EEET 3046 Control Systems / EEET 5148 Control Systems M (2019)
Lecture 9: Frequency response

Dr. Yee Wei Law ⟨yeewei.law@unisa.edu.au⟩

Contents

Figure 1: National Instruments' PXI, Compact RIO, Compact DAQ, and USB hardware platforms (shown from left to right) can be used with NI LabVIEW to provide stimulus and acquire signals from physical systems. NI LabVIEW can then be used to analyze data, determine the mathematical model, and prototype and deploy a controller for the physical system. More information at `http://www.ni.com/en-au/shop/select/spectrum-and-signal-analyzers-category`.

1 Introduction

Similar to the root locus approach, the frequency response approach is a graphical one, but which uses *Bode plots*[1] rather than root locus plots. The frequency response approach complements the root locus approach, but compared to the root locus approach, it has the following advantages:

- Frequency response can be experimentally obtained using a signal generator for the input, varying the input frequency, and recording the output — this can be achieved using an oscilloscope capable of frequency response analysis, such as the Keysight InfiniiVision MSOX3104T[2]; or a *vector network analyzer*, such as Omicron Lab's Bode 100[3] and Tektronix's TTR500 series[4]; or a *spectrum analyzer*, such as National Instruments' PXI Vector Signal Analyzer (see Figure 1). This allows modeling of transfer functions directly from physical data [VV07].
- It enables stability analysis of systems with *nonrational* transfer functions.

Detail: Why approximating nonrational transfer functions is not enough

Consider the rational transfer function $G(s) = N(s)/D(s)$, where $N(s)$ is the numerator and $D(s)$ is the denominator. $D(s)$ is called the *characteristic polynomial* of $G(s)$, and the roots of $D(s)$ are called the *poles* of $G(s)$. Up to this point, we can tell whether $G(s)$ is stable by either of the following methods:

- If none of the poles of $G(s)$ have a positive real part, then its natural response is stable.
- If the characteristic polynomial of $G(s)$ satisfies the Routh-Hurwitz stability criterion, then its natural response is stable. As shown in Lecture 4, the Routh-Hurwitz criterion is useful for determining the constraint on a system parameter for which the system is stable.

However, the methods above break down if $G(s)$ is nonrational, such as when it has an exponential term for representing time delay. As discussed in the previous lecture, plants with time delay are common in the process control industry. Techniques like Padé approximation[a] can be applied to approximate the exponential as a rational function, but the approximation is only valid at low frequencies, and generally not very good. We need a better solution.

The starting point of Padé approximation is the Taylor series expansion of the exponential:

$$e^{-Ls} = 1 - Ls + \frac{L^2 s^2}{2!} - \frac{L^3 s^3}{3!} + \cdots,$$

but the infinite series is not a polynomial fraction. To approximate the infinite series with a fraction, the 1/1 Padé approximant [SMEDI10, Sect. 6.2.1] expresses the exponential as

$$e^{-Ls} \approx \frac{1 - \frac{Ls}{2}}{1 + \frac{Ls}{2}} = 1 - Ls + \frac{L^2 s^2}{2} - \frac{L^3 s^3}{4} + \cdots, \tag{1}$$

[1] MATLAB engineer answers "Why use Bode plots?" on YouTube: `https://youtu.be/F6-EaZobHNk`

[2] `https://youtu.be/ADmVAs6vqLU`

[3] `https://www.omicron-lab.com/products/vector-network-analysis/bode-100/`

[4] `https://www.tek.com/vna/ttr500`

which is consistent with the Taylor series expansion up to the second-order term. The 2/2 Padé approximant expresses the exponential as

$$e^{-Ls} \approx \frac{1 - \frac{Ls}{2} + \frac{L^2 s^2}{12}}{1 + \frac{Ls}{2} + \frac{L^2 s^2}{12}} = 1 - Ls + \frac{L^2 s^2}{2} - \frac{L^3 s^3}{6} + \frac{L^4 s^4}{24} - \frac{L^5 s^5}{144} + \cdots, \tag{2}$$

which is consistent with the Taylor series expansion up to the fourth-order term. Figure 2 compares the step responses of e^{-Ls} and the 2/2 Padé approximant in Eq. (2). Observe the approximation is only valid at low frequencies, and generally unimpressive.

[a] `http://mathworld.wolfram.com/PadeApproximant.html`

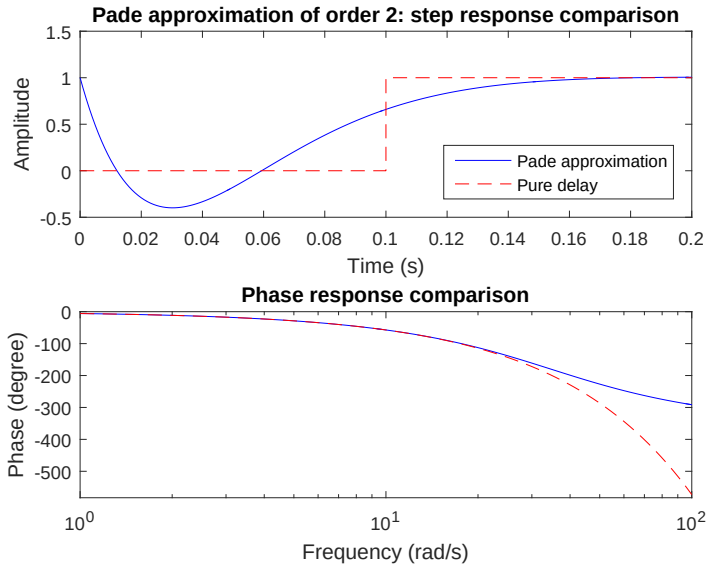


Figure 2: Plots indicating the level of approximation provided by the 2/2 Padé approximant as defined in Eq. (2). These plots are generated with the MATLAB command `pade(0.1, 2)`, where the first argument is the time delay, and the second argument is the order of the Padé approximant. In the plot, the response labeled "Pure delay" is $u(t - 0.1)$, whereas the response labeled "Pade approximation" is the inverse Laplace transform of the 2/2 Padé approximant. Phase response will be discussed later.

- It is more efficient than the root locus approach for designing lead compensators by taking into account transient response and steady-state error simultaneously.

Attention: Frequency response of nonlinear systems

The textbook [Nis15, p. 526] claims frequency response methods have "distinct advantage" when used for "finding the stability of nonlinear systems." This is not true, because the frequency response methods discussed here and in the textbook are only applicable to linear systems. For nonlinear systems, so-called *generalized frequency response functions* [BJ90] can be used, but these are advanced mathematical techniques that share very little in common with the frequency response methods.

In this lecture, we will learn

- what the frequency response of a SISO LTI system is;
- how to represent a frequency response using a Bode plot — the most important tool for frequency response-based controller design;
- how to interpret a Bode plot in terms of the *stability*, *steady-state error* and *transient response* of

the closed-loop system associated with the Bode plot.

It is important to realize from the outset that like the root locus plot, the Bode plot is a graphical tool for *presenting closed-loop information using open-loop information*, as a way of simplifying controller design. Unlike the root locus plot however, reading (closed-loop) stability and transient response information from the Bode plot is not as straightforward. This lecture paves way for the next lecture, where we will learn how to use Bode plots to design PID controllers and lag-lead compensators.

Attention: Complex numbers

Knowledge of complex numbers is essential for this lecture. A complex number $a + jb$ has the magnitude

$$|a + jb| = \sqrt{a^2 + b^2},$$

and phase

$$\angle(a + jb) = \text{atan2}(b, a).$$

A ratio of complex numbers c_1/c_2 has the magnitude $|c_1|/|c_2|$ and phase $\angle c_1 - \angle c_2$. According to Euler's formula,

$$a + jb = re^{j\theta} = r\cos\theta + jr\sin\theta,$$

where $r = |a + jb|$, and $\theta = \angle(a + jb)$.

2 Frequency response

Since a transfer function is just a complex-valued function of the Laplace variable s, we can represent a SISO LTI system with its magnitude $M(s)$ and phase $\phi(s)$. In frequency-domain analysis, we are concerned with the study of sinusoidal signals — and only these, because Fourier theory says any signal can be represented by sinusoids. When considering only sinusoidal signals, we can write M and ϕ as just functions of the frequency ω (as we do with Fourier transform). When the system, denoted $M(\omega)\angle\phi(\omega)$ in polar form, is subjected to the sinusoidal input $M_i(\omega)\angle\phi_i(\omega)$, the steady-state output as a consequence of the Convolution Theorem is the sinusoid

$$M_o(\omega)\angle\phi_o(\omega) = M_i(\omega)M(\omega)\angle[\phi_i(\omega) + \phi(\omega)].$$

When $M_i(\omega)\angle\phi_i(\omega) = 1\angle 0$, we get the *frequency response* of the system. A formal definition follows.

Definition: Frequency response [DB11, p. 554], [Nis15, p. 528]

. . . is the *steady-state* response of a system to a *sinusoidal* input signal of unit magnitude and zero phase. Given frequency response $M(\omega)\angle\phi(\omega)$, $M(\omega)$ is called the magnitude (frequency) response, and $\phi(\omega)$ the phase (frequency) response.

It is important to remember that frequency response is only defined for the steady state. If $G(s)$ is a transfer function, then the associated frequency response is simply $G(j\omega)$, that is, the frequency response $G(j\omega)$ is simply the transfer function $G(s)$ evaluated on the imaginary axis [Nis15, pp. 528–529]. $G(j\omega)$ is sometimes called the *sinusoidal transfer function*.

The Bode plot or Bode diagram is the standard tool for depicting a frequency response. The Bode plot has two sub-plots:

1. A **magnitude plot** of $|G(j\omega)|$ in decibels (dB) against ω in rad/s: With this plot, we study the magnitude or gain of the system. $|G(j\omega)|$ in decibels (dB) is defined as

$$20 \log |G(j\omega)|.$$

In the context of frequency response, every log should be understood as $\log_{10}$. The advantage of using dB is that it converts multiplication to addition, e.g.,

$$20 \log |KG(j\omega)| = 20 \log |K| + 20 \log |G(j\omega)|.$$

Quiz 1

If $G(j\omega)$ has a dc gain (i.e., gain at $\omega = 0$) of 0 dB, and $KG(j\omega)$ has a dc gain of 20 dB, how much is K in absolute units?

2. A **phase plot** of $\angle G(j\omega)$ in degrees against ω in rad/s: With the phase plot, we study the phase shift induced by the system.

For both plots, the vertical axis is linearly spaced, whereas the horizontal axis is logarithmically spaced.

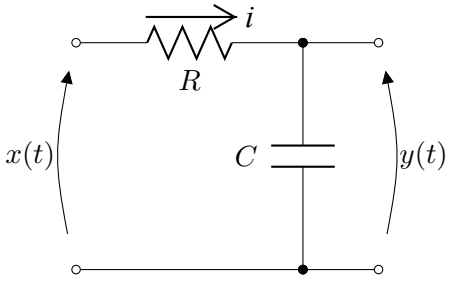


Figure 3: RC circuit from Lecture 2.

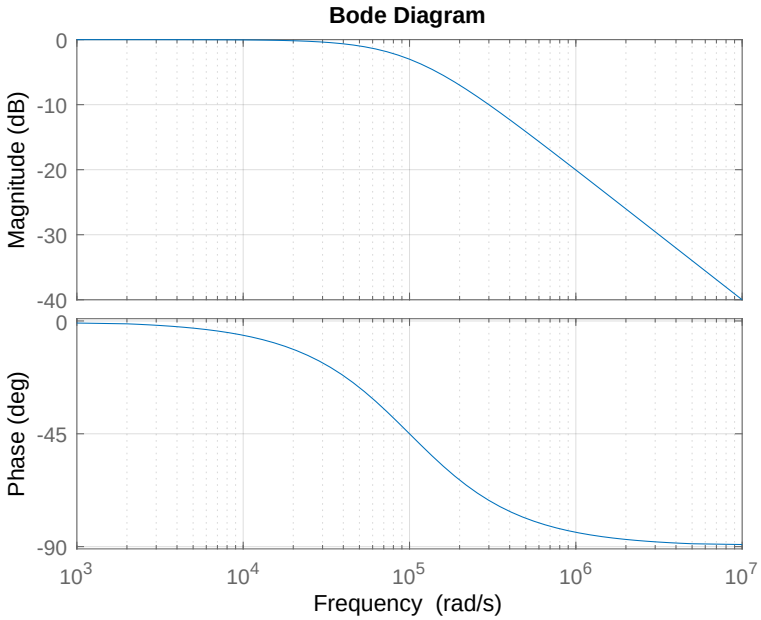


Figure 4: Bode diagram for the RC circuit in Figure 3.

Example 1

The RC circuit from Lecture 2 (see Figure 3) is a first-order low-pass filter. The Bode diagram in Figure 4, which you have most likely seen in one of your earlier circuit courses, is readily obtainable using the following MATLAB code:

```
R = 10; C = 1e-6; G = tf(1, [R*C 1]); bode(G)
```

The magnitude plot clearly shows how input signals with very low frequencies are passed through at (near) unity gain, while input signals with very high frequencies are passed through

attenuated — the higher the frequency, the deeper the attenuation. For example, an input signal at $\omega = 10^6$ rad/s is attenuated at -20 dB, which is equivalent to being multiplied by $10^{-20/20} = 0.1$. In fact, the filtering is considered to be cut off at $\omega = RC = 10^5$ rad/s — this is the so-called *cutoff frequency*, which we shall define later.

Figure 5: A NASA-Army-Bell XV-15 tiltrotor research aircraft. Image by NASA - commons file, Public Domain, `https://commons.wikimedia.org/w/index.php?curid=66531306`.

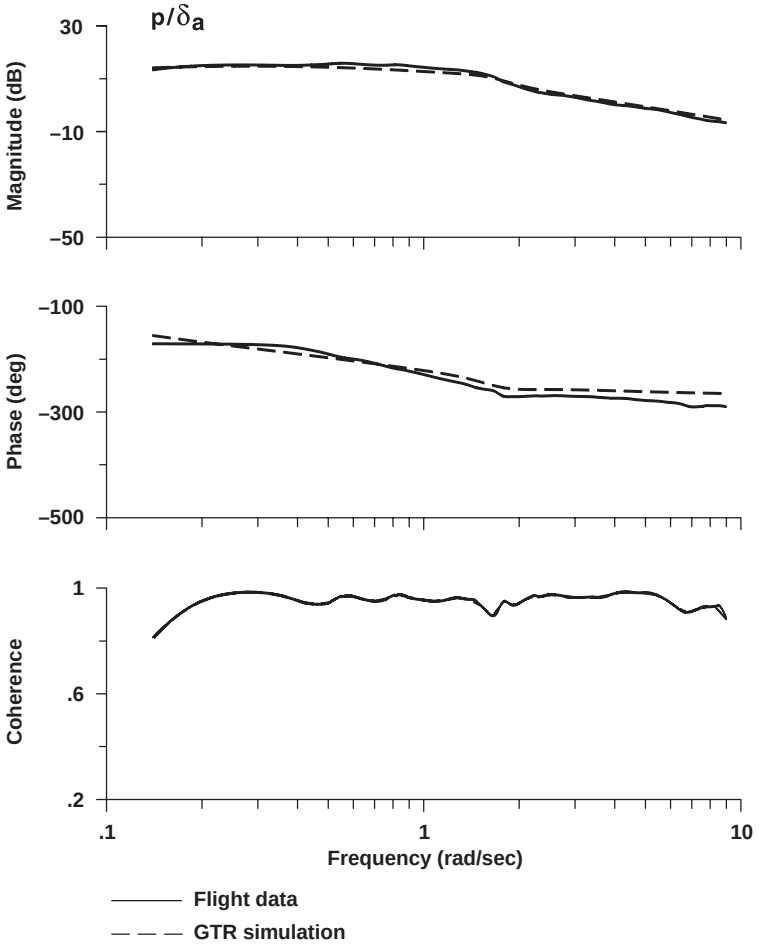


Figure 6: Bode diagram [TR06, Fig. 2.5, p. 38] for the roll rate (p) of the aircraft in Figure 5 in hover mode, with aileron angle (δ_a) as the input. In the legend, the acronym "GTR" is short for "generic tiltrotor". The coherence measure in the third plot measures the fit between the data and model, where a fit of more than 0.6 is considered good.

Example 2

Frequency response is routinely used for system identification, especially in aerospace engineering [TR06]. For example, the frequency response of the roll rate (p) of the Bell-NASA XV-15 tiltrotor in hover mode (see Figure 5), with the aileron angle (δ_a) as the input, can be

experimentally found (see Figure 6). Roll rate refers to how fast the aircraft rolls. Based on the experimentally obtained frequency response data, the δ_a-to-p transfer function can be estimated (see [TR06, Eq. (2.9)]). Figure 6 shows a good fit between Bell Helicopter's generic tiltrotor mathematical model and the flight-test data, but such a good fit is not always achievable in practice.

3 Sketching frequency responses

Just as we can sketch root loci, we can sketch frequency responses — this was very useful before the proliferation of computers. Being able to sketch frequency responses is still useful as a way of understanding frequency responses and troubleshooting system design. The focus here is identifying the salient characteristics of Bode plots, and not so much on the sketching. Consider the frequency response

$$G(j\omega) = K\frac{\Pi_{k=1}^{m}(j\omega + z_k)}{\Pi_{l=1}^{n}(j\omega + p_l)}. \tag{3}$$

- The log-magnitude is given by

$$20\log|G(j\omega)| = 20\left[\log|K| + \frac{1}{2}\sum_{k=1}^{m}\log(z_k^2+\omega^2) - \frac{1}{2}\sum_{l=1}^{n}\log(p_l^2+\omega^2)\right],$$

  which is simply a sum of the log-magnitude of the factors of the numerator minus the sum of the log-magnitude of the factors of the denominator — being able to convert multiplication/division into addition/subtracting is one of the advantages of using unit dB.

- The phase is given by

$$\angle G(j\omega) = \sum_{k=1}^{m}\text{atan2}(\omega, z_k) - \sum_{l=1}^{n}\text{atan2}(\omega, p_l).$$

  which is simply a sum of the phase of the factors of the numerator minus the sum of the phase of the factors of the denominator.

Thus, to sketch the frequency response of $G(j\omega)$ in Eq. (3), we only need to know how to sketch the frequency responses of

- K (magnitude plot: $20\log|K|$, phase plot: 0, trivial),
- $j\omega$ (magnitude plot: $20\log\omega$, phase plot: 90°, see Figure 7),
- $j\omega + a$, where $a \in \mathbb{R}\setminus\{0\}$ (see Sect. 3.1),
- $j\omega + a$, where $a \in \mathbb{C}\setminus\{0\}$ (a comes in a conjugate pair, see Sect. 3.2).

Note how given $G(j\omega)$, inverting $G(j\omega)$ amounts to

- flipping the magnitude curve of $G(j\omega)$ about the 0-dB line, because $20\log(G(j\omega))^{-1} = -20\log G(j\omega)$;
- flipping the phase curve of $G(j\omega)$ about the 0-degree line, because $\angle(G(j\omega))^{-1} = -\angle G(j\omega)$.

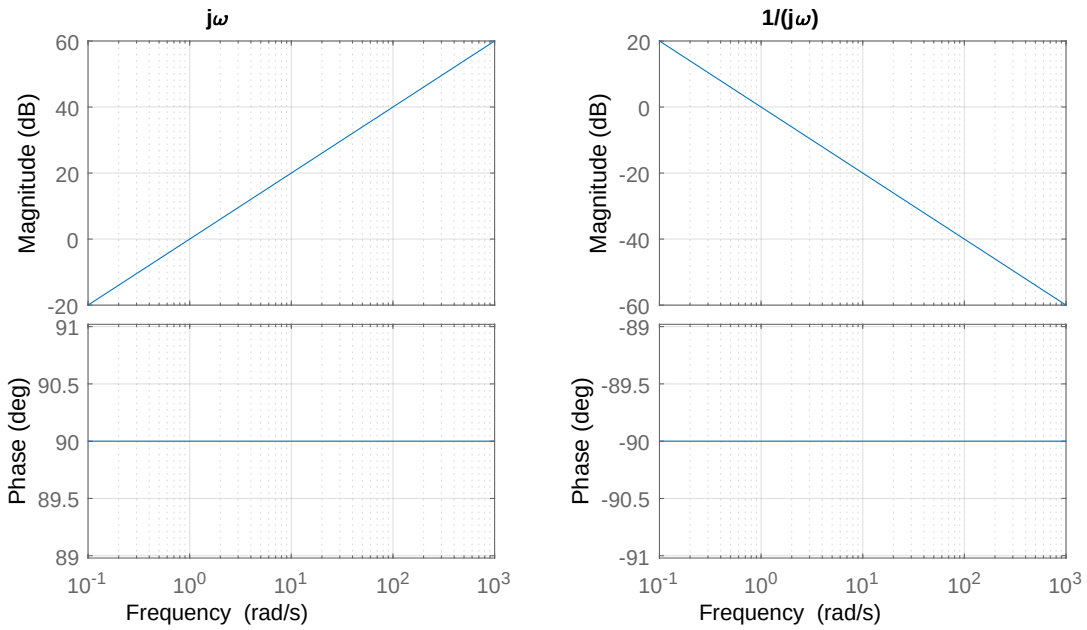


Figure 7: Bode diagrams for $j\omega$ and $1/(j\omega)$. The magnitude gradients are 20 dB per decade and -20 dB per decade respectively.

Quiz 2

Sketch the frequency response of $G(s) = \sqrt{2}/s$.

3.1 $j\omega + a$, where $a \in \mathbb{R} \setminus \{0\}$

Let $G(j\omega) = j\omega + a$, where $a \in \mathbb{R} \setminus \{0\}$. The log-magnitude of $G(j\omega)$ is given by

$$20 \log |G(j\omega)| = 20 \log \sqrt{\omega^2 + a^2}.$$

There are three significant cases:

1. When $\omega \ll a$,

   $$20 \log |G(j\omega)| \approx 20 \log |a|,$$

   providing the *low-frequency asymptote*. So at low frequencies, the log-magnitude behaves like a constant.

2. When $\omega \gg a$,

   $$20 \log |G(j\omega)| \approx 20 \log \omega,$$

   providing the *high-frequency asymptote*. So at high frequencies, the log-magnitude behaves like a ramp. Along the high-frequency asymptote, whenever we increase ω by 10 times, the log-magnitude gains an additional $20 \log 10 = 20$ dB. This is why we say the gradient of the high-frequency asymptote is 20 dB per decade.

3. When $\omega = a$,

   $$20 \log |G(j\omega)| = 10 \log(2a^2) = 10 \log 2 + 20 \log |a| \approx 3 + 20 \log |a|.$$

The frequency $\omega = a$ is called the *break frequency*, *corner frequency*, or *cutoff frequency*. The break frequency can be seen as where the low- and high-frequency asymptotes meet. At the break frequency, the actual gain of $G(j\omega) = j\omega + a$ is given exactly by $\sqrt{2}$ times its dc gain in absolute units, or roughly 3 plus its dc gain in dB.

Quiz 3

Given $G(j\omega) = 1/(j\omega + a)$, where a is real and nonzero. If $G(j\omega)$ has a gain of -3 dB at $\omega = a$, what is the dc gain of $G(j\omega)$? Furthermore, what is the value of a?

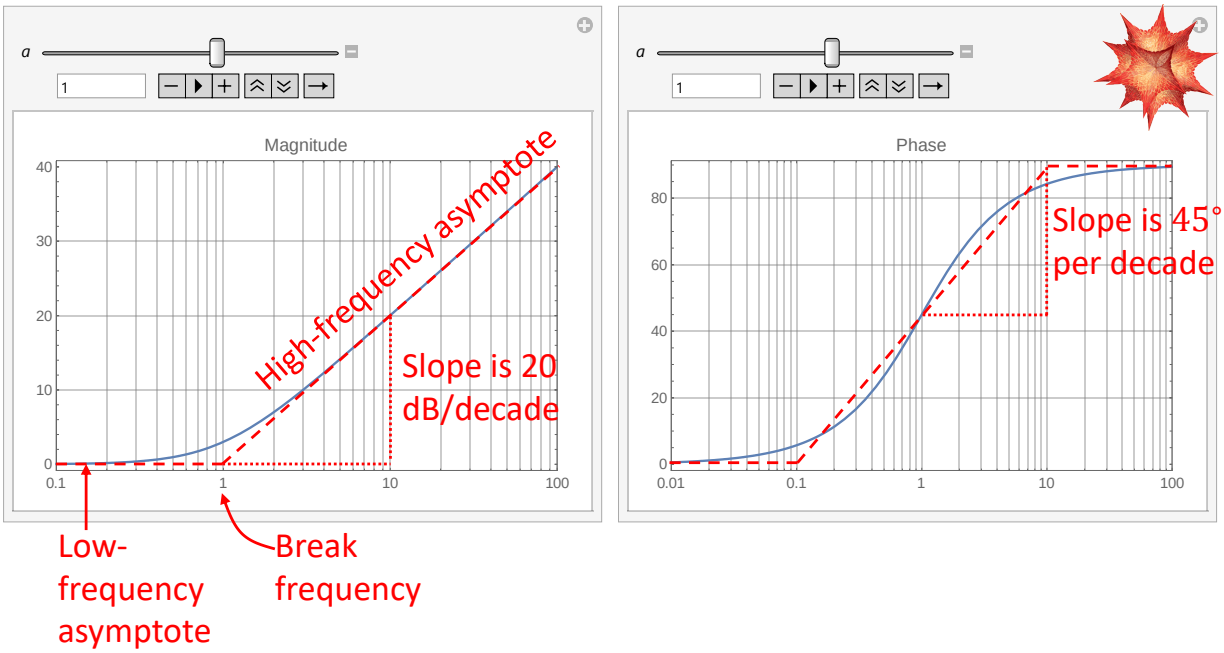


Figure 8: Bode diagram for $j\omega + 1$, with sketch shown as dashed (not dotted) lines.

While the sign of the zero a does not affect the magnitude plot of $G(j\omega) = j\omega + a$, it does affect the phase plot. The phase of $G(j\omega)$ is given

$$\angle G(j\omega) = \text{atan2}(\omega, a).$$

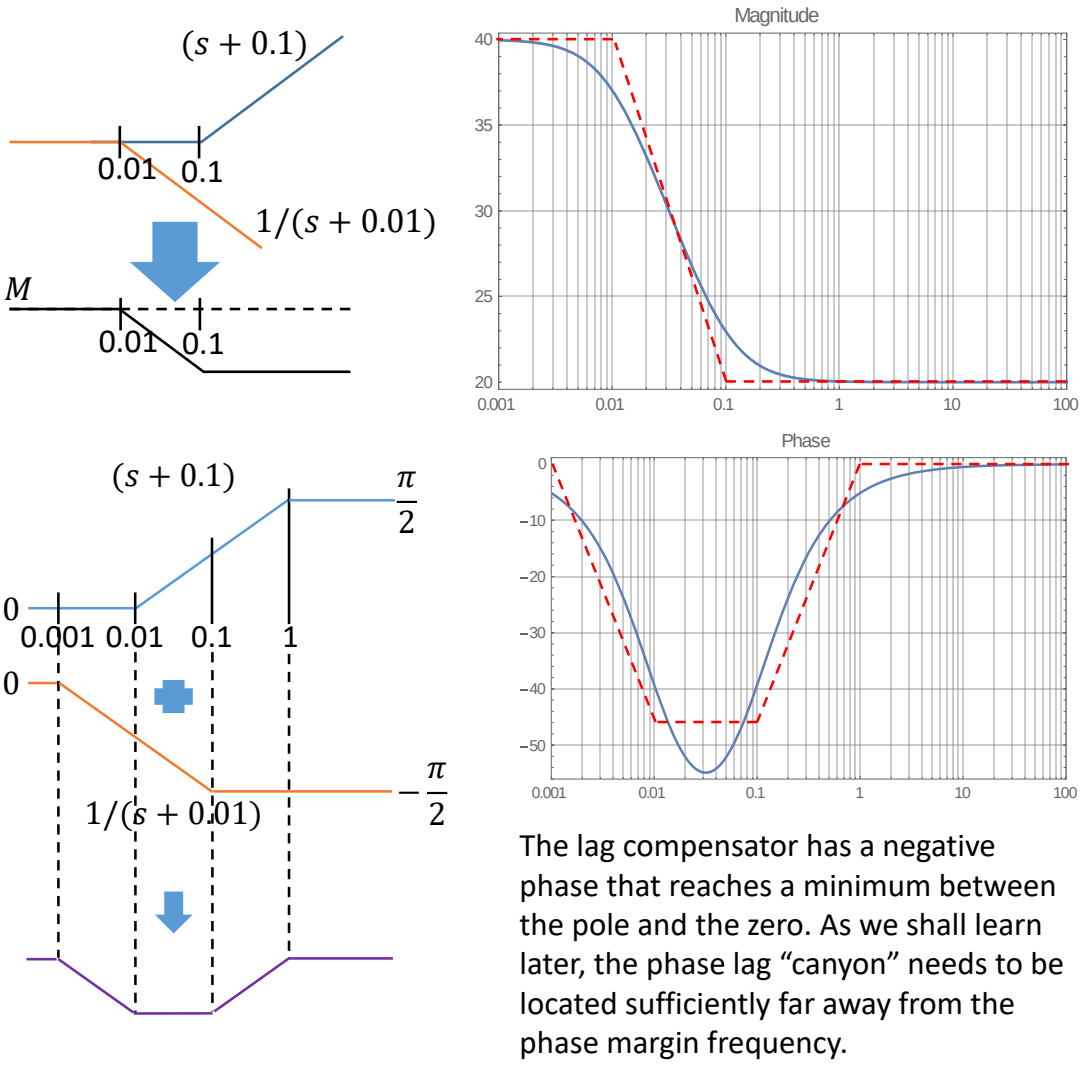
If $a > 0$, then

- when $\omega \ll a$, $\angle G(j\omega) \approx \lim_{\omega \to 0^+} \text{atan2}(0, a) = 0$;
- when $\omega = a$, $\angle G(j\omega) \approx \text{atan2}(a, a) = 45°$;
- when $\omega \gg a$, $\angle G(j\omega) \approx \lim_{\omega \to \infty} \text{atan2}(\omega, a) = 90°$.

If $a < 0$, then

- when $\omega \ll |a|$, $\angle G(j\omega) \approx \lim_{\omega \to 0^+} \text{atan2}(0, a) = 180°$;
- when $\omega = -a$, $\angle G(j\omega) \approx \text{atan2}(-a, a) = 135°$;
- when $\omega \gg |a|$, $\angle G(j\omega) \approx \lim_{\omega \to \infty} \text{atan2}(\omega, a) = 90°$.

As an example, Figure 8 shows a sketch for $j\omega + 1$, along with the exact plots.



The lag compensator has a negative phase that reaches a minimum between the pole and the zero. As we shall learn later, the phase lag "canyon" needs to be located sufficiently far away from the phase margin frequency.

Figure 9: Sketching the frequency response of the lag compensator in Example 3.

Example 3

Sketch the frequency response of the lag compensator $G(s) = 10(s + 0.1)/(s + 0.01)$.

Solution: The magnitude curve for $(s + 0.1)$ has the blue asymptotes in Figure 9, while the magnitude curve for $1/(s+0.01)$ has the orange asymptotes in Figure 9 — overlapping the two sets of asymptotes, we get the black asymptotes labeled M.

The low-frequency asymptote of M is given by

$$20 \log 10 + 20 \log(0.1) - 20 \log 0.01 = 40,$$

while the high-frequency asymptote is given by

$$40 - 20 \left(\frac{0.1/0.01}{10} \right) = 20,$$

all in unit dB.

Sketching the phase response is more troublesome, but follows the same principle, as shown in Figure 9. Luckily, the low- and high-frequency phase asymptotes are easily obtained as $0° + 0° = 0°$ and $90° - 90° = 0°$ respectively.

3.2 $(j\omega)^2 + j2\zeta\omega_n\omega + \omega_n^2$

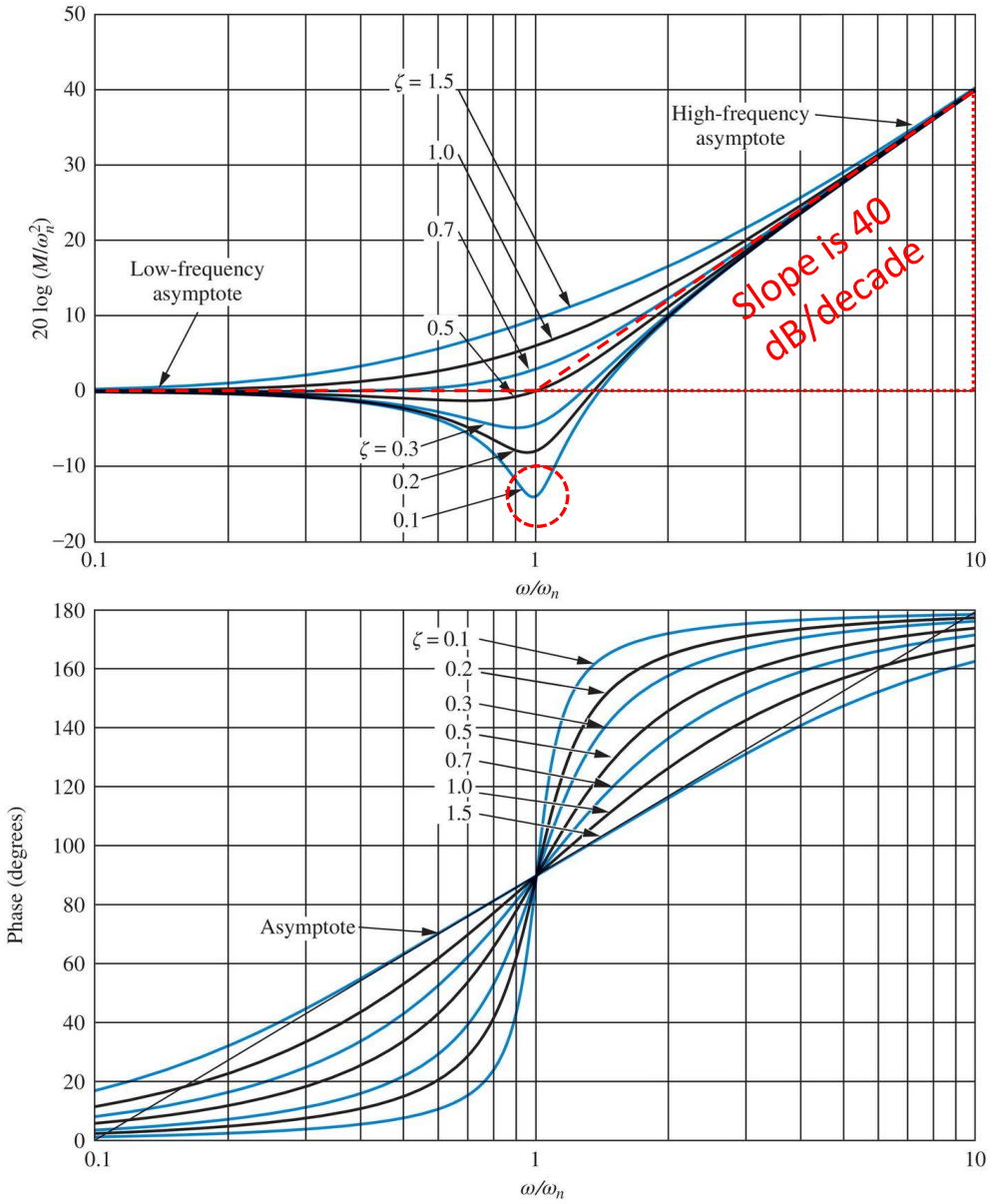


Figure 10: Frequency responses of underdamped second-order systems: the more underdamped the system, the more pronounced the resonant peak, and the more the actual magnitude curve differs from the asymptotes near the resonant frequency. The phase approaches 0 for $\omega \ll \omega_n$, reaches exactly $90°$ at $\omega = \omega_n$, and approaches $180°$ for $\omega \gg \omega_n$. Figure adapted from [Nis15, Fig. 10.14–10.15].

If $H(j\omega)$ is physically realizable and has a polynomial factor $j\omega + a$, where $a \in \mathbb{C} \setminus \{0\}$, then $H(j\omega)$ must also have polynomial factor $(j\omega + \bar{a})$ where $\bar{a}$ is a's conjugate. In this case, we are interested in the Bode plot for

$$G(j\omega) \stackrel{\text{def}}{=} (j\omega + a)(j\omega + \bar{a}) = (j\omega)^2 + j(a + \bar{a})\omega + a\bar{a}.$$

Setting $a \stackrel{\text{def}}{=} \zeta\omega_n + j\omega_n\sqrt{1 - \zeta^2}$, then we get the more familiar-looking expression

$$G(j\omega) = (j\omega)^2 + j2\zeta\omega_n\omega + \omega_n^2. \tag{4}$$

It is possible to sketch the frequency response associated with Eq. (4), but it gets tedious quickly because the difference between the asymptotic approximation and the actual frequency response varies with ζ, as shown in Figure 10. Where the magnitude peaks (downward), the frequency is called the *resonant frequency* — NOT the same as the natural frequency. We have heard a lot about resonance and the need to understand it, for example, on YouTube: `https://youtu.be/_7jENWKgMPY`, where the failure to take resonance into account resulted in the mechanical disintegration of an aircraft structure, costing the life of the pilot. Let us see how we can calculate the resonant frequency. For $G(j\omega)$ defined in Eq. (4),

$$|G(j\omega)|^2 = (\omega_n^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2.$$

Using a bit of calculus, $|G(j\omega)|^2$ reaches a stationary point, when

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\omega}\left\{(\omega_n^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2\right\} = 0 \implies & -4\omega(\omega_n^2 - \omega^2) + 8\zeta^2\omega_n^2\omega = 0 \\ \implies & -4\omega\left\{(\omega_n^2 - \omega^2) - 2\zeta^2\omega_n^2\right\} = 0.\end{aligned}$$

We know the resonant frequency $\neq 0$, so

$$(\omega_n^2 - \omega^2) - 2\zeta^2\omega_n^2 = 0 \implies \omega^2 = \omega_n^2(1 - 2\zeta^2).$$

Therefore, the resonant frequency is

$$\omega_r = \begin{cases}\omega_n\sqrt{1-2\zeta^2}, & \text{if } 0 \le \zeta < 1/\sqrt{2}; \\ \text{undefined}, & \text{otherwise.}\end{cases} \tag{5}$$

For $\zeta \ge 1/\sqrt{2} \approx 0.7071$, the system is well damped, and there is no resonant peak; this explains why in Control System Designer, the default desired damping ratio is 0.7071.

Suppose $G(j\omega)$ takes the standard second-order form:

$$G(s) = \frac{K}{s^2 + 2\zeta\omega_n s + \omega_n^2} \implies G(j\omega) = \frac{K}{\omega_n^2 - \omega^2 + j2\zeta\omega_n\omega}.$$

When $\omega = \omega_r$, $|G(j\omega)|$ reaches its peak and

$$\begin{aligned}|G(j\omega_r)|^2 &= \frac{K^2}{|\omega_n^2 - \omega_r^2 + j2\zeta\omega_n\omega_r|^2} = \frac{K^2}{\left|\omega_n^2 - \omega_n^2(1-2\zeta^2) + j2\zeta\omega_n^2\sqrt{1-2\zeta^2}\right|^2} \\ &= \frac{K^2}{\omega_n^4} \cdot \frac{1}{(2\zeta^2)^2 + \left(2\zeta\sqrt{1-2\zeta^2}\right)^2} = \frac{K^2}{\omega_n^4} \cdot \frac{1}{4\zeta^2(1-\zeta^2)}.\end{aligned} \tag{6}$$

Therefore, the resonant peak has magnitude

$$|G(j\omega_r)| = \frac{K}{\omega_n^2} \cdot \frac{1}{2\zeta\sqrt{1-\zeta^2}}. \tag{7}$$

Quiz 4

If $G(s) = \dfrac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$, where $\zeta < 1/\sqrt{2}$, determine the magnitude of the resonant peak.

At this point, we know how to capture first- and second-order frequency responses in Bode plots, the basic characteristics of these frequency responses, and how to assemble the frequency responses of high-order systems using these first- and second-order frequency responses. The next thing we need to find out is how to extract closed-loop information from Bode plots of open-loop frequency responses, the most pertinent of which being stability, steady-state error, and transient response. We visit these topics in turn in the subsequent sections.

4 Bode plots: Stability

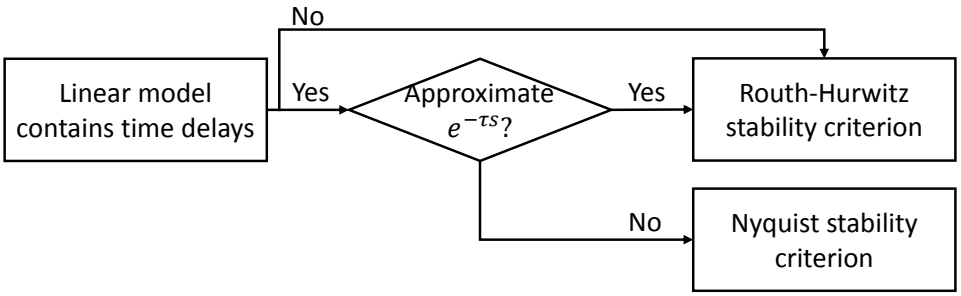


Figure 11: Flowchart for choosing stability criterion.

In the 1930s, as part of Bell Labs' program to develop negative feedback amplifiers for transcontinental telephony, Harry Nyquist investigated the stability of linear feedback systems [Bis09]. The result is the *Nyquist stability criterion*, which has been called "the heart of classical control" [Mac13]. The Nyquist criterion is applicable to transfer functions with dead time, and in fact to any frequency response model — this is especially useful when we cannot derive a mathematical model but can measure the frequency response of a plant. The Nyquist criterion lies in the core of the frequency response approach to controller design. The flowchart in Figure 11 suggests how we can choose between the Routh-Hurwitz and Nyquist stability criteria.

Due to its applicability to plants with time delays, the Nyquist criterion is widely used in the process control industry and for networked control systems, where time delays are prevalent. The concept of *relative stability* in terms of *gain margin* and *phase margin* originates from the Nyquist criterion, and is a pillar of robust control. Robust control is the design of controllers for plants for which there is no accurate model. In aerospace engineering, where model uncertainties are common and the stability requirement is especially stringent, robust control and hence the Nyquist criterion are essential.

4.1 Cauchy's principle of the argument

The Nyquist criterion is grounded in Cauchy's *principle of the argument* ("argument" here refers to the angle of a complex number), a theorem from complex maths that is related to the mapping of contours.

Definition: Contour [PS06]

A *contour* is a curve having a sectionally continuous derivative. A function is *sectionally continuous* on an interval if it has at most a finite number of discontinuities, with right- and left-hand limits at each point in the interval.

By definition, a contour does not have to be continuous, but for our purpose, we can confine ourself to continuous contours. Contours arise in complex maths because of integration on the complex plane, as opposed to integration on a real interval. Think of a contour as a set of complex numbers. Just as a complex function maps a complex number to another complex number, a complex function maps a contour to another contour.

Example 4

The complex function $F(s) = \frac{1}{s}$ maps the complex number $a + jb$ to

$$\frac{1}{a+jb} = \frac{a-jb}{(a+jb)(a-jb)} = \frac{a}{a^2+b^2} - j\frac{b}{a^2+b^2},$$

where a and b are real numbers that are not zero at the same time.

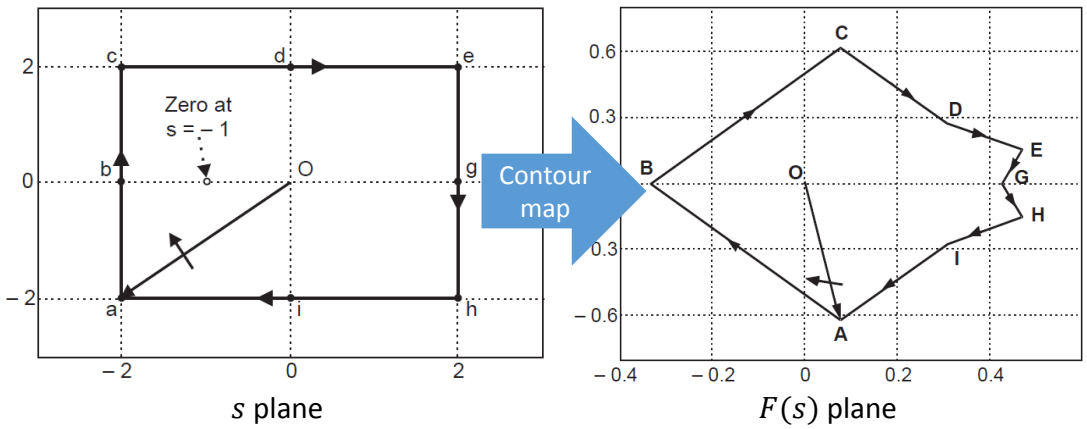


Figure 12: A complex function maps a contour in the s plane to a contour in the $F(s)$ plane. Point a in the s plane is mapped to point A in the $F(s)$ plane. Other points are similarly mapped.

Example 5

This example is taken from [Man06, Sect. 6.1.2]. In Figure 12, the complex function $F(s) = \frac{s+1}{s+5}$ maps a contour in the s plane to a contour in the $F(s)$ plane. For example, point $a = -2 - j2$ is mapped to point

$$A = F(a) = \frac{a+1}{a+5} = \frac{-1-j2}{3-j2} = \frac{(-1-j2)(3+j2)}{13} = \frac{1}{13} - j\frac{8}{13} = 0.0769 - j0.6154.$$

Other mappings can be similarly deduced.

Suppose contour Γ_s is mapped through $F(s)$ to contour Γ_F, then we have

Theorem 1: Cauchy's principle of the argument [DB11, p. 639]

If Γ_s encircles Z zeros and P poles of $F(s)$, and *not* pass through them, in the *clockwise* direction, Γ_F encircles the origin $N = Z - P$ times in the *clockwise* direction.

- Changing the direction of encirclement (clockwise $\leftrightarrow$ counterclockwise) changes the sign of N.
- [DB11, pp. 639-641] gives an intuitive explanation of the theorem.
- Figure 13 shows some sample applications of the theorem.

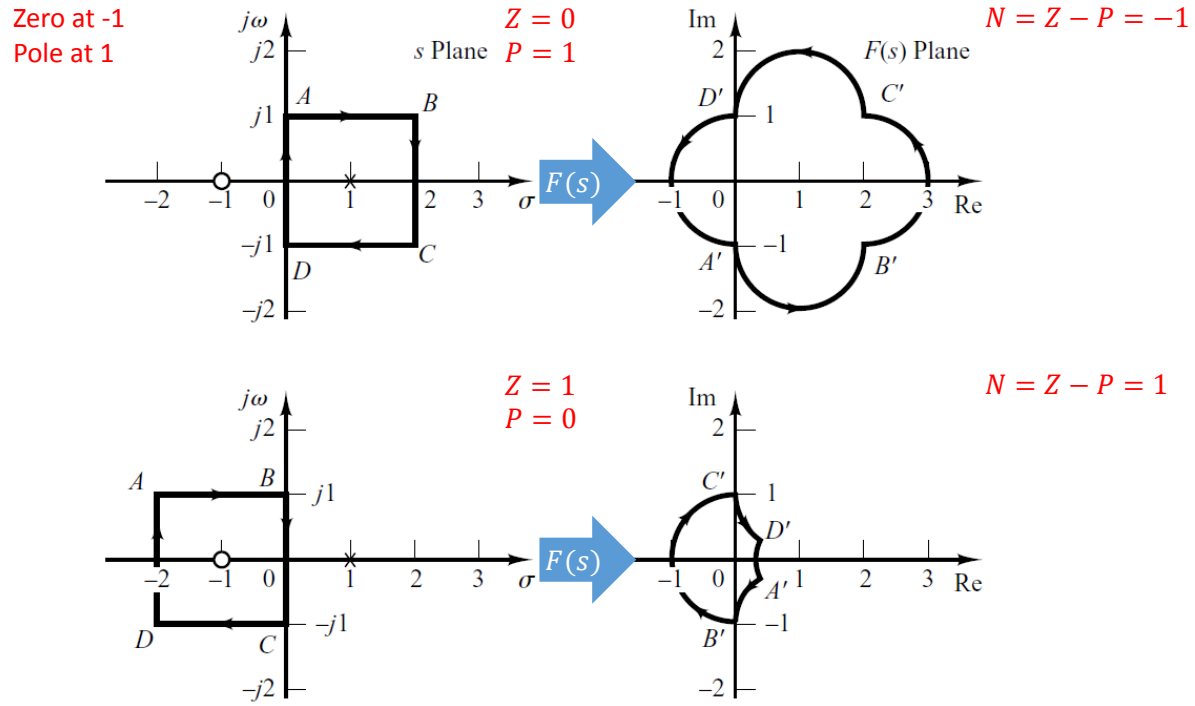


Figure 13: In these examples, $F(s) = \dfrac{s+1}{s-1}$. It is straightforward to verify the examples satisfy Cauchy's principle of the argument. Image from [Oga10, Figure 7-46].

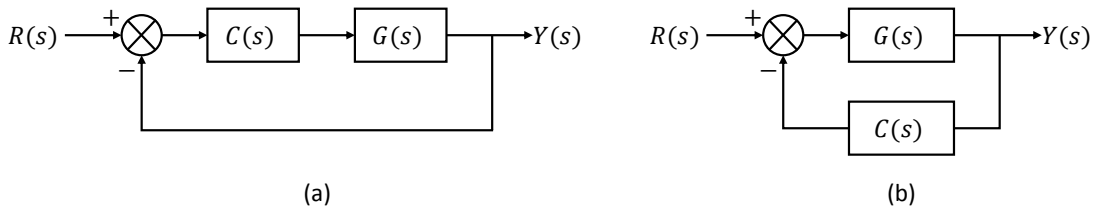


Figure 14: Control system configurations: (a) cascade compensation, (b) feedback compensation.

Interesting observations can be made if the complex function happens to be a *characteristic polynomial*. Consider the cascade-compensated system in Figure 14(a), with the closed-loop transfer function

$$T(s) = \frac{C(s)G(s)}{1 + C(s)G(s)}. \tag{8}$$

Let $C(s) \stackrel{\text{def}}{=} \dfrac{C_N(s)}{C_D(s)}$ and $G(s) \stackrel{\text{def}}{=} \dfrac{G_N(s)}{G_D(s)}$, then

$$1 + C(s)G(s) = \frac{C_D(s)G_D(s) + C_N(s)G_N(s)}{C_D(s)G_D(s)}, \tag{9}$$

and hence the characteristic polynomial is

$$C_D(s)G_D(s) + C_N(s)G_N(s). \tag{10}$$

Since the closed-loop transfer function of the feedback-compensated system in Figure 14(b) is

$$T(s) = \frac{G(s)}{1 + C(s)G(s)}, \tag{11}$$

it is clear that both systems in Figure 14 have the same characteristic polynomial. In fact, we knew this from Lecture 5. From Eq. (9), we can deduce the following facts which we will use later:

- The zeros of $1 + C(s)G(s)$ are the closed-loop poles of the system $T(s)$.
- The poles of $1 + C(s)G(s)$ are the open-loop poles of the system $T(s)$.

Now, define

$$F(s) \stackrel{\text{def}}{=} 1 + L(s), \qquad \text{where } L(s) \stackrel{\text{def}}{=} C(s)G(s), \tag{12}$$

then

- The zeros of $F(s)$ are the closed-loop poles.
- The poles of $F(s)$ are the open-loop poles.
- The origin of the $F(s)$ plane corresponds to the point $(-1, 0)$ on the $L(s)$ plane. Note that $L(s)$ is the *open-loop transfer function* or *loop gain*.

Suppose $F(s)$ maps a contour Γ_s in the s plane to a contour Γ_F in the $F(s)$ plane. Applying the principle of the argument, we can deduce

- If Γ_s encircles clockwise Z zeros and P poles of $F(s)$ in the s plane, then Γ_F encircles clockwise the origin of the $F(s)$ plane N times, where $N = Z - P$.

Since Γ_F in the $F(s)$ plane is equivalent to Γ_F left-shifted by 1 in the $L(s)$ plane, we can equivalently deduce

- If Γ_s encircles clockwise Z closed-loop poles and P open-loop poles of the system $T(s)$ in the s plane, then Γ_L encircles clockwise the point $(-1, 0)$ of the $L(s)$ plane N times, where $N = Z - P$. We shall call $(-1, 0)$ the *critical point*.

The immediately preceding statement is in a form that is very close to the Nyquist stability criterion.

4.2 Nyquist stability criterion

Suppose the s-plane contour Γ_s encircles clockwise the entire right-half plane, then Γ_s should encircle all possible unstable closed-loop poles and open-loop poles. Suppose furthermore we can determine the number of clockwise encirclements of the critical point in the $L(s)$ plane by Γ_L, where $L(s)$ is the open-loop transfer function. Then, we can determine the number of unstable closed-loop poles as $Z = N + P$, where N is the number of clockwise encirclements and P is the number of unstable open-loop poles.

- The preimage contour Γ_s in the s plane that encircles the entire right-half plane is called the *Nyquist path* (see Figure 15(a)).
  - By convention, Γ_s is clockwise.
  - If $L(s)$ has one or more poles/zeros on the imaginary axis, then Γ_s needs to be modified to bypass those poles/zeros in order for Cauchy's principle of the argument to be applicable. Figure 15(b) shows an example where $L(s)$ has a pole/zero at the origin.
- The image contour Γ_L in the $L(s)$ plane is called the *Nyquist curve*.
  - Since $Z = N + P \geq 0 \implies N \geq -P$, N can be positive or negative, i.e., Γ_L can be clockwise or counterclockwise.
  - The Nyquist curve is symmetrical about the real axis.
- The plot/diagram containing a Nyquist curve in a complex plane is called a *Nyquist diagram*, *Nyquist plot* or *polar plot*.

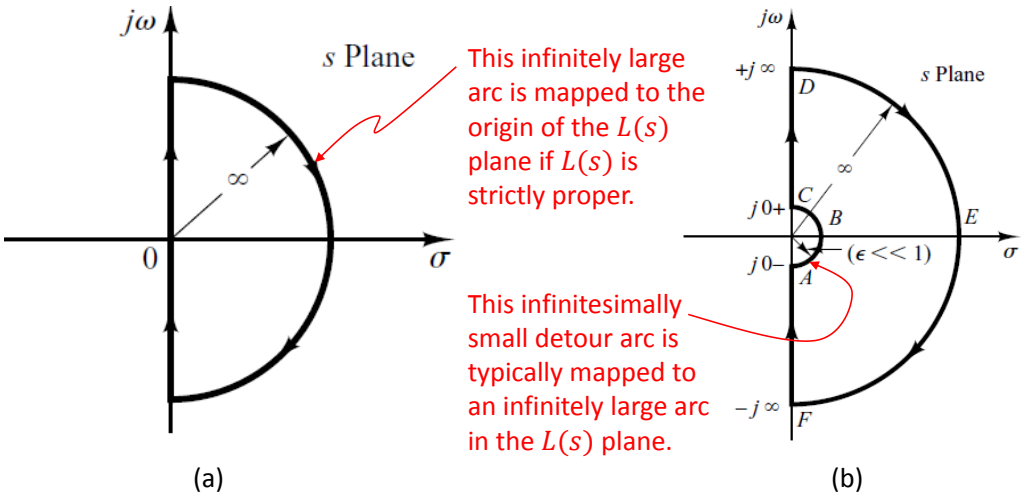


Figure 15: Nyquist path for (a) the case where $L(s)$ has no pole/zero on the imaginary axis; (b) the case where $L(s)$ has a pole/zero on the imaginary axis. Images from [Oga10, Figures 7-47 and 7-51].

At this point, we can state

Theorem 2: Nyquist stability criterion [DB11, Sect. 9.3]

A closed-loop system is stable if and only if, for the contour Γ_L, where $L(s)$ is the open-loop transfer function, the number of counterclockwise encirclements of the critical point $(-1, 0)$

is equal to the number of open-loop poles with positive real parts. Specifically, $Z = N + P$, where

- Z is the number of unstable poles of the *closed-loop* transfer function;
- P is the number of unstable poles of the *open-loop* transfer function;
- N is the number of clockwise encirclements of $(-1, 0)$.

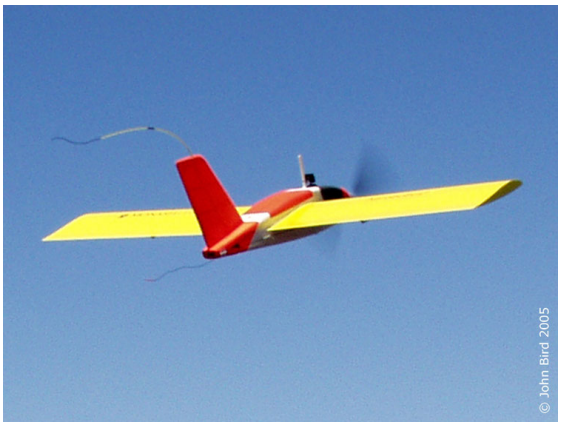


Figure 16: The Monash Aerobotics Research Group's P15035 UAV (see Example 6). Image from `http://www.ctie.monash.edu.au/hargrave/bairnsdale05.html`

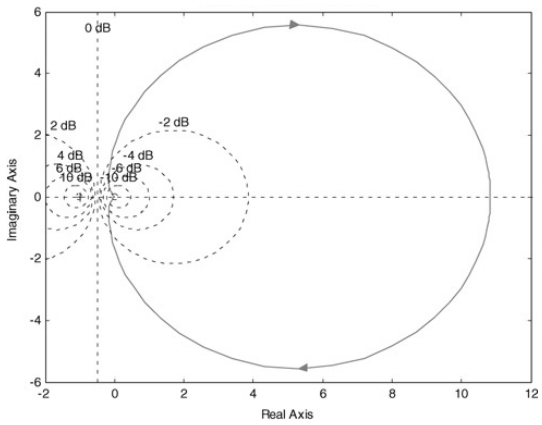


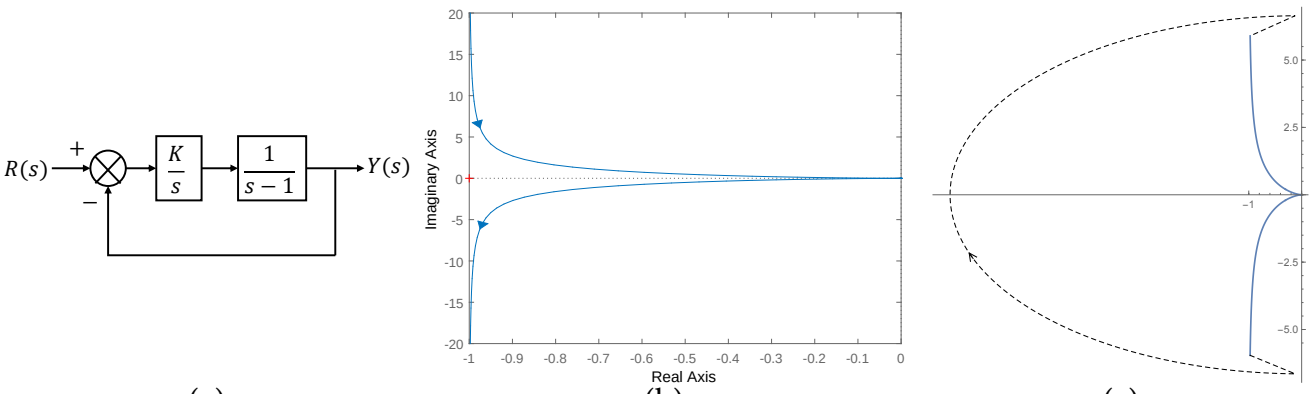
Figure 17: A Nyquist plot for the $\mathcal{H}_\infty$-based altitude control system of the P15035 (see Example 6). Image from [SLE08, Figure 8]

Example 6

The P15035 shown in Figure 16 is a fixed-wing UAV from the Monash Aerobotics Research Group. Like NASA's HiMAT, it uses elevons (elevator + aileron) for flight control; this makes the UAV nimble and mechanically simpler, but also underactuated with significant coupling between roll and yaw, making control challenging [LES15]. As with most aerospace applications, the precise model of the plant is difficult to obtain analytically, and it is common to use system identification techniques to estimate the transfer functions. For the inaccuracies that exist in these transfer functions, robust control has to be used to synthesize controllers for them. For the P15035, the elevon-average-to-altitude transfer function was estimated as,

$$\frac{h(s)}{\delta_A(s)} = \frac{-0.011659(s^2 + 11.88s + 42.46)(s^2 + 9.723s + 99.83)}{(s + 0.4633)(s + 0.1087)(s + 0.01552)(s^2 + 4.887s + 83.12)},$$

where δ_A is the average of the left and right elevon angles. For the estimated transfer function, an $\mathcal{H}_\infty$ controller was synthesized [SLE08]. We do not cover $\mathcal{H}_\infty$ techniques in this course, but it suffices to know they are widely used in aerospace engineering. The $\mathcal{H}_\infty$-compensated altitude control system has the Nyquist plot in Figure 17. Since the Nyquist curve does not encircle $(-1, 0)$, and there is no unstable open-loop pole, based on the Nyquist criterion, the number of unstable closed-loop poles is $Z = N + P = 0 + 0 = 0$, and hence the system is stable.



(a) (b) (c)

Figure 18: (a) The system, (b) its MATLAB-generated Nyquist plot, and (c) its Mathematica-generated Nyquist plot for Example 7.

Example 7

Consider the open-loop unstable system in Figure 18(a). Due to the open-loop pole at the origin, the Nyquist path needs to make a detour around the origin, as shown in Figure 15(b). What is the detour arc of the Nyquist path mapped to? Show the Nyquist plot for $K = 1$. Use the Nyquist criterion to determine the closed-loop stability for this value of K.

Solution: The open-loop transfer function is

$$L(s) = \frac{K}{s(s-1)}.$$

Due to the pole of $L(s)$ at the origin, the Nyquist path must take a detour around the origin; this detour arc is represented by $re^{j\theta}$, where $r \to 0$, $\theta \in [-\pi/2, \pi/2]$. So,

$$L(re^{j\theta}) = \frac{K}{re^{j\theta}(re^{j\theta}-1)} = \frac{K}{[r^2\cos(2\theta) - r\cos\theta] + j[r^2\sin(2\theta) - r\sin\theta]}.$$

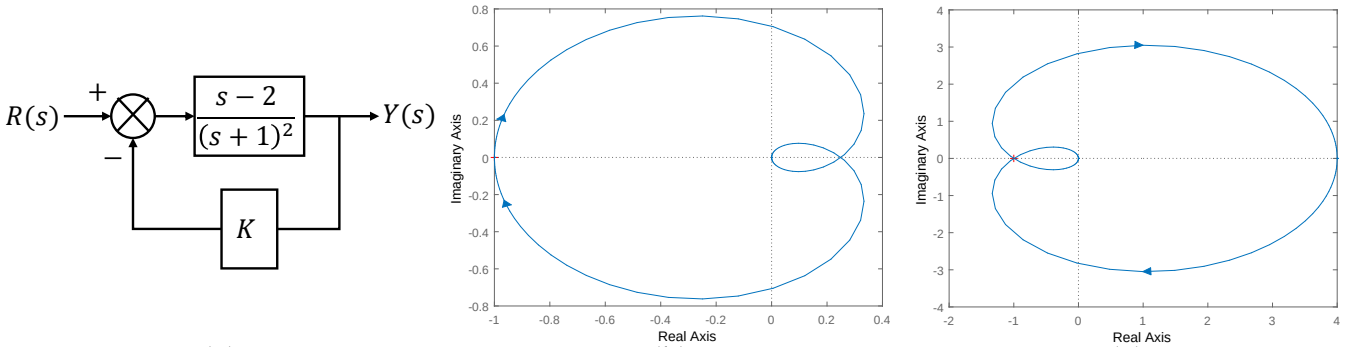
$$\therefore \lim_{r\to 0} |L(re^{j\theta})| = \infty.$$

$$\begin{aligned}\therefore \lim_{r\to 0} \angle L(re^{j\theta}) &= -\lim_{r\to 0} \text{atan2}(r^2\sin(2\theta) - r\sin\theta, r^2\cos(2\theta) - r\cos\theta) \\ &= -\lim_{r\to 0} \text{atan2}(r\sin(2\theta) - \sin\theta, r\cos(2\theta) - \cos\theta) \\ &= -\text{atan2}(-\sin\theta, -\cos\theta).\end{aligned}$$

- When $\theta \in [-\pi/2, 0]$ (fourth quadrant), $\sin\theta < 0$, $\cos\theta > 0$, $-\text{atan2}(+\text{ve}, -\text{ve})$ is in the 3rd quadrant.
- When $\theta \in [0, \pi/2]$ (first quadrant), $\sin\theta > 0$, $\cos\theta > 0$, $-\text{atan2}(-\text{ve}, -\text{ve})$ is in the 2nd quadrant.

Therefore, as θ ranges from $-\pi/2$ to $\pi/2$, $\lim_{r\to 0} \angle L(re^{j\theta})$ ranges from $3\pi/2$ to $\pi/2$. In other words, the detour arc is mapped to an infinite arc sweeping from $3\pi/2$ to $\pi/2$. The Nyquist plot for $K = 1$ is shown in Figure 18(b). It is unfortunate that MATLAB does not indicate the infinite arc, but based on our analysis, we know that the infinite arc causes the Nyquist curve to encircle $(-1, 0)$ clockwise once. The number of unstable closed-loop poles is thus $Z = N + P = 1 + 1 = 2$, thus the system is unstable.

Example 7 highlights an inconvenient fact: all software tools have their limitations, even the "standard" ones. This is why it pays to be knowledgeable in other tools besides the most popular one. For example, unlike MATLAB, Mathematica and WolframAlpha are able to indicate the infinite arc of any Nyquist curve. As evidence, compare the MATLAB-generated Nyquist plot in Figure 18(b) with the Mathematica version in Figure 18(c). MathWorks has presumably left this problem with Nyquist plots unfixed because control designers can get the stability information by other means, e.g., Bode plots and Nichols chart. Nevertheless, there exists third-party MATLAB code[5] for plotting complete Nyquist curves.



(a) (b) (c)

Figure 19: (a) The system, (b) its Nyquist plot for $K = 0.5$, and (c) its Nyquist plot for $K = -2$ in Example 8.

Example 8

Consider the nonminimum-phase system in Figure 19(a). Since there is no open-loop pole or zero on the imaginary axis, a detour-less Nyquist path can be used. Using the Nyquist criterion, determine the range of values of K for which the closed-loop system is stable.

Solution: Making the Nyquist plot for various values of K, we can see that beyond a certain value of K, the closed-loop system becomes unstable; our goal is to find this threshold value of K, denoted K_m. Since there is no unstable open-loop pole,

- when $K < K_m$, the Nyquist curve does not encircle $(-1, 0)$;
- when $K = K_m$, the Nyquist curve intersects $(-1, 0)$ exactly;
- when $K > K_m$, the Nyquist curve encircles $(-1, 0)$ at least once.

The detour-less Nyquist path consists mainly of two parts: one part on the imaginary axis, and another on the infinite arc. On the imaginary axis, $s = j\omega$, and

$$L(s) = L(j\omega) = K\frac{j\omega - 2}{(j\omega + 1)^2} = K\frac{-2 + j\omega}{1 - \omega^2 + j2\omega} = K\frac{(-2 + j\omega)(1 - \omega^2 - j2\omega)}{(1 - \omega^2)^2 + 4\omega^2}$$
$$= K\frac{4\omega^2 - 2 + j(5\omega - \omega^3)}{(\omega^2 + 1)^2} = \frac{K}{(\omega^2 + 1)^2}\left[4\omega^2 - 2 + j(5\omega - \omega^3)\right].$$

When the Nyquist curve intersects $(-1, 0)$, the real part $\Re\{L(j\omega)\} = -1$, and the imaginary

[5] https://au.mathworks.com/matlabcentral/fileexchange/43768-closed-logarithmic-nyquist-plot

part $\Im\{L(j\omega)\} = 0$. Let us deal with the imaginary part first:

$$\Im\{L(j\omega)\} = \frac{K}{(\omega^2+1)^2}(5\omega - \omega^3) = 0 \implies \omega(5-\omega^2) = 0 \implies \omega = 0, \pm\sqrt{5}.$$

- When $\omega = 0$,
$$\Re\{L(j\omega)\} = \Re\{L(0)\} = -2K = -1 \implies K = 0.5.$$
For $K > 0.5$, $\Re\{L(0)\} < -1$, i.e., the Nyquist curve encircles $(-1, 0)$.
- When $\omega = \pm\sqrt{5}$,
$$\Re\{L(\pm j\sqrt{5})\} = \frac{18K}{36} = -1 \implies K = -2.$$
For $K < -2$, $\Re\{L(\pm j\sqrt{5})\} < -1$, i.e., the Nyquist curve encircles $(-1, 0)$

Therefore, for $-2 < K < 0.5$, the system is stable, and for $K > 0.5$ or $K < -2$, the system is unstable. In Figs. 19(b)-(c), the Nyquist curves for $K = 0.5$ and $K = -2$ intersect $(-1, 0)$ exactly.

There are two important messages from Example 8:

1. The closed-loop stability of nonminimum-phase systems is sensitive to the gain.
2. Relative stability can be measured by observing how close a Nyquist curve is to intersecting the critical point. This measure comes in the form of two parameters: gain margin and phase margin, which are the topic of the next section.

4.3 Gain margin and phase margin

In Figure 20, so that it does not intersect or encircle the critical point, the Nyquist curve must cross the real axis to the right of the critical point. A Nyquist path has two main parts: one part on the imaginary axis, and another on the infinite arc. Assuming the infinite arc in the s plane is mapped to the origin of the $L(s)$ plane. Then, a point on the imaginary axis in the s plane is mapped to the negative-real-axis crossing of the Nyquist curve in the $L(s)$ plane. Suppose this point is $s = j\omega_{pc}$, then by definition,

$$\angle L(j\omega_{pc}) = -\pi.$$

The subscript "pc" means "phase crossover", i.e., crossing over a phase angle of $-\pi$ radians. Note that $-\pi$ is the same as π but by convention we use $-\pi$. ω_{pc} is called a *phase crossover frequency* [Oga10, Sect. 7.7]. At this point, we can define the gain margin.

Definition: Gain margin [DB11, Sect. 9.4], [Nis15, Sect. 10.6 and Glossary]

The **gain margin** is the multiplicative change in the open-loop gain required at the phase crossover frequency to make the closed-loop system marginally stable. Mathematically, if G_M is the gain margin, then $|G_M L(j\omega_{pc})| = 1$, i.e.,

$$G_M = \frac{1}{|L(j\omega_{pc})|} \implies G_{M,\text{dB}} = -20\log|L(j\omega_{pc})| \text{ dB}. \tag{13}$$

The phase crossover frequency ω_{pc} at which the gain margin is measured is called the *gain margin frequency*.

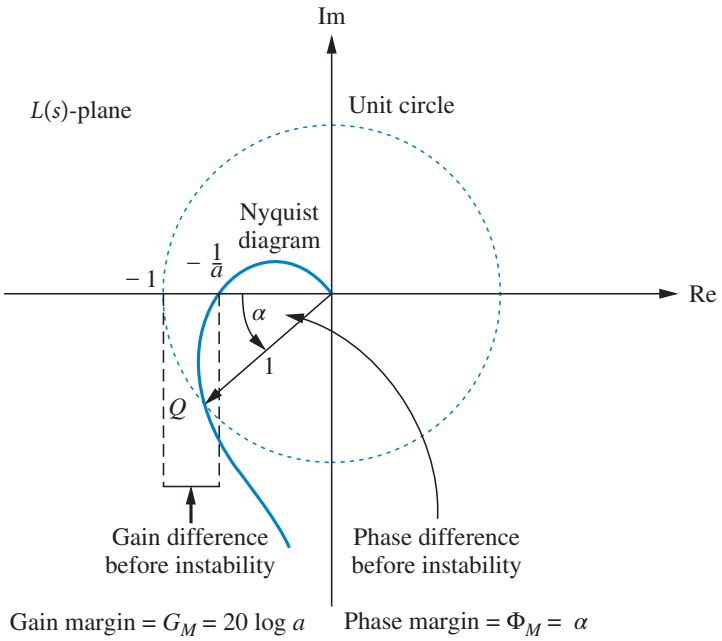


Figure 20: Gain and phase margins. Remember the Nyquist plot is symmetrical about the real axis. Image from [Nis15, FIGURE 10.35].

Quiz 5

A marginally stable system has a gain margin of how many in absolute value and in dB?

Again in Figure 20, where the Nyquist curve crosses the unit circle centered at the origin, the angle α must be positive so that the critical point is not intersected or encircled. Suppose the point $s = j\omega_{gc}$ in the s plane is mapped to the point Q in the $L(s)$ plane, then by definition,

$$|L(j\omega_{gc})| = 1.$$

The subscript "gc" means "gain crossover", i.e., crossing over a gain of one. ω_{gc} is called the *gain crossover frequency* [Oga10, Sect. 7.7]. At this point, we can define the phase margin.

Definition: Phase margin [DB11, Sect. 9.4], [Nis15, Sect. 10.6 and Glossary]

The **phase margin** is the amount of additional phase lag required at the gain crossover frequency to make the closed-loop system marginally stable. Mathematically, if Φ_M is the phase margin, then $\angle L(j\omega_{gc}) - \Phi_M = -\pi$, i.e.,

$$\Phi_M = \angle L(j\omega_{gc}) + \pi. \tag{14}$$

The gain crossover frequency ω_{gc} at which the phase margin is measured is called the *phase margin frequency*.

The following example shows how the gain and phase margins can be calculated. In MATLAB, we can also read these margins off a Nyquist plot, or by calling the function `margin`.

Example 9

This example is adapted from [Nis15, Example 10.8]. Consider the plant

$$L(s) = \frac{K}{(s^2+2s+2)(s+2)}, \qquad K > 0.$$

For the unity feedback system with $L(s)$ as the transfer function of the plant,

(a) Derive an expression for the gain margin (not in dB), and based on the expression, determine the upper bound of K for the system to be stable.

(b) Determine the gain margin (in dB) and phase margin (in degrees) if $K = 10$.

(c) Verify the gain and phase margins obtained earlier using the MATLAB functions `nyquist`, `margin` and `bode`.

(d) Verify the upper bound of K obtained in (a) using the Routh-Hurwitz criterion.

Solution:

(a) We start by deriving an expression for $L(j\omega)$:

$$\begin{aligned} L(j\omega) &= \frac{K}{(-\omega^2+j2\omega+2)(j\omega+2)} = \frac{K}{-j\omega^3-2\omega^2-2\omega^2+j4\omega+j2\omega+4} \\ &= \frac{K}{4-4\omega^2+j(6\omega-\omega^3)} = \frac{K}{(4-4\omega^2)^2+\omega^2(6-\omega^2)^2}\left[4-4\omega^2-j(6\omega-\omega^3)\right]. \end{aligned}$$

To find the gain margin G_M, we have to determine the phase crossover frequency ω_{pc}. By definition,

$$\angle L(j\omega_{pc}) = -\pi \implies 6\omega_{pc}-\omega_{pc}^3 = 0 \implies \omega_{pc}(6-\omega_{pc}^2) = 0 \implies \omega_{pc} = 0, \pm\sqrt{6}.$$

- When $\omega_{pc} = 0$,

$$\Re\{L(j\omega_{pc})\} = \Re\{L(0)\} = \frac{K}{4} > 0,$$

so $\omega_{pc} = 0$ is not the gain margin frequency.

- When $\omega_{pc} = \pm\sqrt{6}$,

$$\Re\{L(j\omega_{pc})\} = K\frac{4-4\omega_{pc}^2}{(4-4\omega_{pc}^2)^2+\omega_{pc}^2(6-\omega_{pc}^2)^2} = \frac{K}{4-4\omega_{pc}^2} = -\frac{K}{20} < 0,$$

so $\omega_{pc} = \pm\sqrt{6}$ is the gain margin frequency. So the gain margin is

$$G_M = \frac{1}{|L(j\omega_{pc})|} = \frac{20}{K}.$$

For the system to be stable

$$G_M = \frac{20}{K} > 1 \implies K < 20,$$

so the upper bound of K is 20.

(b) Based on the results so far, the gain margin for $K = 10$ is

$$G_{M,\text{dB}} = -20\log\left|L\left(\pm j\sqrt{6}\right)\right| = -20\log\left|-\frac{K}{20} + j0\right| = -20\log\left(\frac{10}{20}\right) = 6.0206 \text{ dB}.$$

To find the phase margin Φ_M, we have to determine the gain crossover frequency ω_{gc}. By definition,

$$\begin{aligned} |L(j\omega_{gc})| = 1 &\implies \frac{K}{\sqrt{(4-4\omega_{gc}^2)^2 + (6\omega_{gc} - \omega_{gc}^3)^2}} = 1 \\ &\implies (4-4\omega_{gc}^2)^2 + (6\omega_{gc} - \omega_{gc}^3)^2 = K^2 \\ &\implies \omega_{gc}^6 + 4\omega_{gc}^4 + 4\omega_{gc}^2 - 94 = 0. \end{aligned}$$

Using MATLAB, we can determine the only positive real solution to the equation above to be $\omega_{gc} = 1.7770$. The phase margin is thus

$$\begin{aligned} \Phi_M &= \angle L(j\omega_{gc}) + \pi = -\operatorname{atan2}(6\omega_{gc} - \omega gc^3, 4 - 4\omega_{gc}^2) + \pi \\ \implies \Phi_M &= -\operatorname{atan2}(6\cdot 1.7770 - (1.7770)^3, 4 - 4\cdot 1.7770^2) + \pi = 30.3357^\circ. \end{aligned}$$

(c) Using the code below, we can get the Nyquist diagram and Bode diagram in Figure 21, validating the gain and phase margins obtained analytically earlier.

```
L = tf(10, conv([1 2 2], [1 2]));
nyquist(L);
[Gm,Pm,Wgm,Wpm] = margin(L)
GmdB = 20*log10(Gm)
bode(L);
```

(d) To verify the upper bound of K obtained in (a) using the Routh-Hurwitz criterion, we need the characteristic polynomial, which is

$$(s^2 + 2s + 2)(s + 2) + K = s^3 + 4s^2 + 6s + (4 + K).$$

The Routh table is

s^3	1	6
s^2	4	$4 + K$
s^1	$(20 - K)/4$	
s^0	$4 + K$	

So that there is no sign change in the first column, we must have

$$(20 - K)/4 > 0 \implies K < 20 \qquad \text{and} \qquad 4 + K > 0 \implies K > -4.$$

The upper bound $K < 20$ is consistent with the result in (a). However, we did not get $K > -4$ earlier because we did not determine the range of values of K such that the phase margin is positive.

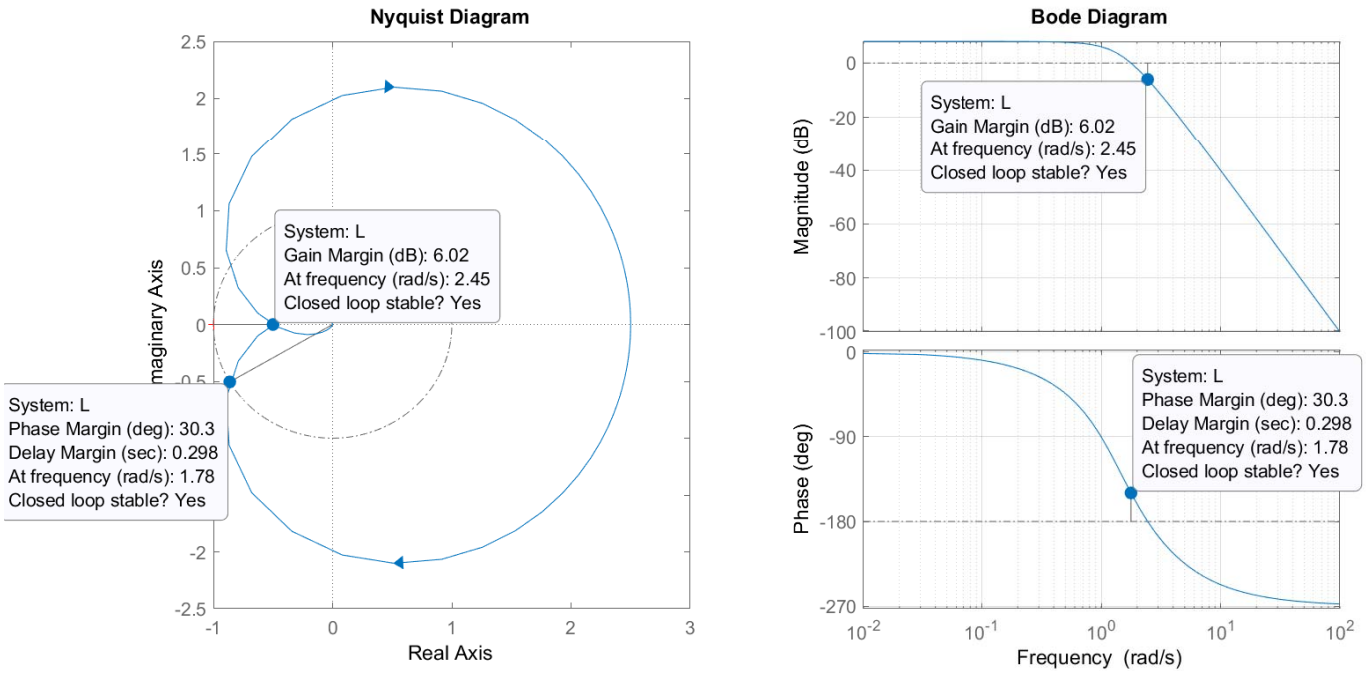


Figure 21: Nyquist and Bode diagrams for Example 9 showing gain and phase margins.

5 Bode plots: Steady-state error

Given a type-0 system of the form

$$G(s) = K\frac{\Pi_{i=1}^{n}(s+z_i)}{\Pi_{j=1}^{m}(s+p_j)},$$

where K, z_i, p_j are positive, the position error constant

$$K_p = \lim_{s\to 0} G(s) = K\frac{\Pi_{i=1}^{n} z_i}{\Pi_{j=1}^{m} p_j}$$

is finite and the same as the low-frequency asymptote of $|G(j\omega)|$.

Similarly, given a type-1 system of the form

$$G(s) = K\frac{\Pi_{i=1}^{n}(s+z_i)}{s\Pi_{j=1}^{m}(s+p_j)},$$

where K, z_i, p_j are positive, the velocity error constant is

$$K_v = \lim_{s\to 0} sG(s) = K\frac{\Pi_{i=1}^{n} z_i}{\Pi_{j=1}^{m} p_j}.$$

Note that when $|G(j\omega_1)| = 1$, so

$$K\frac{\Pi_{i=1}^{n} z_i}{\omega_1 \Pi_{j=1}^{m} p_j} = 1 \implies \omega_1 = K\frac{\Pi_{i=1}^{n} z_i}{\Pi_{j=1}^{m} p_j} = K_v.$$

In other words, K_v is simply the frequency where the magnitude curve intersects the 0-dB line, provided K, z_i, p_j are all positive.

Derivation for the acceleration error constant follows the same principle and is omitted here. Details can be found in [Nis15, Sect. 10.11]. Figure 22 summarizes the ways of reading static error constants from a Bode plot.

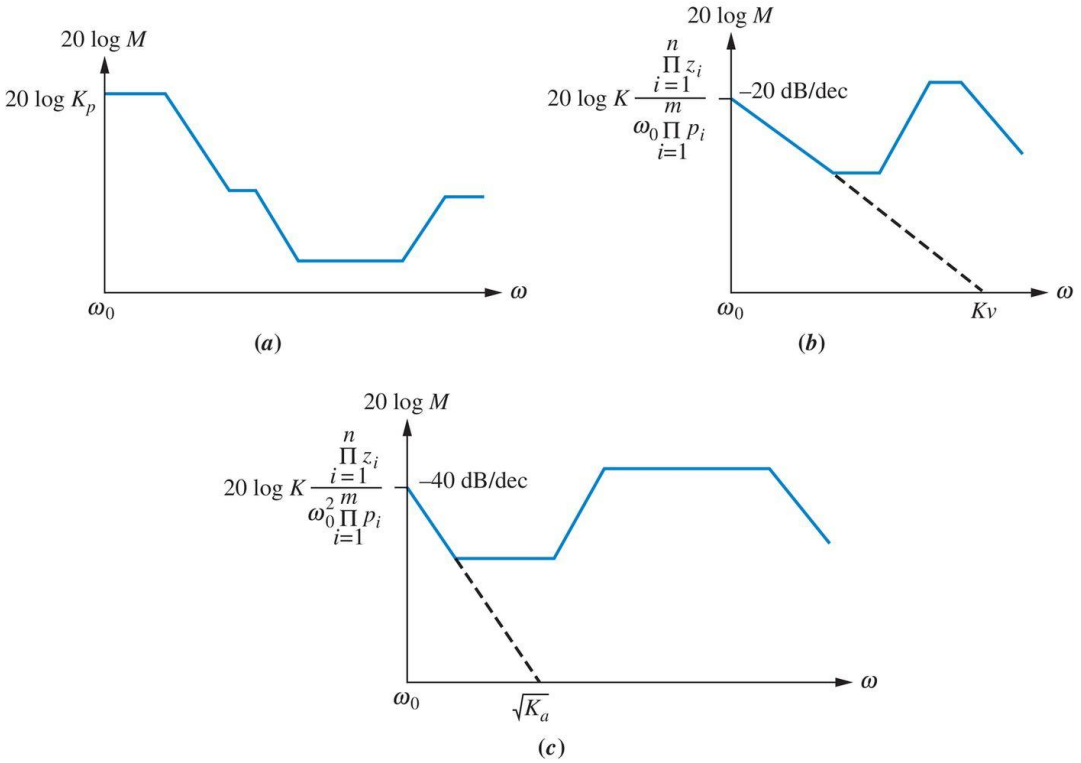


Figure 22: Reading static error constant from the Bode plot for (a) a type-0 system, (b) a type-1 system, (c) a type-2 system. Note all K, z_i, p_i are assumed to be positive. Image from [Nis15, FIGURE 10.51].

6 Bode plots: Transient response

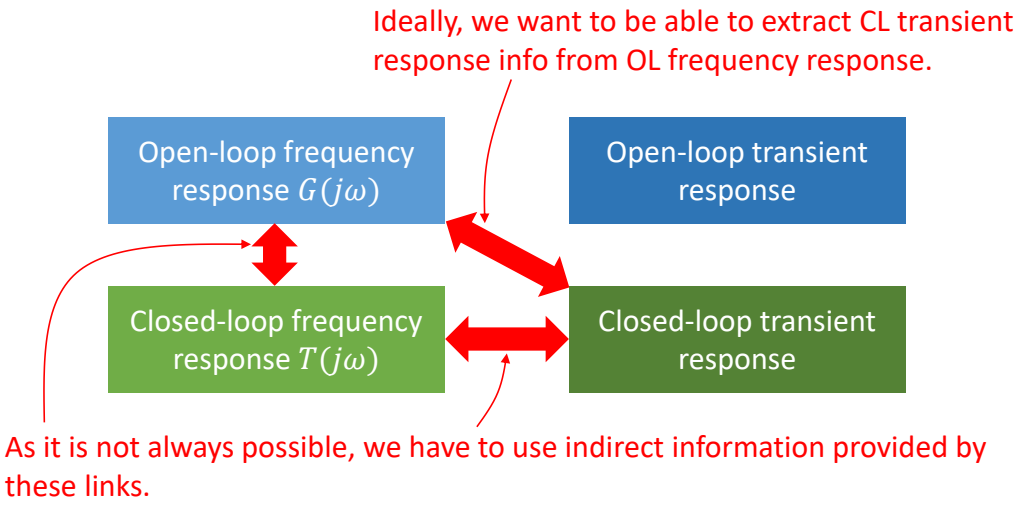


Figure 23: The links to be explored in order to extract CLTR information from OLFR.

The power of the root locus lies in its ability to express information about the closed-loop system based on the open-loop transfer function alone. To unleash the power of frequency response, we shall attempt to extract closed-loop information from open-loop information as well. Ideally, we would like to extract all *closed-loop transient response* (CLTR) information from the *open-loop frequency response* (OLFR) alone, but this is not possible. Instead, we shall find some of the CLTR

information embedded in the *closed-loop frequency response* (CLFR). In summary, we will explore the links

- between OLFR and CLTR,
- between CLFR and CLTR,
- between OLFR and CLFR,

as depicted in Figure 23.

6.1 Link between OLFR and CLTR

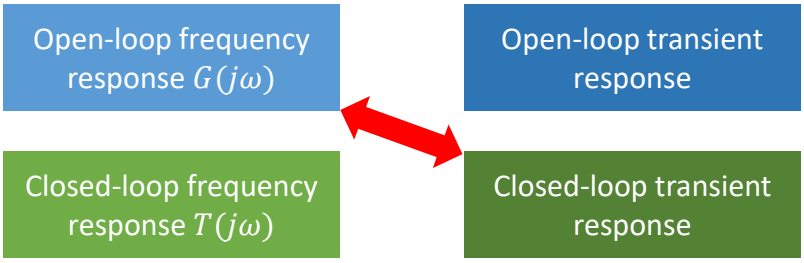


Figure 24: Sect. 6.1 explores the link between OLFR and CLTR.

A relationship between the *open-loop* phase margin and the *closed-loop* damping ratio / overshoot can be found. If we define the open-loop transfer function to be

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s}, \tag{15}$$

then the closed-loop transfer function of a unity feedback system, with $G(s)$ as the plant, is

$$T(s) = \frac{G(s)}{1 + G(s)} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \tag{16}$$

Notice the closed-loop transfer function is in the underdamped second-order form. Let us investigate the open-loop phase margin, i.e., phase margin of $G(j\omega)$, by determining the gain crossover frequency, ω_{gc}. At $\omega = \omega_{gc}$,

$$\begin{aligned} |G(j\omega_{gc})| = \left| \frac{\omega_n^2}{-\omega_{gc}^2 + j2\zeta\omega_n\omega_{gc}} \right| = 1 \implies & \omega_{gc}^4 + 4\zeta^2\omega_n^2\omega_{gc}^2 = \omega_n^4 \\ \implies & \omega_{gc}^4 + 4\zeta^2\omega_n^2\omega_{gc}^2 - \omega_n^4 = 0 \\ \implies & \omega_{gc}^2 = -2\zeta^2\omega_n^2 \pm \omega_n^2\sqrt{4\zeta^4 + 1} \\ \implies & \omega_{gc} = \omega_n\sqrt{-2\zeta^2 + \sqrt{4\zeta^4 + 1}}. \end{aligned} \tag{17}$$

At $\omega = \omega_{gc}$,

$$\begin{aligned} \angle G(j\omega_{gc}) &= -\operatorname{atan2}(2\zeta\omega_n\omega_{gc}, -\omega_{gc}^2) = -\operatorname{atan2}(2\zeta\omega_n, -\omega_{gc}) \\ &= -\left[\pi - \arctan\left(\frac{2\zeta\omega_n}{\omega_{gc}}\right)\right] = \arctan\left(\frac{2\zeta}{\sqrt{-2\zeta^2 + \sqrt{4\zeta^4 + 1}}}\right) - \pi. \end{aligned} \tag{18}$$

The open-loop phase margin is thus

$$\Phi_M = \angle G(j\omega_{gc}) + \pi = \arctan\left(\frac{2\zeta}{\sqrt{-2\zeta^2 + \sqrt{4\zeta^4 + 1}}}\right). \tag{19}$$

Eq. (19) relates the *open-loop phase margin* (Φ_M) to the *closed-loop damping ratio* (ζ), or equivalently, the overshoot of the closed-loop transient response. Figure 25 plots closed-loop ζ against open-loop Φ_M.

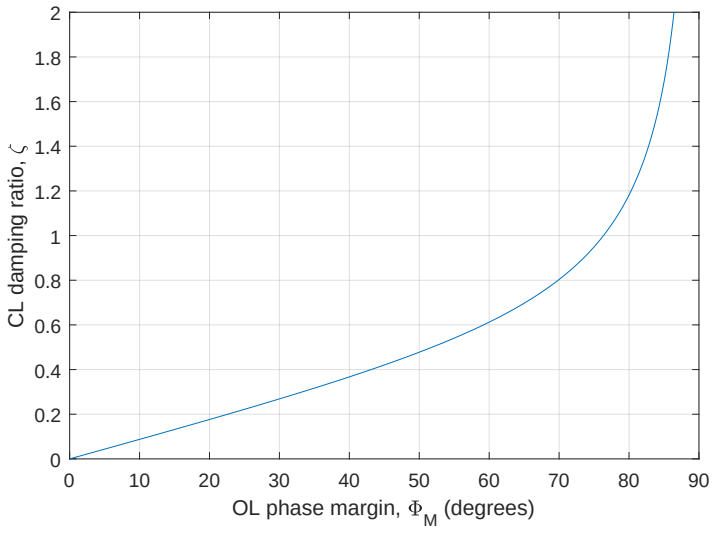


Figure 25: Plot of closed-loop ζ against open-loop Φ_M. Recall that when $\zeta \geq 1/\sqrt{2}$, the magnitude frequency response does not peak. When $\zeta = 1/\sqrt{2}$, the phase margin is about $65.52°$. So if a system's phase margin is more than $65.52°$, the system's magnitude does not peak. This is why it is common to specify a desired phase margin of $60°$.

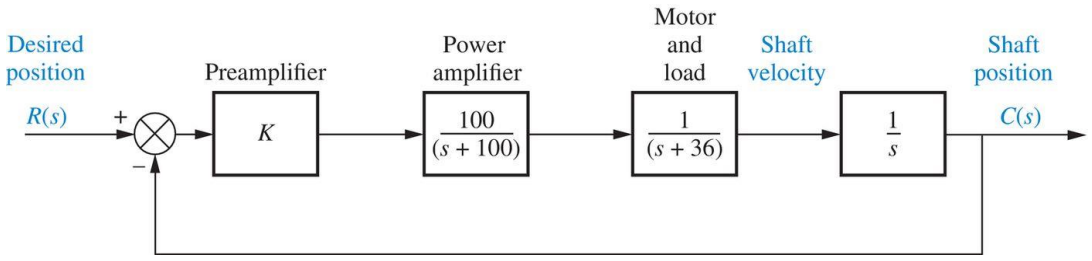


Figure 26: A motor position control system for Example 10. Image from [Nis15, Fig. 11-2].

Example 10

This example is adapted from [Nis15, Example 11.1]. For the position control system in Figure 26, use the frequency response method to find the value of the preamplifier gain, K, to yield a 10% overshoot in the closed-loop step response.

Solution: The loop gain without K is

$$G(s) = \frac{100}{s(s+100)(s+36)}.$$

A 10% overshoot for the closed-loop system corresponds to

$$\zeta = -\frac{\ln(0.1)}{\sqrt{\pi^2 + \ln^2(0.1)}} \approx 0.5912,$$

which in turn corresponds to an open-loop phase margin of

$$\Phi_M = \arctan\left(\frac{2\zeta}{\sqrt{-2\zeta^2+\sqrt{4\zeta^4+1}}}\right) \approx 58.5931^\circ.$$

There are multiple ways of determining a suitable value for K.

- A quick but imprecise way is to launch the Control System Designer using code:

  ```
  G = zpk([], [0 -100 -36], 100); controlSystemDesigner('bode', G);
  ```

  By moving the magnitude curve of $G(j\omega)$ up/down in Control System Designer until the phase margin reaches the calculated value, we can quickly find a suitable value for K (see Figure 27).
- A more precise way is to use the function `fzero`, for example through the following code fragment:

  ```
  fun_K4PM = @(K) PM-phaseMargin(K*G);
  K = fzero(fun_K4PM, 597)
  ```

  where `PM` contains the desired phase margin, and the function `phaseMargin` can be defined as such:

  ```
  function PM = phaseMargin(sys)
    [~,PM,~,~] = margin(sys);
  end
  ```

  A value of 597.7336 for K can be found in this way.

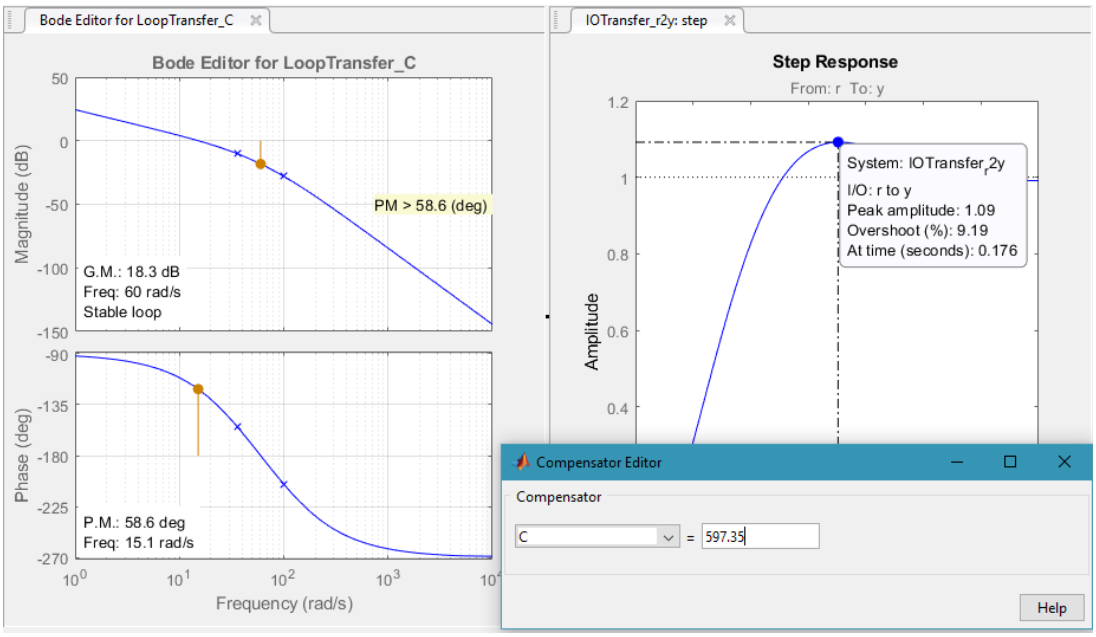


Figure 27: Using Control System Designer to determine K for a desired open-loop phase margin.

6.2 Link between CLFR and CLTR

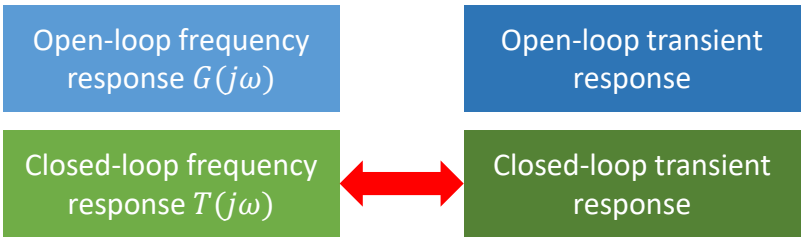


Figure 28: Sect. 6.2 explores the link between CLFR and CLTR.

The concept of *closed-loop bandwidth* plays an important role in frequency response methods.

Definition: Bandwidth [SMEDI10, Appendix J, Sect. J.5.2], [Nis15, p. 571]

. . . of a frequency response is the frequency at which the magnitude is $1/\sqrt{2}$ (i.e., about 3.0103 dB below) its value at zero frequency.

- In terms of gain, the bandwidth is the maximum frequency for a sinusoidal set-point to be attenuated by no more than a factor of $1/\sqrt{2}$, or equivalently, reduced by no more than about 3.0103 dB. For first-order systems of the form $b/(s+a)$, where $a > 0$, it is clear — as Figure 4 for example shows — that the break frequency is also the bandwidth.
- In terms of phase, the bandwidth indicates the frequency range for which satisfactory set-point tracking occurs. At frequencies well above the bandwidth, the phase lag induced by the system can cause the output to be significantly out of phase with the input (see Figure 29).

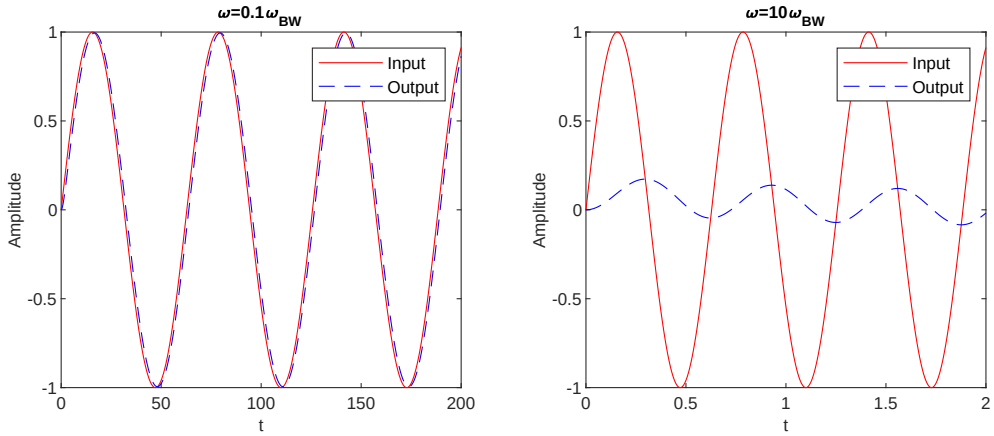


Figure 29: The output can be seen to track the input well for a frequency that is well below the bandwidth, but not for a frequency that is well above the bandwidth.

More importantly, the closed-loop bandwidth is also related to closed-loop transient response. Skipping derivation, the bandwidth, denoted ω_{BW}, is related to the damping ratio and natural frequency by

$$\omega_{BW} = \omega_n \sqrt{(1 - 2\zeta^2) + \sqrt{4\zeta^4 - 4\zeta^2 + 2}}. \tag{20}$$

If we approximate the square-root term by a straight line for $0.3 \leq \zeta \leq 0.8$, then we can find [DB11, Eq. (8.63)]

$$\omega_{BW} \approx \omega_n(-1.1961\zeta + 1.8508).$$

Furthermore, ω_{BW} is related to ζ and the settling time, $t_s \approx 4/(\zeta\omega_n)$, by

$$\omega_{BW} \approx \frac{4}{t_s\zeta}\sqrt{(1-2\zeta^2) + \sqrt{4\zeta^4 - 4\zeta^2 + 2}}. \tag{21}$$

We need Eqs. (20)–(21) for designing lead compensators later.

Quiz 6

Find the closed-loop bandwidth required for a 10% overshoot and 2.2-second settling time [Nis15, Exercise 10.7].

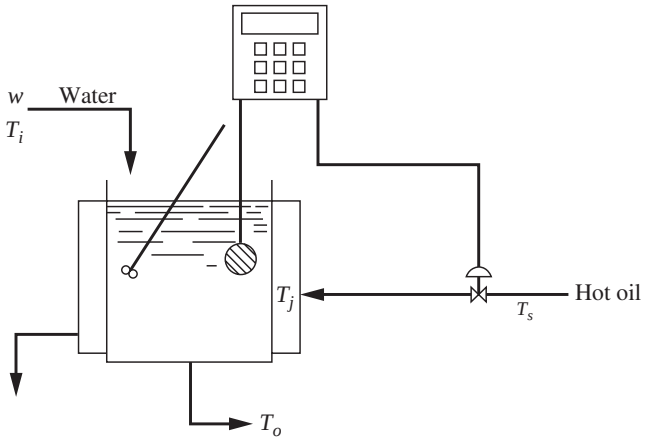


Figure 30: Single-loop control of a jacketed kettle. Image from [CL09, FIGURE 17-1(a)].

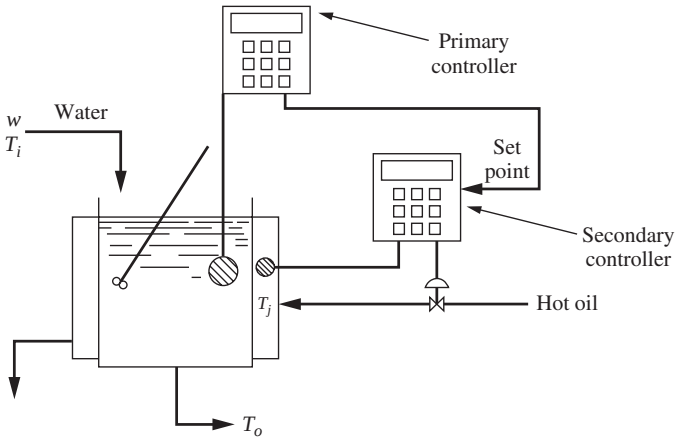


Figure 31: Cascade control of a jacketed kettle. Image from [CL09, FIGURE 17-1(b)].

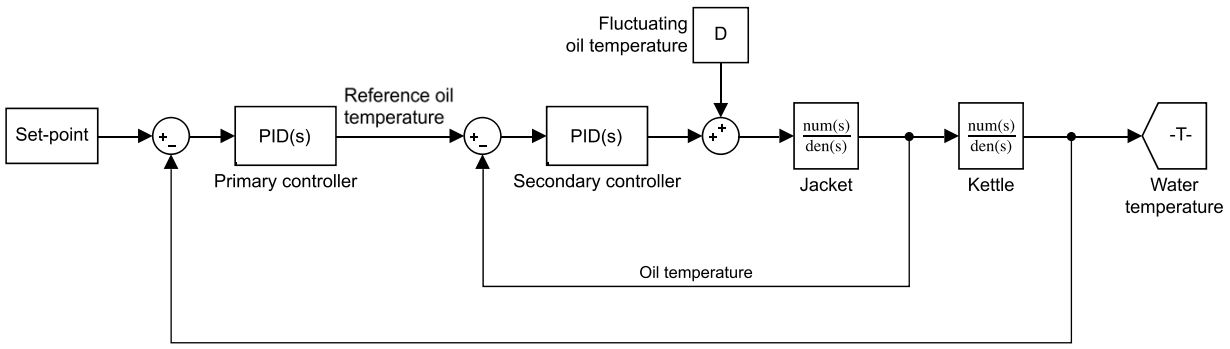


Figure 32: Block diagram for the cascade control system in Figure 31.

Example 11

In the process control industry, it is common to heat a kettle of water through a jacket of hot oil enveloping the kettle [CL09, Sect. 17.1]. This may seem straightforward but heat takes time to propagate, so any excursion in the oil temperature takes time to manifest in the water temperature reading, so suppose only the water temperature is controlled, then by the time the water temperature controller decides to increase/decrease heating, the oil temperature might have gone up/down. This time-delay process is challenging for a single-loop PID controller (see Figure 30). Enter *cascade control*, which involves a feedback loop within another feedback loop (see Figure 31 and Figure 32).

In this cascade control configuration, the *primary controller* in the outer loop tries to maintain water temperature at the set-point, by producing a reference oil temperature for the secondary controller. The *secondary controller* in the inner loop tries to maintain the oil temperature at that reference value. Essentially, the secondary controller plays the role of compensating for temperature propagation delay. In terms of frequency response, the inner control loop must have a *higher bandwidth* than the outer control loop.

6.3 Link between OLFR and CLFR

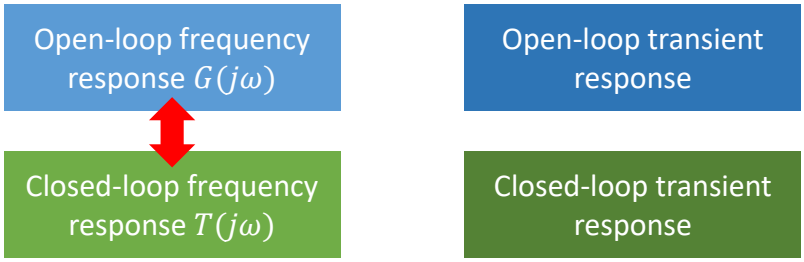


Figure 33: Sect. 6.3 explores the link between OLFR and CLFR.

The link between open-loop frequency response and closed-loop frequency response can be established with the help of the *Nichols chart*, concisely characterized by:

- an x-axis showing the *open-loop phase* (degrees),
  an y-axis showing the *open-loop gain* (dB),
- an origin at $(-180°, 0$ dB),
- grid lines showing *closed-loop gain* (dB) and *closed-loop phase* (degrees) — called the M *circles* and N *circles* respectively.

Compared to Bode plots, Nichols charts have the advantage of superimposing closed-loop information on open-loop information, but are unsketchable.

Example 12

The sample Nichols chart in Figure 34 is for a single-mass oscillator (mass-spring-damper) controlled with a robust PID controller. The gain margin is 17 dB and the phase margin is $52°$. At $\omega = 2.6$,

- OL gain: 0 dB
- OL phase: $-128°$

- CL gain: 1 dB
- CL phase: $-67.5°$ (hard to tell without the associated N circle)

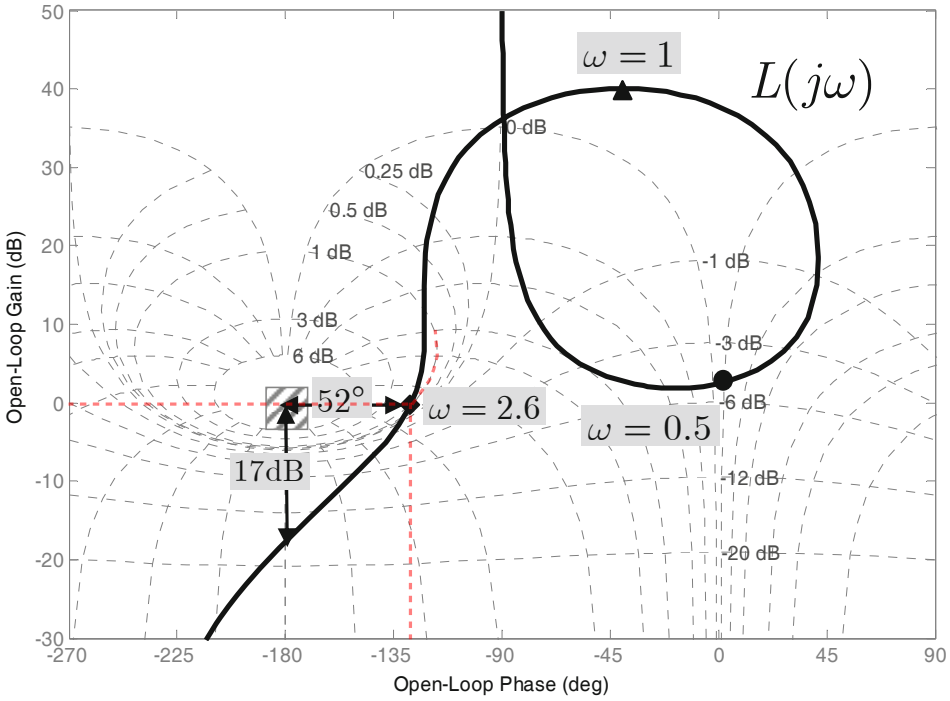


Figure 34: Nichols chart for Example 12 taken from [Jan12, Fig. 10.22].

Example 13

For the open-loop transfer function $G(s) = 100/(s(s+100)(s+36))$, the Nichols chart as shown in Figure 35 can be plotted using the MATLAB code below:

```
G = zpk([], [0 -100 -36], 100);
nichols(G); ngrid;
```

The main usage of the chart is to determine the closed-loop bandwidth, which is related to the settling time through Eq. (21). The closed-loop dc gain is 0 dB, so the closed-loop bandwidth is the ω corresponding to a closed-loop gain of -3 dB. Some quirks of the `nichols` GUI:

- We can only click on the curve in Full View.
- Zooming in while keeping the grid labels visible is problematic. The workaround: restrict the plot to a smaller frequency range, e.g., `nichols(G, 0.0001:0.0001:1)`.
- The MATLAB command `linearSystemAnalyzer('nichols', G)` actually brings up the same interface as `nichols(G)`.
- The MATLAB command `controlSystemDesigner('nichols', G)` provides an alternative interface for compensator design, but does not allow clicking on the curve.

Despite all the trouble with the Nichols chart, we could have obtained the closed-loop bandwidth using the command `bandwidth(feedback(G,1))`.

Nevertheless, the Nichols chart is an important tool for Quantitative Feedback Theory, a robust

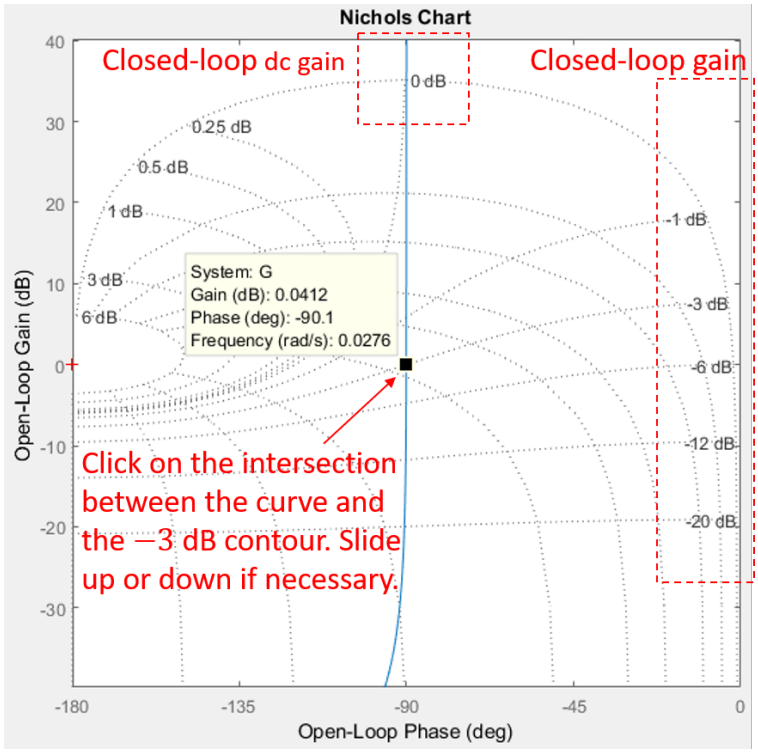


Figure 35: Nichols chart for Example 13.

control design technique. Otherwise, the Nichols chart is mainly used as an alternative way of displaying the Nyquist plot, and generally not frequently/widely used [FPEN15, p. 416], [SMEDI10, Appendix J].

At this point, we have finished exploring the direct and indirect linkes between open-loop frequency response and closed-loop transient response, as summarized in Figure 36.

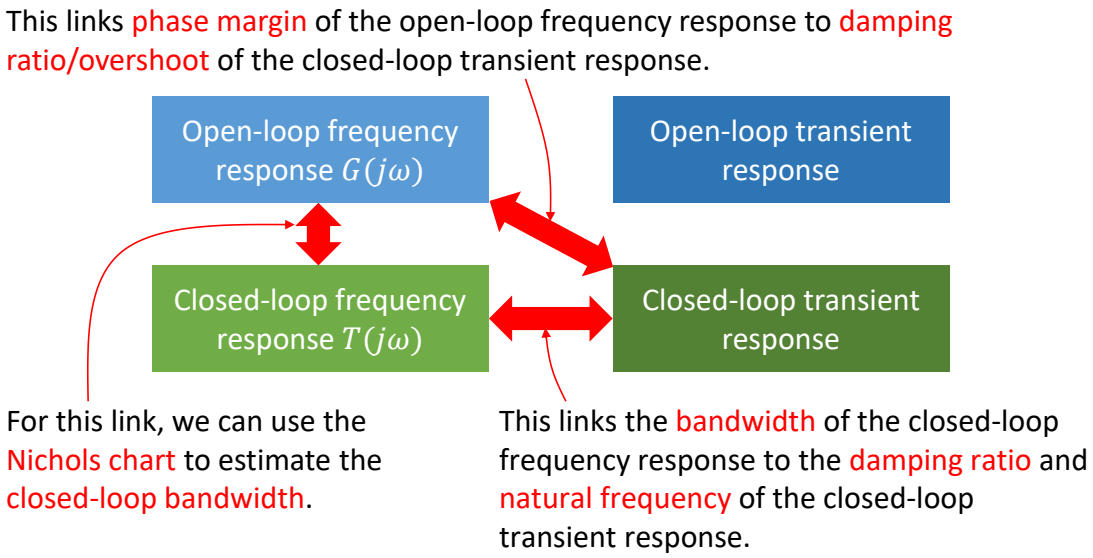


Figure 36: Summary of Sects. 6.1-6.3.

Acknowledgement

David Challender proofread the 2016 version of the manuscript and provided feedback. Kyle Hartzenberg recommended third-party MATLAB code for plotting complete Nyquist curves.

References

[Bis09] Christopher Bissell. A history of automatic control. In *Springer Handbook of Automation*, pages 53–69. Springer, 2009.

[BJ90] S. A. BILLINGS and J. C. PEYTON JONES. Mapping non-linear integro-differential equations into the frequency domain. *International Journal of Control*, 52(4):863–879, 1990.

[CL09] Donald R. Coughanowr and Steven E. LeBlanc. *Process Systems Analysis and Control*. McGraw-Hill, third edition, 2009.

[DB11] R.C. Dorf and R.H. Bishop. *Modern Control Systems*. Pearson, 12th edition, 2011.

[FPEN15] G.F. Franklin, J.D. Powell, and A. Emami-Naeini. *Feedback Control of Dynamic Systems*. Pearson Education Limited, 7th edition, 2015.

[Jan12] Klaus Janschek. *Mechatronic systems design: methods, models, concepts*. Springer-Verlag Berlin Heidelberg, 2012.

[LES15] M. Liu, G. K. Egan, and F. Santoso. Modeling, Autopilot Design, and Field Tuning of a UAV With Minimum Control Surfaces. *IEEE Transactions on Control Systems Technology*, 23(6):2353–2360, November 2015.

[Mac13] Uwe Mackenroth. *Robust control systems: theory and case studies*. Springer Science & Business Media, 2013.

[Man06] Ajit K Mandal. *Introduction to Control Engineering: Modeling, Analysis and Design*. New Age International, 2006.

[Nis15] Norman S. Nise. *Control Systems Engineering*. Wiley, 7th edition, 2015.

[Oga10] Katsuhiko Ogata. *Modern Control Engineering*. Prentice Hall, 5th edition, 2010.

[PS06] Saminathan Ponnusamy and Herb Silverman. *Complex variables with applications*. Birkhäuser Boston, 2006.

[SLE08] Fendy Santoso, M Liu, and Gregory Egan. $\mathcal{H}_2$ and $\mathcal{H}_\infty$ robust autopilot synthesis for longitudinal flight of a special unmanned aerial vehicle: a comparative study. *IET Control Theory & Applications*, 2(7):583–594, 2008.

[SMEDI10] Dale E Seborg, Duncan A Mellichamp, Thomas F Edgar, and Francis J Doyle III. *Process dynamics and control*. John Wiley & Sons, 3rd edition, 2010.

[TR06] Mark B. Tischler and Robert K. Remple. *Aircraft and rotorcraft system identification: engineering methods with flight-test examples*. AIAA education series. AIAA, 2006.

[VV07] Michael Verhaegen and Vincent Verdult. *Filtering and System Identification: A Least Squares Approach*. Cambridge University Press, 2007.