1956 NATIONAL CONVENTION

IRE Convention Record

Part 2

Circuit Theory

SESSIONS ON

Circuits I—Symposium on Application of Recent Network Ideas to Feedback System Problems

Circuits II—Design and Application of Active Networks

Circuits III—Network Synthesis Techniques

SPONSORED BY

IRE PROFESSIONAL GROUP ON

Circuit Theory

Presented at the IRE National Convention, New York, N. Y., March 19-22, 1956

The Institute of Radio Engineers

IRE CONVENTION RECORD

1956 NATIONAL CONVENTION

PART 2 - CIRCUIT THEORY

TABLE OF CONTENTS

Network Theory in the Practical Design of Control Systems

J. G. Truxal

Microwave Research Institute
Polytechnic Institute of Brooklyn
Brooklyn, New York

Network theory in large measure provides the theoretical foundations for the design of feedback control systems. Network theory stands as the language of communication between the applied mathematician and theoretically-minded engineer, on the one hand, and, on the other hand, the practicing engineer who must make a feedback system work, whether it be a control system, a feedback amplifier, or another application. In particular, three facets of network theory have proved of very considerable importance to the control system engineer: realization techniques, feedback theory, and measurement techniques. In the following paragraphs we indicate typical applications of each of these facets, then describe briefly a few of the problems which are of interest to the control engineer, but which are still not satisfactorily solved.

Thus, we are discussing here the correlation of network theory and feedback control system design. There is some danger in any attempt to recite the importance of network theory in feedback control, since there is considerable question in the case of many of the applications as to whether the primary development arose from network theory or from automatic control. This is an insignificant question, however; we shall not even attempt a definition of network theory, but rather assume that network theory includes the study of the properties of networks and techniques for the synthesis of networks with specified characteristics.

REALIZATION TECHNIQUES

To most engineers, network theory means first of all the body of realization techniques - methods for the determination of a network with a specified transfer function, driving-point impedance, etc. The network may contain all three passive elements R, L, and C and, possibly, also active elements such as tubes or transistors.

The first interest of the control system designer is ordinarily in the synthesis of networks to yield a prescribed transfer function. The classical (though not always practical) problem of control system design is illustrated in Fig. 1(a); the components represented by the transfer function $G_2(s)$ are assumed given and the designer must select a $G_1(s)$ to yield satisfactory characteristics for the overall system. If the root-locus method of design is used, the poles and zeros of $G_1(s)$ must be selected.

The loci are first plotted for $G_1 = K$, as shown for a possible G_2 in Fig. 1(b). These loci represent the motion of the poles of T as the gain K is increased from zero. The loci start from the poles of G_2, shown by crosses, and move to the zeros of G_2, here all at infinity. When $K = K_1$ the closed-loop poles are moving into the right half plane and the system is becoming unstable. The problem of control system design is essentially the problem of selecting a set of poles and zeros for $G_1(s)$ such that the loci are modified and move to the left from the complex poles of G_1 instead of to the right. The set of poles and zeros for the controller transfer function $G_1(s)$ is not unique, but suitable sets can be determined in a straightforward procedure.

Once the transfer function $G_1(s)$ is determined, system design devolves to realization of $G_1(s)$ by a physically realizable network. If $G_1(s)$ is a simple transfer function with perhaps no more than two poles and two zeros, the realization techniques of network theory are hardly necessary; rather, we simply call on our experience, intuition, or an appropriate handbook to evoke an appropriate network. Even if G_1 is of greater complexity, isolating amplifiers using vacuum tubes allow us to break G_1 into simple factors, each to be realized by a single stage. Even with transistors the absence of isolation is not a particularly severe handicap at the low frequencies we are considering here.

Unfortunately, actual design of control systems is seldom quite so idealistic. Commonly a complete feedback control system involves a multiplicity of loops, with various transducers measuring relevant signals and the transducer outputs combined in a controller or analog computer. The interconnection of the various signals involves the realization of specified transfer with specified loading and gain levels. In addition, noise considerations commonly limit the form of realization utilized.

There are other applications of realization techniques which should be mentioned. The entire control system may be designed in terms of realization techniques. For example, the control system can be made stable under any passive load if the output admittance is made a positive real function of s. In this situation, we have the realization problem of designing a two-port active network for a specified transfer function with one driving-point function required to be positive real.

Alternatively, the entire feedback system can be designed with the philosophy of the network synthesist, as the overall transfer function, T(s) in Fig. 1, is selected at the outset of the design to meet system specifications, the open-loop transfer function (G_1G_2) is determined analytically or graphically, and the required G_1 is then realized. Such a direct approach to control system design is particularly important in that it demonstrates that the designer can be the master of system performance; in practice it is perhaps most useful in very simple cases or in very complex situations where conventional techniques break down (e.g., in the design of sampled-data systems where the flexibility of a digital controller is available or in complex, multi-loop systems).

Finally, in the category of realization techniques we should include the use of such methods of network theory in analog-computer work. The intelligent utilization of analog computers as aids in the design of feedback systems, particularly nonlinear feedback systems, requires that the simulation of actual components of the system be effected on a transfer-function basis, rather than on the basis of the integro-differential equations. The transfer function of the basic operational-amplifier circuit of Fig. 2 is simply $-Z_2/Z_1$; hence simulation of a specified transfer function involves realization of appropriate Z_2 and Z_1, usually by RC networks. Even more elegant realization problems arise in the design of basically passive analog simulators such as the simulators developed at California Institute of Technology for aircraft flutter studies or in the simulation of specified nonlinear characteristics.

FEEDBACK THEORY

Largely because of the nature of the great number of books in the servo field, feedback theory is nearly synonymous with control-system design theory. The very heavy emphasis on stability theory is embodied in the root-locus plotting, the Routh test, the Nyquist diagram, the asymptotic gain and phase plots, and the Nichols charts. The control engineers emphasis on frequency-response analysis and design has been based directly on network theory.

Yet even here the carry-over from the theoretician to the practicing designer has not always been complete. The designer is quite at home with the asymptotic gain plots, for example, and commonly plots gain in terms of the straight line asymptotes, as in Fig. 3. Yet the similar straight-line approximation for the phase characteristic is apparently not well known.

In the presence of the great emphasis on stability, the control system engineer all too often loses sight of the basic objective of the control system. Stability is really only an annoying by-product; the fundamental motivation for the use of feedback in the first place is the desire to realize a high return difference or loop gain around a portion of the system. If this high loop gain is realized around the motor, for example, the system characteristics are insensitive to changes in motor parameters. Thus, in order to evaluate his system, the control engineer must turn to the basic principles of the theory of active or feedback networks.

Nearly twenty years ago Bode demonstrated that, unless conditionally stable systems are used, there is a definite relation between the loop gain and the excess bandwidth required in a feedback system. In other words, if the loop gain is to be maintained high over a specified bandwidth, beyond this bandwidth it cannot be cut off at an average rate greater than about 33db/decade if satisfactory relative stability is to be realized. In the single-loop system of Fig. 1, the required loop gain is established by the relation

$$S_{G_2}^{T} = \frac{dT/T}{dG_2/G_2} = \frac{1}{1 + G_1 G_2} \qquad (1)$$

In other words, the sensitivity of T with respect to G_2 (the percentage change in G_2 divided into the resulting percentage change in T) is inversely proportional to the loop gain $(-G_1G_2)$ if the latter is much greater than unity. Thus, a sensitivity of 0.001 requires a loop gain of 10^3 or 60db. The magnitude of the loop gain measures the self-calibration properties of the system. The above equation imposes a definite correlation between the sensitivity and the magnitude of the stability problem: the lower the sensitivity, the higher the loop gain, and the worse the stability problem.

One method, largely unexploited, for the feedback designer to circumvent this fundamental limitation is to consider a different measure of sensitivity, a measure more appropriate for the particular problem at hand. For example, the above definition measures the sensitivity of the complex transmission T; in certain feedback systems only the gain $|T|$ is of interest. In such cases, we should consider as a figure of merit for our system

$$R = \frac{d|T| / |T|}{d G_2 / G_2} = \text{Re } S \tag{2}$$

Figure 4 shows two systems for realizing identical overall transfer functions with the same motor and load. The sensitivity S is very nearly the same in the two cases, but, as indicated, the sensitivity of the gain is far better in the case of the minor-loop compensation. A more careful comparison of the two systems demonstrates that we are effectively trading gain sensitivity for phase sensitivity; in other words, although the sensitivity is fixed by the loop gain, the manner in which S is divided between real and imaginary parts (gain and phase sensitivities) depends on the system configuration.

Actually what we commonly want in feedback system design is control over the sensitivity of the transient response - e.g., the sensitivity of the overshoot to changes in a system parameter. We would like to choose a configuration such that, when a parameter changes, the poles and zeros move in such a fashion that the essential characteristics of the transient response remain unaltered. Figure 5 shows three pole-zero configurations which can have nearly identical transient responses. In (a) the response is established almost entirely by the location of the pair of complex poles. As these poles move toward the negative real axis in (b), the zero is moved inward to maintain overshoot approximately constant. In (c) the real-axis pole counteracts the tendency for greater overshoot due to the motion of the complex poles toward the imaginary axis. Thus, if transient-response sensitivity is important, we might achieve a satisfactory design with a system in which variation in a specified parameter caused the transitions among the three configurations of Fig. 5.

The theory of feedback networks is, accordingly, important for the techniques of stability analysis and also as a basis for the comparative evaluation of alternate designs, and block-diagram configurations.

MEASUREMENT TECHNIQUES

The third aspect of network theory of significance to the control engineer is the field of measurement techniques. In particular, network theory has demonstrated the correlations between, on the one hand, the frequency characteristics (gain and phase) so useful in feedback-system design and, on the otherhand, the transient response and the response to random input signals.

Pulse, step-function, and square-wave testing of control-system components has become commonplace since the refinement of techniques for determining frequency characteristics from transient response and the availability of computing equipment for effecting the conversion. The use of low-level, random signals for determining system response by crosscorrelating input and output or by measuring the cross power spectrum provides an alternate approach for measuring system dynamics, although the method also raises certain questions: e.g., how much data is required for a specified accuracy; is it more satisfactory to crosscorrelate and then take the Fourier transform or to determine the cross power spectrum directly?

Thus, network theory has indicated to the control engineer a variety of techniques for the description of his components in mathematical terms useful in system design. A comparatively unexploited field is the characterization and description of typical input signals, but as time-varying and nonlinear systems are used more and more to improve system performance we can anticipate an increasingly strong interest in signal theory. Certainly it is only through a more precise and appropriate description of our input signals that we will be able intelligently to design systems specifically for these signals. The extensive work on the design of nonlinear systems for optimum response to step functions provides one example of the fruitfulness of such an approach.

PRESENT PROBLEMS

Although in a meeting of this type it is customary to emphasize that the feedback art is several millenniums old, we might rather point out that the glamorous age of feedback really was born only about a decade ago. In that short time, enormous strides have been made in the theory pertinent to feedback-system design. In large measure network theory has been responsible for or associated with these strides. The control engineer has built his field on a foundation of the mathematics of network theory and the physics of control elements as they are drawn from mechanical, electrical, aeronautical, chemical engineering, etc.

In each of the three categories we discussed above, however, there are certain problems which at the present time the average control engineer is unable to handle conveniently. Perhaps some of these problems have been solved, but we shall list them briefly here. We consider only linear, time-invariant systems, even though the time-variable or nonlinear systems pose even more important problems.

Realization Techniques

(1) The problem of correlating transient and frequency responses is still troublesome, in spite of the extensive work in this field. A common problem in control system design, for example, involves meeting certain specifications on the step function response (e.g., time delay, overshoot, and settling time) and simultaneously realizing adequate noise filtering. Simple, approximate relations between time delay, for example, and frequency characteristics all too often break down in practical problems. Specifically, we commonly find time delay T_d given as $-d\beta/dw$, where β is the phase characteristic. If the factor $(s^2 + a^2)$ is added in the transfer function, β is unchanged for $w < a$, yet the time delay for step-function inputs may be modified radically in a practical system. Likewise the relation that T_d is $1/K_v$, where K_v is the velocity constant is an inadequate approximation for many systems in which the overshoot is not negligible.

(2) Specific realization problems commonly arise for which there seems to be no simple solution. For example, we frequently would like an RC network with all capacitors a specified size or at least limited in size. Conventionally, we synthesize a network, then try to manipulate the matrix to a satisfactory form. Such matrix manipulation can be tedious and frustrating, particularly since we are not sure a solution exists.

(3) The realization of active networks including transistors as the active elements is still largely a matter of working with familiar configurations. The basic problem of active network synthesis - the sensitivity problem - is still largely unsolved, as, if we wish to design with a reasonable number of elements, we generally can maintain little if any control over the sensitivity in the realization procedure.

Feedback Theory

(4) Other than the problems mentioned earlier, the most important problem in feedback theory is the design of multi-loop systems. In a complex feedback system, we may need three Nyquist diagrams (or three root-locus plots) to determine system performance. Design should guide the choice of the innermost-loop transfer function in such a way as to yield desirable characteristics when we reach the outermost loop. At the present stage of the art, a trial - and - error approach is necessary.

(5) We still know very little about the characteristics of interlocked feedback systems: a set of systems where the signals in one system are fed into the other systems. Such interlocking occurs commonly in complex overall systems. We can make the various systems independent, but this is perhaps primarily an analysis expedient.

Measurements

(6) There is still an important need for study of the effects of errors in measurement, of finite amounts of data, and of various techniques for processing the data. For example, if gain and phase characteristics are measured only over the useful band of frequencies, how can these data be extrapolated to include the larger bandwidth of importance because of stability considerations? or, what is the effect of small errors in transient response measurements?

CONCLUDING COMMENT

To attempt to justify the importance of network theory in feedback-control-system design seems unnecessary, since the design theory is in large measure a direct application of the fundamental concepts of network theory. In the light of the history of feedback control, we can anticipate that the network-theory research and studies of today will provide the foundations for the control-system design techniques of a decade hence.

Acknowledgment

The research included in this paper was supported by the United States Air Force, through the Office of Scientific Research of the Air Research and Development Command.

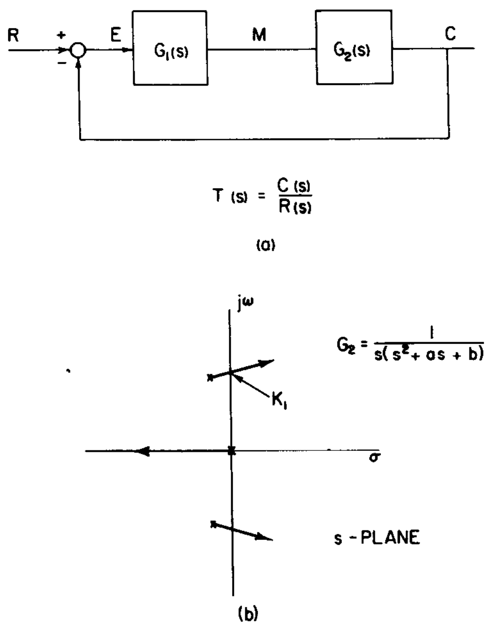

Fig. 1
Basic feedback system

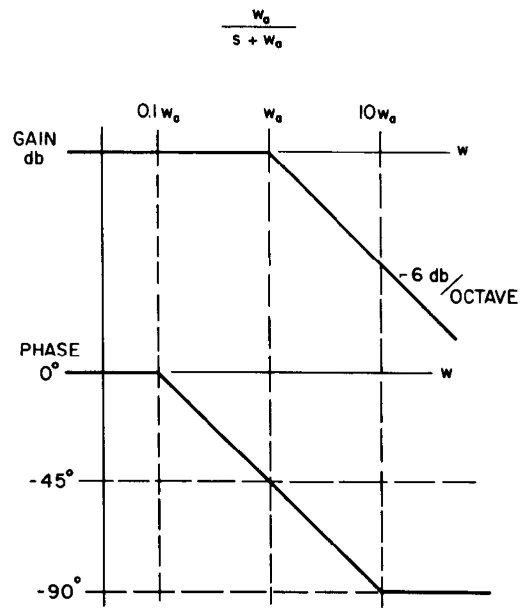

Fig. 3
Gain and phase approximations for linear factor

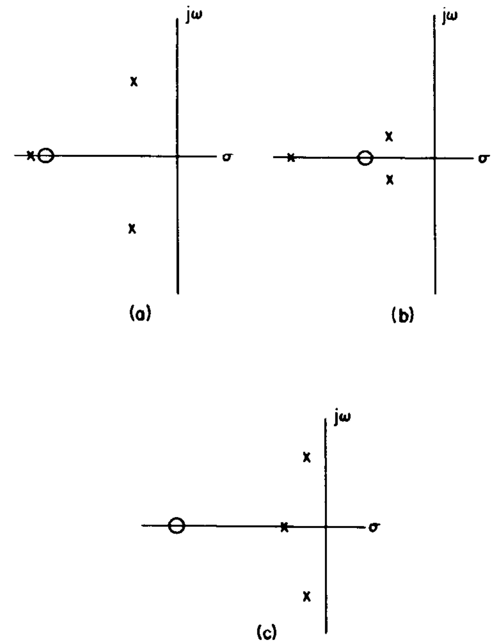

Fig. 5
Three pole configurations with similar transient responses

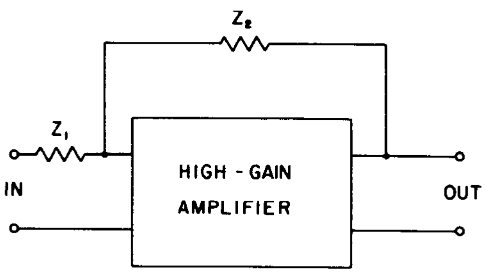

Fig. 2
Basic operational-amplifier circuit

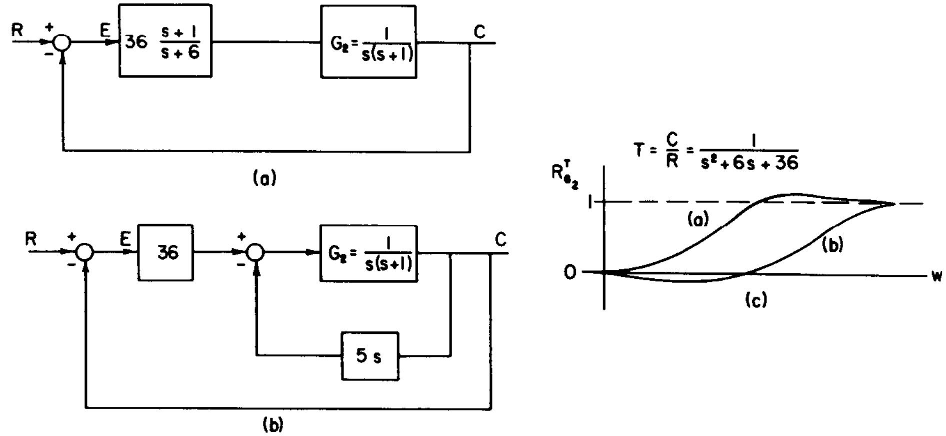

Fig. 4
Two systems with same transmission

SOME THEOREMS APPLICABLE TO THE PROBLEM OF
STABILITY IN LINEAR SYSTEMS

by
John L. Bower, Autonetics Division
North American Aviation, Inc.
Downey, California

Summary

The equivalence of the frequency-domain and time-domain criteria for the stability of linear systems fails in the case of certain oscillatory weighting functions. By excluding such functions on the basis that they can have no physical existence, it becomes possible to prove that the frequency-domain criteria are valid.

Introduction

In the design of feedback control systems, no single aspect has received as extensive treatment as the determination of the degree of stability. Most of the theory of stability of linear systems is based on the criterion that a system is stable if its response to a hypothetical impulse driving function is of such form as to be integrable in absolute value up to $t = \infty$. From this criterion others are derived, applying generally to the Laplace transform of the impulse response, and leading to a determination of whether or not the singularities of the transform lie entirely within the right half-plane.

While the two criteria are equivalent for many classes of impulse response, examples can be found that satisfy the conditions on the transform but not those on the time function. The most troublesome functions of this type are those values alternate with a steadily diminishing period as $t \rightarrow \infty$. Since the "frequency" of such a function increases without limit it clearly violates the restriction of physical realizability imposed on loop gain in most frequency-domain criteria: that the loop gain must vanish at infinite frequency. This restriction, however, is ordinarily injected in such a way as to leave open the question as to whether it is necessary only because we otherwise would not be able to overcome certain purely mathematical obstacles.

In the use of criteria based on the location of singularities of the transform of the impulse response, one encounters awkward limitations on the form of the transform and of the impulse response.[1] In an attempt to avoid these limitations and still retain the validity and convenience of stability criteria based on the location of singularities of the transform, the theorems of this paper were proven. It is hoped that the manner in which the physical realizability restriction is introduced may shed further light on the real need for the restriction in order to validate the frequency-response stability criteria.

We shall find it convenient to define the following:

$w(t)$ is a function of t integrable and bounded over every finite interval of t for $t > 0$, and such that where $(t_{k+1} - t_k) < T$,

$$\frac{1}{t_{k+1}-t_k}\int_{t_k}^{t_{k+1}} dt\,|w(t)| \leq G(t_{k+1}-t_k)^{\alpha}, \qquad (1)$$

$\left(\alpha > 0,\ \frac{\pi}{2} > T > 0,\ \infty > G > 0,\right.$

t_k and t_{k+1} are respectively the kth and (k+1)th points of reversal of sign of $w(t)$ defined by the following for each integral value of k:

$$\left[\int_{t_{k-1}}^{t_k} dt\ w(t)\right] \cdot \left[\int_{t_k}^{t_{k+1}} dt\ w(t)\right] < 0; \qquad (a)$$

and

$$\left|\int_{t_k}^{t_{k+1}} dt\ w(t)\right| = \int_{t_k}^{t_{k+1}} dt\ |w(t)| \qquad \left.(b)\right);$$

$$H(s) \equiv \int_0^{\infty} dt\ e^{-st}\ w(t);\ \text{and} \qquad (2)$$

$$J \equiv \int_0^{\infty} dt\,|w(t)|\,. \qquad (3)$$

Lebesgue integration is assumed throughout. The first theorem is:

(I) A SUFFICIENT CONDITION FOR THE CONVERGENCE OF J TO A FINITE LIMIT IS THAT H(s) CONVERGE AND BE REGULAR ON AND TO THE RIGHT OF THE IMAGINARY AXIS.

(II) IF J CONVERGES, THEN H(s) IS BOUNDED ON THE IMAGINARY AXIS, AND REGULAR IN THE HALF-PLANE $Re(s) > 0$.

Before entering upon the proof, it is pertinent to note that some restriction on $w(t)$ of the nature of (1) is necessary, since functions exist that, in the absence of such restriction, would violate part I of the theorem.

- -

1- See "Theory of Servomechanisms" by James, Nichols, & Phillips, McGraw Hill, 1947.

Such a function is $w(t) = e^{(h-a)t}\sin e^{ht}$, $(0 < a < h)$, whose abscissa of convergence is $-a$, but whose abscissa of absolute convergence is[2] $h-a$. For this function, $H(s)$ is regular in the right half-plane and on the axis[2] although the integral J clearly does not converge. The importance of the restriction (1) will be appreciated more fully after the discussion of application of the theorem to stability problems.

We now turn to the proof of the theorem. We make use of the existence of the derivatives of $H(s)$ of all orders, $n \geq 0$:

$$\frac{d^n H(s)}{ds^n} = \int_0^{\infty} dt\, e^{-st}(-t)^n\, w(t) \,. \qquad (4)$$

By our assumption in (I), the integral in (4) must converge for values of s in the neighborhood of the origin, or,

$$\left| \int_0^{\infty} dt\, (-t)^n\, w(t) \right| = \left| \sum_{k=0}^{K-1} I_k + I_{\infty} \right| < \infty . \qquad (5)$$

In this operation, we have broken up the infinite integral into a series of finite integrals, I_k and I_{∞}, between the K successive points t_k, at which the weighting function, $w(t)$, changes sign, except that $t_0 = 0$. The terms of the series are:

$$I_k = \int_{t_k}^{t_{k+1}} dt(-t)^n w(t) \; ; \; I_{\infty} = \int_{t_K}^{\infty} dt(-t)^n w(t). \qquad (6)$$

Note that I_{∞} does not occur when $K = \infty$.

There are three mutually exclusive cases, corresponding to the possible types of weighting functions. They are as follows:

(A) $w(t)$ is zero identically beyond some finite value of t (regardless of the number of changes of sign of $w(t)$);

(B) for every finite t' there exists an interval of non-zero extent on which $t > t'$ and $w(t) \neq 0$, and either

1 K is finite, or

2 K is infinite ($w(t)$ oscillates in sign to ∞).

In case (A), the integrability of $w(t)$ over every finite interval immediately gives us $J < \infty$. In case (B-1), the integral of the absolute value of $w(t)$ between the limits 0 and t_K must be bounded as in case (A), and we see from (5) with $n = 0$ that

$$\left| I_{\infty} \right| = \left| \int_{t_K}^{\infty} dt\, w(t) \right| = \int_{t_K}^{\infty} dt\, \left| w(t) \right| < \infty \,. \qquad (7)$$

This completes the proof of the first part of the theorem for cases (A) and (B-1).

In case (B-2), the I_k must form a convergent alternating series, but this alone does not suffice to complete the proof. We note that the integral J can be written in series form similar to (5), with the same upper and lower limits of integration for each term. These terms are limited by one of two restrictions, depending upon whether or note the sequence, t_k, tends to diverge slowly or rapidly. For our purposes, "slow divergence" exists when, for positive numbers p and q to be defined, $(t_{k+1} - t_k) < pk^{q-1}$, while "rapid divergence" describes the other cases. We shall choose a suitable value of n for use in (5) and show that the selection of the largest possible terms for the series under the restrictions: (a) of integrability over every finite range; (b) of (1), and (c) of (5) must yield a convergent result.

For the convergent series (5), there must exist an upper bound on the terms, M_n, dependent on n. Then

$$M_n \geq \left| I_k \right| \geq t_k^n \int_{t_k}^{t_{k+1}} dt \left| w(t) \right| \,. \qquad (8)$$

Division of both sides of this equation by t_k^n yields the first of two restrictions on the $|I_k|$:

$$\int_{t_k}^{t_{k+1}} dt \left| w(t) \right| \leq \frac{M_n}{t_k^n} \qquad (9)$$

We now select n such that $\frac{1}{n-1} < \alpha$, and are able, therefore, to take q such that $\frac{1+\alpha}{1+\alpha+n} < q < \frac{\alpha}{1+\alpha}$. Further we take

$$p = \left(\frac{M_n}{G}\right)^{\frac{1}{n+\alpha+1}} \cdot \left(\frac{3z - z^2}{2}\right)^{\frac{n}{n+\alpha+1}} ,$$

where $z = (1+\alpha)(1-q)/n$.

2- See D. V. Widder, "The Laplace Transform", Princeton University Press, 1946, Pages 46 - 57.

Now, by the use of (1), for a "slowly diverging" sequence of t_k, we have

$$\int_{t_k}^{t_{k+1}} dt\,|w(t)| \leq \frac{Gp^{1+\alpha}}{k^{(1+\alpha)(1-q)}}, \text{ when } (t_{k+1}-t_k) < T, \quad (10)$$
$$\text{and } (t_{k+1}-t_k) \leq pk^{q-1}.$$

The lower of the two bounds on the integral in (9) and (10) is found to be provided by (9) only when

$$t_k > \left(\frac{M_n}{Gp^{1+\alpha}}\right)^{1/n} k^z , \qquad (11)$$

with z defined as before. By taking the values t_k less than this limit, for each k, only the greater of the two bounds, given by (10), applies. This selection of t_k is possible for $k > 1$, since for t_k equal to the right-hand member of (11), and for p and q chosen above,

$$t_{k+1}-t_k = \left(\frac{M_n}{Gp^{1+\alpha}}\right)^{1/n} \cdot \left[zk^{z-1} + \frac{z(z-1)k^{z-2}}{2!} + \ldots\right]. (12)$$

The series in brackets here is an alternating one, since $z < \frac{1+\alpha}{1+\alpha+n} < 1$ and converges for $k > 1$, so that the error in neglecting all terms beyond the first is less than the second term (the terms being monotonically decreasing in magnitude). For $k > 1$, therefore,

$$t_{k+1}-t_k < \left(\frac{M_n}{Gp^{1+\alpha}}\right)^{1/n} zk^{z-1}. \left(1 - \frac{z-1}{2k}\right) < pk^{q-1}, \quad (13)$$

according to our choice of n, q, and p.

Thus, we have shown that the largest possible terms of the series I_k have their values limited by (10), and so must form a convergent series, since each bound is proportional to k^{-r}, where $r > 1$.

For the "rapidly diverging" case where $(t_{k+1}-t_k) > pk^{q-1}$, we have $t_{k+1} > \sum_{r\,o}^{k} pr^{q-1}$.

Now, the series $\sum_{k=o}^{\infty} \frac{M_n}{\left(\sum_{r\,o}^{k-1} pr^{q-1}\right)^n}$, has terms less, term by term, than those of (9). Further, since

$$\frac{\sum_{r=o}^{k} pr^{q-1}}{\sum_{r=o}^{k-1} pr^{q-1}} = 1 + \frac{k^{q-1}}{\sum_{r\,o}^{k-1} r^{q-1}} \text{ , which for}$$

$k > N_o$ is greater than $\left[1 + \frac{N_o^{q-1}}{\sum_{r=o}^{k-1} r^{q-1}}\right]$.

Then both series converge by application of the ratio test.

The terms for $(t_{k+1}-t_k) \geq T$ pose no problem, since they must be finite in number and represent integration over a finite interval of t, and therefore must be bounded in their sum according to the original restrictions on the form of w(t). Thus, we have shown that J converges in case (B-2), and have completed the proof of part I of the theorem. While we have made direct use of the properties of H(s) only at the origin, the convergence and analyticity of H(s) there demand that it be regular elsewhere on the imaginary axis and in the right half-plane,[3] so that the hypothesis cannot actually be weakened.

In proving part II, we note that if J converges, H(s) must converge for $Re(s) \geq 0$, and by the reference just cited, must be regular to the right of the imaginary axis. Finally, since

$$|H(s)| \leq J \qquad (14)$$

for $Re(s) \geq 0$, it is clear that the transform, H(s), must be bounded on the imaginary axis. This completes the proof of the theorem.

THEOREM II: THERE EXISTS AT LEAST ONE SINGULARITY ON THE ABSCISSA OF CONVERGENCE OF H(s) IF THE ABSCISSA IS FINITE. Here again, we have three cases to consider, depending upon the forms of w(t), and will classify the cases as in the proof of Theorem I. For case A, the theorem has already been proven.[4] Case B1 will be discussed later. Case B2 calls for the use of a theorem of Delange[5] stated for the case in which $\eta(t)$ is a given complex function defined for $t \geq 0$, and having bounded variation on every finite interval. The integral

$\int_o^{\infty} e^{-st} d\,\eta(t)$ is assumed to have a finite

- -

3- See Widder (2) page 57.
4- See Widder, page 58; also Doetsch, "Handbuch der Laplace-Transformation Band I Theorie der Laplace-Transformation" Birkhauser, Basel, 1950, page 153.
5- Delange, H., "Sur les singularites des integrales de Laplace", Comptes Rendu Acad. Sci., Paris, v 233, pp 1413-1414 (1951).

abscissa of convergence, σ_c, and the function represented by the integral is designated f(s). An approximate translation of the theorem is as follows: We suppose that there exists a real function $\psi(t)$ continuous for t greater than or equal to a certain t_o positive or zero, and a real number ϕ satisfying $0 \leq \phi < \pi/2$, such that: (1) for whatever t' and t" satisfying $t_o \leq t' < t"$, the quantity $\int_{t'}^{t"} e^{-i\psi(u)} d\eta(u)$ has an argument of absolute value at most equal to ϕ, or entirely zero; (2) for whatever t' and t" greater than or equal to t_o, $|\psi(t") - \psi(t')| \leq \gamma |t" - t'|$. Then, f(s) possesses at least one singular point on the segment $[\sigma_c - ikC(\phi), \sigma_c + ikC(\phi)]$, where $C(\phi)$ is a positive number uniquely dependent on ϕ. Delange adds that if $C(\phi)$ is taken as small as possible, it must be a non-decreasing function of ϕ, which tends to oo as ϕ tends to $\pi/2$.

It is necessary only to show that there exists a real continuous function $\psi(t)$ and a real constant ϕ, $0 \leq \phi < \pi/2$ such that the argument of $\int_{t'}^{t"} du\, e^{-i\psi(u)} w(u)$ does not exceed ϕ in absolute value when t' and t" are large; and such that $|\psi(t") - \psi(t')| \leq \gamma |t" - t'|$, where γ is some positive real constant. In searching for a suitable $\psi(t)$ we are guided by the fact that the argument of w(t) itself (considered as a complex function) oscillates between zero and π at the points of alternation in sign, t_k. We shall assume, arbitrarily, that when k is odd, the sign of w(t) changes from positive to negative. A satisfactory function is, then,

$$\psi(t) = \pi/2 + (-1)^{k+1} A \sin \pi \frac{t - t_k}{t_{k+1} - t_k}, \quad t_k < t < t_{k+1} \tag{15}$$

with $A = T$ when $t_{k+1} - t_k \geq T$, and $A = (t_{k+1} - t_k)$ when $t_{k+1} - t_k < T$.
To see that the magnitude of the argument of the above integral does not exceed in magnitude some $\phi < \pi/2$ it is only necessary to note that the argument of the entire integrand is less than $\pi/2$ in magnitude for all t except on a set of zero measure, and therefore the argument of the integral cannot exceed some number less than $\pi/2$.

In regard to the second condition on ψ, we have

$$\left|\psi(t") - \psi(t')\right| \leq \int_{t'}^{t"} dt \left|\frac{d\psi}{dt}\right| \leq \pi |t" - t'| \tag{16}$$

Therefore, this condition is met, with $\gamma = \pi$

Case B1 calls for the same proof as B2 except that for $t > t_K$,

$$\psi(t) = \pi/2 + (-1)^{K+1} T \tanh (t - t_K)$$

Thus, the theorem is proved for all cases.

Physical Restrictions on w(t)

The restriction (1) on the form of w(t) can be shown to be equivalent to a restriction on the nature of the observable response of a physical system to an impulse function. If, for example, y(t) represents an observable such as pressure, current, voltage, velocity, etc., and x denotes a disturbance on the system; and if y is uniquely and linearly related to the history of x, we can write

$$y(t) = \int_0^{\infty} d\tau\, w(\tau) . x(t - \tau). \tag{18}$$

It is a justifiable assumption that for every system, by taking sufficient pains to produce a pulse of sufficiently short duration, x(t) can be made to approximate to any desired degree the effect of an impulse of finite area but of duration approaching zero. For such an excitation on the system, it is clear that as the duration of the pulse of x(t) is made to diminish,

$$y(t) \rightarrow w(t). \tag{19}$$

For each type of observable quantity mentioned above, as well as for others, there is an associated energy storage at the point of observation. For example, in reading the value of a voltage on a pair of terminals, one must recognize the existence of an electric field across them, whose stored energy at any instant is proportional to the square of the voltage. If we now compute the average value of y over the interval between two reversals of sign of w(t), we shall have

$$\bar{y} \equiv \frac{1}{t_{k\,1} - t_k} \int_{t_k}^{t_{k+1}} dt\, w(t) . \tag{20}$$

Now, the average energy associated with y over this interval equals or exceeds $\bar{y}^2$. Therefore, the peak power, P, involved in the transfer of energy into and out of the energy storage exceeds the average energy divided by the time interval $t_{k+1} - t_k$, or

$$P \geq \frac{\bar{y}^2}{t_{k+1} - t_k} . \tag{21}$$

The possibility of occurrence of a true impulse or Dirac delta function in w(t) is eliminated immediately by the requirement that the energy storage at the terminal be finite. Further, an argument similar to the above assures the continuity of $w(t)$. An obvious manipulation of (20) and (21) allows us to state the following THEOREM III: FOR A TIME FUNCTION REPRESENTING THE RESPONSE OF A LINEAR SYSTEM TO A PULSE EXCITATION MEASURED AT A POINT INVOLVING FINITE ENERGY STORAGE IN PROPORTION TO THE SQUARE OF THE OBSERVED QUANTITY, THE PROPERTIES (1) OBTAIN WITH $\alpha = 1/2$ AS THE PULSE APPROACHES THE FORM OF A DIRAC DELTA FUNCTION.

Application to Stability Problems

It is possible to apply the foregoing results to the problem of determining whether or not $w(t)$ represents the impulse response of a stable system, when the available information is contained in the transform, $H(s)$.

In accordance with widely accepted definitions, we say that A STABLE SYSTEM IS ONE WHOSE RESPONSE TO ANY BOUNDED INPUT IS A BOUNDED OUTPUT FUNCTION. Linearity in the system is assumed, as are the other broad restrictions of Theorem III, throughout the following discussion. It can be shown[-6] that the necessary and sufficient condition for stability of a linear system under the above definition is that the integral of the absolute value of $w(t)$ over the interval $(0,\infty)$ converge. In making use of these principles, one frequently finds it desirable to have the means of relating the property of stability to characteristics of $H(s)$, the transform of $w(t)$. Many mathematical and engineering problems proceed most conveniently in the complex frequency domain to the point of solution, and to take full advantage of this convenience it is necessary to have available a criterion of stability stated in that domain. In short, given a function of the complex frequency s, one must have the means of knowing whether or not it is the transform of a stable impulse response. To provide such a criterion, we shall assemble the results of Theorems I, II, and III in the following form: THEOREM IV: GIVEN A FUNCTION $F(s)$, A SUFFICIENT CONDITION FOR THE FUNCTION OF WHICH IT IS THE TRANSFORM TO BE A STABLE WEIGHTING FUNCTION IS THAT $F(s)$ BE REGULAR IN THE RIGHT HALF OF THE s-PLANE AND ON THE IMAGINARY AXIS. THE TRANSFORM OF A STABLE WEIGHTING FUNCTION IS BOUNDED ON THE IMAGINARY AXIS AND REGULAR IN THE RIGHT HALF OF THE s-PLANE.

It should be noted that the statement of the theorem assumes that $F(s)$ is of such form that it is the transform of some time function. To prove the first statement, we note by the second theorem that the abscissa of convergence must be negative unless $F(s)$ is entire. Thus, by Theorem I, the weighting function must be stable, even when $F(s)$ is entire. (We recall that the hypotheses of Theorem I are fulfilled by assumption of the conditions of Theorem III.) This completes the proof of the theorem, since the second part is a re-statement of the second part of Theorem I.

The relation between the criterion of stability just derived and those already in existence deserves some comment. A large body of theory has been developed[-7] to provide graphical and other methods of determining whether or not the function $F(s)$ possesses singularities in the right half of the s-plane, and dealing, in particular, with this problem in the case when the system under study comprises one or more feedback loops. In general these methods must proceed on the assumption of the existence of a theorem such as Theorem IV above, or must place such restrictions on the time functions under consideration as to insure that all singularities of the closed-loop transfer function in the right half of the s-plane are poles. An exception is the classical work of Nyquist[-7] which concerns itself with the determination of stability or lack thereof in a feedback system having a loop impulse response of bounded variation and absolutely integrable over the positive infinite time domain. A derived condition is that the loop gain function with harmonic excitation vanish ultimately with increasing frequency. This condition must hold for any physical system, whether considered as a loop transfer function or as the general response under consideration in Theorem III of this paper. The restriction of greatest practical importance is that which demands that the loop be stable when opened. In view of these facts it would appear that a method based on Theorem IV and capable of showing whether or not the closed-loop transfer function possesses the right properties in the right half-plane and on the imaginary axis would provide greater generality in certain respects.

Acknowledgment

The author wishes to express his indebtedness to Dr. J. L. Barnes for his many suggestions and criticisms on an earlier version of this paper. Thanks to the careful reading and comments by Dr. R. S. Phillips some earlier errors in the statement and proof of Theorem I were located and corrected, and clarifications were made in other parts. The comments of Dr. R. E. Roberson, especially in regard to the application of Theorem I, are very much appreciated.

- -

6- See H. H. James, N. B. Nichols, and R. S. Phillips "Theory of Servomechanisms" McGraw-Hill, 1947, page 38.

7- See L. S. Dzung, "The Stability Criterion" p 13-23 of "Automatic and Manual Control" edited by A. Tustin, Academic Press, 1952. This provides a brief summary of the material.

7- H. Nyquist, "Regeneration Theory", Bell Telephone System Monograph B-642, January 1932.

FEEDBACK SYSTEM SYNTHESIS BY THE INVERSE ROOT-LOCUS METHOD

John A. Aseltine
Systems Research Corporation
Van Nuys, California

Abstract

A concise procedure for root-locus analysis of negative and positive feedback systems is presented. The root-locus method is extended to include the synthesis problem and is illustrated by example. A new definition of feedback sign based on dynamic system behavior is proposed.

Introduction

The root-locus method has been used extensively in the analysis of closed-loop systems.[1,2] It will be seen below that the same techniques used for analysis can be applied to the synthesis problem. The qualitative use of the root-locus method will be emphasized here, since the method is often most useful as a first step prior to computer study of a problem.

The conventional definitions of feedback sign are based on static system performance and become ambiguous in complex systems and in the discussion of synthesis. A new convention for the sign of the feedback has been adopted for the reasons set forth in Appendix A. We will call the feedback by the same sign as the open loop transfer function when the latter contains terms of the form $(s - s_1)$ with all coefficients of $\underline{s}$ terms positive.

The Root-Locus Method

Like all special methods of linear feedback system analysis, the root-locus method is used to determine the positions of the roots of

$$1 \pm KG(s) = 0, \quad K \geq 0 \tag{1}$$

where s is a complex variable. This problem arises in connection with the analysis of the system shown in Figure 1. The closed loop transfer function is

$$Y(s) = \frac{K_1 G(s)}{1 \pm K_1 G(s)} \tag{2}$$

The denominator of (2) is of the form of the left side of (1), the + sign in (1) corresponding to negative feedback, and the minus to positive feedback. Y(s) has zeros at the zeros of G(s) and poles at the roots of equation (1). Since G(s) is in practice usually available in factored form, the principal problem arises in solving equation (1). The root-locus method is a graphical method for doing this. Evidently, the requirements are that s satisfy

$$|G(s)| = \frac{1}{K} \tag{3}$$

and, for negative feedback,

$$\angle[G(s)] = n\pi, \quad n \text{ odd} \tag{4}$$

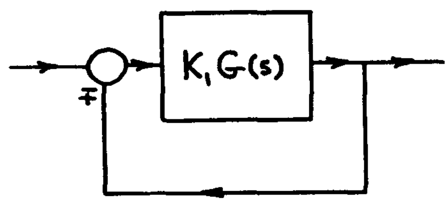

Fig. 1 Feedback System

The greatest utility in the method is that equations (3) and (4) can be solved graphically to yield loci of roots of (1) with K as a parameter. These loci can be constructed rapidly and yield information relating both to steady-state and transient behavior of the closed-loop system.

One of the most useful features of the method is that it provides rapid qualitative evaluation of a closed loop system. Of the many rules available[2] for construction of loci from a plot of the poles and zeros of G(s), the following form a set which allows reasonably accurate sketching of a root-locus diagram for <u>negative feedback</u> in almost every case. The open loop poles and zeros are first plotted in the s-plane.

1. Along the real axis of s-plane, a locus exists wherever total number of real poles and real zeros to the right is odd.

2. Arrows indicating increasing K should be placed on loci so that locus starts (K = 0) on the poles of G(s) and ends ($K \to \infty$) on the zeros (which may be at ∞).

3. Whenever a locus leaves the real axis, it must be accompanied by its complex conjugate.

4. Far from the origin, the loci are straight lines with angles given by:

[No. of finite poles] - [No. of finite zeros]	angles
1	π
2	$\pm \pi/2$
3	$\pm \pi/3, \pi$
⋮	⋮
n	$m\pi/n$, m odd

These asymptotes pass through the centroid of the poles and zeros with the poles counted as positive and zeros counted as negative masses.

In addition, it is useful to know that the loci follow flow lines with poles taken as sources and zeros as sinks. Examples are shown in Figure 2. For a given value of K, the poles of the closed-loop function Y(s) lie at some set of positions on the loci.

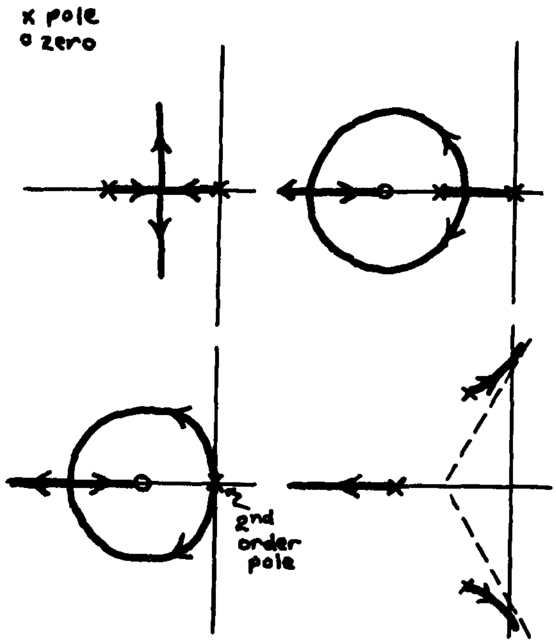

Fig. 2 Root Locus Examples - Negative Feedback

Having sketched the root-locus diagram, the behavior of the closed-loop system can be predicted since the loci represent positions of poles of the latter. At this point, a more accurate plot can be constructed[2] or the information used to guide a computer study of the system.

Positive Feedback

In case the feedback is positive (i.e., positive open-loop transfer function when all $(s - s_1)$ terms have positive coefficients for s), the rules are modified slightly. The changes are in rules 1 and 4, which become for positive feedback:

(1a) Loci exist on the real axis when the number of real poles and real zeros to the right is even.

(4a) Angles of asymptotes are given by:

[No. of finite poles] - [No. of finite zeros]	angles
1	0
2	$0, \pi$
3	$0, \pm 2\pi/3$
⋮	⋮
n	$m\pi/n$, m even

Examples of root-locus plots for positive feedback systems are shown in Figure 3.

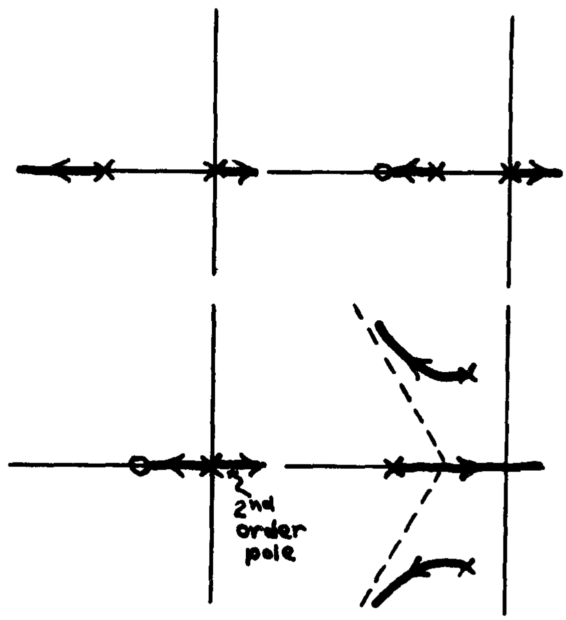

Fig. 3 Root Locus Examples - Positive Feedback

Inverse Root-Locus Method

In the previous sections, we discussed the problem of determining the closed-loop behavior of a system, given the open loop transfer function. The synthesis problem is to find G(s), given the closed-loop transfer function. To this end we write the identity:

$$KG(s) \equiv \frac{\dfrac{KG(s)}{1 \pm KG(s)}}{1 \mp \dfrac{KG(s)}{1 \pm KG(s)}} \qquad (5)$$

and note that it has the same form as (2). We conclude that KG(s) has the same zeros as KG(s) /(1 $\pm$ KG(s)) and that the poles of KG(s) are at the roots of the denominator of (5). The factoring problem presented here is the same as the one connected with equation (1), and the locus of roots of the denominator of (5) can be drawn using rules previously described.

The synthesis procedure for a transfer function with more poles than zeros begins with the plotting of the poles and zeros of the desired closed-loop system on the s-plane. The locus can be drawn for either positive or negative feedback, a positive feedback plot corresponding to the synthesis of a negative feedback system, and vice versa. For a given value of gain K, a set of roots on the loci will be obtained. When these are used to characterize the open-loop function G(s), the feedback system will have the specified poles and zeros.

When the number of zeros of the system to be synthesized is equal to or greater than the number of poles the procedure is changed. The changes are derived in the Appendix and are tabulated below. We assume that the transfer function to be synthesized is of the form

$$\frac{G(s)}{1 \pm KG(s)} = \frac{s^n + a_1 s^{n-1} + \cdots + a_n}{s^m + b_1 s^{m-1} + \cdots + b_m} \qquad (6)$$

Then the synthesis procedure can be summarized as follows:

	positive plot yields system with	negative plot yields system with
$n < m$	neg. feedback	pos. feedback
$n = m$*	neg. feedback, $K < 1$ pos. feedback, $K > 1$	pos. feedback
$m < n$	pos. feedback	pos. feedback

*When $n = m$ the locus passes through the point at ∞ when $K = 1$.

It should be pointed out that when $n \geq m$ the amplitude response of the synthesized system does not diminish at high frequencies, making such systems subject to noise saturation.

Illustrative Example

To illustrate the synthesis technique, let it be desired to obtain a negative feedback system with transfer function,

$$Y(s) = \frac{K(s + 2)}{(s + 4)\,(s^2 + 8s + 32)} \qquad (7)$$

The poles and zeros of Y(s) have been plotted in Figure 4, and an inverse root-locus drawn. Using conventional techniques,[1,2] the values of K for various points along the locus have been found. Depending on the value of K specified, the open-loop poles and zeros can be chosen. A set of poles and zeros for K = 30 have been drawn in Figure 5, and a conventional root-locus drawn. At the K = 30 point on the conventional plot the closed-loop roots have the specified values of (7).

It can be shown that the loci in Figures 4 and 5 between the pole and the point K = 30 would coincide if the two plots were drawn on the same diagram.

Synthesis with Feedback-Path Transfer Function

The more general feedback system with transfer function H(s) in the feedback path can be synthesized by a simple extension of the preceding method. The closed-loop transfer function is

$$Y_1(s) = \frac{KG(s)}{1 \pm KG(s)H(s)} \qquad (8)$$

Now, if a function H(s) is chosen, G(s) can be found as before by starting with poles and zeros of Y_1 (s)H(s) from which KG(s)H(s) is determined. Then, since H(s) has been specified previously, it can be used to find KG(s) simply by superimposing the poles and zeros of 1/H(s) on the KG(s)H(s) plot.

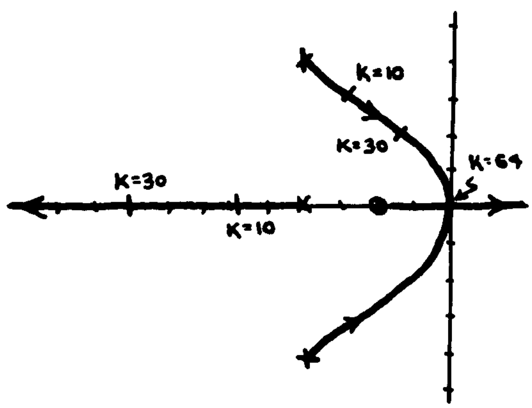

Fig. 4 Synthesis Example

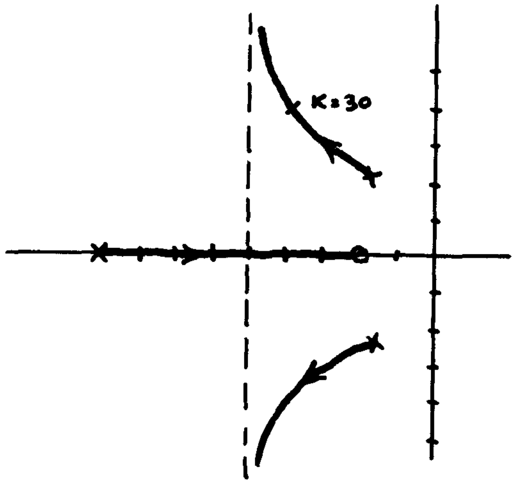

Fig. 5 Root-Locus Plot for Synthesized System

To illustrate the procedure, we consider the problem of finding G(s) in the previous example, but with H(s) = (s + 6). The inverse plot is sketched in Figure 6, and a typical G(s) might have poles and zeros as indicated in Figure 7, where superimposing 1/H(s) = 1/(s + 6) has removed the zero at s = -6.

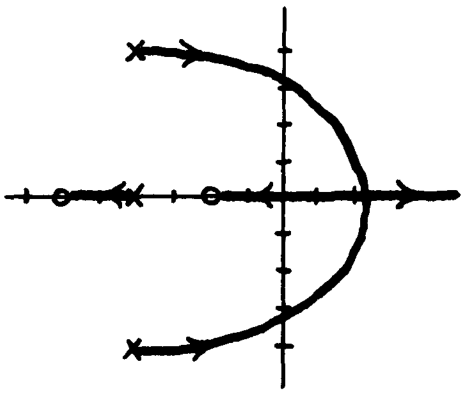

Fig. 6 Synthesis with H(s)

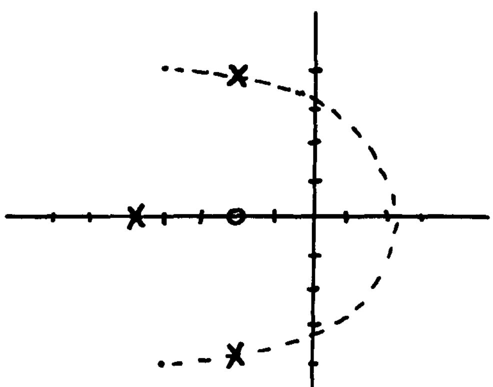

Fig. 7 Open-loop G(s) for Example of Fig. 6

Conclusions

The use of inverse root-locus for synthesis is especially attractive from the point of view of simplicity and amount of information which can be obtained quickly. In addition to its usefulness in synthesizing feedback amplifiers and other devices, it is also useful in determining open-loop characteristics from measured closed-loop data (e.g., airframe characteristics from measured autopilot-airframe response). Like root-locus, the inverse method keeps the system poles and zeros apparent during analytical work, and provides the designer with a geometric picture lacking in algebraic methods. Also like root-locus, it is a trial and error method with the rules presented here giving a first (usually good) approximation.

Appendix

A. Feedback Sign Convention

The angle condition (4) is

$$\angle[G(s)] = \angle\left[\frac{(s-s_1)(s-s_3)\cdots(s-s_n)}{(s-s_2)(s-s_4)\cdots(s-s_m)}\right] = n\pi \qquad (9)$$

The rules for plotting the root-locus are based on the measurement of angles like θ in the s-plane as shown in Figure 8:

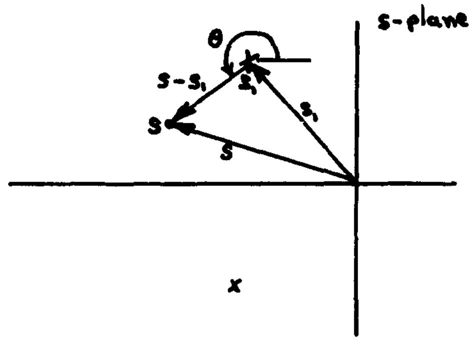

Fig. 8 s-plane plot for angle condition

Therefore, we require that G(s) contain terms of the form $(s - s_1)$ with positive coefficient for s. If terms like $(-s - s_1)$ occur,* they can be put in proper form by changing the sign of the transfer function. This leads us to establish the convention:

> The sign of the feedback is the same as the sign of the open-loop transfer function when all terms in the latter are of the form $(s - s_1)$ with all coefficients of s positive.

The angle condition now can be written

$$\angle[G(s)] = n\pi, \begin{cases} n \text{ odd for neg. feedback} \\ n \text{ even for pos. feedback} \end{cases} \qquad (10)$$

*Terms like this may occur, for example, in multi-loop systems.