega)$ as the Fourier transform of $\phi_{aa}(\tau)$ and the fact that the autocorrelation function, $\phi_{aa}(\tau)$ is an even function of τ . The defining integral for $\phi_{aa}(j\omega)$ is

$$\phi_{aa}(j\omega) = \int_{-\infty}^{\infty} \phi_{aa}(\tau) e^{-j\omega\tau} d\tau \quad . \quad (177)$$

The exponential can be replaced by the trigonometric form:

$$\phi_{aa}(j\omega) = \int_{-\infty}^{\infty} \phi_{aa}(\tau) \cos \omega\tau d\tau - j \int_{-\infty}^{\infty} \phi_{aa}(\tau) \sin \omega\tau d\tau \quad .$$

Since the second integrand is an odd function of τ , the integral is zero and

$$\phi_{aa}(j\omega) = \int_{-\infty}^{\infty} \phi_{aa}(\tau) \cos \omega\tau d\tau \quad ,$$

which is an even function of ω .

(4) $\phi_{aa}(j\omega)$ is nonnegative at all frequencies. A negative $\phi_{aa}(j\omega)$ would indicate power being taken from the system.

(5) If the signal contains a periodic component such that the Fourier series for the component contains terms representing frequencies, $\omega_1, \omega_2, \dots, \omega_n$, $\phi_{aa}(j\omega)$ contains impulses at $\omega_1, -\omega_1, \omega_2, -\omega_2, \dots, \omega_n, -\omega_n$.

This characteristic is apparent from the equation

$$\phi_{aa}(j\omega) = \int_{-\infty}^{\infty} \phi_{aa}(\tau) \cos \omega \tau d\tau \quad .$$

If $f_a(t)$ contains a periodic component of frequency ω_1 , $\phi_{aa}(\tau)$ will contain a term of the form $a_1 \cos \omega_1 \tau$. The corresponding part of $\phi_{aa}(j\omega)$ is

$$\phi_{aa}(j\omega_1) = a_1 \int_{-\infty}^{\infty} \cos \omega_1 \tau \cos \omega \tau d\tau \quad .$$

The right-hand side of this equation is zero for all ω other than ω_1 or $-\omega_1$ (by orthogonality) and is infinite at these two frequencies. The area under $\phi_{aa}(j\omega)$, however, is finite and equal to the power contained in the sinusoidal component. This infinite spike of finite area is just the definition of an impulse.

Crosscorrelation Function in the Frequency Domain

By the definition given above, the crosscorrelation function is given as

$$\phi_{io}(\tau) = \int_0^{\infty} h(\lambda) \phi_{ii}(\tau - \lambda) d\lambda \quad . \quad (178)$$

Taking the Fourier transform of Equation (178) gives

$$\phi_{io}(j\omega) = H(j\omega) \phi_{ii}(j\omega) \quad , \quad (179)$$

which is similar to Equation (175) for the power-density spectrum, however, in this case the phase of the transfer function is included.

Suggested References

1. G. C. Newton, L. A. Gould, and J. F. Kaiser, Analytic Design of Linear Feedback Controls. New York: John Wiley & Sons, 1957.
2. John G. Truxal, Automatic Feedback Control System Synthesis. New York: McGraw-Hill Book Company, 1955.
3. Mischa Schwartz, Information Transmission, Modulation, and Noise. New York: McGraw-Hill Book Company, 1959.

LECTURES NOS. VI AND VII

APPLICATIONS OF GEOMETRIC THEORY TO NONLINEAR REACTOR DYNAMICS

For the first five lectures we assumed that linear systems were of primary concern. Although many systems can be analyzed by the linear methods just described, in the field of dynamics, and particularly in nuclear reactor systems, the equations involved in attempting to describe the systems are nonlinear. The linear approach is not completely useless for these systems however since, as we will see later, under certain conditions there may be a relationship between the general solution of a set of nonlinear differential equations and the solution of its linear approximation.

Nuclear reactor dynamics can be fairly well represented by a set of first order, space independent nonlinear differential equations with respect to time (1, 2). The complete solution of this set of equations is in general a formidable, if not impossible task. However, under various simplifying assumptions explicit solutions of different reactor dynamics problems have been found and have been reported (3 - 5).

One of the most prominent simplifications that is repeatedly used is the linearization of the dynamic equations, which immediately leads to closed form solutions. Such solutions have the important property that they afford experimental verification by means of oscillation or other "small signal" tests, without any hazards (6, 7).

In view of the nonlinear character of the dynamic equations the justified question is often raised about the real value of the linearized or transfer function approach to the problem of reactor dynamics analysis, or, stated differently, about the connection between the "exact" solution and the one derived from the linearized model.

Mathematically speaking this question has a well defined answer. However, there seems to be some misunderstanding in the nuclear reactor field. Conflicting and unqualified statements like "the linearized equations are a very good approximation" and "the linearized equations are an inadequate representation" appear very often in the nuclear literature.

The purpose of the discussion today is twofold. First, it gives a brief summary of some basic notions of the geometric theory of differential equations which unambiguously answer the previous question. This theory is well known in the mathematical literature (8) and various of its aspects pertaining to nuclear reactor dynamics have already been presented (9 - 11). However, it is felt that the power of the geometric theory is not yet fully appreciated. The power of the method lies in the fact that the properties of the solutions of a system of nonlinear differential equations can be visualized in terms of straightforward geometric or topological relationships which yield information about the existence of critical points, the boundedness and stability of the solutions, the existence of periodic solutions and the gross interrelated features of the solutions (maxima, minima, directions of variation, etc.).

Second, we will discuss the dynamic behavior of two reactors describable by third order nonlinear differential equations by means of purely geometric methods. The first is a xenon controlled reactor. This reactor has been analyzed by Chernick (12) by means of numerical and "classical" procedures but the present approach does not require lengthy computations or simplifying approximations. The second is a heterogeneous reactor with two temperature coefficients of reactivity which may have opposite signs. This problem had not been treated so far in full detail.

The investigation of these two reactors brings out the following important points:

- a. It clearly indicates the elegance and simplicity of the geometric

theory by means of which conditions for boundedness and stability in the large and existence of periodic solutions are established.

b. It definitely shows that the linearized model of a reactor does not necessarily contain all the information required for the large signal performance.

c. It implies that even though the solutions may be bounded or periodic under certain conditions, this does not necessarily mean that the upper bounds are tolerable.

A. GEOMETRIC THEORY OF AUTONOMOUS DIFFERENTIAL EQUATIONS

1. The Problem in General

Consider that we wish to investigate a system for which x_1, x_2, \dots, x_n are variables of the system; i.e., they could represent temperature, reactivity, coolant flow rates, etc. Let's define \bar{x} to be the column matrix (vector)

$$\bar{x} \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \quad (180)$$

We will also assume that the x_i 's are related in some way on the basis of fundamental principles of physics; i.e., conservation laws, etc. This means that the variables are interrelated and we may write

$$\frac{d\bar{x}}{dt} = X(\bar{x}) \quad , \quad (181)$$

where X is a matrix (function of \bar{x}) and Equation (181) is a shorthand notation for the set of equations

$$\begin{aligned} \frac{dx_1}{dt} &= a_1 x_1 + a_2 x_2 + \dots + b_{12} x_1 x_2 + \dots + b_{11} x_1^2 + \dots \\ &\vdots \\ &\vdots \\ &\vdots \\ \frac{dx_n}{dt} &= a_1 x_1 + a_2 x_2 + \dots + b_{12} x_1 x_2 + \dots + b_{11} x_1^2 + \dots \end{aligned}$$

Let's assume that we can write Equation (181) as a sum of linear and non-linear terms

$$\frac{d\bar{x}}{dt} = X_1(\bar{x}) + X_2(\bar{x}) \quad , \quad (182)$$

where the first term $X_1(\bar{x})$ contains only the lowest order powers of x_i and the second term $X_2(\bar{x})$ contains all higher powers of x_i . Then there are values, $|\bar{x}| < A$, such that the magnitude of the nonlinear part of Equation (182) is less than the magnitude of the linear part. As an example, consider $X_1(\bar{x}) = x + y$ and $X_2(\bar{x}) = xy$. Then

$$|xy| < |x| + |y| \quad ,$$

which is true if x or y or both are less than unity.

If we consider only the first term of Equation (182) and form the equation

$$\frac{d\bar{x}}{dt} = X_1(\bar{x}) \quad , \quad (183)$$

this is known as the first approximation. If $X_1(\bar{x})$ contains only powers of x_i to the first order, then a lot of information about the solution of Equation (181) can be obtained from a detailed study of its first approximation. However, if $X_1(\bar{x})$ contains powers of x_i equal to or greater than two, very little is known about the relationship or if one even exists.

Assuming then that $X_1(\bar{x})$ is a linear function, there are two types of problems to be considered;

(1) The first approximation is linear with time dependent coefficients; i.e., the elements of the X_1 matrix are time dependent, and

(2) The first approximation is linear but the coefficients are constants with respect to time.

Both problems have been extensively treated, the first by Picard (14), Poincaré (15), and others, and the second by Liapunov (13). We will investigate only the problem with constant coefficients and will follow Liapunov's treatment, as described by Lefschetz (8).

2. Stability in the Small of a System with Linear First Approximation

The problem of stability for small amplitude displacements is treated by what is known as Liapunov's first method and is as follows:

We are concerned with the solution of the system of equations

$$\frac{d\bar{x}}{dt} = X(\bar{x}), \quad (181)$$

which have constant coefficients. Systems of equations with these characteristics are called autonomous systems of equations. Autonomous systems are particularly applicable to the types of experiments performed here at Spert. The introduction of a step change in reactivity means that the external disturbance is essentially independent of time and the reactor is autonomous.

The question of primary importance is stability. Let's rewrite Equation (181) in the form

$$\frac{d\bar{x}}{dt} = P\bar{x} + Q(\bar{x}), \quad (184)$$

where P is a constant matrix made up of the coefficients of first order terms only, and has characteristic roots λ_i , and $Q(\bar{x})$ is some function of the vector \bar{x} and contains second and higher order terms.

Consider only the linear first approximation

$$\frac{d\bar{x}}{dt} = P\bar{x} \quad (185)$$

Assume that $\bar{x} = 0$ is a solution of both Equation (184) and Equation (185);
i.e.

$$\left. \frac{d\bar{x}}{dt} \right|_{\bar{x}=0} = 0 \quad \text{and} \quad \left. P\bar{x} \right|_{\bar{x}=0} = 0$$

therefore

$$\left(\frac{d\bar{x}}{dt} \right)_{\bar{x}=0} = 0 \quad (186)$$

We then call $\bar{x} = 0$ a "critical point" or "equilibrium point". Thus, we can consider $\bar{x} = 0$ as the origin of an n-dimensional space. In reactor experiments if we introduce a small step change in reactivity, the reactor will finally settle out at some fairly constant power level. The reactor is then critical at this power level and the value of the power is called a critical point for the existing conditions of temperature, flow, etc., in the system at that time. It would appear that for a reactor the critical point involves a value of \bar{x} different from zero since the power and other variables are not zero at the critical point. This can be taken care of very easily by shifting the coordinates; i.e., by making the variables be the sum of the critical value and some incremental value. As an example, let $x_1 = A_1$ and $x_2 = A_2$.

Then we can write

$$\begin{aligned} x_1 &= A_1 + x_1' \\ x_2 &= A_2 + x_2' \end{aligned}$$

and the reactor has a critical point when $x_1' = 0$ and $x_2' = 0$. Doing this for all the variables, we would have

$$\bar{x} \equiv \begin{bmatrix} x'_1 \\ x'_2 \\ \cdot \\ \cdot \\ x'_n \end{bmatrix},$$

and a critical point would be for $\bar{x} = 0$, since $x'_1, x'_2, x'_3, \dots, x'_n = 0$.

We will now consider some theorems from Liapunov's first method.

Theorem I

From Equation (185) if the characteristic roots of the matrix P have negative real parts, then there is a range $|\bar{x}| < A$, such that $|Q(\bar{x})| < |\bar{x}|$ and in this range the system described by Equation (184) is stable.

Here is the first statement of a relationship between the solution of Equation (185) and the solution of Equation (184). This says that, if from the linear first approximation of a nonlinear system we determine that the characteristic roots are in the left-half-plane, there is always a small range of values of \bar{x} such that, if the linear approximation is stable in this range, the nonlinear system is also stable in this range.

Thus we have found a means of determining if a complicated, nonlinear system is stable merely by examining the behavior of the simpler linear approximation to the system.

Theorem II

Assume that $Q(\bar{x}) = \sum_2^N a_i \bar{x}^i$ and that

$$\lambda_j \neq \sum_i m_i \lambda_i \quad ; \quad m_i \geq 0 \quad ; \quad \sum m_i = 1 .$$

Then if all the characteristic roots of the matrix P have negative real parts, there is a spheroid region $R(A)$, $|a| \leq \rho$, in which the solution of Equation (184) is asymptotically stable at the origin. If all the characteristic roots have positive real parts, the solution is unstable (which implies that the

equilibrium point is a point of unstable equilibrium) and if some of the characteristic roots have positive and some negative real parts, the solution is conditionally stable. For all practical purposes, however, since the behavior for this latter condition is usually peculiar, the system may be considered unstable.

The important conclusion that theorems I and II bring out is that the study of the linear equation

$$\frac{d\bar{x}}{dt} = P\bar{x} \quad (185)$$

yields useful information about the solution of Equation (184) provided that one uses this information within the amplitude range for which it is established. It is exactly this range which qualifies the validity or insufficiency of the linear approximation and one should not expect the results derived from an investigation of Equation (185) to have any meaning for large amplitude displacements.

3. Boundedness and Stability in the Large for Analytical Systems with Linear First Approximation

Consider again Equation (195) with $Q(\bar{x})$ an analytic function. The boundedness and stability of the solutions for large amplitude displacements can be inferred from a geometric interpretation of Liapunov's second method which consists in the following (13, 8):

Theorem III

Given the set of Equations (184), where $Q(\bar{x})$ is analytic, if there exists a scalar function $V(\bar{x})$ which is definite positive (i.e., is positive for all values of \bar{x} in the range of interest) and if the derivative $\frac{dV(\bar{x})}{dt}$ is negative, then in a region $R(A)$ of the phase space, the origin is stable. Furthermore, if $V(\bar{x}) \rightarrow 0$ for $\bar{x} \rightarrow 0$, the origin is asymptotically stable.

Theorem IV

If a scalar function $V(\bar{x})$ is defined (not necessarily definite) and approaches zero for $\bar{x} \rightarrow 0$ and such that $\frac{dV(\bar{x})}{dt}$ is definite positive and for $|\bar{x}| < \eta$, no matter how small η is, $V(\bar{x})$ may take the sign of $\frac{dV(\bar{x})}{dt}$, then the origin is unstable.

These theorems are equivalent to theorem II. However, they afford a simple geometric interpretation extremely useful for the purposes of this discussion. Suppose theorem III is applicable; i.e., $V(\bar{x}) > 0$. Let $V(\bar{x}) = \epsilon$ be a constant greater than zero and small at a given time. Then $\epsilon = V(\bar{x})$ is a closed surface which for different values of ϵ , represent concentric ovals which tend to the origin as $\epsilon \rightarrow 0$. If $\frac{dV(\bar{x})}{dt} = \frac{d\epsilon}{dt} < 0$, the vector $\frac{dx_1}{dt}$ points inward along every point of $V(\bar{x}) = \epsilon$, and hence the surface collapses toward the origin. This is the meaning of stability.

A surface collapsing means that the trajectories $x_1(t)$ must have a negative slope; i.e., the tangent to the trajectories must point inwardly toward the origin. If the surface were to expand, the tangent to the trajectories would point outwardly. When theorem IV is applicable the vector $\frac{dx_1}{dt}$ points sometimes outward and sometimes inward. Thus, the system is manifestly unstable.

Reversing the argument, one might state that if there exists a surface surrounding the critical point which is large enough to enclose all possible displacements, and is such that the vector field $d\bar{x}/dt$ crosses it everywhere inwardly, then the solutions of Equation (184) are bounded. If the critical point is stable in the sense of Liapunov, the trajectories $x_1(t)$ coalesce to the critical point. If the critical point is unstable, the system may admit periodic solutions.

The power and elegance of this interpretation will become more evident when the problem of boundedness and stability of the xenon controlled reactor and the dynamics of reactors with two temperature coefficients are discussed.

In summary, if the solution of the linear approximation of Equation (184) is stable and the solution of Equation (184) is bounded, then the solution of the latter is also stable. If the solution of the first approximation is unstable, this does not necessarily mean that the solution of Equation (184) is unbounded or lacking periodic closed paths. These results are self-evident since two nonlinear systems may have identical linear approximations but different nonlinear terms, and the large amplitude behavior is determined by the nonlinear terms and not the linear approximations.

4. Existence of Periodic Solutions of Analytical Systems with Linear First Approximation

Conditions for the existence of periodic solutions of autonomous systems have been established by Liapunov (13), Malkin (16), and Poincaré (17). They are based on the principle of analytic continuation and are extremely difficult to implement in any practical case. These conditions will not be discussed here, but are presented in Reference (8). Suffice to note only that the existence of periodic solutions is based on some characteristic properties of the coefficient matrix of the linear approximation and boundedness of the solutions. This indicates that the linear approximation or transfer function approach is useful in determining the existence of periodic solutions but not adequate by itself.

As a substitute for the general conditions for the existence of periodic solutions, Poincaré's method of sections and Brouwer's fixed point theorem will be considered because they are most pertinent to the purposes of this discussion.

Consider a closed region, topologically equivalent to a solid torus, free of critical points and such that the vector field dx_i/dt points inwardly at every point of the surface enclosing the region. In other words, consider a toroidal trap for trajectories $x_i(t)$ that lie inside it. Assume that the trajectories intersect a certain cross section S_1 of the torus without contact, that is without ever being tangent to it. This means that the vector field dx_i/dt intersects S_1 at points Q, Q', \dots and defines a topological mapping $Q \rightarrow Q'$ of S_1 into itself. If the cross section S_1 has a fixed point P , namely if P is mapped into itself, then the particular trajectory that corresponds to $P \rightarrow P$ is closed and therefore periodic. This is Poincaré's method of sections.

In addition, if a simply connected section S_1 is mapped into itself by means of a continuous function, the mapping possesses at least one fixed point and consequently it admits at least one closed or periodic path. This is Brouwer's fixed point theorem (18).

Poincaré's method of sections and Brouwer's fixed point theorem prove very useful in the geometric analysis of nonlinear differential equations as will be emphasized in the subsequent examples.

B. DYNAMICS OF XENON CONTROLLED REACTORS

1. The Model

The reactor model is the same as the one considered by Chernick (12) and is describable by the following set of equations:

$$\frac{d\phi}{dt} = \frac{\rho - \beta}{\tau} \phi + \sum_n \lambda_n C_n \quad (186)$$

$$\rho = \rho_0 - \frac{\sigma_x X}{c \sigma_f} \quad (187)$$

$$\frac{dC_n}{dt} = \frac{\beta_n}{\tau} \phi - \lambda_n C_n \quad (188)$$

$$\frac{dX}{dt} = y_x \sigma_f \phi - \sigma_x X \phi + \lambda_i I - \lambda_x X \quad (189)$$

$$\frac{dI}{dt} = y_i \sigma_f \phi - \lambda_i I \quad (190)$$

All symbols are defined in Reference (12). The system of Equations (186-190) admits a critical point (excluding the one at the origin):

$$\phi_\infty = \frac{c \delta_0 \lambda_x}{\sigma_x (y - c \delta_0)} \quad X_\infty = \frac{c \delta_0 \sigma_f}{\sigma_x}$$

$$I_\infty = \frac{y_i \sigma_f \phi_\infty}{\lambda_i} \quad C_{n\infty} = \frac{\beta_n \phi_\infty}{\lambda_n \tau}$$

provided that $y = y_x + y_i > c \delta_0$. No critical point exists when $y < c \delta_0$.

If the delayed neutrons are considered only through their effects on the neutron mean lifetime (12) and the other variables are measured in terms of their equilibrium values, the system (186-190) reduces to:

$$\frac{d\phi}{dt} = \omega_0 [1 - X] \phi \quad (191)$$

$$\frac{dX}{dt} = \lambda_x [a\phi + \beta I - \gamma\phi X - X] \quad (192)$$

$$\frac{dI}{dt} = \lambda_i [\phi - I] \quad (193)$$

where:

$$\omega_0 = \frac{\delta_0}{\tau_e}$$

τ_e = equivalent neutron lifetime

$$\alpha = \frac{y_x}{y - c\delta_0} \quad \beta = \frac{y_i}{y - c\delta_0} \quad \gamma = \frac{c\delta_0}{y - c\delta_0}$$

$$\alpha + \beta - \gamma = 1$$

2. Stability of the Critical Point

The type of stability at the critical point can be investigated by considering the linear approximation of Equations (191-193).

The characteristic equation of the linear approximation is:

$$s^3 + [\lambda_x(\alpha + \beta) + \lambda_i]s^2 + \lambda_x[\lambda_i(\alpha + \beta) + \omega_0(\alpha - \gamma)]s + \lambda_i\lambda_x\omega_0 = 0 \quad (194)$$

or, what is equivalent:

$$1 + \omega_0\lambda_x(\alpha - \gamma) \frac{s + \frac{\lambda_i}{\alpha - \gamma}}{s[s + \lambda_x(\alpha + \beta)][s + \lambda_i]} = 0 \quad (195)$$

The roots of Equation (195) can be determined by means of the root locus method. Consider two cases:

a. $\alpha > \gamma$ ($y_x > c\delta_0$)

The root locus is shown in Figure 72. The critical point is stable when the asymptote is in the left-half s-plane (Figure 72a). This is true when:

$$\frac{\lambda_i}{\alpha - \gamma} \leq \lambda_x(\alpha + \beta) + \lambda_i \quad (196)$$

Equation (196) is fulfilled when:

$$\psi = \frac{y_x}{y} \geq \frac{\lambda_i}{\lambda_i + \lambda_x} = \frac{1}{1 + \Lambda} \quad ; \quad (197)$$

$$\Delta = \frac{c\delta_0}{y} \leq \frac{(1 + \Lambda)\psi - 1}{\psi - 1 + \Lambda} \quad (198)$$

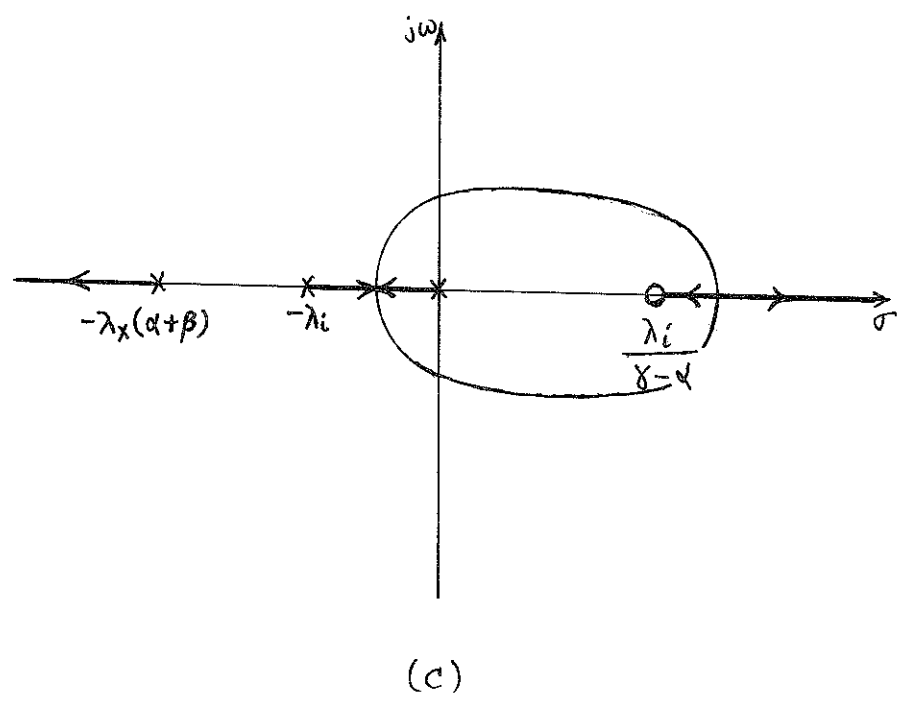
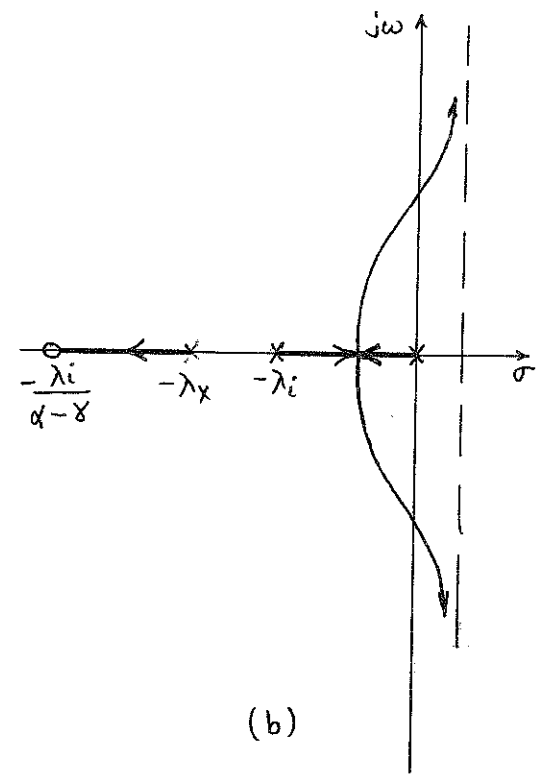
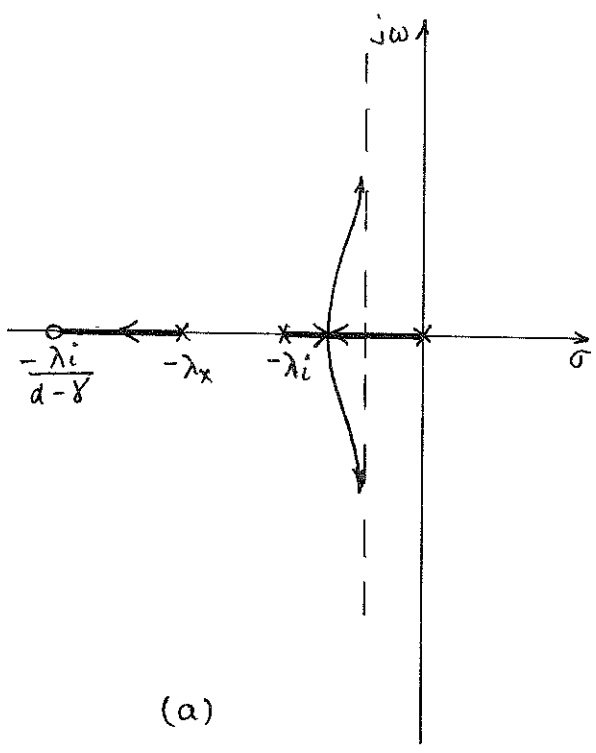


Fig. 72

If condition (196) is not true, the asymptote is in the right-half-plane and the critical point becomes unstable when (Figure 72b):

$$\omega^2 \geq \lambda_i \lambda_x \frac{1}{1 - \psi + \Lambda(\beta - 1)} \quad ; \quad (199)$$

$$\Delta \geq \frac{c \lambda_i \tau_e}{\gamma} \frac{1 + \Lambda - \Delta}{1 - (1 + \Lambda)\psi + (\psi + 1 - \Lambda)\Delta} \quad . \quad (200)$$

Equations (198) and (200) are plotted in Figure 73 for $\lambda_x = 2.09 \times 10^{-5} \text{ sec}^{-1}$, $\lambda_i = 2.87 \times 10^{-5} \text{ sec}^{-1}$, $\gamma = 6.4 \times 10^{-2}$, $\tau_e = 0.1 \text{ sec.}$, $c = 1.5$. Notice that the two plots are practically identical.

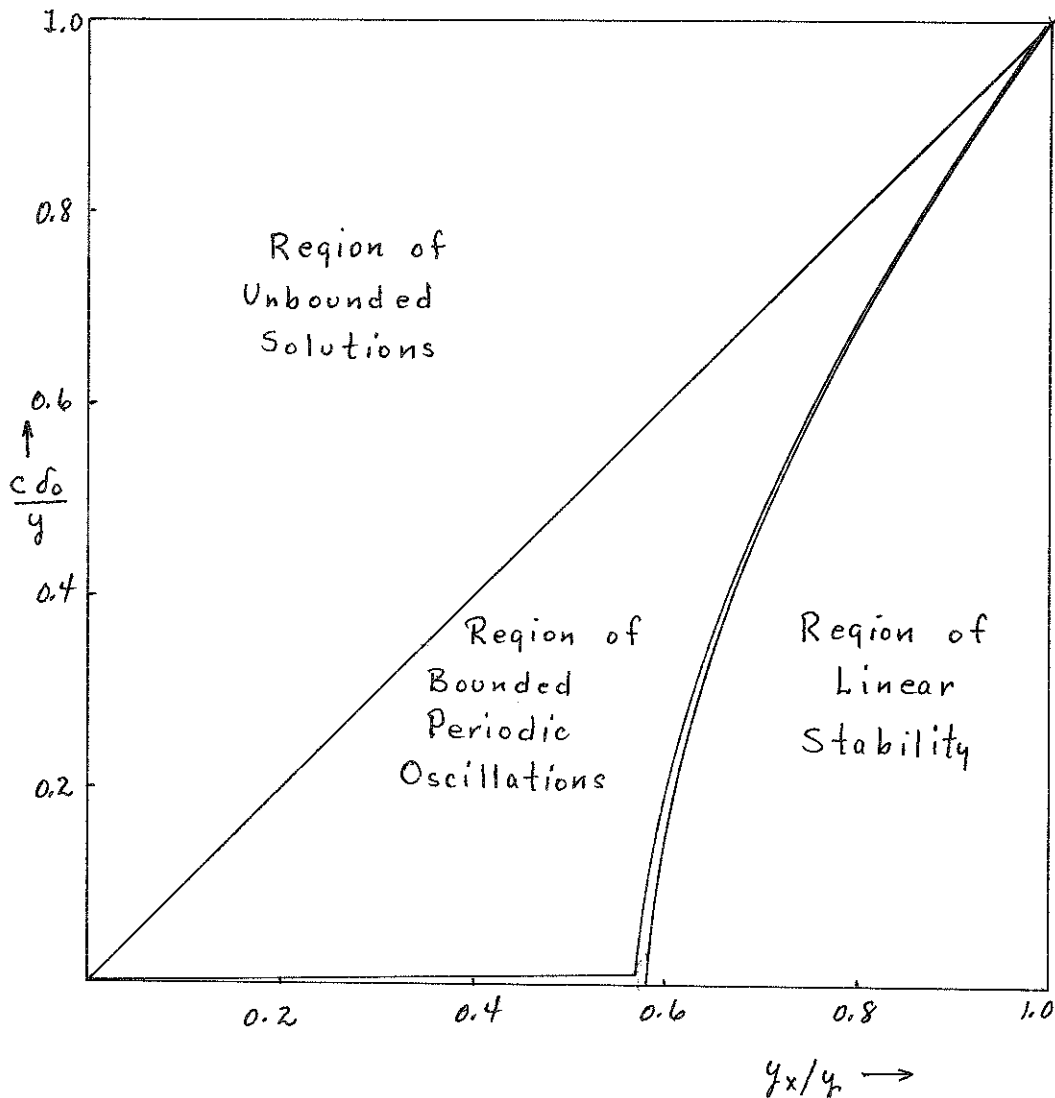


Fig. 73

b. $\alpha < \gamma$ ($\gamma_x < c d_0 < \gamma$)

The root locus is shown in Figure 72c. The conditions for instability are given again by Equations (199) and (200). Simple inspection of Figure 73 indicates that since the critical $\Delta < \Psi$, the critical point is always unstable when $\alpha < \gamma$.

In summary, the linear approximation yields an unstable critical point when the reactivity controlled by xenon is greater than what is given by Equation (200) (Figure 73) and in particular when it is greater than the prompt xenon yield. Most of these results are also given in Reference (12).

3. Boundedness and Stability in the Large

The boundedness of the solutions for very large displacements can be found using the geometric interpretation of Liapunov's second method. More precisely, one investigates the existence of a closed surface in the phase space (ϕ, I, X) which encloses the critical point (1, 1, 1) and is intersected by the trajectories inwardly.

Consider first the case when $\alpha > \gamma$ and the surface shown in Figure 74a. This surface consists of eight mutually intersecting plane surfaces defined as follows:

Take the arbitrary point A(a,a,a) with $a > 1$. Define the plane surfaces:

ABCDFGA: Plane $E_1 // I = 0$ through point A

ABKA: Plane E_2 defined by $\phi - I = 0$ ($X \gg a$)

ALKA: Plane E_3 defined by $X = a$ ($\phi \gg I$)

ALMGA: Plane E_4 defined by $\phi = a$ ($a \gg X \gg 1, \phi \gg I$)

BKCB: Plane E_5 defined by $\alpha \phi + \beta I - X = -a$ ($X \leq a, I \leq \phi$)

CDFONKC: Plane E_6 defined by $\phi = 0$

KLIMONK: Plane E_7 defined by $I = 0$

FGMOF: Plane E_8 defined by $\phi = aX$ ($X \leq 1$)

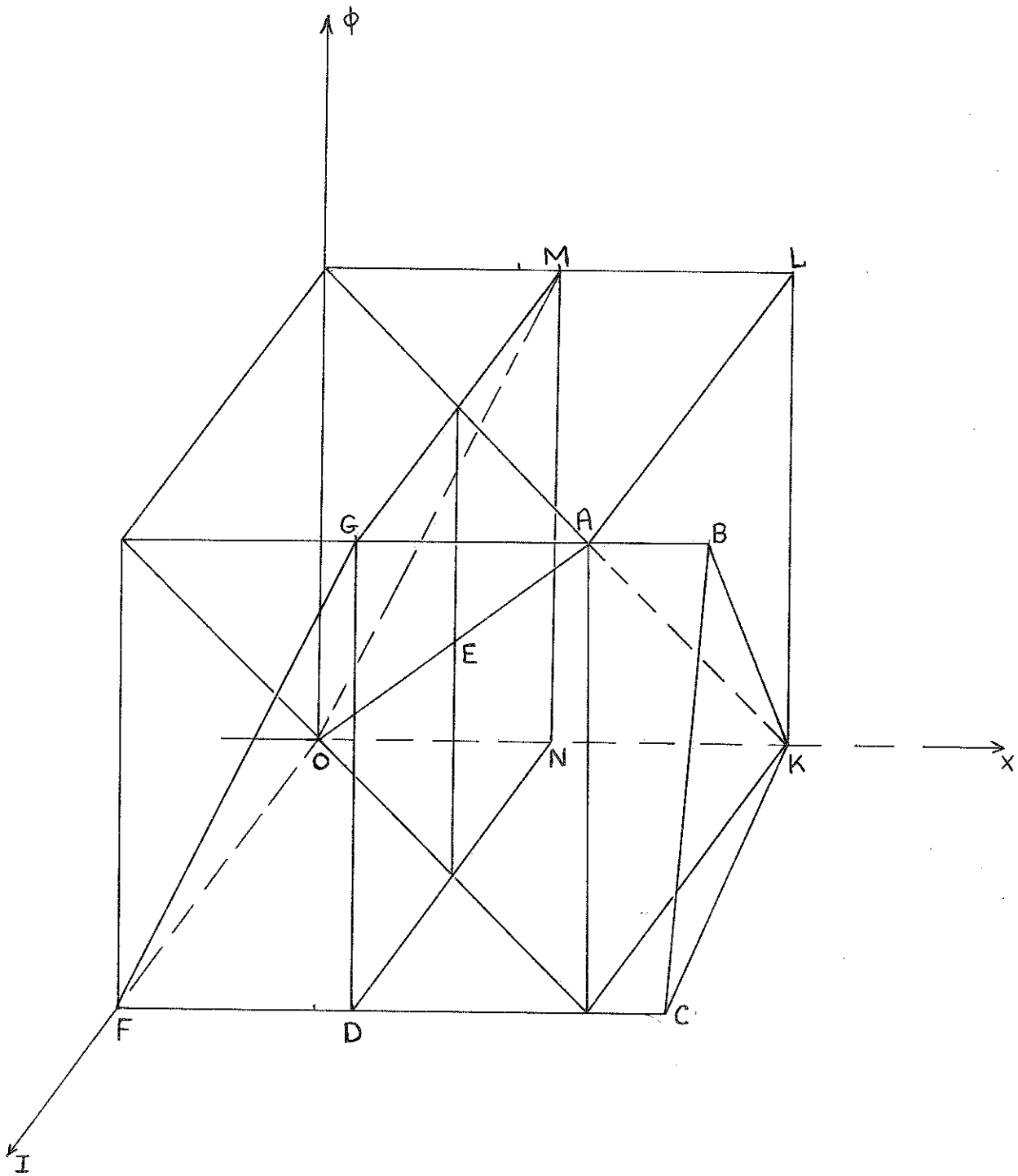


Fig. 74a.

It is evident that this surface does enclose the critical point $E(1, 1, 1)$.

All the trajectories cross the surface inwardly because:

$$\frac{dI}{dt} < 0 \quad \text{on } E_1$$

$$\frac{d\phi}{dt} < 0, \quad \frac{dI}{dt} = 0 \quad \text{on } E_2$$

$$\frac{dX}{dt} < 0 \quad \text{on } E_3 \text{ provided that } a \text{ is large}$$

$$\frac{d\phi}{dt} < 0 \quad \text{on } E_4$$

$$\frac{d}{dt}(\phi^2 + X^2 + I^2) < 0 \quad \text{on } E_5$$

$\phi = 0 \quad \frac{d\phi}{dt} = 0 \quad \text{on } E_6$. The trajectories come only arbitrarily close to E_6 , because $\phi < 0$ has no physical meaning.

$$\frac{dI}{dt} > 0 \quad \text{on } E_7$$

Finally on E_8 observe the following: The surface S :

$$\alpha\phi + \beta I - \gamma\phi X - X = 0 \quad (201)$$

crosses:

$$\text{the line: } I = 0, X = 1 \quad \text{at } \phi = 1/(\alpha - \gamma) > 0$$

$$\text{the line: } \phi = I = X \quad \text{at } \phi = I = X = 1$$

$$\text{the line: } \phi = 0, X = 1 \quad \text{at } I = 1/\beta > 1$$

If a is chosen large, the plane E_8 lies to the left of the surface S and therefore:

$$\frac{dX}{dt} > 0 \quad \text{on } E_8.$$

In addition, the projection of the vector $\frac{d\phi}{dt}, \frac{dI}{dt}, \frac{dX}{dt}$ on the normal of E_0 is:

$$P_n = \omega_0 (1-x)\phi - a \lambda_x [\alpha \phi + \beta I - \gamma \phi X - X]$$

$$= a \left[(\lambda_x a \gamma - \omega_0) X^2 - (\lambda_x \alpha a - \omega_0 - \lambda_x) X - \lambda_x \beta I \right] < 0 \quad (202)$$

provided that $\alpha > \gamma$ and a large. Consequently the trajectories cross E_0 inwardly.

In view of the fact that the only requirement for the existence of the surface of Figure 74a is that a be large, it is evident that such a surface can be made to include all trajectories and since it is intersected inwardly by all trajectories it constitutes a trajectory trap. Consequently, when $\alpha > \gamma$, the solutions of the system of Equations (191-193) are bounded. Furthermore, if the critical point is stable the solutions are asymptotically stable while if the critical point is unstable, the system admits, in general, periodic solutions as it will be shown in the next section.

Next, consider the case when $\alpha < \gamma$ and distinguish the following ranges:

a.
$$\frac{\omega_0 (\gamma - \alpha)}{\lambda_i \gamma} < 1 \quad (\gamma_x < c \sigma_0 < \gamma_x + c \lambda_i \tau_e)$$

Consider the closed surface shown in Figure 74b which consists of seven mutually intersecting plane surfaces defined as follows:

Take the arbitrary point $A(\phi = b, I = b, X = \beta b)$ with $b > 1$. Define the plane surfaces:

- ABCD: Plane E_1 defined by $X = \beta b$
- ADFGHA: Plane E_2 defined by $I = b$ in the region $I > \phi$
- CDFOC: Plane E_3 defined by $\phi = 0$
- CBKLOC: Plane E_4 defined by $I = 0$

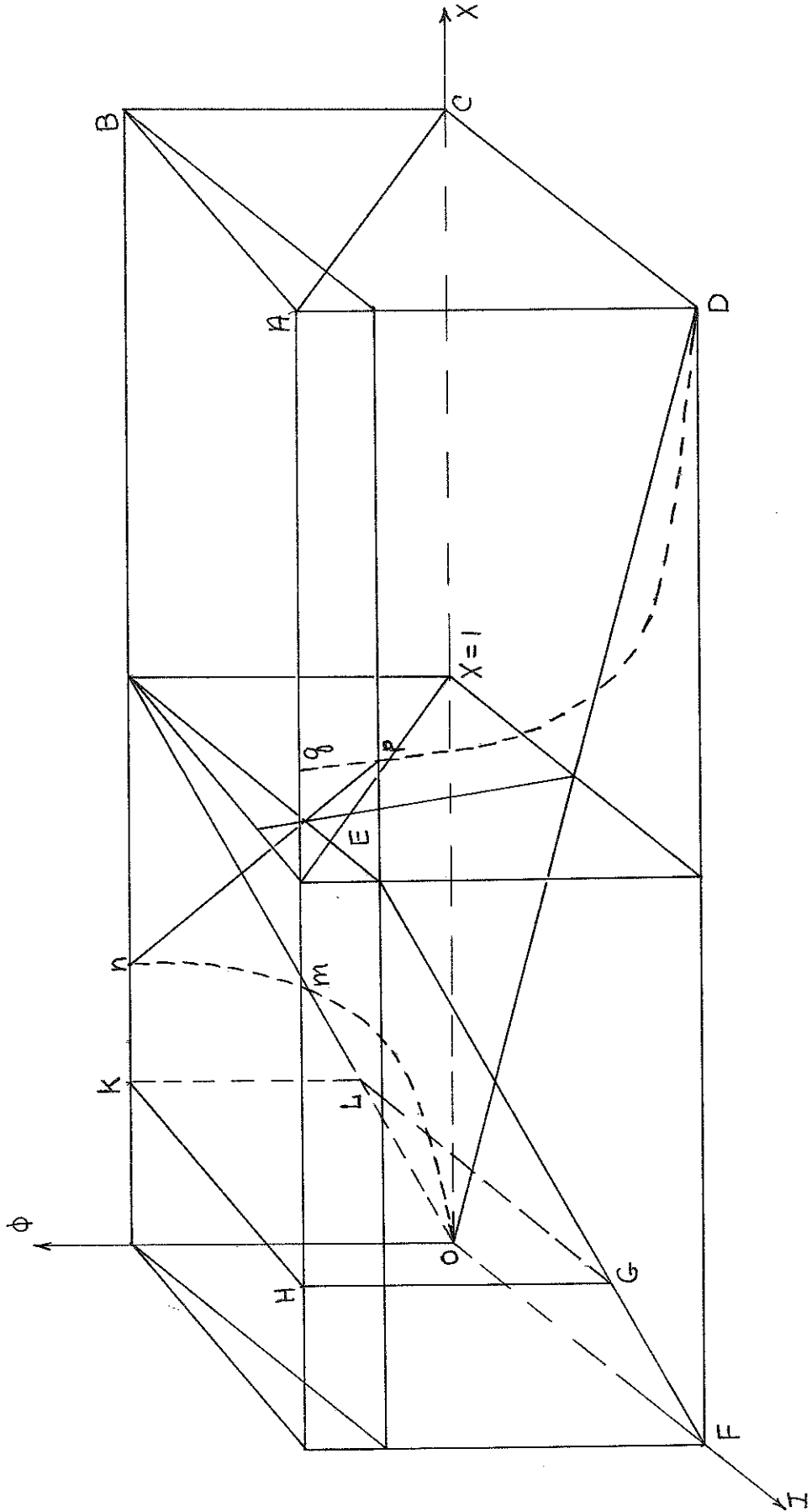


Fig. 74b

FGLOF: Plane E_5 defined by $\phi = aX$ ($a < b$, $X < \frac{d}{\gamma}$)

GHKLG: Plane E_6 defined by $X = d < \frac{d}{\gamma}$

ABKHA: Plane E_7 defined by $\phi - \frac{b-a}{b} I = a$

This surface does enclose the critical point $E(1,1,1)$. All the trajectories cross the surface inwardly because:

$$\frac{dX}{dt} < 0 \quad \text{on} \quad E_1 \quad .$$

To see this clearly consider again the ruled hyperboloid S:

$$a\phi + \beta I - \gamma\phi X - X = 0 \quad . \quad (201a)$$

This hyperboloid has an asymptotic plane at $X = \frac{d}{\gamma} < 1$ and intersects the plane $I = 0$ along the hyperbolic branch Omn and the plane $I = b$ along the hyperbolic branch Dpq . For all points to the right of this surface $dX/dt < 0$ and therefore the same is true for plane E_1 .

$$\frac{dI}{dt} < 0 \quad \text{on} \quad E_2$$

$\frac{d\phi}{dt} = 0$, $\phi = 0$ on E_3 . The trajectories come only arbitrarily close to E_3 because $\phi < 0$ has no physical meaning

$$\frac{dI}{dt} > 0 \quad \text{on} \quad E_4$$

The projection of the vector $(\frac{d\phi}{dt}, \frac{dI}{dt}, \frac{dX}{dt})$ on the normal of plane E_5 is negative provided that a is large enough and a range of X smaller but arbitrarily close to d/γ is considered (see Equation (202)). Consequently, the trajectories cross E_5 inwardly.

$$\frac{dX}{dt} > 0 \quad \text{on} \quad E_6 \quad \text{since it is to the left of the hyperboloid S.}$$

Finally, the projection of the vector $(\frac{d\phi}{dt}, \frac{dI}{dt}, \frac{dX}{dt})$ on the normal of plane

E_7 is:

$$P_n' = \omega_0(1 - X)\phi - \frac{b-a}{b} \lambda_i(\phi - I) < 0 \quad (202a)$$

provided that:

$$\frac{\omega_0(1-X)}{\lambda_i} < \frac{b-a}{b} < 1 \text{ for } X > d \quad (202b)$$

In view of the fact that the range of values α, β, γ under consideration are such that:

$$\frac{\omega_0(\gamma-\alpha)}{\lambda_i \gamma} < 1$$

and d can be taken arbitrarily close to α/γ , conditions (202a,b) are readily satisfied, and the trajectories cross plane E_7 inwardly.

Since the only requirements for the existence of the surface of Figure 74b are that a and $b > a$ be large, it can be immediately concluded that the reactor power is bounded for $c \delta_0 < \gamma_X + c \lambda_i \bar{\tau}_e$.

$$b. \quad \frac{\omega_0(\gamma-\alpha)}{\lambda_i \gamma} > 1 \quad (c \delta_0 > \gamma_X + c \lambda_i \bar{\tau}_e)$$

In this case the reactor power is unbounded because no closed surface surrounding the critical point can be found. In fact, the power diverges either monotonically or in an oscillatory manner.

Monotonic divergence is possible only when the vector $(\frac{d\phi}{dt}, \frac{dI}{dt}, \frac{dX}{dt})$ has positive or zero components asymptotically. Inspection of Equations (191-193) reveals that the only possibility is:

$$X = \text{constant} < 1 \quad \frac{dX}{dt} = 0$$

The solution $X = \text{constant}$ is admissible when the cross-section of the ruled hyperboloid S (Equation 201a) by the plane $X = \text{constant} < 1$ has a slope equal to the asymptotic value of $d\phi/dI$. Consequently:

$$\omega_0(1-X) \phi(\alpha - \gamma X) + \lambda_i(\phi - I) \beta = 0$$

or

$$X^2 - X \left[1 + \frac{\lambda_i}{\omega_0} + \frac{\alpha}{\gamma} \right] + \frac{\lambda_i(\alpha + \beta)}{\omega_0 \gamma} - \frac{\lambda_i}{\omega_0} \frac{X}{\phi} = 0.$$

For $\phi \rightarrow \infty$ this equation admits positive solutions smaller than unity only when:

$$c \delta_0 = 2 \sqrt{c \lambda_i \tau_e \gamma_i} + y_x - c \lambda_i \tau_e$$

When $y_x + c \lambda_i \tau_e < c \delta_0 < 2 \sqrt{c \lambda_i \tau_e \gamma_i} + y_x - c \lambda_i \tau_e$, monotonic divergence is not consistent with the set of Equations (191-193) and all variables ϕ, X, I diverge in an oscillatory manner.

It should be emphasized that all the previous results have been derived without any approximations or tedious computations, as opposed to other approaches to the problem. Furthermore, the existence of bounds does not necessarily imply that the bounds are tolerable. In fact, they may be extremely large.

4. Existence of Periodic Solutions

The question of existence of periodic solutions can be established by means of Poincaré's method of sections and Brouwer fixed point theorem. To this effect investigate the existence of a toroidal region, not containing the critical point, whose bounding surface is intersected inwardly by the trajectories.

Consider first $\alpha > \gamma$. Notice that for all values of $\alpha > \gamma$ one of the characteristic roots of Equation (195) is always real negative (Figure 72a,b), say $-s_1 (s_1 > 0)$. This implies that no closed surface can be found in the neighborhood of the critical point which is crossed outwardly by the trajectories because, for $-s_1 < 0$, there are always two trajectories approaching the critical point. However, a small open ended cylindrical surface around the critical point, intersected outwardly by the trajectories does exist when the former is unstable. To prove this, proceed as follows.

Consider the system of Equations (191-193) and transfer the origin of the phase space to the critical point. Thus find:

$$\frac{d\phi}{dt} = -\omega_0 X - \omega_0 \phi X \quad (203)$$

$$\frac{dX}{dt} = \lambda_x [(\alpha - \gamma)\phi + \beta I - (\gamma + 1)X - \gamma\phi X] \quad (204)$$

$$\frac{dI}{dt} = \lambda_i [\phi - I] \quad (205)$$

If the critical point is unstable, the characteristic roots of the linear approximation are in general:

$$s = -s_1 \quad s = u \pm jv \quad s_1, u, v > 0 \quad (206)$$

A linear transformation of ϕ, I, X into ϕ_1, I_1, X_1 , by means of the modal matrix that corresponds to the characteristic roots reduces the linear approximation to its normal form and Equations (203-205) to:

$$\frac{d\phi_1}{dt} = -s_1\phi_1 + f_1 \quad (207)$$

$$\frac{dX_1}{dt} = uX_1 + vI_1 + f_2 \quad (208)$$

$$\frac{dI_1}{dt} = -vX_1 + uI_1 + f_3 \quad (209)$$

where f_i are second order polynomials in (ϕ_1, I_1, X_1) .

Define the cylinder

$$C = X_1^2 + I_1^2 > 0 \quad (210)$$

Notice that:

$$\frac{dC}{dt} = 2uX_1^2 + 2uI_1^2 + 2(X_1f_2 + I_1f_3) > 0 \quad (211)$$

because $X_1f_2 + I_1f_3$ is of third order in ϕ_1, I_1, X_1 and for sufficiently small values of the latter, the first two terms in the right hand side of Equation (211) dominate. The meaning of Equations (210) and (211) is that there is a small

neighborhood around the critical point in which the cylinder C is intersected outwardly by the trajectories.

The direction of the axis of the cylinder is determined by the directional cosines with respect to ϕ , I, X of the principal axis ϕ_1 , which corresponds to the characteristic root $-s_1$. These cosines are:

$$\begin{aligned} \cos \alpha_\phi &= \frac{[s_1 - \lambda_x(\alpha + \beta)][s_1 - \lambda_i]}{\sqrt{[s_1 - \lambda_x(\alpha + \beta)]^2[s_1 - \lambda_i] + \lambda_i^2[s_1 - \lambda_x(\alpha + \beta)]^2 + \lambda_x^2[\lambda_i - s_1(\alpha - \gamma)]^2}} \\ &= \frac{[s_1 - \lambda_x(\alpha + \beta)][s_1 - \lambda_i]}{D} \quad (212) \\ \cos \alpha_I &= \frac{-\lambda_i[s_1 - \lambda_x(\alpha + \beta)]}{D} \\ \cos \alpha_X &= \frac{\lambda_x[\lambda_i - s_1(\alpha - \gamma)]}{D} \end{aligned}$$

From Figure 72b it is evident that when the critical point is unstable:

$$\lambda_x(\alpha + \beta) < s_1 < \frac{\lambda_i}{\alpha - \gamma} \quad (213)$$

Therefore, the principal axis ϕ_1 and the cylinder are oriented as shown in Figure 75.

Next, extend the cylinder by two funnel-like surfaces beyond the equilibrium point as shown in Figure 75. The funnels consist of three mutually intersecting planes:

EPQE and EP₁Q₁E: Plane E₉ defined by $\phi = 1$

EQRE and EQ₁R₁E: Plane E₁₀ defined by $bI + X = b + 1$ ($b > 0$)

ERRE and EP₁R₁E: Plane E₁₁ defined by $-c\phi + X = 1 - c$ ($c > 0$)

Require that the slopes of planes E₁₀ and E₁₁ be such that the principal direction ϕ_1 is inside the funnels, a condition that is easily fulfilled.

Now, all trajectories cross the funnels outwardly because:

$$\frac{d\phi}{dt} \geq 0 \quad \text{for } X \lesssim 1 \quad \text{and} \quad \frac{dI}{dt} = 0 \quad \text{on } E_9$$

$$\frac{dX}{dt} \geq 0 \quad \text{and} \quad \frac{d\phi}{dt} \geq 0 \quad \text{for } X \lesssim 1 \quad \text{on } E_{10}$$

$$\frac{dX}{dt} \geq 0 \quad \text{and} \quad \frac{d\phi}{dt} \geq 0 \quad \text{for } X \lesssim 1 \quad \text{on } E_{11}, \text{ however the slope of } E_{11} \text{ can be}$$

decreased as in the case of E_9 (Figure 74a) to have the trajectories intersecting outwardly. This does not conflict with the requirement of the directional cosines.

Superposition of the surfaces shown in Figures 74a and 75 results in the critical point free toroidal region that was sought, if the volumes of Figure 74a falling into the funnels and the cylinder as well as the origin are excluded. The exclusion of the origin is straightforward because one of the characteristic roots there is positive. Simple review of the behavior of the trajectories on the planes E_1 through E_{11} and the cylinder C immediately reveals that the bounding surface of the torus is crossed by the trajectories inwardly everywhere. Therefore the topological torus constitutes a trajectory trap.

A typical cross section of the toroidal region by the plane $I = 1$ is shown in Figure 76. Two simply connected sections S_1 and S_2 result. Observe that any trajectory that is originally in the torus is trapped there. Furthermore, it intersects the section S_1 towards the plane of the figure, along the positive I direction, ($dI/dt > 0$ on S_1) and the section S_2 away from the plane of the figure along the negative I direction ($dI/dt < 0$ on S_2). This implies that a trajectory starting from a point $(\phi_0, I_0 = 1, X_0)$ on S_1 moves away and cannot return to S_1 along the negative I direction, etc. Similar considerations of the signs of the vector field $(\frac{d\phi}{dt}, \frac{dI}{dt}, \frac{dX}{dt})$ in the various regions of the torus lead to the overall conclusion that the trajectories circulate around the torus. Therefore, the simply connected section S_1 is topologically mapped into itself by a continuous vector field which circulates in a region free of critical points.

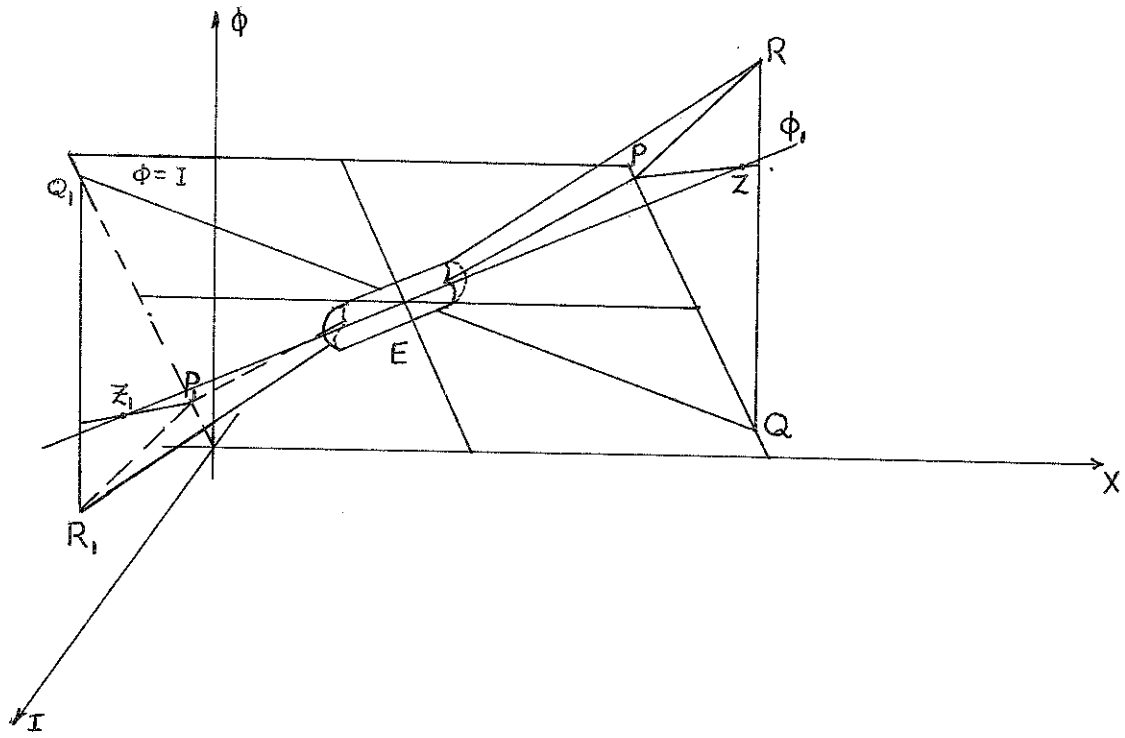


Fig. 75

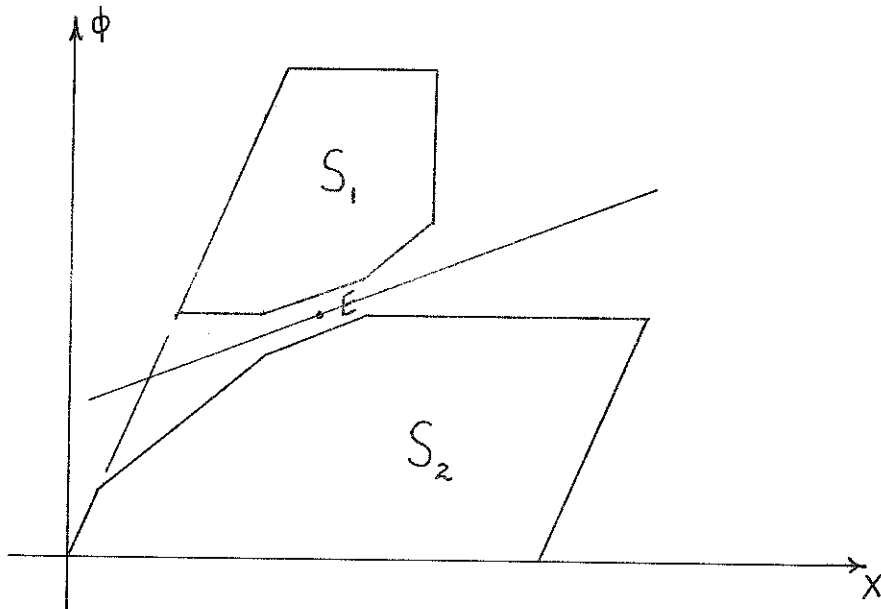


Fig. 76

According to Brouwer's fixed point theorem, the mapping possesses a fixed point, or, the reactor admits periodic solutions.

Similar arguments apply when $y_x < c \rho_0 < y_x + c \lambda_1 \tau_e$ and it is concluded that the reactor admits periodic solutions.

In summary, when $c \rho_0 < y_x + c \lambda_1 \tau_e$ and the critical point is unstable, the xenon controlled reactor oscillates. The oscillations may be sinusoidal or of the relaxation type as discussed in (12). It should be pointed out that the existence of the solid torus is not adequate topology to guarantee either the uniqueness or the stability of the periodic solutions. Such questions can be examined by means of the general theorem of existence of periodic solutions which is beyond the scope of this communication. Also it should be noted that the existence of a negative characteristic root implies that some exceptional trajectories may indeed converge to the critical point even when the latter is unstable.

On the other hand, when $y_x + c \lambda_1 \tau_e < c \rho_0$, no closed surface surrounds the critical point and the reactor power is always unbounded (Figure 73). The unboundedness manifests itself either by diverging oscillations or by a monotonically increasing power with a bounded xenon concentration as already discussed in section B.3.

C. DYNAMICS OF REACTORS WITH TWO TEMPERATURE COEFFICIENTS

1. The Reactor Model

Two-region reactors with two temperature coefficients of reactivity have been already analyzed (9) by means of Liapunov's second method. However the requirement of existence of a Liapunov function may be over restricting. Here the problem is treated in all generality.

The reactor model is assumed independent of spatial coordinates and delayed neutrons are neglected. The reactor dynamics, with respect to step changes of reactivity, are describable by:

$$\frac{d\phi}{dt} = \rho_1 \phi \quad (214)$$

$$\epsilon_1 \frac{dT_1'}{dt} = \eta_1(\phi - \phi_0) - h(T_1' - T_2') \quad (215)$$

$$\epsilon_2 \frac{dT_2'}{dt} = \eta_2(\phi - \phi_0) + h(T_1' - T_2') - wT_2' \quad (216)$$

$$\rho_1 = \rho_0 + r_1 T_1' + r_2 T_2' \quad (217)$$

where:

ϵ_i = heat capacity of i^{th} region

h = overall heat transfer coefficient between regions (1) and (2)

η_i = fractional power delivered to i^{th} region ($\eta_1 + \eta_2 = 1$)

r_1 = temperature coefficient of reactivity over neutron lifetime

ρ_0 = step input over neutron lifetime

T_i' = average temperature increment of i^{th} region

wT_2' = power removal

ϕ = power

ϕ_0 = steady state power before step ρ_0 is applied

A simple change of variable:

$$T_i = T_1' + b_i T_2' \quad (218)$$

where:

$$b_{1,2} = \frac{\frac{h}{\epsilon_1} - \frac{h}{\epsilon_2} - w \pm \sqrt{\left[\frac{h}{\epsilon_1} - \frac{h}{\epsilon_2} - w\right]^2 + 4 \frac{h^2}{\epsilon_1 \epsilon_2}}}{2 \frac{h}{\epsilon_2}} \quad (219)$$

reduces the system of Equations (214 - 217) into the form:

$$\frac{d\phi}{dt} = \rho_1 \phi \quad (220)$$

$$\frac{dT_1}{dt} = a_1(\phi - \phi_0) - g_1 T_1 \quad (221)$$

$$\frac{dT_2}{dt} = a_2(\phi - \phi_0) - g_2 T_2 \quad (222)$$

$$\rho_1 = \rho_0 + r_1 T_1 + r_2 T_2 \quad (223)$$

with:

$$g_i = \frac{h}{\varepsilon_1} - b_i \frac{h}{\varepsilon_2} \quad a_i = \frac{\eta_1}{\varepsilon_1} + b_i \frac{\eta_2}{\varepsilon_2}$$

$$r_1 = \frac{r_2' - r_1' b_2}{b_1 - b_2} \quad r_2 = \frac{r_1' b_1 - r_2'}{b_1 - b_2}$$

The coefficients g_i are always positive. The coefficients a_i can also be assumed positive because, if a_i were not, a simple change of variable $T_i \rightarrow -T_i$ would result in a system with positive coefficients.

The system of Equations (220 - 223) admits a critical point:

$$\phi_\infty = \phi_0 - \frac{\rho_0 g_1 g_2}{r_1 a_1 g_2 + r_2 a_2 g_1} \quad \rho = 0 \quad (224)$$

$$T_{1\infty} = \frac{a_1(\phi_\infty - \phi_0)}{g_1} \quad T_{2\infty} = \frac{a_2(\phi_\infty - \phi_0)}{g_2}$$

If the variables are measured in terms of their equilibrium values, Equations (220 - 223) reduce to:

$$\frac{d\phi}{dt} = \rho \phi \quad (225)$$

$$\frac{dT_1}{dt} = g_1(\phi - T_1) + g_1 \frac{\phi_0}{\phi_\infty - \phi_0} (\phi - 1) \quad (226)$$

$$\frac{d\Gamma_2}{dt} = g_2(\phi - \Gamma_2) + g_2 \frac{\phi_0}{\phi_\infty - \phi_0} (\phi - 1) \quad (227)$$

$$\rho = \left[\frac{r_1 a_1}{g_1} (\Gamma_1 - 1) + \frac{r_2 a_2}{g_2} (\Gamma_2 - 1) \right] (\phi_\infty - \phi_0) \quad (228)$$

For the purposes of the subsequent discussion ϕ_0 is assumed equal to zero. This is done for mathematical expediency and does not involve any loss of generality.

2. Stability of the Critical Point

Proceeding as in the case of the xenon controlled reactor, it is found that the characteristic equation of the linear approximation of Equations (225 - 228) is ($\phi_0 = 0$):

$$s(s+g_1)(s+g_2) - \phi_\infty (r_1 a_1 + r_2 a_2) \left[s + \frac{r_1 a_1 g_2 + r_2 a_2 g_1}{r_1 a_1 + r_2 a_2} \right] = 0 \quad (229)$$

The root loci of this equation are shown in Figure 77a, b, c, which indicates that:

a. The critical point is stable when

$$r_1 a_1 g_1 + r_2 a_2 g_2 < 0 \qquad r_1 a_1 g_2 + r_2 a_2 g_1 < 0 \quad (230)$$

b. The critical point is conditionally stable when

$$\begin{aligned} r_1 a_1 g_1 + r_2 a_2 g_2 > 0 & \qquad r_1 a_1 g_2 + r_2 a_2 g_1 < 0 \\ g_1 > g_2 & \qquad \phi_\infty < \frac{g_1 g_2 (g_1 + g_2)}{r_1 a_1 g_1 + r_2 a_2 g_2} \end{aligned} \quad (231)$$

$$\rho_0 < - \frac{(g_1 + g_2)(r_1 a_1 g_2 + r_2 a_2 g_1)}{r_1 a_1 g_1 + r_2 a_2 g_2}$$

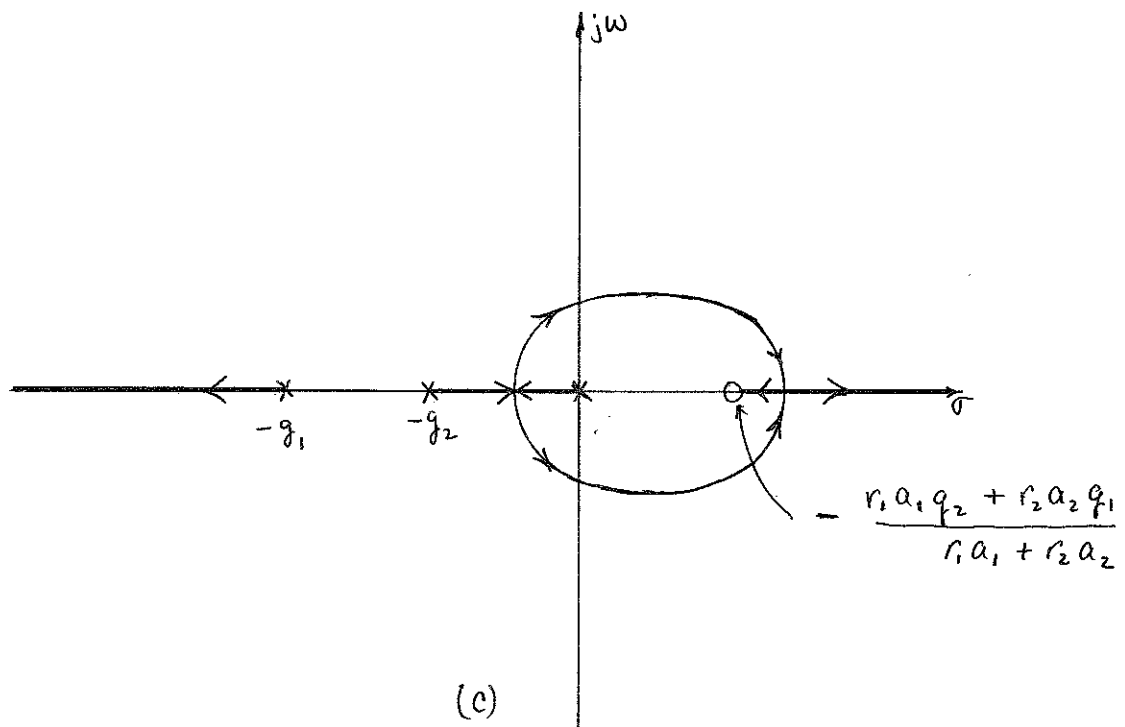
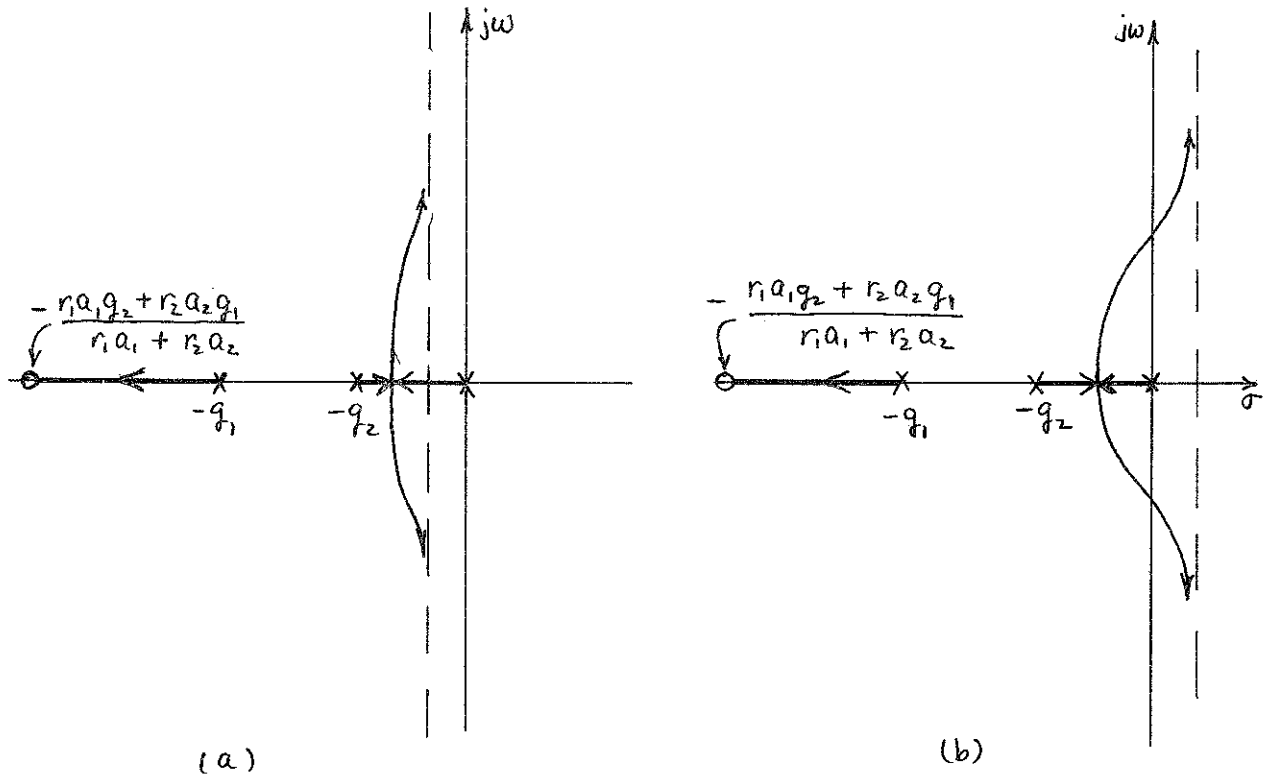


Fig. 77

c. No critical point exists when:

$$r_1^a g_2 + r_2^a g_1 > 0 \quad (232)$$

In summary, when r_1, r_2 are positive, the linear approximation admits unstable solutions, when r_1, r_2 are negative it admits stable solutions and when $r_1 > 0, r_2 < 0$: if conditions (230) are satisfied, the linear approximation admits stable solutions for all values of ϕ_∞ (viz: ρ_0), while if conditions (231) are true the linear approximation admits stable solutions only for a limited range of values ϕ_∞ . These results have also been presented in (19). In an actual case the previous conditions can of course be expressed in terms of the temperature coefficients of reactivity, etc.

3. Boundedness and Stability in the Large

The boundedness and stability of some solutions have already been investigated. More precisely, when $r_1, r_2 < 0$ or conditions (230) are satisfied the reactor is asymptotically stable (9). However, nothing has been reported on boundedness and stability when conditions (231) are satisfied. In this case geometric theory is very helpful.

Assume $r_1 > 0, r_2 < 0$. Consider the phase space (ϕ, T_1, T_2) and the arbitrary point $A(b, b, b)$ with $b > 1$, shown in Figure 78. Define a closed region by the surfaces:

- ABCD: Plane $E_1 // T_1 = 0$ through point A
- ABFGNA: Plane $E_2 // \phi = 0$ through point A
- ADMLNA: Plane $E_3 // T_1 = 0$ through point A
- BCOHFB: Plane E_4 defined by $T_1 = 0$
- KLMOHK: Plane E_5 defined by $T_2 = 0$
- CDMOC: Plane E_6 defined by $\phi = 0$
- KLNGK: Ruled paraboloid E_7 defined by:

$$L_1 = b - 1 - \ln b = \left[\phi - 1 - \ln \phi - \frac{r_1^a g_1}{2g_1^2} (T_1 - 1)^2 - \frac{r_2^a g_2}{2g_2^2} (T_2 - 1)^2 \right] \quad (233)$$

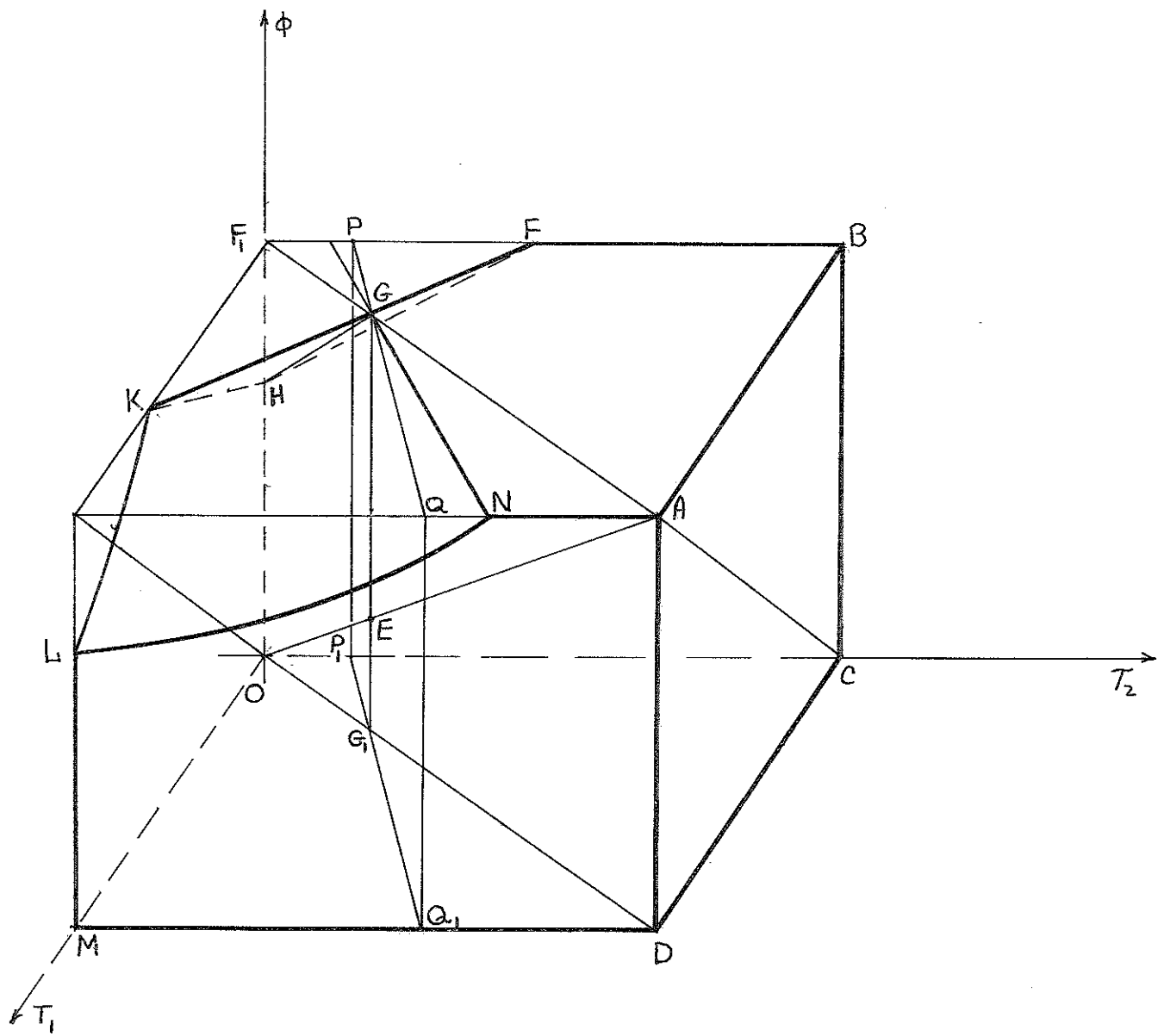


Fig. 78

FGKHF: Plane E_8 to be determined.

Notice that this region does enclose the critical point and is intersected by the trajectories inwardly because:

$$\frac{dT_2}{dt} < 0 \quad \text{on } E_1$$

$\frac{d\phi}{dt} < 0$ on E_2 because the boundaries GN and GF of this plane are to the right of the plane PGQQ₁G₁P₁P defined by:

$$\frac{r_1 a_1}{g_1} (T_1 - 1) + \frac{r_2 a_2}{g_2} (T_2 - 1) = 0 \quad (234)$$

$$\frac{dT_1}{dt} < 0 \quad \text{on } E_3$$

$$\frac{dT_1}{dt} > 0 \quad \text{on } E_4$$

$$\frac{dT_2}{dt} > 0 \quad \text{on } E_5$$

$\phi = 0, \frac{d\phi}{dt} = 0$ on E_6 . The trajectories come only arbitrarily close to $\phi = 0_+$, because $\phi < 0$ has no physical meaning.

The ruled paraboloid E_7 is intersected inwardly if the following conditions are satisfied:

a. L_1 be positive. Since $\phi - 1 - \ln \phi$ is always positive, this is fulfilled if:

$$\left| \frac{T_1 - 1}{T_2 - 1} \right| < \sqrt{\frac{r_2 a_2 g_1^2}{r_1 a_1 g_2^2}} \quad (235)$$

b. dL_1/dt be negative. It can be easily shown that:

$$\frac{dL_1}{dt} = \frac{r_1 a_1}{g_1} (T_1 - 1)^2 + \frac{r_2 a_2}{g_2} (T_2 - 1)^2 \quad (236)$$

This is negative if:

$$\left| \frac{T_1 - 1}{T_2 - 1} \right| < \sqrt{\frac{r_2 a_2 g_1}{r_1 a_1 g_2}} \quad (237)$$

Notice that in view of conditions (231):

$$-\frac{r_2 a_2 g_1}{r_1 a_1 g_2} > \sqrt{-\frac{r_2 a_2 g_1^2}{r_1 a_1 g_2^2}} > \sqrt{-\frac{r_2 a_2 g_1}{r_1 a_1 g_2}} \quad (238)$$

provided that $r_1 a_1 + r_2 a_2 < 0$. The intersections of the paraboloid with the plane $\phi = b$ are the lines:

$$\left| \frac{T_1 - 1}{T_2 - 1} \right| = \sqrt{\frac{r_2 a_2 g_1^2}{r_1 a_1 g_2^2}} \quad (239)$$

Consequently conditions (235) and (237) are readily satisfied and furthermore GN and GF do indeed lie to the right of plane PGQQ₁G₁P₁P (see inequalities (238)).

Finally, for b sufficiently large, the slope of plane E_ϕ can be adjusted so that the trajectories cross it inwardly. Indeed, the directional cosine with respect to ϕ , of the vector field $(d\phi/dt, dT_1/dt, dT_2/dt)$ for large b is:

$$\cos \alpha_\phi \cong \frac{\rho}{\sqrt{\rho^2 + q_1^2 + q_2^2}} \quad (240)$$

and has a maximum value when ρ is evaluated at the point

$$K(\phi = b, T_1 = 1 + \sqrt{-r_2 a_2 g_1^2 / r_1 a_1 g_2}, T_2 = 0).$$

If the plane E_ϕ forms an angle with ^{the} ϕ -axis smaller than the one corresponding to $\cos \alpha_\phi \max$, then E_ϕ is crossed everywhere inwardly by the trajectories.

This completes the determination of the closed region. Taking b as large as desired, all trajectories can be included in the region and cannot escape from it. That is, all solutions are bounded.

In conclusion, when $r_1 > 0$, $r_2 < 0$, and $r_1 a_1 g_2 + r_2 a_2 g_1 < 0$, $r_1 a_1 + r_2 a_2 < 0$ the solutions of the system (225 - 228) (with $\phi_0 = 0$) are bounded, regardless of whether the critical point is stable or not. In fact, when the critical point is stable, the two region reactor is asymptotically stable and when the critical point is unstable periodic solutions may exist. The latter problem is discussed in the next section.

When $r_1 a_1 + r_2 a_2 > 0$ it can be easily shown that no closed surface can be found which is intersected everywhere inwardly by the trajectories. Consequently, the solutions are unbounded. It is interesting to note that both in the two region reactor and in the xenon controlled reactor, when the critical point is unstable and the characteristic loci are as shown in Figures 72c and 77c, the reactors are unstable in the large. This problem seems to be related to the structural stability of third order nonlinear systems and will be discussed in a future communication.

4. Existence of Periodic Solutions

The existence of periodic solutions, when the reactor is bounded in the large, is again investigated by means of Poincaré's method of sections.

Notice again that one of the characteristic roots is always real and negative, say $-s_1$ ($s_1 > 0$). Clearly, the existence of a toroidal region free of critical points is to be investigated.

Consider the system of Equations (225 - 228) and transfer the origin to the critical point. Thus find:

$$\frac{d\phi}{dt} = \phi_\infty \left[\frac{r_1 a_1}{g_1} T_1 + \frac{r_2 a_2}{g_2} T_2 \right] (\phi + 1) \quad (241)$$

$$\frac{dT_1}{dt} = g_1(\phi - T_1) \quad (242)$$

$$\frac{dT_2}{dt} = g_2(\phi - T_2) \quad (243)$$

When the critical point is unstable, the characteristic roots are

$$s = -s_1 \quad s = u \pm jv \quad s_1, u, v > 0 \quad (244)$$

The directional cosines of the principal direction corresponding to the characteristic root $-s_1$ are:

$$\begin{aligned} \cos \alpha_\phi &= \frac{(s_1 - g_1)(s_1 - g_2)}{\sqrt{(s_1 - g_1)^2(s_1 - g_2)^2 + g_1^2(s_1 - g_2)^2 + g_2^2(s_1 - g_1)^2}} \\ &= \frac{(s_1 - g_1)(s_1 - g_2)}{D} \end{aligned} \quad (245)$$

$$\cos \alpha_{T_1} = - \frac{g_1(s_1 - g_2)}{D}$$

$$\cos \alpha_{T_2} = - \frac{g_2(s_1 - g_1)}{D}$$

Notice that since $g_1 > g_2$, $(r_2 a_2 g_1 + r_2 a_2 g_1) / (r_1 a_1 + r_2 a_2) > s_1 > g_1 + g_2$, this direction, drawn from the critical point E, lies in the region:

$$\frac{T_1}{T_2} > - \frac{r_2 a_2 g_1}{r_1 a_1 g_2} \quad \phi > 0 \quad (246)$$

and is shown in Figure 79 along with some typical planar cross sections of the phase space.

Plane ① // to $T_1 = 0$

Plane ② // to $T_2 = 0$

Plane ③ // to $\phi = 0$

Plane ④ // to $\frac{r_1 a_1}{g_1} (T_1 - 1) + \frac{r_2 a_2}{g_2} (T_2 - 1) = 0$

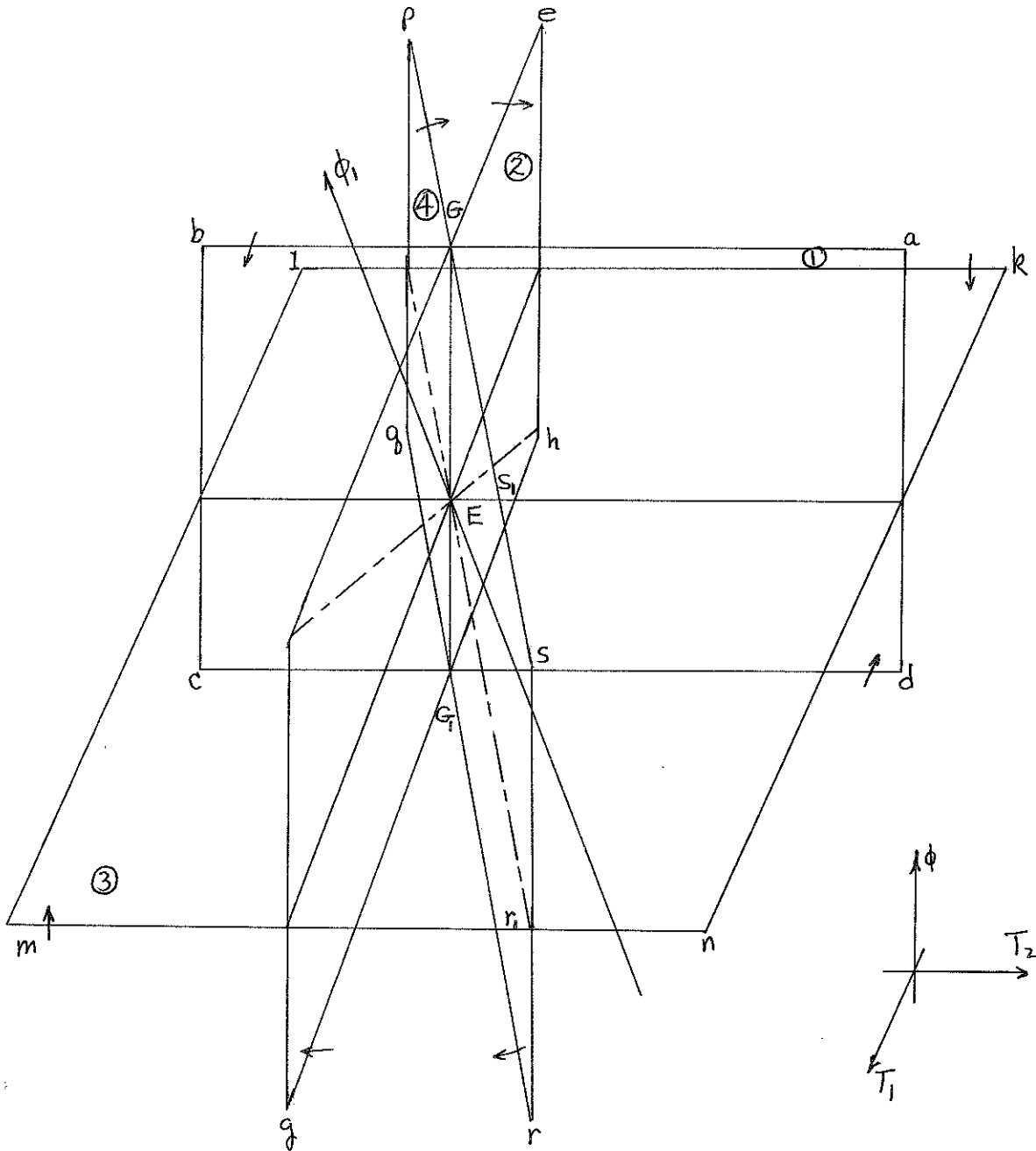


Fig. 79

These cross-sections indicate the general pattern of direction of the vector field at various regions of the phase space. On each planar cross-section there is a dividing line (dotted line) which divides the plane into two regions intersected in opposite directions by the trajectories. The direction of crossing is indicated in the figure. For example cross-section ① (plane abcd) is crossed along the positive T_1 direction over the region abc and along the negative T_1 direction along the region acd.

Simple inspection of the directional pattern reveals that no surface can be found which encircles the principal direction ϕ_1 and is crossed outwardly by the trajectories. Consequently, no toroidal surface free of critical points can be defined and no periodic solutions exist. The boundedness of the solutions implies that they eventually converge to the critical point along the principal direction ϕ_1 in spite of the fact that the critical point is unstable in the small.

In summary, when $r_1 a_1 + r_2 a_2 < 0$ and the critical point is unstable, the reactor variables are bounded but no periodic solutions exist. When $r_1 a_1 + r_2 a_2 > 0$ the solutions are unbounded. The results for the linear and nonlinear behavior of the reactor both for small and large variations are shown in Figure 80.

It should be pointed out again that boundedness of solutions does not imply tolerable solutions.

D. CONCLUSIONS

A very brief review of the geometric theory of differential equations has been presented in an attempt to clarify the relationship that exists between the "small" and "large" signal behavior of nuclear reactor systems. The theory is illustrated by two specific examples, the xenon controlled reactor and a two-region reactor with two temperature coefficients of reactivity.

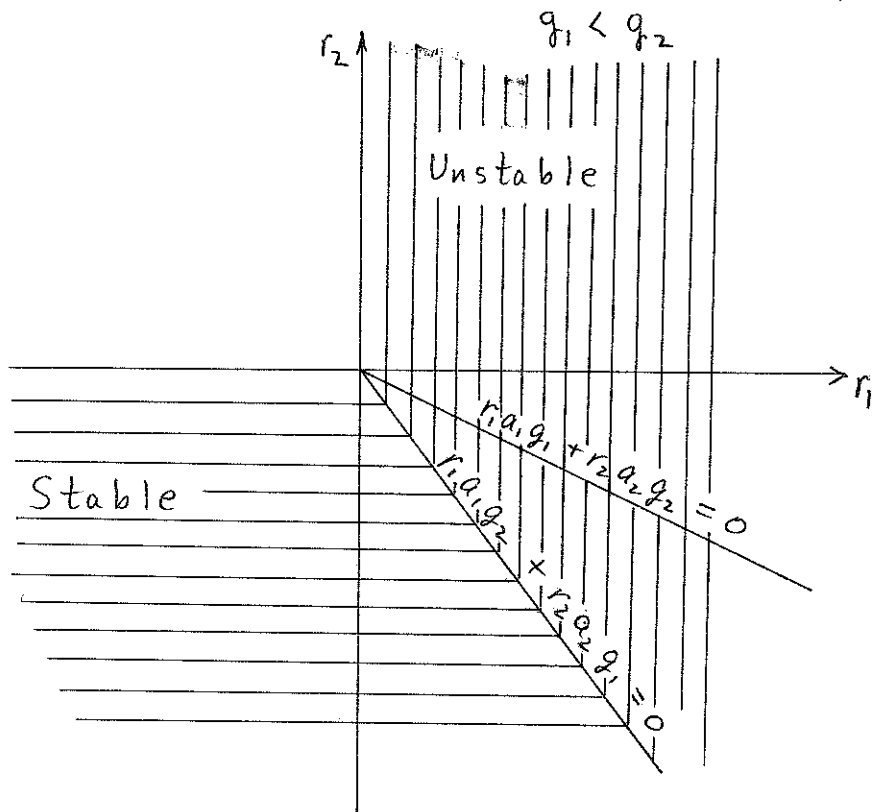
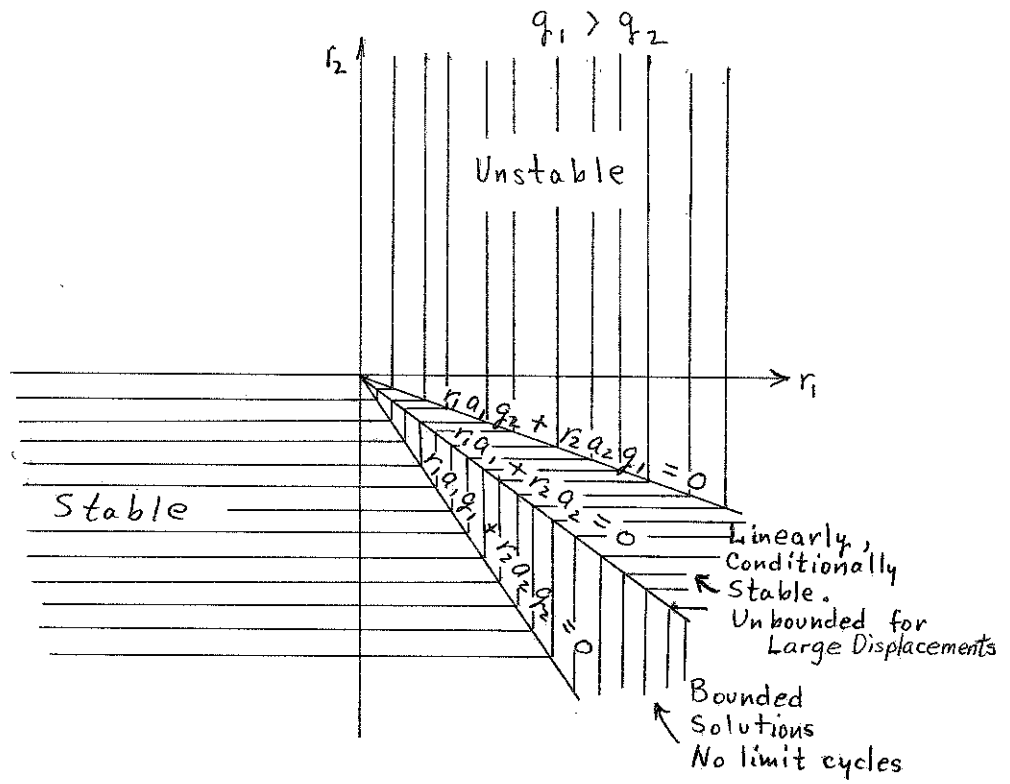


Fig. 80.

It is emphasized that the "small" signal or linear approximation is useful when used within a definite range of amplitude variations of the dependent variables, the magnitude of the range being defined by the nonlinear terms.

It is shown that the boundedness of the "large" signal behavior is in general dependent on the nonlinear terms while the stability of the solutions as well as the existence of periodic oscillations depend both on the linear approximation and the nonlinear terms.

In summary, the linear or transfer function approach to reactor dynamics does not contain enough information to predict the performance of reactors when large power level changes are involved.

Both for the xenon controlled reactor and the reactor with two temperature coefficients of reactivity, conditions for boundedness and stability and existence of periodic solutions are derived by simple geometric considerations and without any approximations or lengthy computations. The entire range of characteristic parameters and pertinent dependent variables is covered.

The analysis of these two reactor types clearly indicates the usefulness and elegance of the geometric theory in the field of nuclear reactor dynamics.

One way to perform this transformation is through orthogonal expansions. We consider a complete set of orthonormal functions and denote each member of the set by $\phi_i(t)$. Orthonormality means that if the range of orthonormality is from 0 to T, then

$$\int_0^T \phi_n(t) \phi_m(t) dt = \delta_{nm} = \begin{cases} 0 & ; n \neq m \\ 1 & ; n = m \end{cases}$$

where δ_{nm} is the Kronecker delta. Completeness means that if we expand a function in terms of this set of orthonormal functions, as more terms are used the error is decreased in the least square error sense. As was said before, we want to express a function of the past time in terms of a function of the present time. Then we can write

$$x(t - \tau) = \sum_i^{\infty} u_i(t) \phi_i(\tau) \quad , \quad (248)$$

where $u_i(t)$ is a function of the present time. Thus we have written the function $x(t - \tau)$ as a series, each term of which contains one member of the orthonormal set and a coefficient. The coefficient $u_i(t)$ is a function of the present time, and if we want this expansion to be complete the coefficients have to be the Fourier coefficients which are given by

$$u_i(t) = \int_0^T x(t - \tau) \phi_i(\tau) d\tau \quad . \quad (249)$$

Now what have we accomplished? First, we have succeeded in determining a set of coefficients $u_i(t)$ which are functions of the present time. Second, if we do not want to take an infinite number of terms in the series, then when we truncate the series we have a least squares approximation and can

estimate the error due to truncating the series by

$$\varepsilon = \overline{f^2} - \sum \overline{u_i^2(t)}$$

Third, instead of using values of $x(t - \tau)$ we can equivalently use the $u_i(t)$'s. These functions are equivalent to $x(t - \tau)$ in the same sense that the Fourier coefficients are equivalent to a function or in the sense that a logarithm is equivalent to a number, etc. Finally, we realize that we can reproduce the coefficients $u_i(t)$ experimentally. As an example, consider the following: Suppose we have a linear system whose system function is one of the members of the complete orthonormal set. Thus we can build up a linear system whose system function $H(t)$ is a member of the complete set. Suppose we excite this system by an input $x(t)$. What will be the output? We can visualize the system as shown in the diagram (Figure 81). The output is the coefficient $u_i(t)$. This

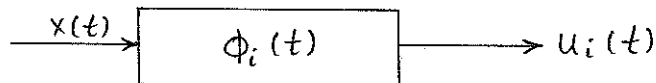


Fig. 81

is true because the convolution of the input with the system function is the output, and from Equation (249) this is the coefficient $u_i(t)$.

To illustrate that it is possible to reproduce the coefficients $u_i(t)$ experimentally, consider the orthonormal set to be the Laguerre polynomials, given by the equations

$$L_n^{(\alpha)}(x) = (-1)^n x^{-\alpha} e^x \frac{d^n}{dx^n} (x^{\alpha+n} e^{-x}) ; (n=0, 1, \dots)$$

for each $\alpha > -1$. In Figure 82 diagrams (a), (b), and (c) are electrical networks whose system functions are the first, second and third order Laguerre polynomials, respectively.

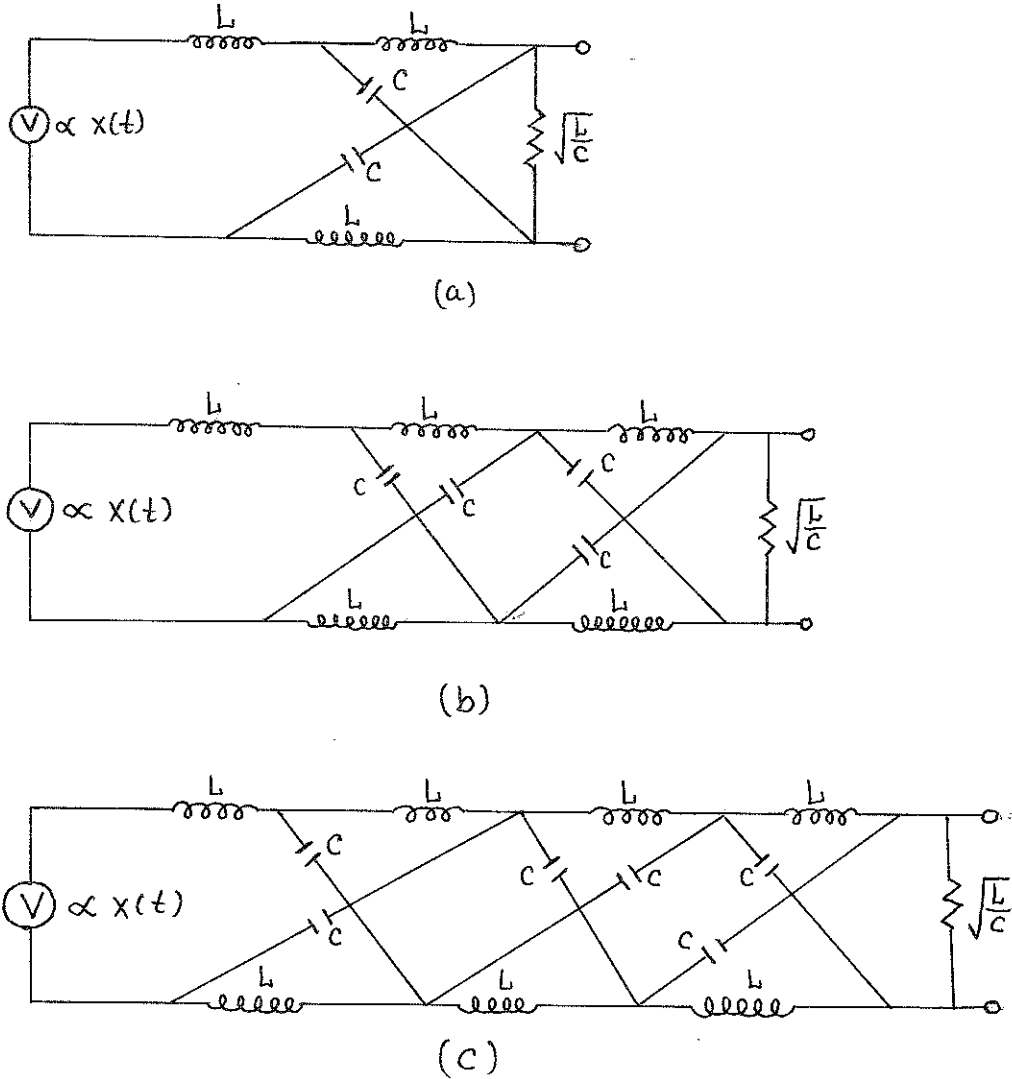


Fig. 82

In Figure 82, if V is proportional to the input $x(t)$, the output of the circuit (b) is the second coefficient $u_2(t)$. The output of the circuit (c) is the third coefficient $u_3(t)$. Thus, if we built up many circuits in this manner, we could reproduce all the coefficients $u_1(t)$.

We see then that it is possible to make a transformation from $x(t - \tau)$ to $u_1(t)$ and this transforms Equation (247) to a relationship of the form

$$y(t) = F [u_1(t), u_2(t), \dots, u_n(t) \dots] \quad (250)$$

Thus the present values of the output depend on an infinite number of

coefficients of functions of the present time. We can represent this situation by the diagram of Figure 83.

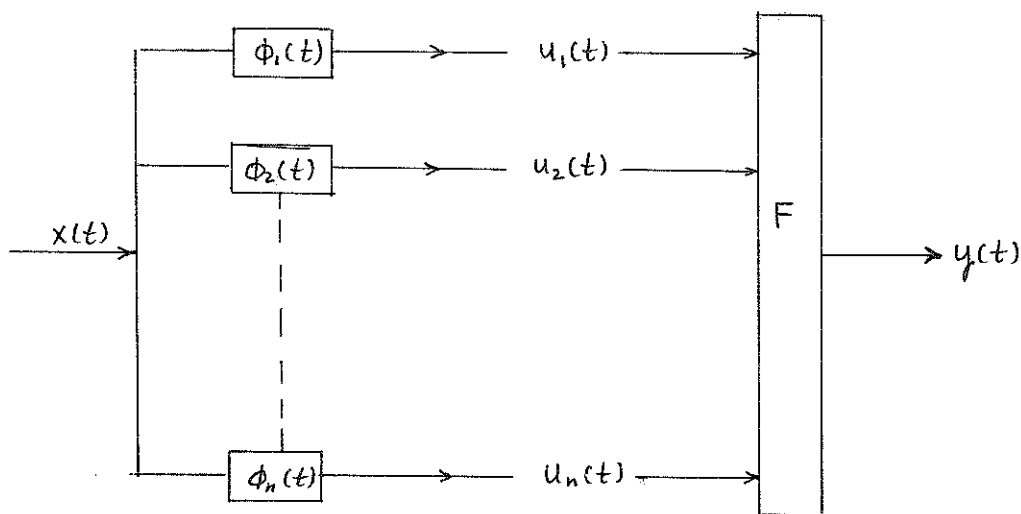


Fig. 83

The diagram of Figure 83 is interpreted in the following manner: The input $x(t)$ is fed into an infinite number of linear systems with system functions $\phi_1(t), \phi_2(t), \dots, \phi_n(t)$ to produce the $u_1(t)$'s. This part of the diagram represents a system with memory. Then the functional F takes the $u_1(t)$'s and combines them in some nonlinear manner to yield the output $y(t)$. This means then that even though we started with a nonlinear system, the response of which depended upon the past time, we now have an equivalent system in which we have separated the nonlinearity and memory.

Now let's forget for the present time about physical systems and consider only the relationship given by Equation (250). Mathematically we can consider the output $y(t)$ as being a function of many variables; i.e., we can consider the $u_1(t)$'s as variables. Then, since we know that many times it is helpful to express functions as power series, and since the theory of power series is well established, we might consider the function $y(t)$ as the Taylor series

expansion

$$y(t) = a_0 + \sum_i a_i u_i(t) + \sum_{i,j} a_{i,j} u_i(t) u_j(t) + \dots + \sum_{i,j,\dots,n} a_{i,\dots,n} u_i(t) \dots u_n(t). \quad (251)$$

For the present, let's forget about the restrictions involved and assume that we know what conditions must exist for this expansion to be valid. The $u_i(t)$'s are functions of the input only and do not depend upon the system. The a_i 's characterize the system. Thus, for different physical systems the a_i 's are different. The output $y(t)$ is then a function of quantities which depend only on the input and quantities which depend only on the system. If we could determine the a_i 's for a given physical system then we could describe the system in terms of a model.

Since we have already required that

$$u_i(t) = \int_0^T x(t-\tau) \phi_i(\tau) d\tau, \quad ,$$

then by writing (from Equation (251))

$$y(t) = h_0 + \int_0^T h_1(\tau) x(t-\tau) d\tau + \int_0^T \int_0^T h_2(\tau_1, \tau_2) x(t-\tau_1) x(t-\tau_2) d\tau_1 d\tau_2 + \dots + \int_0^T \dots \int_0^T h_n(\tau_1, \dots, \tau_n) x(t-\tau_1) \dots x(t-\tau_n) d\tau_1 \dots d\tau_n \quad (252)$$

we have

$$\begin{aligned} h_0 &= a_0 \\ h_1(\tau) &= \sum_i a_i \phi_i(\tau) \\ h_2(\tau_1, \tau_2) &= \sum_{i,j} a_{i,j} \phi_i(\tau_1) \phi_j(\tau_2) \\ &\vdots \\ h_n(\tau_1, \dots, \tau_n) &= \sum_{i,\dots,n} a_{i,\dots,n} \phi_i(\tau_1) \dots \phi_n(\tau_n) \end{aligned} \quad (253)$$

Thus we have expressed the output $y(t)$ as a power series in terms of the input $x(t - \tau_i)$. The system is now represented by the kernels $h_n(\tau_1, \dots, \tau_n)$ which are described in terms of the orthonormal set $\phi_i(\tau)$. Although we have been talking in terms of an infinite number of terms, unless abrupt changes are involved (such as square waves, etc.) it may be possible to use only a few terms in the expansion.

The method described above suggests one approach to the problem of analysis of nonlinear systems. If the integrals of the expansion are orthogonal to each other, it may be possible to determine the individual kernels by use of the orthogonality. Thus it would be worthwhile to see if an expansion such as Equation (252) could be made in terms of orthogonal functionals. Before we can investigate further the feasibility of an expansion in terms of orthogonal functionals, the problem of what kinds of inputs can be used must be answered.

To be able to use a single input, it should be of such a nature as to display all the properties that any other conceivable input would give. Thus we need to use an input which can reproduce any other possible input.

This problem has been investigated and the solution is that the most desirable signal is a Gaussian white noise input. The reasons for this result are as follows:

Gaussian distributed means that the function is not predictable in time but that there is a finite probability that x will lie between x_i and $x_i + dx_i$. The distribution of these probabilities is the normal probability distribution. Wiener has shown that there exists a finite probability that a Gaussian signal will represent any well behaved curve, provided the Lebesgue integral

$$L^2 = \int_{-\infty}^{\infty} x^2 dt < M$$

exists.

White noise means that if we take $x(t_1)$ for a Gaussian distribution and $x(t_2)$ for another Gaussian distribution, there is no correlation between the two values. Thus the crosscorrelation gives a delta function at $t_1 = t_2$ and in the frequency domain the transformation is frequency independent. (Note: Because of the large errors introduced if the Gaussian signal is not considered over infinite time, from a practical standpoint we cannot use the Gaussian input. However, as we will discuss later, efforts are being made to adapt the principles of this theory for use with other signals.)

Now that we know that theoretically the most desirable input signal should be a Gaussian white noise signal, it remains to show that if we use this signal we can expand the nonlinear functional in terms of orthonormal functionals which are functions of the input. Thus, we want to write

$$y(t) = G_0 + G_1(K_1, x, t) + G_2(K_2, x, t) + \dots \quad (254)$$

where each G_n is a functional which depends on a kernel $K_n(\tau_1, \tau_2, \dots, \tau_n)$, the input x , and time. It will be assumed that the kernels are symmetrical.

This means that

$$K_n(\tau_1, \tau_2, \dots, \tau_n) = K_n(\tau_n, \tau_{n-1}, \dots, \tau_1), \quad (255)$$

for all permutations of τ_i . All systems can be considered to have symmetrical kernels since, if the kernel is not symmetrical we can obtain a corresponding symmetrical kernel by the formula

$$K_n^* = \frac{1}{n!} \sum K_n \quad (256)$$

where K_n^* is the symmetrical kernel and K_n is the original, unsymmetrical kernel. The term $\sum K_n$ means that we take all permutations of the variables τ_i in K_n , of which there are $n!$, and add them up. As an example, if

$K_2(\tau_1, \tau_2)$ is not symmetrical, we can write

$$K_2^* = \frac{K_2(\tau_1, \tau_2) + K_2(\tau_2, \tau_1)}{2!},$$

which is symmetrical.

Our next step is to determine the G's. We know that we want the G's to be ordered; i.e., we want G_0 to be a constant, G_1 to be a first order functional, etc., and we want the functionals to be orthogonal. For convenience we will normalize G_0 by setting it equal to unity, that is, the integral of G_0^2 is equal to unity. Thus,

$$G_0 = 1 \quad . \quad (257)$$

We write G_1 as

$$G_1 = \int K_1(\tau) x(t-\tau) d\tau + K_0, \quad (258)$$

and we want G_1 to be orthogonal to G_0 . We will assume that the Gaussian input $x(t - \tau)$ has an average value equal to zero; i.e., both positive and negative values are equally likely. If we make G_1 orthogonal to unity it can be made orthogonal to all constants, thus, by the definition of orthogonality we multiply G_1 by unity, average over the time and set it equal to zero. Then

$$\begin{aligned} 0 &= \int (1) G_1(\tau) d\tau = \int K_0 dt + \int dt \int K_1(\tau) x(t-\tau) d\tau \\ &= K_0 + \int K_1(\tau) d\tau \int x(t-\tau) dt, \end{aligned} \quad (259)$$

since $\frac{1}{a-b} \int_a^b K_0 dt = K_0$. Also, since $\int x(t-\tau) dt = 0$, we have

$$K_0 = 0 \quad . \quad (260)$$

Thus, for G_1 to be orthogonal to a constant, $K_0 = 0$ and we now have

$$G_1 = \int k_1(\tau) x(t-\tau) d\tau \quad . \quad (261)$$

To normalize G_1 , we must have

$$\int G_1^2 dt = 1 \quad . \quad (262)$$

Then, since $k_0 = 0$, we have

$$\int G_1^2 dt = \int dt \int k_1(\tau_1) x(t-\tau_1) d\tau_1 \int k_1(\tau_2) x(t-\tau_2) d\tau_2 = 1 \quad (263)$$

Since the kernels $k_1(\tau_1)$ and $k_1(\tau_2)$ are independent of time, we can rearrange Equation (263) to get

$$\int d\tau_1 \int d\tau_2 k_1(\tau_1) k_1(\tau_2) \int x(t-\tau_1) x(t-\tau_2) dt = 1 \quad (264)$$

Now, because the input is Gaussian and $x(t - \tau_1)$ and $x(t - \tau_2)$ are statistically independent, $\int x(t-\tau_1) x(t-\tau_2) dt$ is just the autocorrelation function of the input, and for $\tau_1 \neq \tau_2$ there is no correlation; i.e., the integral above is zero. However, for $\tau_1 = \tau_2$ the autocorrelation function is a Dirac delta function and Equation (264) becomes

$$\int \int \delta(\tau_1 - \tau_2) k_1(\tau_1) k_1(\tau_2) d\tau_1 d\tau_2 = 1 \quad . \quad (265)$$

Integrating over τ_2 gives

$$\int k_1^2(\tau) d\tau = 1 \quad . \quad (266)$$

Thus, G_1 is normalized, and we have two functionals G_0 and G_1 which are orthonormal.

Now let's consider the functional G_2 which we write as

$$G_2 = \iint K_2(\tau_1, \tau_2) x(t-\tau_1) x(t-\tau_2) d\tau_1 d\tau_2 + \int K_1(\tau) x(t-\tau) d\tau + k_0, \quad (267)$$

where $K_1(\tau)$ is not necessarily the same as was used in G_1 , but only indicates a first order functional in the input $x(t)$. We want to define G_2 so that it is normal to any constant and to any functional of first degree.

To determine orthogonality to any constant, as for G_1 we multiply by unity, average over time, and set equal to zero. Thus;

$$0 = \int \langle G_2 \rangle dt = \iint K_2(\tau_1, \tau_2) \overline{x(t-\tau_1) x(t-\tau_2)} d\tau_1 d\tau_2 + \int K_1(\tau) \overline{x(t-\tau)} d\tau + k_0, \quad (268)$$

where the bar denotes averaging over time. As before, $\overline{x(t-\tau)} = 0$,

and $\overline{x(t-\tau_1) x(t-\tau_2)} = \delta(\tau_1 - \tau_2)$. Thus,

$$k_0 = - \int K_2(\tau, \tau) d\tau. \quad (269)$$

To be orthogonal to any first degree functional we must multiply by an arbitrary first degree functional $\int C(\tau) x(t-\tau) d\tau$, average over time, and set equal to zero. Thus,

$$0 = \iiint C(\tau) K_2(\tau_1, \tau_2) \overline{x(t-\tau) x(t-\tau_1) x(t-\tau_2)} d\tau d\tau_1 d\tau_2 + \iint C(\tau) K_1(\tau_1) \overline{x(t-\tau) x(t-\tau_1)} d\tau d\tau_1 + k_0 \int C(\tau) \overline{x(t-\tau)} d\tau, \quad (270)$$

which becomes

$$\int C(\tau) K_1(\tau) d\tau = 0, \quad (271)$$

since $\overline{x(t-\tau) x(t-\tau_1) x(t-\tau_2)}$ and $\overline{x(t-\tau)}$ vanish for a Gaussian distribution and $\overline{x(t-\tau) x(t-\tau_1)} = \delta(\tau - \tau_1)$. It can be shown that the average value of the product of n Gaussian distributed signals;

i.e., $\overline{\chi(t-\tau)\chi(t-\tau_1)\cdots\chi(t-\tau_n)}$ is zero for n odd and reduces to delta functions for n even. Now, since $C(\tau)$ is arbitrary, to satisfy Equation (271), $K_1(\tau)$ must be zero. Thus,

$$K_1(\tau) = 0,$$

and

$$G_2 = \iint K_2(\tau_1, \tau_2) \chi(t-\tau_1) \chi(t-\tau_2) d\tau_1 d\tau_2 - \int K_2(\tau, \tau) d\tau, \quad (272)$$

and G_2 is orthogonal to all constants, any first order functional, and any combination of the two.

Finally, for normalization we want

$$\int G_2^2 dt = 1, \quad (273)$$

thus

$$\begin{aligned} \int G_2^2 dt = & \iint K_2(\tau_1, \tau_2) \chi(t-\tau_1) \chi(t-\tau_2) \iint K_2(\tau_3, \tau_4) \chi(t-\tau_3) \chi(t-\tau_4) d\tau_1 \cdots d\tau_4 - \\ & - \int K_2(\tau, \tau) d\tau \iint K_2(\tau_3, \tau_4) \chi(t-\tau_3) \chi(t-\tau_4) d\tau_3 d\tau_4 - \\ & - \int K_2(\tau_5, \tau_5) d\tau_5 \iint K_2(\tau_1, \tau_2) \chi(t-\tau_1) \chi(t-\tau_2) d\tau_1 d\tau_2 + \\ & + \int K_2(\tau_5, \tau_5) d\tau_5 \int K_2(\tau, \tau) d\tau = 1. \quad (274) \end{aligned}$$

Averaging over time gives

$$\begin{aligned} & \iiint K_2(\tau_1, \tau_2) K_2(\tau_3, \tau_4) \overline{\chi(t-\tau_1) \chi(t-\tau_2) \chi(t-\tau_3) \chi(t-\tau_4)} d\tau_1 \cdots d\tau_4 - \\ & - \int K_2(\tau, \tau) d\tau \iint K_2(\tau_3, \tau_4) \overline{\chi(t-\tau_3) \chi(t-\tau_4)} d\tau_3 d\tau_4 - \\ & - \int K_2(\tau_5, \tau_5) d\tau_5 \iint K_2(\tau_1, \tau_2) \overline{\chi(t-\tau_1) \chi(t-\tau_2)} d\tau_1 d\tau_2 + \\ & + \int K_2(\tau_5, \tau_5) d\tau_5 \int K_2(\tau, \tau) d\tau = 1. \quad (275) \end{aligned}$$

Now, as mentioned above, it can be shown that

$$\overline{\chi(t-\tau_1) \chi(t-\tau_2) \cdots \chi(t-\tau_n)} = \begin{cases} 0 & ; \text{ for } n \text{ odd} \\ \sum \prod \overline{x_i x_j} & ; \text{ for } n \text{ even} \end{cases}$$

where $\overline{x_i x_j}$ is the delta function for all combinations of i, j , and \prod denotes the product of pairs of combinations. For example, consider

$\overline{\chi(t-\tau_1) \chi(t-\tau_2) \chi(t-\tau_3) \chi(t-\tau_4)}$. This is written as

$$\begin{aligned} & \left[\overline{\chi(t-\tau_1) \chi(t-\tau_2)} \right] \left[\overline{\chi(t-\tau_3) \chi(t-\tau_4)} \right] + \left[\overline{\chi(t-\tau_1) \chi(t-\tau_3)} \right] \left[\overline{\chi(t-\tau_2) \chi(t-\tau_4)} \right] + \\ & + \left[\overline{\chi(t-\tau_1) \chi(t-\tau_4)} \right] \left[\overline{\chi(t-\tau_3) \chi(t-\tau_2)} \right] \end{aligned}$$

This may also be written as

$$\left[\delta(\tau_1-\tau_2) \delta(\tau_3-\tau_4) + \delta(\tau_1-\tau_3) \delta(\tau_2-\tau_4) + \delta(\tau_1-\tau_4) \delta(\tau_3-\tau_2) \right] \cdot$$

Using this notation, Equation (275) becomes

$$\begin{aligned} & \iiint \int K_2(\tau_1, \tau_2) K_2(\tau_3, \tau_4) \left[\delta(\tau_1-\tau_2) \delta(\tau_3-\tau_4) + \delta(\tau_1-\tau_3) \delta(\tau_2-\tau_4) + \delta(\tau_1-\tau_4) \delta(\tau_3-\tau_2) \right] d\tau_1 \cdots d\tau_4 - \\ & - \iiint \int K_2(\tau, \tau) K_2(\tau_3, \tau_4) \delta(\tau_3-\tau_4) d\tau d\tau_3 d\tau_4 - \\ & - \iiint \int K_2(\tau_5, \tau_5) K_2(\tau_1, \tau_2) \delta(\tau_1-\tau_2) d\tau_5 d\tau_1 d\tau_2 + \\ & + \iiint \int K_2(\tau_5, \tau_5) K_2(\tau, \tau) d\tau_5 d\tau = 1 \end{aligned} \quad (276)$$

Now consider each term of Equation (276) individually. Integrating the second term on the left with respect to τ_4 gives

$$- \iint K_2(\tau, \tau) K_2(\tau_3, \tau_3) d\tau d\tau_3 = - \left[\int K_2(\tau, \tau) d\tau \right]^2 \cdot$$

Integrating the third term with respect to τ_2 gives

$$- \iint K_2(\tau_5, \tau_5) K_2(\tau_1, \tau_1) d\tau_5 d\tau_1 = - \left[\int K_2(\tau, \tau) d\tau \right]^2 \cdot$$

The fourth term may be written as

$$+ \left[\int K_2(\tau, \tau) d\tau \right]^2 \quad .$$

Thus, adding the second, third and fourth terms gives

$$- \left[\int K_2(\tau, \tau) d\tau \right]^2 \quad . \quad (277)$$

For the first term on the left-hand side of Equation (276), we have to integrate four times; i.e., with respect to τ_1 , τ_2 , τ_3 , and τ_4 . Consider only the integral

$$\iiint \int K_2(\tau_1, \tau_2) K_2(\tau_3, \tau_4) \delta(\tau_1 - \tau_2) \delta(\tau_3 - \tau_4) d\tau_1 \dots d\tau_4 \quad .$$

Integrate first with respect to τ_4 and then with respect to τ_2 . This gives

$$\iint K_2(\tau_1, \tau_2) K_2(\tau_3, \tau_3) d\tau_1 d\tau_3 \quad . \quad (278)$$

Consider only the integral

$$\iiint \int K_2(\tau_1, \tau_2) K_2(\tau_3, \tau_4) \delta(\tau_1 - \tau_3) \delta(\tau_2 - \tau_4) d\tau_1 \dots d\tau_4 \quad .$$

Integrating first with respect to τ_4 and then with respect to τ_3 gives

$$\iint K_2(\tau_1, \tau_2) K_2(\tau_1, \tau_2) d\tau_1 d\tau_2 \quad . \quad (279)$$

Finally, consider the integral

$$\iiint \int K_2(\tau_1, \tau_2) K_2(\tau_3, \tau_4) \delta(\tau_1 - \tau_4) \delta(\tau_3 - \tau_2) d\tau_1 \dots d\tau_4 \quad .$$

Integrating first with respect to τ_4 and then with respect to τ_3 gives

$$\iint K_2(\tau_1, \tau_2) K_2(\tau_2, \tau_1) d\tau_1 d\tau_2 \quad . \quad (280)$$

But $K_2(\tau_1, \tau_2) = K_2(\tau_2, \tau_1)$ by symmetry, and Equations (279) and (280) are equivalent.

From Equations (276), (277), (278), (279) and (280) then we get

$$\iint K_2(\tau_1, \tau_1) K_2(\tau_3, \tau_3) d\tau_1 d\tau_3 + 2 \iint K_2^2(\tau_1, \tau_2) d\tau_1 d\tau_2 - \left[\int K_2(\tau, \tau) d\tau \right]^2 = 1. \quad (281)$$

Now $K_2(\tau_1, \tau_1)$ and $K_2(\tau_3, \tau_3)$ are equivalent to $K_2(\tau, \tau)$. Thus the first integral on the left of Equation (281) is equivalent to

$$\int K_2(\tau_1, \tau_1) d\tau_1 \int K_2(\tau_3, \tau_3) d\tau_3 = \left[\int K_2(\tau, \tau) d\tau \right]^2,$$

and Equation (281) becomes

$$2 \iint K_2^2(\tau_1, \tau_2) d\tau_1 d\tau_2 = 1. \quad (284)$$

This means then that G_2 is normalized, and we can write G_2 as in Equation (272)

as

$$G_2 = \iint K_2(\tau_1, \tau_2) X(t-\tau_1) X(t-\tau_2) d\tau_1 d\tau_2 - \int K_2(\tau, \tau) d\tau. \quad (285)$$

We are now in a position to write a general formula for G_n as

$$G_n = \sum_{\nu=0}^{\left[\frac{n}{2} \right]} a_{n-2\nu}^{(n)} \int \cdots \int_n K_n(\tau_1, \tau_2, \dots, \tau_n) X(t-\tau_1) \cdots X(t-\tau_{n-2\nu}) \cdot \delta(\tau_{n-2\nu+1} - \tau_{n-2\nu+2}) \cdots \delta(\tau_{n-1} - \tau_n) d\tau_1 \cdots d\tau_n, \quad (286)$$

where

$$\left[\frac{n}{2} \right] = \begin{cases} n/2 & \text{for } n \text{ even} \\ \frac{n-1}{2} & \text{for } n \text{ odd} \end{cases}$$

and

$$a_{n-2\nu}^{(n)} = (-1)^\nu \frac{n!}{2^\nu (n-2\nu)! \nu!}.$$

And the generalized normalization formula is

$$\int G_n^2 dt = n! \int \cdots \int_n K_n^2(\tau_1, \tau_2, \dots, \tau_n) d\tau_1 \cdots d\tau_n = I_{(287)}$$

Thus we have shown that we can expand the nonlinear functional F in terms of orthonormal functionals G_n which depend on the kernels K_n , the input $x(t - \tau)$ and time, where the input is Gaussian.

Wiener proved that the set of G 's is complete in the following manner:

There is a definite difference in the expansion of a functional in terms of orthonormal functionals and in the expansion of a function in terms of orthogonal functions. To understand this difference, we will use examples. Suppose we have a function of time $f(t)$ and expand it in terms of orthogonal functions $\phi_n(t)$. Then we can write

$$f(t) = \sum_{n=0}^N a_n \phi_n(t) \quad .$$

If we limit the number of terms to describe the function, the mean square error is given by

$$m.s.e. = \overline{f^2(t)} - \sum_{n=0}^N a_n^2$$

which approaches zero as N approaches infinity. This defines completeness of a set of orthogonal functions.

In the case of functionals as described by Wiener, completeness of the functionals is defined for the expansion

$$y(t) = \sum_{n=0}^N G_n = G_0 + G_1 + \cdots + G_n$$

if we have

$$\int y^2(t) dt \longrightarrow \sum_{n=0}^N \int G_n^2 dt$$

where the G's are orthogonal but not normalized. However, if the G's are also normalized

$$\int y^2(t) dt \longrightarrow \sum_{n=0}^N a_n^2 \int G_n^2 dt \longrightarrow \sum_{n=0}^N a_n^2$$

where the a's are constants. Thus for functionals expanded in terms of functionals, the approximation of $y(t)$ by the G's will be complete only in a statistically average sense.

Now, all this is very nice but how does it relate to our problem of analyzing specific physical systems? We will now discuss Wiener's application of the above theoretical development.

Suppose we have a physical system into which we put a Gaussian white noise (we will also call this a stochastic function) input and get an output. The Gaussian input for example may be produced by an electron tube since emission of electrons in a tube is Gaussian. Now let's put a system, the characteristics of which are known, in parallel with the unknown physical system.

Now feed the outputs of both the unknown physical system and the known system into a multiplier, and feed the output of the multiplier into an integrator to obtain average values. The diagram for this system is shown in Figure 84. Wiener claims that the output of the integrator provides a means of defining the physical system. We will see how he arrives at this conclusion.

By Wiener's interpretation, the output of the unknown physical system can be represented by $\sum_n G_n(K_n, x, t)$. We can also consider the output of the known box in terms of (for the present unknown) G functionals as $\sum_m G_m(H_m, x, t)$ where the kernels are of course different since the systems are different. Now we can multiply these two outputs and integrate the product with respect to time to get the average values.

However, to completely describe the physical system we also need the non-linear functionals of second and higher order. These are more difficult to obtain experimentally although the procedure is quite similar. To determine the second-order functional we can define the output of the known box as

$$\sum_m G_m(H_m, X, t) = \iint \delta(\tau_1 - \tau_3) \delta(\tau_2 - \tau_4) X(t - \tau_1) X(t - \tau_2) d\tau_1 d\tau_2 \quad (293)$$

Thus, the kernel is $H_2 = \delta(\tau_1 - \tau_3) \delta(\tau_2 - \tau_4)$. Integrating the right-hand side of Equation (293) first with respect to τ_2 and then with respect to τ_1 we get

$$\begin{aligned} \sum_m G_m(H_m, X, t) &= \int \delta(\tau_1 - \tau_3) X(t - \tau_1) X(t - \tau_4) d\tau_1 \\ &= X(t - \tau_3) X(t - \tau_4) \quad (294) \end{aligned}$$

Thus, the known box consists of two delay terms and a multiplier. A diagram of the known box is shown in Figure 85.

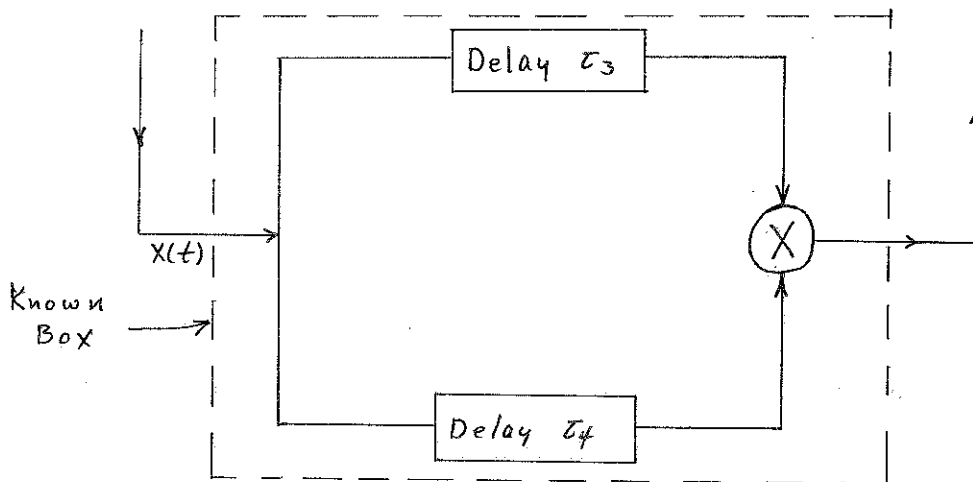


Fig. 85

The second order functional for the physical system is given by

$$G_2(K_n, X, t) = \iint K_2(\tau_5, \tau_6) X(t - \tau_5) X(t - \tau_6) d\tau_5 d\tau_6 - \int K_2(\tau, \tau) d\tau \quad (295)$$

Multiplying Equations (294) and (295) and averaging with respect to time gives

$$\begin{aligned} \overline{\sum_{n=m=2} G_n G_m} &= \iint K_2(\tau_5, \tau_6) \overline{x(t-\tau_5)x(t-\tau_6)x(t-\tau_3)x(t-\tau_4)} d\tau_5 d\tau_6 - \\ &\quad - \int K_2(\tau, \tau) \overline{x(t-\tau_3)x(t-\tau_4)} d\tau \\ &= \iint K_2(\tau_5, \tau_6) [\delta(\tau_5-\tau_6)\delta(\tau_3-\tau_4) + \delta(\tau_5-\tau_6)\delta(\tau_6-\tau_4) + \delta(\tau_4-\tau_5) \cdot \\ &\quad \cdot \delta(\tau_3-\tau_6)] d\tau_5 d\tau_6 - \int K_2(\tau, \tau) \delta(\tau_3-\tau_4) d\tau. \end{aligned} \quad (296)$$

The last term on the right side of Equation (296) vanishes since $\tau_3 \neq \tau_4$ and the delta function vanishes for these conditions. Integrating Equation (296) with respect to τ_5 gives

$$\begin{aligned} \overline{\sum_{n,m} G_n G_m} &= \int K_2(\tau_5, \tau_6) \delta(\tau_3-\tau_4) d\tau_6 + \int K_2(\tau_3, \tau_6) \delta(\tau_6-\tau_4) d\tau_6 + \\ &\quad + \int K_2(\tau_4, \tau_6) \delta(\tau_3-\tau_6) d\tau_6. \end{aligned} \quad (297)$$

Integrating Equation (297) with respect to τ_6 gives

$$\begin{aligned} \overline{\sum_{n,m} G_n G_m} &= 0 + K_2(\tau_3, \tau_4) + K_2(\tau_4, \tau_3) \\ &= 2 K_2(\tau_3, \tau_4), \end{aligned} \quad (298)$$

since K_2 is symmetrical. Thus for the Gaussian input $x(t)$ and the known box as described in Figure 85, the kernel for the second order functional for the physical system can be determined by experiment. Therefore we can write, for the physical system,

$$G_2(K_2, x, t) = \iint K_2(\tau_3, \tau_4) x(t-\tau_3)x(t-\tau_4) d\tau_3 d\tau_4 - \int K_2(\tau, \tau) d\tau, \quad (299)$$

where $K_2(\tau_1, \tau_2)$ can be determined from experiment.

Theoretically, this method of determining the kernels of the physical system applies to all higher order kernels also. Experimentally, however, it is extremely difficult to implement this method because of the necessity of using a Gaussian signal. First, a Gaussian signal requires an infinitely long time to determine average values and second, if an infinitely long time is not considered, the errors involved are found to be large compared with the kernel values that are to be measured. The theory does, however, suggest a way of studying dynamics and people are trying to find easier ways to implement the basic principles.

Suggested References

1. Volterra, Theory of Functionals, Reprint, Dover, 1959.
2. N. Wiener, Nonlinear Problems in Random Theory; Massachusetts Institute of Technology Press, 1958.

LECTURE NO. IX

FURTHER METHODS OF ANALYSIS OF NONLINEAR SYSTEMS

Let's review briefly what we discussed in Lecture VIII. We realize that we are ignorant about many things, but we are curious about them and we experiment and try to learn. Many devices are conceived and designed by man to perform a certain desired function but which are not thoroughly understood at the time they are built. To illustrate this point, consider the concept of a nuclear reactor. On the basis of the fundamental principle of "fission", tremendous devices have been designed and constructed to produce heat. But as yet we cannot describe completely the detailed processes occurring in the reactor under all possible conditions. Because of our insatiable desire to better understand what we have created, however, we experiment with these devices in an effort to determine their behavior when subjected to various operating conditions with the specific object in mind that the experimental results will assist us in obtaining fundamental information about the system.

We know from our experiences with simple (linear) physical systems that if we disturb the system with an input of some kind, the relationship between the input and the resulting output contains information about the system. This is illustrated by the fact that if we put the same input into two different physical systems we get different outputs. Applying this method to the more complicated (nonlinear) systems and admitting that all the pertinent information is contained in the input applied and the output observed, we need to determine how to extract this information. The effort to extract this information is called "analysis" of the physical system.

We begin our analysis in a qualitative manner by expressing the relationship between input and output in the following abstract form:

$$y(t) = F[x(t - \tau)] \quad ,$$

where $y(t)$ is the output, $x(t)$ is the input, and the relationship says that the output of the system depends on the input at all past times τ (but not at future times). This means that the system has a memory. Since the input is a function of time itself, we call the relationship between input and output a functional relationship. Thus F represents a functional.

Our second effort is to expand this functional in an infinite series because we know how to work with series. In so doing we shall attempt also to compartmentalize the effects of the system memory and of the system non-linearity. Thus we write

$$y(t) = h_0 + \int h_1(\tau) x(t-\tau) d\tau + \iint h_2(\tau_1, \tau_2) x(t-\tau_1) x(t-\tau_2) d\tau_1 d\tau_2 + \dots$$

which, in general, involves an infinite number of terms. However, for practical application we shall wish to be able to use only two or three terms; any larger number would make the effort involved almost prohibitive.

Now, whenever we expand a function in a series we like to impose the property of orthogonality. This is very important for the following reasons: First, by so doing we make the error involved on truncating the series at any point a minimum in the least mean square sense. Second, the error decreases as the number of terms in the series increases. And third, for an expansion in terms of orthogonal functions, each term is independent of the others. Thus we like to expand in terms of orthogonal functions.

To better illustrate what we mean by orthogonal functions, consider the output of the system as a vector in an n -dimensional space. In this space (as in 3-dimensional space) the axes or vectors describing the system are all orthogonal or perpendicular to each other. Then a series expansion of the output vector is nothing more than expressing the vector in terms of its projections along the directions of the n -dimensional space. For each set of orthogonal functions there corresponds a set of vectors in the n -dimensional

space. Therefore, which axes are applicable depends upon which orthogonal set we wish to expand the vector in terms of.

As was mentioned in the first lecture, one must be discreet in choosing which set of orthogonal functions to use in describing the output of a particular system for a particular input. In the expansion above, we expressed the output as a series of functionals involving the input. The next step then is to determine if we can make this expansion in terms of orthogonal functionals which are functions of the input.

In Lecture No. VIII we discussed the feasibility of doing this and described how it could be done for a Gaussian white-noise input function. We realized, however, that the required Gaussian input was not particularly useful from a practical standpoint. Thus, the need still exists to be able to perform the desired expansion in terms of a more practical input function. Today's lecture (the last in this series) will describe several methods of attempting to implement the theory of Lecture No. VIII.

The fact that the Gaussian white-noise signal is the best signal to use (to permit the required expansion mentioned above) is a result of some special features of the Gaussian signal. First, the Gaussian signal has a finite probability of reproducing any conceivable signal (provided the Lebesgue integral of that signal exists). Second, it makes sense to talk about stochastic processes for many systems. Third, and perhaps the most convenient property of the Gaussian signal, is that when terms involving products of the input function by itself involve an odd number of factors, $x(t - \tau_i)$, the average value of such terms is zero. This greatly simplifies the possibility of implementing the theory and also results in the fact that the terms of the expansion are homogeneous. Homogeneity in this case means that for each orthogonal functional in the expansion, the terms in that functional are all of the same order in the input $x(t)$, so to speak. This property is not always possible to

achieve with other signals.

Today we will approach the problem of determining the G functionals in a slightly different manner. This approach will emphasize the importance of the Gaussian signal and will also indicate what requirements are necessary if we attempt to use a signal other than Gaussian.

I. ORTHOGONAL POLYNOMIALS

Instead of talking about orthogonal functionals in general, let's talk about orthogonal polynomials. We approach this problem in the following manner: If we have the sequence of variables $1, \bar{z}, z^2, \dots, z^n$, we can build up a set of orthogonal polynomials $P_n(z)$ where $P_n(z)$ is a polynomial of n^{th} power in z , and such that $P_n(z)$ is orthogonal to all other polynomials in the set. The expression for this type of polynomial is

$$P_n(z) = (D_{n-1}, D_n)^{-1/2} \begin{vmatrix} C_0 & C_1 & \dots & C_n \\ C_1 & C_2 & \dots & C_{n+1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ C_{n-1} & \dots & \dots & C_{2n-1} \\ 1 & \bar{z} & \dots & \bar{z}^n \end{vmatrix}, \quad (300)$$

with

$$C_m = \overline{\bar{z}^m} = \frac{1}{B-A} \int_A^B \bar{z}^m d\bar{z}, \quad (301)$$

where the interval from A to B is the region of interest, and D_n is a determinant with elements $C_{\nu+\mu}$ given by

$$D_n = |C_{\nu+\mu}| \quad ; \quad \nu, \mu = 0, 1, \dots, n. \quad (302)$$

From the orthogonality condition

$$\int_A^B P_n(z) P_m(z) d\bar{z} = 0 \quad ; \quad \text{for } n \neq m. \quad (303)$$

As an example of how we can make the polynomials orthogonal, consider the set

$$P_0 = 1$$

$$P_1(z) = z - \bar{z} = z - c_1$$

where c_1 is a constant, and

$$P_2(z) = az^2 + bz + c \quad .$$

Then to make $P_1(z)$ orthogonal to P_0 , we multiply $P_1(z)$ by unity, integrate from A to B and set equal to zero. This gives $\bar{z} = \frac{A+B}{2} = c_1$. Thus

$$P_1(z) = z - \frac{A+B}{2} \quad .$$

To make $P_2(z)$ orthogonal to both P_0 and $P_1(z)$, perform the same operations by multiplying $P_2(z)$ by unity and integrating, and then multiplying $P_2(z)$ by $P_1(z)$ and integrating. This will give two equations in three unknowns. We can solve these equations for two of the unknowns in terms of the other, and then, by normalization, solve for the third unknown. Thus the procedure is quite similar to that we followed for the G functionals and the kernels. If we assume that z is a random variable with some specified probability distribution, we can generalize the above procedure by multiplying by a weighting function $w(z)$ to give

$$\int_A^B P_n(z) P_m(z) w(z) dz = 0 \quad . \quad (304)$$

Depending on the form of the weighting function, we can determine various sets of orthogonal polynomials, such as Legendre polynomials, Bessel functions, Jacobi polynomials, Hermite polynomials, Laguerre polynomials, etc. Thus, from a given sequence there is no difficulty in defining a set of orthogonal polynomials.

Suppose we have a different sequence which has many variables; for example,

$$1, z_1, z_2, \dots, z_n, z_1^2, z_1 z_2, z_1 z_3, \dots, z_1 z_n, z_2 z_n, \dots$$

Following a similar procedure to that outlined above, we can determine a set of orthogonal polynomials which are functions of the variables. Here we took discrete variables, but suppose we don't consider discrete variables but variables which are determined from a function of time. That is, suppose we have an $x(t)$ and let $x(t_1) = z_1, x(t_2) = z_2, \dots$. We can therefore build up the above sequence in this manner and the sequence is made up of time dependent functions evaluated at different times.

In the case of a Gaussian signal with a white noise spectrum, Grad [] has shown that the following polynomials are obtained:

$$\begin{aligned} P_0 &= 1 \\ P_1 &= x(t) \\ P_2 &= x(t-\tau_1)x(t-\tau_2) - \overline{x(t-\tau_1)x(t-\tau_2)} \\ &= x(t-\tau_1)x(t-\tau_2) - \int(\tau_1-\tau_2) \\ P_3 &= x(t-\tau_1)x(t-\tau_2)x(t-\tau_3) - x(t-\tau_1)\int(\tau_2-\tau_3) - \\ &\quad - x(t-\tau_2)\int(\tau_1-\tau_3) - x(t-\tau_3)\int(\tau_2-\tau_1) \\ &\quad \vdots \\ &\quad \cdot \end{aligned}$$

By multiplying each successive expression by the previous ones and integrating according to the definition of orthogonality, it is easy to show that all the terms are orthogonal. Actually, this is a way of defining the G functionals we spoke of in Lecture No. VIII. To show this, suppose we take a kernel K_n and operate on the n^{th} polynomial, P_n . Then we postulate

$$G_n = \int \dots \int_n K_n(\tau_1, \tau_2, \dots, \tau_n) P_n d\tau_1 d\tau_2 \dots d\tau_n \quad (305)$$

We will now prove that this is the G_n functional.

It is evident that $K_0 P_0 = K_0$, thus we generate the G_0 functional, $G_0 = K_0$ which can be normalized to give $G_0 = 1$. For the first order functional G_1 (from Equation (305)), we have

$$\int K_1(\tau) \chi(t) d\tau = G_1,$$

which is the same relationship obtained before. Continuing, we have

$$\begin{aligned} G_2 &= \iint K_2(\tau_1, \tau_2) [\chi(t-\tau_1) \chi(t-\tau_2) - \delta(\tau_1-\tau_2)] d\tau_1 d\tau_2 \\ &= \iint K_2(\tau_1, \tau_2) \chi(t-\tau_1) \chi(t-\tau_2) d\tau_1 d\tau_2 - \iint K_2(\tau_1, \tau_2) \delta(\tau_1-\tau_2) d\tau_1 d\tau_2. \end{aligned}$$

Integrating the second integral on the right with respect to τ_2 gives

$$G_2 = \iint K_2(\tau_1, \tau_2) \chi(t-\tau_1) \chi(t-\tau_2) d\tau_1 d\tau_2 - \int K_2(\tau_1, \tau_1) d\tau_1,$$

which is also the same as was obtained in Lecture No. VIII. We could continue this process and build up all of the G functionals, however, we can already recognize certain implications. Instead of generating the G functionals, we can leave the results in the form $K_n P_n$. All the polynomials P_n are orthogonal to each other; thus, let's define an n^{th} order functional to be

$$\int \dots \int_n K_n P_n d\tau_1 \dots d\tau_n$$

which has an n^{th} order kernel operating on an n^{th} order polynomial, as yet unspecified, and also has the property of being orthogonal to all other polynomials of the same character. We can show that the functionals are also orthogonal and we obtain the relationship

$$\int dt \int \dots \int_{n+m} K_n K_m P_n P_m d\tau_1 \dots d\tau_{n+m} = 0 \quad ; \quad n \neq m \quad (306)$$

since the kernels are independent of time and we have already specified that the polynomials are orthogonal. Thus, if we want to use a signal other than the Gaussian signal, the signal we use must have the following properties:

1. It is easily produced
2. It can be shifted easily (i.e., delayed)
3. It must lend itself to the creation of orthogonal polynomials which are relatively simple.

We see then that we are now in the same position as we were with Wiener's method where the specified signal was the Gaussian white-noise signal. However, we now have the advantage that the polynomials describing our signal need be orthogonal only over a limited time range. This latter property is very desirable since for practical application, any experiment must be performed during a finite time interval, and all that is required is that the polynomials be orthogonal over this period of time. Of course the time interval must be long enough to obtain all the information about the system.

In principle, regardless of what signal we use we can always determine the required polynomials describing the signal. However, if some forethought and judgment is not used in selection of the signal, we may wind up having to solve very involved equations. As an example, suppose we have a signal $x(t)$, the average value of which is zero; i.e., $\bar{x}(t) = 0$. Then let

$$P_0 = 1$$

$$P_1 = x(t - \tau_0)$$

$$P_2 = x(t - \tau_1)x(t - \tau_2) - g_2(\tau_1, \tau_2)x(t - \tau_0) - \overline{x(t - \tau_1)x(t - \tau_2)}$$

where $g_2(\tau_1, \tau_2)$ is a kernel which is a function of the signal, not of the system. Then for orthogonality, it is readily seen that P_1 is orthogonal to P_0 since $\int x(t - \tau_0)dt = 0$ by definition. For P_2 to be orthogonal to P_0

$$\int dt \left[x(t-\tau_1)x(t-\tau_2) - g_2(\tau_1, \tau_2)x(t-\tau_0) - \overline{x(t-\tau_1)x(t-\tau_2)} \right] = 0 .$$

Performing the integration gives

$$\overline{x(t-\tau_1)x(t-\tau_2)} - g_2(\tau_1, \tau_2) \int x(t-\tau_0) dt - \overline{x(t-\tau_1)x(t-\tau_2)} = 0 .$$

Since $\int x(t-\tau_0) dt = 0$, this orthogonality condition is satisfied. For P_2 to be orthogonal to P_1 , performing the integration with respect to time gives

$$\overline{x(t-\tau_0)x(t-\tau_1)x(t-\tau_2)} - g_2(\tau_1, \tau_2) \overline{x(t-\tau_0)x(t-\tau_0)} - (\overline{x(t-\tau_1)x(t-\tau_2)}) \overline{x(t-\tau_0)} = 0$$

and since $\overline{x(t-\tau_0)}$ is zero, the last term on the left vanishes and

$$g_2(\tau_1, \tau_2) = \frac{\overline{x(t-\tau_0)x(t-\tau_1)x(t-\tau_2)}}{\overline{x(t-\tau_0)x(t-\tau_0)}} . \quad (307)$$

Thus, for an arbitrary input $x(t)$ with $\bar{x}(t) = 0$, it is possible to satisfy the orthogonality conditions, but the equations to be solved are unduly complicated, and the second and higher order kernels involved in the polynomials may be quite difficult to determine. In principle, however, $g_n(\tau_1, \dots, \tau_n)$ can be evaluated from the average values of products of the input signal, as indicated by Equation (307). Thus, for any input it is conceptually possible to make an expansion of the output in terms of orthogonal functionals involving orthogonal polynomials which are functions of the input.

We will now discuss some experiments which have been performed with nuclear reactor systems. There are five main experimental methods used in conjunction with nuclear reactors; crosscorrelation and autocorrelation measurements, oscillation experiments (where the driving source varies sinusoidly), ramp tests (where reactivity is added to the system linearly with time), and step tests (where excess reactivity insertions approximate a step function). As a first attempt to use the principles of the theories discussed in Lectures VIII and IX, we will try to compare what one measures in a reactor by using each of the three methods; crosscorrelation using a binary signal, autocorrelation, and crosscorrelation using an oscillatory signal. We will first discuss

crosscorrelation by describing an experiment that was performed using a binary signal; i.e., the signal consisted of positive and negative values of equal amplitude. The crosscorrelation technique was of course applied to the binary input and the output of the reactor. We will show that the crosscorrelation method reveals only the linear response of the nonlinear system.

The second approach describes application of the autocorrelation technique to nonlinear systems. No external signal is required for this application. Autocorrelation is applied to the output of a critical reactor. The internal source is considered Gaussian white noise, although this assumption may not be entirely valid. It will be shown that the autocorrelation technique applied in this manner involves not only the linear functionals, but also the higher-order functionals.

Finally, we will show that the oscillation technique describes the response in terms of the odd kernels only.

Crosscorrelation Method

The input signal used in this experiment was a binary signal, the characteristics of which are shown in Figure 86. The interpretation of Figure 86 is

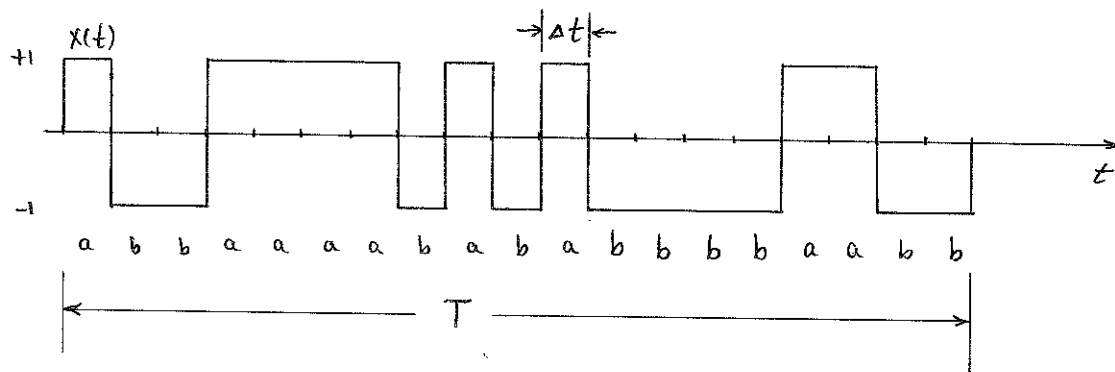


Fig. 86

as follows: the signal has positive and negative values of equal magnitude after the start of time; a time interval is denoted by Δt and for this pictorial example there are 19 time intervals for a period T ; it is possible for the signal to change sign after each Δt , although it may not necessarily do so.

This special signal also has a special autocorrelation property. The autocorrelation function of the signal $x(t)$ is as shown in Figure 87. When $-\Delta t < \tau < \Delta t$, $\phi_{xx}(\tau)$ varies linearly, having a maximum value of unity

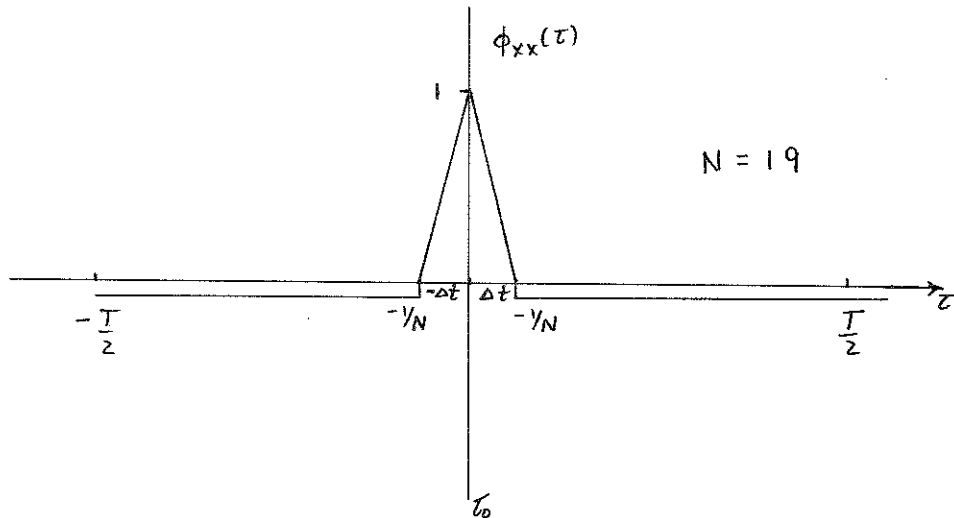


Fig. 87

when τ equals τ_0 . It is zero after an interval Δt and then for larger values of time it is $-1/N$ where N is the number of intervals in one period T . Because it is a specially designed signal it is not a purely random signal, although if N is large it approaches a random signal. Not all binary signals have this particular autocorrelation property. One requirement of this property is that N is primary and $N = 4K - 1$. Some values of N which work are 19, 251, and 1019. Thus the autocorrelation function for this particular input $x(t)$ is very close to a delta function; i.e., the bandwidth of the signal is very broad. In fact, for this particular case the half-amplitude values of the bandwidth are from $\frac{0.56}{T}$ to $\frac{0.33}{\Delta t}$.

Now suppose that we take a physical system (such as a nuclear reactor), excite it with this input and measure the output. According to what we discussed before, it is always possible to make an orthogonal power series expansion of the output in terms of the input. Then we can express the output of

the reactor, or whatever the physical system is, as

$$y(t) = \sum_{i=0}^n G_n(H_n, x, t) \quad , \quad (308)$$

where G_n is an arbitrary functional depending on the kernel H_n , the input $x(t)$, and time. We have already discussed how to crosscorrelate the input and output for this system. In Lecture No. V, in the section on crosscorrelation, the experiment described in Figure 71 is exactly the same experiment we wish to describe here. Thus, the diagram of our experiment is just a repeat of Figure 71 and is shown in Figure 88. By analyses incorporating the principles of Lecture No. VIII, we will determine what the output $z(\tau)$ of the integrator is.

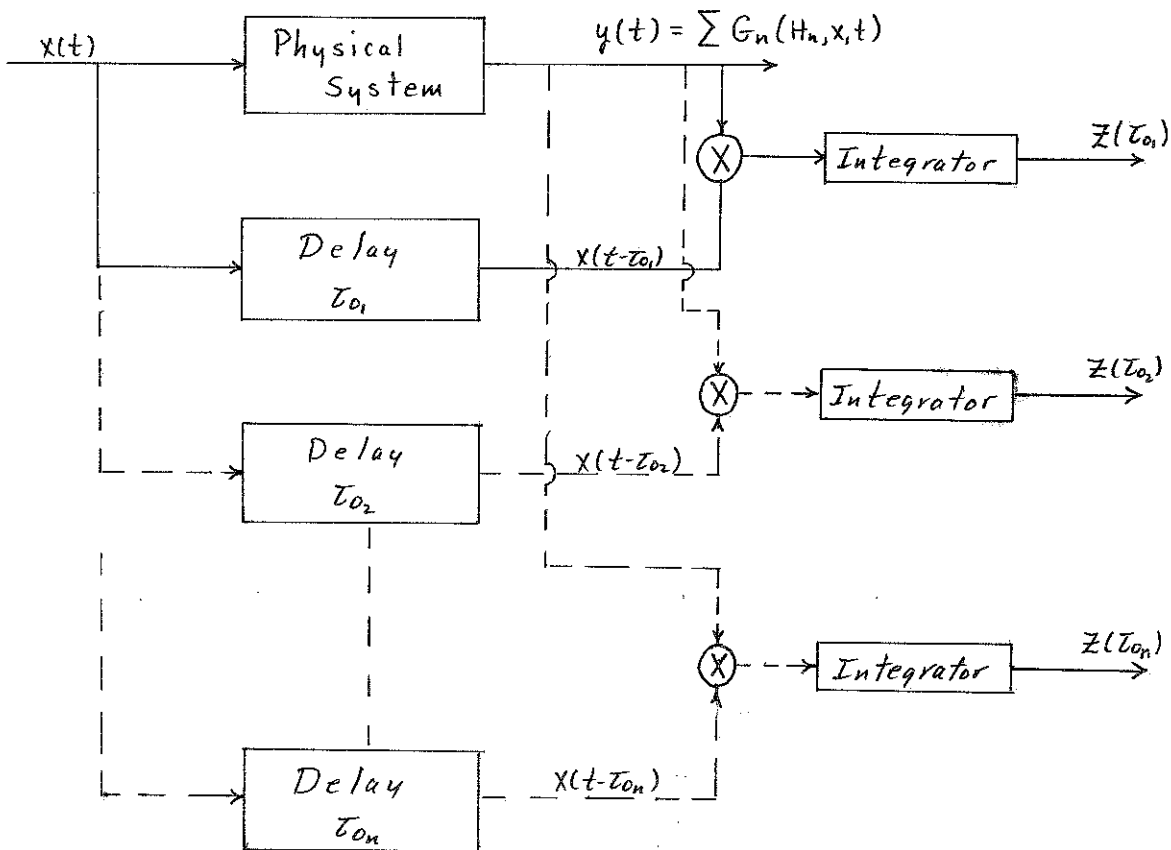


Fig. 88

As we discussed in Lecture No. VIII, by feeding the same input to a known box which is in parallel with the reactor, multiplying the outputs of the two systems and averaging, we can conceptually determine experimentally the kernels associated with the unknown physical system. This information is then the output of the integrators. We will now see if we can do this.

We have chosen the known box to be a simple delay line with output $x(t - \tau_{01})$. Since we can express this output as a first order functional which is orthogonal to all constants and all higher order polynomials involved in describing the input, we can write

$$\begin{aligned} Z(\tau_{01}) &= \frac{1}{T} \int_0^T dt \int_0^T H_1(\tau) x(t-\tau) x(t-\tau_{01}) d\tau \\ &= \frac{1}{T} \int_0^T H_1(\tau) d\tau \int_0^T x(t-\tau) x(t-\tau_{01}) dt, \end{aligned} \quad (309)$$

where it is assumed that it is necessary to consider integration only over one period T . Now the integral $\frac{1}{T} \int_0^T x(t-\tau) x(t-\tau_{01}) dt$ is simply the autocorrelation function $\phi_{xx}(\tau)$ of the input $x(t)$. Thus, we can rewrite Equation (309) as

$$Z(\tau_{01}) = \int_0^T H_1(\tau) \phi_{xx}(\tau) d\tau. \quad (310)$$

From Figure 87 and the fact that the interval $2\Delta t$ is small compared with T for large N , we can effectively divide the integral of Equation (310) into two parts. The first part considers $\phi_{xx}(\tau)$ has the value $-1/N$ over the interval from 0 to T , and the second part involves the product of $H_1(\tau)$ and the triangle of Figure 87 with height $\phi_{xx}(\tau) = 1$ and base $2\Delta t$. Assuming $H_1(\tau)$ does not vary much over the $2\Delta t$ interval, we can call it $H_1(\tau_{01})$ over this interval,

and we have

$$Z(\tau_{01}) = -\frac{1}{N} \int_0^T H_1(\tau) d\tau + H_1(\tau_{01}) \Delta t \quad (311)$$

The first term on the right-hand side of Equation (311) arises because $\phi_{xx}(\tau)$ is equal to $-1/N$, and the second term is explained in the following manner: Since $H_1(\tau)$ is considered a constant over the $2\Delta t$ interval, its value is $H_1(\tau_{01})$. Since the spike in the autocorrelation function is very sharp, the value of $\phi_{xx}(\tau)$ in the $2\Delta t$ interval is effectively the height of the triangle. Integrating from 0 to T then merely results in multiplying $H_1(\tau_{01})$ by the area of the triangle which is $\frac{1}{2}(2\Delta t)(1)$ or Δt . In order for this to be valid the interval Δt must be appropriately chosen. That is, if $H_1(\tau)$ is a fast changing function then Δt must be very small.

Equation (311) is to be interpreted so that the term $H_1(\tau_{01})\Delta t$ is the system function at the value of the delay τ_{01} and is the major contribution to the value of $z(\tau_{01})$. The term $-\frac{1}{N} \int_0^T H_1(\tau) d\tau$ is a correction term and is a constant. We can evaluate the correction in the following manner. The variable τ_{01} is at our disposal; i.e., we can put it anywhere we like. Thus, let's shift τ_{01} by $-\Delta t$. Thus our autocorrelation function, which is multiplied by $H_1(\tau)$, looks as shown in Figure 89. For an autocorrelation function with the

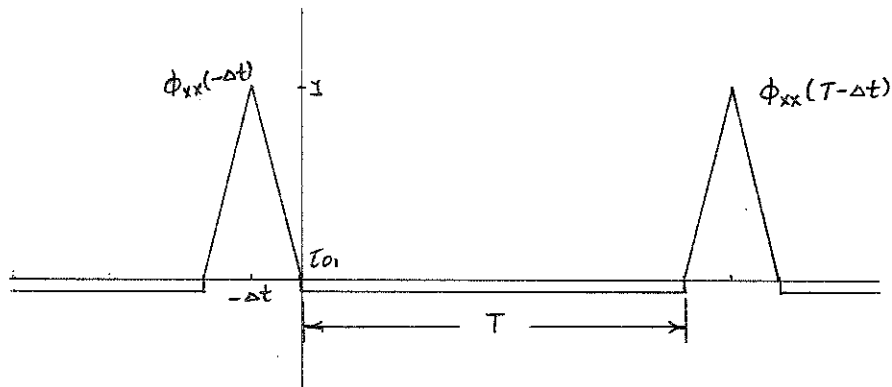


Fig. 89

shape shown in Figure 89, when we form the output z of the integrator we obtain only the correction term. This gives

$$\bar{z}(-\Delta t) = -\frac{1}{N} \int_0^T H_1(\tau) d\tau \quad (312)$$

Thus, the output of the integrator is a measure of the correction term. As was stated before, for a given N the value of the correction term is a constant. From Equation (312) we see that increasing the value of N decreases the magnitude of the correction term.

In reviewing what we have done analytically, we see that for a known box which is a delay line we have obtained the value of the kernel $H_1(\tau_0)$ of the physical system, plus a correction term. By performing only one more experiment we can evaluate the magnitude of the correction term. Thus, we determine the linear portion of the response of the physical system for a particular delay time in the known system. Repeating this procedure for different delay times would give us values of the linear system function as a function of time. We see then that we get only the linear portion of the response by the method of crosscorrelation.

An experiment of this type was actually performed with the KIWI A3 reactor. With this reactor, the time available for the experiment was only one minute. A block diagram of the experimental set-up is shown in Figure 90.

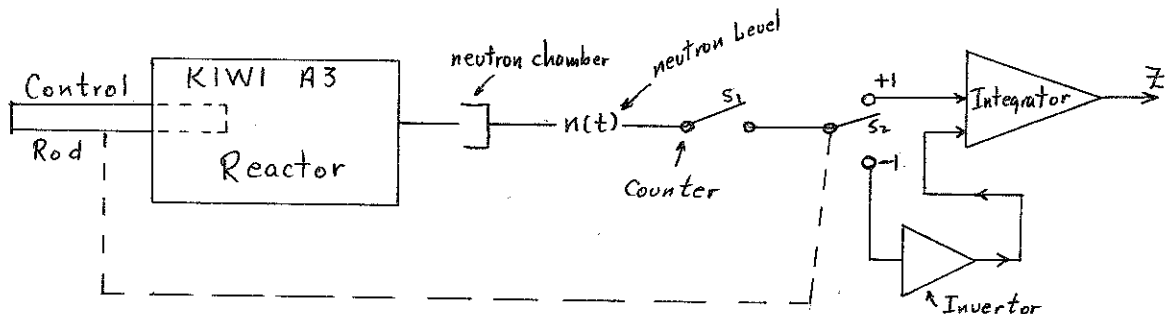


Fig. 90

The input signal was fed to the system by means of a paper tape. A representative diagram of the tape is shown in Figure 91 for a value of N equal to 19, although in the actual experiment, N was 251 and 1019. In Figure 91 only the input signal and the signal repeated for a shift of $3\Delta t$ are shown. Thus, in this representative case, τ_0 is equal to $3\Delta t$. The holes in the tape indicate

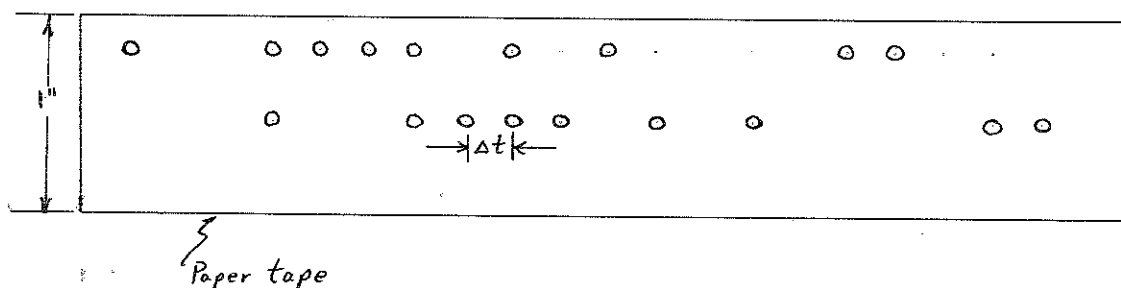


Fig. 91

$+1$ and each Δt where there is no hole indicates -1 . During the experiment, the paper tape moves through a reader which contains a light source. As the tape passes through, if light passes through a hole in the input signal channel a signal is fed, via a photomultiplier tube, to the control rod. This moves the control rod into the reactor. When no light passes through the input signal channel of the tape, the control rod is actuated to be pulled out of the reactor. Thus, the motion of the control rod changes the neutron level of the reactor. The delayed signal is also monitored, but by a separate photomultiplier tube. When light passes through the delayed signal channel the switch S_2 is in position $+1$ and the output of the reactor $n(t)$ goes directly into the integrator. When light does not pass through the delayed signal channel, the switch S_2 goes to position -1 . The output $n(t)$ of the reactor is then fed into an amplifier which inverts the sign of the signal. The output of the inverter is then fed to the integrator which averages over the time T . During the time T the switch S_1 is closed. After T ; i.e., when 251 or 1019 time increments have elapsed, S_1 opens and no signal is fed to the integrator. For this experiment then, the output of the integrator is $\sim H_1(\tau_0)$, with

$$\begin{aligned}
 & - \frac{1}{16} e^{-2j\omega t} H_3(-\omega, \omega, -\omega) + \frac{1}{16} H_3(\omega, -\omega, -\omega) - \frac{1}{16} e^{-2j\omega t} H_3(\omega, -\omega, -\omega) - \\
 & - \frac{1}{16} e^{-2j\omega t} H_3(-\omega, -\omega, -\omega) + \frac{1}{16} e^{-4j\omega t} H_3(-\omega, -\omega, -\omega) + \dots \quad (322)
 \end{aligned}$$

Averaging over one period (or any integral number of periods) gives

$$\frac{\omega}{2\pi} \int_0^{\frac{2\pi}{\omega}} e^{nj\omega t} dt = \frac{1}{jn2\pi} e^{nj\omega t} \Big|_0^{\frac{2\pi}{\omega}} = \frac{e^{nj2\pi} - 1}{jn2\pi} = 0$$

for n an integer equal to or greater than unity. Thus, all terms involving $e^{\pm nj\omega t}$ vanish and the time average of $y(t) \sin \omega t$ is given by

$$\begin{aligned}
 y(t) \sin \omega t &= \frac{1}{4} H_1(\omega) + \frac{1}{4} H_1(-\omega) + \frac{1}{16} H_3(-\omega, \omega, \omega) + \frac{1}{16} H_3(\omega, -\omega, \omega) + \\
 &+ \frac{1}{16} H_3(-\omega, -\omega, \omega) + \frac{1}{16} H_3(\omega, \omega, -\omega) + \frac{1}{16} H_3(-\omega, \omega, -\omega) + \\
 &+ \frac{1}{16} H_3(\omega, -\omega, -\omega) + \dots \quad (323)
 \end{aligned}$$

By Equation (323) and the forms of the previous equations, it is easily seen that if this method were carried out even further, one would obtain an expression involving only the odd kernels. Therefore, all orders of the odd harmonics would appear.

Suggested Reference

1. N. Wiener, Nonlinear Problems in Random Theory; Massachusetts Institute of Technology Press, 1958.