UNCLASSIFIED

AD 626 450

I. SERVOMECHANISMS THEORY

The Johns Hopkins University
Silver Spring, Maryland

May 1958

UNCLASSIFIED

NOTICE TO DEFENSE DOCUMENTATION CENTER USERS

This document is being distributed by the Clearinghouse for Federal Scientific and Technical Information, Department of Commerce, as a result of a recent agreement between the Department of Defense (DOD) and the Department of Commerce (DOC).

The Clearinghouse is distributing unclassified, unlimited documents which are or have been announced in the Technical Abstract Bulletin (TAB) of the Defense Documentation Center.

The price does not apply for registered users of the DDC services.

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

May 1958

I. SERVOMECHANISMS THEORY

A Familiarization Course

Given by

The Bumblebee Controls Group

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

FOREWORD

This report documents the first of a series of Familiarization Courses sponsored by the Applied Physics Laboratory Committee on Education.

The aim of the series is the instruction of staff members in operations and techniques outside their own areas of specialization leading to a ready integration of ideas between fields and to the establishment of sound bases for further inquiry.

The four lectures included herein were designed to provide tools basic to an understanding of servomechanisms theory. A class numbering 36 was in attendance at the talks given by members of the Laboratory's Bumblebee Controls Group during July and August of 1957.

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

TABLE OF CONTENTS

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

Lecture 1

BUMBLEBEE MISSILE CONTROL SYSTEMS

by

W. A. Good

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

Bumblebee Missile Control Systems

by W. A. Good

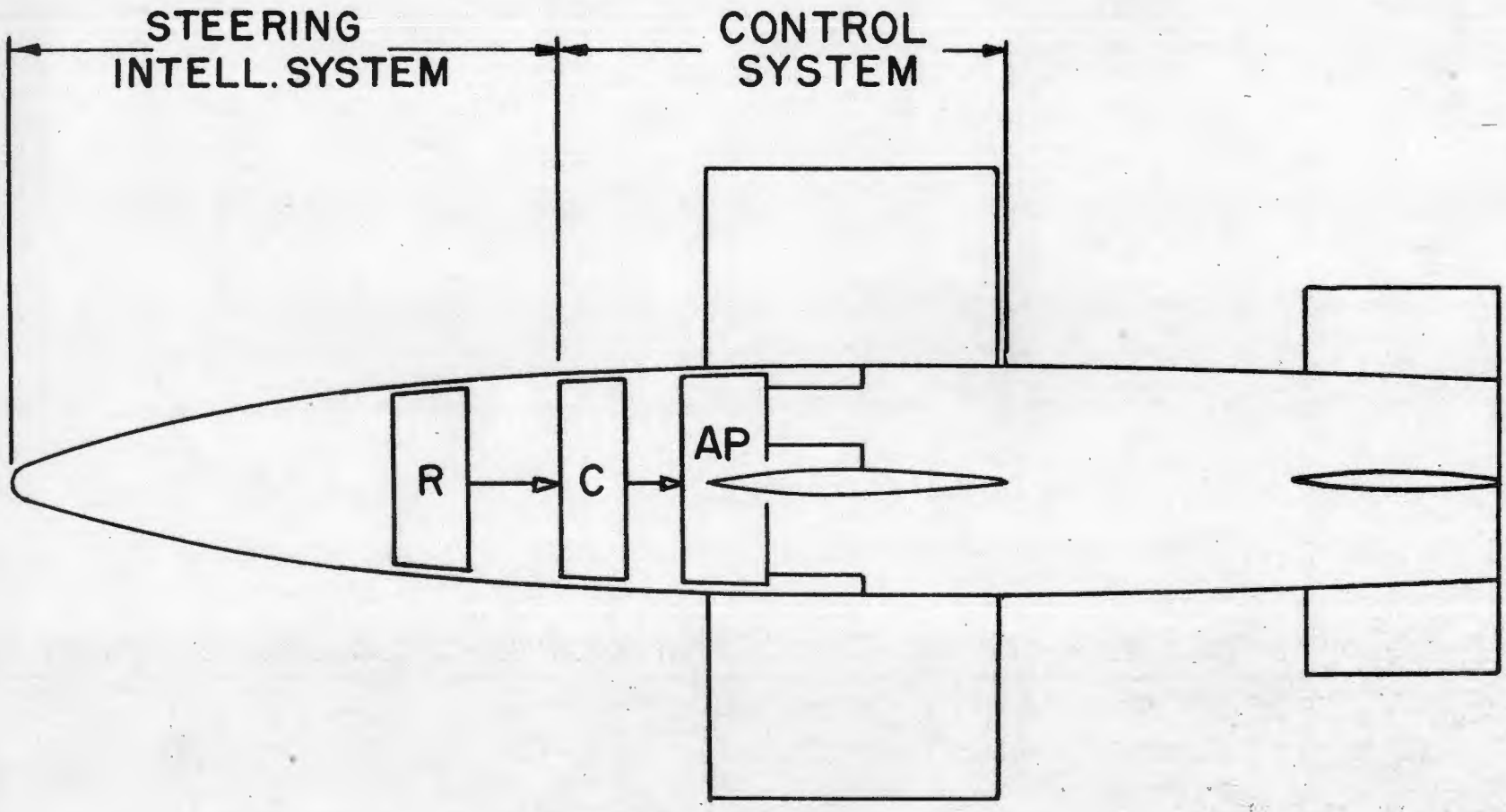
The control system of a missile is usually defined as that portion of the system which accepts signals from the intelligence section, such as the receiver, and converts this information into proper wing deflections to steer the missile in the desired manner. See Figure 1.

A natural division exists within the control system which divides it into two sections, the computer and the autopilot. The computer must modify the intelligence signals into suitable lateral acceleration commands. The autopilot accepts the acceleration commands and alters them into appropriate wing deflections such that the airframe will provide the called-for accelerations. Thus, it is seen that the properties of the computer are related to the overall guidance loop whereas the autopilot characteristics are involved with the aerodynamics features of the airframe.

The roll stabilization system is usually considered as a part of the missile control system but will not be considered in this paper.

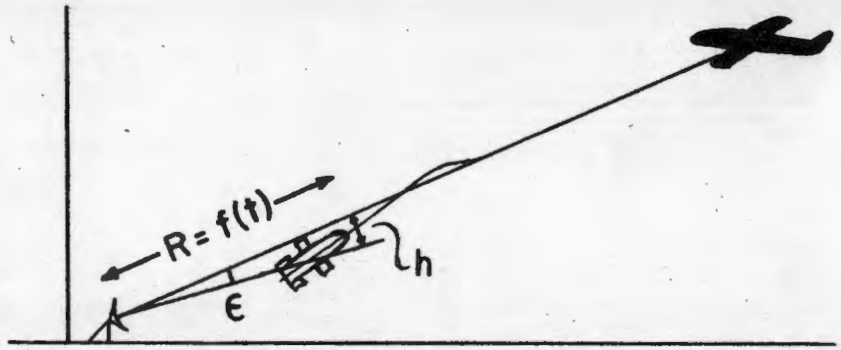
Let us first turn to the computer section and examine its duties in a beamrider guidance system. Figure 2 presents the basic elements of the beamrider path and shows that the input to the computer is a voltage whose amplitude is a measure of the off-beam angular error. The computer must modify this angular error and produce a lateral acceleration command. First, the angular error is translated to a distance off-beam error by multiplying by range or actually by a function of time since the expected missile velocity is reasonably well known. This would be the only role of the computer if the problem of path stability did not arise. However, since we have an acceleration proportional to off-beam distance and the distance, in turn, is equal to the second integral of the acceleration, the result will be an undamped oscillation about the beam, analogous to a mass and spring without damping. This is undesirable and the necessary path damping may be supplied by a shaping network with a derivative characteristic shown in Figure 2. Unfortunately, the derivative is accompanied by low gain at low frequencies and high gain at high frequencies. This condition is just the reverse of what is desired since noise signals at high frequencies should be attenuated and the gain at low frequencies should be high to minimize biases due to gravity and other misalignments. This situation is corrected to some extent by filter networks. The noise filter in this case is a simple network operating at frequencies above the derivative region and the low frequency gain is provided by an integral network reaching back to as low as 1/50 rad/sec. Typical overall gain values in g's per 100 feet vary from 1/2 to perhaps 5 in some cases.

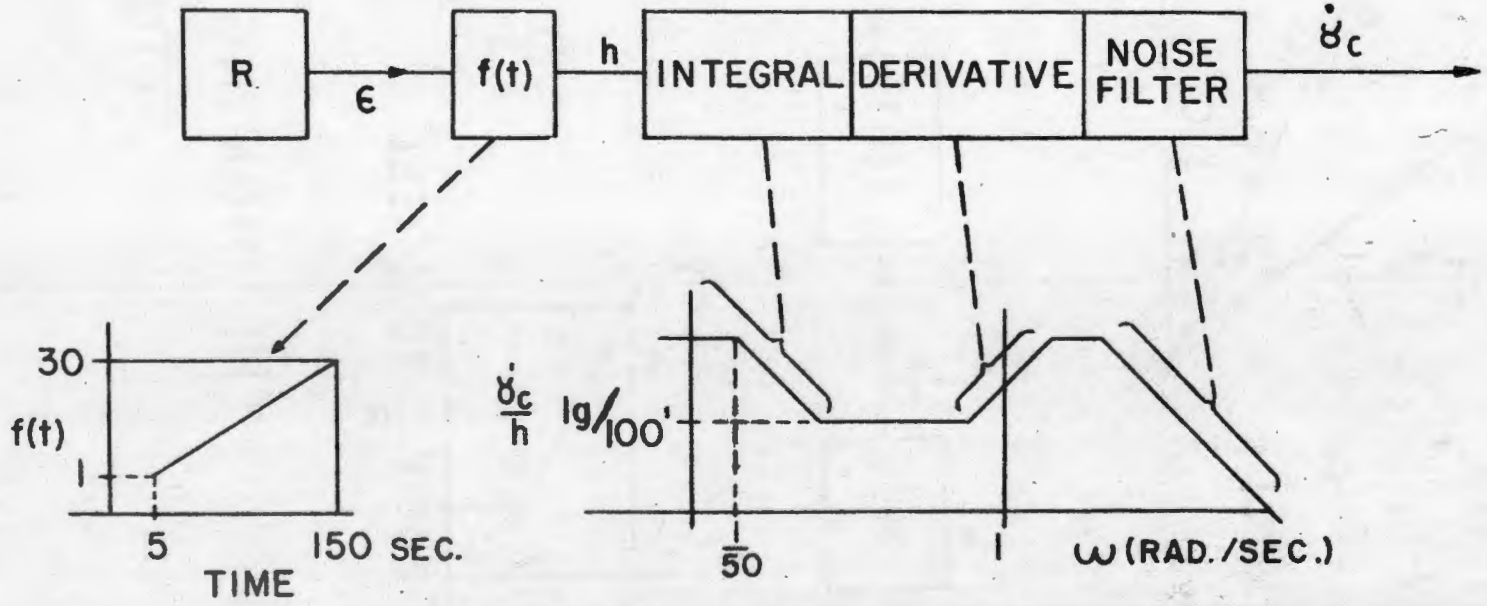
Next let us look at the type of computer used in a homing missile. Figure 3 shows the essential parameters of the homing process. The quantity sigma represents the angle between the missile-target line-of-sight and a fixed reference. An ideal homing intelligence system is capable of producing a measure of the rate of change of sigma with respect to time. If the missile can be made to align its velocity vector such that sigma-dot is always zero, a perfect intercept course will result. Ideally, the computer needs only to produce an acceleration command proportional to sigma-dot. However, in actual



MISSILE CONTROL SYSTEM

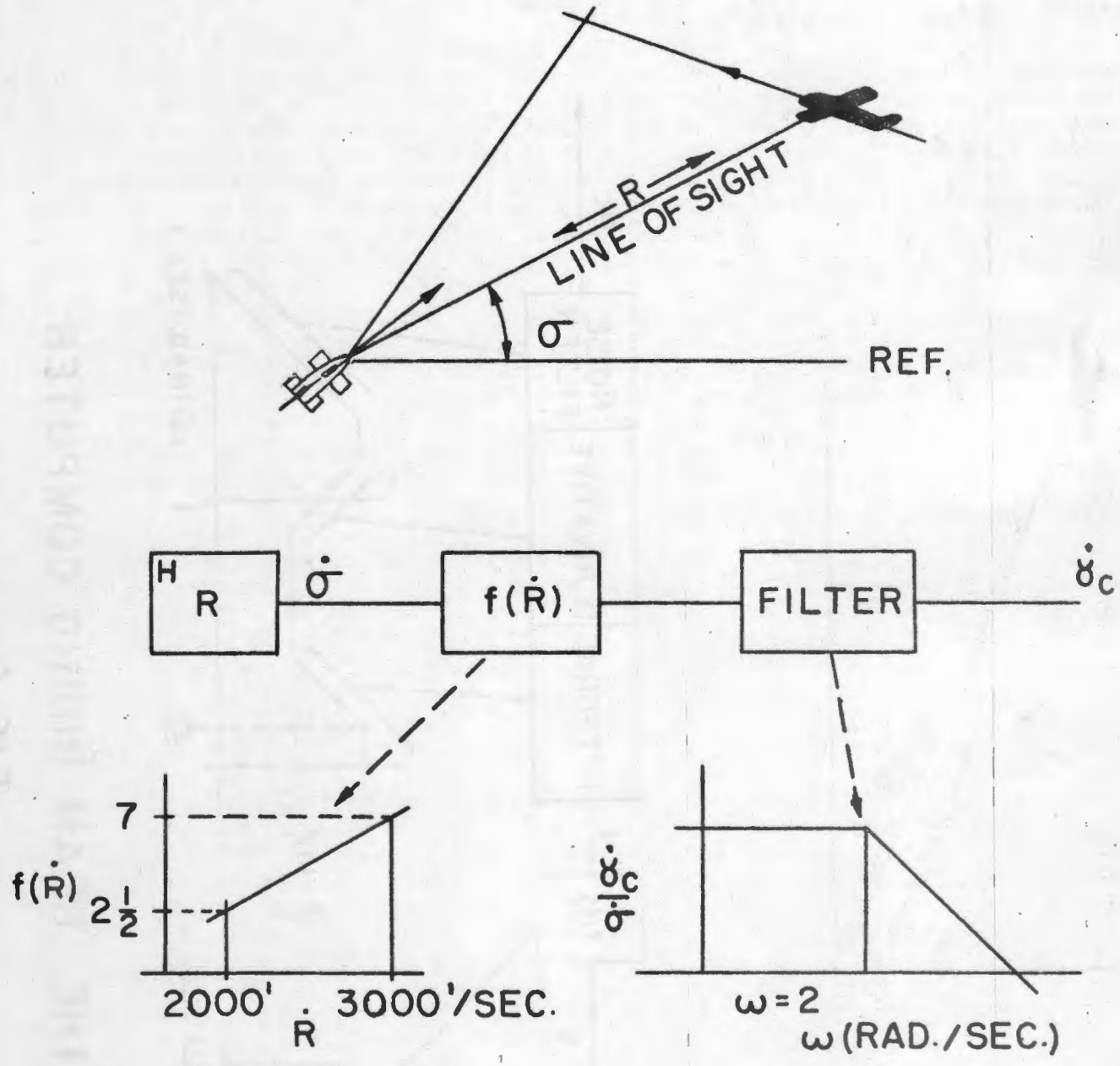
FIG. 1





THE BEAM RIDING COMPUTER

FIG. 2



THE HOMING COMPUTER

FIG. 3

practice two modifications to this simple requirement are usually employed. Since the sigma-dot signal contains unwanted noise signals, a filter of about one-half second time constant is used. If this filter is too large, path instability or inaccurate homing will result; if too small, excessive noise will reach the autopilot. The other modification is to vary the proportionality between sigma-dot and acceleration command as a function of the range-rate between the missile and target. Thus, for a rapidly closing attack, the homing gain will be higher and the missile will be more responsive since it has less time to cope with the target.

In many ways the autopilot has a much tougher task to perform than the computer. Regardless of altitude, speed, center of gravity shift or type of airframe the autopilot must properly direct the control surfaces to produce the called-for acceleration in a rapid and stable manner.

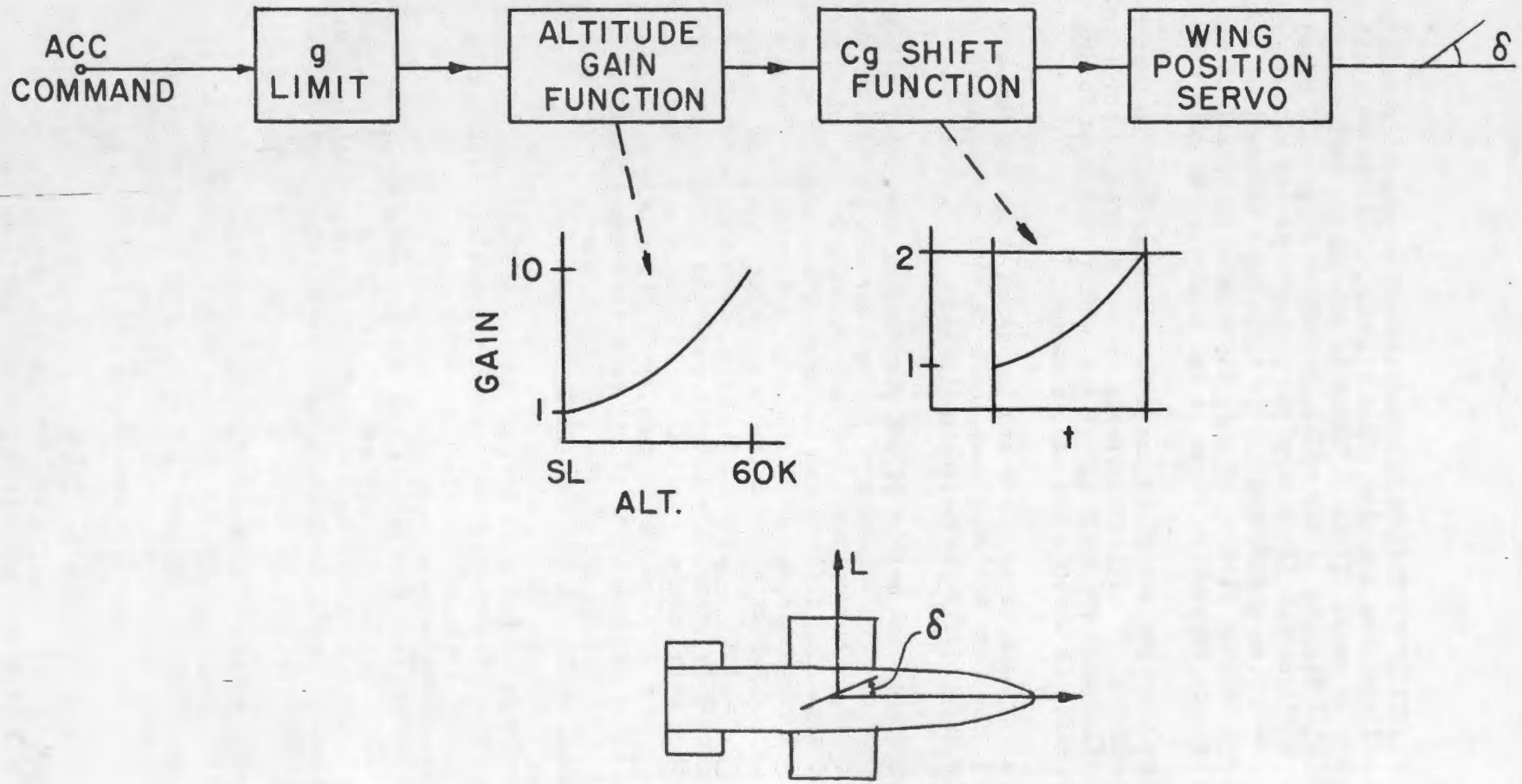
Several types of autopilots are in common usage and three of these will be described. They are the wing position type, the acceleration feedback type (AFB) and the sensitivity feedback type (SFB).

The wing position method is the simplest of the three systems, but requires a great deal of foreknowledge of the aerodynamic characteristics of the missile and of the flight conditions. One must be able to predict under all flight conditions exactly what wing deflection will provide the desired acceleration or aerodynamic gain. For example, a one degree wing deflection produces a certain "g" turn at sea level, whereas a ten degree wing angle is needed to cause the same "g" turn at 60,000 feet altitude. Figure 4 shows this autopilot where a gain adjustment is made to compensate for the altitude effect. Pressure gauges may be used to effect this change. Another variation in aerodynamic gain (g's per degree of wing) is due to center of gravity (cg) shift as a result of fuel consumption. A missile may start its flight in a tail heavy condition and end with a nose heavy distribution due to the burning of the solid rocket. Thus, it becomes less maneuverable at the end of flight and a gain change must be made to compensate. In this case the cg shift as a function of time is well known and hence a corresponding gain change as a function of time will suffice. Although the missile speed also influences the aerodynamic gain, it is more difficult to measure and will not be included in this discussion.

Another important function of the autopilot is not only to produce the called-for acceleration but to preserve the integrity of the airframe in the process. A g-limiter is added to the front of the autopilot to clip any excessive commands which might otherwise cause unusual wing deflections and subsequent structural damage.

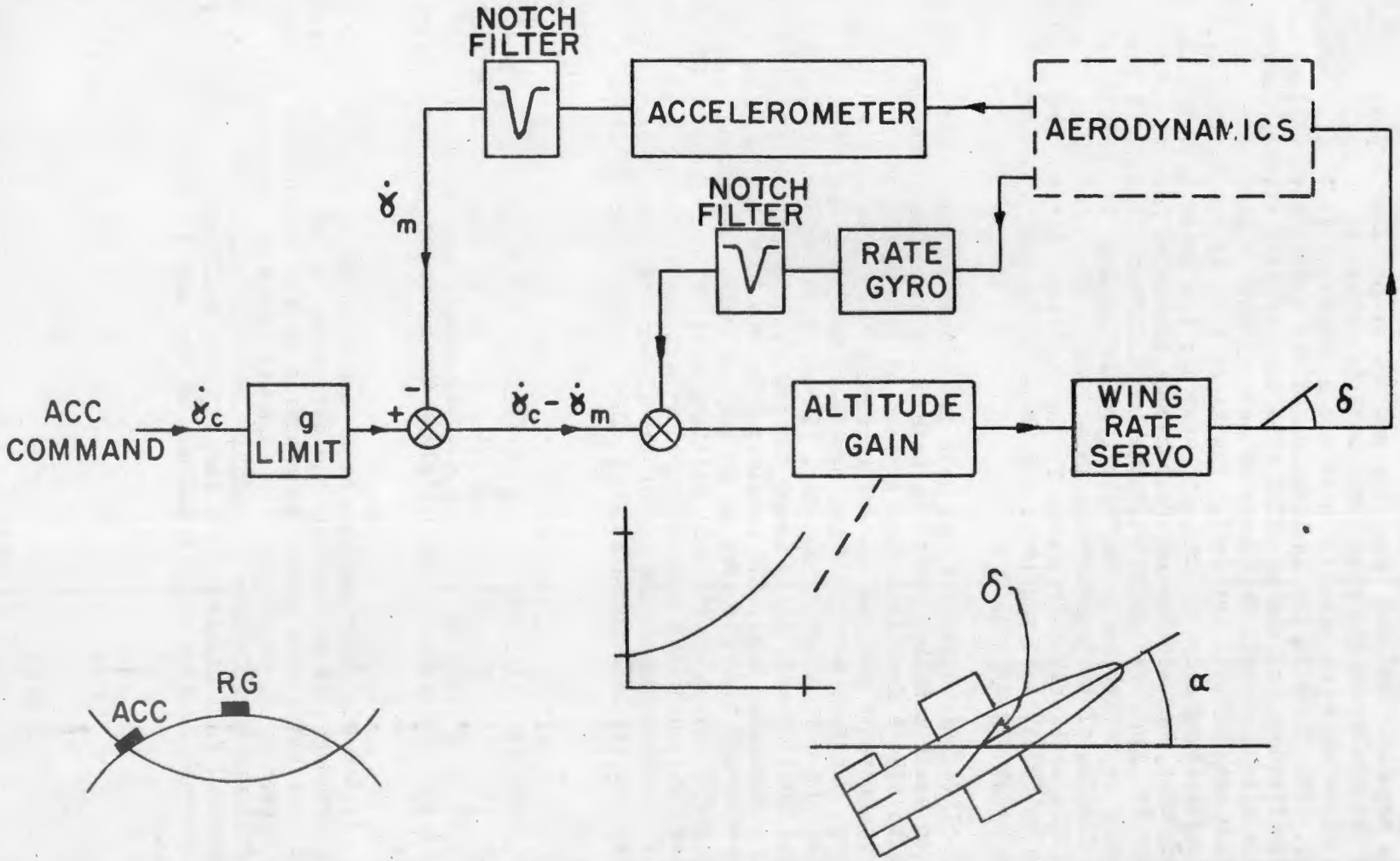
The brief review of the wing position autopilot should infer that it is most effective when used with an airframe with linear aerodynamic characteristics such as a wing controlled missile with small body angle of attack and over restricted speed range. Outside of these conditions, it becomes more difficult to closely match the output acceleration to the command.

An autopilot in which the actual acceleration of the missile maneuver is fed back and matched against the command would seem to be the most logical system. This is the acceleration feedback (AFB) autopilot and is very effective, but not without limitations. In the simplest form shown in Figure 5,



THE WING POSITION AUTOPILOT

FIG. 4



THE ACCELERATION FEEDBACK AUTOPILOT

FIG. 5

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

the wing servo should be of the rate type such that for low frequencies the loop gain through the servo is very high. Under these conditions the AFB system will act like a negative feedback amplifier and the output acceleration will closely match the command since the wing angle will be driven to an angle which will cause the mismatch to be small. There are a number of stability problems which now appear. The first is the inclination of the system to oscillate at the weathercock frequency of the missile. The uncontrolled missile behaves like a large arrow with a natural frequency of several cycles per second and a damping of less than one-tenth of critical. Now since part of the missile lift comes from the oscillating body, it is quite easy for the phase lags through the servo, weathercock mode, and accelerometer to become excessive near the sharply resonant weathercock frequency and cause sustained oscillations in the AFB loop. The standard cure is to include a yaw rate gyro in the loop to provide deliberate synthetic damping of the weathercock mode. Now a step command will produce a rapidly damped acceleration response in the missile. A rise time of about 0.1 second can be obtained at sea level and about 0.3 second at 60,000 feet altitude with a well-adjusted system.

Into this autopilot, we have introduced two instruments which are sensitive to their mounting positions within the airframe structure. Since the airframe acts like a free-free bending beam at frequencies which the system can pass, the instruments can pick up and introduce the extraneous vibration into the AFB loop. In fact certain steps must be taken to prevent the whole system from oscillating at the body frequency. In a typical case this could be about 30 cps. The ideal remedy is to mount the instruments at body stations where the pick-up is a minimum. Thus, the accelerometer should be attached at the positional node and the rate gyro at an angular node. When other requirements deny these positions, a second best palliative is the insertion of electrical notch filters following the instrument to reject the vibration frequency. Of course, the low frequency phase lag of the filter gets into the low frequency AFB loop and eats up valuable phase margin. This is the penalty paid for the notch filter.

It should be pointed out that the AFB autopilot must also work properly for the same altitude and cg variations discussed for the first autopilot. This calls for the addition of at least an altitude-gain device, but it need not be as precise as before since the closed loop effect helps minimize the internal gain variations. Nevertheless the altitude-gain device cannot be omitted.

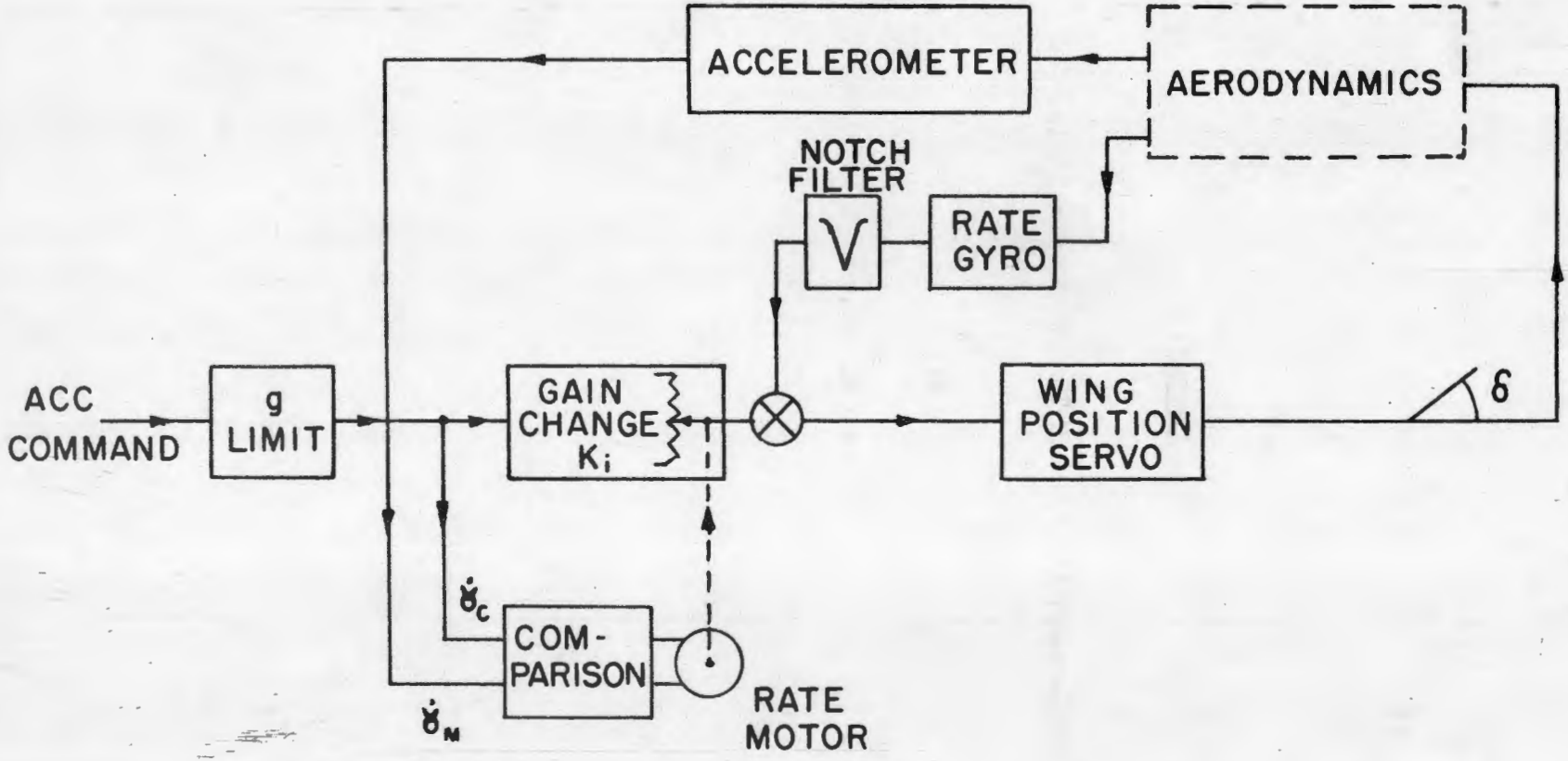
If the wing servo cannot be a pure rate servo, then the low frequency match between the command and the output may be poor and more sensitive to internal loop gain variations. In practice this effect may be produced due to the necessity of using position type wing servos. The position type servo was required to eliminate a possible ambiguity of the four wing positions when steering and roll commands were zero. There would exist a number of wing position combinations which would not be zero and still give zero roll and zero steering.

A nonlinear aspect of the AFB loop is the maximum rate of the wing servo. If this is too high the servo power consumption is excessive; if it is too low, the AFB loop may participate in a destructive limit-cycle type of oscillation. The choice of wing rate is pressed between these two conditions. Typical values might be 100 - 200°/sec. as an adequate compromise.

The third autopilot to be discussed is the Sensitivity Feedback (SFB) type. It retains some of the features of the other two. The wing position servo is retained along with a single gain change device which is the heart of the system. The rate gyro still provides weathercock damping but is less critical than before because the accelerometer signal is fed into a different place and does not destabilize as much as before. Note that the accelerometer signal ($\dot{\gamma}_M$) is now compared with the command ($\dot{\gamma}_c$). The smoothed difference signal causes the rate motor to run in one direction or the other depending on whether the accelerometer is reading larger or smaller than the command. Aerodynamic variation may demand a 20:1 gain change which compensates for the altitude and other variations in the aerodynamic gain. Ideally the SFB adapts its position to always provide the correct gain to match the output acceleration with the command. In practice the absence of a command allows no information to the rate motor and it tends to remain at its last position until sufficient commands do arise. Also the gain may seek an improper value if biases of an aerodynamic nature exist in the loop. (See Fig. 6.)

An attempt has been made in this discussion to isolate the autopilot from the overall guidance loops. In the main this is possible for beamriding but not completely so in the homer. In the latter case certain receiver signals come through the computer and into the autopilot and tend to interfere with the rate gyro contribution such that the weathercock damping is disturbed. Normally additional overall system design studies are required to resolve this type of problem.

In summary a missile control system has been broken into its computer and autopilot sections. The computer has been related to the overall guidance loop problem, whereas the autopilot has been confined to the job of providing accurate and rapid acceleration response in spite of the variations in aerodynamic gain. The three autopilots have been compared on functional and conceptual bases and their relative merits have been described.



THE SENSITIVITY FEEDBACK AUTOPILOT

FIG. 6

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

Lecture 2

THE LAPLACE AND FOURIER TRANSFORMS AND THEIR USES

by

John M. LeGaré

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

THE LAPLACE AND FOURIER TRANSFORMS AND THEIR USES

by John M. LeGaré

The Laplace and Fourier transforms are no more than mathematical tools whose inherent characteristics facilitate operations involving differential equations, but as such they have become indispensible to the servomechanisms engineer. The principal advantages of these transforms are that first, by transforming functions of a real variable (usually time) into functions of a complex variable, mathematical manipulation becomes considerably simplified and, secondly, there are certain incidental but highly useful parallels between these resulting functions of a complex variable and certain more conventional concepts. This presentation will for the most part restrict itself to the Laplace transform with a concluding portion summarizing the essential differences between the two transform types.

In order to more clearly depict the role of the LaPlace transform, a brief review of the linear differential equation is appropriate. The study of dynamic systems including closed-loop servo mechanisms is basically the study of differential equations of time variables. For the purpose of using the Laplace transform in the study of such systems, we shall restrict ourselves to those characterized by linear differential equations with constant coefficients. The general form of equation representing such a system is as follows:

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + \text{---} + a_n x = \frac{d^m y}{dt^m} + b_1 \frac{d^{m-1} y}{dt^{m-1}} + \text{--} + b_m y, \quad (1)$$

where x may be the output of some dynamic device and x its input function of time.

It will be recalled that the differential obeys the distributive and associative laws of algebra, so that the equation may be written in the form

$$\left(\frac{d^n}{dt^n} + a_1 \frac{d^{n-1}}{dt^{n-1}} + \text{---} + a_n\right) x = \left(\frac{d^m}{dt^m} + b_1 \frac{d^{m-1}}{dt^{m-1}} + \text{---} + b_m\right) y \quad (2)$$

or, in symbolic notation,

$$\left(D^n + a_1 D^{n-1} + \text{---} + a_n\right) x = \left(D^m + b_1 D^{m-1} + \text{---} + b_m\right) y \quad (3)$$

Also recall that the solution of such a differential equation consists of two parts, a so-called transient or complimentary solution plus the particular integral. The transient solution, furthermore, always has the form.

$$x_c = A_1 e^{s_1 t} + A_2 e^{s_2 t} + \text{---} + A_n e^{s_n t} \quad (4)$$

This may be shown by substituting $x = e^{st}$ into the left-half of the differential equation, and remembering that the n-th derivative of e^{st} is $s^n e^{st}$, we find

$$s^n e^{st} + a_1 s^{n-1} e^{st} + a_2 s^{n-2} e^{st} + \text{---} a_n e^{st} = 0 \text{ or,}$$

$$(s^n + a_1 s^{n-1} + a_2 s^{n-2} + \text{---} a_n) e^{st} = 0, \text{ and since}$$

e^{st} is not identically zero, we may divide both sides by it, leaving

$$s^n + a_1 s^{n-1} + a_2 s^{n-2} + \text{---} + a_n = 0 \quad (5)$$

Now we have a polynomial in s for which the roots $s_1, s_2, \cdots s_n$ may be found, giving the values in the complimentary solution (4) above. The solution of the particular integral may be found by classical means such as the method of undetermined coefficients.

As an example of this process, consider a simple spring, mass, and damper to which a force, y(t), is applied. Then the motion of the mass, x (t), is expressed in the following manner:

$$M \frac{d^2x}{dt^2} + C \frac{dx}{dt} + Kx = y(t) \qquad (6)$$

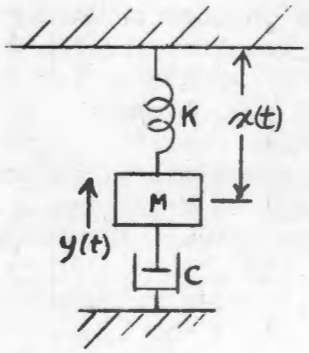


Figure 1

Where M = mass

C = damping coefficient

K = spring constant

Then the transient solution is

$$(M D^2 + CD + K)\ x = 0 \qquad (7)$$

$$D = \frac{-C \pm \sqrt{C^2 - 4MK}}{2M}$$

$$x = A_1\ e^{\left(-\frac{C}{2M} + \sqrt{\frac{C^2}{4M^2} - \frac{K}{M}}\right)t} + A_2\ e^{\left(-\frac{C}{2M} - \sqrt{\frac{C^2}{4M^2} - \frac{K}{M}}\right)t} \qquad (8)$$

If the discriminant is positive then the exponent will be real, and the solution of the form

$$x = A_1\ e^{\sigma_1 t} + A_2\ e^{\sigma_2 t}\ , \qquad (9)$$

and if the discriminant is negative, then the exponents are complex, as below:

$$x = A_1\ e^{(\sigma_1 + j\omega_1)t} + A_2\ e^{(\sigma_1 - j\omega_1)t} \qquad (10)$$

This equation may be interpreted to say that if the spring-mass-damper system above is given any set of initial conditions, such as an initial position and velocity, the transient behavior without any forcing function will behave as in equation (8) with appropriate constants A_1 and A_2. Also recall the Euler relation which allows us to interpret the exponents as having the following significance.

$$A_1\ e^{(\sigma_1 + j\omega_1)t} = A_1\ e^{\sigma_1 t}\ e^{j\omega_1 t} = A_1\ e^{\sigma_1 t}\ (\cos \omega_1 t + j \sin \omega_1 t)$$

Thus the transient solution, equation 10, represents a damped sinusoid (if the value of σ_1 is negative) whose decrement factor σ_1 and frequency ω_1 are functions of the system constants M, C and K as in equation (8), and whose phase and amplitude are functions of the initial conditions.

The forced solution or particular integral will be linearity added to this transient solution, and will be of the same form as the input time function and its derivatives.

Similarly, if one examines the differential equation of the simple electrical circuit consisting of an inductance, resistance and capacitance, the differential equation of the output voltage resulting from an input voltage is determine as follows:

$$e_o = \frac{Q}{C}$$

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{Q}{C} = e_i$$

$$LC\frac{d^2e_o}{dt^2} + RC\frac{de_o}{dt} + e_o = e_i \qquad (11)$$

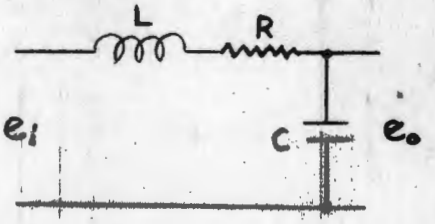


Figure 2

We notice immediately that the differential equation (11) of this system is identical in form to that of the spring-mass-damper system equation (6) and that thus the solution would be of the same form, with only the various constants involved taking on a different significance. A slightly different result exists when L = 0. Under these conditions, the transient solution becomes

$$e_o = A_1 e^{-\frac{t}{RC}} \qquad (12)$$

Also under these conditions, if one were to impress sinusoidal driving voltage into this network, then by classical a.c. theory the ratio of output voltage to input voltage would be, in vector form,

$$\frac{e_o}{e_i} = \frac{-jx_c}{R - jx_c} = \frac{\frac{1}{jwc}}{R + \frac{1}{jwc}} = \frac{\frac{1}{RC}}{jw + \frac{1}{RC}} \qquad (13)$$

Let us keep the results (12) and (13) in mind to observe their relation to the Laplace transform method.

Now what is the Laplace transform and what relation has it to these differential equations? The mathematical definition of the direct Laplace transform is stated mathematically as -

$$\mathcal{L}[f(t)] \triangleq \int_0^\infty f(t)\, e^{-st}\, dt \qquad (14)$$

where s is the complex variable $\sigma + jw$.

Similarly the inverse transformation is

$$\mathcal{L}^{-1}[F(s)] (=) \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} F(s)\, e^{st}\, ds \qquad (15)$$

where s is the complex variable $\sigma + jw$ and c in this integration is the abscissa of convergence whose significance will be discussed later.

By means of the above relations, one may tabulate Laplace transforms corresponding to various useful time functions, some of which are given in the appendix. In order to observe one relation between the Laplace transform and the differential equation solutions considered before, let us take the Laplace transform of the simple decaying exponential term

$$f(t) = Ae^{-at} \tag{16}$$

the transform becomes

$$\mathcal{L}[Ae^{-at}] = \int_0^\infty Ae^{-at}e^{-st}dt = \int_0^\infty A\,e^{-(s+a)t}dt$$

$$= -\frac{A}{s+a}\,e^{-(s+a)t}\Big|_0^\infty = \frac{A}{s+a} \tag{17}$$

Note that the time function (16) is similar to that of the transient solution (12) if a = 1/RC and the Laplace transform (17) bears a similar relation to the steady state sinusoidal response (13) if s = jw. This similarity will be examined further after one observes another fortunate relation between the Laplace transform and the differential equation. In order to point out this relation, let us take the Laplace transform of three highly useful time functions, the unit ramp, the unit step, and the unit impulse function.

The unit ramp is a time function which may be defined thus:

$$r(t-T) = \begin{cases} 0, & t<T \\ t-T, & t>T \end{cases}$$

$$\mathcal{L}[r(t-T)] = \int_0^\infty r(t-T)e^{-st}dt = \int_T^\infty (t-T)\,e^{-st}dt$$

This integration may be accomplished by means of integration by parts, using the formula

$$\int u\,dv = uv - \int v\,du.$$

Let $u = t - T$, then $du = dt$

$dv = e^{-st}\,dt$, $\quad v = -\frac{1}{s}\,e^{-st}$.

$$\mathcal{L}[r(t-T)] = \left| -\frac{t-T}{s}\,e^{-st} + \int \frac{1}{s}\,e^{-st}dt \right|_T^\infty$$

$$= \left| -\frac{t-T}{s}\,e^{-st} - \frac{1}{s^2}\,e^{-st} \right|_T^\infty$$

$$= \frac{1}{s^2}\,e^{-sT}$$

(18)

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

We are usually interested in the special case where T = 0.

$$\mathcal{L}[r(t)] = \frac{1}{s^2} \tag{19}$$

Next, finding the Laplace transform of the unit step function, defined mathematically as

$$u(t-T) = \begin{cases} 0, & t<T \\ 1, & t>T \end{cases} .$$

$$\mathcal{L}[u(t-T)] = \int_0^\infty u(t-T)e^{-st}dt = \int_T^\infty 1\, e^{-st}dt$$

$$= -\frac{1}{s}e^{-st}\Big|_T^\infty = \frac{1}{s}e^{-sT} . \tag{20}$$

Again, we are usually interested in the case where T = o,

$$\mathcal{L}[u(t)] = \frac{1}{s} . \tag{21}$$

Lastly, let us find the Laplace transform of the Dirac delta-function or unit impulse function.

$$\delta(t-T) = \begin{cases} o & t<T \\ \infty & t=T \\ o & t>T \end{cases} \quad \text{and} \quad \int_{-\infty}^{\infty}\delta(t-T)\,dt = 1$$

$$\mathcal{L}[\delta(t-T)] = \int_0^\infty \delta(t-T)\,e^{-st}dt$$

Here, because of the extremely narrow range over which the δ function is non-zero, we may write

$$\mathcal{L}[\delta(t-T)] = e^{-sT}\int_{T-\epsilon}^{T+\epsilon}\delta(t-T)dt = 1\,e^{-sT} \tag{22}$$

And again, when T = 0, $\mathcal{L}[\delta(t)] = 1.$ (23)

Now, one notices a relation between the last three time function and between their Laplace transforms; i.e., the successive functions are each the derivative of the preceding or integral of the succeeding function, and the successive transforms are each s times the preceding or 1/s times the succeeding.

Hence, one is inclined to conclude an equivalence between differentiation and multiplication by s and between integration and multiplication by 1/s. Let us see what the precise relation is.

Here we denote the Laplace transform of f(t) by F(s)

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

Now let us accomplish the integration by parts, letting u = f(t) and dv = e^{-st} dt

$$F(s) = -\frac{1}{s} f(t)\, e^{-st}\Big|_0^\infty + \frac{1}{s}\int_0^\infty \left[\frac{df(t)}{dt}\right] e^{-st}\, dt$$

$$sF(s) = f(0+) + \int_0^\infty \left[\frac{df(t)}{dt}\right] e^{-st}\, dt$$

or,
$$\mathcal{L}\left[\frac{df(t)}{dt}\right] = sF(s) - f(0+) \tag{24}$$

This process may be extended to higher order derivatives yielding the general form

$$\mathcal{L}\left[f^{(n)}(t)\right] = s^n F(s) - \sum_{k=1}^{n} f^{(k-1)}(0+)\, s^{n-k}. \tag{25}$$

Similarly, one may arrive at the relation between Laplace transform and the process of integration in the time domain

$$\mathcal{L}\left[\int f(t)\, dt\right] = \frac{F(s)}{s} + \frac{f^{(-1)}(0+)}{s}, \tag{26}$$

or in its general form,

$$\mathcal{L}\left[f^{(-n)}(t)\right] = \frac{F(s)}{s^n} + \sum_{k=1}^{n} \frac{f^{(-k)}(0+)}{s} \tag{27}$$

These two important theorems, that of real differentiation and real integration, together with the fact that the Laplace transform is a linear operator

$$\left(\mathcal{L}[k f(t)] = k F(s) \quad \text{and} \quad \mathcal{L}[f_1(t) + f_2(t)] = F_1(s) + F_2(s)\right)$$

makes the Laplace transform a useful tool in the solution and interpretation of linear differential equations.

Let us again examine the general linear differential equation with constant coefficients.

$$\left(\frac{d^n}{dt^n} + a_1 \frac{d^{n-1}}{dt^{n-1}} + \text{---} + a\right) x(t) = \left(\frac{d^m}{dt^m} + b_1 \frac{d^{m-1}}{dt^{m-1}} + b_m\right) y(t)$$

Now for the particular condition where the initial conditions of the dependent variable, x, and its successive derivatives are zero, a particular disturbance y whose initial conditions are also zero is applied, then the Laplace transform of the equation becomes simply

$$(s^n + a_1 s^{n-1} + \text{---} + a_n) X(s) = (s^m + b_1 s^{m-1} + \text{---} + b_m) Y(s)$$

$$\text{or } X(s) = \frac{s^m + b_1 s^{m-1} + \text{---} + b_m}{s^n + a_1 s^{n-1} + \text{---} + a_n} \; Y(s). \qquad (28)$$

The rational fraction in s which operates upon Y (s) may be given the notation $H(s) = \frac{B(s)}{A(s)}$ and is called the system transfer function, indicating that an input Y(s) is transferred into an output X(s). The significance of H(s) is further clarified if the input y(t) is the Dilac delta-function $\delta(t)$, in which case Y(s) = 1. Under these conditions

$$X(s) = \frac{s^m + b_1 s^{m-1} + \text{---} b_m}{s^n + a_1 s^{n-1} + \text{---} a_n} \triangleq H(s). \qquad (29)$$

Thus H(s) is the Laplace transform of the system response to a unit impulse, and its inverse transform h(t) is the system impulse response in the time domain. The system impulse response or its equivalent transfer function completely characterize any dynamic system which may be represented classically by a linear differential equation with constant coefficients.

Further insight into the significance of this transfer function may be obtained by making a partial fraction expansion of this rational function of s and taking the inverse Laplace transform to obtain the impulse response, h(t). The partial fraction expansion will consist of linear factors of the denominator, some of which possess complex roots which will always occur in conjugate pairs.

$$H(s) = \frac{A_1}{(s-s_1)} + \frac{A_2}{(s-s_2)} + \text{----} \frac{A_n}{(s-s_n)} \qquad (30)$$

Note that the roots s_1, s_2, etc. are determined solely by the coefficients of the denominator of H(s) while the numerator contributes only to the A's. The A_k's may be determined for any term by multiplying both sides of the equation by $(s-s_k)$ and evaluation at $s = s_k$. Thus one may perform the inverse Laplace transform directly, or simply recognize each term of the expansion to be the Laplace transform of $A_k e^{s_k t}$, yielding a solution

$$h(t) = A_1 e^{s_1 t} + A_2 e^{s_2 t} + \text{---} A_n e^{s_n t} \qquad (31)$$

Note that a root $s_k = \sigma_k + jw_k$ will, if the value of σ_k is negative, produce a decaying oscillatory time function, or if $\omega_k = 0$, a simple decaying exponential. Conversely, if σ_k is zero or positive, the time function will continue undiminished or grow without bounds as time increases. Such a system is defined as unstable. A convenient and frequently used concept is that of the locations of the complex roots of the numerator function of s, B(s), and denominator function, A(s), in the complex s-plane. Here the roots of the numerator are referred to as the zeros of H(s) and denoted by o, while the roots of the denominator are referred to as poles and denoted by x. The reason for this nomenclature is more apparent if one thinks of an axis perpendicular to the and jw axes, which is the coordinate of the modulus of H(s) as s is varied over its entire plane. The H(s) resulting would be a three dimensional surface which would touch the s-plane at the zeros and which would project upward to infinity at the poles. Again note that the growth or decay rates of the h(t) are determined only by the location of the poles in the complex s-plane.

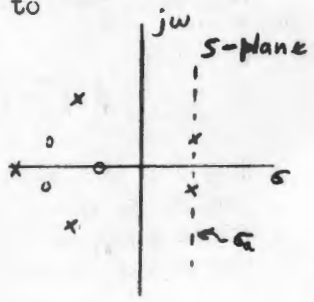


Figure 3

The behavior of H(s) as the value of a traverses the jw axis from the origin to plus infinity is of particular interest, since from the equations (13) and (17) one might suspect a relation between s = jw and the steady-state sinusoidal response of the system. This is indeed the case, and it may be shown that H(jw) is the system steady-state response in complex form to a sinusoidal driving function of unity amplitude. This relation may be proven by substitution of the Laplace transform of a unit sinusoid $\frac{w}{s^2 + w^2}$ for X(s) into equation (28).

$$X(s) = H(s) \frac{w}{s^2 + w^2}$$

Then the partial fraction expansion of X(s) will contain two additional terms.

$$--- + \frac{K_{jw}}{s-jw} + \frac{K_{-jw}}{s + jw} \qquad (32)$$

And since in the steady state, all the transient terms of the time solution will diminish to zero for a stable system, the remaining response will ultimately be due only to the two forcing terms above.

$$x_{ss}(t) = K_{jw} e^{+jwt} + K_{-jw} e^{-jwt}$$

where $K_{jw} = (s-jw) H(s) \frac{w}{s^2 + w^2} \Big|_{s=+jw}$

$$= \frac{H(jw)}{2j}$$

and $K_{-jw} = (s+jw) H(s) \frac{w}{s^2 + w^2} \Big|_{s=-jw}$

$$= \frac{H(-jw)}{-2j}$$

$$x_{ss}(t) = \frac{H(j\omega)\, e^{j\omega t} - H(-j\omega)\, e^{-j\omega t}}{2j}$$

$$= |H(j\omega)| \frac{e^{j(\omega t+\theta)} - e^{-j(\omega t+\theta)}}{2j}$$

$$= |H(j\omega)| \sin(\omega t+\theta), \text{ where } \theta = \tan^{-1} \frac{\text{Im } H(j\omega)}{\text{Re } H(j\omega)} \qquad (33)$$

Thus, the system transfer function H(s) is particularly useful for determining the system transient time behavior for various time inputs, and for determining the sinusoidal steady-state response of the system without taking the inverse transform. In cases where inverse transformation is desirable in determining a time response, the transformation can usually be more easily obtained by application of the Cauchy residue theorem.

The time function input most frequently used, other than the sinusoid, is the unit step function since this function is one whose resultant response in a dynamic system is most readily compatible with intuition. In Laplace notation,

$$X(s) = Y(s)\, H(s) = \frac{1}{s}\, H(s) \qquad (34)$$

From this x(t) may be determined by inverse transformation. But also recall that the function 1/s is conceptually related to integration in the time domain. Hence, if the initial conditions of the system are zero,

$$\mathcal{L}^{-1}[X(s)] = x(t) = \int_0^t h(t)\, dt \qquad (35)$$

So if one knows the system impulse response h(t) analytically, then a simple integration of the system impulse response yields the system response to a unit step input.

Two other useful theorems allow certain interpretations of time response from the system transfer function without going through the details of inverse transformation. These are the initial and final value theorems.

The initial value theorem may be stated as -

$$\lim_{s \to \infty} sF(s) = \lim_{t \to 0} f(t) \qquad (36)$$

and the final value theorem as -

$$\lim_{s \to o} sF(s) = \lim_{t \to \infty} f(t) \tag{37}$$

A particular simplication is achieved by use of the latter relation (37) to determine the steady-state response of a system to a unit step input.

$$X(s) = \frac{1}{s} H(s)$$

$$\lim_{t \to \infty} x(t) = \lim_{s \to o} sX(s) = \lim_{s \to o} \frac{s}{s} H(s)$$

$$= \lim_{s \to o} H(s)$$

In conclusion, a comparison of the Laplace and Fourier transforms is appropriate. Recall that we have placed no restrictions upon whether the real part of the exponents of epsilon in the time functions are negative or positive (convergent or divergent) and that time functions have been limited to positive time, in using the Laplace transform. The capability of the Laplace transform to handle divergent exponential time functions is permitted by the fact that the real part of s in the direct transformation must be greater than σ_a; the abscissa of absolute convergence in order to insure that the integral converge. Divergent transient system behavior corresponds to poles in the right-half s-plane. Also, in the inverse transform, integrating from $c-j\infty$ to $c+j\infty$ where c is equal to or greater than the abscissa of absolute convergence, means that traversing a line in the right-half s-plane parallel to the jw axis and to the right of all poles. In the Fourier integral, the mathematical definitions may be stated as

$$\mathcal{F}[f(t)] = F(w) = \int_{-\infty}^{\infty} f(t)\, e^{-jwt}\, dt$$

and

$$\mathcal{F}^{-1}[F(w)] = f(t) \; (=) \; \frac{1}{2\pi} \int_{-\infty}^{\infty} F(w)\, e^{jwt}\, dw$$

Note that the time domain for negative time is included in the transform, but that no provision exists for handling divergent time functions. Thus, the right-half $j\omega$-plane corresponds instead to the description of time behavior prior to t=o. These properties make the Laplace transform particularly applicable to time functions where stability is in question, and the Fourier transform applicable to time functions where stability or convergence is assured, and is particularly useful in statistical noise theory. The transfer function of a stable, physically realizable system is identical in both.

A Brief Table of Laplace Transforms

Transform	Function
$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t)\,dt$	$f(t)$
$aF(s) + bG(s)$	$af(t) + bg(t)$
$sF(s) - f(0)$	$\frac{df(t)}{dt}$
$s^2F(s) - sf(0) - \frac{df(0)}{dt}$	$\frac{d^2f(t)}{dt^2}$
$s^nF(s) - \sum_{k=1}^{n} s^{n-k}\frac{d^{k-1}f(0)}{dt^{k-1}}$	$\frac{d^nf(t)}{dt^n}$
$\frac{1}{s^n}F(s)$	$\int_0^t \cdots \int_0^t f(t)\,\underbrace{dt \cdots dt}_{n \text{ times}}$
$(-)^n \frac{d^nF(s)}{ds^n}$	$t^n f(t)$
$F(s-a)$	$e^{-at}f(t)$
$e^{-as}F(s)$	$\begin{cases} f(t-a) & \text{when } t>a \\ 0 & \text{when } t<a \end{cases}$
$F(s)G(s)$	$\int_0^t f(\tau)\,g(t-\tau)\,d\tau$
$\frac{1}{s}$	1
$\frac{1}{s+a}$	e^{-at}
$\frac{1}{(s+a)(s+b)}$	$\frac{1}{a-b}(e^{-bt} - e^{-at})$
$\frac{s}{(s+a)(s+b)}$	$\frac{1}{b-a}(be^{-bt} - ae^{-at})$
$\frac{a}{s^2+a^2}$	$\sin at$
$\frac{s}{s^2+a^2}$	$\cos at$
$\frac{a}{s^2-a^2}$	$\sinh at$
$\frac{s}{s^2-a^2}$	$\cosh at$

Transform	Function	
$\frac{2as}{(s^2+a^2)^2}$	$t\sin at$	
$\frac{s^2-a^2}{(s^2+a^2)^2}$	$t\cos at$	
$\frac{2a^3}{(s^2+a^2)^2}$	$\sin at - at\cos at$	
$\frac{2as}{(s^2-a^2)^2}$	$t\sinh at$	
$\frac{s^2+a^2}{(s^2-a^2)^2}$	$t\cosh at$	
$\frac{2a^3}{(s^2-a^2)^2}$	$at\cosh at - \sinh at$	
$\frac{a}{(s+b)^2+a^2}$	$e^{-bt}\sin at$	
$\frac{s+b}{(s+b)^2+a^2}$	$e^{-bt}\cos at$	
$\frac{4a^3}{s^4+4a^4}$	$\sin at\cosh at - \cos at\sinh at$	
$\frac{2a^2s}{s^4+4a^4}$	$\sin at\sinh at$	
1	$\mathcal{I}(t)$	unit impulse
e^{-st_1}	$\mathcal{I}(t-t_1)$	
s	$\mathcal{D}(t)$	unit doublet
se^{-st_1}	$\mathcal{D}(t-t_1)$	
$\frac{1}{s^n}$	$\frac{t^{n-1}}{(n-1)!}$	$n>0$
$\frac{n!}{s^{n+1}}$	t^n	$n>-1$
$\frac{1}{(s+a)^n}$	$\frac{t^{n-1}e^{-at}}{(n-1)!}$	$n>0$
$\frac{s}{(s+a)^n}$	$\frac{(n-1)-at}{(n-1)!}t^{n-2}e^{-at}$	$n>1$

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

Lecture 3

GRAPHICAL REPRESENTATION OF TRANSFER FUNCTIONS

by

R. J. Martin

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

GRAPHICAL REPRESENTATION OF TRANSFER FUNCTIONS

by R. J. Martin

INTRODUCTION:

Today's discussion will deal with the various methods of graphically representing transfer functions. These graphical pictures of system responses are the tools with which control systems engineers can design and analyze control circuits. They are useful in showing the frequency response of a system, determining stability of systems from their open loop characteristics, analyzing systems in the presence of nonlinearities, finding closed loop responses, and obtaining a feel for system response due to noise inputs. The three types of plotting in wide use are the Bode, Polar, and Nichols plots. Each of these will be discussed separately with appropriate examples to demonstrate their use.

BODE PLOTS

The type of graphical plots used most often in control system work are the Bode plots. These are frequency response curves and consist of two curves; a gain vs. frequency curve, and a phase vs. frequency curve. The gain axis is calibrated in decibels according to the definition.

$$\text{Db} = 20 \log \left[\text{Voltage Ratio}\right] . \tag{1}$$

The frequency scale is logarithmic, whereas Db gain and phase are plotted on linear scales.

Example #1: $F(s) = K$ (2)

The gain and phase of this transfer do not vary with frequency. If $K = 2$, the gain from equation (1) is +6 Db, making the gain plot a straight line at + 6 Db. The phase curve is a straight line at 0°.

Example #2: $F(s) = \frac{1}{Ts}$ (3)

If the simple substitution of $j\omega$ is made for s, as discussed in previous lecture, the equation (3) becomes $F(j\omega) = \frac{1}{j\omega T} = \frac{1}{\omega T} \angle -90^\circ$. Thus, the phase curve does not vary with frequency and is a straight line at -90°. The gain term does vary with frequency as $\frac{1}{T\omega}$. At $\omega = \frac{1}{T}$, the gain is unity, or from equation (1), 0 Db.

At one octave above this, $\omega = \frac{2}{T}$, and the gain equals $\frac{1}{2}$. This, from equation (1), corresponds to -6 Db. Likewise, at one octave below the original frequency, $\omega = \frac{1}{2T}$ and the gain is 2, which gives +6 Db. The gain curve is then a straight line with a slope of -6 Db/octave, and intersecting 0 Db at $\omega = \frac{1}{T}$.

If the transfer had been $F(s) = Ts$, the gain and phase plots would be inversed. That is, the phase would be constant at $+90^\circ$, and the gain slope would be +6 Db/octave. The point of intersection with 0 Db would remain at $\omega = \frac{1}{T}$.

Example #3: $F(s) = \frac{1}{Ts + 1}$. (4)

Again making the $j\omega$ substitution, equation (4) becomes $F(j\omega) = \frac{1}{j\omega T + 1}$. This has been evaluated for various values of ω in the table below:

TABLE 1

$$F(j\omega) = \frac{1}{j\omega T + 1}$$

ω	GAIN RATIO	Db Gain	Phase
1/16T rad.	1	0	- 3 5/16°
1/8T	1	0	- 6 5/8°
1/4T	0.98	-¼	-13 1/4°
1/2T	0.90	-1	-26 1/2°
1/T	0.707	-3	-45°
2/T	0.445	-7	-63 1/2°
4/T	0.245	-12¼	-76 3/4°
8/T	0.125	-18	-83 3/8°
16/T	0.0625	-24	-86 11/16°

Bode plots of this data are given in Figure 1. It can be seen that the gain curve approaches two assymptotes on either side of the corner frequency. At low frequencies, the curve approaches unity gain, whereas at high frequencies, it approaches $\frac{1}{\omega T}$. Some rounding occurs in the region of the corner such that at the corner frequency, the gain is -3 Db. The phase curve also approaches two assymptotes, 0° at low frequencies, and -90° for higher frequencies. At the corner frequency the phase is -45°. One octave below the corner, the phase is -26½°. For each successive octave reduction in frequency, the phase shift is half the previous phase shift. That is, at $\omega = \frac{1}{4T}$, the phase is $\frac{-26\frac{1}{2}^\circ}{2} = -13\frac{1}{4}^\circ$, and at $\omega = \frac{1}{8T}$, the phase is $\frac{-13\frac{1}{4}^\circ}{2} = -6\ 5/8^\circ$, etc. Also, one octave above the corner, the phase is $-63\frac{1}{2}^\circ$, or $-(90^\circ - 26\frac{1}{2}^\circ)$. At each successive octave increase in frequency, the phase is half the amount of phase from 90° as for the previous octave. That is, at $\omega = 4/T$, the phase is $-\left(90^\circ - \frac{26\frac{1}{2}^\circ}{2}\right) = -(90^\circ - 13\frac{1}{4}^\circ) = -76\ 3/4^\circ$, and at $\omega = 8/T$, the phase is $-(90^\circ - \frac{13\frac{1}{4}^\circ}{2}) = -(90^\circ - 6\ 5/8^\circ) = -83\ 3/8^\circ$, etc. Thus, keeping these relations in mind, the design engineer need not evaluate the transfer function at several frequencies, but merely has to draw the assymptotes and round off in the vicinity of the corner.

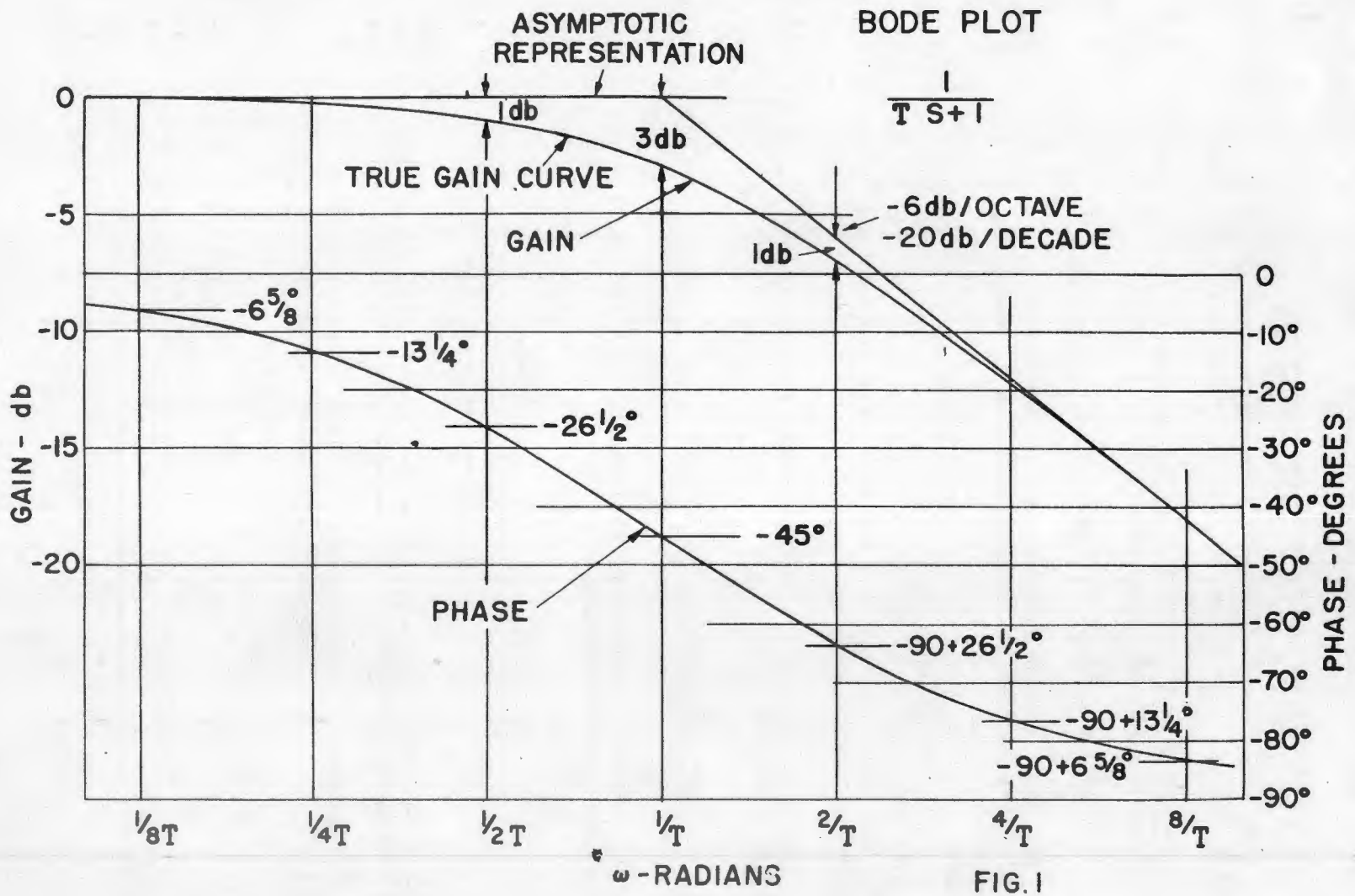

BODE PLOT
ASYMPTOTIC REPRESENTATION
1 / (T S+1)
0
-5
-10
-15
-20
1db
3db
TRUE GAIN CURVE
GAIN
-6db/OCTAVE
-20db/DECADE
1db
-6 5/8°
-13 1/4°
-26 1/2°
-45°
PHASE
-90+26 1/2°
-90+13 1/4°
-90+6 5/8°
0
-10°
-20°
-30°
-40°
-50°
-60°
-70°
-80°
-90°
GAIN - db
PHASE - DEGREES
1/8T
1/4T
1/2T
1/T
2/T
4/T
8/T
ω - RADIANS

FIG. 1

It should also be noted that if $F(s) = Ts + 1$, the Bode plots obtained are the inverse of those just discussed, in that the gain curve approaches a +6 Db/octave slope at high frequencies, and the phase approaches $+90^\circ$.

$$\text{Example \#4:} \quad F(s) = \frac{1}{T^2s^2 + 2\zeta Ts + 1} \tag{5}$$

This is a second order lag term where T is the time constant, and ζ is the damping factor. The value of ζ lies between zero and unity.

Case I: $\zeta = 1$ If $\zeta = 1$, the quadratic factors into two equal first order terms, and equation (5) becomes

$$F(s) = \frac{1}{Ts + 1} \circ \frac{1}{Ts + 1} = \frac{1}{(Ts + 1)^2}$$

The Bode plot is then the product of two first order lags as discussed in Example #3. However, since the gain scale is in decibels, the gain curves are merely added together. Also, remembering that when taking the product of two vectors, their phases are added, the phase curves are also added together. Thus, the Bode plot of a second order equation with unity damping is merely double the gain and phase curves which make up the first order factors. A plot of this is shown in Figure 2. The gain approaches a -12 Db/octave slope at high frequencies, and the phase curve approaches -180°.

Case II: $0 < \zeta < 1$ If $j\omega$ is substituted for s, the transfer of equation (5) becomes

$$\frac{1}{1 - T^2\omega^2 + j2\zeta T\omega} .$$

The following tables are evaluations of this function for $\zeta = 0.5$, and $\zeta = 0.1$ at various frequencies:

TABLE 2

$$\frac{1}{(1 - \omega^2T^2) + j2\zeta T\omega} \quad \zeta = 0.5.$$

ω	Gain Ratio	Gain in Db	Phase
1/8T rad.	1.0	0	-9°
1/4T	1.0	0	-15°
1/2T	1.12	+1	-35°
1/T	1.0	0	-90°
2/T	.28	-11	-145°
4/T	.0625	-24	-165°
8/T	.0167	-36	-171°

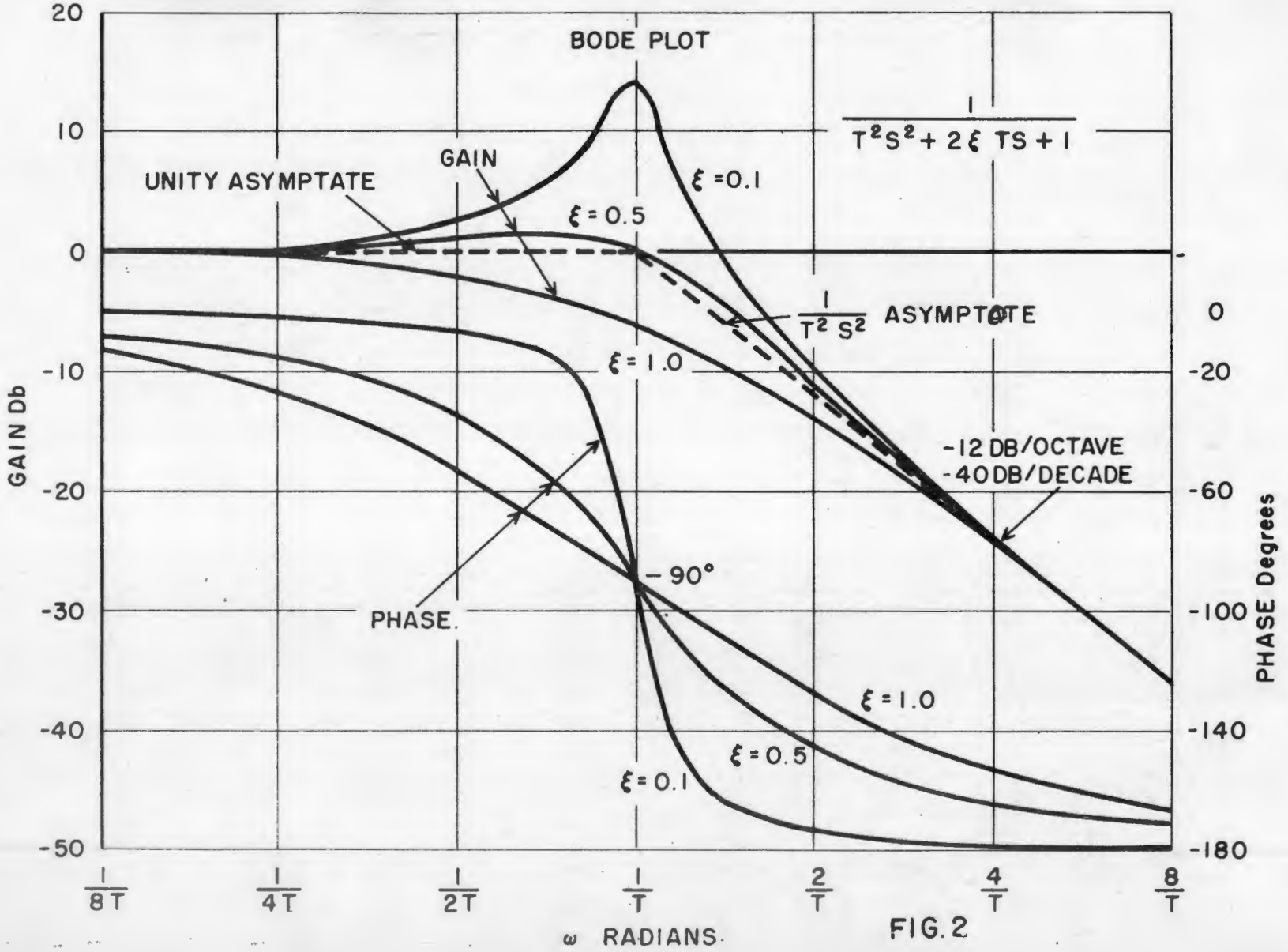

BODE PLOT
$\frac{1}{T^2S^2 + 2\xi TS + 1}$
UNITY ASYMPTATE
GAIN
$\xi = 0.1$
$\xi = 0.5$
$\xi = 1.0$
$\frac{1}{T^2 S^2}$ ASYMPTOTE
-12 DB/OCTAVE
-40 DB/DECADE
-90°
PHASE
$\xi = 1.0$
$\xi = 0.5$
$\xi = 0.1$
GAIN Db
PHASE Degrees
20
10
0
-10
-20
-30
-40
-50
0
-20
-60
-100
-140
-180
$\frac{1}{8T}$
$\frac{1}{4T}$
$\frac{1}{2T}$
$\frac{1}{T}$
$\frac{2}{T}$
$\frac{4}{T}$
$\frac{8}{T}$
ω RADIANS

FIG. 2

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

TABLE 3

$$\frac{1}{(1 - T^2\omega^2) + j2\zeta T\omega} \qquad \zeta = 0.1$$

ω	Gain Ratio	Gain in DB	Phase
1/8T rad.	1.0	0	-1
1/4T	1.0	0	-3
1/2T	1.32	+2.5	-7
1/T	5.0	+14	-90
2/T	.34	-9.5	-173
4/T	.0625	-24	-177
8/T	.0167	-36	-179

Bode plots of these functions are also shown in Figure 2. It can be seen that as ζ is decreased, a peak is developed in the gain curve at the corner frequency. Also, the phase plot crosses -90° at the corner with a steeper slope, thereby giving reduced phase lag prior to the corner, but increased lags after the corner. A ζ of 0.707 is the magic number for which the gain curve does not go positive.

Case III: $\zeta = 0$ If $\zeta = 0$, equation (5) becomes $\frac{1}{T^2s^2 + 1}$. Making the $j\omega$ substitution for s, we obtain $F(j\omega) = \frac{1}{1 - T^2\omega^2}$. At the corner frequency, the gain value goes to infinity, and the phase jumps from 0° to -180°. This is the transfer function of an oscillator. Needless to say, one does not try to construct such a function in control system work.

Case IV: $\zeta > 1.0$ If ζ is greater than one, the quadratic of equation (5) factors into two first order corners. Each can be plotted separately, and combined as mentioned under Case I above.

POLAR PLOTS

Another form of representing a transfer function graphically is by means of the polar plot, or as is sometimes called, a Nyquist plot. This is a plot of gain ratio vs. phase angle of a network as frequency is varied. It is plotted on polar coordinates. Although frequency is not a coordinate of this type of graph, every point on the curve does correspond to a particular frequency, and this information is often added to the plot.

To familiarize ourselves with this form of plotting, let us again refer to our examples:

Example #1: $F(s) = K$ (2)

Seeing as gain and phase are constant, the plot of this function is merely a point at gain equals K, and zero phase angle.

Example #2: $F(s) = \frac{1}{Ts}$ (3)

Here the gain varies but the phase is constant at -90°. The polar plot consists of a line from origin to infinity along the -90° axis. The gain is infinite at zero frequency, and approaching zero at high frequencies. At $\omega = 1/T$, the gain is unity.

$$\text{Example \#3:} \quad F(s) = \frac{1}{Ts + 1} \tag{4}$$

At zero frequency, the value of this transfer is unity with zero phase angle. As frequency is increased, the phase angle increases in a negative direction and the gain decreases until at infinite frequency, the gain approaches zero along the -90° axis. As can be seen in Figure 3, the curve is actually a semi-circle with a radius of 0.5, and whose center is 0.5 $\angle 0°$.

$$\text{Example \#4:} \quad F(s) = \frac{1}{T^2s^2 + 2\zeta Ts + 1} \tag{5}$$

Case I: $\zeta = 1.0$ The quadratic factors into two first order terms. The plot is merely the product of two curves as shown in Example #3. At zero frequency, the gain is unity, with zero phase, and at high frequencies, it approaches zero along the -180° axis. The Polar plot of this function appears in Figure 4.

Case II: $0 < \zeta < 1.0$ The data from Tables 2 and 3 are plotted in Figure 4. As can be seen, as ζ decreases, the gain in the region of $\omega = 1/T$ is increased.

Case III: $\zeta = 0$ This function would plot as a line along the 0° axis from 1.0 to infinity. At infinite gain, it rotates through -180°, and reappears along the -180° axis where it continues to zero gain.

NICHOLS PLOTS

The last method of plotting transfer function characteristics is the Nichols plot. It is a plot of gain in decibels versus phase shift of a network as frequency varies. However, it is plotted on rectangular coordinates.

Considering again the four examples discussed in the previous sections, we shall discuss the Nichols form of plotting.

$$\text{Example \#1:} \quad F(s) = K \tag{2}$$

Since neither gain nor phase vary with frequency, plot of this function is a point at zero phase shift, and gain in Db. corresponding to K.

$$\text{Example \#2:} \quad F(s) = \frac{1}{Ts} \tag{3}$$

The phase is constant at -90°, therefore, the plot appears as a straight line along the -90° axis from a decibel gain of plus infinity to minus infinity.

$$\text{Example \#3:} \quad F(s) = \frac{1}{Ts + 1} \tag{4}$$

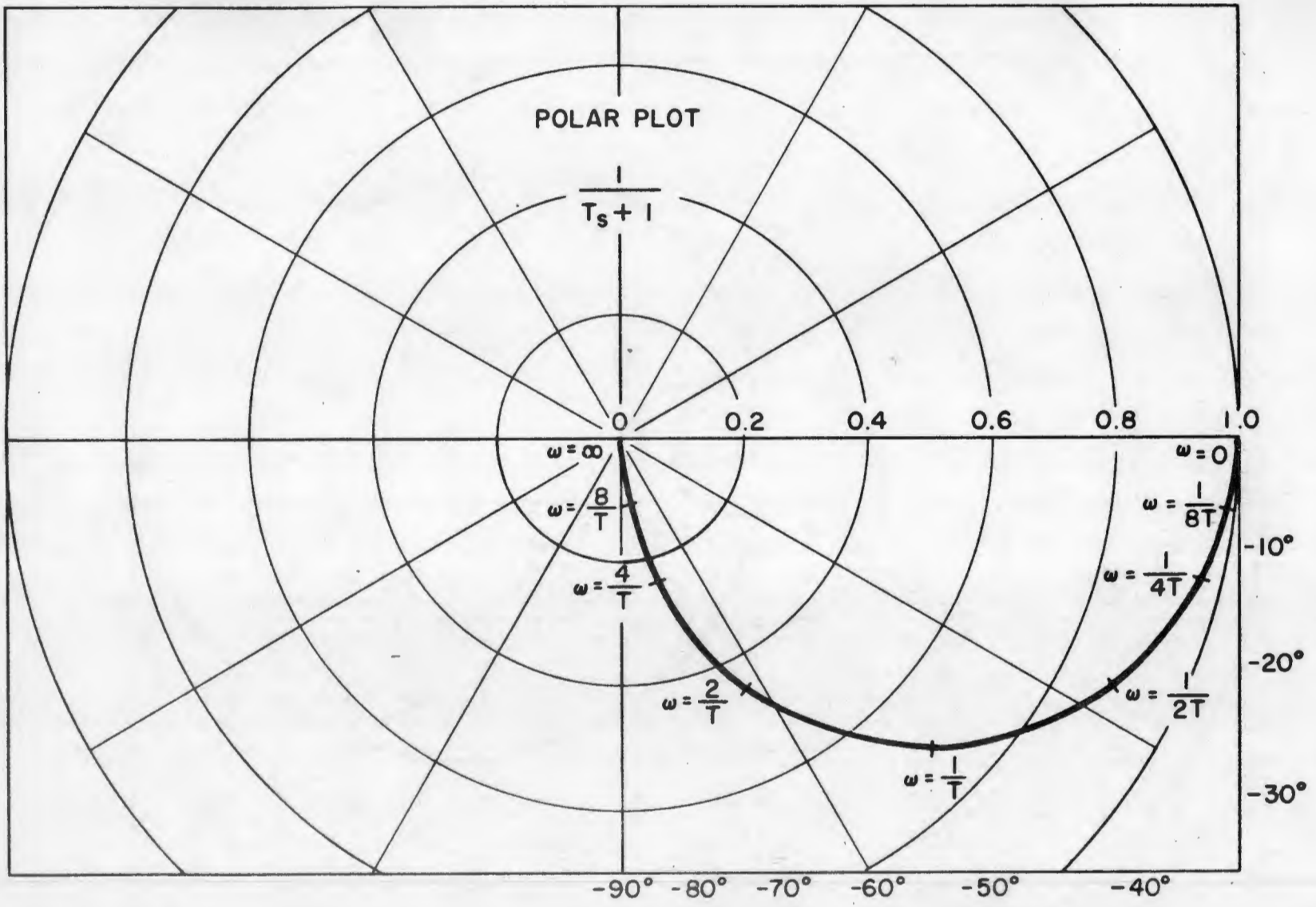

POLAR PLOT
1
Ts + 1
0
0.2
0.4
0.6
0.8
1.0
ω = ∞
ω = 8/T
ω = 4/T
ω = 2/T
ω = 1/T
ω = 1/2T
ω = 1/4T
ω = 1/8T
ω = 0
-10°
-20°
-30°
-90°
-80°
-70°
-60°
-50°
-40°

FIG. 3

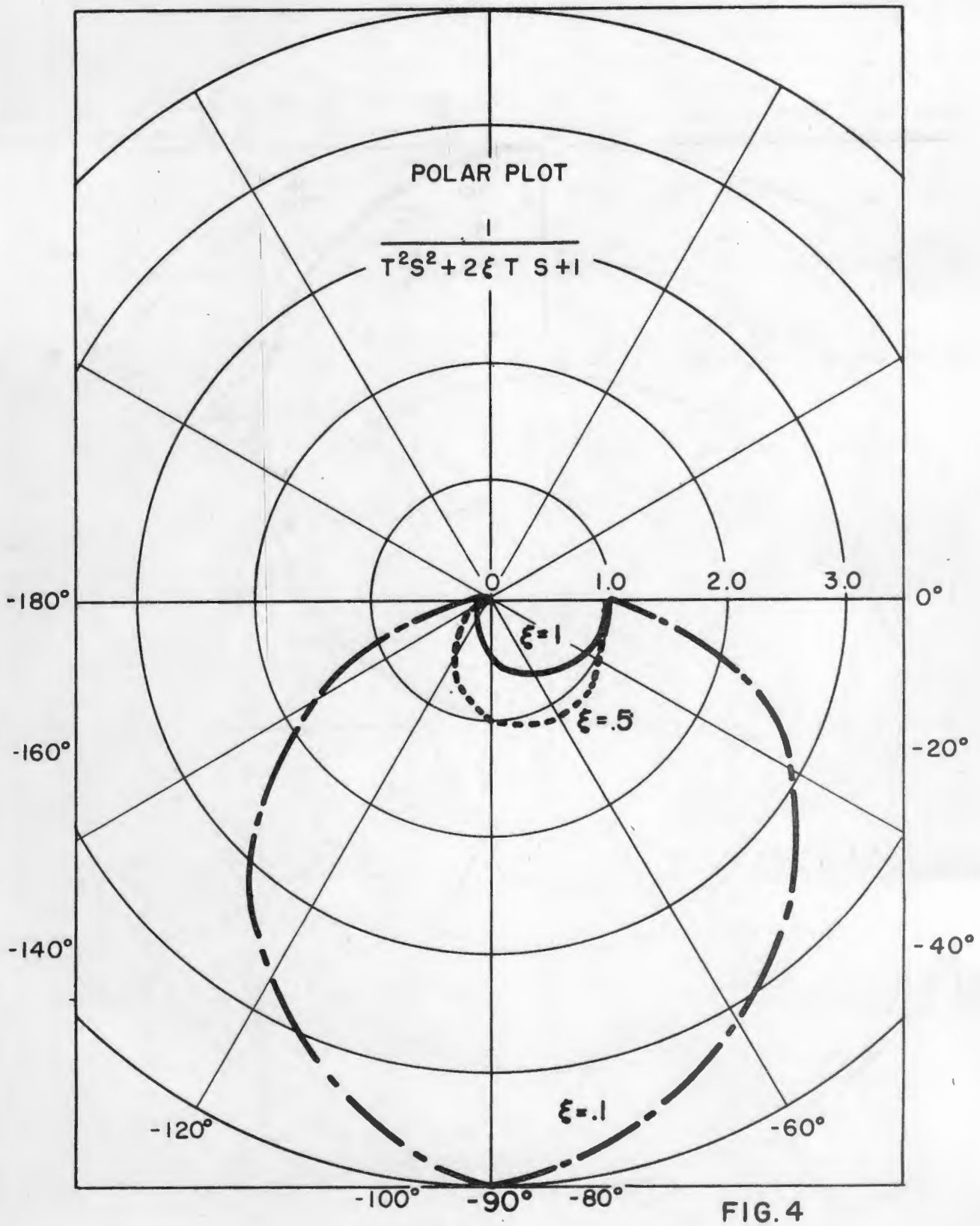

POLAR PLOT
1
T²S² + 2ξ T S + 1
0
1.0
2.0
3.0
-180°
0°
-160°
-20°
-140°
-40°
-120°
-60°
-100°
-90°
-80°
ξ=1
ξ=.5
ξ=.1

FIG. 4

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

A plot of this function is shown in Figure 5. At low frequencies, the curve starts at 0 Db, and no phase lag. As frequency is increased, the gain decreases, and the phase approaches -90°.

$$\text{Example \#4: } F(s) = \frac{1}{T^2s^2 + 2\zeta Ts + 1} \qquad (5)$$

Case I: $\zeta = 1.0$ This is merely the product of two first order terms as discussed in previous sections. A Nichols plot of this transfer is shown in Figure 5.

Case II: $0 < \zeta < 1$ As ζ is decreased, the gain peak at the corner is also noticed in the Nichols plots. This can be seen in Figure 5.

Case III: $\zeta = 0$ A Nichols plot of this would be a line up the 0° phase axis from 0 Db to infinity, reappearing at -180°, where it goes from plus infinity to minus infinity in decibel gain.

FURTHER EXAMPLE

Now, consider the more complicated transfer function,

$$F(s) = K \frac{(T_1s + 1)}{s(T_2s + 1)(T_3^2s^2 + 2\zeta T_3s + 1)}$$

where $K = 1.0$
$T_1 = 2.0$
$T_2 = 0.5$
$T_3 = 0.1$
$\zeta = 0.5$

This function consists of a gain term, K; an integral term, 1/s; a first order lead corner $(T_1s + 1)$; a first order lag corner, $\frac{1}{(T_2s + 1)}$; and a second order lag term, $\frac{1}{(T_3^2s^2 + 2\zeta T_3s + 1)}$. The total transfer is the product of all these terms. Therefore, the complete Bode plot is the sum of all the gain and phase curves. The first step in plotting the function is to plot each individual curve.

K - This gives a gain curve along the 0 Db axis. It contributes no phase shift so the phase plot is along the 0° axis.

1/s - This is a constant -6 Db per octave slope intersecting the 0 Db axis at $\omega = 1.0$ radian. The phase shift is a constant at -90°.

$(T_1s + 1)$ - The assymptotic gain curve is 0 Db until the corner frequency of $\omega_1 = \frac{1}{T_1} = 0.5$ radians. Here it approaches a +6 Db per octave slope. Proper rounding in the region of the corner should be included. The phase curve starts at 0°, crosses +45° at 0.5 radians, and approaches +90° at high frequencies. It follows the relations of phase shift per octave above and below the corner as described in the section on Bode plots.

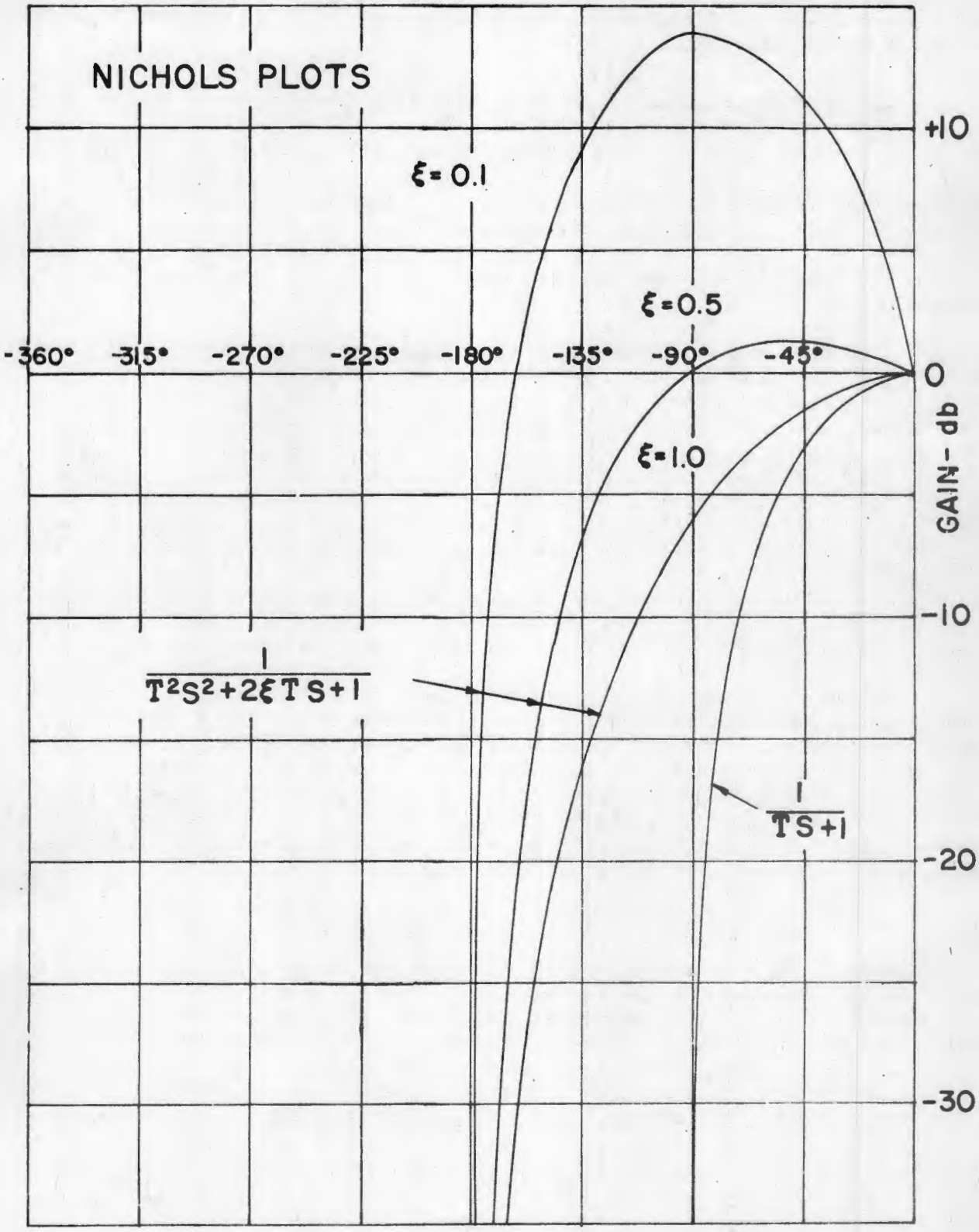

NICHOLS PLOTS
ξ = 0.1
ξ = 0.5
ξ = 1.0
-360°
-315°
-270°
-225°
-180°
-135°
-90°
-45°
+10
0
-10
-20
-30
GAIN - db
1 / T²S²+2ξTS+1
1 / TS+1

FIG. 5

$\frac{1}{(T_2 s + 1)}$ - The gain curve is again 0 Db until the corner of $\omega_2 = \frac{1}{T_2} = 2$ radians. Here the gain curve approximates a -6 Db per octave slope. Again, proper rounding in the vicinity of the corner should be considered. The phase curve starts at 0°, crosses -45° at 2 radians, and approaches -90° at high frequencies according to the octave relationships.

$\frac{1}{T_3 2 s 2 + 2\zeta T_3 s + 1}$ - In this case also the gain curve starts at 0 Db, and at the corner, $\omega_3 = \frac{1}{T_3} = 10$ radians, breaks into a -12 Db per octave slope. The phase curve starts at 0°, and referring to Figure 2, crosses 90° at $\omega = 10$ radians. At high frequencies it approaches -180°.

The above five curves are then added point by point to obtain the total transfer plot.

A quicker method of obtaining the total transfer is to begin by drawing the low frequency gain and phase curves. For this example, it is K/s; a -6 Db per octave slope through 0 Db at $\omega = 1$, with a constant -90° phase lag. Then, at the first corner, $\omega_1 = 0.5$ radians, a 6 Db per octave gain slope is added to the -6 Db per octave slope, giving a net 0 Db per octave slope. The phase leads associated with the lead corner are added directly to the -90° curve so that at high frequencies, the phase curve approaches 0°. Next, at the first order lag corner, $\omega_2 = 2.0$ radians, a -6 Db per octave slope is added to the gain curve, and the appropriate phase lags are added to the phase curve. The gain slope is then -6 Db per octave, and the phase curve approaches -90°. Finally, at $\omega = 10.0$, the second order lag term offers a -12 Db per octave corner which must be added to the gain curve producing a net slope of -18 Db per octave. The proper second order phase lags are also added to the phase curve making it approach -270°. Thus, the total gain and phase curves are obtained as shown in Fig. 6.

If one wishes to make a Polar or Nichols plot of this type transfer, he could make the $j\omega$ substitution for s, and determine the value of the transfer at various frequencies. However, this is a very burdensome operation. Another method would be to sketch the plots for each factor and combine them appropriately. This is also somewhat laborious for anything but the Bode plot. The easiest method to obtain Polar and Nichols plots is to first sketch the Bode plots as discussed above, and then, from the gain and phase of the Bode plot, sketch the Polar and Nichols plots. Figure 7 and 8 show these plots for our example problem.

CLOSED LOOP

All of our discussion to this point has pertained to open loop transfer functions. If, however, we close a loop around one of these open loop transfers, the net closed loop transfer is extremely different. Consider, for example, a transfer function G(s) whose output is subtracted from a reference signal to produce an error signal which is then applied to the transfer function as an input. This is shown in the block diagram of Figure 9 where R is the reference signal, C is the controlled signal output, and E is the error signal into the transfer.

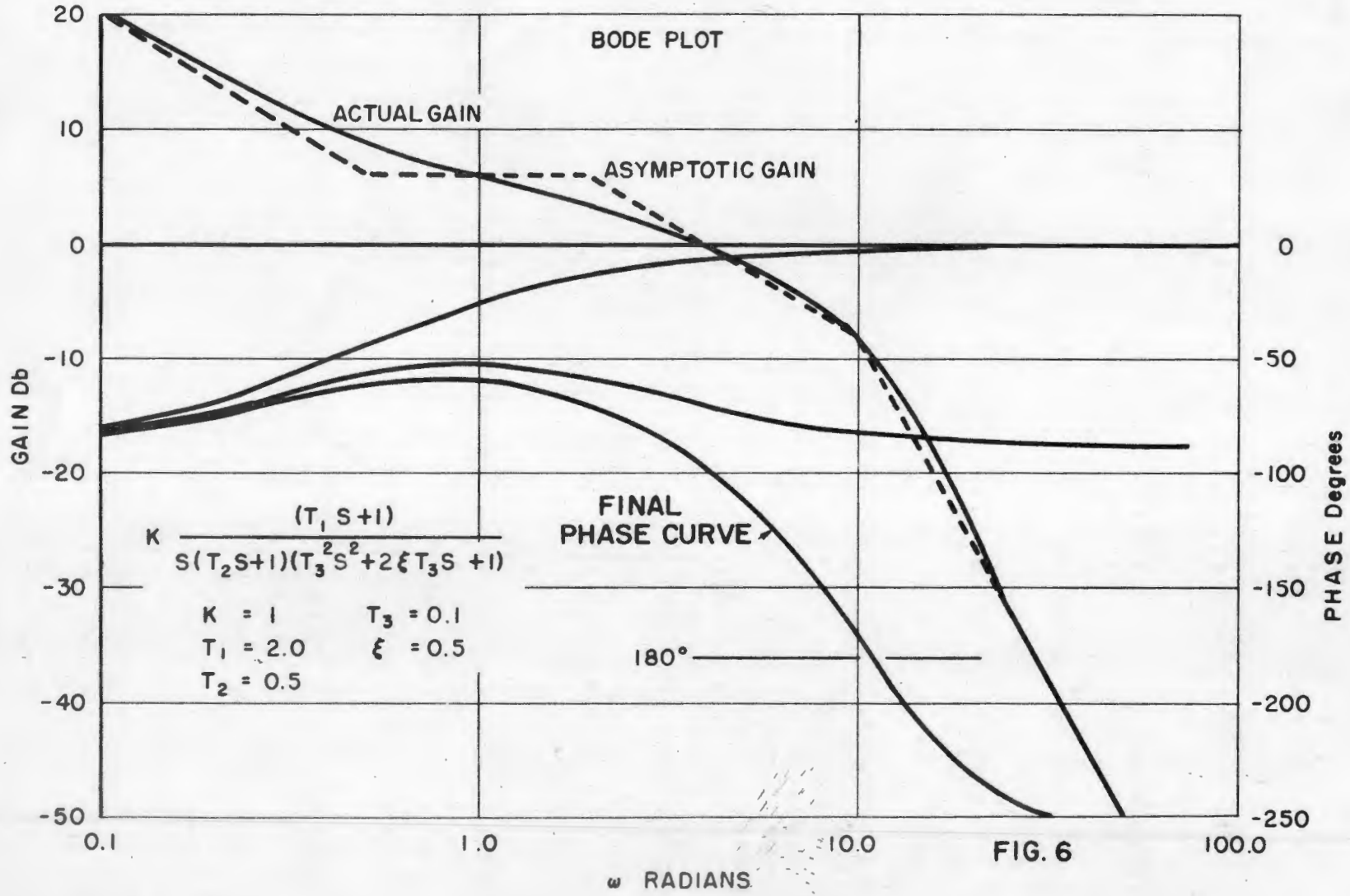

BODE PLOT
ACTUAL GAIN
ASYMPTOTIC GAIN
FINAL PHASE CURVE
$K \frac{(T_1 S+1)}{S(T_2 S+1)(T_3^2 S^2+2\xi T_3 S+1)}$
K = 1
T_1 = 2.0
T_2 = 0.5
T_3 = 0.1
ξ = 0.5
180°
GAIN Db
PHASE Degrees
ω RADIANS
20
10
0
-10
-20
-30
-40
-50
0
-50
-100
-150
-200
-250
0.1
1.0
10.0
100.0

FIG. 6

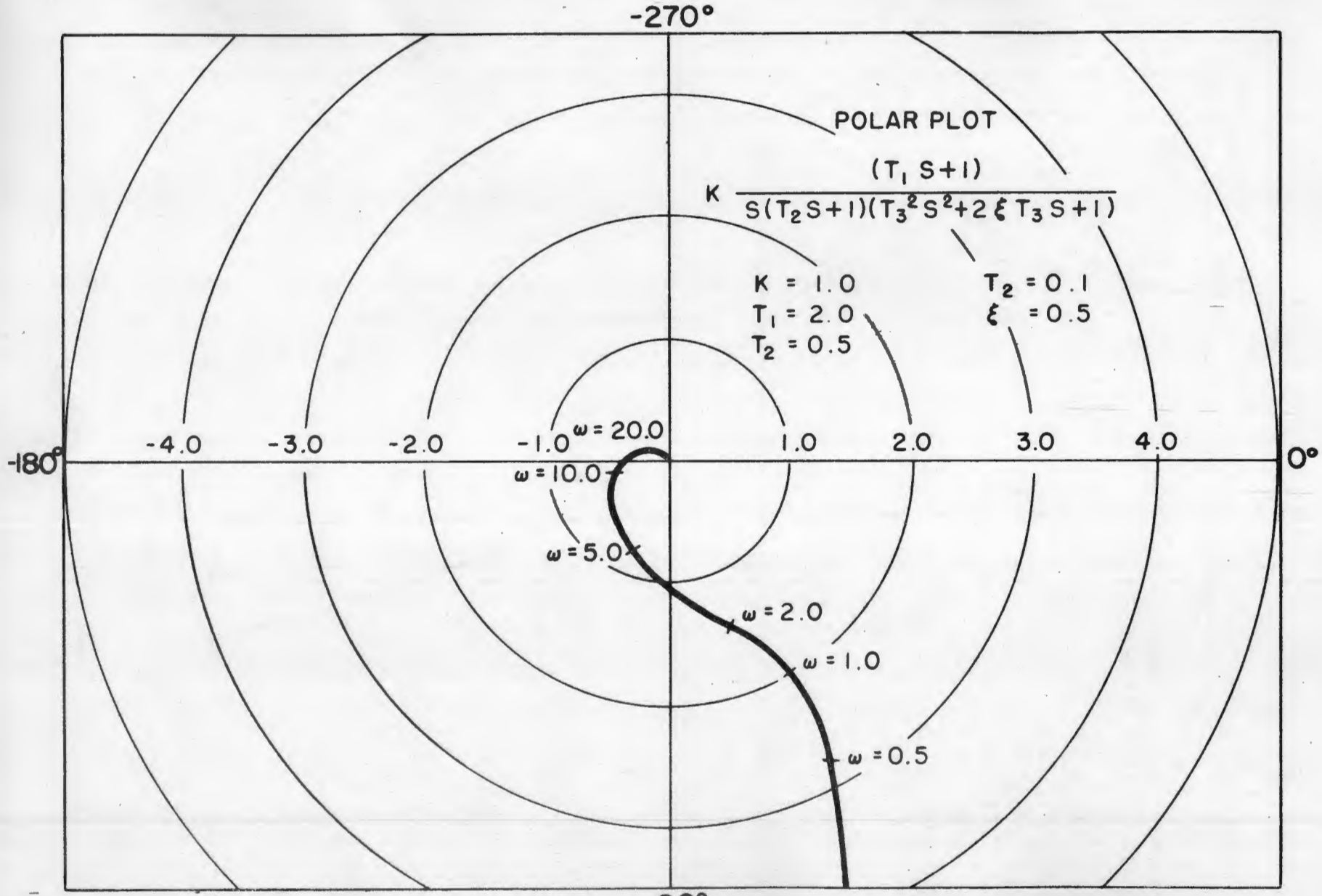

-270°
POLAR PLOT
$K \frac{(T_1 S+1)}{S(T_2 S+1)(T_3^2 S^2+2\xi T_3 S+1)}$
K = 1.0
T1 = 2.0
T2 = 0.5
T2 = 0.1
ξ = 0.5
-180°
0°
-4.0
-3.0
-2.0
-1.0
1.0
2.0
3.0
4.0
ω = 20.0
ω = 10.0
ω = 5.0
ω = 2.0
ω = 1.0
ω = 0.5
-90°

FIG. 7

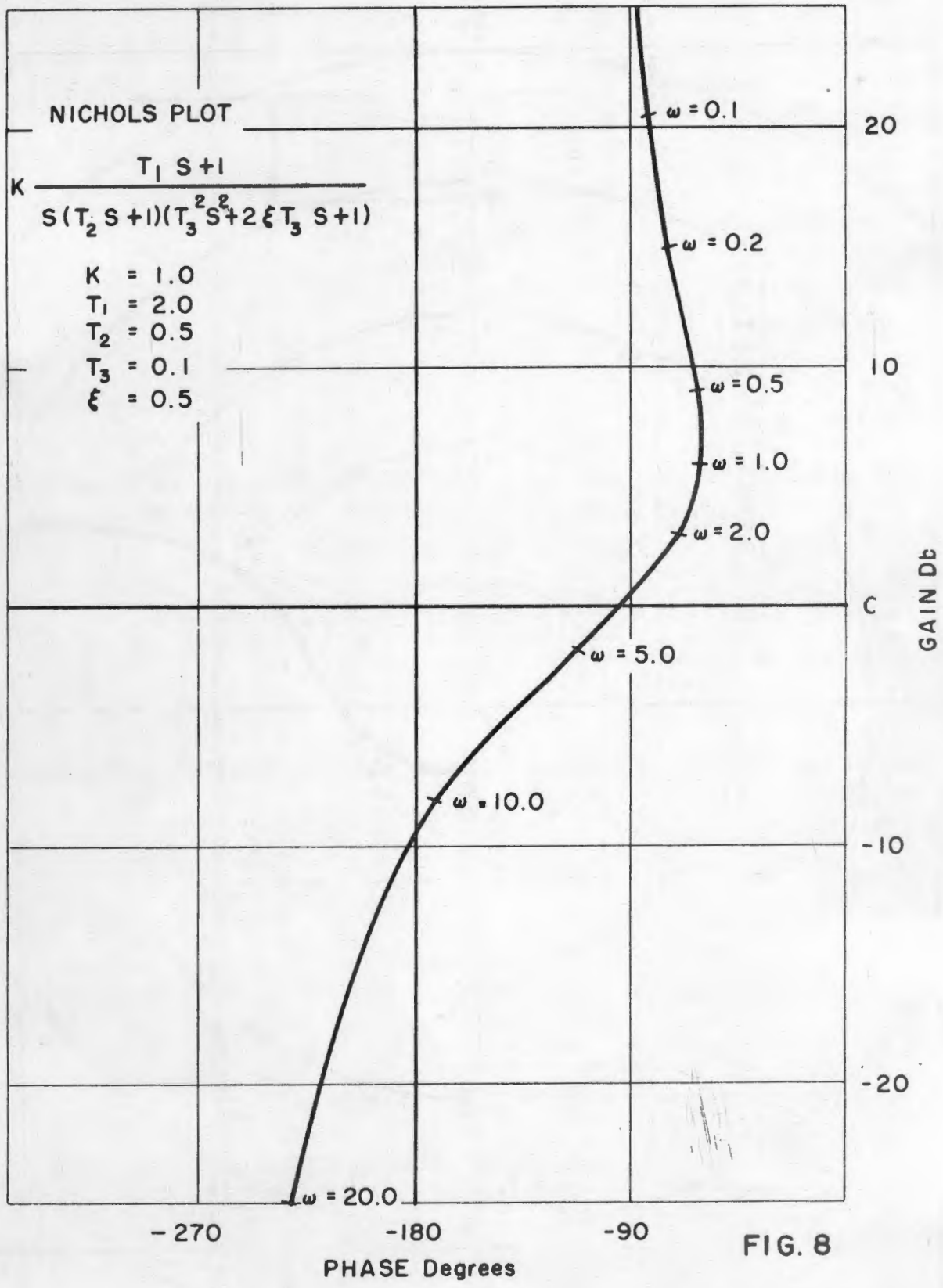

NICHOLS PLOT
$K \frac{T_1 S+1}{S(T_2 S+1)(T_3^2 S^2+2\xi T_3 S+1)}$
K = 1.0
T_1 = 2.0
T_2 = 0.5
T_3 = 0.1
ξ = 0.5
ω = 0.1
ω = 0.2
ω = 0.5
ω = 1.0
ω = 2.0
ω = 5.0
ω = 10.0
ω = 20.0
20
10
0
-10
-20
GAIN Db
-270
-180
-90
PHASE Degrees

FIG. 8

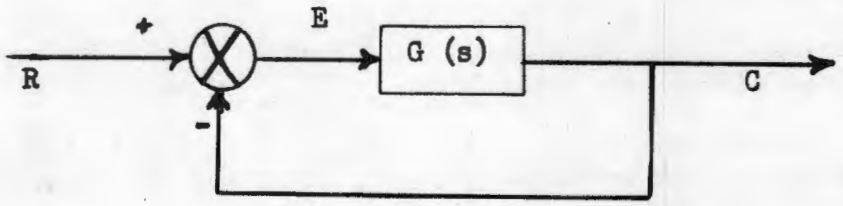


FIGURE 9

The following relations exist:

$$R - C = E \tag{6}$$

$$E\,G(s) = C \tag{7}$$

It is desired to find the closed loop transfer; that is, the transfer from R to C. Thus, eliminating E from the above equations, and solving for $\frac{C}{R}$,

$$R - C = \frac{C}{G(s)}$$

$$R = C\left[1 + \frac{1}{G(s)}\right] = C\left[\frac{G(s) + 1}{G(s)}\right]$$

$$\frac{C}{R} = \frac{G(s)}{1 + G(s)} \tag{8}$$

This, then, is the closed loop transfer. If the open loop transfer, G(s), consists of two polynomials in s, such that

$$G(s) = \frac{a(s)}{b(s)}, \text{ then } \frac{C}{R} = \frac{\frac{a(s)}{b(s)}}{1 + \frac{a(s)}{b(s)}} = \frac{a(s)}{b(s) + a(s)} \tag{9}$$

In our previous example,

$$a(s) = K\,(T_1 s + 1)$$

$$b(s) = s\,(T_2 s + 1)\,(T_3^2 s^2 + 2\zeta T_3 s + 1)$$

Hence, the closed loop transfer of our example is

$$\frac{C}{R} = \frac{K\,(T_1 s + 1)}{s(T_2 s + 1)\,(T_3^2 s^2 + 2\zeta T_3 s + 1) + K\,(T_1 s + 1)}$$

$$= \frac{K\,(T_1 s + 1)}{T_2 T_3^2 s^4 + (T_3^2 + 2\zeta T_3 T_2)\,s^3 + (T_2 + 2\zeta T_3)\,s^2 + (KT_1 + 1)\,s + K} \tag{10}$$

The denominator of equation (10) consists of a fourth order polynomial. This is somewhat laborious to factor into first and second order terms. Oftentimes, in closed loop analysis of this type, polynomials of a much higher order are found making

the task of factoring almost prohibitive. It is, therefore, desirable to obtain a faster method of obtaining closed loop system transfers. This is easily accomplished using graphical methods.

In the case of the Polar plot, this becomes a straightforward operation. We wish to obtain the closed loop expression $\frac{G(s)}{1 + G(s)}$. On the Polar plot, we already have a plot of the function G(s). The (1 + G(s)) term can easily be obtained by adding $1/0^{\circ}$ to each point on the G(s) curve. Then the two curves may be divided, point by point, to obtain the closed loop response.

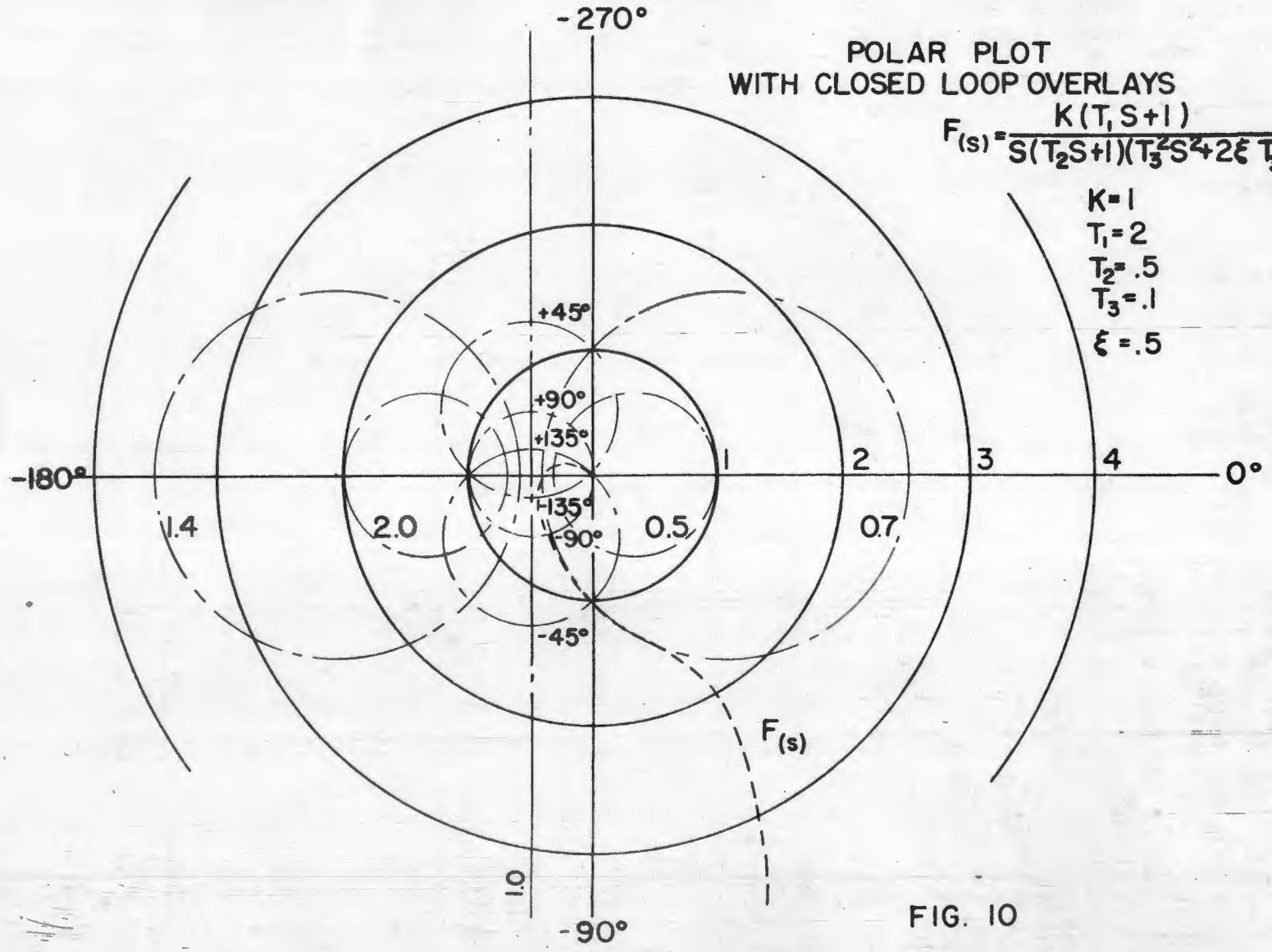
However, every point P on the polar coordinates has a corresponding (P + 1) point. By dividing these points, one obtains the closed loop gain and phase of the point P regardless of whether it lies on the curve of a transfer function. Thus, a family of equal closed loop gain curves and equal closed loop phase curves could be constructed on the Polar plot. Then, one merely has to plot an open loop transfer on the Polar coordinates, and by referring to this closed loop gridwork, determine the closed loop system response. Figure 10 shows these closed loop overlays along with a plot of our example problem. From this we can see that the closed loop response starts at approximately unity with small phase lags. As frequency is increased, the gain drops to about 0.7, levels off there for a short while, and eventually drops off to zero gain. The phase decreases gradually also, reaching -270° as the gain approaches zero.

This same type of closed loop overlay can be constructed for Nichols plots. Figure 11 shows the Nichols plot closed loop overlays along with our example problem. Again, we see the gain starts at about 0 Db, with about 0° phase lag. As frequency is increased, the gain drops to about -3 Db for a while, and then decreases rapidly in the vicinity of -180°.

These closed loop overlays can be used to find closed loop response when only experimental open loop data are available, and an analytical expression cannot be obtained. Likewise, if only closed loop experimental data are available for a system, the overlays can be used in reverse to obtain the systems open loop transfer.

CONCLUSION

In closing, we would do well to review some of the merits and draw-backs of the various methods of plotting transfer function characteristics. The Bode plots are by far the quickest and easiest to draw, but are poor for determining closed loop response. Closed loop overlays, which are available for the Nichols and Polar plots, make the task of determining closed loop transfers relatively simple. Determination of system stability from open loop plots can be somewhat confusing when working with Bode and Nichols plots, whereas definite rules which can be applied to the Polar plot remove all doubt of stability. The area of great interest on open loop plots is the vicinity of unity gain when the phase is -180°. In the Polar plot, this portion of the graph is somewhat compressed. The Nichols plot essentially takes the Polar plot, opens it up, plots gain in Db, and thereby expands the scale in the region of interest. The Nichols plot also finds applications when working with nonlinearities. It should also be pointed out that, even though the Bode plot requires two curves, the frequency information contained in this form of plotting is more obvious than that offered with the Polar and Nichols plots.



POLAR PLOT
WITH CLOSED LOOP OVERLAYS

$$F_{(s)} = \frac{K(T_1S+1)}{S(T_2S+1)(T_3^2S^2+2\xi T_3S+1)}$$

$K = 1$
$T_1 = 2$
$T_2 = .5$
$T_3 = .1$
$\xi = .5$

FIG. 10

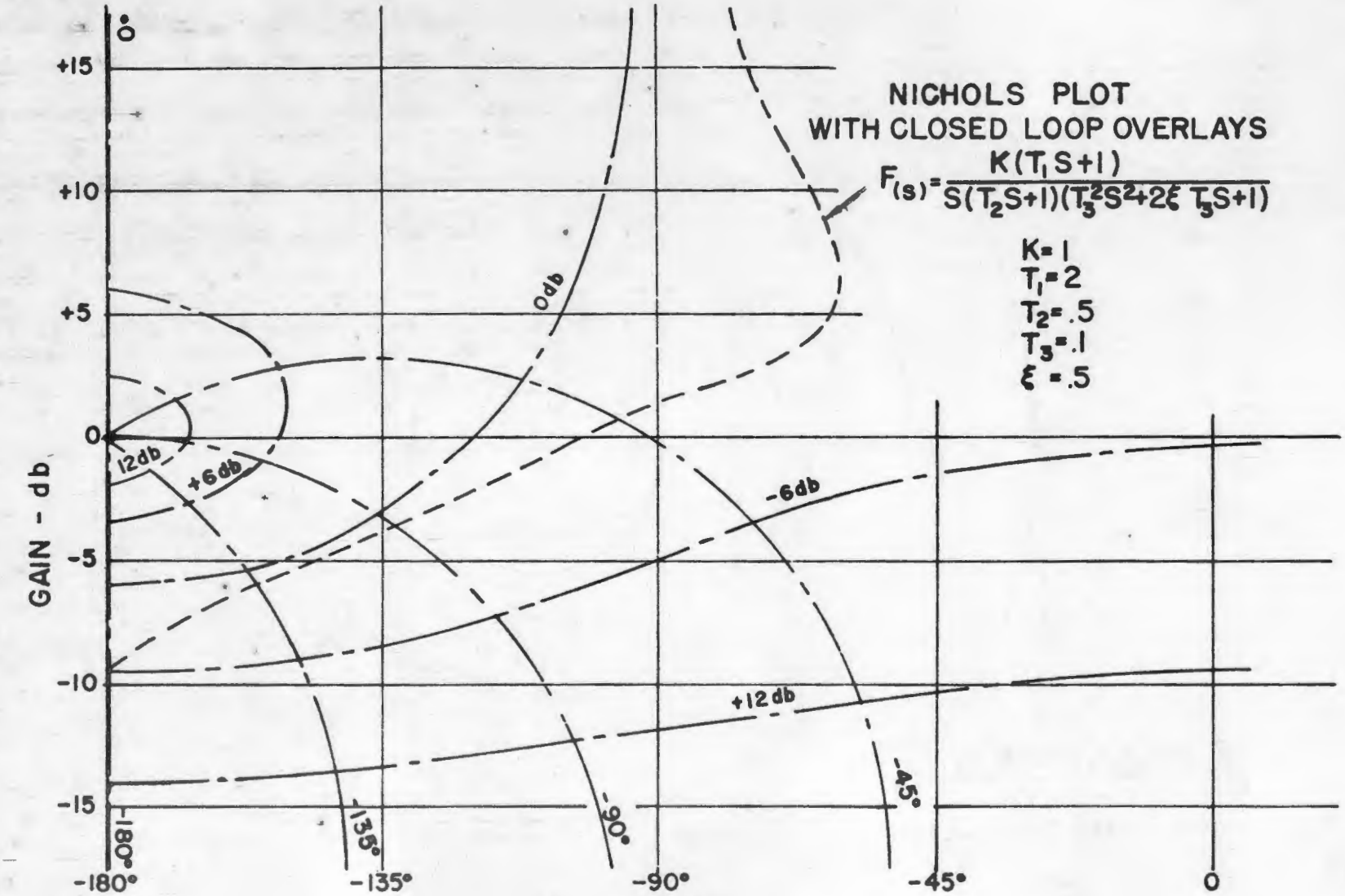

NICHOLS PLOT
WITH CLOSED LOOP OVERLAYS
$F_{(s)}=\frac{K(T_1S+1)}{S(T_2S+1)(T_3^2S^2+2\xi T_3S+1)}$
K=1
T_1=2
T_2=.5
T_3=.1
ξ=.5
0db
12db
+6db
-6db
+12db
-45°
-90°
-135°
-180°
+15
+10
+5
0
-5
-10
-15
GAIN - db
-180°
-135°
-90°
-45°
0
PHASE

FIG. 11

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

Lecture 4

STABILITY AND COMPENSATION

by

B. E. Amsler

APPLIED PHYSICS LABORATORY
THE JOHNS HOPKINS UNIVERSITY
SILVER SPRING MARYLAND

STABILITY AND COMPENSATION

by B. E. Amsler

Thus far in this series the basic theory commonly employed in control system analysis work has been discussed and some of the "tricks of the trade" have been given. In this final section, we will draw upon this previous work and show how it is applied in control system analysis work. In this section the main emphasis will be on the system stability analysis problem. However, the closed loop performance obtained from a given system function will also be given some attention.

Let us consider a simplified system as defined by the diagram of Figure 1:

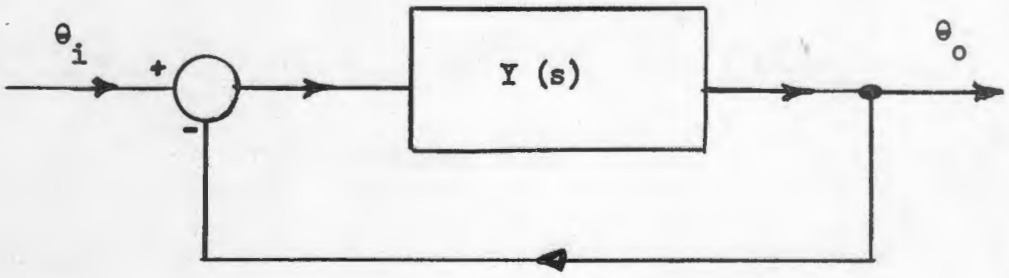


FIGURE 1

where Y(s) is any linear transfer function. Note that, in general,

$$Y(s) = K\frac{a(s)}{b(s)} \qquad (1)$$

where a(s) and b(s) are polynomials in "s" and are of the form

$$a_n s^n + a_{n-1} s^{n-1} - - - - - as + 1$$

while K is a constant.

STABILITY AND COMPENSATION

In control system work "K" is normally defined as the system gain while $\frac{a(s)}{b(s)}$ defines the frequency variant portion of the system transfer function. Note that $\frac{a(s)}{b(s)}$ is always factorable and can only be made up of terms of the form

$$(Ts + 1) \text{ and} (T^2s^2 + 2\zeta Ts + 1).$$

Thus the Bode plot of Y(s) can easily be constructed. The resultant closed loop transfer function for the system is given by:

$$\frac{\theta_o(s)}{\theta_i(s)} = \frac{Y(s)}{1 + Y(s)}$$

Substituting for Y(s) we obtain

$$\frac{\theta_o(s)}{\theta_i(s)} = \frac{Ka(s)}{Ka(s) + b(s)} = \frac{a(s)}{a(s) + \frac{b(s)}{K}} \qquad (2)$$

Thus the closed loop transfer function is also defined by the ratio of two polynomials in "s". It is evident that these polynomials are also factorable into a product of terms $(Ts + 1)$ and $(T^2s^2 + 2\zeta Ts + 1)$.

Now in Lecture 2 of this series, it was pointed out that a transfer function defined by the ratio of two polynomials could be separated into a sum of independent terms by partial functions, the denominator of each term being one factor of the denominator of the transfer function.

Thus

$$\frac{\theta_o(s)}{\theta_i(s)} = \frac{a(s)}{a(s) + \frac{b(s)}{K}} = \frac{A_1}{T_1s + 1} + \frac{A_2}{T_2s + 1} \cdots + \frac{B_1s + C_1}{T_{10}^2s^2 + 2\zeta T_{10}s + 1} + \cdots \qquad (3)$$

where

$$(T_1 s + 1)(T_2 s + 1) - - - - (T_{10}^2 s^2 + 2\zeta T_{10} s + 1) - - - -$$

are the factors of the denominator $a(s) + \frac{b(s)}{K}$.

The inverse transform of $\frac{\Theta_o}{\Theta_i}(s)$ is the impulse response of the system (see Lecture 2). Thus the impulse response is the sum of the exponentials given by the inverse transforms of the individual parts of (3).

A closed loop system is unstable if the impulse response is divergent.

Now

$$L^{-1}\left[\frac{A}{Ts + 1}\right] \triangleq \frac{A}{T} e^{-t/T}$$

$$L^{-1}\left[\frac{Bs + C}{T^2 s^2 + 2\zeta T s + 1}\right] \triangleq M e^{-\zeta t/T} \sin\left(\frac{1}{T} t + \Phi\right)$$

where M and Φ are also functions of T, ζ, B, and C.

Thus it is seen that the system is unstable if any of the first order denominator factors has a negative time constant T or if any quadratic factor has a negative damping coefficient ζ since only this could result in a positive exponent of "e". Note also that a negative T in the first order term gives simple divergence whereas negative damping in the quadratic gives divergent oscillation. In short, instability of the closed loop can result only if the denominator of $\frac{\Theta_o}{\Theta_i}(s)$ contains terms of the form $(-Ts + 1)$ or $(T^2 s^2 - 2\zeta T s + 1)$. Thus the question of system stability resolves itself to determining whether or not terms of this form exist in the denominator of the closed loop transfer function. In most

practical cases, the oscillatory type of instability is the troublesome one; simple divergence generally results from an incorrect feedback polarity.

It should be noted also that the frequency of oscillation for the quadratic (oscillatory) type instability is defined by the value of $\omega = \frac{1}{T}$ for the unstable quadratic factor.

There exist several methods of determining whether or not factors of this divergent type exist in the denominator. One method is Routh's criterion. This criterion relates the existence of these terms to the inter-relationship among the coefficients of the denominator polynomial. However, the usefulness of this method in control system work is limited since it does not define degree of stability. That is, we do not know whether one of the system parameters is considerably or only slightly out of line as regards stability requirements. Similarly, we are unable to tell by Routh's criterion whether a system is or is not "almost unstable". For example, the effects of parameter variations cannot be considered directly by using Routh's criterion.

Another obvious method of determining instability is simply to factor the denominator polynomial associated with any given system. If negative roots of the polynomial (that is, terms of the divergent form) exist, then the system is unstable. This straightforward method is in fact commonly applied in the case of basic system analysis work where the higher order terms associated with the practical hardware are not considered. However, the labor involved in factoring the equation for more than third order makes this method also prohibitive for evaluation of practical

systems where the important denominator terms may run to tenth and higher order. The method should not, however, be ignored, and ability to quickly factor third and perhaps fourth order expressions is extremely useful.

There does exist one method which is applicable to a system of any order and which defines not only the absolute stability of the system but also the degree of stability. This method is based on a criterion derived by Nyquist. The proof of this criterion is rather involved requiring the application of a complex variable theory and will not be covered here. However, a discussion can be found in almost any good servo mechanism text such as Truxal or James, Nichols and Phillips. The Nyquist criterion is in fact a graphical type criterion and its application employs those graphical techniques outlined in the previous section. Specifically, the Nyquist criterion in its complete form may be stated as follows:

> If the open loop response function of a negative feedback type control system is plotted in polar form for all values of ω from $-\infty$ to $+\infty$, the system will be stable if and only if the net number of counterclockwise encirclements of the point $(-1 + j0)$ is equal to the number of poles in the right half plane of the open loop transfer function.

Here the words "negative feedback type system" mean that the system block diagram shows one inherent sign reversal as in Figure 1.

As outlined in Lecture 2, the poles of a transfer function are simply the roots of the denominator. For s equal to a root of the denominator, the denominator vanishes and thus the total expression approaches infinity. Poles in the right half

plane mean roots for positive values of s. Consider a term, $Ts + 1$. It has a root and thus a pole at $s = -1/T$. This is on the negative real axis and is thus in the left half plane. But if T were itself negative, then the root is positive and we have a pole in the right half plane. Thus positive roots of a denominator mean unstable terms in the expression. That is, if our open loop contains a term of the form $\frac{1}{-Ts + 1}$, then we must encircle the $(-1 + j0)$ point one time counterclockwise. If the denominator contains a term $\frac{1}{T^2s^2 - 2\zeta Ts + 1}$ we have a complex conjugate root and thus an oscillating type instability in the open loop for positive s (in the right half plane). Thus we must make two counterclockwise encirclements. However, if our open loop transfer function is itself stable (the most common case) then the net number of encirclements of the point $(-1 + j0)$ must be zero.

In order to illustrate the use of the Nyquist criterion in this simple form, let us assume a specific open loop system which will allow direct factoring of the closed loop denominator. We will then compare the results obtained by direct factoring with the corresponding results indicated by the Nyquist criterion.

Specifically, let us assume a system of the form given in Figure 2:

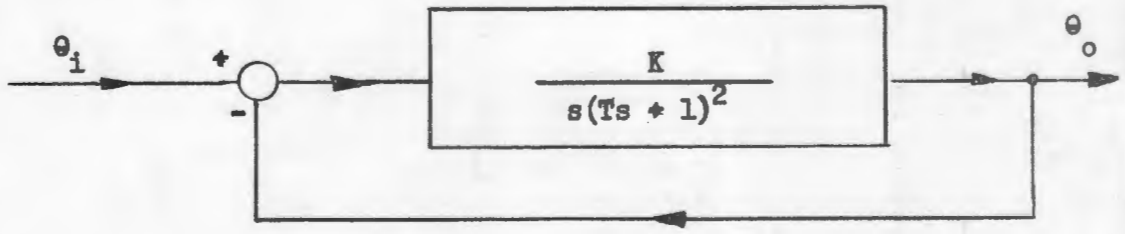


FIGURE 2

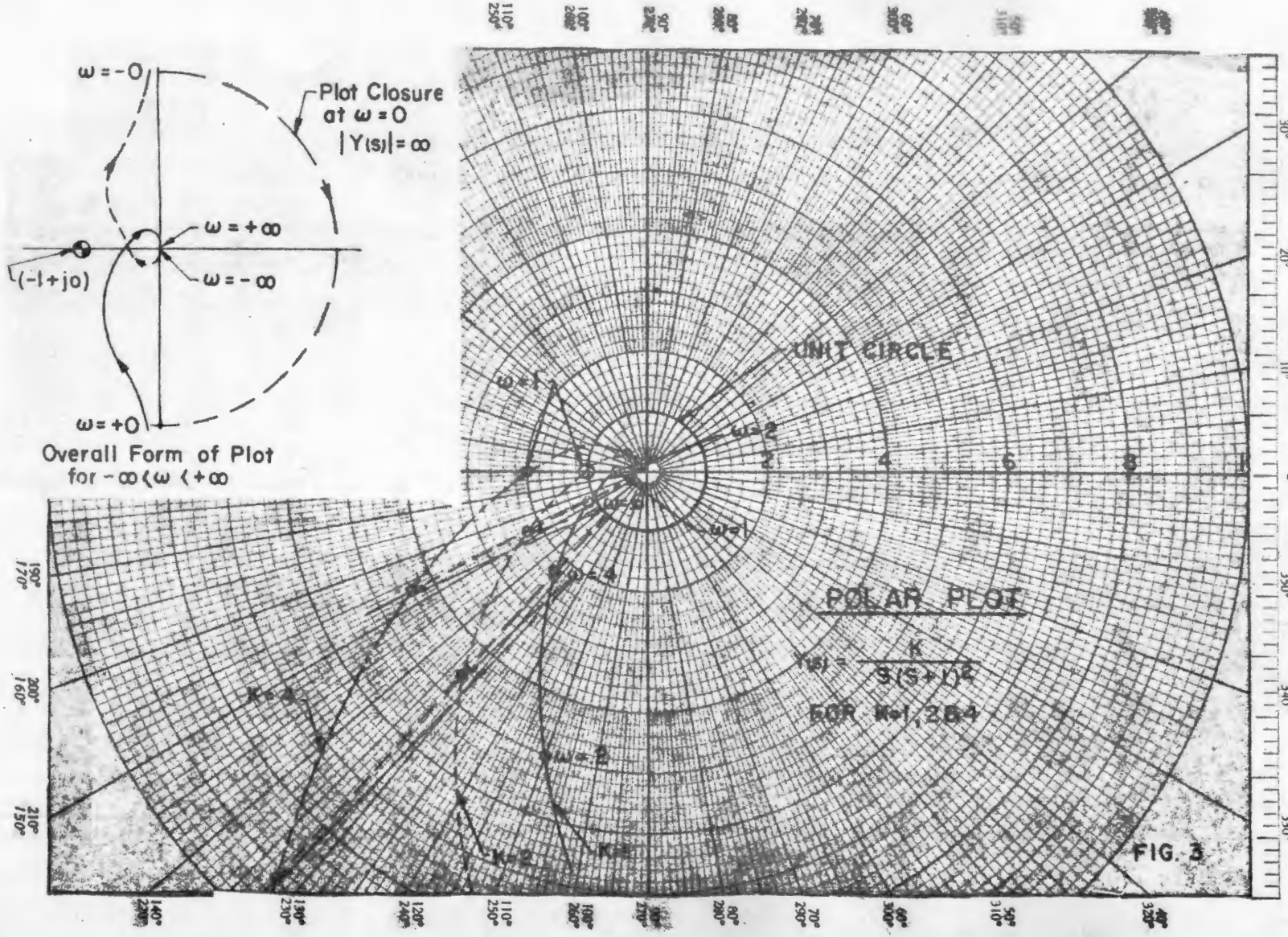
For this case we obtain

$$\frac{\theta_o}{\theta_i}(s) = \frac{1}{\frac{T^2}{K}s^3 + \frac{2T}{K}s^2 + \frac{1}{K}s + 1} \qquad (4)$$

In order to simplify the analysis, we will assume T = 1 second. The detailed polar plots for $\omega > 0$ for three values of gain K are shown in Figure 3. Note that for K = 1 the plot passes inside (to the right) of the point -1 + j0; for K = 2 the plot passes directly thru the point -1 + j0; for K = 4 the plot passes to the left of the point. In order to determine whether or not the -1 + j0 point is actually encircled, we must complete the plot for all values of ω from $-\infty$ to $+\infty$. The form of this plot is shown in the inset of Figure 3. The portion for $-\infty < \omega < 0$ is simply the mirror image of the portion $0 < \omega < +\infty$. Since the transfer function includes an integral term $\frac{1}{s}$, the amplitude open loop response is infinite at $\omega = \pm 0$. Thus the plot is closed at infinite amplitude for $\omega = 0$. In this type of system, the plot closure at $\omega = 0$ is always clockwise with 180° of rotation at $\omega = 0$ for every order of integration in the system.

Thus we see that the -1 + j0 point is not encircled for values of $K < 2.0$ but is encircled for values greater than this. Thus a value of K = 2 should define a boundary between stability and instability. Intuitively, we would expect the system to be neutrally stable for this value of K which would infer an oscillation that would be constant in amplitude.

In order to check the point, we will set K = 2, T = 1 in Equation 4. We can then determine the factored form of the denominator.


ω = -0
Plot Closure
at ω = 0
|Y(s)| = ∞
ω = +∞
(-1+jo)
ω = -∞
ω = +0
Overall Form of Plot
for -∞ < ω < +∞
UNIT CIRCLE
POLAR PLOT
Y(s) = K / S(S+1)²
FOR K=1, 2&4
K=4
K=2
K=1
FIG. 3

STABILITY AND COMPENSATION

Thus for K = 2

$$\frac{\theta_o}{\theta_i}(s) = \frac{1}{\frac{1}{2}s^3 + s^2 + \frac{1}{2}s + 1}$$

$$= \frac{1}{\left(\frac{1}{2}s + 1\right)\left(s^2 + 1\right)}$$

The term $(s^2 + 1)$ is in fact a special case of the general form $T^2s^2 + 2\zeta Ts + 1$ where ζ, the damping coefficient, is identically zero. In short, it represents a perfect sinusoidal oscillation term of constant amplitude. Thus the system is just neutrally stable. The frequency of oscillation is $\omega_o = \frac{1}{T} =$ 1 rad./sec. Referring to the polar plot, it is seen that this is just the frequency where $Y(s) = -1 + j0$.

Now let us substitute the other two values of gain into the Equation 4 and determine the resultant forms of $\frac{\theta_o}{\theta_i}(s)$.

For K = 1 we obtain

$$\frac{\theta_o}{\theta_i}(s) = \frac{1}{s^3 + 2s^2 + s + 1}$$

$$= \frac{1}{(.57\,s + 1)\,(1.325^2s^2 + 2 \times .162 \times 1.325\,s + 1)}$$

Note that the second term is a quadratic of the form $T^2s^2 + 2\zeta Ts + 1$ where the natural resonant frequency is $\omega_n = \frac{1}{T} = \frac{1}{1.325} = .755$ rad./sec. and the damping ratio ζ is 0.162.

STABILITY AND COMPENSATION -

Referring to Figure 2 of Lecture 3, we see that $\zeta = .162$ infers a rather high peak in the quadratic frequency response. We might, therefore, infer that for this value of gain K, the servo would be stable but would tend to ring badly at $\omega = .755$ rad./sec.

Similarly, for $K = 4$, we obtain

$$\frac{\theta_o}{\theta_i}(s) = \frac{1}{\frac{1}{4}s^3 + \frac{1}{2}s^2 + \frac{1}{4}s + 1}$$

$$= \frac{1}{(.432s + 1)\left(.785^2 s^2 + 2 \times (-.116) \times .785s + 1\right)}$$

For this case we find an unstable (divergent) response form since the damping coefficient associated with the quadratic is negative. The frequency of this divergent oscillation would be $\omega = 1/.785 = 1.275$ rad./second.

Thus, in summary, we see that the Nyquist criterion has indeed predicted the stability condition of the closed loop.

Referring again to the polar plot of Figure 3, it is seen that in order to prepare such a plot, we needed to define the phase and gain of the open loop transfer function at several frequencies so that the plot could be constructed. In the case considered, this is not difficult. However, in a higher order system, considerable labor is involved in computing the points by directly substituting $j\omega$ for s, and then determining magnitude and phase. On the other hand, as shown in Lecture 3, the Bode type plots (db gain and phase angle vs. frequency) are rather easily constructed. Thus if this form of plot can be used to determine

stability directly, considerable time and effort might be saved. The answer is, of course, that the plots can be used in this way provided that the form of the overall Nyquist (polar plot) is kept in mind. Referring to the inset in Figure 3, we see that this particular system is stable so long as the plot for $0 < \omega < \infty$ passes between the $-1 + j0$ point and the origin. In short, stability for this case requires only that the gain be less than unity (zero db) at a phase angle of -180°.

Corresponding Bode plots for the three values of gain are given in Figure 4. It is seen that the gain is exactly unity (zero db) at -180° phase for $K = 2$, is greater than unity for the case $K = 4$, and less than unity for the case $K = 1$. Thus stability can be directly determined from the Bode plot. Note also that changing the loop gain K simply effects the plot by moving the amplitude plot up and down. In the case considered, the gains considered were in 6 db increments. Thus the plots are each separated by 6 db.

In the case considered, the criterion for stability is rather obvious by direct inspection of the Bode plot. However, in more complex system problems, this is often not the case. A general rule is to plot the detailed open loop transfer function on a Bode plot. From this plot sketch (as per the inset of Figure 3) a polar plot to determine the form of the phase and gain relationship necessary for stability. The detailed evaluation is then performed on the Bode plot.

To illustrate this point, consider the open loop transfer function Y(s) and resultant Bode and polar plots sketched in Figure 5. In this case, the open loop transfer is rather complex, and direct factoring of the resultant fifth order closed loop would be extremely difficult. However, the Bode plot could be rather

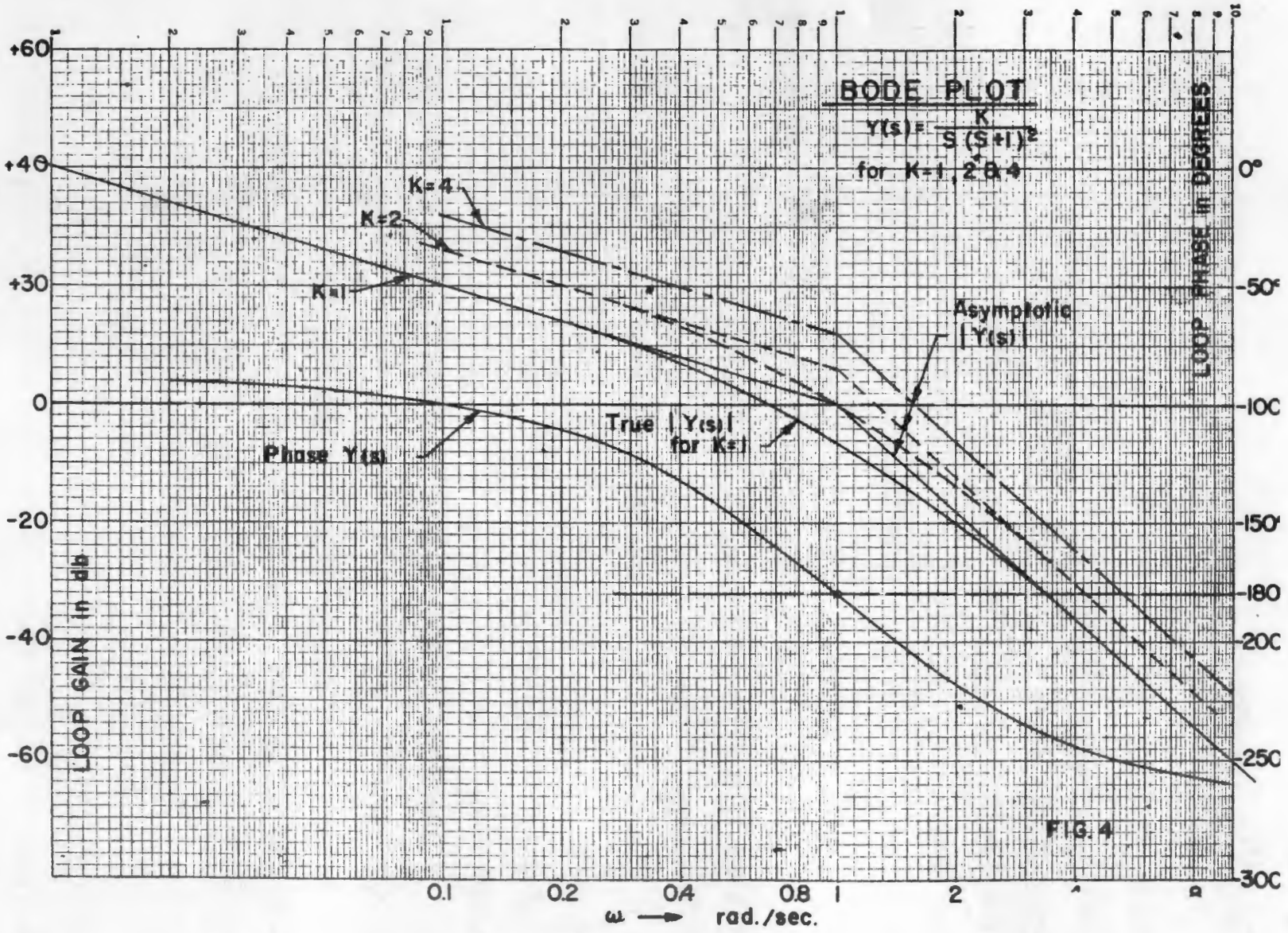
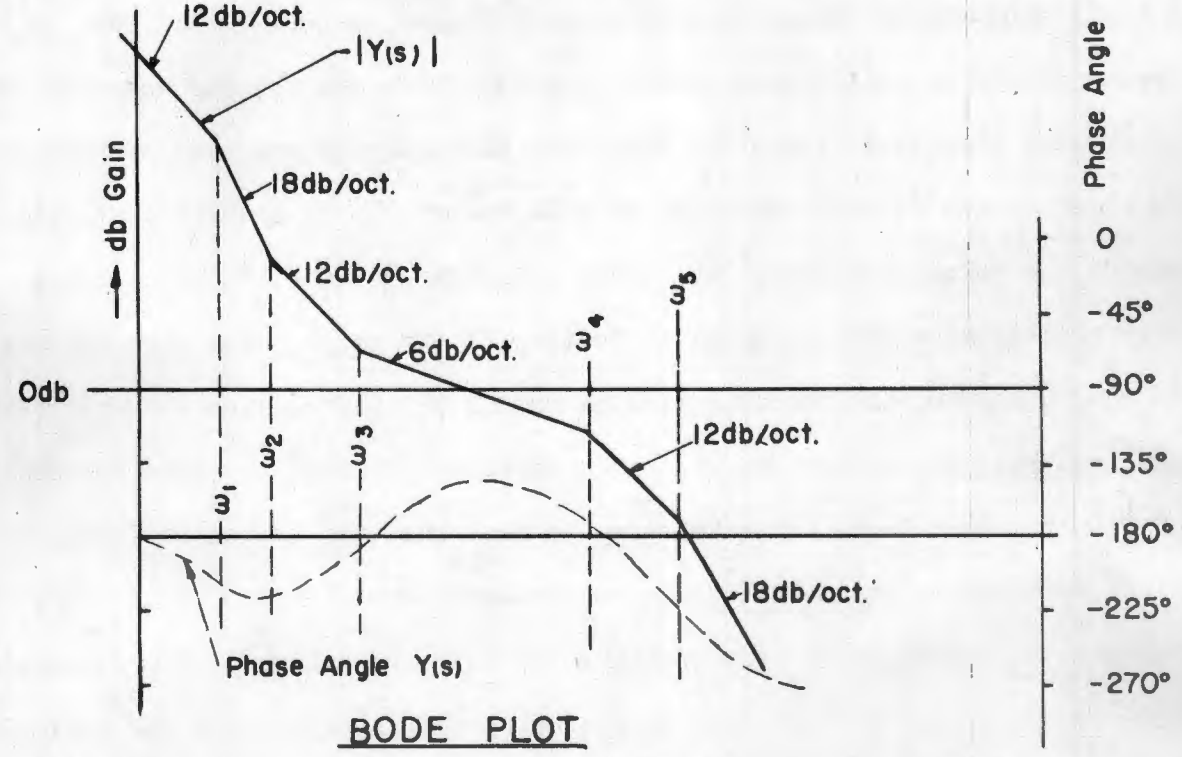
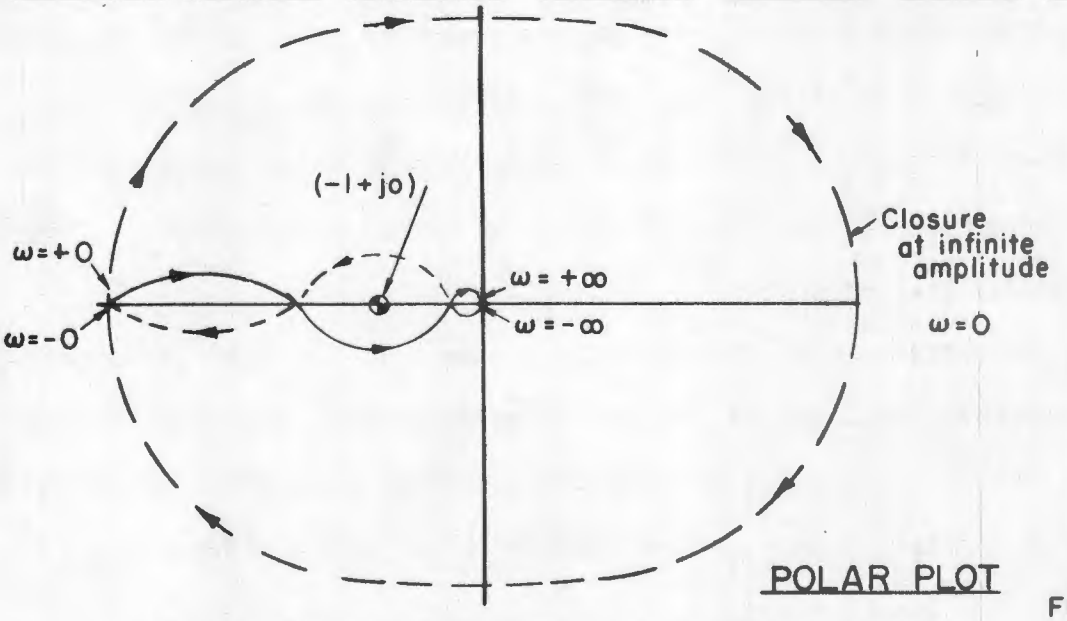

BODE PLOT
$Y(s) = \frac{K}{S(S+1)^2}$
for K=1, 2 & 4
K=4
K=2
K=1
Asymptotic |Y(s)|
True |Y(s)| for K=1
Phase Y(s)
LOOP GAIN in db
LOOP PHASE in DEGREES
+60
+40
+30
0
-20
-40
-60
0°
-50°
-100
-150°
-180
-200
-250
-300
0.1
0.2
0.4
0.8
1
2
ω → rad./sec.

FIG. 4



BODE PLOT

$$Y(s) = \frac{K}{S^2} \frac{\left(\frac{1}{\omega_2}s+1\right)\left(\frac{1}{\omega_3}s+1\right)}{\left(\frac{1}{\omega_1}S+1\right)\left(\frac{1}{\omega_4}S+1\right)\left(\frac{1}{\omega_5}S+1\right)}$$



POLAR PLOT

FIG. 5

BODE & POLAR PLOTS for a TYPICAL Y(s)

quickly constructed using the techniques outlined in Lecture 3. But, is the system stable or unstable as shown? Clearly there are regions where the gain is greater than unity (zero db) while the phase equals and even exceeds -180°. In order to decide this question, we will resort to the Nyquist criterion and sketch the polar plot for $\omega > 0$. With a little practice, these sketches can be very quickly made without precise plotting of any point. The plot for $\omega < 0$ is then completed and suitable closure made a $\omega = 0$ following the previously outlined rule where requires closure clockwise with 180° rotation for each power of $\frac{1}{s}$ in the open loop. In this case, we find that the system is stable as shown since we make one counterclockwise encirclement over the frequency range roughly between ω_3 and ω_4 when both positive and negative values of ω are considered. However, we also make one complete clockwise encirclement at $\omega = 0$ (infinite amplitude). Thus the net number of encirclements is zero, and since we have no poles in the positive half plane of Y(s), the system is stable as shown. Referring back to the Bode plot, it is seen that, in order to be stable, the open loop gain must cross the unity gain value within the frequency range of the "phase bulge" where the phase angle is less than 180°. If the gain is either raised or lowered so that gain crossover (point where the loop gain is unity) occurs outside this region, the system is unstable. This type of system is commonly referred to as a "conditionally stable" system.

In passing it might be noted that the open loop form just described is the type normally employed in the overall guidance loop of a typical beamriding missile. The simpler type previously considered in some detail is rather typical of the type employed in the control surface servos of the missile.

STABILITY AND COMPENSATION -

Thus far we have considered in detail the criterion of absolute stability. However, further consideration is required in that we are concerned not only with the question "Is it stable?" but also with the questions relating to the nature of the input to output transfer function resulting from a given open loop characteristic. We must also determine the allowable tolerances required within the system to insure maintenance of stability.

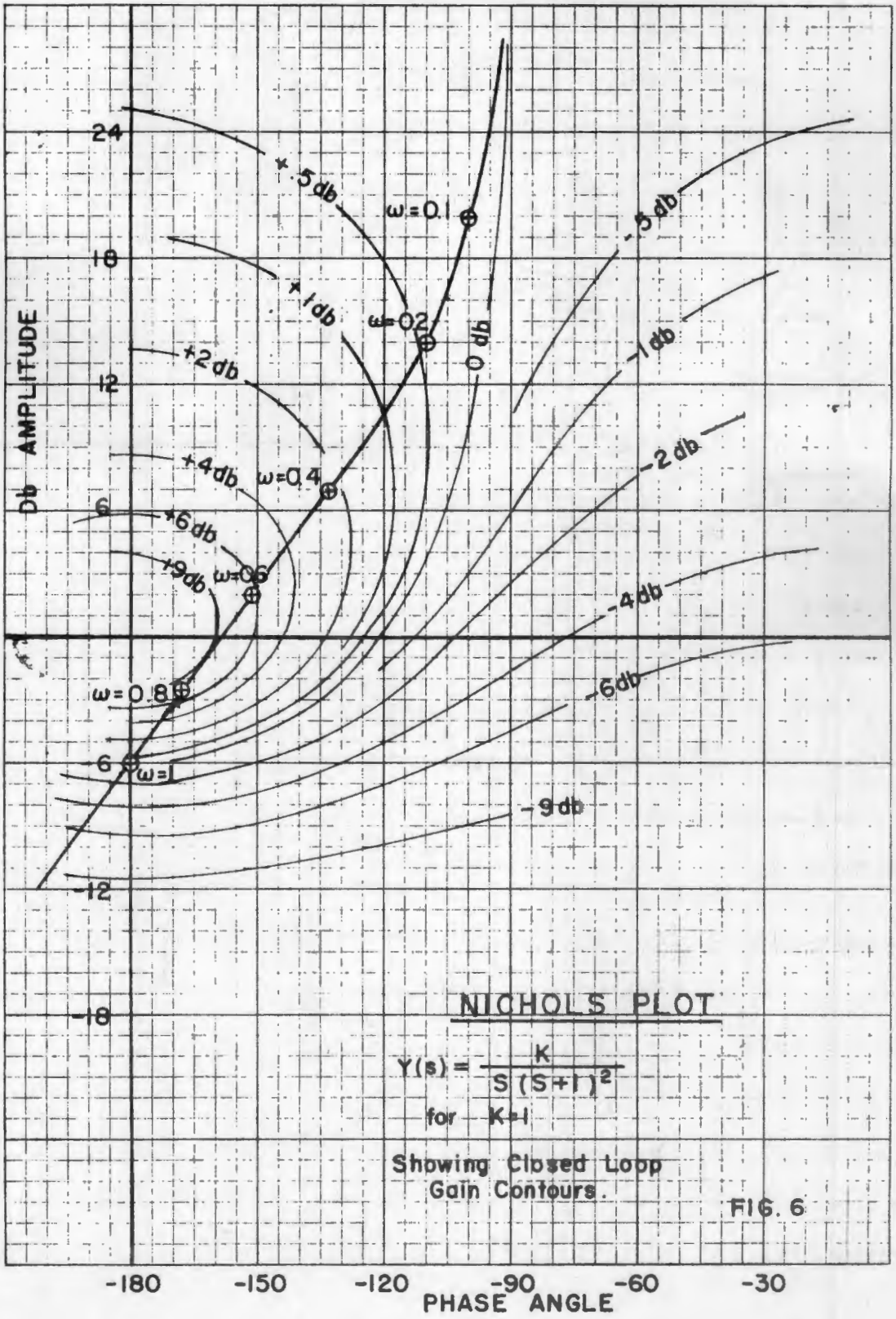
Toward these ends we often employ two more or less arbitrarily selected measurements of "stability margin". Stability margin is generally specified in two parts, specifically, the gain margin and the phase margin, which are defined as follows:

> Gain margin is the amount of system gain increase that is required to just produce instability. Thus it is given by the negative of the db gain (or inverse of numerical gain) at the frequency where the phase angle is -180°. In our example, the gain margin is +6 db for K = 1, 0 db for K = 2, and -6 db for K = 4.
>
> Phase Margin is the amount of additional phase lag that is required to just produce loop instability at the frequency where loop gain is unity. It is given by $180 - \Phi(\omega_o)$ where $\Phi(\omega_o)$ is the loop phase lag in degrees at ω_o, the frequency where loop gain is unity. In our previous example the phase margin was about $+22^{\circ}$ for K = 1, 0° for K =2, and -18° for K = 4.

These values are easily determined from a Bode plot.

Clearly negative gain and/or phase margin infers an unstable system and vice versa. Zero phase margin condition and zero gain margin condition occur simultaneously. Furthermore, as might be expected, the higher the value of margin, the "more stable" the system. Obviously, the value of the stability margin gives us some feel for system tolerances. It will also give us some insight into the expected closed loop performance characteristics even though no direct comparative set of values can be assigned.

A feel for the relationship between stability margin and closed loop performance can best be gained by considering the open loop plot of Y(s) on the Nichols plot together with an overlay of closed loop phase and gain contours outlined in Lecture 3. A Nichols plot of the previously used Y(s) is shown in Figure 6 for the K = 1. It is seen that changes in K correspond to moving the plot directly up or down as in the Bode plot. The gain margin is seen to be 6 db and the phase margin 22 degrees as defined by the intersections of the plot on the gain and phase coordinate axes. Note that for the case considered (K = 1), the curve just touches the +9 db closed loop gain contour $\frac{\theta_o}{\theta_i}$ (s); the value of closed loop gain falls off rather sharply on either side of this value. This sharp peaking is simply the peak associated with the rather small value of damping coefficient (ζ = 0.18) previously determined by factoring the resultant cubic. An inspection of the closed loop gain contours shown in Figure 6 shows that in order to have a closed loop response peak of less than 3 db, the phase margin must be about 40°. A phase margin of less than this must surely result in a peak of greater than 3 db since it would require that we cross inside of the 3 db contour.



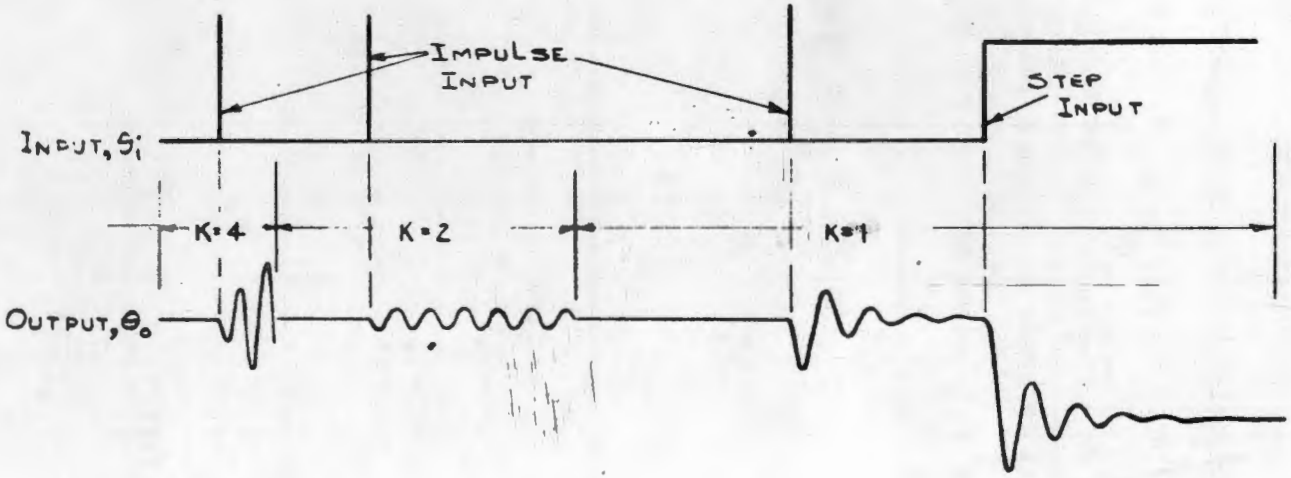
NICHOLS PLOT

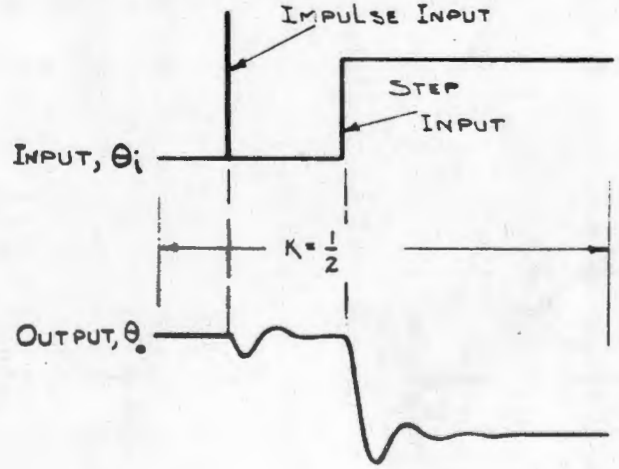
$$Y(s) = \frac{K}{S(S+1)^2}$$

for K=1

Showing Closed Loop Gain Contours.

FIG. 6





CLOSED LOOP
TRANSIENT RESPONSES
FOR VARIOUS VALUES OF K

OPEN LOOP TRANSFER

$$Y(s) = \frac{K}{s(s+1)^2}$$

K = 4, 2, 1 & 1/2

FIG 7

STABILITY AND COMPENSATION --

In order to reduce our peak for the case considered to less than 3 db we would require an additional loop gain reduction of 6 db which would give a phase margin of 43° and a gain margin of 12 db. It will be noted that for a simple quadratic term a 6 db peak corresponds to $\zeta \cong .25$, 3 db gives $\zeta \cong 0.4$, while a 9 db peak infers $\zeta \cong 0.17$. This infers that in order to obtain a reasonably well damped response characteristic $\theta_o/\theta_i(s)$, we must reduce the loop gain to approximately $K = \frac{1}{2}$.

Substituting this reduced value of loop gain ($K = \frac{1}{2}$) into Equation 4, we obtain

$$\frac{\theta_o}{\theta_i} = \frac{1}{2s^3 + 4s^2 + 2s + 1}$$

$$\cong \frac{1}{(0.635s + 1)\ (1.78^2 s^2 + 2 \times .375 \times 1.78s + 1)}$$

It will be noted that $\zeta = 0.375$ defines a quadratic peak of approximately 3 db. Thus we find that the results predicted by the closed loop contour overlay on the Nichols plot are verified by the direct factoring of the closed loop expression.

The closed loop results obtained with all the various assumed values of gain, K, are tabulated below for the general resulting form

$$\frac{\theta_o}{\theta_i}(s) = \frac{1}{\frac{s^3}{K} + \frac{2s^2}{K} + \frac{s}{K} + 1} = \frac{1}{(T_1 s + 1)\ (T_2^2 s^2 + 2\zeta T_2 s + 1)}$$

STABILITY AND COMPENSATION -

Gain K	T_1	T_2	ζ	Phase Margin	Gain Margin
4	.43	.785	-.116	-18°	-6 db
2	.5	1	0	0	0
1	.57	1.325	+.162	21°	6 db
$\frac{1}{2}$	.635	1.78	+.375	43°	12 db

Note that, in general, as K is decreased, the damping coefficient is raised, and the values of T_1 and T_2 are increased indicating decreased closed loop bandwidth. The corresponding increase in phase and gain margin is also given. This particular closed loop system was simulated on the analog computer and transient responses recorded for the assumed values of gain. A copy of the actual recorded run is given in Figure 7. Comparison of this Figure with the above table shows the correlation between transient (both impulse and step) response and the results obtained directly from the analysis. Note particularly the increase in damping and decrease in resonant frequency obtained as K is decreased.

The distribution of this document has been made in accordance with a list on file in the APL/JHU Technical Reports Group.