Chapter 11

Loop Shaping

Quotation

Authors, citation.

In this chapter we continue to explore the use of frequency domain techniques for design of feedback systems. We begin with a more thorough description of the performance specifications for control systems, and then introduce the concept of "loop shaping" as a mechanism for designing controllers in the frequency domain. We also introduce some fundamental limitations to performance for systems with right half plane poles and zeros.

11.1 A Basic Feedback Loop

In the previous chapter, we considered the use of PID feedback as a mechanism for designing a feedback controller for a given process. In this chapter we will expand our approach to include a richer repertoire of tools for shaping the frequency response of the closed loop system.

One of the key ideas in this chapter is that we can design the behavior of the closed loop system by studying the open loop transfer function. This same approach was used in studying stability using the Nyquist criterion: we plotted the Nyquist plot for the *open* loop transfer function to determine the stability of the *closed* loop system. From a design perspective, the use of loop analysis tools is very powerful: since the loop transfer function is $L = PC$, if we can specify the desired performance in terms of properties of L, we can directly see the impact of changes in the controller C. This is much easier, for example, than trying to reason directly about the response
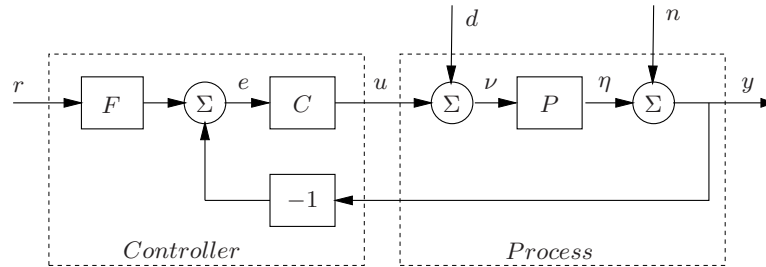
Figure 11.1: Block diagram of a basic feedback loop.

of the closed loop system, whose transfer function is given by

$$G_{yr} = \frac{PC}{1 + PC}$$

(assuming $F = 1$).

We will start by investigating some key properties of the feedback loop. A block diagram of a basic feedback loop is shown in Figure 11.1. The system loop is composed of two components, the process and the controller, and the controller has two blocks: the feedback block C and the feedforward block F. There are two disturbances acting on the process, the *load disturbance*, d, and the *measurement noise*, n. The load disturbance represents disturbances that drive the process away from its desired behavior, while the measurement noise represents the uncertainty in sensing the output of the system. In the figure, the load disturbance is assumed to act on the process input. This is a simplification, since disturbances often enter the process in many different ways, but allows us to streamline the presentation without significant loss of generality.

The process output η is the real physical variable that we want to control. Control is based on the measured signal y, where the measurements are corrupted by measurement noise n. The process is influenced by the controller via the control variable u. The process is thus a system with three inputs—the control variable u, the load disturbance d and the measurement noise n—and one output—the measured signal. The controller is a system with two inputs and one output. The inputs are the measured signal y and the reference signal r and the output is the control signal u. Note that the control signal u is an input to the process and the output of the controller, and that the measured signal is the output of the process and an input to the controller.

The feedback loop in Figure 11.1 is influenced by three external signals, the reference r, the load disturbance d and the measurement noise n. There are at least three signals, η, y and u that are of great interest for control, giving nine relations between the input and the output signals. Since the system is linear, these relations can be expressed in terms of the transfer functions. The following relations are obtained from the block diagram in Figure 11.1:

$$
\begin{pmatrix} w \\ y \\ u \end{pmatrix} = \begin{pmatrix} \frac{P}{1+PC} & -\frac{PC}{1+PC} & \frac{PCF}{1+PC} \\ \frac{P}{1+PC} & \frac{1}{1+PC} & \frac{PCF}{1+PC} \\ -\frac{PC}{1+PC} & -\frac{C}{1+PC} & \frac{CF}{1+PC} \end{pmatrix} \begin{pmatrix} d \\ n \\ r \end{pmatrix}. \tag{11.1}
$$

To simplify notations we have dropped the arguments of all transfer functions.

There are several interesting conclusions we can draw from these equations. First we can observe that several transfer functions are the same and that all relations are given by the following set of six transfer functions, which we call the *Gang of Six*:

$$
\begin{array}{ccc}
\dfrac{PCF}{1+PC} & \dfrac{PC}{1+PC} & \dfrac{P}{1+PC} \\[3mm]
\dfrac{CF}{1+PC} & \dfrac{C}{1+PC} & \dfrac{1}{1+PC}.
\end{array} \tag{11.2}
$$

The transfer functions in the first column give the response of the process output and control signal to the setpoint. The second column gives the same signals in the case of pure error feedback when $F = 1$. The transfer function $P/(1 + PC)$, in the third column, tells how the process variable reacts to load disturbances and the transfer function $C/(1 + PC)$, in the second column, gives the response of the control signal to measurement noise. Notice that only four transfer functions are required to describe how the system reacts to load disturbances and the measurement noise, and that two additional transfer functions are required to describe how the system responds to setpoint changes.

The linear behavior of the system is determined by six transfer functions in equation (11.2) and specifications can be expressed in terms of these transfer functions. The special case when $F = 1$ is called a system with (pure) error feedback. In this case all control actions are based on feedback from the error only and the system is completely characterized by four transfer functions, namely the four rightmost transfer functions in equation (11.2),

which have specific names:

$$S = \frac{1}{1 + PC} \qquad \text{sensitivity function}$$

$$T = \frac{PC}{1 + PC} \qquad \text{complementary sensitivity function}$$

$$PS = \frac{P}{1 + PC} \qquad \text{load sensitivity function}$$

$$CS = \frac{C}{1 + PC} \qquad \text{noise sensitivity function}$$

(11.3)

These transfer functions and their equivalent systems are called the *Gang of Four*. The load disturbance sensitivity function is sometimes called the input sensitivity function and the noise sensitivity function is sometimes called the output sensitivity function. These transfer functions have many interesting properties that will be discussed in detail in the rest of the chapter and good insight into these properties is essential for understanding feedback systems.

The procedure for designing a controller for the system in Figure 11.1 can be divided into two independent steps:

1. Design the feedback controller C that reduces the effects of load disturbances and the sensitivity to process variations without introducing too much measurement noise into the system.

2. Design the feedforward F to give the desired response to the reference signal (or setpoint).

The properties of the system can be expressed in terms of properties of the transfer functions (11.3), as illustrated in the following example.

Example 11.1. Consider the process

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

with a PI feedback controller

$$C(s) = 0.775 + \frac{1}{2.05s}$$

and a feedforward controller
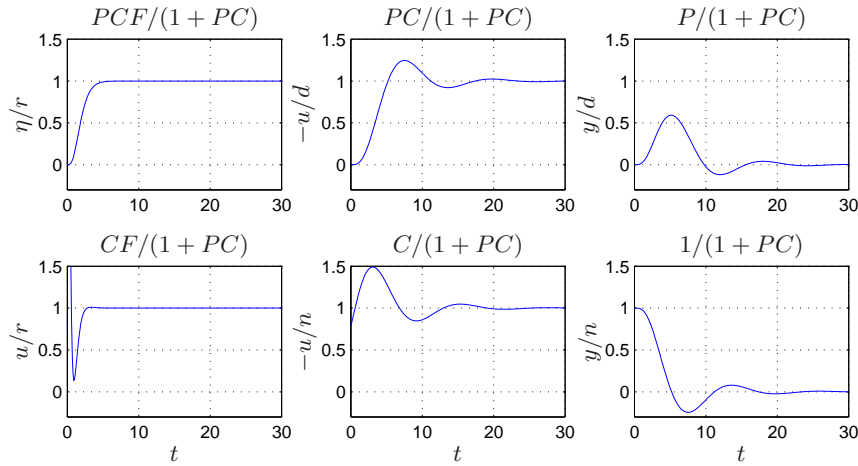
$$F(s) = \frac{1}{(0.5s + 1)^4}.$$

Figure 11.2: Step responses of the Gang of Six for PI control $k = 0.775$, $T_i = 2.05$ of the process $P(s) = (s + 1)^{-4}$. The feedforward is designed to give the transfer function $(0.5s + 1)^{-4}$ from reference r to output y.

Figures 11.2 and 11.3 show the step and frequency responses for the Gang of Six and give useful insight into the properties of the closed loop system.

The time responses in Figure 11.2 show that the feedforward gives a substantial improvement of the response speed as seen by the differences between the first and second columns. The settling time is substantially shorter with feedforward, 4 s versus 25 s, and there is no overshoot. This is also reflected in the frequency responses in Figure 11.3, which show that the transfer function with feedforward has higher bandwidth and that it has no resonance peak.

The transfer functions $CF/(1 + PC)$ and $-C/(1 + PC)$ represent the signal transmission from reference to control and from measurement noise to control. The time responses in Figure 11.2 show that the reduction in response time by feedforward requires a substantial control effort. The initial value of the control signal is out of scale in Figure 11.2 but the frequency response in Figure 11.3 shows that the high frequency gain of $PCF/(1+PC)$ is 16, which can be compared with the value 0.78 for the transfer function $C/(1 + PC)$. The fast response thus requires significantly larger control signals.

There are many other interesting conclusions that can be drawn from Figures 11.2 and 11.3. Consider for example the response of the output to load disturbances expressed by the transfer function $P/(1 + PC)$. The
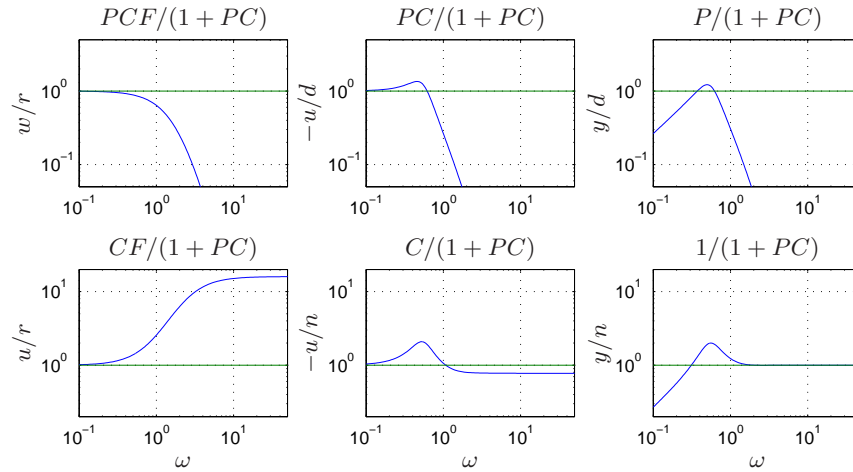
Figure 11.3: Gain curves of frequency responses of the Gang of Six for PI control $k = 0.775$, $T_i = 2.05$ of the process $P(s) = (s+1)^{-4}$ where the feedforward has been designed to give the transfer function $(0.5s+1)^{-4}$ from reference to output.

frequency response has a pronounced peak 1.22 at $\omega_{max} = 0.5$ and the corresponding time function has its maximum 0.59 at $t_{max} = 5.2$. Notice that the peaks are of the same magnitude and that the product of $\omega_{max}t_{max} = 2.6$. Similar relations hold for the other responses. ∇

11.2   Performance Specifications

A key element of the control design process is how we specify the desired performance of the system. Inevitably the design process requires a tradeoff between different features of the closed loop system and specifications are the mechanism by which we describe the desired outcome of those tradeoffs.

Frequency Domain Specifications

One of the main methods of specifying the performance of a system is through the frequency response of various input/output pairs. Since specifications were originally focused on setpoint response, it was natural to consider the transfer function from reference input to process output. For a system with error feedback, the transfer function from reference to output is equal to the complementary transfer function, $T = PC/(1+PC)$. A typical gain curve for this response is shown in Figure 11.4. Good performance
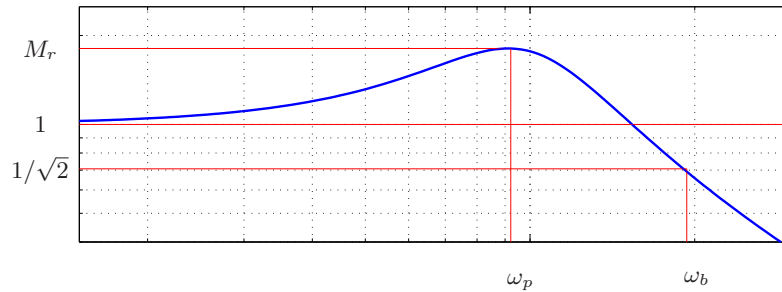
Figure 11.4: Gain curve for transfer function from setpoint to output.

requires that the zero frequency gain is one (so that the output tracks the reference). Typical specification measures include:

- The *resonance peak*, M_r, is the largest value of the frequency response.

- The *peak frequency*, ω_p, is the frequency where the maximum occurs.

- The *bandwidth*, ω_b, is the frequency where the gain has decreased to $1/\sqrt{2}$.

Specifications can also be related to the loop transfer function, $L = PC$. Useful features that have been discussed previously are:

- The *gain crossover frequency*, ω_{gc}, is the lowest frequency where the loop transfer function L has unit magnitude. This is roughly equal to the frequency where the closed loop gain drops to below $1/\sqrt{2}$.

- The *gain margin*, g_m, is the amount that the loop gain can be increased before reaching the stability limit. A high gain margin insures that errors in modeling the gain of the system do not lead to instability.

- The *phase margin*, φ_m, is the amount of phase lag required to reach the stability limit. A phase margin of $30°$ to $60°$ is typically required for robustness to modeling errors and non-oscillatory response.

These concepts were given in more detail in Section 9.3.

In addition to specifications on the loop transfer function, there are also a number of useful specifications on the sensitivity function and the complementary sensitivity function:

- The *maximum sensitivity*, M_s, is the peak value of the magnitude of sensitivity function and indicates the maximum amplification from the reference to the error signal.

- The *maximum sensitivity frequency*, ω_{ms}, is the frequency where the sensitivity function has its maximum.

- The *sensitivity crossover frequency*, ω_{sc}, is the frequency where the sensitivity function becomes greater than 1 for the first time. Disturbances are attenuated below this frequency and can be amplified above this frequency.

- The *maximum complementary sensitivity*, M_t, is the peak value of the magnitude of the complementary sensitivity function. It provides the maximum amplification from the reference signal to the output signal.

- The *maximum complementary sensitivity frequency*, ω_{mt}, is the frequency where the complementary sensitivity function has its maximum.

As we will see in the rest of the chapter, these various measures can be used to gain insights into the performance of the closed loop system and are often used to specify the desired performance for a control design.

Although we have defined different specifications for the loop transfer function L, the sensitivity function S and the complementary sensitivity function T, these transfer functions are all related through a set of algebraic relationships:

$$S = \frac{1}{1 + L} \qquad T = \frac{L}{1 + L} \qquad S + T = 1.$$

These relationships can limit the ability to independently satisfy specifications for the quantities listed above and may require tradeoffs, as we shall see.

Relations between Time and Frequency Domain Features

In Section 5.3 we described some of the typical parameters that described the step response of a system. These included the rise time, steady state error, and overshoot. For many applications, it is natural to provide these time domain specifications and we can relate these to the eigenvalues of the closed loop system, which are equivalent to the poles of the transfer function $T = PC/(1 + PC)$.

There are approximate relations between specifications in the time and frequency domain. Let $G(s)$ be the transfer function from reference to output. In the time domain the response speed can be characterized by the rise time T_r and the settling time T_s. In the frequency domain the response time

can be characterized by the closed loop bandwidth ω_b, the gain crossover frequency ω_{gc}, the sensitivity frequency ω_{ms}. The product of bandwidth and rise time is approximately constant $T_r \omega_b \approx 2$, so decreasing the rise time corresponds to increasing the closed loop bandwidth.

The overshoot of the step response M_p is related to the resonant peak M_r of the frequency response in the sense that a larger peak normally implies a larger overshoot. Unfortunately there is no simple relation because the overshoot also depends on how quickly the frequency response decays. For $M_r < 1.2$ the overshoot M_p in the step response is often close to $M_r - 1$. For larger values of M_r the overshoot is typically less than $M_r - 1$. These relations do not hold for all systems: there are systems with $M_r = 1$ that have a positive overshoot. These systems have transfer functions that decay rapidly around the bandwidth. To avoid overshoot in systems with error feedback it is advisable to require that the maximum of the complementary sensitivity function is small, say $M_t = 1.1 - 1.2$.

Response to Load Disturbances

The sensitivity function in equation (11.3) shows how feedback influences disturbances. Disturbances with frequencies that are lower than the sensitivity crossover frequency ω_{sc} are attenuated by feedback and those with $\omega > \omega_{sc}$ are amplified by feedback. The largest amplification is the maximum sensitivity M_s.

Consider the system in Figure 11.1. The transfer function from load disturbance d to process output w is

$$G_{wd} = \frac{P}{1 + PC} = PS = \frac{T}{C}. \tag{11.4}$$

Since load disturbances typically have low frequencies, it is natural that the criterion emphasizes the behavior of the transfer function at low frequencies. Filtering of the measurement signal has only marginal effect on the attenuation of load disturbances because the filter typically only attenuates high frequencies. For a system with $P(0) \neq 0$ and a controller with integral action, the controller gain goes to infinity for small frequencies and we have the following approximation for small s:

$$G_{wd} = \frac{T}{C} \approx \frac{1}{C} \approx \frac{s}{k_i}. \tag{11.5}$$

Figure 11.5 gives the gain curve for a typical case and shows that the approximation is very good for low frequencies.
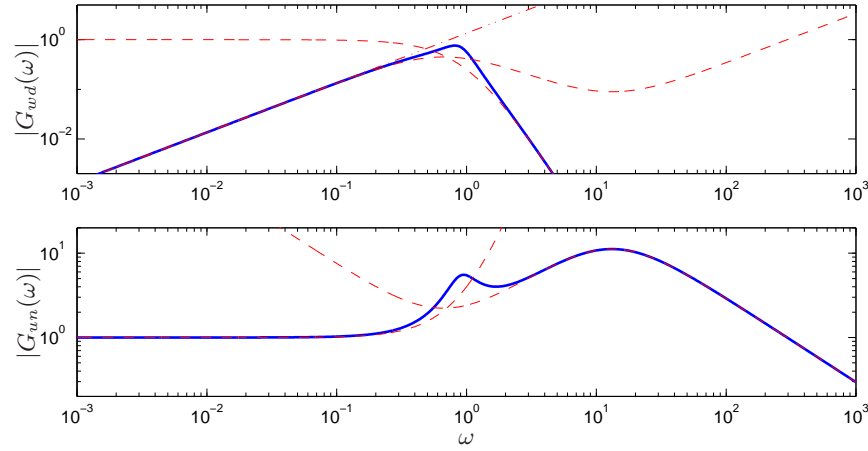
Figure 11.5: Gains of the transfer functions G_{wd} and G_{un} for PID control ($k = 2.235$, $T_i = 3.02$, $T_i = 0.756$ and $T_f = Td/5$) of the process $P = (s+1)^{-4}$. The gain of the transfer functions P, C, $1/C$ are shown with dashed lines and s/k_i with dash-dotted lines.

Measurement noise, which typically has high frequencies, generates rapid variations in the control variable that are detrimental because they cause wear in many actuators and they can even saturate the actuator. It is thus important to keep the variations in the control signal at reasonable levels— a typical requirement is that the variations are only a fraction of the span of the control signal. The variations can be influenced by filtering and by proper design of the high frequency properties of the controller.

The effects of measurement noise are captured by the transfer function from measurement noise to the control signal,

$$G_{un} = \frac{C}{1 + PC} = CS = \frac{T}{P}. \tag{11.6}$$

Figure 11.5 shows the gain curve of G_{un} for a typical system. For low frequencies the transfer function the sensitivity function equals 1 and equation (11.6) can be approximated by $1/P$. For high frequencies is is approximated as $G_{un} \approx C$. A simple measure of the effect of measurement noise is the high frequency gain of the transfer function G_{un},

$$M_{un} := \|G_{un}\|_\infty = \sup_\omega |G_{un}(j\omega)|. \tag{11.7}$$
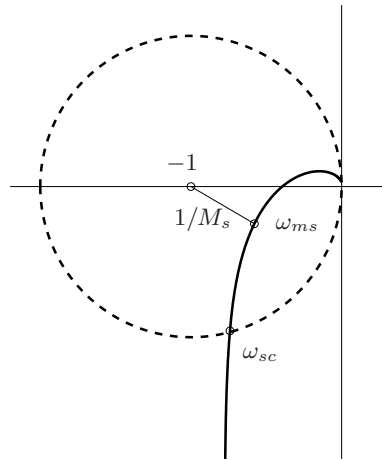
Figure 11.6: Nyquist curve of loop transfer function showing graphical interpretation of maximum sensitivity. The sensitivity crossover frequency ω_{sc} and the frequency ω_{ms} where the sensitivity has its largest value are indicated in the figure. All points inside the dashed circle have sensitivities greater than 1.

The sensitivity function can be written as

$$S = \frac{1}{1 + PC} = \frac{1}{1 + L}. \tag{11.8}$$

Since it only depends on the loop transfer function it can also be visualized graphically using the Nyquist plot of the loop transfer function. This is illustrated in Figure 11.6. The complex number $1 + L(j\omega)$ can be represented as the vector from the point -1 to the point $L(j\omega)$ on the Nyquist curve. The sensitivity is thus less than one for all points outside a circle with radius 1 and center at -1. Disturbances of these frequencies are attenuated by the feedback. If a control system has been designed based on a given model, it is straightforward to estimate the potential disturbance reduction simply by recording a typical output and filtering it through the sensitivity function.

Example 11.2. Consider the same system as the previous example

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

with a PI controller. Figure 11.7 shows the gain curve of the sensitivity function for $k = 0.8$ and $k_i = 0.4$. The figure shows that the sensitivity crossover frequency is 0.32 and that the maximum sensitivity 2.1 occurs at
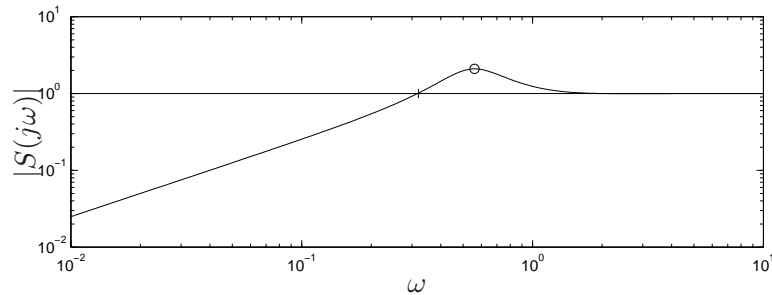
Figure 11.7: Gain curve of the sensitivity function for PI control ($k = 0.8$, $k_i = 0.4$) of process with the transfer function $P(s) = (s + 1)^{-4}$. The sensitivity crossover frequency is indicated by $+$ and the maximum sensitivity by o.

$\omega_{ms} = 0.56$. Feedback will thus reduce disturbances with frequencies less than 0.32 rad/s, but it will amplify disturbances with higher frequencies. The largest amplification is 2.1. ∇

11.3   Feedback Design via 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 $L = PC$. 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 different design methods, collectively 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.

Design Considerations

We will first discuss suitable forms of a loop transfer function that give 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 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
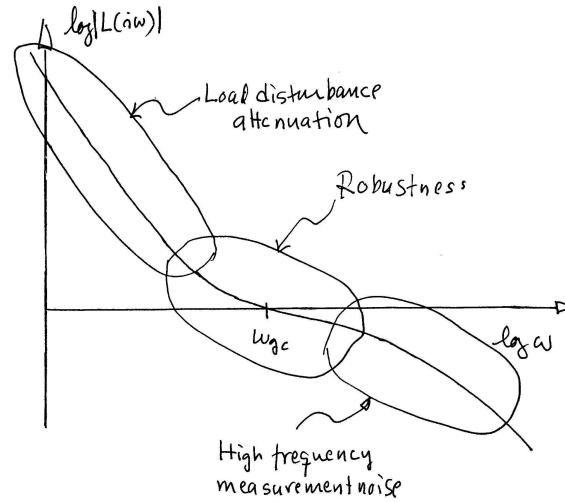
Figure 11.8: Gain curve of the Bode plot for a typical loop transfer function. The gain crossover frequency ω_{gc} and the slope n_{gc} of the gain curve at crossover are important parameters.

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 transfer function have a low gain at frequencies higher than the crossover frequencies. The loop transfer function should thus have the shape indicated in Figure 11.8.

Bode's relations (see Section 9.4) impose restrictions on the shape of the loop transfer function. Equation (9.5) 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_{gc} and phase margin φ_m:

$$n_{gc} = -2 + \frac{2\varphi_m}{\pi}. \tag{11.9}$$

This formula holds approximately when the gain curve does not deviate too much from a straight line. It follows from equation (11.9) that the phase margins 30°, 45° 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 the basic approach by the design of a PI controller.

Example 11.3 (Design of a PI controller)**.** Consider a system with the

transfer function

$$P(s) = \frac{1}{(s+1)^4}.$$ (11.10)

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 lag smaller than $180 - \varphi_m$, where φ_m is the desired phase margin. For the process (11.10) we have

$$\angle P(j\omega) = -4\arctan\omega$$

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 a PI controller.

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

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

which implies

$$k_p = \frac{\sqrt{1 + \omega_{gc}^2 T_i^2}}{\omega_{gc}T_i P(j\omega_{gc})}.$$

We have one equation for the unknowns k and T_i. An additional condition can be obtained by requiring that the PI controller have a phase lag of $45°$ at the gain crossover, hence $\omega T_i = 0.5$. Figure 11.9 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 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 11.9 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
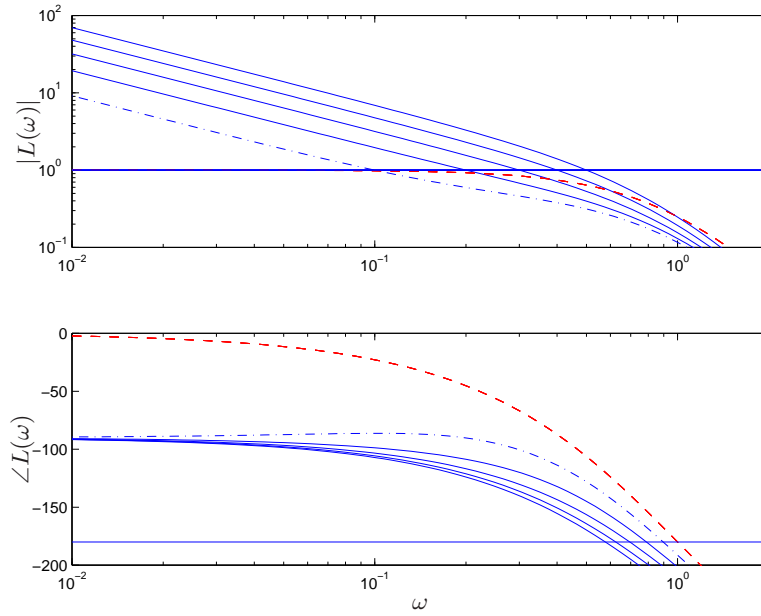
Figure 11.9: Bode plot of the loop transfer function for PI control of a process with the transfer function $P(s) = 1/(s + 1)^4$ with $\omega_{gc} = 0.1$ (dash-dotted), 0.2, 0.3, 0.4 and 0.5. The dashed line in the figure is the Bode plot of the process.
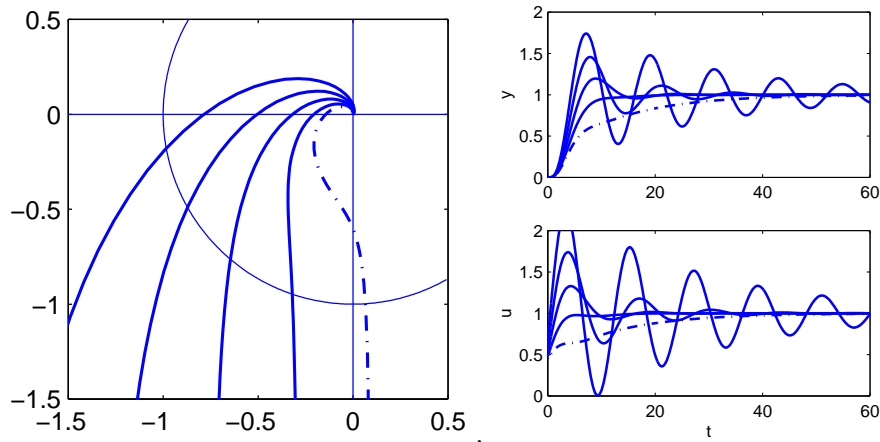

Figure 11.10: Nyquist plot of the loop transfer function for PI control of a process with the transfer function $P(s) = 1/(s + 1)^4$ with $\omega_{gc} = 0.1$ (dash-dotted), 0.2, 0.3, 0.4 and 0.5 (left) and corresponding step responses of the closed loop system (right).

a lower phase margin. Figure 11.10 shows the Nyquist plots of the loop transfer functions and the step responses of the closed loop system. The responses to command signals show that the designs with large 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 $\omega_{gc} = 0.25$, 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 11.10 that the corresponding step responses are also very sluggish. ∇

Lead Compensation

A common problem in design of feedback systems is that the phase lag of the system at the desired crossover frequency is not high enough to allow either proportional or integral feedback to be used effectively. Instead, one may have a situation where you need to add phase *lead* to the system, so that the crossover frequency can be increased.

A standard way to accomplish this is to use a *lead compensator*, which has the form

$$C(s) = k\frac{s+a}{s+b} \qquad a < b. \tag{11.11}$$

The transfer function corresponding to this controller is shown in Figure 11.11. A key feature of the lead compensator is that it adds phase *lead* in the frequency range between the pole/zero pair (and extending approximately 10X in frequency in each direction). By appropriately choosing the location of this phase lead, we can provide additional phase margin at the gain crossover frequency.

Because the phase of a transfer function is related to the slope of the magnitude, increasing the phase requires increasing the gain of the loop transfer function over the frequency range in which the lead compensation is applied. Hence we can also think of the lead compensator as changing the slope of the transfer function and thus shaping the loop transfer function in the crossover region (although it can be applied elsewhere as well).

Example 11.4 (Pitch control for a ducted fan)**.** Consider the control of the pitch (angle) of a vertically oriented ducted fan, as shown in Figure 11.12. We model the system with a second order transfer function of the form

$$P = \frac{r}{Js^2 + ds + mgl},$$

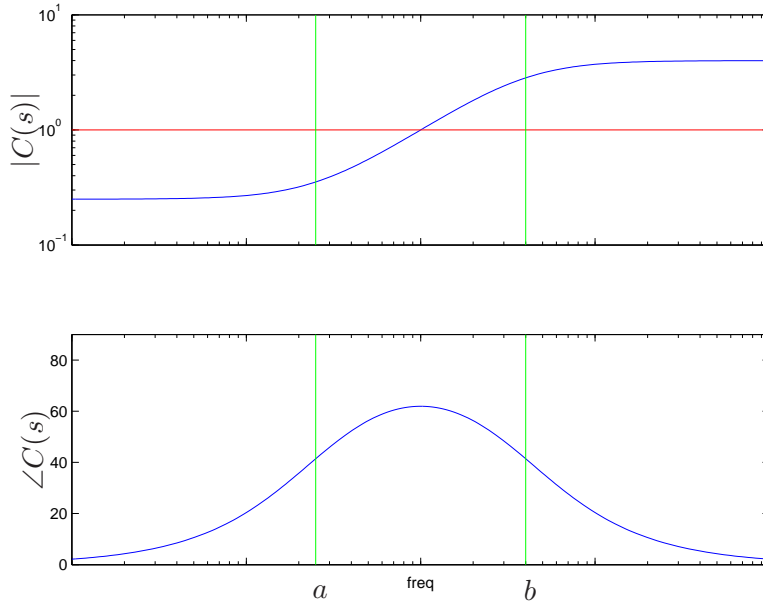with the parameters given in Table 11.1.   We take as our performance

Figure 11.11: Frequency response for a lead compensator, $C(s) = k(s+a)/(s+b)$.

specification that we would like less than 1% error in steady state and less than 10% tracking error up to 10 rad/sec.

The open loop transfer function is shown in Figure 11.13a. To achieve our performance specification, we would like to have a gain of at least 10 at a frequency of 10 rad/sec, requiring the gain crossover frequency to be at a higher frequency. We see from the loop shape that in order to achieve the
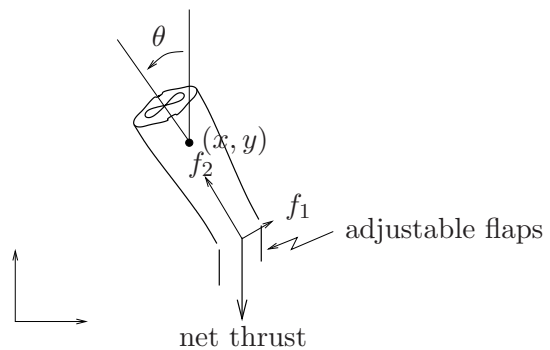

Figure 11.12: Caltech ducted fan with support stand.

Symbol	Description	Value	
m	inertial mass of fan, x axis	4.0	kg
J	fan moment of inertia, φ_3 axis	0.0475	kg m^2
r	nominal distance of flaps from fan pivot	26.0	cm
d	angular damping factor	0.001	kg m/s
g	gravitational constant	9.8	m/sec^2

Table 11.1: Parameter values for the planar ducted fan model which approximate the dynamics of the Caltech ducted fan.

desired performance we cannot simply increase the gain, since this would give a very low phase margin. Instead, we must increase the phase at the desired crossover frequency.

To accomplish this, we use a lead compensator (11.11) with $a = 2$ and $b = 50$. We then set the gain of the system to provide a large loop gain up to the desired bandwidth, as shown in Figure 11.13b. We see that this system has a gain of greater than 10 at all frequencies up to 10 rad/sec and that it has over 40° degrees of phase margin. ∇

The action of a lead compensator is essentially the same as that of the derivative portion of a PID controller. As described in Section 10.5, we often use a filter for the derivative action of a PID controller to limit the high frequency gain. This same effect is present in a lead compensator through the pole at $s = b$.

Equation (11.11) is a first order lead compensator and can provide up to 90° of phase lead. Higher levels of phase lead can be provided by using a second order lead compensator:

$$C = k \frac{(s+a)^2}{(s+b)^2} \qquad a < b.$$

11.4 Fundamental Limitations

Although loop shaping gives us a great deal of flexibility in designing the closed loop response of a system, there are certain fundamental limits on what can be achieved. We consider here some of the primary performance limitations that can occur; additional limitations having to do with robustness are considered in the next chapter.

One of the key limitations of loop shaping occurs when we have the possibility of cancellation of right half plane poles and zeros. The canceled
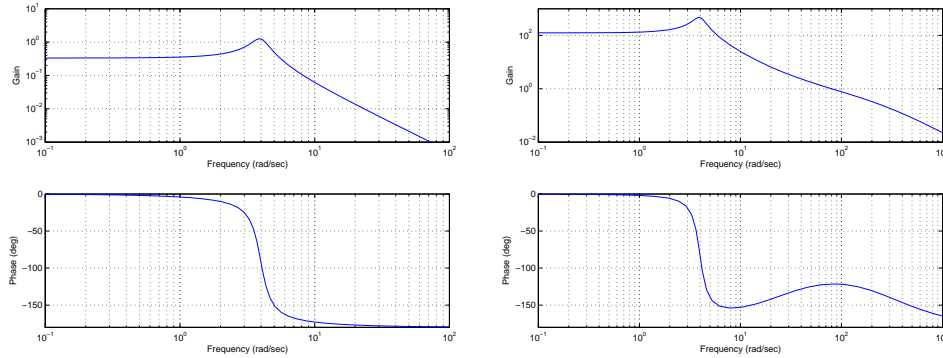
Figure 11.13: Control design using a lead compensator: (a) Bode plot for P and (b) Bode plot for $L = PC$ using a lead compensator.

poles and zeros do not appear in the loop transfer function but they can appear in the transfer functions from disturbances to outputs or control signals. Cancellations can be disastrous if the canceled factors are unstable, as was shown in Section 7.5. 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

$$P(s) = P_{mp}(s)P_{nmp}(s), \tag{11.12}$$

where P_{mp} is the minimum phase part and P_{nmp} is the non-minimum phase part. The factorization is normalized so that $|P_{nmp}(j\omega)| = 1$ and the sign is chosen so that P_{nmp} has negative phase. Requiring that the phase margin is φ_m we get

$$\arg L(j\omega_{gc}) = \arg P_{nmp}(j\omega_{gc}) + \arg P_{mp}(j\omega_{gc}) + \arg C(j\omega_{gc}) \geq -\pi + \varphi_m, \tag{11.13}$$

where C is the controller transfer function. Let n_{gc} be the slope of the gain curve at the crossover frequency; since $|P_{nmp}(j\omega)| = 1$ it follows that

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

The slope n_{gc} is negative and larger than -2 if the system is stable. It follows from Bode's relations, equation (9.5), that

$$\arg P_{mp}(j\omega) + \arg C(j\omega) \approx n_{gc}\frac{\pi}{2}$$

Combining this with equation (11.13) gives the following inequality for the allowable phase lag

$$\varphi_\ell = -\arg P_{nmp}(j\omega_{gc}) \le \pi - \varphi_m + n_{gc}\frac{\pi}{2}. \tag{11.14}$$

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 $\varphi_\ell = \pi/2 = 1.57$ rad and a phase margin of 45° and $n_{gc} = -1$ gives and admissible phase lag $\varphi_\ell = \pi/4 = 0.78$ rad. 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 radians, or roughly 30° to 90°.

The crossover frequency inequality shows 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 $\Delta P(j\omega_{gc})$. As we shall see in the next chapter, the admissible phase lag is then reduced by $\arg \Delta P(j\omega_{gc})$.

A straightforward 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, shown in Figure 11.14, will immediately show the permissible gain crossover frequencies.

As an illustration we will give some analytical examples.

Example 11.5 (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

$$P_{nmp}(s) = \frac{z - s}{z + s}. \tag{11.15}$$

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

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

Since the phase of P_{nmp} decreases with frequency, the inequality (11.14) gives the following bound on the crossover frequency:

$$\frac{\omega_{gc}}{z} \le \tan\frac{\varphi_\ell}{2}. \tag{11.16}$$

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. ∇
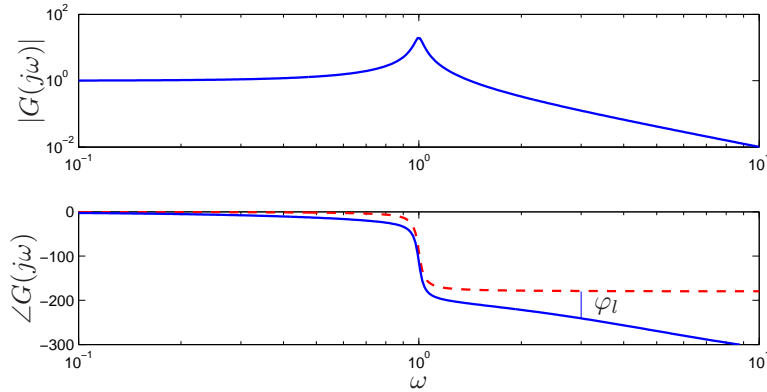
Figure 11.14: Bode plot of process transfer function (full lines) and corresponding minimum phase transfer function (dashed). The permissible gain crossover frequencies are those for which the difference in phase between the two curves satisfies equation (11.14).

Example 11.6 (Time delay). The transfer function of a time delay is

$$P(s) = e^{-sT}. \tag{11.17}$$

This is also the non-minimum phase part P_{nmp} and the corresponding phase lag is

$$\varphi_\ell = -\arg P_{nmp}(j\omega) = \omega T \qquad \Longrightarrow \qquad w_{gc} \le \frac{\varphi_l}{T}.$$

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 right half plane zero $z = 2/T$. A slow zero thus corresponds to a long time delay. ∇

Example 11.7 (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

$$P_{nmp}(s) = \frac{s + p}{s - p}, \tag{11.18}$$

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

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

Table 11.2: 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

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.   ▽

Example 11.8 (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

$$P_{nmp}(s) = \frac{(z-s)(s+p)}{(z+s)(s-p)}. \qquad (11.19)$$

The phase lag of this transfer function is

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

The mininum value of the right hand side is given by

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

which is achieved at $\omega = \sqrt{pz}$. The crossover frequency inequality (11.14) becomes

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

or

$$\frac{p}{z} \leq \tan\frac{\varphi_\ell}{4}.$$

The design choices $\varphi_m = \pi/4$ and $n_{gc} = -1/2$ gives $p < 0.17z$. Table 11.2 shows the admissible pole-zero ratios for different phase margins.   The

phase-margin that can be achieved for a given ratio p/z is

$$\varphi_m < \pi + n_{gc}\frac{\pi}{2} - 4\arctan\sqrt{\frac{p}{z}}.$$ (11.20)

A pair of poles and zeros in the right half plane thus 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. ∇

Avoiding Difficulties with RHP Poles and Zeros

As the examples above show, right half plane poles and zeros significantly limit the achievable performance of a system, hence one would like to avoid these whenever possible. The poles of a system depend on the intrinsic dynamics of the system and are given by 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 the 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 aircraft.

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.

11.5   Design Example

In this section we carry out a detailed design example that illustrates the main techniques in this chapter.

11.6   Further Reading

A more complete description of the material in this chapter is available in the text by Doyle, Frances and Tannenbaum [DFT92] (out of print, but available online).

11.7   Exercises

1. Regenerate the controller for the system in Example 11.4 and use the frequency responses for the Gang of Four to show that the performance specification is met.