Paper No.
53—A-11

Frequency-Response Data Presentation Standards and Design Criteria

By RUFUS OLDENBURGER,[1] ROCKFORD, ILL.

This paper gives standards for the presentation of frequency-response data. The standards are recommended by the ASME-IRD Dynamic Systems Committee to facilitate the exchange of information directly and through the medium of publications. The Committee recommends that magnitude curves be plotted on logarithmic co-ordinates, and phase on the linear scale versus frequency on the log scale. Further recommendations involve transfer functions, measurements, nonlinearity, ambient effects, and other factors. The paper was written with the technician in mind, as well as the automatic control engineer and scientist. For this reason the reader with little or no mathematical background is carried as far as possible into the theory of the frequency-response field. Basic design criteria, in common use, as well as others, are given, and it is shown that control design, as far as dynamic performance is concerned, can be reduced, at least in rough analyses, to simple properties of curves that may be either theoretically or experimentally obtained. From the slope of one such curve the designer can often say whether or not a controller will be stable when placed on the system to be controlled. It is hoped that this paper will enable as many readers as possible to start on the scientific design of controls, and where they do so to use the standards here recommended. It is not implied that the reading of this paper is a substitute for a knowledge of some of the vast body of mathematics on which it is based. One objective of the paper is to give management, at least technical management, some insight into how automatic controls can be designed scientifically. The paper concludes with a discussion of the role the frequency-response approach plays among the design techniques available to the worker in the automatic control field.

Introduction

The ASME-IRD Dynamic Systems Committee, hereafter referred to as the "DS-Committee," was formed by the IRD Executive Committee November 26, 1951, to recommend standards for the presentation of frequency-response data. The necessity for such standards become apparent simultaneously and independently to several members of the ASME who were employing frequency-response methods in the design of automatic controls. Some confusion had resulted because different companies presented the same frequency-response information in different ways. Thus one company would graph magnitude ratio versus frequency, and another the logarithm of the magnitude ratio versus the logarithm of the frequency. From a glance at the frequency-response curves for a physical component a great deal can be said about the dynamic properties of the component, provided that one is familiar with the co-ordinates used. In order that the curves of one company be readily interpreted by another it is highly desirable that they plot their curves on standard co-ordinates. These and other considerations led to the formation of the DS-Committee. Since its founding, this committee has made a thorough study of the matter of standards. The standards it wishes to recommend are incorporated in this paper.

To facilitate the understanding of the DS-Committee recommendations, and the frequency-response approach to control design, some of the background will be covered in this paper by the author, and design criteria given so that the technician with little, or no mathematical training, can with this information at least begin the scientific design of automatic controls. Most of the design criteria have appeared in various places in the servomechanism literature. One of the objectives of the paper is to give technical management some insight into how automatic controls can be designed scientifically.

Summaries of design criteria and committee recommendations are given near the end of the paper.

Recommendations for Magnitude and Phase Curves

Consider a physical system as in Fig. 1, with an input m and an output c.

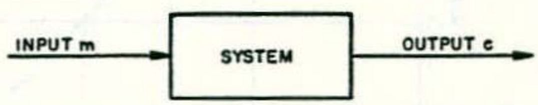

Fig. 1 Block Diagram of a Physical System With an Input m and Output c

The input is a variable whose value m controls the value c of the output. An example is the case of a diesel engine where the throttle controls the speed of engine. The variable m may be taken to be the throttle position in inches and c the speed of the engine in revolutions per minute. As is customary, the input and output are measured as deviations from equilibrium values, that is, values for which the engine speed and the throttle position are steady.

As the input is varied sinusoidally, so that $m = A \sin \omega t$ for the *magnitude* (amplitude) A, constant ω, and time t, the output eventually will vary sinusoidally if the system is what the mathematicians call linear. Experience and theory show that it is advisable for the reader to assume that the system he is studying can be treated as a linear system, unless proved otherwise.

The quantity ω is the product of 2π by the frequency f of the oscillation, that is, approximately 6.28 f. For steady oscillations the output of the linear system is $B \sin (\omega t + \phi)$ for a magnitude B and phase angle ϕ. For a linear system and a given frequency f the quantity B is proportional to the input magnitude A.

The quotient B/A is the *magnitude ratio* of the output to the input. In Fig. 2 the "magnitude frequency-response curve" of a diesel engine is plotted for the case of a load with small damping. For the example of Fig. 2 the quantity c is measured in units of a hundred rpm and the input m is in inches.

[1] Chief Mathematician, Woodward Governor Company. Mem. ASME.

Contributed by the Instruments and Regulators Division for presentation at the Annual Meeting, New York, N. Y., November 29–December 3, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 3, 1953. Paper No. 53—A-11.

Discussion on this paper will be accepted at ASME Headquarters until January 11, 1954

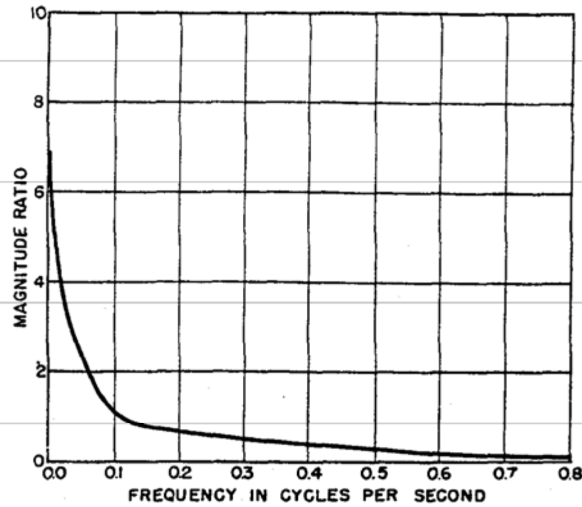

Fig. 2 Magnitude Curve of a Diesel Engine

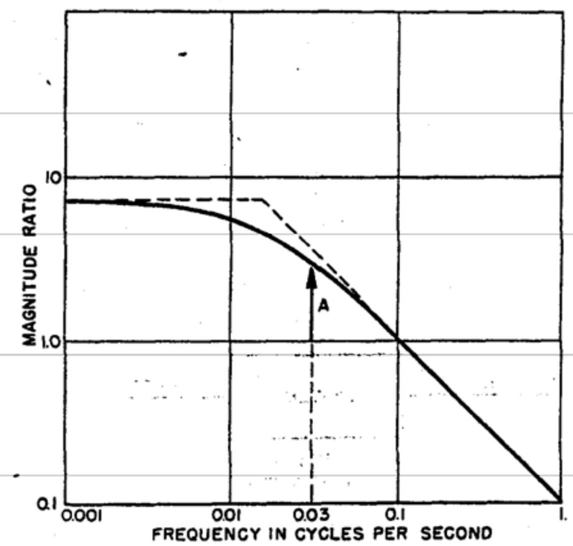

Fig. 3 Logarithmic Plot of Magnitude Response of Diesel Engine

The reader familiar with the frequency-response field will recognize the curve in Fig. 2 as that of a single-capacity system. Plotting the information of Fig. 2 on log co-ordinates,[2] one obtains the curve in Fig. 3. It will be noted that the curve in Fig. 3 can be approximated by two asymptotic straight-line rays as shown dashed in Fig. 3. The reader will recall that if the part of a straight line to one side of a point on the line is removed, what is left is a "ray." The curve formed by these rays is called the *asymptotic magnitude curve* of the engine. Since logarithms of real numbers to one base are proportional to the logarithms of these numbers to another base, the base associated with the log co-ordinates is immaterial.

Suppose that the engine speed is measured and that the difference between this speed and a speed setting, which it is desired to maintain, actuates a governor that positions the throttle. The difference can be considered to be the input to the governor and the throttle position to be the output. If the input is oscillated sinusoidally, the output will do the same, assuming again that the system is linear. A plot of a typical governor is given in Fig. 4.

[2] These plots are often called "Bode plots."

The curve in Fig. 4 can be approximated by the asymptotic curve shown dashed in Fig. 4, composed of two straight-line asymptotic rays represented by the first and last segments and two intermediate straight-line segments. To obtain this asymptotic curve it is really necessary to have the mathematical formula, that is "transfer function" for this curve. Where the technician is not acquainted with these functions, he will probably have to work with the actual response curves.

Suppose now that the governor is placed on the engine, but that the loop is open as shown in Fig. 5. Here the engine speed-measuring element is disconnected from the rest of the governor, theoretically if not actually. If the input e (actuating signal) to the governor is oscillated sinusoidally the output c (controlled variable) of the engine, namely, the engine speed (measured from equilibrium), will vary sinusoidally (for practical purposes). By mathematics it can be shown that the magnitude curve of the system composed of the governor and engine is obtained by merely adding ordinates on the magnitude curves for the components, plotted on logarithmic co-ordinates. Thus the vectors A and B in Figs. 3 and 4, at 0.03 cycle per second, are added to get the vector $A + B$ in Fig. 6 where the magnitude curve of the governor-engine open loop is plotted. It is convenient to think of the horizontal line labeled 1 as the horizontal axis because the logarithm of one is zero. The vectors A, B, and $A + B$ are thus measured from this axis.

To use the method of addition of ordinates it is necessary for the output of the controller and the input to the process (controlled system) to be measured in the same units. For the analysis of this paper the input e to the controller and output c of the process must be measured in the same units.

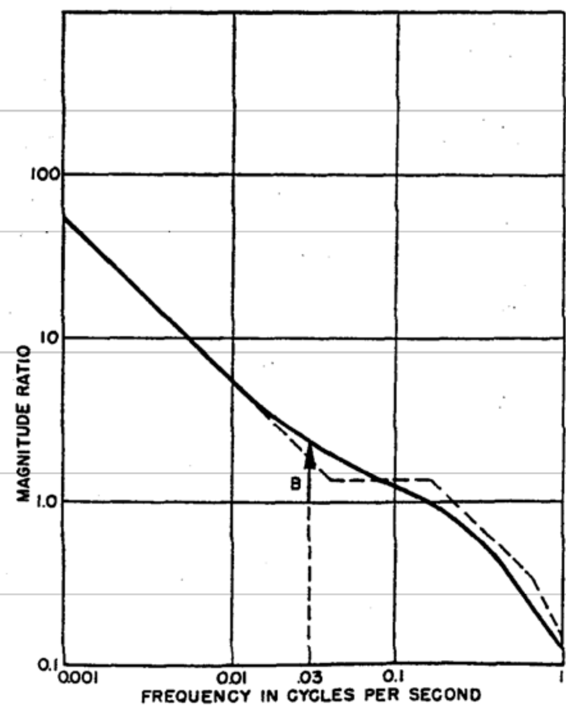

Fig. 4 Magnitude Curve of an Engine Governor

Fig. 5 Magnitude Curve for the Governor-Engine Open Loop

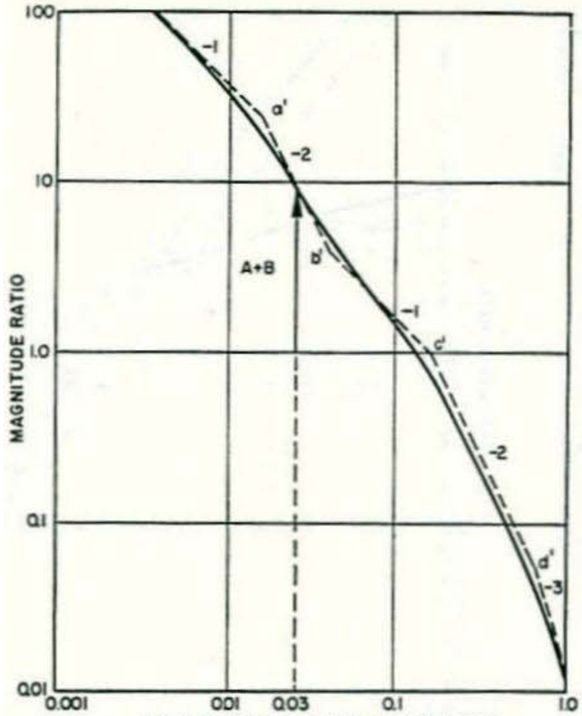

Fig. 6 Magnitude Curve for Governor-Engine Open Loop

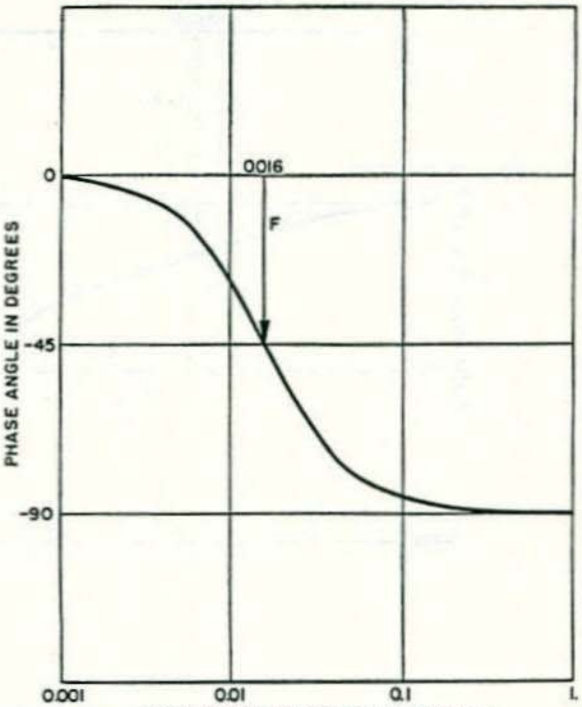

Fig. 7 Phase Curve for Diesel Engine

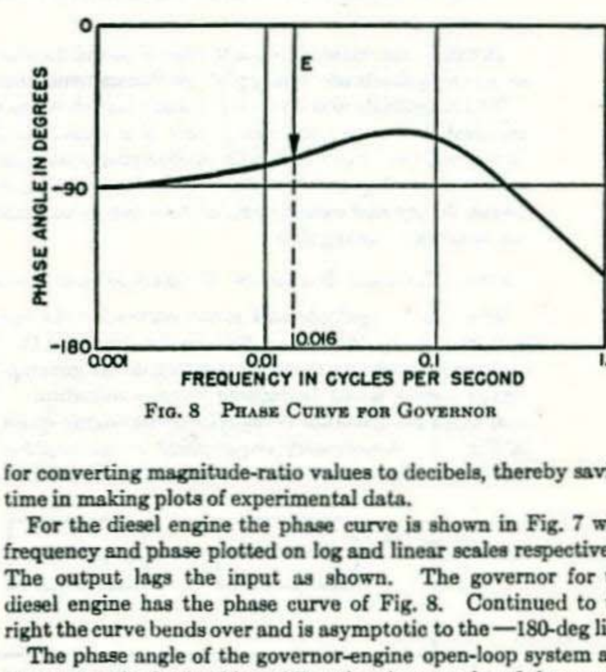

Fig. 8 Phase Curve for Governor

If the corresponding ordinates of the asymptotic curves in Figs. 3 and 4 are added, the broken line in Fig. 6 is obtained. This is the asymptotic magnitude curve of the governor-engine open loop. It is seen that the resultant curve obtained by using the straight-line approach is not much different from the correct curve.

Increasing the magnitude ratio of the governor-engine system by a factor K, regardless of frequency, is said to "raise the gain" of the system by K. The definition of "gain" is beyond the scope of this paper, and will be omitted (5).[3]

Raising the gain of the governor-engine system corresponds merely to shifting the magnitude curve of Fig. 6 upward. Gain is often readily adjustable in a physical system. When logarithmic co-ordinates are used one can often work with straight-line asymptotic approximations to the magnitude curves. Because of these considerations, the following recommendation is made by the DS-Committee:

Recommendation 1. It is recommended that magnitude curves be plotted on logarithmic co-ordinates with the "magnitude ratio" as the vertical co-ordinate and the "frequency" in cycles per unit time as the horizontal co-ordinate.

Except in England the frequency f is used in preference to the period P as the co-ordinate on the horizontal scale. Oscillators are generally marked for cycles per unit time, such as cycles per second, and it takes an additional computation to obtain radians ω per unit time. For these reasons the use of frequency is recommended.

It may be desirable to plot frequency on a nondimensional scale.

Let R be the magnitude ratio. It will be recalled that the magnitude ratio in *decibels* is $20 \log_{10} R$. The use of decibels is standard in the Bell Laboratories approach. However, the use of the logarithmic scale for magnitude ratio eliminates the need for converting magnitude-ratio values to decibels, thereby saving time in making plots of experimental data.

For the diesel engine the phase curve is shown in Fig. 7 with frequency and phase plotted on log and linear scales respectively. The output lags the input as shown. The governor for the diesel engine has the phase curve of Fig. 8. Continued to the right the curve bends over and is asymptotic to the —180-deg line.

The phase angle of the governor-engine open-loop system at a frequency f is obtained by adding the phase angles of the components at the frequency f. Thus the vectors E and F in Figs. 8 and 7 add up to the vector $E + F$ in Fig. 9 where the phase curve is plotted for the system composed of the governor and the engine.

The use of a linear scale in angular degrees to express the phase difference between the input and output of a physical system is well established. The following is therefore recommended by the DS-Committee:

Recommendation 2. It is recommended that "phase" be plotted in degrees on a linear vertical scale and "frequency" in cycles per unit time on a horizontal logarithmic scale.

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

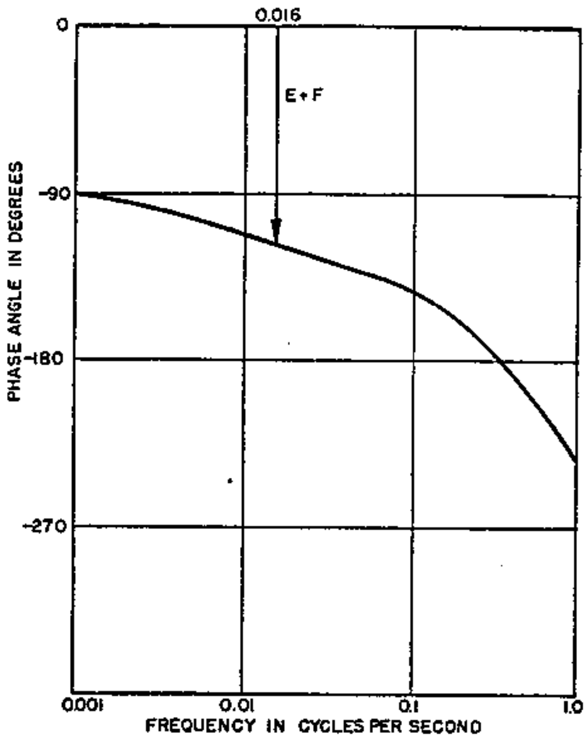

Fig. 9 Phase Curve for Governor-Engine Open Loop

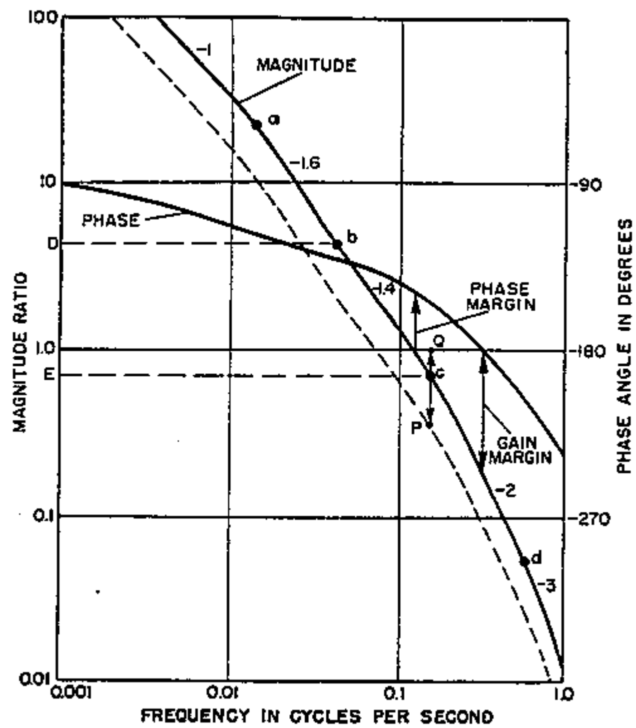

Fig. 11 Response Curves for Governor-Engine Open Loop

It will be convenient to use the same unit of measure for cycles on the magnitude and phase plots of Recommendations 1 and 2.

The magnitude and phase curves are just about as typical of a physical device as fingerprints are of a human being. These curves can be used to obtain the differential equations of physical systems, to diagnose troubles when there is disagreement between theory and experiment, and to design automatic controls for systems to be regulated.

Design Criteria Based on Magnitude and Phase Curves

From the magnitude and phase curves for the open loop the performance of the closed loop is determined (1). Now let r (reference input) be the speed setting in the governor, and c the actual engine speed (measured from equilibrium). The situation when the governor is regulating the engine speed is pictured in Fig. 10. A necessary requirement of the control is that the system of Fig. 10 be stable. From the magnitude and phase

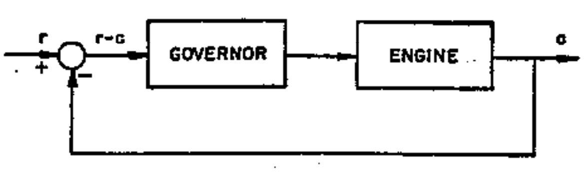

Fig. 10 Closed Loop

curves for the open loop of Fig. 5 a great deal can be said about the stability of the closed loop, and rough design rules can be stated. To save space it is convenient to plot both magnitude and phase curves on the same sheet of paper. For the engine example the curves are given in Fig. 11.

Dr. Harry Nyquist of the Bell Laboratories wrote the pioneering paper (2) of the frequency-response field. By advanced mathematics he solved the problem of stability for closed loops and linear systems by reducing the problem to the study of the open-loop response. Dr. Hendrik Bode of the Bell Laboratories extended this work and introduced (3) the concepts of gain and phase margin to be defined immediately. In terms of these concepts some simple design rules can be given.

Suppose that a small disturbance δ is applied to a system S, where δ eventually dies out. If for each such disturbance δ the resulting response of S dies out, the system S is said to be *stable.* Otherwise, S is *unstable.*

When the open-loop magnitude ratio is 1, we are at *gain crossover*. Suppose for the moment that there is only one point of gain crossover. This may be said to be the normal case. If we subtract the phase lag at gain crossover from 180 deg, the *phase margin* is obtained; that is, the phase margin is the sum of 180 deg and the phase angle. It is to be recalled that the phase angle is negative for a lag, and positive for a lead. If there is more than one point of gain crossover the angular difference just defined is formed for each point, and the minimum is defined to be the phase margin. The phase margin is indicated in Fig. 11 for the example under discussion. In this case the phase margin is about 30 deg. The reader will note that when the magnitude ratio (gain in the Bode sense) is 1, its logarithm is zero, and the magnitude curve crosses what may be conveniently thought of as the horizontal axis.

If we are at a condition where the phase lag is 180 deg (phase angle is −180 deg) we are at a point of *phase-crossover*. Suppose first that there is only one such point. This may be considered to be the normal case. Let R be the magnitude ratio at phase crossover. The gain margin is then $1/R$ if R is less than 1, and equal to R if R is greater than or equal to 1. The gain margin is often measured in decibels. If there is more than one point of phase crossover, the "gain margin" is by definition the minimum of the quantities $\{R\}$ with $R > 1$ and $\{1/R\}$ with $R \leq 1$ for the R's at phase-crossover points. In Fig. 11 a vector has been drawn to indicate the gain margin. *It is not the length of the vector.* For the example of Fig. 11 the magnitude ratio R at phase crossover is approximately 0.2. The gain margin is then 1/0.2, that is, 5. In decibels this is 20 $\log_{10}$ 5, or simply, 14 db.

The lag for the open loop may never reach 180 deg. In this case the gain margin is not defined. Similarly, the open-loop magnitude ratio may not reach 1, in which event the phase margin is not defined. Where a margin is not defined, and it is referred to in the design rules, the designer does not have to concern himself with it.

For the purpose of exposition it will be convenient to extend the definitions of gain and phase margins to some limiting cases. If

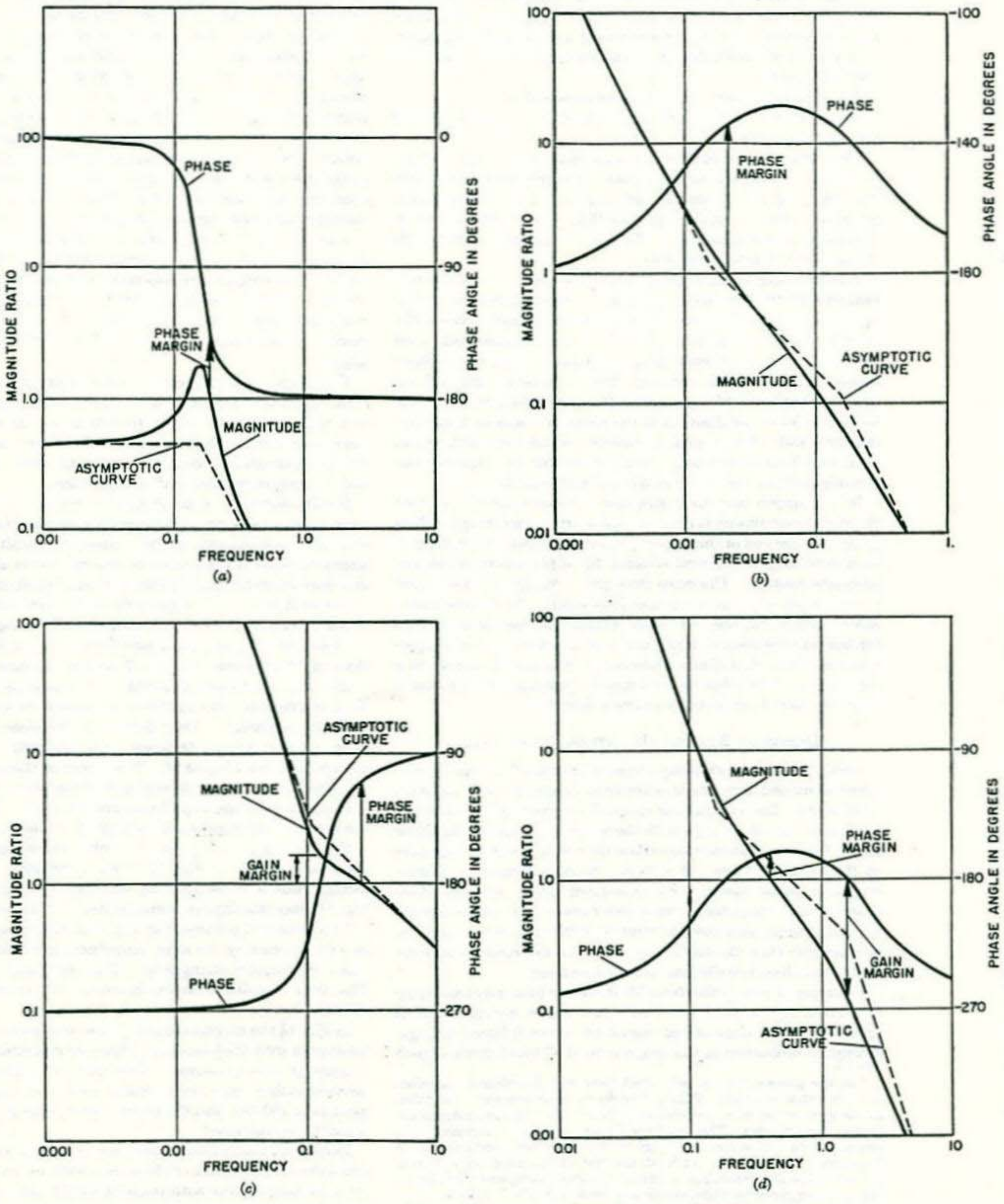

Fig. 12 Gain and Phase Margins for Some Physical Systems

the lag never reaches 180 deg but approaches 180 deg as the magnitude ratio goes to zero, the gain margin will be said to be infinite, and will then be denoted by ∞. Similarly, if the magnitude ratio never reaches 1 but approaches 1 as the lag goes to zero, the phase margin is said to be 180 deg.

From the laws of physics (4) and energy limitations it is known that as the frequency increases beyond all bounds the magnitude ratio of the open loop approaches zero or a finite positive number.

In Fig. 12 are shown gain and phase margins for some open-loop physical systems.

The following design rule is in common use (5–9):

Design Rule 1. The phase margin should be at least 30 deg and the gain margin at least 2.5 (8 decibels).

These rules can be justified on mathematical grounds. It can be shown by complex variable theory that the condition of 180-deg phase lag and magnitude ratio equal to 1 for the open loop is to be avoided. Rule 1 insures that this condition will be avoided by a definite margin. This will be clarified further in the discussion of transfer locus plots.

The foregoing design rule should be considered a rough criterion that enables the technician untrained in mathematics, as well as one who is trained, to take the frequency-response curves of the control and the system to be controlled, and eliminate some poor designs. It can be proved that, in general, it is not safe to allow smaller phase and gain margins. Thus a phase margin of 5 deg or gain margin of 1.05 would place the system in a sense so close to the border of instability that the solution would be highly oscillatory, and if the frequency-response curves were obtained all or in part from theory rather than experiment the neglected factors might throw the system into complete instability.

It can happen that the design rule above is satisfied, but that the closed-loop system is still too oscillatory. An example will be given near the end of this paper. It can be shown, however, that for a wide range of physical systems the rules provide completely adequate control. The rules thus give a margin of safety that is often sufficient, but sometimes falls short. To be sure that a control will work, the technician without the background to go further into the theory might make use of a library of frequency-response curves that characterize good performance, and see to it that the curves he obtains for a specific problem are identical or close to a pair of curves known to be acceptable.

Design on Basis of Magnitude Curves Only

Bode (3) treated primarily physical systems that can be represented by ordinary linear differential equations with constant coefficients. Let a magnitude curve W be given and consider the systems of the above type with the same W. Among the phase curves for these systems there is one for which the phase lags take on the smallest values. A system with such a curve is called a *minimum phase system*. For minimum phase systems Bode showed that magnitude curve determines the phase curve.[4] Since it is sometimes more difficult to obtain the phase curve experimentally than the magnitude curve, it is desirable to have design criteria based on the magnitude curve alone.

From one of Bode's theorems (3) it follows that the phase angle at a given frequency f, for a minimum phase system, depends primarily on the slope of the magnitude curve (plotted on logarithmic co-ordinates) in the neighborhood of f and, if the slope is reasonably constant in the neighborhood of f, the phase angle is proportional to the slope of the magnitude curve. In fact, the phase angle is approximately ($90\,m$) deg for the slope m.

To illustrate this relationship the reader is referred to the solid magnitude curve of Fig. 11. The part of the curve to the left of the point a has the slope −1 approximately. The reader will recall that a line with slope −1 falls one unit for each unit movement to the right; that is, the line makes an angle of 135 deg with the horizontal axis. By Bode's result the phase angle should be approximately −90 deg for low frequencies, which checks with the figure. For the segment c-d of the magnitude curve the slope is approximately −2 (the line drops two units for each unit movement to the right) and the corresponding phase angle should be about −180 deg, which also agrees with the figure. To the right of the point d the slope is approximately −3, whence the corresponding phase angle should be about −270 deg. For high frequencies the phase curve is asymptotic to the −270-deg line.

The slope of the curve from a to b is approximately −1.6, and from b to c about −1.4. Therefore we would expect the phase angle for the frequencies associated with the arc from a to c to be about −135 deg, which is correct. It is evident that there is a rough correlation between the slopes −1, −2, and −3 of the segments of the asymptotic curve of Fig. 6 and the actual phase angles.

For points on a long section of the magnitude curve, but away from the ends of this section where the slope m along the section is a fairly constant integer, the formula $-90\ m$ deg is a good approximation to the phase angle. Near the ends of the section the approximation becomes a very rough one. For a short section the approximation is also very rough.

If gain crossover occurs at a point interior to a long section of constant slope −2, the lag at gain crossover will be about 180 deg, and the phase margin will be numerically small. To insure an adequate phase margin the following rule (7) is used for the common case where the magnitude curve is falling at gain crossover:

Design Rule 2. The slope of the magnitude curve on logarithmic co-ordinates at and near gain crossover should be about −1.

The slope −1 has been chosen because −1 is definitely greater than the "dangerous" value −2, and at the same time it is generally, or at least often, desirable to stay as close to the slope −2 as is practical in order to obtain a transient response that is sufficiently oscillatory. Let f denote the frequency at gain crossover. *It is customary to design for a slope near −1 over the frequency range from $^1/_2 f$ to $2f$.* The extent of the region over which the slope should be −1 to each side of the point of gain crossover depends on the slope away from gain crossover. With experience the designer will know about how much to allow for this.

The dashed curve of Fig. 11 is obtained by moving the solid magnitude curve of Fig. 11 down a distance equal to log 2. Design Rule 2 is satisfied by the dashed magnitude curve of Fig. 11 since the slope is approximately −1.4 (this is near enough −1 for practical purposes) at and near gain crossover. It is not so well satisfied by the solid magnitude curve since the slope on this curve rapidly changes to −2 as one passes gain crossover. The solid magnitude curve, however, still represents a stable system.

In Fig. 13 the magnitude and phase curves are shown for a stable system with the loop open, where the magnitude curve rises at a point of gain crossover. This may be considered to be the unusual rather than the common case. The phase margin is more than 180 deg, and the system has a large margin of stability when the loop is closed.

In plotting the frequency-response curves, it is not necessary to know the precise nature of these curves for all frequencies. Thus for a physical system with transients that endure 1 or 2 min it is not normally necessary to know the response at 100 cycles per

[4] In the process-control field dead time and distributed lag often play an important role. Where they do one does not have minimum phase systems, in fact, the ordinary linear differential-equations approach breaks down. The reader will then need both magnitude and phase curves. However, although in this case the design theory of the present section does not hold, the rest of the paper still applies. An extensive mathematical treatment of dead time and distributed lag has been given by Oldenbourg and Sartorius (10). Also see Bode (3), chapter 13.

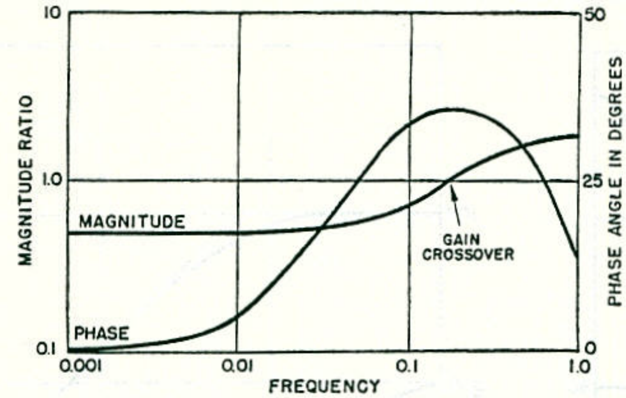

Fig. 13 Magnitude and Phase Curves of Stable System

minute (cpm) or 0.01 cpm. In fact, a range of 10 cpm to 0.1 cpm is often sufficient.

The frequency-response curves should be plotted for an intermediate range of frequencies, and not necessarily for extremely low frequencies. Where gain and phase margins are defined these curves should be known for a range of frequencies that extend well beyond the frequencies at the crossover points. A common case is that where the asymptotic magnitude curve in the neighborhood of gain crossover is composed of three sections with slopes —2, —1, and —2, respectively, as shown in Fig. 14. The magnitude curve (solid or dashed) of Fig. 11 is an example to which Fig. 14 applies. To insure adequate stability margins the

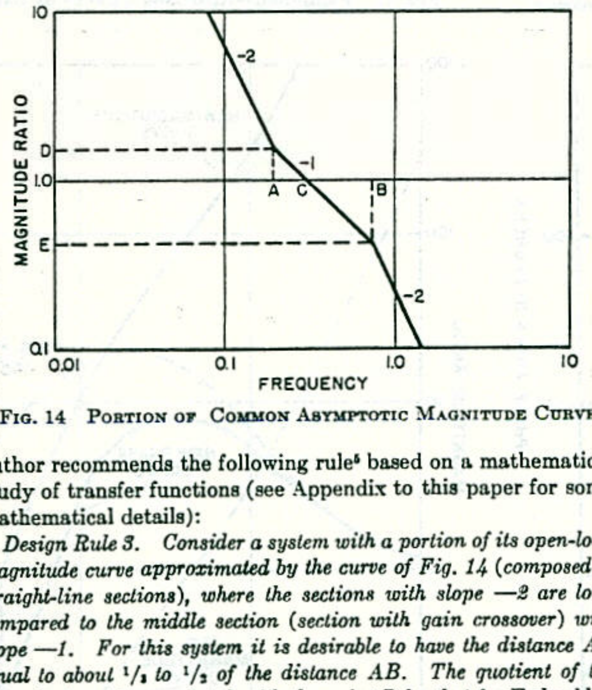

Fig. 14 Portion of Common Asymptotic Magnitude Curve

author recommends the following rule[5] based on a mathematical study of transfer functions (see Appendix to this paper for some mathematical details):

Design Rule 3. Consider a system with a portion of its open-loop magnitude curve approximated by the curve of Fig. 14 (composed of straight-line sections), where the sections with slope —2 are long compared to the middle section (section with gain crossover) with slope —1. For this system it is desirable to have the distance AC equal to about $^1/_3$ *to* $^1/_2$ *of the distance AB. The quotient of the magnitude ratio associated with the point D by that for E should be at least 5 (ratio of frequency associated with B to that for A should be at least 5).*

It can be easily shown that the magnitude ratio associated with the point D of Fig. 14, divided by the magnitude ratio associated with E, is equal to the frequency associated with B divided by the frequency associated with A.

It follows from Bode's phase-magnitude relationship that increasing the slope of a magnitude curve (making it less steep where the curve is falling) increases the phase angle. If now one or both sections adjoining the middle section with slope —1 are short sections with slope —2 adjoining sections with greater slope (not as steep), *the ratio of the frequencies* associated with the points B and A *can be somewhat less than* 5, such as 4.

For the governor-engine example of Fig. 11, the middle section *bc* of the solid magnitude curve has a slope about —1.4 while the adjoining sections *ab* and *cd* have the approximate slopes —1.6 and —2, respectively. The points *a, b, c,* and *d* marked on this figure are points of demarcation between sections with essentially different slopes, and may be called "break points." In the strict sense the expression *break point* applies to a point where straight-line sections of different slopes meet. For purposes of exposition we shall use the more liberal interpretation given here. On the magnitude curve in Fig. 11 the break points are not unique, but are chosen visually.[6] For the asymptotic magnitude curve they are unique. In Fig. 6 they are the points *a', b', c',* and *d'*. In a very rough way the section *ab* may be treated as one with a slope approximately equal to —2 and the section *bc* as one with the approximate slope —1. The vertical distance from *c* to *b* is the log 6.5, whence the quotient of the magnitude ratios corresponding to *DE* of Fig. 11 is 6.5. Design rule 3 is satisfied in this respect.

The sections *a'b', b'c',* and *c'd'*, of the asymptotic magnitude curve for the governor-engine example have the slopes —2, —1, and —2, respectively, Fig. 6. The vertical distance between *b'* and *c'* is log 4, whence the corresponding ratio is 4. This does not satisfy Design rule 3. However, the section *a'b'* is not long compared to *b'c'*, and the ray on the asymptotic curve to the left of *a'* has the slope —1. It follows that the *quotient of the magnitude* ratios associated with the points *b'* and *c' can be somewhat less* than the 5 of Design rule 3, and the actual value 4 is quite satisfactory.

The point of gain crossover on the solid magnitude curve, Fig. 11, is near the right end of the middle section. For the asymptotic curve, Fig. 6, it is at the right end. Design Rule 3 is thus violated in this respect. Let the logarithms be taken to the base 10. Moving the magnitude curve down a distance 0.3, as shown in the dashed curve in Fig. 11, brings gain crossover at a point in agreement with Design Rule 3, namely, to or near the mid-point of the middle section. This is true whether we work with the magnitude curve or its asymptotic approximation.

As noted before, displacing the magnitude curve on logarithmic co-ordinates up or down *changes* the gain of the open loop accordingly. This is one of the major advantages of plotting magnitude on these co-ordinates. Doubling the gain merely multiplies each amplitude ratio by 2 and thus adds $\log_{10} 2$, that is, 0.3, to each ordinate. Halving the gain corresponds to moving the magnitude curve down a distance 0.3 since $\log_{10} 0.5 = -0.3$. As noted before, in the example of Fig. 11, we have halved the gain in going from the heavy to the dashed curve. *Changing the gain does not affect the phase curve.*

The foregoing application of Design Rule 3 may appear to lack precision, but this rule has been illustrated with an actual arbitrarily chosen example. Nevertheless, the rough methods given here yield good quantitative results.

The desire to have an adequate but not excessive phase margin is a reason for the foregoing rule. The author will introduce here an adjunct intended to help the designer secure adequate gain margin and stability. This adjunct is based on a study of transfer functions. It is assumed that we have the common case where the magnitude curve is falling at gain crossover.

Let R be the magnitude ratio at a point P. The *magnitude*

[5] In (9), p. 185, the ratio AC/AB equal to $^1/_3$ is recommended.

[6] The writer asked six technicians one at a time to select the break points on this curve, given only the curve. All of them came remarkably close to the four break points of Fig. 11.

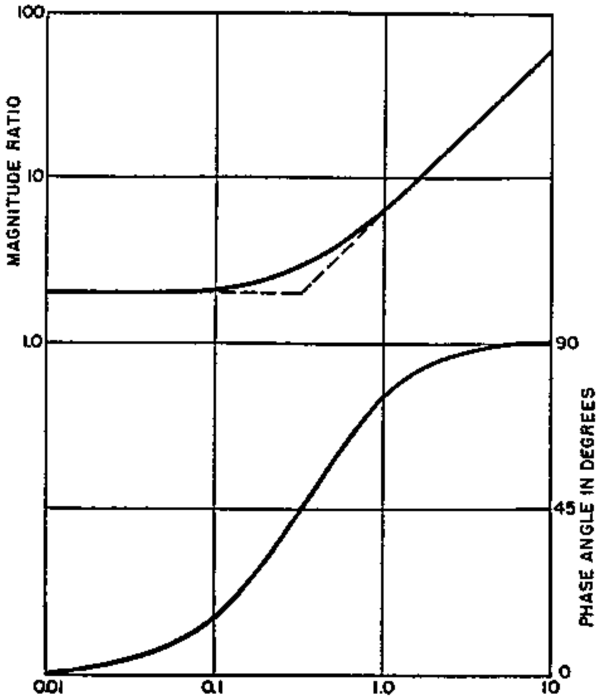

Fig. 15 Frequency-Response Curves of Lead Component

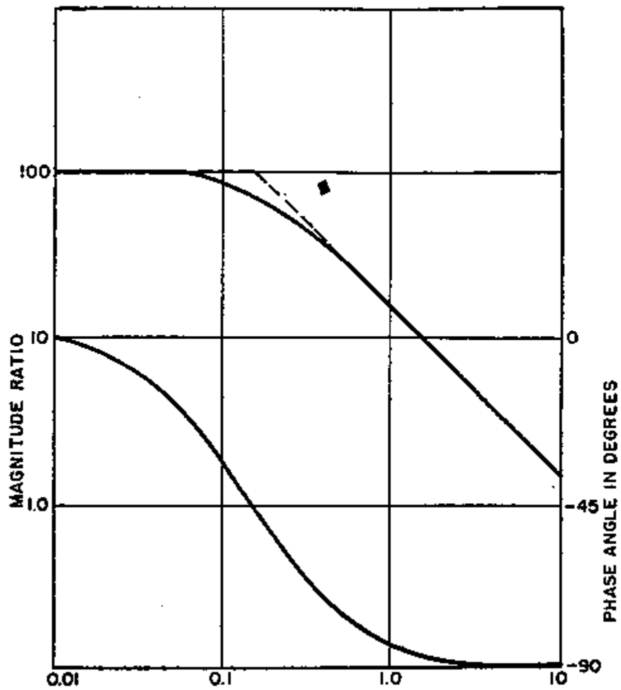

Fig. 17 Frequency-Response Curves of Lag Component

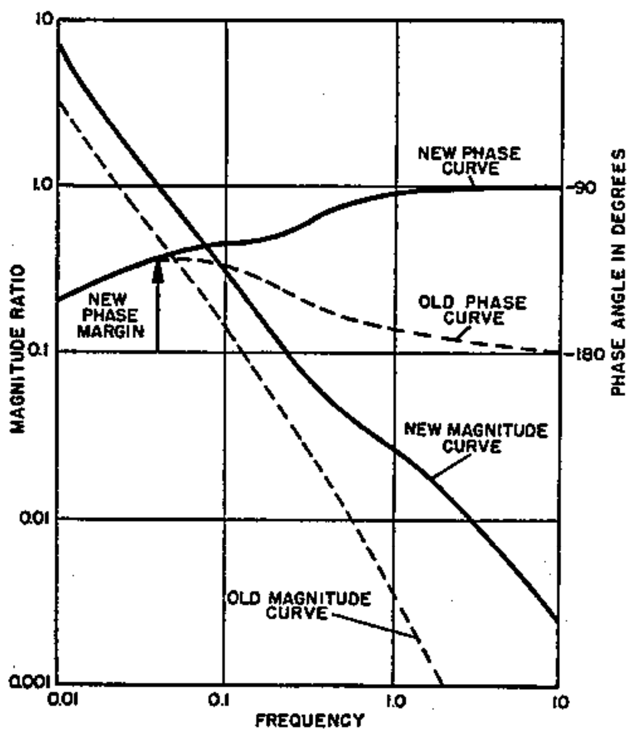

Fig. 16 Frequency-Response Curves Modified by Lead Component

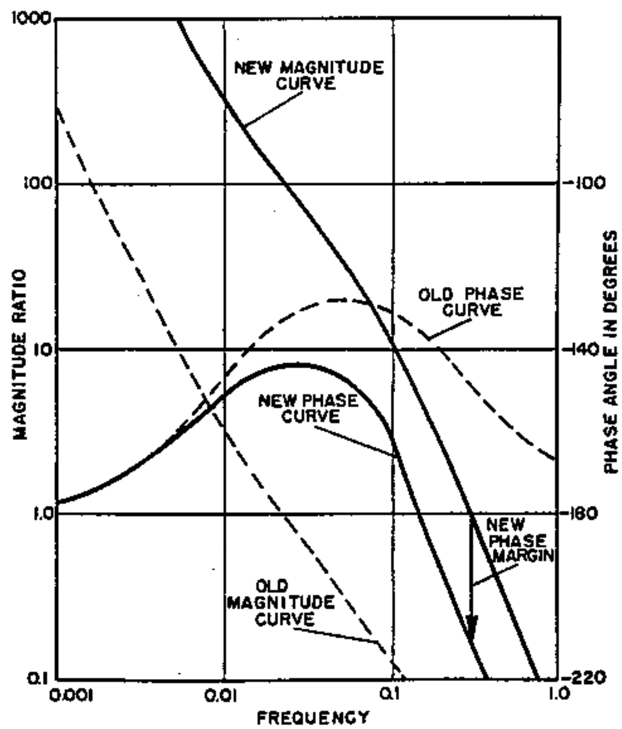

Fig. 18 Frequency-Response Curves Modified by Lag Component

margin at P is by definition $1/R$ if $R < 1$ and R if $R \geqq 1$.

Design Rule 4. If after gain crossover the slope of the magnitude curve is more than —2 (curve not as steep as when the slope is —2) until a point P is reached, and the slope eventually is less than —2 (steeper) the magnitude margin at P should be 2 (6 decibels) or more.

The point P is shown on the dashed magnitude curve in Fig. 11, and the magnitude margin QP at P is 3 (9 decibels). On the solid curve this margin is inadequate.

On an asymptotic curve the point P is uniquely determined and Design Rule 4 is then particularly easy to apply.

We noted previously that Bode's phase-magnitude relationship implies that increasing the slope of the magnitude curve (making it less steep where the curve is falling) increases the phase angle. Thus if the point of gain crossover with which the phase margin is associated is kept fixed and the slope of the magnitude curve is increased (less steep where the curve is falling), *the phase margin is improved.* From the same Bode relationship it follows that if the slope of the magnitude curve never is less than —2 (no steeper on the falling portions), the phase lag never exceeds 180 deg, and there is no gain margin. Any portion of a magnitude curve where the slope is less than —2 (steeper) tends to increase the phase lags and deteriorate the phase margin.

Innumerable design rules like those of this section can be formulated. No group of such rules can replace the direct mathematical study of a given problem, but they can help the engineer in his rough analysis of a control problem.

For higher frequencies lag components tend to decrease phase angles whereas lead components tend to increase them.

The frequency-response curves for an example of a simple lead component are shown in Fig. 15. Adding ordinates, the response curves of the system of Fig. 12(*b*) change to those of Fig. 16 when this lead component is incorporated.

For a lag component we have the corresponding curves in Fig. 17, and the result of adding ordinates to those of Fig. 12(*b*) yields the curves of Fig. 18 for the system obtained. At higher frequencies the new magnitude curve drops below the old one.

The system of Fig. 16 is stable, whereas that of Fig. 18 is unstable.

Systems Without Unity Feedback

Consider the system of Fig. 19 where there is a component in the feedback path so that the input to the control is not $r - c$ as in Fig. 10.

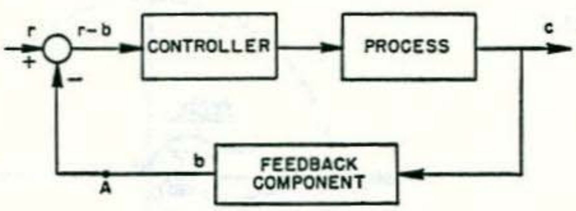

Fig. 19 System With Component in Feed-Back Path

The previous theory applies here, but to obtain the open-loop magnitude-frequency response curve, as when we open the loop at A, we must add ordinates of the three magnitude curves for the controller, process, and feed-back component. The same is true for the phase curve. The case of Fig. 19 is a common one since the variable c to be controlled must be measured and the measuring element will generally involve leads and lags. The controller responds to the difference r-b between the input r and the output b of the feed-back component. Thus a temperature control with a thermocouple measuring element is sensitive to the difference between the temperature setting and the temperature as measured by the thermocouple. The thermocouple has a thermal lag.

If for the range of frequencies involved in the stability of the loop shown in Fig. 19, the response of the feed-back component can be neglected, the loop of Fig. 19 reduces to that of Fig. 10. This could be the case for a temperature control with a fast thermocouple.

Transfer Locus (Nyquist) Plots

The DS-Committee does not recommend standardizing on any one type of frequency-response replot. The type of replot that most clearly demonstrates the effects of a component on over-all performance depends on the location and character of the component and upon the criteria of performance desired. The application of frequency response to some common problems is well understood. However, its application to complex control problems, particularly process control, has not been evaluated fully. Industrial processes are subject to multiple disturbances and interacting effects, many of which are known to exist, but few of which have been quantitatively determined. Means of demonstrating these effects will be developed by the practising engineer. In publishing his findings, he should be free to choose and devise the replots best-suited to his problem. However, since the committee recommends that the magnitude and phase frequency-response curves be given on logarithmic co-ordinates, the reader will fin d it convenient to use the log magnitude-phase charts to be discussed later. The DS-Committee therefore makes the following suggestion:

Recommendation 3. Where it is convenient to do so the use of log magnitude-phase charts (also referred to as Black or Nichols charts) in preference to other types of replots is recommended.

The *transfer locus plot* is a commonly used and widely recognized replot. This is the well-known Nyquist plot and combines in one curve the information supplied by the magnitude and frequency-response curves of an open loop (the other replots are also called "transfer locus plots" in the literature, but we prefer to reserve this terminology for the Nyquist plots). This plot is obtained by drawing a curve on polar co-ordinates where the vectorial angle for a frequency f is the open-loop phase angle, and the length of the radius vector is the open-loop magnitude ratio for the frequency f. The frequency is thus the parameter along the curve.

The solid curves in Fig. 11 for the governor-engine open loop combine to give the transfer locus of Fig. 20. Thus when $f = 0.064$, the magnitude ratio is 3 and the phase angle —136 deg as shown for the point P. The light portions of the curves in Fig. 20 to the left of the vertical axis are distorted to show how the curves are asymptotic to the vertical axis. Also, the curves are enlarged in the neighborhood of the origin to make the small loop visible.

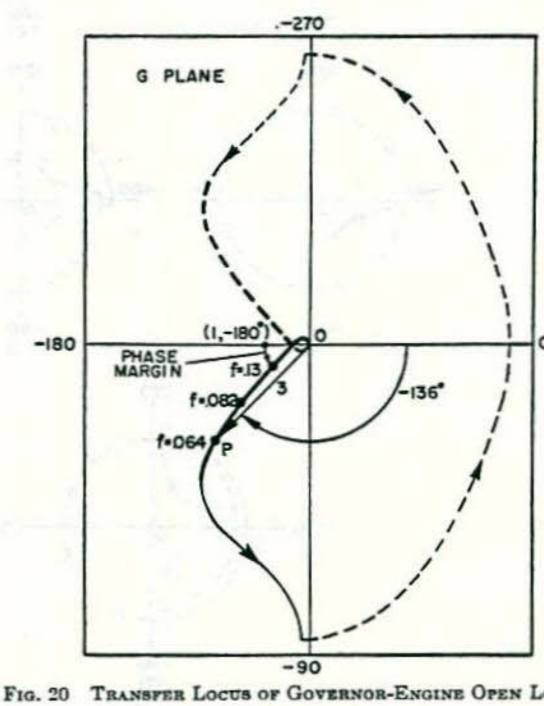

Fig. 20 Transfer Locus of Governor-Engine Open Loop

To enlarge the scale of the pertinent portion of the plane the DS-Committee makes the following recommendation:

Recommendation 4. Where readability and accuracy of data extraction can be improved by moving the co-ordinate axes to one side or corner of the transfer locus (Nyquist) plot, this is recommended.

Mathematical Background

It is hoped that even though the mathematically untrained reader may not understand the details of the present section he will at least obtain a general idea of the theory involved.

The dependency of the output of a linear system on the input (11) is described by the "transfer function" $G(s)$ of a complex variable s, where $G(s)$ is related to the concept of impedance in electrical theory. It can be shown that the transfer locus is a plot of this function for $s = j\omega$, where $j = \sqrt{-1}$, and $\omega = 2\pi f$ for the frequency f, while f varies from 0 to ∞. By drawing a symmetrical branch with respect to the horizontal axis the conjugate curve is obtained, associated with $G(j\omega)$ for ω ranging from 0 to $-\infty$. For the governor-engine example the conjugate curve is shown dashed in Fig. 20 as a branch running from the origin upward and asymptotic to the —270 deg ray. For the transfer functions normally treated (such as those analytic except for poles) theory shows that if the magnitude of the radius vector increases beyond all bounds as ω goes to zero, the transfer locus is asymptotic to one of the vertical or horizontal rays, emanating from the origin O. If the transfer locus terminates at finite points, the locus and its conjugate form a closed curve. The end of the transfer locus associated with $\omega = \infty$ is a finite point (4). If the other end ($\omega \to 0$) is asymptotic to one of the rays emanating from the origin, the curve formed by the transfer locus and its conjugate is closed by drawing a properly chosen arc. In physical systems the locus is often asymptotic to the —90-deg ray and sometimes to the —180-deg ray. It is fairly uncommon to have the locus asymptotic to the —270-deg ray. It therefore will be sufficient for what follows to show the reader how to close the curve formed by a transfer locus and its conjugate in the cases where the —90, —180, or —270-deg ray is the asymptote.

From complex variable theory involving the Nyquist stability criterion the curves must be closed in a certain way (the closed curve is to be traversed once as one goes once, say, clockwise, "around" the right half of the s-plane) and the arcs for closing the curves can be taken to be circular arcs (they may be more than 360 deg) of very great ("infinite") radius as indicated in Fig. 21. In Figs. 21 (a) and (b) the —90-deg ray is the asymptote. In Figs. 21 (c) and (d) the asymptote is the —180-deg ray. Finally, Fig. 21(e) shows the method of closure in the case of a —270-deg ray. In Fig. 21(e) the transfer locus is asymptotic to the —270-deg ray in the third quadrant, and it takes $1^1/_2$ revolutions in the clockwise sense to close the curve. The curve formed by the transfer locus, its conjugate, and closure arc will be called the "closed transfer locus."

In Fig. 20 the closure is shown by the dashed arc in the right half plane. The closed transfer locus can be traversed in a counterclockwise direction as shown by the arrows. As we traverse this locus we do not encircle point (1, —180-deg).

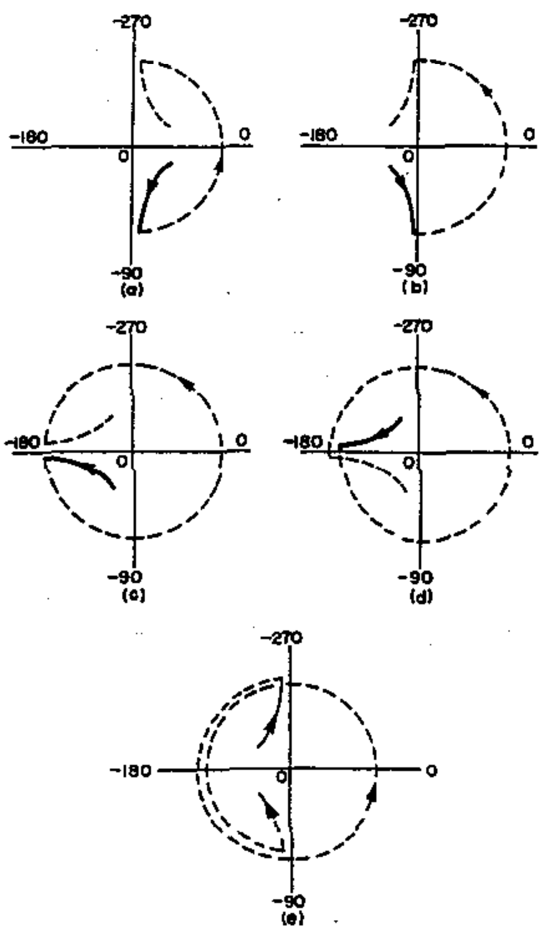

Fig. 21 Methods of Closing a Curve Formed by Transfer Locus and Its Conjugate

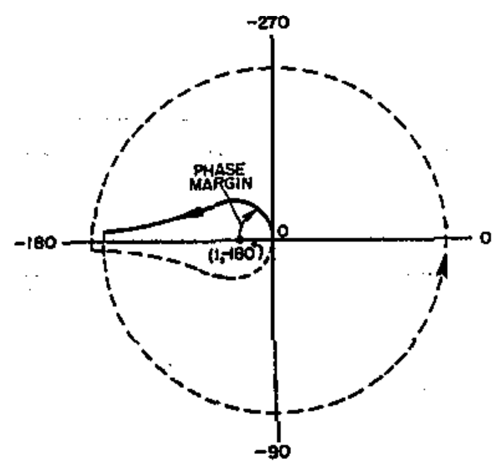

Fig. 22 Closed Transfer Locus of Unstable System

For the transfer locus of Fig. 22 plotted as a solid line the curve is closed as shown. When we traverse the curve once as indicated by the arrows we encircle the point (1, —180-deg) twice. The closed transfer locus for the example of Fig. 12(d) is shown in Fig. 23. The *net* encirclement of the point (1, —180-deg) for this case is zero.

To get the *net encirclement* draw a vector from the point (1, —180-deg) to a point P on the transfer locus. Let P traverse the entire closed curve. The vector will rotate through an angle ($360n$) deg, where n is the number of times the locus encircles the point (1, —180-deg).

It is assumed that the closed-loop linear systems under discussion are stable if opened. For such systems Nyquist's famous stability criterion follows:

Nyquist Stability Criterion. Let S be a closed-loop system that is stable when the loop is opened. The system S is stable if and only if

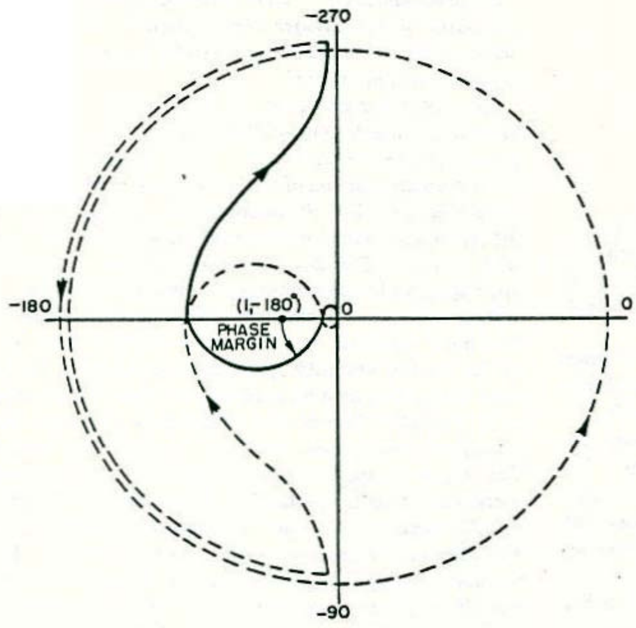

Fig. 23 Closed Transfer Locus for Case of Fig. 12(d)—a Stable System

in one traversal of the closed transfer locus the net encirclement of the point (1, —180-deg) is zero.

The cases of Figs. 20 and 23 are stable by this criterion whereas Fig. 22 is the locus of an unstable system. The system S is unstable if the locus passes through the point (1, —180-deg).

The Nyquist stability criterion can be stated in a more general form to cover systems with unstable open loops.

Design With Transfer Locus Plots

For the practicing engineer, Nyquist's criterion can be replaced by the following:

Practical Stability Criterion. A system is stable if it has a positive phase margin. Otherwise it is unstable.

This criterion has been proved to hold for a system S that is "absolutely stable" (3), that is, a system that is stable up to some value of gain and unstable for larger values, or that is stable for all gains.

Multiplying the gain of the open loop by a factor k corresponds merely to multiplying each radius vector by the factor k. Thus raising the gain by the factor 10 in the governor-engine example changes the transfer locus of Fig. 20 to that of Fig. 24. The locus now loops around the point (1, —180-deg), and the closed system is unstable. The dashed portion of the curve in Fig. 24 is a distorted section to show how the curve is asymptotic to the vertical axis. The phase margin is seen to be negative as shown.

M and N Loci

With the aid of auxiliary curves the frequency response of the closed loop of Fig. 10 can be obtained directly from the transfer locus of the open loop.

Let M be the magnitude ratio of the closed loop (see Fig. 10 with r as the input and c as the output). It can be shown that the curves of constant M are circles on the plane of the transfer locus, as shown in Fig. 25. Here the transfer locus of Fig. 20 is drawn on the same graph. When $f = 0.083$ the locus crosses the curve $M = 1.5$. Hence the magnitude ratio of output c to input r at $f = 0.083$ is 1.5 for the closed loop. Note that the "maximum value" M_p of M for the points on the transfer locus is 2.23. It can be shown that if the maximum value M_p of M

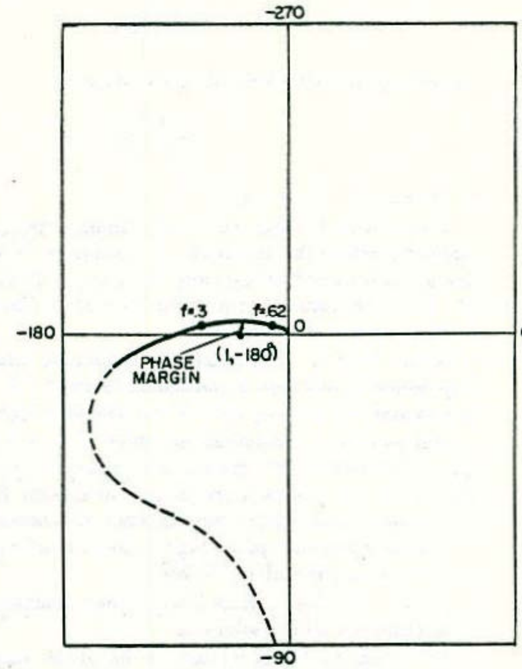

Fig. 24 Transfer Locus of Unstable System With High Gain

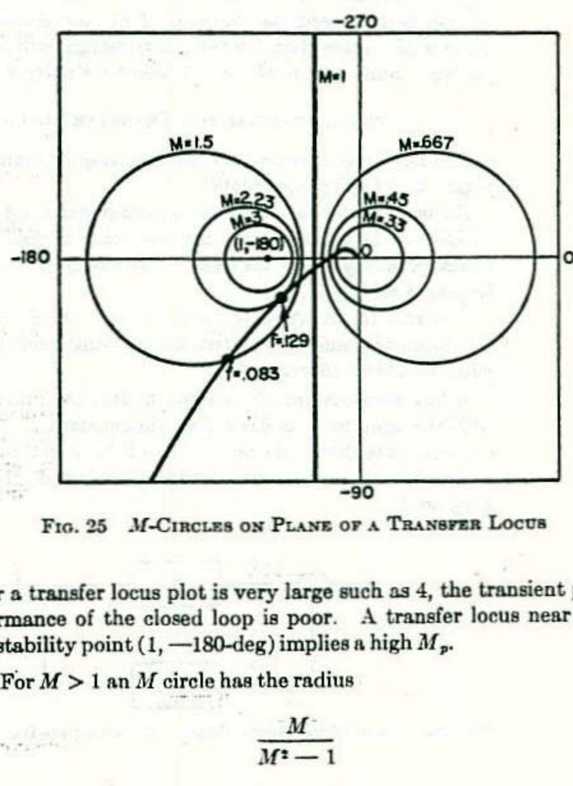

Fig. 25 *M*-Circles on Plane of a Transfer Locus

for a transfer locus plot is very large such as 4, the transient performance of the closed loop is poor. A transfer locus near the instability point (1, —180-deg) implies a high M_p.

For $M > 1$ an M circle has the radius

$$\frac{M}{M^2 - 1}$$

and center on the horizontal axis a distance

$$\frac{M^2}{M^2 - 1}$$

to the left of the origin. If $M = 1$ the M-circle reduces to a straight line $1/2$ unit to the left of the vertical axis. If $M < 1$ the M-circle has the radius

$$\frac{M}{1 - M^2}$$

and center on the horizontal axis a distance

$$\frac{M^2}{1 - M^2}$$

to the right of the origin.

The following design rule is in common use (12). If the frequency at which the maximum M is attained is positive, this frequency is termed the resonant frequency. If at zero frequency $M = 1$, the value of the frequency at which M first falls to 0.707 is termed the *bandwidth.*

Design Rule 5. The maximum magnitude ratio M_p of a closed loop should be less than 2 and should be about 1.3. The bandwidth or resonant frequency (if there is one) should be high as possible.

Mathematical considerations show that a low resonant frequency or bandwidth means slow transient performance. The bandwidth (or resonant frequency) in Design Rule 5 is limited by "noise" suppression requirements and other considerations.

Curves of constant phase lag for closed-loop frequency response are circles, and are called "*N loci.*"

NOTE: By Design Rule 5 the dashed magnitude curve of Fig. 11 is preferred to the solid one.

The quantity M_p is related to the phase margin (see Design Rule 1). Suppose that gain crossover occurs at a point where the phase margin is 30-deg. The value of M at this point is 1.93 (approximately), whence M_p is at least 1.93. As the phase margin is decreased the value of M at gain crossover increases. Thus if M_p is less than 1.9 the phase margin will be greater than 30 deg. Similarly, if $M_p = 1.3$ the phase margin is over 45 deg.

Recommendations for Transfer Locus Plots

The DS-Committee makes the following recommendation with regard to transfer locus plots:

Recommendation 5. It is recommended that a sufficient number of frequencies be shown on the transfer locus to indicate the rate at which frequency affects the locus. The use of M and N loci should be kept to minimum.

The recommendation to keep the use of M and N-loci to a minimum is made to prevent the drawing from being crowded with too many curves.

It has been customary to write the transfer function associated with the open loop as KG for a gain constant K. It is often inconvenient to determine and remove K from the transfer function, especially in experimental studies. Instead, the following is suggested:

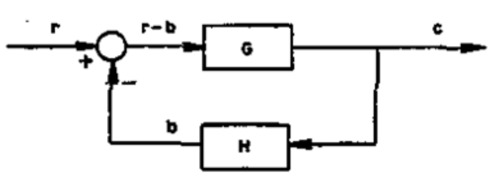

FIG. 26 LOOP WITH BOTH FORWARD AND FEED-BACK COMPONENTS

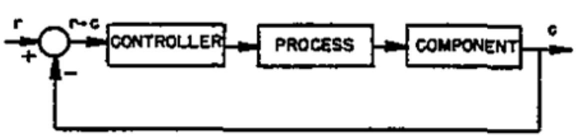

FIG. 27 SYSTEM OF FIG. 19 WITH FEED-BACK COMPONENT PLACED IN FORWARD PORTION

Recommendation 6. Let G denote the open-loop transfer function. The plane of the transfer locus should be labeled as the G-plane (see Fig. 20) rather than as the KG-plane for a gain constant K and frequency variant factor G.

The preceding discussion of transfer locus plots has been restricted to closed-loop systems without frequency variant components in the feed-back path. Let G denote the transfer function of the forward portion of the loop, namely, of the controller-process in Fig. 19. Let H designate the corresponding function for the feed-back component so that GH is the transfer function of the open loop (see Fig. 26). As noted before, the magnitude curve *(on logarithmic co-ordinates) of the open loop is obtained* by adding ordinates of the magnitude curves of the forward and feed-back portions, and the same thing is true of the phase. As far as the stability question is concerned, the reader may treat the system as if all components were in the forward portion, as in Fig. 27. The M and N-loci now no longer designate the magnitude and phase response of the closed loop of Fig. 19. The designer can, however, design the elements involved in the problem so that good stability is secured for the arrangement of Fig. 27, using the criterion for maximum M given in the foregoing. The system of Fig. 19 will now be equally stable. However, the frequency response of the two loops will not be identical. This may be a factor, especially where it is necessary to have the output c follow the input r very precisely, as in aiming an anti-aircraft gun.

Thus far we have focused our attention on the response c to input settings r of the controller. This is of prime importance in the design of follow-up servos. In the usual regulator applications, however, a variable is kept at a constant or reasonably constant value and the system is subject to disturbances not arising from the controller setting. These are often load disturbances, and will be referred to as such. The transfer function relating the load L and the output c will, in general, be different from that relating r and c. However, the stability problem is the same. Thus, instead of treating the response of the closed loop to sinusoidal oscillations of the load, we consider the response to sinusoidal oscillations of the controller setting. There are, of course, situations where the designer will wish to study the response to load oscillations directly.

Inverse Nyquist Plots

The transfer locus plot is the plot of the open-loop transfer function G. On the other hand, the inverse Nyquist plot (13) is the polar plot of $1/G$, and is obtained as follows: The magnitude of the radius vector on the inverse Nyquist plot is the magnitude (amplitude) of the input to the open loop divided by the magnitude of the output (the reciprocal is used for the transfer locus plot) and the vectorial angle is the negative of the phase angle used for the transfer locus plot.

For the case of unity feedback shown in Fig. 10, the inverse Nyquist plot for the closed loop (relating r and c) is obtained by a *simple displacement of the points of this plot for the open loop.* The inverse plot makes it easy to modify the plot for the incorporation of feed-back components. These considerations and its emphasis on the low-frequency region often make the inverse plot a convenient one to work with. The inverse Nyquist plot for the governor-engine example of Fig. 20 is shown as the solid curve in Fig. 28. The reader will note that the locus is above the (1, 180-deg) point instead of below it as for the direct Nyquist plot. The phase margin is indicated in the figure. To insure stability it is required that this margin be positive, and therefore that the curve, assuming a phase margin exists, must pass *above* the (1, 180-deg) point as shown. The inverse plot for the closed loop is given by the dashed line in Fig. 28, obtained by displacing each point one unit to the right. For the case of unity feedback

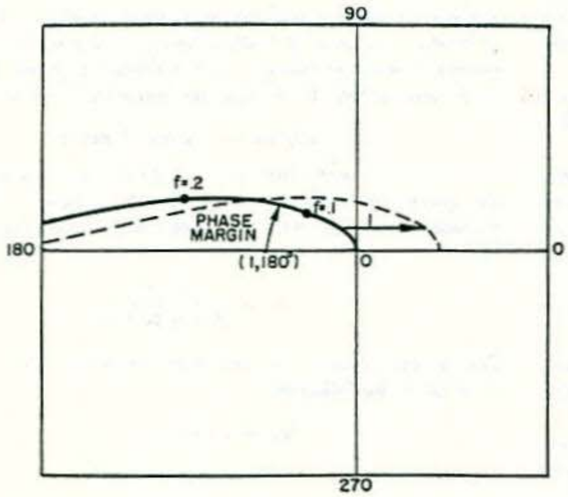

FIG. 28 INVERSE NYQUIST PLOT FOR GOVERNOR-ENGINE CASE

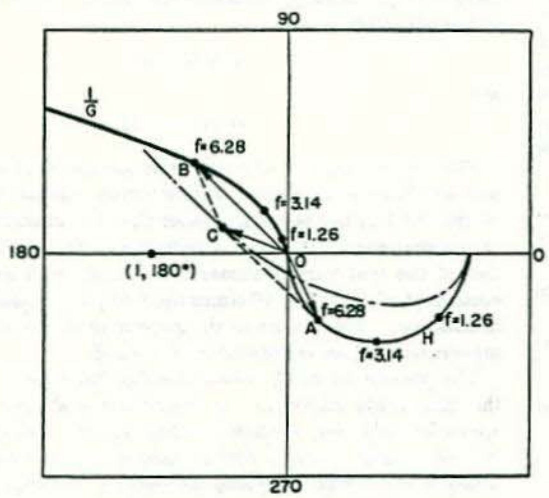

FIG. 29 INVERSE NYQUIST PLOTS

this displacement thus yields the closed-loop plot from the open loop.

In Fig. 29 are shown inverse Nyquist plots for a physical system with G and H as forward and feed-back transfer functions (see Fig. 26). The heavier curve marked $1/G$ is the inverse Nyquist plot of the forward part of the loop, and the lighter one marked H is the transfer locus for the feed-back portion. The points marked A and B are points for the same frequency f on the H and $1/G$ curves, respectively, where $f = 6.28$. Adding the vectors OB and OA, we obtain the point C on the inverse Nyquist plot of the closed loop by drawing the diagonal OC of the parallelogram with sides OA and OB. The closed-loop inverse locus is the curve with long and short dashes in the figure.

The M loci are concentric circles on the inverse plot. The distance from the origin to a point on the inverse Nyquist plot for the closed loop is simply $1/M$ for the magnitude ratio M defined as for the direct Nyquist plot (output magnitude over input). The inverse transfer function for the closed loop is simply

$$\frac{1}{G} + H$$

Recommendation 7. Let G and H be the respective transfer functions of the forward and feed-back portions of a closed loop. On the inverse Nyquist plot the loci of the forward and feed-back parts of the loop should be labeled 1/G and H, respectively (see Fig. 29). The use of M-circles should be minimized.

Log Magnitude-Phase Plots

The log magnitude-phase (9) plot is obtained by graphing magnitude ratio as the vertical co-ordinate on a logarithmic scale, and phase angle on a horizontal linear scale. The log magnitude-phase plot has the following advantages:

(*a*) The basic frequency-response data, presented with the aid of logarithmic co-ordinates as recommended in this paper, may be transferred directly to the log magnitude-phase plot with dividers.

(*b*) The products of several component frequency responses may be combined by graphical addition and thus the effects of components on over-all performance may be easily demonstrated.

(*c*) The log magnitude-phase plot has the advantage that the gain of the open loop can be changed by merely sliding the log magnitude-phase plot up or down. Changing the gain does not affect the phase angle but moves the plot up or down as described for the magnitude curve on logarithmic co-ordinates. Normally, a co-ordinate system with M and N-curves on a template is slid up and down over the log magnitude-phase plot to change the gain of the loop.

The log magnitude-phase plot has the distinct disadvantage that M contours are not circles: hence special construction, or templates, are necessary for drawing them. For the governor-engine example of Fig. 11 the log magnitude-phase plot is shown in Fig. 30.

Recommendation 8. It is recommended that the log magnitude-phase plot be graphed with the magnitude ratio as the vertical co-ordinate on a logarithmic scale and phase angle in degrees on a linear horizontal scale.

Transfer Functions

Suppose that a physical system S is at rest before the instant

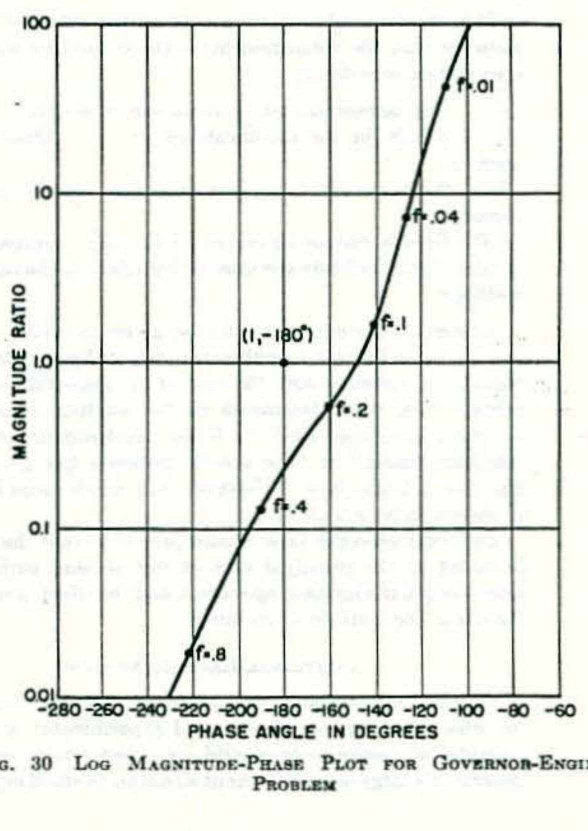

FIG. 30 LOG MAGNITUDE-PHASE PLOT FOR GOVERNOR-ENGINE PROBLEM

$t = 0$. Let the input m to S be one which vanishes identically for $t < 0$. Let c denote the output of S, and assume that the Laplace transforms of m and c exist.

The *transfer function* of S is by definition the Laplace transform of the output divided by that of the input, provided that this quotient is unique.

Recommendation 9. The independent variable in a transfer function should be denoted by s. Thus G(s) is a transfer function of the complex variable s. In actual computation jω is to be substituted for s.

Recommendations Regarding Presentation of Measurement Data

It is recommended that from the point of view of the presentation of measurements of frequency responses, the least common denominator of technical audiences should govern.

Since attention should be focused on the diagram of the system studied, it is not desirable to load up this drawing. If the measurement equipment is simple, this equipment may be indicated on the diagram with symbols different from those used to describe the system components, or by the use of dotted lines, indicating at most the real or equivalent circuit of the measurement system. If the measurement equipment is complicated, a simple diagram may be repeated, with the measurement system put in by dotted lines. The characteristics and type of measurement systems used may be tabulated either in the text or in the diagram for the measurement system. Most extensively, *these data could include:*

(*a*) *The input impedance* in mechanical or electrical terms of the measurement system relative to the impedance of the primary system across the points of measurement.

(*b*) *The input-output sensitivity* of the measurement system as a whole or by components.

(*c*) *The equivalent circuit of the measurement system*, its *frequency response*, or its *time constant* if this exists.

The DS-Committee has been concerned with the problem of assuring that the measurement methods used in obtaining frequency-response data:

(*a*) Can be reproduced to obtain the same data.

(*b*) Result in no modifications of the "true" frequency response.

(*c*) If they modify the true response, that it may be corrected for.

(*d*) Or if it cannot be corrected for, that suitable discussion be included to indicate the qualitative effect of the measurement methods.

As much information should be given as would permit the casual reader to have a qualitative idea of the significance of the measuring systems, and the expert a quantitative idea. To achieve this, such statements as "a rotating strut-type wire fatigue testing machine," "a linear accelerometer with a wide frequency band," or "the system response has no appreciable lag" are not adequate. However, not much more information is needed to be sufficient.

The common-sense view should prevail so that the description is suited to the principal uses of the system, with particular references to the limits of operation, and the effect of uncontrolled factors in the ambient environment.

Additional Recommendations

The frequency-response data should include the range of inputs for which the data are linear, and experimental or theoretical *estimates of nonlinearity* should be given where possible. In general, if a large sinusoidal input signal and a small input sinusoid produce the same magnitude and phase responses, the process can be treated as linear, at least at the control point about which the measurement was taken. If the linearity is not known, a detailed description of the experimental setup should be given.

Failure of Design Criteria

Commonly used criteria for stability may be satisfied while the system is unstable for practical purposes. To illustrate, consider the system with the open-loop transfer function G given by

$$G = \frac{0.486}{s^2 + 0.3s + 1} \qquad [1]$$

The system is taken to have unity feedback. For the function G we have the following:

$$M_p = 1.34 \text{ at } \omega_r = 1.2$$

$$\text{Phase margin: } 45°$$

$$\text{Gain margin: } \infty$$

Here ω_r is 2π times the resonant frequency. However, the roots of the equation

$$1 + G = 0 \qquad [2]$$

are

$$-0.15 \pm 1.21\,j \qquad [3]$$

The reader familiar with the root method of treating stability will see that the transients will be highly oscillatory because the coefficient 1.21 of j is much greater than the numerical value 0.15 of the real part. It is the experience at the author's company that if the real part is numerically small, such as 0.1, and the coefficient of j is about 10 times the real part or more, the system is unstable. This occurs in the governing of the speeds of prime movers where 1 sec is the unit of time used.

The present example violates Design Rule 2 since the slope of the magnitude curve on log co-ordinates at gain crossover is approximately —4, as shown in Fig. 12 (*a*). However, *this rule was made only to insure that the phase margin would be adequate*, which it is. Thus *satisfying the stability criteria will not insure adequate stability or adequate transients*, but may help to achieve this.

The Role of Frequency-Response Approach

The frequency-response approach has attained its present position in the automatic control field partly because many engineers prefer working with curves rather than algebraic formulas, and there is a general lack of familiarity with simple methods (14) for solving algebraic equations of high degree. Where possible it is desirable to derive the actual transients, or at least obtain the characteristic roots. An analyst familiar with the theory of such roots can tell from them what the curves for the transients will look like. Where an analog computer is available the problem of obtaining the transients is an easy one. Considerable use is now being made of the Evans root-locus method (15) of solving an algebraic equation. Like the frequency-response design methods it has the disadvantage that it involves the graphing of curves. All of the curve-drawing approaches to the design of automatic controls, including the analog computer, suffer from the fact that a curve represents only a numerical case. Correct and rapid design can often, if not generally, be accomplished only by manipulating the mathematical formulas with general symbols for some of the constants. To do this a considerable amount of mathematical training is indicated. Actually, both numerical and abstract studies are needed in the analysis of automatic controls.

The great power of the frequency-response approach for the control engineer or scientist, with a considerable mathematical background, lies in the application of frequency-response methods to areas, such as nonlinear phenomena, where the usual algebraic methods fail (16). Hysteresis, saturation, and variable damping are examples. The use of frequency-response curves for diagnostic purposes rather than design is also invaluable in that factors that should not have been neglected in a problem often show up on these curves.

For the beginner the frequency-response approach is a boon. This is especially true because of the correlation that exists between the transient response and the frequency response of a physical system (7).

Appendix

To clarify the use of transfer functions in making design rules we shall consider a system with the open-loop transfer function

$$\frac{k(s + a)}{s^2(s + 1)} \qquad [4]$$

where a and k are positive real constants, and $a < 1$. The asymptotic magnitude curve in this case is composed of three falling straight-line sections with slopes —2, —1, and —2, respectively, Fig. 14. The magnitude ratio R for this function is given by

$$R = \frac{k\sqrt{a^2 + \omega^2}}{\omega^2\sqrt{1 + \omega^2}} \qquad [5]$$

The ratio p of the value of R at $\omega = a$ to the value of R at $\omega = 1$ is given by

$$p = \frac{2}{a(1 + a^2)} \qquad [6]$$

The characteristic equation for the closed loop is

$$x^3 + x^2 + kx + ak = 0 \qquad [7]$$

Routh's stability criterion (6) implies that $1 > a$. For a value of a "near" 1, such as 0.5, the roots of the foregoing equation are much too oscillatory. For $a = 0.5$ we have $p = 3.2$. For $a = 0.2$ the roots are satisfactory and $p = 9.6$. For these and other reasons we have recommended in Design Rule 3 that p be chosen equal to or greater than 5.

BIBLIOGRAPHY

1 "Communication Networks," by Ernst A. Guillemin, John Wiley & Sons, Inc., New York, N. Y., vol. 2, 1935, p. 475.

2 "Regeneration Theory," by Harry Nyquist, *Bell System Technical Journal*, vol. 11, 1932, pp. 126–147.

3 "Network Analysis and Feedback Amplifier Design," by Hendrik Bode, D. Van Nostrand Company, Inc., New York, N. Y., 1945.

4 "Mathematical Engineering Analysis," by Rufus Oldenburger, The Macmillan Company, New York, N. Y., 1950.

5 "Principles of Servomechanisms," by G. S. Brown and D. P. Campbell, John Wiley & Sons, Inc., New York, N. Y., 1948.

6 "Servomechanism Analysis," by G. J. Thaler and R. G. Brown, McGraw-Hill Book Company, Inc., New York, N. Y., 1953.

7 "Servomechanisms and Regulating System Design," by Harold Chestnut and R. W. Mayer, John Wiley & Sons, Inc., New York, N. Y., 1951.

8 "Introduction to Servomechanisms," by A. Porter, Methuen & Company, Ltd., London, England, 1950.

9 "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

10 "The Dynamics of Automatic Controls," by R. C. Oldenbourg and H. Sartorius, The American Society of Mechanical Engineers, New York, N. Y., 1948.

11 "Transient in Linear Systems," by M. F. Gardner and J. L. Barnes, John Wiley & Sons, Inc., New York, N. Y., vol. 1, 1942.

12 "Application of Circuit Theory to the Design of Servomechanisms," by A. C. Hall, *Journal of The Franklin Institute*, vol. 242, October, 1946, p. 287.

13 "Parallel Circuits in Servomechanisms," by H. T. Marcy, Trans. AIEE, vol. 65, 1946, pp. 521–529.

14 "Practical Computational Methods in the Solution of Equations," by Rufus Oldenburger, *American Mathematical Monthly*, vol. 45, June-July, 1948, pp. 335–342.

15 "Control Synthesis by Root Locus Method," by W. R. Evans, Trans. AIEE, vol. 69, part 1, 1950, pp. 1–4.

16 "Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements," by E. Calvin Johnson, Trans. AIEE, vol. 71, part 2, July, 1952, pp. 169–181.