STABILITY THEORY, ASYMPTOTIC

STABILITY CRITERIA

Automatic control is an essential part of engineering and science. It finds applications in many areas from space vehicles and missiles to industrial processes and medicine. Automatic control devices, laboratory equipment, design and analysis tools, and complete automatic processes and systems are offered by many companies, some of which are listed in Table 1. Basically, a control system consists of interconnected components that achieve a desired response. In order to meet the objectives effectively, the system must be understood fully and properly modeled mathematically. When the system is mathematically represented, it may be designed appropriately and the performance may be examined and analyzed. For the performance analysis, many methods are available. For example, the classic control theory is the earliest and one of the most established methods, mainly applied in simple systems.

Although a nonlinear approach is available, in classic control theory, the foundations of analysis are mainly based on linear system theory. The linear system approach assumes a cause–effect relationship between the components of the system and expresses this relationship as differential equations.

Table 1. List of Manufacturers

Automated Applications Inc. 680 Flinn Ave. Unit 36 Moorpark, CA 93021 Tel: 800-893-4374 Fax: 805-529-8630	FTI International, Inc. Ha'shikma St. Ind. Zone P.O.B. 87 Kfar Saba, Israel Tel: 052-959152-4 Fax: 052-959162
Automation Technologies International 17451 W. Dartmoor Drive Grayslake, IL 60030-3014 Tel: 708-367-3347 Fax: 708-367-1475	Kuntz Automation Engineering 402 Goetz Street, Dept. 7 Santa Ana, CA 92707 Tel: 714-540-7370 Fax: 714-540-6287
Capitol Technologies, Inc. 3613 Voorde Drive South Bend, IN 46628 Tel: 219-233-3311 Fax: 219-233-7082	The Math Works, Inc. 24 Prime Park Way Natic, MA 01760-1500 Tel: 508-647-7000 Fax: 508-647-7001
CEI Automation 15250 E. 33rd Place Aurora, CO 80011 Tel: 303-375-0050 Fax: 303-375-1112	Munck Automation Technology 161-T Enterprise Drive Newport News, VA 23603 Tel: 804-887-8080 Fax: 804-887-5588
Control Engineering Company 8212 Harbor Springs Road Harbor Springs, MI 49740 Tel: 800-865-3591	Portech Pathfinder Operations 1610 Fry Avenue, Dept. T Canon City, CO 81212 Tel: 800-959-0688 Fax: 719-269-1157
Design Technology Corporation 5 Suburban Park Drive Billerica (Boston), MA 01821 Tel: 508-663-7000 Fax: 508-663-6841	Prime Automation, Inc. 1321 Capital Drive Rockford, IL 61109-3067 Tel: 815-229-3800 Fax: 815-229-5491
Fata Automation 37655 Interchange Drive Farmington, MI 48335 Tel: 810-478-9090 Fax: 810-478-9557	

Once the system performance is stated as differential equations in the time domain, Laplace transforms are commonly, used for frequency analysis. Some examples of Laplace transforms are given in Table 2.

A control system can be an open loop or closed loop. Figure 1 illustrates an open-loop system in which a controller controls the process without using any feedback. In this case, the output is not compared with the input; therefore, deviations of the output from the desired value cannot be automatically corrected. This method finds limited application since it does not lead to fully automatic systems.

In a closed-loop system, the actual output is compared with a desired reference input by a suitably arranged feedback mechanism. Figure 2 illustrates a single-input–single-output feedback control system. In this system, a prescribed relationship of one system variable to another is maintained by comparing the two and using the difference as a means of control. Using system modeling and mathematical representations, a closed loop control system with a single-input single-output may be represented as shown in Fig. 3. In this case, the relationship between the input and output of the single-input–single-output system can be expressed as

$$e(t) = r(t) - c(t)H \tag{1}$$

$$c(t) = e(t)G \tag{2}$$

where

$r(t)$ = reference input
$e(t)$ = error signal
$c(t)$ = output signal
G = forward path gain or process transfer function
H = feedback gain

Eliminating the error $e(t)$ and rearranging Eqs. (1) and (2) gives the closed-loop gain

$$M = \frac{c(t)}{r(t)} = \frac{G}{1+GH}$$

or in the Laplace transform domain,

$$M(s) = \frac{C(s)}{R(s)} = \frac{G(s)}{1+G(s)H(s)} \tag{3}$$

The main effect of the feedback is that it reduces the error between the reference input and the system output, thus forcing the system output to follow or track the reference input. It also affects the forward gain G by a factor of $1/(1 + GH)$. This effect is one of the main subjects of study in classic control theory. For example, when $GH = -1$ the gain M will be infinite; hence, C will increase without bound, thus leading to unstable conditions.

In many cases, the control systems are much more complex than a single-input–single-output arrangement. They may have many inputs and outputs and controlled variables that are interrelated as shown in Fig. 4. These systems necessitate a multivariable control system approach for analysis and design. However, the feedback concept of the single-input–single-output linear system will be the main concern of this article.

A control system needs to be designed carefully with a suitable configuration with clearly identified specifications to achieve the desired performance. In this process, the identification of key parameters, mathematical representation of the system, and system analysis play important roles. After having mathematically represented the system, it is possible to employ analytical tools to describe the characteristics of the feedback control system. Important characteristics, such as the transient and steady state performance, frequency response, sensitivity, and robustness can be studied in detail. When these characteristics are known, the desired response can be obtained by adjusting the system parameters. Therefore, a good understanding and effective use of stability theory in control systems is very important. If the system is not stable, it will display an erratic and destructive response and will get out of bounds and disintegrate.

The transient response of a system is related to its stability. Typical responses times of second-order systems are illustrated in Fig. 5. In this system, the outputs are bounded by the decaying oscillations. Therefore, a stable system may be defined as a system with bounded response. If the oscillations increase with time, the system is said to be unstable. The stability of a dynamic system can be described by its response to an input disturbance. The output response can be either decreasing, increasing, or neutral, giving an indication of sta-

Table 2. Laplace Transform Table

Time Domain Function	Laplace Transform
$u(t)$ (unit step)	$\frac{1}{s}$
t	$\frac{1}{s^2}$
t^n, for any positive integer n	$\frac{n!}{s^{n+1}}$
e^{-at}	$\frac{1}{s+a}$
$\frac{e^{-at} - e^{-bt}}{b-a}$	$\frac{1}{(s+a)(s+b)}$
$\frac{\omega_n}{\sqrt{1-\xi^2}} e^{-\xi\omega_n t} \sin(\omega_n \sqrt{1-\xi^2}\, t)$	$\frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2}$
$\sin(\omega_n t)$	$\frac{\omega_n}{s^2 + \omega_n^2}$
$\cos(\omega_n t)$	$\frac{s}{s^2 + \omega_n^2}$
$1 + \frac{1}{\sqrt{1-\xi^2}} e^{-\xi\omega_n t} \sin(\omega_n \sqrt{1-\xi^2}\, t - \varphi)$ where $\varphi = \tan^{-1}\left(\frac{\sqrt{1-\xi^2}}{-\xi}\right)$	$\frac{\omega_n^2}{s(s^2 + 2\xi\omega_n s + \omega_n^2)}$
$1 - \cos(\omega_n t)$	$\frac{\omega_n^2}{s(s^2 + \omega_n^2)}$
$1 - e^{-t/T}$	$\frac{1}{s(1+Ts)}$
$\frac{1}{2\omega_n}[\sin(\omega_n t) + \omega_n t \cos(\omega_n t)]$	$\frac{s^2}{(s^2 + \omega_n^2)^2}$
$1 - 2T + (1 + 2T)e^{-t/T}$	$\frac{1}{s^2(1+Ts)^2}$

bility as defined by stability criteria. There are three types of stability of control systems: the bounded-input–bounded-output (BIBO) stability, asymptotic stability, and marginal stability (stability in the sense of Lyapunov).

BOUNDED-INPUT–BOUNDED-OUTPUT STABILITY

A system is stable if a bounded input yields to a bounded output. When the closed-loop transfer function of a linear system is expressed as a Laplace transform, the stability may be defined in terms of the locations of the poles of the system in the complex plane or the s-plane. An s-plane is shown in Fig. 6 indicating the right half plane and left half plane. Take a single-input–single-output system and express the input and output relationship as

$$M(s) = \frac{C(s)}{R(s)} = \frac{K\prod_{i=1}^{L}(s+z_i)}{\prod_{k=1}^{U}(s+p_k)\prod_{j=1}^{R}(s^2 + 2\alpha_j s + \alpha_j^2 s + \alpha_j^2 + \omega_j^2)} \quad (4)$$

where the denominator of $M(s)$ is the characteristic equation of the system. The roots of the characteristic equation are the poles of the closed-loop system. The time response of the output is a function of the roots of this characteristic equation.

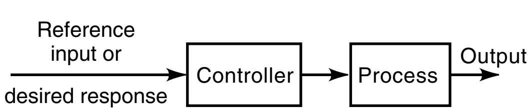

Figure 1. An open-loop control system. An open-loop control system does not compare the output and the input; therefore, any deviations between the two cannot be corrected automatically. This system is applicable only in simple cases in which the process characteristics are fully known and the outputs from the desired values are not all very important.

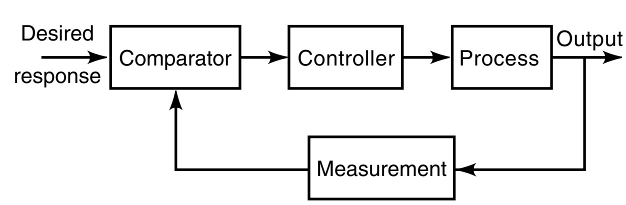

Figure 2. A closed-loop control system. In a single-input–single-output system, the output of the system is compared with the input, and any deviation between the two is corrected by suitably designed controllers. However, the use of feedback can lead to instability. Closed-loop control arrangements are used extensively in automatic processes and devices.

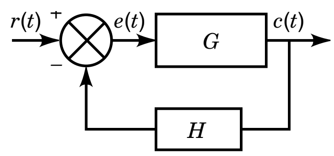

Figure 3. Block diagram of a closed-loop system. A closed-loop system may be represented mathematically by a forward loop transfer function G and a feedback loop transfer function H. The relation between the input and output can be expressed in terms of these two terms in the form of a system transfer function, which is extremely useful in system design and analysis.

For example, the time response of this system for an impulse function may be written as

$$c(t) = \sum_{k=1}^{U}(A_k e^{-P_k t}) + \sum_{j=1}^{R}\left(\frac{B_j e^{-\alpha_j t}\sin(\omega_j t)}{\omega_j}\right) \tag{5}$$

Left-Half-Plane Poles

To clarify the important concept of locations of poles on the s-plane, as used in stability analysis, see Eq. (4). As we can see in this equation, the poles of the closed-loop system may be real or complex, and simple or repeated. It is often convenient to plot the poles and zeros of the closed-loop transfer function on the s-plane. The s-plane can be considered to be in three parts, the right half plane (RHP), the left half plane (LHP), and the pure imaginary axis or $j\omega$-axis. If a pole lies inside the open LHP, then the pole has a negative real part. If it lies inside the closed RHP, then it has a positive or zero repeated real part.

Consider poles $1/(s + p)^n$ on the s-plane. For a real p, a portion of the time domain response of the system will be proportional to

$$c(t) \propto 1/n!\,t^{n-1}e^{-pt} \tag{6}$$

If $p < 0$, it lies on the RHP, and its response increases exponentially. If it is at the origin, $p = 0$ and simple, its response is a step function. When $p = 0$ and repeated with multiplicity $n > 2$, then its response approaches infinity as $t \to \infty$. If the poles are in the LHP or $p < 0$, then the response e^{-pt} approaches zero as $t \to \infty$.

Therefore,

1. The time response of a pole, simple or repeated, approaches zero as $t \to \infty$ if and only if the pole lies inside the open-loop LHP or has a negative real part.

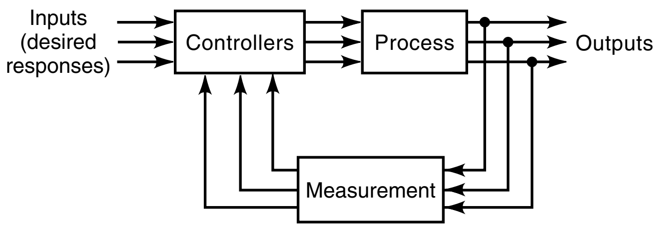

Figure 4. A multivariable control system. Many control systems have multiple inputs and multiple outputs. In these cases, using the multivariable control theory and matrix approach is applicable. In the design and analysis, most of the theories developed for single-input–single-output systems can still be used.

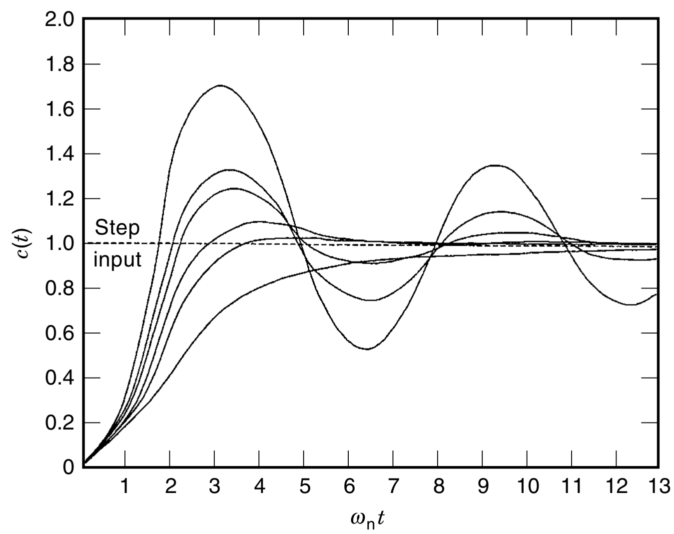

Figure 5. Time response of a second-order system. The output of a second-order system for a unit step input contains a transient and steady state response. The sinusoidals frequency and amplitude components depend on the natural frequency and the damping ratio of the system. If the oscillation increases without bound, the system is said to be unstable. The stability can be related to the locations of the poles on the s-plane.

2. The time response of a pole approaches a nonzero constant as $t \to \infty$ if and only if the pole is simple and located at $s = 0$.

As indicated earlier, in order to obtain a bounded response to a bounded input, the poles of the closed-loop system must be in the left-hand portion of the s-plane. That is, $p_k > 0$ and $\alpha j > 0$ so that the exponential terms e^{-pkt} and $e^{-\alpha jt}$ decay to zero as the time goes to infinity. A necessary and sufficient condition is that a feedback system is stable if all the poles of the system transfer function have negative real parts. If the characteristic equation has simple roots on the imaginary axis ($j\omega$) with all other roots on the left half plane, the steady state output is sustained oscillations for a bounded input. If the input is sinusoidal with a frequency equal to the magni-

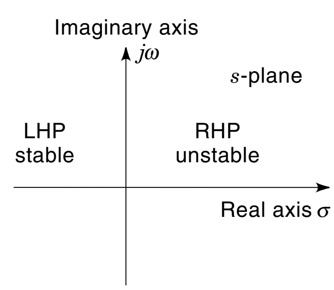

Figure 6. A complex plane. The response of control system depends on the locations of poles and zeros of the characteristic equation on the complex plane also known as the s-plane. The poles in the closed-loop transfer function on the right half plane lead to instability because the exponential term in the time domain representation increases as time increases. The poles in the LHP indicate stable conditions, whereas poles on the imaginary axis may lead to stable or unstable conditions. In this case, repetition of roots on the imaginary axis and the inputs must be taken into account.

tude of the $j\omega$-axis pole, the output becomes unbounded. This is called marginal stability because only certain bounded inputs cause the output to become unbounded. For an unstable system, the characteristic equation has at least one root in the right half of the s-plane, that is at least one of the exponential terms e^{-pkt} and/or $e^{-\alpha jt}$ will increase indefinitely as the time increases. Repeated $j\omega$-axis roots will also result in an unbounded output for any input.

ASYMPTOTIC STABILITY

In general, the output response of a linear time-invariant system may be divided into two components.

1. The *forced response* is the part of the response that has the same form as the input.
2. The *natural response* is the part of the response that follows a form, which is dictated by the poles of the characteristic equation.

In some cases, the investigation of stability by using only the transfer function $M(s)$ is not sufficient. Hence, the nature of the input signal must be taken into account. For example, a plant output $c(t)$ is said to track or follow the reference input $r(t)$ asymptotically if

$$\lim_{t\to\infty} |c(t) - r(t)| \to 0 \tag{7}$$

Suppose that the transfer function of the overall control system is $M(s)$; if $M(s)$ is not stable the system cannot follow any reference signals. If $M(s)$ is stable, in order for the system to be asymptotically stable, it is an additional requirement that the system be capable of following all inputs. This is important because, in some cases, the output may be excited by nonzero initial conditions such as noise or disturbance. As a result, the stability conditions may be generalized as follows.

1. The system is stable if the natural response approaches zero as $t \to \infty$.
2. The system is unstable if the natural response grows without bound as $t \to \infty$.
3. The system is marginally stable or marginally unstable if the natural response neither grows nor decays as $t \to \infty$.
4. The system is stable if bounded inputs result in bounded outputs.
5. The system is unstable if bounded inputs result in unbounded outputs.

Here, in order to explain asymptotic stability and to lay a firm background for the following theories, a rigorous mathematical approach may be introduced. To observe the natural response $h(t)$ of a linear time-invariant system, a dirac delta function (impulse) $\delta(t)$ may be applied to the input to give the system internal energy upon which to act. The dirac delta function has a rectangular shape with a height of $1/\epsilon$ and a width of ϵ. ϵ is made vanishingly small so that the function has infinite height and zero width and unit area. The ensuing response is the natural response. Its Laplace transform is identical to the transfer function of the system, which can be written in the general partial fraction form as

$$\begin{aligned} H(s) &= \sum_{i=1}^{n} F(s,\, p_i,\, r_i) \\ F(s,\, p_i,\, r_i) &= \frac{K_1}{(s-p_i)^{r_i}} + \frac{K_2}{(s-p_i)^{r_i-1}} + \cdots + \frac{K_r}{(s-p_i)} \end{aligned} \tag{8}$$

where there are n sets of poles, located at $s = p_i$, each of multiple of r_i.

The impulse response may be written in terms of the system poles by taking the inverse Laplace transform of $H(s)$. The general expression is

$$\begin{aligned} h(t) &= \sum_{i=1}^{n} f(t,\, p_i,\, r_i) \\ f(t,\, p,\, r) &= e^{pt}(k_1 t^{r-1} + k_2 t^{r-2} + \cdots + 1) \end{aligned} \tag{9}$$

The behavior of $h(t)$ is dictated by the behavior of $f(t, p, r)$. As $t \to \infty$, $f(t, p, r)$ becomes dominated by the $e^{pt}t^{r-1}$ term, so the behavior of $h(t)$ as t becomes large may be investigated by the following limit:

$$\begin{aligned} L &= \lim_{t\to\infty} (t^{r-1}|e^{pt}|) \\ &= \lim_{t\to\infty} \left(\frac{t^{r-1}}{e^{-\alpha}}\right) \end{aligned} \tag{10}$$

where $p = \sigma + j\omega$. The limit is in the infinity divided by infinity indeterminate form. Applying L'Hopital's rule $r - 1$ times results in

$$\begin{aligned} |L| &= \lim_{t\to\infty} \left(\frac{(r-1)!}{\sigma^{r-1}e^{-\sigma t}}\right) \\ &= \lim_{t\to\infty} \left(\frac{(r-1)!}{\sigma^{r-1}} e^{\sigma t}\right) \end{aligned} \tag{11}$$

There are several regions of interest.

1. $\sigma > 0$. The $e^{\sigma t}$ term forces the limit to approach infinity.
2. $\sigma < 0$. The $e^{\sigma t}$ term forces the limit to approach zero.
3. $\sigma = 0$, $r = 1$. In this case, we have a zero divided by zero indeterminate form with three independent variables. The solution is obtained by allowing the limit $(\sigma, r - 1, t) \to (0, 0, \infty)$ to be approached from an arbitrary trajectory.

$$\begin{aligned} L &= \lim_{(\sigma,\, r-1,\, t)\to(0,\,0,\,\infty)} \left(\frac{(r-1)!}{\sigma^{r-1}} e^{\sigma t}\right) \\ &= \frac{1}{1} e^{k} \\ &= C \end{aligned} \tag{12}$$

   where k depends of the trajectory of approach and C is a bounded constant.

4. $\sigma = 0$, $r > 1$. This time we have a limit with two independent variables, $(\sigma, t) \to (0, \infty)$.

$$\begin{aligned} L &= \lim_{(\sigma,\, t)\to(0,\,\infty)} \left(\frac{(r-1)!}{\sigma^{r-1}} e^{\sigma t}\right) \\ &= \lim_{\sigma\to 0} \left(\frac{(r-1)!}{\sigma^{r-1}} e^{k}\right) \\ &\to \infty \end{aligned} \tag{13}$$

   where k depends on the trajectory of approach.

We can now summarize the stability of a system depending of the location of the poles.

Pole Locations	Stability
Poles on the left half plane ($\sigma < 0$) only.	The natural response approaches zero, so the system is stable.
Any pole on the right half plane ($\sigma > 0$), or pole of multiplicity greater than one on the $j\omega$-axis ($\sigma = 0$ and $r > 1$).	The natural response approaches infinity, so the system is unstable.
Any pole on the $j\omega$-axis of multiplicity equal to one ($\sigma = 0$ and $r = 1$).	The natural response approaches neither zero nor infinity. It is, however, bounded. The system is called marginally stable or marginally unstable.

Consider the case of a pole of multiplicity equal to one on the $j\omega$-axis, if the input were to be a sinusoid of a frequency equal to the distance of this pole from the origin. This would have the same effect on the total response as if the input were zero and the system had a pole of multiplicity equal to two on the $j\omega$-axis. The output would then approach infinity even though the inputs were bounded. Consequently, a bounded input function that will produce an unbounded output exists. A system classified as marginally stable under the asymptotic stability definition is, therefore, classified as unstable under the bounded-input–bounded-output definition.

ROUTH–HURWITZ CRITERION

To determine the stability of a system, we need to know whether any poles are located in the RHP. It is always possible to calculate these pole locations by direct computational methods, but it is not necessary. For determining system stability, it is enough just to know whether there are any poles on the RHP or not. This can be investigated using the Routh–Hurwitz criterion.

A generalized nth-order characteristic polynomial may be represented as

$$\begin{aligned} P(s) &= \sum_{i=0}^{n} a_i s^i \\ &= k\prod_{i=0}^{n}(s+p_i) \end{aligned} \tag{14}$$

where a_i are the polynomial coefficients, k is a constant and $s = -p_i$ are the roots of the polynomial.

The Routh–Hurwitz criterion is based on the Mikhailov criterion, which states that if a system is characterized by an nth order polynomial $P(s)$, then it is necessary and sufficient for stability that the following condition be satisfied.

The contour traced in the $P(s)$ domain by $P(j\omega)$, $0 \le \omega < \infty$, must proceed counterclockwise around the origin and $\lim_{\omega\to\infty}\arg[P(j\omega)]$ must tend toward $(n\pi/2)$. If the Mikhailov criterion is applied algebraically to the generalized form of $P(s)$ given previously, then the Routh–Hurwitz criterion results are based on the determinants of the coefficients. From these determinants, it is possible to derive a set of polynomials, known as subsidiary functions. If the coefficients of these functions are listed, we are left with what is known as Routh's array, which is the tabular technique presented next.

The Routh–Hurwitz criterion may be expressed as follows. There are two necessary, but not sufficient, conditions for no RHP poles.

1. All the polynomial coefficients a_i must have the same sign. The coefficients are determined by cross-multiplication of roots p_i. If two particular coefficients were of opposite sign, it would mean that one cross multiplication yielded a positive result whereas another yielded a negative result. This is possible only if there exist at least two p_i of opposite sign, which means that one of them must be on the right half plane.
2. No a_i can be zero. Cancellation of terms in the cross-multiplication implies one LHP pole and one RHP pole.

If either of these two criteria is violated, it is immediately clear that the system is unstable. Otherwise, further analysis is required by the formation of the Routh's array.

Let's express Eq. (14) in the following form:

$$P(s) = a_n s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots + a_1 s + a_0 \tag{15}$$

To form the Routh's array, the highest order coefficient a_n, followed by every second coefficient is listed in the first row, labeled s^n. The rest of the coefficients are listed in the second row, labeled s^{n-1}. More rows are added all the way down to s^0 as illustrated.

$$\begin{array}{c|lll}
s^n & a_n & a_{n-2} & a_{n-4} \\
s^{n-1} & a_{n-1} & a_{n-3} & a_{n-5} \\
s^{n-2} & b_{n-1} = \dfrac{a_{n-1}a_{n-2} - a_n a_{n-3}}{a_{n-1}} & b_{n-3} = \dfrac{a_{n-1}a_{n-4} - a_n a_{n-5}}{a_{n-1}} & b_{n-5} = \dfrac{a_{n-1}a_{n-6} - a_n a_{n-7}}{a_{n-1}} \\
s^{n-3} & c_{n-1} = \dfrac{b_{n-1}a_{n-3} - a_{n-1}b_{n-3}}{b_{n-1}} & c_{n-3} = \dfrac{b_{n-1}a_{n-5} - a_{n-1}b_{n-5}}{b_{n-1}} & c_{n-5} = \dfrac{b_{n-1}a_{n-7} - a_{n-1}b_{n-7}}{b_{n-1}} \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
s^1 & x_{n-1} & x_{n-3} & \\
s^0 & a_0 & &
\end{array}$$

To simplify manual calculation of the table, it is useful to note that multiplication of any row by a positive constant will not affect the end result.

The number of sign changes in the first column gives the number of poles located on the right half plane.

As an example, take the polynomial

$$P(s) = 3s^3 + s^2 + 4s + 2 \tag{16}$$

The Routh array is constructed as follows:

s^3	3	4
s^2	1	2
s^1	−2	0
s^0	2	

There are two changes of sign in the first column, which means that there are two poles located on the RHP. We have

$$P(s) = 3(s + 0.476)(s - 0.0712 + 1.18j)(s - 0.0712 - 1.18j) \tag{17}$$

which confirms the fact that there are two poles on the right half plane.

Special Cases

In forming Routh's array, there are three special cases that need further consideration.

First Case. The first column of a row is zero, but the rest of the row is not entirely zero.

Take the following polynomial as an example.

$$P(s) = s^5 + s^4 + 2s^3 + 2s^2 + 3s + 4 \tag{18}$$

When the zero appears, replace it with a variable ϵ, to complete the table. Then take the limit as $\epsilon \to 0$, both from above and below, to determine if there are any sign changes.

				$\epsilon \to 0^+$	$\epsilon \to 0^-$
s^5	1	2	3	+	+
s^4	1	2	4	+	+
s^3	ϵ	−1		+	−
s^2	$2 + \frac{1}{\epsilon}$	4		+	−
s^1	$-1 - \frac{4\epsilon^2}{2\epsilon + 1}$	0		−	−
s^0	4			+	+

The table shows that there are two changes of sign in the first column, regardless of whether ϵ approaches zero from above or below in this case. Consequently, there are two roots in the right half plane.

The poles are located at $s_1 = 0.6672 \pm 1.1638j$, $s_2 = -0.5983 \pm 1.2632j$ and at $s_3 = -1.1377$, confirming the result.

Second Case. A whole row consists of zeros only.

When an entire row of zeros is encountered in row s^m, an auxiliary polynomial of order $m + 1$ is formed by using the s^{m+1} row as the coefficient and by skipping every second power of s. The row containing zeros is then replaced with the coefficients resulting from the derivative of the auxiliary polynomial.

Consider the following polynomial:

$$\begin{aligned} P(s) &= (s^2 + 4)(s + 1)(s + 3) \\ &= s^4 + 4s^3 + 7s^2 + 16s + 12 \end{aligned} \tag{19}$$

We proceed to build a Routh array as

s^4	1	7	12
s^3	4	16	0
s^2	3	12	
s^1	0	0	

The auxiliary polynomial is formed and differentiated.

$$\begin{aligned} A(s) &= 3s^2 + 12 \\ \therefore\ A'(s) &= 6s \end{aligned} \tag{20}$$

We may then replace the s^1 row and proceed.

s^4	1	7	12
s^3	4	16	0
s^2	3	12	
s^1	6	0	
s^0	12		

Because there are no changes in sign in the first column, there are no roots on the RHP. The presence of this row of zeros, however, indicates that the polynomial has an even polynomial as a factor. An even polynomial has only terms with even powers of s. It has the property that all its roots are symmetrical about both the real and the imaginary axis. Consequently, an even polynomial must have either roots in both the left and right half planes or only on the $j\omega$-axis. In this case, there are no right half plane roots, so they must be located on the $j\omega$-axis. In addition, the auxiliary polynomial $A(s)$ is the same even polynomial that caused the row of zeros, so we can tell that these roots are located at $s = \pm 2j$.

Third Case. There is a repeated root on the $j\omega$-axis.

The Routh–Hurwitz criterion indicates the existence of roots on the imaginary axis, but it does not indicate whether they are of multiplicity greater than one, which is essential knowledge if the distinction between marginal stability and instability is required.

Take the following polynomial as an example:

$$\begin{aligned} P(s) &= (s + 1)(s^2 + 4)^2 \\ &= s^5 + s^4 + 8s^3 + 8s^2 + 16s + 16 \end{aligned} \tag{21}$$

				Auxillary polynomial	Derivative	Final table entry		
s^5	1	8	16			1	8	16
s^4	1		16			1	8	16
s^3	0	8		$s^4 + 8s^2 + 16$	$4s^3 + 16s$	4	16	
s^2	4					4	16	
s^1	0	16		$4s^2 + 16$	$8s$	8		
s^0	16	16				16		

Even though none of the signs in the first column have changed sign, there are two roots located at $s = 2j\omega$ and two at $s = -2j\omega$. A system having $P(s)$ as a characteristic equation must be considered unstable, even though the Routh–Hurwitz algorithm did not predict it.

Routh developed this criterion in 1877. In 1893, Hurwitz, apparently unaware of Routh's work, developed a similar technique based on determinants, from which the Routh–Hurwitz criterion is derivable. In 1892, Lyapunov developed a more general technique that is applicable to both linear and nonlinear systems, called the direct method of Lyapunov.

NYQUIST CRITERION

Given the characteristic equation of a system, the Routh–Hurwitz criterion enables a system analyst to determine whether or not the system is stable without actually solving the roots of the characteristic equation. Unfortunately, the method still requires the characteristic equation, which may be somewhat cumbersome, to be derived. The Nyquist criterion offers a graphical method of solution based on the open-loop transfer function, thereby saving some algebraic manipulation. In addition, Nyquist's method is quite capable of handling pure time delays, which Routh's method and the root locus method can handle only clumsily, at best.

The Nyquist criterion is based on the following principal argument. Suppose a contour Γ_1 is traced arbitrarily in the s-plane as shown in Fig. 7(a). If each point s, comprising Γ_1 were to be transformed by a polynomial function of s, $L(s)$, then a new contour Γ_2 would result in the $L(s)$-plane, as illustrated in Fig. 7(b). Provided that Γ_1 does not pass through any poles or zeros of $L(s)$, the contour Γ_2 does not encircle the origin. The principal argument relates the number of times that the new contour Γ_2 encircles the origin to the number of poles and zeros of $L(s)$ encircled by Γ_1. In other words, Γ_2 encircles the origin by the difference between the number of poles and number of zeros in contour Γ_1

$$N = Z - P \tag{22}$$

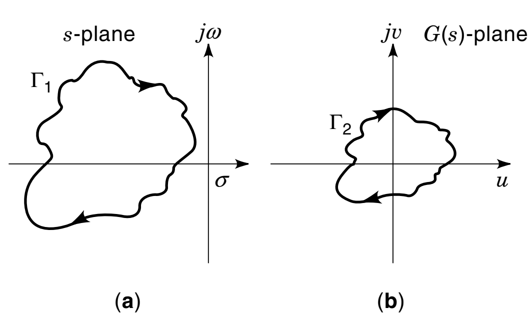

Figure 7. Contours in the s-plane and $G(s)H(s)$-plane. Every closed contour on the s-plane traces a closed contour on the $G(s)H(s)$-plane. If there are any poles or zeros (but not equal in numbers) of $G(s)H(s)$ in the contour in the s-plane, the contour in the $G(s)H(s)$ will encircle the origin at least once. If the number of poles of $G(s)H(s)$ inside the contour in the s-plane is greater than zero, the contour in the $G(s)H(s)$-plane goes in the opposite direction of the contour on the s-plane. If the zeros are greater than poles, the contours are in the same direction.

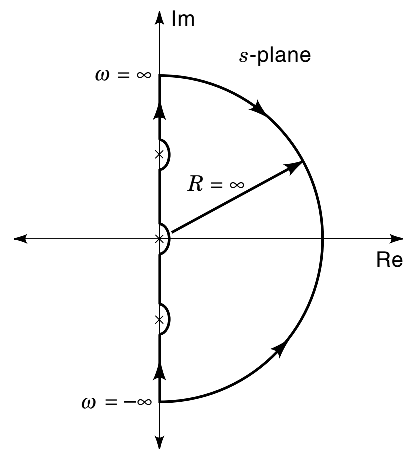

Figure 8. The Nyquist contour. If a contour is traced on the s-plane covering the entire RHP in the clockwise direction and if the number of zeros of $G(s)H(s)$ are greater than number of poles then the corresponding contour on the $G(s)H(s)$ will encircle the origin at least once in the same direction. The poles of $G(s)H(s)$ can usually be determined easily from the mathematical model. The number of zeros of $G(s)H(s)$ can be determined by the Nyquist plot.

where

N is the number of encirclements of the origin by Γ_2,
Z is the number of zeros of $L(s)$ encircled by Γ_1, and
P is the number of poles of $L(s)$ encircled by Γ_1.

A positive N indicates that Γ_2 and Γ_1 both travel in the same direction (i.e., clockwise or counterclockwise), whereas negative N indicates opposite directions. Some examples of contours Γ_2 and Γ_1 and encirclements are given in the section dedicated for Nyquist. Interested readers should refer to this section.

The Nyquist Contour

Consider the transfer function of a closed-loop system

$$\begin{aligned} \frac{C(s)}{R(s)} &= \frac{G(s)}{1 + G(s)H(s)} \\ &= \frac{G(s)}{1 + L(s)} \end{aligned} \tag{23}$$

From this equation, the following points should be clear.

1. The poles of $1 + L(s)$ are the poles of $L(s)$, the open-loop transfer function. This makes identification of the poles of $1 + L(s)$ possible simply by inspection in most cases.
2. Most importantly, the zeros of $1 + L(s)$ are the poles of $C(s)/R(s)$, the closed-loop transfer function.

The problem of determining the closed-loop pole locations may then be reduced to determining the zeros of $1 + L(s)$. To do this, the Nyquist contour, which covers the entire right half s-plane, as shown in Fig. 8, is used. It consists of a section along the $j\omega$-axis covering $-\infty < \omega < \infty$ and a semicircle described by $|s| \to \infty$, $-\pi/2 \le \arg(s) \le \pi/2$. The diagram also shows how the contour skips around simple (not repeated) poles of $L(s)$ located on the $j\omega$-axis to avoid discontinuities when the contour is mapped onto the $L(s)$ plane. The direction of the contour has been arbitrarily drawn as clockwise.

If the Nyquist contour is mapped onto the $1 + L(s)$-plane, we would find that the resulting contour would encircle the origin $N = Z - P$ times (in the clockwise sense). It should be emphasized that Z is the variable under investigation because it concerns the poles of the closed-loop system. The requirement for stability is that $1 + L(s)$ contain no zeros on the right half plane, or $Z = 0$. That is to say, if the (clockwise) Nyquist contour were mapped onto the $1 + L(s)$-plane, it is a requirement for closed loop stability that the resulting contour encircle the origin *counterclockwise* exactly the same number times as the number of poles of $L(s)$ in the RHP.

The contour shown in Fig. 8 skips to the right around $j\omega$-axis poles. Consequently, these $j\omega$-axis poles are not considered to be right-half-plane poles. It is perfectly feasible for the contour to skip to the left around these poles, in which case they should be included in the count of right-half-plane poles. It is emphasized that the poles of $L(s)$ are easily obtainable.

A further refinement of the Nyquist criterion is that it is unnecessary to plot the contour on the $1 + L(s)$-plane and observe the number of encirclements about the origin. The plot on the $L(s)$ plane is in fact identical to that of the $1 + L(s)$-plane, except that it is shifted left by one unit. It will therefore suffice to plot the contour on the $L(s)$-plane and observe the number of encirclements about the Cartesian point $(-1, 0)$.

Simplified Nyquist Plot

Suppose that the function $1 + L(s)$ contains P poles on the RHP, P' poles on the $j\omega$-axis, and Z zeros on the right half plane and that two contours are to be mapped onto the $L(s)$-plane.

1. The Nyquist contour skips to the right around the P' poles on the $j\omega$-axis. When mapped on the $L(s)$-plane, it is found to encircle the Cartesian point $(-1, 0)$ point N_1 times.
2. The Nyquist contour skips to the left around the P' poles on the $j\omega$-axis. When mapped on the $L(s)$-plane, it is found to encircle the Cartesian point $(-1, 0)$ point N_2 times.

Each contour may be considered to consist of three sections, each contributing a certain number of turns about the Cartesian point $(-1, 0)$.

1. The section consisting of $-\infty < \omega < \infty$, excluding the skips around $j\omega$-axis poles. Because of symmetry about the real axis in both the contour and the location of poles and zeros, this section may be divided into two halves—the positive imaginary axis and the negative imaginary axis, each contributing N_A turns.
2. The section consisting of the infinite semicircle, contributing N_B turns. If the order of the numerator of $L(s)$ is less than or equal to the denominator, then as $s \to \infty$, $L(s)$ corresponds to a point on the real axis or an encirclement of the origin. The contribution to the number of turns about the Cartesian point $(-1, 0)$ in either case is, therefore, $N_B = 0$.
3. The skips around the $j\omega$-axis poles. Because the two contours skip around these poles in opposite directions, if contour 1 were to contribute N_C turns, then contour 2 would contribute $-N_C$ turns.

Combining all these sections,

$$N_1 = 2N_A + N_B + N_C \tag{24}$$

$$N_2 = 2N_A + N_B - N_C \tag{25}$$

From the principal argument, it is also known that $N_1 = Z - P$; therefore,

$$N_2 = Z - P - P' \tag{26}$$

Eliminating N_1, N_2, and N_C and realizing that $N_B = 0$, we find that

$$N_A = \frac{2Z - 2P - P'}{4} \quad \text{or} \quad \phi = \left(Z - P - \frac{P'}{2}\right)\pi \tag{27}$$

where ϕ is the angle of rotation (in the clockwise sense) about the point $(-1, 0)$ when the line $\sigma = 0$, $\omega \geq 0$ is mapped onto the $L(s)$-plane, where $s = \sigma + j\omega$.

For stability, we require that $Z = 0$, from which the modified form of the Nyquist stability criterion may be expressed as

$$\phi = -\left(P + \frac{P'}{2}\right)\pi \tag{28}$$

That is to say, if the open-loop transfer function's frequency response is plotted on polar coordinates and is found to encircle the Cartesian point $(-1, 0)$ in a *counterclockwise* direction by an angle of exactly $\pi(P + P'/2)$ radians. In this case, where P is the number of open-loop transfer function poles on the right half plane and P' is the number of open-loop transfer function poles on the imaginary axis, the closed-loop system is stable.

For the illustration of Nyquist stability criterion, let's take an example of a system having a open-loop transfer function $G(s)H(s)$ of

$$\begin{aligned} G(s)H(s) &= L(s) \\ &= \frac{30}{(s+1)(s+2)(s+3)} \\ &= \frac{30}{s^3 + 6s^2 + 11s + 6} \end{aligned} \tag{29}$$

The Nyquist plot of $L(s)$ of Eq. (29) can be obtained in a number of ways (e.g., polar plots) by substituting $s \to j\omega$. By calculating the real and imaginary components of $L(j\omega)$, the Ny-

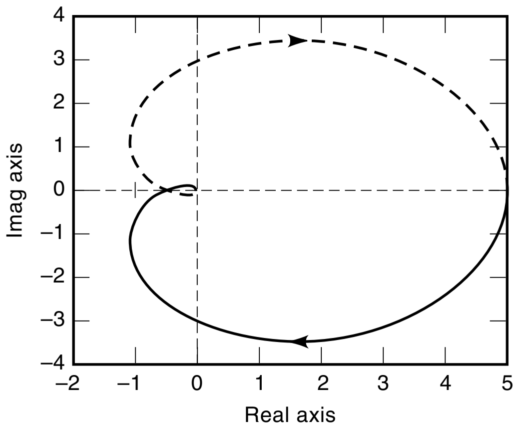

Figure 9. A typical Nyquist plot. This is the plot of a third-order system and hence it traces three quadrants. The curve cuts the real axis on the negative side. If the gain is increased sufficiently, the curve will encircle the −1 point hence indicating instability. This means that at least one of the roots of the characteristic equation, poles of the closed loop system, will be on the right half of the s-plane.

quist plot of Eq. (29) may be plotted as shown in Fig. 9. It can be seen that the contour does not encircle the point (−1, 0), so the system is stable.

Further examples of Nyquist plots are given in Fig. 10.

From the Nyquist plots, it is possible to find phase and gain margins of the system. The gain margin is defined to be the amount of gain that can be allowed before the system becomes unstable, and the phase margin is the angle at unity gain. It is also possible to find the phase crossover frequency ω_c and the gain crossover frequency ω_g either from the graph or mathematically. From the graph the phase and gain margins of the preceding example are 25° and 6 dB, respectively. It is also possible to design the system to obtain desired responses by varying the margins.

THE ROOT LOCUS

Often engineers want to see how changes in some parameters such as loop gain will affect the performance and the stability of a system. The root locus is a widely practiced method in this regard. It gives information about how the closed-loop poles of the system vary as the parameter in question is changed. This is particularly useful in determining the range the parameter may cover while keeping the system stable. As discussed previously, the relative stability of a system is largely determined by the location of poles, which the root locus approach clearly confirms.

Formulation of Root Locus

Figure 11 shows a block diagram of a system with a variable loop gain. It has a closed-loop transfer function given by

$$\frac{C(s)}{R(s)} = \frac{kG(s)}{1+kG(s)H(s)} \tag{30}$$

In order to investigate how $G(s)$ and $H(s)$ contribute poles and zeros to the closed loop system, it is informative to let $G(s) = N_G/D_G$ and $H(s) = N_H/D_H$. Equation (30) then reduces to

$$\frac{C(s)}{R(s)} = \frac{kN_G D_H}{D_G D_H + kN_G N_H} \tag{31}$$

Equation (31) reveals that the zeros of the closed-loop system are independent of k and correspond to the zeros of $G(s)$ and the poles of $H(s)$. However, as $k \to 0$, there is pole/zero cancellation of the $H(s)$ pole term D_H, and as $k \to \infty$ there is pole/zero cancellation of the $G(s)$ zero term N_G. The location of the closed-loop poles, or the roots of the characteristic equation, is the subject of the remaining discussion.

The root locus is a plot in the s-plane of the poles of the closed-loop transfer function as the parameter k varies from 0 to ∞. From Eq. (1), it should be clear that these poles correspond to the zeros of the $1 + kG(s)H(s)$ denominator. The root locus is therefore a plot in the s-plane of Eq. (31)

$$kG(s)H(s) = -1 \tag{32}$$

Equation (32) may be expressed in its polar form as

$$\frac{\prod_{i=1}^{u} A_i}{\prod_{i=1}^{v} B_i} = \frac{1}{k} \tag{33a}$$

$$\sum_{i=1}^{u} \theta_i - \sum_{i=1}^{v} \phi_i = \pi(1+2n) \tag{33b}$$

where A_i is the distance between a point on the root locus and the ith zero of the loop transfer function $L(s)$; B_i is the distance between a point on the root locus and the ith pole; θ_i is the angle about the ith zero from the positive real axis to a point on the root locus; ϕ_i is the angle about the ith pole from the positive real axis to a point on the root locus; u is the number of zeros in the loop transfer function; v is the number of poles; k is the loop gain; and n is any integer. It should be evident that Eq. (33b) determines whether a point is on the root locus, and Eq. (33a) just determines the value of k corresponding to that point.

It is always possible to solve Eq. (32) for an array of values for k, but that would be too time consuming. Evans developed a set of rules for sketching the root locus, reducing the problem to a few simple calculations.

Figure 12 shows the root locus of a typical system with an open-loop transfer function given by

$$\begin{aligned} T(s) &= \frac{k}{s(s+2)(s+5)} \\ &= \frac{k}{s^3+7s^2+10s} \end{aligned} \tag{34}$$

Further examples of root locus are given in Fig. 10.

The Root Locus Method of Evans

A number of rules may be applied to sketch the root locus.

- The root locus starts with $k = 0$ at the poles of $G(s)H(s)$ and finishes with $k \to \infty$ at the zeros of $G(s)H(s)$.

This can be seen from the magnitude condition

$$|G(s)H(s)| = \frac{1}{k} \tag{35}$$

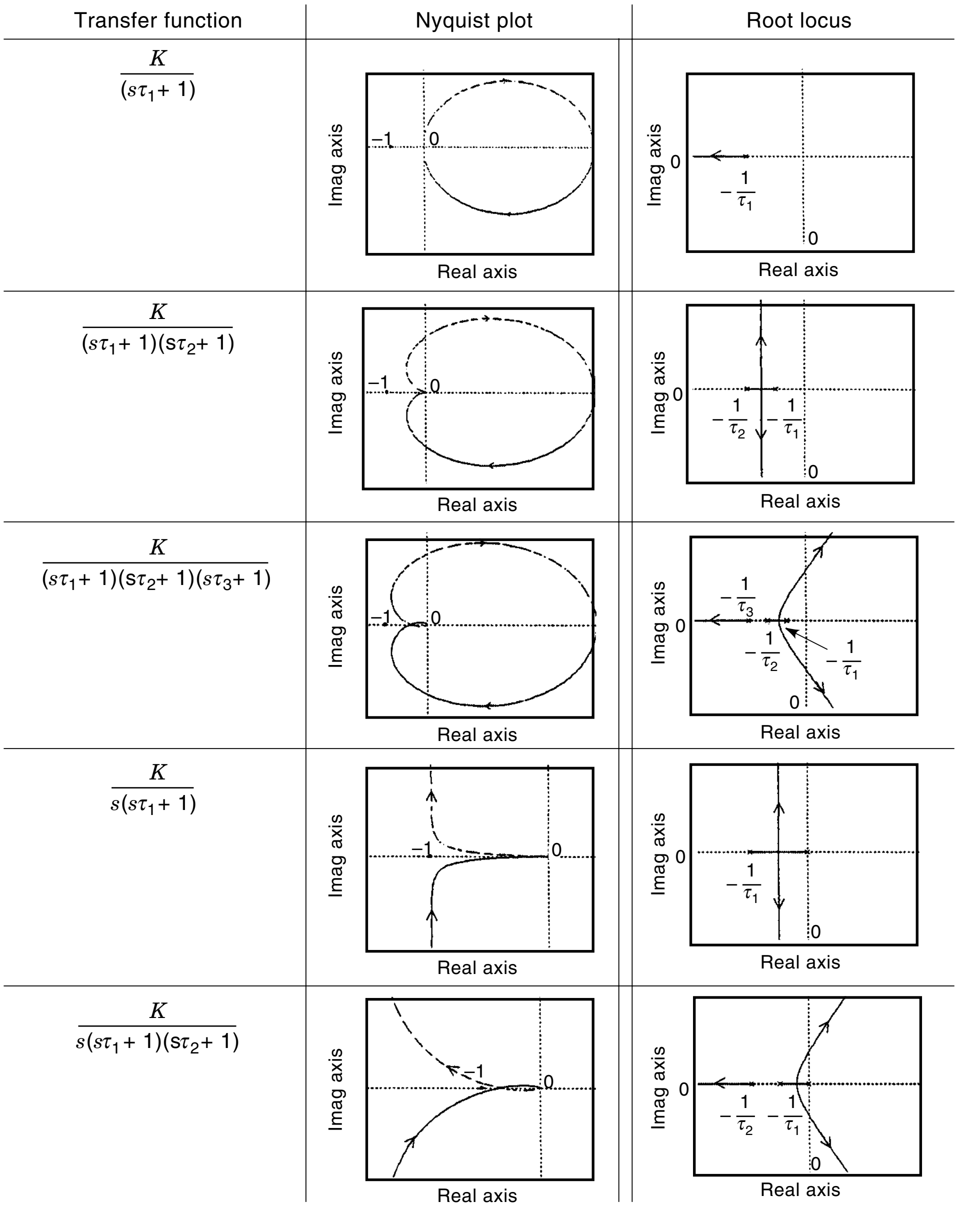

Figure 10. Examples of Nyquist and root locus plots. The stability of control systems can be determined by various methods as exemplified here. In obtaining these examples, a popular software called MATLAB was used.

As k approaches zero, the magnitude of the loop transfer function becomes infinite, corresponding to a pole. For k becoming infinitely large, the loop transfer function becomes infinitesimally small, corresponding to a zero.

Actually, inspection of Eq. (31) reveals that the poles of $H(s)$ and zeros of $G(s)$ never actually appear as poles of the closed-loop transfer function because of pole/zero cancellation for $k = 0$ and $k \to \infty$. This point should be kept in mind when designing systems with the root locus; for very high and very low gains, there may be significant pole/zero cancellation.

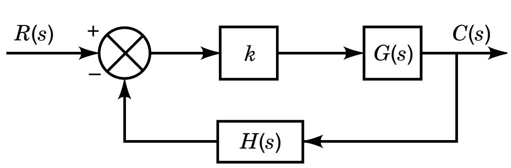

Figure 11. Block diagram of a closed-loop system with variable k. In many systems, one of the parameters of the system is varied to achieve the desired response. In this case, the variation of k in the forward path will relocate the roots of the characteristic equation on the s-plane. The suitable locations of the roots lead to appropriate system design.

The zeros of $G(s)H(s)$ include both the finite zeros found in the denominator terms and the infinite zeros at $|s| \to \infty$ caused by a denominator of higher order than the numerator. The result is that there are always the same number poles and zeros and that the root locus will always have enough zeros at which to terminate, be they finite or infinite.

- The root locus plot is symmetrical about the real axis. All physically realizable transfer functions have real coefficients. Transfer functions with real coefficients always produce complex poles and zeros in conjugate pairs, which means that if Eq. (32) locates a point $s = \sigma + j\omega$ as being on the root locus, then $s = \sigma - j\omega$ must also be on the root locus.

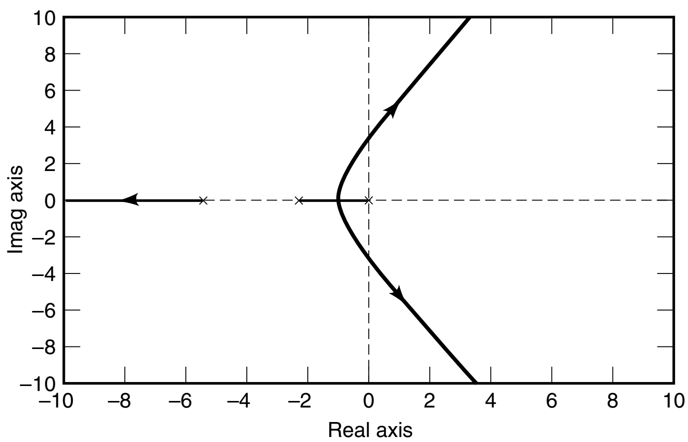

Figure 12. The root locus of a system with a characteristic equation = 1 + $ks(s + 2)(s + 5)$. This is a typical example of root locus. Roots start from poles of the characteristic equation when $k = 0$ and approaches zeros as $k \to \infty$. In this example, all three zeros are at ∞. At some value of k, the root loci crosses the imaginary axis to the RHP, thus indicating unstable conditions. With the aid of root locus a suitable value of k can be determined to locate the roots at the desired points on the s-plane.

- A point on the real axis is on the root locus if and only if there is an odd number of poles and zeros on the right-hand side of it. The angular contribution to Eq. (33b) of a pole or zero on the real axis to the left of a point on the real axis is always zero. In the case of a complex conjugate pair of poles or zeros, if one member of the pair contributes an angle of θ then the other will contribute $2\pi - \theta$. The total contribution from the pair is then $2\pi \equiv 0$ rad. Similarly, any pole or zero to the right of a point on the real axis will end up contributing π rad to the angular equation. Consequently, an odd number of poles or zeros is required to satisfy Eq. (33).
- Branches terminating at infinite zeros approach an asymptotic line that is described by

$$\psi_n = \frac{\pi(2n+1)}{v-u} \tag{36a}$$

$$\sigma_A = \frac{\sum_{i=1}^{v} p_i - \sum_{i=1}^{u} z_i}{v-u} \tag{36b}$$

  where ψ_n is the angle between the positive real axis and the nth asymptote; (σ_A, 0) is the point at which the asymptotes intersect the real axis; p_i is the ith open-loop transfer function pole location; z_i is the ith open-loop transfer function zero location; v is the number of open-loop transfer function poles; and u is the number of zeros.

Applying these rules will provide a reasonable sketch of the root locus. There are several significant points on the sketch that may be of interest to locate in terms of their precise location and the value of k required to achieve them.

- Breakaway points. These are the points at which multiple roots meet and then diverge. This most commonly occurs on the real axis, but it may occur anywhere, as shown in Fig. 13, plotted for $G(s)H(s) = (s + 1)(s + 3)(s^2 + 4s + 8)$. All breakaway points s must satisfy the following conditions:

$$\frac{d}{ds}G(s)H(s) = 0 \tag{37a}$$

$$\arg[G(s)H(s)] = \pi(1+2n) \tag{37b}$$

For real values of s, Eq. (37a) implies Eq. (37b), but for complex s, there is no such implication.

If there are n poles involved in a breakaway point, then there are always n branches entering and n branches leaving. The angle ψ between the entering and leaving branches is given by

$$\psi = \frac{\pi}{n} \tag{38}$$

- Angle of departure from a complex root. The angle at which a branch leaves a pole or arrives at a zero may be determined by assuming a point s infinitesimally close to the singularity. Because s is infinitesimally close to the pole or zero, the angular contributions to the angle Eq. (33b) from all the other poles and zeros are known, and the only unknown quantity is the contribution from the pole or zero in question. This angle, the angle of departure, is easily found by Eq. (33b).
- Imaginary axis intersection. This is a very important point to know because it reveals the value of k that will result in a marginally stable closed loop system. Forming a Routh table with the unknown k as a parameter and then solving for k to give a row of zeros in the table is one of the most common methods. For higher-order systems, the table may become too cumbersome. In such a situation, it may be more desirable to solve Eq. (32) as

$$G(j\omega)H(j\omega) = -\frac{1}{k} \tag{39}$$

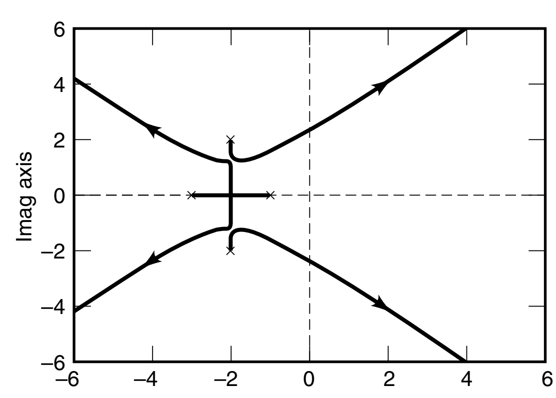

Figure 13. Breakaway points. The root locus plotted for $G(s)H(s) = (s + 1)(s + 3)(s^2 + 4s + 8)$ shows typical breakaway points at which multiple roots meet and then diverge. The breakaways generally occur on the real axis, but they may occur anywhere. In this example, a breakaway has happened on the real axis where as two others have taken place on the s-plane, in which the corners of the two root loci have met.

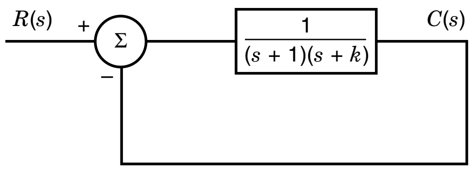

Figure 14. A representation of a variable component other than the root gain. In this case, one of the open-loop poles is the variable. This can be handled by forming an equivalent loop transfer function to construct the root locus.

In this case, the root locus is solved for an imaginary axis intercept, $s = j\omega$.

System Parameters Other Than Gain

In many situations the loop gain is not the parameter that is variable. It may be that the position of the open loop poles is the variable, as in Fig. 14. The way to handle this is to form an equivalent loop transfer function for the purpose of constructing the root locus as

$$\frac{C(s)}{R(s)} = \frac{1}{(s+1)(s+k)+1} \tag{40}$$

After some algebraic manipulation, Eq. (40) may be expressed in the form of

$$\begin{aligned}\frac{C(s)}{R(s)} &= \frac{\dfrac{1}{s^2+s+1}}{1+k\dfrac{s+1}{s^2+s+1}} \\ &= \frac{G(s)}{1+kG(s)H(s)}\end{aligned} \tag{41}$$

where $G(s) = 1/(s^2 + s + 1)$ and $H(s) = s + 1$. The root locus may now be constructed in the normal manner.

It may also occur that there are two parameters that are variable. Then the root locus may be represented by a set of contours or a surface plots.

THE RICCATI EQUATION

Many matrix equations naturally arise in linear control system theory. One of the most applied equations is the Riccati equation, which can be expressed as

$$\frac{dx}{dt} + P(t)x = Q(t)x^2 + R(t) \tag{42}$$

The equation was first developed and applied by Count Riccati and Jacopo Francesco in the early eighteenth century. In recent times, the Riccati equation finds wide application, particularly in the area of optimal control and filtering. In these applications, the matrix Riccati equation depicts a system of Riccati equations given by

$$\dot{\mathbf{X}}(t) + \mathbf{X}(t)\mathbf{A}(t) - \mathbf{D}(t)\mathbf{X}(t) - \mathbf{B}(t)\mathbf{X}(t) + \mathbf{C}(t) = 0 \tag{43}$$

Solution of the Riccati Equation

A closed form solution to Eq. (42) cannot be guaranteed depending of the functions $P(t)$, $Q(t)$, and $R(t)$. However, assuming that there is a known solution x_1, then the equation may be reduced to a first-order linear equation by letting

$$x = x_1 + \frac{1}{u} \tag{44}$$

Taking the derivative with respect to t results in

$$\frac{dx}{dt} = \frac{dx_1}{dt} - \frac{1}{u^2}\frac{du}{dt} \tag{45}$$

Substituting Eqs. (44) and (45) into Eq. (41) gives

$$\frac{dx_1}{dt} - \frac{1}{u^2}\frac{du}{dt} + P(t)\left(x_1 + \frac{1}{u}\right) = Q(t)\left(x_1^2 + \frac{2x_1}{u} + \frac{1}{u^2}\right) + R(t) \tag{46}$$

But since it is known that

$$\frac{dx_1}{dt} + P(t)x_1 = Q(t)x_1^2 + R(t) \tag{47}$$

Eq. (46) reduces to

$$\frac{du}{dt} + [2x_1(t)Q(t) - P(t)]u = -Q(t) \tag{48}$$

which is a linear first-order differential equation and is simple to solve particularly if $P(t)$, $Q(t)$, and $R(t)$ are constants, as would the case if Eq. (41) were to describe a time-invariant system.

The Matrix Riccati Differential Equation

Consider the dynamic system given by the state-space description as

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{F}\mathbf{x}(t) + \mathbf{G}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{H}\mathbf{x}(t)\end{aligned} \tag{49}$$

where $\mathbf{x}(t)$ is the state matrix (n by 1), $\mathbf{u}(t)$ is the control matrix (q by 1), and $\mathbf{y}(t)$ is the matrix (p by 1) of output variables to be controlled and $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ are matrices of appropriate dimensions that characterize the system.

In the optimal control applications, the object of optimal control is to find $\mathbf{u}(t)$ over an interval $t \in [t_1, t_2]$ such that some cost function is optimized. One of the popularly used cost function is the quadratic cost function that can be generalized by

$$F(t_1, t_2, T) = \int_{t_1}^{t_2} [\mathbf{y}'(t)\mathbf{y}(t) + \mathbf{u}'(t)\mathbf{u}(t)]\,dt + \mathbf{x}'(t_2)\mathbf{T}\mathbf{x}(t_2) \tag{50}$$

where $\mathbf{T}$ is a constant, real, symmetric ($\mathbf{T} = \mathbf{T}'$) and nonnegative definite ($|\mathbf{T}| \geq 0$) matrix.

It can be proven that there exists a unique optimal control for finite $t_2 - t_1 > 0$, which has the form of

$$\mathbf{u}(t) = -\mathbf{G}'\mathbf{P}(t_1, t_2, T)\mathbf{x}(t) \tag{51}$$

where $\mathbf{P}(t_1, t_2, T)$ can be described by

$$\dot{\mathbf{P}}(t) + \mathbf{P}(t)\mathbf{F} + \mathbf{F}'\mathbf{P}(t) - \mathbf{P}(t)\mathbf{G}\mathbf{G}'\mathbf{P}(t) + \mathbf{H}'\mathbf{H} = 0 \tag{52}$$

with the terminal condition that $\mathbf{P}(t_2) = \mathbf{T}$.

Correlation of Eq. (52) together with Eq. (43) reveals that the optimum control problem does indeed reduce to the problem of solving the Riccati matrix equation with constant, real **A**, **B**, and **C** matrices and **D** = **A**′. Furthermore, because any matrix **UU**′ is a symmetric, nonnegative definite matrix for **U** with real elements, it follows that **B** and **C** in are symmetric and nonnegative definite matrices, which are necessary conditions for solutions.

Solution of the Riccati Matrix Differential Equation

The following equation gives the form of the Riccati equation of interest applied in optimal control problems:

$$\dot{\mathbf{X}}(t) + \mathbf{X}(t)\mathbf{A} - \mathbf{A}'\mathbf{X}(t) - \mathbf{X}(t)\mathbf{B}\mathbf{X}(t) + \mathbf{C} = 0 \qquad (53)$$

with **B** and **C** matrices being symmetric and nonnegative definite.

The solution of Eq. (53) may be found by setting up the linear Hamiltonian matrix differential system as

$$\begin{bmatrix} \dot{\mathbf{U}}(t) \\ \dot{\mathbf{V}}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & -\mathbf{B} \\ -\mathbf{C} & -\mathbf{A}' \end{bmatrix} \begin{bmatrix} \mathbf{U}(t) \\ \mathbf{V}(t) \end{bmatrix} \qquad (54)$$

where $\mathbf{V}(t_2) = \mathbf{T}\mathbf{U}(t_2)$, and **T** is as defined in Eq. (50).

Then if the solution of Eq. (54) may be found by using Eq. (53)

$$\begin{bmatrix} \mathbf{U}(t) \\ \mathbf{V}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{21} & \mathbf{w}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{U}(t_2) \\ \mathbf{V}(t_2) \end{bmatrix} \qquad (55)$$

and $\mathbf{w}_{11} + \mathbf{w}_{12}\mathbf{T}$ is invertible, then the solution of Eq. (53) is given by

$$\mathbf{X}(t) = (\mathbf{w}_{21} + \mathbf{w}_{22}\mathbf{T})(\mathbf{w}_{11} + \mathbf{w}_{12}\mathbf{T})^{-1} \qquad (56)$$

As an example, consider

$$\frac{dx}{dt} + 2x - x^2 + 1 = 0 \qquad (57)$$

which corresponds to **A** = [1], **B** = [1], and **C** = [1] in Eq. (53).

The associated linear Hamiltonian matrix is then given by

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \qquad (58)$$

which has the solution

$$\begin{bmatrix} u \\ v \end{bmatrix} = a \begin{bmatrix} 1 \\ 2.414 \end{bmatrix} e^{-1.414t} + b \begin{bmatrix} 1 \\ -0.414 \end{bmatrix} e^{1.414t} \qquad (59)$$

For some arbitrary constants a and b, the solution becomes

$$\begin{aligned} u(0) &= a + b \\ v(0) &= 2.414a - 0.414b \end{aligned} \qquad (60)$$

Also, Eq. (59) may be expressed as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0.8536e^{1.414t} + 0.1464e^{-1.414t} & -0.3536e^{1.414t} + 0.3536e^{-1.414t} \\ -0.3534e^{1.414t} + 0.3534e^{-1.414t} & 0.1464e^{1.414t} + 0.8536e^{1.414t} \end{bmatrix} \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} \qquad (61)$$

which corresponds to Eq. (55). Applying Eq. (56), the final solution to Eq. (57) is found to be

$$x(t) = \frac{(-0.3534 + 0.1464T)e^{1.414t} + (0.3534 + 0.8536T)e^{1.414t}}{(0.8536 - 0.3536T)e^{1.414t} + (0.1464 + 0.3536Te^{-1.414t}} \qquad (62)$$

As expected, $x(t_2) = x(0) = T$.

Riccati Algebraic Equation and Infinite Horizon

In regulating control systems, it is not convenient to restrict control to a finite time period, $t \in [t_1, t_2]$. Even though it is possible to let t_1 approach negative infinity, it is customary to let t_2 approach infinity. Assuming that the system is time invariant [i.e., the matrices **A**, **B** and **C** in Eq. (53) are constant], the two intervals yield the same result.

Suppose that (**A**, **B**) are stabilizable. Then there must exist some stabilizing control that is not necessarily optimal. The resulting cost function associated with this stabilizing control will then dominate the optimal cost function, and it must be finite. Consequently, the solution of the Riccati equation with infinite horizon must be bounded.

The solution $X(t)$, as t approaches infinity, will either approach a constant value or become periodic, depending on the system and the value of **T** chosen. The value of $X(t)$ for $t \to \infty$ in the case where it is constant may be found by substituting $\dot{\mathbf{X}}(t) = 0$ into Eq. (53) to give the Riccati algebraic equation.

$$\mathbf{XA} + \mathbf{A}'\mathbf{X} - \mathbf{XBX} + \mathbf{C} = 0 \qquad (63)$$

If **A**, **B**, and **C** are 1 by 1 matrices (i.e., there are just scalar numbers), the solution of Eq. (63) is a trivial task. The general solution is a little more complex, involving Jordan chains.

A Jordan chain of the matrix **H** is a set of vectors, $x_1, x_2, \ldots, x_n$ such that

$$\begin{aligned} \mathbf{H}\mathbf{x}_1 &= \lambda \mathbf{x}_1 \\ \mathbf{H}\mathbf{x}_j &= \lambda \mathbf{x}_j + \mathbf{x}_{j-1} \end{aligned} \qquad (64)$$

where λ is an eigenvalue of **H** and $\mathbf{x}_1$ is the associated eigenvector.

Equation (63) has a solution for X if and only if there exists a set of Jordan chains of

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & -\mathbf{B} \\ -\mathbf{C} & -\mathbf{A}' \end{bmatrix}$$

given by $x_1, x_2, \ldots, x_n$. If we let

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix}$$

and

$$\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_n]$$

and

$$\mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_n]$$

then the solutions of Eq. (63) are given by $\mathbf{X} = \mathbf{V}\mathbf{U}^{-1}$.

We can verify this by using it to derive the quadratic formula. Let

$$ax^2 + bx + c = 0 \tag{65}$$

where $\mathbf{A} = -b/2$, $C = -c$, and $B = a$, to form the Hamiltonian matrix,

$$\mathbf{H} = \begin{bmatrix} -\frac{b}{2} & -a \\ c & \frac{b}{2} \end{bmatrix} \tag{66}$$

which has two eigenvalues given by

$$\begin{aligned} \lambda &= \pm\sqrt{\left(\frac{b}{2}\right)^2 - ac} \\ &= \pm\frac{\sqrt{b^2 - 4ac}}{2} \end{aligned} \tag{67}$$

The eigenvector associated with the eigenvalue of $\lambda \neq 0$ is

$$\mathbf{w} = \begin{bmatrix} 1 \\ \dfrac{-b}{2a} - \dfrac{\lambda}{a} \end{bmatrix} \tag{68}$$

The solution of Eq. (65) is then found as

$$\begin{aligned} x &= -\frac{b}{2a} - \frac{\lambda}{2a} \\ &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \end{aligned} \tag{69}$$

which is the familiar quadratic equation.

For $\lambda = 0$, there exists one Jordan chain,

$$\begin{aligned} x_1 &= \begin{bmatrix} 1 \\ -\dfrac{b}{2a} \end{bmatrix} \\ x_2 &= \begin{bmatrix} 0 \\ -\dfrac{1}{a} \end{bmatrix} \end{aligned} \tag{70}$$

giving one solution, $x = -b/2a$.

LYAPUNOV STABILITY

Lyapunov studied the question of motion, basing his theories on the nonlinear differential equations of motion. His equations for linear motion are equivalent to Routh's criterion. The Lyapunov's theorem determines the stability in the small region about an equilibrium point. The stability in the large may be determined by the global stability techniques. In this article, stability in the small will be introduced, and some references will be made to global stability methods.

Lyapunov considers the stability of general systems described by ordinary differential equations expressed in the state-variable form as

$$\dot{\mathbf{X}} = \mathbf{f}(\mathbf{x}) \tag{71}$$

Suppose that in a system all states have settled to constant values, hence achieved an equilibrium state. If the system is in the equilibrium state, that is, no states are varying in time, the equilibrium state may be described by

$$\dot{\mathbf{X}}_e = \mathbf{f}(\mathbf{x}_e) = 0 \tag{72}$$

In order to seek solutions to Eq. (71), Lyapunov introduced a continuously differentiable scalar function $V(x)$ with the following properties.

1. Positive definite if $V(0) = 0$ for all $t \geq t_0$ and $V(x) > 0$ for all $t \geq t_0$.
2. Positive semidefinite if $V(x) > 0$ for all x.
3. Negative definite or negative semidefinite if $-V(x)$ positive definite or positive semidefinite.

These conditions ensure that V is positive if any state is different from zero but equals zero when the state is zero. These conditions ensure that V is a smooth function and the trajectory does not expand indefinitely but rather is drawn to the origin. This can be explained with the aid of Fig. 15. Lyapunov stability states that an equilibrium state $\mathbf{x}_e$ of a dynamic system is stable if for every $\epsilon > 0$, there exists a $\sigma > 0$, where σ depends only on ϵ, such that $(\mathbf{x}_0 - \mathbf{x}_e) < \sigma$ results $(x(t;\, \mathbf{x}_0) - \mathbf{x}_e) < \epsilon$ for all $t > t_0$.

This statement of stability in the sense of Lyapunov indicates that if an equilibrium state is stable, the trajectory will remain within a given neighborhood of the equilibrium point if the initial state is close to the equilibrium point. Likewise, an equilibrium state $\mathbf{x}_e$ of a dynamic system is unstable if there exists an ϵ, such that a corresponding σ value cannot be found.

From the preceding explanations, asymptotic stability may be defined. An equilibrium state $\mathbf{x}_e$ of a dynamic system is asymptotically stable if

a. it is Lyapunov stable.
b. there is a number $\sigma_a > 0$ such that every motion starting within σ_a in the neighborhood of $\mathbf{x}_e$ converges to $\mathbf{x}_e$ as $t \to \infty$.

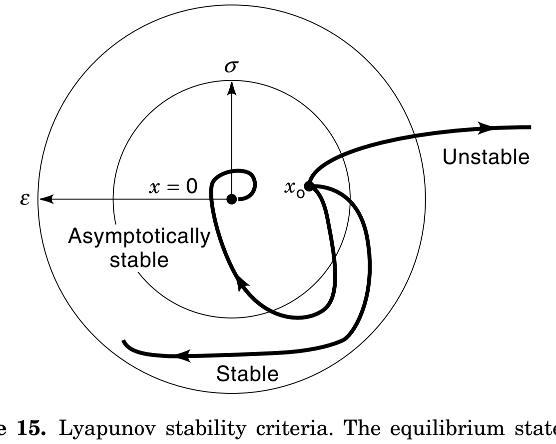

Figure 15. Lyapunov stability criteria. The equilibrium state $\mathbf{x}_e$ is stable if for every $\epsilon > 0$, there exists a $\sigma > 0$, where σ depends only on ϵ, such that the trajectory remains within a given neighborhood of the equilibrium point. Likewise, an equilibrium state $\mathbf{x}_e$ of a dynamic system is unstable if there exists an ϵ, such that a corresponding σ value cannot be found.

A simplified version of Lyapunov's first theorem of stability, Eq. (71), may be explained. Suppose that $\delta\dot{\mathbf{X}} = \mathbf{A}\delta\mathbf{x}$ is a valid model about the equilibrium point $\mathbf{x}_e$ and the roots of the characteristic equation may be expressed in matrix form as

$$s\mathbf{I} - \mathbf{A} = 0 \tag{73}$$

All eigenvalues of **A** have negative real parts if for any symmetric positive definite matrix **N**, the Lyapunov equation

$$\mathbf{A}^T\mathbf{M} + \mathbf{A}\mathbf{M} = -\mathbf{N} \tag{74}$$

has a symmetric positive definite solution.

For example, suppose that Eq. (71) can be expressed as

$$\dot{\mathbf{X}} = \mathbf{A}\mathbf{x}(t) \tag{75}$$

Its solution is $\mathbf{x}(t) = e^{At}\, x(0)$. If the eigenvalues of **A** are k_1, k_2, and k_3, then every component of $\mathbf{x}(t)$ is a linear combination of $e^{k1t,}$ e^{k2t}, and e^{k3t}. These time functions will approach zero as $t \to \infty$ if and only if k_i has negative real parts. Thus it can be concluded that any nonzero initial state will approach to zero only if **A** is stable. This can be generalized as follows.

1. If the characteristic values all have negative real parts, the equilibrium point is asymptotically stable.
2. If at least one of the values has a positive real part, the equilibrium point is unstable.
3. If one or more characteristic values have zero real parts, with all other values having negative real parts, the system stability cannot be determined with the current method.

Testing the stability by considering the linear part is referred to as Lyapunov's first or indirect method. Using the Lyapunov function directly on the nonlinear equations themselves is called the second or direct method. The argument is as follows. Lyapunov showed that in the case of nonlinear systems Eq. (71) may be extended as

$$\dot{\mathbf{X}} = \mathbf{F}\mathbf{x} + \mathbf{g}(\mathbf{x}) \tag{76}$$

where $\mathbf{g}(\mathbf{x})$ contains all the higher powers of **x**. If $\mathbf{g}(\mathbf{x})$ goes to zero faster than **x** does, then the system is stable if all the roots of **F** are strictly inside the left half plane and will be unstable if at least one root is in the right-half plane. For the system with roots in the left half plane and on the imaginary axis, the stability depends on the terms in the function **g**.

For global stability analysis of linear constant systems, quadratic functions are often used. Consider the function $\mathbf{V} = \mathbf{x}^T\,\mathbf{P}\mathbf{x}$, where **P** is a symmetric positive matrix. The **V** is the sum of squares of $\mathbf{x}_i$. In general, if **P** is positive, we can find a matrix **T** such that $\mathbf{P} = \mathbf{T}^T\mathbf{T}$ and $\mathbf{V} = \Sigma\mathbf{z}_i$, where $\mathbf{z} = \mathbf{T}\mathbf{x}$. For the derivative of V the chain rule can be used as

$$\begin{aligned} V &= d\mathbf{x}^T\mathbf{P}\mathbf{x}/\mathrm{dt} \\ &= \mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{x}^T\mathbf{P}\mathbf{x} \\ &= \mathbf{x}^T(\mathbf{F}^T\mathbf{P} + \mathbf{P}\mathbf{F})\mathbf{x} \\ &= -\mathbf{x}^T Q\mathbf{x} \end{aligned} \tag{77}$$

where

$$\mathbf{Q} = -(\mathbf{F}^T\mathbf{P} + \mathbf{P}\mathbf{F}) \tag{78}$$

For any positive **Q**, the solution of **P** of the Lyapunov equation is positive if and only if all the characteristic roots of **F** have negative real parts. That is, if a system matrix **F** is given, it is possible to select a positive **Q**, solve the Lyapunov equation in $n(n-1)/2$ unknowns, and test to see if **P** is positive by looking at the determinants of the n principal minors. From this, the stability may be determined from the equations without either solving them or finding the characteristic roots.

The study of nonlinear systems is vast, here only the basic principles of the methods have been discussed. Also, Lyapunov methods are applied in many diverse areas of control engineering; therefore, it is impossible to cover them all here. Interested readers should refer to the reading list given at the end of this article.

ROBUST STABILITY

Consider a linear time-invariant feedback system with a plant transfer function $G(s)$ and a compensator with $G_c(s)$ cascaded as shown in Fig. 16. In many applications, the plant model will not accurately represent the actual physical system because of (1) unmodeled dynamics and time delays, (2) changes in equilibrium points, (3) nonlinear characteristics of the plant, (4) noise and other disturbance inputs, and (5) parameter drift. The aim of a robust system is to assure that performance is maintained in spite of model inaccuracies and parameter changes.

The closed-loop transfer function of the system in Fig. 16 may be written as

$$M(s) = \frac{C(s)}{R(s)} = \frac{G(s)G_c(s)}{1 + G(s)G_c(s)} \tag{79}$$

The sensitivity of the system to changes in $G(s)$ or $G_c(s)$ can be expressed by the sensitivity function

$$S = \frac{\delta M/M}{\delta G/G} = \frac{1}{1 + G(s)G_c(s)} \tag{80}$$

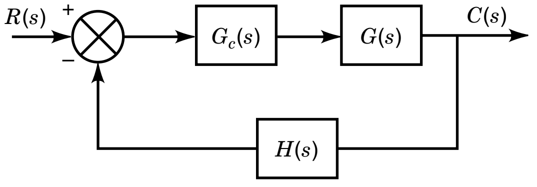

Figure 16. Block diagram of a closed-loop control system for robust stability analysis. In the many mathematical representation of systems, a full account of all affected parameters may not be taken into consideration because of unmodeled dynamics and time delays. Also, during the operations, the equilibrium points may change, parameters may drift, and noise and disturbances may become significant. The aim of a robust system is to assure that performance is maintained in spite of model inaccuracies and parameter changes.

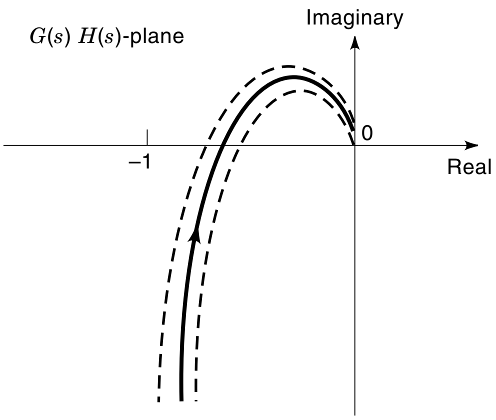

Figure 17. An example of a closed-loop system resulting from parameter drifts in Nyquist plots. This diagram indicates that because of uncertainties in modeling and changes in parameters the gain and phase margins may be altered. These alterations may lead to unstable conditions if these margins are close to critical values.

As can be seen from Eq. (80), the sensitivity function has the same characteristic equation $[1 + G(s)G_c(s)]$ as the closed-loop transfer function $M(s)$. For sensitivity to be small, it is necessary to have a high value for loop gain $L(s) = G(s)\ G_c(s)$. The high gain is obtained at high frequencies of $L(j\omega)$. But as we know, high gain could cause instability and poor responsiveness of $M(s)$. Now, the design problem becomes a matter of selecting $G_c(s)$ such that the closed-loop sensitivity is small, and the closed-loop transfer function has a wide bandwidth. At the same time, the desired gain and phase margins must be achieved.

The stability of the control system depends on the open-loop transfer function $L(s) = G(s)G_c(s)$. Because of the uncertainties outlined here, the transfer function may be written as

$$L(s) = G_c(s)[G(s) + \Delta G(s)] \tag{81}$$

For system stability, Nyquist's stability condition must always be satisfied. That is the -1 point must not be encircled by the $L(j\omega)$ under any circumstances. An example of uncertainty in a typical Nyquist plot resulting from $\Delta G(s)$ is illustrated in Fig. 17. To guarantee stability, a safe gain and phase margin must be ensured. Many methods are available to deal with the robustness of the system including classical methods linked to the use various compensators and PID controllers. The H_∞ technique is one method that finds extensive application in robust control design and analysis.

In general, feedback reduces the effect of disturbances and moderate modeling errors or parameter changes in the control system. In the presence of disturbances and sensor noises, systems are designed such that they keep the tracking errors and outputs small for disturbance inputs. In order to achieve this, the sensitivity to modeling errors and sensor noise must be made small, thus making the system robust. In this case, the plant output will follow any reference input asymptotically even if there are variations in the parameters of disturbance and noise. Briefly, it can be said that the system is more robust if it can tolerate larger perturbations in its parameters.

EXPONENTIAL STABILITY

The study of exponential signals (e^{at}) is important in linear system analysis. They contain a variety of signals such as constants, sinusoids, or exponentially decaying or increasing sinusoids.

A system with an n-dimensional state model is said to be an exponential system if its state-transition matrix $\Phi(t, \tau)$ can be written in matrix exponential form

$$\Phi(t,\ \tau) = e^{\Gamma(t,\ \tau)} \tag{82}$$

where $\Gamma(t, \tau)$ is an $n \times n$ matrix function of t and τ.

A sufficient condition for the system to be uniformly exponentially stable is that the eigenvalues of the of the $n \times n$ matrix $(1/t)\ \Gamma(t, \tau)$ be bounded as functions of t and have real parts $< -v$ for all $t > \tau$ and for some $v > 0$ and τ.

In many applications the stochastic components and random noises are included in the dynamical system models. The stochastic aspects of the model are used to capture the uncertainty about the environment in which the system is operating. The analysis and control of such systems involve evaluating the stability properties of the random dynamical systems. The stability of the system can be studied by stochastic stability approaches.

BIBLIOGRAPHY

1. W. S. Levine, *The Control Handbook,* Boca Raton, FL: CRC Press, 1996.
2. K. Ogata, *Modern Control Engineering,* 3rd ed., Upper Saddle River, NJ: Prentice-Hall, 1997.
3. B. J. Kuo, *Automatic Control Systems,* 6th ed., Englewood Cliffs, NJ: Prentice-Hall, 1991.
4. K. Watanabe, *Adaptive Estimation and Control: Partitioning Approach,* Hertfordshire, UK: Prentice-Hall, 1992.

HALIT EREN
ALEXANDER LOWE
Curtin University of Technology,
Australia