CHAPTER

FOUR

FREQUENCY-DOMAIN ANALYSIS

Control System Design
Friedland, B., 1986
McGraw-Hill

4.1 STATUS OF FREQUENCY-DOMAIN METHODS

For a period of about twenty years—from the early 1940s through the early 1960s—frequency-domain methods were the only systematic tools for the analysis and design of control systems. These methods were developed by physicists and electrical engineers in response to the World War II need for improved servomechanisms to be used in various weapons systems, and were based upon the frequency response/operational calculus methods then in use for designing electrical networks for communication systems. It is no coincidence that the pioneering work of Nyquist[1] and Bode[2] in the early part of the century, and even the very invention of the feedback amplifier by Black,[3] all products of the Bell Telephone Laboratories, were done in the interest of improved communication systems.

(The connection between frequency-domain methods and communication systems is a possible explanation of why the development of control theory took place and still continues mostly in academic departments of electrical engineering, even though the electrical hardware in many control systems is all but negligible.)

Through the interdisciplinary activities of individuals such as the late Rufus Oldenburger, a mechanical engineer who understood and appreciated the significance of frequency-domain methods, these techniques were introduced to other branches of engineering and became widely used throughout the entire field of automatic control.

Just at the time that frequency-domain methods had reached their peak of development, in the late 1950s and early 1960s, the alternative state-space

methods began to make their appearance. But while the new state-space methods developed rapidly in the decades following and found new adherents and apostles, the vigor of frequency-domain methods hardly diminished. Notwithstanding the level to which state-space methods have been developed, most control systems continue to be analyzed and designed by frequency-domain methods. Concepts such as "bandwidth," "phase and gain margins," and "corner frequencies" are entrenched in control system technology and are not likely to be displaced. They continue to be useful.

Starting in the mid 1970s, new impetus was imparted to frequency-domain methods for multivariable systems through the efforts of a number of investigators centered in Great Britain around Rosenbrock and MacFarlane. (See Note 4.1.) Among the fruits of this effort was a growing recognition that frequency-domain methods and state-space methods enhance and complement each other. The burgeoning theory of robust control systems, which was started only in the past few years, is further evidence of the symbiosis of frequency-domain and state-space methods.

4.2 FREQUENCY-DOMAIN CHARACTERIZATION OF DYNAMIC BEHAVIOR

The fundamental concept of frequency-domain analysis is the "transfer function" which expresses the relationship between the Laplace transform $y(s)$ of the system output $y(t)$ and the Laplace transform $u(s)$ of the input $u(t)$

$$y(s) = H(s)u(s) \tag{4.1}$$

where $H(s)$ is the transfer function of the system. This relationship is valid for any time-invariant linear system, even when the system cannot be represented by sets of ordinary differential equations of finite order. The representation (4.1) is valid, for example, for systems whose physical properties are described by partial differential equations, or by pure "transport" delays.

The validity of (4.1) is a consequence of the linearity and time invariance of the system under examination. In the time domain such a system can be represented by the convolution integral

$$y(t) = \int_0^t H(t-\tau)u(\tau)\,d\tau \tag{4.2}$$

where $H(t)$ is the "impulse-response" (matrix) of the system.

The basic frequency-domain relation (4.1) follows from (4.2) as a result of the well-known "convolution theorem" proved in many texts (see[4], for example) which asserts that the Laplace transform of a convolution of two functions is the product of the respective Laplace transforms of these functions. Thus, the transfer function $H(s)$ is the Laplace transform of the impulse response:

$$H(s) = \mathscr{L}[H(t)] = \int_0^\infty e^{-st}H(t)\,dt \tag{4.3}$$

When the number of inputs and/or outputs is greater than 1, then H(s) is a matrix of appropriate dimension: if there are m inputs and l outputs, then H(s) is an l-by-m matrix, the elements of which are the transfer functions from the individual components of the input vector to the individual components of the output vector.

When the system of interest has the standard state-space representation

$$\dot{x} = Ax + Bu$$

$$y = Cx + Du$$

then, as shown in Chap. 3, the transfer function (matrix) is given explicitly by

$$\begin{aligned} H(s) &= C(sI - A)^{-1}B + D \\ &= \frac{C(E_1 s^{k-1} + E_2 s^{k-2} + \cdots + E_k)B}{s^k + a_1 s^{k-1} + \cdots + a_k} + D \end{aligned} \tag{4.4}$$

where the denominator of H(s) is the characteristic polynomial

$$D(s) = |sI - A| = s^k + a_1 s^{k-1} + \cdots + a_k \tag{4.5}$$

and $E_1 = I, E_2, \ldots, E_k$ are the coefficient matrices of the adjoint matrix for the resolvent $(sI - A)^{-1}$, as discussed in Chap. 3. The roots of the *characteristic equation* $|sI - A| = 0$ are called the *characteristic roots* or *eigenvalues* of the system.

If the D matrix is nonzero, there is a direct path from some input to some output. The transfer functions from those inputs that are directly connected to the output will be polynomials of degree k in s. All the other transfer functions are *proper* rational functions, that is, ratios of polynomials in s in which the degree of the numerator is strictly less than the degree of the denominator.

The variable s of the Laplace transform is a complex variable

$$s = \sigma + j\omega \qquad j = \sqrt{-1}$$

called *complex frequency. Frequency-domain* analysis owes its name to this identification of s with complex frequency.

A transfer function H(s) which is a proper rational function of s can be expanded in partial fractions

$$\begin{aligned} H(s) &= \frac{N_1 s^{k-1} + \cdots + N_k}{s^k + d_1 s^{k-1} + \cdots + d_k} \\ &= H_1(s) + H_2(s) + \cdots + H_{\bar{k}}(s) \end{aligned} \tag{4.6}$$

where

$$H_i(s) = \frac{R_{1i}}{s - s_i} + \frac{R_{2i}}{(s - s_i)^2} + \cdots + \frac{R_{\nu_i i}}{(s - s_i)^{\nu_i}} \tag{4.7}$$

The complex frequencies $s_i (i = 1, 2, \ldots, \bar{k} < k)$ are the *distinct* roots of the denominator of (4.6) and the ν_i are corresponding multiplicities of these roots. These roots of the denominator are called the *poles* of the transfer function because H(s) becomes infinite at these complex frequencies and a contour map of the complex plane appears as if it has poles sticking up from these points.

If H(s) is a matrix, then the coefficients N_i of the numerator polynomial of (4.6) are matrices and so are the coefficient matrices R_{ij} of the partial fraction expansion.

The impulse response $H(t)$ of the system is given by the inverse Laplace transform of (4.6):

$$H(t) = H_1(t) + H_2(t) + \cdots + H_{\bar{k}}(t) \tag{4.8}$$

where

$$H_i(t) = (R_{1i} + R_{2i}t + \cdots + R_{\nu_i}t^{\nu_i - 1}/(\nu_i - 1)!)\,e^{s_i t} \tag{4.9}$$

Thus the impulse response of a time-invariant linear system having a proper rational function of s as its transfer function is a sum of time-weighted exponentials of the form of (4.9). The exponents of the exponentials are the poles of the transfer function, and the time-weighting functions are polynomials in t of one degree less than the multiplicity of the corresponding poles.

If the numerator of the transfer function (4.6) is the same degree as the denominator, the constant term can be removed and the remainder written as a proper rational function, i.e.,

$$H(s) = \frac{N_0 s^k + N_1 s^{k-1} + \cdots + N_k}{s^k + d_1 s^{k-1} + \cdots + d_k} = N_0 + \frac{\bar{N}_1 s^{k-1} + \cdots + \bar{N}_k}{s^k + d_1 s^{k-1} + \cdots + d_k} \tag{4.10}$$

$$\bar{N}_i = N_i - N_0 d_i \qquad (i = 1, 2, \ldots, k) \tag{4.11}$$

The corresponding impulse response has the form

$$H(t) = N_0\delta(t) + \sum_{i=1}^{\bar{k}} (R_{1i} + \cdots + R_{\nu_i}t^{\nu_i - 1}/(\nu_i - 1)!)\,e^{s_i t} \tag{4.12}$$

where $\delta(t)$ is the unit impulse function (*Dirac delta* function).

It is certainly possible to conceive of systems having transfer functions in which the degree of the numerator is higher than the denominator. For example, an electrical inductor has the transfer function (complex impedance)

$$\frac{v(s)}{i(s)} = z(s) = Ls$$

when the voltage $v(t)$ is regarded as the output and the current is regarded as the input. The impulse response of such systems, in general, contains not only impulses, but various derivatives (doublets, etc.) of impulses. These are bothersome and can generally be avoided by suitable reformulation of the problem. If the voltage, in the case of the inductor, is regarded as the input and the current is regarded as the output, then the transfer function is the admittance

$$\frac{i(s)}{v(s)} = y(s) = \frac{1}{Ls}$$

which is a perfectly acceptable, proper rational function.

The general form of the transfer function (4.10) is consistent with the transfer function of the state-space representation given by (4.4). In particular

$$N_0 = D$$

and

$$\begin{aligned} \bar{N}_i &= CE_iB \\ d_i &= a_i \end{aligned} \qquad i = 1, 2, \ldots, k$$

Thus the impulse response of a system in the standard state-space representation is a sum of time-weighted exponentials $e^{s_i t}$ with the exponents s_i being the roots of the characteristic polynomial, i.e.,

$$|sI - A| = s^k + a_1 s^{k-1} + \cdots + a_k = (s - s_i)^{r_i} \cdots (s - s_k)^{r_k} \qquad (4.13)$$

Multiple poles (i.e., repeated characteristic roots) occur quite frequently at the origin ($s = 0$). For example a pure mass with the transfer function $H(s) = 1/ms^2$ has a double pole at $s = 0$. But multiple poles at other complex frequencies rarely occur in practical problems. To simplify a derivation it is often convenient to assume that multiple poles of a system occur only at the origin.

4.3 BLOCK-DIAGRAM ALGEBRA

One reason for the popularity of frequency-domain analysis is that the dynamic behavior of a system can be studied using only algebraic operations. The transfer functions of subsystems can be combined algebraically to yield the transfer function of the overall system, and its response to various inputs can be obtained by multiplying the Laplace transform of the input by the transfer function, as prescribed by (4.1), to obtain the Laplace transform of the output. The actual output time function, if needed, is calculated by finding the inverse Laplace transform of $y(s)$, using algebraic techniques (partial fractions) in conjunction with a table of Laplace transforms. Nowhere in this analysis, except possibly in deriving the transfer functions of the subsystems, is it necessary to have any dealings with differential equations.

The basic techniques of manipulating block diagrams consist of combining transfer functions in parallel and in tandem and eliminating feedback loops. The three operations are illustrated in Fig 4.1.

Figure 4.1(a) shows a system comprising two subsystems with transfer functions (matrices) $H_1(s)$ and $H_2(s)$. The summing junction, represented by the circle, makes sense only when each subsystem has the same number of outputs, i.e., dimensions of y_1 and y_2 are equal. Then

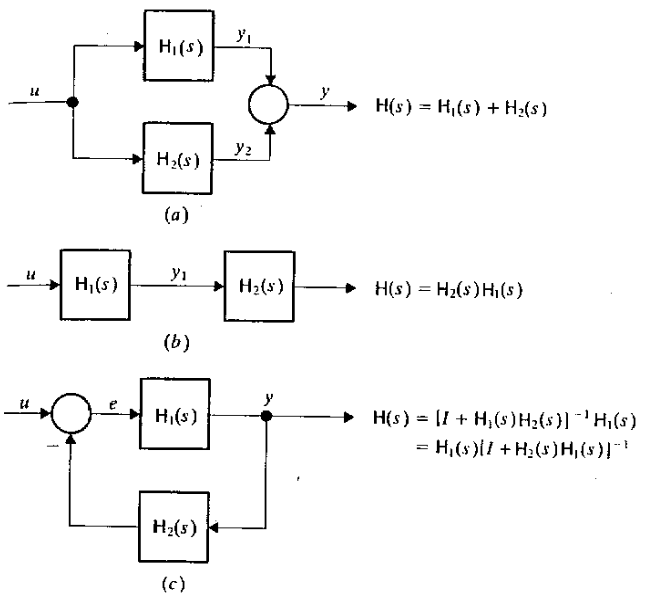

Figure 4.1 Subsystems in combination. (a) Subsystems in parallel; (b) Subsystems in tandem; (c) Single-loop feedback system.

$$\begin{aligned} y(s) &= y_1(s) + y_2(s) = H_1(s)u(s) + H_2(s)u(s) \\ &= [H_1(s) + H_2(s)]u(s) \end{aligned} \qquad (4.14)$$

Thus the transfer function of a parallel combination of subsystems is the sum of the transfer functions.

The *tandem* (or *series*) combination of two subsystems is shown in Fig. 4.1(b). For this combination

$$y_1(s) = H_1(s)u(s)$$

and

$$y(s) = H_2(s)y_1(s)$$

Thus

$$y(s) = H_2(s)H_1(s)u(s)$$

and the transfer function of the tandem combination is the *product* of the transfer functions:

$$H(s) = H_2(s)H_1(s) \qquad (4.15)$$

Note that the order in which the factors of $H(s)$ are placed depends on the order in which the subsystems are connected. In general $H_2(s)H_1(s) \neq H_1(s)H_2(s)$, except when H_1 and H_2 are 1-by-1 matrices.

A system containing a feedback loop is shown in Fig. 4.1(c). The transfer function $H_1(s)$ is called the *forward transmission* and the transfer function $H_2(s)$ is called the *feedback transmission*. The minus sign at the summing junction indicates that the signal e is the *difference* between the system input u and the feedback signal z. This corresponds to *negative* feedback. The transfer function for Fig. 4.1(c) is obtained by tracing the signal flow through the system:

$$y(s) = H_1(s)e(s) = H_1(s)[u(s) - z(s)]$$

But

$$z(s) = H_2(s)y(s)$$

Thus

$$y(s) = H_1(s)[u(s) - H_2(s)y(s)]$$

or

$$[I + H_1(s)H_2(s)]y(s) = H_1(s)u(s)$$

and, finally,

$$y(s) = [I + H_1(s)H_2(s)]^{-1}H_1(s)u(s)$$

Thus, the transfer function (matrix) of the system containing a feedback loop is

$$H(s) = [I + H_1(s)H_2(s)]^{-1}H_1(s) \tag{4.16}$$

The matrix

$$F(s) = I + H_1(s)H_2(s)$$

which may be called the *return-difference* (matrix)—a generalization of the terminology introduced by Bode[2]—has an inverse except at isolated values of s at which the transfer matrix becomes infinite. These values of s are the poles of the system. Since

$$[I + H_1(s)H_2(s)]^{-1} = \frac{\text{adj}\,[I + H_1(s)H_2(s)]}{|I + H_1(s)H_2(s)|}$$

it follows that the characteristic equation of a single-loop feedback system is

$$|I + H_1(s)H_2(s)| = 0 \tag{4.17}$$

In words: the zeros of the determinant of the return difference are the poles of the system.

Alternative expressions for the transfer function are obtained by following different sequences of steps. In particular,

$$y(s) = H_1(s)e(s)$$

and

$$e(s) = u(s) - z(s) = u(s) - H_2(s)H_1(s)e(s)$$

Thus

$$[I + H_2(s)H_1(s)]e(s) = u(s)$$

or

$$e(s) = [I + H_2(s)H_1(s)]^{-1}u(s) \tag{4.18}$$

Finally

$$y(s) = H_1(s)[I + H_2(s)H_1(s)]^{-1}u(s)$$

Thus

$$H(s) = H_1(s)[I + H_2(s)H_1(s)]^{-1} \tag{4.19}$$

From (4.19) it is seen that another form of the characteristic equation of the system is

$$|I + H_2(s)H_1(s)| = 0 \tag{4.20}$$

One should not make the mistake of assuming that $H_1(s)$ and $H_2(s)$ commute just because the order in which they are multiplied does not matter in setting up the characteristic equation. It does follow, however, that $H_1(s)$ and $H_2(s)$ are *conformable*, in whatever order they are multiplied. Since $H_1(s)H_2(s)$ may be a higher-dimension (or lower-dimension) matrix than $H_2(s)H_1(s)$, calculations can be simplified by working with the product having the smaller dimension.

When $H_1(s)$ and $H_2(s)$ are transfer functions of single-input, single-output systems, then both (4.17) and (4.19) reduce to the well-known formula

$$H(s) = \frac{H_1(s)}{1 + H_1(s)H_2(s)} \tag{4.21}$$

and the return difference is

$$F(s) = 1 + H_1(s)H_2(s)$$

By repeated combination of subsystems in parallel, in tandem, and with feedback loops it is often possible to obtain the transfer function of a fairly complex system without performing a great deal of matrix algebra. Instead of by the repeated combination of elements, the block diagram of a single-input, single-output system can be reduced in a single operation by the use of the general-gain formula developed by S. J. Mason.[5] Mason's rule is fraught with possibility of error, however, unless the user is very careful with bookkeeping.

Example 4A Distillation column The fourth-order dynamic model of a distillation column, as developed by Gilles and Retzbach, was given in Chap. 2 (Example 2G on p. 47). The transfer functions from the inputs to the state variables and outputs can be obtained using the matrix calculations described in Chap. 3. But in this example the transfer functions are more readily calculated by block-diagram manipulations.

The block diagram corresponding to the differential equations is shown in Fig. 4.2. The overall system has been subdivided into two subsystems as shown in Fig. 4.3, each corresponding to a different physical aspect of the process. The first subsystem, having a single input Δu_1 and a single output x_2, represents the boiler. The second subsystem then represents the inner operation of the distillation column. The integrators have been represented by their transfer functions, $1/s$. Subsystem 1 itself comprises two single-loop feedback systems, separated by a gain element. Thus, by (4.15) and (4.21)

$$\mathrm{H}_1(s) = \frac{\mathrm{x}_2(s)}{\Delta \mathrm{u}_1(s)} = \frac{1/s}{1-(1/s)a_{22}} a_{21} \frac{1/s}{1-(1/s)a_{11}} \cdot b_{11} = \frac{a_{21}b_{11}}{(s-a_{22})(s-a_{11})} \qquad (4A.1)$$

The second subsystem has two inputs, x_2 and Δu_2, and two outputs, Δz_1 and Δz_2. The input-output relation can be expressed as

$$\begin{bmatrix} \Delta z_1(s) \\ \Delta z_2(s) \end{bmatrix} = \mathrm{H}_{21}(s)\Delta s(s) + \mathrm{H}_{22}(s)\mathrm{x}_2(s) \qquad (4A.2)$$

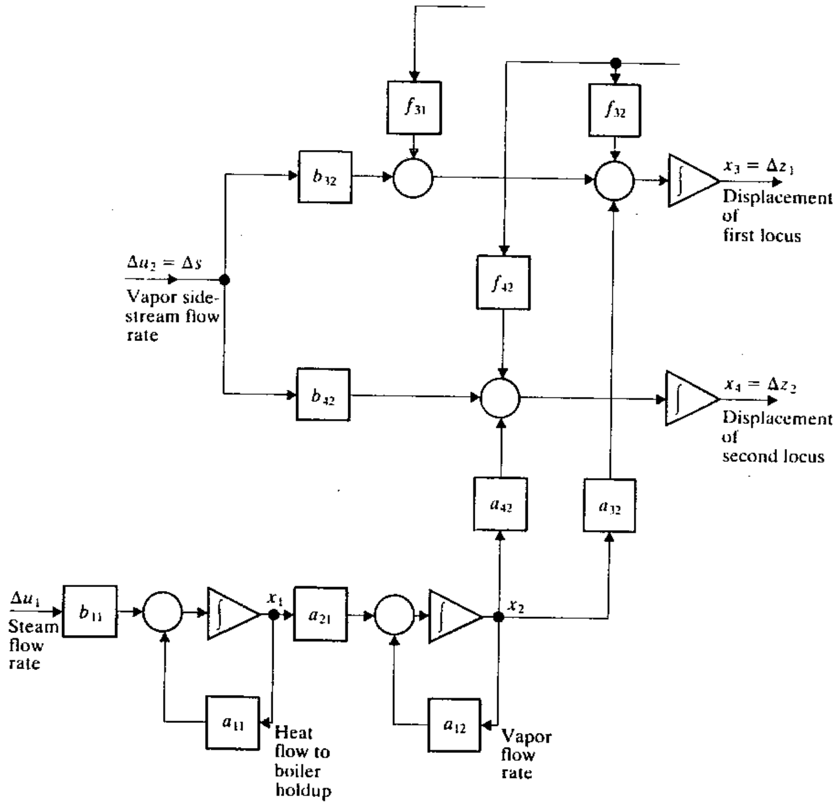

Figure 4.2 Dynamic model of distillation column.

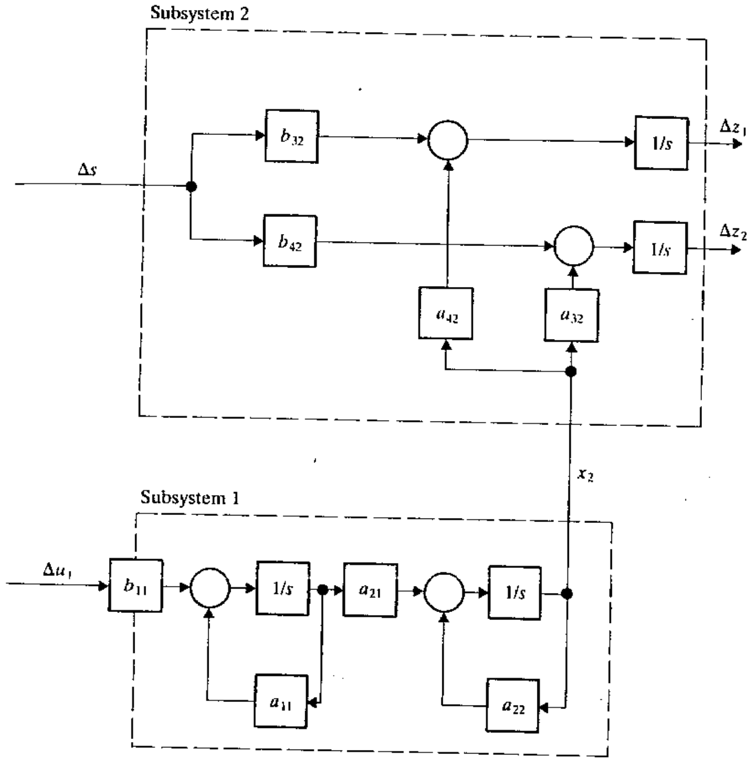

Figure 4.3 Representation of distillation column as two subsystems.

where

$$\mathrm{H}_{21}(s) = \begin{bmatrix} b_{32}/s \\ b_{42}/s \end{bmatrix} \qquad \mathrm{H}_{22}(s) = \begin{bmatrix} a_{32}/s \\ a_{42}/s \end{bmatrix}$$

Substituting $\mathrm{x}_2(s) = \mathrm{H}_1(s)\Delta\mathrm{u}_1(s)$, as given by (4A.1) into (4A.2) gives

$$\begin{bmatrix} \Delta z_1(s) \\ \Delta z_2(s) \end{bmatrix} = H_{21}(s)\Delta s(s) + \begin{bmatrix} b_{11}a_{32}a_{21}/s(s-a_{22})(s-a_{11}) \\ b_{11}a_{42}a_{21}/s(s-a_{22})(s-a_{11}) \end{bmatrix} \Delta \mathrm{u}_1(s)$$

$$= \left[\begin{array}{c|c} \dfrac{b_{32}}{s} & \dfrac{b_{11}a_{32}a_{21}}{s(s-a_{22})(s-a_{11})} \\ \hline \dfrac{b_{42}}{s} & \dfrac{b_{11}a_{42}a_{21}}{s(s-a_{22})(s-a_{11})} \end{array}\right] \begin{bmatrix} \Delta s(s) \\ \Delta \mathrm{u}_1(s) \end{bmatrix}$$

Note that the poles of the system are located at

$$s = 0 \qquad s = a_{11} \qquad s = a_{22}$$

There are only three different poles, although the system is fourth-order. The reason for this is

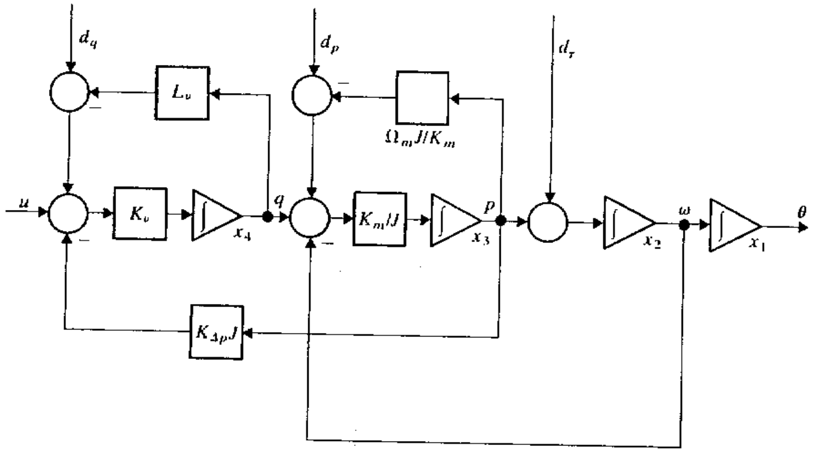

Figure 4.4 Dynamic model of hydraulically actuated tank gun turret.

that subsystem 2 contains two identical dynamic subsystems, namely integrators, in parallel. This has important implications with regard to controllability, as will be discussed in Chap. 5.

Example 4B Hydraulically actuated gun turret A block diagram corresponding to the dynamic model of the hydraulically actuated tank gun turret of Example 2D is shown in Fig 4.4. After simplification by combining the feedback loops around the integrators, the equivalent block diagram, with the disturbances omitted, has the appearance of Fig. 4.5(*a*). The picture is a bit complicated because of the two crossed feedback paths, (1) from ω to the summer after q, and (2) from p to the summer following u.

A trick often used in block-diagram simplification, however, reduces the block diagram of 4.5(*a*) to 4.5(*b*). The trick is to move the starting point for feedback path (1) from ω to p, compensating for the transfer function of $1/s$ from p to ω by placing that transfer function in the moved path (1) as shown in Fig. 4.5(*b*). The transfer function from q to p is given by

$$\frac{p(s)}{q(s)} = \frac{\dfrac{K_m/J}{s+\Omega_m}}{1+\dfrac{K_m/J}{s+\Omega_m}\dfrac{1}{s}} = \frac{(K_m/J)s}{s(s+\Omega_m)+K_m/J} \tag{4B.1}$$

The block-diagram resulting in this simplification is shown in Fig. 4.5(*c*). From this figure it is seen that the transfer function from u to p is

$$\frac{p(s)}{u(s)} = \frac{\dfrac{K_v}{s+K_vL_v}\,\dfrac{(K_m/J)s}{s(s+\Omega_m)+K_m/J}}{1+K_{\Delta p}J\dfrac{K_v}{s+K_vL_v}\,\dfrac{(K_m/J)s}{s(s+\Omega_m)+K_m/J}}$$

$$= \frac{(K_vK_m/J)s}{(s+K_vL_v)[s(s+\Omega_m)+K_m/J]+K_{\Delta p}K_vK_ms} \tag{4B.2}$$

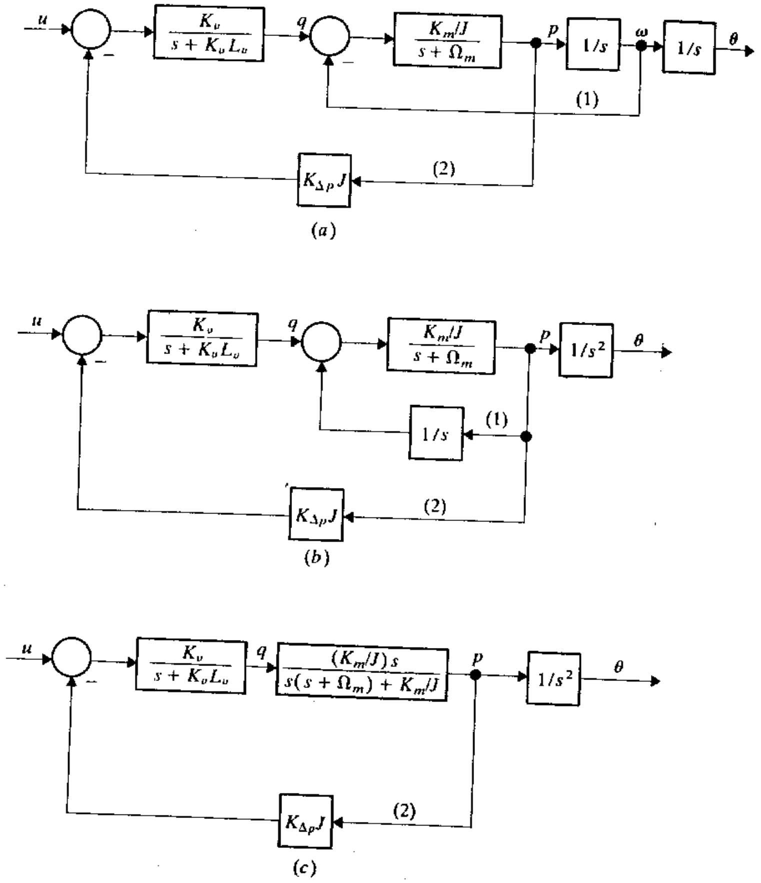

Figure 4.5 Block-diagram simplification of model of hydraulically actuated tank gun turret. (*a*) Figure 4.4 after reduction of loops around integrators; (*b*) Path from ω to q moved to p and integrator added; (*c*) Final simplification.

And the transfer function from the system input u to the angle θ is $1/s^2$ times $p(s)/u(s)$. Thus

$$H(s) = \frac{\theta(s)}{u(s)} = \frac{K_vK_m/J}{s\{(s+K_vL_v)[s(s+\Omega_m)+K_m/J]+K_{\Delta p}K_vK_ms\}}$$

The denominator of $H(s)$ is the characteristic polynomial $D(s)$ of the open-loop system. On expansion it is found that

$$D(s) = s\left[s^3+(\Omega_m+K_vL_v)s^2+\left(\frac{K_m}{J}+\Omega_mK_vL_v+K_{\Delta p}K_vK_m\right)s+\frac{K_mK_vL_v}{J}\right]$$

4.4 STABILITY

The quintessential requirement of a closed-loop dynamic system is stability: the ability of the system to operate under a variety of conditions without "self-destructing."

Two categories of stability are of interest. The first category relates to the ability of the system to return to equilibrium after an arbitrary displacement away from equilibrium, and the second relates to the ability of the system to produce a bounded output for any bounded input. For nonlinear or time-varying systems these categories are distinct: a system may possess one kind of stability without possessing the other. Detailed discussions of these categories and theorems giving conditions for stability can be found in various textbooks on system theory, such as [4].

If we confine our attention to linear, time-invariant systems, however, the situation regarding stability is much simpler. Both categories of stability are all but equivalent. Moreover, the basic stability criterion is directly determined by the locations of the system poles, i.e., the roots of the characteristic equation of the system.

Ability of a system to return to equilibrium relates to the unforced system

$$\dot{x} = Ax \tag{4.22}$$

For the initial state $x(0) = x_0$, the unforced differential equation (4.22) has the solution

$$x(t) = e^{At}x_0 \tag{4.23}$$

where e^{At} is the state-transition matrix, given by

$$e^{At} = \mathscr{L}^{-1}[(sI - A)^{-1}] = \sum_{i=1}^{k}\left(\sum_{j=1}^{\nu_i} R_{ji}t^j \Big/ j!\right) e^{s_i t} \tag{4.24}$$

in accordance with the discussion of Sec. 4.2. The following properties can be directly inferred from the form of the state transition matrix as given by (4.24):

1. If the real parts of *all* the characteristic roots are *strictly* negative (i.e., not zero or positive), then e^{At} tends asymptotically to zero. Hence, no matter how large the initial state x_0 is, $x(t) \to 0$ as $t \to \infty$. The system is said to be *asymptotically stable.*
2. If *any* characteristic root has a strictly positive real part, the state-transition matrix given by (4.24) will have at least one term which will tend to infinity as $t \to \infty$. In this case it is always possible to find some initial state which will cause $x(t)$ to become infinite. The system is said to be *unstable.*
3. If *all* the characteristic roots have nonpositive real parts, but one or more of the characteristic roots has a *zero* real part, the situation is somewhat more complicated: if all the characteristic roots having zero real parts are simple roots, then the corresponding terms in the state-transition matrix are of the form

unit circle
in complex plane

$$R_i e^{j\omega_i t} \qquad j = \sqrt{-1}$$

Since $|e^{j\omega_i t}| = 1$, it is clear that these terms in the state transition matrix are bounded. Hence the state $x(t)$ that evolves from any initial state x_0 will also remain bounded. But there will be some initial states from which the subsequent solution will *not* approach zero asymptotically. Systems of this type are said to be *stable*, but not asymptotically stable. If, on the other hand, any of the characteristic roots that has a zero real part is a *repeated* root, then, owing to the polynomial in t that multiplies $e^{j\omega_i t}$, there will be at least one term in e^{At} which will tend to infinity as $t \to \infty$. ⊗ Hence there will be some initial state for which $x(t) \to \infty$, and the system is unstable. (In the strict sense, the multiplicity of the roots of the *minimum* equation, i.e., the equation of lowest degree satisfied by the matrix A, as discussed in the Appendix, rather than the multiplicity of the roots of the characteristic equation, must be examined to test for the stability of systems with such roots on the imaginary axis.)

The above conclusions are summarized in Table 4.1.

Stability of the second category: bounded-input bounded-output (BIBO) stability is determined using the convolution integral (4.2). Consider only a single-input, single-output system, having a scalar impulse response $h(t)$. For this system

$$y(t) = \int_0^t h(t-\tau)u(\tau)\,d\tau \tag{4.25}$$

It is easy to show that

$$|y(t)| \leq \int_0^t |h(t-\tau)||u(\tau)|\,d\tau \tag{4.26}$$

The meaning of the input $u(t)$ being bounded is that there is a constant c such that

$$|u(t)| \leq c \qquad \text{for all } t \tag{4.27}$$

Table 4.1 Stability conditions for linear systems

Condition		Implication
1. $\mathrm{Re}(s_i) < 0$	for all i	System is asymptotically stable
2. $\mathrm{Re}(s_i) > 0$	for some i	System is unstable
3. $\mathrm{Re}(s_i) = 0$	for some $i = j$, and	
(a) s_i is simple root for all such j		System is stable, but not asymptotically stable
(b) s_i is multiple root for some such j		System is unstable

⊗: this is because no $e^{\alpha t}$ $(s = \alpha + i\beta)$ exists to force it to $\to 0$: , ?

In this case, by (4.26),

$$|y(t)| \leq c \int_0^t |h(t-\tau)|\, d\tau \tag{4.28}$$

In accordance with (4.8) and (4.9), the impulse response of a time-invariant system is a sum of time-weighted exponentials. If the system is asymptotically stable, then the exponentials all tend asymptotically to zero; no matter how large the time-weighting on the exponentials, the integral in (4.28) will be finite for all t (including $t \to \infty$), and hence $|y(t)|$ will be finite. Thus we see that an asymptotically stable time-invariant system produces a bounded output for every bounded input. On the other hand, suppose the system produces an unbounded output for some bounded inputs. This output must result from some term in the impulse response that does not tend asymptotically to zero, which implies that the system is not asymptotically stable. Thus a linear time-invariant system in which a bounded input produces an unbounded output cannot be asymptotically stable. (Although the system is not asymptotically stable, it may still be stable. The simplest example is an integrator for which $h(t) = 1$. For a bounded input, say $u(t) = 1$, $y(t) = t$ which tends to infinity with t, so for this example a bounded input does not produce a bounded output. There are many similar examples.)

The foregoing discussion may be summarized as follows:

Asymptotically stable system ⇒ every bounded input produces
a bounded output

Unstable system ⇒ some bounded input produces
an unbounded output

Note that the implications go only in one direction. We may *not* conclude that a system for which every bounded input produces a bounded output is asymptotically stable. As we shall see in the next chapter, it is possible that some unstable state variables are not excited by the input. It is also not permissible to conclude that if some bounded input produces an unbounded output, the system is unstable. An ideal integrator, already cited, is an example of a stable system for which a bounded input (say a step function) produces an unbounded output (a ramp).

Example 4C Aircraft longitudinal motion The linear dynamic equations for the longitudinal motion of an aircraft were given in (2.36). Using the state and control definitions

$$x = \begin{bmatrix} \Delta u \\ \alpha \\ q \\ \theta \end{bmatrix} \qquad u = \delta_E$$

we obtain the dynamics and control matrices

$$A = \begin{bmatrix} X_u & X_\alpha & 0 & -g \\ Z_u/V & Z_\alpha/V & 1 & 0 \\ M_u & M_\alpha & M_q & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} X_E \\ Z_E/V \\ M_E \\ 0 \end{bmatrix} \tag{4C.1}$$

The resolvent for this system is

$$\Phi(s) = (sI - A)^{-1} = \begin{bmatrix} s - X_u & -X_\alpha & 0 & g \\ -Z_u/V & s - Z_\alpha/V & -1 & 0 \\ -M_u & -M_\alpha & s - M_q & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}^{-1} \tag{4C.2}$$

From the resolvent we obtain the characteristic polynomial

$$|sI - A| = s^4 + a_1 s^3 + a_2 s^2 + a_3 s + a_4 \tag{4C.3}$$

where $a_1 = -\dfrac{Z_\alpha}{V} - M_q - X_u$

$$\begin{aligned} a_2 &= \frac{Z_\alpha}{V} M_q - M_\alpha + X_u\left(\frac{Z_\alpha}{V} + M_q\right) - \frac{Z_u}{V} X_\alpha \\ a_3 &= -X_u\left(\frac{Z_\alpha}{V} M_q - M_\alpha\right) + X_\alpha\left(\frac{Z_u}{V} M_q - M_u\right) \\ a_4 &= g\frac{Z_u M_\alpha - Z_\alpha M_u}{V} \end{aligned} \tag{4C.4}$$

Contribution to the characteristic equation of the terms due to the change in speed (those with the subscript u) are usually quite small relative to the other terms. Thus, as an approximation, the characteristic polynomial is

$$\begin{aligned} |sI - A| &\approx s^2\left(s^2 - \left(\frac{Z_\alpha}{V} + M_q\right)s + \frac{Z_\alpha}{V} M_q - M_\alpha\right) \\ &= s^2[s^2 + 2\zeta_s\Omega_s s + \Omega_s^2] \end{aligned} \tag{4C.5}$$

The double pole at the origin is due to the translation of the aircraft as an ideal mass, and the quadratic factor is due to the rotation of the aircraft about the center of mass. This motion is seen to be that of a mass-spring-damper system with a damping factor ζ_s and a natural frequency Ω_s, and is called the *short-period* motion of the aircraft.

If Ω_s^2 and ζ_s are both positive, the poles of the short-period motion lie in the left half-plane and the short-period motion is stable. The aircraft in this case is said to be *aerodynamically stable.* Until very recently, it was the responsibility of the aerodynamicist to design the aircraft to ensure aerodynamic stability for all operating regimes of the aircraft. It is of course possible to stabilize an unstable aircraft by means of a properly designed control system, but the hardware (i.e., sensors and actuators) used to implement the control system must be extremely reliable—as reliable as the airframe itself. With the advent of multiply-redundant hardware, it is possible to achieve a very high degree of reliability, and it is now considered safe to operate aerodynamically unstable aircraft having suitable multiply redundant stability augmentation systems.

As an example we consider the numerical data for an actual aircraft, the AFTI-16 (a modified version of the F-16 fighter) in the landing approach configuration, as given in Table 4C.1 [6].

Using the data in Table 4C.1 we find that

$$\begin{aligned} \frac{Z_\alpha}{V} + M_q &= 2\zeta_s\Omega_s = 1.01 \\ \frac{Z_\alpha}{V} M_q - M_\alpha &= \Omega_s^2 = -1.1621 \end{aligned} \tag{4C.6}$$

Since $\Omega_s^2 < 0$, the aircraft is aerodynamically unstable in this regime, having poles at

$$s_1 = -1.695 \qquad \text{and} \qquad s_2 = 0.685 \tag{4C.7}$$

Table 4C.1 Aerodynamic parameters for AFTI-16 on landing approach

	$V = 139$ Kt		
$X_u = -0.0507$	$X_\alpha = -3.861$		$X_E = 0$
$Z_u/V = -0.00117$	$Z_\alpha/V = -0.5164$		$Z_E/V = -0.0717$
$M_u = -0.000129$	$M_\alpha = 1.4168$	$M_q = -0.4932$	$M_E = -1.645$

The short-period poles as given by (4C.7) are only approximate, since the effects of the speed changes have not been accounted for. To take these effects into account we must calculate the coefficients of the characteristic polynomial using (4C.4). For the data of Table 4C.1 the coefficients are found to be

$$a_1 = 1.0603$$

$$a_2 = -1.1154$$

$$a_3 = -0.0565$$

$$a_4 = -0.0512$$

Numerical solution of the characteristic equation yields the pole locations

$$s_1 = -1.705 \qquad s_2 = 0.724 \qquad s_{3,4} = -0.0394 \pm j0.200$$

We observe that the short-period poles s_1 and s_2, when speed changes are accounted for, are located very close to the approximate locations given in (4C.7). Another pair of poles (which are at the origin in the approximate analysis), with a natural frequency of $[(0.200)^2 + (0.0394)^2]^{1/2} = 0.204$ and a damping factor of 0.19, also appears due to speed changes. The motion due to these poles is known as *phugoid* motion and is manifest as a slight oscillation in altitude. (See Note 4.2.)

4.5 ROUTH-HURWITZ STABILITY ALGORITHMS

In the previous section we saw that the imaginary axes of the complex frequency plane (*the s plane*) separates the region of stability from the region of instability. If all poles lie in the left half-plane the system is asymptotically stable; otherwise the system is not asymptotically stable.

It is now a routine exercise for a digital computer to find the roots of a polynomial of very high degree. Before the advent of digital computers, however, testing the stability of a system by calculating the zeros of the characteristic equation was not practical. Methods were needed that did not require actual calculation of these roots. The earliest contribution to this problem was the Adams Prize Essay (1874-1877) of E. J. Routh[7] who developed a simple tabular algorithm by which it is possible to determine whether a given polynomial has all its roots in the left half-plane without finding the roots. A different algorithm was developed by A. Hurwitz[8] in 1895. And in 1962, P. C. Parks[9] showed that these algorithms could be derived by use of a stability theorem that M. A. Liapunov developed[10] in 1892-1907.

The algebraic criteria are derived in textbooks such as Schwarz and Friedland[4] on linear systems, and will not be repeated here. For convenience of the reader, the resulting algorithms are presented here without proof.

Table 4.2 Routh table

	1 a_1	a_2 a_3	a_4 a_5	$a_6 \cdots$ $a_7 \cdots$
$\alpha_1 = \dfrac{1}{a_1}$	$b_1 = a_2 - \alpha_1 a_3$	$b_2 = a_4 - \alpha_1 a_5$	$b_3 = a_6 - \alpha_1 a_7$	$\cdots\cdots$
$\alpha_2 = \dfrac{a_1}{b_1}$	$c_1 = a_3 - \alpha_2 b_2$	$c_2 = a_5 - \alpha_2 b_3$	$\cdots\cdots\cdots\cdots$	
$\alpha_3 = \dfrac{b_1}{c_1}$	$d_1 = b_2 - \alpha_3 c_2$	$\cdots\cdots\cdots\cdots$		
$\alpha_4 = \dfrac{c_1}{d_1}$	$\cdots\cdots\cdots\cdots$			
$\vdots$				

The characteristic polynomial of the system to be tested for stability is assumed to be of the form

$$D(s) = s^k + a_1 s^{k-1} + \cdots + a_{k-1}s + a_k$$

The Routh table corresponding to $D(s)$ is constructed as shown in Table 4.2. The first two rows are obtained by transcribing the coefficients of $D(s)$ in alternate rows as shown. Each succeeding row of the table is completed using entries in the two preceding rows, until there are no more terms to be computed. In the left margin are found a column of exactly k numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$ for a kth-order system. The theorem of the Routh algorithm is that the roots of $D(s) = 0$ lie in the left half-plane, excluding the imaginary axis, if and only if all the α's are strictly positive.

The Hurwitz criterion, which is equivalent to the Routh algorithm, is based on the construction of a $k \times k$ Hurwitz matrix

$$H = \underbrace{\left[\begin{array}{cccc} a_1 & a_3 & \cdots & \cdots \\ 1 & a_2 & \cdots & \cdots \\ \hline 0 & a_1 & a_3 & \cdots \\ 0 & 1 & a_2 & \cdots \\ \hline 0 & 0 & a_1 & \cdots \\ 0 & 0 & 1 & \cdots \end{array}\right]}_{k \text{ columns}} \left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\\1\end{array}}\right\} k \text{ rows}$$

The first two rows of H are formed from the coefficients of $D(s)$, with zeros used for a_{k+1} through a_{2k-1}. Each row following is obtained by shifting one step to the right the entries of the row two positions above, and padding the empty positions with zeros. This process is continued until the $k \times k$ matrix is completed. The stability theory based on the Hurwitz matrix is that the zeros of $D(s)$ are in the left half-plane, excluding the imaginary axis, if and only if the determinants

$$D_1 = a_1$$

$$D_2 = \begin{vmatrix} a_1 & a_3 \\ 1 & a_2 \end{vmatrix}$$

$$D_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ 1 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$D_k = |H|$$

are all strictly positive.

Example 4D Distillation column—continued A closed-loop control system for the distillation column of the previous example is proposed by making the change in the steam flow rate Δu_1 proportional to the error between Δz_1 and some desired set point value, say $\overline{\Delta z}$. Thus, as shown in Fig 4.6,

$$\Delta u_1(s) = u(s) = K[\overline{\Delta z} - \Delta z_1(s)] \tag{4D.1}$$

and from the analysis of Example 4A

$$u(s) = \frac{s(s - a_{11})(s - a_{22})}{b_{11}a_{32}a_{21}}\Delta z_1(s) \tag{4D.2}$$

The closed-loop transfer function $H_c(s)$ is obtained by substituting (4D.2) into (4D.1) and finding the ratio of Δz_1 to $\overline{\Delta z}$

$$H_c(s) = \frac{\bar{K}}{s(s - a_{11})(s - a_{22}) + \bar{K}} \tag{4D.3}$$

where $\bar{K} = b_{11}a_{32}a_{21}K$.

The characteristic polynomial of the closed-loop system is

$$s^3 - (a_{11} + a_{22})s^2 + a_{11}a_{22}s + \bar{K} \tag{4D.4}$$

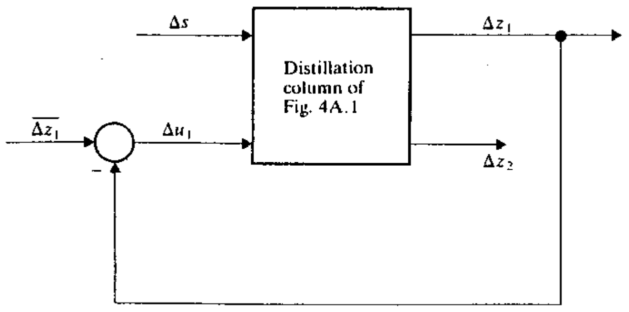

Figure 4.6 Single-loop control of distillation column.

Thus,

$$a_1 = -a_{11} - a_{22}$$
$$a_2 = a_{11}a_{22}$$
$$a_3 = \bar{K}$$

The Routh table for this example is

$$\begin{array}{c|cc}
 & 1 & a_2 \\
 & a_1 & \bar{K} \\
\hline
\alpha_1 = \dfrac{1}{a_1} & a_2 - \dfrac{\bar{K}}{a_1} & \\
\hdashline
\alpha_2 = \dfrac{a_1}{a_2 - \dfrac{\bar{K}}{a_1}} & \bar{K} & \\
\hdashline
\alpha_3 = \dfrac{a_2 - \dfrac{\bar{K}}{a_1}}{\bar{K}} & &
\end{array}$$

Thus, for stability of the closed-loop system we must have

$$a_1 > 0 \tag{4D.5}$$

$$a_2 - \frac{\bar{K}}{a_1} > 0 \quad \text{or} \quad \bar{K} < a_1a_2 \tag{4D.6}$$

$$\bar{K} > 0 \tag{4D.7}$$

The first condition is a requirement on the open-loop dynamics. From the data given about the process a_{11} and a_{22} are both negative, so (4D.5) is automatically satisfied. The second and third conditions are combined to give

$$0 < \bar{K} < a_1a_2 \tag{4D.8}$$

which means that the gain $\bar{K}$ (which is a negative feedback gain) must be positive—i.e., only negative feedback is permissible, and that $\bar{K}$ must be smaller than a fixed positive number.

The Hurwitz matrix for this example is

$$H = \begin{bmatrix} a_1 & \bar{K} & 0 \\ 1 & a_2 & 0 \\ 0 & a_1 & \bar{K} \end{bmatrix}$$

and the stability requirements are

$$D_1 = a_1 > 0$$
$$D_2 = a_1a_2 - \bar{K} > 0$$
$$D_3 = \bar{K}D_2 > 0$$

which are the same conditions as obtained using the Routh algorithm.

In the root-locus method to be studied in the next section we will be concerned with the variation of the closed-loop poles with the loop gain $\bar{K}$. By the methods to be explained more fully in that section, we find that the roots move from the open-loop poles to infinity. The open-loop poles occur at

$$s = 0$$
$$s = a_{11} \qquad (4D.9)$$
$$s = a_{22}$$

and the loci of the closed-loop poles have the appearance shown in Fig 4.7. One locus moves along the negative real axis, and the other two, after moving together, separate from the real axis and move to asymptotes at angles of ±60 degrees from the positive real axis. The gain $\bar{K}$ at which the loci cross the imaginary axis is the gain at which (4D.6) is an equality:

$$\bar{K} = a_1 a_2 \qquad (4D.10)$$

The frequency ω at which the crossing occurs is obtained by substituting $s = j\omega$ into (4D.4)

$$-j\omega^3 - \omega^2 a_1 + j\omega a_2 + \bar{K} = 0 \qquad (4D.11)$$

The real and the imaginary parts of (4D.11) must simultaneously be zero:

$$\omega^3 - \omega a_2 = 0 \qquad (4D.12)$$

$$-\omega^2 a_1 + \bar{K} = 0 \qquad (4D.13)$$

From (4D.12) we obtain $\omega = 0$ (corresponding to the open loop pole at the origin) and $\omega^2 = a_2$. From (4D.13) we obtain $\omega^2 = \bar{K}/a_1$. Since by (4D.10) the critical gain $\bar{K} = a_1 a_2$, the second expression for ω^2 is consistent with the first.

To obtain the "breakaway frequency" $s = -c_2$ at which the root loci join before leaving the real axis, we note that at that point, there is a double pole, so the characteristic equation must be

$$(s + c_1)(s + c_2)^2 = s^3 + (c_1 + 2c_2)s^2 + c_2(2c_1 + c_2)s + c_1 c_2^2$$

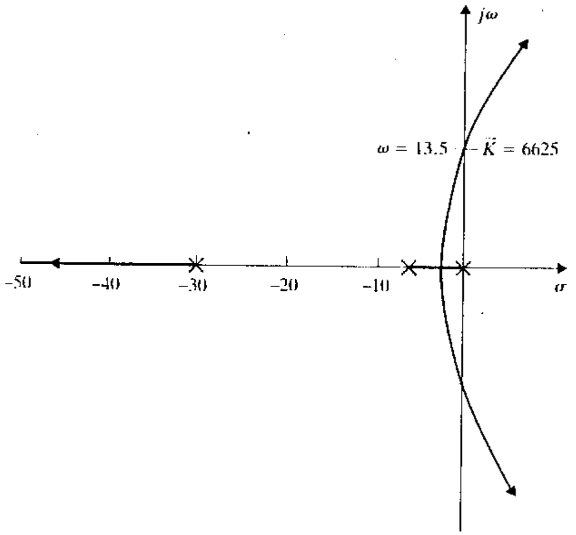

Figure 4.7 Root-locus for feedback of steam flow rate.

Thus we must have

$$a_1 = c_1 + 2c_2$$
$$a_2 = c_2(2c_1 + c_2)$$
$$\bar{K} = c_1 c_2^2$$

which can be solved simultaneously to give c_1, c_2, and $\bar{K}$. Another method of finding the breakaway frequency is given in discussion of the root locus method of the next section. Using the numerical data for the parameters of this process as given in Example 2G, namely

$$a_{11} = -30.3 \qquad a_{22} = -6.02$$

gives

$$a_1 = 36.32 \qquad a_2 = 182.4$$

Thus the gain at which the roots cross into the right half-plane is

$$\bar{K} = 6625$$

and the frequency at the crossing of the axis is

$$\omega = 13.5$$

The root loci separate from the real axis at

$$s = -c_2 = -2.84$$

and this occurs for a gain $\bar{K} = 247$.

4.6 GRAPHICAL METHODS

The algebraic tests of Routh and Hurwitz give the precise range(s) of parameter(s) for which a system is stable, and do not require the calculation of the closed-loop poles. They are most useful for testing whether a design is satisfactory but are not as convenient as some of the graphical methods (root-loci, Bode and Nyquist plots) for design purposes. Since frequency-domain *design* methods are not considered in this book, we will not dwell at length on these graphical methods, but refer the reader instead to one of the standard textbooks on the subject.[4, 11, 12] On the other hand, graphical representations can often serve as an aid to interpreting the design results that are obtained by state-space methods. For this reason, it is worth considering them at least briefly.

Except for the recent extensions to multivariable systems (as typified by the work of Rosenbrock and MacFarlane) the graphical methods are addressed to a single-loop system having a return-difference function

$$T(s) = 1 + KG(s) \qquad (4.29)$$

where K is a scalar gain (the "loop gain") and $G(s)$ is a rational function known as the "open-loop" transfer function. A return difference of the form of (4.29) arises directly in the systems shown in Fig. 4.8, but it is always possible to manipulate the block diagram of a system so that the characteristic equation of the system appears in this form for any system parameter represented by K. The graphical methods are devices for elucidating the dynamic characteristics of

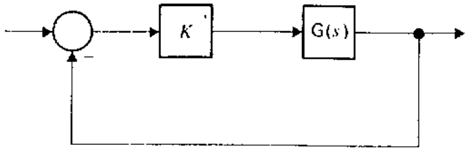

Figure 4.8 Single-loop feedback system return difference T(s) = 1 + KG(s).

a system having an open-loop transfer function G(s) as the loop gain K is varied.

Root-locus method The root-locus method, developed by Evans[13] in 1948 is simply a plot of the locations in the complex plane of the roots of $F(s) = 0$, (i.e., the poles of the closed-loop system) as the loop gain is varied. The open-loop transfer function is assumed to be a rational function of s, i.e.,

$T = 1 - KC\frac{n}{d} \overset{!}{=} 0$

$d = k \cdot C\, n \Rightarrow k \to 0 \quad d \to 0$

$k \to \infty \quad C \cdot n = 0$

$$G(s) = C\frac{\prod_{i=1}^{n_z}(s - z_i)}{\prod_{i=1}^{n_p}(s - p_i)} = C\frac{n}{d} \qquad (4.30)$$

where C is a real constant, z_i $(i = 1, \ldots, n_z)$ are the open-loop zeros and p_i $(i = 1, \ldots, n_p)$ are open-loop poles. If desired the constant C can be absorbed in the loop gain, by defining $\bar{K} = KC$. It is seen that as $K \to 0$, the closed-loop poles, which are the roots of (4.29), tend to the open-loop poles p_i. On the other hand, as the gain K tends to infinity, the closed-loop poles tend to the open-loop zeros. If G(s) is a proper rational function, however, there are fewer open-loop zeros than open-loop poles. Since the number of closed-loop poles does not change as K is varied, where do the closed-loop poles go that do not go to the open-loop zeros? They go to infinity. The manner in which they go to infinity depends on the excess of poles over zeros. Imagine viewing the complex plane from a great distance. From this vantage point all the poles and zeros appear to be at the origin and G(s) looks like $1/s^{(n_p - n_z)}$. Thus, from this vantage point the root-locus equation looks like

$$1 + K\frac{1}{s^e} = 0 \qquad e = n_p - n_z$$

or

$$s^e + K = 0 \qquad (4.31)$$

where e is the excess of poles over zeros in the open-loop transfer function. Thus, as K becomes very large, the root loci that do not terminate at the open-loop zeros tend to infinity in the same way as the solutions of (4.31) tend to infinity, namely as the eth roots of $-K$. Since there are exactly e such roots at equal angles around a circle, these lines are the asymptotes of the root loci that tend to infinity. Figure 4.9 illustrates the asymptotic behavior for large values of loop gain K of those branches of the root loci that tend to infinity.

The asymptotic behavior of the root loci can be rationalized another way: we can say that the number of poles and zeros are always equal and that the root loci always go from the poles to the zeros, but that those zeros (e in number) which are not in the finite part of the s plane lie at infinity.

$T(s) = (1 + KG(s)) \overset{!}{=} 0$

→ this is what I'm interested in, in order to invert the matrix T(s) (4.17)

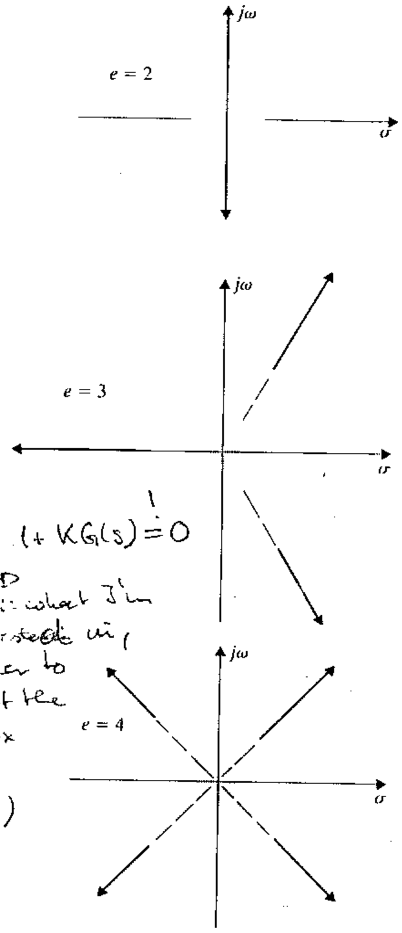

Figure 4.9 Asymptotes of root loci for several values of excess poles.

Figure 4.9 shows that whenever the excess of poles over zeros is greater than 2, the root loci must eventually cross the imaginary axis into the right half of the s plane. Consequently, no system having an excess of two or more can be stable for all values of gain. Since the excess is two or more in most practical systems, the implication is that in practice there is a finite upper limit to the loop gain. The ratio of the loop gain at which a system is designed to operate to the gain at which it becomes unstable (expressed logarithmically), is known as the *gain margin* of the system. Gain margin is an important consideration in systems in which the loop gain may change during the operating lifetime of the process (due possibly to aging of components).

It must not be inferred that an excess of poles over zeros of 2 or less guarantees stability, since the root loci may cross into the right half-plane for finite values of gain, and remain there (when $e = 2$) or cross back into the left half-plane when K becomes large enough.

The root loci cross into the right half-plane through the imaginary axis. Except in the trivial case when the crossing of the imaginary axis is through the origin ($s = 0$), the loci cross the imaginary axis at points $s = \pm j\omega$. This means that the nature of the unforced dynamic response changes from being sinusoidal with slight positive damping to sinusoidal with slight negative damping. At the dividing line, the response becomes purely sinusoidal: that of a harmonic oscillator. The gains that cause the root loci to cross the imaginary axis and the frequencies at which they occur are significant parameters in the root-locus method. These frequencies and the gains at which they occur can be obtained by setting $s = j\omega$ into the characteristic equation and equating the real parts and the imaginary parts to zero. The calculation is facilitated by the fact that the gains at which the crossings occur also make exact equalities out of the inequalities that result from the Routh (or Hurwitz) algorithm. This was already illustrated in Example 4C.

The basic rules for drawing the root loci, as already illustrated are

The loci move continuously from the open-loop poles to the open-loop zeros or to infinity.

The loci approach infinity at lines which are in the direction of the eth roots of $-1 = e^{-j\pi}$ from the origin.

Many other rules for constructing root-locus plots are obtained from the basic root locus equation

$$1 + K\mathrm{G}(s) = 1 + K\frac{\prod_{i=1}^{n_z}(s - z_i)}{\prod_{i=1}^{n_p}(s - p_i)} = 0 \tag{4.32}$$

Each factor $s - z_i$ or $s - p_i$ is represented in the complex plane by a vector (a "phasor" in electrical engineering parlance) from the zero z_i or pole p_i to the point x. If s is a point on the root locus, then (4.32) must hold, i.e.,

$$\frac{\prod_{i=1}^{n_z}(s - z_i)}{\prod_{i=1}^{n_p}(s - p_i)} = -\frac{1}{K} \tag{4.33}$$

which means that the product of the lengths of all the vectors from the zeros to the point s divided by the product of the lengths of all the vectors from the poles must come out to be $1/K$ and that the sum of all the angles of the vectors from the zeros to the point s minus the angles of the vectors from the poles must add to $-180°$. Rules obtained from this general principle, such as the directions of departure of the loci from the open-loop poles or of arrival at the open-loop zeros, can be found in various textbooks [11, 12] that concentrate on frequency-domain analysis.

The points at which the loci leave the real axis are known as *breakaway points*. To find these points consider the root-locus equation

$$1 + K\frac{\mathrm{N}(s)}{\mathrm{D}(s)} = 0$$

Multiply by the open-loop denominator to obtain the characteristic equation

$$\mathrm{P}(s) = \mathrm{D}(s) + K\mathrm{N}(s) = 0$$

The breakaway points are those at which $\mathrm{P}(s)$ has a multiple root. Thus if $s = a$ is a breakaway point we can write

$$\mathrm{P}(s) = (s - a)^2\mathrm{P}_1(s)$$

where $\mathrm{P}_1(s)$ is a polynomial of degree $k - 2$ obtained by multiplying all the factors except the factor $(s - a)^2$ arising because of the multiple root. ($\mathrm{P}_1(s)$ could conceivably have more than a double root at $s = a$, in which case $\mathrm{P}_1(s)$ could contain other $(s - a)$ factors. This is of no concern.)

The derivative of $\mathrm{P}(s)$ with respect to s is

$$\mathrm{P}'(s) = 2(s - a)\mathrm{P}_1(s) + (s - a)^2\mathrm{P}_1'(s)$$

Thus, at $s = a$

$$\mathrm{P}'(a) = 2(a - a)\mathrm{P}(a) + (a - a)^2\mathrm{P}'(s) = 0$$

In other words at a breakaway point $s = a$ the derivative of $\mathrm{P}(s)$ is zero. If $s = a$ is any other point on the root locus, we can write

$$\mathrm{P}(s) = (s - a)\mathrm{P}_1(s)$$

where $\mathrm{P}_1(a) \neq 0$. It thus follows that

$$\mathrm{P}'(s) = \mathrm{P}_1(s) + (s - a)\mathrm{P}_1'(s)$$

and hence

$$\mathrm{P}'(a) = \mathrm{P}_1(a) \neq 0$$

Thus the breakaway points on the real axis are distinguished from all other points on the real axis by the property that the derivative of $\mathrm{P}(s)$ goes to zero at the breakaway points. Note that $\mathrm{P}'(s)$ is a polynomial of degree $k - 1$ in s, and hence finding its roots poses a numerical problem only slightly less complicated than finding the roots of $\mathrm{P}(s)$ themselves. If the latter are to be found with the aid of a computer, it is hardly worth the trouble of finding the breakaway points by solving for the roots of

$$\mathrm{P}'(s) = 0 \tag{4.34}$$

The reader may wish to verify that $\mathrm{P}'(a) = 0$ at the breakaway points in the previous examples.

Nyquist diagram The earliest graphical method investigating the stability of linear systems was developed by H. Nyquist in 1932[1] and is based on the polar plot of the loop transmission transfer function. To understand Nyquist's method, recall that the condition for *instability* is that

$$1 + K\mathrm{G}(s) = 0 \qquad \text{or} \qquad \mathrm{G}(s) = -\frac{1}{K} \tag{4.35}$$

for some value of s in the *right* half of s plane. Conversely, if there does not exist a value s in the right half-plane for which $\mathrm{G}(s) = 1/K$, then we are assured that the system is stable.

For every point s in the right half-plane, there is a point $z = \mathrm{G}(s)$ in the z plane. (If $\mathrm{G}(s)$ is a rational function, then for each value of s there is *only* one value of $z = \mathrm{G}(s)$.) Thus the function $\mathrm{G}(s)$ "maps" the right half of the s plane into some region of the z plane. (Since $\mathrm{G}(s)$ is a continuous function and the right half-plane is a contiguous region, the map of the right half-plane by the function $\mathrm{G}(s)$ is also contiguous.) If the region of the z plane that is the map of the right half of the s plane under the function $\mathrm{G}(s)$ covers the point $-1/K$, the system is unstable; if the map does not cover the point $-1/K$, the system is stable. The two cases are depicted in Fig. 4.10.

The basic principle of the method of Nyquist is thus to determine whether or not the map of the right half of the s plane created by the function $\mathrm{G}(s)$ covers the point $-1/K$. Is it necessary to find $\mathrm{G}(s)$ for every s in the right half-plane? The answer, fortunately, is no. There is a theorem in complex variables which asserts that the map of the boundary of a region in the s plane is the boundary of the map of that region in the z plane. Thus, to find the map of the right half of the s plane under $\mathrm{G}(s)$ we need only find the map of the *boundary* of the right half of the s plane. The entire right half of the s plane is unbounded, of course. We get around that difficulty by finding the map of the large semicircular region bounded by the imaginary axis between $-j\Omega$ and $j\Omega$ in the semicircle of radius Ω in the right half-plane. Then we pass to the limit as $\Omega \to \infty$. If we are dealing with a proper rational function (i.e., the numerator degree is lower than the denominator degree) then as $\Omega \to \infty$, $\mathrm{G}(s) \to 0$, so the whole semicircle maps into just one point: $\mathrm{G}(s) = 0$.

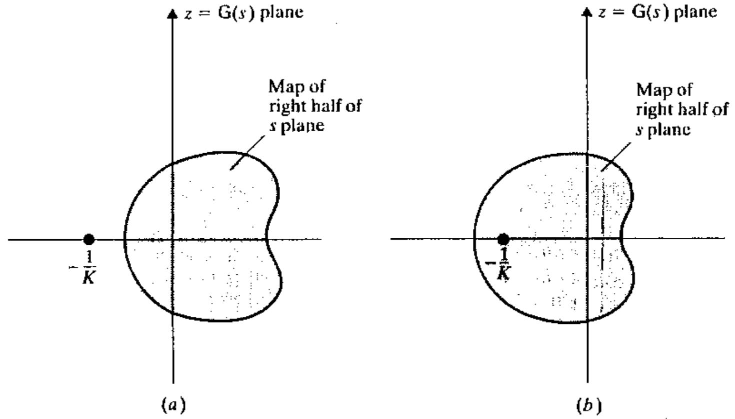

Figure 4.10 System with return difference $\mathrm{F}(s) = 1 + K\mathrm{G}(s)$ is unstable if map of right half of s plane covers the point $-1/K$. (a) Stable system; (b) Unstable system.

To construct the map we start an excursion at the origin O and "walk" up the imaginary axis to the point A as shown in Fig. 4.11(a) at which $s = j\Omega$. The map of this portion of the imaginary axis may have the appearance of the curve $O' - A'$ in the $\mathrm{G}(s)$ plane as shown in Fig. 4.11(b). Then we walk around the semicircle to the point B. The map of the semicircle $A - B$ is the arc $A' - B'$. Finally we return to the origin O upward along the imaginary axis along the path $B - O$ and obtain the corresponding arc $B' - O'$ in the $\mathrm{G}(s)$ plane. The map of the entire right half-plane is obtained by letting $\Omega \to \infty$ which has the effect of shrinking the arc $A' - B'$ to a single point. Since we know that the semicircular arc maps into just one point, there is no need to bother with that arc. It is enough just to walk up the imaginary axis.

The map of the imaginary axis separates the map of the right half of the s plane from the map of the left half of the s plane. It is necessary of course to know which points on the $\mathrm{G}(s)$ plane correspond to the points on the right half of the s plane, and which correspond to points on the left half of the s plane. We are aided in this process by the fact that the transformation $z = \mathrm{G}(s)$ is "conformal": angles are locally preserved.[14] Thus if we take our excursion

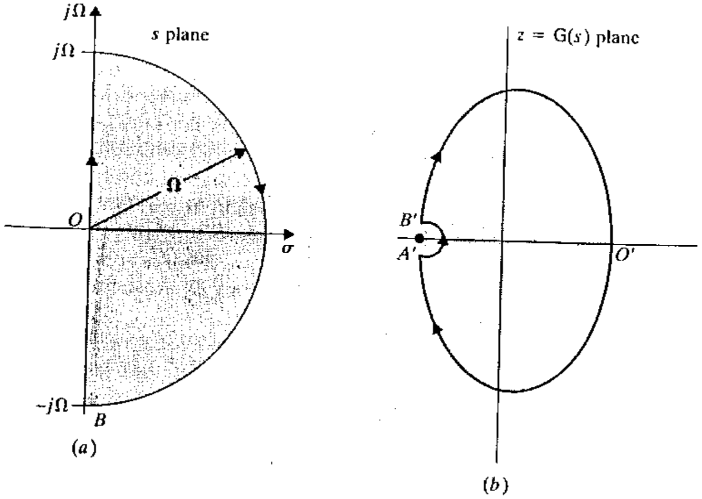

Figure 4.11 How to map right half of s plane into $z = \mathrm{G}(s)$ plane. (a) Semicircle approximates all of right half-plane; (b) Map of semicircle.

along the imaginary axis with our right hand extended so that it lies over the right half-plane, the corresponding excursion over the map of the imaginary axis with the right hand outstretched is over the map of the right half-plane. This principle is sufficient to identify the map of the right half-plane in all cases, and is equivalent to Nyquist's "encirclement rule" which we shall give later on.

In Figs. 4.10 and 4.11 we drew the maps of the right half-plane (an infinite region) as a finite region, because the entire right half-plane outside a semicircular arc shrinks down to a single point. But what happens if G(s) has poles on the imaginary axis? In that case, of course, the map of the region near a pole will result in very large values of z. Since our excursion along the imaginary axis is not permitted, we might consider an excursion along a line in the left half-plane parallel to and slightly left of the imaginary axis. This places the imaginary axis itself into the right half-plane—the region of instability which is where our previous classification of the region of instability would rightfully place it. On the other hand one might, with some justification, argue that a physical open-loop system is bound to have some damping present and hence that the open-loop poles are near but not exactly upon the imaginary axis. This means that an excursion up the imaginary axis is permitted, and that the poles encountered on the excursion are to our left. Each approach will result in a different Nyquist diagram. But there is no practical difference, because only the part of the Nyquist diagram that is remote from the open-loop poles is needed to assess the stability of a system. We adopt the approach of keeping the imaginary axis poles to our right. Thus, suppose for example, the loop transmission G(s) has a pole at the origin and a pair of complex poles on the imaginary axis at $\omega = \omega_c$ (as well as poles and zeros elsewhere in the s plane). As we proceed along a path parallel to and near the imaginary axis, starting on the real axis and going upward, we find that the map starts with a large real number and then rapidly becomes a large complex number with phase angle of nearly 90°. As the excursion continues upward the map of the line continues to evolve in accordance with the total constellation of poles and zeros until the line brings us near the pole and $j\omega_c$. In the vicinity of this pole the phase angle goes from a large number B at a phase angle of +90° to a large number C at a phase angle of zero to a large number D at a phase angle of −90° and then to zero as $\omega \to \infty$. The mirror image of the contour shown in Fig. 4.12 is the map of the lower half of the line parallel to the imaginary axis. As the line in the s plane approaches the imaginary axis the points A', B', and D' move toward $\pm\infty$ as indicated.

Although we have been concerned with the map of the entire right half of the s plane, it is apparent that the boundary of the map produced by the imaginary axis is usually the one feature of the map that is needed to determine whether or not a system is stable. If the point $-1/K$ is covered by the map of the right half of the s plane, the map of the imaginary axis "encircles" the point $-1/K$ in a clockwise direction as ω increases from 0 to ∞. But if the map does not cover the point $-1/K$, the map of the imaginary axis does not encircle the point $-1/K$ in a clockwise direction. Thus, in most cases, only the map of the imaginary axis is drawn and this curve is called the Nyquist diagram. The customary stability criterion is thus familiarly stated as follows:

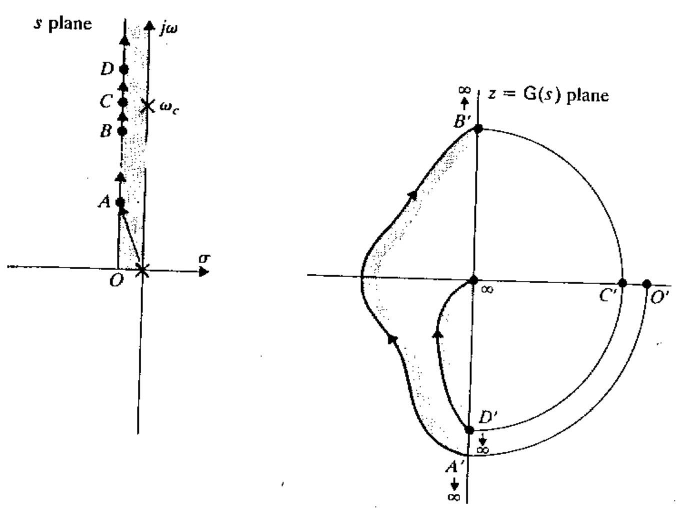

Figure 4.12 Nyquist diagram for transmission with poles on imaginary axis.

> **Nyquist stability criterion** A system having a return difference $1 + KG(s)$ is stable if and only if the Nyquist diagram, i.e., the map of the imaginary axis, does not encircle the point $-1/K$ in the clockwise direction.

It must be noted that the encirclement test must be performed very carefully when the loop transmission G(s) itself has poles or zeros in the right half-plane, as discussed in various texts on complex variables and systems.[14] If G(s) has poles and/or zeros in the right half-plane it is safer to map the entire right half-plane by G(s) and check whether or not it covers the point $-1/K$.

The behavior of the Nyquist plot as $\omega \to 0$ depends on the order of the pole at the origin. If there is no pole at the origin G(0) is finite and is a real number. It is positive unless, for some perverse reason, the dc gain G(0) is defined to be negative. If there is a simple pole at the origin then as $\omega \to 0$, $G(j\omega) \to C/j\omega$ which tends to infinity in magnitude and −90° in phase. Similarly if there is a double pole at the origin then, as $\omega \to 0$, $G(j\omega) \to C/(j\omega)^2 = -C/\omega^2$ which tends to infinity in magnitude and −180° in phase. And so forth. The order of the pole at the origin is known as the system "type" and, as will be discussed in Sec. 4.7, governs the ability of the system to track an input in the form of a polynomial time function without steady state error.

As $\omega \to \infty$, the behavior of the Nyquist plot depends on the excess of poles over zeros. If the excess is one, the plot approaches the origin along the negative imaginary axis because the G(s) behaves as $C/j\omega \to 0\angle{-90°}$. If the excess is two, the Nyquist plot approaches zero along the negative real axis because G(s) behaves as $C/(j\omega)^2 \to 0\angle{-180°}$. And so forth.

For a system of high order it is possible for the Nyquist diagram to have the appearance shown in Fig. 4.13 in which the map of the imaginary axis of the s plane crosses the real axis of the G(s) plane several times. It is not immediately obvious which of the enclosed regions are maps of portions of the right half of the s plane and which are maps of the left half. The rule about walking up the map of the imaginary axis with the right hand outstretched is helpful in this case. Following that rule we see that regions ②, ③, and ⑤ belong to the right half-plane but regions ① and ④ belong to the left half-plane. This means that as K is increased ($-1/K \to 0$ along the negative real axis) the system is stable until $-1/K$ crosses into region ③ when the system becomes unstable. It remains unstable until K is raised sufficiently to make $-1/K$ fall into region ④, which is a region in which the system is stable. It remains stable until K is further increased to bring $-1/K$ into region ⑤, when the system again becomes unstable and remains so as $K \to \infty$. If the gain K is chosen to put $-1/K$ in

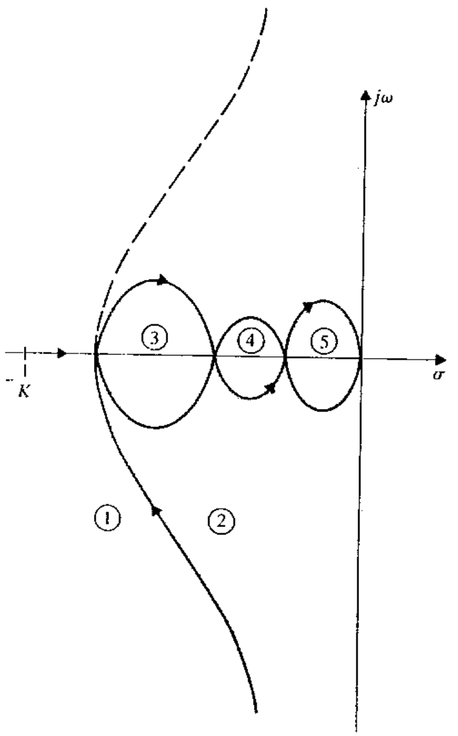

Figure 4.13 Nyquist diagram of a conditionally stable system. (System is stable if $-1/K$ is in regions 1 or 4.)

region ④ the system is said to be *conditionally stable.* A conditionally stable system is generally undesirable because of the danger that a reduction of gain as well as an increase can make the system unstable. Sometimes there is no way to avoid conditionally stable systems, but it is often possible to design a compensator to shape the Nyquist diagram to avoid having conditional stability. Methods that can be used to accomplish the required shaping are discussed in textbooks on frequency-domain methods of control system design.[11, 12]

The Nyquist diagram is used not only to assess the stability of a system (by determining whether the map of G(s) covers the point $-1/K$, but also to investigate system "robustness" which is a measure of how much the system can change without becoming unstable. The further away the point $-1/K$ is from the map of the right half-plane, the more the system transmission (i.e., the map of the right half-plane under G(s)) can change without endangering stability. Hence it is desirable that this distance be substantial. A quantitative measure of this distance is the *gain margin* as discussed in Sec. 4.9 which deals with robustness in general.

Bode plots The Nyquist diagram can be regarded as a polar plot of the magnitude and phase of G(s) when $s = j\omega$, that is, a polar plot of the magnitude and phase of G($j\omega$) with the frequency ω serving as a parameter. The same information can be presented in a pair of plots: one of the magnitude and the other of the phase of G($j\omega$), each as a function of frequency ω. These are known as the Bode plots of G(s). In particular let

$$G(j\omega) = |G(j\omega)|\, e^{j\theta_G(\omega)}$$

where $|G(j\omega)|$ and $\theta_G(\omega)$ are known as the *magnitude* and *phase* functions of the loop transmission G(s). Instead of plotting $|G(j\omega)|$ it is customary to plot

$$D(\omega) = 20 \log_{10} |G(j\omega)|$$

Regardless of the units of G(s), the units of $D(\omega)$ are invariably *decibels* (abbreviated dB). Since there is no physical significance to the logarithm of a quantity that is not dimensionless, i.e., the ratio of two physical variables of the same type (e.g., voltage out/voltage in, etc.) it is not strictly proper to use the decibel notation unless G(s) is a dimensionless ratio, i.e., unless the input and the output are the same physical type. But this improper usage is universally condoned and accepted.

The plot of $D(\omega)$ vs. ω is known as the Bode amplitude plot and the plot of $\theta_G(\omega)$ vs. ω is known as the Bode phase plot.

The Bode plot for a transfer function that has only real poles and zeros is particularly easy to construct graphically. In particular, consider a system having a loop transfer function

$$G(s) = G_0 \frac{\left(1 + \frac{s}{z_1}\right) \cdots \left(1 + \frac{s}{z_l}\right)}{\left(1 + \frac{s}{p_1}\right) \cdots \left(1 + \frac{s}{p_k}\right)} \qquad (4.36)$$

This form of G(*s*) is especially convenient for Bode plots because each factor in the numerator and the denominator is unity at $s = 0$ and hence the dc gain, $G(0) = G_0$ is explicitly exhibited. When $s = j\omega$

$$G(j\omega) = G_0 \frac{\left(1+\frac{j\omega}{z_1}\right)\cdots\left(1+\frac{j\omega}{z_l}\right)}{\left(1+\frac{j\omega}{p_1}\right)\cdots\left(1+\frac{j\omega}{p_k}\right)}$$

and hence

$$|G(j\omega)| = |G_0| \frac{\left[1+\left(\frac{\omega}{z_1}\right)^2\right]^{1/2}\cdots\left[1+\left(\frac{\omega}{z_l}\right)^2\right]^{1/2}}{\left[1+\left(\frac{\omega}{p_1}\right)^2\right]^{1/2}\cdots\left[1+\left(\frac{\omega}{p_k}\right)^2\right]^{1/2}} \tag{4.37}$$

Thus

$$D(\omega) = 20\log|G_0| + 10\log\left[1+\left(\frac{\omega}{z_1}\right)^2\right] + \cdots + 10\log\left[1+\left(\frac{\omega}{z_l}\right)^2\right]$$
$$-10\log\left[1+\left(\frac{\omega}{p_1}\right)^2\right] - \cdots - 10\log\left[1+\left(\frac{\omega}{p_k}\right)^2\right] \tag{4.38}$$

and (for $G_0 > 0$)

$$\theta_G(\omega) = \tan^{-1}\left(\frac{\omega}{z_1}\right) + \cdots + \tan^{-1}\left(\frac{\omega}{z_l}\right) - \tan^{-1}\left(\frac{\omega}{p_1}\right) - \cdots - \tan^{-1}\left(\frac{\omega}{p_k}\right) \tag{4.39}$$

(If the dc gain G_0 is negative, a 180° phase shift must be added to (4.39).) These results may be interpreted as follows: (Fig. 4.14.)

The log-magnitude plot $D(\omega)$ is the sum of the log-magnitude plot of each contributing factor and the phase plot is the sum of the phase plots of each contributing factor.

With increasing frequency, the contribution of a zero is an increase in both the log-magnitude and the phase; the contribution of a pole is a decrease† in both log-magnitude and phase.

The contribution of a typical zero or pole is shown in Fig. 4.12. It is seen that at the frequency $\omega = z_i$ or $\omega = p_i$ the magnitude is exactly twice its value at dc ($\omega = 0$) and the phase shift is exactly 45°. As the frequency is further increased

† The phase relation is valid only when the contributing pole or zero is in the left half plane, that is, p_i or z_i is positive. If the zero or pole is in the *right* half plane, then z_i or p_i is negative, and the phase contribution is opposite. Bode[2] has called such poles or zeros *nonminimum phase*. Nonminimum phase *poles* are indicative of an unstable open loop system, of course. The effect of open loop zeros is more subtle, however, and is discussed in greater detail in Note 4.7.

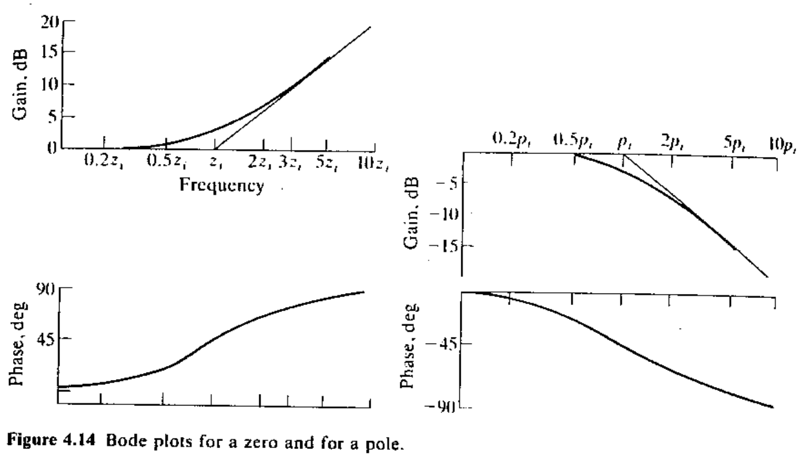

Figure 4.14 Bode plots for a zero and for a pole.

(*a*) $G(s) = 1 + \frac{s}{z_i}$ (*b*) $G(s) = \frac{1}{1 + (s/p_i)}$

the contributions to log-magnitude plots are

$$D_i(\omega) \to 10\log\left(\frac{\omega}{z_i}\right)^2 = 20\log\left(\frac{\omega}{z_i}\right) \qquad \text{for a zero}$$

$$D_i(\omega) \to -10\log\left(\frac{\omega}{p_i}\right)^2 = -20\log\left(\frac{\omega}{p_i}\right) \qquad \text{for a pole}$$

Thus, if a logarithmic frequency scale is used D_i is asymptotic to a line having a slope of 20 (dB) for each tenfold increase in frequency, that is, "20 dB per decade." The slope is positive for a zero, and negative for a pole. The asymptote intersects the logarithmic frequency axis at $\omega = z_i$ or $\omega = p_i$. At these frequencies, known as the "corner" frequencies, the exact gain is $\pm 10 \log 2 = 3.010$ dB, so these are also known as the "3 dB" frequencies.

The phase contribution from each factor tends to ±90°. (Positive for a zero; negative for a pole.)

The log-magnitude and phase curves for the overall system are obtained by simply adding the curves of each contributing factor. Thus, for example, the log-magnitude and phase curve for

$$G(s) = \frac{(1+s)\left(1+\frac{s}{5}\right)}{\left(1+\frac{s}{2}\right)\left(1+\frac{s}{10}\right)\left(1+\frac{s}{20}\right)}$$

has the appearance shown in Fig. 4.15. The maximum deviation from the straight line approximation is 3 dB.

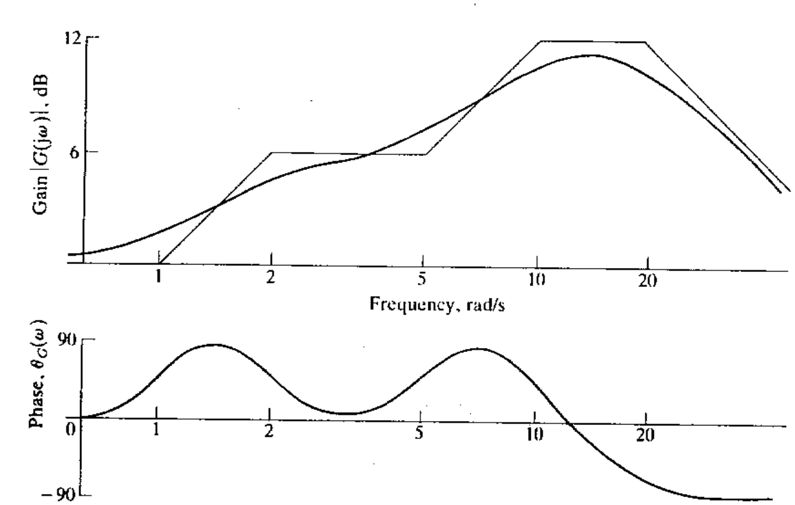

Figure 4.15 Bode plot for

$$G(s) = \frac{(s+1)(s/5+1)}{(s/2+1)(s/10+1)(s/20+1)}$$

A pole or a zero at the origin is treated slightly differently, because the log magnitude is not finite as $\omega \to 0$. A zero at the origin means that $D(\omega) \to -\infty$; a pole at the origin means that $D(\omega) \to +\infty$. These are the plots for a corner frequency of 0; in other words a zero at the origin contributes an increasing log-magnitude line with a constant slope of +20 dB/decade; a pole at the origin contributes a decreasing log-magnitude line with a slope of −20 dB/decade. Each passes through 0 dB at $\omega = 1$. The phase angle due to a pole is a constant of −90 degrees and the phase angle contribution of a zero is a constant +90 degrees.

The Bode plots for a transfer function G(s) that has complex poles or zeros is more complicated, because the straight-line approximation as illustrated in Fig. 4.13 is not applicable since a transfer function with a complex-conjugate pair of poles will include a factor of the form

$$G_i(s) = \frac{1}{1 + 2\zeta(s/\omega_0) + (s/\omega_0)^2} \tag{4.40}$$

The log-magnitude and phase functions corresponding to (4.40) are

$$D_i(\omega) = -20\log\left[1 + (4\zeta^2 - 2)\left(\frac{\omega}{\omega_0}\right)^2 + \left(\frac{\omega}{\omega_0}\right)^4\right]^{1/2}$$
$$\theta_i(\omega) = \tan^{-1}\left(\frac{-(\omega/\omega_0)\zeta}{1 - (\omega/\omega_0)^2}\right) \tag{4.41}$$

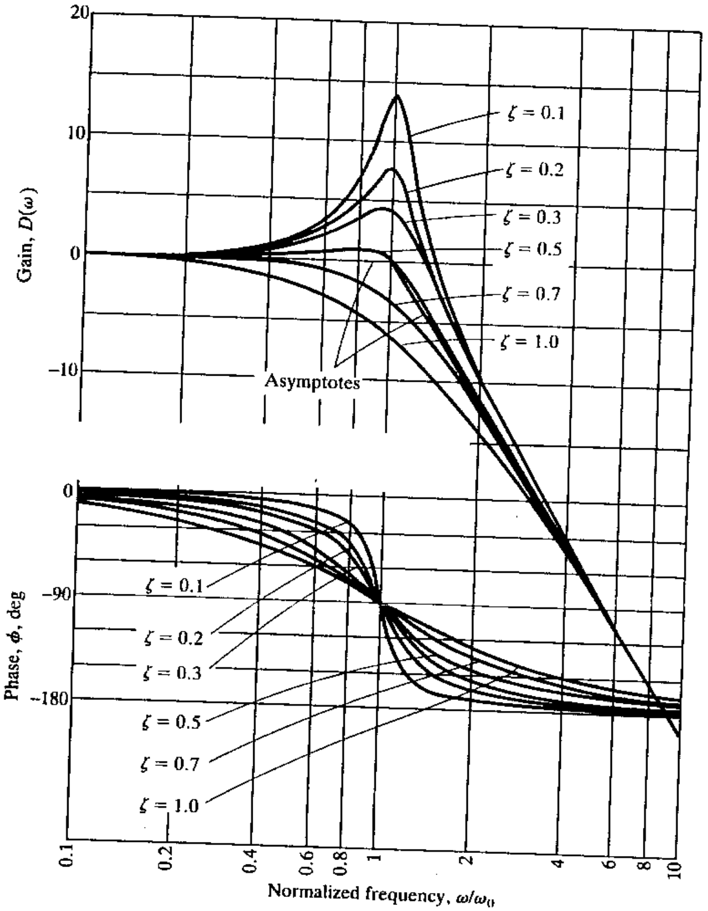

Figure 4.16 Bode plots for second-order system

$$G(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

The log-magnitude and phase curves vs. normalized frequency ω/ω_0 are shown in Fig. 4.16 for various damping factors ranging from $\zeta = 0.1$ (lightly damped) to $\zeta = 1.0$. It is seen that as the damping becomes very small, the log-magnitude becomes very large in the vicinity of the natural frequency $\omega \approx \omega_0$, and the phase shift rapidly changes from angles close to zero to angles close to 180°, crossing through exactly 90° at $\omega/\omega_0 = 1$.

The frequency ω_r at which the peak (often called a *resonance peak*) in the log-magnitude plot occurs can be found by taking the derivative of (4.41) with

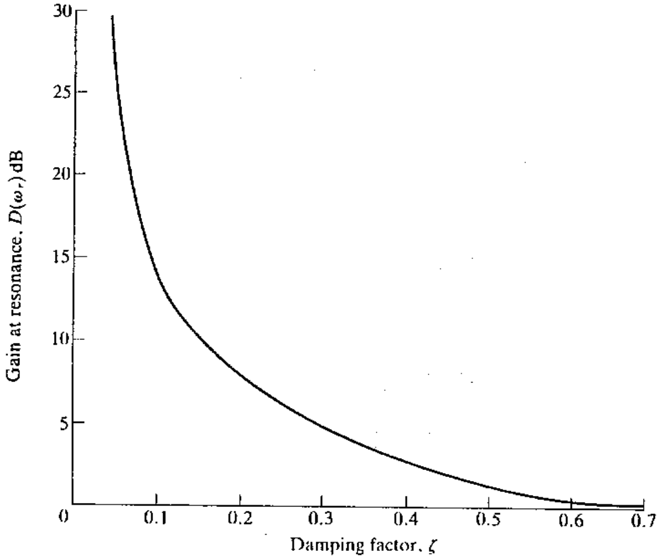

Figure 4.17 Gain at resonance of

$$G(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \qquad (\omega_r = \sqrt{1 - 2\zeta}\omega_0)$$

respect to $x = (\omega/\omega_0)^2$ and setting it to zero. It is found that this frequency is given

$$\omega_r = \sqrt{1 - 2\zeta^2}\omega_0$$

which means that there is no resonance peak for $\zeta > 1/\sqrt{2}$. For $\zeta < 1/\sqrt{2}$, the gain at resonance is given by

$$D_i(\omega_r) = -20 \log [4\zeta^2(1 - \zeta^2)] \qquad \text{(dB)} \tag{4.42}$$

a graph of which is shown in Fig. 4.17.

Example 4E Hydraulically actuated gun turret (continued) In Example 4B we found the transfer function between the input u and the output angle θ of the hydraulically actuated gun turret. Using the numerical data given in Example 2D for azimuth control:

$$K_v = 94.3 \qquad L_v = 1.0 \qquad J = 7900 \qquad K_m = 8.46 \times 10^6$$

$$\omega_m = 45.9 \qquad \text{and} \qquad K_{\Delta p} = 6.33 \times 10^{-6}$$

we find numerically that

$$H(s) = \frac{100\,980}{s(s^3 + 140.2s^2 + 10\,449s + 100\,980)} \tag{4E.1}$$

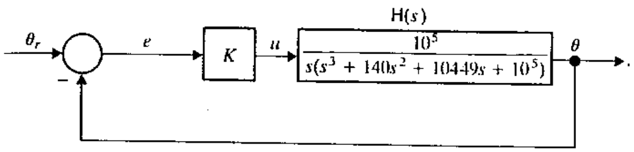

Figure 4.18 Closed-loop control of gun turret.

The root-locus equation for the closed-loop process (see Fig. 4.18) is

$$1 + \frac{100\,980K}{s(s^3 + 140.2s^2 + 10\,449s + 100\,980)} \tag{4E.2}$$

The open-loop poles of the process are at $s = 0$ and at the roots of the cubic factor $s^3 + 140.2s^2 + 10\,449s + 100\,980$. The latter are found numerically to be located at

$$s = -11.2$$

$$s = -64.5 \pm j69.6$$

Since there are four poles and no (finite) zeros of the transfer function, the root loci all go to ∞ parallel to lines at ±45° and ±135° angles from the real axis.

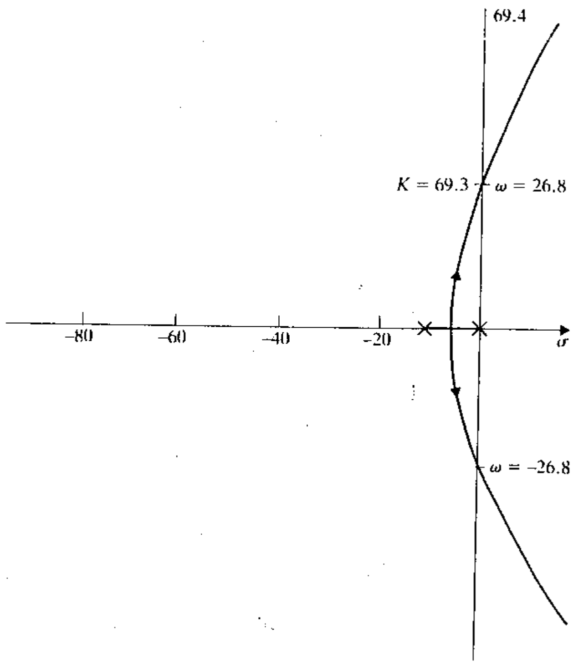

Figure 4.19 Root-locus plot for feedback control of hydraulically actuated gun turret.

To find the frequency and gain for crossings of the imaginary axis we set $s = j\omega$ in the characteristic equation

$$s^4 + 140.2s^3 + 10\,449s^2 + 100\,980s + 100\,980K = 0 \qquad (4E.3)$$

with $s = j\omega$ this becomes

$$\omega^4 - j140.2\omega^3 - 10\,449\omega^2 + j100\,980\omega + 100\,980K = 0$$

or, on equating the real and the imaginary parts to zero,

$$\omega^4 - 10\,449\omega^2 + 100\,980K = 0 \qquad (4E.4)$$

$$-140.2\omega^3 + 100\,980\omega = 0 \qquad (4E.5)$$

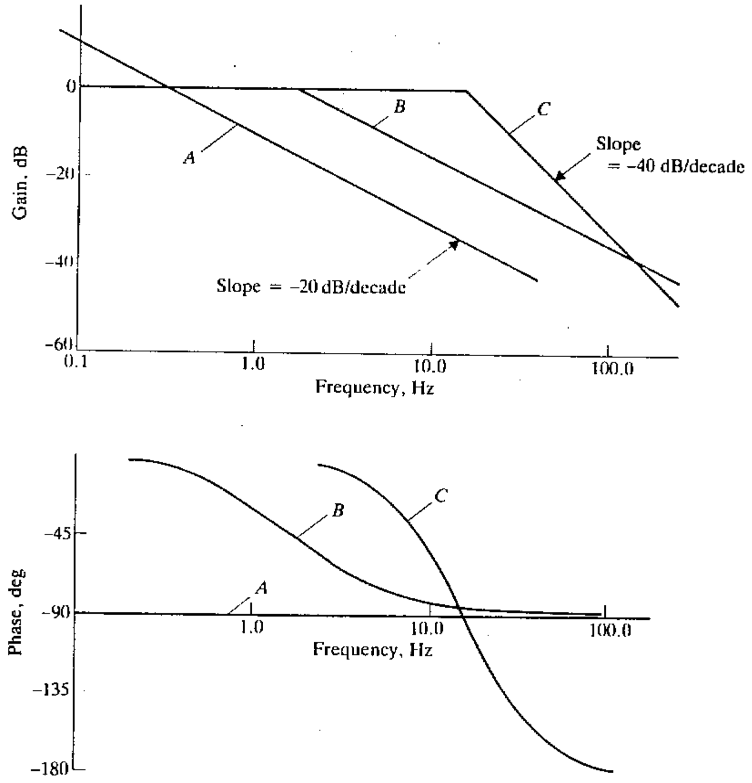

Figure 4.20(*a*) Bode plots for factors of G(*s*).

$$G(s) = \frac{1}{s\left(1 + \frac{s}{11.2}\right)\left[1 + 2(0.68)\frac{s}{95} + \left(\frac{s}{95}\right)^2\right]}$$

The second equation gives $\omega = 0$, the starting point of the locus, and

$$\omega = \sqrt{100\,980/140.2} = 26.8$$

and this value of ω when substituted into (4E.5) gives $K = 69.4$. The same value of K could have been obtained by use of the Routh or the Hurwitz algorithm. (See Prob. 4.9.)

The root locus plot for this system is shown in Fig. 4.19.

The transfer function, in factored form, is

$$G(s) = \frac{100\,980}{s(s + 11.2)(s^2 + 129s + 9016)}$$

$$= \frac{1}{s(1 + s/11.2)[1 + 2(0.68)(s/95) + (s/95)^2]}$$

The Bode plots for each factor in G(*s*) are shown in Fig. 4.20(*a*); the composite is shown in Fig. 4.20(*b*).

The Nyquist plot corresponding to G(*s*) is shown in Fig. 4.21.

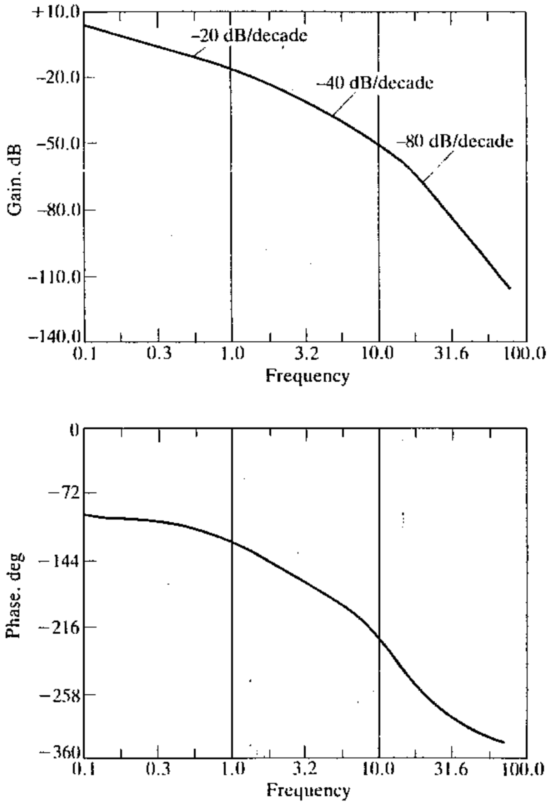

Figure 4.20(*b*) Bode plots for gun turret.

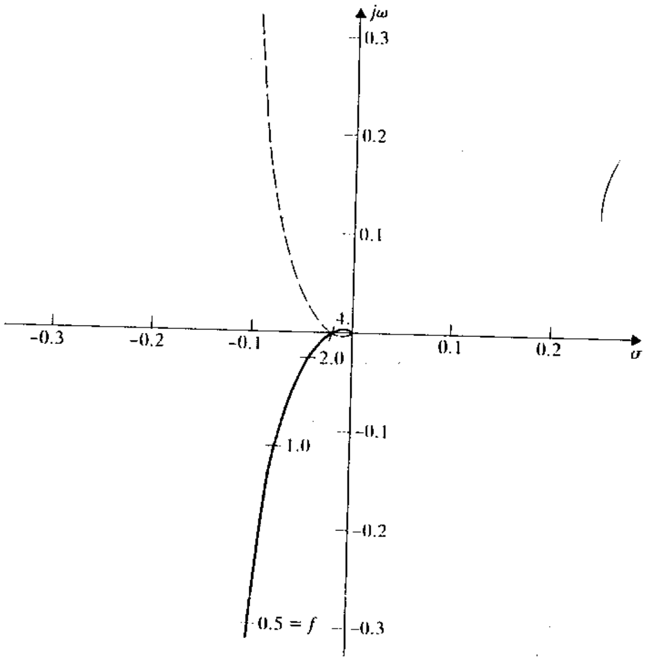

Figure 4.21 Nyquist diagram for hydraulically actuated gun turret.

Example 4F Missile dynamics. The motion of a missile about its pitch axis was shown in Example 3F to be given by

$$\dot{\alpha} = q + \frac{Z_\alpha}{V}\alpha + \frac{Z_\delta}{V}\delta$$

$$\dot{q} = M_\alpha \alpha + M_\delta \delta \qquad \text{(assuming } M_q \approx 0\text{)} \tag{4F.1}$$

where α is the angle of attack
q is the pitch rate
δ is the control surface deflection

The control surface is rotated by means of an actuator, the dynamics of which is typical of a first-order lag: (Fig. 4.22)

$$\dot{\delta} = \frac{1}{\tau}(u - \delta) \tag{4F.2}$$

where u is the input to the actuator.

A missile guidance system typically issues a guidance command in the form of the desired acceleration a_{N_c} normal to the missile velocity vector. The function of the autopilot, the design of which shall be considered in several examples later in the book, is to make the achieved normal acceleration a_N "track" the commanded acceleration with good fidelity. It is thus appropriate to deal with the error e between the commanded and the achieved normal

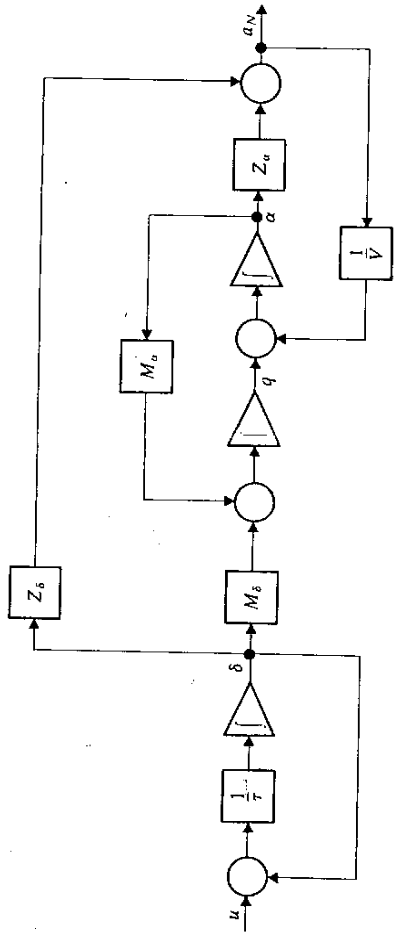

Figure 4.22 Missile attitude dynamics with normal acceleration as output.

acceleration. The latter is given by

$$a_N = Z_\alpha \alpha + Z_\delta \delta \tag{4F.3}$$

The transfer function from u to a_N is determined to be:

$$\mathrm{H}(s) = \frac{1}{\tau s + 1} \frac{Z_\delta s^2 + Z_\alpha M_\delta - Z_\delta M_\alpha}{s^2 + \dfrac{Z_\alpha}{V} s - M_\alpha} \tag{4F.4}$$

A representative set of numerical values for a hypothetical highly-maneuverable missile are:

$$V = 1253 \text{ ft/s}$$
$$Z_\alpha = -4170 \text{ ft/s}^2$$
$$Z_\delta = -1115 \text{ ft/s}^2$$
$$M_\alpha = -248 \text{ rad/s}^2$$
$$M_\delta = -662 \text{ rad/s}$$
$$\tau = .01 \text{ s}$$

For these values we obtain

$$\mathrm{H}(s) = \frac{-1115(s^2 - 2228)}{(0.01s + 1)(s^2 + 3.33s + 248)} \tag{4F.5}$$

The zeros of the denominator are at

$$s = -100$$

and at

$$s = -1.67 \pm j15.65$$

and the zeros of the numerator are at

$$s = \pm 47.2$$

Note that the dc gain of H(s) is positive: A positive input produces a positive response. But for high frequencies $\mathrm{H}(s) \to -1115/(0.01s + 1)$ which produces a negative response for a positive input. The change of sign in the transfer function as the frequency is increased is another consequence of the right half-plane zero of H(s) and is the source of apparent paradoxical behavior of the system. One paradox is in the root locus, shown in Fig. 4.23. It is observed that as the gain is increased from zero in the positive direction the one branch of the root locus crosses the imaginary axis at $s = 0$ and then continues to the real root as $s = 47.1$. This behavior is clear from the characteristic equation:

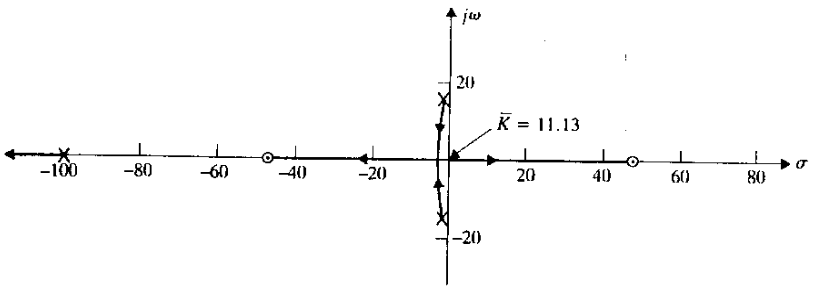

Figure 4.23 Root loci for missile dynamics. $\bar{K}$ increasing.

$$s^3 + (100.33 + \bar{K})s^2 + 581s + 24\,800 - 2228\bar{K} = 0$$

where

$$\bar{K} = 111\,500K$$

The coefficient of s^0 vanishes at

$$\bar{K} = \frac{24\,800}{2228} = 11.13$$

and hence there is a pole at the origin for this value of $\bar{K}$.

The graphical methods of Nyquist and Bode have an advantage over the algebraic Routh-Hurwitz methods: They are not restricted to rational transfer functions and thus not limited to systems characterized by ordinary differential equations. Thus they are applicable to systems characterized by partial-differential equations and pure delays. The following example provides a frequency-domain explanation for the instability exhibited (as found in Chap. 1) by a system with a delayed output.

Example 4G Pure delay In Chap. 1 we considered a system whose output $y(t)$ is a faithful, but delayed, version of the input $u(t)$

$$y(t) = u(t - T) \tag{4G.1}$$

The Laplace transform of the delayed input is

$$\mathrm{y}(s) = \int_0^\infty e^{-st} u(t - T)\, dt = \int_{-T}^\infty e^{-s(\tau + T)} u(\tau)\, d\tau \tag{4G.2}$$

On the assumption that $u(t)$ is zero for $t < 0$, (4G.2) becomes

$$\mathrm{y}(s) = e^{-sT} \int_0^\infty e^{-s\tau} u(\tau)\, d\tau = e^{-sT} \mathrm{u}(s)$$

Thus, the transfer function of a pure delay is

$$\mathrm{G}(s) = e^{-sT}$$

with $s = j\omega$

$$\mathrm{G}(j\omega) = e^{-j\omega T}$$

Thus

$$|\mathrm{G}(j\omega)| = 1$$

and

$$\theta_G(\omega) = -\omega T$$

The Nyquist diagram is thus a circle centered at the origin as shown in Fig. 4.24(a), and the closed-loop system, having the return difference

$$\mathrm{T}(s) = 1 + K\mathrm{G}(s) = 1 + K e^{-sT} \tag{4G.3}$$

is unstable for $K > 1$, as was found in Chap. 1. The Bode diagram has a constant amplitude of 1 (0 dB) and a linearly decreasing phase (which does not look linear on a logarithmic frequency axis as shown in Fig. 4.24 for $T = 0.01$ s).

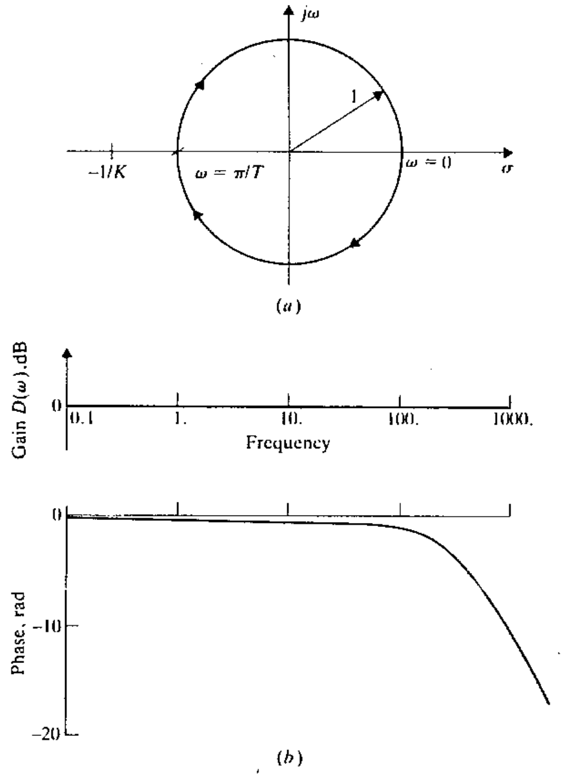

Figure 4.24 Stability plots for pure delay $G(s) = e^{-Ts}$. (a) Nyquist plot; (b) Bode plot.

4.7 STEADY STATE RESPONSES: SYSTEM TYPE

Stability is the control system designer's first concern. With stability assured, interest shifts to the nature of the response of the system to various types of reference inputs. (For the present we consider only single-input, single-output processes. The general multiple-input, multiple-output case is treated in Sec. 4.10.)

A system designed to follow a reference input, rather than merely to return to equilibrium, is generally known as a "tracking" system, and has the configuration shown in Fig. 4.25. The output of the system is y and the input to the open-loop plant is u. The difference between the desired reference input is called the *system error*

$$e = y_r - y$$

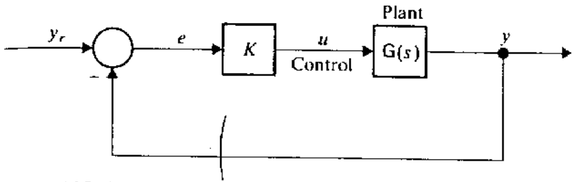

Figure 4.25 Error driven feedback control system.

In the simplest type of control system, the system error is multiplied by a gain K to produce the control input u. Since the input to the plant is proportional to the error, if the error becomes large it will produce a large input which, because of negative feedback, will tend to drive the error to zero.

To analyze the behavior of the system quantitatively, consider the transfer functions from the input to the error e and to the output y. Using the block-diagram algebra of Sec. 4.3, or any other convenient method, one can easily determine these transfer functions

$$H_E(s) = \frac{1}{1 + K G(s)} = \frac{e(s)}{y_r(s)} \tag{4.43}$$

$$H_C(s) = \frac{K G(s)}{1 + K G(s)} = \frac{y(s)}{y_r(s)} \tag{4.44}$$

Note that the return difference $1 + KG(s)$ is in the denominator of the transfer function (4.43) between the reference input and the error. To make the error small, we would like the return difference to be large. With the plant transfer function $G(s)$ given, the return difference $1 + KG(s)$ can be increased by increasing K. For reasons of stability, however, we usually can't make K arbitrarily large. Therefore it is not possible to design a system that can track *every* reference input with arbitrarily small errors. As a matter of fact, we might not be very happy with a system that does so. The reference input typically has rapid changes and noise which often one might not want to track with perfect fidelity.

One measure of the performance of a system is its steady state behavior when the reference input to the system is a polynomial time function:

$$y_r(t) = p_m(t) = C_1 + \frac{C_2}{2}t + \cdots + \frac{C_m}{m!}t^m \tag{4.45}$$

It is useful to formalize this measure of performance with the following:

Definition A system is of "type m" if it can track a polynomial input of degree m with finite but nonzero steady state error.

We shall shortly discover that a system of type m can track a polynomial of degree $m - 1$ (or less) with zero steady state error, but that the error in tracking a polynomial reference input of degree $m + 1$ (or greater) becomes infinite.

The steady state behavior of the error is determined with the aid of the Laplace transform final-value theorem [4]:

$$\text{Steady state error} = \lim_{t\to\infty} e(t) = \lim_{s\to 0} s\mathrm{e}(s)$$

where, by (4.43)

$$\mathrm{e}(s) = \frac{1}{1 + K\mathrm{G}(s)}\mathrm{y}_r(s) \qquad (4.46)$$

When the input $y_r(t)$ is the polynomial time function $p_m(t)$, then its Laplace transform

$$\mathrm{y}_r(s) = \mathrm{p}_m(s) = \frac{C_1}{s} + \frac{C_2}{s^2} + \cdots + \frac{C_{m+1}}{s^{m+1}} \qquad (C_{m+1} \neq 0)$$
$$= \frac{C_1 s^{m+1} + C_2 s^m + \cdots + C_{m+1}}{s^{m+1}} \qquad (4.47)$$

Thus, from (4.46),

$$s\mathrm{e}(s) = \frac{1}{1 + K\mathrm{G}(s)} \frac{C_1 s^{m+1} + \cdots + C_{m+1}}{s^m} \qquad (4.48)$$

The limit as $s \to 0$ of $s\mathrm{e}(s)$ is infinite if G(0) is finite, because of the presence of the factor s^m in the denominator of $s\mathrm{e}(s)$ given by (4.48). The only way that, as $s \to 0$, lim $s\mathrm{e}(s)$ can be finite is if G(s) has a pole of the proper order at $s = 0$, that is, if G(s) is of the form

$$\mathrm{G}(s) = \frac{\mathrm{N}(s)}{s^p \mathrm{D}(s)} \qquad (4.49)$$

where neither N(s) nor D(s) have zeros at $s = 0$. When G(s) is of the form (4.49) then

$$s\mathrm{e}(s) = \frac{\mathrm{D}(s)}{1 + K\dfrac{\mathrm{N}(s)}{s^p\mathrm{D}(s)}} \frac{C_1 s^{m+1} + \cdots + C_{m+1}}{s^m}$$
$$= \frac{s^{p-m}\mathrm{D}(s)}{s^p\mathrm{D}(s) + K\mathrm{N}(s)}(C_1 s^{m+1} + \cdots + C_{m+1}) \qquad (4.50)$$

From (4.50) we infer the following:

$$\begin{aligned} &\text{If} \quad p > m, \quad \lim_{t\to\infty} e(t) = \lim_{s\to 0} s\mathrm{e}(s) = 0 \\ &\text{If} \quad p = m, \quad \lim_{t\to\infty} e(t) = \lim_{s\to 0} s\mathrm{e}(s) \text{ is finite but nonzero} \qquad (4.51) \\ &\text{If} \quad p < m, \quad \lim_{t\to\infty} e(t) = \lim_{s\to 0} s\mathrm{e}(s) \text{ is infinite} \end{aligned}$$

We thus conclude that:

The system type is equal to the order of the pole of G(s) at $s = 0$.

Since a pole at the origin represents a perfect integrator, the system type is often defined as the number of cascaded integrators in the system.

It should be noted that the presence of a single integrator in the open-loop plant implies that it is stable, but not asymptotically stable, and that more than one integrator means that the open-loop plant is unstable. Thus we see that a closed-loop system, having the ability to track a polynomial input, cannot result when the open-loop plant is asymptotically stable! This should come as no surprise. The closed-loop system, after all, is error-driven, and we are insisting that the steady-state error go to zero. This means that the control input u also becomes zero in the steady state. But, at the same time we are demanding that the output be nonzero! To sustain a nonzero output with a zero input is not one of the properties of an asymptotically stable system.

It is emphasized that the system type is determined by the order of the pole at the origin in the open-loop process, and not the closed-loop process. A properly designed closed-loop process must invariably be asymptotically stable, and thus the closed-loop transfer function must have all its poles in the left half-plane and none on the imaginary axis, which includes the origin.

The steady state error that results when the polynomial input is the same degree as the system type is determined with the aid of (4.50)

$$\lim_{t\to\infty} e(t) = \lim_{s\to 0} s\mathrm{e}(s) = \begin{cases} \dfrac{C_1\mathrm{D}(0)}{\mathrm{D}(0) + K\mathrm{N}(0)} & \text{for } p = m = 0 \\ \dfrac{C_{m+1}\mathrm{D}(0)}{K\mathrm{N}(0)} & \text{for } p = m \geqq 1 \end{cases} \qquad (4.52)$$

If we define the "fractional error," or "error ratio," as

$$r_m = \frac{1}{C_{m+1}} \lim_{t\to\infty} e(t)$$

then, from (4.52) we determine the fractional error for a type m system (when the input is a polynomial of degree m)

$$r_m = \begin{cases} \dfrac{1}{1 + K\mathrm{N}(0)/\mathrm{D}(0)} = \dfrac{1}{1 + K\mathrm{G}(0)} & m = 0 \\ \dfrac{1}{K\mathrm{N}(0)/\mathrm{D}(0)} = \dfrac{1}{K\mathrm{G}(0)} & m \geqq 1 \end{cases} \qquad (4.53)$$

Thus the steady state error ratio decreases as the open-loop dc gain KG(0) increases. Hence if the requisite number of integrators is not present to make the plant of the desired type, the steady state error ratio can be reduced (but not brought to zero) by making the loop gain KG(0) high. But of course it can't be made arbitrarily high without compromising stability.

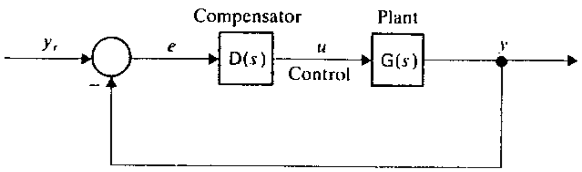

Figure 4.26 Error-driven feedback control system with compensator.

If closed-loop system tracking performance of a particular type is required, but the existing open-loop plant type is not sufficiently high, the required number of integrators is supplied by means of a "compensator": a dynamic system having the transfer function D(s) placed between the measured error and the input to the original plant, as shown in Fig. 4.26.

The most common type of compensator is the so-called PI (proportional + integral) compensator

$$D(s) = \frac{K_1}{s} + K_2 = \frac{K_1 + K_2 s}{s}$$

The pole in D(s) raises the type of the open-loop system.

Before the era of digital computers, it was difficult to construct a perfect integrator: a device which maintains an absolutely constant output indefinitely in the absence of any input. The output of a physical device called an integrator would tend to drift, either to zero or, worse still, to infinity. The quality of an integrator was measured by the time it could be expected to hold a constant output. High-quality integrators needed to control processes whose natural time constants were of the order of several hours (typical in process control) were very expensive and marginally reliable, hence PI controls were bothersome. The digital computer now provides a means for realizing a perfect integrator, and the hardware problems of PI control have disappeared.

In principle, the compensator can provide a double pole at the origin thus raising the open-loop system type by 2. This is generally infeasible in practice, however, because the output of such a compensator is an increasing function of time even when the error is zero. This output is the input u to the original plant. Hence the physical input to the plant is constantly increasing in magnitude. Sooner or later a limit will be reached at which point the physical input to the plant will have to stop increasing: the control input "saturates." The input demanded by the compensator will not be physically attainable. This fact of saturation needs to be taken into account in the system design. The system should not be required to exhibit behavior of a type of which it is physically incapable.

In classical system designs, the integrators needed to raise the system type are frequently included in the plant model. The designer is concerned with shaping the dynamic behavior of that part of the compensator having no poles

at the origin. With the state variable approach to be developed in this book, the integrators required in the compensator emerge in the normal design process.

4.8 DYNAMIC RESPONSE: BANDWIDTH

Another consideration in control system design is *dynamic response.* Not only must the closed-loop system be stable and reach the required steady state value eventually, but it cannot take forever to get to where it's going, and it should not be too oscillatory.

The dynamic characteristics of the system are typically defined in terms of the response to a unit-step input. The "step response" of a typical system is shown in Fig. 4.27. The parameters of major interest are as follows:

Overshoot Difference between peak value and steady state value;

Rise time Time for output to reach a specified fraction (usually $1 - e^{-1} = 0.632$) of the steady state value.

In addition to these parameters, often other parameters are of interest such as delay time (the time that it takes for the output to "get started"—to reach say 10 percent of its steady state value), "peak-time" (time to reach first peak in a system with positive overshoot).

There is no universal agreement on the definitions. For example, in some fields (e.g., process control) rise time is defined as the time it takes for the process to go from 10 to 90 percent of its final value. Or it may be defined as the time it takes to get to its steady state value the first time. (This is only meaningful in a system with overshoot.)

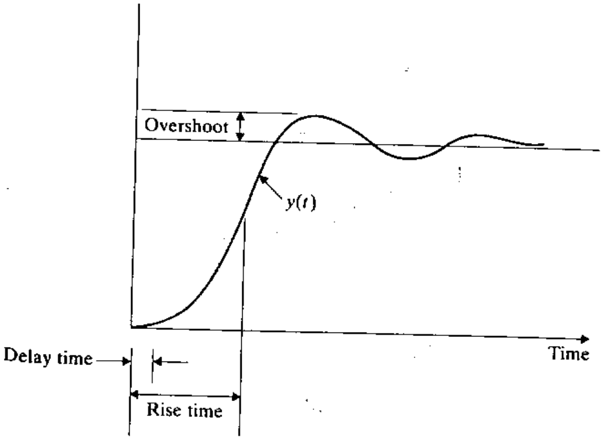

Figure 4.27 Characteristics of dynamic response.

Dynamic performance requirements of a system are typically specified in terms of the maximum permissible rise time and overshoot. These parameters are readily measured on a time-history plot of the output of an actual system or, for design purposes, on the output of a simulation of the actual system. But they are not readily calculated from the transfer function of the system. One way of avoiding the calculation problem is to define the response time as the *centroid* $\bar{t}$ of the impulse response of the system

$$\bar{t} = \frac{\int_0^\infty t h(t)\, dt}{\int_0^\infty h(t)\, dt} \tag{4.54}$$

The physical significance of the centroidal response time $\bar{t}$ is that it is the location of a single impulse which has the same effect as the actual system. An impulse located at $t = \bar{t}$ would give rise to a step occurring at that time, as shown in Fig. 4.28. This interpretation of the response time has a certain intuitive appeal.

The response time $\bar{t}$ cannot be determined by simply inspecting the step response of the system. Precise calculation of $\bar{t}$ from recorded data would require numerical integration of the step response: not a difficult task for a digital computer but not as easy as picking one or two points off a curve. The definition of response time by $\bar{t}$ has decided advantages with regard to analysis, however: It can be calculated directly from the transfer function H(s) without the need for first determining the expression for the step response or the impulse response. To see this, note that the transfer function is the Laplace transform of the impulse response

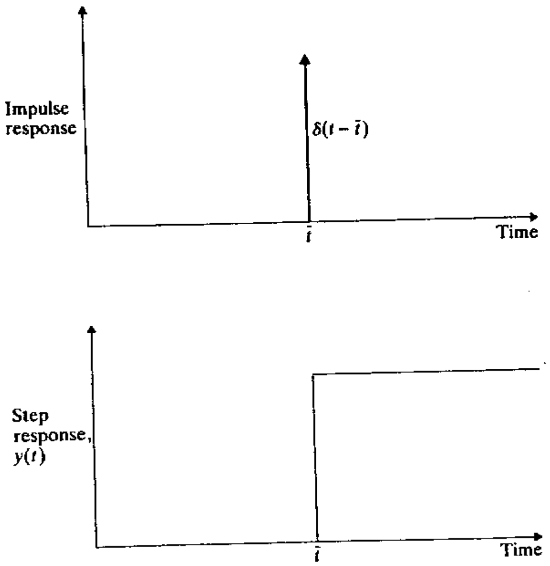

Figure 4.28 Definition of centroidal response time.

$$\mathrm{H}(s) = \int_0^\infty e^{-st} h(t)\, dt \tag{4.55}$$

Thus

$$\mathrm{H}(0) = \int_0^\infty h(t)\, dt \tag{4.56}$$

Now take the derivative of both sides of (4.55) with respect to s

$$\mathrm{H}'(s) = d\mathrm{H}(s)/ds = -\int_0^\infty t\, e^{-st} h(t)\, dt$$

Thus

$$-\mathrm{H}'(0) = \int_0^\infty t h(t)\, dt$$

Thus

$$\bar{t} = -\frac{\mathrm{H}'(0)}{\mathrm{H}(0)} \tag{4.57}$$

In words, the centroidal response time $\bar{t}$ is the ratio of the Laplace transforms of the derivative of the transfer function at the origin to the transfer function itself; $\bar{t}$ is thus a measure of how fast the transfer function decreases at the origin.

It is very easy to calculate $\bar{t}$ using (4.57). For this reason it is a useful definition notwithstanding the possible difficulty of determining it from measured data. In typical systems, moreover, $\bar{t}$ is very close to the rise time calculated using other definitions. In a first-order system, for example, with

$$\mathrm{H}(s) = \frac{\omega_0}{s + \omega_0} \tag{4.58}$$

it is found that (4.57) gives

$$\bar{t} = \frac{1}{\omega_0} \tag{4.59}$$

The step response corresponding to this transfer function is

$$a(t) = \mathscr{L}^{-1}\left[\frac{\omega_0}{s(s+\omega_0)}\right] = 1 - e^{-\omega_0 t}$$

The step response reaches $1 - e^{-1}$ of its final value at $t_{0.63}$ given by

$$\omega_0 t_{0.63} = 1 \tag{4.60}$$

Thus, for a first-order system, the centroidal response time and the 63 percent response time are exactly equal. (The response of a first-order system is the basis of the definition of the 63 percent response time.)

The centroidal response time is easy to calculate for a second-order system with

$$H(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{4.61}$$

Using (4.57) it is found that

$$\bar{t} = \frac{2\zeta}{\omega_0} \tag{4.62}$$

This means that the response time increases with the damping factor ζ. For two systems having equal natural frequencies, the system with the larger damping factor has the larger response time, which agrees with our intuition.

Calculation of the step response corresponding to H(s) as given by (4.61) and then solving for $t_{0.63}$ is a messy business. The step-response curves themselves have the appearance shown in Fig. 4.29. The 63 percent and the centroidal response times are both shown graphically vs. ζ in Fig. 4.30. They agree exactly at a damping factor $\zeta = 1$ and are within about 20 percent of each other for the range of damping factors of practical interest: $\zeta > 0.7$.

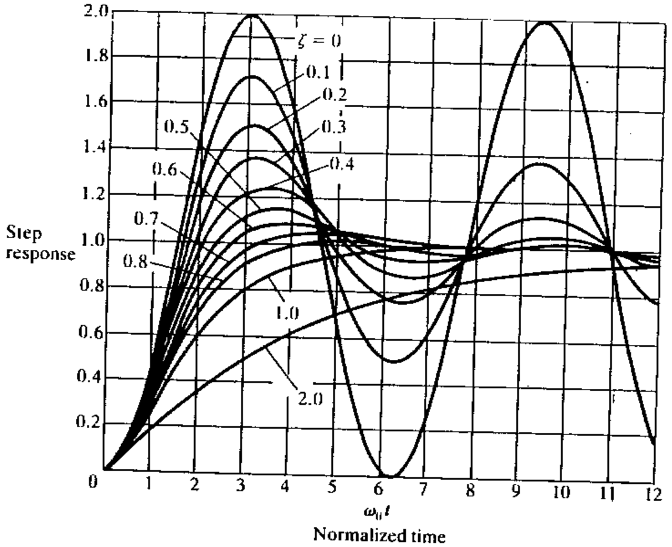

Figure 4.29 Step response of second-order system

$$G(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

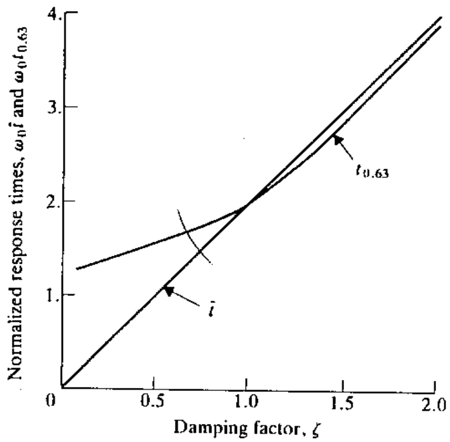

Figure 4.30 Response time of second-order system

$$H(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 + \omega_0^2}$$

For very low damping factors, the centroidal rise time is much smaller than the 63 percent rise time, but neither gives a very good indication of system behavior.

Calculation of the centroidal rise time of systems in tandem (cascade) is instructive. Suppose a system H(s) comprises two subsystems H$_1$(s) and H$_2$(s) in tandem

$$H(s) = H_1(s)H_2(s)$$

Then

$$H'(s) = H_1'(s)H_2(s) + H_2'(s)H_1(s)$$

Thus

$$\frac{H'(s)}{H(s)} = \frac{H_1'(s)H_2(s) + H_2'(s)H_1(s)}{H_1(s)H_2(s)} = \frac{H_1'(s)}{H_1(s)} + \frac{H_2'(s)}{H_2(s)} \tag{4.63}$$

Thus, on evaluating (4.63) at $s = 0$, we obtain

$$\bar{t} = \bar{t}_1 + \bar{t}_2 \tag{4.64}$$

i.e., the centroidal response time of a tandem combination of systems is the sum of the centroidal response times of each. This also agrees with our intuition: There is a lag in going through the first system, and the second system adds an additional lag, so we would expect that a formula like (4.64) will hold, at least approximately, for any reasonable definition of response time.

Finally, consider the effect on response time due to feedback. Suppose H(s) is the transfer function of a closed-loop error-driven control system

$$\mathrm{H}(s)=\frac{K\mathrm{G}(s)}{1+K\mathrm{G}(s)}$$

Then

$$\mathrm{H}'(s)=\frac{[1+K\mathrm{G}(s)]K\mathrm{G}'(s)-K\mathrm{G}(s)K\mathrm{G}'(s)}{[1+K\mathrm{G}(s)]^2}=\frac{K\mathrm{G}'(s)}{[1+K\mathrm{G}(s)]^2}$$

and

$$\frac{\mathrm{H}'(s)}{\mathrm{H}(s)}=\frac{K\mathrm{G}'(s)}{[1+K\mathrm{G}(s)]^2}\,\frac{1+K\mathrm{G}(s)}{K\mathrm{G}(s)}=\frac{1}{1+K\mathrm{G}(s)}\,\frac{\mathrm{G}'(s)}{\mathrm{G}(s)} \tag{4.65}$$

Thus, on evaluating (4.65) at $s=0$, we find the closed-loop centroidal response time given by

$$\bar{t}_{CL}=\frac{1}{1+K\mathrm{G}(0)}\bar{t}_G \tag{4.66}$$

where $\bar{t}_G$ is the response time of the open-loop plant. We see that another benefit of feedback is the reduction of the system response time: The closed-loop response time is equal to the open-loop response time divided by the dc return difference.

The centroidal response time can readily be calculated from the state-space representation of the transfer function

$$\mathrm{H}(s)=C(sI-A)^{-1}B \tag{4.67}$$

where C is a $1\times k$ matrix and B is a $k\times 1$ matrix, so the product $C(sI-A)^{-1}B$ is a 1×1 matrix. The dc transmission is

$$\mathrm{H}(0)=-CA^{-1}B$$

(We assume that the system (4.67) is asymptotically stable, which assures that A^{-1} exists.) To find H$'(0)$ we write

$$\mathrm{H}(s)=C\Phi(s)B$$

where

$$\Phi(s)=(sI-A)^{-1}$$

is the resolvent of A. Then

$$\mathrm{H}'(0)=C\Phi'(0)B$$

Now

$$(sI-A)\Phi(s)=I$$

Differentiate with respect to s

$$\Phi(s)-A\Phi'(s)=0$$

or

$$\Phi'(0)=A^{-1}\Phi(0)=-A^{-2}$$

and

$$\mathrm{H}'(0)=-CA^{-2}B$$

Finally

$$\bar{t}=\frac{\mathrm{H}'(0)}{\mathrm{H}(0)}=\frac{CA^{-2}B}{CA^{-1}B} \tag{4.68}$$

(It is legal to divide by $CA^{-1}B$ because $CA^{-1}B$ is a 1×1 matrix.)

One is tempted to cancel C and B in the numerator and denominator of (4.68) and thereby obtain an expression for $\bar{t}$ in terms of only the dynamics matrix A. This is not legal, of course. Even if it were legal, the final result wouldn't make any sense because the result would be $\bar{t}=A^{-1}$ which is equating a matrix to a scalar. Is there any way that something like this equation could make sense? The answer is yes, provided that an appropriate generalization is used. The quantity

$$\bar{t}=(\|A^{-1}\|/\|A\|)^{1/2} \tag{4.69}$$

has been studied by Bass.[15] For a further discussion of this, see Note 4.3.

In the frequency domain, the dynamic behavior of a system is characterized by its *bandwidth.* The reader has very likely formed an intuitive notion of bandwidth by reasoning as follows: Every input can be regarded as comprising components at various frequencies; a rapidly changing input means that it has a large high frequency content, while a smooth, slowly varying input has a relatively smaller high frequency content. Thus if the control system is to faithfully reproduce an input that is changing rapidly, it must be capable of faithfully reproducing inputs at high frequencies, i.e., to have a large bandwidth. On the other hand, if the input is slowly varying, the control system does not need a high bandwidth. A step change (i.e., a discontinuity) in the input has a large amount of high-frequency content: to reproduce it faithfully (with a short rise time) requires a high bandwidth. If a short rise time is not required then neither is a high bandwidth. Reasoning thus one reaches the conclusion that there is an inverse relationship between bandwidth and response time.

In communication systems where data is transmitted by modulation of a carrier, the bandwidth of a system is generally defined as the width of the resonance curve between the "half power" frequencies—i.e., the frequencies at which the gain is "3 dB down" from the value at resonance. In a control system, however, the oscillatory response characteristic of a resonant system is highly undesirable. If well designed, a closed-loop system will have a frequency response characteristic of the Bode plots of Fig. 4.16 with a damping factor ζ greater than about 0.4 or 0.5. The resonance peak, if any, is very small and the frequency response typically is very flat until a critical frequency is reached, at which point it begins falling off quite rapidly. A system of this type is called *low* *pass*—it passes low frequencies without substantial attenuation—and its bandwidth is usually defined as the frequency at which the gain is below the dc gain by a factor of $\sqrt{2}$. In other words

$$\left|\frac{H(jW)}{H(0)}\right| = \frac{1}{\sqrt{2}}$$

or

$$\frac{|H(jW)|^2}{H^2(0)} = \frac{1}{2} \qquad (4.70)$$

where W is the bandwidth of the system.

In a first-order system

$$H(s) = \frac{\omega_0}{s+\omega_0}$$

(4.70) becomes

$$\frac{\omega_0^2}{W^2+\omega_0^2} = \frac{1}{\left(\frac{W}{\omega_0}\right)^2+1} = \frac{1}{2}$$

or

$$W = \omega_0 \qquad (4.71)$$

We already knew this from the nature of the Bode plot for a first-order system. For a second-order system

$$H(s) = \frac{\omega_0^2}{s^2+2\zeta\omega_0 s+\omega_0^2}$$

The bandwidth W is the solution of

$$\frac{1}{\left[1-\left(\frac{W}{\omega_0}\right)^2\right]^2+4\zeta\left(\frac{W}{\omega_0}\right)^2} = \frac{1}{2}$$

or

$$\left(\frac{W}{\omega_0}\right)^2 = 1-2\zeta^2+\sqrt{(1-2\zeta^2)^2+1} \qquad (4.72)$$

A plot of the bandwidth of a second-order system vs. damping factor is shown in Fig. 4.31. For purposes of comparison, the reciprocal of the centroidal response time $\bar{t}$ is also shown. It is observed that for the useful range of damping factors ($\zeta > 0.4$) the reciprocal of $\bar{t}$ is a lower bound on the bandwidth:

$$W\bar{t} > 1$$

In other words the product of the bandwidth and the centroidal response is

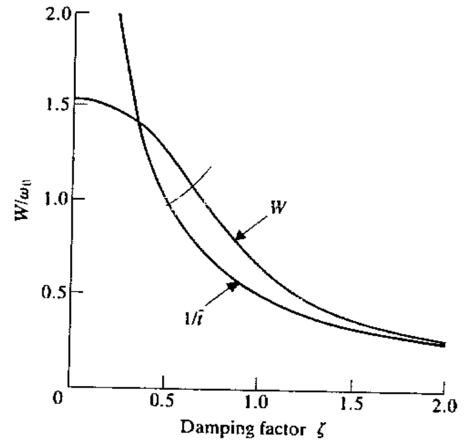

Figure 4.31 Bandwidth of second-order system

$$H(s) = \frac{\omega_0^2}{s^2+2\zeta\omega_0+\omega_0^2}$$

always greater than 1—a sort of time-frequency uncertainty principle. Note, however that $W\bar{t}$ is never *much* greater than 1. Thus we can safely say that the product of bandwidth and response time, which is exactly 1 for a first-order system, is approximately 1 for properly damped second-order systems. The relationship

$$W\bar{t} \approx 1 \qquad (4.73)$$

turns out to hold for a higher order as well, and reinforces our intuitive conception of the reciprocal relationship between bandwidth and response time.

4.9 ROBUSTNESS AND STABILITY (GAIN AND PHASE) MARGINS

In designing the control law for some process we deal not with the physical process per se but rather with a mathematical model of the process. The control law will be acceptable only if the mathematical model predicts the behavior of the physical process reasonably well. But no mathematical model can predict the behavior of a physical process *exactly*; there will always be some discrepancy between the actual behavior of the physical process and that predicted by the mathematical model. And the discrepancy may increase with time owing to normal aging and deterioration. Some uncertainty of the physical process is a reality that the control system engineer must contend with.

One of the well-known advantages of feedback is that it confers a degree of "robustness" or immunity to uncertainty or changes in the process. "Sensitivity" S of a process to a change in one of the parameters of the process in a way of quantifying the advantage of feedback. Suppose that the transfer function of the process is $H(s;\alpha)$ where α is some parameter that can change in the process. We define the *sensitivity* of the transfer function $H(s;\alpha)$ to a change in α by

$$S(\alpha)=\frac{1}{\mathrm{H}(s)}\frac{\partial \mathrm{H}(s)}{\partial\alpha}=\frac{1}{\mathrm{H}(s;\alpha)}\lim_{\Delta\alpha\to 0}\frac{\mathrm{H}(s;\alpha+\Delta\alpha)-\mathrm{H}(s;\alpha)}{\Delta\alpha} \qquad (4.74)$$

The sensitivity is thus the fractional change in the transfer function due to a change in the parameter and corresponds to our intuitive understanding of sensitivity.

Let us compare the sensitivity of the open-loop system consisting of an amplifier of gain K and a plant $\mathrm{G}(s)$ in tandem (as shown in Fig. 4.32(*a*)) with the closed-loop system shown in Fig. 4.32(*b*). The transfer function of the open-loop system is

$$\mathrm{H}_0(s)=K\mathrm{G}(s) \qquad (4.75)$$

and hence the sensitivity to a change in K is

$$S_0(K)=\frac{1}{\mathrm{H}_0}\frac{\partial \mathrm{H}_0}{\partial K}=\frac{1}{K}$$

The closed-loop transfer function, on the other hand, is

$$\mathrm{H}_c(s)=\frac{K\mathrm{G}(s)}{1+K\mathrm{G}(s)} \qquad (4.76)$$

and the corresponding sensitivity is

$$S_c=\frac{1+\mathrm{G}(s)K}{\mathrm{G}(s)K}=\frac{[1+\mathrm{G}(s)K]\mathrm{G}(s)-\mathrm{G}(s)K\mathrm{G}(s)}{[1+\mathrm{G}(s)K]^2}=\frac{1}{K(1+\mathrm{G}(s)K)} \qquad (4.77)$$

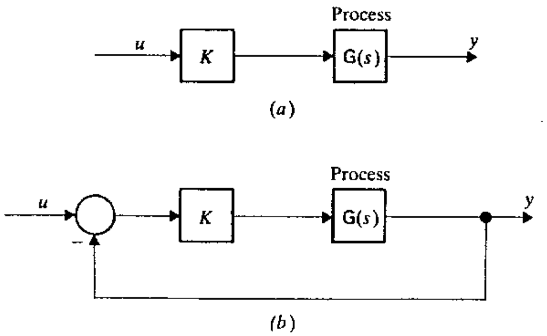

Figure 4.32 Open-loop and closed-loop processes for sensitivity comparisons. (*a*) Open-loop process $\mathrm{H}_0(s)=K\mathrm{G}(s)$; (*b*) Closed-loop process $\mathrm{H}_c(s)=K\mathrm{G}(s)/[1+K\mathrm{G}(s)]$.

The ratio of the closed-loop sensitivity to the open-loop sensitivity is

$$\frac{S_c}{S_0}=\frac{1}{1+\mathrm{G}(s)K} \qquad (4.78)$$

Feedback thus has the effect of reducing the sensitivity to gain variations by the reciprocal of the return difference $1+\mathrm{G}(s)K$. The higher the return difference, the lower the sensitivity to parameter changes. Thus a high return difference not only speeds up the dynamic response, as shown in (4.66), but also tends to immunize the system to changes in parameters of the open-loop system. (In the above example, the sensitivity to changes in the gain K of the amplifier was computed. But the principle is equally valid for changes in other parameters of the plant $\mathrm{G}(s)$.)

From the standpoint of speed of response and immunity to parameter variations, we should like the return difference of the system to be large at all frequencies. There are various reasons, however, why this is not a practical goal. The most important reason is that the transfer function of every practical plant is "low pass," tending to zero (in magnitude) as frequency becomes infinite. If the amplifier has a fixed gain K, the loop transmission will tend to zero at high frequencies, and hence the return difference will tend to unity. Instead of an amplifier with a constant gain of K one might conceive of using a dynamic "compensator" $\mathrm{D}(s)$ with a gain that increases with frequency to counteract the decrease in the plant transfer function. Such compensators are feasible but not very desirable in practice because they amplify the inherent high-frequency noise in the system. Moreover, a *physical* compensator cannot sustain a transfer function that increases indefinitely with frequency. As with every physical device, the transfer function of any compensator must eventually "roll off" with frequency.

Thus, even with dynamic compensation, the loop transmission of a system must ultimately become zero, and the return difference must ultimately approach unity. The practical design problem is not how to keep the return difference large at all frequencies, but rather how to make the return difference tend to unity in a graceful, well-behaved manner.

The problem is phase shift. The decrease in amplitude of the loop gain is accompanied by a phase shift. It is possible for the loop gain to reach unity in *magnitude*, and to have a phase shift of 180° at some frequency. In this case the return difference is zero and the transfer function of the system becomes infinite: the system is unstable. In order for the system to be stable, it is not permissible for the return difference to go to zero at any frequency. Moreover, because of possible differences between the transfer function used for purposes of design and the true transfer function, it is imprudent to permit the return difference to come close to zero. In a practical design it is necessary to provide reasonable *stability margins*.

Gain margin and *phase margin* are the stability margins in common use. The gain margin is the amount that the loop gain can be changed, at the frequency at which the phase shift is 180°, without reducing the return difference to zero.

The phase margin is the amount of phase lag that can be added to the open-loop transfer function, at the frequency at which its magnitude is unity, without making the return difference zero. These margins can be illustrated on a Nyquist plot or a Bode plot for a typical transfer function, as shown in Fig. 4.33.

The Nyquist diagram corresponding to a typical loop transmission is shown in Fig. 4.33(*a*). The loop gain *K* is taken to be unity, as is customary in gain and phase margin analyses. The system as shown is stable. At the frequency ω_2 at which the phase shift is 180°, the magnitude of the loop transfer function $|G(j\omega_2)|$ is less than unity by the gain margin γ. This means that the loop transmission can be raised by an amount γ without causing the system to become unstable. In most instances the gain margin is expressed as a logarithmic ratio in $\text{dB} = 20 \log \gamma$. The Nyquist diagram of Fig. 4.33 also shows the phase margin ϕ which is the angle that the phasor $G(j\omega_1)$ makes with the negative real axis at the frequency ω_1 at which $|G(j\omega)|$ first reaches unity. If a phase lag less than the phase margin ϕ were added to each point on the plot of $G(j\omega)$, the Nyquist diagram would not encircle the $-1 + j0$ point and the closed-loop system would remain stable.

The gain and phase margins are shown in the Bode plot of the loop transmission in Fig. 4.33(*b*). Note again that the phase margin is the difference between 180° and the actual phase shift at the "gain crossover" frequency ω_1 and that the gain margin is the amount that the log magnitude plot falls below 0 dB at the "phase crossover" frequency ω_2.

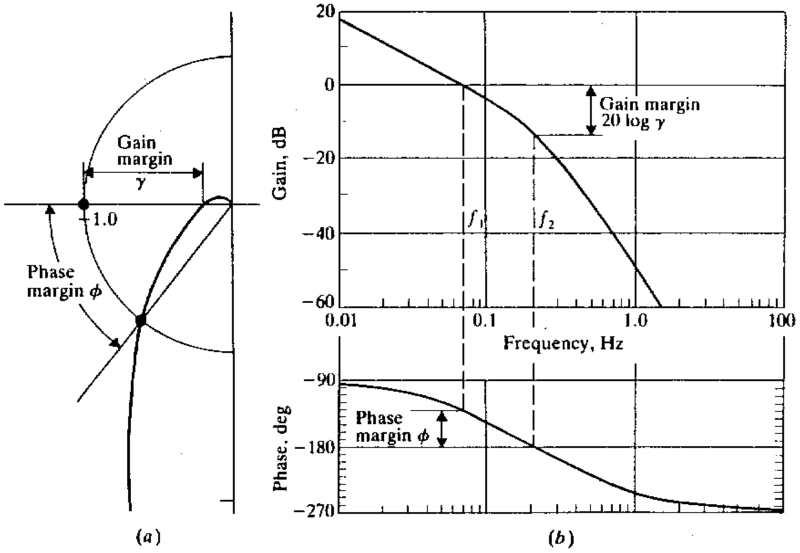

Figure 4.33 Gain and phase margins defined. (*a*) Nyquist diagram; (*b*) Bode plots.

The gain and phase margins are conveniently expressed in terms of the magnitude of the return difference

$$T(j\omega) = 1 + G(j\omega) \tag{4.79}$$

which is the phasor from the point $-1 + j0$ to the phasor $G(j\omega)$ on the Nyquist plot, as shown in Fig. 4.34(*a*). It is seen directly from Fig. 4.34(*b*) that

$$\gamma = |T(j\omega_2)| = |1 + G(j\omega_2)| \qquad \theta_G(\omega_2) = 180° \tag{4.80}$$

And, by a simple geometric construction (see Fig. 4.34(*c*))

$$\phi = 2 \sin^{-1} \left| \frac{T(j\omega_1)}{2} \right| \qquad |G(j\omega_1)| = 1 \tag{4.81}$$

Since the gain and phase margins are directly related to the variation with frequency ω of the return difference $T(j\omega)$, it would not be very surprising to find that the return difference has an important role to play in the assessment of the robustness of a control system. The return difference retains its importance even in multiple-input, multiple-output systems, in which the concepts of gain and phase margin become problematic.

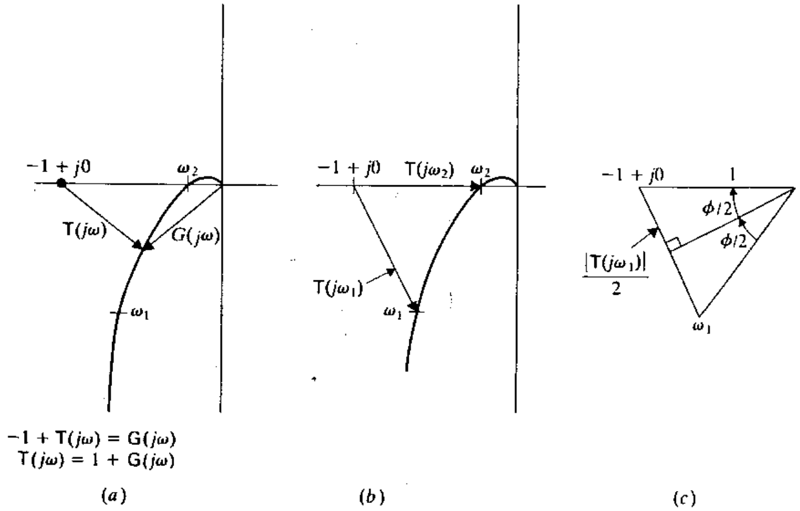

Figure 4.34 Use of return difference with phase and gain margins. (*a*) Return difference is phasor from $-1 + j0$ to phase $G(j\omega)$; (*b*) Return differences at frequencies of gain and phase margin; (*c*) Construction for phase-margin formula.

4.10 MULTIVARIABLE SYSTEMS: NYQUIST DIAGRAM AND SINGULAR VALUES

Most of this chapter has been concerned with single-input, single-output "scalar" systems: The plant under examination has one input and one output and hence, in the frequency domain, it is characterized by a single transfer function. Most of the systems encountered in practice fall into this class. But there are many systems in which there are more than one control input and/or more than one output of concern. An example of a multiple-input, multiple-output (or, more simply, *multivariable*) system is the lateral channel of an aircraft (in which the inputs are the rudder and aileron deflections and in which possible outputs would be the roll and yaw angles). Another example is the distillation column discussed in Example 4A and earlier.

With state-space methods, the focus is on the *state* of the process more than on the inputs or the outputs. Hence there is less need for making a distinction between scalar systems and multivariable systems than there is with frequency-domain methods. (This is one of the advantages of state-space methods.) Nevertheless, a familiarity with some of the concepts of multivariable frequency-domain analysis cannot but be helpful to the user of state-space design methods.

The intention of this section is certainly not to provide an introduction to multivariable frequency-domain *design* methods. Several textbooks on this subject, as discussed in Sec. 4.1 and Note 4.1, are available to the interested reader. This section is intended rather to introduce some of the concepts of multivariable frequency-domain analysis that are useful to the engineer who is using state-space methods for design. In particular, it is often necessary to assess the robustness of a design—to find the gain and phase margins as discussed in Sec. 4.9. But what are the gain and phase margins in a multivariable system? This section is addressed to such questions.

Poles and zeros The difference between a scalar system and a multivariable system becomes apparent when we try to define the poles and, more importantly, the *zeros* of a multivariable system. In a scalar system, having the transfer function

$$H(s) = \frac{b_0 s^k + b_1 s^{k-1} + \cdots + b_k}{s^k + a_1 s^{k-1} + \cdots + a_k} = \frac{N(s)}{D(s)} \tag{4.82}$$

the poles of the system are the frequencies s_p at which the denominator is zero

$$D(s) = s^k + a_1 s^{k-1} + \cdots + a_k = 0 \tag{4.83}$$

and the zeros of the system are the frequencies at which the numerator is zero

$$N(s) = b_0 s^k + b_1 s^{k-1} + \cdots + b_k = 0$$

A multivariable system, however, is represented by a *matrix* of transfer functions

$$H(s) = \begin{bmatrix} H_1(s) & \cdots & H_{1m}(s) \\ \cdots\cdots & \cdots & \cdots\cdots \\ H_{l1}(s) & \cdots & H_{lm}(s) \end{bmatrix} \tag{4.84}$$

where m is the number of inputs to the system and l is the number of outputs. Every entry in the matrix represents the transfer function between one of the inputs and one of the outputs, and is the ratio of polynomials in s in the form of (4.82). Since each of these transfer functions has its own set of poles and zeros, how are the poles and zeros of the entire *system* to be defined? The issue is readily settled for poles: The poles of the system can be defined as the totality of the poles of all the transfer functions in the matrix (4.84). This definition of the system poles is equivalent to expressing the transfer functions as (possibly different) numerator polynomials in s all over a single common denominator polynomial. If the matrix H(s) of the system is obtained starting from a state-space model, as described in Sec. 3.5, this common denominator is the characteristic polynomial $|sI - A|$.

It is reasonable to define the poles of the system as the collection of all the poles of the transfer functions in H(s) because at any pole at least one transfer function becomes infinite and hence the matrix H(s) cannot be said to exist. On the other hand, the entire transfer matrix does not become zero at a zero of one of the elements in H(s). Thus it would not be appropriate to define the zeros of an entire system as the collection of the zeros of all the individual transfer function $H_{ij}(s)$. On the other hand, it would not make much sense to define the zeros of the system as those frequencies at which the matrix H(s) becomes zero, for then unless all the transfer functions have a common factor (a very unusual condition), the system H would have no zeros at all!

What is needed here is a definition of the zeros of a multivariable system which is a natural generalization of the zeros of a scalar system. One way of interpreting the zeros of a scalar system is as the poles of the *inverse* system, the transfer function of which is

$$H^{-1}(s) = \frac{D(s)}{N(s)}$$

If the multivariable system is "square," i.e., if there are exactly the same number of outputs as inputs, then H(s) will be a square matrix and, in general, will an inverse except at isolated (complex) frequencies. It is appropriate to call these frequencies the zeros of the system. Since the condition for H(s) to have an inverse is that the determinant of the transfer matrix be nonzero

$$|H(s)| = \begin{vmatrix} H_{11}(s) & \cdots & H_{1m}(s) \\ \cdots\cdots & \cdots & \cdots\cdots \\ H_{l1}(s) & \cdots & H_{lm}(s) \end{vmatrix} \neq 0$$

we can say that the zeros of a square system are the zeros of the determinant $|H(s)|$.

A "nonsquare" system, i.e., one in which the number of inputs is not equal to the number of outputs, is more of a problem. One way of addressing the problem, when there are more inputs than outputs, is to define several more independent outputs (assuming that the order of the system is high enough to permit this). This is done by adding rows to the observation matrix C, independent of those already present, so that the number of rows in C equals the number of columns in B. Although these added outputs may not be of particular interest, this artifice permits use of any of the theory developed for square systems. This technique will not avail, however, when there are fewer inputs than outputs. Adding inputs that are not physically present is not permissible since the resulting mathematical model would no longer represent the physical process. A more general definition of the zeros of a nonsquare system avoids the need for the artifice of adding inputs or outputs. We define the zeros as those (complex) frequencies at which the *rank* of the transfer function matrix $H(s)$ is reduced. Normally the transfer-function matrix will be of "full rank," i.e., $\text{rank}[H(s)] = \min[l, m]$ except at the zeros of the system at which $H(s)$ will drop to less than full rank. (See Note 4.4.)

The return difference and the multivariable Nyquist diagram In Sec. 4.9 we found that the phase and gain margins of a scalar system, traditional measures of system robustness, can be determined by an examination of the behavior of the return difference as a function of frequency. It turns out that the manner in which the return difference varies with frequency also provides a means of measuring robustness in a multivariable system. To develop the concept of robustness for multivariable systems we need to clarify the notion of return difference and find a useful and convenient way of characterizing its behavior with frequency.

In Sec. 4.3 (block-diagram algebra) we saw that it is generally possible to express the transfer function from the input of a feedback system to its output in the form of the product of a "forward transmission matrix" and the inverse of another matrix of the form $T(s) = I + G(s)$, where $G(s)$ is the "loop transmission." The matrix $T(s)$ was called the "return difference." This is the matrix that we investigate to assess the robustness of a multivariable system.

An important feature of the return difference $T(s)$ is that it is the sum of an identity matrix (which of course is square) and the loop transmission $G(s)$. In order for the sum of I and $G(s)$ to be defined, $G(s)$ must also be a square matrix of the same dimension as I. Hence the return difference $T(s)$ is a square matrix and there is no problem defining its zeros: they are the zeros of the determinant

$$|T(s)| = |I + G(s)|$$

Since the inverse of the return-difference matrix appears as a factor in the transfer-function matrix of the closed-loop system, the poles of the latter will include the zeros of the return difference. Naturally we expect the return difference to be zero at some (complex) values of s. If the system is stable, these zeros will occur only in the left half-plane; if the system is unstable, however, one or more zeros of the return difference will occur in the right half-plane. The question that an analysis of robustness seeks to answer is how much a parameter of a stable system can be permitted to vary before the system becomes unstable. In a scalar system, we can estimate the robustness by using the classical stability margins which can be determined by examining the behavior of $T(s)$. But by Nyquist's method, we only need to investigate $T(s)$ for $s = j\omega$, that is, for s on the imaginary axis. In particular, the minimum value of the magnitude of the return difference indicates how close the Nyquist diagram comes to the "critical" $-1 + j0$ point (see Fig. 4.31) and is a pretty good measure of robustness: The larger the minimum value of the magnitude of the return difference, the more robust (i.e., tolerant of loop gain variations) the system.

The determinant of the return-difference matrix in a multivariable system plays the role of the return difference itself of a scalar system. It would be quite natural to assume that the robustness of a multivariable system can be determined by studying the Nyquist diagram for the determinant of the return difference of the system. A polar plot of the determinant of the return difference $|T(j\omega)|$ as ω is varied from $-\infty$ to ∞ may be termed the *multivariable Nyquist diagram* for the closed-loop system. To make the plot resemble a scalar Nyquist diagram we can place the origin at the critical point $-1 + j0$. This is the same as writing

$$|T(j\omega)| = 1 + G(j\omega) \tag{4.85}$$

and obtaining the Nyquist plot for $G(j\omega)$. In the multivariable case, however, $G(j\omega)$ cannot be interpreted at the "open-loop" transmission.

Singular value analysis Although the multivariable Nyquist plot is fairly easy to obtain, especially with the aid of a computer, it often does not tell enough about the robustness of the system under investigation, because the determinant of a matrix is not always a good indicator of how near that matrix comes to being singular. And that is what we want to determine for the return difference matrix $T(s)$.

To see why the determinant of a matrix may be a poor indicator of how near the matrix comes to being singular, consider the matrix

$$M = \begin{bmatrix} 1 & 0 \\ 1/\varepsilon & 1 \end{bmatrix} \tag{4.86}$$

This matrix has a determinant of unity independent of ε, yet is only an ε away from being singular: Replace the zero in the upper right-hand corner of M by ε and M is singular. The eigenvalues of a matrix are scarcely a better measure of the incipient singularity. In the case of M of (4.86), for example, the eigenvalues of M are both unity, and do not provide an inkling into the near-singularity of M.

A better measure of the near-singularity of a matrix is the set of *singular values* of the matrix, defined as the square-roots of the eigenvalues of M^HM where M^H is the transpose of the complex conjugate of M. (These eigenvalues are always positive and real. See Note 4.5.) When M is a real matrix $M^H = M'$. But since the matrices we will be considering are in general complex functions of frequency (when $s = j\omega$), it is necessary to make use of the more general form. Thus, in a multivariable system, instead of investigating the behavior of the determinant of $\mathrm{T}(j\omega)$ or even of the eigenvalues of $\mathrm{T}(j\omega)$, it is more appropriate to investigate the singular values of $\mathrm{T}(j\omega)$, that is, the square roots of the eigenvalues of

$$\mathrm{S}(\omega) = \mathrm{T}^H(j\omega)\mathrm{T}(j\omega) = \mathrm{T}'(-j\omega)\mathrm{T}(j\omega) \qquad (4.87)$$

(Note that $\mathrm{T}^H(j\omega) = \mathrm{T}'(-j\omega)$ because $\mathrm{T}(s)$ is a real function of s, that is, every element of $\mathrm{T}(s)$ is real when s is real.)

As an illustration of the advantage of singular values over eigenvalues as a measure of incipient singularity, we find the singular values of the matrix M in (4.86). Since M is real, we need the eigenvalues of

$$S = M'M = \begin{bmatrix} 1 & 1/\varepsilon \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1/\varepsilon & 1 \end{bmatrix} = \begin{bmatrix} 1 + 1/\varepsilon^2 & 1/\varepsilon \\ 1/\varepsilon & 1 \end{bmatrix}$$

The characteristic equation of S is

$$|\lambda I - S| = \begin{bmatrix} \lambda - 1 - 1/\varepsilon^2 & -1/\varepsilon \\ -1/\varepsilon & \lambda - 1 \end{bmatrix} = \lambda^2 - \lambda\left(2 + \frac{1}{\varepsilon^2}\right) + 1$$

and the eigenvalues are

$$\lambda_1 = 1 + \frac{1}{2\varepsilon^2} + \frac{1}{2\varepsilon^2}\sqrt{1 + 4\varepsilon^2}$$

$$\lambda_2 = 1 + \frac{1}{2\varepsilon^2} - \frac{1}{2\varepsilon^2}\sqrt{1 + 4\varepsilon^2}$$

The larger eigenvalue λ_1 tends to infinity as $\varepsilon \to 0$. But the smaller eigenvalue λ_2 tends to zero, and as a consequence, one of the singular values $\sigma_2 = \lambda^{1/2} \to 0$ as $\varepsilon \to 0$, thus providing an indication that M becomes singular as $\varepsilon \to 0$. Note also that the two singular values $\sigma_1 = \lambda_1^{1/2}$ and σ_2 separate more and more as $\varepsilon \to 0$. This is an indication of the "ill-conditioning" of M: Not only does it become singular as $\varepsilon \to 0$, but it also grows (i.e., one of its elements grows) to infinity.

The singular values of the return difference matrix $\mathrm{T}(j\omega) = I + \mathrm{G}(j\omega)$ can be used to estimate the gain and phase margins of a multivariable system. A number of theorems have been developed since the late 1970s for estimating these margins (see Note 4.6). A typical theorem is the following [16]:

Let $\underline{\sigma}[\mathrm{T}(j\omega)]$ denote the smallest of the singular values of the return difference at the frequency ω. Suppose that there is a constant $\alpha \leqq 1$ such that $\underline{\sigma}[\mathrm{T}(j\omega)] \geqq \alpha$ for all frequencies. Then there is a *guaranteed gain margin*

$$GM = \frac{1}{1 \pm \alpha} \qquad (4.88)$$

and a *guaranteed phase margin*

$$PM = \pm 2 \sin^{-1}\left(\frac{\alpha}{2}\right) \qquad (4.89)$$

A number of theorems related to the one cited above have been developed during the early 1980s. This subject is likely to continue to receive a great deal of attention by researchers of the decade.

One of the major problems of singular value analysis is that the stability margins that it guarantees are *extremely conservative*, because they allow for the simultaneous magnitude and phase variations of *any* of the gains at which the loop is closed. In other words,.if the off-nominal return difference is

$$\bar{\mathrm{T}}(s) = I + D\mathrm{G}(s)$$

the singular value analysis tends to seek out the least favorable variation of the matrix D that represents the departure of the plant from its nominal, or design, value. The matrix D in practice, however, is anything but arbitrary. The singular value analysis may predict very small gain or phase margins on the basis of unlikely, or even impossible, plant variations.

A more reasonable assessment of robustness would take into account only those variations in system parameters that can actually occur in reality. If the total range of variation of the system can be represented by a single parameter, say μ, which ranges between 0 and 1, that is, the plant transfer matrix is $\mathrm{G}(s, \mu)$ with $\mathrm{G}(s, 0) = \bar{\mathrm{G}}(s)$ being the "nominal" system and $\mathrm{G}(s, 1)$ being the transfer matrix for the largest possible variation of the physical parameters, then there is a general theorem [16] that asserts that the system is stable for all values of $0 \leqq \mu \leqq 1$ if, and only if, the multivariable Nyquist plot for $1 + \mathrm{G}(s, 0)$ deforms "smoothly" into the Nyquist plot for $1 + \mathrm{G}(s, 1)$ and at no time covers the critical point $-1 + j0$.

Example 4H Two-loop control of distillation column The distillation column introduced in Example 2G and discussed later in Examples 4A and 4C has two control inputs u_1 (the steam flow rate) and Δs (the vapor side-stream flow rate) and may be regarded as having two outputs Δz_1 and Δz_2, the positions of the "interphase fronts." This process thus seems a natural candidate for a two-loop control system, as shown in Fig. 4.35. The side-stream flow rate Δs is controlled by the displacement Δz_1 of the front between the water and the propanol, and the steam flow rate Δu_1 is controlled by the displacement Δz_2 of the front between the water and the glycol. A more general control in which each control input is a linear combination of the two control outputs would probably be employed in practice, but the resulting complexity of the overall system in this case would obscure the analysis of this example, the objective of which is to illustrate the multivariable Nyquist diagram and singular-value estimation of gain and phase margins.

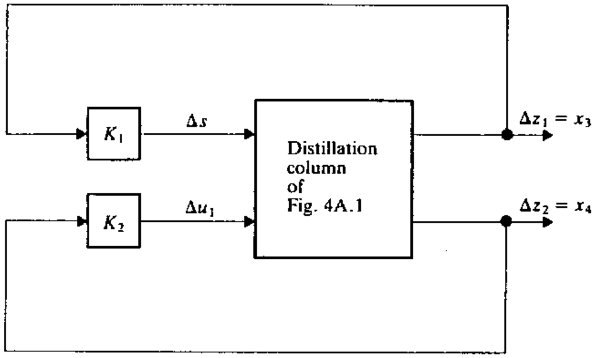

Figure 4.35 Two-loop control of distillation column.

Using the numerical data of Example 2G in the transfer matrix determined in Example 4A, we find that the transfer matrix for the plant is

$$H(s) = \begin{bmatrix} \dfrac{3.04}{s} & \dfrac{-278.2}{s(s+6.02)(s+30.3)} \\ \dfrac{0.052}{s} & \dfrac{-206.6}{s(s+6.02)(s+30.3)} \end{bmatrix} \tag{4H.1}$$

with the diagonal matrix

$$K = \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix}$$

implied by the structure of Fig. 4.35, and we find the return-difference matrix

$$T(s) = I + KH(s) = \begin{bmatrix} 1 + \dfrac{3.04K_1}{s} & -\dfrac{278.2K_1}{s(s+6.02)(s+30.3)} \\ \dfrac{0.052K_2}{s} & 1 - \dfrac{206.6K_2}{s(s+6.02)(s+30.3)} \end{bmatrix} \tag{4H.2}$$

The closed-loop poles are determined by setting the determinant of the return difference to zero

$$0 = |T(s)| = \left(1 + \frac{3.04K_1}{s}\right)\left(1 - \frac{206.6K_2}{s(s+6.02)(s+30.3)}\right) + \frac{(278.2)(0.052)K_1K_2}{s^2(s+6.02)(s+30.3)} \tag{4H.3}$$

From (4H.3) we obtain the characteristic equation of the system

$$\begin{aligned} \Delta(s) &= s^2(s+6.02)(s+30.3) + 3.04K_1 s(s+6.02)(s+30.3) - 206.6K_2 s - 613.9K_1K_2 \\ &= s^4 + (36.62 + 3.04K_1)s^3 + (182.4 + 110.4K_1)s^2 + (554.5K_1 - 206.2K_2)s - 613.9K_1K_2 \\ &= 0 \end{aligned} \tag{4H.4}$$

The ranges of K_1 and K_2 for which the closed-loop system is stable can be found using the Routh array or the Hurwitz matrix described in Sec. 4.5. In particular, the Hurwitz matrix for this example is

$$H = \begin{bmatrix} 36.32 + 3.04K_1 & 554.5K_1 - 206.6K_2 & 0. & 0. \\ 1. & 182.4 + 110.4K_1 & -613.6K_1K_2 & 0. \\ 0. & 36.32 + 3.04K_1 & 554.5K_1 - 206.6K_2 & 0. \\ 0. & 1. & 182.4 + 110.4K_1 & -613.6K_1K_2 \end{bmatrix}$$

By the Hurwitz matrix criterion, stability of the closed loop system is assured if

$$\begin{aligned} D_1 &= 36.32 + 3.04K_1 > 0 \\ D_2 &= \begin{vmatrix} 36.32 + 3.04K_1 & 554.5K_1 - 206.6K_2 \\ 1. & 182.4 + 110.4K_1 \end{vmatrix} > 0 \\ D_3 &= (554.5K_1 - 206.6K_2)D_2 + 613.6K_1K_2(36.32 + 3.04K_1)^2 > 0 \\ D_4 &= -613.6K_1K_2D_3 > 0 \end{aligned} \tag{4H.5}$$

Numerical analysis (Prob. 4.6) reveals that the four inequalities of (4H.5) are simultaneously satisfied in a region that is nearly rectangular and given by

$$0 \leqq K_1 \leqq 11.94$$
$$0 \leqq -K_2 \leqq \sim 32.1$$

(See Fig. 4.36.)

A "comfortable" operating point would be at

$$K_1 = 5 \qquad K_2 = -10$$

which provides a gain margin of over 2 for K_1 and over 3 for K_2. The return-difference matrix at this operating point is given by

$$\begin{aligned} T(s) &= \begin{bmatrix} 1 + \dfrac{15.2}{s} & -\dfrac{1392.5}{s(s+6.02)(s+30.3)} \\ -\dfrac{0.52}{s} & 1 + \dfrac{2066.}{s(s+6.02)(s+30.3)} \end{bmatrix} \\ &= \begin{bmatrix} \dfrac{s+15.2}{s} & \dfrac{-1392.5}{s(s+6.02)(s+30.3)} \\ \dfrac{-0.52}{s} & \dfrac{s^3 + 36.32s^2 + 182.4s + 2066.}{s(s+6.02)(s+30.3)} \end{bmatrix} \end{aligned}$$

This is the matrix we will analyze using the methods of this section.

First we obtain the "multivariable Nyquist plot," i.e., the magnitude and phase of the determinant of the return difference:

$$\begin{aligned} |T(s)| &= \frac{(s+15.2)(s^3 + 36.32s^2 + 182.4s + 2066.) - (0.52)(1392.5)}{s^2(s+6.02)(s+30.3)} \\ &= \frac{s^4 + 5.52s^3 + 734.5s^2 + 4839s + 30\,686.}{s^4 + 36.32s^3 + 182.4s^2} \\ &= 1 + G(s) \end{aligned}$$

where

$$G(s) = \frac{15.2s^3 + 552.06s^2 + 4839.s + 30\,686.}{s^4 + 36.32s^3 + 182.4s^2} \tag{4H.6}$$

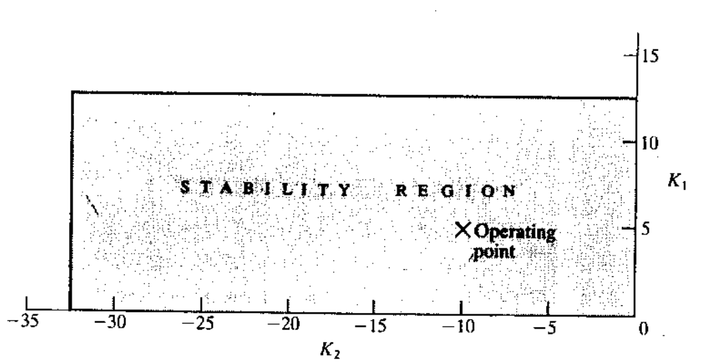

Figure 4.36 Stability region for two-loop control of distillation column.

The (multivariable) Nyquist plot for G(s) as given by (4H.6) is shown in Fig. 4.37. The plot starts with a phase shift of slightly more than 180° and crosses the 180° line (at a frequency of about 6 cycles per hour) at $\sigma \approx -4.7$. A scalar system with this Nyquist plot would be only conditionally stable (with a gain reduction margin of 4.7). This is somewhat misleading, because we know that the plant is open-loop stable and that both gains can be reduced to zero (individually and simultaneously) without compromising stability. The Nyquist plot also seems to imply an infinite margin for gain increase and this again is misleading because we have already determined that there is a finite upper limit on each of the gains. Finally, we note that the Nyquist diagram suggests a phase margin of about 65°. But how is this phase margin to be interpreted?

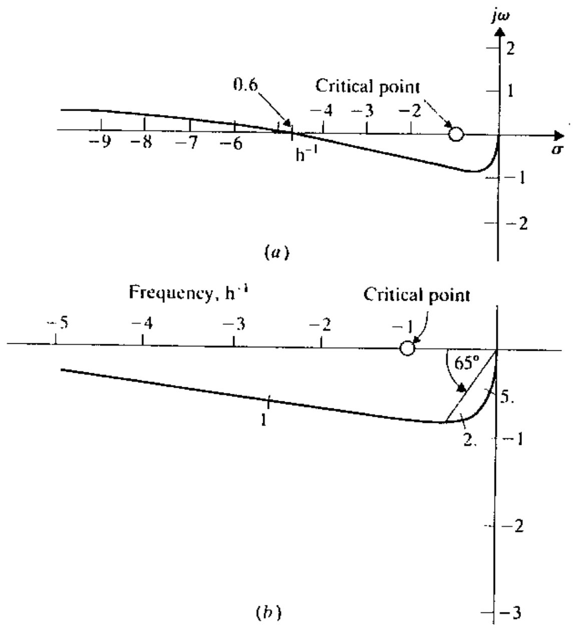

Figure 4.37 Multivariable Nyquist plot for two-loop control of distillation column. (a) System appears to be conditionally stable; (b) Apparent phase margin of 65°.

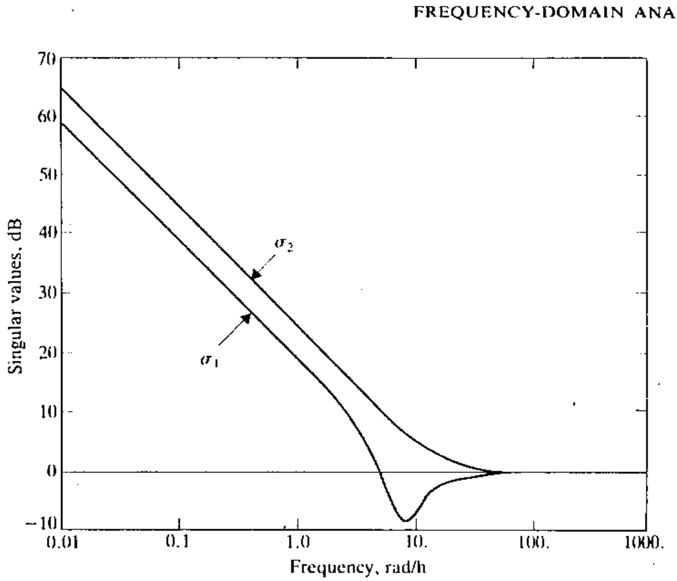

Figure 4.38 Singular values for two-loop control of distillation column.

If we examine how G(s) is computed from T(s) we can readily explain this apparent paradox. The characteristic equation (4H.3) contains K_1, K_2, and the product K_1K_2. We are trying to assess the effect of changes in K_1 and K_2 as if they resulted from the variation of a single parameter. It ought not to be too surprising to find that an apparent gain reduction margin in the Nyquist diagram is the effect of positive gain margins on K_1 and K_2.

The numerically computed plots (using the customary dB vs. log frequency scale) of the two singular values of the return difference are shown in Fig. 4.38. The smaller of the two singular values reaches a minimum of −8.16 dB at the frequency of 7.9 rad/h (1.26 cycles per hour). The corresponding value of $\alpha = 0.391$, and hence by (4.88) and (4.89) the guaranteed gain and phase margins, are

$$GM = \begin{cases} \dfrac{1}{1+0.391} = 0.718 \\ \dfrac{1}{1-0.391} = 1.64 \end{cases} \tag{4H.7}$$

and

$$PM = \begin{cases} 2\sin^{-1}\dfrac{0.391}{2} = 22.5° \\ -2\sin^{-1}\dfrac{0.391}{2} = -22.5° \end{cases} \tag{4H.8}$$

Considering that the gain and phase margin estimates are known to be very conservative, the predictions of (4H.7) and (4H.8) are not too bad. In particular, the upper margin of 1.64 compares quite favorably with the true margin of 11.94/5 = 2.39.

PROBLEMS

Problem 4.1 Closed-loop transfer function for state feedback

Consider the control process of Fig. P4.1, corresponding to the process

$$\dot{x} = Ax + Bu$$

$$y = Cx$$

"State-variable" feedback is used

$$u = u_0 - Gx$$

(*a*) Aided by block-diagram analysis, show that the transfer function from u_0 to y is given by

$$H(s) = C\Phi(s)B[I + G\Phi(s)B]^{-1} \qquad \text{(P4.1)}$$

where

$$y(s) = H(s)u(s)$$

and $\Phi(s) = (sI - A)^{-1}$ is the resolvent.

(*b*) Show that (P4.1) implies that the transmission zeros of the process, i.e., the zeros of $|C\Phi(s)B|$ are not altered by state-variable feedback.

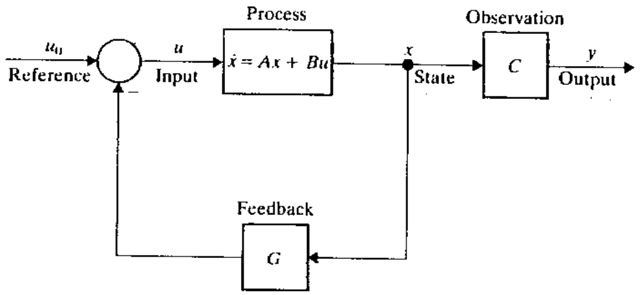

Figure P4.1 Control system with state-variable feedback.

Problem 4.2 Three-capacitance thermal system: feedback control

It is desired to control the temperature at point 1 on the insulated rod of Prob. 2.3 et seq. For this purpose a temperature sensor (e.g., thermocouple) is attached to the rod, and the input power is varied in proportion to the difference $d = T_e - T_1$. In the electrical analog the feedback law is

$$e_0 = u = gd \qquad d = e_R - v_3$$

(*a*) Find the closed loop transfer functions to the output from the reference input and from the disturbance

$$H_1(s) = \frac{v_3(s)}{e_R(s)} \qquad H_2(s) = \frac{v_3(s)}{v_0(s)}$$

respectively.

(*b*) Find the range of the feedback gain g for which the closed-loop system is stable. (Why is it not stable for all g?)

(*c*) Draw the root locus of the system.

(*d*) Draw the Nyquist and Bode diagrams.

(*e*) As a function of g, determine the steady state error $d_{ss} = e_R(\infty)$ when $v_0 \neq e_R$.

Use the following numerical data

$$R = 1 \qquad C = 2$$

Problem 4.3 Three-capacitance thermal system: PI control

The presence of a steady state error in the system of Prob. 4.2 makes the control law unsuited for precise temperature control. To improve the steady state performance, proportional + integral (PI) control is to be used. The transfer function of the "compensator" is to be

$$\frac{u(s)}{d(s)} = g_1 + \frac{g_2}{s}$$

instead of g as used in Prob 4.2.

(*a*) Find the range of gains g_1 and g_2 for which the closed-loop system is stable.

(*b*) For $g_2/g_1 = 1$, draw the root locus of the system with g_1 as the variable gain.

(*c*) Draw the Nyquist and Bode diagrams corresponding to part (*b*).

Problem 4.4 Aircraft lateral dynamics: modes and transfer functions

The aerodynamics data for a fighter aircraft are as follows

$$\frac{Y_\beta}{V} = -0.746 \quad \frac{Y_p}{V} = 0.006 \quad \frac{Y_r}{V} = 0.001 \quad \frac{g}{V} = 0.0369 \quad \frac{Y_A}{V} = 0.0012 \quad \frac{Y_R}{V} = 0.0092$$

$$L_\beta = -12.9 \quad L_p = -0.746 \quad L_r = 0.387 \quad L_A = 6.05 \quad L_R = 0.952$$

$$N_\beta = 4.31 \quad N_p = 0.024 \quad N_r = -0.174 \quad N_A = -0.416 \quad N_R = -1.76$$

(*a*) Using the state vector $x = [\beta, p, r, \phi]'$ as given in (2.46), write the A and B matrices.

(*b*) The eigenvalues for the lateral motion of an aircraft consist, typically, of two complex poles with relatively low damping, and a pair of real poles. The complex pair defines a mode called *dutch roll*. One real pole, relatively far from the origin, defines a mode called *roll subsidence*, and a real pole near the origin defines the *spiral* mode. (The latter is sometimes unstable—spiral divergence.) Using the data given above find the four modes for this aircraft.

(*c*) A stability augmentation system (SAS) is to be designed for this aircraft using two-rate gyros, each of which measures one of the bodies rated p and r. Find the transfer functions

$$\frac{p(s)}{\delta_A(s)} \quad \frac{r(s)}{\delta_A(s)} \quad \frac{p(s)}{\delta_R(s)} \quad \frac{r(s)}{\delta_R(s)}$$

Is it apparent from these transfer functions why the ailerons are used for roll (p) control and the rudder is used for yaw (r) control?

(*d*) Find the transmission zeros of the process.

Problem 4.5 Aircraft longitudinal dynamics, simplified

The aerodynamic coefficients for an aircraft are approximated by

$$\frac{Z_\alpha}{V} = -1 \qquad \frac{Z_E}{V} = -0.1$$

$$M_q = -0.5 \qquad M_\alpha = -5. \qquad M_E = -9.$$

$$\frac{X_\alpha}{V} = -14. \qquad \frac{X_E}{V} = -1.$$

All other coefficients are negligible.

(*a*) Find the open-loop poles (short-period and phugoid modes) of the aircraft.

(*b*) Find the transfer functions of the aircraft from δ_E to θ and u.

Problem 4.6 Distillation column

By use of the Hurwitz matrix, verify that the range of gains for stability of the closed-loop system of Example 4G is as shown in Fig. 4.36.

Problem 4.7 Double-effect evaporator

Consider the double-effect evaporator introduced in Example 2H, with the dynamics matrices as defined by (2H.5)-(2H.7).

(*a*) Find the open-loop poles (eigenvalues) of the system.

(*b*) The observed quantities ("outputs") are $y_1 = x_1$ (first-effect holdup) and $y_2 = x_4$ (second-effect holdup). What is the observation matrix? Find the transfer functions from the controls to the outputs.

Problem 4.8 Double-effect evaporator: feedback control

A two-loop control system is proposed for the double-effect evaporator of Prob. 4.7 in which the first-effect holdup x_1 is controlled by the steam-flow rate u_1 and the second-effect holdup x_4 is controlled by the first-effect bottoms flow rate u_2. The resulting system has the structure shown in Fig. P4.8.

(*a*) Find the range of gains g_1 and g_2 for which the closed-loop system is stable.

(*b*) Let the return difference for the process be given by

$$T(s) = I + GH(s)$$

where $G = [g_1, g_2]$ and H(s) is the 2×2 transfer-function matrix. Plot the singular values of the system as a function of frequency, with the loops opened at the input. Use a nominal value $G = \bar{G} = [-40, 40]$.

(*c*) Repeat part (*b*) with a nominal gain matrix $\bar{G} = [-20, 10]$.

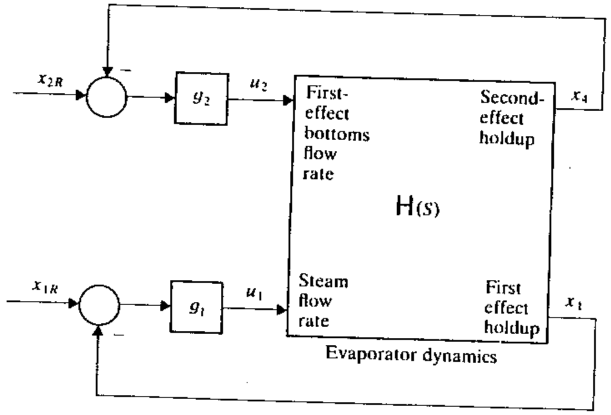

Figure P4.8 Control system for double-effect evaporator.

Problem 4.9 Gun turret: range of gain for stability

Using the Routh table or the Hurwitz matrix criterion, find the range of gains for which the turret control system of Example 4D is asymptotically stable.

Problem 4.10 Missile autopilot: Bode and Nyquist diagrams

Draw thę Bode and Nyquist diagrams for the missile autopilot of Example 4F.

Problem 4.11 Missile autopilot: Acceleration and angular rate feedback

In addition to the feedback of the normal acceleration error $e = a_{NC} - a_N$, a a missile autopilot will frequently also make use of the pitch rate. This will thus result in a control law

$$u = K_1(a_{NC} - a_N) - K_2 q \qquad \text{(P4.11}a\text{)}$$

(*a*) Using this control law, find the range of K_1 and K_2 for which the closed-loop system is stable.

(*b*) In order to implement the control law of (P4.11*a*) an additional sensor (a rate gyro to measure q) is needed. What are the benefits that such a control law might confer on the system that would justify the additional cost of the sensor?

NOTES

Note 4.1 Frequency-domain analysis

The frequency-domain approach to control system analysis and design which was developed during the 1940s and 1950s is often called the "classical" approach to distinguish it from the "modern" state-space approach which had its beginnings in the late 1950s and early 1960s.

Not all investigators agreed on the advantages of the state-space approach over the frequency-domain approach, and a significant minority remain unconvinced to the present day. The complaint of the frequency-domain advocates is that the reason for the use of feedback is the uncertainties in the dynamic process, and that when these uncertainties are present, the qualitative methods of frequency-domain analysis are more appropriate. Qualitative system properties such as bandwidth, stability margins, etc., were regarded as difficult to study by state-space methods. To answer the need for computational design tools for multivariable systems that would rival the state-space tools in power, the classicists developed such techniques as multivariable root loci, multivariable Nyquist plots, and various subsidiary techniques. Much of the theoretical results and most of the design software is the product of the efforts of investigators of the United Kingdom, led by H. H. Rosenbrock[17] of the University of Manchester and A. J. G. MacFarlane[18] of the University of Cambridge. I. Horowitz of the Weitzmann Institute (Israel) is another leading exponent of the classical approach. Since state-space concepts have been included in many of the newer frequency-domain methods, it might be appropriate to call this work "neoclassical." Neoclassical frequency-domain activity of the western hemisphere is represented by the theoretical work of C. A. Desoer of the University of California (Berkeley) and G. Zames of McGill University.

Note 4.2 Aircraft dynamic modes

The terminology of aircraft dynamics stems from the 1930s and 1940s. The longitudinal modes are called "short-period" and "phugoid"; the lateral modes are "dutch-roll," "spiral," and "roll subsidence." The background of this terminology is given by Etkin.[19]

Note 4.3 Bandwidth of multivariable systems

R. W. Bass[15] has defined the bandwidth of a system in state-space form as

$$W = (\|A\|/\|A^{-1}\|)^{1/2}$$

and has proved that $\|\Phi(j\omega)\Phi^{-1}(0)\| \leqq 1/\sqrt{2}$ for $\omega \gg W$, where $\Phi(j\omega) = (j\omega I - A)^{-1}$, i.e., the resolvent of the system at $s = j\omega$. These inequalities are generalizations of the concept of bandwidth for scalar systems.

A shortcoming of this definition of bandwidth is that neither the control distribution matrix B nor the observation matrix C enter into the definition. Thus this definition depends only on the poles, and not on the zeros, of the system.

Note 4.4 Transmission zeros

It is said with considerable justification that state-space methods are concerned primarily with the poles of a system rather than with its zeros. This is surely one of the reasons that transmission zeros play a much larger role within the classical and neoclassical (frequency-domain) methodology than they do in the state-space methodology. In the latter the compensator is designed by the separation principle: first a "full-state" feedback law is designed to estimate those states that are not directly measured; then an observer is designed to estimate the missing states. In the first step only the A and B matrices are used; in the second step only the A and C matrices are used. The only place A, B, and C are brought together is when the full-state feedback law is combined with the observer to yield the required compensator. Since the transmission zeros depend on A, B, and C together, through $H(s) = C(sI - A)^{-1}B$, the state-space approach obscures the transmission zeros. Since the behavior of a system depends not only on its poles, but also on its zeros, the absence of a clear connection in the state-space methodology between the compensator design and the transmission zeros is a shortcoming of this methodology and suggests a possible direction for future research.

As one might expect, transmission zeros receive the greatest attention in books emphasizing the frequency-domain methodology. In particular, see [17] and [18].

Note 4.5 Singular-value analysis

The singular values of a matrix are of particular importance in determining whether a matrix is relatively easy to invert ("well conditioned") or difficult to invert ("ill conditioned"). They are consequently of special interest in the branch of numerical analysis that is concerned with algorithms for the manipulation of large matrices. Singular-value analysis is prominent in books on numerical methods, such as Householder.[20]

Note 4.6 Robustness of multivariable control systems

The study of robustness of multivariable control systems by means of singular-value analysis is represented by the work of a number of investigators centered at the Massachusetts Institute of Technology beginning in the late 1970s with the doctoral research of M. G. Safonov.[21] A number of papers that make use of singular-value analysis as an analytical tool for multivariable control systems are included in the Special Issue on Linear Multivariable Control Systems of the *IEEE Transactions on Automatic Control*.[22]

Note 4.7 Nonminimum phase poles and zeros

It is readily seen that a pole or a zero at $s = -\sigma_0 + j\omega_0$ or at $s = +\sigma_0 + j\omega_0$ will have the same effect on the Bode amplitude characteristic of a system, but will have different effects on the phase characteristic. The phase associated with the pole or zero in the left half-plane is $\phi_1 = \tan^{-1}(\sigma_0/\omega + \omega_0)$ while the phase associated with the pole or zero in the left half-plane is $\phi_2 = \tan^{-1}(-\sigma_0/\omega + \omega_0)$; ϕ_1 is always less than 90 degrees while ϕ_2 is greater than 90 degrees. Thus of the two, the left half-plane pole or zero is the one of *minimum-phase*, a term first used by Bode.[2] A nonminimum phase pole is always indicative of an unstable system. Nonminimum phase zeros, on the other hand, can occur in a stable system, but if they do occur they are often a source of difficulty to the control system designer.

REFERENCES

1. Nyquist, H., "Regeneration Theory," *Bell System Technical Journal*, vol. 11, 1932, pp. 126–147.
2. Bode, H. W., *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand, New York, 1945.
3. Black, H. S., "Inventing the Negative Feedback Amplifier," *IEEE Spectrum*, vol. 14, no. 1, January 1977, pp. 54–60.
4. Schwarz, R. J., and Friedland, B., *Linear Systems*, McGraw-Hill Book Co., New York, 1965.
5. Mason, S. J., "Feedback Theory: Some Properties of Signal Flow Graphs," *Proceedings of the IRE*, vol. 41, no. 9, September 1953.
6. Rynaski, E. J., "Flight Control Synthesis Using Robust Output Observers," *Proc. AIAA Guidance and Control Conference*, San Diego, CA, September 1982, pp. 825–831.
7. Routh, E. J., *A Treatise on the Stability of a Given State of Motion*, Macmillan & Co., London, 1877.
8. Hurwitz, A., "Über die Bedingungen, unter welchen einer Gleichung nur Wurzeln mit negativen reelen Teilen besitzt," *Math. Ann.*, vol. 146, 1895, pp. 273–284.
9. Parks, P. C., "A New Proof of the Routh-Hurwitz Stability Criterion Using the Second Method of Lyapunov," *Proc. Cambridge Philosophical Society*, vol. 58, pt. 4, 1962, pp. 694–702.
10. Lyapunov, M. A., "Le problème général de la stabilité du mouvement," *Ann. Fac. Sci. Toulouse*, vol. 9, 1907, pp. 203–474.
11. D'Azzo, J. J., and Houpis, C. H., *Linear Control System Analysis and Design: Conventional and Modern*, McGraw-Hill Book Co., New York, 1981.
12. Ogata, K., *Modern Control Engineering*, Prentice-Hall Inc., Englewood Cliffs, NJ, 1970.
13. Evans, W. R., "Graphical Analysis of Control Systems," *Trans. AIEE, Pt. II*, vol. 67, 1948, pp. 547–551.
14. Churchill, R. V., *Introduction to Complex Variables and Applications*, McGraw-Hill Book Co., New York, 1948.
15. Bass, R. W., "Robustified LQG Synthesis to Specifications," *Proc. 5th Meeting of Coord. Group On Modern Control Theory, Part II*, Dover, NJ, October 1983, pp. 11–93.
16. Lehtomaki, N. A., Sandell, N. R., Jr., and Athans, M., "Robustness Results on LQG Based Multivariable Control System Designs," *IEEE Trans. on Automatic Control*, vol. AC-26, no. 1, February 1981, pp. 75–93.
17. Rosenbrock, H. H., *Computer Aided Control System Design*, Academic Press, New York, 1974.
18. MacFarlane, A. J. G. (ed.), *Frequency Response Methods in Control Systems*, IEEE Press, New York, 1979.
19. Etkin, B., *Dynamics of Flight*, John Wiley & Sons, New York, 1959.
20. Householder, A. S., *The Theory of Matrices in Numerical Analysis*, Blaisdell Publishing Co., Waltham, MA, 1964.
21. Safonov, M. G., *Stability and Robustness of Multivariable Feedback Systems*, MIT Press, Cambridge, MA, 1980.
22. Special Issue on Linear Multivariable Control Systems, *IEEE Trans. on Automatic Control*, vol. AC-26, no. 1, February 1981.