5
Linear Systems, Theory, and Design: A Brief Review

5.1 Introduction

Generally, a dynamical system is characterized by a differential equation that gives a relation between the input and the output of that system. A dynamical system may be linear or nonlinear. It is said to be linear if the differential equation that characterizes the system is linear.

A differential equation is linear if the coefficients are constants or functions of only the independent variable and not that of the dependent variable. The most important property of the linear systems is the applicability of the principle of superposition, i.e, if $y_1(t)$ and $y_2(t)$ are two solutions to inputs $r_1(t)$ and $r_2(t)$, then the solution to the new input $r(t) = c_1 r_1(t) + c_2 r_2(t)$ is given by $y(t) = c_1 y_1(t) + c_2 y_2(t)$. This feature enables us to build system response to any complex input function by expressing it as a sum of several simple input functions.

A system is said to be nonlinear if the differential equation that characterizes it is nonlinear. A differential equation is nonlinear if it contains products or powers of the dependent variable or its derivatives. Nonlinear differential equations, in general, are quite difficult to solve. Furthermore, the property of superposition does not hold for nonlinear systems.

A control system may be an open- or closed-loop system. An open-loop system is one in which the output has no effect on the input. In other words, in an open-loop system the output is not fed back for comparison with the input for regulation. An open-loop control can be used in practice if the relation between the output and the input is precisely known and the system is not subject to internal parameter variations or external disturbances. A closed-loop system is one in which the output is measured and is fed back to the input for comparison and system regulation. An advantage of the closed-loop or feedback system is that the system response will be relatively insensitive to internal parameter variations or external disturbances. For open-loop systems, stability is not a major concern. However, it is of major concern for a closed-loop system because a closed-loop system may tend to overcorrect itself and in that process develop instability.

In this chapter, we will review the basic principles of linear time-invariant systems and their representation in the transfer function form using Laplace transform. We will also discuss system response to standard inputs such as unit-step function and derive expressions for steady-state errors. Furthermore, we will briefly discuss the frequency response and stability of closed-loop systems and the design of compensators. Finally, we will give a brief exposure to modern state-space analyses and design methods.

5.2 Laplace Transform

For linear systems, the application of Laplace transform enables us to express the given differential equation in an algebraic form that greatly simplifies the analyses of control systems. Obviously this type of simplification is not possible for nonlinear systems. In view of this, one often introduces what is called an equivalent linear system. Such a linearized system is valid for only a limited range of parameter values. The linearization process may have to be repeated several times to cover the entire range of parameter values of interest.

In this section, we will briefly review the main results on Laplace transform that are useful in the analyses and design of aircraft control systems. For more information, the reader may refer to the standard texts on linear systems.[1-3]

The Laplace transform of a function $f(t)$ is defined as

$$L[f(t)] = \bar{f}(s) = \int_0^\infty f(t)e^{-st}\,dt \tag{5.1}$$

where s is a complex variable, equal to $\sigma + j\omega$. Quite often, s is also called the Laplace variable. We assume that $f(t)$ satisfies all the conditions for the existence of its Laplace transform.

The inverse of a Laplace transform is defined as

$$f(t) = L^{-1}[\bar{f}(s)] = \frac{1}{2\pi j}\int_0^\infty \bar{f}(s)e^{st}\,ds \tag{5.2}$$

Some of the important theorems on Laplace transform are summarized in the following.

1) Translated function. The Laplace transform of the translated function $f(t-\alpha)$ (Fig. 5.1), where $f(t-\alpha) = 0$ for $0 < t < \alpha$ can be obtained as follows.

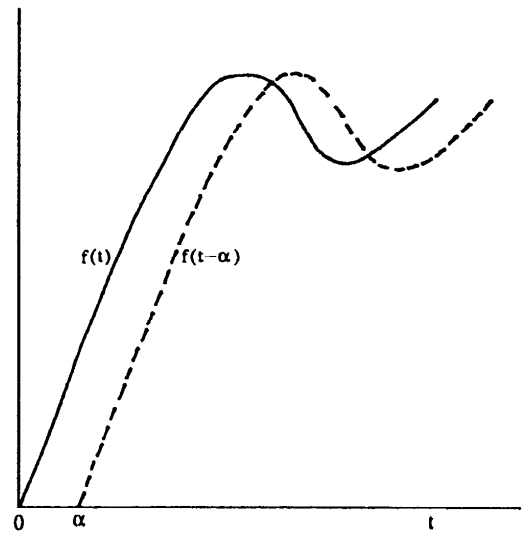
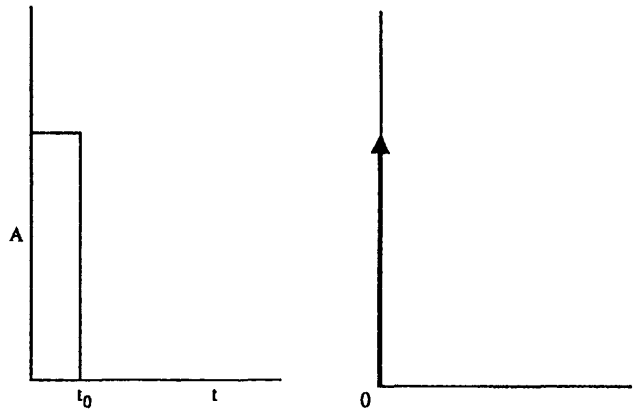


Fig. 5.1 Translated function.



a) Pulse function **b) Impulse function**

Fig. 5.2 Pulse and impulse functions.

We have

$$L[f(\tau)] = \bar{f}(s) = \int_0^\infty f(\tau)e^{-s\tau}\,d\tau \tag{5.3}$$

Let $\tau = t - \alpha$. Then,

$$\bar{f}(s) = \int_0^\infty f(t-\alpha)e^{-s(t-\alpha)}dt = e^{\alpha s}\int_0^\infty f(t-\alpha)e^{-st}dt = e^{\alpha s}L[f(t-\alpha)] \tag{5.4}$$

Therefore,

$$L[f(t-\alpha)] = e^{-\alpha s}\bar{f}(s) \tag{5.5}$$

This theorem states that the translation of the time function $f(t)$ by α corresponds to a multiplication of its transform by $e^{-\alpha s}$.

Using this theorem, one can obtain the Laplace transform of a pulse function as follows. A pulse function (Fig. 5.2a) is given by

$$\begin{aligned} f(t) &= A \qquad 0 \le t \le t_0 \\ &= 0 \qquad t < 0, t_0 < t \end{aligned} \tag{5.6}$$

The given pulse function can be expressed as a sum of two-step functions, each of height A, one positive step function beginning at $t = 0$ and the other negative step function at $t = t_0$. Thus,

$$f(t) = Al(t) - Al(t - t_0) \tag{5.7}$$

where $l(t)$ and $l(t - t_0)$ are the unit step functions originating at $t = 0$ and $t = t_0$,

respectively. The Laplace transform of a unit step function is given by

$$L[f(t)] = \int_0^\infty l(t)e^{-st}\,dt = \int_0^\infty 1e^{-st}\,dt = \frac{1}{s} \tag{5.8}$$

Then,

$$L[f(t)] = L[Al(t)] - L[Al(t-t_0)] = \tfrac{A}{s}(1 - e^{-st_0}) \tag{5.9}$$

The impulse function is a special limiting case of a pulse function. Consider the pulse function given by

$$f(t) = \lim_{t \to 0}\left(\frac{A}{t_0}\right) \qquad 0 \le t \le t_0 \tag{5.10}$$

$$= 0 \qquad t < 0 \qquad t_0 < t \tag{5.11}$$

The height of the impulse is A/t_0 and its duration is t_0. Therefore, the area under the impulse is equal to A. As the duration t_0 approaches zero, the height of the pulse approaches infinity giving us an impulse (Fig. 5.2b). Note that, even though the height of the impulse tends to infinity, the area remains finite.

The Laplace transform of an impulse function is given by

$$L[f(t)] = \lim_{t_0 \to 0}\frac{A}{t_0 s}(1 - e^{-st_0}) = \lim_{t_0 \to 0}\frac{\frac{d}{dt_0}[A(1 - e^{-st_0})]}{\frac{d}{dt_0}(t_0 s)} = A \tag{5.12}$$

An impulse function of infinite magnitude and zero duration is a mathematical fiction. However, if the magnitude of a pulse input is very large and its duration is very small, then we can approximate it as an impulse input. An impulse function whose area is equal to unity is called a unit-impulse function or Dirac delta function. The Laplace transform of a unit-impulse function is unity. A unit-impulse function occurring at $t = t_1$ is usually denoted by $\delta(t - t_1)$, which has the following properties:

$$\delta(t - t_1) = 0 \qquad t \ne t_1 \tag{5.13}$$

$$= \infty \qquad t = t_1 \tag{5.14}$$

$$\int_{-\infty}^{\infty}\delta(t - t_1)\,dt = 1 \tag{5.15}$$

The concept of the unit-impulse is very useful in differentiating discontinuous functions. For example,

$$\delta(t) = \frac{d}{dt}l(t) \tag{5.16}$$

where $\delta(t)$ and $l(t)$ are the unit-impulse and unit-step functions, respectively, and both occur at the origin. Thus, integrating a unit-impulse function, we get a unit-step function. The concept of impulse function helps us represent functions involving multiple discontinuities. Such a representation will have that many impulse

functions as the number of discontinuities and the magnitude of each impulse function will be equal to the magnitude of the corresponding discontinuity.

*2) **Multiplication of** $\mathbf{f(t)}$ **by** $\mathbf{e^{-\alpha t}}$.* The Laplace transform of the function $e^{-\alpha t} f(t)$ is given by

$$L[e^{-\alpha t} f(t)] = \int_0^\infty f(t) e^{-\alpha t} e^{-st}\, dt = \bar{f}(s+\alpha) \tag{5.17}$$

Thus, multiplying the function $f(t)$ by $e^{-\alpha t}$ has the effect of replacing the Laplace variable s by $s+\alpha$. Here, α may be real or complex.

*3) **Change of time scale.*** Suppose the time t is changed to t/α, then

$$L\left[f\left(\frac{t}{\alpha}\right)\right] = \int_0^\infty f\left(\frac{t}{\alpha}\right) e^{-st} dt \tag{5.18}$$

Let $t_1 = t/\alpha$ and $s_1 = \alpha s$. Then,

$$L\left[f\left(\frac{t}{\alpha}\right)\right] = \int_0^\infty f(t_1) e^{-s_1 t_1} d(\alpha t_1) = \alpha \bar{f}(s_1) = \alpha \bar{f}(\alpha s) \tag{5.19}$$

As an example, consider $f(t) = e^{-t}$ so that $f(t/4) = e^{-0.25t}$. We have $\alpha = 4$ and $s_1 = 4s$. Then,

$$L[f(t)] = \bar{f}(s) = \frac{1}{s+1} \tag{5.20}$$

and

$$L\left[f\left(\frac{t}{4}\right)\right] = \alpha \bar{f}(\alpha s) = \frac{4}{4s+1} \tag{5.21}$$

*4) **Differentiation.*** The Laplace transform of the derivative of a function is given by

$$L\left[\frac{d}{dt} f(t)\right] = s\bar{f}(s) - f(0) \tag{5.22}$$

To prove this theorem, we proceed as follows:

$$\int_0^\infty f(t) e^{-st} dt = f(t) \frac{e^{-st}}{-s}\bigg|_{t=0}^{t=\infty} - \int_0^\infty \frac{df(t)}{dt} \frac{e^{-st}}{-s}\, dt \tag{5.23}$$

Then,

$$\bar{f}(s) = \frac{f(0)}{s} + \frac{1}{s} L\left[\frac{df(t)}{dt}\right] \tag{5.24}$$

so that

$$L\left[\frac{df(t)}{dt}\right] = s\bar{f}(s) - f(0) \tag{5.25}$$

In a similar fashion, we can obtain the Laplace transforms of higher order derivatives of $f(t)$. For example,

$$L\left[\frac{d^2 f(t)}{dt^2}\right] = s^2 \bar{f}(s) - sf(0) - \dot{f}(0) \tag{5.26}$$

where $\dot{f}(0)$ is the value of $df(t)/dt$ at $t = 0$.

*5) **Final value theorem.*** This theorem gives

$$f(\infty) = \lim_{t\to\infty} f(t) = \lim_{s\to 0} [s\bar{f}(s)] \tag{5.27}$$

To prove this theorem, take the limit as s approaches zero in Eq. (5.25).

$$\lim_{s\to 0} \int_0^\infty \left[\frac{df(t)}{dt}\right] e^{-st} dt = \lim_{s\to 0} [s\bar{f}(s)] - f(0) \tag{5.28}$$

Because $e^{-st} = 1$ as $s \to 0$, we have

$$\int_0^\infty \left[\frac{df(t)}{dt}\right] dt = f(\infty) - f(0) = \lim_{s\to 0} [s\bar{f}(s)] - f(0) \tag{5.29}$$

Hence,

$$f(\infty) = \lim_{s\to 0} [s\bar{f}(s)] \tag{5.30}$$

This theorem is very useful in determining the steady-state value of a given function using its Laplace transform.

*6) **Initial value theorem.*** The initial value of a function $f(t)$ is given by

$$f(0_+) = \lim_{s\to\infty} s\bar{f}(s) \tag{5.31}$$

To prove this theorem, consider the Laplace transform of $df(t)/dt$ and take the limit as $s \to \infty$ in Eq. (5.25).

$$\lim_{s\to\infty} \int_0^\infty \left[\frac{df(t)}{dt}\right] e^{-st}\, dt = \lim_{s\to\infty} [s\bar{f}(s) - f(0)] \tag{5.32}$$

As s approaches infinity, e^{-st} approaches zero. Hence,

$$f(0) = \lim_{s\to\infty} [s\bar{f}(s)] \tag{5.33}$$

*7) **Integration theorem.*** The Laplace transform of the integral of $f(t)$ is given by

$$L\left[\int f(t)\, dt\right] = \frac{\bar{f}(s)}{s} + \frac{f^{-1}(0)}{s} \tag{5.34}$$

where $f^{-1} = \int f(t)\, dt$ evaluated at $t = 0$.

To prove this theorem, we proceed as follows:

$$L\left[\int f(t)\,dt\right] = \int_0^\infty \left[\int f(t)\,dt\right] e^{-st}\,dt \tag{5.35}$$

$$= \left[\int f(t)\,dt\right]\frac{e^{-st}}{-s}\Bigg|_0^\infty + \frac{1}{s}\int_0^\infty f(t)e^{-st}dt \tag{5.36}$$

$$= \frac{1}{s}\int f(t)\,dt|_{t=0} + \frac{1}{s}\int_0^\infty f(t)e^{-st}dt \tag{5.37}$$

$$= \frac{f^{-1}(0)}{s} + \frac{\bar{f}(s)}{s} \tag{5.38}$$

Hence, the theorem is proved.

*8) **Convolution integral.*** The integral of the form $\int_0^t f_1(t-\tau)f_2(\tau)\,d\tau$ is called the convolution integral and is frequently encountered in the study of control systems.

The Laplace transform of the convolution integral is given by

$$L\left[\int_0^t f_1(t-\tau)f_2(\tau)\,d\tau\right] = \bar{f}_1(s)\bar{f}_2(s) \tag{5.39}$$

For the proof of this theorem, the reader may refer elsewhere.[1,3]

5.3 Transfer Function

Let us consider a linear system represented by the following differential equation:

$$\ddot{y} + a\dot{y} + by = kr(t) \tag{5.40}$$

where a is the damping constant, b is frequency parameter, and $r(t)$ is the input function. We assume $\dot{y}(0) = y(0) = 0$.

Taking the Laplace transform of both sides and using the initial conditions $\dot{y}(0) = y(0) = 0$,

$$s^2\bar{y}(s) + as\bar{y}(s) + b\bar{y}(s) = k\bar{r}(s) \tag{5.41}$$

$$\frac{\bar{y}(s)}{\bar{r}(s)} = \frac{k}{s^2 + as + b} \tag{5.42}$$

Let

$$G(s) = \frac{\bar{y}(s)}{\bar{r}(s)} \tag{5.43}$$

so that

$$G(s) = \frac{k}{s^2 + as + b} \tag{5.44}$$

Here, $G(s)$ is the ratio of the Laplace transform of the output to the Laplace transform of the input and is called the transfer function of the given system. If

the input is a unit-impulse function whose Laplace transform is unity, then the transfer function is equal to the Laplace transform of the output. In other words, by measuring the output for a unit-impulse function input, one can deduce the information on the system transfer function.

5.4 System Response

The system response depends on the order of the system. The order of the system refers to the order of the differential equation representing the physical system or the degree of the denominator of the corresponding transfer function. For example,

$$m\ddot{x} + c\dot{x} + kx = u(t) \tag{5.45}$$

is a second-order differential equation. If

$$G_1(s) = \frac{k}{(s+a)} \tag{5.46}$$

$$G_2(s) = \frac{k}{s^2 + as + b} \tag{5.47}$$

then $G_1(s)$ is a first-order system and $G_2(s)$ is a second-order system.

Generally, the output or response of a system consists of two parts: 1) the natural or free response and 2) forced response. In the following, we discuss the unit-step response of typical first- and second-order systems.

5.4.1 *Response of First-Order Systems*

Consider a typical first-order system

$$G(s) = \frac{s+b}{s+a} \tag{5.48}$$

The response of this system to a unit-step function whose Laplace transform $1/s$ is given by

$$\bar{y}(s) = G(s)\bar{r}(s) \tag{5.49}$$

$$= \left(\frac{s+b}{s+a}\right)\left(\frac{1}{s}\right) \tag{5.50}$$

It is convenient to use the method of partial fractions to factor the right-hand side. Let

$$\frac{s+b}{s(s+a)} = \frac{A}{s} + \frac{B}{s+a} \tag{5.51}$$

Multiply throughout by $s(s+a)$ so that

$$s + b = A(s+a) + Bs \tag{5.52}$$

This identity is supposed to hold for all values of s. Therefore, with $s = -a$, we get $B = (a-b)/a$ and, with $s = 0$, we get $A = b/a$. Then,

$$\bar{y}(s) = \frac{b}{as} + \frac{(a-b)}{a(s+a)} \tag{5.53}$$

Taking the inverse Laplace transform,

$$y(t) = \frac{b}{a} + \left(\frac{a-b}{a}\right)e^{-at} \tag{5.54}$$

We note that $y(0) = 1$ and $y(\infty) = b/a$.

The first term on the right-hand side represents forced response and the second term represents the natural response. The forced response is also known as steady-state response, and the natural response is also known as transient response. Observe that the pole at $s = 0$ corresponding to the input unit-step function generates the forced response. The transient response is generated by the system pole at $s = -a$ and is of the form e^{-at}. Thus, the farther to the left the pole is located on the negative real axis, the faster the transient response will decay to zero. On the other hand, if this pole is located on the positive real axis, then the response will be of diverging nature because the output will increase steadily with time.

The zeros of the system and the input function influence the amplitude of both the steady-state and the transient response. In this case, we have only one system zero at $s = -b$, and there is no zero because of the input function. The effect of this system zero on the amplitude of the response can be seen in Eq. (5.54).

The quantity $1/a$ is called the time constant of the given first-order system. The time constant is a measure of the speed with which a system responds to an external input. The lower the value of the time constant (or higher the value of a), the faster will be the system response. Sometimes, a is also called the exponential decay frequency. For $t = 1/a$, $y(t) = 0.63$ times its final rise above the initial value. In other words, at time equal to one time constant, the output rises to 63% of its steady-state value above the initial value.

The rise time T_r is the time for the output to increase from 0.1 to 0.9 times its final or steady-state value. However, it may be noted that some authors define it as the time for the output to rise from 0.1 to 100% of the final value. However, this alternative definition is not used in this text.

For a first-order system,

$$T_r = \frac{2.2}{a} \tag{5.55}$$

The settling time T_s is defined as the time required for the output to reach, for the first occurrence, within 2% of its final or steady-state value. For the first-order system, this value is approximately given by

$$T_s = \frac{4}{a} \tag{5.56}$$

5.4.2 Response of Second-Order Systems

Now let us consider a second-order system given by

$$G(s) = \frac{b}{s^2 + as + b} \tag{5.57}$$

The response to a unit-step input is

$$\bar{y}(s) = G(s)\bar{r}(s) \tag{5.58}$$

$$= \frac{b}{} \tag{5.59}$$

A second-order system has two poles. In general, the response of a second-order system can be any one of the four types of responses as shown in Fig. 5.3. Suppose the system poles, which depend on the values of a and b, are both real and negative as shown in Fig. 5.3a; then the corresponding response is a steady rise, without any overshoot, to the final value. This type of response is called an "overdamped" response. If the poles are purely imaginary, then the response is a constant amplitude sinusoid that will continue forever because there is no damping in the system (Fig. 5.3b). This type of response is called "oscillatory response." The frequency of this undamped oscillation is called the natural frequency of the system. If the system poles are a pair of complex conjugate numbers with negative real parts, then the transient response will be oscillatory and is characterized by overshoots as shown in Fig. 5.3c. This type of response is called "underdamped response" and the frequency of this oscillation is called the exponential decay frequency or the damped frequency. If the poles are real, negative, and equal to each other, then the response is said to be critically damped as shown in Fig. 5.3d.

The transfer function of the second-order system given by the Eq. (5.57) can be expressed in the standard form as follows:

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{5.60}$$

where

$$\omega_n = \sqrt{b} \tag{5.61}$$

$$\zeta = \frac{a}{2\omega_n} = \frac{a}{2\sqrt{b}} \tag{5.62}$$

Here, ω_n is the natural frequency of the system, and ζ is the damping ratio of the system. The damping ratio is defined as the ratio of the existing damping to that required for critical damping. For $\zeta > 1$, the second-order system has two real, negative, and unequal roots, and the system has an overdamped response as in Fig. 5.3a. When $\zeta = 1$, the two real negative roots become equal, and the motion associated with this case is called critically damped motion as shown in Fig. 5.3d. When $\zeta < 1.0$, the second-order system has a pair of complex roots with negative real parts, and the system displays a damped oscillatory motion as in Fig. 5.3c. Thus, the condition $\zeta = 1$ represents the boundary between the overdamped exponential motion and the damped oscillatory motion.

The damping ratio ζ and the natural frequency ω_n are two important parameters that characterize a second-order system. The response of a second-order system depends on the values of these two parameters. The system poles, in terms of frequency and damping ratio, are given by

$$s_{1,2} = -\sigma_d \pm j\omega_d \tag{5.63}$$

We have

$$\sigma_d = \zeta\omega_n \qquad \omega_d = \omega_n\sqrt{1-\zeta^2} \tag{5.64}$$

so that

$$s_{1,2} = -\zeta\omega_n \pm j\omega_n\sqrt{1-\zeta^2} \tag{5.65}$$

Here, ω_d is called the damped frequency.

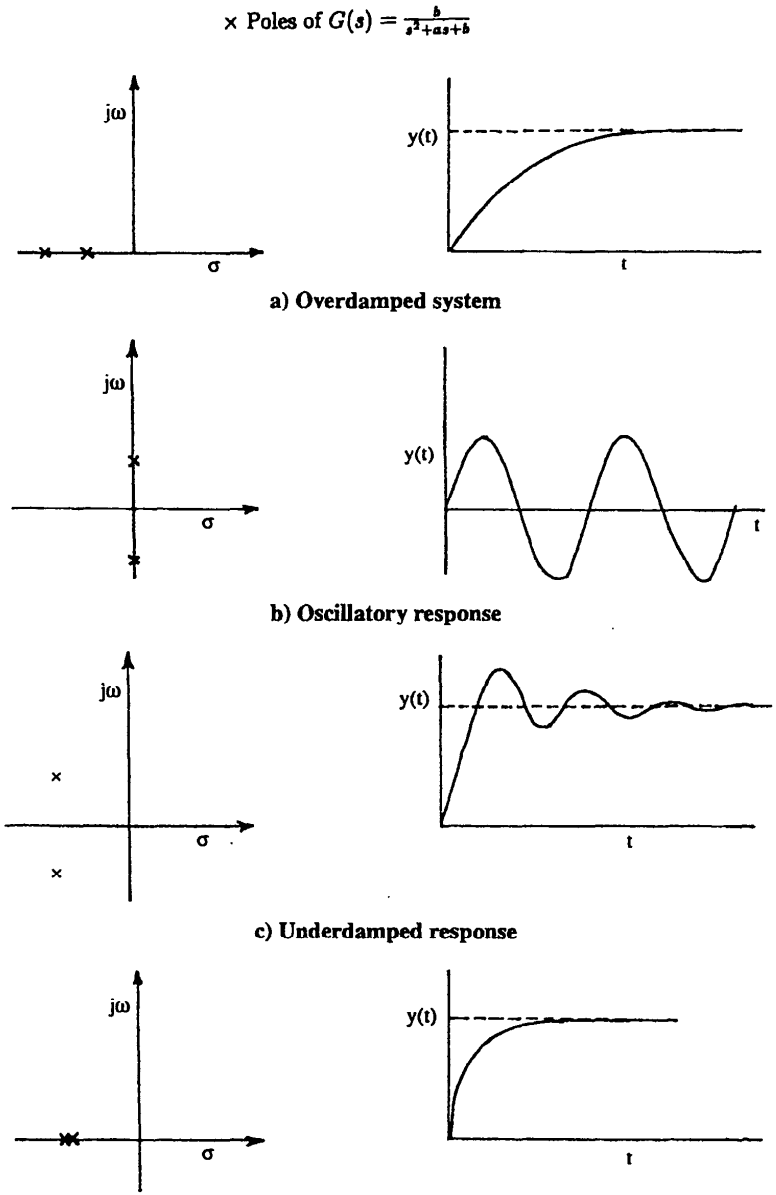


Fig. 5.3 Second-order system response.

The unit-step input response of a second-order system of Eq. (5.60) is given by

$$\bar{y}(s) = \frac{\omega_n^2}{s\left(s^2 + 2\zeta\omega_n s + \omega_n^2\right)} \tag{5.66}$$

$$= \frac{k_1}{s} + \frac{k_2 s + k_3}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{5.67}$$

where k_1, k_2, and k_3 are constants. Expanding the partial fractions, taking the inverse Laplace transforms, and simplifying, we obtain

$$y(t) = 1 - \frac{1}{\sqrt{1-\zeta^2}} e^{-\zeta\omega_n t}\cos(\omega_d t - \phi) \tag{5.68}$$

$$= 1 - \frac{1}{\sqrt{1-\zeta^2}} e^{-\zeta\omega_n t}\cos\left(\omega_n t\sqrt{1-\zeta^2} - \phi\right) \tag{5.69}$$

where the phase angle ϕ is given by

$$\phi = \tan^{-1}\frac{\zeta}{\sqrt{1-\zeta^2}} \tag{5.70}$$

The typical response for various values of the damping parameter ζ are shown in Fig. 5.4. Because the time appears as a product $\omega_n t$ in Eq. (5.69), it is convenient

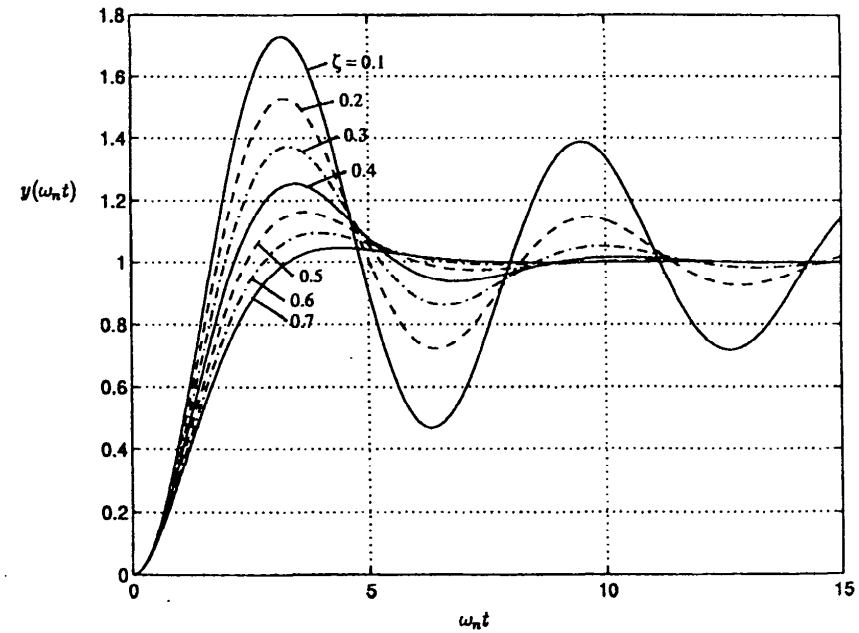


Fig. 5.4 Typical second-order system response.

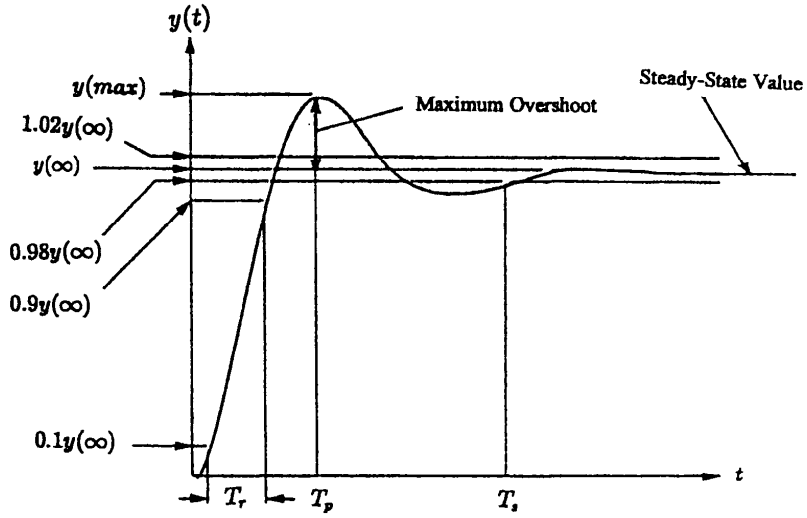


Fig. 5.5 Characteristics of second-order system response.

to plot $y(\omega_n t)$ vs $\omega_n t$, which has the effect of normalizing the time with respect to the system natural frequency ω_n. We observe that the lower the value of ζ, the more oscillatory is the response and the larger is the overshoot.

The parameters that characterize the response of a second-order system (see Fig. 5.5) are as follows:

*1) **Peak time** $\mathbf{T_p}$.* It is the time required to reach the first or maximum peak $y(\max)$.

*2) **Percent overshoot** $\mathbf{O_s}$.* It is the maximum overshoot above the final or steady-state value and expressed as a percentage of steady-state value $y(\infty)$.

*3) **Settling time** $\mathbf{T_s}$.* It is the time required for the transient response to come and stay within $\pm 2\%$ of the steady-state value.

*4) **Rise time** $\mathbf{T_r}$.* This is the time required for the response to rise from 0.1 to 0.9 of the final or steady-state value at its first occurrence.

Notice that the settling time and rise time are basically the same as those defined for first-order systems. The above definitions are general in nature and as such apply to systems of any order.

In the following, we will derive analytical expressions for T_p, O_s, and T_s for second-order systems. However, for the rise time T_r, it is not possible to obtain a simple analytical expression.

The peak time T_p can be obtained by differentiating Eq. (5.69) with respect to time t and finding the first zero crossing for $t \geq 0$ as

$$T_p = \frac{\pi}{\omega_d} \tag{5.71}$$

$$= \frac{\pi}{\omega_n \sqrt{1-\zeta^2}} \tag{5.72}$$

The percent overshoot O_s is given by

$$O_s = \frac{y(\max) - y(\infty)}{y(\infty)} \times 100 \tag{5.73}$$

where $y(\max)$ is the value of $y(t)$ at $t = T_p$. For the unit-step input, $y(\infty) = 1$. Then,

$$O_s = e^{\frac{-\zeta\pi}{\sqrt{1-\zeta^2}}} \times 100 \tag{5.74}$$

This relation states that the percent overshoot depends uniquely on the damping ratio ζ. The value of the damping ratio corresponding to a given percent overshoot is given by

$$\zeta = \frac{-\ell_n (O_s/100)}{\sqrt{\pi^2 + \ell_n^2 (O_s/100)}} \tag{5.75}$$

The settling time T_s is the value of time t when the amplitude of the damped response comes within ± 0.02 for the first occurrence. Using Eq. (5.69),

$$\frac{1}{\sqrt{1-\zeta^2}} e^{-\zeta \omega_n T_s} = 0.02 \tag{5.76}$$

Solving, we get

$$T_s = \frac{-\ell_n \left(0.02\sqrt{1-\zeta^2}\right)}{\zeta \omega_n} \tag{5.77}$$

However, this expression is somewhat complex for frequent use. Instead, the following simple approximation is used to evaluate T_s. The numerator of the above equation varies from 3.91 to 4.74 as ζ varies from 0 to 0.9. For typical underdamped second-order systems, the numerator is usually close to 4. In view of this, the following approximation is often used:

$$T_s = \frac{4}{\zeta \omega_n} \tag{5.78}$$

5.4.3 Nonminimum Phase Systems

If all the poles and zeros of a system lie in the left half of the s-plane, then the given system is called a minimum phase system. If a system has at least one pole or one zero in the right half of the s-plane, then such a system is called a nonminimum phase system. A characteristic property of a nonminimum phase system is that the transient response may start out in the opposite direction to the input but comes back eventually in the same direction.

For the first-order system given by Eq. (5.48), if $b \leq 0$, the system becomes a nonminimum phase system. The steady-state value will be negative, whereas the response starts out with an initial value equal to $+1.0$.

5.5 Steady-State Errors of Unity Feedback Systems

Ideally, control systems are designed so that the output follows the reference input all the time. In other words, it is desired that the steady-state value of the output be equal to the value of the reference input as closely as possible. However, it may not always be possible to achieve this goal and, in reality, the steady-state value of the output differs from the value of the reference input.

The steady-state error is the difference between the steady-state value of the output and the reference input. Usually, unit-step, unit-ramp, or parabolic functions are used as test inputs to determine the steady-state error. In the following, we will derive expressions for steady-state error for unity feedback systems as shown in Fig. 5.6. It may be noted that any given nonunity feedback system (Fig. 5.7) can be expressed as an equivalent unity feedback system by adding and subtracting a unity feedback loop as shown in Fig. 5.8a and obtaining an equivalent unity feedback system as shown in Fig. 5.8b.

Let $e(t)$ be the error signal that is the difference between the output and the input. For steady-state error to be zero, $e(t) \to 0$ as $t \to \infty$.

We have

$$e(t) = r(t) - y(t) \tag{5.79}$$

Taking Laplace transforms,

$$\bar{e}(s) = \bar{r}(s) - \bar{y}(s) \tag{5.80}$$

$$= \bar{r}(s) - G(s)\bar{e}(s) \tag{5.81}$$

$$= \frac{\bar{r}(s)}{1 + G(s)} \tag{5.82}$$

Using the final value theorem in Eq. (5.30), we can obtain the steady-state error $e(t)$ as follows:

$$e(\infty) = \lim_{s \to 0}[s\bar{e}(s)] \tag{5.83}$$

$$= \lim_{s \to 0}\left[\frac{s\bar{r}(s)}{1 + G(s)}\right] \tag{5.84}$$

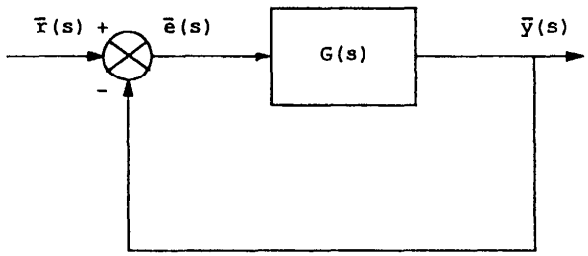


Fig. 5.6 Unity feedback system.

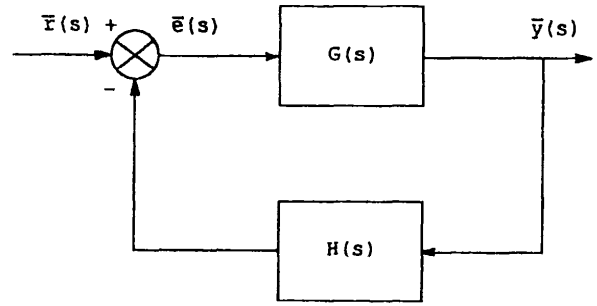
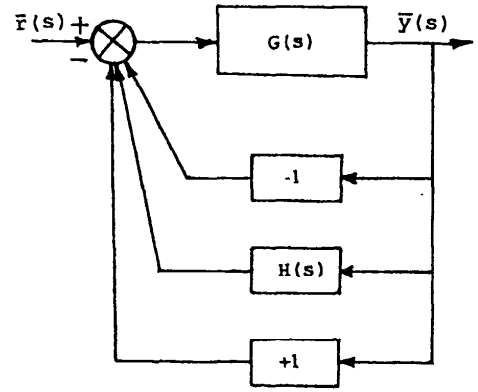
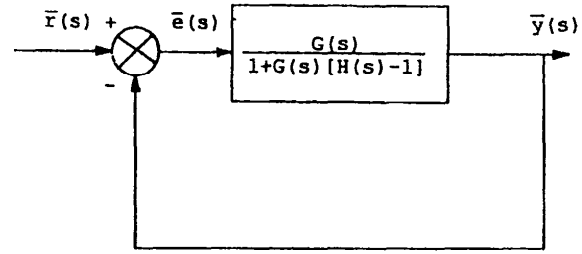


Fig. 5.7 Nonunity feedback system.



a) Addition and subtraction of unity feedback



b) Equivalent unity feedback system

Fig. 5.8 Equivalent unity feedback system for a given nonunity feedback system.

The steady-state error to a unit-step function $\bar{r}(s) = 1/s$ is given by

$$e(\infty) = \frac{1}{1 + \lim_{s \to 0} G(s)} \tag{5.85}$$

In other words, for steady-state error to a unit-step function to be zero, $\lim_{s \to 0} G(s) = \infty$.

Generally, we have

$$G(s) = \frac{(s + z_1)(s + z_2) \cdots (s + z_m)}{s^q (s + p_1)(s + p_2) \cdots (s + p_n)} \tag{5.86}$$

For rational transfer functions, $n \geq m$, i.e., the number of poles exceeds the number of zeros. The value of the index q designates the type of the system. For example, if $q = 0$, the system is said to be a type "0" system and, for a type "0" system,

$$\lim_{t \to 0} G(s) = \frac{z_1 z_2 \cdots z_m}{p_1 p_2 \cdots p_n} \tag{5.87}$$

which is finite. Hence, the steady-state error for a type "0" system to a unit-step input is nonzero and is given by

$$e(\infty) = \frac{1}{1 + K_p} \tag{5.88}$$

Here, K_p is called the position constant and is given by

$$K_p = \frac{z_1 z_2 \cdots z_m}{p_1 p_2 \cdots p_n} \tag{5.89}$$

For type "1" or higher systems ($q \geq 1$), $K_p = \infty$, and the steady-state error to a unit-step input approaches zero. A type "1" system is said to have one integrator in the forward path. In other words, the integer value of q corresponds to the number of integrators in the forward path.

It can be shown that the steady-state error to a unit-ramp function $r(t) = t$ or $\bar{r}(s) = 1/s^2$ is given by

$$e(\infty) = \frac{1}{K_v} \tag{5.90}$$

where the velocity error coefficient K_v is given by

$$K_v = \lim_{s \to 0} [sG(s)] \tag{5.91}$$

Thus, if the steady-state error for a unit-ramp function is to vanish, the velocity constant K_v must be very large, which implies that we must have $q \geq 2$. In other words, we must have at least two integrators in the forward path. Thus, for a type "0" system, the steady-state error for a unit-ramp function is infinity; for a type "1" system, it is finite; and for systems of type "2" or higher, it is zero.

Similarly, the steady-state error to a unit-parabolic input function, $r(t) = t^2$ or $\bar{r}(s) = 1/s^3$, can be obtained as

$$e(\infty) = \frac{1}{K_a} \tag{5.92}$$

where

$$K_a = \lim_{s \to 0} [s^2 G(s)] \tag{5.93}$$

We observe that for a given system to have zero steady-state error to a unit-parabolic input, we must have at least have three integrators in the forward path. Therefore, the steady-state error to unit-parabolic input of type "0" and type "1" systems is infinite; for type "2" systems, it is finite; and for systems of type "3" or higher, it is zero.

5.6 Frequency Response

In steady-state, a sinusoidal input to a linear system generates a sinusoidal response (output) of the same frequency. However, the magnitude and phase angles of the response are generally different from those of the input and also vary with the frequency of the applied input. In the following, we will determine the steady-state response (magnitude and phase angle) to sinusoidal inputs.[2]

In general, a sinusoid input function can be represented as

$$r(t) = A \cos \omega t + B \sin \omega t \tag{5.94}$$

$$= M_i \cos(\omega t + \phi_i) \tag{5.95}$$

Here, M_i and ϕ_i are the magnitude and phase angle of the input sinusoid function and are given by the following expressions:

$$M_i = \sqrt{A^2 + B^2} \tag{5.96}$$

$$\phi_i = -\tan^{-1} \frac{B}{A} \tag{5.97}$$

In phasor notation,

$$r(t) = M_i \angle \phi_i \tag{5.98}$$

Furthermore, we assume that we can represent $r(t)$ as a complex number, $r(t) = A - jB$ so that $A - jB = M_i e^{j\phi_i}$ and $A + jB = M_i e^{-j\phi_i}$. Taking the Laplace transform of Eq. (5.94), we get

$$\bar{r}(s) = \left(\frac{As + B\omega}{s^2 + \omega^2} \right) \tag{5.99}$$

The response to a sinusoidal input is given by

$$\bar{y}(s) = \left(\frac{As + B\omega}{s^2 + \omega^2} \right) G(s) \tag{5.100}$$

$$= \left(\frac{As + B\omega}{(s + j\omega)(s - j\omega)} \right) G(s) \tag{5.101}$$

$$= \frac{k_1}{s + j\omega} + \frac{k_2}{s - j\omega} + \cdots \tag{5.102}$$

Because we are interested in only the steady-state response, we have ignored the terms corresponding to $G(s)$, which generate the transient response. Recall that

the steady-state response comes from the poles because of input function, which in this case are at $s = \pm j\omega$. Using partial fraction method, we get

$$k_1 = \left[\frac{As + B\omega}{s - j\omega} G(s)\right]_{s=-j\omega} \tag{5.103}$$

$$= \frac{A + jB}{2} G(-j\omega) \tag{5.104}$$

Because $G(j\omega)$ is a complex number, we can write $G(j\omega) = M_g e^{j\phi_g}$ or $G(-j\omega) = M_g e^{-j\phi_g}$. With $A + jB = M_i e^{-j\phi_i}$, we have

$$k_1 = \frac{1}{2} M_i e^{-j\phi_i} M_g e^{-j\phi_g} \tag{5.105}$$

$$= \frac{M_i M_g}{2} e^{-j(\phi_i + \phi_g)} \tag{5.106}$$

Similarly,

$$k_2 = \frac{M_i M_g}{2} e^{j(\phi_i + \phi_g)} \tag{5.107}$$

$$= k_1^* \tag{5.108}$$

where * denotes the complex conjugate. Then,

$$\bar{y}_\infty(s) = \frac{M_i M_g}{2}\left[\frac{e^{-j(\phi_i + \phi_g)}}{(s + j\omega)} + \frac{e^{j(\phi_i + \phi_g)}}{(s - j\omega)}\right] \tag{5.109}$$

where the suffix ∞ denotes the steady-state value ($t \to \infty$) and M_g and ϕ_g are the magnitude and phase angle of the transfer function $G(s)$ (with $s = j\omega$) and are given by

$$M_g = |G(j\omega)| \tag{5.110}$$

$$\angle\phi_g = \angle G(j\omega) \tag{5.111}$$

Taking the inverse Laplace transforms in Eq. (5.109), we get

$$y_\infty(t) = \frac{M_i M_g}{2}\left[e^{-j(\phi_i + \phi_g + \omega t)} + e^{j(\phi_i + \phi_g + \omega t)}\right] \tag{5.112}$$

$$= M_i M_g \cos(\phi_i + \phi_g + \omega t) \tag{5.113}$$

or, in phasor notation,

$$M_\infty \angle\phi_\infty = M_i M_g \angle(\phi_i + \phi_g) \tag{5.114}$$

Thus, at any frequency, the magnitude of the steady-state output is the product of the magnitude of the input and the magnitude of the transfer function. The phase of the steady-state output is the sum of the phase of the input and the phase of the transfer function. Therefore, if we want to know how the magnitude and phase of the system response vary with frequency of a given sinusoidal input, it is sufficient to know the variation of M_g and ϕ_g with frequency because M_i and ϕ_i are supposed to be known. This process of determining the variation of M_g and

ϕ_g with frequency is called the frequency response of the system. In other words, the frequency response of a system whose transfer function is $G(s)$, $s = j\omega$ is nothing but the variation of M_g and ϕ_g with frequency ω.

One of the most widely used methods of obtaining the frequency response of a transfer function is the Bode plot. It consists of two parts: the magnitude plot in decibels where one decibel of $M = 20 \log_{10} M$ and the phase plot in degrees, both plotted against frequency ω, which is usually expressed in radians/second.

Generally, the Bode plot is drawn for open-loop transfer function $G(s)$. Furthermore, if the transfer function contains a variable gain k, then the Bode plot is made for $k = 1$. For any other value of k, the corresponding Bode plot can be easily obtained by shifting the entire Bode plot by $20 \log_{10} k$. The plot shifts upward if $k > 0$ and downward if $k < 0$. We will illustrate the method of drawing a Bode plot with the help of Example 5.1.

Example 5.1

Draw the Bode plot for a system given by

$$G(s) = \frac{k(s + 3)}{s(s + 1)(s + 2)}$$

Solution. The first step is to assume $k = 1$ and rewrite the given transfer function in the following form:

$$G(s) = \frac{\frac{3}{2}\left(\frac{s}{3} + 1\right)}{s(s + 1)\left(\frac{s}{2} + 1\right)}$$

$$= \frac{\frac{3}{2} G_1(s)}{G_2(s) G_3(s) G_4(s)}$$

where

$$G_1(s) = \left(\frac{s}{3} + 1\right)$$

$$G_2(s) = s$$

$$G_3(s) = (s + 1)$$

$$G_4(s) = \left(\frac{s}{2} + 1\right)$$

Substituting $s = j\omega$, taking logarithms on both sides, multiplying by 20 to convert to decibels, and taking absolute values, we get

$$20 \log_{10} |G(j\omega)| = 20 \log_{10} \tfrac{3}{2} + 20 \log_{10} |G_1(j\omega)|$$

$$- 20 \log_{10} |G_2(j\omega)| - 20 \log_{10} |G_3(j\omega)| - 20 \log_{10} |G_4(j\omega)|$$

Now let us consider the magnitude plot of each one of the terms on the right-hand side separately. The magnitude plot of the term $20 \log_{10} \frac{3}{2} = 3.5218$ for all values

of the frequency ω. For the second term,

$$20 \log_{10} |G_1(j\omega)| = 20 \log_{10} \left| \left(\frac{j\omega}{3} + 1 \right) \right|$$

$$= 20 \log_{10} \left| \sqrt{\frac{\omega^2}{9} + 1} \right|$$

For smaller values of ω (low-frequency approximation), we assume

$$20 \log_{10} |G_1(j\omega)| = 20 \log_{10} 1$$

$$= 0$$

and, for higher values of ω (high-frequency approximation), we assume

$$20 \log_{10} |G_1(j\omega)| = 20 \log_{10} \frac{\omega}{3}$$

Thus, for $\omega = 3$ rad/s, $20 \log_{10} |G_1(j\omega)| = 0$ and, for $\omega = 30$ rad/s, $20 \log_{10} \times |G_1(j\omega)| = 20$ db. The slope of the high-frequency approximation of this term is +20 db/decade. Here, one decade means a tenfold increase in frequency.

The frequency at which the low-frequency approximation intersects the high-frequency approximation is called the corner frequency. For the magnitude plot of $G_1(j\omega)$, the corner frequency is 3 rad/s. The low-frequency approximation holds for frequencies that are below the corner frequency, and the high-frequency approximation holds for frequencies that are above the corner frequency. Proceeding in a similar way, we find that the magnitude plots of $G_3(j\omega)$ and $G_4(j\omega)$ have corner frequencies of 1.0 and 2.0 rad/s, respectively. The magnitude plot of $G_2(j\omega)$ is a straight line and hence has no corner frequency. Each of the magnitude plots of $G_2(j\omega)$, $G_3(j\omega)$, and $G_4(j\omega)$ have a slope of −20 db/decade. The component magnitude plots of $G_1(j\omega)$, $G_2(j\omega)$, $G_3(j\omega)$, and $G_4(j\omega)$ are shown in Fig. 5.9.

The phase plot can be drawn using the relation

$$\angle G(j\omega) = \angle \tfrac{3}{2} + \angle G_1(j\omega) - \angle G_2(j\omega) - \angle G_3(j\omega) - \angle G_4(j\omega)$$

Note that $\angle \frac{3}{2} = 0$. As before, let us consider the terms on the right-hand side one by one. The phase of the second term is given by

$$\angle G_1(j\omega) = \angle \left(\frac{j\omega}{3} + 1 \right) = \tan^{-1} \left(\frac{\omega}{3} \right)$$

For small values of ω (low-frequency approximation), $\angle G_1(j\omega) = 0$. As $\omega \to \infty$ (high-frequency approximation), $\angle G_1(j\omega) = 90$ deg. For $\omega = 3$ rad/s, $\angle G_1(j\omega) =$ 45 deg and, for $\omega = 30$ rad/s, $\angle G_1(j\omega) \simeq 90$ deg so that the slope of high-frequency approximation is 45 deg/decade. We assume that the low-frequency approximation holds for frequencies that are one decade below the frequency at which the phase angle is 45 deg. For $\angle G_1(j\omega)$, this value is 0.3 rad/s.

With these approximations, the phase plot $\angle G_1(j\omega)$ and those of other terms are shown in Fig. 5.10.

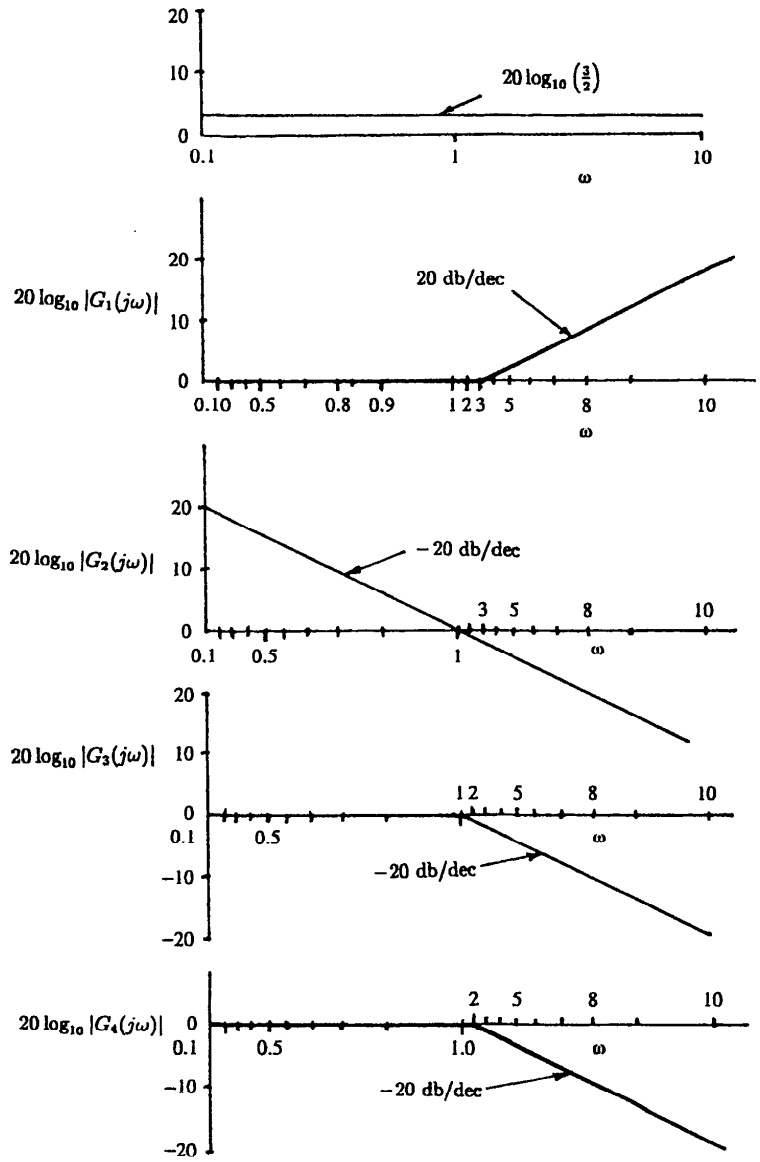


Fig. 5.9 Component Bode-magnitude plots for Example 5.1.

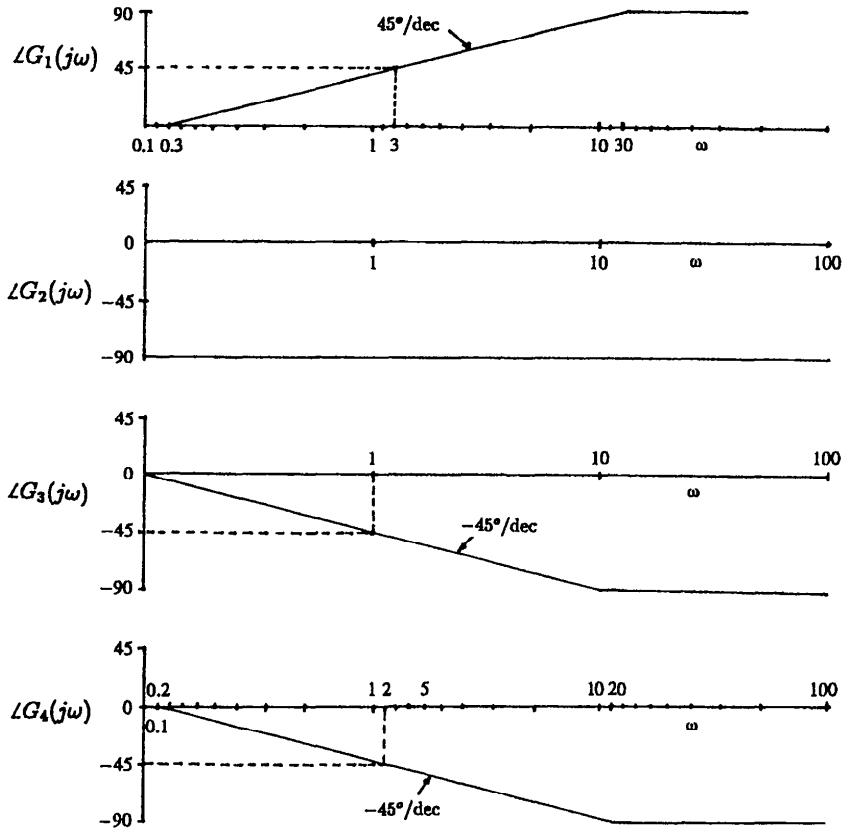


Fig. 5.10 Component Bode-phase plots for Example 5.1.

The combined magnitude and phase plots, which are the sum of component plots, are shown in Fig. 5.11.

MATLAB[4] is a convenient tool for control system analysis and design. We assume that the reader has access to this or a software with similar capabilities. Using MATLAB,[4] the magnitude and phase plots of the given transfer function are drawn as shown in Fig. 5.12. It is interesting to observe that the approximate method that involves the concept of corner frequencies comes close to the more accurate plots given by MATLAB.[4]

5.7 Stability of Closed-Loop Systems

One of the most important requirements for a control system is stability. A linear, time-invariant system is said to be stable if a bounded input produces a bounded output. In other words, for a stable system the output should reach a steady state.

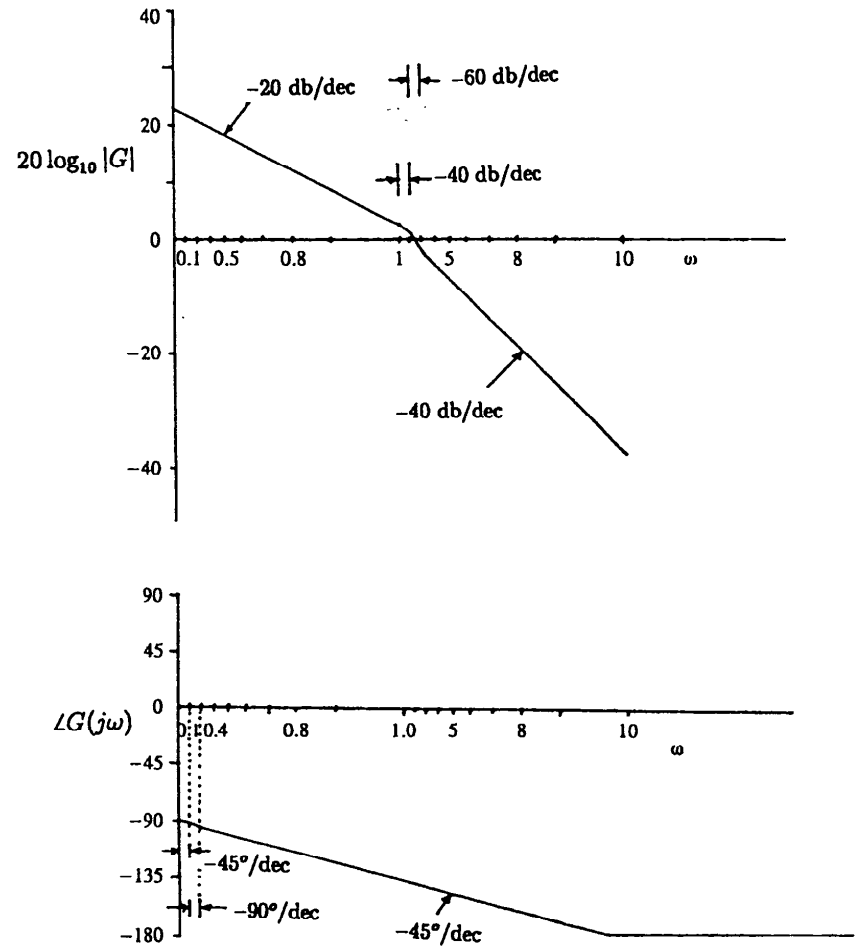


Fig. 5.11 Combined Bode plots for Example 5.1.

If the input is zero, then the transient or free response must decay or go to zero as the time approaches infinity. Therefore, if the output to a bounded input is not bounded and the free response does not decay, the system is said to be unstable.

The transient response depends on the location of the system poles in the s-plane. If all poles are on the left half of the s-plane—i.e., all poles are negative if real or have negative real parts if complex—then the transient response is one of exponential decay or damped oscillation, and the system is stable. On the other hand, if any one or more of the system poles are located on the right half of the s-plane—i.e., are positive if real or have positive real parts if complex—then the transient response is one of exponential divergence or an oscillatory motion with ever-increasing amplitude. Such a system is said to be unstable. Thus, a stable

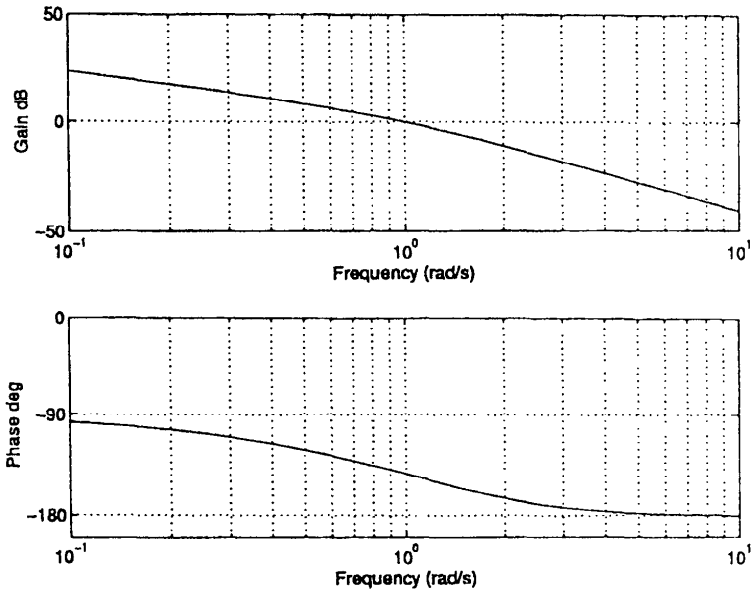


Fig. 5.12 Bode plots using MATLAB[4] for Example 5.1.

system has all poles located in the left half of the s-plane, and an unstable system has one or more of its poles located in the right half of the s-plane. If some or all the system poles are located on the imaginary axis, the transient response will be consisting of pure oscillatory motion in which the amplitude of oscillation neither increases nor decreases. Such a system is said to be neutrally stable.

The task of determining the stability of an open-loop system is simple and straightforward because the open-loop poles are known. However, it is not so straightforward for the closed-loop systems because the closed-loop poles, which are the roots of the characteristic equation $1 + kGH = 0$, are not known. Moreover, the exercise of finding the closed-loop poles has to be repeated many times if a system parameter like the gain k is a variable. It is a simple task to determine the roots of the characteristic equation $1 + kGH = 0$ if this expression is a polynomial in s of degree lower than three. For fourth- or higher degree polynomials, the analytical determination of the roots is not a simple task. In the following, we will discuss methods that help us determine the stability of the closed-loop systems without actually solving the characteristic equation $1 + kGH = 0$. These are 1) Routh's stability criterion, 2) the root-locus method, and 3) Nyquist stability criterion.

5.7.1 Routh's Stability Criterion

Routh's stability criterion helps us determine whether any of the closed-loop poles are positive if real or have positive real parts if complex without actually solving the closed-loop characteristic equation.

The procedure is as follows.

1) Express the characteristic polynomial in the following form:

$$a_0 s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_{n-1} s + a_n = 0 \tag{5.115}$$

where the coefficients $a_0, a_1, \ldots, a_n$ are real quantities. We assume that $a_n \neq 0$ so that any zero root is removed.

2) Examine the value of each coefficient. If any coefficient is zero or negative when at least one other coefficient is positive, then Routh's criterion states that there will be at least one root of the characteristic polynomial that is imaginary or has a positive real part. In such a case, the system is not stable. Therefore, for stability, all the coefficients must be positive or must have the same sign. This forms the necessary condition for stability.

3) To check whether the sufficiency condition is satisfied, form the Routh's array as follows:

$$\begin{array}{cccccc}
s^n & a_0 & a_2 & a_4 & a_6 & . \\
s^{n-1} & a_1 & a_3 & a_5 & a_7 & . \\
s^{n-2} & b_1 & b_2 & b_3 & b_4 & . \\
s^{n-3} & c_1 & c_2 & c_3 & c_4 & . \\
s^{n-4} & d_1 & d_2 & d_3 & d_4 & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
s^2 & e_1 & e_2 & & & \\
s^1 & f_1 & & & & \\
s^0 & g_1 & & & &
\end{array} \tag{5.116}$$

where

$$b_1 = \frac{a_1 a_2 - a_0 a_3}{a_1}, \quad b_2 = \frac{a_1 a_4 - a_0 a_5}{a_1}, \quad b_3 = \frac{a_1 a_6 - a_0 a_7}{a_1}, \quad \ldots \tag{5.117}$$

$$c_1 = \frac{b_1 a_3 - a_1 b_2}{b_1}, \quad c_2 = \frac{b_1 a_5 - a_1 b_3}{b_1}, \quad c_3 = \frac{b_1 a_7 - a_1 b_4}{b_1}, \quad \ldots \tag{5.118}$$

and

$$d_1 = \frac{c_1 b_2 - b_1 c_2}{c_1}, \quad d_2 = \frac{c_1 b_3 - b_1 c_3}{c_1} \tag{5.119}$$

$$\ldots \quad \ldots \tag{5.120}$$

This process is continued until nth row has been completed. The complete array of coefficients is triangular. Note that the evaluation of b_i, c_i, and d_i, etc., is continued

until the remaining ones are zero. For example, for a fourth-degree polynomial in s, $a_5 = a_6 = \cdots = 0$ so that $b_3 = b_4 = \cdots = 0$, $c_2 = c_3 = \cdots = 0$, $d_2 = d_3 = \cdots = 0$, $e_2 = e_3 = \cdots = 0$, and $f_2 = f_3 = \cdots = 0$.

Routh's stability criterion states that the number of roots of the characteristic polynomial with positive real parts is equal to the number of changes in sign of the coefficients of the first column of Routh's array. It is important to note that the exact values of these coefficients need not be known; instead only the signs are required. Thus, the sufficiency condition for a closed-loop system to be stable is that all the elements of the first column of Routh's array must be positive or must have the same sign.

To summarize, the necessary and sufficient condition for the stability of a closed-loop system is that all the coefficients of the characteristic polynomial and the elements of the first column of Routh's array must be positive or must have the same sign.

If any of the coefficients of the characteristic polynomial or any element of the first column in the Routh's array is zero, then replace that term by a very small positive number ϵ and proceed as usual with the evaluation of the rest of the elements of Routh's array.

For a fourth-order polynomial, the Routh's stability criterion reduces to the following:

1) All coefficients a_0, a_1, a_2, a_3, and a_4 must be positive.
2) The Routh's discriminant $(a_1a_2 - a_0a_3)a_3 - a_1^2a_4$ must be positive.

Example 5.2

Using Routh's criterion, determine the stability of the system represented by the following characteristic polynomial:

$$s^4 + 2s^3 + 5s^2 + 2s + 2 = 0$$

Solution. This is a fourth-degree polynomial in s. We have $a_0 = 1$, $a_1 = 2$, $a_2 = 5$, $a_3 = 2$, $a_4 = 2$, and $a_5 = 0$. Because all the coefficients are positive and none of the coefficients a_0 to a_4 is zero, the necessary condition for stability is satisfied. To examine whether the sufficiency condition is satisfied, we form Routh's array as follows:

s^4:	1	5	2
s^3:	2	2	0
s^2:	4	2	0
s^1:	1	0	
s^0:	2		

We observe that all the elements of the first column of this table are positive; hence the sufficiency condition is also satisfied. Hence, the characteristic polynomial has no positive real root or a complex root with positive real part and the given system is stable.

Example 5.3

Given the characteristic polynomial

$$s^4 + 3s^3 + 2s^2 + 4s + 1 = 0$$

Examine the stability of the system using Routh's stability criterion.

Solution. Because all of the coefficients of this fourth-degree polynomial are positive, the necessary condition is satisfied. To see whether the sufficiency condition is satisfied, form Routh's array as follows:

s^4:	1	2	1
s^3:	3	4	0
s^2:	$\frac{2}{3}$	1	0
s^1:	$-\frac{1}{2}$	0	
s^0:	1		

There are two sign changes in the first column starting with the row corresponding to s^2. Hence, there will be two roots that are either positive or have positive real parts and the given system is unstable.

Example 5.4

For the system whose characteristic polynomial is given by

$$s^4 + 2s^2 + 5s + 2 = 0$$

Examine the stability of the system using Routh's criterion.

Solution. Notice that the s^3 term is missing. Hence, we rewrite the given polynomial as follows:

$$s^4 + \epsilon s^3 + 2s^2 + 5s + 2 = 0$$

where ϵ is a small positive number, say 0.0001. With this, we observe that the necessary condition is satisfied. To see whether the sufficiency condition is satisfied, we form Routh's array as follows:

s^4:	1	2	2
s^3:	ϵ	5	0
s^2:	$-50{,}000$	2	0
s^1:	5	0	
s^0:	2		

There are two sign changes. Hence, there will be two roots that are either positive or have positive real parts. Hence, the given system is unstable. Note that if ϵ appears in any expression, we have to evaluate the value of that expression by taking the limit as ϵ tends to zero.

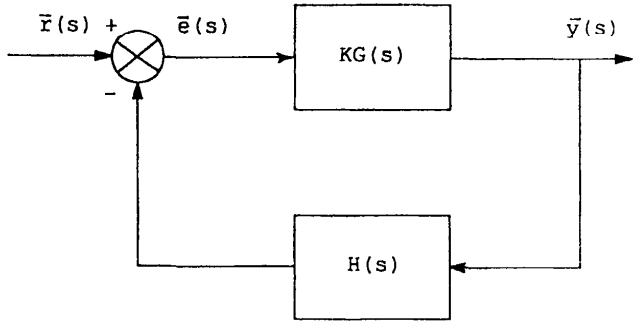


Fig. 5.13 Feedback control system.

5.7.2 Root-Locus Method

The root-locus is a powerful method of determining the nature of transient response and stability of control systems. It is a graphical method and is particularly well suited for application to those problems where any parameter or the loop-gain is a variable.

Consider a closed-loop system as shown in Fig. 5.13. The closed-loop transfer function is given by

$$T(s) = \frac{kG(s)}{1 + kG(s)H(s)} \tag{5.121}$$

The equation

$$1 + kG(s)H(s) = 0 \tag{5.122}$$

is known as the characteristic equation of the given closed-loop system. The roots of this equation are also called the eigenvalues of the closed-loop system. In other words, the poles of $T(s)$ are the eigenvalues of the closed-loop system. The root-locus is a plot of the variation of roots of Eq. (5.122) as the parameter k is varied from zero to infinity. Using Eq. (5.122), we can deduce the following two conditions for a given point to lie on the root-locus.

*1) **Magnitude condition.*** If a given point s is to lie on the root-locus, we must have

$$|kG(s)H(s)| = 1 \tag{5.123}$$

*2) **Phase condition.*** For a given point to lie on the root-locus, we must have

$$\angle kG(s)H(s) = (2n + 1)180 \tag{5.124}$$

where $n = 0, \pm 1, \pm 2, \ldots$. Note that the expression on the right-hand side of Eq. (5.124) is an odd multiple of 180 deg with either positive or negative sign. These two conditions form the basis of sketching the root-locus as the parameter k varies from zero to infinity.

To understand the meaning of Eqs. (5.123) and (5.124), let us refer to Fig. 5.14. Suppose P is to be a point on the root-locus; then according to the magnitude

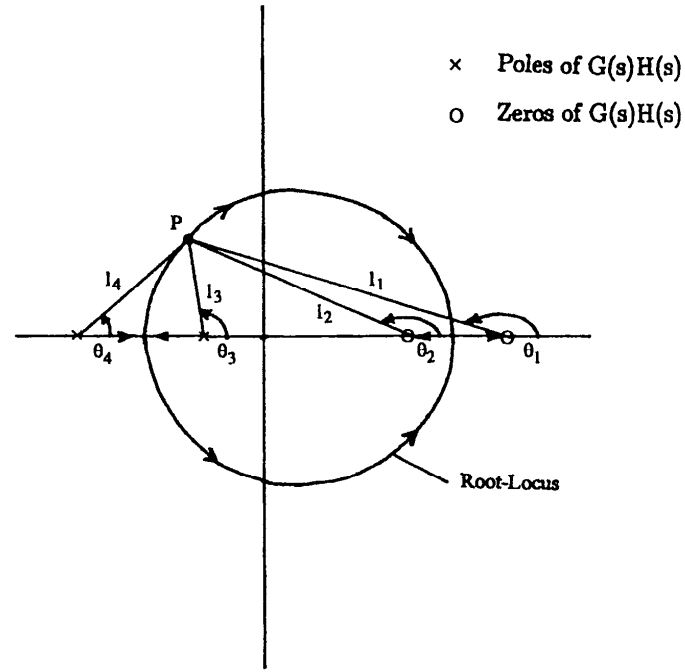


Fig. 5.14 Magnitudes and angles of vectors for a point on the root-locus.

condition

$$k = \frac{1}{|G(s)H(s)|} = \frac{\prod l_{i,p}}{\prod l_{i,z}} \tag{5.125}$$

where $l_{i,p}$ and $l_{i,z}$ are the magnitudes of the vectors drawn from each of the poles and zeros to the given point on the root-locus. For the root-locus shown in Fig. 5.14, the magnitude and phase conditions are

$$k = \frac{l_3 l_4}{l_1 l_2} \tag{5.126}$$

$$\theta_1 + \theta_2 - \theta_3 - \theta_4 = (2n + 1)180 \tag{5.127}$$

Rules for sketching root-locus.

*1) **Number of branches of root-locus.*** Note that each of the closed-loop poles moves in the s-plane as the parameter k varies. Therefore, the number of branches of the root-locus is equal to the number of closed-loop poles.

*2) **Symmetry.*** For all physical systems, the coefficients of the characteristic equation are real. As a result, if any of its roots are complex, then they occur in pairs as complex conjugates. All the real roots lie on either the positive or negative real axes. Hence, the root-locus of a physical system is always symmetric with respect to the real axis.

*3) **Real axis segments.*** Whether a given segment of the real axis forms a part of the root-locus depends on the angle condition shown in Eq. (5.124), i.e., the algebraic sum of the angles subtended at that point because all the poles and zeros must be equal to an odd multiple of 180 deg. The net angle contribution of the complex poles or zeros is zero because they always occur as complex conjugate pairs. Furthermore, the angle contribution of a real axis pole or zero located to the right of a point on the real axis is zero. The angle contribution to a point on the root-locus comes only from those real axis poles and zeros that are located on the left side and is equal to -180 deg for poles and 180 deg for zeros. Because the sum of all such contributions has to be an odd multiple of 180 deg, it is clear that only that part of the real axis segment forms a branch of the root-locus that lies to the left of odd number of poles and/or zeros.

*4) **Starting and ending points of root-locus.*** To understand where the root-locus begins and where it ends as the parameter k is varied from zero to infinity, let

$$G(s) = \frac{N_g(s)}{D_g(s)} \tag{5.128}$$

$$H(s) = \frac{N_h(s)}{D_h(s)} \tag{5.129}$$

Note that $N_g = 0$ and $D_g = 0$ give us, respectively, the zeros and poles of the open-loop transfer function $G(s)$. Similarly, $N_h(s) = 0$ and $D_h(s) = 0$ give, respectively, the zeros and poles of $H(s)$.

Then,

$$T(s) = \frac{kN_g(s)D_h(s)}{D_g(s)D_h(s) + kN_g(s)N_h(s)} \tag{5.130}$$

When the parameter $k \to 0$, the closed-loop transfer function $T(s)$ can be approximated as

$$T(s) = \frac{kN_g(s)D_h(s)}{D_g(s)D_h(s)} \tag{5.131}$$

i.e., when $k \to 0$, the poles of $T(s)$ coincide with the combined open-loop poles of $G(s)$ and $H(s)$. Therefore, the root-locus starts at the open-loop poles of the system.

When $k \to \infty$, we have

$$T(s) = \frac{N_g(s)D_h(s)}{N_g(s)N_h(s)} \tag{5.132}$$

That is, when $k \to \infty$, the poles of $T(s)$ approach the combined zeros of $G(s)$ and $H(s)$. In other words, the root-locus ends at the open-loop zeros of the system. Summarizing, the root-locus starts at the open-loop poles and ends at the open-loop zeros. This statement implies that the system should have equal number of poles and zeros, which is true if we assume that the missing zeros and poles are located at infinity. To understand this point, consider

$$G(s) = \frac{k}{s(s+3)(s+5)} \tag{5.133}$$

We have three poles at $s = 0, -3, -5$ and no finite zeros. Therefore, the missing zeros are located at $s = \infty$.

As $s \to \infty$, $G(s) = 1/s^3 = 0$, i.e., $G(s)$ has three zeros at $s = \infty$.

Consider

$$G(s) = s \tag{5.134}$$

This system has a zero at $s = 0$ and a pole at infinity because, as $s \to \infty$, $G(s) \to \infty$. Similarly, $G(s) = 1/s$ has a zero at infinity because $G(s) \to 0$ as $s \to \infty$.

*5) **Asymptotes.*** The asymptotes give the behavior of the root-locus as the parameter k approaches infinity. The point of intersection of the asymptotes with the real axis σ_o (see Fig. 5.15a) and the slopes of the asymptotes M at this point are given by

$$\sigma_o = \frac{\sum \text{poles} - \sum \text{zeros}}{n_p - n_z} \tag{5.135}$$

$$M = \tan\frac{(2n+1)\pi}{n_p - n_z} \tag{5.136}$$

where n_p and n_z are the number of open-loop poles and zeros, respectively, and $n = 0, \pm 1, \pm 2, \ldots$. The running index n gives the slopes of the asymptotes that form the branches of the root-locus as $k \to \infty$.

Imaginary axis crossing. Another characteristic feature that is of interest in the root-locus method is the point where the root-locus crosses the imaginary axis because the system stability changes at this point. If the imaginary axis crossing is from right to left of the s-plane, the closed-loop becomes stable as the gain is increased. If it is from left to right, then the closed-loop system becomes unstable on increasing the gain.

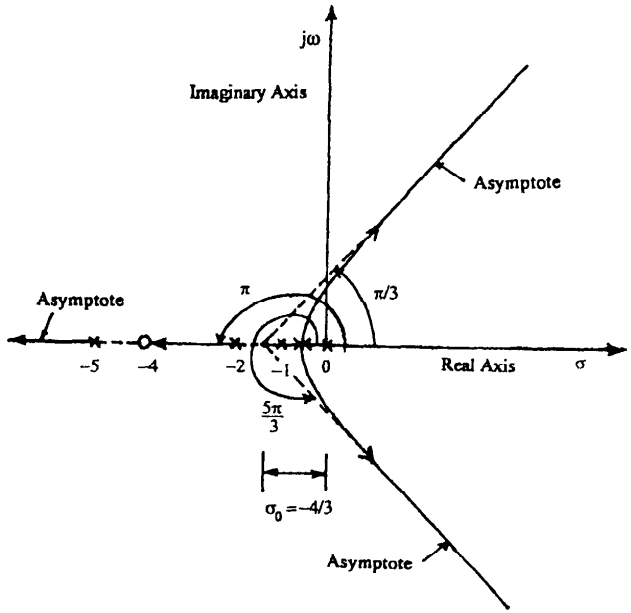
The point(s) where the root-locus crosses the imaginary axis can be determined by 1) using the Routh's criterion and finding the values of the gain k that give all the zeros in any one row of the Routh's table or 2) substituting $s = j\omega$ in the characteristic equation, setting both real and imaginary parts to zero and solving for the gain k and frequency ω. We will illustrate this second procedure in the following example.

Example 5.5

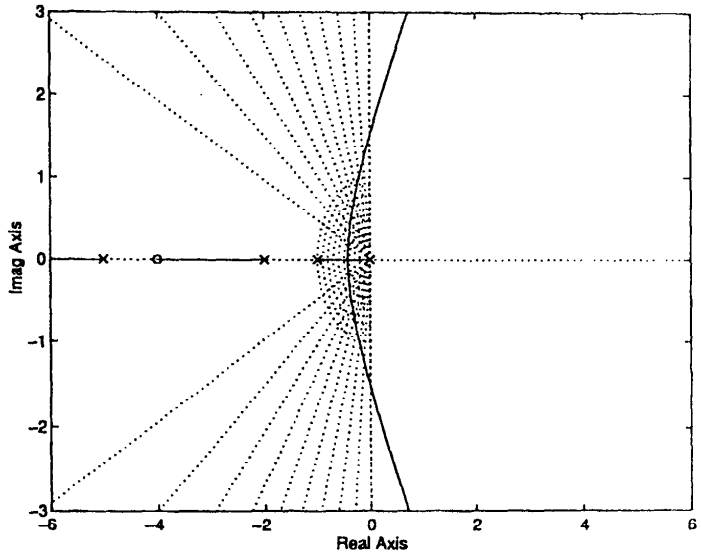
Sketch the root-locus for the unity feedback system with

$$G(s) = \frac{k(s+4)}{s(s+1)(s+2)(s+5)}$$

Solution. We have four poles at $s = 0, -1, -2, -5$ and only one finite zero at $s = -4$. Therefore, the other three missing zeros are at infinity. We have four branches of the root-locus. Furthermore, the root-locus will be symmetrical with respect to the real axis. That segment of the real axis forms a part of the root-locus, which lies to the left of the odd number of poles and/or zeros. Thus, the real axis segment between the poles at 0 and -1 and between the pole at -2 and zero at -4 and all the real axis that is to the left of the pole at -5 forms the branches of the root-locus.



a)



b)

Fig. 5.15 Root-locus for Example 5.5.

We should have three asymptotes corresponding to three branches of the root-locus, which seek zeros at infinity. We have

$$\sigma_o = \frac{\sum \text{poles} - \sum \text{zeros}}{n_p - n_z}$$

$$= \frac{(0-1-2-5)-(-4)}{4-1}$$

$$= -\frac{4}{3}$$

$$M = \tan\frac{(2n+1)\pi}{n_p - n_z}$$

$$= \tan\frac{(2n+1)\pi}{3}$$

$$= \tan\frac{\pi}{3} \qquad n = 0$$

$$= \tan\pi \qquad n = 1$$

$$= \tan\frac{5\pi}{3} \qquad n = 2$$

With this information, the root-locus can be sketched as shown in Fig. 5.15a. The root-locus crosses the imaginary axis from the left half to the right half of the s-plane, i.e., the closed-loop system becomes unstable as the value of the gain k is increased beyond this point.

The value of the gain k and frequency ω where the root-locus crosses the imaginary axes can be obtained as follows.

The characteristic equation is

$$s^4 + 8s^3 + 17s^2 + s(10+k) + 4k = 0$$

Substituting $s = j\omega$, we obtain

$$\omega^4 - 17\omega^2 + 4k + j(-8\omega^3 + \omega[10+k]) = 0$$

Equating real and imaginary parts to zero, we get $k = 8.4856$ and $\omega = \pm 1.5201$.

MATLAB[4] is a convenient tool for plotting the root-locus. The MATLAB command RLOCUS sketches the root-locus, and the command RLOCFIND enables us to find the value of the gain k and the location of the closed-loop poles corresponding to any point on the root-locus. Using RLOCFIND, we find that $k = 8.8$ and $\omega = 1.55$ when the root-locus crosses the imaginary axis. These values are in good agreement with the analytical values. Furthermore, the corresponding locations of the closed-loop poles are -5.15, -2.88, and 0 ± 1.55.

The root-locus obtained using MATLAB[4] is shown in Fig. 5.15b.

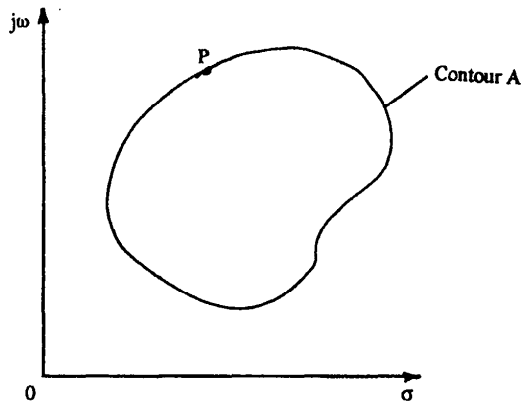
We can also find other information using MATLAB.[4] For example, we can find the value of the gain k so that the closed-loop system is stable and operates with a damping ratio ζ of 0.4. Using RLOCFIND, we obtain $k = 2.0$ and closed-loop poles $p = -5.04$, -2.4, and $-0.3 \pm j0.8$.

5.7.3 Nyquist Stability Criterion

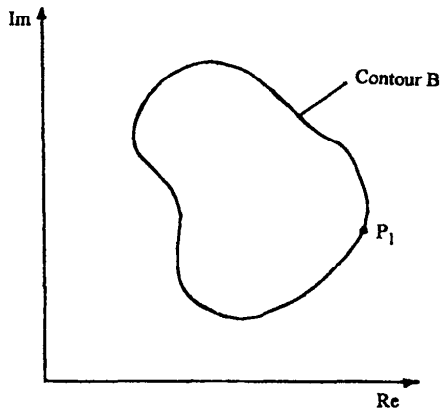
Concept of mapping. Before we discuss the Nyquist stability criterion, let us briefly review the concept of mapping. Suppose we are given a contour A in the the s-plane as shown in Fig. 5.16a and a function $F(s) = s^2 + 2s + 1$. Consider a point P on the contour A in the s-plane, and let the coordinates of point P be $4 + j3$. If we substitute this complex number into the given function $F(s)$, we get another complex number

$$F(s) = (4 + j3)^2 + 2(4 + j3) + 1 = 16 + j30 \qquad (5.137)$$

Suppose we plot the real and imaginary parts of this number in another plane, called the F-plane; we get point P_1 as shown in Fig. 5.16b. The point P_1 in the F-plane is said to be the image of the point P in the s-plane. Here, $F(s)$ is called the mapping function. In a similar way, we can map all other points on contour A to

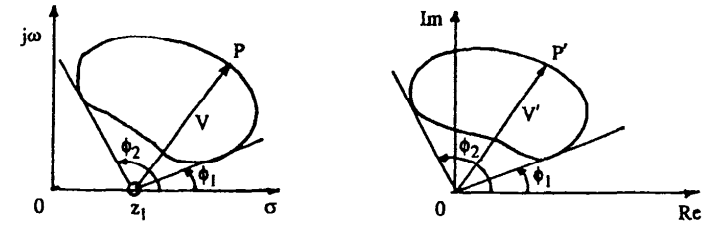


a) s-plane

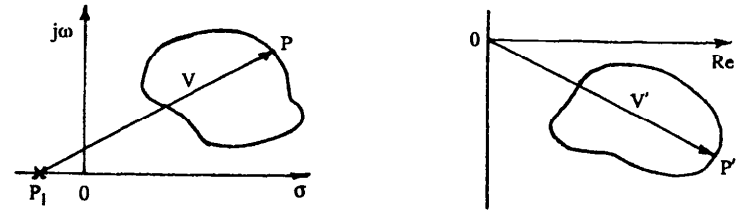


b) F-plane

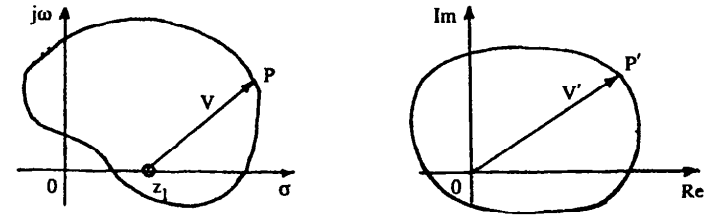
Fig. 5.16 Concept of mapping.



a) $F(s) = s - z_1$



b) $F(s) = 1/(s - p_1)$



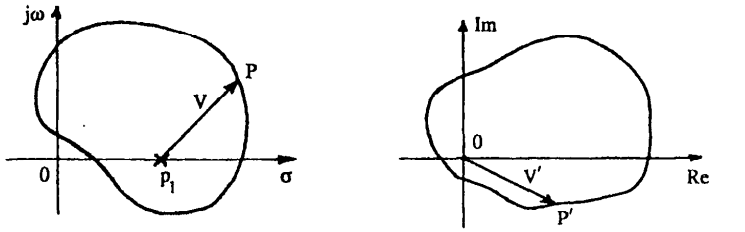
c) $F(s) = s - z_1$

Fig. 5.17 Illustration of contour mapping.

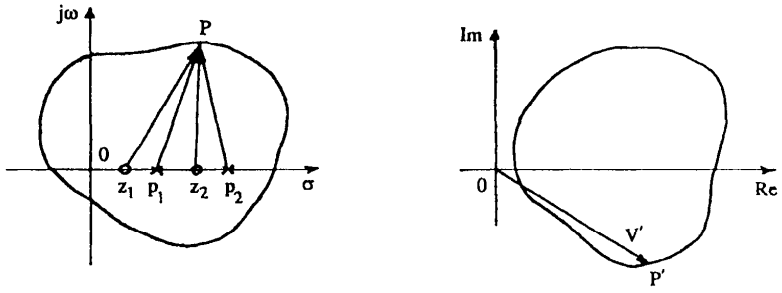
corresponding points in the F-plane and obtain the contour B. Then the contour A in the s-plane is said to be mapped to the contour B in the F-plane. We assume that the mapping is one to one, i.e., for every point in the s-plane, there is one and only one corresponding point in the F-plane and vice versa.

To understand the concept of mapping further, let $F(s) = s - z_1$ and let the point $s = z_1$, which is the zero of $F(s)$, lie outside the contour A as shown in Fig. 5.17a. Instead of using the coordinates of point P, let us use the vector approach. Every point P on the contour A is associated with a vector V. Let V' be the image vector in the F-plane. For this case, $|V'| = |V|$ and $\angle V' = \angle V$. Now as we move clockwise along the contour A, the magnitude and phase of the vector V vary. The phase oscillates between the two limiting values ϕ_1 and ϕ_2. In this case, a clockwise movement along the contour A corresponds to a clockwise movement along the image contour B in the F-plane.

Now let $F(s) = 1/(s - p_1)$ and let the pole $s = p_1$ lie outside the contour A as shown in Fig. 5.17b. For this case, $|V'| = 1/|V|$ and $\angle V' = -\angle V$. As a result, the contour A in the first quadrant maps to contour B in the fourth quadrant. Observe



d) $F(s) = 1/(s - p_1)$



e) $F(s) = (s - z_1)(s - z_2)/(s - p_1)(s - p_2)$

Fig. 5.17 Illustration of contour mapping, continued.

that a clockwise movement along the contour A in the s-plane corresponds to a counterclockwise movement along the contour B in the F-plane because the phase angle in the F-plane is negative of that in the s-plane.

Suppose the zero of $F(s) = s - z_1$ lies inside the contour A as shown in Fig. 5.17c. Then the vector V' makes one complete rotation of 360 deg in the F-plane so that the contour B encloses the origin. Similarly, we have a pole of the mapping function $F(s) = 1/(s - p_1)$ that lies inside the contour A; then the image contour B in the F-plane also encloses the origin as shown in Fig. 5.17d. If the contour A encloses an equal number of poles and zeros of the mapping function $F(s)$, then clockwise encirclement of the origin due to the zeros cancels the counterclockwise encirclement due to poles, and the image contour B does not enclose the origin as shown in Fig. 5.17e.

Nyquist plot. Suppose the contour A in the s-plane is a semicircle of infinite radius covering the entire right half of the s-plane, then the corresponding image contour in the F-plane is said to be the Nyquist plot of right half of the s-plane through the given mapping function $F(s)$. If we have zeros and/or poles in the right half of the s-plane, then the image contour B in the F-plane will encircle the origin n times where $n = n_z - n_p$ and n_z and n_p are, respectively, the number of zeros and poles of the mapping function $F(s)$ located in the right half of the s-plane. If $n > 0$, then we will have n clockwise encirclements and, if $n < 0$, we will have that many counterclockwise encirclements of the origin.

Nyquist criterion of stability. The transfer function of a closed-loop system is given by

$$T(s) = \frac{G(s)}{1 + G(s)H(s)} \tag{5.138}$$

$$= \frac{N_g D_h}{D_g D_h + N_g N_h} \tag{5.139}$$

where $G(s) = N_g/D_g$ and $H(s) = N_h/D_h$.

The poles of the closed-loop transfer function $T(s)$ are generally not known and have to be determined actually by solving the closed-loop characteristic equation $1 + G(s)H(s) = 0$. Note that the poles of $T(s)$ are the zeros of $1 + G(s)H(s)$. The closed-loop system will be unstable if any of the poles of $T(S)$ are located in the right half of the s-plane. The usefulness of the Nyquist stability criterion is that it enables us to know whether any of the poles of $T(s)$ are located in the right half of the s-plane without actually solving the closed-loop characteristic equation $1+G(s)H(s) = 0$. In this way, it gives us an idea whether the given closed-loop system is stable or not without actually knowing the location of the closed-loop poles. This information is very useful in evaluating the stability of a closed-loop system as a certain system parameter, say the gain k, is varied.

Suppose we make a Nyquist plot of the function $F(s) = 1+G(s)H(s)$. Note that the zeros of this function $F(s)$ are the poles of the closed-loop transfer function $T(s)$ and the poles of $F(s)$ are the combined poles of the open-loop transfer function $G(s)H(s)$, which are known. Then, the Nyquist criterion for stability centers around the determination of the parameter, $N = P - Z$, where N is the number of encirclements of the origin in the F-plane, P is the number of poles of $F(s)$ located in the right half of the s-plane, and Z is the number of zeros of $F(s)$ that are located in the right half of the s-plane. Note that a positive value of N corresponds to counterclockwise encirclement of the origin and a negative value to clockwise encirclement. Here, P is known but Z is not known. Therefore, unless we have a method to determine Z, we cannot sketch a Nyquist plot and determine the system stability. As said before, we do not have an easy method of finding Z.

Suppose we use the function $G(s)H(s)$ as the mapping function instead of $1 + G(s)H(s)$ because all the poles and zeros of the function $G(s)H(s)$ are known. The resulting Nyquist plot is the same as that of $1+G(s)H(s)$ except that it is displaced by one unit to the left of the origin. Then, instead of counting the encirclement of the origin, we can count the encirclement of the point -1. Everything else remains the same, and we can now use the Nyquist plot to determine the system stability. With this modification, the Nyquist criterion for the stability of a closed-loop system can be restated as follows.

If a contour A in the s-plane that covers the entire right half of the s-plane is mapped to the F-plane with the mapping function $F(s) = G(s)H(s)$, then the number of closed-loop poles Z that lie in the right half of the s-plane equals the number of open-loop poles P that are in the right half of the s-plane minus the number of counterclockwise rotations N of the Nyquist plot around the point -1 in the F-plane, i.e., $Z = P - N$. For stability of the closed-loop system, Z must be equal to zero.

To understand the Nyquist criterion, let us study two cases shown in Fig. 5.18. Let us assume that somehow we know the zeros of $1 + G(s)H(s)$, which are

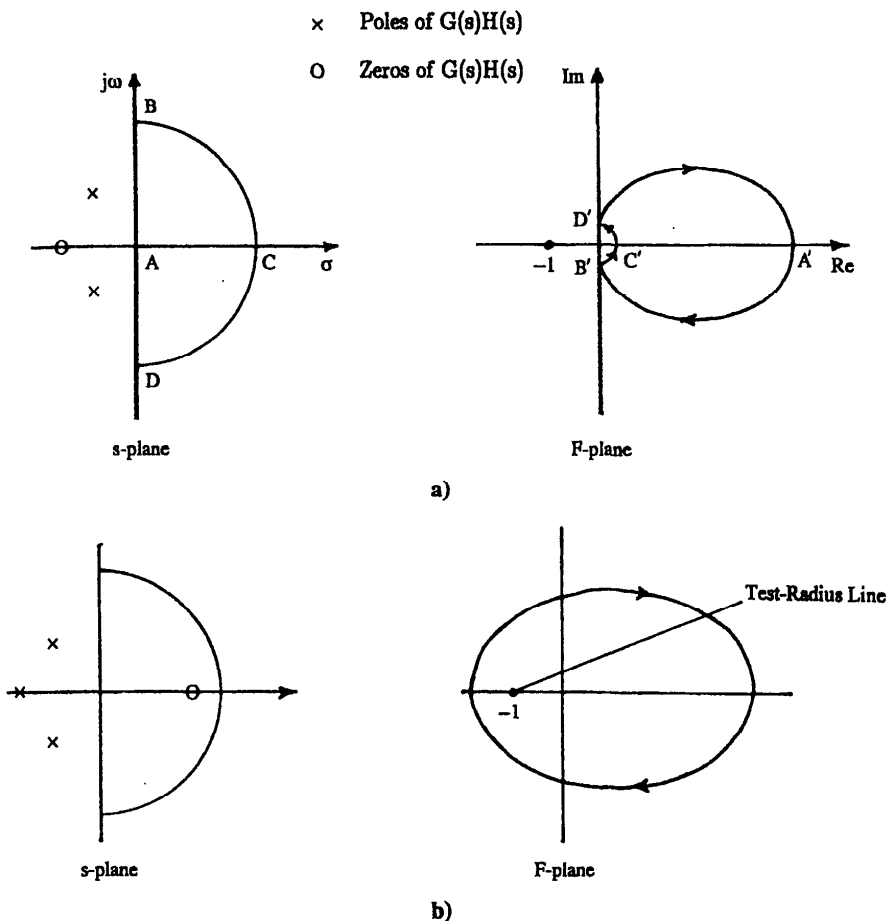


Fig. 5.18 Nyquist plots for mapping function $1 + G(s)H(s)$.

poles of the closed-loop transfer function $T(s)$. The poles of $1 + G(s)H(s)$ are the combined poles of the open-loop transfer function $G(s)H(s)$ and are known. Let the open circles denote the zeros of $1+G(s)H(s)$ and cross the poles of $G(s)H(s)$. For Fig. 5.18a, there are no poles or zeros of $1 + G(s)H(s)$ in the right half of the s-plane, i.e., $P = 0$ and $Z = 0$. Hence, the Nyquist plot will not encircle the point -1 in the F-plane as shown in Fig. 5.18b. For this case, $N = P - Z = 0$ and the system is stable. For Fig. 5.18b, we have one zero of $1 + G(s)H(s)$ located in the right half of the s-plane (unstable system). Therefore, $Z = 1$. Furthermore, $P = 0$ because there are no poles of $1 + G(s)H(s)$ located in right half of the s-plane. Hence, according to the Nyquist criterion, $N = P - Z = -1$, i.e., the Nyquist plot in the F-plane will encircle the point -1 once in the clockwise direction as shown in Fig. 5.19b.

The number of encirclements can be conveniently determined by drawing a radial line from the point -1 and counting the number of intersections with the

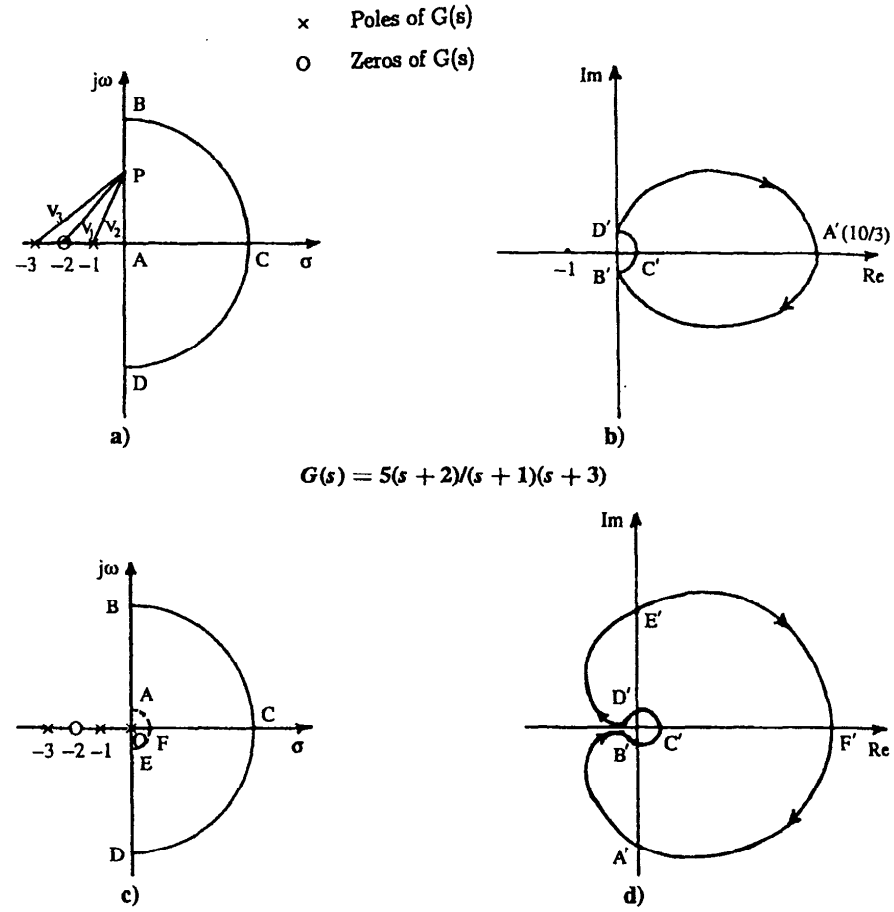


Fig. 5.19 Nyquist plots for Examples 5.6 and 5.7.

Nyquist plot as shown. However, the reader should keep in mind that, in a given problem, the locations of closed-loop poles are not known as assumed in this discussion.

Example 5.6

Draw the Nyquist plot for a unity feedback system with

$$G(s) = \frac{5(s+2)}{(s+1)(s+3)}$$

Solution. The first step is to select some points along the Nyquist contour in the s-plane as shown in Fig. 5.19a. Consider an arbitrary point P. Let V_1, V_2, and

V_3 be the vectors drawn to point P from the zeros and poles as shown. Then,

$$|V_P'| = \frac{5|V_1|}{|V_2||V_3|}$$

$$\angle V_P' = \angle V_1 - \angle V_2 - \angle V_3$$

If point P coincides with A, $|V_1| = 2$, $|V_2| = 1$, and $|V_3| = 3$ so that $|V_A'| = 10/3$. The phase angles $\angle V_1 = \angle V_2 = \angle V_3 = 0$ so that $\angle V_A' = 0$. Thus the point A in the s-plane maps to point A' on the real axis in the F-plane with abscissa equal to $10/3$. In a similar fashion, we find the magnitude and phase angles at other image points such as $|V_B'| = 0$, $\angle V_B' = -90$ deg; $|V_C'| = 0$, $\angle V_C' = 0$; and $|V_D'| = 0$, $\angle V_D' = 90$ deg.

Based on this information, the Nyquist plot can be sketched as shown in Fig. 5.19b. Observe that the Nyquist plot does not encircle the origin but just goes around it in a semicircle of "zero" radius. As we move clockwise in the s-plane starting at point A, we move in the counterclockwise direction in the F-plane from A'.

In this example we didn't have any poles of the mapping function on the imaginary axis. If we did, then we have to draw semicircles of infinitesimally small radii around each one to prevent a breakdown of the mapping procedure at these points. We illustrate the procedure of drawing such Nyquist plots with the help of Example 5.7.

Example 5.7

Draw the Nyquist plot for the system with

$$G(s) = \frac{5(s+2)}{s(s+1)(s+3)}$$

Solution. Here, we have a pole at $s = 0$ at the origin. As said above, we draw a semicircle of infinitesimally small radius around it as shown in Fig. 5.19c. The magnitudes and phase angles of the image points A'–F' are as follows: $|V_F'| = \infty$, $\angle V_F' = 0$; $|V_A'| = \infty$, $\angle V_A' = -90$ deg; $|V_B'| = 0$, $\angle V_B' = -180$ deg; $|V_C'| = 0$, $\angle V_C' = 0$; $|V_D'| = 0$, $\angle V_D' = 180$ deg; and $|V_E'| = \infty$, $\angle V_E' = 90$ deg.

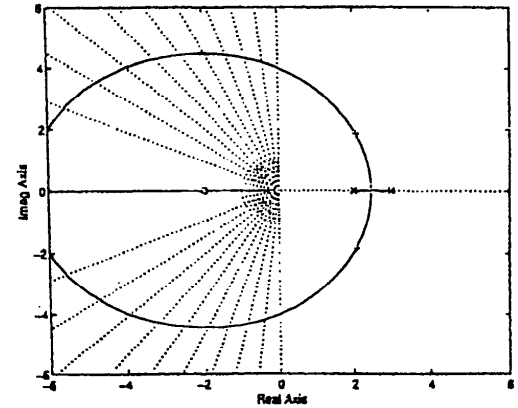
The Nyquist plot is shown in Fig. 5.19d. Observe that the small circle of "zero" radius encircles the origin in the F-plane in the counterclockwise direction because the phase angle changes from -180 deg at B' to $+180$ deg at D'.

Example 5.8

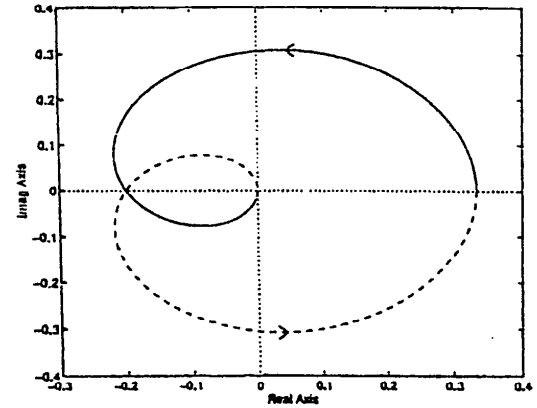
Determine the stability of a unity feedback system given by

$$G(s) = \frac{k(s+2)}{(s-2)(s-3)}$$

Solution. Here, $H(s) = 1$. Furthermore, assume that the gain k is a variable. We use MATLAB[4] and the root-locus is drawn as shown in Fig. 5.20a. We observe that the root-locus starts in the right half of the s-plane, implying that the closed-loop system is unstable for small values of the gain k and, for $k \geq 4.9420$, the



a) Root-locus



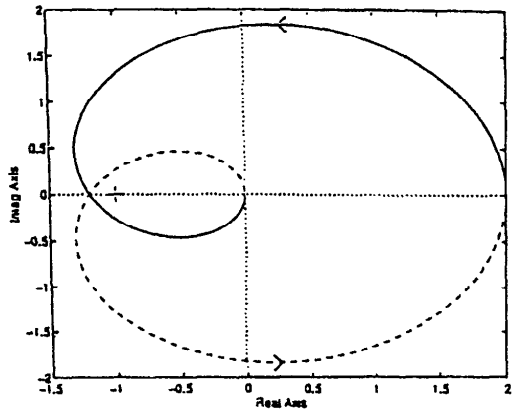
b) Nyquist plot for $k = 1$

Fig. 5.20 Root-locus and Nyquist plots for Example 5.8.

root-locus crosses over to the left half of the s-plane, indicating that the closed-loop system becomes stable for $k \geq 4.9420$.

Now using MATLAB,[4] let us draw the Nyquist plot as shown in Fig. 5.20b assuming $k = 1$. We observe that the Nyquist plot in the F-plane does not encircle the point -1, which means that $N = 0$. Instead, it intersects the negative real axis at $s = -0.2$. We have $P = 2$ because the poles $s = 2$ and $s = 3$ of $G(s)$ are located in the right half of the s-plane. According to Nyquist criterion, we get $Z = P - N = 2$. In other words, the Nyquist criterion predicts that there are two poles of the closed-loop transfer function $T(s)$ located in the right half of the s-plane and, therefore, the system is unstable. From the root-locus of Fig. 5.20a, we find this to be true.

Suppose we increase the gain k beyond unity. Then the Nyquist plot will expand and eventually touch the critical point -1. When this happens, the value of k is



c) Nyquist plot for $k = 6$

Fig. 5.20 Root-locus and Nyquist plots for Example 5.8, continued.

equal to $1/0.2 = 5$, which is quite close to that predicted by the root-locus method. For higher values of gain k, the Nyquist plot will expand further and will encircle the critical point -1 twice in a counterclockwise direction as shown in Fig. 5.20c for $k = 6$. We then have $N = 2$ and $Z = P - N = 2 - 2 = 0$, which indicates that the closed-loop system has become stable.

This example has illustrated an important concept that the stability of closed-loop systems depends on the value of the gain. Feedback systems that are unstable for low values of gain can become stable for higher values of gain, and those that are stable for low values of gain can become unstable for higher values of gain. The Nyquist criterion can be used to determine the gain at the crossover point. This kind of dependence of the system stability on the value of the gain leads to the concepts of gain and phase margins as discussed in the next section.

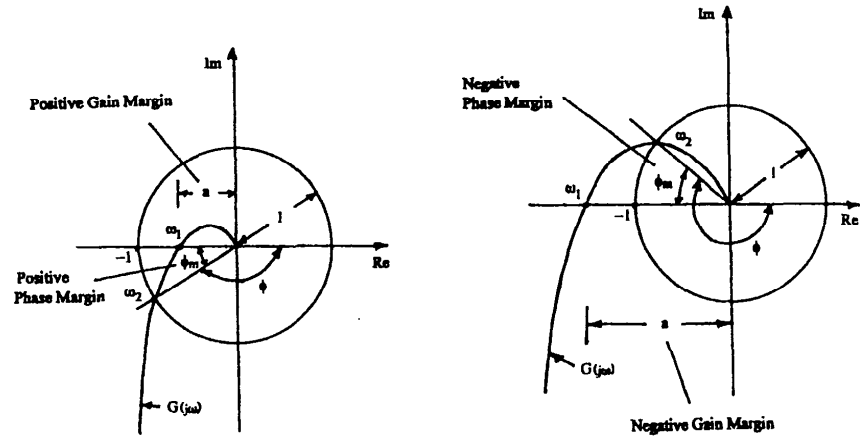
5.7.4 Gain and Phase Margins

The Nyquist stability criterion enables us to define two quantities that are measures of the level of stability of a given closed-loop system. These quantities are the so-called gain and phase margin that are widely used in the control system analyses and design. Generally, the systems with higher values of gain and phase margins have better capability to withstand changes in the system parameters before becoming unstable.

The concepts of gain margin and phase margin are illustrated in Fig. 5.21.

The gain margin G_M is the reciprocal of the magnitude $|G(j\omega)|$ at the phase crossover frequency. The phase crossover frequency is the frequency ω_1 at which the phase angle of the open-loop transfer function $G(j\omega)$ is -180 deg. The gain margin is given by

$$G_M = \frac{1}{|G(j\omega_1)|} \tag{5.140}$$



a) Stable system b) Unstable system

Fig. 5.21 Gain and phase margins.

If $a = |G(j\omega_1)|$ (see Fig. 5.21), then

$$G_M = \frac{1}{a} \tag{5.141}$$

The gain margin is usually expressed in decibels as

$$G_M(db) = 20 \log_{10}\left(\frac{1}{a}\right) \tag{5.142}$$

$$= -20 \log_{10} a \tag{5.143}$$

The gain margin expressed in decibels is positive if $a < 1$ (Fig. 5.21a) and is negative if $a > 1$ as shown in Fig. 5.21b. A positive gain margin (in decibels) means that the system is stable, and a negative gain margin (in decibels) means that the system is unstable. For a stable minimum phase system, the value of the gain margin indicates how much the open-loop gain can be increased before the closed-loop system becomes unstable. For example, a gain margin of 30 db implies that the open-loop gain can be increased by a factor of 31.6228 before the closed-loop system becomes unstable. On the other hand, if the gain margin is -30 db, then the closed-loop system is already unstable, and the gain has to be reduced by a factor of 31.6228 to make the closed-loop system stable.

The phase margin is defined as the amount of additional phase lag at the gain crossover frequency that can be introduced in the open-loop system to make the closed-loop system unstable. The gain crossover frequency ω_2 is that frequency when the magnitude of the open-loop transfer function $G(j\omega)$ is unity.

The phase margin is usually denoted by ϕ_M and is expressed in degrees and is given by

$$\phi_M = 180 + \phi \tag{5.144}$$

where $\phi = \angle G(j\omega_2)$ is the open-loop phase angle at the gain crossover frequency as shown in Fig. 5.21. Note that the value of the phase angle ϕ is negative in Fig. 5.21 because it is measured in the clockwise direction. Thus, for a stable system (Fig. 5.21a), the phase margin is positive because $|\phi| < 180$ deg, and for an unstable system the phase margin is negative because $|\phi| > 180$ deg as indicated in Fig. 5.21b. For example, a phase margin of 30 deg indicates that the open-loop phase lag can be increased further by 30 deg before making the system unstable. On the other hand, a phase margin of -30 deg indicates that the system is already unstable, and the open-loop phase lag has to be reduced by 30 deg to make the system stable.

It is important to bear in mind that the Nyquist plots shown in Fig. 5.21 are drawn for unity gain. For the purpose of estimating the gain and phase margins, a simplified Nyquist plot mapping only the positive imaginary axis in the s-plane is usually sufficient. The part of the Nyquist plot in the F-plane that corresponds to the semicircle of infinite radius in the s-plane is usually the circle(s) of "zero" radius around the origin in the F-plane and hence is not needed in evaluating the gain and phase margins. Furthermore, the mapping of the positive imaginary axis is equivalent to studying the frequency response of the system because the Nyquist diagram is a polar plot of the magnitude vs phase of the open-loop transfer function with frequency as an implicit variable.

The Bode plot also offers an alternative and a convenient method to estimate the gain and phase margins as illustrated in Fig. 5.22.

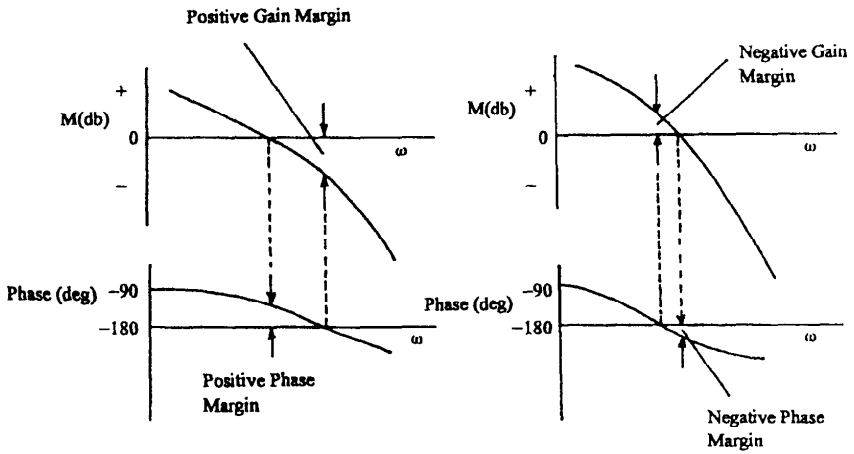


Fig. 5.22 Gain and phase margins using Bode plots.

5.8 Relations Between Time-Domain and Frequency-Domain Parameters

Generally, the performance requirements for control systems are specified in terms of time-domain parameters like rise time T_r, settling time T_s, time for peak amplitude T_p, and percent overshoot O_s. In the following, we present some relations between these time-domain parameters and frequency-domain parameters. These relations will be useful in the analyses and design of control systems using frequency-domain methods.

Consider a second-order system whose open-loop and unity feedback closed-loop transfer functions are given by

$$G(s) = \frac{\omega_n^2}{s(s + 2\zeta\omega_n)} \tag{5.145}$$

$$T(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{5.146}$$

Let M denote the magnitude of the closed-loop frequency response. Then,

$$M = |T(j\omega)| = \frac{\omega_n^2}{\sqrt{\left(\omega_n^2 - \omega^2\right)^2 + 4\zeta^2\omega_n^2\omega^2}} \tag{5.147}$$

A typical plot of M vs ω is shown in Fig. 5.23.

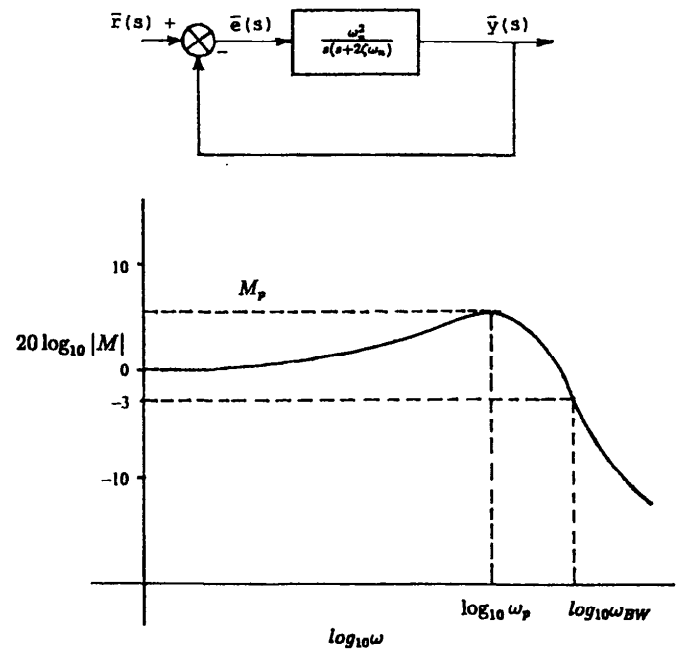


Fig. 5.23 Frequency-response parameters.

To determine the peak amplitude M_p and the corresponding frequency ω_p, take the squares on both sides of Eq. (5.147), differentiate with respect to ω, and equate the resulting expression to zero to obtain

$$M_p = \frac{1}{2\zeta\sqrt{1-\zeta^2}} \tag{5.148}$$

$$\omega_p = \omega_n\sqrt{1-2\zeta^2} \tag{5.149}$$

Equations (5.148) and (5.149) show that M_p and ω_p of the closed-loop frequency response are directly related to the damping ratio ζ. The percent overshoot O_s and damping ratio ζ are related through Eq. (5.75), which is reproduced here in the following:

$$\zeta = \frac{-\ell_n(O_s/100)}{\sqrt{\pi^2 + \ell_n^2(O_s/100)}} \tag{5.150}$$

Using these relations, given the value of M_p, we can determine the percent overshoot O_s of the given closed-loop system and vice versa.

The bandwidth ω_{BW} is another important characteristic of the closed-loop frequency response. The bandwidth is defined as that frequency at which the magnitude M drops to 0.707 or $1/\sqrt{2}$ of its value at $\omega = 0$. This is also equivalent to a drop by 3 db. From Eq. (5.147), we find that $M = 1$ when $\omega = 0$. Therefore, substituting $M = 1/\sqrt{2}$ and $\omega = \omega_{BW}$ in Eq. (5.147), we obtain

$$\omega_{BW} = \omega_n\sqrt{(1-2\zeta^2) + \sqrt{4\zeta^4 - 4\zeta^2 + 2}} \tag{5.151}$$

The settling time T_s and time for peak amplitude T_p given by Eqs. (5.77) and (5.72) are reproduced in the following:

$$T_s = \frac{4}{\zeta\omega_n} \tag{5.152}$$

$$T_p = \frac{\pi}{\omega_d} \tag{5.153}$$

$$= \frac{\pi}{\omega_n\sqrt{1-\zeta^2}} \tag{5.154}$$

Using Eq. (5.151), we can rewrite these relations in terms of bandwidth ω_{BW} as follows:

$$T_s = \left(\frac{4}{\omega_{BW}\zeta}\right)\sqrt{(1-2\zeta^2) + \sqrt{4\zeta^4 - 4\zeta^2 + 2}} \tag{5.155}$$

$$T_p = \left(\frac{\pi}{\omega_{BW}\sqrt{1-\zeta^2}}\right)\sqrt{(1-2\zeta^2) + \sqrt{4\zeta^4 - 4\zeta^2 + 2}} \tag{5.156}$$

The above equations give relations between the time-domain parameters T_s and T_p and the frequency-response parameter ω_{BW} for second-order systems. These relations contain the two basic system parameters, ζ and ω_n.

Another important parameter of the frequency-domain design method is the phase margin ϕ_M. A relation between phase margin ϕ_m and the damping parameter ζ can be obtained as follows.

Let $\omega = \omega_1$ when the magnitude of the open-loop frequency response is unity, i.e., $|G(j\omega)| = 1$, or

$$|Gj(\omega_1)| = \frac{\omega_n^2}{|(-\omega_1^2 + j2\zeta\omega_n\omega_1)|} = 1 \tag{5.157}$$

so that

$$\omega_1 = \omega_n\sqrt{-2\zeta^2 + \sqrt{1+4\zeta^4}} \tag{5.158}$$

The phase angle of $G(j\omega)$ at $\omega = \omega_1$ is given by

$$\angle G(j\omega_1) = -90 - \tan^{-1}\frac{\omega_1}{2\zeta\omega_n} \tag{5.159}$$

$$= -90 - \tan^{-1}\left(\frac{\sqrt{-2\zeta^2 + \sqrt{1+4\zeta^4}}}{2\zeta}\right) \tag{5.160}$$

The phase margin is given by

$$\phi_M = 180 + \angle G(j\omega_1) = 90 - \tan^{-1}\left(\frac{\sqrt{-2\zeta^2+\sqrt{1+4\zeta^4}}}{2\zeta}\right) \tag{5.161}$$

$$= \tan^{-1}\left(\frac{2\zeta}{\sqrt{-2\zeta^2 + \sqrt{1+4\zeta^4}}}\right) \tag{5.162}$$

The variation of ϕ_M with ζ is shown in Fig. 5.24.

5.9 Design of Compensators

The response characteristics of a given system depend on the internal physical nature of the system and may not meet the specified performance requirements. A simple modification of the plant dynamics is an obvious first choice to meet the performance specifications. However, such a thing may not be possible in practice because the plant can be quite complex so that it may not be easy to affect the necessary modifications. In such cases, the adjustment of the gain is the next obvious step. However, in many cases, this alone may not be sufficient, and one may have to redesign the entire plant. Such a process can be quite expensive and time consuming. A simpler alternative is to introduce a compensator into the system that compensates for the deficiencies of the original plant so that the overall system, including the compensator, meets the specified performance requirements.

A compensator is also called a controller. Compensators that employ pure integration to improve steady-state error or pure differentiation to speed up the transient

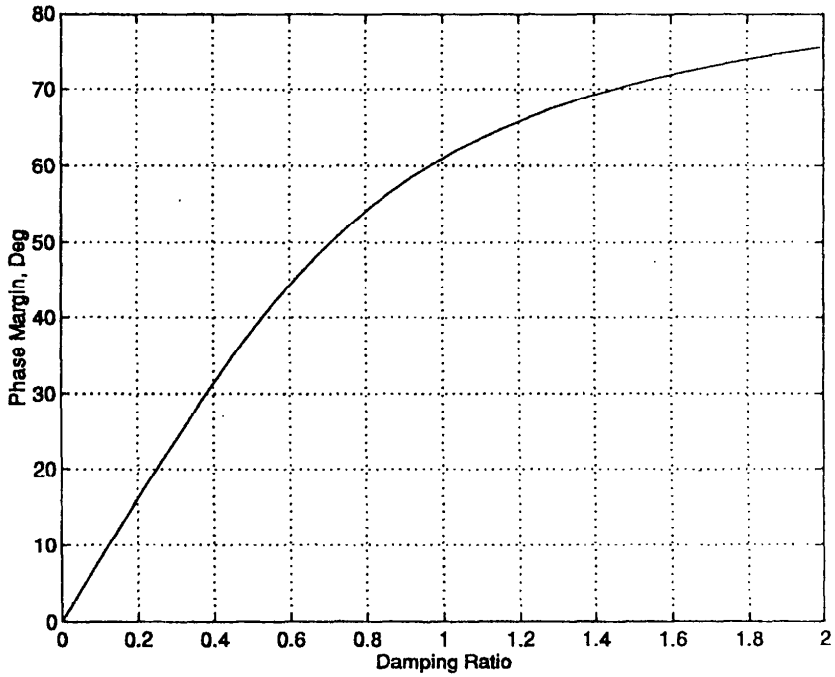


Fig. 5.24 Relation between phase margin and damping ratio.

response are called ideal compensators. However, a disadvantage of ideal compensators is that their implementation requires active networks like operational amplifiers. It is possible to construct passive networks involving resistors, inductors, and capacitors and to achieve performances close to those of ideal compensators. Such compensators are called either lead or lag or lead-lag or lag-lead compensators depending on their type. We will not be dealing with the issues concerning hardware implemention of the compensator designs discussed here. The interested reader may refer elsewhere.[1-3]

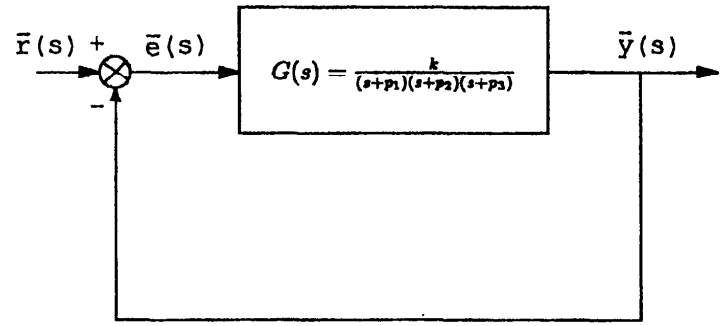
In this section, we will discuss the design of compensators to obtain the specified transient response or steady-state error or both for single-input-single-output systems. Basically, there are two design methods: 1) the root-locus method and 2) the frequency-response method. The frequency-response method has the advantage that the explicit knowledge of the plant transfer function is not needed. All that is needed is the plant frequency response. However, the main disadvantage of the frequency-response method is that the quantities one deals with are not directly related to the time-response parameters, which are specified as design requirements. Hence, the design becomes more of trial and error, and the number of iterations depends on the knowledge and experience of the designer. On the other hand, the root-locus method has a clear advantage in that the quantities it deals with are directly related to the design requirements. Furthermore, the correlation of the root-locus with time response is quite good. Also, the effect of changing compensator

parameters can be easily observed by studying the root-locus. With the availability of tools like MATLAB,[4] the root-locus method becomes very attractive for control system design. However, a disadvantage of the root-locus method is that it becomes more complex as the order of the system increases.

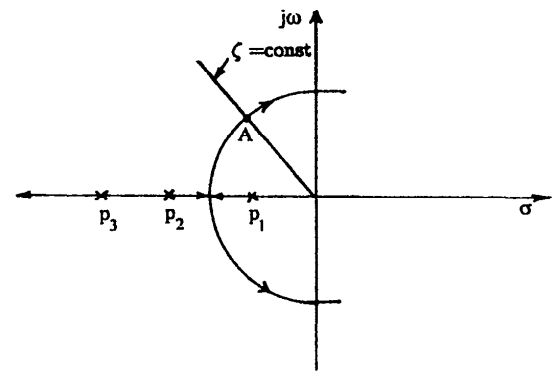
Here, we will use the root-locus method for compensator design of single-input-single-output systems. Readers interested in using frequency-domain methods may refer elsewhere.[1,2] For multi-input-multi-output systems, the modern state-space methods are quite convenient. We will discuss these approaches in Section 5.10.

5.9.1 Proportional–Integral Compensator

To understand the basic principles of designing an integral compensator, consider a type "0" system with unity feedback as shown in Fig. 5.25a. The root-locus for this system is sketched in Fig. 5.25b. Let us assume that the system is operating at point A, having the desired transient response. Recall that the transient response is characterized by the settling time T_s, time for peak amplitude T_p, and the rise time T_r, all of which depend on the damping ratio ζ and frequency ω. At point A,

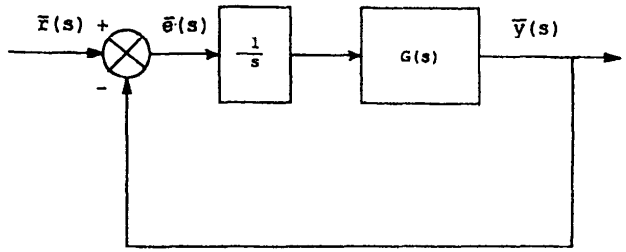


a) Given unity feedback system

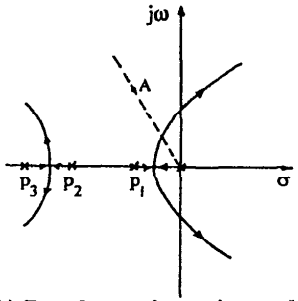


b) Root-locus

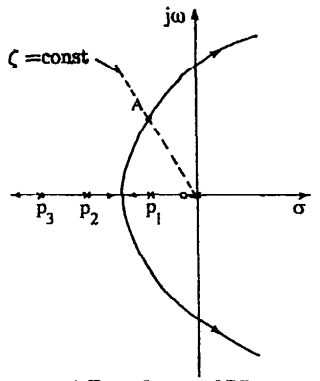
Fig. 5.25 System operating at the desired point A.



a) Integral compensator



b) Root-locus of pure integral compensator



c) Root-locus of PI compensator

Fig. 5.26 Proportional–integral compensator.

the closed-loop poles are a pair of complex roots and one real root. The steady-state error of this system is equal to $\epsilon(\infty) = 1/(1 + K_p)$, where $K_p = k/p_1 p_2 p_3$. Because K_p is finite, the steady-state error is nonzero.

To drive the steady-state error to zero, let us make it a type "1" system by adding a pure integrator in the forward path as shown in Fig. 5.26a. This amounts to adding a pole at the origin as shown in Fig. 5.26b. The root-locus of the entire system is now changed and does not go through point A as shown in Fig. 5.26b. In other words, the steady-state error is driven to zero, but the transient response has changed. To solve this problem, add a compensator zero at $s = -z_c$, which is close to the origin so that this zero almost cancels the compensator pole. Such a compensator is called a proportional–integral (PI) compensator. Now, the root-locus with PI compensation (Fig. 5.26c) is nearly the same as that of the basic uncompensated system (Fig. 5.25b). Therefore, the transient response will remain unaffected while the steady-state error is driven to zero.

The transfer function of a PI compensator is given by

$$G_c(s) = \frac{s + z_c}{s} \tag{5.163}$$

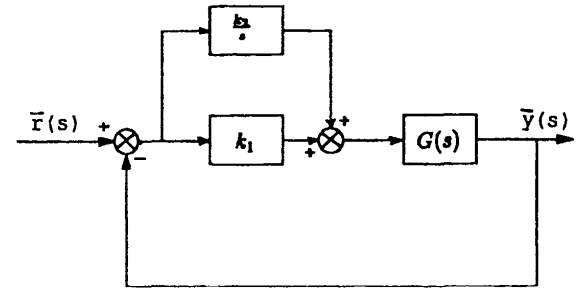


Fig. 5.27 Implementation of a PI compensator.

The schematic implementation of a PI compensator is shown in Fig. 5.27 with

$$G_c(s) = k_1 + \left(\frac{k_2}{s}\right) \tag{5.164}$$

$$= k_1\left[1 + \frac{k_2}{k_1}\left(\frac{1}{s}\right)\right] \tag{5.165}$$

In this implementaion, $k_1 = 1$ and $k_2 = z_c$. Note that the first term on the right-hand side is the "proportional" part and the second term is the "integral" part.

5.9.2 Proportional–Derivative Compensator

Generally, the derivative compensation is used when a simple gain adjustment alone cannot give the desired transient response of the closed-loop system. This concept is illustrated in Fig. 5.28. In Fig. 5.28a, a simple gain adjustment is sufficient because the root-locus passes through the desired operating point A. However, in Fig. 5.28b, the root-locus cannot pass through A for any value of the gain k. The addition of a compensator zero close to the origin modifies the root-locus so that it is made to pass through point A as shown in Fig. 5.28c. Such a compensator that produces a zero in the forward path is called a proportional–derivative (PD) compensator.

The transfer function of a PD compensator is of the form

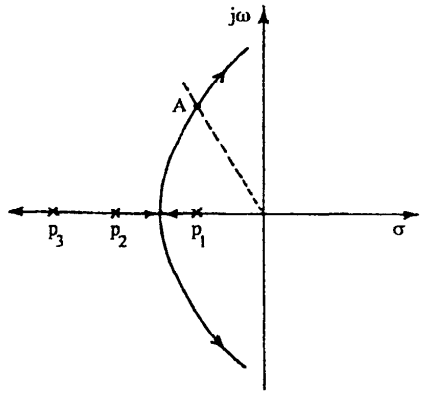
$$G_c(s) = s + z_c \tag{5.166}$$

which is essentially the sum of a differentiator s and a gain z_c.

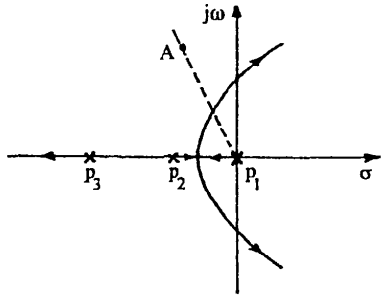
The implementation of a PD compensator is schematically shown in Fig. 5.29. The transfer function of such a compensator can also be written in the following form:

$$G_c(s) = k_1 + k_2 s = k_2\left(\frac{k_2}{k_1} + s\right) \tag{5.167}$$

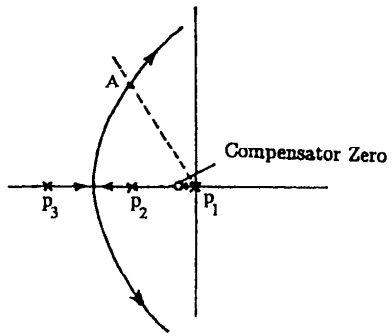
The first term on the right-hand side is the "proportional" part, and the second term is the "derivative" part. The gains k_1 and k_2 are design variables to be determined so that the system attains the specified transient response.



a)



b)



c)

Fig. 5.28 Concept of proportional-derivative compensation.

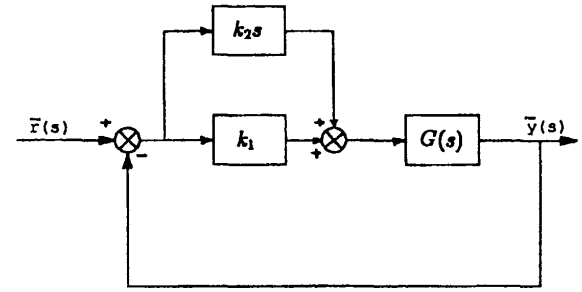


Fig. 5.29 Implementation of PD compensator.

5.9.3 Lead/Lag Compensator

The implementation of a PI or a PD compensator requires active elements. Instead, we can use passive networks to achieve nearly the same objective. A passive network usually produces a pole-zero combination.

The transfer function of a lead/lag compensator is of the form

$$G_c(s) = \frac{s + z_c}{s + p_c} \tag{5.168}$$

A schematic diagram of a lead/lag compensator is shown in Fig. 5.30. By a suitable choice of the locations of pole and zero, we can have either a lag or lead compensator. For example, if the zero is close to the origin ($z_c \simeq 0$) and the pole is farther away to the left of the zero, then it is a lead compensator. On the other hand, if both the pole and zero are located close to the origin with the pole located to the right of the zero, then it is a lag compensator.

Note that for both the lead and lag compensators, the pole and the zero are supposed to be located in the left half of the s-plane.

5.9.4 Proportional–Integral and Derivative Controller

Suppose we want to improve the transient response as well as reduce the steady-state error, then we use the PID (proportional–integral and derivative) controller.

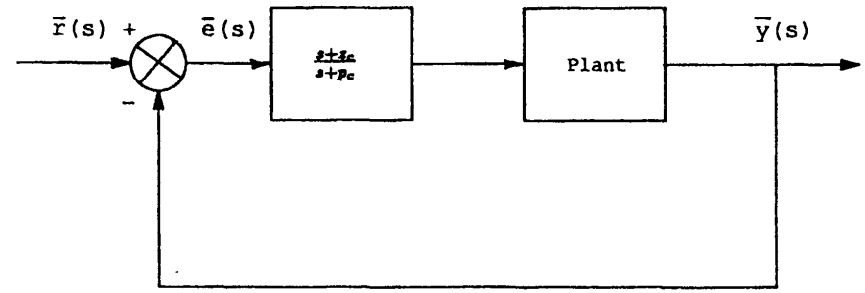


Fig. 5.30 Lead/lag compensator.

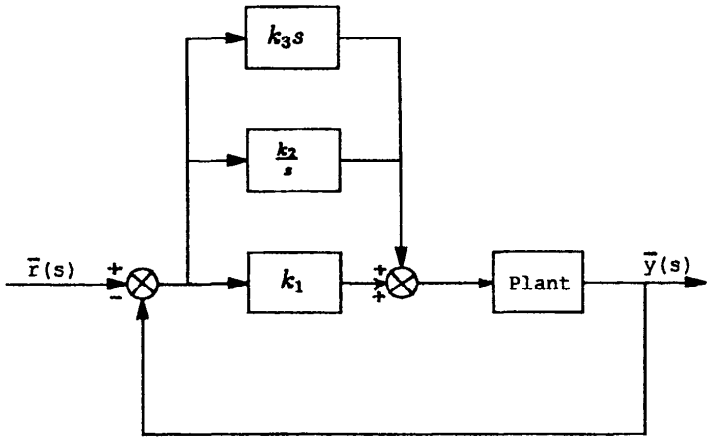


Fig. 5.31 Proportional–integral and derivative controller.

There are two ways of designing a PID controller: 1) design a PD controller first to improve the transient response and then add a PI controller to improve the steady-state error or 2) design a PI controller first and then add a PD controller. Both methods are iterative because one affects the other.

The schematic diagram of a PID controller is shown in Fig. 5.31.

5.9.5 Feedback Compensation

The desired transient response can also be obtained by feedback compensation. With a proper choice of the feedback-loop transfer function, the root-locus can be modified to obtain the desired transient response. This method offers an added advantage that the parts of the system can be isolated for improvement in transient response prior to closing the major loop. This approach is also equivalent to relocating the open-loop poles of the system so that the root-locus is reshaped to obtain the desired closed-loop poles.

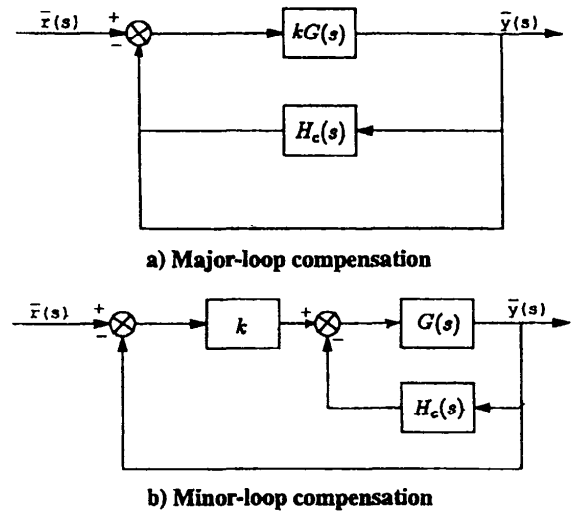
Feedback compensation can be accomplished in two ways: 1) major-loop compensation and 2) minor-loop compensation as schematically shown in Fig. 5.32.

Major-loop compensation. Let $H_c(s) = k_h s$ be the transfer function of the major loop. This form of transfer function is representative of a tachometer or a rate gyro whose input is a displacement and output is the time rate of change of displacement or velocity. For example, if the input is bank angle, then the output will be roll rate.

The closed-loop transfer function is given by

$$T(s) = \frac{kG(s)}{1 + kk_h G(s)\left(s + \frac{1}{k_h}\right)} \tag{5.169}$$

The characteristic equation of this major-loop feedback compensated system is



a) Major-loop compensation

b) Minor-loop compensation

Fig. 5.32 Feedback compensation.

given by

$$1 + kk_h G(s)\left(s + \frac{1}{k_h}\right) = 0 \tag{5.170}$$

Thus, in principle, the major-loop feedback compensation introduces a zero into the system at $s = -1/k_h$ so that the root-locus is reshaped to pass through the desired operating point. By varying the parameter k_h, we can vary the gain as well as the location of this zero. Even though this concept is similar to the PD compensation, there is a difference. The compensator zero in the case of a PD compensator is an open-loop zero, whereas the zero introduced in major-loop feedback compensation is not an open-loop zero.

Minor-loop compensation. With $H_c(s) = k_h s$, the open-loop transfer function of the minor loop is $G_c(s)k_h s$. Thus, the addition of a zero at the origin of the minor-loop root-locus considerably speeds up the response of the minor loop and also has an effect on the overall system performance. Once the gain k_h is adjusted to obtain the desired performance of the minor loop, the outer loop is closed, and the gain k is adjusted to obtain the specified overall system performance. This method of compensation is usually used in aircraft control systems to improve the response to individual degrees of freedom like pitch, roll, or yaw before closing the outer loop as we will discuss in Chapter 6.

Example 5.9

For the system shown in Fig. 5.33, 1) design a PI compensator to reduce the steady-state error to zero and 2) a lag compensator to reduce the steady-state error by a factor of 10 for a step input without affecting the transient response. Assume that the system is required to operate with a damping ratio of $\zeta = 0.2$.

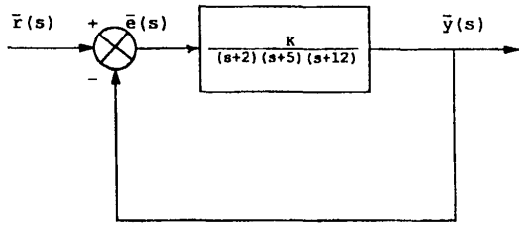


Fig. 5.33 Control system of Example 5.9.

Solution. We have

$$G(s) = \frac{k}{(s+2)(s+5)(s+12)}$$

The first step is to draw the root-locus of the basic (uncompensated) system and determine the value of the gain for operation at $\zeta = 0.2$. Using MATLAB,[4] the root-locus is drawn as shown in Fig. 5.34a. For operation with $\zeta = 0.2$, the values of the gain and closed-loop pole locations are $k = 679.086$ and $p = -16.2416$, $-1.3792 \pm j6.8773$.

The position constant K_p and the steady-state error $e(\infty)$ are given by

$$K_p = \frac{k}{p_1 p_2 p_3} = \frac{679.086}{2 * 5 * 12} = 5.6590$$

$$e(\infty) = \frac{1}{1 + k_p} = \frac{1}{1 + 5.6590} = 0.1502$$

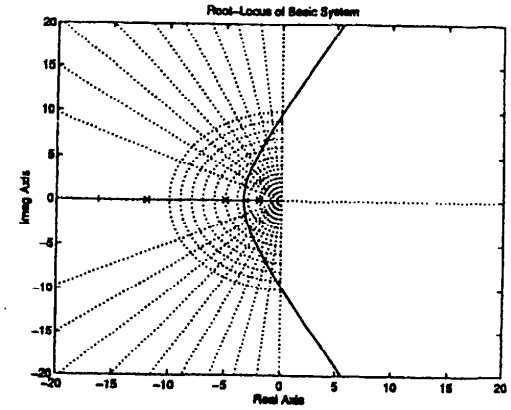
With this, we get the steady-state value of the output (for a unit-step input), $y(\infty) = 1 - e(\infty) = 0.8498$.

Design of a PI compensator. The PI compensator is characterized by a pole at the origin and a zero close to it. Let us choose the zero at $s = -0.05$. With this, the open-loop transfer function of the PI-compensated system is

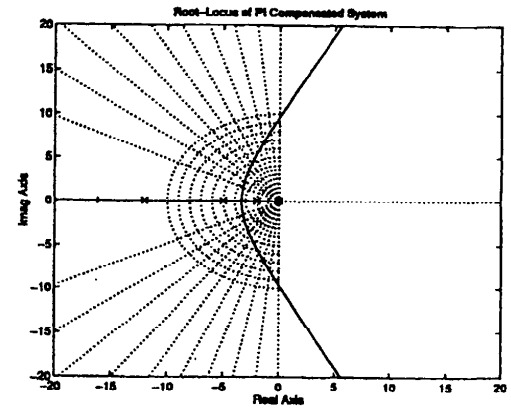
$$G_c(s) = \frac{k(s + 0.05)}{s(s+2)(s+5)(s+12)}$$

Now let us determine the value of the gain k so that the PI-compensated system operates at a damping ratio of 0.2, while the steady-state error is driven to zero. Using MATLAB,[4] we draw the root-locus for the PI-compensated system as shown in Fig. 5.34b. For operating with $\zeta = 0.2$, we obtain $k = 673.175$ and $p = -16.2118$, $-1.3728 \pm j6.8408$, and -0.0426.

Comparing these results with those obtained earlier for the basic system, we observe that the dominant second-order complex poles that determine the transient response are virtually unchanged because the pole at $s = -0.0426$ almost cancels with the zero at $s = -0.05$. The pole at $s = -16.2118$ is so far away on the left-hand side of the s-plane that its influence is negligible. In view of this, the system will essentially behave like a second-order system with dominant poles at $-1.3728 \pm j6.8408$.



a)



b)

Fig. 5.34 Root-locii for the control system of Example 5.9.

Design of the lag compensator. We have to design the lag compensator to achieve a reduction in the steady-state error by a factor of 10, i.e.,

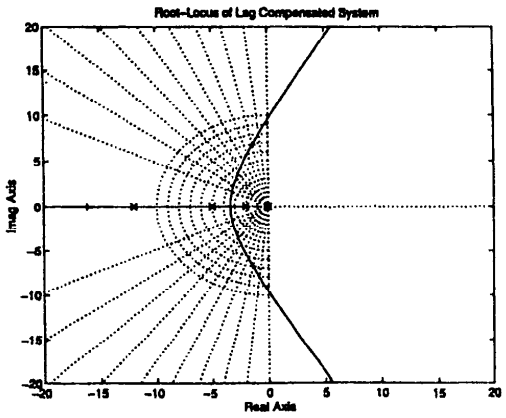
$$e(\infty) = \frac{0.1502}{10}$$
$$= 0.01502$$

Then,

$$K_p = \frac{1 - e(\infty)}{e(\infty)} = \frac{1 - 0.01502}{0.01502} = 64.7895$$

For the lag-compensated system,

$$K_p = \frac{z_c k}{p_c * 2 * 5 * 12}$$



c)

Fig. 5.34 Root-locii for the control system of Example 5.9, continued.

Here, we have three unknowns k, z_c, and p_c and one relation as above. To begin with, let us assume $k = 679.086$ (uncompensated gain). Then, we can choose one of the two remaining unknowns arbitrarily. Let us choose $z_c = 0.05$ so that we get $p_c = 0.0044$. The open-loop transfer function of the lag-compensated system is given by

$$G(s) = \frac{k(s + 0.05)}{(s + 0.0044)(s + 2)(s + 5)(s + 12)}$$

Now we can draw the root-locus as shown in Fig. 5.34c and determine the value of the gain and closed-loop poles for operating with $\zeta = 0.2$. We get $k = 670.3530$ and $p = -16.2019, -1.3796 \pm j6.8304$, and -0.0433.

Thus, the pole locations are similar to those observed for the PI compensator.

With this new value of the gain $k = 670.3530$, the steady-state error is slightly changed. We have

$$\begin{aligned} K_p &= \frac{z_c k}{p_c * 2 * 5 * 12} \\ &= \frac{0.05 * 670.3530}{0.0044 * 2 * 5 * 12} \\ &= 63.4804 \\ e(\infty) &= \frac{1}{1 + K_p} \\ &= \frac{1}{1 + 63.4804} \\ &= 0.0155 \end{aligned}$$

which is close to the target value of 0.01502. Hence, we need not repeat the design process.

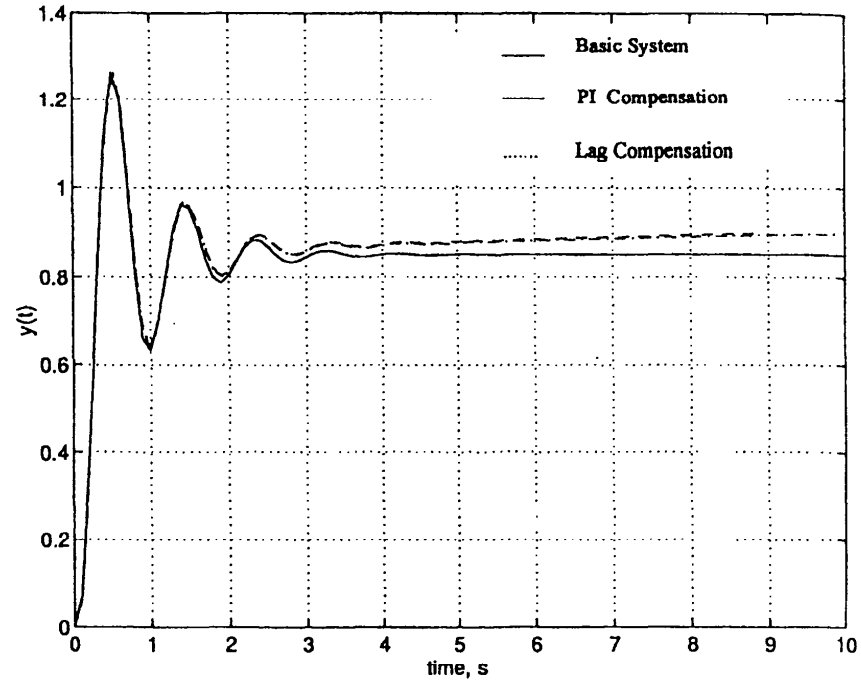


Fig. 5.35 Unit-step responses of Example 5.9.

Now to test the designs of PI- and lag-compensated systems, we have obtained unit-step responses of the basic, PI-, and lag-compensated systems as shown in Fig. 5.35. We observe that the design objectives are met. The transient response of the PI- and lag-compensated systems are almost identical to that of the basic system. Furthermore, as t assumes large values, the steady-state error for the PI compensator approaches zero and that for the lag compensator approaches the target value of 0.01502.

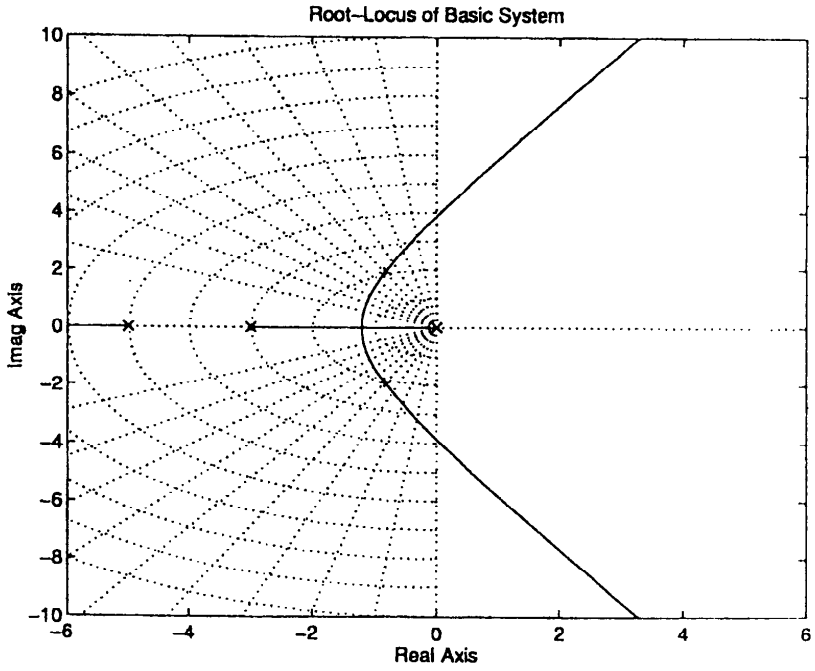
Example 5.10

For the following system, design 1) a PD compensator and 2) a lead-lag compensator so that the peak time is reduced by a factor of 3, while the percent overshoot remains unchanged at 25.38%.

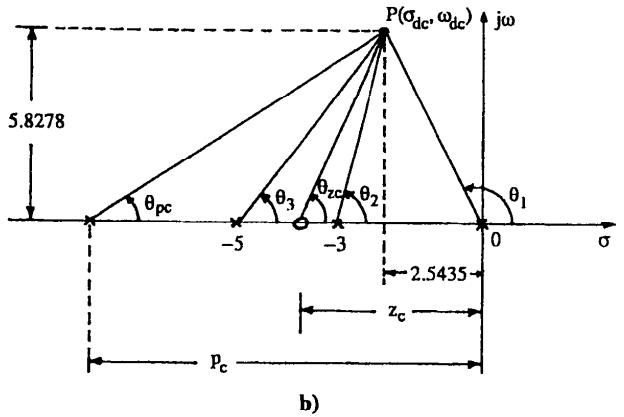
$$G(s) = \frac{k}{s(s + 3)(s + 5)}$$

Solution.

PD compensator. The first step is to draw the root-locus of the basic system as shown in Fig. 5.36a. Using Eq. (5.75), we find that the damping ratio that corresponds to 25.38% overshoot is equal to 0.4. For the basic (uncompensated) system, the value of the gain that corresponds to $\zeta = 0.4$ is equal to 28.3825, and

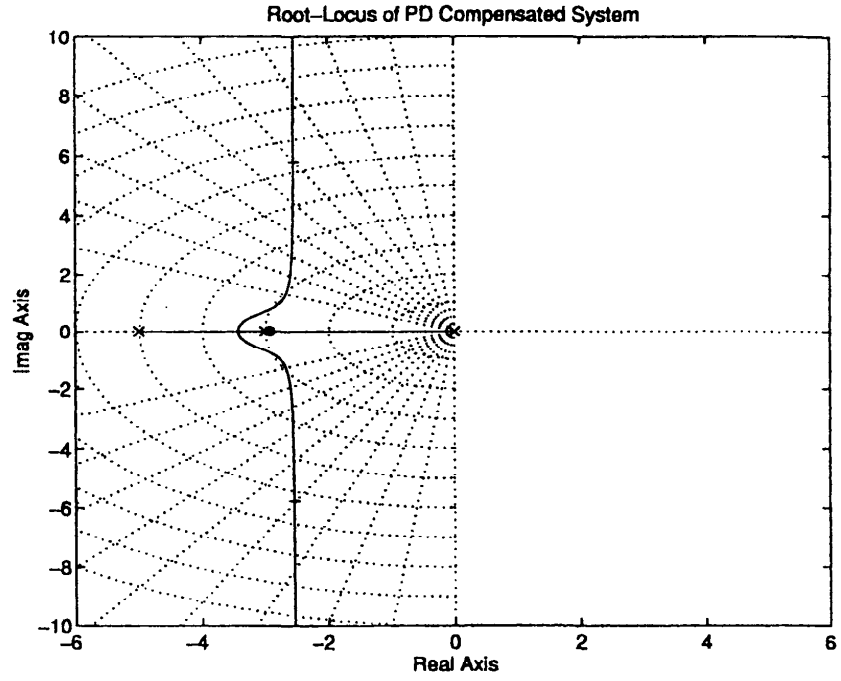


a)



b)

Fig. 5.36 Root-locii for the control system of Example 5.10.



c)

Fig. 5.36 Root-locii for the control system of Example 5.10, continued.

the closed-loop pole locations are $-0.8299 \pm j1.9426$ and -6.3402. Because the third pole located at -6.3402 is farther from the second-order poles, we can use the second-order approximation. With this assumption, the times for peak amplitude for the basic system T_p and that for the compensated system T_{pc} are given by

$$T_p = \frac{\pi}{\omega_d}$$

$$= \frac{\pi}{1.9426} = 1.6172\,\text{s}$$

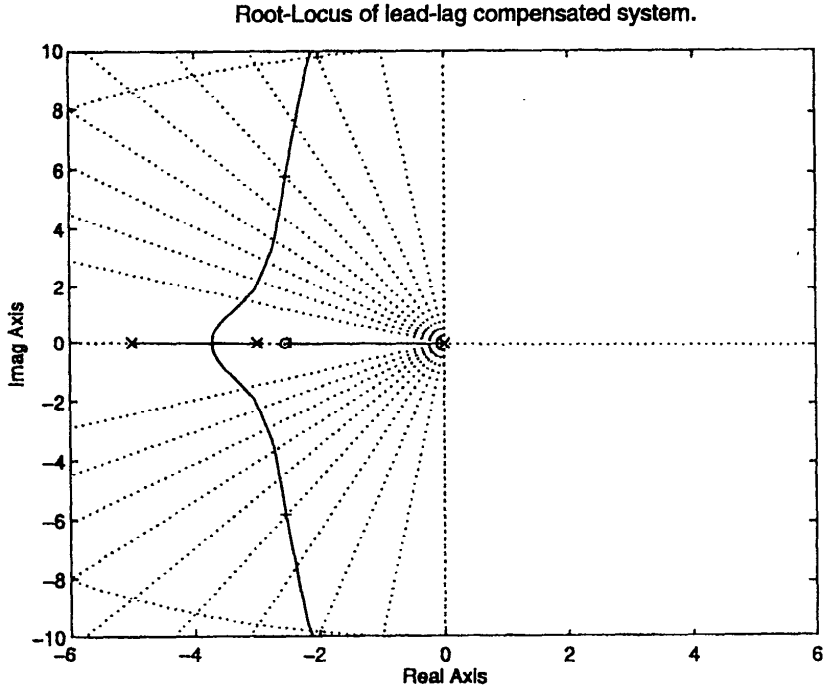
$$T_{pc} = \frac{T_p}{3} = 0.5391\,\text{s}$$

Then,

$$\omega_{dc} = \frac{\pi}{T_{pc}}$$

$$= \frac{\pi}{0.5391}$$

$$= 5.8278$$



d)

Fig. 5.36 Root-locii for the control system of Example 5.10, continued.

$$\omega_{nc} = \frac{\omega_{dc}}{\sqrt{1-\zeta^2}}$$
$$= \frac{5.8278}{\sqrt{1-0.4^2}}$$
$$= 6.3587$$
$$\sigma_{dc} = -\zeta\omega_{nc}$$
$$= -0.4 * 6.3587$$
$$= -2.5435$$

Here, σ_{dc} and ω_{dc} are the real and imaginary parts for the dominant second-order poles.

The transfer function of the PD compensator is given by

$$G_c(s) = s + z_c$$

Now we have to determine the location of the compensating zero z_c so that the root-locus passes through the point $(\sigma_{dc}, \omega_{dc})$. The value of z_c is determined by

the angle condition of Eq. (5.124), which in this case leads to

$$\theta_{zc} - (\theta_1 + \theta_2 + \theta_3) = (2n+1)180$$

Referring to Fig. 5.36b (ignoring the pole at $s = p_c$), we find $\theta_1 = 113.5736$ deg, $\theta_2 = 85.5274$ deg, and $\theta_3 = 67.1488$ deg. Choosing $n = -1$, we obtain $\theta_{zc} = 86.2498$ deg and $z_c = 2.9261$. Then, the transfer function of the PD-compensated system is given by

$$G_c(s) = \frac{k(s+2.9261)}{s(s+3)(s+5)}$$

Then we draw the root-locus using MATLAB[4] as shown in Fig. 5.36c and obtain $k = 40.2971$ and $p = -2.5435 \pm j5.8317$ and -2.9130 for operating at $\zeta = 0.4$.

Lead-lag compensator. The transfer function of the lead compensator is given by Eq. (5.168) as

$$G_c(s) = \frac{s+z_c}{s+p_c}$$

We have to find the locations of the zero z_c and the pole p_c on the real axis so that the design objectives are met.

From the analysis of PD compensator as in 1) above, we know that the net angle contribution due to the zero at $s = -z_c$ and pole at $s = -p_c$ must be equal to 86.2498 deg. Let us assume that the angle contribution due to zero at $s = -z_c$ is $\theta_{zc} = 90$ deg so that $z_c = 2.5435$. Then, the angle contribution due to the pole at $s = -p_c$ is $\theta_p = 3.7502$ deg so that $p_c = 88.9$. Then, the transfer function of the lead-compensated system is given by

$$G_c(s) = \frac{k(s+2.5435)}{s(s+3)(s+5)(s+88.90)}$$

Now we draw the root-locus of the lead-lag compensated system as shown in Fig. 5.36d (the root-locus around the origin is shown in this figure) and select the operating point for $\zeta = 0.4$. We get $k = 3451.4$ and closed-loop poles at -89.3603, $-2.5405 \pm j5.7879$, and -2.4588.

Let us verify the designs by performing the simulation, i.e., we determine the unit-step response using MATLAB.[4] The results are shown in Fig. 5.37. We observe that the peak amplitudes (hence the percent overshoot O_s) for all three cases are nearly equal. For the basic system, $T_p = 1.80$ s and, for PD- and lead-compensated systems, $T_p \simeq 0.6$ s. Thus, the design objectives have been realized.

Example 5.11

Design a PID controller for a unity feedback system with

$$G(s) = \frac{k(s+15)}{(s+1)(s+3)(s+9)}$$

to operate at a peak time, which is 50% of the basic system and has a zero steady-state error for a unit-step input while continuing to operate at a damping ratio of 0.3.

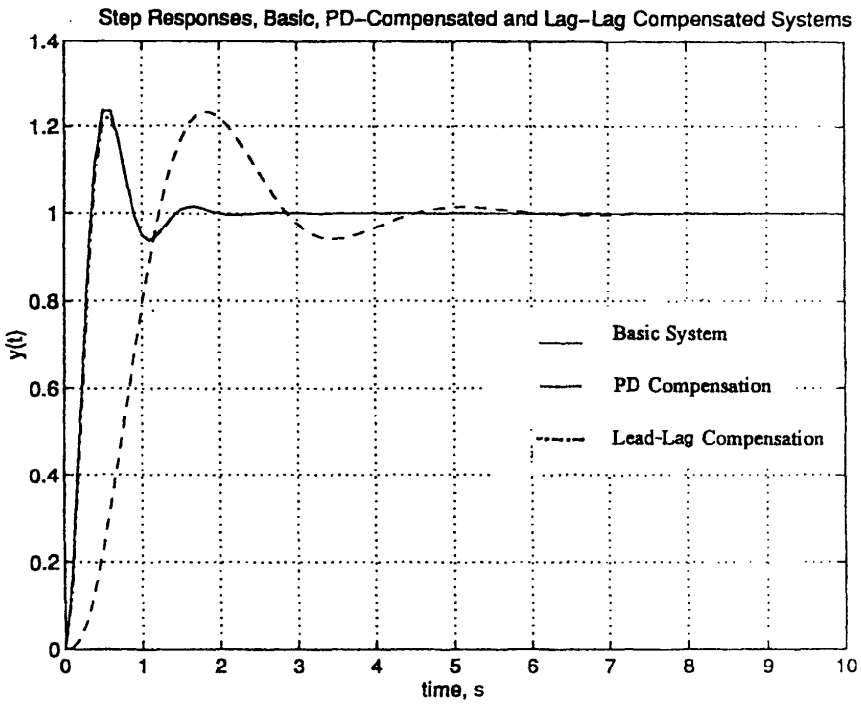


Fig. 5.37 Unit-step responses for Example 5.10.

Solution. The approach we take here is to design the PD controller first and then add a PI controller.

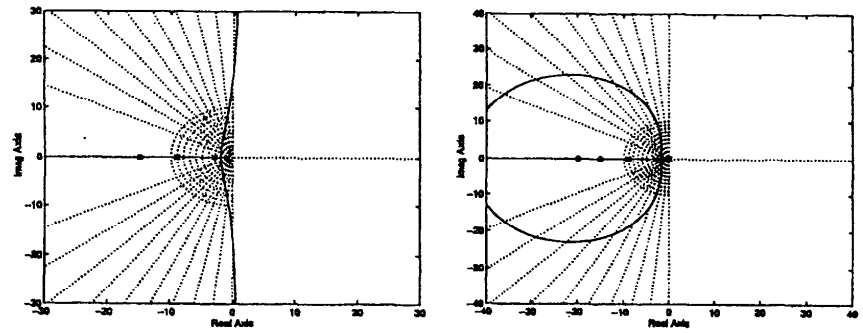
Draw the root-locus for the basic system using MATLAB[4] as shown in Fig. 5.38a and pick the point on the root-locus corresponding to $\zeta = 0.3$. We get $k = 14.2869$ and $p = -10.089, -1.4555 \pm j4.6689$. Using this information, we get $T_p = 0.6729$ s. Then, for the compensated system, $T_{pc} = T_p/2 = 0.3365$ s, which corresponds to $\omega_{dc} = 9.3378$, $\omega_{nc} = \omega_{dc}/\sqrt{1-\zeta^2} = 9.7887$, and $\sigma_{dc} = \zeta\omega_{nc} = 2.9366$.

Now we calculate the angle contributions. Proceeding as before in Example 5.10, we get $\theta_1 = 101.1851$, $\theta_2 = 89.6355$, $\theta_3 = 39.0601$, and $\theta_4 = 58.2290$ so that $\theta_{zc} = 29.9885$ giving $z_c = 19.9014$. With this, the transfer function of the PD-compensated system is given by

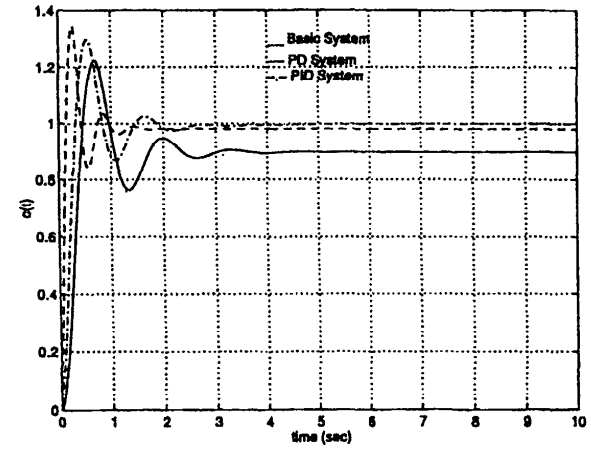
$$G(s) = \frac{k(s+15)(s+19.9014)}{(s+1)(s+3)(s+9)}$$

Next, we add the PI controller. Select a pole at $s = 0$ and a zero at $s = -0.5$ so that the transfer function of the PID controller is given by

$$G(s) = \frac{k(s+0.5)(s+15)(s+19.9014)}{s(s+1)(s+3)(s+9)}$$



a) Root-locus of basic system **b) Root-locus of PID-compensated system**



c) Unit-step responses

Fig. 5.38 PID controller for Example 5.11.

Now we draw the root-locus of the PID system as shown in Fig. 5.38b and pick the point corresponding to $\zeta = 0.3$. We get $k = 6.1112$ and closed-loop poles at $-3.6846 \pm j12.2343$, -11.2460, and -0.4969.

The unit-step responses of the basic, PD-, and PID-compensated systems are shown in Fig. 5.38c. It may be observed that the PID-compensated system meets the design requirements.

Example 5.12

For the system of Example 5.10, design a major-loop feedback to achieve the same performance.

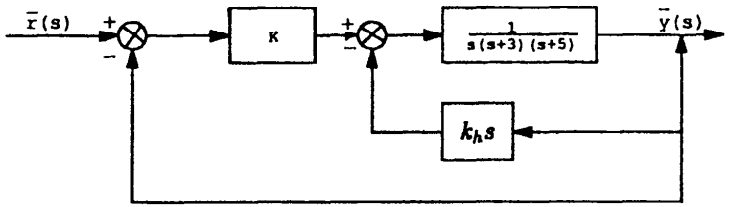
Solution. We have found in Example 5.10 that $\theta_{zc} = 86.2498$. With this, we obtain the equivalent pole location, $z_c = 2.9261$, and $k_h = 1/z_c = 0.3418$. We then

plot the root-locus using MATLAB and obtain the value of the gain as 40.2971, which is equal to kk_h so that $k = 117.8967$. The reader may verify that this response is identical to that of the PD controller of Example 5.10.

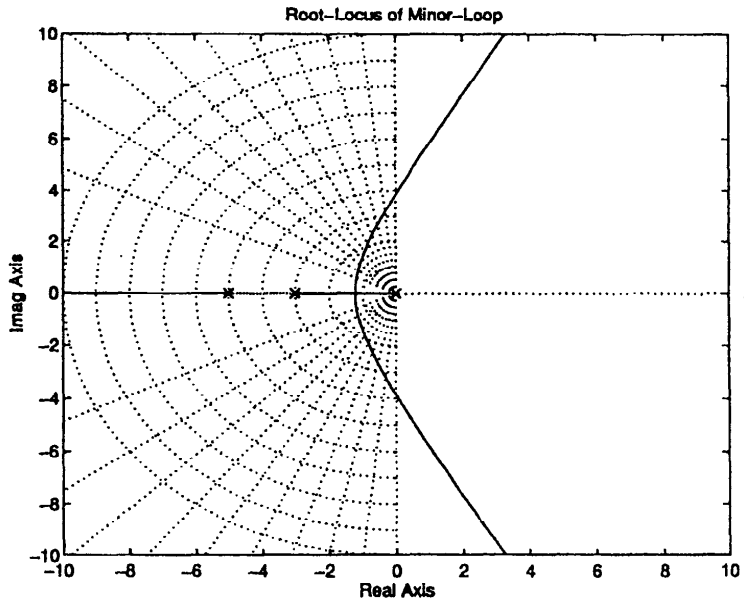
Example 5.13

For the control system shown in Fig. 5.39a, determine the gain k_h so that the minor loop operates with a damping ratio of 0.707 and the complete system has a damping ratio of 0.4.

Solution. Consider the minor loop. We draw the root-locus using MATLAB[4] as shown in Fig. 5.39b and pick the point on the root-locus corresponding to $\zeta = 0.707$. Then, we get $k_h = 14.1018$ and $p = 5.8471, -1.0765 \pm j1.1194$. Having designed the minor loop and knowing the value of k_h, we can simplify the system

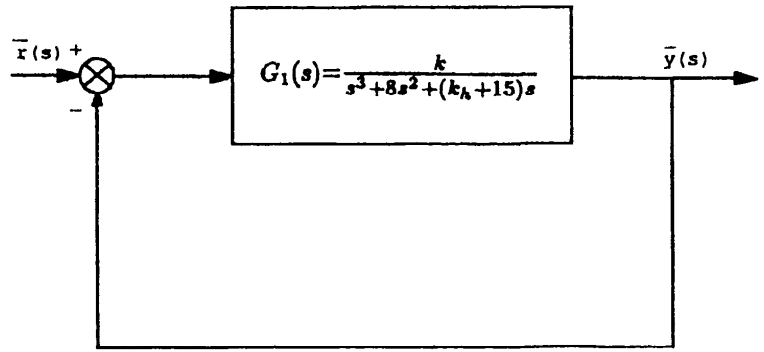


a) Control system

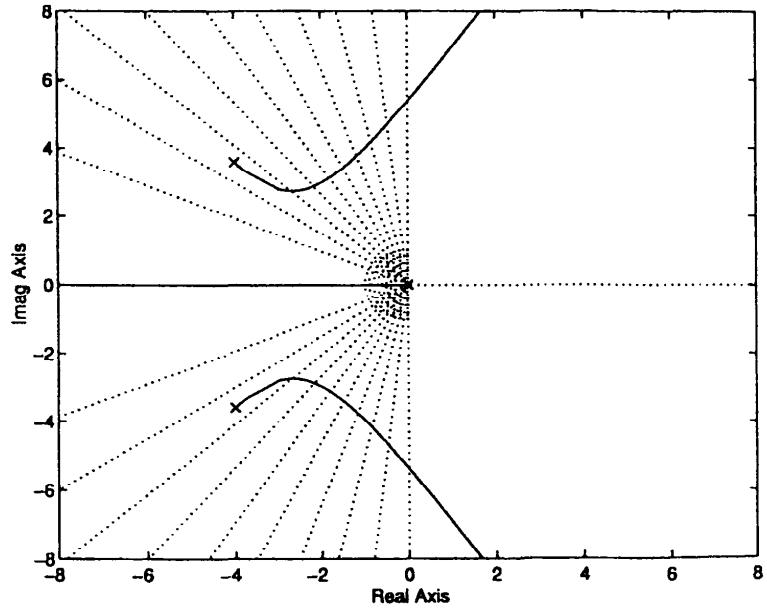


b)

Fig. 5.39 Minor-loop design for Example 5.13.



c) Outer-loop block diagram



d) Root-locus of outer loop

Fig. 5.39 Minor-loop design for Example 5.13, continued.

block diagram as shown in Fig. 5.39c. We have

$$G_1(s) = \frac{k}{s^3 + 8s^2 + (k_h + 15)s}$$

For this system, we draw the root-locus as shown in Fig. 5.39d and obtain $k = 69.5452$ and $p = -4.9495$, $-1.5202 \pm j3.4222$ for operating at $\zeta = 0.4$. This completes the design.

5.10 State-Space Analysis and Design

The classical method of analyses discussed in the previous sections is called the frequency-domain technique because it is based on system representation in the form of transfer function. The main advantage of this approach is that the governing differential equation of the system is replaced by an algebraic transfer function. However, a disadvantage of this classical method is that it is limited to linear time-invariant systems with zero initial conditions. The modern state-space approach is more general in nature because it can be used to represent nonlinear, time-varying systems with nonzero initial conditions. The state-space method can also handle multi-input-multi-output systems in a compact manner. Furthermore, the state-space approach becomes very attractive because it is based on matrix algebra, and powerful matrix analyses tools like MATLAB[4] are commercially available.

5.10.1 Concept of State Variable

The choice of a set of variables to be designated as state variables for a given system is somewhat arbitrary. In other words, there is no unique method of defining what should be a set of state variables for a given system. However, the state variables have to meet some requirements, which can be stated as follows.

1) The variables selected as state variables must be linearly independent, i.e., it should not be possible to express any one or more of the state variables in terms of the remaining state variables. Mathematically, if x is an n dimensional vector with components x_i, $i = 1, n$, then the components x_i are said to be linearly independent if $\alpha_i x_i \neq 0$, for all $\alpha_i \neq 0$ and all $x_i \neq 0$. Here, α_i, Λ, $\dot{e} = 1$, and n are arbitrary constants.

2) Given all the initial conditions, the input for $t \geq 0$, and the solution of the governing differential equation in terms of the selected state variables, one must be able to describe uniquely any physical parameter (state and output) of the system for all $t \geq 0$. In other words, if any of the physical parameters of the system cannot be described in this manner, then the selected variables do not qualify to be designated as state variables.

5.10.2 State-Space Representation

A state vector is a vector whose elements are the state variables satisfying the above requirements. If n is the dimension of a state vector, then the state-space is an n dimensional space whose axes are the state variables. For example, if x_1, x_2, and x_3 are the elements of the state vector x, then $n = 3$, and the state-space is a three-dimensional space with x_1, x_2, and x_3 as three axes.

The state equation is a set of n simultaneous first-order differential equations involving n state variables and m inputs. Usually, $m < n$. The output equation is a set of algebraic equations that relate the outputs of the system to the state variables.

For example,

$$\dot{x} = Ax + Bu \tag{5.171}$$

$$y = Cx + Du \tag{5.172}$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{5.173}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdot & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdot & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdot & a_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & a_{n3} & \cdot & a_{nn} \end{bmatrix} \tag{5.174}$$

$$B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \tag{5.175}$$

is a state-space representation of an nth order system. Here, the order of the system is equal to the number of simultaneous first-order differential equations. In Eqs. (5.171) and (5.172), x is the state vector of dimension n, A is the system matrix of dimension $n \times n$, B is an $n \times m$ input coupling matrix, u is an $m \times 1$ input vector, y is the output vector of dimension q, C is a $q \times n$ output matrix, and D is a $q \times m$ feed forward matrix. The term feed forward is used when a part of the input directly appears at the output. A schematic diagram of the state-space representation is shown in Fig. 5.40.

5.10.3 State Transition Matrix

Consider the homogeneous part ($u = 0$) of state Eqs. (5.171) and (5.172) as given by

$$\dot{x} = Ax \tag{5.176}$$

$$y = Cx \tag{5.177}$$

The state transition matrix $\Phi(t)$ is defined as a matrix that satisfies the equation

$$x(t) = \Phi(t)x(0) \tag{5.178}$$

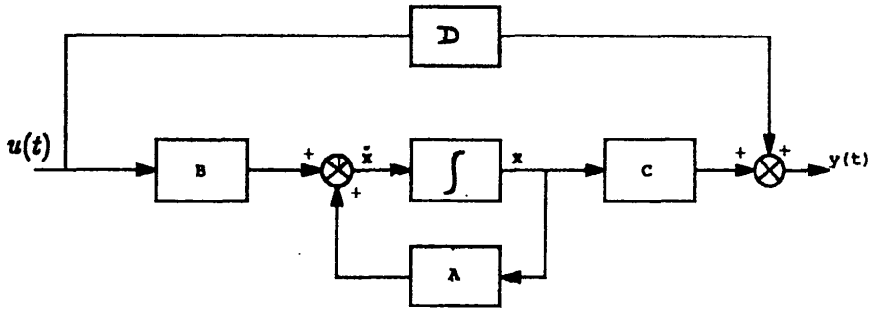


Fig. 5.40 Schematic diagram of state-space representation.

In other words, given the initial conditions $x(0)$, the state transition matrix enables us to predict the state vector at $t \geq 0$.

Substituting for $x(t)$ from Eq. (5.178) into state Eq. (5.176), we obtain

$$\dot{\Phi}(t)x(0) = A\Phi(t)x(0) \tag{5.179}$$

$$\left[\dot{\Phi}(t) - A\Phi(t)\right]x(0) = 0 \tag{5.180}$$

If this identity is to hold for all arbitrary values of $x(0)$, we must have

$$\dot{\Phi}(t) - A\Phi(t) = 0 \tag{5.181}$$

This shows that the state transition matrix $\Phi(t)$ is a solution to the homogeneous state Eq. (5.176).

Determination of state transition matrix. Take the Laplace transformation of Eq. (5.176),

$$s\bar{x}(s) - x(0) = A\bar{x}(s) \tag{5.182}$$

$$\bar{x}(s) = (sI - A)^{-1}x(0) \tag{5.183}$$

Here, we assume that $(sI - A)^{-1}$ exists, i.e., $(sI - A)$ is nonsingular. Then,

$$x(t) = L^{-1}\left[(sI - A)^{-1}\right]x(0) \tag{5.184}$$

for $t \geq 0$. Comparing Eqs. (5.184) and (5.178), we get

$$\Phi(t) = L^{-1}\left[(sI - A)^{-1}\right] \tag{5.185}$$

Let

$$x(t) = e^{At}x(0) \tag{5.186}$$

The matrix exponential is given by

$$e^{At} = I + At + \frac{A^2t^2}{2!} + \cdots + \frac{A^nt^n}{n!} + \cdots \tag{5.187}$$

where I is the identity matrix. We note that Eq. (5.186) satisfies the homogeneous state Eq. (5.176). Hence,

$$\Phi(t) = e^{At} = I + At + \frac{A^2t^2}{2!} + \cdots + \frac{A^nt^n}{n!} + \cdots \tag{5.188}$$

Using this, the solution of the complete nonhomogeneous state Eq. (5.171) can be expressed as[1,3]

$$x(t) = \Phi(t)x(0) + \int_0^t \Phi(t-\tau)Bu(\tau)\,d\tau \tag{5.189}$$

and the output

$$y(t) = C\left[\Phi(t)x(0) + \int_0^t \Phi(t-\tau)Bu(\tau)\,d\tau\right] + Du \tag{5.190}$$

The integral in Eq. (5.189) is the convolution integral, which was introduced earlier in Eq. (5.39). The first term on the right-hand side of Eq. (5.189) represents the solution to the homogeneous part of the state equation and gives the free (transient) response. The second term represents the forced response and is independent of the initial conditions $x(0)$.

Properties of state transition matrix. The state transition matrix $\Phi(t)$ has the following properties. The proof of these identities is left as an exercise to the reader.

$$\Phi(0) = I \tag{5.191}$$

$$\Phi^{-1}(t) = \Phi(-t) \tag{5.192}$$

$$\Phi(t_2 - t_1)\Phi(t_1 - t_0) = \Phi(t_2 - t_0) \tag{5.193}$$

$$[\Phi(t)]^k = \Phi(kt) \tag{5.194}$$

Characteristic equation. Given a square matrix A, the equation

$$\Delta(\lambda) = |\lambda I - A| = 0 \tag{5.195}$$

is called the characteristic equation of matrix A. Here, $|.|$ denotes the determinant of the argument (square) matrix.

Eigenvalues and eigenvectors. The roots of characteristic Eq. (5.195) are called the eigenvalues of the matrix A and are usually denoted by λ_i, $i = 1, \ldots, n$, where n is the number of rows or columns of the matrix A. As an example, let

$$A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \tag{5.196}$$

so that

$$\lambda I - A = \begin{bmatrix} \lambda - 1 & -1 \\ -4 & \lambda - 1 \end{bmatrix} \tag{5.197}$$

and

$$\Delta(\lambda) = |\lambda I - A| = \lambda^2 - 2\lambda - 3 = 0 \tag{5.198}$$

Solving, we get $\lambda_1 = 3$ and $\lambda_2 = -1$.

An important property of the eigenvalues is that they remain invariant under any linear transformation. A direct consequence of this property is that the closed-loop characteristic equation also remains invariant under any linear transformation. To make this point clear, suppose we are given a linear, time-invariant system $\dot{x} = Ax + Bu$ and we transform this system using a linear transformation $x = Pz$ so that the transformed system is $\dot{z} = P^{-1}APz + P^{-1}Bu$. Then, the eigenvalues of matrix A and those of $P^{-1}AP$ are identical. Interested readers may verify this statement by working out the details.

If λ_i is an eigenvalue of the square matrix A, then any vector x that satisfies the equation

$$Ax = \lambda_i x \tag{5.199}$$

is called the eigenvector corresponding to the eigenvalue λ_i. In other words, every eigenvalue will have an associated eigenvector. Returning to the above example,

$$\begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (-1) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{5.200}$$

or

$$2x_1 + x_2 = 0 \tag{5.201}$$

$$4x_1 + 2x_2 = 0 \tag{5.202}$$

Note that the two equations are identical, stating that the eigenvector is not unique and depends on our choice of one of the two variables. Let $x_2 = 1$ so that $x_1 = -1/2$. The eigenvector corresponding to $\lambda = -1$ is $[-1/2\ 1]^T$. Here, the superscript "T" denotes the matrix transpose. It can be shown that the eigenvector obtained by choosing any other value for x_2 would be a scalar multiple of this eigenvector. Similarly, the eigenvector corresponding to $\lambda_2 = 3$ is $[1\ 2]^T$.

To understand the physical meaning of eigenvalues and eigenvectors, consider a system with two first-order, coupled linear differential equations

$$\dot{x}_1 = x_1 + x_2 \tag{5.203}$$

$$\dot{x}_2 = 4x_1 + x_2 \tag{5.204}$$

Assume that the solution is given by $x_1(t) = u_1 e^{\lambda t}$ and $x_2(t) = u_2 e^{\lambda t}$. Substituting, we get

$$\lambda u_1 e^{\lambda t} = u_1 e^{\lambda t} + u_2 e^{\lambda t} \tag{5.205}$$

$$\lambda u_2 e^{\lambda t} = 4u_1 e^{\lambda t} + u_2 e^{\lambda t} \tag{5.206}$$

Because $e^{\lambda t} \geq 0$ for all $t \geq 0$, we can write

$$\lambda u_1 = u_1 + u_2 \tag{5.207}$$

$$\lambda u_2 = 4u_1 + u_2 \tag{5.208}$$

or

$$\begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \lambda \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{5.209}$$

which is of the form $Au = \lambda u$. This equation is of the same form as Eq. (5.199) with λ as the eigenvalue of the matrix A. Furthermore, we recognize that the above matrix A is the same as matrix A in Eq. (5.196), which has the eigenvalues of -1 and 3 and eigenvectors of $[-1/2\ \ 1]^T$ and $[1\ \ 2]^T$. This means, with $\lambda = -1$, $u_1 = -1/2$, and $u_2 = 1$, we get the first solution as

$$x_1 = -\tfrac{1}{2}e^{-t} \qquad x_2 = e^{-t} \tag{5.210}$$

and, with $\lambda = 3$, $u_1 = 1$, and $u_2 = 2$, we get the second solution as

$$x_1 = e^{3t} \qquad x_2 = 2e^{3t} \tag{5.211}$$

Therefore, the general solution is given by

$$x_1 = -\tfrac{1}{2}e^{-t} + e^{3t} \qquad x_2 = e^{-t} + 2e^{3t} \tag{5.212}$$

Thus, we observe that the eigenvalues determine the nature of the transient response and the eigenvectors determine the amplitude of this response.

5.10.4 Controllability and Observability

The concept of controllability is linked to the question of whether the input u affects or controls the variation of each one of the state variables x_i. If it does, then we say that the given system is controllable. On the other hand, if any of the state variables are not influenced by the input u, then the system is said to be uncontrollable. Alternatively, if we can take the system from a given initial state $x(0)$ to a specified final state $x(t_f)$ using the available control u, then the system is said to be controllable. If not, the system is uncontrollable.

The given nth order linear, time-invariant system

$$\dot{x} = Ax + Bu \tag{5.213}$$

$$y = Cx \tag{5.214}$$

is said to be controllable if the matrix

$$Q_c = [B \quad AB \quad A^2B \quad \cdots \quad A^{n-1}B] \tag{5.215}$$

is nonsingular or has full rank n. The matrix Q_c is called the controllability matrix.[1-3]

The concept of observability is related to the question whether each one of the state variables affects or controls the variation of the output y. If the answer to this question is yes, then the system is said to be observable. If not, the system is

unobservable. Thus, for an unobservable system, the input does not affect some or all of the output variables.

A given linear, time-invariant system in the state-space form is said to be observable if the matrix

$$Q_0 = \begin{bmatrix} C \\ CA \\ CA^2 \\ .. \\ .. \\ CA^{n-1} \end{bmatrix} \tag{5.216}$$

is nonsingular or has the full rank n. The matrix Q_0 is called the observability matrix.[1-3]

5.10.5 Phase-Variable Form

Let us suppose that we have a state-space representation in the form

$$\dot{x} = Ax + Bu \tag{5.217}$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ x_n \end{bmatrix} \tag{5.218}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & . & 0 \\ 0 & 0 & 1 & 0 & . & 0 \\ 0 & 0 & 0 & 1 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ -a_0 & -a_1 & -a_2 & . & . & -a_{n-1} \end{bmatrix} \tag{5.219}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ . \\ . \\ 1 \end{bmatrix} \tag{5.220}$$

Equation (5.217) with matrices A and B given by Eqs. (5.219) and (5.220) is called the phase-variable form of state Eq. (5.171). The advantage of the phase-variable

form is that there is a minimum amount of coupling between the state variables. For example, given all the initial conditions, $x_1(0), x_2(0), \ldots, x_n(0)$, we can first solve $\dot{x}_1(0) = x_2(0)$ and obtain $x_1(\Delta t)$ by an integration over a small time step Δt. Similarly, we can get $x_2(\Delta t), x_3(\Delta t), \ldots, x_n(\Delta t)$ by successively solving the other equations.

The phase-variable form of representation has another advantage that the elements of the last row constitute the coefficients of the characteristic equation as follows:

$$\Delta(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = 0 \tag{5.221}$$

This property is useful in the design of compensators using the pole-placement method, which we will discuss a little later.

5.10.6 Conversion of Differential Equations to Phase-Variable Form

Let the given dynamical system be represented by the following linear differential equation:

$$\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_1\frac{dy}{dt} + a_0 y = u(t) \tag{5.222}$$

Let us select a set of state variables such that each subsequent state variable is defined as the derivative of the previous state variable. That is,

$$x_1 = y \qquad x_2 = \dot{y} = \dot{x}_1, \ldots, x_n = \frac{d^{n-1}y}{dt^{n-1}} = \dot{x}_{n-1} \tag{5.223}$$

Then,

$$\dot{x}_1 = x_2 \tag{5.224}$$

$$\dot{x}_2 = x_3 \tag{5.225}$$

. . .

. . .

. . .

$$\dot{x}_n = -a_0 x_1 - a_1 x_2 - a_2 x_3 - \cdots - a_{n-1}x_n + u(t) \tag{5.226}$$

Or, in matrix form,

$$\dot{x} = Ax + Bu \tag{5.227}$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{5.228}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -a_0 & -a_1 & -a_2 & \cdot & \cdot & -a_{n-1} \end{bmatrix} \tag{5.229}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \tag{5.230}$$

The output is given by

$$y = Cx = [1 \quad 0 \quad \cdots \quad 0] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{5.231}$$

Thus, by choosing each successive state variable to be the derivative of the previous one, we are able to express the given differential equation in the state-space, phase-variable form.

5.10.7 Conversion of General State-Space Representation to Phase-Variable Form

A system given in a general state-space form can be expressed in the phase-variable form if the system is controllable. Let

$$\dot{x} = Ax + Bu \tag{5.232}$$

$$y = Cx \tag{5.233}$$

be the given plant, which is not in phase-variable form. We assume that this system is controllable. Furthermore, let us assume that there exists a matrix P, which is defined as

$$z = Px \tag{5.234}$$

which transforms the given system into phase-variable form,

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \cdot \\ \cdot \\ \dot{z}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ -a_0 & -a_1 & \cdot & \cdot & \cdot & -a_{n-1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \cdot \\ \cdot \\ z_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix} u \tag{5.235}$$

The transformation matrix P has the form

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdot & \cdot & \cdot & p_{1n} \\ p_{21} & p_{22} & \cdot & \cdot & \cdot & p_{2n} \\ p_{31} & p_{32} & \cdot & \cdot & \cdot & p_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{n1} & p_{n2} & \cdot & \cdot & \cdot & p_{nn} \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ \cdot \\ \cdot \\ P_n \end{bmatrix} \tag{5.236}$$

We have

$$z_1 = P_1 x \tag{5.237}$$

so that

$$\dot{z}_1 = P_1 Ax + P_1 Bu \tag{5.238}$$

Because this transformed equation is supposed to be in phase-variable form, we must have $\dot{z}_1 = z_2$. This gives $z_2 = P_1 Ax$ and $P_1 B = 0$. From Eq. (5.234), we have $z_2 = P_2 x$. Therefore, $P_2 = P_1 A$. Continuing this further, we find that $\dot{z}_2 = P_1 A^2 x$, $P_1 AB = 0$, $P_3 = P_1 A^2, \ldots, \dot{z}_{n-1} = P_1 A^{n-1} x$, $P_1 A^{n-2} B = 0$, and $P_{n-1} = P_1 A^{n-2}$. Finally, we have

$$\dot{z}_n = P_1 A^n x + P_1 A^{n-1} Bu \tag{5.239}$$

Comparing this with Eq. (5.235), we find that $P_1 A^{n-1} B = 1$. With this, we can construct the following matrix:

$$\begin{bmatrix} P_1 B & P_1 AB & P_1 A^2 B & \cdot & P_1 A^{n-1} B \end{bmatrix} = [0 \quad 0 \quad 0 \quad \cdot \quad 1] \tag{5.240}$$

or

$$P_1 = [0 \quad 0 \quad 0 \quad \cdot \quad 1][B \quad AB \quad A^2 B \quad \cdot \quad A^{n-1} B]^{-1} \tag{5.241}$$

$$= [0 \quad 0 \quad 0 \quad \cdot \quad 1] Q_c^{-1} \tag{5.242}$$

Because we have assumed that the given system is controllable, the controllability matrix Q_c is nonsingular and Q_c^{-1} exists. Once P_1 is known, then $P_2, P_3, \ldots, P_n$ can be calculated using the relations derived above. Then, the phase-variable form of the given system is

$$\dot{z} = (PAP^{-1})z + (PB)u \tag{5.243}$$

5.10.8 Conversion of Transfer Function Form to Phase-Variable Form

Suppose the relation between the input and output of a system is given in the form of a transfer function; we can convert this to state-space phase-variable representation using a number of different approaches. Here, we will discuss a method based on decomposition of the transfer function. To illustrate the method, consider the system shown in Fig. 5.41a.

Let the open-loop transfer function of a system be given by

$$G(s) = \frac{k(s^2 + a_1 s + a_2)}{s^3 + b_1 s^2 + b_2 s + b_3} \tag{5.244}$$

The first step is to decompose the given system into two blocks, one for the denominator with transfer function $G_1(s)$ and the other for the numerator with transfer function $G_2(s)$ as shown in Fig. 5.41b. Let the output of the first block be denoted as $\bar{x}_1(s)$. Then, for the first block,

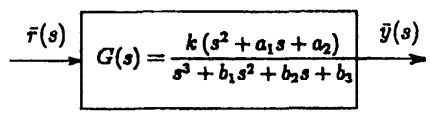
$$G_1(s) = \frac{\bar{x}_1(s)}{\bar{r}(s)} = \frac{k}{s^3 + b_1 s^2 + b_2 s + b_3} \tag{5.245}$$

so that

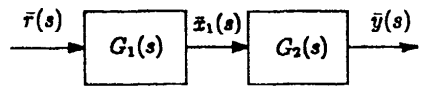
$$\bar{x}_1(s)\left(s^3 + b_1 s^2 + b_2 s + b_3\right) = k\bar{r}(s) \tag{5.246}$$

Taking the inverse Laplace transforms,

$$\frac{d^3 x_1}{dt^3} + b_1 \frac{d^2 x_1}{dt^2} + b_2 \frac{dx_1}{dt} + b_3 x_1 = kr(t) \tag{5.247}$$



a)



b)

Fig. 5.41 Decomposition of a given control system in phase-variable form.

Let $x_2 = \dot{x}_1$ and $x_3 = \dot{x}_2 = \ddot{x}_1$ so that

$$\dot{x}_1 = x_2 \tag{5.248}$$

$$\dot{x}_2 = x_3 \tag{5.249}$$

$$\dot{x}_3 = -b_1 x_3 - b_2 x_2 - b_3 x_1 + kr(t) \tag{5.250}$$

In matrix form,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -b_3 & -b_2 & -b_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} kr(t) \tag{5.251}$$

which is the required phase-variable representation.

Consider the second block. This block gives the output matrix. We have

$$\bar{y}(s) = (s^2 + a_1 s + a_2)\bar{x}_1(s) \tag{5.252}$$

Taking the inverse Laplace transform, we get

$$y(t) = \ddot{x}_1 + a_1 \dot{x}_1 + 2x_1 \tag{5.253}$$

$$= x_3 + a_1 x_2 + a_2 x_1 \tag{5.254}$$

$$= [a_2 \quad a_1 \quad 1] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{5.255}$$

5.10.9 Pole-Placement Method

The root-locus method discussed earlier is essentially a pole-placement method in frequency-domain analyses. The term pole refers to the poles of the closed-loop transfer function. When we consider higher order systems greater than two, the classical PD or PI type of controllers will not be able to place all the poles as desired because there are only two free variables at our disposal in PD or PI controllers. Therefore, for higher order systems, the pole-placement method of the state-space approach becomes very attractive because it can place all the closed-loop poles arbitrarily, but subject to the conditions that all the states are available for feedback and the given plant satisfies the controllability condition.

Consider the system represented in state-space form as given by

$$\dot{x} = Ax + Bu \tag{5.256}$$

$$y = Cx \tag{5.257}$$

With full-state feedback, $u = r(t) - Kx$, where $r(t)$ is the $m \times 1$ input vector and K is an $m \times n$ matrix of feedback gains. Then,

$$\dot{x} = (A - BK)x + Br(t) \tag{5.258}$$

$$y = Cx \tag{5.259}$$

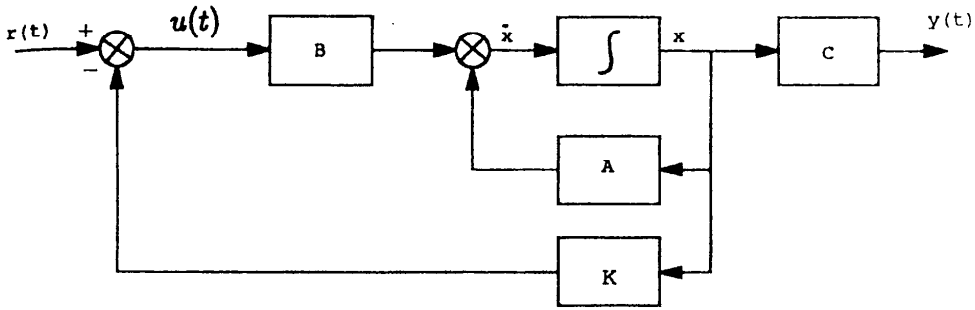


Fig. 5.42 Schematic diagram of pole-placement method.

The block diagram implementation of the full-state feedback control system is shown in Fig. 5.42.

For simplicity, consider a system with single input, which means K is of dimension $1 \times n$. Then, the design procedure is as follows.

1) Represent the given plant in phase-variable form.

$$\dot{z} = A_p z + B_p u \tag{5.260}$$

2) Feed back each state variable to the input of the plant with gains k_i so that

$$A_p - B_p K = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -(a_0 + k_1) & -(a_1 + k_2) & \cdot & \cdot & \cdot & -(a_{n-1} + k_n) \end{bmatrix} \tag{5.261}$$

3) Write down the characteristic equation for the plant as follows:

$$|sI - (A_p - B_p K)| = s^n + (a_{n-1} + k_n)s^{n-1} + \cdots + (a_1 + k_2)s + (a_0 + k_1) = 0 \tag{5.262}$$

4) Decide on all the closed-loop pole locations that give the desired system response, i.e., the desired characteristic equation is given by

$$s^n + d_{n-1}s^{n-1} + \cdots + d_1 s + d_0 = 0 \tag{5.263}$$

5) Equate coefficients of the two characteristic equations

$$d_{n-1} = a_{n-1} + k_n, \ldots, d_1 = a_1 + k_2, d_0 = a_0 + k_1 \tag{5.264}$$

so that

$$k_1 = d_0 - a_0, k_2 = d_1 - a_1, \ldots, k_n = d_{n-1} - a_{n-1} \tag{5.265}$$

and

$$K = [k_1 \quad k_2 \quad \cdot \quad \cdot \quad k_n] \tag{5.266}$$

The full-state feedback law in the transformed z-space is given by

$$u = -Kz + r(t) \tag{5.267}$$

or, in the original state-space,

$$u = -KPx + r(t) \tag{5.268}$$

so that the given system with full-state feedback is given by

$$\dot{x} = Ax + Bu \tag{5.269}$$

$$= (A - KPx)x + Br(t) \tag{5.270}$$

6) Perform a simulation to verify the design.

The advantages of expressing the given plant in the phase-variable form is that equations for the gains k_i are uncoupled and k_i can be easily obtained as given in Eq. (5.265). However, if the plant is not controllable, then it is not possible to represent it in the phase-variable form. For such a case, the above design procedure remains same except for the fact the equations for k_i will be coupled. Then the gains k_i have to be obtained by solving the n coupled algebraic equations.

5.10.10 Dual Phase-Variable Form

The state-space representation, which is in the form,

$$\dot{x} = Ax + BU \tag{5.271}$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad A = \begin{bmatrix} -a_{n-1} & 1 & 0 & 0 & \cdot & 0 \\ -a_{n-2} & 0 & 1 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -a_2 & 0 & 0 & \cdot & 1 & 0 \\ -a_1 & \cdot & \cdot & \cdot & \cdot & 1 \\ -a_0 & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \tag{5.272}$$

is said to be in dual phase-variable form. Similar to the phase-variable form, the elements of the first column of the matrix A in dual phase-variable form constitute the coefficients of the characteristic equation as follows:

$$s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots + a_1 s + a_0 = 0 \tag{5.273}$$

Furthermore, this form of representation of the system matrix A is very useful in the design of state observers, which we will be discussing a little later.

5.10.11 Conversion of Transfer Function Form to Dual Phase-Variable Form

We will illustrate this method with the help of an example. Consider once again the system (see Fig. 5.41) given by

$$G(s) = \frac{k(s^2 + a_1 s + a_2)}{s^3 + b_1 s^2 + b_2 s + b_3} \tag{5.274}$$

Rewrite this in the following form:

$$G(s) = \frac{\frac{k}{s} + \frac{a_1 k}{s^2} + \frac{a_2 k}{s^3}}{1 + \frac{b_1}{s} + \frac{b_2}{s^2} + \frac{b_3}{s^3}} \tag{5.275}$$

$$= \frac{\bar{y}(s)}{\bar{r}(s)} \tag{5.276}$$

or

$$\bar{y}(s)\left(1 + \frac{b_1}{s} + \frac{b_2}{s^2} + \frac{b_3}{s^3}\right) = \bar{r}(s)\left(\frac{k}{s} + \frac{a_1 k}{s^2} + \frac{a_2 k}{s^3}\right) \tag{5.277}$$

Then,

$$\bar{y}(s) = \frac{1}{s}\left[-b_1\bar{y}(s) + k\bar{r}(s) + \frac{1}{s}\left([\bar{r}(s)ka_1 - b_2\bar{y}(s)] + \frac{1}{s}[ka_2\bar{r}(s) - b_3\bar{y}(s)]\right)\right] \tag{5.278}$$

Let

$$s\bar{x}_3(s) = ka_2\bar{r}(s) - b_3\bar{y}(s) \tag{5.279}$$

$$s\bar{x}_2(s) = ka_1\bar{r}(s) - b_2\bar{y}(s) + \bar{x}_3(s) \tag{5.280}$$

$$s\bar{x}_1(s) = -b_1\bar{y}(s) + \bar{x}_2(s) + k\bar{r}(s) \tag{5.281}$$

so that

$$\bar{y}(s) = \bar{x}_1(s) \tag{5.282}$$

Taking the inverse Laplace transforms, we obtain the desired dual phase-variable form as follows:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}-b_1 & 1 & 0\\ -b_2 & 0 & 1\\ -b_3 & 0 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}1\\ a_1\\ a_2\end{bmatrix} kr(t) \tag{5.283}$$

and

$$y(t) = x_1(t) \tag{5.284}$$

5.10.12 Observer Design

The pole-placement design method requires that all the state variables are accurately measured and are available for feedback. If this requirement is met and the system is controllable, then a complete control over all the eigenvalues is possible. A problem arises if some or all of the states are not actually measured or are not available for state feedback. An obvious solution would be to add more sensors that can measure the missing states. However, this approach may not always be feasible and often can be quite expensive. The other option is to estimate the unavailable states using a subsystem called a state observer. An observer that estimates all the states, including those that are actually measured, is called a full-state observer, and one that estimates only those states that are not measured is called a reduced-state observer. Here, we will discuss the procedure for the design of a full-state observer.

The design of an observer is based on the knowledge of a mathematical model of the plant, input(s), and output(s). The basic idea is to make the estimated states as close to the actual states as possible, but the problem is that all the actual states are not available for comparison. However, we do know the output of the given plant, and we can compare it with the estimated output of the observer. The design objective is then to drive the error between the actual and estimated outputs to zero as rapidly as possible so that, in the limit, the estimated states approach the actual states. The schematic diagram of such a full-state observer is shown in Fig. 5.43.

Suppose the dynamics of the plant and output are given by

$$\dot{x} = Ax + Bu \tag{5.285}$$

$$y = Cx \tag{5.286}$$

Let the observer dynamics and the output be given by

$$\dot{\hat{x}} = A\hat{x} + Bu + L(y - \hat{y}) \tag{5.287}$$

$$\hat{y} = C\hat{x} \tag{5.288}$$

so that

$$y - \hat{y} = C(x - \hat{x}) \tag{5.289}$$

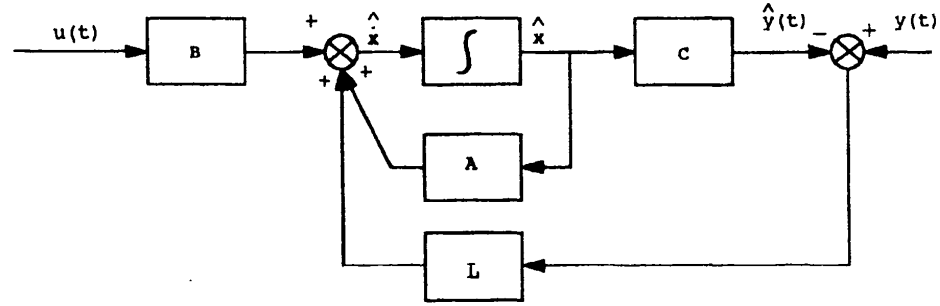


Fig. 5.43 Schematic diagram of full-state observer.

Here, $\hat{x}$ is the estimated state vector, and $\hat{y}$ is the estimated output. Let $e_x = x - \hat{x}$ be the error between the actual states and the estimated states and $e_y = y - \hat{y}$ be the error between the measured and estimated outputs. Then,

$$\dot{e}_x = \dot{x} - \dot{\hat{x}} = A(x - \hat{x}) - L(y - \hat{y}) \tag{5.290}$$

$$= (A - LC)(x - \hat{x}) \tag{5.291}$$

$$e_y = C(x - \hat{x}) \tag{5.292}$$

The objective of the design is to choose the observer gain matrix L such that the errors e_x and e_y approach zero as rapidly as possible. In other words, we choose the observer gain matrix L so that the closed-loop eigenvalues produce the desired transient response of the observer.

Plant given in dual phase-variable form. Suppose that the given plant A is in dual phase-variable form

$$A = \begin{bmatrix} -a_{n-1} & 1 & 0 & 0 & \cdot & 0 \\ -a_{n-2} & 0 & 1 & 0 & \cdot & 0 \\ -a_{n-3} & 0 & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -a_1 & \cdot & \cdot & \cdot & \cdot & 1 \\ -a_0 & 0 & 0 & 0 & \cdot & 0 \end{bmatrix} \tag{5.293}$$

Then, let us assume that the matrix $A - LC$ has the following form

$$A - LC = \begin{bmatrix} -(a_{n-1} + l_1) & 1 & 0 & 0 & \cdot & 0 \\ -(a_{n-2} + l_2) & 0 & 1 & 0 & \cdot & 0 \\ -(a_{n-3} + l_3) & 0 & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -(a_1 + l_n) & \cdot & \cdot & \cdot & \cdot & 1 \\ -(a_0 + l_n) & 0 & 0 & 0 & \cdot & 0 \end{bmatrix} \tag{5.294}$$

The characteristic equation of the observer system is then given by

$$s^n + (a_{n-1} + l_1)s^{n-1} + (a_{n-2} + l_2)s^{n-2} + \cdots + (a_0 + l_n) = 0 \tag{5.295}$$

Let the characteristic equation that gives the desired transient response of the observer be given by

$$s^n + d_{n-1}s^{n-1} + d_{n-2}s^{n-2} + \cdots + d_0 = 0 \tag{5.296}$$

Equating the coefficients of like powers of s, we obtain

$$l_1 = d_{n-1} - a_{n-1}, l_2 = d_{n-2} - a_{n-2}, \ldots, l_n = d_0 - a_0 \tag{5.297}$$

Here, $l_1, l_2, \ldots, l_n$ are the elements of the feedback gain matrix to achieve the desired performance of the full-state observer.

Plant not given in dual phase-variable form. Suppose the given plant

$$\dot{x} = Ax + BU \tag{5.298}$$

$$y = Cx \tag{5.299}$$

is not in dual phase-variable form. Then we assume that there exists a matrix P where $x = Pz$ that transforms this plant into a dual phase-variable form as given by

$$\dot{z} = A_z z + B_z u \tag{5.300}$$

$$y = C_z z \tag{5.301}$$

where $A_z = P^{-1}AP$, $B_z = P^{-1}B$, and $C_z = CP$. The observability matrix of the original system is given by

$$Q_{0.x} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix} \tag{5.302}$$

and that of the transformed system in phase-variable form is given by

$$Q_{0.z} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix} P \tag{5.303}$$

so that

$$P = Q_{0.x}^{-1} Q_{0.z} \tag{5.304}$$

Let

$$e_z = z - \hat{z} \tag{5.305}$$

Then,

$$\dot{e}_z = (A_z - L_z C_z)e_z \tag{5.306}$$

$$y - \hat{y} = C_z e_z \tag{5.307}$$

where $\hat{z}$ and $\hat{y}$ are the estimated state and the output vectors in the transformed system. Thus, the design procedure is as follows: 1) design the observer in the transformed space and obtain the gain matrix L_z and 2) transform back to the original system to get the corresponding gain matrix L_x as follows.

We have $z = P^{-1}x$, $\hat{z} = P^{-1}\hat{x}$, so that $e_z = z - \hat{z} = P^{-1}(x - \hat{x}) = P^{-1}e_x$ and $\dot{e}_z = P^{-1}\dot{e}_x$. Then,

$$\dot{e}_x = (A - L_x)e_x \tag{5.308}$$

$$y - \hat{y} = Ce_x \tag{5.309}$$

where

$$L_x = PL_z \tag{5.310}$$

Here, L_x is the gain matrix of the full-state observer corresponding to the given system.

Now one question that remains to be answered is how to construct the transformation matrix P. We will discuss one method that is based on knowing the characteristic equation or the eigenvalues of the given system. This approach is feasible because we have software tools like MATLAB that are available. We will illustrate this method in a later example.

Example 5.14

Given the plant,

$$\dot{x} = \begin{bmatrix} 1 & 2 & 1 \\ 3 & 5 & 2 \\ 4 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u$$

Convert it to the phase-variable form.

Solution. For the given system, the controllability matrix is given by

$$Q_c = [B \quad AB \quad A^2B] = \begin{bmatrix} 0 & 4 & 28 \\ 1 & 9 & 69 \\ 2 & 6 & 34 \end{bmatrix}$$

The rank of this matrix is three because all the columns are linearly independent. The inverse of this matrix exists and is given by

$$Q_c^{-1} = \begin{bmatrix} -1.35 & 0.4 & 0.3 \\ 1.30 & -0.7 & 0.35 \\ -0.15 & 0.10 & -0.05 \end{bmatrix}$$

Then,

$$P_1 = [0 \quad 0 \quad 1][B \quad AB \quad A^2B]^{-1}$$

$$= [0 \quad 0 \quad 1] \begin{bmatrix} -1.35 & 0.4 & 0.3 \\ 1.30 & -0.7 & 0.35 \\ -0.15 & 0.10 & -0.05 \end{bmatrix}$$

$$= [-0.15 \quad 0.10 \quad -0.05]$$

Similarly,

$$P_2 = P_1A$$

$$= [-0.05 \quad 0.20 \quad -0.10]$$

$$P_3 = P_1A^2$$

$$= [0.15 \quad 0.90 \quad 0.05]$$

so that

$$P = \begin{bmatrix} -0.15 & 0.10 & -0.05 \\ -0.05 & 0.20 & -0.10 \\ 0.15 & 0.90 & 0.05 \end{bmatrix}$$

$$PAP^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -7 & -13 & 9 \end{bmatrix}$$

$$PB = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Then, the phase-variable form of the given system is given by

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -7 & -13 & 9 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u$$

where $z = Px$.

Example 5.15

For the plant,

$$G(s) = \frac{25(s+2)}{(s+1)(s+3)(s+5)}$$

1) Represent the plant in phase-variable, state-space form.

2) Design a phase-variable, full-state feedback controller to yield 15% overshoot with a settling time of 1 s.

Solution. For 15% overshoot, from Eq. (5.75), we find that $\zeta = 0.5169$. Then, $\omega_n = 4/\zeta = 7.7384$, $\omega_d = \omega_n\sqrt{1-\zeta^2} = 6.6245$, and $\sigma = \zeta * \omega_n = 4.0$. Therefore, the dominant poles are at $-4 \pm j6.6245$. Because the given system is a third-order system, we must choose one more pole. Let this pole be located at -2.1 so that it nearly cancels the zero at -2 justifying the second-order approximation.

Therefore, the characteristic equation that gives the desired response is given by

$$(s+4-j6.6245)(s+4+j6.6245)(s+2.1)=0$$

or

$$s^3+10.1s^2+76.6s+125.7543=0$$

The next step is to express the given plant in state-space, phase-variable form. For this purpose, let us decompose the transfer function into two blocks (one for the numerator and another for the denominator) in cascade as schematically shown earlier in Fig. 5.41b.

For the first block,

$$G_1(s)=\frac{\bar{x}_1(s)}{\bar{r}(s)}$$

$$=\frac{1}{(s+1)(s+3)(s+5)}$$

$$=\frac{1}{s^3+9s^2+23s+15}$$

or

$$(s^3+9s^2+23s+15)\bar{x}_1(s)=\bar{r}(s)$$

Taking the inverse Laplace transform,

$$\frac{d^3x_1}{dt^3}+\frac{9d^2x_1}{dt^2}+\frac{23dx_1}{dt}+15x_1=r(t)$$

Let $\dot{x}_1=x_2$ and $\dot{x}_2=x_3$ so that

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -15 & -23 & -9\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}r(t)$$

Consider the second block. Following a similar procedure,

$$y(t)=\begin{bmatrix}50 & 25 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

The phase-variable form with full-state feedback system is given by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -(15+k_1) & -(23+k_2) & -(9+k_3)\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}r(t)$$

The characteristic equation of the given system is given by

$$s^3+(9+k_3)s^2+(23+k_2)s+(15+k_1)=0$$

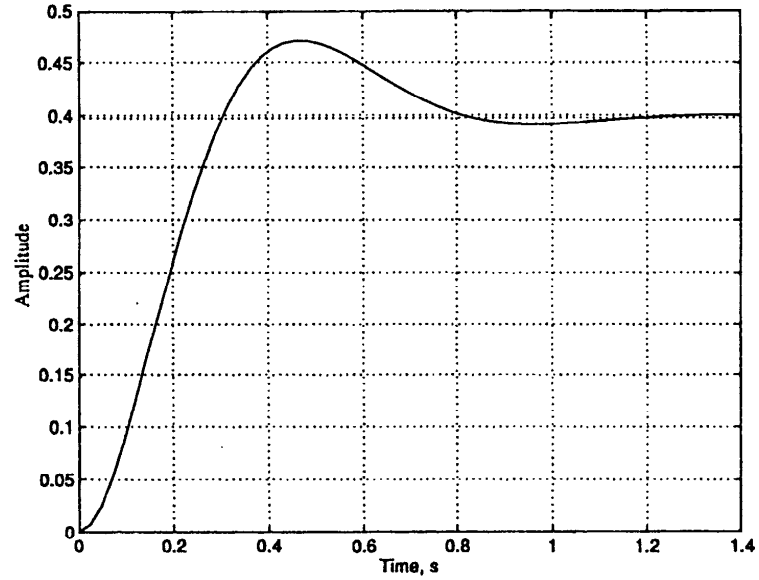


Fig. 5.44 Unit-step response of full-state feedback design of Example 5.15.

Comparing the coefficients of the above characteristic equation with that of the desired characteristic equation, we get $k_1=110.7543$, $k_2=53.6$, and $k_3=1.10$.

Thus, the given system with full-state feedback is given by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -125.7543 & -76.6 & -10.10\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}r(t)$$

Now, we verify the design by simulating a response to a unit-step input using MATLAB.[4] The results of the simulation are shown in Fig. 5.44. It can be observed that the design requirements have been met.

Example 5.16

Given the system

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}-5 & 2 & 0\\ 1 & -3 & 1\\ 0 & 1 & -1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}u(t)$$

$$y=\begin{bmatrix}1 & 0 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

1) Express the given plant in dual phase-variable form. 2) Design a full-state observer so that the closed-loop characteristic equation is given by $s^3 + 100s^2 + 1500s + 40{,}000 = 0$. 3) Verify the design for $u(t) = 25t$ assuming $x_1(0) = 5$, $x_2(0) = 1.5$, and $x_3(0) = 0.25$.

Solution. We have the given system

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

$$y = Cx$$

where

$$A = \begin{bmatrix} -5 & 2 & 0 \\ 1 & -3 & 1 \\ 0 & 1 & -1 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

It is convenient to use MATLAB[4] to compute the observability matrix, which for this system is found to be

$$Q_{0,x} = \begin{bmatrix} 1 & 0 & 0 \\ -5 & 2 & 0 \\ 27 & -16 & 2 \end{bmatrix}$$

Using MATLAB,[4] we find that this system has full rank of three; hence the system is observable. Furthermore,

$$Q_{0,x}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2.5 & 0.5 & 0 \\ 6.5 & 4.0 & 0.5 \end{bmatrix}$$

Next step is to express the given plant in the dual phase-variable form. We know that the elements of the first column of the matrix in dual phase-variable form are the coefficients of the characteristic equation. Conversely, given the coefficients of the characteristic equation, we can directly write down the dual phase-variable form of matrix A.

As we know, the characteristic equation is given by

$$\Delta(s) = |sI - A| = 0$$

However, expanding the determinant in the characteristic equation is simple if it is of an order lower than three. However, for determinants of orders greater than three, this is a tedious job. For this purpose, MATLAB[4] comes out very handy. Using MATLAB,[4] we get the characteristic equation

$$s^3 + 9s^2 + 20s + 8 = 0$$

Then, the dual phase-variable form of the given system is

$$A_z = \begin{bmatrix} -9 & 1 & 0 \\ -20 & 0 & 1 \\ -8 & 0 & 0 \end{bmatrix}$$

Furthermore, we assume that

$$C_z = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

Then,

$$Q_{0.z} = \begin{bmatrix} C_z \\ C_z A_z \\ C_z A_z^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ -9 & 1 & 0 \\ 61 & -9 & 1 \end{bmatrix}$$

and

$$P = Q_{0.x}^{-1} Q_{0.z}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 0.5 & 0 \\ 1 & -0.5 & 0.5 \end{bmatrix}$$

Now let us design the full-state observer for the system transformed in dual phase-variable form. We have

$$A_z - L_z C_z = \begin{bmatrix} -(9 + l_1) & 1 & 0 \\ -(20 + l_2) & 0 & 1 \\ -(8 + l_3) & 0 & 0 \end{bmatrix}$$

The characteristic equation of the observer is given by

$$s^3 + (9 + l_1)s^2 + (20 + l_2)s + 8 + l_3 = 0$$

The desired characteristic equation is

$$s^3 + 100s^2 + 1500s + 40{,}000 = 0$$

Comparing the coefficients of like powers of s, we get $l_1 = 91$, $l_2 = 1480$, and $l_3 = 39{,}992$. Then, the gain matrix corresponding to the given system is obtained by transforming back as

$$L_x = PL_z = \begin{bmatrix} 91 \\ 558 \\ 19{,}347 \end{bmatrix}$$

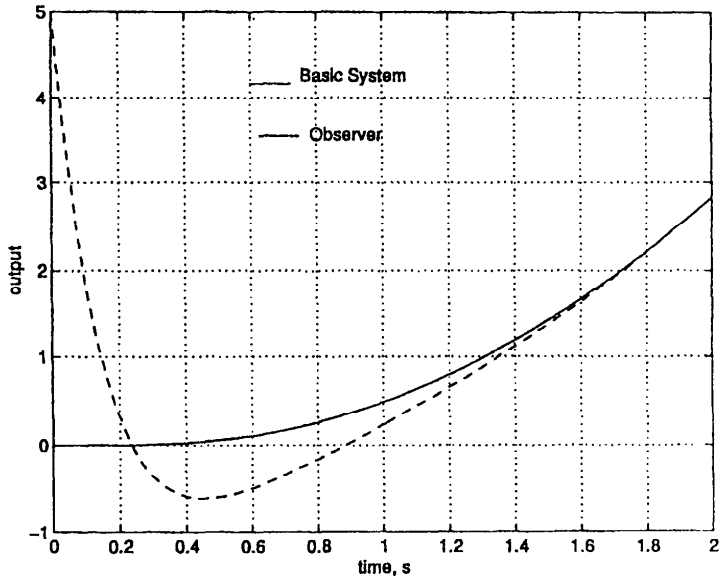


Fig. 5.45 Response of full-state observer of Example 5.16.

The performance of the observer for the given initial conditions of $x_1(0) = 5$, $x_2(0) = 1.5$, and $x_3(0) = 0.25$ and to the input $r(t) = 25t$ is shown in Fig. 5.45 using MATLAB.[4] We note that the observer performs as expected.

5.11 Summary

In this chapter, we have reviewed the basic principles of linear systems and illustrated the theory with a number of solved examples. This background will be useful in the study of aircraft dynamics and control. We will derive longitudinal and lateral-directional transfer functions and study the free response of the aircraft. We will also use the design methods we have learned here for the design of stability augmentation systems and automatic flight control systems of the aircraft to obtain the desired handling qualities. It was not possible to go into all the details of linear systems theory. Readers interested in getting more detailed information on control systems may refer elsewhere.[1–3]

References

[1]Ogata, K., *Modern Control Engineering*, Prentice–Hall, Englewood Cliffs, NJ, 1986.

[2]Norman, S., and Nise, N. S., *Control System Engineering*, Benjamin/Cummings, 1992.

[3]Kuo, B., *Automatic Control Systems*, 4th ed., Prentice–Hall, Englewood Cliffs, NJ, 1982.

[4]*Pro-MATLAB for Sun Workstations*, The MathWorks, Natick, MA, Jan. 1990.

Problems

5.1 Sketch the Bode plot for the open-loop systems with

$$\text{(a)}\quad G(s) = \frac{50}{s(s+5)}$$

$$\text{(b)}\quad G(s) = \frac{25(s+2)}{(s+3)(s+5)(s+7)}$$

5.2 Using Routh's criterion, examine the stability of the closed-loop system with a characteristic polynomial given by

$$\text{(a)}\quad s^4 + 5s^3 + 3s^2 + s + 2 = 0$$

$$\text{(b)}\quad s^4 + 2s^3 + 0.001s^2 + 3s + 4 = 0$$

$$\text{(c)}\quad s^4 + 4s^3 + 7s + 2 = 0$$

[Answer: (a) Two sign changes, unstable; (b) Two sign changes, unstable; and (c) Two sign changes, unstable.]

5.3 Sketch the root-locus for a unity feedback system with

$$\text{(a)}\quad G(s) = \frac{k(s+3)}{s(s+2)(s+4)}$$

$$\text{(b)}\quad G(s) = \frac{k(s+1)}{s(s+2)(s+3)(s+7)}$$

Determine the value of the gain k when the closed-loop system in (b) becomes unstable.

5.4 Sketch the root-locus for a unity feedback system with

$$G(s) = \frac{k(s+3)(s+7)}{s(s+1)}$$

Find the value of the gain k so that the closed-loop system is stable and is operating with a damping ratio of 0.7.

5.5 For the following unity feedback systems, sketch the Nyquist plot.

$$\text{(a)}\quad G(s) = \frac{25}{(s+1)(s+3)}$$

$$\text{(b)}\quad G(s) = \frac{k(s+2)}{(s-3)(s-4)}$$

For the system in (b), determine the value of the gain for which the closed-loop system becomes unstable.

5.6 Using the Nyquist plot, determine the gain and phase margins of the system given by

$$G(s) = \frac{k(s-2)(s-3)}{s(s+5)(s+9)}$$

5.7 For the system in Exercise 5.6, use Bode plots to obtain the gain and phase margins.

5.8 For a unity feedback system with

$$G(s) = \frac{k}{(s+2)(s+4)}$$

design a PI controller to reduce the steady-state error to zero for a unit-step input. Assume that the system is operating with a damping ratio of 0.6. Plot the unit-step response to verify your design. Compare the values of T_s and T_p for the basic and compensated systems.

5.9 For the system given in Example 5.8, design a lag compensator to reduce the steady-state error by a factor of 15.

5.10 For the unity feedback system given by

$$G(s) = \frac{k}{(s+1)(s+2)(s+3)}$$

(a) Determine the value of the gain k for 15% overshoot. Determine the corresponding values of T_s and T_p.

(b) Design a PD controller for reducing T_p by a factor of 2 and T_s by 50%, while operating at 15% overshoot in both cases.

5.11 Given the unity feedback system with

$$G(s) = \frac{k}{s(s+3)(s+5)}$$

determine the value of gain k for the system to operate with a damping ratio of 0.51. Find the corresponding locations of closed-loop poles. If a lead compensator is to be designed to reduce the time for peak amplitude T_p by 50%, with compensator zero placed at -2.5, find the compensator pole location. How does the performance of the compensated system compare with that of the basic system?

5.12 For the unity feedback system with

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

(a) show that the system cannot be made to operate with time for peak amplitude of 2.0 s and 23.38% overshoot by simple gain adjustment and (b) design a suitable compensator to achieve this performance.

5.13 For the unity feedback system with

$$G(s) = \frac{k}{(s+1)(s+4)(s+7)}$$

design a PID controller that will give time for peak amplitude of 1.2 s and 15% overshoot with zero steady-state error for a unit-step input. Plot the unit-step response for the basic and PID-compensated systems.

5.14 For the system shown in Fig. P5.14, determine the values of gain k_h and k so that the minor loop operates with a damping ratio of 0.707 and the entire closed-loop system has 15% overshoot.

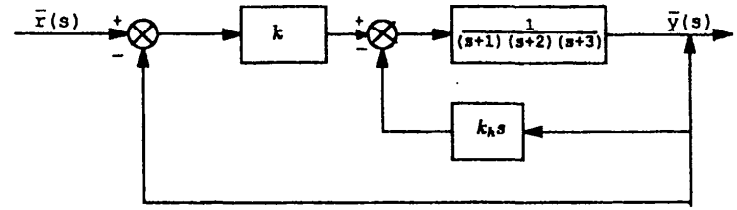


Fig. P5.14 Control system for Exercises 5.14 and 5.15.

5.15 Determine the rate gyro gain k_h for the system shown in Fig. P5.14 so that the compensated system operates at one-third the settling time compared to the basic system while continuing to have the same 15% overshoot.

5.16 Given the linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

find (a) eigenvalues of the matrix A, (b) the state transition matrix $\Phi(t)$, and (c) state vector $x(t)$ for the following cases:

(i) $A = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix}$ $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $x(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

(ii) $A = \begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix}$ $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $x(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

(iii) $A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$ $B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ $x(0) = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$

5.17 Given the state equation

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 3 & 5 & 2 \\ 4 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} u(t)$$

$$y = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Can this system be transformed into phase-variable form? If so, find the transformation $z = Px$ so that the transformed system $\dot{z} = A_z z + B_z u(t)$, $y = C_z z$ is in phase-variable form.

5.18 Represent the following system in state-space, phase-variable form:

$$\frac{d^3x}{dt^3} + \frac{2d^2x}{dt^2} + \frac{3dx}{dt} + 5x = u(t)$$

5.19 Given the state equation

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1 \\ 1 & 2 & -3 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u(t)$$

$$Y = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Can this system be transformed to dual-phase variable form? If so, find the transformation $z = Px$ such that the transformed system $\dot{z} = A_z z + B_z u(t)$, $y = C_z z$ is in dual phase-variable form.

5.20 Design a phase-variable, full-state feedback controller for the plant given by

$$G(s) = \frac{10(s + 0.8)}{(s + 2)(s + 3)(s + 5)}$$

to yield a 15% overshoot with a settling time of 0.8 s.

5.21 Design a full-state feedback observer for the plant

$$G(s) = \frac{1}{s(s + 3)(s + 6)}$$

so that the closed-loop characteristic equation of the observer system is given by

$$s^3 + 90s^2 + 2000s + 10{,}000 = 0$$