Chapter 7

Loop Analysis

Quotation

Authors, citation.

This chapter describes how stability and robustness can be determined by investigating how sinusoidal signals propagate around the feedback loop. The Nyquist stability theorem is a key result which gives a new way to analyze stability. It also make it possible to introduce measures degrees of stability. Another important idea, due to Bode, makes it possible to separate linear dynamical systems into two classes, minimum phase systems that are easy to control and non-minimum phase systems that are difficult to control.

7.1 Introduction

The basic idea of loop analysis is to trace how a sinusoidal signal propagates in the feedback loop. Stability can be explored by investigating if the signal grows or decays when passes around the feedback loop. This is easy to do because the transmission of sinusoidal signals through a dynamical system is characterized by the frequency response of the system. The key result is the Nyquist stability theorem, which is important for several reasons. The concept of Lyapunov stability is binary, a system is either stable or unstable. From an engineering point of view it is very useful to have a notion of degrees of stability. Stability margins can be defined based on properties of the Nyquist curve. The Nyquist theorem also indicates how an unstable system should be changed to make it stable.

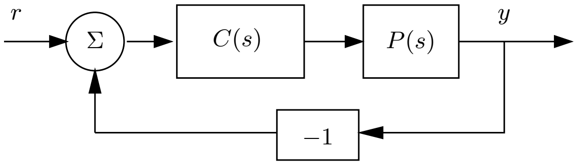


Figure 7.1: Block diagram of a simple feedback system.

7.2 The Basic Idea

Consider the system in Figure 7.1. 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 $C(s) = n_c(s)/d_c(s)$ the closed loop system has the transfer function.

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

and the characteristic polynomial is

$$d_{cl}(s) = d_p(s)d_c(s) + n_p(s)n_c(s)$$

This approach is straight forward 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 investigate conditions for maintaining oscillations in a feedback loop. Consider the system in Figure 7.1. Introduce

$$L(s) = P(s)C(s)$$

which is called the *loop transfer function*. The system can then be represented by the block diagram in Figure 7.2. In this figure we have also cut the loop open. We will first determine conditions for having a periodic oscillation in the loop. 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 could then connect A to B. Tracing signals around the loop we find that the signals at A and B are identical if

$$L(i\omega_0) = -1, \tag{7.1}$$

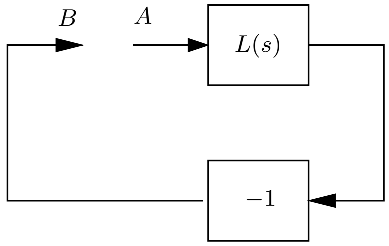


Figure 7.2: Block diagram of feedback system with the loop opened at AB.

which is the condition for maintaining an oscillation. To explore this further we will introduce a graphical representation of the loop transfer function.

The Nyquist Plot

The frequency response of the loop transfer function can be represented by plotting the complex number $L(i\omega)$ as a function of ω. Such a plot is called the *Nyquist plot* and the curve is called the *Nyquist curve*. An example of a Nyquist plot is given in Figure 7.3. The magnitude $|L(i\omega)|$ is called the *loop gain* because it tells how much the signal is amplified as is passes around the feedback loop and the angle $\arg L(i\omega)$ is called the *phase*.

The condition for oscillation (7.1) implies that the Nyquist curve of the loop transfer function goes through the point $L = -1$, which is called the *critical point*.

Intuitively it seems reasonable that the system is stable if $L(i\omega_c)| < 1$, which means that the critical point -1 is on the left hand side of the Nyquist curve as indicated in Figure 7.3. This means that the signal at point B will have smaller amplitude than the injected signal. This is essentially true, but there are several subtleties that requires a proper mathematics to clear up. This will be done later. The precise statement is given by the Nyquist stability theorem.

For loop transfer functions which do not have poles in the right half plane the precise condition is that the complete Nyquist plot does not encircle the critical point -1. The complete Nyquist plot is obtained by adding the plot for negative frequencies shown in the dashed curve in Figure 7.4. This plot is the mirror image of the Nyquist curve in the real axis. The following procedure can be used to determine that there are no encirclements. Fix a pin at the critical point orthogonal to the plane. Attach a rubber string

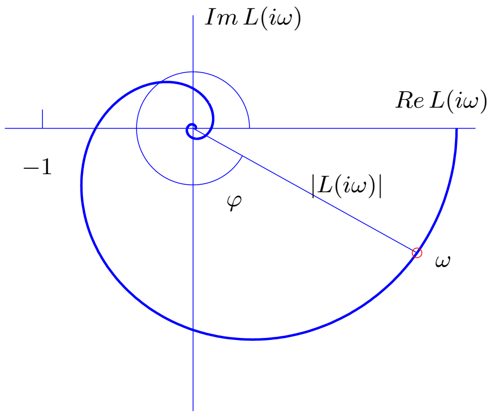


Figure 7.3: Nyquist plot of the transfer function $L(s) = 1.4e^{-s}/(s+1)^2$. The gain and phase at the frequency ω are $g = |L(i\omega)|$ and $\varphi = \arg L(i\omega)$.

with one end in the pin and another to the Nyquist plot. Let the end of the string attached to the Nyquist curve traverse the whole curve. There are no encirclements if the cord does not wind up on the pin when the curve is encircled.

One nice property of the Nyquist stability criterion is that it can be applied to infinite dimensional systems as is illustrated by the following example.

Example 7.1 *(Heat Conduction).* Consider a temperature control system shown in where the process has the transfer function

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

and the controller is a proportional controller with gain k. The loop transfer function is $L(s) = ke^{-\sqrt{s}}$, its Nyquist plot for $k = 1$ is shown in Figure **??**. We have

$$P(i\omega) = e^{-\sqrt{i\omega}} = e^{-\sqrt{\omega/2} - i\sqrt{\omega/2}}.$$

Hence

$$\log P(i\omega) = -\sqrt{i\omega} = -\frac{\omega\sqrt{2}}{2} - i\frac{\omega\sqrt{2}}{2},$$

We have

$$\arg L(i\omega) = -\frac{\omega\sqrt{2}}{2}.$$

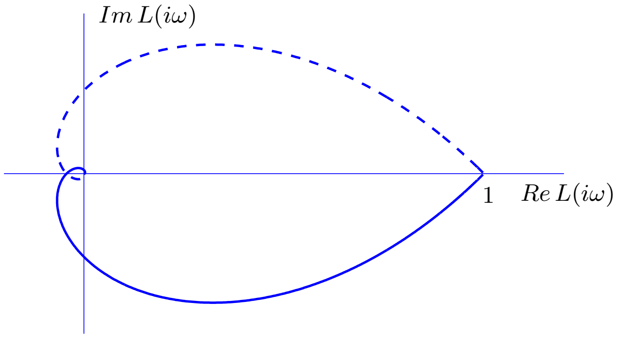


Figure 7.4: Nyquist plot of the transfer function $L(s) = e^{-\sqrt{s}}$

The phase is $-\pi$ for $\omega = \omega_c = \pi/\sqrt{2}$ and the gain at that frequency is $ke^{-\pi} = 0.0432k$. The Nyquist plot for a system with gain k is obtained simply by multiplying the Nyquist curve in the figure by k. The Nyquist curve reaches the critical point $L = -1$ for $k = e^{\pi} = 23.1$. The complete Nyquist curve in Figure 7.4 shows that the Nyquist curve does not encircle the critical point if $k < e^{\pi}$ which is the stability condition.

7.3 Nyquist's Stability Theorem

We will now state and prove the Nyquist stability theorem. This will require results from the theory of complex variables. Since precision is needed we will also use a more mathematical style of presentation. The key result is the following theorem about functions of complex variables.

Theorem 7.1 *(Principle of Variation of the Argument). Let D be a closed region in the complex plane and let Γ be the boundary of the region. Assume the function $f : \mathbb{C} \to \mathbb{C}$ is analytic in D and on Γ except at a finite number of poles and zeros, then*

$$w_n = \frac{1}{2\pi}\Delta_\Gamma \arg f(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f'(z)}{f(z)}dz = N - P$$

where N is the number of zeros and P the number of poles in D. Poles and zeros of multiplicity m are counted m times.

Proof. Assume that $z = a$ is a zero of multiplicity m. In the neighborhood of $z = a$ we have

$$f(z) = (z - a)^m g(z)$$

where the function g is analytic and different form zero. We have

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

The second term is analytic at $z = a$. The function f'/f thus has a single pole at $z = a$ with the residue m. The sum of the residues at the zeros of the function is N. Similarly we find that the sum of the residues of the poles of is $-P$. Furthermore we have

$$\frac{d}{dz} \log f(z) = \frac{f'(z)}{f(z)}$$

which implies that

$$\int_\Gamma \frac{f'(z)}{f(z)} dz = \Delta_\Gamma \log f(z)$$

where Δ_Γ denotes the variation along the contour Γ. We have

$$\log f(z) = \log |f(z)| + i \arg f(z)$$

Since the variation of $|f(z)|$ around a closed contour is zero we have

$$\Delta_\Gamma \log f(z) = i \Delta_\Gamma \arg f(z)$$

and the theorem is proven. □

Remark 7.1. The number $\Delta_\Gamma \arg f(z)$ is the variation of the argument of the function f as the curve Γ is traversed in the positive direction.

Remark 7.2. The number w_n is called the winding number. It equals the number of encirclements of the critical point when the curve Γ is traversed in the positive direction.

Remark 7.3. The theorem is useful to determine the number of poles and zeros of an function of complex variables in a given region. To use the result we must determine the winding number. One way to do this is to investigate how the curve Γ is transformed under the map f. The variation of the argument is the number of times the map of Γ winds around the origin in the f-plane. This explains why the variation of the argument is also called the winding number.

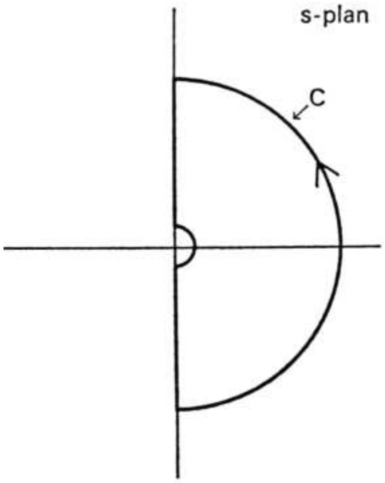


Figure 7.5: The Nyquist contour Γ.**Should be redrawn with poles on imaginary axis as you have in your lecture notes.**

Theorem 7.1 can be used to prove Nyquist's stability theorem. To do that we choose Γ as the Nyquist contour shown in Figure 7.5, 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. The Nyquist contour is obtained by letting $R \rightarrow \infty$ and $r \rightarrow 0$.

The contour consists of a small half circle to the right of the origin, the imaginary axis and a large half circle to the right with with the imaginary axis as a diameter. To illustrate the contour we have shown it drawn with a small radius r and a large radius R. The Nyquist curve is normally the map of the positive imaginary axis. We call the contour Γ the full Nyquist contour.

Consider a closed loop system with the loop transfer function $L(s)$. The closed loop poles are the zeros of the function

$$f(s) = 1 + L(s)$$

To find the number of 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 clockwise direction. The winding number can conveniently be determined from the Nyquist plot. A direct application of the Theorem 7.1 gives the following result.

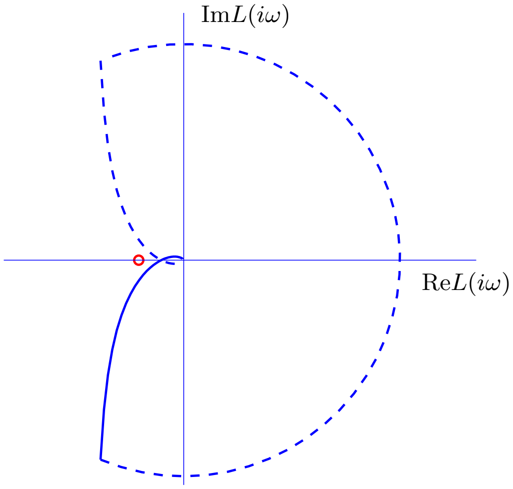


Figure 7.6: The complete Nyquist curve for the loop transfer function $L(s) = \frac{k}{s(s+1)^2}$. The curve is drawn for $k < 2$. The map of the positive imaginary axis is shown in full lines, the map of the negative imaginary axis and the small semi circle at the origin in dashed lines.

Theorem 7.2 *(Nyquist's Stability Theorem). Consider a closed loop system with the loop transfer function $L(s)$, which which has P poles in the region enclosed by the Nyquist contour. Let the winding number of the function $f(s) = 1 + L(s)$ when s encircles the Nyquist contour Γ be w_n. The closed loop system then has $w_n + P$ poles in the right half plane.*

There is a subtlety with the Nyquist plot when the loop transfer function has poles on the imaginary axis because the gain of is infinite at the poles. This means that the map of the small semicircles are infinitely large half circles. When plotting Nyquist curves in Matlab correct results are obtained for poles at the origin but Matlab does not deal with other poles on the imaginary axis. We illustrate Nyquist's theorem by an examples.

Example 7.2 *(A Simple Case).* Consider a closed loop system with the loop transfer function

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

Figure 7.6 shows the image of the contour Γ under the map L. The loop transfer function does not have any poles in the region enclosed by the Nyquist contour. The Nyquist plot intersects the imaginary axis for $\omega = 1$ the intersection is at $-k/2$. It follows from Figure 7.6 that the winding

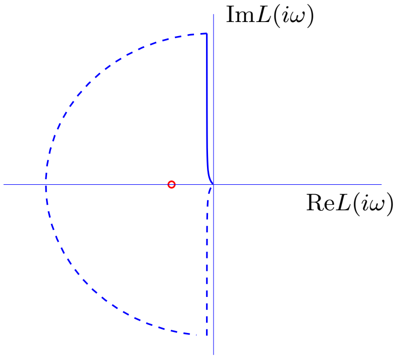


Figure 7.7: Complete Nyquist plot for the loop transfer function $L(s) = \frac{k}{s(s-1)(s+5)}$. The curve on the right shows the region around the origin in larger scale. The map of the positive imaginary axis is shown in full lines, the map of the negative imaginary axis and the small semi circle at the origin in dashed lines.

number is zero if $k < 2$ and 2 if $k > 2$. We can thus conclude that the closed loop system is stable if $k < 2$ and that the closed loop system has two roots in the right half plane if $k > 2$.

Next we will consider a case where the loop transfer function has a pole inside the Nyquist contour.

Example 7.3 *(Loop Transfer Function with RHP Pole).* Consider a feedback system with the loop transfer function

$$L(s) = \frac{k}{s(s-1)(s+5)}$$

This transfer function has a pole at $s = 1$ which is inside the Nyquist contour. The complete Nyquist plot of the loop transfer function is shown in Figure 7.7. Traversing the contour Γ in clockwise we find that the winding number is $w_n = 1$. It follows from the principle of the variation of the argument that the closed loop system has $w_n + P = 2$ poles in the right half plane.

Conditional Stability

The normal situation is that an unstable system can be stabilized simply by reducing the gain. There are however situations where a system can be

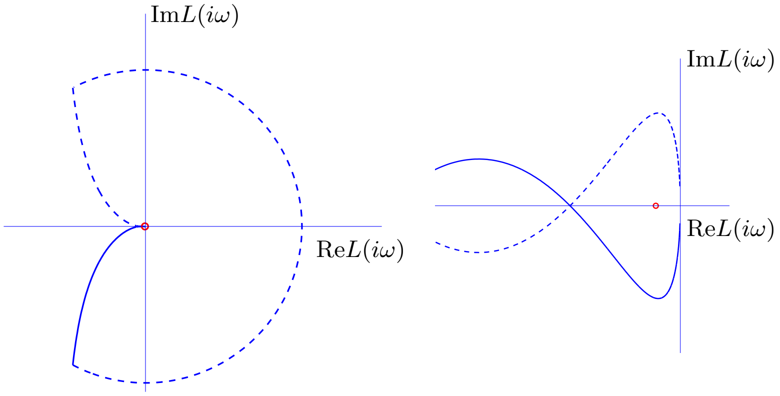


Figure 7.8: Nyquist curve for the loop transfer function $L(s) = \frac{3(s+1)^2}{s(s+6)^2}$. The plot on the right is an enlargement of the area around the origin of the plot on the left.

stabilized increasing the gain. This was first encountered 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. We will illustrate by an example.

Example 7.4 *(Conditional Stability).* Consider a feedback system with the loop transfer function

$$L(s) = \frac{3(s+1)^2}{s(s+6)^2} \tag{7.2}$$

The Nyquist plot of the loop transfer function is shown in Figure 7.8. 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 **??** is strongly misleading in this case. Injection of a sinusoid with frequency 2 rad/s and amplitude 1 at A gives in steady state give an oscillation at B that is in phase with the input and has amplitude 12. Intuitively it is seems unlikely that closing of the loop will result a stable system. It follows from Nyquist's stability criterion that the system is stable because the critical point is to the left of the Nyquist curve when it is traversed for increasing frequencies. It was actually systems of this type which motivated much of the research that led Nyquist to develop his stability criterion.

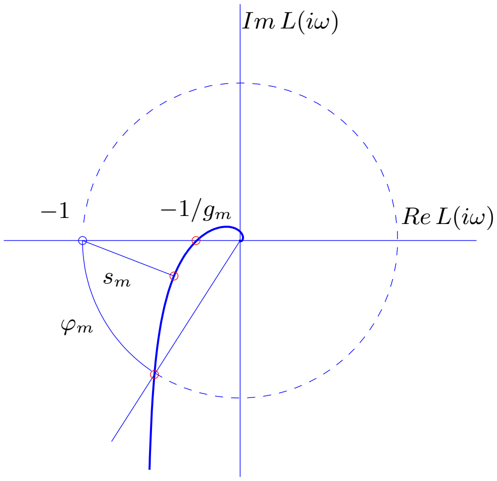


Figure 7.9: Nyquist plot of the loop transfer function L with gain margin g_m, phase margin φ_m and stability margin s_m.

7.4 Stability Margins

In practice it is not enough that the system is stable. There must also be some margins of stability. There are many ways to express this. Many of the criteria are inspired by Nyquist's stability criterion. The key idea is that it is easy plot of the loop transfer function $L(s)$. An increase of controller gain simply expands the Nyquist plot radially. An increase of the phase of the controller twists the Nyquist plot clockwise, see Figure 7.9. Let ω_{180} be the *phase crossover frequency*, which is the smallest frequency where the phase of the loop transfer function $L(s)$ is $-180°$. The gain margin is

$$g_m = \frac{1}{|L(i\omega_{180})|}. \tag{7.3}$$

It tells how much the controller gain can be increased before reaching the stability limit.

Let ω_{gc} be the gain crossover frequency, the lowest frequency where the loop transfer function $L(s)$ has unit magnitude. The phase margin is

$$\varphi_m = \pi + \arg L(i\omega_{gc}), \tag{7.4}$$

the amount of phase lag required to reach the stability limit. The margins have simple geometric interpretations in the Nyquist diagram of the loop transfer function as is shown in Figure 7.9. A drawback with gain and phase margins is that it is necessary to give both of them in order to guarantee that the Nyquist curve not is close to the critical point. One way to express margins by a single number the *stability margin* s_m, which is the shortest distance from the Nyquist curve to the critical point. This number also has other nice interpretations as will be discussed in Chapter 9. Reasonable values of the margins: phase margin $\varphi_m = 30^\circ - 60^\circ$, 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 amplitude ratio of the loop transfer function has several peaks at high frequencies the delay margin is a more relevant measure. A more detailed discussion of robustness measures is given in Chapter 9.

Gain and phase margins can be determined from the Bode plot of the loop transfer function. A change of controller gain translates the gain curve vertically and it has no effect on the phase curve. To determine the gain margin we first find the phase crossover frequency ω_{180} 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 one. The phase margin is the phase of the loop transfer function at that frequency plus 180°. Figure 7.10 illustrates how the margins are found in the Bode plot of the loop transfer function. The stability margin cannot easily be found from the Bode plot of the loop transfer function. There are however other Bode plots that will give s_m. This will be discussed in Chapter 9.

Example 7.5 *(Pupillary light reflex dynamics).* The pupillary light reflex dynamics was discussed in Example 6.9. Stark found a clever way to artificially increase the loop gain by focusing a narrow beam at the boundary of the pupil as shown in Figure 6.7. It was possible to increase the gain so much that the pupil started to oscillate. The Bode plot in Figure 6.8 shows that the phase crossover frequency is $\omega_{gc} = 8$ rad/s. This is in good agreement with Stark's experimental investigations which gave an average frequency of 1.35 Hz or 8.5 rad/s. indicates that oscillations will

7.5 Bode's Relations

An analysis of the Bode plots reveals that there appears to be be a relation between the gain curve and the phase curve. Consider e.g. the Bode plots for the differentiator and the integrator in Figure 6.3. For the differentiator

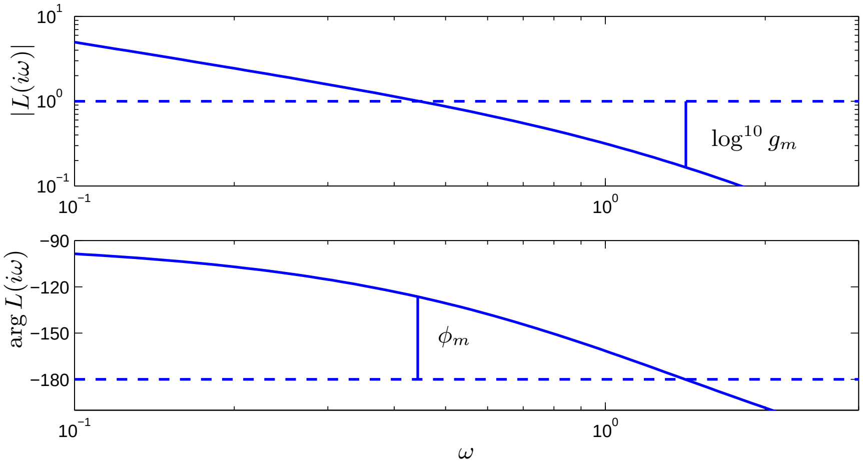


Figure 7.10: Finding gain and phase margins from the Bode plot of the loop transfer function. The loop transfer function is $L(s) = 1/(s(s+1)(s+2))$, the gain margin is $g_m = 6.0$, the gain crossover frequency $\omega_{gc} = 1.42.$, the phase margin is $\phi_m = 53°$ at the phase crossover frequency $\omega = 0.44$.

the slope is +1 and the phase is constant $\pi/2$ radians. For the integrator the slope is -1 and the phase is $-\pi/2$. For the 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 curves for systems with no poles and zeros in the right half plane. He found that the phase was a uniquely given by the gain and vice versa.

$$\begin{aligned}\arg G(i\omega_0) &= \frac{1}{\pi}\int_0^\infty \frac{d\log|G(i\omega)|}{d\log\omega}\log\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right| d\log\omega \\ &= \frac{\pi}{2}\int_0^\infty f(\omega)\frac{d\log|G(i\omega)|}{d\log\omega} d\log\omega \approx \frac{\pi}{2}\frac{d\log|G(i\omega)|}{d\log\omega}\end{aligned} \tag{7.5}$$

where f is the weighting kernel

$$f(\omega) = \frac{2}{\pi^2}\log\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right| = \frac{2}{\pi^2}\log\left|\frac{\frac{\omega}{\omega_0}+1}{\frac{\omega}{\omega_0}-1}\right|$$

The phase curve is thus a weighted average of the derivative of the gain curve. The weight w is shown in Figure 7.11. Notice that the weight falls off rapidly, it is practically zero when frequency has changed by a factor of

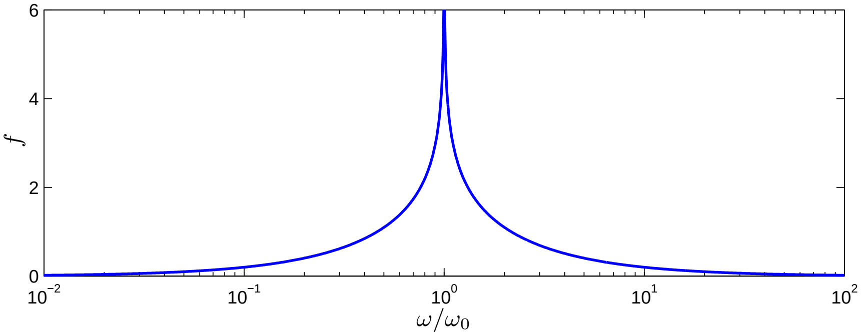


Figure 7.11: The weighting kernel f in Bodes formula for computing the phase curve from the gain curve for minimum phase systems.

ten. It follows from (7.5) that a slope of +1 corresponds to a phase of $\pi/2$ or 90°. Compare with Figure 6.3, where the Bode plots have constant slopes -1 and +1.

Non-minimum Phase Systems

Bode's relations hold for systems that do not have poles and zeros in the left half plane. Such systems are called *minimum phase systems* because systems with poles and zeros in the right half plane have larger phase lag. The distinction is important in practice because minimum phase systems are easier to control than systems with larger phase lag. We will now give a few examples of non-minimum phase.

Example 7.6 *(A Time Delay).* The transfer function of a time delay of T units is $G(s) = e^{-sT}$. This transfer function has unit gain, $|G(i\omega)| = 1$, and the phase is

$$\arg G(i\omega) = -\omega T$$

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 ωT. Notice that the phase lag increases linearly with frequency. Figure 7.12 shows the Bode plot of the transfer function.

It seems intuitively reasonable it is impossible to make a system with a time delay respond faster than the time delay. The presence of a time delay will thus limit the response speed of a system. Next we will consider a system with a zero in the right half plane.

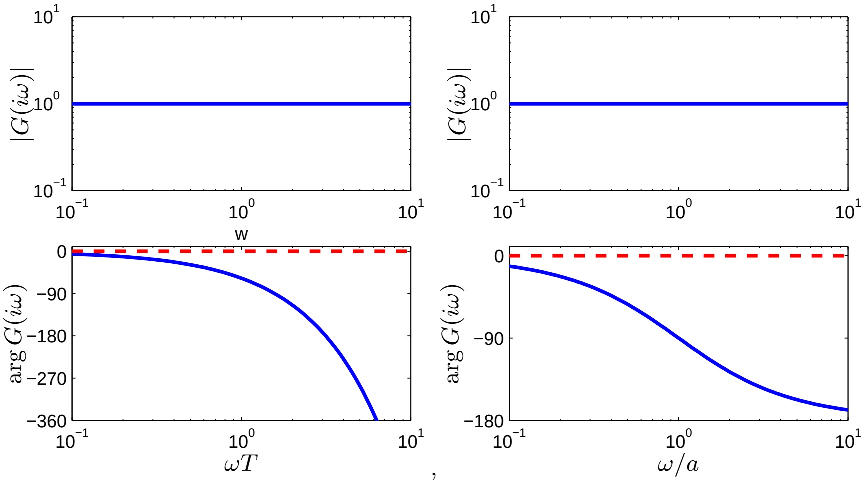


Figure 7.12: Bode plots of a time delay $G(s) = e^{-sT}$ (left) and a system with a right half plane zero $G(s) = (a-s)/(a+s)$ (right). The dashed lines show the phase curves of the corresponding minimum phase systems.

Example 7.7 *(System with a RHP zero).* Consider a system with the transfer function

$$G(s) = \frac{a-s}{a+s}, \quad a > 0$$

which has a zero $s = a$ in the right half plane. The transfer function has unit gain, $|G(i\omega)| = 1$, and

$$\arg G(i\omega) = -2 \arctan \frac{\omega}{a}$$

The corresponding minimum phase system with unit gain has the transfer function $G(s) = 1$. Figure 7.12 shows the Bode plot of the transfer function. The Bode plot resembles the Bode plot for a time delay which is not surprising because the exponential function e^{-sT} can be approximated by

$$e^{-sT} \approx \frac{1 - sT/2}{1 + sT/2}.$$

As far as minimum phase properties are concerned a right half plane zero at $s = a$ is thus similar to a time delay of $T = 2/a$. Since long time delays create difficulties in controlling a system we may expect that systems with zeros close to the origin are also difficult to control.

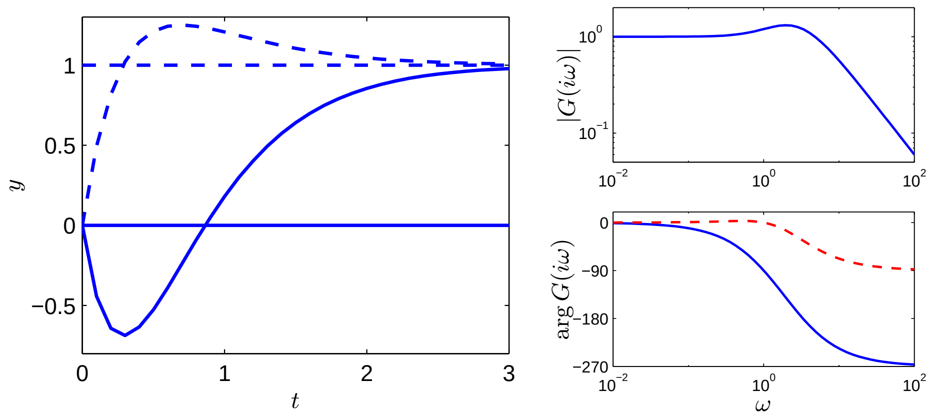


Figure 7.13: Step responses (left) and Bode plots (right) of a system with a zero in the right half plane (full lines) and the corresponding minimum phase system (dashed).

shows the step response of a system with the transfer function

$$G(s) = \frac{6(-s+1)}{s^2+5s+6},$$

which a zero in the right half plane. Notice that the output goes in the wrong direction initially, which is also referred to as inverse response. The figure also shows the step response of the corresponding minimum phase system which has the transfer function

$$G(s) = \frac{6(s+1)}{s^2+5s+6}.$$

The curves show that the minimum phase system responds much faster. It thus appears that a the non-minimum phase system is more difficult to control. This is indeed the case as will be shown in Section 7.7.

The presence of poles and zeros in the right half plane imposes severe limitation on the achieveable performance. Dynamics of this type should be avoided by redesign of the system. 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 not uncommon in practice. We end this section by giving a few examples of such systems.

Example 7.8 *(Backing a Car).* Consider backing a car close to a curb. The transfer function from steering angle to distance from the curve is non-minimum phase. This is a mechanism that is similar to the aircraft.

Example 7.9 *(Flight Control).* The transfer function from elevon to height in an airplane is non-minimum phase. When the elevon is raised there will be a force that pushes the rear of the airplane down. This causes a rotation which gives an increase of the angle of attack and an increase of the lift. Initially the aircraft will however loose height. The Wright brothers understood this and used control surfaces in the front of the aircraft to avoid the effect.

Example 7.10 *(Revenue from Development).* The time behavior of the profits from a new project typically has the behavior shown in Figure 7.13 which indicates that such a system is difficult to control tightly.

7.6 Loop Shaping

One advantage the the Nyquist stability theorem is that it is based on the loop transfer function which is related to the controller transfer function through $L(s) = P(s)C(s)$. It is thus easy to see how the controller influences the loop transfer function. To make an unstable system stable we simply have to bend the Nyquist curve away from the critical point. This simple idea is the basis of several design method called *loop shaping.* The methods are based on the idea of choosing a compensator that gives a loop transfer function with a desired shape. One possibility is to start with the loop transfer function of the process and modify it by changing the gain, and adding poles and zeros to the controller until the desired shape is obtained. Another more direct method is to determine a loop transfer function L_0 which gives the desired robustness and performance. The controller transfer function is then given by

$$C(s) = \frac{L_0(s)}{P(s)}. \tag{7.6}$$

We will first discuss suitable forms of a loop transfer function which gives good performance and good stability margins. Good robustness requires good gain and phase margins. This imposes requirements on the loop transfer function around the crossover frequencies ω_{pc} and ω_{gc}. The gain of L_0 at low frequencies must be large in order to have good tracking of command signals and good rejection of low frequency disturbances. This can be achieved by having a large crossover frequency and a steep slope of the gain curve for the loop transfer function at low frequencies. To avoid injecting too much measurement noise into the system it is desirable that the loop

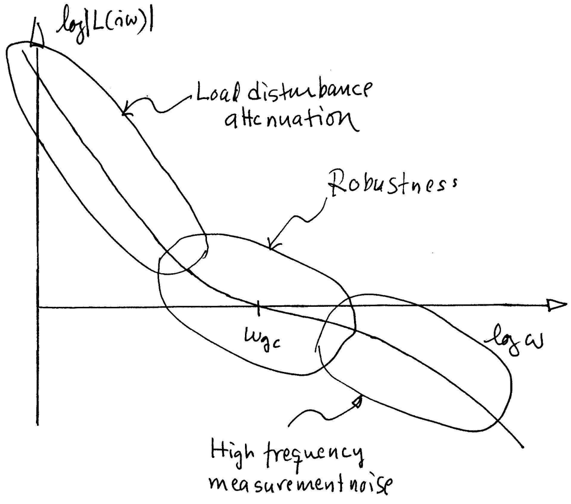

log|L(iω)|
Load disturbance attenuation
Robustness
ω_{gc}
log ω
High frequency measurement noise

Figure 7.14: Gain curve of the Bode plot for a typical loop transfer function. The gain crossover frequency ω_{gc} and the slope n_{gc} of the gain curve at crossover are important parameters.

transfer function has a low gain at frequencies higher than the crossover frequencies. The loop transfer function should thus have the shape indicated in Figure 7.14.

Bodes relations (7.5) impose restrictions on the shape of the loop transfer function. Equation (7.5) implies that the slope of the gain curve at gain crossover cannot be too steep. If the gain curve is constant we have the following relation between slope n and phase margin φ_m

$$n_{gc} = -2 + \frac{\varphi_m}{90}. \tag{7.7}$$

This formula holds approximately when the gain curve does not deviate too much from a straight line. It follows from(7.7) that the phase margins 30°, $45^c irc$ and 60° corresponds to the slopes -5/3, -3/2 and -4/3.

There are many specific design methods that are based on loop shaping. We will illustrate by a design of a PI controller.

Design of a PI Controller

Consider a system with the transfer function

$$P(s) = \frac{1}{(s+1)^4} \tag{7.8}$$

A PI controller has the transfer function

$$C(s) = k + \frac{k_i}{s} = k\frac{1 + sT_i}{sT_i}$$

The controller has high gain at low frequencies and its phase lag is negative for all parameter choices. To have good performance it is desirable to have high gain and a high gain crossover frequency. Since a PI controller has negative phase the gain crossover frequency must be such that the process has phase of at lag smaller than $180 - \varphi_m$, where φ_m is the desired phase margin. For the process (7.8) we have

$$\arg P(i\omega) = -4 \arctan \omega$$

If a phase margin of $\pi/3$ or 60° is required we find that the highest gain crossover frequency that can be obtained with a proportional controller is $\omega_{gc} = \tan \pi/6 = 0.577$. The gain crossover frequency must be lower with PI control.

A simple way to design a PI controller is to specify the gain crossover frequency to be ω_{gc}. This gives

$$L(i\omega) = P(i\omega)C(i\omega) = \frac{kP(i\omega)\sqrt{1 + \omega_{gc}^2 T_i^2}}{\omega_{gc} T_i} = 1$$

which implies

$$k = \frac{\sqrt{1 + \omega_{gc}^2 T_i^2}}{\omega_{gc} T_i P(i\omega_{gc})}$$

We have one equation for the unknowns k and T_i. An additional condition can be obtained by requiring that the PI controller has a phase lag of arctan 0.5 or 45° at the gain crossover, hence $\omega T_i = 0.5$. Figure 7.15 shows the Bode plot of the loop transfer function for $\omega_{gc} = 0.1$, 0.2, 0.3, 0.4 and 0.5. The phase margins for corresponding to these crossover frequencies are 94°, 71°, 49°, 29° and 11°. The gain crossover frequency must be less than 0.26 to have the desired phase margin 60°. Figure 7.15 shows that the controller increases the low frequency gain significantly at low frequencies and that the the phase lag decreases. The figure also illustrates the tradeoff between performance and robustness. A large value of ω_{gc} gives a higher low frequency gain and a lower phase margin. Figure 7.16 shows the Nyquist plots of the loop transfer functions and the step responses of the closed loop system. The response to command signals show that the designs with large

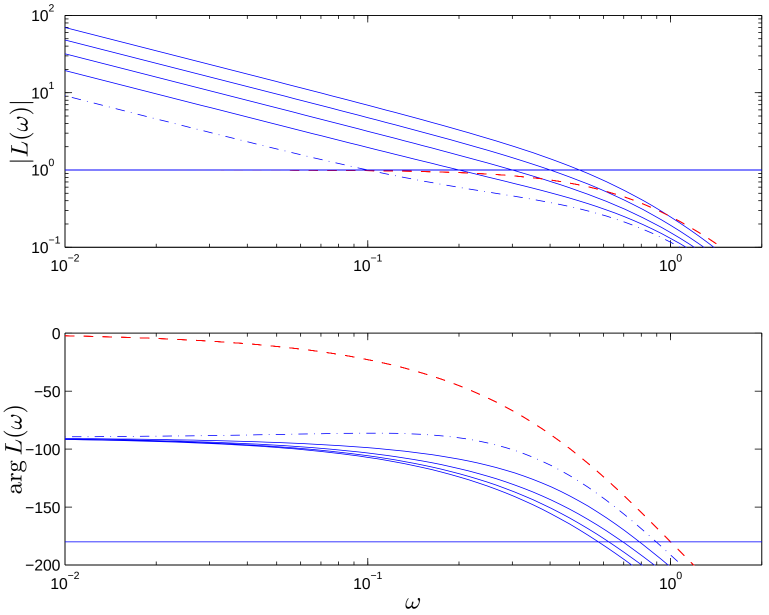


Figure 7.15: Bode plot of the loop transfer function for PI control of a process with the transfer function $P(s) = 1/(s+1)^4$ with $\omega_{gc} = 0.1$ (dash-dotted), 0.2, 0.3, 0.4 and 0.5. The dashed line in the figure is the Bode plot of the process.

ω_{gc} are too oscillatory. A reasonable compromise between robustness and performance is to choose ω_{gc} in the range 0.2 to 0.3. For $\omega_{gc} = 0.25$ we the controller parameters are $k = 0.50$ and $T_i = 2.0$. Notice that the Nyquist plot of the loop transfer function is bent towards the left for low frequencies. This is an indication that integral action is too weak. Notice in Figure 7.16 that the corresponding step responses are also very sluggish.

Bode's Loop Transfer Function

Any loop transfer function can be obtained by proper compensation for processes that have minimum phase. It is then interesting to find out if there is a best form of loop transfer function. When working with feedback amplifiers Bode suggested the following loop transfer function.

$$L(s) = \left(\frac{s}{\omega_{gc}}\right)^{\alpha}. \tag{7.9}$$

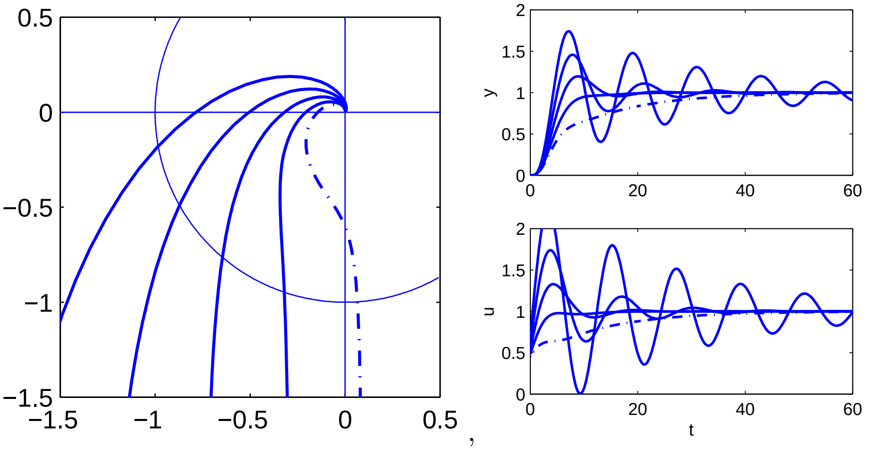


Figure 7.16: Nyquist plot of the loop transfer function for PI control of a process with the transfer function $P(s) = 1/(s+1)^4$ with $\omega_{gc} = 0.1$ (dash-dotted), 0.2, 0.3, 0.4 and 0.5 (left) and corresponding step responses of the closed loop system (right).

The Nyquist curve for this loop transfer function is simply a straight line through the origin with $\arg L(i\omega) = \alpha\pi/2$, see Figure 7.17. Bode called (7.9) the ideal cut-off characteristic. We will simply call it *Bode's loop transfer function*. The reason why Bode choose this particular loop transfer function is that it gives a closed-loop system where the phase margin is constant $\varphi_m = \pi(1 + \alpha/2)$ even if the gain changes and that the amplitude margin is infinite. The slopes $\alpha = -4/3$, -1.5 and $-5/3$ correspond to phase margins of 60°, 45° and 30°. Compare with (7.7). Notice however that the crossover frequency changes when the gain changes.

The transfer function given by Equation (7.9) is an irrational transfer function for non-integer n. It can be approximated arbitrarily close by rational frequency functions. Bode realized that it was sufficient to approximate L over a frequency range around the desired crossover frequency ω_{gc}. Assume for example that the gain of the process varies between k_{min} and k_{max} and that it is desired to have a loop transfer function that is close to (7.9) in the frequency range $(\omega_{min}, \omega_{max})$. It follows from (7.9) that

$$\frac{\omega_{max}}{\omega_{min}} = \left(\frac{k_{max}}{k_{min}}\right)^{1/\alpha}$$

With $\alpha = -5/3$ and a gain ratio of 100 we get a frequency ratio of about 16

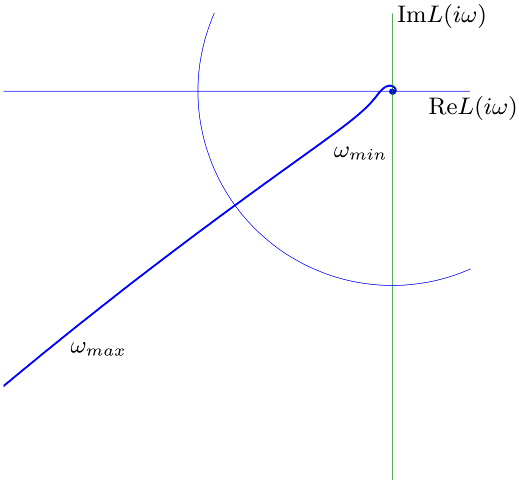


Figure 7.17: Nyquist curve for Bode's loop transfer function. The unit circle is also shown in the figure.

and with $\alpha = -4/3$ we get a frequency ratio of 32. To avoid having too large a frequency range it is thus useful to have α as small as possible. There is, however, a compromise because the phase margin decreases with decreasing α and the system becomes unstable for $\alpha = -2$.

The operational amplifier is a system that approximates the loop transfer function (7.9) with $\alpha = 1$. This amplifier has the useful property that it is stable under a very wide range of feedbacks.

Fractional Systems

It follows from Equation (7.9) that the loop transfer function is not a rational function. We illustrate this with a process having the transfer function.

$$P(s) = \frac{k}{s(s+1)}$$

Assume that we would like to have a closed loop system that is insensitive to gain variations with a phase margin of 45°. Bode's ideal loop transfer function that gives this phase margin is

$$L(s) = \frac{1}{s\sqrt{s}} \tag{7.10}$$

Since $L = PC$ we find that the controller transfer function is

$$C(s) = \frac{s+1}{\sqrt{s}} = \sqrt{s} + \frac{1}{\sqrt{s}} \tag{7.11}$$

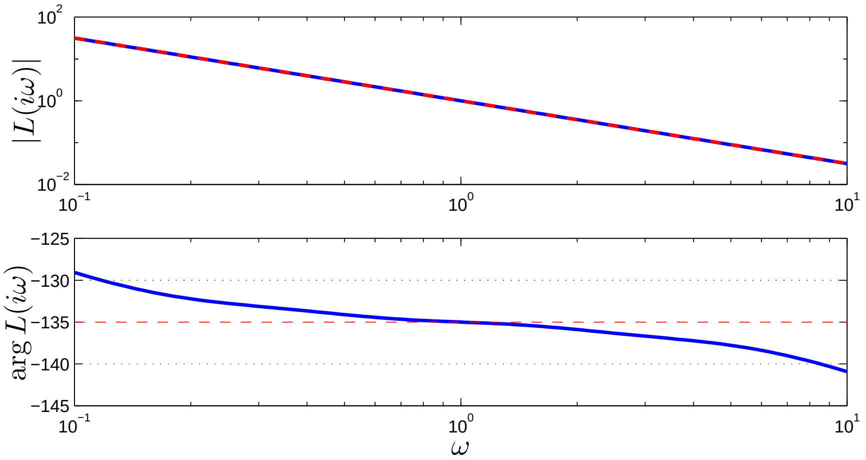


Figure 7.18: Bode diagram of the loop transfer function obtained by approximating the fractional controller (7.11) with the rational transfer function (7.12). The fractional transfer function is shown in full lines and the approximation in dashed lines.

To implement a controller the transfer function is approximated with a rational function. This can be done in many ways. One possibility is the following

$$\hat{C}(s) = k\frac{(s+1/16)(s+1/4)(s+1)^2(s+4)(s+16)}{(s+1/32)(s+1/8)(s+1/2)(s+2)(s+8)(s+32)} \qquad (7.12)$$

where the gain k is chosen to equal the gain of $\sqrt{s}+1/\sqrt{s}$ for $s=i$. Notice that the controller is composed of sections of equal length having slopes 0, +1 and -1 in the Bode diagram. Figure 7.18 shows the Bode diagrams of the loop transfer functions obtained with the fractional controller (7.11) and its approximation (7.12). The differences between the gain curves is not visible in the graph. Figure 7.18 shows that the phase margin is close to 45° even if the gain changes substantially. The differences in phase is less than 2° for a gain variation of 2 orders of magnitude. The range of permissible gain variations can be extended by increasing the order of the approximate controller (7.12). Even if the closed loop system has the same phase margin when the gain changes the response speed will change with the gain. The performance will however change with gain because the response time will vary significantly with gain.

The example shows that robustness is obtained by increasing controller complexity. The range of gain variation that the system can tolerate can be increased by increasing the complexity of the controller.

7.7 Fundamental Limitations

It follows from (7.6) that loop shaping design may cause cancellations of process poles and zeros. The canceled poles and zeros do not appear in the loop transfer function but they appear in the Gang of Four. Cancellations be disastrous if the canceled factors are unstable as was shown in Section 6.7. This implies that there is a major difference between minimum phase and non-minimum phase systems.

To explore the limitations caused by poles and zeros in the right half plane we factor the process transfer function as

$$P(s) = P_{mp}(s)P_{nmp}(s) \tag{7.13}$$

where P_{mp} is the minimum phase part and P_{nmp} is the non-minimum phase part. The factorization is normalized so that $|P_{nmp}(i\omega)| = 1$ and the sign is chosen so that P_{nmp} has negative phase. Let the controller transfer function be $C(s)$, the loop transfer function is then $L(s) = P(s)C(s)$. Requiring that the phase margin is φ_m we get

$$\arg L(i\omega_{gc}) = \arg P_{nmp}(i\omega_{gc}) + \arg P_{mp}(i\omega_{gc}) + \arg C(i\omega_{gc}) \geq -\pi + \varphi_m. \tag{7.14}$$

Let n_{gc} be the slope of the gain curve at the crossover frequency, since $|P_{nmp}(i\omega)| = 1$ it follows that

$$n_{gc} = \left.\frac{d\log|L(i\omega)|}{d\log\omega}\right|_{\omega=\omega_{gc}} = \left.\frac{d\log|P_{mp}(i\omega)C(i\omega)|}{d\log\omega}\right|_{\omega=\omega_{gc}}.$$

The slope n_{gc} is negative and larger than -2 if the system is stable. It follows from Bode's relations Equation 7.5 that

$$\arg P_{mp}(i\omega) + \arg C(i\omega) \approx n\frac{\pi}{2}$$

Combining this with Equation (7.14) gives the following inequality

$$\varphi_\ell = -\arg P_{nmp}(i\omega_{gc}) \leq \pi - \varphi_m + n_{gc}\frac{\pi}{2} \tag{7.15}$$

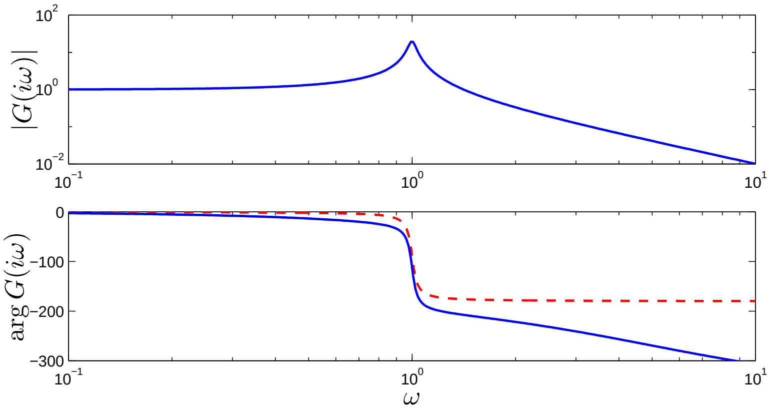


Figure 7.19: Bode plot of process transfer function (full lines) and corresponding minimum phase transfer function (dashed).

This condition which we call the *crossover frequency inequality* shows that the gain crossover frequency must be chosen so that the phase lag of the non-minimum phase component is not too large. To find numerical values we will consider some reasonable design choices. A phase margin of 45° ($\varphi_m = \pi/4$), and a slope $n_{gc} = -1/2$ gives an admissible phase lag of $\varphi_\ell = \pi/2 = 1.57$ and a phase margin of 45° and $n_{gc} = -1$ gives and admissible phase lag $\varphi_\ell = \pi/4 = 0.78$. It is thus reasonable to require that the phase lag of the non-minimum phase part is in the range of 0.5 to 1.6.

The crossover frequency inequality implies that non-minimum phase components impose severe restrictions on possible crossover frequencies. It also means that there are systems that cannot be controlled with sufficient stability margins.

The conditions are more stringent if the process has an uncertainty $\Delta P(i\omega_{gc})$. The admissible phase lag is then reduced by $\arg \Delta P(i\omega_{gc})$.

A straight forward way to use the crossover frequency inequality is to plot the phase of the transfer function of the process and the phase of the corresponding minimum phase system. Such a plot which is shown in Figure 7.19 will immediately show the permissible gain crossover frequencies. As an illustration we will give some analytical examples.

A Zero in the Right Half Plane

The non-minimum phase part of the plant transfer function for a system with a right half plane zero is

$$P_{nmp}(s) = \frac{z-s}{z+s}. \tag{7.16}$$

where $z > 0$. The phase lag of the non-minimum phase part is

$$\varphi_\ell = -\arg P_{nmp}(i\omega) = 2\arctan\frac{\omega}{z}$$

Since the phase of P_{nmp} decreases with frequency the inequality (7.15) gives the following bound on the crossover frequency.

$$\frac{\omega_{gc}}{z} \leq \tan\frac{\varphi_\ell}{2}. \tag{7.17}$$

With reasonable values of φ_ℓ we find that the gain crossover frequency must be smaller than the right half plane zero. It also follows that systems with slow zeros are more difficult to control than system with fast zeros.

A Time Delay

The transfer function of a time delay is

$$P(s) = e^{-sT}. \tag{7.18}$$

This is also the non-minimum phase part P_{nmp}. The phase lag of the non-minimum phase part is

$$\varphi_\ell = -\arg P_{nmp}(i\omega) = \omega T.$$

If the transfer function for the time delay is approximated by

$$e^{-sT} \approx \frac{1-sT/2}{1+sT/2}$$

we find that a time delay T corresponds to a RHP zero $z = -2/T$. A slow zero thus corresponds to a long time delay.

A Pole in the Right Half Plane

The non-minimum phase part of the transfer function for a system with a pole in the right half plane is

$$P_{nmp}(s) = \frac{s+p}{s-p} \tag{7.19}$$

where $p > 0$. The phase lag of the non-minimum phase part is

$$\varphi_\ell = -\arg P_{nmp}(i\omega) = 2\arctan\frac{p}{\omega}.$$

and the crossover frequency inequality becomes

$$\omega_{gc} > \frac{p}{\tan(\varphi_\ell/2)}.$$

With reasonable values of φ_ℓ we find that the gain crossover frequency should be larger than the unstable pole.

A Pole and a Zero in the Right Half Plane

The non-minimum phase part of the transfer function for a system with both poles and zeros in the right half plane is

$$P_{nmp}(s) = \frac{(z-s)(s+p)}{(z+s)(s-p)}. \tag{7.20}$$

The phase lag of this transfer function is

$$\varphi_\ell = -\arg P_{nmp}(i\omega) = 2\arctan\frac{\omega}{z} + 2\arctan\frac{p}{\omega} = 2\arctan\frac{\omega_{gc}/z + p/\omega_{gc}}{1-p/z}.$$

The right hand side has its minimum

$$\min_{\omega_{gc}}\Bigl(2\arctan\frac{\omega_{gc}/z + p/\omega_{gc}}{1-p/z}\Bigr) = 2\arctan\frac{2\sqrt{p/z}}{1-p/z} = 4\arctan\sqrt{\frac{p}{z}},$$

for $\omega = \sqrt{pz}$, and the crossover frequency inequality (7.15) becomes

$$\varphi_\ell = -\arg P_{nmp}(i\omega) \le 4\arctan\sqrt{\frac{p}{z}},$$

or

$$\frac{p}{z} \le \tan\frac{\varphi_\ell}{4}$$

Table 7.1: Achievable phase margin for for $\varphi_m = \pi/4$ and $n_{gc} = -1/2$ and different pole-zero ratios p/z.

p/z	0.45	0.24	0.20	0.17	0.12	0.10	0.05
z/p	2.24	4.11	5.00	5.83	8.68	10	20
φ_m	0	30	38.6	45	60	64.8	84.6

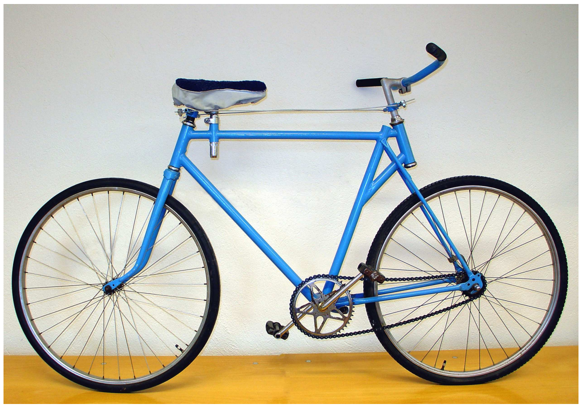


Figure 7.20: Bicycle with rear wheel steering.

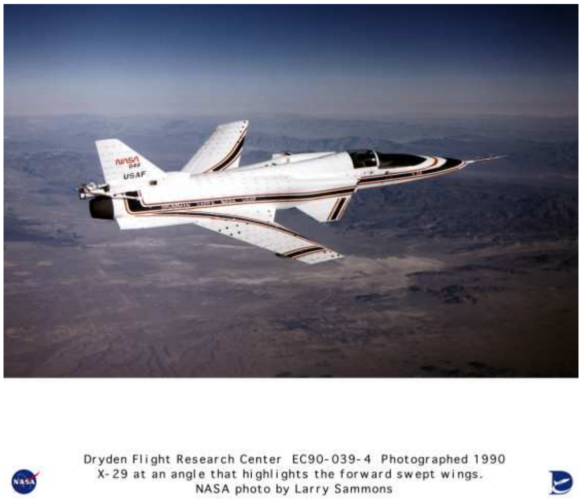
The design choices $\varphi_m = \pi/4$ and $n_{gc} = -1/2$ gives $p < 0.17z$. Table 7.1 shows the admissible pole-zero ratios for different phase margins. The phase-margin that can be achieved for a given ratio p/z is

$$\varphi_m < \pi + n_{gc}\frac{\pi}{2} - 4\arctan\sqrt{\frac{p}{z}}. \tag{7.21}$$

A pair of poles and zeros in the right half plane imposes severe constraints on the gain crossover frequency. The best gain crossover frequency is the geometric mean of the unstable pole and zero. A robust controller does not exist unless the pole-zero ratio is sufficiently small.

Example 7.11 *(The Rear Steered Bicycle).* A bicycle with rear wheel steering is shown in Figure 7.20. The transfer function from steering angle δ to tilt angle θ is

$$G_{\theta\delta}(s) = \frac{mhV}{Jc}\frac{V - as}{s^2 - mgh/J}$$



Dryden Flight Research Center EC90-039-4 Photographed 1990
X-29 at an angle that highlights the forward swept wings.
NASA photo by Larry Sammons

Figure 7.21: The X-29 aircraft.

where m is the total mass of the bicycle and the rider, J the total moment of inertia of bicycle and rider with respect to the contact line of the wheels and the ground, h the height of the center of mass from the ground, a the vertical distance from the center of mass to the contact point of the front wheel, V the forward velocity, and g the acceleration of gravity. The system has a pole at $s = p = \sqrt{mgh/J}$ in the right half plane, caused by the pendulum effect. The system also has a zero at $s = z = V/a$ in the right half plane. The pole-zero ratio is

$$\frac{p}{z} = \frac{a}{V}\sqrt{\frac{mgh}{J}}$$

Typical values are $m = 70$ kg, $h = 1.2$ m, $a = 0.7$, $J = 120$ kgm^2 and $V = 5$ m/s, give $z = V/a = 7.14$ rad/s and $p = \omega_0 = 2.6$ rad/s. The pole-zero ratio is $p = 0.36z$, which is too large. Table 7.1 indicates that the achievable phase margin is much to small. This explains why it practically impossible to read the bicycle.

Example 7.12 *(The X-29).* The X-29 is an experimental aircraft with forward swept wings, see Figure 7.21. Considerable design effort has been devoted to the design of the flight control system for the aircraft. One of the design criteria was that the phase margin should be greater than 45° for all flight conditions. At one flight condition the model has the following non-minimum phase component

$$P_{nmp}(s) = \frac{s-26}{s-6}$$

Since the pole/zero ratio is $6/26 = 0.23$, it follows from Table 7.1 that the achievable phase margin is less than 45°.

A Pole in the Right Half Plane and Time Delay

The non-minimum phase part of the transfer function for a a system with one pole in the right half plane and a time delay T is

$$P_{nmp}(s) = \frac{s+p}{s-p}e^{-sT}. \tag{7.22}$$

This transfer function has the phase lag

$$\varphi_\ell = -\arg P(i\omega) = 2\arctan\frac{p}{\omega_{gc}} + \omega_{gc}T = \pi - 2\arctan\omega p + \omega T. \tag{7.23}$$

The right hand side is larger than π if $pT > 2$ and the system cannot be stabilized. If $pT < 2$ the smallest phase lag

$$\min_{\omega}\Big(2\arctan\frac{p}{\omega_{gc}} + \omega_{gc}T\Big) = 2\arctan\sqrt{\frac{2}{pT}-1} - pT\sqrt{\frac{2}{pT}-1},$$

is obtained for

$$\omega_{gc} = p\sqrt{\frac{2}{pL}-1}.$$

The design choice $\varphi_m = \pi/4$ and $n_{gc} = -0.5$ gives

$$pT \leq 0.326. \tag{7.24}$$

To control the system robustly the product of the time delay and the unstable pole must sufficiently small.

Example 7.13 *(Pole balancing).* Consider balancing of an inverted pendulum. A pendulum of length ℓ has a right half plane pole $\sqrt{g/\ell}$. Assuming that the neural lag of a human is 0.07 s. The inequality (7.24) gives $\ell > 0.45$. The calculation thus indicate that it should possible to balance a pendulum whose length is 0.5 m. To balance a pendulum whose length is 0.1 m the time delay must be less than 0.03s.

Pendulum balancing has also been done using video cameras as angle sensors. The limited video rate imposes strong limitations on what can be achieved. With a video rate of 20 Hz it follows from (**??**) that the shortest pendulum that can be balanced robustly with $\varphi_m = 45°$ and $n_{gc} = -0.5$ is $\ell = 0.23$m.

Avoiding Difficulties with RHP Poles and Zeros

The poles of a system depend on the intrinsic dynamics of the system. They are the eigenvalues of the dynamics matrix A of a linear system. Sensors and actuators have no effect on the poles. The only way to change poles is to redesign a system. Notice that this does not imply that unstable systems should be avoided. Unstable system may actually have advantages, one example is high performance supersonic aircrafts.

The zeros of a system depend on the how sensors and actuators are coupled to the states. The zeros depend on all the matrices A, B, C and D in a linear system. The zeros can thus be influenced by moving sensors and actuators or by adding sensors and actuators. Notice that a fully actuated system $B = I$ does not have any zeros.

7.8 The Small Gain Theorem

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 one for all frequencies. This result can be extended to much more general situations. To do so we need a general concept of gain of a system. For this purpose we first define appropriate classes of input and output signals, $u \in \mathcal{U}$ and $u \in \mathcal{Y}$, where $\mathcal{U}$ and $\mathcal{Y}$ are spaces where a notion of magnitude is defined. The gain of a system is defined as

$$\gamma = \sup_{u \in \mathcal{U}} \frac{||y||}{||u||}.$$

A system is *input-output stable* if the gain is finite. This definition also works for nonlinear systems. Now consider the closed system in Figure 7.22. Let the gains of the systems H_1 and H_2 be γ_1 and γ_2. The *small gain theorem* says that the closed loop system is input output stable if $\gamma_1\gamma_1 <$, and the gain of the closed loop system is

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

Notice that if systems H_1 and H_2 are linear it follows from the Nyquist stability theorem that the closed loop is stable, because if $\gamma_1\gamma_2 < 1$ the Nyquist curve is always inside the unit circle. The small gain theorem is thus an extension of the Nyquist stability theorem.

It also follows from the Nyquist stability theorem that a closed loop system is stable if the phase of the loop transfer function is between $-\pi$ and

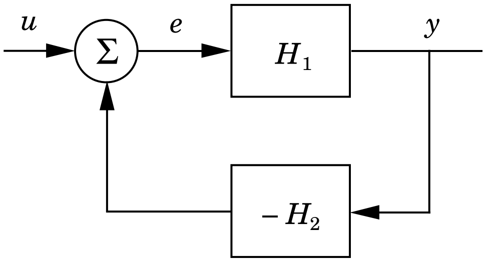


Figure 7.22: A simple feedback loop.

π. This result can also be extended to nonlinear system. The result which is called the *passivity theorem* is closely related to the small gain theorem.

7.9 Further Reading

Nyquists original paper in Bellmans monograph. Bodes book, James Nichols Phillips, Chestnut Meyer, Truxal, kja limitations paper.

7.10 Exercises

1. Backing a car
2. Boeing 747
3. Consider a closed loop system for stabilization of an inverted pendulum with a PD controller. The loop transfer function is

$$L(s) = \frac{s+2}{s^2-1} \tag{7.25}$$

This transfer function has one pole at $s = 1$ in the right half plane. The Nyquist plot of the loop transfer function is shown in Figure 7.23. Traversing the contour Γ in clockwise we find that the winding number is -1. Applying Theorem 1 we find that

$$N - P = -1$$

Since the loop transfer function has a pole in the right half plane we have $P = 1$ and we get $N = 0$. The characteristic equation thus has no roots in the right half plane and the closed loop system is stable.

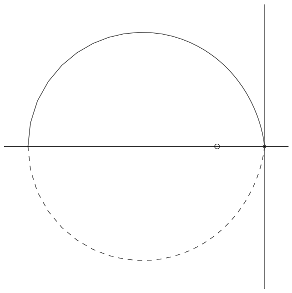


Figure 7.23: Map of the contour Γ under the map $L(s) = \frac{s+2}{s^2-1}$ given by (7.25). The map of the positive imaginary axis is shown in full lines, the map of the negative imaginary axis and the small semi circle at the origin in dashed lines.