

Feedback Systems

An Introduction for Scientists and Engineers

SECOND EDITION

Karl Johan Åström
Richard M. Murray

Version v3.0j (2019-08-15)

This is the electronic edition of *Feedback Systems* and is available from <http://fbsbook.org>. Hardcover editions may be purchased from Princeton University Press, <http://press.princeton.edu/titles/8701.html>.

This manuscript is for personal use only and may not be reproduced, in whole or in part, without written consent from the publisher (see <http://press.princeton.edu/permissions.html>).

Chapter Ten

Frequency Domain Analysis

Mr. Black proposed a negative feedback repeater and proved by tests that it possessed the advantages which he had predicted for it. In particular, its gain was constant to a high degree, and it was linear enough so that spurious signals caused by the interaction of the various channels could be kept within permissible limits. For best results the feedback factor $\mu\beta$ had to be numerically much larger than unity. The possibility of stability with a feedback factor larger than unity was puzzling.

Harry Nyquist, “The Regeneration Theory,” 1956 [Nyq56].

In this chapter we study how the stability and robustness of closed loop systems can be determined by investigating how sinusoidal signals of different frequencies propagate around the feedback loop. This technique allows us to reason about the closed loop behavior of a system through the frequency domain properties of the *open loop* transfer function. The Nyquist stability theorem is a key result that provides a way to analyze stability and introduce measures of degrees of stability.

10.1 THE LOOP TRANSFER FUNCTION

Understanding how the behavior of a closed loop system is influenced by the properties of its open loop dynamics is tricky. Indeed, as the quote from Nyquist above illustrates, the behavior of feedback systems can often be puzzling. However, using the mathematical framework of transfer functions provides an elegant way to reason about such systems, which we call *loop analysis*.

The basic idea of loop analysis is to trace how a sinusoidal signal propagates in the feedback loop and explore the resulting stability by investigating if the propagated signal grows or decays. This is easy to do because the transmission of sinusoidal signals through a linear dynamical system is characterized by the frequency response of the system. The key result is the Nyquist stability theorem, which provides a great deal of insight regarding the stability of a system. Unlike proving stability with Lyapunov functions, studied in Chapter 5, the Nyquist criterion allows us to determine more than just whether a system is stable or unstable. It provides a measure of the degree of stability through the definition of stability margins. The Nyquist theorem also indicates how an unstable system should be changed to make it stable, which we shall study in detail in Chapters 11–13.

Consider the system in Figure 10.1a. The traditional way to determine if the closed loop system is stable is to investigate if the closed loop characteristic polynomial has all its roots in the left half-plane. If the process and the controller have rational transfer functions $P(s) = n_p(s)/d_p(s)$ and $C(s) = n_c(s)/d_c(s)$, then the

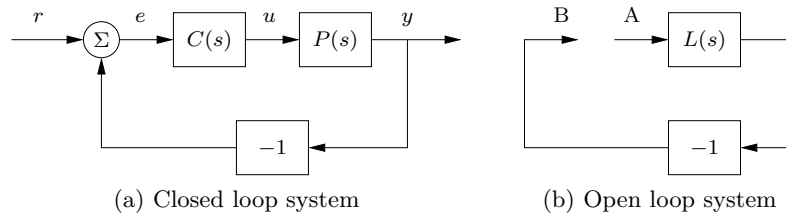


Figure 10.1: The loop transfer function. The stability of the feedback system (a) can be determined by tracing signals around the loop. Letting $L = PC$ represent the loop transfer function, we break the loop in (b) and ask whether a signal injected at the point A has the same magnitude and phase when it reaches point B.

closed loop system has the transfer function

$$G_{yr}(s) = \frac{PC}{1 + PC} = \frac{n_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)},$$

and the characteristic polynomial is

$$\lambda(s) = d_p(s)d_c(s) + n_p(s)n_c(s).$$

To check stability, we simply compute the roots of the characteristic polynomial and verify that they each have negative real part. This approach is straightforward but it gives little guidance for design: it is not easy to tell how the controller should be modified to make an unstable system stable.

Nyquist's idea was to first investigate conditions under which oscillations can occur in a feedback loop. To study this, we introduce the *loop transfer function* $L(s) = P(s)C(s)$, which is the transfer function obtained by breaking the feedback loop, as shown in Figure 10.1b. The loop transfer function is simply the transfer function from the input at position A to the output at position B multiplied by -1 (to account for the usual convention of negative feedback).

Assume that a sinusoid of frequency ω_0 is injected at point A. In steady state the signal at point B will also be a sinusoid with the frequency ω_0 . It seems reasonable that an oscillation can be maintained if the signal at B has the same amplitude and phase as the injected signal because we can then disconnect the injected signal and connect A to B. Tracing signals around the loop, we find that the signals at A and B are identical if there is a frequency ω_0 such that

$$L(i\omega_0) = -1, \tag{10.1}$$

which then provides a condition for maintaining an oscillation. The condition in equation (10.1) implies that the frequency response goes through the value -1 , which is called the *critical point*. Letting ω_c represent a frequency at which $\angle L(i\omega_c) = 180^\circ$, we can further reason that the system is stable if $|L(i\omega_c)| < 1$, since the signal at point B will have smaller amplitude than the injected signal. This is essentially true, but there are several subtleties that require a proper mathematical analysis, leading to Nyquist's stability criterion. Before discussing the details we give an example of calculating the loop transfer function.

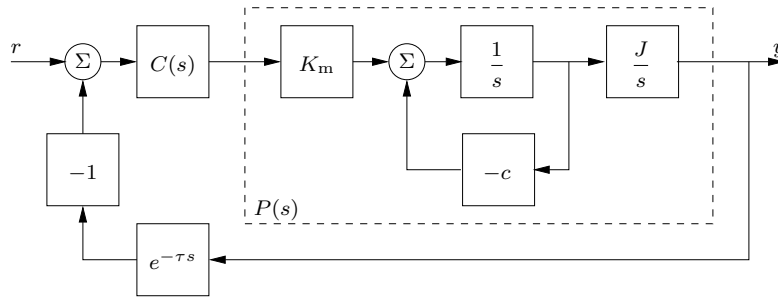


Figure 10.2: Block diagram of a DC motor control system with a short delay in the sensed position of the motor.

Example 10.1 Electric motor with proportional controller and delay

Consider a simple direct current electric motor with inertia J and damping (or back EMF) c . We wish to control the position of the motor using a feedback controller, and we consider the case where there is a small delay in the measurement of the motor position (a common case for controllers implemented on a computer with a fixed sampling rate). A block diagram for the motor with a controller $C(s)$ is shown in Figure 10.2. Using block diagram algebra, the process dynamics can be shown to be

$$P(s) = \frac{k_I}{Js^2 + cs}.$$

We now use a proportional controller of the form

$$C(s) = k_p.$$

The loop transfer function for the system control system is given by

$$L(s) = P(s)C(s)e^{-\tau s} = \frac{k_I k_p}{Js^2 + cs} e^{-\tau s},$$

where τ is the delay in sensing of the motor position. We wish to understand under which conditions the closed loop system is stable.

The condition for oscillation is given by equation (10.1), which requires that the phase of the loop transfer function must be 180° at some frequency ω_0 . Examining the loop transfer function we see that if $\tau = 0$ (no delay) then for s near 0 the phase of $L(s)$ will be 90° while for large s the phase of $L(s)$ will approach 180° . Since the gain of the system decreases as s increases, it is not possible for the condition of oscillation to be met in the case of no delay (the gain will always be less than 1 at arbitrarily high frequency).

When there is a small delay in the system, however, it is possible that we might get oscillations in the closed loop system. Suppose that ω_0 represents the frequency at which the magnitude of $L(i\omega)$ is equal to 1 (the specific value of ω_0 will depend on the parameters of the motor and the controller). Notice that the magnitude of the loop transfer function is not affected by the delay, but the phase increases as τ increases. In particular, if we let θ_0 be the phase of the undelayed system at frequency ω_0 , then a time delay of $\tau_c = (\pi + \theta_0)/\omega_0$ will cause $L(i\omega_0)$ to be equal to -1 . This means that as signals traverse the feedback loop, they can return in phase with the original signal and an oscillation may result.

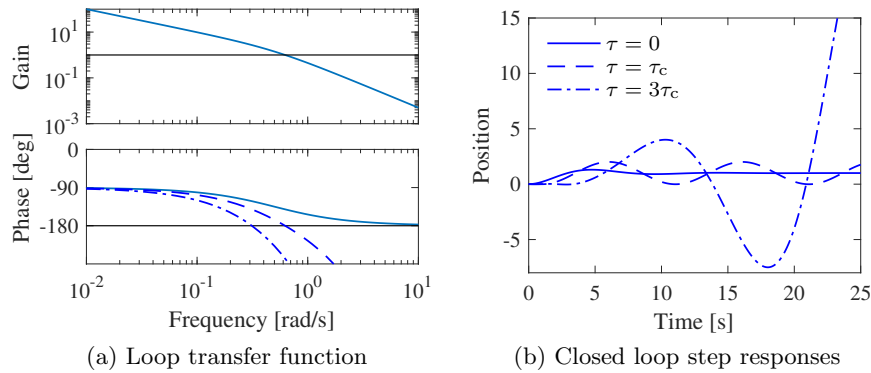


Figure 10.3: Loop transfer function and step response for the DC motor control system. The system parameters are $k_I = 1$, $J = 2$, $c = 1$ and the controller parameters are $k_p = 1$ and $\tau = 0, 1$, and 3 .

Figure 10.3 shows three controllers that result in stable, oscillatory, and unstable closed loop performance, depending on the amount of delay in the system. The instability is caused by the fact that the disturbance signals that propagate around the feedback loop can be in phase with the original disturbance due to the delay. If the gain around the loop is greater than or equal to one, this can lead to instability. ∇

One of the powerful concepts embedded in Nyquist's approach to stability analysis is that it allows us to study the stability of the feedback system by looking at properties of the loop transfer function $L = PC$. The advantage of doing this is that it is easy to see how the controller should be chosen to obtain a desired loop transfer function. For example, if we change the gain of the controller, the loop transfer function will be scaled accordingly and the critical point can be avoided. A simple way to stabilize an unstable system is thus to reduce the gain or to otherwise modify the controller so that the critical point -1 is avoided. Different ways to do this, called loop shaping, will be developed and discussed in Chapter 12.

10.2 THE NYQUIST CRITERION

In this section we present Nyquist's criterion for determining the stability of a feedback system through analysis of the loop transfer function. We begin by introducing a convenient graphical tool, the Nyquist plot, and show how it can be used to ascertain stability.

The Nyquist Plot

We saw in the previous chapter that the dynamics of a linear system can be represented by its frequency response and graphically illustrated by a Bode plot. To study the stability of a system, we will make use of a different representation of the frequency response called a *Nyquist plot*. The Nyquist plot of the loop transfer

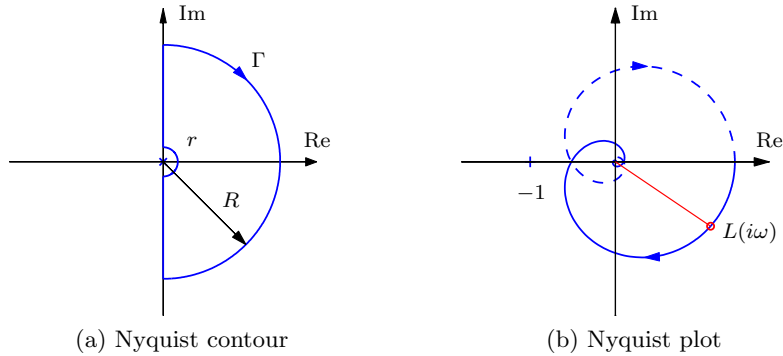


Figure 10.4: The Nyquist contour and the Nyquist plot. (a) The Nyquist contour Γ encloses the right half-plane, with a small semicircle around any poles of $L(s)$ at the origin or on the imaginary axis (illustrated here at the origin) and an arc whose radius R extends towards infinity. (b) The Nyquist plot is the image of the loop transfer function $L(s)$ when s traverses Γ in the clockwise direction. The solid curve corresponds to $\omega > 0$, and the dashed curve to $\omega < 0$. The gain and phase at the frequency ω are $g = |L(i\omega)|$ and $\varphi = \angle L(i\omega)$. The curve is generated for $L(s) = 1.4 e^{-s}/(s + 1)^2$.

function $L(s)$ is formed by tracing $s \in \mathbb{C}$ around the *Nyquist contour*, consisting of the imaginary axis combined with an arc at infinity connecting the endpoints of the imaginary axis. This contour, sometimes called the “Nyquist D contour” is denoted as $\Gamma \subset \mathbb{C}$ and is illustrated in Figure 10.4a. The image of $L(s)$ when s traverses Γ gives a closed curve in the complex plane and is referred to as the Nyquist plot for $L(s)$, as shown in Figure 10.4b. Note that if the transfer function $L(s)$ goes to zero as s gets large (the usual case), then the portion of the contour “at infinity” maps to the origin. Furthermore, the portion of the plot corresponding to $\omega < 0$, shown in dashed lines in Figure 10.4b, is the mirror image of the portion with $\omega > 0$.

There is a subtlety in the Nyquist plot when the loop transfer function has poles on the imaginary axis because the gain is infinite at the poles. To solve this problem, we modify the contour Γ to include small deviations that avoid any poles on the imaginary axis, as illustrated in Figure 10.4a (assuming a pole of $L(s)$ at the origin). The deviation consists of a small semicircle to the right of the imaginary axis pole location. Formally the contour Γ is defined as

$$\Gamma = \lim_{\substack{r \rightarrow 0 \\ R \rightarrow \infty}} (-iR, -ir) \cup \{re^{i\theta} : \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]\} \cup (ir, iR) \cup \{Re^{-i\theta} : \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]\} \tag{10.2}$$

for the case with a pole at the origin.

We now state the Nyquist criterion for the special case where the loop transfer function $L(s)$ has no poles in the right half-plane and no poles on the imaginary axis except possibly at the origin.

Theorem 10.1 (Simplified Nyquist criterion). *Let $L(s)$ be the loop transfer function for a negative feedback system (as shown in Figure 10.1a) and assume that L has no poles in the closed right half-plane ($\text{Re } s \geq 0$) except possibly at the origin. Then the closed loop system $G_{cl}(s) = L(s)/(1 + L(s))$ is stable if and only if the image of L along the closed contour Γ given by equation (10.2) has no net*

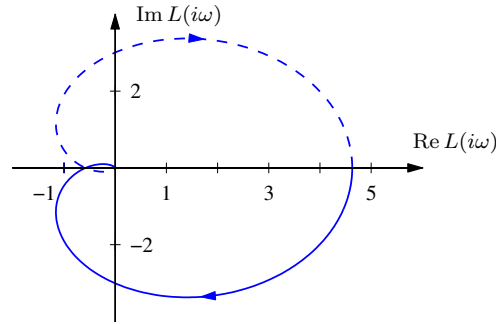


Figure 10.5: Nyquist plot for a third-order transfer function $L(s)$. The Nyquist plot consists of a trace of the loop transfer function $L(s) = 1/(s+a)^3$ with $a = 0.6$. The solid line represents the portion of the transfer function along the positive imaginary axis, and the dashed line the negative imaginary axis. The outer arc of the Nyquist contour Γ maps to the origin.

encirclements of the critical point $s = -1$.

The following conceptual procedure can be used to determine that there are no net encirclements. Fix a pin at the critical point $s = -1$, orthogonal to the plane. Attach a string with one end at the critical point and the other on the Nyquist plot. Let the end of the string attached to the Nyquist curve traverse the whole curve. There are no encirclements if the string does not wind up on the pin when the curve is encircled. The number of encirclements is called the *winding number*.

Example 10.2 Nyquist plot for a third-order system

Consider a third-order transfer function

$$L(s) = \frac{1}{(s+a)^3}.$$

To compute the Nyquist plot we start by evaluating points on the imaginary axis $s = i\omega$, which yields

$$L(i\omega) = \frac{1}{(i\omega + a)^3} = \frac{(a - i\omega)^3}{(a^2 + \omega^2)^3} = \frac{a^3 - 3a\omega^2}{(a^2 + \omega^2)^3} + i \frac{\omega^3 - 3a^2\omega}{(a^2 + \omega^2)^3}.$$

This is plotted in the complex plane in Figure 10.5, with the points corresponding to $\omega > 0$ drawn as a solid line and $\omega < 0$ as a dashed line. Notice that these curves are mirror images of each other.

To complete the Nyquist plot, we compute $L(s)$ for s on the outer arc of the Nyquist contour. This arc has the form $s = Re^{-i\theta}$ for $\theta \in [-\pi/2, \pi/2]$ and $R \rightarrow \infty$. This gives

$$L(Re^{-i\theta}) = \frac{1}{(Re^{-i\theta} + a)^3} \rightarrow 0 \quad \text{as } R \rightarrow \infty.$$

Thus the outer arc of the Nyquist contour Γ maps to the origin on the Nyquist plot. ∇

An alternative to computing the Nyquist plot explicitly is to determine the plot from the frequency response (Bode plot), which gives the Nyquist curve for $s = i\omega$,

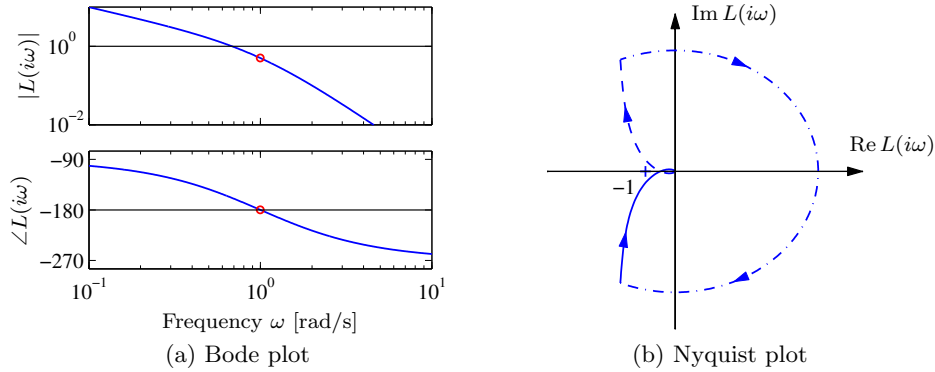


Figure 10.6: Sketching Nyquist and Bode plots. The loop transfer function is $L(s) = 1/(s(s + 1)^2)$. The frequency response (a) can be used to construct the Nyquist plot (b). The large semicircle is the map of the small semicircle of the Nyquist contour around the pole at the origin. The closed loop is stable because the Nyquist curve does not encircle the critical point. The point where the phase is -180° is marked with a circle in the Bode plot.

$\omega > 0$. We start by plotting $L(i\omega)$ from $\omega = 0$ to $\omega = \infty$, which can be read off from the magnitude and phase of the transfer function. We then plot $L(Re^{i\theta})$ with $\theta \in [\pi/2, 0]$ and $R \rightarrow \infty$, which goes to zero if the high-frequency gain of $L(i\omega)$ goes to zero (if and only if $L(s)$ is strictly proper). The remaining parts of the plot can be determined by taking the mirror image of the curve thus far (normally plotted using a dashed line). The plot can then be labeled with arrows corresponding to a clockwise traversal around the Nyquist contour (the same direction in which the first portion of the curve was plotted).

Example 10.3 Third-order system with a pole at the origin

Consider the transfer function

$$L(s) = \frac{k}{s(s + 1)^2},$$

where the gain has the nominal value $k = 1$. The Bode plot is shown in Figure 10.6a. The system has a single pole at $s = 0$ and a double pole at $s = -1$. The gain curve of the Bode plot thus has the slope -1 for low frequencies, and at the double pole $s = 1$ the slope changes to -3 . For small s we have $L \approx k/s$, which means that the low-frequency asymptote intersects the unit gain line at $\omega = k$. The phase curve starts at -90° for low frequencies, it is -180° at the breakpoint $\omega = 1$, and it is -270° at high frequencies.

Having obtained the Bode plot, we can now sketch the Nyquist plot, shown in Figure 10.6b. It starts with a phase of -90° for low frequencies, intersects the negative real axis at the breakpoint $\omega = 1$ where $L(i) = -0.5$ and goes to zero along the imaginary axis for high frequencies. The small half-circle of the Nyquist contour at the origin is mapped on a large circle enclosing the right half-plane. The Nyquist curve does not encircle the critical point $s = -1$, and it follows from the simplified Nyquist theorem that the closed loop system is stable. Since $L(i) = -k/2$, we find the closed loop system becomes unstable if the gain is increased to $k = 2$ or beyond. ∇

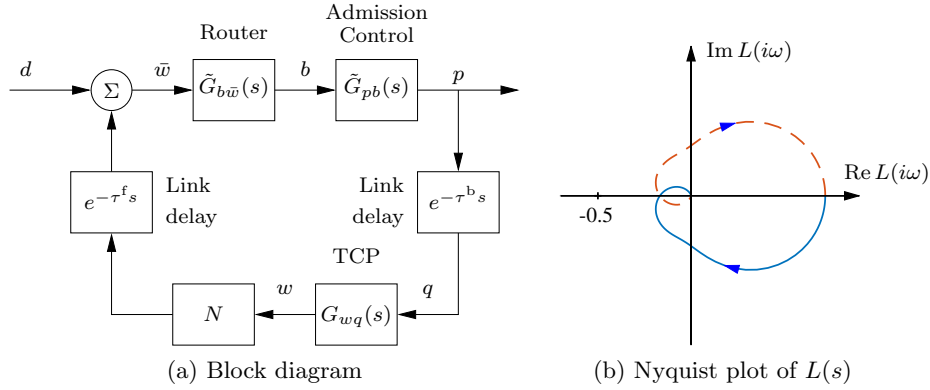


Figure 10.7: Internet congestion control. A set of N sources using TCP/Reno send messages through a single router with admission control (left). Link delays are included for the forward and backward directions. The Nyquist plot for the loop transfer function is shown on the right.

The Nyquist criterion does not require that $|L(i\omega_c)| < 1$ for all ω_c corresponding to a crossing of the negative real axis. Rather, it says that the number of encirclements must be zero, allowing for the possibility that the Nyquist curve could cross the negative real axis and cross back at magnitudes greater than 1. The fact that it was possible to have high feedback gains surprised the early designers of feedback amplifiers, as mentioned in the quote in the beginning of this chapter.

One advantage of the Nyquist criterion is that it tells us how a system is influenced by changes of the controller parameters. For example, it is very easy to visualize what happens when the gain is changed since this just scales the Nyquist curve.

Example 10.4 Congestion control

Consider the Internet congestion control system described in Section 4.4. Suppose we have N identical sources and a disturbance d representing an external data source, as shown in Figure 10.7a. We let w represent the individual window size for a source, q represent the end-to-end probability of a dropped packet, b represent the number of packets in the router's buffer, and p represent the probability that a packet is dropped by the router. We write \bar{w} for the total number of packets being received from all N sources. We also include forward and backward propagation delays between the router and the senders.

To analyze the stability of the system, we use the transfer functions computed in Exercise 9.9:

$$\tilde{G}_{b\bar{w}}(s) = \frac{1}{\tau_e^p s + e^{-\tau^f s}}, \quad G_{wq}(s) = -\frac{1}{q_e(\tau_e^p s + q_e w_e)}, \quad \tilde{G}_{pb}(s) = \rho,$$

where (w_e, b_e) is the equilibrium point for the system, N is the number of sources, τ_e^p is the steady-state round-trip time, and τ^f and τ^b are the forward and backward propagation times. We use $\tilde{G}_{b\bar{w}}$ and \tilde{G}_{qp} to represent the transfer functions with the forward and backward time delays removed since this is accounted for as a separate blocks in Figure 10.7a. Similarly, $G_{wq} = G_{\bar{w}q}/N$ since we have pulled out the multiplier N as a separate block as well.

The loop transfer function is given by

$$L(s) = \rho \cdot \frac{N}{\tau_e^p s + e^{-\tau^f s}} \cdot \frac{1}{q_e(\tau_e^p s + q_e w_e)} e^{-\tau_e^t s},$$

where $\tau^t = \tau^p + \tau^f + \tau^b$ is the total round trip delay time. Using the fact that $w_e = b_e/N = \tau_e^p c/N$ and $q_e = 2/(2 + w_e^2) \approx 2/w_e^2 = 2N^3/(\tau_e^p c)^2$ from equation (4.17), we can show that

$$L(s) = \rho \cdot \frac{N}{\tau_e^p s + e^{-\tau^f s}} \cdot \frac{c^3(\tau_e^p)^3}{2N^2(c(\tau_e^p)^2 s + 2N)} e^{-\tau_e^t s}.$$

Note that we have chosen the sign of $L(s)$ to use the same sign convention as in Figure 10.1b.

The Nyquist plot for the loop transfer function is shown in Figure 10.7b. To obtain an analytic stability criterion we can approximate the transfer function close to the intersection with the negative real axis, which occurs at the “phase crossover” frequency ω_{pc} . The second factor is stable if $\tau_e^p > \tau^f$ and has fast dynamics, so we approximate it by its zero frequency gain N . The third factor has slow dynamics (it can be shown that $2N \ll c(\tau_e^p)^2 \omega_{pc}$), and we can approximate it by an integrator. We thus obtain the following approximation of the loop transfer function around the frequency ω_{pc} :

$$L(s) \approx \rho \cdot N \cdot \frac{c^3(\tau_e^p)^3}{2N^2 c(\tau_e^p)^2 s} e^{-\tau_e^t s} = \frac{\rho c^2 \tau_e^p}{2Ns} e^{-\tau_e^t s}.$$

The integrator has a phase lag of $\pi/2$ and the transfer function $L(s)$ has the phase crossover frequency $\omega_{pc} = \pi/(2\tau_e^p)$. A necessary condition for stability is thus $|L(i\omega_{pc})| < 1$, which gives the condition

$$\frac{\rho c^2 (\tau_e^p)^2}{\pi N} < 1.$$

Using the Nyquist criterion, the closed loop system will be unstable if this quantity is greater than 1. In particular, for a fixed processing time τ_e^p , the system will become unstable as the link capacity c . This indicates that the TCP protocol may not be scalable to high-capacity networks, as pointed out by Low *et al.* [LPD02]. Exercise 10.9 provides some ideas of how this might be overcome. ∇

The General Nyquist Criterion



Theorem 10.1 requires that $L(s)$ has no poles in the closed right half-plane, except possibly at the origin. In some situations this is not the case and we need a more general result. This requires some results from the theory of complex variables, for which the reader can consult Ahlfors [Ahl66]. Since some precision is needed in stating Nyquist’s criterion properly, we will use a more mathematical style of presentation. We also follow the mathematical convention of counting encirclements in the counterclockwise direction for the remainder of this section. The key result is the following theorem about functions of complex variables.

Theorem 10.2 (Principle of variation of the argument). *Let Γ be a closed contour in the complex plane and let D represent the interior of Γ . Assume the function*

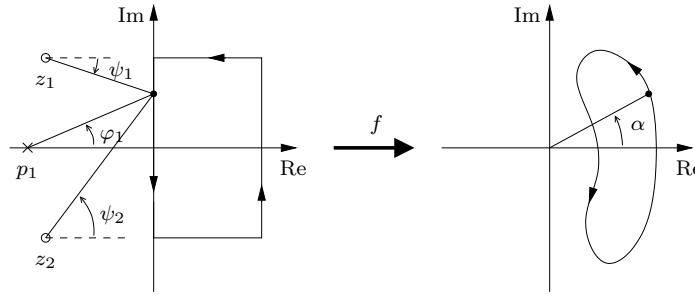


Figure 10.8: Graphical proof of the principle of the variation of the argument.

$f : \mathbb{C} \rightarrow \mathbb{C}$ is analytic on Γ and D except at a finite number of poles and zeros in D . Then the winding number $n_{w,\Gamma}(f(s))$ of the function $f(s)$ as s traverses the contour Γ in the counterclockwise direction is given by

$$n_{w,\Gamma}(f(s)) = \frac{1}{2\pi} \Delta \arg_{\Gamma} f(s) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f'(s)}{f(s)} ds = n_{z,D} - n_{p,D},$$

where $\Delta \arg_{\Gamma}$ is the net variation in the angle when s traverses the contour Γ in the counterclockwise direction, $n_{z,D}$ is the number of zeros of $f(s)$ in D , and $n_{p,D}$ is the number of poles of $f(s)$ in D . Poles and zeros of multiplicity m are counted m times.

To understand why the principle of variation of the argument is true, we keep track of how the argument (angle) of a function varies as we traverse a closed contour. Figure 10.8 illustrates the basic idea. Consider a function $f : \mathbb{C} \rightarrow \mathbb{C}$ of the form

$$f(s) = \frac{(s - z_1) \cdots (s - z_m)}{(s - p_1) \cdots (s - p_n)}, \quad (10.3)$$

where z_i are zeros and p_i are poles. We can rewrite the factors in this function by keeping track of the distance and angle to each pole and zero:

$$f(s) = \frac{r_1 e^{i\psi_1} \cdots r_m e^{i\psi_m}}{\rho_1 e^{i\theta_1} \cdots \rho_n e^{i\theta_n}}.$$

The argument (angle) of $f(s)$ at any given value of s can be computed by adding the contributions for the zeros and subtracting the contributions from the poles,

$$\arg(f(s)) = \sum_{i=1}^m \psi_i - \sum_{i=1}^n \theta_i.$$

We now consider what happens if we traverse a closed loop contour Γ . If all of the poles and zeros in $f(s)$ are outside of the contour, then the net contribution to the angle from terms in the numerator and denominator will be zero since there is no way for the angle to “accumulate.” Thus the contribution from each individual zero and pole will integrate to zero as we traverse the contour. If, however, the zero or pole is inside the contour Γ , then the net change in angle as we transverse the contour will be 2π for terms in the numerator (zeros) or -2π for terms in the denominator (poles). Thus the net change in the angle as we traverse the contour

is given by $2\pi(n_{z,D} - n_{p,D})$, where $n_{z,D}$ is the number of zeros inside the contour and $n_{p,D}$ is the number of poles inside the contour.

Formal proof. Assume that $s = a$ is a zero of multiplicity m . In the neighborhood of $s = a$ we have

$$f(s) = (s - a)^m g(s),$$

where the function g is analytic and different from zero. The ratio of the derivative of f to itself is then given by

$$\frac{f'(s)}{f(s)} = \frac{m}{s - a} + \frac{g'(s)}{g(s)},$$

and the second term is analytic at $s = a$. The function f'/f thus has a single pole at $s = a$ with the residue m . The sum of the residues at the zeros of this function is $n_{z,D}$. Similarly, we find that the sum of the residues for the poles is $-n_{p,D}$, and hence

$$n_{z,D} - n_{p,D} = \frac{1}{2\pi i} \int_{\Gamma} \frac{f'(s)}{f(s)} ds = \frac{1}{2\pi i} \int_{\Gamma} \frac{d}{ds} \log f(s) ds = \frac{1}{2\pi i} \Delta \arg_{\Gamma} \log f(s),$$

where $\Delta \arg_{\Gamma}$ again denotes the variation along the contour Γ . We have

$$\log f(s) = \log |f(s)| + i \arg f(s),$$

and since the variation of $|f(s)|$ around a closed contour is zero it follows that

$$\Delta \arg_{\Gamma} \log f(s) = i \Delta \arg_{\Gamma} \arg f(s),$$

and the theorem is proved. □

This theorem is useful in determining the number of poles and zeros of a function of a complex variable in a given region. By choosing an appropriate closed region D with boundary Γ , we can determine the difference between the number of zeros and poles through computation of the winding number.

Theorem 10.2 can be used to obtain a general version of Nyquist's stability theorem by choosing Γ as the Nyquist contour shown in Figure 10.4a, which encloses the right half-plane. To construct the contour, we start with part of the imaginary axis $-iR \leq s \leq iR$ and a semicircle to the right with radius R . If the function f has poles on the imaginary axis, we introduce small semicircles with radii r to the right of the poles as shown in the figure to avoid crossing through a singularity. The Nyquist contour is obtained by selecting R large enough and r small enough so that all open-loop right half-plane poles are enclosed. Ⓢ

Note that Γ has orientation *opposite* that shown in Figure 10.4a. The convention in engineering is to traverse the Nyquist contour in the clockwise direction since this corresponds to increasing frequency moving upwards along the imaginary axis, which makes it easy to sketch the Nyquist contour from a Bode plot. In mathematics it is customary to define the winding number for a curve with respect to a point so that it is positive when the contour is traversed counterclockwise. This difference does not matter as long as we use the same convention for orientation when traversing the Nyquist contour and computing the winding number.

To use the principle of variation of the argument (Theorem 10.2) to obtain an improved stability criterion we apply it to the function $f(s) = 1 + L(s)$, where $L(s)$ is the loop transfer function of a closed loop system with negative feedback. The generalized Nyquist criterion is given by the following theorem.

Theorem 10.3 (General Nyquist criterion). *Consider a closed loop system with loop transfer function $L(s)$ that has $n_{p,\text{rhp}}$ poles in the region enclosed by the Nyquist contour Γ . Let $n_{w,\Gamma}(1 + L(s))$ be the winding number of $f(s) = 1 + L(s)$ when s traverses Γ in the counterclockwise direction. Assume that $1 + L(i\omega) \neq 0$ for all ω on Γ and that $n_{w,\Gamma}(1 + L(s)) + n_{p,\text{rhp}} = 0$. Then the closed loop system has no poles in the closed right half-plane and it is thus stable.*

Proof. The proof follows directly from the principle of variation of the argument, Theorem 10.2. The closed loop poles of the system are the zeros of the function $f(s) = 1 + L(s)$. It follows from the assumptions that the function $f(s)$ has no zeros on the contour Γ . To find the zeros in the right half-plane, we investigate the winding number of the function $f(s) = 1 + L(s)$ as s moves along the Nyquist contour Γ in the *counterclockwise* direction. The winding number n_w can be determined from the Nyquist plot. A direct application of Theorem 10.2 shows that since $n_{w,\Gamma}(1 + L(s)) + n_{p,\text{rhp}}(L(s)) = 0$, then $f(s)$ has no zeros in the right half-plane. Since the image of $1 + L(s)$ is a shifted version of $L(s)$, we usually express the Nyquist criterion as net encirclements of the -1 point by the image of $L(s)$. \square

The condition that $1 + L(i\omega) \neq 0$ on Γ implies that the Nyquist curve does not go through the critical point -1 for any frequency. The condition that $n_{w,\Gamma}(1 + L(s)) + n_{p,\text{rhp}}(L(s)) = 0$, which is called *the winding number condition*, implies that the Nyquist curve encircles the critical point as many times as the loop transfer function $L(s)$ has poles in the right half-plane.

As noted above, in practice the Nyquist criterion is most often applied by traversing the Nyquist contour in the *clockwise* direction, since this corresponds to tracing out the Nyquist curve from $\omega = 0$ to ∞ , which can be read off from the Bode plot. In this case, the number of net encirclements of the -1 point must also be counted in the *clockwise* direction. If we let P be the number of unstable poles in the loop transfer function, N be the number of clockwise encirclements of the point -1 , and Z be the number of unstable stable zeros of $1 + L$ (and hence the number of unstable poles of the closed loop) then the following relation holds:

$$Z = N + P.$$

Note also that when using small semicircles of radius r to avoid poles on the imaginary axis these will generate a section of the Nyquist curve with large magnitude, requiring care in computing the winding number.

Example 10.5 Stabilized inverted pendulum

The linearized dynamics of a normalized inverted pendulum can be represented by the transfer function $P(s) = 1/(s^2 - 1)$, where the input is acceleration of the pivot and the output is the pendulum angle θ , as shown in Figure 10.9 (Exercise 9.5). We attempt to stabilize the pendulum with a proportional-derivative (PD) controller having the transfer function $C(s) = k(s + 2)$. The loop transfer function is

$$L(s) = \frac{k(s + 2)}{s^2 - 1}.$$

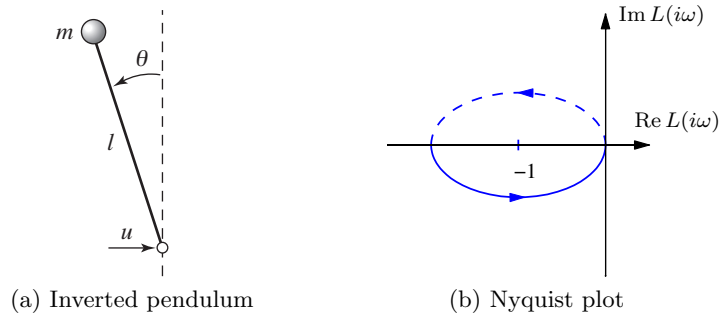


Figure 10.9: PD control of an inverted pendulum. (a) The system consists of a mass that is balanced by applying a force at the pivot point. A proportional-derivative controller with transfer function $C(s) = k(s + 2)$ is used to command u based on θ . (b) A Nyquist plot of the loop transfer function for gain $k = 1$. There is one counterclockwise encirclement of the critical point, giving $N = -1$ clockwise encirclements.

The Nyquist plot of the loop transfer function is shown in Figure 10.9b. We have $L(0) = -2k$ and $L(\infty) = 0$. If $k > 0.5$, the Nyquist curve encircles the critical point $s = -1$ in the counterclockwise direction when the Nyquist contour γ is encircled in the clockwise direction. The number of encirclements is thus $N = -1$. Since the loop transfer function has one pole in the right half-plane ($P = 1$), we find that $Z = N + P = 0$ and the system is thus stable for $k > 0.5$. If $k < 0.5$, there is no encirclement and the closed loop will have one pole in the right half-plane. Notice that the system is unstable for small gains but stable for high gains. ∇

Conditional Stability

An unstable system can often be stabilized simply by reducing the loop gain. However, as Example 10.5 illustrates, there are situations where a system can be stabilized by *increasing* the gain. This was first encountered by electrical engineers in the design of feedback amplifiers, who coined the term *conditional stability*. The problem was actually a strong motivation for Nyquist to develop his theory. The following example further illustrates this concept.

Example 10.6 Conditional stability for a third-order system

Consider a feedback system with the loop transfer function

$$L(s) = \frac{3k(s + 6)^2}{s(s + 1)^2}. \tag{10.4}$$

The Nyquist plot of the loop transfer function is shown in Figure 10.10 for $k = 1$. Notice that the Nyquist curve intersects the negative real axis twice. The first intersection occurs at $L = -12$ for $\omega = 2$, and the second at $L = -4.5$ for $\omega = 3$. The intuitive argument based on signal tracing around the loop in Figure 10.1b is misleading in this case. Injection of a sinusoid with frequency 2 rad/s and amplitude 1 at A gives, in steady state, an oscillation at B that is in phase with the input and has amplitude 12. Intuitively it seems unlikely that closing of the loop will result in a stable system. Evaluating the winding number for the Nyquist plot in

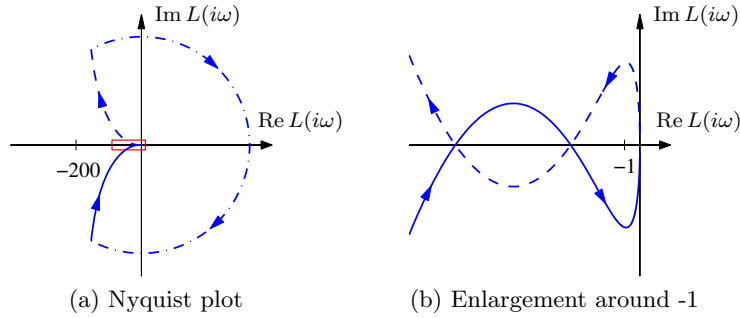


Figure 10.10: Nyquist curve for the loop transfer function $L(s) = (3(s + 6)^2)/(s(s + 1)^2)$. The plot on the right is an enlargement of the box around the origin of the plot on the left. The Nyquist curve intersects the negative real axis twice but has no net encirclements of -1 .

Figure 10.10 shows that the winding number is zero and the system is thus by using the version of Nyquist's stability criterion in Theorem 10.3. The closed loop system is stable for any $k > 2/9$. It becomes unstable if the gain is reduced to $1/12 < k < 2/9$, and it will be stable again for gains less than $1/12$. ∇

10.3 STABILITY MARGINS

In practice it is not enough that a system is stable. There must also be some margins of stability that describe how far from instability the system is and its robustness to perturbations. Stability is captured by Nyquist's criterion, which says that the loop transfer $L(s)$ function should avoid the critical point -1 , while satisfying a winding number condition. Stability margins express how well the Nyquist curve of the loop transfer avoids the critical point. The shortest distance s_m of the Nyquist curve to the critical point is a natural criterion, which is called the *stability margin*. It is illustrated in Figure 10.11a, where we have plotted the portion of the curve corresponding to $\omega > 0$. A stability margin s_m means that the Nyquist curve of the loop transfer function is outside a circle around the critical point with radius s_m .

Other margins are based the influence of the controller on the Nyquist curve. An increase in controller gain expands the Nyquist plot radially. An increase in the phase of the controller turns the Nyquist plot clockwise. Hence from the Nyquist plot we can easily pick off the amount of gain or phase that can be added without causing the system to become unstable.

The *gain margin* g_m of a closed-loop system is defined as the smallest multiplier of the loop gain that makes the system unstable. It is also the inverse of the distance between the origin and the point between -1 and 0 where the loop transfer function crosses the negative real axis. If there are several crossings the gain margin is defined by the intersection that is closest to the critical point. Let this point be $L(i\omega_{pc})$, where ω_{pc} represent this frequency, called the *phase crossover frequency*. The gain

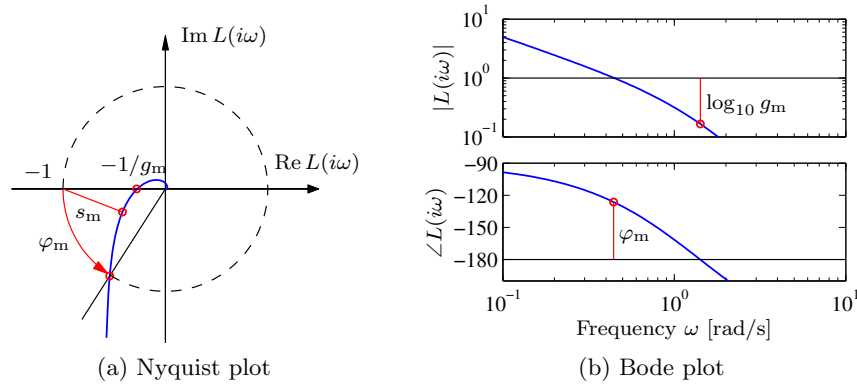


Figure 10.11: Stability margins for a third-order loop transfer function $L(s)$. The Nyquist plot (a) shows the stability margin, s_m , the gain margin g_m , and the phase margin φ_m . The stability margin s_m is the shortest distance to the critical point -1 . The gain margin corresponds to the smallest increase in gain that creates an encirclement, and the phase margin is the smallest change in phase that creates an encirclement. The Bode plot (b) shows the gain and phase margins.

margin for the system is then

$$g_m = \frac{1}{|L(i\omega_{pc})|}. \tag{10.5}$$

This number can be obtained directly from the Nyquist plots as shown in Figure 10.11a.

The *phase margin* is the amount of phase lag required to reach the stability limit. Let ω_{gc} be the *gain crossover frequency*, the frequency where the loop transfer function $L(i\omega_{pc})$ intersects the unit half-circle below the real axis. The phase margin is then

$$\varphi_m = 180^\circ + \angle L(i\omega_{gc}). \tag{10.6}$$

This number can be obtained from the Nyquist plots as shown in Figure 10.11a. If the Nyquist curve intersects the half-circle many times the phase margin is defined by the intersection that is closest to the critical point.

The gain and phase margins can be determined from the Bode plot of the loop transfer function, as illustrated in Figure 10.11b. To find the gain margin we first find the phase crossover frequency ω_{pc} where the phase is -180° . The gain margin is the inverse of the gain at that frequency. To determine the phase margin we first determine the gain crossover frequency ω_{gc} , i.e., the frequency where the gain of the loop transfer function is 1. The phase margin is the phase of the loop transfer function at that frequency plus 180° . Figure 10.11b illustrates how the margins are found in the Bode plot of the loop transfer function. The margins are not well defined if the loop transfer function intersects the lines $|G(i\omega)| = 1$ or $\angle G(i\omega) = -180^\circ \pm n \cdot 360^\circ$ many times.

The gain and phase margins are classical robustness measures that have been used for a long time in control system design. They were particularly attractive because design was often based on the Bode plot of the loop transfer function. The

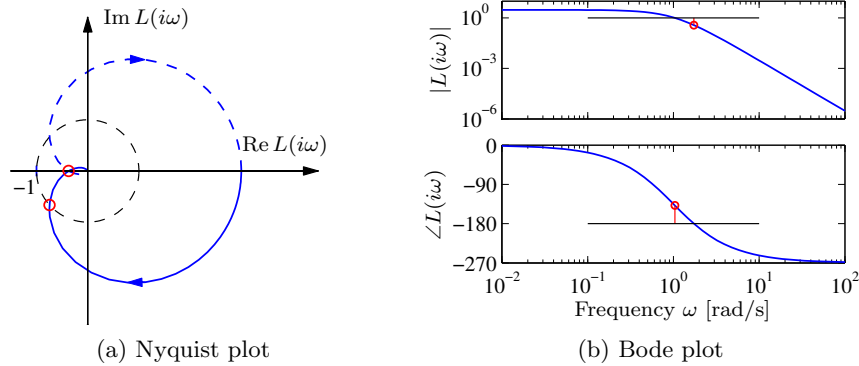


Figure 10.12: Stability margins for a third-order transfer function. The Nyquist plot on the left allows the gain, phase, and stability margins to be determined by measuring the distances of relevant features. The gain and phase margins can also be read off of the Bode plot on the right.

gain and phase margins are related to the stability margin through inequalities

$$g_m \geq \frac{1}{1 - s_m}, \quad \varphi_m \geq 2 \arcsin(s_m/2), \quad (10.7)$$

which follows from Figure 10.12 and the fact that s_m is less than the distance $d = 2 \sin(\varphi_m/2)$ from the critical point -1 to the point defining the gain crossover frequency.

A drawback with the stability margin s_m is that it does not have a natural representation in the Bode plot of the loop transfer function. It can be shown that the peak magnitude M_s of the closed loop transfer function $1/(1 + P(s)C(s))$ is related to the stability margin through the formula $s_m = 1/M_s$, as will be discussed in Chapter 13 together with more general robustness measures. A drawback with gain and phase margins is that both have to be given to guarantee that the Nyquist curve is not close to the critical point. It is also difficult to represent the winding number in the Bode plot. In general, it is best to use the Nyquist plot to check stability since this provides more complete information than the Bode plot.

Example 10.7 Stability margins for a third-order system

Consider a loop transfer function $L(s) = 3/(s + 1)^3$. The Nyquist and Bode plots are shown in Figure 10.12. To compute the gain, phase, and stability margins, we can use the Nyquist plot shown in Figure 10.12. This yields the following values:

$$g_m = 2.67, \quad \varphi_m = 41.7^\circ, \quad s_m = 0.464.$$

The gain and phase margins can also be determined from the Bode plot. ∇

Even if both the gain and phase margins are reasonable, the system may still not be robust, as is illustrated by the following example.

Example 10.8 Good gain and phase margins but poor stability margins

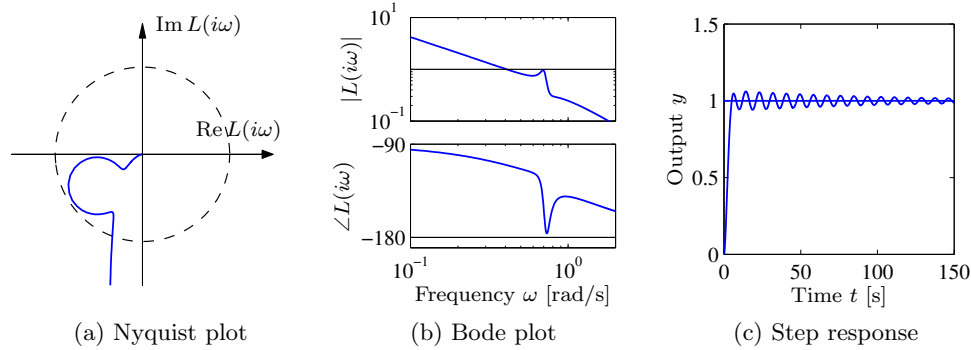


Figure 10.13: System with good gain and phase margins but a poor stability margin. The Nyquist plot (a) and Bode plot (b) of the loop transfer function and step response (c) for a system with good gain and phase margins but with a poor stability margin. The Nyquist plot shows only the portion of the curve corresponding to $\omega > 0$.

Consider a system with the loop transfer function

$$L(s) = \frac{0.38(s^2 + 0.1s + 0.55)}{s(s + 1)(s^2 + 0.06s + 0.5)}.$$

A numerical calculation gives the gain margin as $g_m = 266$, and the phase margin is 70° . These values indicate that the system is robust, but the Nyquist curve is still close to the critical point, as shown in Figure 10.13a. The stability margin is $s_m = 0.27$, which is very low. The closed loop system has two resonant modes, one with damping ratio $\zeta = 0.81$ and the other with $\zeta = 0.014$. The step response of the system is highly oscillatory, as shown in Figure 10.13c. ∇

When designing feedback systems, it will often be useful to define the robustness of the system using gain, phase, and stability margins. These numbers tell us how much the system can vary from our nominal model and still be stable. Reasonable values of the margins are phase margin $\varphi_m = 30^\circ$ – 60° , gain margin $g_m = 2$ – 5 and stability margin $s_m = 0.5$ – 0.8 .

There are also other stability measures, such as the *delay margin*, which is the smallest time delay required to make the system unstable. For loop transfer functions that decay quickly, the delay margin is closely related to the phase margin, but for systems where the gain curve of the loop transfer function has several peaks at high frequencies, the delay margin is a more relevant measure.

Example 10.9 Nanopositioning system for an atomic force microscope

Consider the system for horizontal positioning of the sample in an atomic force microscope, described in more detail in Section 4.5. The system has oscillatory dynamics, and a simple model is a spring–mass system with low damping. The normalized transfer function is given by

$$P(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2}, \tag{10.8}$$

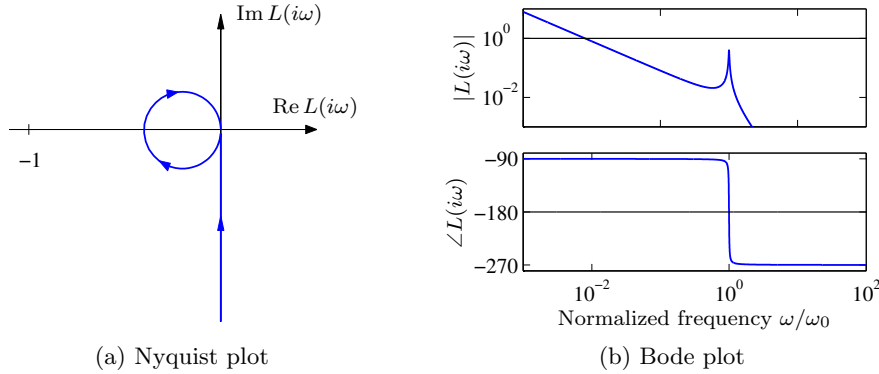


Figure 10.14: Nyquist and Bode plots of the loop transfer function for the AFM system (10.8) with an integral controller. The frequency in the Bode plot is normalized by ω_0 . The parameters are $\zeta = 0.01$ and $k_i = 0.008$.

where the damping ratio typically is a very small number, e.g., $\zeta = 0.1$.

We will start with a controller that has only integral action. The resulting loop transfer function is

$$L(s) = \frac{k_i \omega_0^2}{s(s^2 + 2\zeta \omega_0 s + \omega_0^2)},$$

where k_i is the gain of the controller. Nyquist and Bode plots of the loop transfer function are shown in Figure 10.14. Notice that the part of the Nyquist curve that is close to the critical point -1 is approximately circular.

From the Bode plot in Figure 10.14b, we see that the phase crossover frequency is $\omega_{pc} = \omega_0$, which will be independent of the gain k_i . Evaluating the loop transfer function at this frequency, we have $L(i\omega_0) = -k_i/(2\zeta\omega_0)$, which means that the stability margin is $s_m = 1 - k_i/(2\zeta\omega_0)$. To have a desired stability margin of s_m the integral gain should be chosen as

$$k_i = 2\zeta\omega_0(1 - s_m).$$

Figure 10.14 shows Nyquist and Bode plots for the system with gain margin $g_m = 2.5$ and stability margin $s_m = 0.6$. The gain curve in the Bode plot is almost a straight line for low frequencies and has a resonant peak at $\omega = \omega_0$. The gain crossover frequency is approximately equal to k_i and the phase decreases monotonically from -90° to -270° : it is equal to -180° at $\omega = \omega_0$. The gain curve can be shifted vertically by changing k_i : increasing k_i shifts the gain curve upward and increases the gain crossover frequency. ∇

10.4 BODE'S RELATIONS AND MINIMUM PHASE SYSTEMS

An analysis of Bode plots reveals that there appears to be a relation between the gain curve and the phase curve. Consider, for example, the Bode plots for the differentiator and the integrator (shown in Figure 9.13). For the differentiator the slope is $+1$ and the phase is a constant $\pi/2$ radians. For the integrator the slope is

-1 and the phase is $-\pi/2$. For the first-order system $G(s) = s + a$, the amplitude curve has the slope 0 for small frequencies and the slope $+1$ for high frequencies, and the phase is 0 for low frequencies and $\pi/2$ for high frequencies.

Bode investigated the relations between the gain and phase curves in his plot and he found that for a special class of systems there was indeed a relation between gain and phase. These systems do not have time delays or poles and zeros in the right half-plane and in addition they have the property that $\log |G(s)|/s$ goes to zero as $s \rightarrow \infty$ for $\text{Re } s \geq 0$. Bode called these systems *minimum phase systems* because they have the smallest phase lag of all systems with the same gain curve. For minimum phase systems the phase is uniquely given by the shape of the gain curve, and vice versa:

$$\arg G(i\omega_0) = \frac{\pi}{2} \int_0^\infty f(\omega) \frac{d \log |G(i\omega)|}{d \log \omega} \frac{d\omega}{\omega} \approx \frac{\pi}{2} \left. \frac{d \log |G(i\omega)|}{d \log \omega} \right|_{\omega=\omega_0}, \quad (10.9)$$

where f is the weighting kernel

$$f(\omega) = \frac{2}{\pi^2} \log \left| \frac{\omega + \omega_0}{\omega - \omega_0} \right|, \quad \text{and} \quad \int_0^\infty f(\omega) \frac{d\omega}{\omega} = 1. \quad (10.10)$$

The phase curve for a minimum phase system is thus a weighted average of the derivative of the gain curve. Notice that since $|G(s)| = |-G(s)|$ and $\angle(-G(s)) = \angle G(s) - 180^\circ$, the sign of the minimum phase $G(s)$ must also be chosen properly. We assume that the sign is always chosen so that $\angle G(s) > \angle(-G(s))$.

We illustrate Bode's relation (10.9) with an example.

Example 10.10 Phase of $G(s) = s^n$

We have $\log G(s) = n \log s$ and hence $d \log G(s)/d \log s = n$. Equation (10.9) then gives

$$\arg G(i\omega_0) = \frac{\pi}{2} \int_0^\infty f(\omega) \frac{d \log |G(i\omega)|}{d \log \omega} \frac{d\omega}{\omega} = \frac{\pi}{2} \int_0^\infty n f(\omega) \frac{d\omega}{\omega} = n \frac{\pi}{2},$$

where the last equality follows from equation (10.10). If the gain curve has constant slope n , the phase curve is a horizontal line $\arg G(i\omega) = n\pi/2$. ∇

We will now give a few examples of transfer functions that are not minimum phase transfer functions. The transfer function of a time delay of τ units is $G(s) = e^{-s\tau}$. This transfer function has unit gain $|G(i\omega)| = 1$, and the phase is $\arg G(i\omega) = -\omega\tau$. The corresponding minimum phase system with unit gain has the transfer function $G(s) = 1$. The time delay thus has an additional phase lag of $\omega\tau$. Notice that the phase lag increases linearly with frequency. Figure 10.15a shows the Bode plot of the transfer function. (Because we use a log scale for frequency, the phase falls off exponentially in the plot.)

Consider a system with the transfer function $G(s) = (a - s)/(a + s)$ with $a > 0$, which has a zero $s = a$ in the right half-plane. The transfer function has unit gain $|G(i\omega)| = 1$, and the phase is $\arg G(i\omega) = -2 \arctan(\omega/a)$. The corresponding minimum phase system with unit gain has the transfer function $G(s) = 1$. Figure 10.15b shows the Bode plot of the transfer function. A similar analysis of the transfer function $G(s) = (s + a)/(s - a)$ with $a > 0$, which has a pole in the right half-plane, shows that its phase is $\arg G(i\omega) = -2 \arctan(a/\omega)$. The Bode plot is

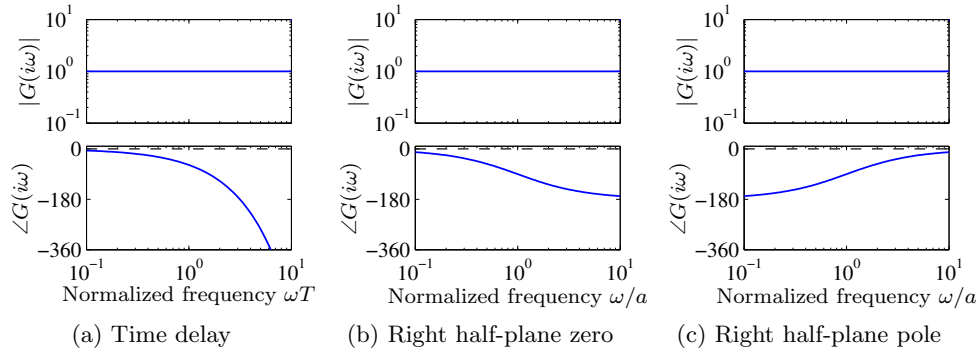


Figure 10.15: Bode plots of systems that are not minimum phase. (a) Time delay $G(s) = e^{-sT}$, (b) system with a right half-plane (RHP) zero $G(s) = (a - s)/(a + s)$ and (c) system with right half-plane pole $G(s) = (s + a)/(s - a)$. The corresponding minimum phase system has the transfer function $G(s) = 1$ in all cases, the phase curves for that system are shown as dashed lines.

shown in Figure 10.15c.

The presence of poles and zeros in the right half-plane imposes severe limits on the achievable performance as will be discussed in Chapter 14. Dynamics of this type should be avoided by redesign of the system. While the poles are intrinsic properties of the system and they do not depend on sensors and actuators, the zeros depend on how inputs and outputs of a system are coupled to the states. Zeros can thus be changed by moving sensors and actuators or by introducing new sensors and actuators. Non-minimum phase systems are unfortunately quite common in practice.

The following example shows that difficulties can arise in the response of non-minimum phase systems.

Example 10.11 Vehicle steering

The vehicle steering model considered in Examples 6.13 and 9.10 has different properties depending on whether we are driving forward or in reverse. The non-normalized transfer function from steering angle to lateral position for the simple vehicle model is

$$P(s) = \frac{av_0s + v_0^2}{bs^2},$$

where v_0 is the velocity of the vehicle and $a, b > 0$ (see Example 6.13). The transfer function has a zero at $s = v_0/a$. In normal (forward) driving this zero is in the left half-plane, but it is in the right half-plane when driving in reverse, $v_0 < 0$. The unit step response is

$$y(t) = \frac{av_0t}{b} + \frac{v_0^2t^2}{2b}.$$

The lateral position thus begins to respond immediately to a steering command through the. For reverse steering v_0 is negative and the initial response is in the wrong direction, a behavior that is representative for non-minimum phase systems (called an *inverse response*).

Figure 10.16 shows the step response for forward and reverse driving. The

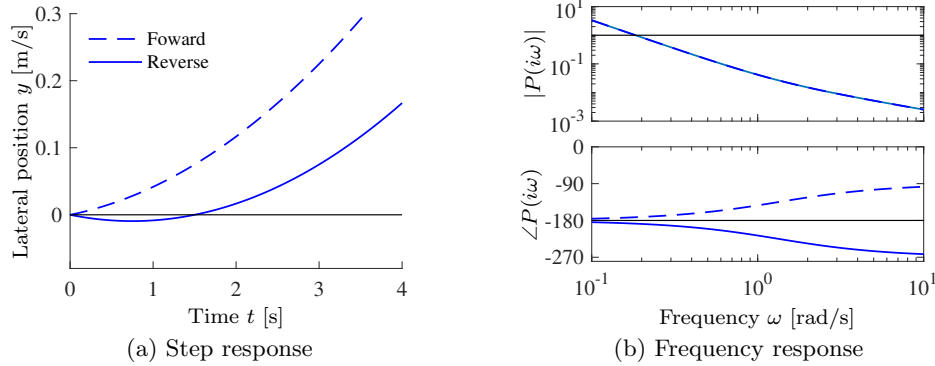


Figure 10.16: Vehicle steering for driving in reverse. (a) Step responses from steering angle to lateral translation for a simple kinematics model when driving forward (dashed) and reverse (solid). With rear-wheel steering the center of mass first moves in the wrong direction and the overall response with rear-wheel steering is significantly delayed compared with that for front-wheel steering. (b) Frequency response for driving forward (dashed) and reverse (solid). Notice that the gain curves are identical, but the phase curve for driving in reverse has non-minimum phase.

parameters are $a = 1.5$ m, $b = 3$ m, $v_0 = 2$ m/s for forward driving, and $v_0 = -2$ m/s for reverse driving. Thus when driving in reverse there is an initial motion of the center of mass in the *opposite* direction and there is a delay before the car begins to move in the desired manner.

The position of the zero v_0/a depends on the location of the sensor. In our calculation we have assumed that the sensor is at the center of mass. The zero in the transfer function disappears if the sensor is located at the rear wheel. Thus if we look at the center of the rear wheels instead of the center of mass, the inverse response is not present and the resulting input/output behavior is simplified. ∇

10.5 GENERALIZED NOTIONS OF GAIN AND PHASE



A key idea in frequency domain analysis is to trace the behavior of sinusoidal signals through a system. The concepts of gain and phase represented by the transfer function are strongly intuitive because they describe amplitude and phase relations between input and output. In this section we will see how to extend the concepts of gain and phase to more general systems, including some nonlinear systems. We will also show that there are analogs of Nyquist’s stability criterion if signals are approximately sinusoidal.

System Gain and Passivity

We begin by considering the case of a static linear system $y = Au$, where A is a matrix whose elements are complex numbers. The matrix does not have to be square. Let the inputs and outputs be vectors whose elements are complex numbers

and use the Euclidean norm

$$\|u\| = \sqrt{\sum |u_i|^2}. \quad (10.11)$$

The norm of the output is

$$\|y\|^2 = u^* A^* A u,$$

where $*$ denotes the complex conjugate transpose. The matrix $A^* A$ is symmetric and positive semidefinite, and the right-hand side is a quadratic form. The square root of eigenvalues of the matrix $A^* A$ are all real, and we have

$$\|y\|^2 \leq \bar{\lambda}(A^* A) \|u\|^2,$$

where $\bar{\lambda}$ denotes the largest eigenvalue. The gain of the system can then be defined as the maximum ratio of the output to the input over all possible inputs:

$$\gamma = \max_u \frac{\|y\|}{\|u\|} = \sqrt{\bar{\lambda}(A^* A)}. \quad (10.12)$$

The square root of the eigenvalues of the matrix $A^* A$ are called the *singular values* of the matrix A , and the largest singular value is denoted by $\bar{\sigma}(A)$.

To generalize this to the case of an input/output dynamical system, we need to think of the inputs and outputs not as vectors of real numbers but as vectors of *signals*. For simplicity, consider first the case of scalar signals and let the signal space L_2 be square-integrable functions with the norm

$$\|u\|_2 = \sqrt{\int_0^\infty |u|^2(\tau) d\tau}.$$

This definition can be generalized to vector signals by replacing the absolute value with the vector norm (10.11). We can now formally define the *gain* of a system taking inputs $u \in L_2$ and producing outputs $y \in L_2$ as

$$\gamma = \sup_{u \in L_2} \frac{\|y\|_2}{\|u\|_2}, \quad (10.13)$$

where sup is the *supremum*, defined as the smallest number that is larger or equal to its argument. The reason for using the supremum is that the maximum may not be defined for $u \in L_2$. This definition of the system gain is quite general and can even be used for some classes of nonlinear systems, though one needs to be careful about how initial conditions and global nonlinearities are handled.

This generalized notion of gain can be used to define the concept of input/output stability for a system. Roughly speaking, a system is called bounded input/bounded output (BIBO) stable if a bounded input gives a bounded output for all initial states. A system is called input to state stable (ISS) if $\|x(t)\| \leq \beta(\|x(0)\|) + \gamma(\|u\|)$ where β and γ are monotonically increasing functions that vanish at the origin.

The norm (10.13) has some nice properties in the case of linear systems. In particular, given a single-input, single-output stable linear system with transfer function $G(s)$, it can be shown that the norm of the system is given by

$$\gamma = \sup_{\omega} |G(i\omega)| =: \|G\|_\infty. \quad (10.14)$$

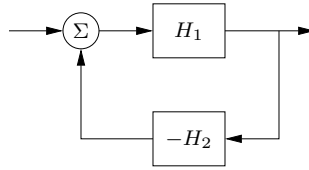


Figure 10.17: Block diagram of feedback connection of two general nonlinear systems H_1 and H_2 .

In other words, the gain of the system corresponds to the peak value of the frequency response. This corresponds to our intuition that an input produces the largest output when we are at the resonant frequencies of the system. $\|G\|_\infty$ is called the *infinity norm* of the transfer function $G(s)$.

This notion of gain can be generalized to the multi-input, multi-output case as well. For a linear multivariable system with a transfer function matrix $G(s)$ we can define the gain as

$$\gamma = \|G\|_\infty = \sup_{\omega} \bar{\sigma}(G(i\omega)). \tag{10.15}$$

Thus we can combine the idea of the gain of a matrix with the idea of the gain of a linear system by looking at the maximum singular value over all frequencies.

In addition to generalizing the system gain, it is also possible to make generalizations of the concept of phase. The angle between two vectors can be defined by the equation

$$\langle u, y \rangle = \|u\| \|y\| \cos(\varphi), \tag{10.16}$$

where the left argument denotes the scalar product. If systems are defined in such a way that we have norms of signals and a scalar product between signals we can use equation (10.16) to define the phase between two signals. For square-integrable inputs and outputs we have the scalar product

$$\langle u, y \rangle = \int_0^\infty u(\tau)y(\tau) d\tau,$$

and the *phase* φ between the signals u and y can now be defined through equation (10.16).

Systems where the phase between inputs and outputs is 90° or less for all inputs are called *passive systems*. Systems where the phase is strictly less than 90° are called *strictly passive*.

Extensions of Nyquist's Theorem

There are many extensions of the Nyquist theorem, and we briefly sketch a few of them here. For linear systems it follows from Nyquist's theorem that the closed loop is stable if the gain of the loop transfer function is less than 1 for all frequencies. Since we have a notion of gain for nonlinear systems given by equation (10.13), we can extend this case of the Nyquist theorem to nonlinear systems:

Theorem 10.4 (Small gain theorem). *Consider the closed loop system shown in Figure 10.17, where H_1 and H_2 are input/output stable systems and the signal spaces and initial conditions are properly defined. Let the gains of the systems H_1*

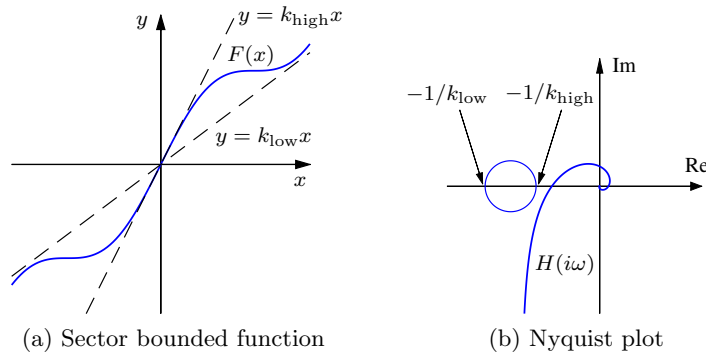


Figure 10.18: Stability using the circle criterion. For a feedback system with a sector-bounded nonlinearity (a), the Nyquist plot (b) must stay outside of a circle defined by $-1/k_{\text{low}} \leq x \leq -1/k_{\text{high}}$

and H_2 be γ_1 and γ_2 . Then the closed loop system is input/output stable if $\gamma_1 \gamma_2 < 1$, and the gain of the closed loop system is

$$\gamma = \frac{\gamma_1}{1 - \gamma_1 \gamma_2}.$$

Another extension of the Nyquist theorem to nonlinear systems can be obtained by investigating the phase shift of the nonlinear systems. Consider again the system in Figure 10.17. It follows from the Nyquist criterion that if the blocks H_1 and H_2 are linear transfer functions, then the closed loop system is stable if the phase of $H_1 H_2$ is always less than 180° . A generalization of this to nonlinear systems is that the closed loop system is stable if both H_1 and H_2 are passive and if one of them is strictly passive. This result is called the *passivity theorem*.

A final useful extension of the Nyquist theorem applies to the system in Figure 10.18 where H_1 is a linear system with transfer function $H(s)$ and the nonlinear block H_2 is a static nonlinearity described by the function $F(x)$, which is *sector-bounded*

$$k_{\text{low}} x \leq F(x) \leq k_{\text{high}} x. \quad (10.17)$$

The following theorem allows us to reason about the stability of such a system.

Theorem 10.5 (Circle criterion). *Consider a negative feedback system consisting of a linear system with transfer function $H(s)$ and a static nonlinearity defined by a function $F(x)$ satisfying the sector bound (10.17). The closed loop system is stable if the Nyquist curve of $H(i\omega)$ is outside a circle with diameter $-1/k_{\text{low}} \leq x \leq -1/k_{\text{high}}$ and the encirclement condition is satisfied.*

The extensions of the Nyquist theorem we have discussed are powerful and easy to apply, and we will use them later to in Chapter 13. Details, proofs, and applications are found in [Kha01].

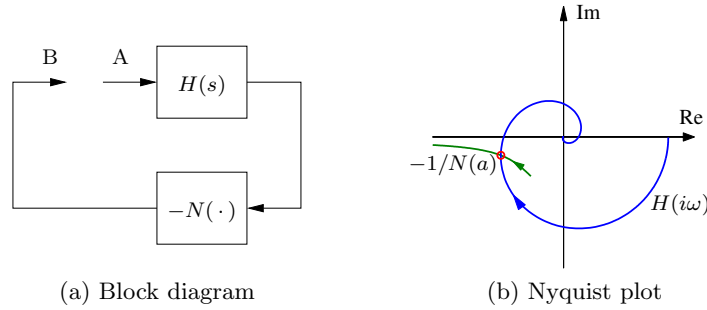


Figure 10.19: Describing function analysis. A feedback connection between a static nonlinearity and a linear system is shown in (a). The linear system is characterized by its transfer function $H(s)$, which depends on frequency, and the nonlinearity by its describing function $N(a)$, which depends on the amplitude a of its input. The Nyquist plot of $H(i\omega)$ and the plot of the $-1/N(a)$ are shown in (b). The intersection of the curves represents a possible limit cycle.

Describing Functions

For special nonlinear systems like the one shown in Figure 10.19a, which consists of a feedback connection between a linear system and a static nonlinearity, it is possible to obtain a generalization of Nyquist’s stability criterion based on the idea of *describing functions*. Following the approach of the Nyquist stability condition, we will investigate the conditions for maintaining an oscillation in the system. If the linear subsystem has low-pass character, its output is approximately sinusoidal even if its input is highly irregular. The condition for oscillation can then be found by exploring the propagation of a sinusoid that corresponds to the first harmonic.

To carry out this analysis, we have to analyze how a sinusoidal signal propagates through a static nonlinear system. In particular we investigate how the first harmonic of the output of the nonlinearity is related to its (sinusoidal) input. Letting $F(x)$ represent the nonlinear function, we expand $F(e^{i\omega t})$ in terms of its harmonics:

$$F(ae^{i\omega t}) = \sum_{n=0}^{\infty} M_n(a)e^{i(n\omega t + \varphi_n(a))},$$

where $M_n(a)$ and $\varphi_n(a)$ represent the gain and phase of the n th harmonic, which depend on the input amplitude since the function $F(x)$ is nonlinear. We define the describing function to be the complex gain of the first harmonic:

$$N(a) = M_1(a)e^{i\varphi_1(a)}. \tag{10.18}$$

The function can also be computed by assuming that the input is a sinusoid and using the first term in the Fourier series of the resulting output.

Neglecting higher harmonics and arguing as we did when deriving Nyquist’s stability criterion, we find that an oscillation can be maintained if

$$H(i\omega)N(a) = -1. \tag{10.19}$$

This equation means that if we inject a sinusoid of amplitude a at A in Figure 10.19,

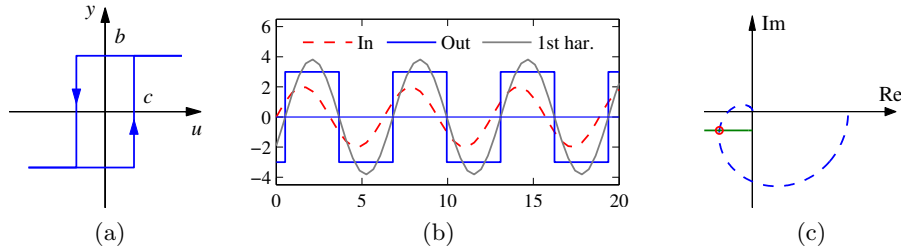


Figure 10.20: Describing function analysis for a relay with hysteresis. The input/output relation of the hysteresis is shown in (a) and the input with amplitude $a = 2$, the output and its first harmonic are shown in (b). The Nyquist plots of the transfer function $H(s) = (s + 1)^{-4}$ and the negative of the inverse describing function for the relay with $b = 3$ and $c = 1$ are shown in (c).

the same signal will appear at B and an oscillation can be maintained by connecting the points. Equation (10.19) gives two conditions for finding the frequency ω of the oscillation and its amplitude a : the phase of $H(i\omega)N(a)$ must be 180° and its magnitude must be unity. A convenient way to solve the equation is to plot $H(i\omega)$ and $-1/N(a)$ on the same diagram as shown in Figure 10.19b. The diagram is similar to the Nyquist plot where the critical point -1 is replaced by the curve $-1/N(a)$ and a ranges from 0 to ∞ .

It is possible to define describing functions for types of inputs other than sinusoids. Describing function analysis is a simple method, but it is approximate because it assumes that higher harmonics can be neglected. Excellent treatments of describing function techniques can be found in the texts by Atherton [Ath75] and Graham and McRuer [GM61].

Example 10.12 Relay with hysteresis

Consider a linear system with a nonlinearity consisting of a relay with hysteresis. The output has amplitude b and the relay switches when the input is $\pm c$, as shown in Figure 10.20a. Assuming that the input is $u = a \sin(\omega t)$, we find that the output is zero if $a \leq c$, and if $a > c$ the output is a square wave with amplitude b that switches at times $\omega t = \arcsin(c/a) + n\pi$. The first harmonic is then $y(t) = (4b/\pi) \sin(\omega t - \alpha)$, where $\sin \alpha = c/a$. For $a > c$ the describing function and its inverse are

$$N(a) = \frac{4b}{a\pi} \left(\sqrt{1 - \frac{c^2}{a^2}} - i \frac{c}{a} \right), \quad \frac{1}{N(a)} = \frac{\pi \sqrt{a^2 - c^2}}{4b} + i \frac{\pi c}{4b},$$

where the inverse is obtained after simple calculations. Figure 10.20b shows the response of the relay to a sinusoidal input with the first harmonic of the output shown as a dashed line. Describing function analysis is illustrated in Figure 10.20c, which shows the Nyquist plot of the transfer function $H(s) = 2/(s + 1)^4$ (dashed line) and the negative inverse describing function of a relay with $b = 1$ and $c = 0.5$. The curves intersect for $a = 1$ and $\omega = 0.77$ rad/s, indicating the amplitude and frequency for a possible oscillation if the process and the relay are connected in a feedback loop. ∇

It follows from the example that the describing function for a relay without

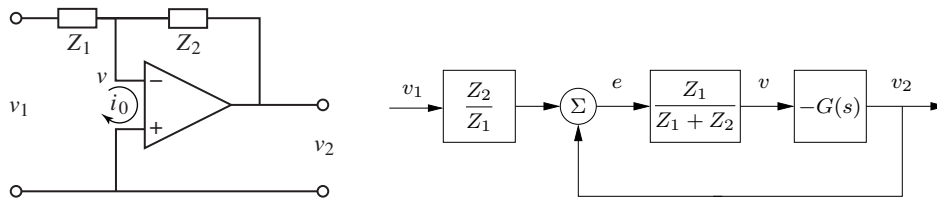
hysteresis is $N(a) = 4b/(a\pi)$ and $-1/N(a)$ is thus the negative real axis. For the saturation function, $-1/N(a)$ is the part of the negative real axis from $-\infty$ to -1 .

10.6 FURTHER READING

Nyquist’s original paper giving his now famous stability criterion was published in the *Bell Systems Technical Journal* in 1932 [Nyq32]. More accessible versions are found in the book [BK64], which also includes other interesting early papers on control. Nyquist’s paper is also reprinted in an IEEE collection of seminal papers on control [Bas01]. Nyquist used +1 as the critical point, but Bode changed it to -1 , which is now the standard notation. Interesting perspectives on early developments are given by Black [Bla77], Bode [Bod60], and Bennett [Ben93]. Nyquist did a direct calculation based on his insight into the propagation of sinusoidal signals through systems; he did not use results from the theory of complex functions. The idea that a short proof can be given by using the principle of variation of the argument is presented in the delightful book by MacColl [Mac45]. Bode made extensive use of complex function theory in his book [Bod45], which laid the foundation for frequency response analysis where the notion of minimum phase was treated in detail. A good source for complex function theory is the classic by Ahlfors [Ahl66]. The extensions of the Nyquist theorem to a closed loop system that is composed of a linear system and a static nonlinearity has received significant attention. An extensive treatment of the passivity and small gain theorems and describing functions is given in the book by Khalil [Kha01]. Describing functions for many nonlinearities are given in the books by Atherton [Ath75] and Graham and McRuer [GM61]. Frequency response analysis was a key element in the emergence of control theory as described in the early texts by James et al. [JNP47], Brown and Campbell [BC48], and Oldenburger [Old56], and it became one of the cornerstones of early control theory. Frequency response methods underwent a resurgence when robust control emerged in the 1980s, as will be discussed in Chapter 13.

EXERCISES

10.1 (Operational amplifier loop transfer function) Consider the operational amplifier circuit shown below, where Z_1 and Z_2 are generalized impedances and the open loop amplifier is modeled by the transfer function $G(s)$.



Show that the system can be modeled as the block diagram on the right, with loop transfer function $L = Z_1G/(Z_1 + Z_2)$ and feedforward transfer function $F =$

$$Z_1/(Z_1 + Z_2).$$

10.2 (Atomic force microscope) The dynamics of the tapping mode of an atomic force microscope are dominated by the damping of the cantilever vibrations and the system that averages the vibrations. Modeling the cantilever as a spring–mass system with low damping, we find that the amplitude of the vibrations decays as $\exp(-\zeta\omega_0 t)$, where ζ is the damping ratio and ω_0 is the undamped natural frequency of the cantilever. The cantilever dynamics can thus be modeled by the transfer function

$$G(s) = \frac{a}{s + a},$$

where $a = \zeta\omega_0$. The averaging process can be modeled by the input/output relation

$$y(t) = \frac{1}{\tau} \int_{t-\tau}^t u(v) dv,$$

where the averaging time is a multiple n of the period of the oscillation $2\pi/\omega$. The dynamics of the piezo scanner can be neglected in the first approximation because they are typically much faster than a . A simple model for the complete system is thus given by the transfer function


$$P(s) = \frac{a(1 - e^{-s\tau})}{s\tau(s + a)}.$$

Plot the Nyquist curve of the system and determine the gain of a proportional controller that brings the system to the boundary of stability.

10.3 (Heat conduction) A simple model for heat conduction in a solid is given by the transfer function

$$P(s) = ke^{-\sqrt{s}}.$$

Sketch the Nyquist plot of the system. Determine the frequency where the phase of the process is -180° and the gain at that frequency. Show that the gain required to bring the system to the stability boundary is $k = e^\pi$.

10.4 (Vectored thrust aircraft) Consider the state space controller designed for the vectored thrust aircraft in Examples 7.9 and 8.7. The controller consists of two components: an optimal estimator to compute the state of the system from the output and a state feedback compensator that computes the input given the (estimated) state. Compute the loop transfer function for the system and determine the gain, phase, and stability margins for the closed loop dynamics. 

10.5 (Vehicle steering) Consider the linearized model for vehicle steering with a controller based on state feedback discussed in Example 8.4. The transfer functions for the process and controller are given by

$$P(s) = \frac{\gamma s + 1}{s^2}, \quad C(s) = \frac{s(k_1 l_1 + k_2 l_2) + k_1 l_2}{s^2 + s(\gamma k_1 + k_2 + l_1) + k_1 + l_2 + k_2 l_1 - \gamma k_2 l_2},$$

as computed in Example 9.10. Let the process parameter be $\gamma = 0.5$ and assume that the state feedback gains are $k_1 = 0.5$ and $k_2 = 0.75$ and that the observer gains are $l_1 = 1.4$ and $l_2 = 1$. Compute the stability margins numerically.

10.6 (Unity gain operational amplifier) Consider an op amp circuit with $Z_1 = Z_2$ that gives a closed loop system with nominally unit gain. Let the transfer function of the operational amplifier be

$$G(s) = \frac{ka_1a_2}{(s+a)(s+a_1)(s+a_2)},$$

where $a_1, a_2 \gg a$. Show that the condition for oscillation is $k < a_1 + a_2$ and compute the gain margin of the system. Hint: Assume $a = 0$.

10.7 (Stability margins for second-order systems) A process whose dynamics is described by a double integrator is controlled by an ideal PD controller with the transfer function $C(s) = k_d s + k_p$, where the gains are $k_d = 2\zeta\omega_0$ and $k_p = \omega_0^2$. Calculate and plot the gain, phase, and stability margins as a function ζ .

10.8 (Kalman's inequality) Consider the linear system (7.20). Let $u = -Kx$ be a state feedback control law obtained by solving the linear quadratic regulator problem. Prove the inequality

$$(I + L(-i\omega))^T Q_u (I + L(i\omega)) \geq Q_u,$$

where

$$K = Q_u^{-1} B^T S, \quad L(s) = K(sI - A)^{-1} B.$$

(Hint: Use the Riccati equation (7.33), add and subtract the terms sS , multiply with $B^T(sI + A)^{-T}$ from the left and $(sI - A)^{-1}B$ from the right.)

For single-input single-output systems this result implies that the Nyquist plot of the loop transfer function has the property $|1 + L(i\omega)| \geq 1$, from which it follows that the phase margin for a linear quadratic regulator is always greater than 60° .

10.9 (Congestion control in overload conditions) A strongly simplified flow model of a TCP loop under overload conditions is given by the loop transfer function

$$L(s) = \frac{k}{s} e^{-s\tau},$$

where the queuing dynamics are modeled by an integrator, the TCP window control is a time delay τ , and the controller is simply a proportional controller. A major difficulty is that the time delay may change significantly during the operation of the system. Show that if we can measure the time delay, it is possible to choose a gain that gives a stability margin of $s_m \geq 0.6$ for all time delays τ .

10.10 (Bode's formula) Consider Bode's formula (10.9) for the relation between gain and phase for a transfer function that has all its singularities in the left half-plane. Plot the weighting function and make an assessment of the frequencies where the approximation $\arg G \approx (\pi/2)d \log |G|/d \log \omega$ is valid.

10.11 (Circle criterion) Consider the system in Figure 10.17, where H_1 is a linear system with the transfer function $H(s)$ and H_2 is a static nonlinearity $F(x)$ with the property $x F(x) \geq 0$. Use the circle criterion to prove that the closed loop system is stable if $H(s)$ is strictly passive.

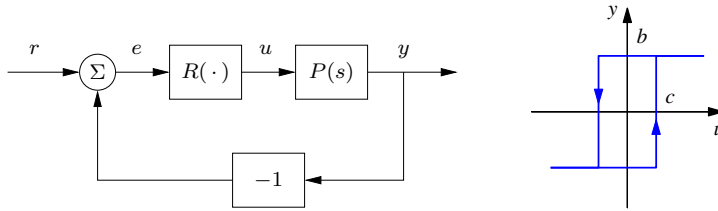
10.12 (Padé approximation to a time delay) Consider the transfer functions

$$G(s) = e^{-s\tau}, \quad G_1(s) = \frac{1 - s\tau/2}{1 + s\tau/2}. \quad (10.20)$$

Show that the minimum phase properties of the transfer functions are similar for frequencies $\omega < 1/\tau$. A long time delay τ is thus equivalent to a small right half-plane zero. The approximation $G_1(s)$ in equation (10.20) is called a first-order *Padé approximation*.

10.13 (Inverse response) Consider a system whose input/output response is modeled by $G(s) = 6(-s + 1)/(s^2 + 5s + 6)$, which has a zero in the right half-plane. Compute the step response for the system, and show that the output goes in the wrong direction initially, which is also referred to as an *inverse response*. Compare the response to a minimum phase system by replacing the zero at $s = 1$ with a zero at $s = -1$.

10.14 (Describing function analysis) Consider the system with the block diagram shown on the left below.



The block R is a relay with hysteresis whose input/output response is shown on the right and the process transfer function is $P(s) = e^{-s\tau}/s$. Use describing function analysis to determine frequency and amplitude of possible limit cycles. Simulate the system and compare with the results of the describing function analysis.

10.15 (Describing functions) Consider the saturation function

$$y = \text{sat}(x) = \begin{cases} -1 & \text{if } x \leq -1, \\ x & \text{if } -1 < x \leq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

Show that the describing function is

$$N(a) = \begin{cases} x & \text{if } |x| \leq 1, \\ \frac{2}{\pi} \left(\arcsin \frac{1}{x} + \frac{1}{x} \sqrt{1 - \frac{1}{x^2}} \right) & \text{if } |x| > 1. \end{cases}$$