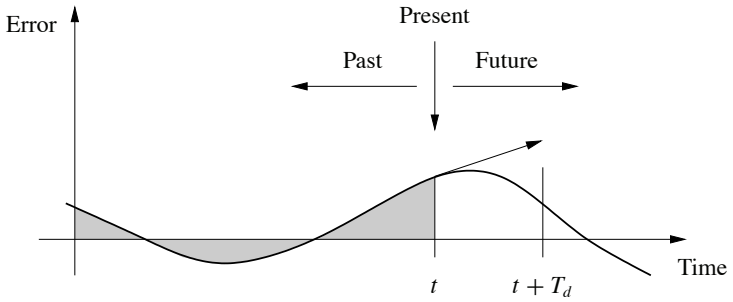


Figure 1.17: Action of a PID controller. At time t, the proportional term depends on the instantaneous value of the error. The integral portion of the feedback is based on the integral of the error up to time t (shaded portion). The derivative term provides an estimate of the growth or decay of the error over time by looking at the rate of change of the error. T_d represents the approximate amount of time in which the error is projected forward (see text).

1.6 Further Reading

The material in this section draws heavily from the report of the Panel on Future Directions on Control, Dynamics and Systems [155]. Several additional papers and reports have highlighted the successes of control [159] and new vistas in control [45, 130, 204]. The early development of control is described by Mayr [148] and in the books by Bennett [28, 29], which cover the period 1800–1955. A fascinating examination of some of the early history of control in the United States has been written by Mindell [152]. A popular book that describes many control concepts across a wide range of disciplines is *Out of Control* by Kelly [121]. There are many textbooks available that describe control systems in the context of specific disciplines. For engineers, the textbooks by Franklin, Powell and Emami-Naeini [79], Dorf and Bishop [61], Kuo and Golnaraghi [133] and Seborg, Edgar and Mellichamp [178] are widely used. More mathematically oriented treatments of control theory include Sontag [182] and Lewis [136]. The book by Hellerstein et al. [97] provides a description of the use of feedback control in computing systems. A number of books look at the role of dynamics and feedback in biological systems, including Milhorn [151] (now out of print), J. D. Murray [154] and Ellner and Guckenheimer [70]. The book by Fradkov [77] and the tutorial article by Bechhoefer [25] cover many specific topics of interest to the physics community.

Exercises

1.1 (Eye motion) Perform the following experiment and explain your results: Holding your head still, move one of your hands left and right in front of your face, following it with your eyes. Record how quickly you can move your hand before you begin to lose track of it. Now hold your hand still and shake your head left to right, once again recording how quickly you can move before losing track of your hand.

a system in observable canonical form, which is given by

$$W_o = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -a_1 & 1 & 0 & \dots & 0 \\ -a_1^2 - a_1 a_2 & -a_1 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ * & * & & \dots & 1 \end{bmatrix},$$

where * represents an entry whose exact value is not important. The rows of this matrix are linearly independent (since it is lower triangular), and hence W_o is full rank. A straightforward but tedious calculation shows that the inverse of the observability matrix has a simple form given by

$$W_o^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1 & 1 & 0 & \cdots & 0 \\ a_2 & a_1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ a_{n-1} & a_{n-2} & a_{n-3} & \cdots & 1 \end{bmatrix}.$$

As in the case of reachability, it turns out that if a system is observable then there always exists a transformation T that converts the system into observable canonical form. This is useful for proofs since it lets us assume that a system is in observable canonical form without any loss of generality. The observable canonical form may be poorly conditioned numerically.

7.2 State Estimation

Having defined the concept of observability, we now return to the question of how to construct an observer for a system. We will look for observers that can be represented as a linear dynamical system that takes the inputs and outputs of the system we are observing and produces an estimate of the system's state. That is, we wish to construct a dynamical system of the form

$$\frac{d\hat{x}}{dt} = F\hat{x} + Gu + Hy,$$

where u and y are the input and output of the original system and $\hat{x} \in \mathbb{R}^n$ is an estimate of the state with the property that $\hat{x}(t) \to x(t)$ as $t \to \infty$.

The Observer

We consider the system in equation (7.1) with D set to zero to simplify the exposition:

$$\frac{dx}{dt} = Ax + Bu, \qquad y = Cx. \tag{7.6}$$

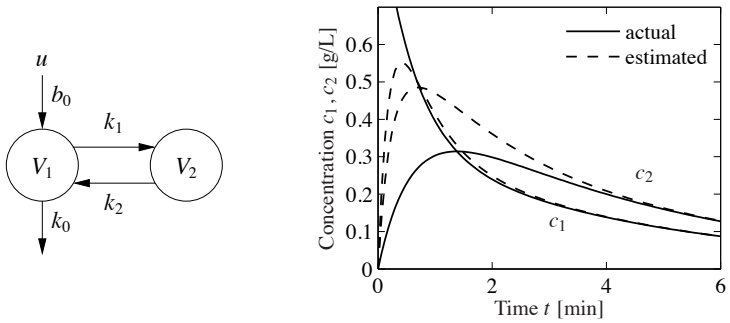


Figure 7.4: Observer for a two compartment system. A two compartment model is shown on the left. The observer measures the input concentration u and output concentration $y = c_1$ to determine the compartment concentrations, shown on the right. The true concentrations are shown by solid lines and the estimates generated by the observer by dashed lines.

Let the desired characteristic polynomial of the observer be $s^2 + p_1 s + p_2$, and equation (7.11) gives the observer gain

$$L = \begin{bmatrix} 1 & 0 \\ -k_0 - k_1 & k_1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ k_0 + k_1 + k_2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} p_1 - k_0 - k_1 - k_2 \\ p_2 - k_0 k_2 \end{bmatrix}$$
$$= \begin{bmatrix} p_1 - k_0 - k_1 - k_2 \\ (p_2 - p_1 k_2 + k_1 k_2 + k_2^2)/k_1 \end{bmatrix}.$$

Notice that the observability condition $k_1 \neq 0$ is essential. The behavior of the observer is illustrated by the simulation in Figure 7.4b. Notice how the observed concentrations approach the true concentrations. ∇

The observer is a dynamical system whose inputs are the process input u and the process output y. The rate of change of the estimate is composed of two terms. One term, $A\hat{x} + Bu$, is the rate of change computed from the model with $\hat{x}$ substituted for x. The other term, $L(y - \hat{y})$, is proportional to the difference $e = y - \hat{y}$ between measured output y and its estimate $\hat{y} = C\hat{x}$. The observer gain L is a matrix that tells how the error e is weighted and distributed among the states. The observer thus combines measurements with a dynamical model of the system. A block diagram of the observer is shown in Figure 7.5.

Computing the Observer Gain

For simple low-order problems it is convenient to introduce the elements of the observer gain L as unknown parameters and solve for the values required to give the desired characteristic polynomial, as illustrated in the following example.

Example 7.3 Vehicle steering
The normalized linear model for vehicle steering derived in Examples 5.12 and 6.4 gives the following state space model dynamics relating lateral path deviation y to

to command signals and disturbances are decoupled. Disturbance responses are governed by the observer and the state feedback, while the response to command signals is governed by the trajectory generator (feedforward).

For an analytic description we start with the full nonlinear dynamics of the process

$$\frac{dx}{dt} = f(x, u), \qquad y = h(x, u). \tag{7.23}$$

Assume that the trajectory generator is able to compute a desired trajectory (x_d, u_{ff}) that satisfies the dynamics (7.23) and satisfies $r = h(x_d, u_{\text{ff}})$. To design the controller, we construct the error system. Let $z = x - x_d$ and $v = u - u_{\text{ff}}$ and compute the dynamics for the error:

$$\begin{aligned}\dot{z} &= \dot{x} - \dot{x}_d = f(x, u) - f(x_d, u_{\text{ff}}) \\ &= f(z + x_d, v + u_{\text{ff}}) - f(x_d, u_{\text{ff}}) =: F(z, v, x_d(t), u_{\text{ff}}(t)).\end{aligned}$$

In general, this system is time-varying. Note that $z = -e$ in Figure 7.10 due to the convention of using negative feedback in the block diagram.

For trajectory tracking, we can assume that e is small (if our controller is doing a good job), and so we can linearize around $z = 0$:

$$\frac{dz}{dt} \approx A(t)z + B(t)v, \quad A(t) = \left.\frac{\partial F}{\partial z}\right|_{(x_d(t), u_{\text{ff}}(t))}, \quad B(t) = \left.\frac{\partial F}{\partial v}\right|_{(x_d(t), u_{\text{ff}}(t)}.$$

It is often the case that $A(t)$ and $B(t)$ depend only on x_d, in which case it is convenient to write $A(t) = A(x_d)$ and $B(t) = B(x_d)$.

Assume now that x_d and u_{ff} are either constant or slowly varying (with respect to the performance criterion). This allows us to consider just the (constant) linear system given by $(A(x_d), B(x_d))$. If we design a state feedback controller $K(x_d)$ for each x_d, then we can regulate the system using the feedback

$$v = -K(x_d)z.$$

Substituting back the definitions of e and v, our controller becomes

$$u = -K(x_d)(x - x_d) + u_{\text{ff}}.$$

This form of controller is called a *gain scheduled* linear controller with *feedforward* u_{ff}.

Finally, we consider the observer. The full nonlinear dynamics can be used for the prediction portion of the observer and the linearized system for the correction term:

$$\frac{d\hat{x}}{dt} = f(\hat{x}, u) + L(\hat{x})(y - h(\hat{x}, u)),$$

where $L(\hat{x})$ is the observer gain obtained by linearizing the system around the currently estimated state. This form of the observer is known as an *extended Kalman filter* and has proved to be a very effective means of estimating the state of a nonlinear system.

and

$$G_{ur}(s) = \frac{k_r}{1 + KG_{\hat{x}u}(s)} = \frac{k_1(s^2 + l_1 s + l_2)}{s^2 + s(\gamma k_1 + k_2 + l_1) + k_1 + l_2 + k_2 l_1 - \gamma k_2 l_2},$$

where k_1 and k_2 are the controller gains.

Finally, we compute the full closed loop dynamics. We begin by deriving the transfer function for the process $P(s)$. We can compute this directly from the state space description of the dynamics, which was given in Example 5.12. Using that description, we have

$$P(s) = G_{yu}(s) = C(sI - A)^{-1}B + D = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} s & -1 \\ 0 & s \end{bmatrix}^{-1} \begin{bmatrix} \gamma \\ 1 \end{bmatrix} = \frac{\gamma s + 1}{s^2}.$$

The transfer function for the full closed loop system between the input r and the output y is then given by

$$G_{yr} = \frac{k_r P(s)}{1 + P(s)G_{uy}(s)} = \frac{k_1(\gamma s + 1)}{s^2 + (k_1\gamma + k_2)s + k_1}.$$

Note that the observer gains l_1 and l_2 do not appear in this equation. This is because we are considering steady-state analysis and, in steady state, the estimated state exactly tracks the state of the system assuming perfect models. We will return to this example in Chapter 12 to study the robustness of this particular approach. ∇

Pole/Zero Cancellations

Because transfer functions are often polynomials in s, it can sometimes happen that the numerator and denominator have a common factor, which can be canceled. Sometimes these cancellations are simply algebraic simplifications, but in other situations they can mask potential fragilities in the model. In particular, if a pole/zero cancellation occurs because terms in separate blocks that just happen to coincide, the cancellation may not occur if one of the systems is slightly perturbed. In some situations this can result in severe differences between the expected behavior and the actual behavior.

To illustrate when we can have pole/zero cancellations, consider the block diagram in Figure 8.7 with $F = 1$ (no feedforward compensation) and C and P given by

$$C(s) = \frac{n_c(s)}{d_c(s)}, \qquad P(s) = \frac{n_p(s)}{d_p(s)}.$$

The transfer function from r to e is then given by

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

If there are common factors in the numerator and denominator polynomials, then these terms can be factored out and eliminated from both the numerator and denominator. For example, if the controller has a zero at $s = -a$ and the process has

a pole at $s = -a$, then we will have

$$G_{er}(s) = \frac{(s+a)d_c'(s)d_p(s)}{(s+a)d_c(s)d_p'(s) + (s+a)n_c'(s)n_p(s)} = \frac{d_c'(s)d_p(s)}{d_c(s)d_p'(s) + n_c'(s)n_p(s)},$$

where $n_c'(s)$ and $d_p'(s)$ represent the relevant polynomials with the term $s + a$ factored out. In the case when $a < 0$ (so that the zero or pole is in the right half-plane), we see that there is no impact on the transfer function G_{er}.

Suppose instead that we compute the transfer function from d to e, which represents the effect of a disturbance on the error between the reference and the output. This transfer function is given by

$$G_{ed}(s) = \frac{d_c'(s)n_p(s)}{(s+a)d_c(s)d_p'(s) + (s+a)n_c'(s)n_p(s)}.$$

Notice that if $a < 0$, then the pole is in the right half-plane and the transfer function G_{ed} is *unstable*. Hence, even though the transfer function from r to e appears to be okay (assuming a perfect pole/zero cancellation), the transfer function from d to e can exhibit unbounded behavior. This unwanted behavior is typical of an *unstable pole/zero cancellation*.

It turns out that the cancellation of a pole with a zero can also be understood in terms of the state space representation of the systems. Reachability or observability is lost when there are cancellations of poles and zeros (Exercise 8.11). A consequence is that the transfer function represents the dynamics only in the reachable and observable subspace of a system (see Section 7.5).

Example 8.7 Cruise control

The input/output response from throttle to velocity for the linearized model for a car has the transfer function $G(s) = b/(s-a), a < 0$. A simple (but not necessarily good) way to design a PI controller is to choose the parameters of the PI controller so that the controller zero at $s = -k_i/k_p$ cancels the process pole at $s = a$. The transfer function from reference to velocity is $G_{vr}(s) = bk_p/(s+bk_p)$, and control design is simply a matter of choosing the gain k_p. The closed loop system dynamics are of first order with the time constant $1/bk_p$.

Figure 8.10 shows the velocity error when the car encounters an increase in the road slope. A comparison with the controller used in Figure 3.3b (reproduced in dashed curves) shows that the controller based on pole/zero cancellation has very poor performance. The velocity error is larger, and it takes a long time to settle.

Notice that the control signal remains practically constant after $t = 15$ even if the error is large after that time. To understand what happens we will analyze the system. The parameters of the system are $a = -0.0101$ and $b = 1.32$, and the controller parameters are $k_p = 0.5$ and $k_i = 0.0051$. The closed loop time constant is $1/(bk_p) = 2.5$ s, and we would expect that the error would settle in about 10 s (4 time constants). The transfer functions from road slope to velocity and control

The situation is more complicated if there is a direct term. If $y = h(x, u)$, then replacing u by $-ky$ gives

$$\frac{dx}{dt} = f(x, -ky), \qquad y = h(x, -ky).$$

To obtain a differential equation for x, the algebraic equation $y = h(x, -ky)$ must be solved to give $y = \alpha(x)$, which in general is a complicated task.

When algebraic loops are present, it is necessary to solve algebraic equations to obtain the differential equations for the complete system. Resolving algebraic loops is a nontrivial problem because it requires the symbolic solution of algebraic equations. Most block diagram-oriented modeling languages cannot handle algebraic loops, and they simply give a diagnosis that such loops are present. In the era of analog computing, algebraic loops were eliminated by introducing fast dynamics between the loops. This created differential equations with fast and slow modes that are difficult to solve numerically. Advanced modeling languages like Modelica use several sophisticated methods to resolve algebraic loops.

8.4 The Bode Plot

The frequency response of a linear system can be computed from its transfer function by setting $s = i\omega$, corresponding to a complex exponential

$$u(t) = e^{i\omega t} = \cos(\omega t) + i \sin(\omega t).$$

The resulting output has the form

$$y(t) = G(i\omega)e^{i\omega t} = Me^{i(\omega t+\varphi)} = M\cos(\omega t + \varphi) + iM\sin(\omega t + \varphi),$$

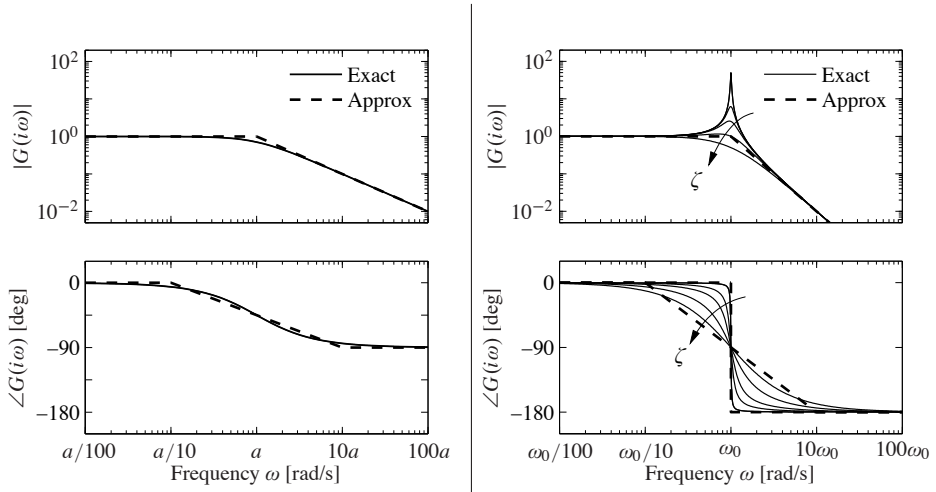
where M and φ are the gain and phase of G:

$$M = |G(i\omega)|, \qquad \varphi = \arctan \frac{\operatorname{Im} G(i\omega)}{\operatorname{Re} G(i\omega)}.$$

The phase of G is also called the *argument* of G, a term that comes from the theory of complex variables.

It follows from linearity that the response to a single sinusoid (sin or cos) is amplified by M and phase-shifted by φ. Note that $-\pi < \varphi \leq \pi$, so the arctangent must be taken respecting the signs of the numerator and denominator. It will often be convenient to represent the phase in degrees rather than radians. We will use the notation $\angle G(i\omega)$ for the phase in degrees and $\arg G(i\omega)$ for the phase in radians. In addition, while we always take $\arg G(i\omega)$ to be in the range $(-\pi, \pi]$, we will take $\angle G(i\omega)$ to be continuous, so that it can take on values outside the range of -180° to 180°.

The frequency response $G(i\omega)$ can thus be represented by two curves: the gain curve and the phase curve. The *gain curve* gives $|G(i\omega)|$ as a function of frequency ω, and the *phase curve* gives $\angle G(i\omega)$. One particularly useful way of drawing these



(a) First-order system (b) Second-order system

Figure 8.13: Bode plots for first- and second-order systems. (a) The first-order system $G(s) = a/(s+a)$ can be approximated by asymptotic curves (dashed) in both the gain and the frequency, with the breakpoint in the gain curve at $\omega = a$ and the phase decreasing by 90° over a factor of 100 in frequency. (b) The second-order system $G(s) = \omega_0^2/(s^2 + 2\zeta\omega_0 s + \omega_0^2)$ has a peak at frequency a and then a slope of -2 beyond the peak; the phase decreases from 0° to −180°. The height of the peak and the rate of change of phase depending on the damping ratio ζ ($\zeta = 0.02, 0.1, 0.2, 0.5$ and 1.0 shown).

the following straight lines

$$\log|G(i\omega)| \approx \begin{cases} 0 & \text{if } \omega < a \\ \log a - \log\omega & \text{if } \omega > a, \end{cases}$$

$$\angle G(i\omega) \approx \begin{cases} 0 & \text{if } \omega < a/10 \\ -45 - 45(\log\omega - \log a) & a/10 < \omega < 10a \\ -90 & \text{if } \omega > 10a. \end{cases}$$

The approximate gain curve consists of a horizontal line up to frequency $\omega = a$, called the *breakpoint* or *corner frequency*, after which the curve is a line of slope -1 (on a log-log scale). The phase curve is zero up to frequency $a/10$ and then decreases linearly by 45°/decade up to frequency $10a$, at which point it remains constant at 90°. Notice that a first-order system behaves like a constant for low frequencies and like an integrator for high frequencies; compare with the Bode plot in Figure 8.12.

Finally, consider the transfer function for a second-order system,

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

for which we have

$$\log|G(i\omega)| = 2\log\omega_0 - \frac{1}{2}\log\left(\omega^4 + 2\omega_0^2\omega^2(2\zeta^2 - 1) + \omega_0^4\right),$$
$$\angle G(i\omega) = -\frac{180}{\pi}\arctan\frac{2\zeta\omega_0\omega}{\omega_0^2 - \omega^2}.$$

The gain curve has an asymptote with zero slope for $\omega \ll \omega_0$. For large values of ω the gain curve has an asymptote with slope -2. The largest gain $Q = \max_\omega |G(i\omega)| \approx 1/(2\zeta)$, called the *Q-value*, is obtained for $\omega \approx \omega_0$. The phase is zero for low frequencies and approaches 180° for large frequencies. The curves can be approximated with the following piecewise linear expressions

$$\log|G(i\omega)| \approx \begin{cases} 0 & \text{if } \omega \ll \omega_0 \\ 2\log\omega_0 - 2\log\omega & \text{if } \omega \gg \omega_0, \end{cases}$$
$$\angle G(i\omega) \approx \begin{cases} 0 & \text{if } \omega \ll \omega_0 \\ -180 & \text{if } \omega \gg \omega_0. \end{cases}$$

The Bode plot is shown in Figure 8.13b. Note that the asymptotic approximation is poor near $\omega = \omega_0$ and that the Bode plot depends strongly on ζ near this frequency.

Given the Bode plots of the basic functions, we can now sketch the frequency response for a more general system. The following example illustrates the basic idea.

Example 8.8 Asymptotic approximation for a transfer function
Consider the transfer function given by

$$G(s) = \frac{k(s+b)}{(s+a)(s^2 + 2\zeta\omega_0 s + \omega_0^2)}, \qquad a \ll b \ll \omega_0.$$

The Bode plot for this transfer function appears in Figure 8.14, with the complete transfer function shown as a solid line and the asymptotic approximation shown as a dashed line.

We begin with the gain curve. At low frequency, the magnitude is given by

$$G(0) = \frac{kb}{a\omega_0^2}.$$

When we reach $\omega = a$, the effect of the pole begins and the gain decreases with slope -1. At $\omega = b$, the zero comes into play and we increase the slope by 1, leaving the asymptote with net slope 0. This slope is used until the effect of the second-order pole is seen at $\omega = \omega_c$, at which point the asymptote changes to slope -2. We see that the gain curve is fairly accurate except in the region of the peak due to the second-order pole (since for this case ζ is reasonably small).

The phase curve is more complicated since the effect of the phase stretches out much further. The effect of the pole begins at $\omega = a/10$, at which point we change from phase 0 to a slope of -45°/decade. The zero begins to affect the phase at $\omega = b/10$, producing a flat section in the phase. At $\omega = 10a$ the phase

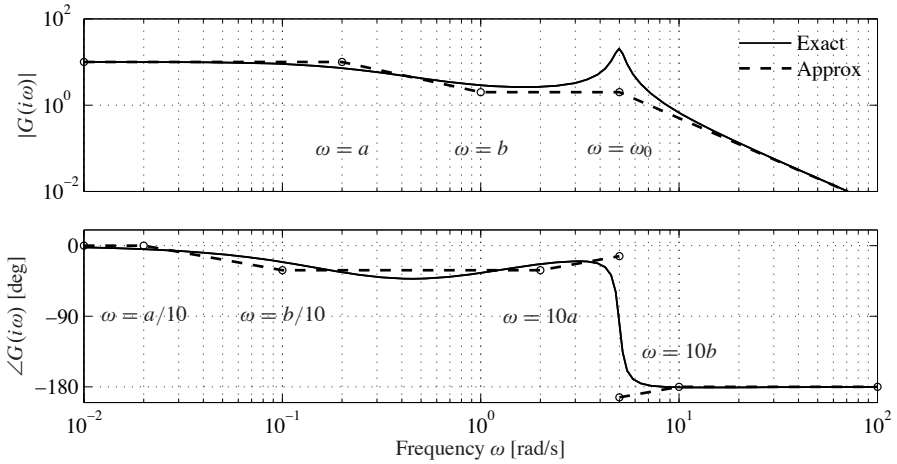


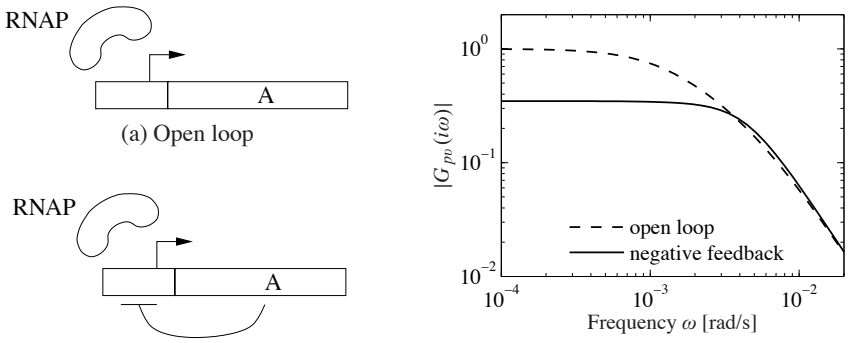
Figure 8.14: Asymptotic approximation to a Bode plot. The thin line is the Bode plot for the transfer function $G(s) = k(s+b)/(s+a)(s^2 + 2\zeta\omega_0 s + \omega_0^2)$, where $a \ll b \ll \omega_0$. Each segment in the gain and phase curves represents a separate portion of the approximation, where either a pole or a zero begins to have effect. Each segment of the approximation is a straight line between these points at a slope given by the rules for computing the effects of poles and zeros.

contributions from the pole end, and we are left with a slope of +45°/decade (from the zero). At the location of the second-order pole, $s \approx i\omega_c$, we get a jump in phase of $-180°$. Finally, at $\omega = 10b$ the phase contributions of the zero end, and we are left with a phase of -180 degrees. We see that the straight-line approximation for the phase is not as accurate as it was for the gain curve, but it does capture the basic features of the phase changes as a function of frequency. ∇

The Bode plot gives a quick overview of a system. Since any signal can be decomposed into a sum of sinusoids, it is possible to visualize the behavior of a system for different frequency ranges. The system can be viewed as a filter that can change the amplitude (and phase) of the input signals according to the frequency response. For example, if there are frequency ranges where the gain curve has constant slope and the phase is close to zero, the action of the system for signals with these frequencies can be interpreted as a pure gain. Similarly, for frequencies where the slope is +1 and the phase close to 90°, the action of the system can be interpreted as a differentiator, as shown in Figure 8.12.

Three common types of frequency responses are shown in Figure 8.15. The system in Figure 8.15a is called a *low-pass filter* because the gain is constant for low frequencies and drops for high frequencies. Notice that the phase is zero for low frequencies and $-180°$ for high frequencies. The systems in Figure 8.15b and c are called a *band-pass filter* and *high-pass filter* for similar reasons.

To illustrate how different system behaviors can be read from the Bode plots we consider the band-pass filter in Figure 8.15b. For frequencies around $\omega = \omega_0$, the signal is passed through with no change in gain. However, for frequencies well



(a) Open loop

(b) Negative feedback

(c) Frequency response

Figure 8.16: Noise attenuation in a genetic circuit. The open loop system (a) consists of a constitutive promoter, while the closed loop circuit (b) is self-regulated with negative feedback (repressor). The frequency response for each circuit is shown in (c).

The resulting transfer function is given by

$$G_{pv}^{\mathrm{cl}}(s) = \frac{\beta}{(s+\gamma)(s+\delta)+\beta\sigma}, \qquad \sigma = \frac{n\alpha_1 k p_e^{n-1}}{(1+kp_e^n)^2}.$$

Figure 8.16c shows the frequency response for the two circuits. We see that the feedback circuit attenuates the response of the system to disturbances with low-frequency content but slightly amplifies disturbances at high frequency (compared to the open loop system). Notice that these curves are very similar to the frequency response curves for the op amp shown in Figure 8.3b. ∇

Transfer Functions from Experiments

The transfer function of a system provides a summary of the input/output response and is very useful for analysis and design. However, modeling from first principles can be difficult and time-consuming. Fortunately, we can often build an input/output model for a given application by directly measuring the frequency response and fitting a transfer function to it. To do so, we perturb the input to the system using a sinusoidal signal at a fixed frequency. When steady state is reached, the amplitude ratio and the phase lag give the frequency response for the excitation frequency. The complete frequency response is obtained by sweeping over a range of frequencies.

By using correlation techniques it is possible to determine the frequency response very accurately, and an analytic transfer function can be obtained from the frequency response by curve fitting. The success of this approach has led to instruments and software that automate this process, called *spectrum analyzers*. We illustrate the basic concept through two examples.

Example 8.10 Atomic force microscope

To illustrate the utility of spectrum analysis, we consider the dynamics of the atomic force microscope, introduced in Section 3.5. Experimental determination of the

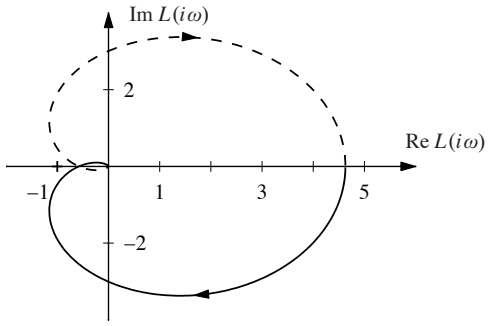


Figure 9.4: Nyquist plot for a third-order transfer function. The Nyquist plot consists of a trace of the loop transfer function $L(s) = 1/(s+a)^3$. The solid line represents the portion of the transfer function along the positive imaginary axis, and the dashed line the negative imaginary axis. The outer arc of the D contour maps to the origin.

Nyquist D contour. This arc has the form $s = Re^{i\theta}$ for $R \to \infty$. This gives

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

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

An alternative to computing the Nyquist plot explicitly is to determine the plot from the frequency response (Bode plot), which gives the Nyquist curve for $s = i\omega$, $\omega > 0$. We start by plotting $G(i\omega)$ from $\omega = 0$ to $\omega = \infty$, which can be read off from the magnitude and phase of the transfer function. We then plot $G(Re^{i\theta})$ with $\theta \in [-\pi/2, \pi/2]$ and $R \to \infty$, which almost always maps to zero. The remaining parts of the plot can be determined by taking the mirror image of the curve thus far (normally plotted using a dashed line). The plot can then be labeled with arrows corresponding to a clockwise traversal around the D contour (the same direction in which the first portion of the curve was plotted).

Example 9.3 Third-order system with a pole at the origin
Consider the transfer function

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

where the gain has the nominal value $k = 1$. The Bode plot is shown in Figure 9.5a. The system has a single pole at $s = 0$ and a double pole at $s = -1$. The gain curve of the Bode plot thus has the slope -1 for low frequencies, and at the double pole $s = 1$ the slope changes to -3. For small s we have $L \approx k/s$, which means that the low-frequency asymptote intersects the unit gain line at $\omega = k$. The phase curve starts at -90° for low frequencies, it is -180° at the breakpoint $\omega = 1$ and it is -270° at high frequencies.

Having obtained the Bode plot, we can now sketch the Nyquist plot, shown in Figure 9.5b. It starts with a phase of -90° for low frequencies, intersects the negative real axis at the breakpoint $\omega = 1$ where $L(i) = 0.5$ and goes to zero along

at $z = a$ with the residue m. The sum of the residues at the zeros of the function is Z. Similarly, we find that the sum of the residues of the poles of is $-P$, and hence

$$Z - P = \frac{1}{2\pi i}\int_\Gamma \frac{f'(z)}{f(z)}\,dz = \frac{1}{2\pi i}\int_\Gamma \frac{d}{dz}\log f(z)\,dz = \frac{1}{2\pi i}\Delta_\Gamma \log f(z),$$

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

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

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

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

and the theorem is proved. □

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

Theorem 9.3 can be used to prove Nyquist's stability theorem by choosing Γ as the Nyquist contour shown in Figure 9.3a, which encloses the right half-plane. To construct the contour, we start with part of the imaginary axis $-jR \le s \le jR$ 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 $R \to \infty$ and $r \to 0$. Note that Γ has orientation *opposite* that shown in Figure 9.3a. (The convention in engineering is to traverse the Nyquist contour in the clockwise direction since this corresponds to moving upwards along the imaginary axis, which makes it easy to sketch the Nyquist contour from a Bode plot.)

To see how we use the principle of variation of the argument to compute stability, consider a closed loop system with the loop transfer function $L(s)$. The closed loop poles of the system 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 *counterclockwise* direction. The winding number can conveniently be determined from the Nyquist plot. A direct application of Theorem 9.3 gives the Nyquist criterion, taking care to flip the orientation. Since the image of $1 + L(s)$ is a shifted version of $L(s)$, we usually state the Nyquist criterion as net encirclements of the -1 point by the image of $L(s)$.

9.3 Stability Margins

In practice it is not enough that a system is stable. There must also be some margins of stability that describe how stable the system is and its robustness to perturbations. There are many ways to express this, but one of the most common is the use of gain and phase margins, inspired by Nyquist's stability criterion. The key idea is that it

[21] M. Atkinson, M. Savageau, J. Myers, and A. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell*, 113(5):597–607, 2003.

[22] M. B. Barron and W. F. Powers. The role of electronic controls for future automotive mechatronic systems. *IEEE Transactions on Mechatronics*, 1(1):80–89, 1996.

[23] T. Basar (editor). *Control Theory: Twenty-five Seminal Papers (1932–1981)*. IEEE Press, New York, 2001.

[24] T. Basar and P. Bernhard. *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Birkhauser, Boston, 1991.

[25] J. Bechhoefer. Feedback for physicists: A tutorial essay on control. *Reviews of Modern Physics*, 77:783–836, 2005.

[26] R. Bellman and K. J. Åström. On structural identifiability. *Mathematical Biosciences*, 7:329–339, 1970.

[27] R. E. Bellman and R. Kalaba. *Selected Papers on Mathematical Trends in Control Theory*. Dover, New York, 1964.

[28] S. Bennett. *A History of Control Engineering: 1800–1930*. Peter Peregrinus, Stevenage, 1979.

[29] S. Bennett. *A History of Control Engineering: 1930–1955*. Peter Peregrinus, Stevenage, 1993.

[30] L. L. Beranek. *Acoustics*. McGraw-Hill, New York, 1954.

[31] R. N. Bergman. Toward physiological understanding of glucose tolerance: Minimal model approach. *Diabetes*, 38:1512–1527, 1989.

[32] D. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall, Englewood Cliffs, 1987.

[33] B. Bialkowski. Process control sample problems. In N. J. Sell (editor), *Process Control Fundamentals for the Pulp & Paper Industry*. Tappi Press, Norcross, GA, 1995.

[34] G. Binnig and H. Rohrer. Scanning tunneling microscopy. *IBM Journal of Research and Development*, 30(4):355–369, 1986.

[35] H. S. Black. Stabilized feedback amplifiers. *Bell System Technical Journal*, 13:1–2, 1934.

[36] H. S. Black. Inventing the negative feedback amplifier. *IEEE Spectrum*, pp. 55–60, 1977.

[37] J. F. Blackburn, G. Reethof, and J. L. Shearer. *Fluid Power Control*. MIT Press, Cambridge, MA, 1960.

[38] J. H. Blakelock. *Automatic Control of Aircraft and Missiles*, 2nd ed. Addison-Wesley, Cambridge, MA, 1991.

[39] G. Blickley. Modern control started with Ziegler-Nichols tuning. *Control Engineering*, 37:72–75, 1990.

[40] H. W. Bode. *Network Analaysis and Feedback Amplifier Design*. Van Nostrand, New York, 1945.

[41] H. W. Bode. Feedback—The history of an idea. *Symposium on Active Networks and Feedback Systems*. Polytechnic Institute of Brooklyn, New York, 1960. Reprinted in [27].

[42] W. E. Boyce and R. C. DiPrima. *Elementary Differential Equations*. Wiley, New York, 2004.

[43] B. Brawn and F. Gustavson. Program behavior in a paging environment. *Proceedings of the AFIPS Fall Joint Computer Conference*, pp. 1019–1032, 1968.

[44] R. W. Brockett. *Finite Dimensional Linear Systems*. Wiley, New York, 1970.

[45] R. W. Brockett. New issues in the mathematics of control. In B. Engquist and W. Schmid (editors), *Mathematics Unlimited—2001 and Beyond*, pp. 189–220. Springer-Verlag, Berlin, 2000.

[46] G. S. Brown and D. P. Campbell. *Principles of Servomechanims*. Wiley, New York, 1948.