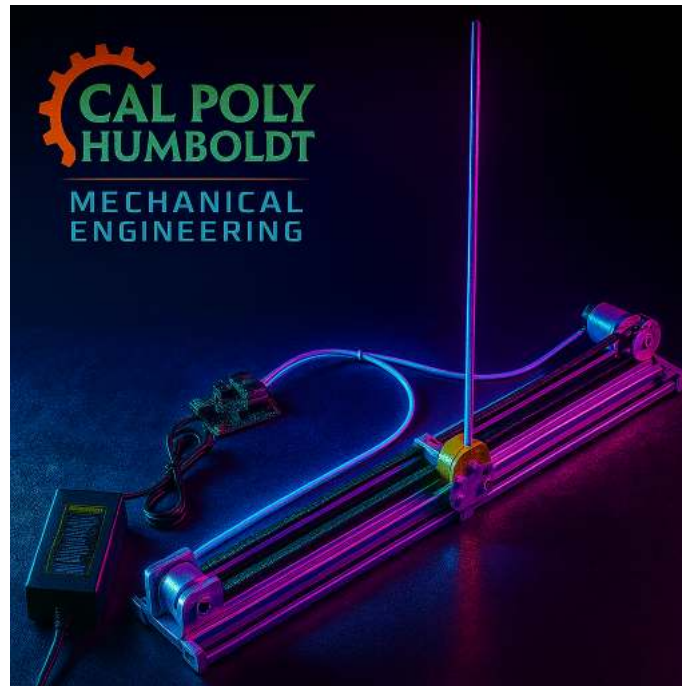


MEASUREMENTS, INSTRUMENTATION, AND CONTROLS



Joshua P. Steimel

California State Polytechnic University,
Humboldt



Measurements, Instrumentation, and Controls

Joshua P. Steimel

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (<https://LibreTexts.org>) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by [NICE CXone Expert](#) and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

This text was compiled on 05/15/2026

TABLE OF CONTENTS

[Licensing](#)

[Chapter 1: Measurement Processes and Confidence Intervals](#)

[Chapter 2: T-tests, P-values, and Hypothesis Tests](#)

[Chapter 3: Propagation of Error and ANOVA](#)

[Chapter 4: Fourier Analysis and Continuous Fourier Transforms](#)

[Chapter 5: Discrete Fourier Transforms, Nyquist Frequency, Signal Aliasing, and Spectral Leakage](#)

[Chapter 6: Strain Gauges and Wheastone Bridges](#)

[Chapter 7: Introduction to Feedback Control, Laplace Transforms, and Transfer Functions](#)

[Chapter 8: System Properties, Poles and Zeroes, System Response and Stability](#)

[Chapter 9: Time Domain Specifications](#)

[Chapter 10: System](#)

[Chapter 11: PID Control](#)

[Chapter 12: Root Locus](#)

[Chapter 13: Frequency Response, Nyquist Stability Criterion, and Bode Plots](#)

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

Licensing

A detailed breakdown of this resource's licensing can be found in [Back Matter/Detailed Licensing](#).

Chapter 1: Measurement Processes and Confidence Intervals

Lecture 1: Measurement Processes and Confidence Intervals

Topics Covered in this Class

Part I: Measurements, Confidence Intervals, T-Tests, P-Values, ANOVA, Fourier Analysis, Wheatstone Bridges

Part II: Laplace Transforms, Stability, Feedback Control, Root Locus, Routh-Hurwitz, Bode Plots, Control Design, PID Control, Nyquist Plots

What is a Measurement?

A **measurement** is a **quantitative** comparison between a predefined **standard** and a **measurand**.

What are some examples of different types of measurements we can make?

- Temperature
- Length
- Time
- Stress
- Strain
- Viscosity
- Force
- Torque
-

The process or act of measurement is a quantitative comparison between a predefined standard and a measurand.

Measurand: the particular physical parameter being observed and quantified, i.e., the input quantity of the measuring process

Standard: Must be of the same character as the measurand and usually recognized by the NIST (National Institute of Standards and Technology)

What are some standards?

- Tape Measure
- Balance
- Speed of Light
- ...

The act of measurement produces a result.

ADD FIGURE

The significance of measurements:

- Obtain quantitative information on the actual state of physical variables and processes (i.e. T, P, V, etc.)
- Vital for control processes which require measured discrepancy between actual and desired output (i.e. temperature control on sheet annealing line)
- Daily operations (i.e. central power station)
- Establishing costs

Measurements are only useful if they are reliable. This brings up the concept of the accuracy or uncertainty of a measurement. No measurement is perfect and therefore engineers must place tolerances on measured values.

There are two basic methods of measurement: i) direct and ii) indirect

Direct Comparison:

- Measurements utilize a primary or secondary standard
- Ex: Measuring length using tape measure, which is a secondary standard, primary standard would use speed of light
- Simple measurement however often we need greater accuracy
- Direct comparison is less common than indirect comparison

Indirect Comparison:

- Utilizes some form of transducing device coupled to a chain of connecting apparatus as the measuring system
- Input or measurand converted into an analogous form which is a necessary conversion so the information is intelligible
- Ex. Human senses cannot detect strain so conversion is necessary
- The analogous signal is processed and can be amplified, filtered, or remotely recorded
- Transducers convert mechanical input into analogous electrical for processing, particularly for dynamic mechanical measurements

Generalized Measuring System

Stage I: Detection-transduction or sensor-transducer stage

Stage II: Intermediate or signal-conditioning stage

Stage III: Terminating or readout-recording stage

ADD FIGURE

Examples:

- Stage I: spring-mass, thermocouple, piezoelectric
- Stage II: gearing, dashpots, bridges, filters
- Stage III: liquid column, digital printout, magnetic recording

Stage 1: Sensor-Transducer Stage

- Primary function is to detect or sense the measurand
- Ideally should be insensitive to all other inputs, ex. insensitive to temperature for strain gauge
- Unwanted sensitivity can lead to error termed noise when it varies rapidly and drift when it varies slowly

Stage 2: Signal-Conditioning Stage

- Primary purpose is to modify the transduced information so it is acceptable to Stage 3
- The transduced signal is conditioned via amplification, filtering, integration, differentiation, etc.

Stage 3: Readout-Recording Stage

- Provides information in a form comprehensible to one of the human senses
- Example: Relative displacement or a digital readout

Ex. 1: Tire Pressure Gauge

ADD FIGURE

Task: Fill out the generalized block diagram for using a tire pressure gauge to measure tire pressure

What is the Measurand/Input?

- Pressure

What is the Sensor-Transducer?

- The piston-cylinder senses pressure and converts it to force. A secondary transducer (spring) converts the force to displacement.

What is the Signal Conditioner?

- None in this case.

What is the Readout-Recording?

- The displacement of the stem, which can be read from the calibrated scale.

Using Correct Terminology When Assessing Measurement Quality/Apparatus Performance

- **Accuracy:** Difference between measured and true values. Manufacturer will specify maximum error but not confidence intervals typically.
- **Precision:** Difference between reported values during repeated measurements of same quantity.
- **Resolution:** Smallest increment of change in the measured value that can be determined by the readout scale. Often the same as precision.
- **Sensitivity:** The change of an instrument's output per unit change in the measured quantity. Higher sensitivity often indicates finer resolution, better precision, and higher accuracy.

Error and Uncertainty:

Error is the difference between the *measured* and *true* value of the quantity being measured. Unfortunately, the true value can never really be known so instead we quantify the **uncertainty** of a measurement at specific levels of confidence. Error can be categorized into two basic categories either **bias** or **precision** error/uncertainty. We will cover uncertainty and error analysis much more in-depth in the next several lectures...but never use percent error in this course.

Error

We can begin with our discussion of error by defining the following equation

$$Error = x_m - x_{true} \quad (\text{Chapter 1.1})$$

where x_m is the measured value and x_{true} is the true value.

When designing an experimental system or executing an experiment the experimentalist's primary objective should be to minimize the error. After the experiment is completed we need to estimate a bound on the error with a degree of certainty or a confidence interval (i.e. 90%, 95%, 99%), hopefully you have encountered confidence intervals before but if not no worries we will discuss this in depth.

There are many different sources of inaccuracy or error however most errors can be classified as:

- (1) **Precision**
- (2) **Bias**
- (3) **Illegitimate**
- (4) **Errors that are sometimes bias or sometimes precision errors**

In this course we will spend a considerable amount of time discussing **1. Precision** and **2. Bias** error in great detail. But we will discuss all these sources of error and how to account for them when designing experiments or presenting your results. Let's start with **Precision Error**.

Precision or Random Errors

I really do not like the term random error but you will occasionally see that term in literature but in this course we will use the term precision primarily.

Precision errors are **different for each successive measurement**. Thus if you create a histogram, which shows the probability of encountering or measuring a specific measurand, the width of your distribution (hopefully it is Gaussian) will be due to precision error if that is the only source of error in your experiment. From this distribution, **statistical analysis can be utilized to estimate the size of the precision error**. Typically bias and precision errors occur simultaneously and contributes to the **total error**, more on this fun calculation later, hope you remember Taylor expansions....

Why Do We Have Precision Error?

- (1) Disturbances to the equipment
- (2) Fluctuating experimental conditions
- (3) Insufficient measuring-system sensitivity
- (4) Stochastic (Random) nature of measurements even in static conditions

Examples of Precision Error:

- Measuring temperature with thermocouple as temperature fluctuates
- Measure the length of a beam with a micrometer you will get slightly different measurements than classmates
- Changes in temperature conditions can affect swimming of bacteria

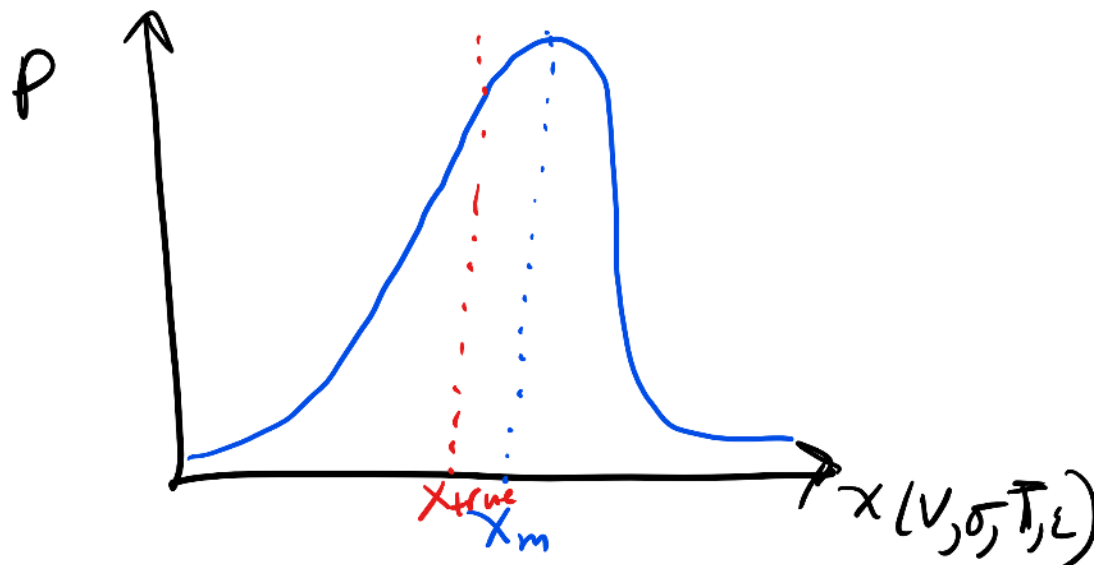


Figure Chapter1.1: Precision Error Distribution.

Bias or Systematic Errors

I like the term bias and systematic much better than random. However, in this course we will primarily use the term Bias error.

Bias errors will occur the same way each time a measurement is made. For example a balance that consistently reads 0.5 mg heavier. In some literature you might see this termed as a fixed offset error. The key thing to note about bias errors is they are fixed and thus **they do not show a distribution**. This is in stark contrast to precision errors which show a distribution, typically Gaussian. Therefore if there is no distribution we cannot apply any **statistical analysis**.

There are multiple types of bias errors:

(1) Calibration errors

- This is the most common error
- Typically calibration errors are **zero-offset errors** which cause all reading to be offset by constant amount or *scale errors*
- I.e. scale constantly light by 1lb or clock runs fast by 1 minute
- Can be reduced via calibration procedures and proving the system via comparison with *known* standard

(2) Consistently recurring human errors

- Ex. experimenter who consistently jumps the gun when recording synchronized readings either time, voltage, etc.

(3) Errors caused by defective equipment

(4) Loading errors

- i.e. not putting a DSC pan at the center of the DSC will skew temperature measurements

(5) System resolution limits

- i.e. Bias error for a ruler with cm tick marks will be cm, compared to a micrometer with $1\mu\text{m}$ resolution, that is 4 order of magnitude difference.

Illegitimate Error

Illegitimate error can and most likely will occur to you at some point as you perform experiments as an engineer. Illegitimate errors include

- Blunders and mistakes
- Computational or calculation errors
- Analysis errors

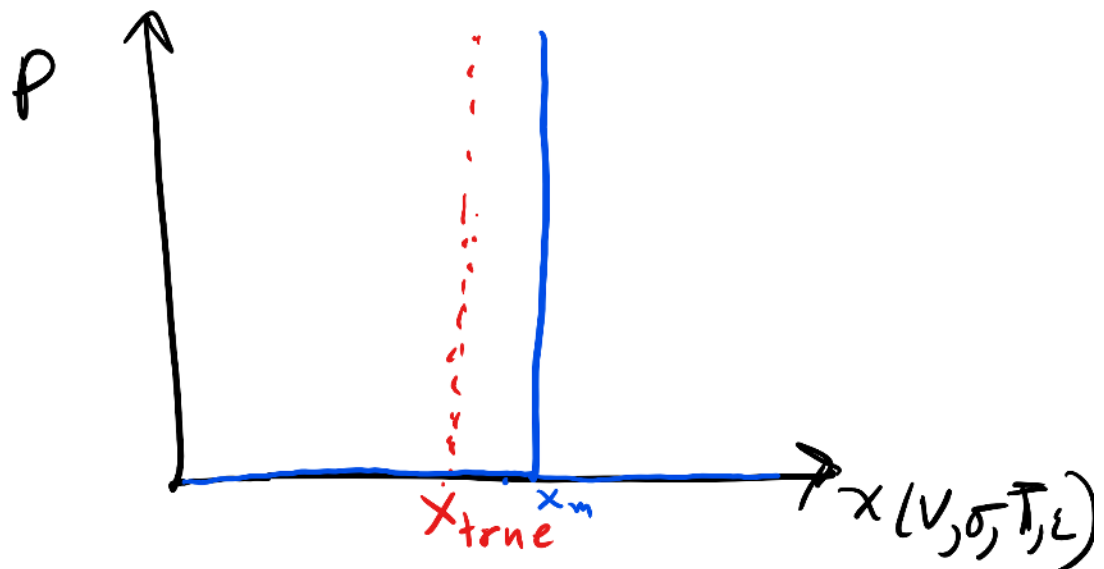


Figure Chapter1.2: Bias Error Distribution

While these things can occur in lab **you must always perform the experiment again**. You cannot simply present your results in either a presentation or a form of technical writing and attribute error to illegitimate error. You must perform the experiment again.

Errors That Are Sometimes Bias and Sometimes Precision Error:

These are the most difficult sources of error that we must analyze as it is difficult to determine what type of error analysis must be done and which error has a distribution and what does not.

- **Instrument Hysteresis**

- Hysteresis error can be seen in stress strain measurements and is path dependent

- **Calibration drift and variation in test or environmental conditions**

- Drift occurs when the response varies over time often due to sensitivity to temperature or humidity

- **Variations in procedure or definition among experimenters**

Quantifying Total Uncertainty, Comprehensive Error Analysis

Again typically we are concerned with **bias** and **precision** error when evaluating experimental data. And this data comes from two classes of experiments: i) *single- sample* and ii) *repeated-sample* experiments.

A **sample** refers to an individual measurement of a specific quantity. Ex. Measuring strain in a material several times under identical loading is a repeated-sample experiment. A single measurement of strain is a single-sample experiment and we cannot extract the distribution of precision error in this case. In this course and in your career as an engineer you will most likely be taking multiple measurements,

i.e. repeated sample experiments. Single sample experiments are more uncommon and typically occur when the measurement you are making is extremely difficult and multiple experiments are non realistic to conduct, typically due to economic or practicality reasons.

The total uncertainty U_x in a measurement of x is

$$U_x = \sqrt{B_x^2 + P_x^2} \quad (\text{Chapter 1.2})$$

where B_x is the bias uncertainty in a measurement of x and P_x is the precision uncertainty. Note that x is just a variable and we can talk about uncertainty in measuring volume (V), pressure (P), stress (σ), etc.

This is a largely empirically derived formula however it does depend on the **assumption that the bias and precision uncertainty are independent sources of error**. Also the bias and precision uncertainty should have the **same confidence level** when calculating the **total uncertainty**. We will come back to this a bit later in excruciating detail for this calculation so please be patient.

Minimizing Error in Designing Experiments

Best time to **minimize experimental error** is in the **design stage**. You should **perform a single-sample uncertainty analysis** of the proposed experimental arrangement **prior to building the apparatus**. This allows one to make a decision on whether the expected uncertainty is acceptable and to identify the sources of uncertainty. So when designing an experiment:

- **Avoid approaches** that require two large numbers to be measured to determine the small difference between them
- **Design experiments or sensors that amplify the signal strength** to improve sensitivity
- Build *null designs* in which the output is measured as a change from zero rather than as a change in nonzero value
- Avoid experiments with large *correction factors* must be applied as part of the data-reduction procedure
- Attempt to **minimize the influence of the measuring system on the measured variable**
- **Calibrate** entire system to minimize calibration related bias errors.

Gaussian and Normal Distribution

One of the most common distributions encountered is the Gaussian distribution, also called the normal distribution or bell curve.

The Gaussian distribution describes an infinite population of data where each datum represents a measurement of a single quantity. Each value x differs due to precision error.

The probability of obtaining a specific value x is described by the probability density function (PDF), $f(x)$.

The total area under the curve equals 1, which represents 100% probability:

$$P = \int_{-\infty}^{\infty} f(x) dx = 1 \quad (\text{Chapter 1.3})$$

Gaussian PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (\text{Chapter 1.4})$$

Where μ = population mean and σ = population standard deviation.

To simplify, define

$$z = \frac{x - \mu}{\sigma} \quad (\text{Chapter 1.5})$$

, so:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (\text{Chapter 1.6})$$

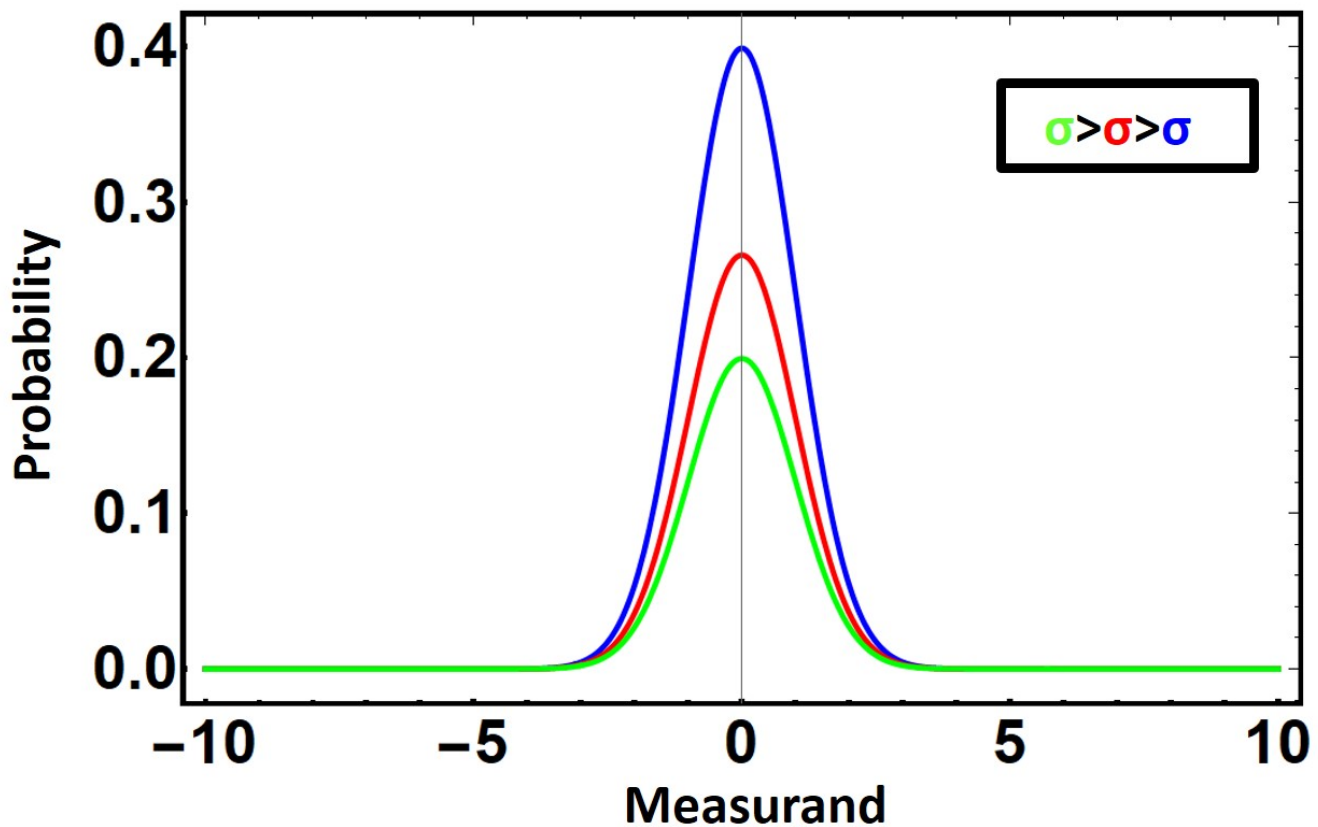


Figure Chapter1.3: Gaussian Distribution

You can use Z-tables to find the area under the curve for a given z . For example, the 95% confidence interval contains most values within $\pm 1.96\sigma$.

Outliers

Outliers are data points that are much higher or lower than the mean.

You can disregard a data point if it exceeds 3σ , but must report it and justify its exclusion.

Confidence Intervals

The area between $\pm z$ under the standard curve defines a confidence interval.

Example: A 95% confidence interval covers ~95% of the population's measurements.

Example Problems - Confidence Intervals

Working with Z Tables

- What is the area under the standard normal distribution curve between $z = -1.43$ and $z = 1.43$?
- What percent of the population lies within this range?
- What range of z contains 90% of the data?

Running and Tumbling *E. coli*

Given: mean = $3 \mu\text{m/s}$, $\sigma = 0.1 \mu\text{m/s}$, range = $[2.857, 3.143] \mu\text{m/s}$

- Assume normal distribution
- Compute z for bounds
- Use standard normal table to compute area
- Estimate number of samples within range
- Determine 95% confidence bounds

Marble Diameters

Given: $N = 765$ marbles, $\mu = 200 \text{ mm}$, $\sigma = 20 \text{ mm}$

- Determine diameter range for 60% of marbles

Ultimate Tensile Strength Measurements

Given: $\mu = 303 \text{ MPa}$, $\sigma = 33 \text{ MPa}$

- What is the probability a sample exceeds 350 MPa?

Voltage Measurements

Given: $\mu = 10 \text{ V}$, $\sigma = 3.4 \text{ V}$

- How many of 150 measurements fall between 10–15 V?

Under Pressure Measurements

Given: $\mu = 39 \text{ Pa}$, $\sigma = 4 \text{ Pa}$, $N = 120$

- How many measurements fall between 35–45 Pa?
- How many below 35 Pa?
- How many below 45 Pa?

This page titled [Chapter 1: Measurement Processes and Confidence Intervals](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 10: System

System Stability and Pole-Based Classification

System stability is directly tied to the locations of the poles of the transfer function $H(s) = \frac{Y(s)}{U(s)}$. The classification is as follows:

Stable: all poles lie in the open left-half of the complex plane (LHP).

Marginally stable: poles on the imaginary axis, but all are simple (non-repeated).

Unstable: at least one pole in the right-half plane (RHP), or repeated poles on the imaginary axis.

Worked Example: First-Order Impulse Response

Given:

$$H(s) = \frac{1}{s + \sigma}, \quad \text{Impulse response: } y(t) = e^{-\sigma t} \quad (\text{Chapter 10.1})$$

If $\sigma > 0$, $y(t) \rightarrow 0$: stable

If $\sigma = 0$, $y(t) = 1$: marginally stable

If $\sigma < 0$, $y(t) \rightarrow \infty$: unstable

The most general definition of stability is that the plant or transfer function is stable if and only if the output function $y(t)$ goes to zero as time goes to infinity. We are marginally stable if and only if the absolute value of $y(t)$ is less than some constant value but also never decays to zero and the system is unstable if and only if $y(t)$ goes to infinity as time goes to infinity.

Pole Stability Conditions

Let's consider a plant function $P(s)$ with poles p_i . Then we can summarize that:

$P(s)$ is stable if and only if all p_i are in the LHP.

$P(s)$ is marginally stable if and only if:

- All p_i are in the LHP or on the imaginary axis, and
- Some non-repeated p_i are on the imaginary axis.

$P(s)$ is unstable if and only if:

- Some p_i are in the RHP, or
- Some p_i are on the imaginary axis and are repeated.

Let's look at some examples to test if they are stable:

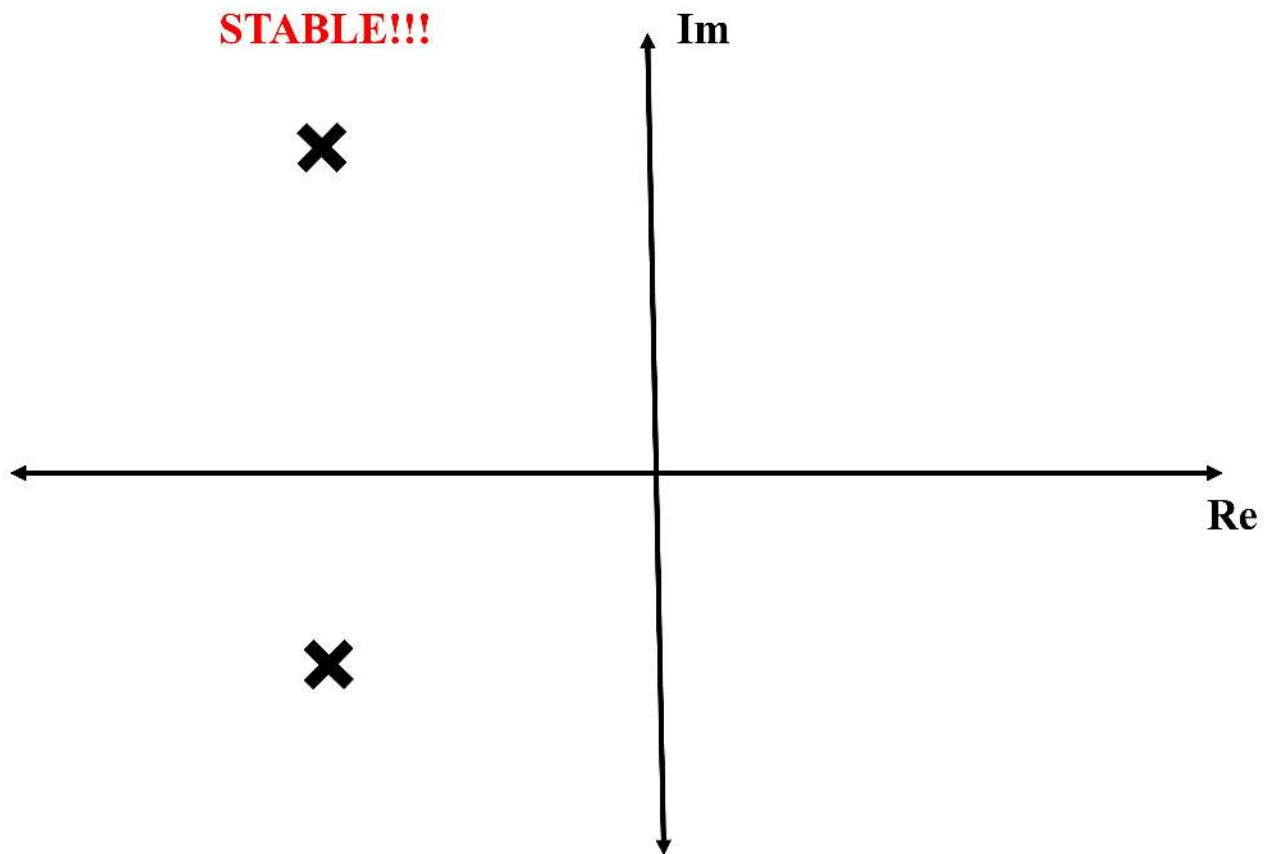


Figure Chapter10.1: Stable, Marginally Stable, or Unstable

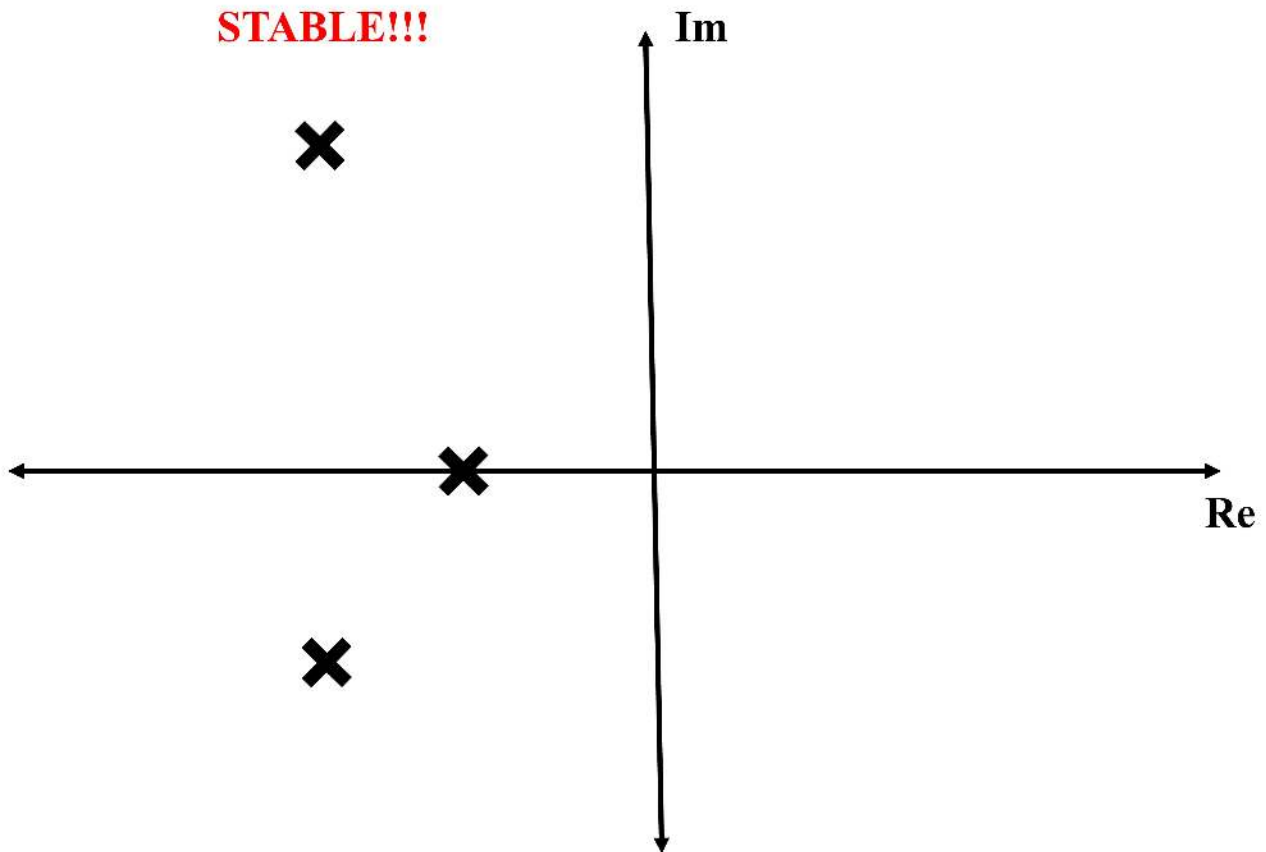


Figure Chapter10.2: Stable, Marginally Stable, or Unstable

MARGINALLY STABLE!!!

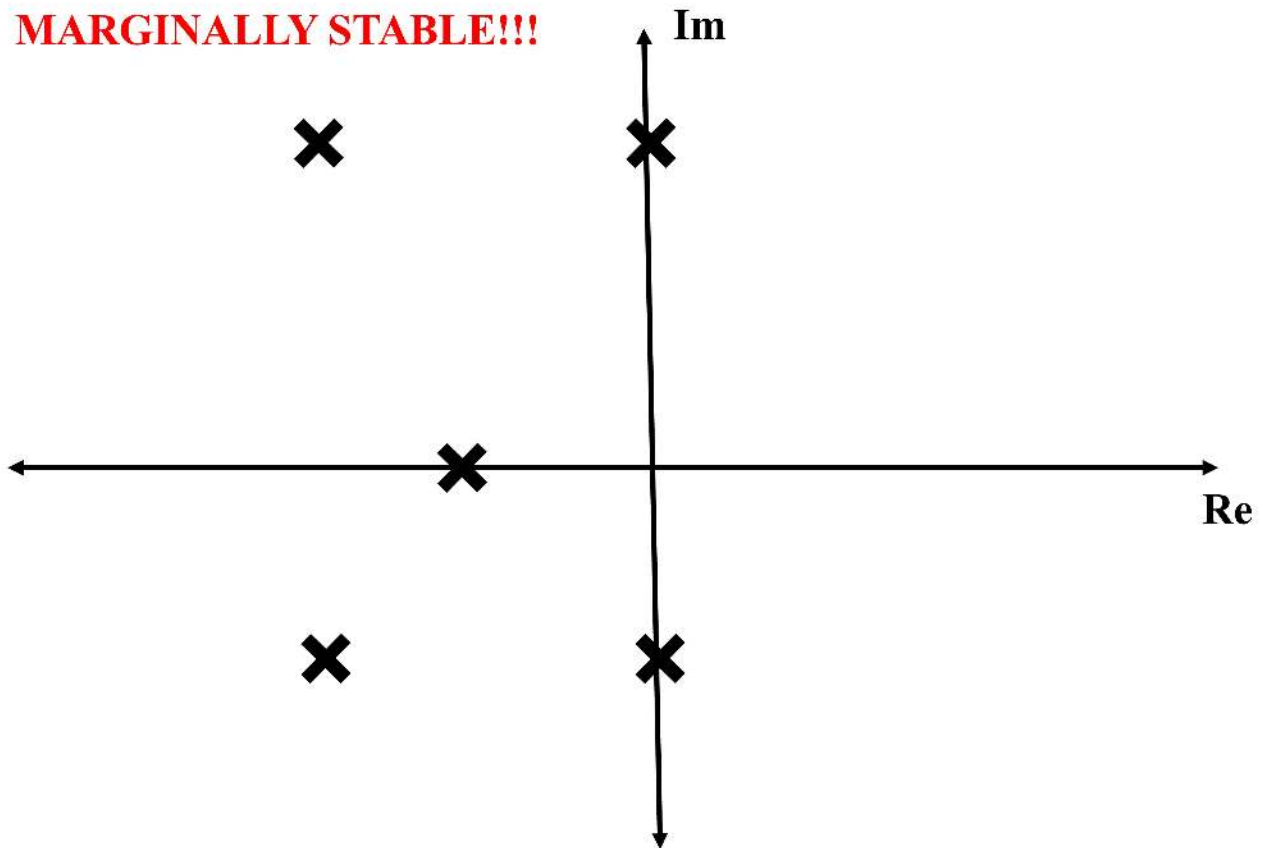


Figure Chapter10.3: Stable, Marginally Stable, or Unstable

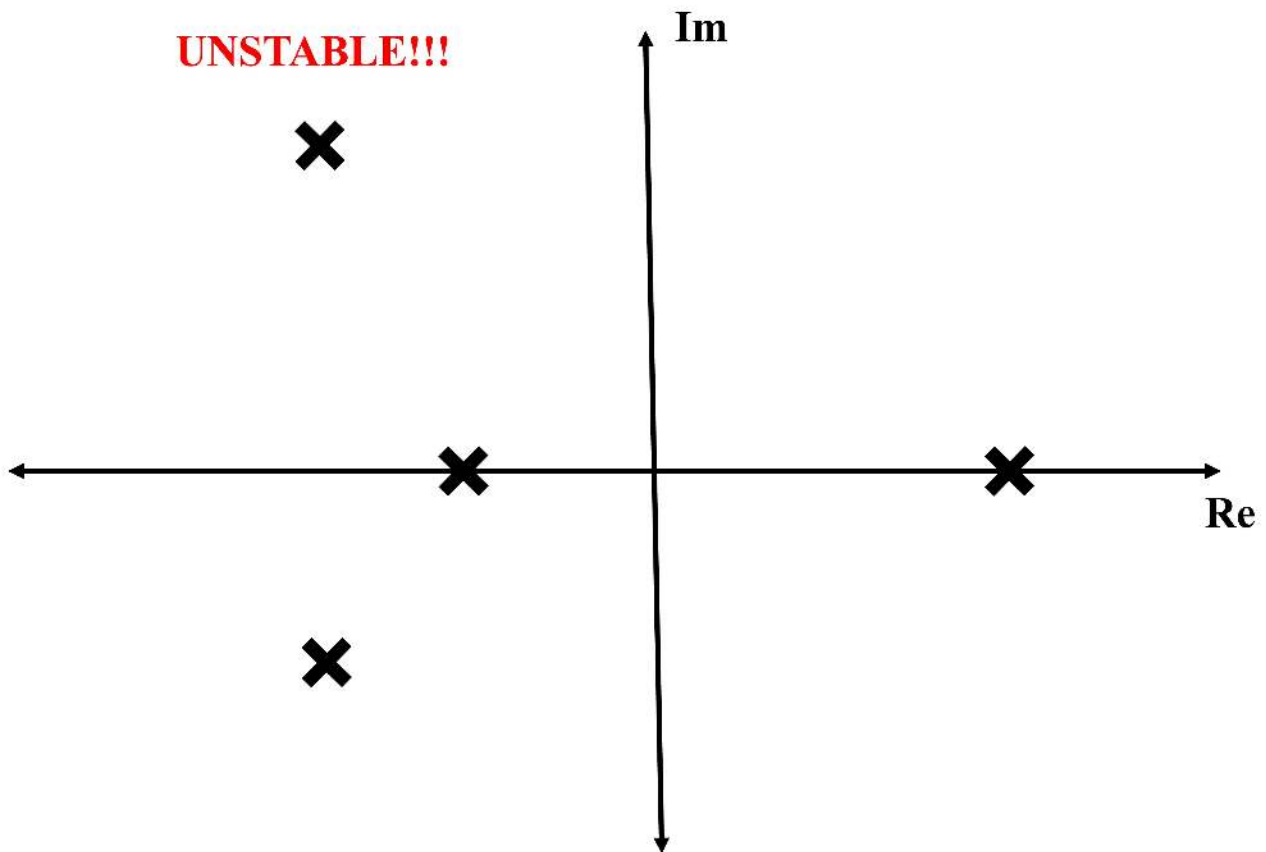


Figure Chapter10.4: Stable, Marginally Stable, or Unstable

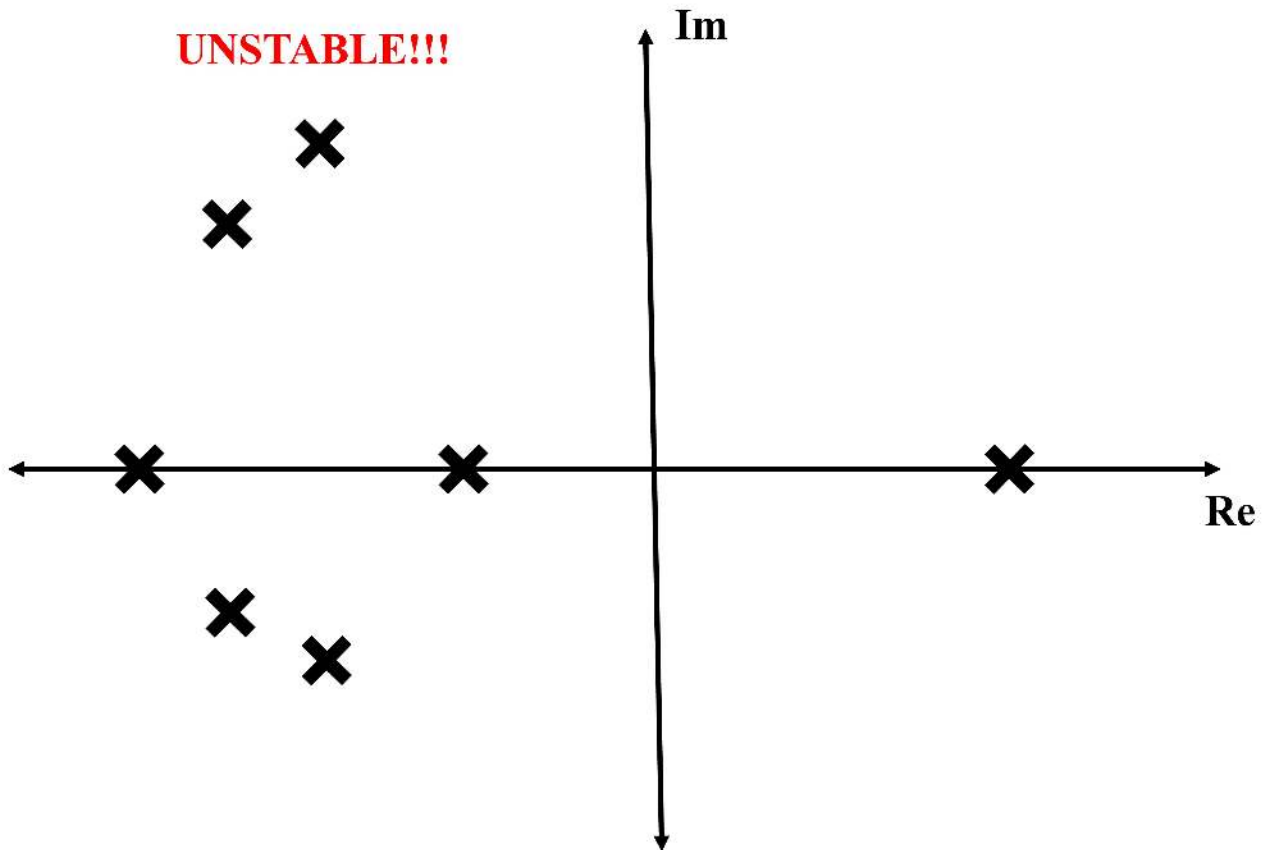


Figure Chapter10.5: Stable, Marginally Stable, or Unstable

We can also look at some transfer functions and determine if they are stable:

$$\frac{1}{s+4}, \quad \frac{s-3}{(s+5)^2+2^2}, \quad \frac{1}{s-5} \quad (\text{Chapter 10.2})$$

The first two transfer functions are stable. The last is unstable.

Stability for First Order Transfer Functions

$$\frac{1}{s+a} \quad (\text{Chapter 10.3})$$

We are stable as long as $a > 0$.

Stability for Second Order Transfer Functions

$$\frac{1}{s^2 + a_1 s + a_2} \quad (\text{Chapter 10.4})$$

We are stable as long as $a_1 > 0$ and $a_2 > 0$.

Stability for nth Order Transfer Functions

$$\frac{1}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \dots + a_n} \quad (\text{Chapter 10.5})$$

We are stable as long as all $a_i > 0$. However, this condition is not comprehensive enough to guarantee stability. To be rigorous for nth-order transfer functions, we must use the Routh-Hurwitz Stability Criterion.

The Routh-Hurwitz Stability Criterion

This technique provides a systematic method to determine the number of poles with positive real parts (i.e., unstable poles) without explicitly calculating roots. The method involves:

1. Writing the characteristic polynomial:

$$P(s) = a_n s^n + a_{n-1} s^{n-1} + \dots + a_0 \quad (\text{Chapter 10.6})$$

2. Constructing the Routh table using these coefficients.

3. Checking signs in the first column.

Stability condition: All entries in the first column must be strictly positive.

Interpretation:

The number of sign changes in the first column equals the number of RHP poles.

A zero in the first column requires special handling (epsilon substitution or auxiliary polynomial).

Assume $a_0 > 0$ for normalization. The Routh array is built row by row as follows:

Step 1: First Two Rows

Row 1 (even powers): $a_0, a_2, a_4, \dots$

Row 2 (odd powers): $a_1, a_3, a_5, \dots$

Step 2: Compute Subsequent Rows

Use the following formula for all subsequent elements:

$$b_1 = \frac{\begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix}}{a_1} = \frac{a_1 a_2 - a_0 a_3}{a_1} \quad (\text{Chapter 10.7})$$

$$b_2 = \frac{\begin{vmatrix} a_1 & a_5 \\ a_0 & a_4 \end{vmatrix}}{a_1} = \frac{a_1 a_4 - a_0 a_5}{a_1} \quad (\text{Chapter 10.8})$$

Rule: If any row begins with zero, replace it with a small ε , or use an auxiliary polynomial if the entire row is zero.

Generic Routh Table (up to 4th order)

$$P(s) = a_0 s^4 + a_1 s^3 + a_2 s^2 + a_3 s + a_4 \quad (\text{Chapter 10.9})$$

$$\begin{array}{c|ccc} s^4 & a_0 & a_2 & a_4 \\ s^3 & a_1 & a_3 & 0 \\ s^2 & b_1 & b_2 & 0 \\ s^1 & c_1 & 0 & 0 \\ s^0 & d_1 & 0 & 0 \end{array} \quad (\text{Chapter 10.10})$$

Where:

$$b_1 = \frac{a_1 a_2 - a_0 a_3}{a_1}, \quad b_2 = \frac{a_1 a_4 - a_0 \cdot 0}{a_1} = a_4 \quad (\text{Chapter 10.11})$$

$$c_1 = \frac{b_1 a_3 - a_1 b_2}{b_1}, \quad d_1 = a_4 \quad (\text{Chapter 10.12})$$

Stability Criterion:

The system is stable if and only if all entries in the first column are strictly positive (no sign changes).

Worked Example

Consider the characteristic polynomial:

$$P(s) = s^4 + 3s^3 + 3s^2 + 2s + 1 \quad (\text{Chapter 10.13})$$

We build the Routh table:

s^4	1	3	1
s^3	3	2	0
s^2	$\frac{3 \cdot 3 - 1 \cdot 2}{3} = \frac{7}{3}$	$\frac{3 \cdot 1 - 1 \cdot 0}{3} = 1$	0
s^1	$\frac{7 \cdot \frac{2}{3} - \frac{7}{3} \cdot 1}{\frac{7}{3}} = \frac{14/3 - 7/3}{7/3} = \frac{5/3}{7/3} = \frac{5}{7}$	0	0
s^0	1	0	0

(Chapter 10.14)

First column:

$1, 3, \frac{7}{3}, \frac{5}{7}, 1$

All entries are positive. Therefore, the system is stable.

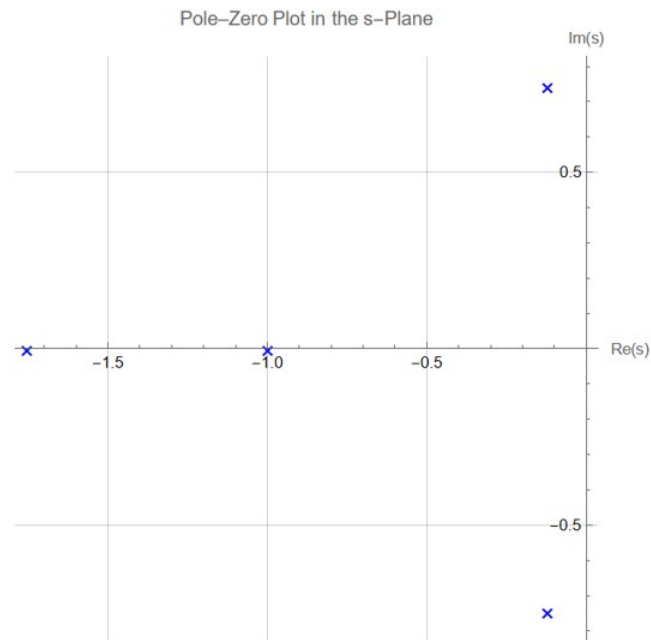


Figure Chapter10.6: Pole and Zero Plot

We can also confirm that this system is stable by plotting the poles as well. The figure confirms that all poles lie in the left half-plane.

This is our first tool to use for control design!

Controller Influence on Closed-Loop Poles and Stability

When designing a feedback control system, the dynamic behavior of the system is ultimately governed by the closed-loop transfer function, not the open-loop plant alone. This closed-loop behavior depends critically on the controller $C(s)$, which appears directly in the closed-loop transfer function.

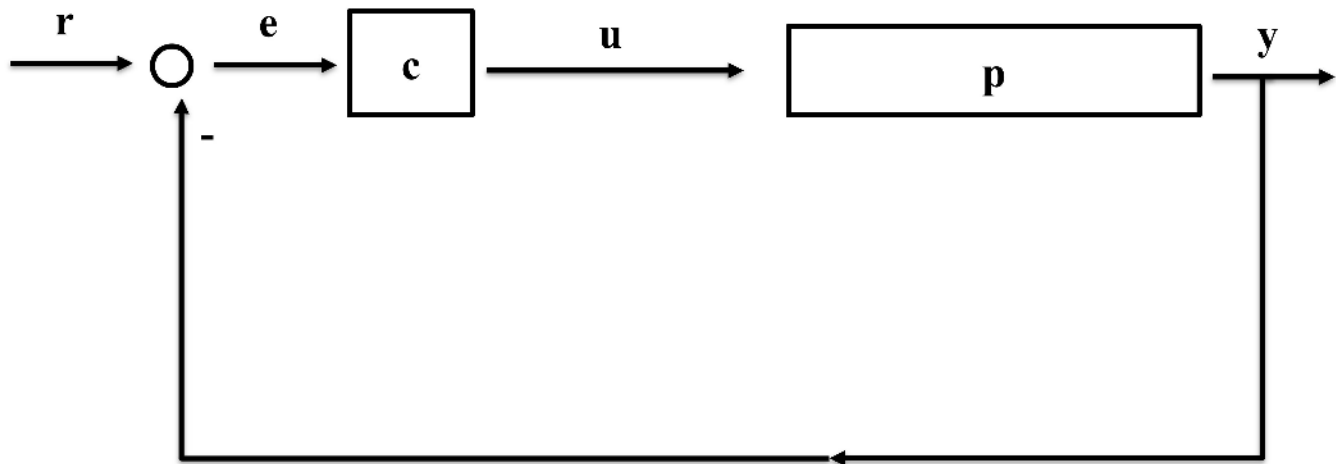


Figure Chapter10.7: Unity Feedback with Controller

Given a plant $P(s)$ and a controller $C(s)$, the closed-loop transfer function with unity feedback can be calculated as follows:

$$G(s) = \frac{Y(s)}{R(s)} \quad (\text{Chapter 10.15})$$

where we know that in this system we can define the following relationships:

$$C(s) = \frac{U(s)}{E(s)}, \quad E(s) = R(s) - Y(s), \quad P(s) = \frac{Y(s)}{U(s)} \quad (\text{Chapter 10.16})$$

Now we can rearrange:

$$U(s) = C(s)[R(s) - Y(s)] \Rightarrow Y(s) = P(s)C(s)[R(s) - Y(s)] \quad (\text{Chapter 10.17})$$

Solve for $Y(s)$:

$$\begin{aligned} Y(s) + P(s)C(s)Y(s) &= P(s)C(s)R(s) \\ \Rightarrow Y(s)[1 + P(s)C(s)] &= P(s)C(s)R(s) \end{aligned} \quad (\text{Chapter 10.18})$$

Closed-loop transfer function:

$$G(s) = \frac{Y(s)}{R(s)} = \frac{C(s)P(s)}{1 + C(s)P(s)} \quad (\text{Chapter 10.19})$$

The poles of this system are the roots of the denominator:

$$1 + C(s)P(s) = 0 \quad (\text{Chapter 10.20})$$

This is known as the characteristic equation of the closed-loop system. Its roots determine the stability and transient performance of the system. Importantly, these poles are directly influenced by the choice of controller $C(s)$. This means we can shift the poles—and therefore the behavior—of the system by changing $C(s)$.

Using Feedback to Stabilize an Unstable System

Suppose the open-loop system (plant) is unstable; that is, it has one or more poles in the right half of the s -plane. Using appropriate feedback, we can relocate these poles to the left half-plane (LHP), thereby stabilizing the system. This is one of the most powerful features of feedback control.

To illustrate this idea, consider a simple case where the controller is a constant gain:

$$C(s) = K \quad (\text{Chapter 10.21})$$

Then the closed-loop transfer function becomes:

$$G_{cl}(s) = \frac{KP(s)}{1 + KP(s)} \quad (\text{Chapter 10.22})$$

Let's assume the plant is a first-order unstable system:

$$P(s) = \frac{1}{s - a}, \quad \text{with } a > 0 \quad (\text{Chapter 10.23})$$

This system has an open-loop pole at $s = a$, which lies in the right half-plane, so it is clearly unstable.

The closed-loop transfer function becomes:

$$G_{cl}(s) = \frac{K}{s - a + K} \quad (\text{Chapter 10.24})$$

The pole of the closed-loop system is at:

$$s = a - K \quad (\text{Chapter 10.25})$$

Now observe that:

- If $K < a$, the pole remains in the right half-plane: the system is still unstable.
- If $K = a$, the pole is at the origin: the system is marginally stable.
- If $K > a$, the pole moves to the left half-plane: the system becomes stable.

This demonstrates that we can use the feedback gain K to shift the pole location and change an unstable system into a stable one.

Worked Example: Stability via Routh-Hurwitz

We are given a closed-loop transfer function of the form:

$$G_{cl}(s) = \frac{(s+1)k}{s^3 + 5s^2 + (k-6)s + k} \quad (\text{Chapter 10.26})$$

The denominator of the transfer function is the characteristic polynomial:

$$P(s) = s^3 + 5s^2 + (k-6)s + k \quad (\text{Chapter 10.27})$$

We will determine the range of values of k that ensure the system is stable. This means all roots of $P(s)$ must lie in the left-half of the s -plane, which is guaranteed if all the entries in the first column of the Routh-Hurwitz table are positive.

Step 1: Construct the Routh Table

$$\begin{array}{c|cc} s^3 & 1 & k-6 \\ s^2 & 5 & k \\ s^1 & \frac{5(k-6)-1 \cdot k}{5} = \frac{5k-30-k}{5} = \frac{4k-30}{5} & 0 \\ s^0 & k & 0 \end{array} \quad (\text{Chapter 10.28})$$

Step 2: Routh-Hurwitz Stability Conditions

For stability, all entries in the first column must be positive:

- $1 > 0$ — always true
- $5 > 0$ — always true
- $\frac{4k-30}{5} > 0 \Rightarrow 4k - 30 > 0 \Rightarrow \boxed{k > 7.5}$
- $k > 0$

Thus, the system is stable when:

$$\boxed{k > 7.5} \quad (\text{Chapter 10.29})$$

This guarantees all the first-column entries of the Routh table are strictly positive, which implies the system is stable.

```
PoleZeroPlotSimple[TransferFunctionModel[Gs /. k -> 5, s]]
```

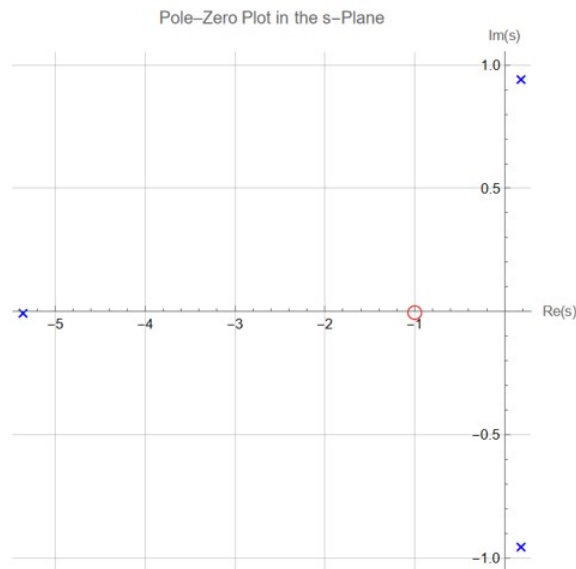


Figure Chapter10.8: Unstable

```
PoleZeroPlotSimple[TransferFunctionModel[Gs /. k -> 10, s]]
```

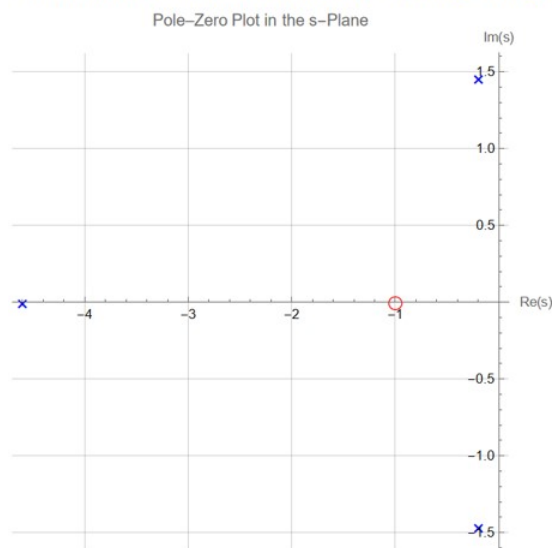


Figure Chapter10.9: Stable

General Implications for Control Design

In practice, the controller $C(s)$ may be more complex than a simple gain—it could include proportional-integral-derivative (PID) terms, lead/lag compensation, or state feedback. Regardless, the controller enters the closed-loop characteristic equation through the term $1 + C(s)P(s)$, and thus:

- All control objectives—such as stability, settling time, overshoot, and steady-state error—depend on the roots of this equation.
- Controller design is fundamentally about shaping the closed-loop pole locations to achieve desired performance.
- Even when the plant has undesirable dynamics (e.g., slow response, high overshoot, or instability), feedback can dramatically alter system behavior.

Caution: Feedback Can Also Destabilize a Stable System

While feedback can be used to stabilize an unstable system, the reverse is also true. Poorly designed feedback can move poles into the right half-plane and make a stable system unstable. For example, selecting a gain K that is too large can push a lightly damped complex-conjugate pair of poles into the unstable region, resulting in oscillations or runaway growth.

Thus, we must always verify closed-loop stability using:

- Analytical pole computation
- Routh-Hurwitz stability test
- Root locus or Nyquist criteria

Feedback Structure and Controller Design

In a unity feedback control system, the goal is for the output $y(t)$ to track the reference $r(t)$ in the presence of:

- Disturbance $w(t)$: an external signal acting on the plant
- Sensor noise $v(t)$: affecting the measured output

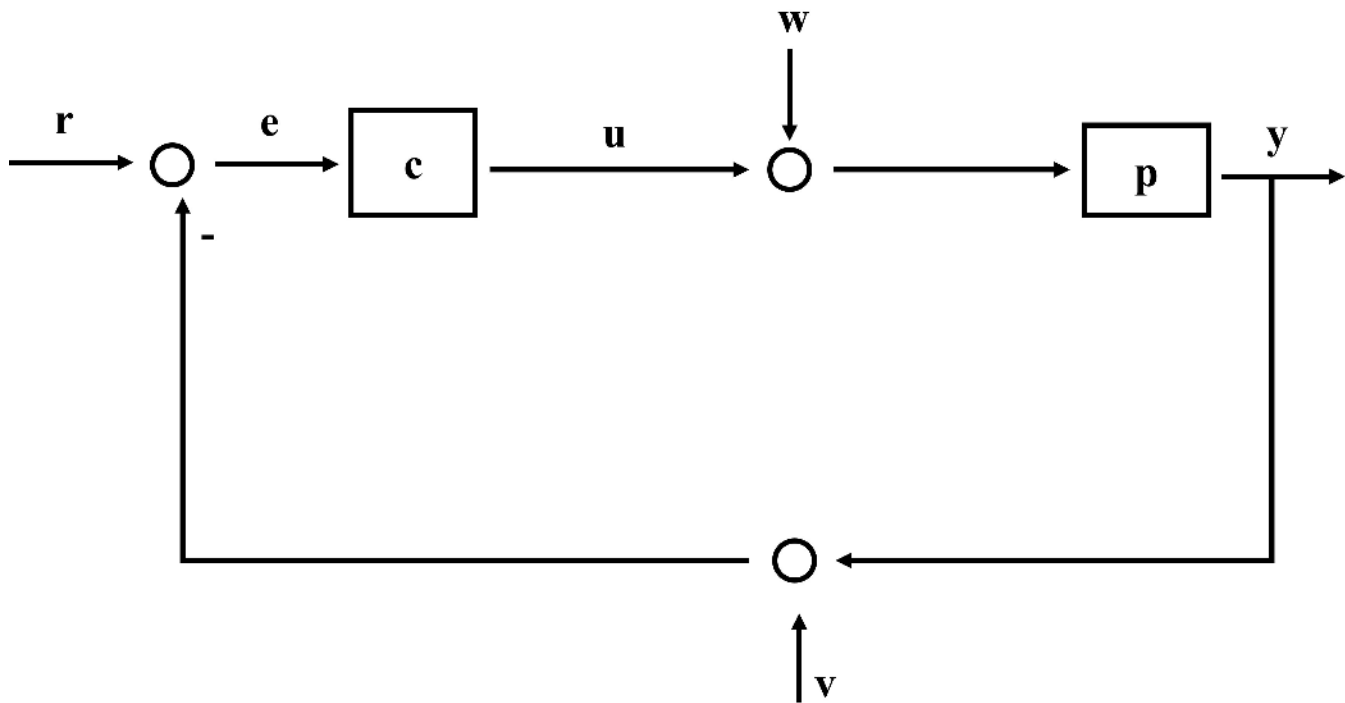


Figure Chapter10.10: Closed Loop Unity Feedback Loop with Disturbance and Noise

The key relationships:

$$e(t) = r(t) - y(t), \quad u(t) = C(s)e(t)$$

Control Design Objectives

1. **Stability:** All internal signals must remain bounded for all bounded inputs.
2. **Tracking:** Output $y(t)$ should follow reference $r(t)$ as $t \rightarrow \infty$, assuming $w(t) = v(t) = 0$.
3. **Disturbance Rejection:** Output should stay close to $r(t)$ even if $w(t) \neq 0$.
4. **Control Effort Minimization:** Minimize $u(t)$ to prevent actuator saturation or damage.

Controller Gain Trade-offs

Let:

$$C(s) = K, \quad P(s) = \frac{1}{s+a}$$

Then:

$$T(s) = \frac{K}{s+a+K} \Rightarrow \text{Pole shifts left with increasing } K$$

Benefits of large K :

- Reduced steady-state error

- Faster response (shorter t_s)
- Improved tracking

Risks of large K :

- Excessive control effort
- Potential actuator limitations
- Risk of destabilizing the system if the plant has lightly damped or unstable modes

Block Diagram Algebra for Closed-Loop Systems

For unity feedback, the closed-loop transfer function from reference $r(t)$ to output $y(t)$ is:

$$\frac{Y(s)}{R(s)} = \frac{C(s)P(s)}{1 + C(s)P(s)} \quad (\text{Chapter 10.30})$$

We can also calculate other relevant transfer functions:

$$\frac{Y(s)}{R(s)} = \frac{C(s)P(s)}{1 + C(s)P(s)} \quad (\text{tracking}) \quad (\text{Chapter 10.31})$$

$$\frac{E(s)}{R(s)} = \frac{1}{1 + C(s)P(s)} \quad (\text{tracking error}) \quad (\text{Chapter 10.32})$$

$$\frac{E(s)}{W(s)} = \frac{-P(s)}{1 + C(s)P(s)} \quad (\text{regulation}) \quad (\text{Chapter 10.33})$$

$$\frac{U(s)}{R(s)} = \frac{C(s)}{1 + C(s)P(s)} \quad (\text{control effort}) \quad (\text{Chapter 10.34})$$

The denominator $1 + C(s)P(s)$ is shared across all closed-loop transfer functions, and its roots determine closed-loop stability.

We can also use this very helpful table to determine the numerator for transfer functions here but remember that this is only for unity feedback.

	R	W	V
Y	CP	P	-CP
E	1	-P	CP
U	C	-CP	-C

Figure Chapter10.11: Unity Feedback Transfer Function Numerator Table

Once the feedback loop is closed, the system no longer behaves like the open-loop plant. Instead, its dynamics are governed entirely by the closed-loop transfer function, which depends directly on the controller $C(s)$.

Remember, to guarantee that the closed-loop system is stable, all poles of the transfer function must lie in the left half of the complex s -plane. For real poles, this means each must satisfy $\text{Re}(s_i) < 0$.

Interpretation of These Transfer Functions

Stability: We want all real values of the poles to be less than zero and to satisfy our previous stability criteria.

Tracking: To achieve good tracking, we want the output $y(t)$ to follow the reference $r(t)$ closely. This implies the tracking error $e(t) = r(t) - y(t)$ should be small. From:

$$\frac{E(s)}{R(s)} = \frac{1}{1 + C(s)P(s)} \quad (\text{Chapter 10.35})$$

we see that increasing the magnitude of $C(s)$ (controller gain) reduces this quantity, leading to better tracking performance.

Regulation: Similarly, to reject disturbances $w(t)$, we require:

$$\frac{E(s)}{W(s)} = \frac{-P(s)}{1 + C(s)P(s)} \quad (\text{Chapter 10.36})$$

to be small in magnitude. Again, a larger $C(s)$ suppresses the impact of disturbances on the error. In this case, we say the controller has high authority, meaning it dominates the behavior of the system.

Control Effort: However, increasing controller gain also increases the required actuation. From:

$$\frac{U(s)}{R(s)} = \frac{C(s)}{1 + C(s)P(s)} \quad (\text{Chapter 10.37})$$

we observe that if $C(s)$ is too large, the control effort $u(t)$ may become excessive. This could lead to actuator saturation, overheating, or inefficiency, especially if $P(s)$ has limited bandwidth or sensitivity.

Trade-Off Between Tracking and Effort

There is a fundamental trade-off in feedback control:

Larger $C(s)$: better tracking, better disturbance rejection, but more control effort

Smaller $C(s)$: less effort, but poorer performance

Example: Gain Controller

Consider a proportional controller: $C(s) = K$, and suppose the plant is:

$$P(s) = \frac{1}{s - 2} \quad (\text{Chapter 10.38})$$

This plant is unstable (it has a pole at $s = 2$ in the right-half plane).

The closed-loop transfer function becomes:

$$G_{cl}(s) = \frac{K}{s - 2 + K} \quad (\text{Chapter 10.39})$$

The closed-loop pole is located at:

$$s = 2 - K \quad (\text{Chapter 10.40})$$

Analysis:

If $K < 2$, the pole remains in the RHP $\Rightarrow$ unstable

If $K = 2$, the pole is at the origin $\Rightarrow$ marginally stable

If $K > 2$, the pole moves to the LHP $\Rightarrow$ system becomes stable

Thus, by selecting $K > 2$, feedback control stabilizes the system.

Assume a unit step input $r(t) = 1 \Rightarrow R(s) = \frac{1}{s}$. Then the output becomes:

$$Y(s) = G(s) \cdot R(s) = \frac{K}{(s - 2 + K)s} \quad (\text{Chapter 10.41})$$

We can see the system response here from unstable, to stable, to improved performance as K increases.

`Manipulate[Plot[yt /. k → 1, {t, 0, 5}, PlotRange → All], {1, 0, 50}]`

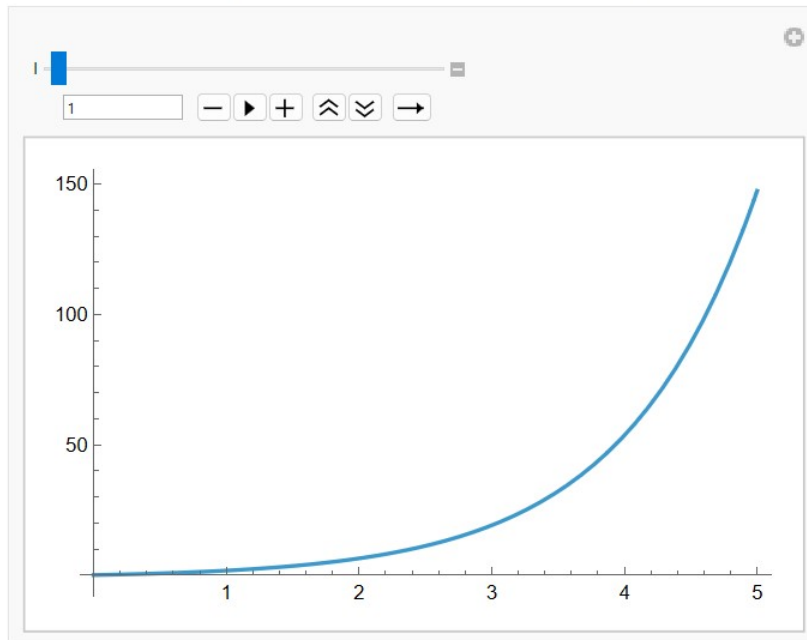


Figure Chapter10.12: Unstable System Response

`Manipulate[Plot[yt /. k → 1, {t, 0, 5}, PlotRange → All], {1, 0, 50}]`

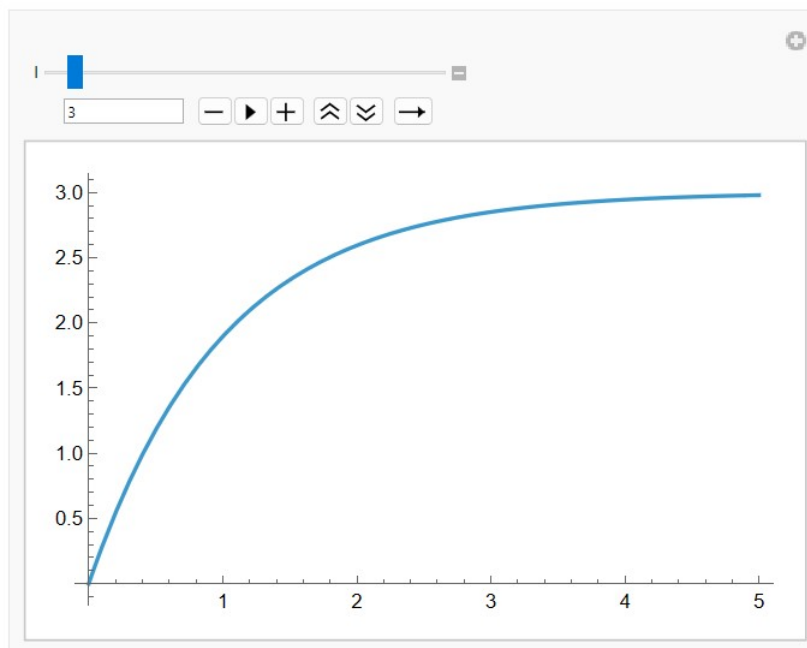


Figure Chapter10.13: Stable but poor performance

Manipulate[Plot[$y(t) \rightarrow 1$, { t , 0, 5}], PlotRange → All], { K , 0, 50}]

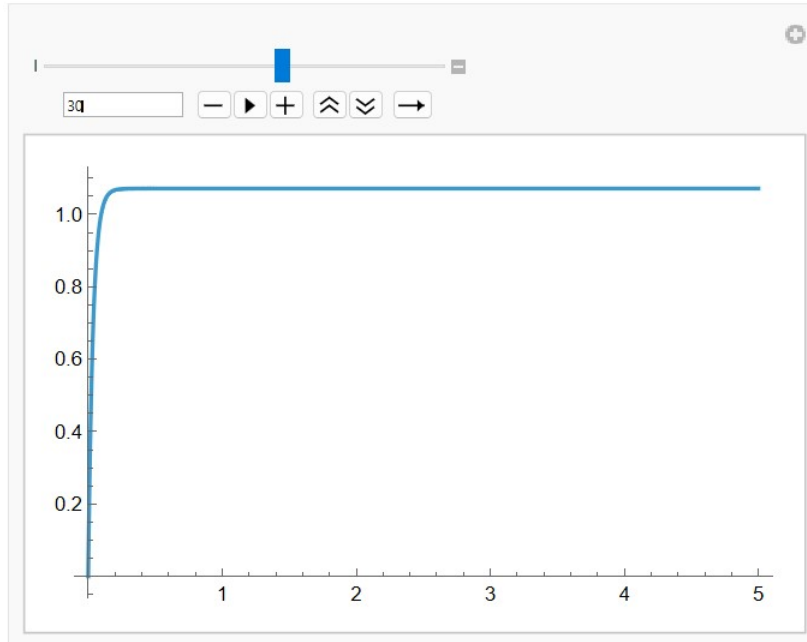


Figure Chapter10.14: Stable improved performance

We define the control effort in the Laplace domain as:

$$U(s) = C(s) \cdot E(s) = C(s) \cdot (R(s) - Y(s)) \quad (\text{Chapter 10.42})$$

Alternatively, we can derive $U(s)$ directly as:

$$U(s) = \frac{C(s)}{1 + C(s)P(s)} \cdot R(s) = \frac{K}{1 + \frac{K}{s-2}} \cdot \frac{1}{s} = \frac{K(s-2)}{(s + (K-2))s} \quad (\text{Chapter 10.43})$$

Thus:

$$U(s) = \frac{K(s-2)}{s(s + (K-2))} \quad (\text{Chapter 10.44})$$

We can see the system response for an unstable, stable, and controller response for larger K here.

`Manipulate[Plot[ut /. k → i, {t, 0, 5}, PlotRange → All], {i, 0, 10}]`

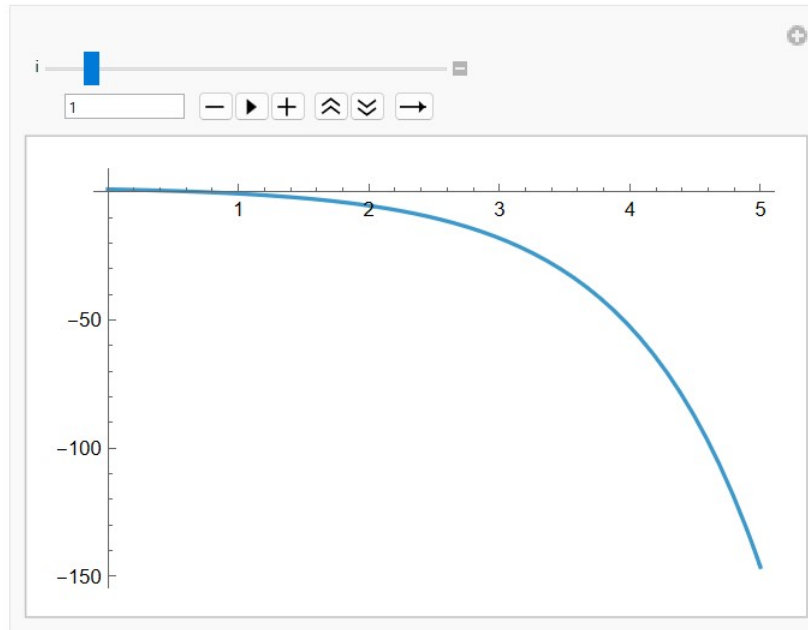


Figure Chapter10.15: Unstable System Response

`Manipulate[Plot[ut /. k → i, {t, 0, 5}, PlotRange → All], {i, 0, 10}]`

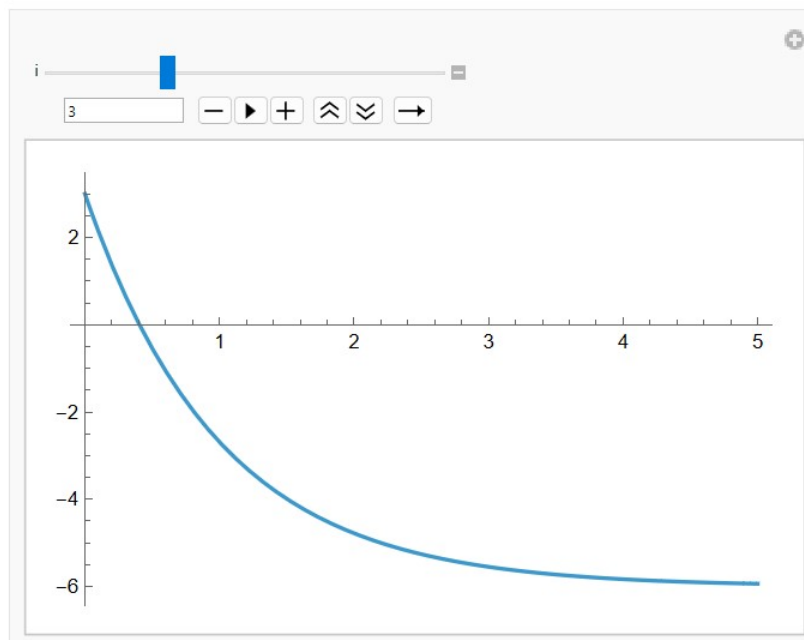


Figure Chapter10.16: Stable but poor performance

Manipulate[Plot[ut /. k → i, {t, 0, 5}, PlotRange → All], {i, 0, 10}]

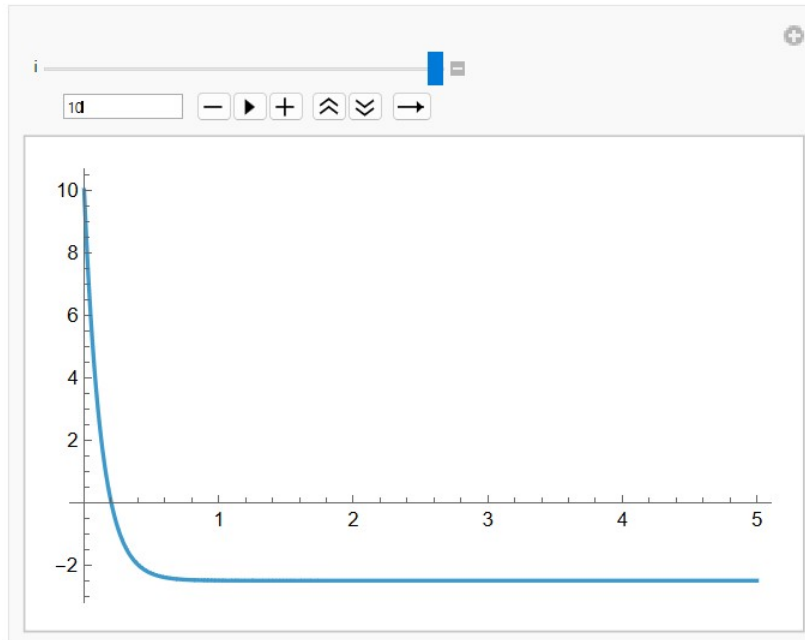


Figure Chapter10.17: Stable improved performance

Trade-Off and Risk of Instability

While increasing K improves tracking and speeds up convergence, it also:

Increases required actuation $u(t)$

May exceed actuator limits (saturation, overheating)

Can destabilize the system if $P(s)$ has fragile dynamics

Even though we started with an unstable plant, we can stabilize it with $K > 2$. However, this comes at the cost of potentially excessive control effort.

If we attempt to apply a similar strategy to a different plant (e.g., a stable but lightly damped one), increasing K without analysis may introduce instability through:

Poorly located poles

Reduced phase margin

Interaction with delays or non-minimum phase elements

Example: Feedback Can Destabilize a Stable Plant

In this example, we explore how a poorly chosen controller can destabilize an otherwise stable system. Let:

$$P(s) = \frac{2}{s+1}, \quad C(s) = \frac{s-k}{s+2} \quad (\text{Chapter 10.45})$$

This represents a stable first-order plant (pole at $s = -1$) and a controller that introduces both a pole at $s = -2$ and a zero at $s = k$, where $k \in \mathbb{R}$ is a tunable parameter.

Closed-Loop Transfer Function

Under unity feedback, the closed-loop transfer function is:

$$G_{cl}(s) = \frac{C(s)P(s)}{1 + C(s)P(s)} = \frac{\left(\frac{s-k}{s+2}\right) \cdot \left(\frac{2}{s+1}\right)}{1 + \left(\frac{s-k}{s+2}\right) \cdot \left(\frac{2}{s+1}\right)} \quad (\text{Chapter 10.46})$$

Multiply numerator and denominator:

$$G_{cl}(s) = \frac{2(s-k)}{(s+2)(s+1) + 2(s-k)} \quad (\text{Chapter 10.47})$$

Now simplify the denominator:

$$(s+2)(s+1) = s^2 + 3s + 2 \quad (\text{Chapter 10.48})$$

$$2(s-k) = 2s - 2k \quad (\text{Chapter 10.49})$$

$$\Rightarrow \text{Denominator: } s^2 + 3s + 2 + 2s - 2k = s^2 + 5s + 2 - 2k \quad (\text{Chapter 10.50})$$

So the closed-loop transfer function becomes:

$$G_{cl}(s) = \frac{2(s-k)}{s^2 + 5s + 2 - 2k} \quad (\text{Chapter 10.51})$$

Characteristic Equation

To analyze stability, we examine the denominator:

$$s^2 + 5s + (2 - 2k) = 0 \quad (\text{Chapter 10.52})$$

Let us denote the characteristic polynomial:

$$P(s) = s^2 + 5s + (2 - 2k) \quad (\text{Chapter 10.53})$$

This is a second-order polynomial. To ensure stability, both roots must lie in the left-half complex plane. This happens if and only if:

All coefficients are positive

The roots have negative real parts

Condition 1: Positive Coefficients

We already have:

$$\text{Coefficient of } s^2 = 1 > 0, \quad \text{Coefficient of } s = 5 > 0 \quad (\text{Chapter 10.54})$$

So we require:

$$2 - 2k > 0 \Rightarrow k < 1 \quad (\text{Chapter 10.55})$$

Condition 2: Discriminant

The roots are given by:

$$\begin{aligned} s &= \frac{-5 \pm \sqrt{25 - 4(2 - 2k)}}{2} \\ &= \frac{-5 \pm \sqrt{25 - 8 + 8k}}{2} = \frac{-5 \pm \sqrt{17 + 8k}}{2} \end{aligned} \quad (\text{Chapter 10.56})$$

This is always real for $k \geq -\frac{17}{8}$, but we care only about whether both roots are in the left half-plane.

From the quadratic formula, regardless of whether the roots are real or complex, the real part of both roots is always:

$$\frac{-5}{2} < 0 \quad (\text{Chapter 10.57})$$

So stability in this case is entirely determined by the constant term:

$$\boxed{k < 1} \quad (\text{Chapter 10.58})$$

We can see the poles here:

```
PoleZeroPlotSimple[TransferFunctionModel[Gs /. k -> 3, s]]
```

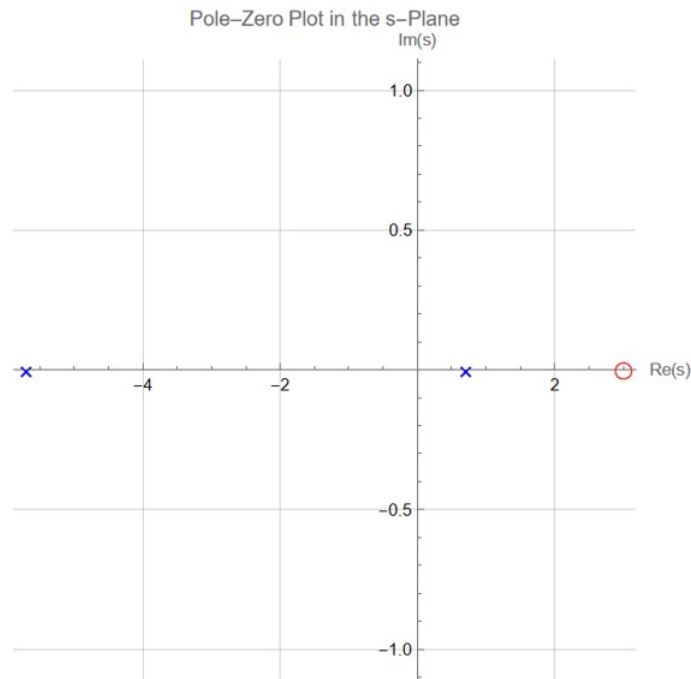


Figure Chapter10.18: Destabilized

```
PoleZeroPlotSimple[TransferFunctionModel[Gs /. k -> 0.1, s]]
```

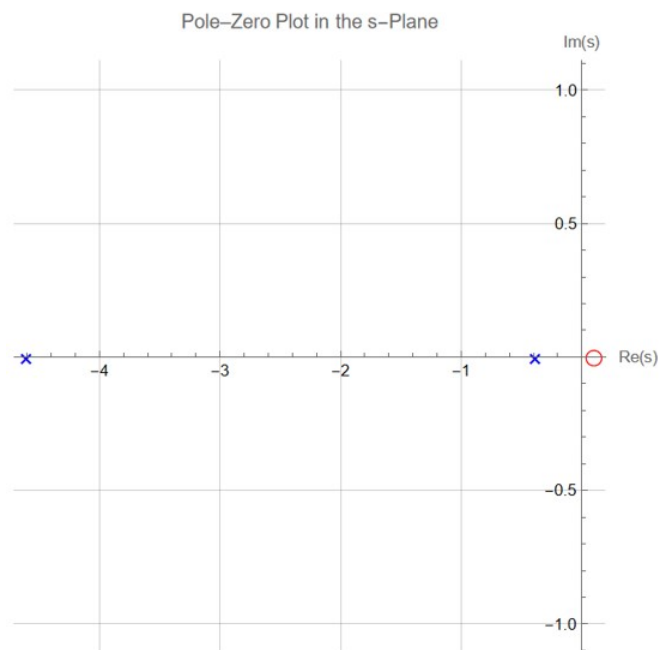


Figure Chapter10.19: Stable System

Conclusion

The closed-loop system becomes unstable when the controller zero is placed to the right of $s = 1$. This illustrates an important concept:

Even though the plant $P(s) = \frac{2}{s+1}$ is stable,

and the controller has stable dynamics $C(s) = \frac{s-k}{s+2}$,

the interaction between the controller and plant introduces a pole-zero interplay that can destabilize the system.

Key Insight

Stability is not guaranteed just because individual components are stable. The location of the controller zero (at $s = k$) can destabilize the system if it disrupts the structure of the closed-loop poles.

Design Rule: Avoid placing controller zeros too far to the right, especially when they affect low-order or lightly damped systems.

This example illustrates a central principle of control design:

Feedback allows us to change pole locations and stabilize unstable systems.

Increased controller gain improves performance (faster $y(t)$, smaller $e(t)$).

However, excessive gain increases $u(t)$ and may destabilize the system.

We must balance performance, control effort, and robustness.

This is why pole placement, root locus, and frequency response tools are essential for rigorous controller design.

Feedback is powerful but can be dangerous if misapplied. For example, if the plant is already stable, adding a high-gain controller without considering phase margin or bandwidth limitations can introduce unwanted dynamics (e.g., unstable pole-zero cancellations, delayed phase lag). Therefore, control design must always verify:

Stability (pole locations in LHP)

Performance (settling time, overshoot)

Feasibility (bounded control effort)

Feedback control enables designers to systematically modify system dynamics, improve performance, and reject disturbances. However, this power must be exercised with care. Poorly tuned controllers may consume excessive energy or destabilize the system. Control design is therefore a process of negotiation between performance, robustness, and practical constraints.

In future topics, we will explore controller structures such as PID, root locus design, and frequency-domain techniques (Bode, Nyquist), all of which build on this foundation of time-domain performance, stability, and block diagram relationships.

This page titled [Chapter 10: System](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 11: PID Control

Overview of PID Control

Proportional-Integral-Derivative (PID) control is one of the most widely used control strategies in both academic and industrial settings. Its versatility, simplicity, and effectiveness make it a go-to solution for systems ranging from robotics and aerospace to HVAC and process control.

The general PID controller has the following transfer function:

$$C(s) = K_p + \frac{K_i}{s} + K_d s \quad (\text{Chapter 11.1})$$

In practical implementations, this controller structure means that the error signal $e(t) = r(t) - y(t)$ is passed through three parallel control blocks:

- A proportional gain K_p
- An integrator with gain K_i
- A differentiator with gain K_d

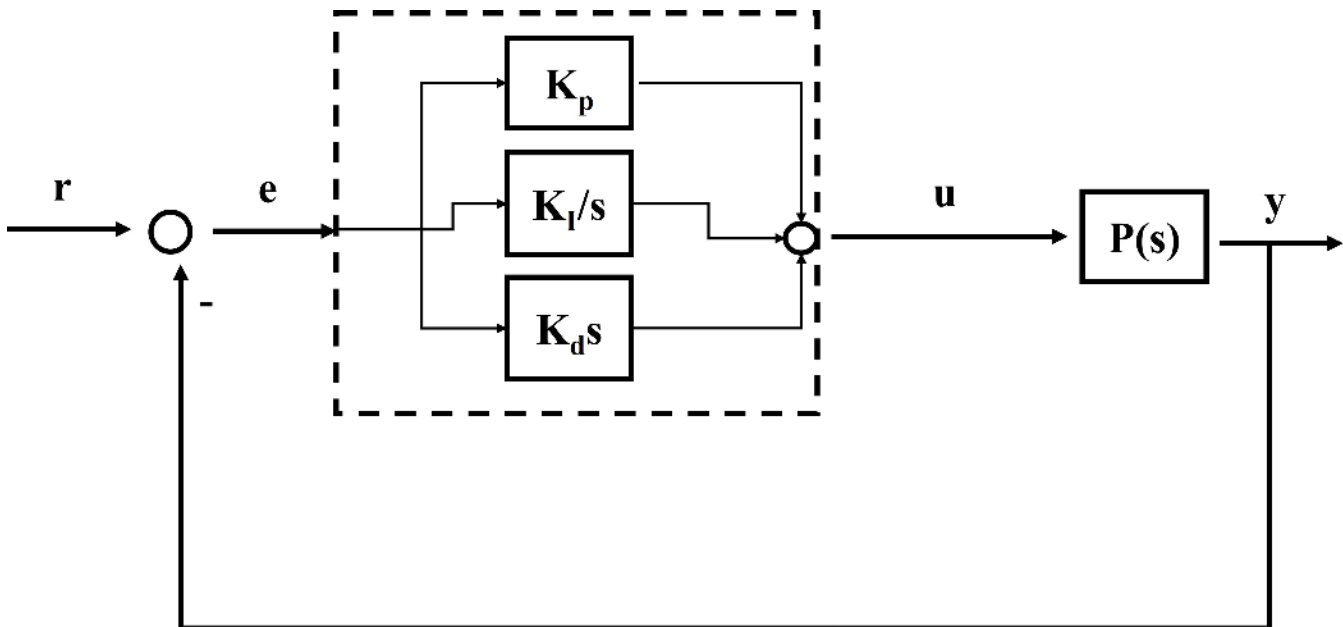


Figure Chapter11.1: PID Control

The outputs of these blocks are then summed to produce the control command $u(t)$. This structure gives us control over the transient and steady-state behavior of the system, and the ability to shift closed-loop poles. However, due to the presence of three tunable gains, the PID parameter space is large and often unintuitive without a structured tuning method or physical insight.

Motivation for PID: Unknown Plant Case

One of the key strengths of PID control is that it does not require an exact mathematical model of the plant $P(s)$. In practice, many industrial systems (e.g., HVAC units, process control loops) operate with minimal knowledge of the underlying dynamics. In such situations, PID controllers are often tuned heuristically or using standard procedures, and can still provide robust and acceptable performance.

This model-free feature is one of the reasons PID control has remained dominant in the field despite the rise of more modern model-based control strategies.

Physical Interpretation Using Mass-Spring-Damper

To build intuition, consider a mass-spring-damper system modeled by:

$$m\ddot{y}(t) + c\dot{y}(t) + ky(t) = u(t)$$

(Chapter 11.2)

This system has:

- A damping term $c\dot{y}(t)$
- A stiffness term $ky(t)$
- A mass term $m\ddot{y}(t)$

We wish to regulate the output $y(t)$ so that it closely follows a reference input $r(t)$, despite the presence of external disturbances. This is a regulation control objective. Now let's examine the effect of each PID term individually.

Proportional Control: K_p

If we apply a pure proportional controller, $C(s) = K_p$, the closed-loop system becomes:

$$G_{cl}(s) = \frac{K_p P(s)}{1 + K_p P(s)} \quad (\text{Chapter 11.3})$$

given

$$P(s) = \frac{1}{ms^2 + cs + k} \quad (\text{Chapter 11.4})$$

We want to determine the poles of the closed-loop system, i.e., the roots of the equation:

$$s = \frac{-c \pm \sqrt{c^2 - 4m(k + K_p)}}{2m} \quad (\text{Chapter 11.5})$$

Closed-Loop Poles

Let us define the poles explicitly:

$$s_{1,2} = \frac{-c \pm \sqrt{c^2 - 4m(k + K_p)}}{2m} \quad (\text{Chapter 11.6})$$

Interpretation:

- If $c^2 > 4m(k + K_p)$: real and distinct (overdamped)
- If $c^2 = 4m(k + K_p)$: repeated real root (critically damped)
- If $c^2 < 4m(k + K_p)$: complex conjugate (underdamped)

As we increase K_p , the term $k + K_p$ increases, which:

- Increases the natural frequency $\omega_n = \sqrt{(k + K_p)/m}$
- Decreases the damping ratio $\zeta = \frac{c}{2\sqrt{m(k + K_p)}}$

Thus, increasing K_p causes:

- Faster oscillation (higher frequency)
- Lower damping (more oscillatory response)
- Settling time remains largely unchanged, because $\zeta\omega_n \approx \text{constant}$ if c and m are fixed

This adds a static gain into the loop, which primarily influences the system's stiffness. Specifically:

- Increasing K_p moves the poles in the complex plane further up the imaginary axis
- ω_n increases (higher natural frequency)
- Damping ratio ζ decreases which leads more oscillatory response
- The real part σ remains unchanged which leads to no improvement in settling time
- Control effort increases and will lead to large control signals

Let's actually visualize this by considering the scenario of our old mass, spring constant, and damping coefficient values we worked with previously for our time domain specs.

Thus, proportional control can speed up the response and improve tracking, but often at the cost of increased oscillations and limited impact on convergence rate.

Numerical Example: $m = 2$, $c = 8$, $k = 200$

We substitute into the closed-loop pole formula:

$$s = \frac{-8 \pm \sqrt{8^2 - 4(2)(200 + K_p)}}{2(2)} = \frac{-8 \pm \sqrt{64 - 16(200 + K_p)}}{4} \quad (\text{Chapter 11.7})$$

The system transitions from overdamped to underdamped when the discriminant becomes negative:

$$64 - 16(200 + K_p) < 0 \Rightarrow K_p > -196 \quad (\text{Chapter 11.8})$$

Since all practical K_p values are nonnegative, the system is always underdamped for all $K_p > 0$. We can see the system response and poles here:

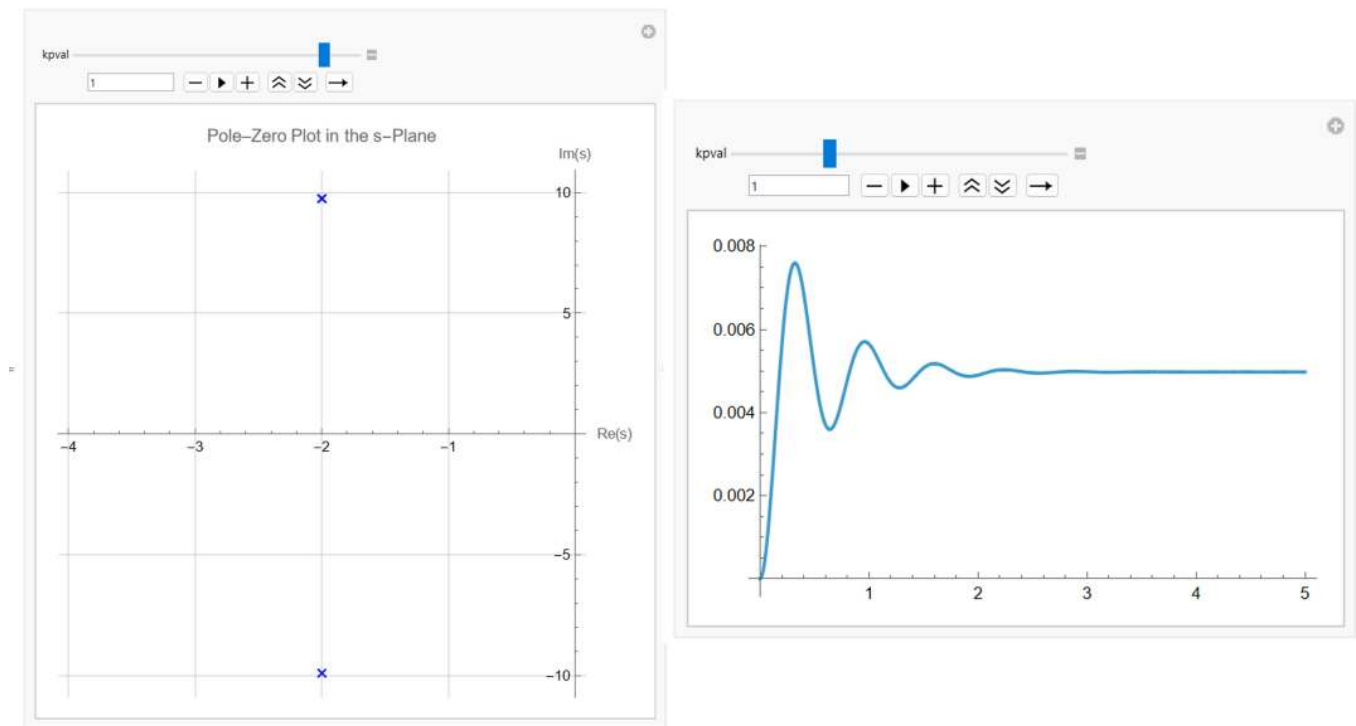


Figure Chapter11.2: P Control $K_p = 1$

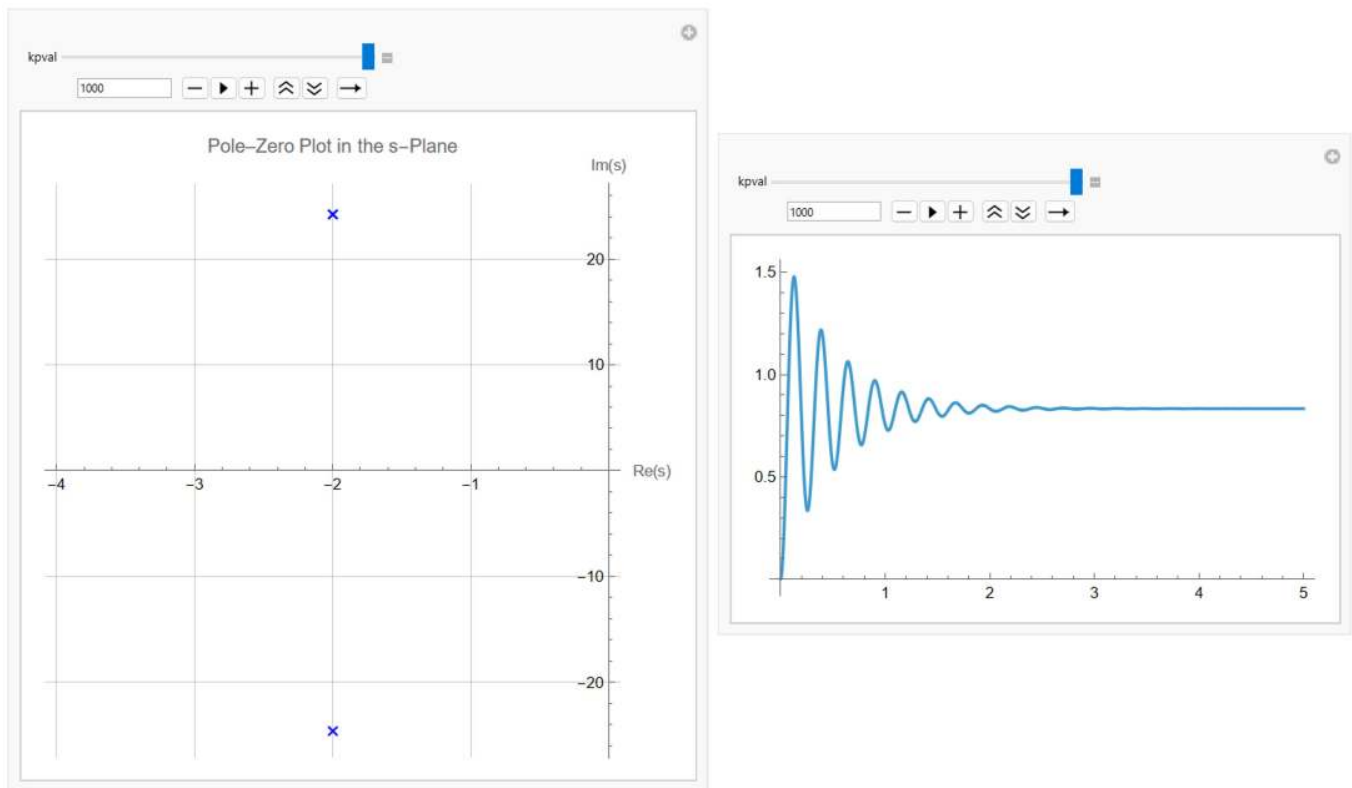


Figure Chapter11.3: P Control $K_p = 1000$

Takeaway: Increasing K_p speeds up the system (higher ω_n) but reduces damping (ζ). This leads to faster but more oscillatory responses.

Derivative Control: K_d

Adding derivative control gives:

$$C(s) = K_p + K_d s \quad (\text{Chapter 11.9})$$

This introduces a velocity-dependent term and increases the effective damping in the system.

- K_d increases the coefficient on s in the denominator which leads to improved damping
- The complex poles move leftward (more negative real part) and down in imaginary axis
- ζ increases which leads to less oscillation
- σ increases which leads to faster settling time
- More phase margin which leads to improved robustness

In effect, derivative control acts like adding a physical damper. It smooths out the oscillations and increases the rate of convergence. When combined with K_p , this forms a PD controller, which gives independent tuning over stiffness and damping.

Closed-Loop Transfer Function

Under unity feedback, the closed-loop transfer function becomes:

$$G(s) = \frac{(K_p + K_d s) \cdot P(s)}{1 + (K_p + K_d s) \cdot P(s)} \quad (\text{Chapter 11.10})$$

Let us collect terms in the denominator:

$$G(s) = \frac{K_p + K_d s}{ms^2 + (c + K_d)s + (k + K_p)} \quad (\text{Chapter 11.11})$$

Characteristic Equation

The denominator is the characteristic polynomial:

$$ms^2 + (c + K_d)s + (k + K_p) \quad (\text{Chapter 11.12})$$

This is again a second-order system, but now both the damping and stiffness terms are affected:

- $c_{\text{eff}} = c + K_d$ — damping increases with K_d
- $k_{\text{eff}} = k + K_p$ — stiffness increases with K_p

Closed-Loop Poles

The roots of the characteristic equation are:

$$s_{1,2} = \frac{-(c + K_d) \pm \sqrt{(c + K_d)^2 - 4m(k + K_p)}}{2m} \quad (\text{Chapter 11.13})$$

Interpretation:

- Increasing K_d leads to more damping and poles move leftward in the s-plane
- Increasing K_p leads to more stiffness and poles move away from the imaginary axis (increased natural frequency)
- You now have independent control over:
 - The rate of convergence (via K_d)
 - The frequency of oscillation (via K_p)

System Parameters in Terms of K_p , K_d

$$\omega_n = \sqrt{\frac{k + K_p}{m}}, \quad \zeta = \frac{c + K_d}{2\sqrt{m(k + K_p)}} \quad (\text{Chapter 11.14})$$

These expressions show how both gains directly shape the closed-loop response. By adjusting K_p and K_d , we can place the poles in the complex plane with a specific damping ratio ζ and natural frequency ω_n , enabling faster convergence and reduced overshoot.

Benefits of PD Control

- Improves response speed via increased natural frequency
- Increases damping to reduce overshoot and oscillations
- Settling time decreases because poles move further left
- Better transient behavior and robustness

Limitations

- No integrator → cannot eliminate steady-state error
- Still a Type 0 system → steady-state error persists for step inputs
- Sensitive to high-frequency noise (derivative amplification)

Therefore:

- PD control introduces a zero into the loop
- K_p and K_d provide two degrees of freedom to shape the transient response
- Poles can be placed to achieve desired performance — fast, well-damped convergence
- Still limited in steady-state performance without an integrator

Numerical Example: $m = 2$, $c = 8$, $k = 200$

The characteristic equation is:

$$2s^2 + (8 + K_d)s + (200 + K_p) \quad (\text{Chapter 11.15})$$

The closed-loop poles are:

$$s_{1,2} = \frac{-(8 + K_d) \pm \sqrt{(8 + K_d)^2 - 8(200 + K_p)}}{4}$$

(Chapter 11.16)

We can also see some of the poles and system response:

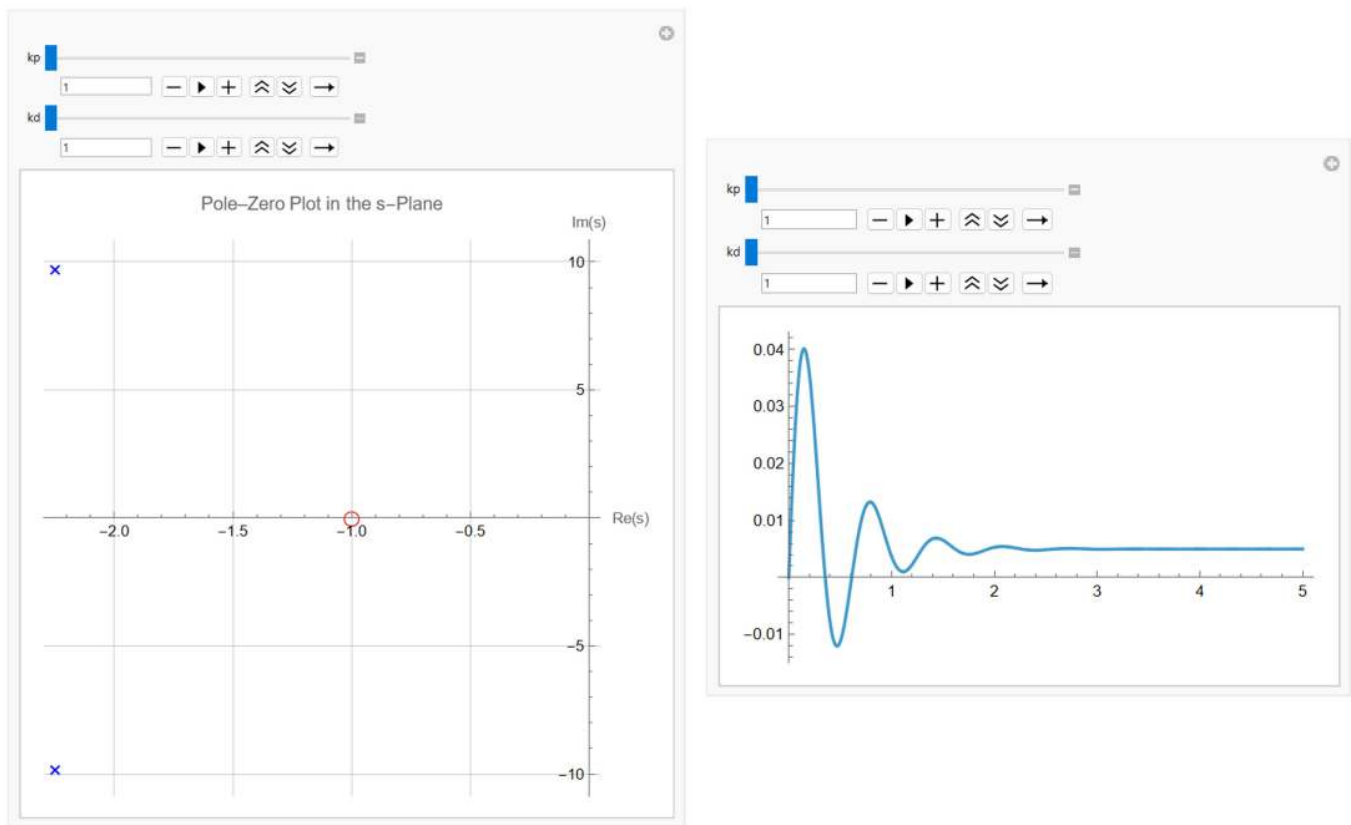


Figure Chapter11.4: PD Control Kp = 1 and Kd = 1

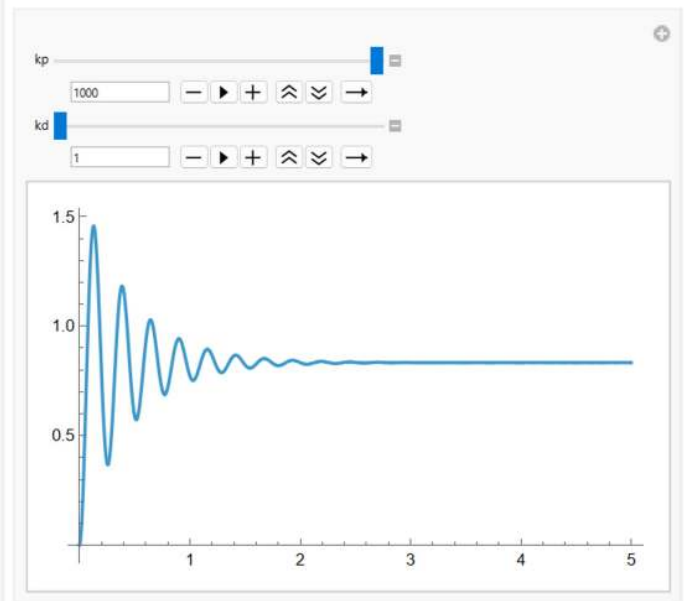
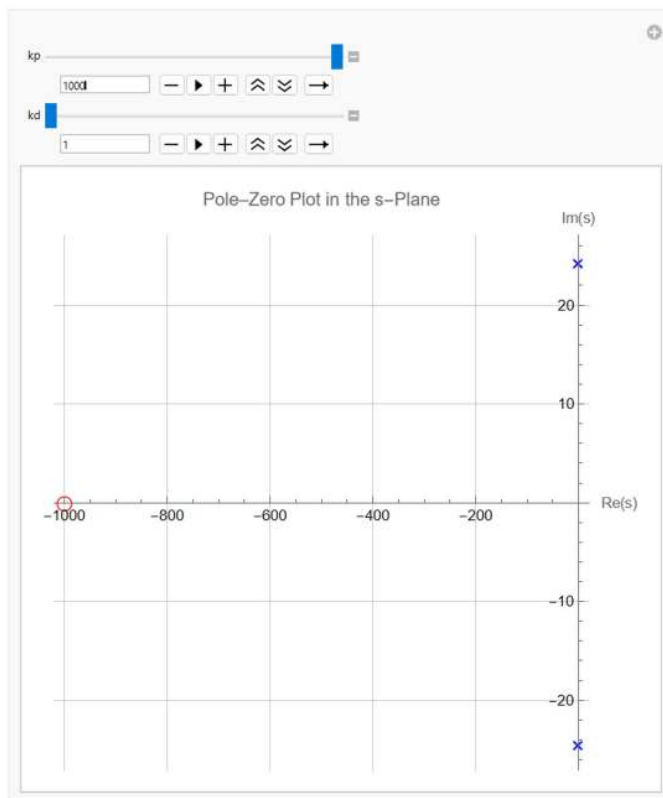


Figure Chapter11.5: PD Control $K_p = 1000$ and $K_d = 1$

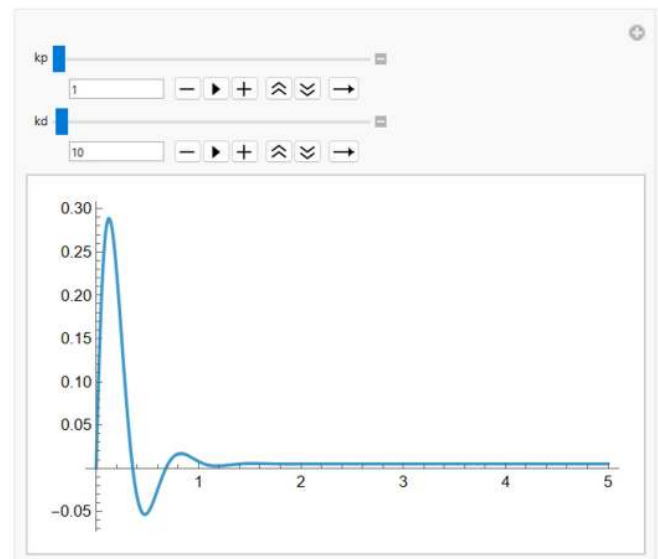
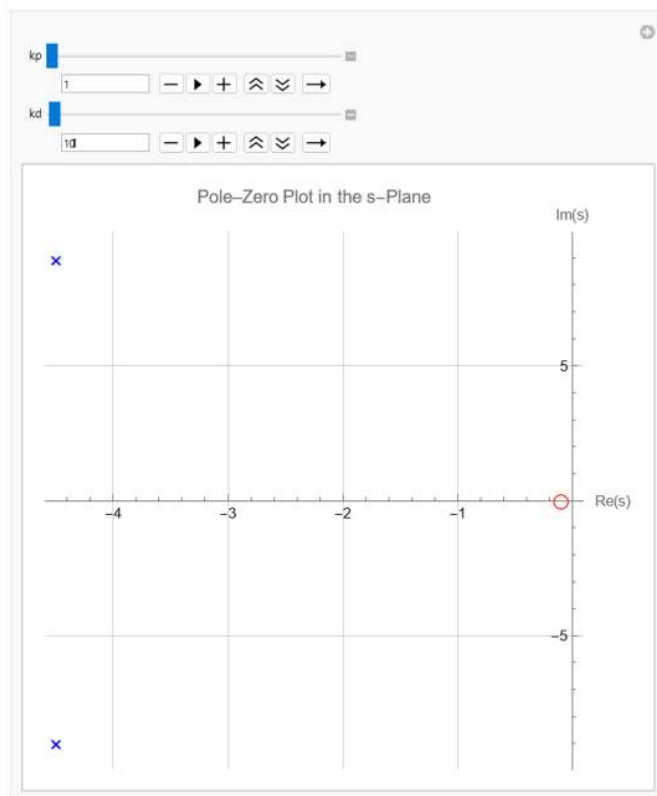


Figure Chapter11.6: PD Control $K_p = 1$ and $K_d = 10$

Conclusion:

As K_d increases:

- ζ increases — better damping
- Poles move further left — faster settling
- Response becomes smoother with less overshoot

Integral Control: K_i

Integral control introduces a pole at the origin:

$$C(s) = \frac{K_i}{s} \quad (\text{Chapter 11.17})$$

- Improves steady-state error and drives it to zero for step inputs
- Increases system type which leads to better tracking of low-order reference signals
- But introduces a new pole which therefore increases system order to third order

When included with K_p and K_d , the denominator becomes a cubic (third-order) polynomial. Now stability analysis requires methods such as Routh-Hurwitz or root locus.

Effect of Integral Gain on Stability

While K_i improves steady-state performance, it can also introduce instability. Consider root locus as K_i increases:

- One root starts at the origin (integrator pole)
- As K_i increases, this pole moves left (stabilizing), but other poles may migrate toward the right-half plane
- If K_i exceeds a critical value, the system becomes unstable

Thus, we must carefully tune K_i to avoid crossing the stability boundary.

Integral Control Applied to a Mass-Spring-Damper System

We now extend our controller by adding integral action. The controller becomes:

$$C(s) = K_p + \frac{K_i}{s} \quad (\text{Chapter 11.18})$$

This structure is known as a Proportional-Integral (PI) controller. The addition of $\frac{K_i}{s}$ introduces an integrator, which has profound effects on the system:

- It introduces a pole at the origin
- It increases the system type from 0 to 1
- It improves steady-state error — eliminates error for step reference input
- It increases the system order (denominator becomes cubic)

Plant Model

As before, the plant is modeled by a second-order mass-spring-damper system:

$$P(s) = \frac{1}{ms^2 + cs + k} \quad (\text{Chapter 11.19})$$

Closed-Loop Transfer Function

The closed-loop transfer function with unity feedback becomes:

$$G(s) = \frac{C(s)P(s)}{1 + C(s)P(s)} = \frac{\left(K_p + \frac{K_i}{s}\right) \cdot \frac{1}{ms^2 + cs + k}}{1 + \left(K_p + \frac{K_i}{s}\right) \cdot \frac{1}{ms^2 + cs + k}} \quad (\text{Chapter 11.20})$$

Simplifying:

$$G(s) = \frac{K_p s + K_i}{s(ms^2 + cs + k) + K_p s + K_i} \quad (\text{Chapter 11.21})$$

Characteristic Equation

Multiply out the denominator:

$$s(ms^2 + cs + k) + K_p s + K_i = ms^3 + cs^2 + ks + K_p s + K_i \quad (\text{Chapter 11.22})$$

Group terms:

$$\boxed{ms^3 + cs^2 + (k + K_p)s + K_i} \quad (\text{Chapter 11.23})$$

This is a third-order system. To assess stability, we must examine the roots of this cubic equation, typically using the Routh-Hurwitz criterion or root locus.

Effect of K_i on Performance and Stability

- The pole at the origin increases system type $\rightarrow$ better tracking for step inputs
- As K_i increases:
- One pole remains near the origin
- The other two poles may become lightly damped and eventually unstable
- There exists a critical value of K_i beyond which the system becomes unstable
- Must use Routh-Hurwitz or root locus to determine safe range of K_i

Numerical Example: PI Control on Mass-Spring-Damper

Let $m = 2$, $c = 8$, $k = 200$

The characteristic equation becomes:

$$2s^3 + 8s^2 + (200 + K_p)s + K_i \quad (\text{Chapter 11.24})$$

We assess stability using the Routh-Hurwitz criterion. The first column of the Routh table must be strictly positive for stability.

Example 1:

Let $K_p = 100$, $K_i = 50$

Characteristic polynomial:

$$2s^3 + 8s^2 + 300s + 50 \quad (\text{Chapter 11.25})$$

All first-column entries are positive $\rightarrow$ system is stable.

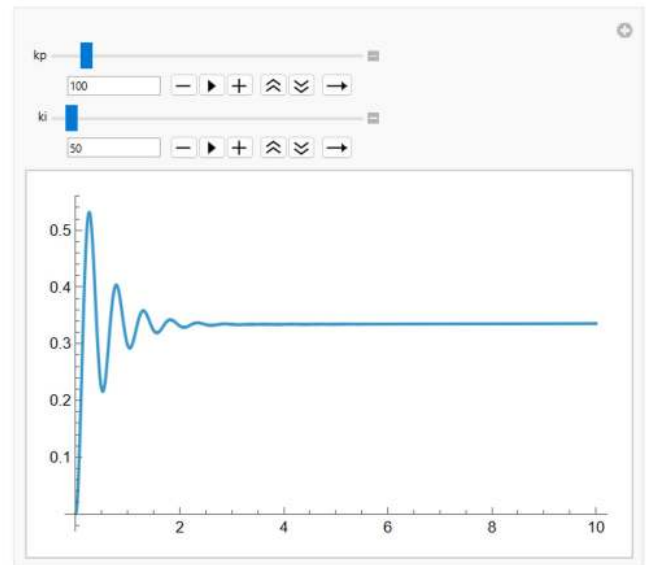
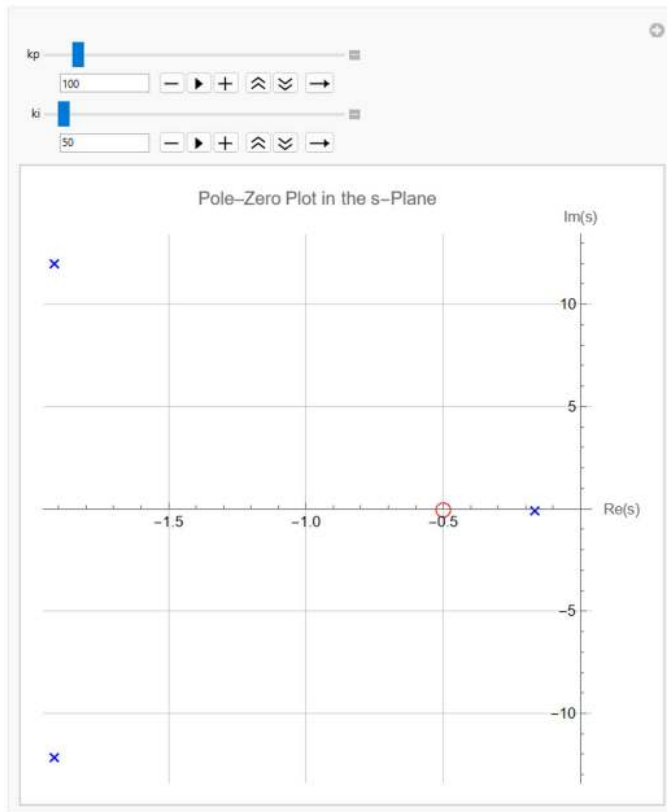


Figure Chapter11.7: PI Control Stable

Example 2:

Let $K_p = 0.1$, $K_i = 1000$

We are no longer stable, and we can see the response here:

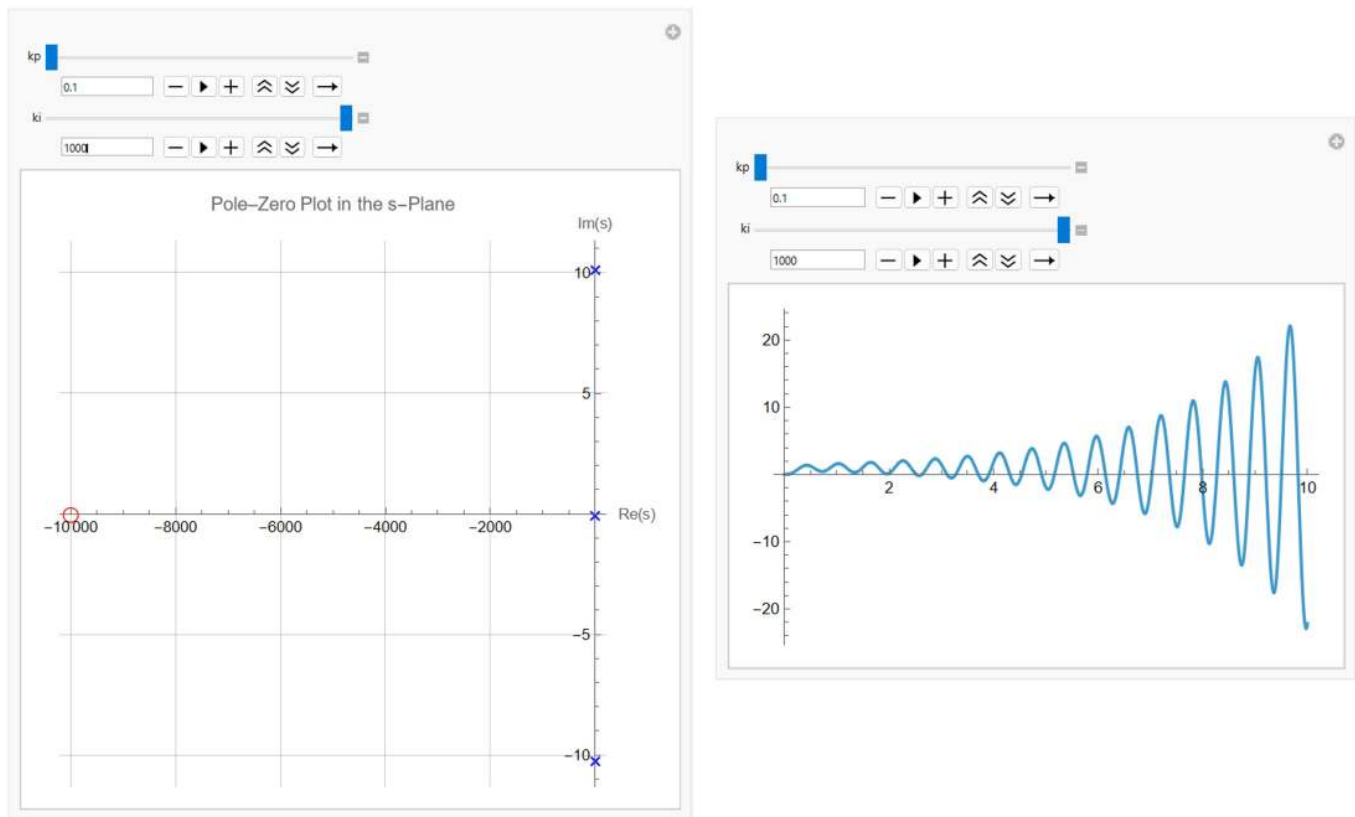


Figure Chapter11.8: PI Control Unstable

Conclusion:

- For this system, modest $K_i \in [0, 300]$ gives stability with fast convergence
- Large K_i risks instability and should be validated via Routh or root locus
- The addition of integral control improves tracking but raises the system order and complexity

Full PID Controller

Now we consider the complete Proportional-Integral-Derivative (PID) controller:

$$C(s) = K_p + \frac{K_i}{s} + K_d s = \frac{K_d s^2 + K_p s + K_i}{s} \quad (\text{Chapter 11.26})$$

Closed-loop transfer function:

$$G(s) = \frac{C(s)P(s)}{1 + C(s)P(s)} = \frac{(K_d s^2 + K_p s + K_i) \cdot \frac{1}{ms^2 + cs + k}}{1 + (K_d s^2 + K_p s + K_i) \cdot \frac{1}{s(ms^2 + cs + k)}} \quad (\text{Chapter 11.27})$$

Simplify the denominator:

$$\begin{aligned} & s(ms^2 + cs + k) + K_d s^2 + K_p s + K_i \\ &= ms^3 + (c + K_d)s^2 + (k + K_p)s + K_i \end{aligned} \quad (\text{Chapter 11.28})$$

Final Characteristic Equation for PID Control

The closed-loop poles are the roots of:

$$s^3 + \frac{c + K_d}{m}s^2 + \frac{k + K_p}{m}s + \frac{K_i}{m} = 0 \quad (\text{Chapter 11.29})$$

This third-order equation determines the system stability and performance. The PID controller allows complete shape control over this cubic polynomial.

Interpretation of Each Gain

- K_p : increases stiffness → improves rise time but may increase overshoot
- K_d : increases damping → reduces overshoot, improves settling time
- K_i : adds integrator → eliminates steady-state error but may destabilize

Tuning Caution

Although a PID controller gives three tunable gains and full control over a second-order system, the third-order dynamics introduced by the integrator can introduce unwanted oscillations or instability.

- Increasing K_i too much can push one or more poles into the right-half plane
- Trade-off exists between steady-state performance and transient robustness
- Tuning must be done carefully (e.g., Routh table, root locus, simulation)

Routh-Hurwitz Stability Analysis for PID Control

We now analyze the stability of the PID-controlled mass-spring-damper system using the Routh-Hurwitz criterion. From the previous derivation, the closed-loop characteristic polynomial is:

$$P(s) = ms^3 + (c + K_d)s^2 + (k + K_p)s + K_i \quad (\text{Chapter 11.30})$$

To assess stability, we construct the Routh table. First, define the normalized polynomial by dividing through by m :

$$P(s) = s^3 + \frac{c + K_d}{m}s^2 + \frac{k + K_p}{m}s + \frac{K_i}{m} \quad (\text{Chapter 11.31})$$

Let:

$$a_3 = 1, \quad a_2 = \frac{c + K_d}{m}, \quad a_1 = \frac{k + K_p}{m}, \quad a_0 = \frac{K_i}{m} \quad (\text{Chapter 11.32})$$

Construct the Routh Table

$$\begin{array}{c|cc} s^3 & 1 & \frac{k+K_p}{m} \\ s^2 & \frac{c+K_d}{m} & \frac{K_i}{m} \\ s^1 & b_1 & 0 \\ s^0 & \frac{K_i}{m} & - \end{array} \quad (\text{Chapter 11.33})$$

Where:

$$b_1 = \frac{\left(\frac{c+K_d}{m} \cdot \frac{k+K_p}{m} - 1 \cdot \frac{K_i}{m} \right)}{\frac{c+K_d}{m}} = \frac{(k + K_p) - \frac{K_i m}{c+K_d}}{m} \quad (\text{Chapter 11.34})$$

Stability Conditions

All the elements in the first column must be strictly positive for stability:

1. $\frac{c+K_d}{m} > 0 \Rightarrow c + K_d > 0$
2. $\frac{K_i}{m} > 0 \Rightarrow K_i > 0$
3. $b_1 > 0 \Rightarrow (k + K_p) > \frac{K_i m}{c+K_d}$

Interpretation

This provides clear guidance on gain selection:

- Derivative gain K_d must be large enough to ensure sufficient damping
- Integral gain K_i must be positive but cannot be too large
- There exists a critical value of K_i beyond which the system becomes unstable

Numerical Example: PID Controller

Let $m = 2$, $c = 8$, $k = 200$.

The closed-loop characteristic polynomial becomes:

$$2s^3 + (8 + K_d)s^2 + (200 + K_p)s + K_i \quad (\text{Chapter 11.35})$$

Normalize:

$$s^3 + \frac{8 + K_d}{2}s^2 + \frac{200 + K_p}{2}s + \frac{K_i}{2} \quad (\text{Chapter 11.36})$$

We define:

$$a_3 = 1, \quad a_2 = \frac{8 + K_d}{2}, \quad a_1 = \frac{200 + K_p}{2}, \quad a_0 = \frac{K_i}{2} \quad (\text{Chapter 11.37})$$

Routh Table:

$$\begin{array}{c|cc} s^3 & 1 & \frac{200+K_p}{2} \\ s^2 & \frac{8+K_d}{2} & \frac{K_i}{2} \\ s^1 & b_1 & 0 \\ s^0 & \frac{K_i}{2} & - \end{array} \quad (\text{Chapter 11.38})$$

Where:

$$b_1 = \frac{\left(\frac{8+K_d}{2} \cdot \frac{200+K_p}{2} - 1 \cdot \frac{K_i}{2}\right)}{\frac{8+K_d}{2}} = \frac{200 + K_p - \frac{K_i}{8+K_d}}{2} \quad (\text{Chapter 11.39})$$

Stability Conditions:

All entries in the first column must be positive:

1. $\frac{8+K_d}{2} > 0 \Rightarrow K_d > -8$
2. $\frac{K_i}{2} > 0 \Rightarrow K_i > 0$
3. $b_1 > 0 \Rightarrow 200 + K_p > \frac{K_i}{8+K_d}$

Worked Example:

Let $K_p = 100$, $K_d = 4$, $K_i = 50$

Then:

$$b_1 = \frac{200 + 100 - \frac{50}{8+4}}{2} = \frac{300 - \frac{50}{12}}{2} = \frac{300 - 4.1667}{2} = 147.92 \quad (\text{Chapter 11.40})$$

All entries in the Routh table are positive therefore system is stable as seen below:

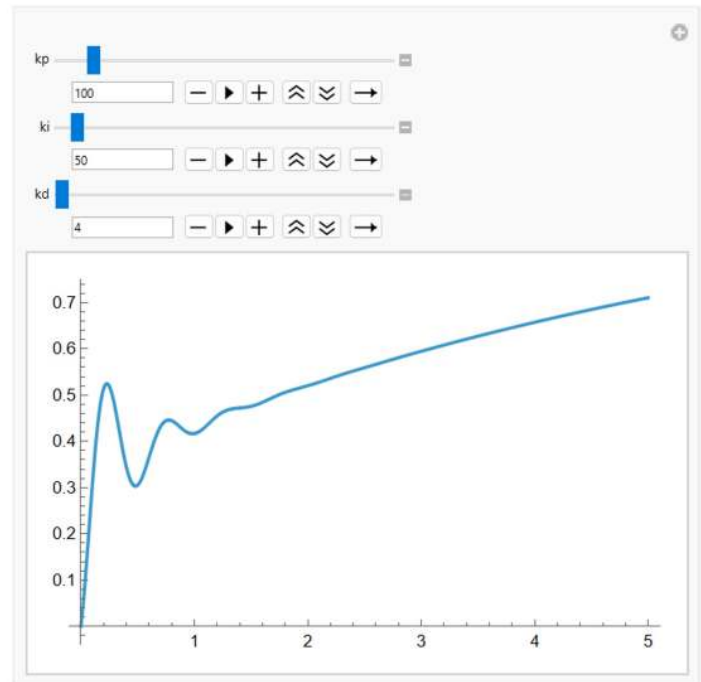
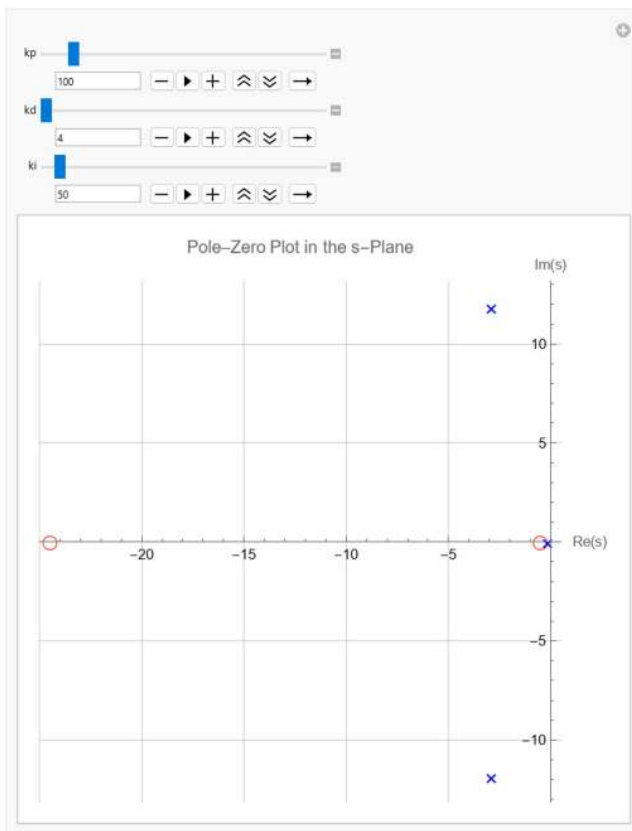


Figure Chapter11.9: PID Control

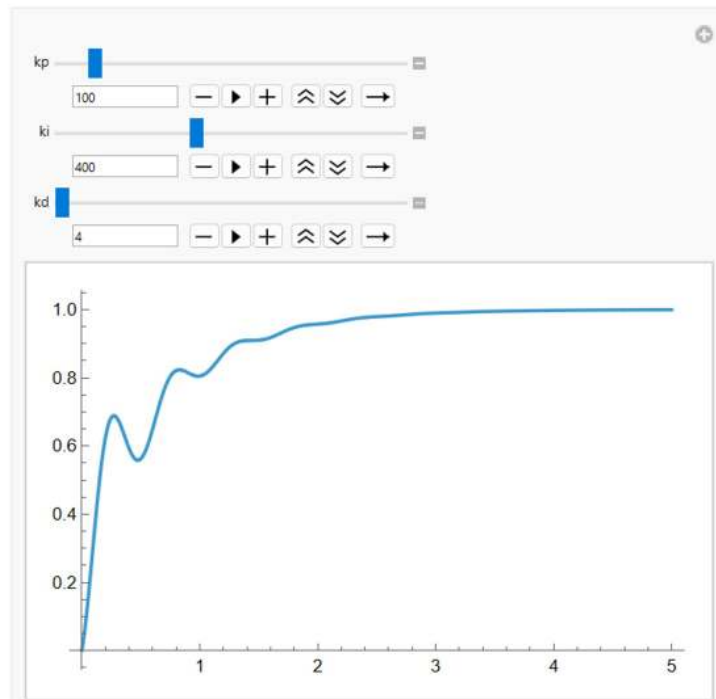
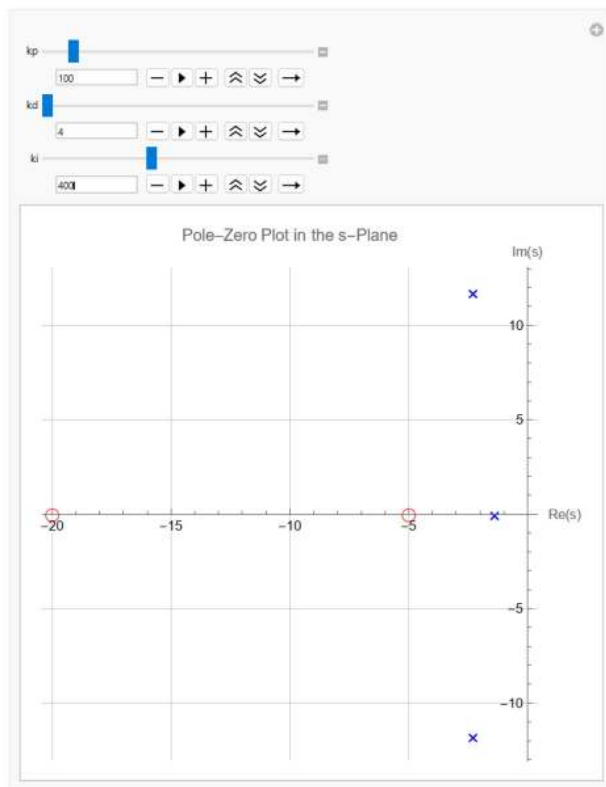


Figure Chapter11.10: PID Control

But if K_i is too large...

Let $K_i = 1200$, then:

$$b_1 = \frac{300 - \frac{1200}{12}}{2} = \frac{300 - 100}{2} = 100 \quad (\text{Chapter 11.41})$$

Still stable. Try $K_i = 2500$:

$$b_1 = \frac{300 - \frac{2500}{12}}{2} = \frac{300 - 208.33}{2} = 45.83 \quad (\text{Chapter 11.42})$$

Still positive. If $K_i > 3600$, $b_1 < 0$ thus instability begins as seen here:

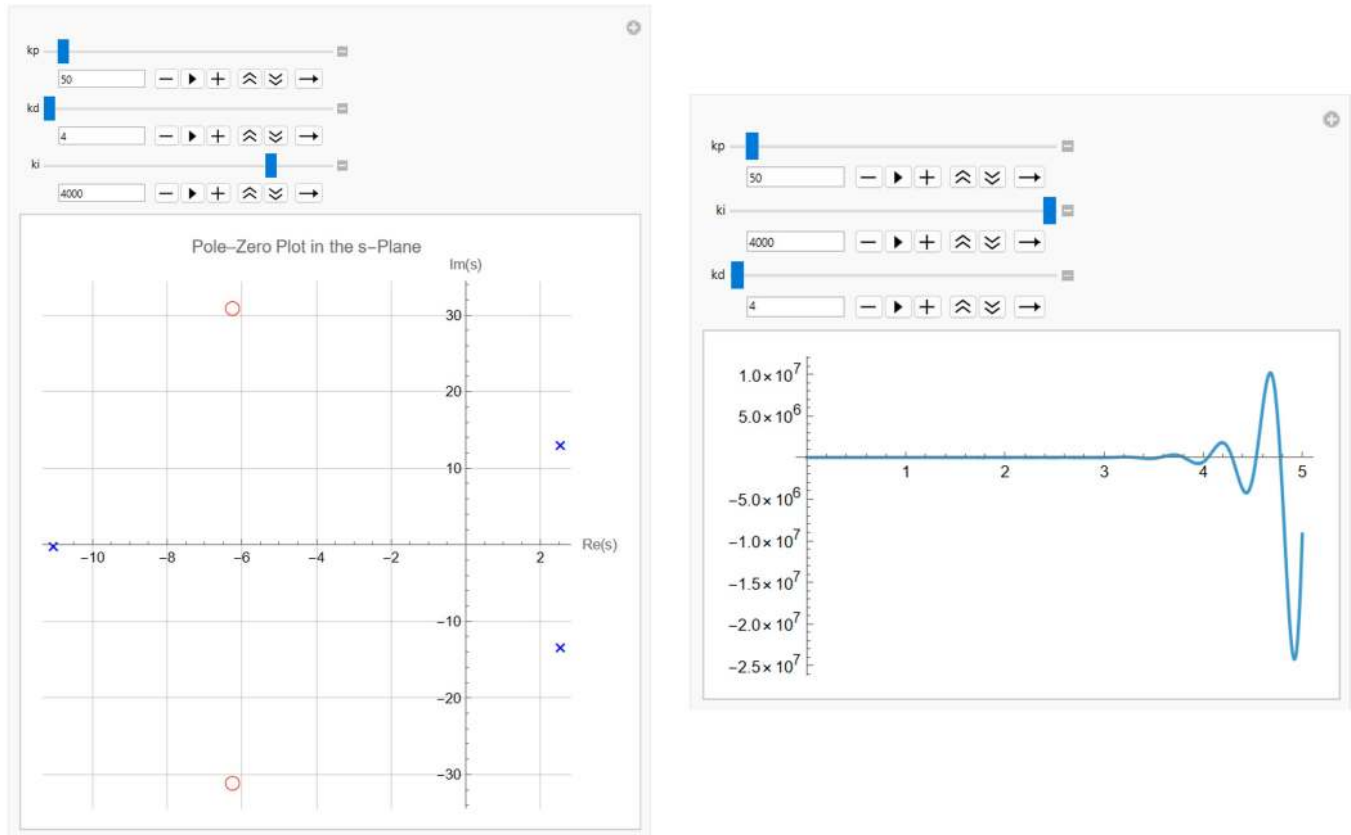


Figure Chapter11.11: PID Control

Conclusion:

- K_p shifts the poles for speed
- K_d adds damping and helps suppress overshoot
- K_i improves steady-state accuracy but must be bounded to preserve stability

Use the inequality below to ensure stability in this system:

$$K_i < (8 + K_d)(200 + K_p) \quad (\text{Chapter 11.43})$$

PID Control with Known and Unknown Plant Models

One of the key advantages of PID control is that it can be applied effectively whether or not the mathematical model of the plant is known. The design approach differs significantly depending on whether the plant model is available.

Case 1: Known Plant Model

If the plant transfer function $P(s)$ is known, either from first-principles modeling (e.g., free-body diagrams, energy methods, circuit laws) or system identification, then we have full access to analysis tools. These include:

- Derivation of the differential equation or transfer function
- Simulation of system dynamics

- Controller design via pole placement, root locus, frequency response, and state-space methods
- Formal analysis of stability (Routh-Hurwitz, Nyquist), transient response, and control effort

In this case, PID parameters can be tuned analytically to meet specified time-domain or frequency-domain performance criteria.

Case 2: Unknown Plant Model

In many real-world systems, particularly in industrial settings, the exact plant dynamics are unknown or difficult to model. For example:

- Thermal systems like HVAC
- Fluid flow control
- Legacy or black-box systems

In these situations, PID control is often applied using heuristic or empirical tuning. Engineers do not assume a precise transfer function but rather tune the controller based on measured input-output behavior.

Empirical PID Tuning Methods

Two widely used model-free PID tuning methods are:

1. Ziegler-Nichols Ultimate Sensitivity Method

This classical tuning method predates computer simulation tools and remains effective for many systems. The process is:

1. Set $K_i = 0$ and $K_d = 0$ (use only proportional control).
2. Gradually increase K_p until the output begins to oscillate with constant amplitude.
3. The gain at which sustained oscillations occur is called the **ultimate gain** K_u .
4. The period of oscillation is called the **ultimate period** T_u .
5. Use the following formulas to determine initial PID gains:

$$K_p = 0.1K_u, \quad K_i = \frac{2K_p}{T_u}, \quad K_d = \frac{K_p T_u}{8} \quad (\text{Chapter 11.44})$$

6. These gains can then be fine-tuned based on system performance.

This method pushes the system to the edge of instability (marginal stability) to extract key dynamic properties and use them as design anchors.

We can see this process here for the following system:

$$P(s) = \frac{1}{s^3 + 9s^2 + 23s + 15} \quad (\text{Chapter 11.45})$$

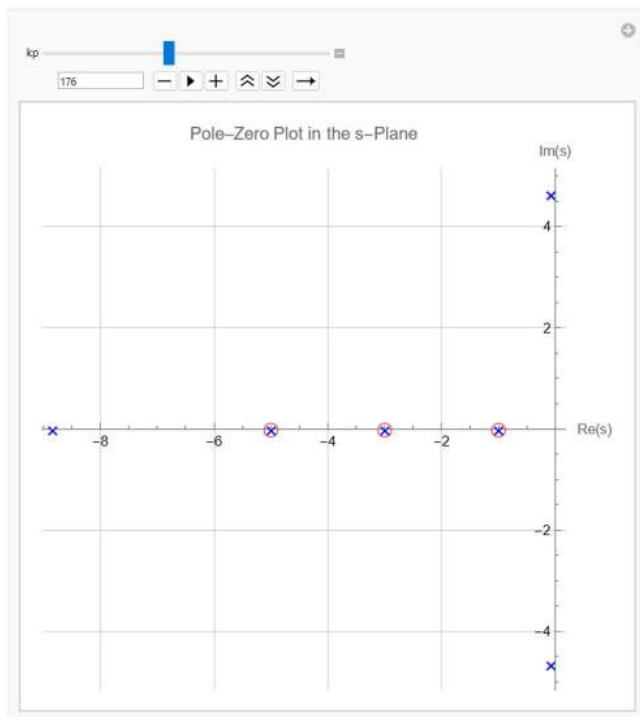
The control for this system will be our full PID controller:

$$C(s) = K_p + K_d s + \frac{K_i}{s} \quad (\text{Chapter 11.46})$$

We will also assume a response function of:

$$R(s) = \frac{1}{s} \quad (\text{Chapter 11.47})$$

Now as mentioned we set $K_d = 0$ and $K_i = 0$. We then have to change and increase K_p and we can see the response where we do not have sustained oscillations, i.e., we are looking for poles that lie on the imaginary plane as seen below.



Plot[y_t /. $K_p \rightarrow 176$, { t , 0, 10}, PlotRange $\rightarrow$ All]

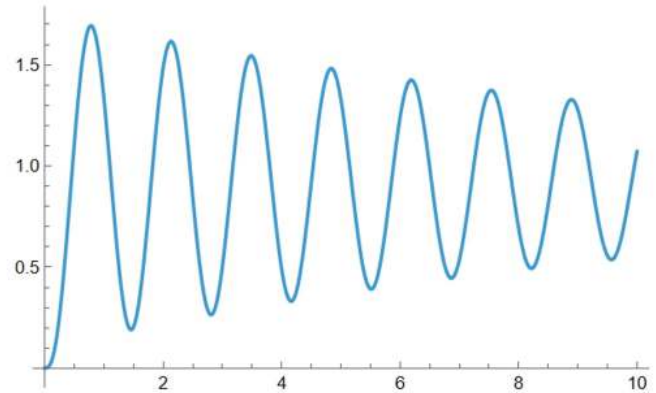
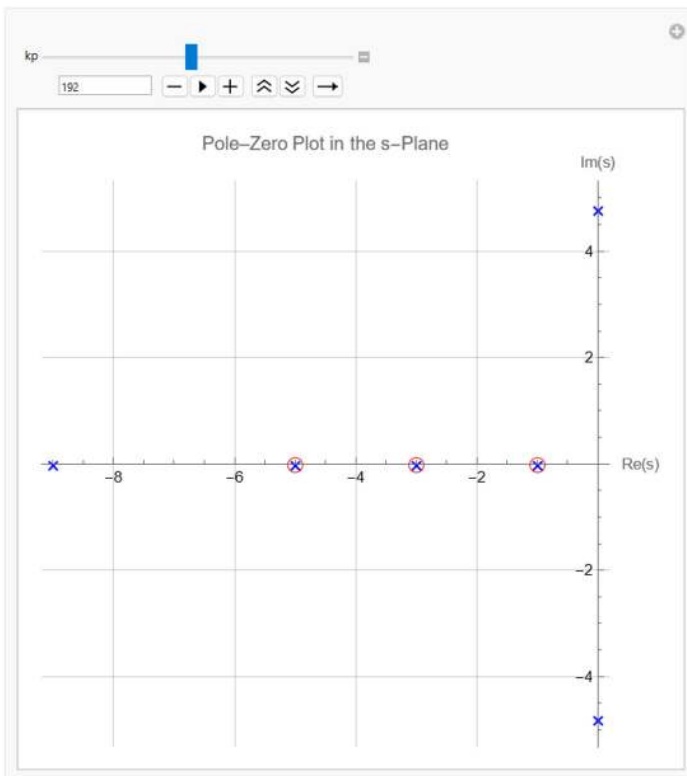


Figure Chapter11.12: Ziegler-Nichols No Sustained Oscillations



Plot[y_t /. $K_p \rightarrow 192$, { t , 0, 10}, PlotRange $\rightarrow$ All]

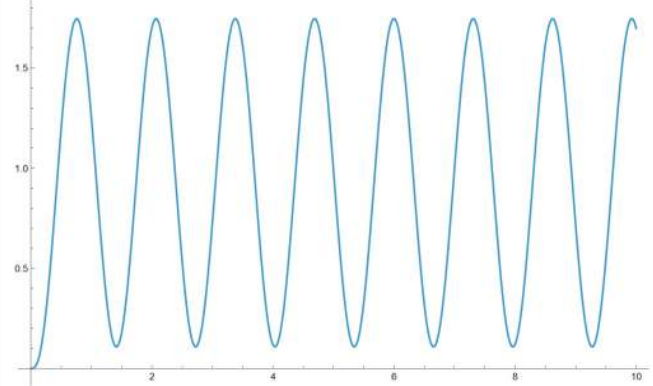


Figure Chapter11.13: Ziegler-Nichols Sustained Oscillation

Once we obtain the sustained oscillations we can calculate the ultimate period and we have found our ultimate gain which is 192. We can then calculate our values for our PID control and then we can see the resulting response to our system and the response looks very good.

PoleZeroPlotSimple[TransferFunctionModel[newGs, s]]

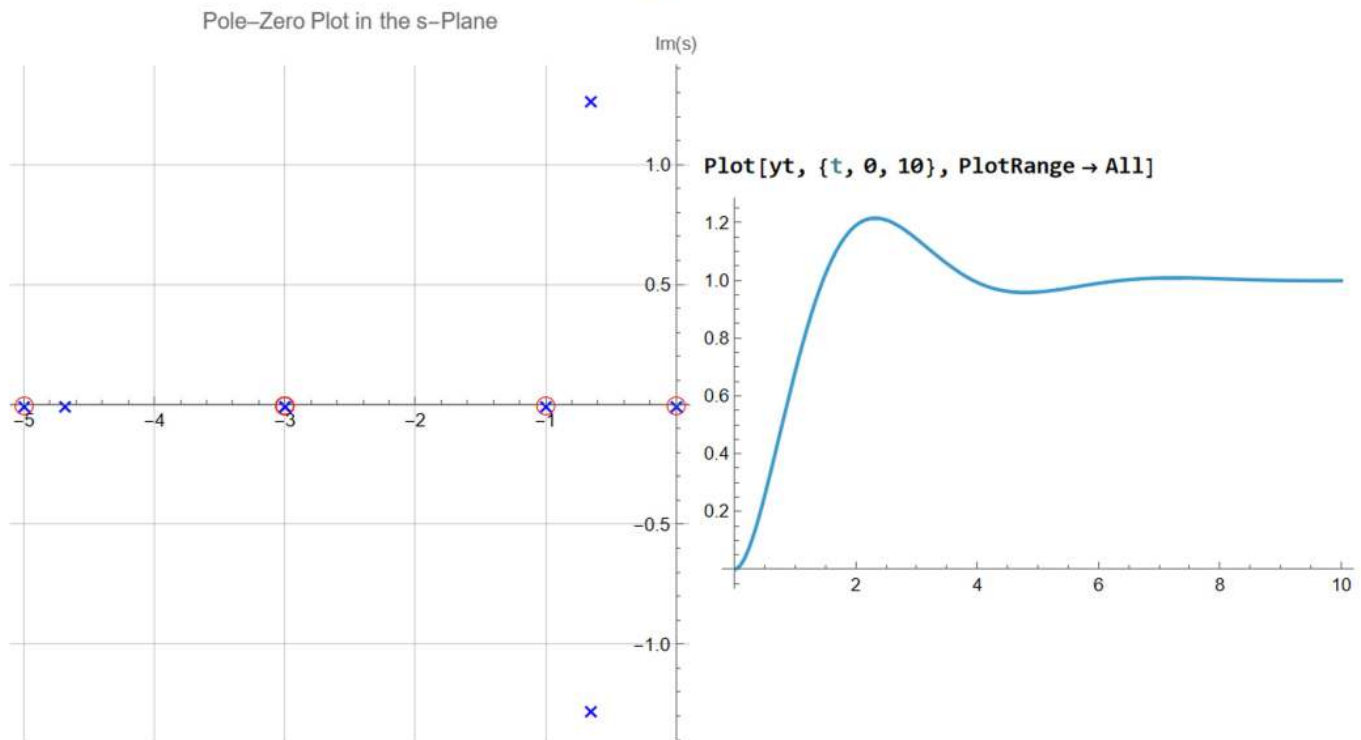


Figure Chapter11.14: Ziegler-Nichols Response

2. Quarter Decay Method (Modified Ziegler-Nichols)

Sometimes it is not possible or desirable to produce sustained oscillations in the system. In such cases, the Quarter Decay method is used:

1. Again, start with $K_i = 0$ and $K_d = 0$.
2. Slowly increase K_p until the output exhibits underdamped oscillations that decay over time.
3. Continue increasing K_p until the ratio of the second peak to the first peak is approximately 1/4. Make sure to look at peak amplitude and subtract the peak height from the steady-state value.
4. At this point, define:

$$K_u = 2K_p, \quad T_u = \text{measured period of oscillation} \quad (\text{Chapter 11.48})$$

5. Use the same Ziegler-Nichols tuning formulas to determine initial values for K_p, K_i, K_d .

This method avoids the need for full marginal stability and instead uses the transient shape to infer system dynamics.

So we can do a similar analysis here and first we have to find the quarter decay value.

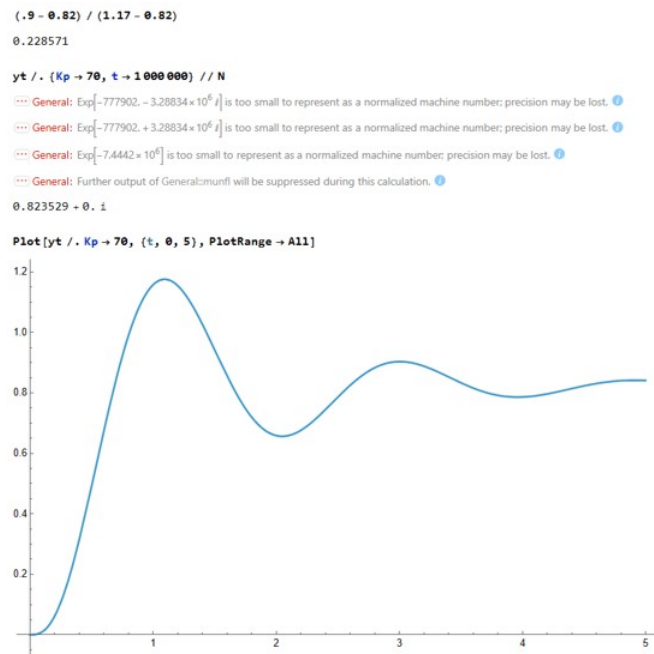


Figure Chapter11.15: Quarter Decay Condition not met

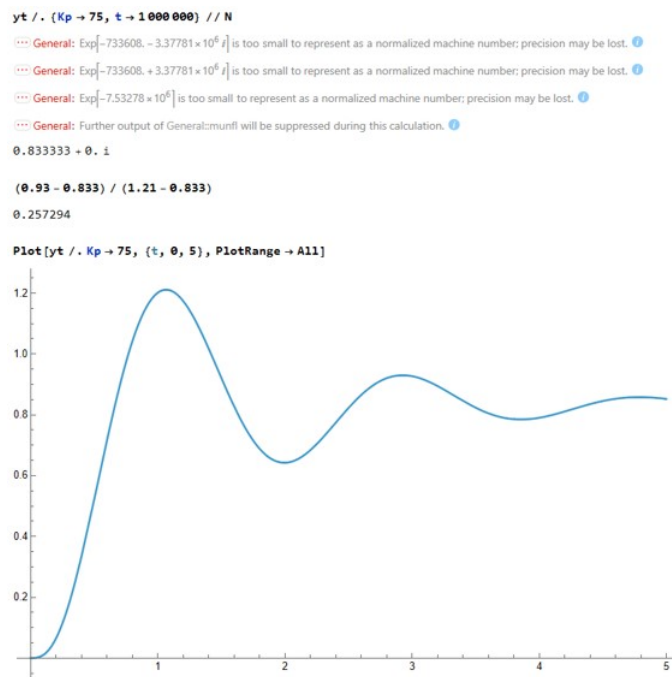


Figure Chapter11.16: Quarter Decay Condition Met

We can also calculate our ultimate gain and our ultimate period. We can then calculate our other values for the PID control and we can see the resulting system response here:

PoleZeroPlotSimple[TransferFunctionModel[newGs, s]]

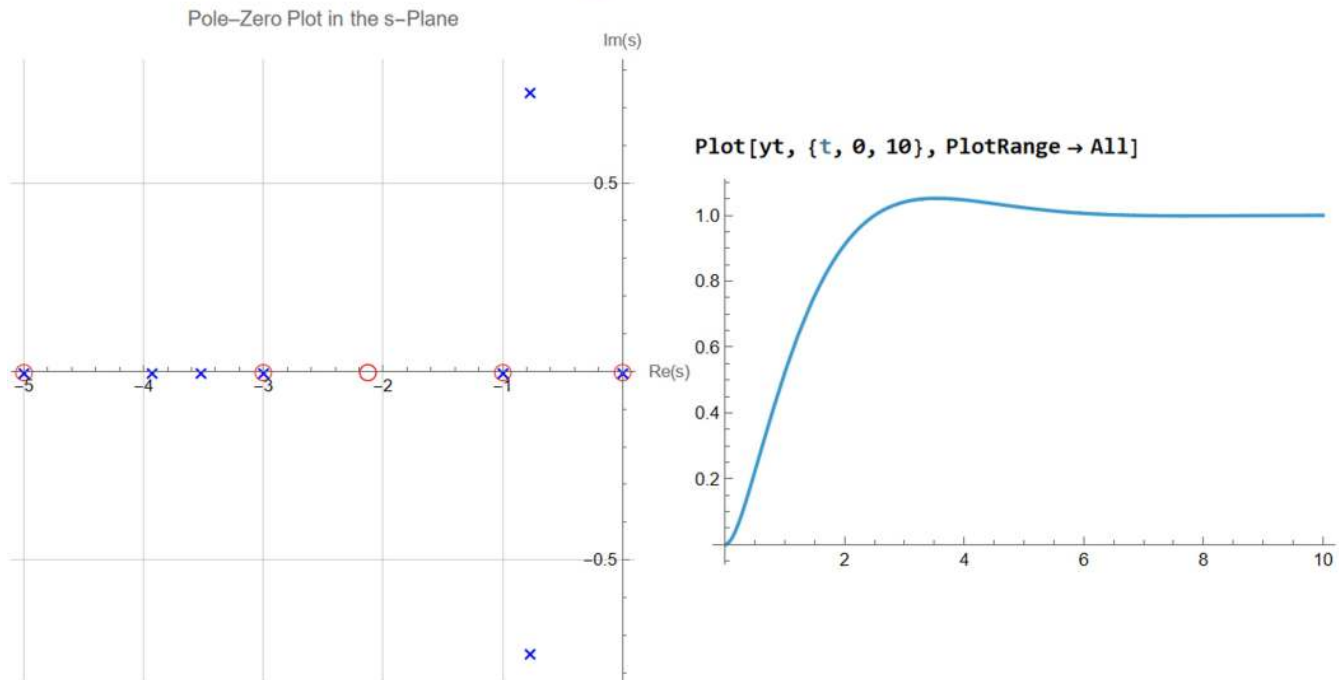


Figure Chapter11.17: System Response Quarter Decay Method

Summary of Empirical PID Tuning

These tuning methods work without knowing the plant model.

They rely on observing how the system responds to increasing K_p under proportional-only control.

Both methods generate estimates of the ultimate gain K_u and period T_u , which define the system's most unstable behavior.

Once baseline gains are chosen, further tuning is often done experimentally or with tools like Simulink.

These techniques are still widely used in HVAC, process control, and embedded control applications where analytical modeling is impractical or unnecessary. PID control remains effective in both model-based and model-free contexts — its power lies in its simplicity, interpretability, and flexibility.

This page titled [Chapter 11: PID Control](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 12: Root Locus

Introduction to Root Locus Method

Root locus is a graphical method for examining how the roots (poles) of the closed-loop transfer function vary as a function of the feedback gain K . It is particularly useful for control design because the closed-loop poles fully determine the dynamic behavior of the system.

Root Locus Form and Characteristic Equation

To apply root locus methods, the characteristic equation must be written in the form:

$$1 + KL(s) = 0 \quad (\text{Chapter 12.1})$$

This form is essential. In some systems, algebraic manipulation is needed to bring the system into this structure. Importantly, root locus is always traced based on this functional form of re-writing the transfer function and poles in this form $L(s)$.

Root Locus Behavior with Proportional Control: Real vs. Complex Poles

Let us consider a simple unity feedback control system with:

$$P(s) = \frac{1}{s+1}, \quad C(s) = K \quad (\text{Chapter 12.2})$$

This is a first-order, stable plant with a pole at $s = -1$. The proportional controller multiplies the plant by a gain K .

Closed-Loop Transfer Function

The closed-loop transfer function is:

$$G(s) = \frac{C(s)P(s)}{1 + C(s)P(s)} = \frac{K}{s+1+K} \quad (\text{Chapter 12.3})$$

This is a first-order system with a single pole located at:

$$\boxed{s = -1 - K} \quad (\text{Chapter 12.4})$$

As $K \rightarrow \infty$, the pole moves toward $s = -\infty$. The system becomes increasingly fast and remains stable for all positive values of K as shown in the figure:

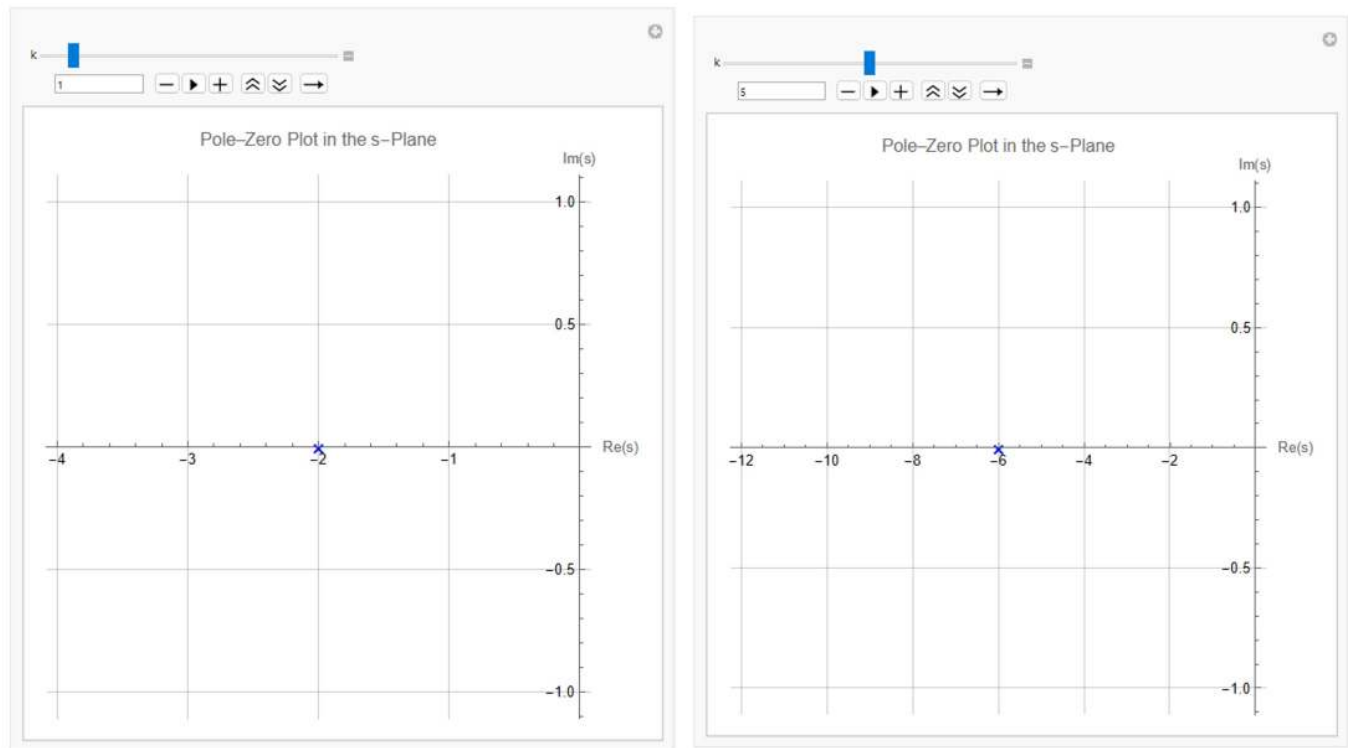


Figure Chapter12.1: Pole Location

We can also visualize the motion of this pole as K changes using Mathematica:

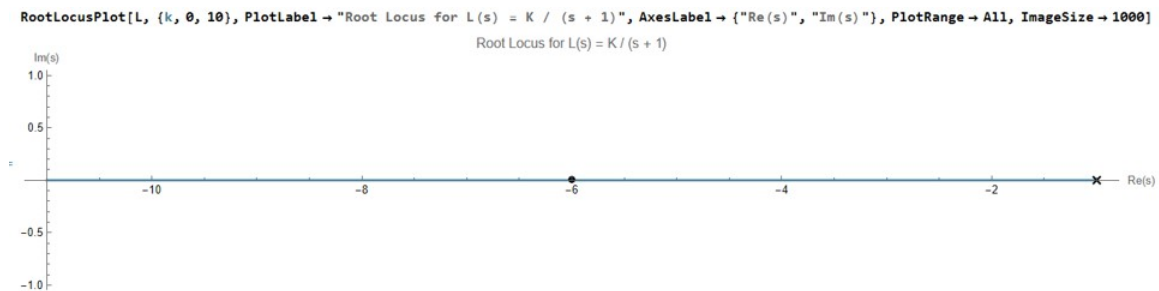


Figure Chapter12.2: Root Locus Plot

Pole Behavior for Real Poles

For this simple case, the root locus consists of a single branch entirely along the real axis. The key behavior is:

- At $K = 0$, the pole is at $s = -1$ (the open-loop pole)
- As K increases, the pole moves leftward on the real axis
- The pole never becomes complex
- As $K \rightarrow \infty$, $s \rightarrow -\infty$

This reflects the typical behavior for systems with a single real pole: increasing gain improves speed and reduces rise/settling time without introducing oscillation.

Now Consider a More Complex System

Suppose instead we have:

$$P(s) = \frac{1}{(s+1)(s+2)} \quad \text{and still } C(s) = K \quad (\text{Chapter 12.5})$$

Then the open-loop transfer function becomes:

$$L(s) = \frac{K}{(s+1)(s+2)} \quad (\text{Chapter 12.6})$$

Characteristic Equation

$$1 + \frac{K}{(s+1)(s+2)} = 0 \Rightarrow (s+1)(s+2) + K = 0$$

$$= 0 \Rightarrow s^2 + 3s + (2+K) = 0 \quad (\text{Chapter 12.7})$$

The closed-loop poles are:

$$s = \frac{-3 \pm \sqrt{1-4K}}{2} \quad (\text{Chapter 12.8})$$

Qualitative Behavior

- When $K < \frac{1}{4}$, the poles are real and distinct
- When $K = \frac{1}{4}$, the poles are repeated (critically damped)
- When $K > \frac{1}{4}$, the poles become complex conjugates
- As $K \rightarrow \infty$, the poles approach:

$$s \approx \frac{-3 \pm j\sqrt{4K-1}}{2} \quad (\text{Chapter 12.9})$$

The real part stays fixed at $-\frac{3}{2}$, but the imaginary part grows.

Thus, for second-order systems, increasing K can cause poles to become complex and oscillatory — but they still remain in the left-half plane because the real part remains negative.

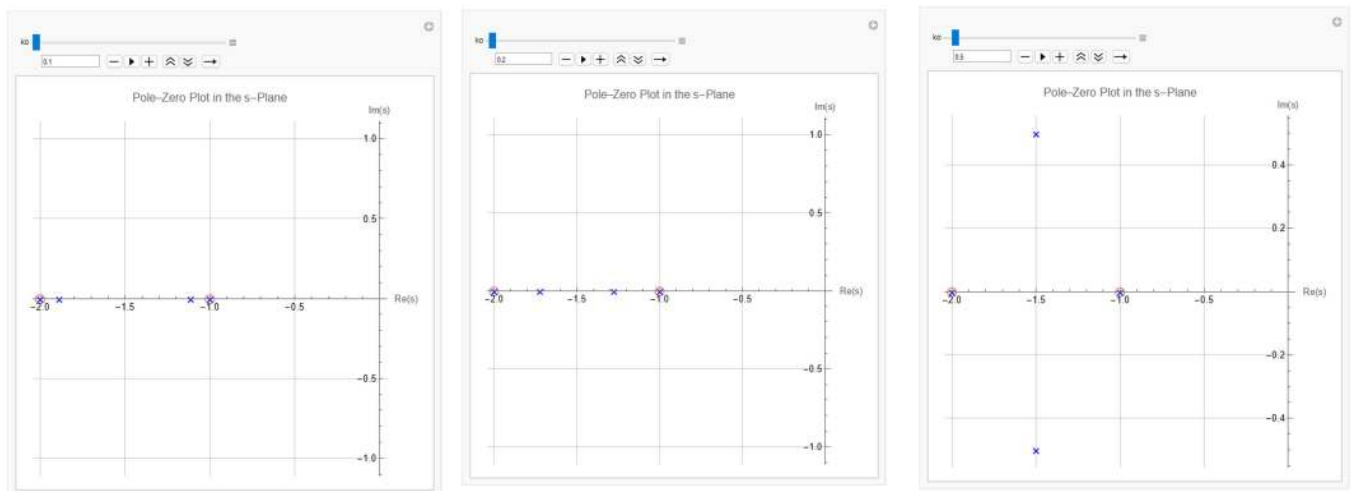


Figure Chapter12.3: Pole Location Motion

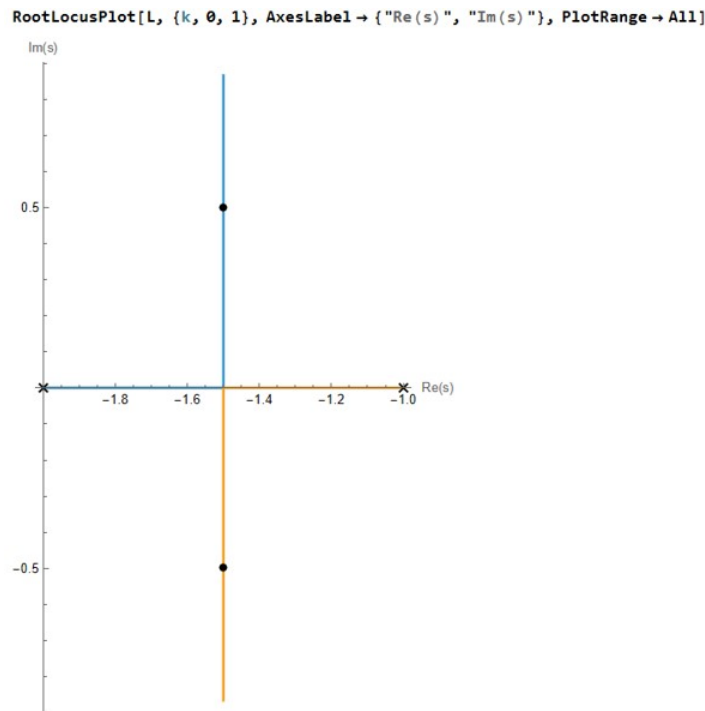


Figure Chapter12.4: Root Locus Plot with Two Poles

Now Consider a Poorly Designed System

Let:

$$P(s) = \frac{s-1}{(s+1)(s+2)}, \quad C(s) = K \quad (\text{Chapter 12.10})$$

This system has:

- A zero at $s = 1$ (in the right-half plane)
- Poles at $s = -1, -2$

Characteristic Equation

$$1 + K \cdot \frac{s-1}{(s+1)(s+2)} = 0 \quad (\text{Chapter 12.11})$$

<h3>Closed-Loop Poles</h3>

$$s = \frac{-(3+K) \pm \sqrt{(3+K)^2 - 4(2-K)}}{2} \quad (\text{Chapter 12.12})$$

As K increases:

- The constant term $2 - K$ becomes negative when $K > 2$
- This implies the product of the roots is negative → one pole must be positive → the system becomes unstable

Key Insight

- For simple systems, increasing K moves poles to $-\infty$
- For higher-order systems with RHP zeros, increasing K can destabilize the system
- Root locus allows us to track how poles move and determine safe values of K

Summary

- In first-order systems like $P(s) = \frac{1}{s+1}$, increasing K moves the pole leftward — faster and more stable
- In second-order systems, poles can become complex as K increases — leading to oscillations

- If the plant has right-half-plane zeros, increasing K can push poles into the RHP — causing instability
- Root locus gives a graphical way to understand this behavior and guide controller design

`RootLocusPlot[L, {K, 0, 4}, AxesLabel -> {"Re(s)", "Im(s)"}, PlotRange -> All, ImageSize -> 1000]`

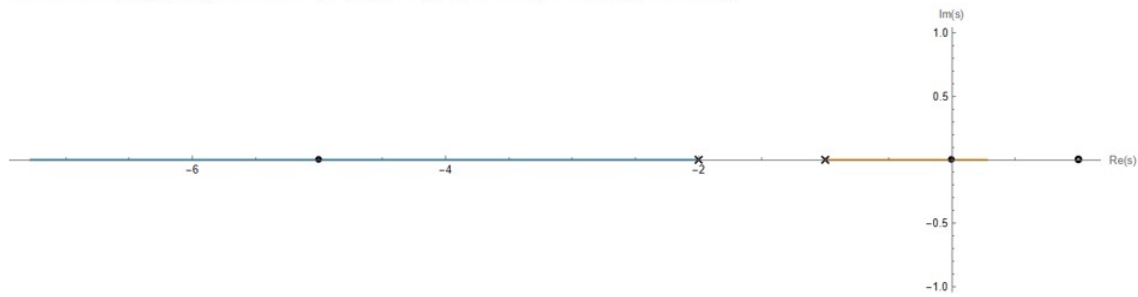


Figure Chapter12.5: Root Locus Plot becoming unstable

Root locus plots are useful because they allow us to:

- Predict closed-loop pole locations without solving for roots numerically
- Design controllers by adding poles/zeros
- Assess how gain K influences stability, damping, and response time

Fundamental Root Locus Rules

Rule 0: Symmetry About the Real Axis

The root locus is always symmetric with respect to the real axis. This is because complex poles always appear in conjugate pairs, so their paths mirror each other about the real axis.

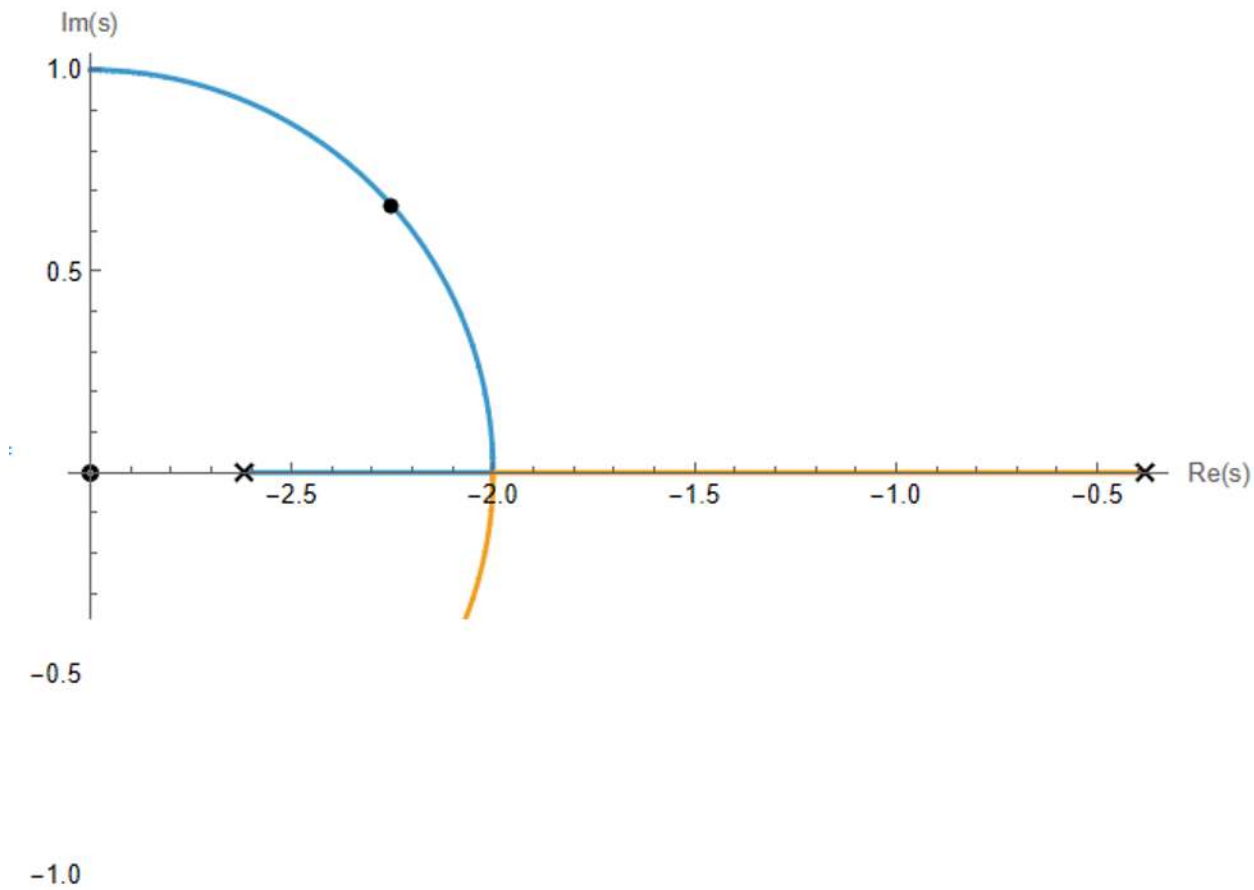


Figure Chapter12.6: Rule 0 Violated

Rule 1: Number of Branches and Endpoints

Let $L(s) = \frac{N(s)}{D(s)}$ with:

- n = number of poles of $L(s)$
- m = number of zeros of $L(s)$

Then:

- The root locus has n branches, one for each pole
- m branches terminate at the finite zeros of $L(s)$
- $n - m$ branches go to infinity, approaching asymptotes

Note: $n \geq m$ must hold to preserve causality.

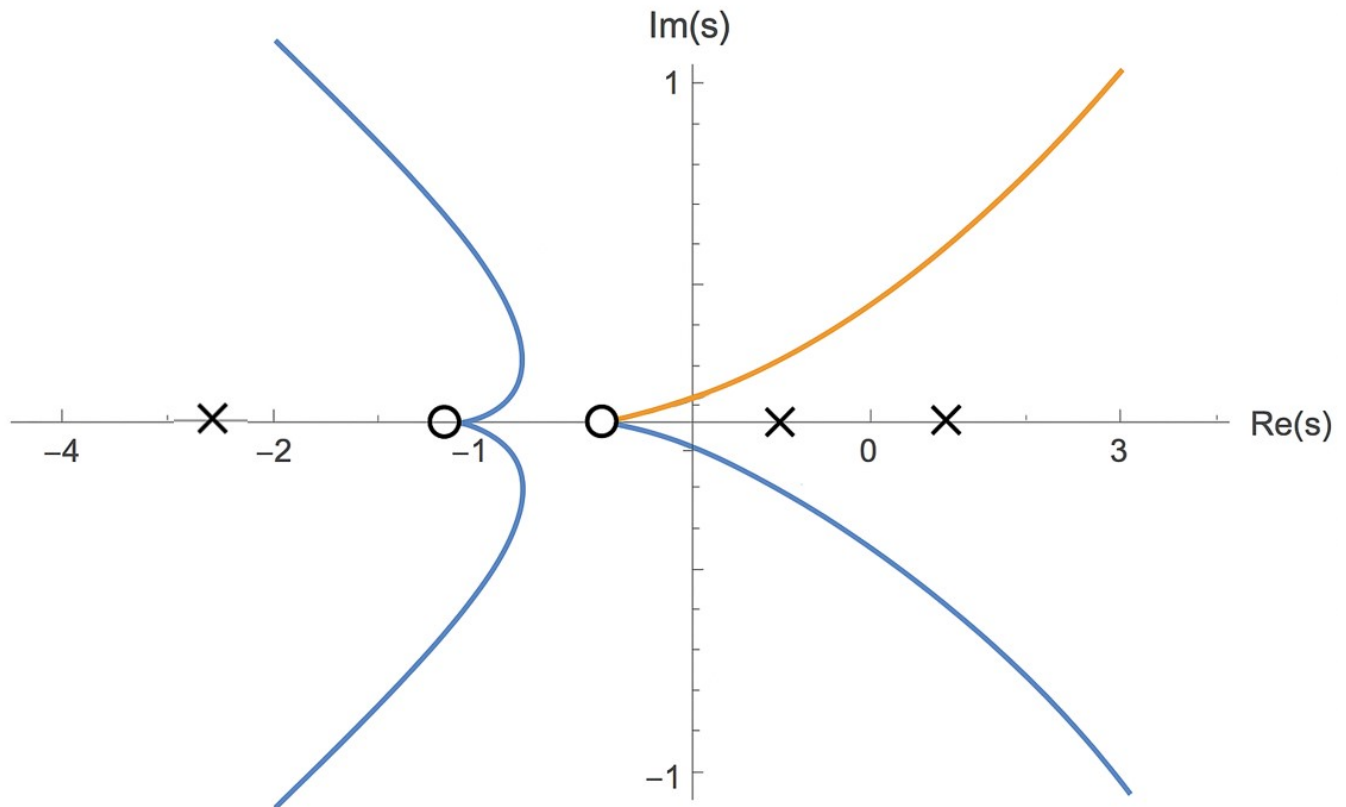


Figure Chapter12.7: Rule 1 Violated

Rule 2: Real-Axis Segments

A point s on the real axis belongs to the root locus if the total number of poles and zeros to the right of s on the real axis is odd.

This rule can be applied visually using a vertical line test.

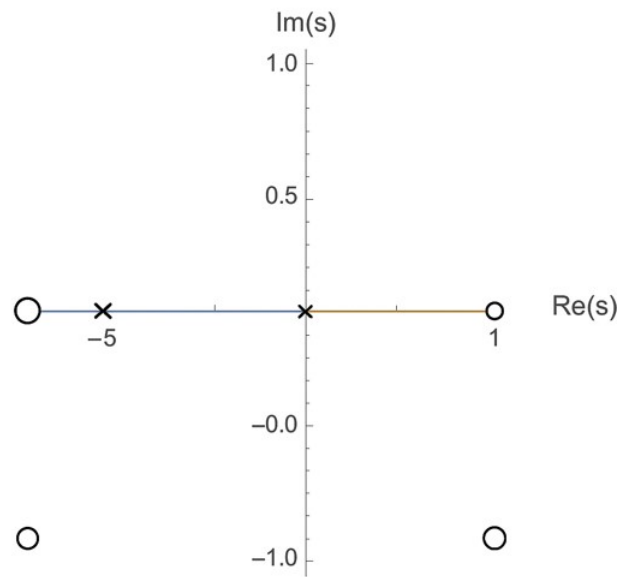


Figure Chapter12.8: Rule 2 Violated

Rule 3: Asymptotes of Branches Going to Infinity

For $n - m$ branches that go to infinity, the asymptotes intersect the real axis at:

$$\alpha = \frac{\sum (\text{poles}) - \sum (\text{zeros})}{n - m} \quad (\text{Chapter 12.13})$$

and depart at angles:

$$\theta_k = \frac{(2k + 1)\pi}{n - m}, \quad k = 0, 1, 2, \dots \quad (\text{Chapter 12.14})$$

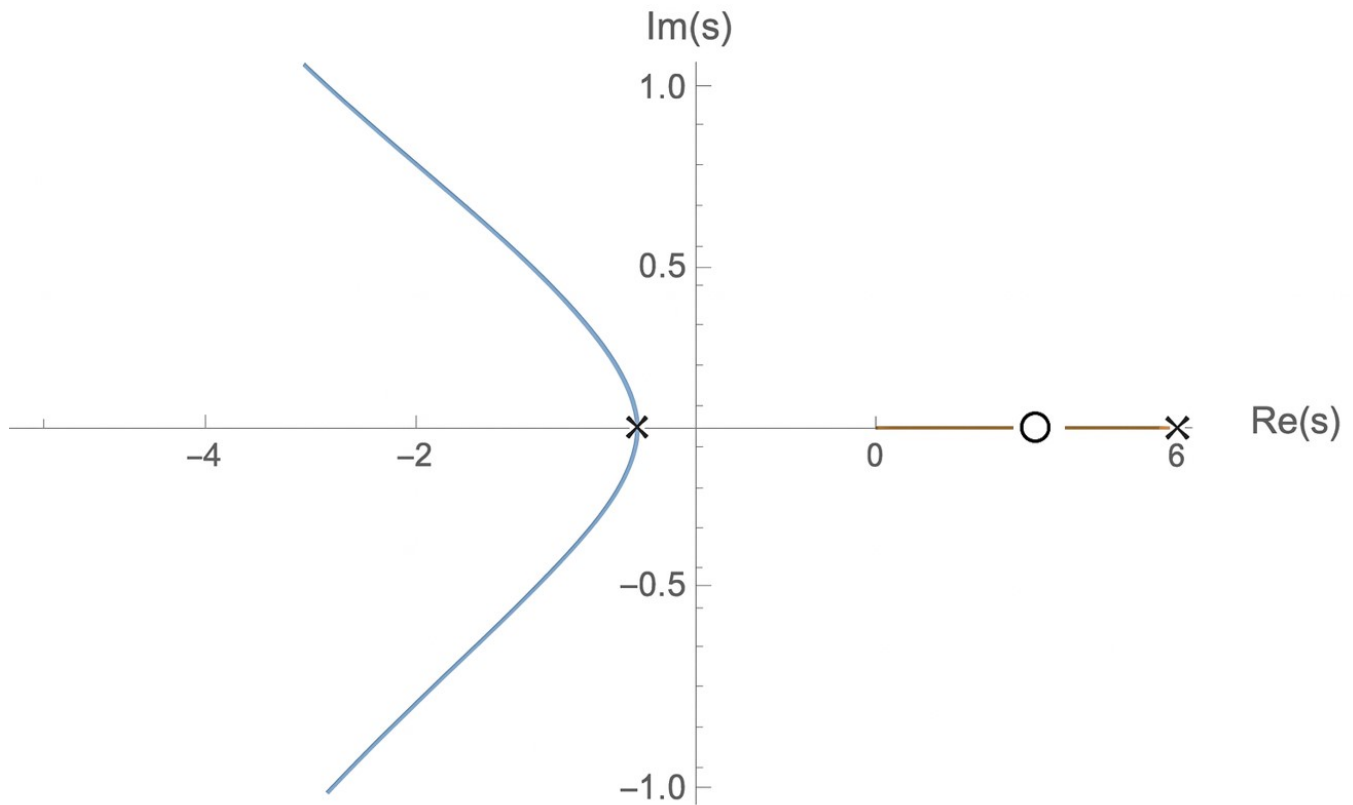


Figure Chapter12.9: Rule 3 Violated

Rule 4: Angle of Departure and Arrival

For complex poles and zeros not on the real axis, the angle of departure from a pole and angle of arrival at a zero can be calculated using:

$$\phi_{\text{departure}} = 180^\circ - \sum \angle(\text{vectors from other poles and zeros}) \quad (\text{Chapter 12.15})$$

This rule is tedious and typically used only for precision analysis but can be necessary to perform.

Root Locus Geometric Formulas with Worked Example

We now apply the root locus geometric rules to a specific example, including the asymptote behavior and the angle of departure from a complex pole. These are essential tools for understanding pole movement and controller design.

Given System

Consider the open-loop transfer function:

$$L(s) = \frac{s + 4}{(s + 2)(s^2 + 4s + 20)} \quad (\text{Chapter 12.16})$$

This system has:

- One finite zero at $z_1 = -4$
- Three poles:
 - One real pole at $p_1 = -2$
 - Two complex poles at $p_{2,3} = -2 \pm j4$

Step 1: Number of Branches and Asymptotes

- Number of poles: $n = 3$
- Number of zeros: $m = 1$
- Number of asymptotes: $n - m = 2$

Centroid of Asymptotes:

$$\alpha = \frac{p_1 + p_2 + p_3 - z_1}{n - m} = \frac{-2 + (-2 + j4) + (-2 - j4) - (-4)}{2} \quad (\text{Chapter 12.17})$$
$$= \frac{-6 + 4}{2} = -1$$

Asymptote Angles:

$$\theta_k = \frac{(2k+1)\pi}{n-m}, \quad k = 0, 1 \Rightarrow \theta_0 = 90^\circ, \quad \theta_1 = 270^\circ \quad (\text{Chapter 12.18})$$

These asymptotes indicate that two branches go to infinity vertically, one upward and one downward from $s = -1$.

Step 2: Angle of Departure from a Complex Pole

Compute the angle of departure from the pole at $p_2 = -2 + j4$:

$$\phi_d = \pi + \sum_j \angle(p - z_j) - \sum_{i \neq p} \angle(p - p_i) \quad (\text{Chapter 12.19})$$

Where:

- $p = -2 + j4$
- $z_1 = -4$
- $p_1 = -2, p_3 = -2 - j4$

Compute angles:

$$\angle(p - z_1) = \arg(2 + j4) = \tan^{-1}\left(\frac{4}{2}\right) = 63.4^\circ \quad (\text{Chapter 12.20})$$

$$\angle(p - p_1) = \arg(j4) = 90^\circ, \quad \angle(p - p_3) = \arg(j8) = 90^\circ \quad (\text{Chapter 12.21})$$

Final calculation:

$$\phi_d = 180^\circ + 63.4^\circ - (90^\circ + 90^\circ) = 180 + 63.4 - 180 = \boxed{63.4^\circ} \quad (\text{Chapter 12.22})$$

So the root locus departs from the complex pole at $-2 + j4$ at an angle of approximately 63.4° .

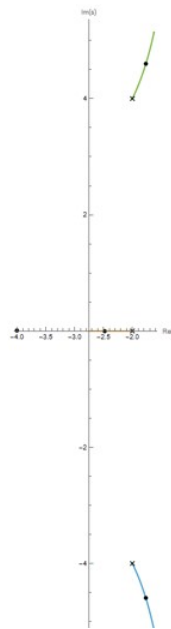


Figure Chapter12.10: Root Locus Plot

Summary of This Example

- The system has 3 poles and 1 zero
- The root locus has 2 asymptotes at $\pm 90^\circ$, centered at $\alpha = -1$
- The root locus leaves the complex pole at $-2 + j4$ at an angle of 63.4°
- Symmetry rule guarantees the conjugate pole departs at 63.4° below the real axis

Rule 5: Breakaway and Break-in Points

When two or more branches of the root locus come together or separate on the real axis, they do so at breakaway (diverging) or break-in (converging) points. These points are always located on the real axis and must lie within segments of the root locus (as determined by Rule 2).

To calculate them, write:

$$L(s) = \frac{n(s)}{d(s)} \quad (\text{Chapter 12.23})$$

Then apply the quotient rule to the open-loop transfer function (excluding the gain k), and solve:

$$\frac{d}{ds} \left(\frac{n(s)}{d(s)} \right) = 0 \quad (\text{Chapter 12.24})$$

That is, compute:

$$\frac{n'(s)d(s) - n(s)d'(s)}{d(s)^2} = 0 \quad (\text{Chapter 12.25})$$

This simplifies to:

$$n'(s)d(s) - n(s)d'(s) = 0 \quad (\text{Chapter 12.26})$$

Only real-valued roots of this equation that lie on valid segments of the real axis are valid breakaway or break-in points.

Root Locus and Control Design

Root locus is a powerful graphical method to design controllers that shift the closed-loop poles into desired locations in the complex s -plane. By adding poles or zeros to the open-loop transfer function, we can shape the root locus and place poles to meet time-domain or frequency-domain specifications.

Worked Example

Let the open-loop transfer function be:

$$L(s) = \frac{k(s+3)}{s^2+3s+2} \quad (\text{Chapter 12.27})$$

Zeros: $s = -3$

Poles: $s = -1, -2$

This system has two poles and one zero, so one branch will go to the zero and one will go to infinity.

Step 1: Use the Quotient Rule

Let:

$$n(s) = s + 3, \quad d(s) = s^2 + 3s + 2 \quad (\text{Chapter 12.28})$$

Then:

$$n'(s)d(s) - n(s)d'(s) = 1 \cdot (s^2 + 3s + 2) - (s + 3)(2s + 3) \quad (\text{Chapter 12.29})$$

Expand both terms:

$$(s^2 + 3s + 2) - (2s^2 + 9s + 9) = -s^2 - 6s - 7 \quad (\text{Chapter 12.30})$$

Solve:

$$-s^2 - 6s - 7 = 0 \Rightarrow s^2 + 6s + 7 = 0 \quad (\text{Chapter 12.31})$$

$$s = \frac{-6 \pm \sqrt{36 - 28}}{2} = \frac{-6 \pm \sqrt{8}}{2} = \frac{-6 \pm 2\sqrt{2}}{2} = -3 \pm \sqrt{2} \quad (\text{Chapter 12.32})$$

$$\Rightarrow s \approx -4.414, \quad s \approx -1.586 \quad (\text{Chapter 12.33})$$

Step 2: Check Real-Axis Segments (Rule 2)

From Rule 2: A point on the real axis is part of the root locus if the number of real poles and zeros to the right of that point is odd.

Real poles: $s = -1, -2$

Real zero: $s = -3$

Segments of the real axis:

- $(-\infty, -3)$: 3 elements to the right $\rightarrow$ on locus
- $(-3, -2)$: 2 to the right $\rightarrow$ not on locus
- $(-2, -1)$: 1 to the right $\rightarrow$ on locus
- $(-1, \infty)$: 0 to the right $\rightarrow$ not on locus

So the breakaway and break-in points:

$$\boxed{-4.414 \text{ and } -1.586} \quad (\text{Chapter 12.34})$$

are both valid — they lie on segments of the real axis where the root locus exists.

Step 3: Asymptote Calculation

Number of poles: $n = 2$, number of zeros: $m = 1$

So there is one branch going to infinity.

Asymptote:

$$\alpha = \frac{\sum \text{poles} - \sum \text{zeros}}{n - m} = \frac{-1 + (-2) - (-3)}{1} = 0 \quad (\text{Chapter 12.35})$$

Angle:

$$\theta_k = \frac{(2k+1)\pi}{n-m} = \pi \Rightarrow 180^\circ \quad (\text{Chapter 12.36})$$

Conclusion

- Root locus starts at poles $s = -1$ and $s = -2$
- One branch terminates at the zero at $s = -3$
- One branch goes to infinity along the real axis at angle 180°
- Breakaway at $s \approx -1.586$, break-in at $s \approx -4.414$

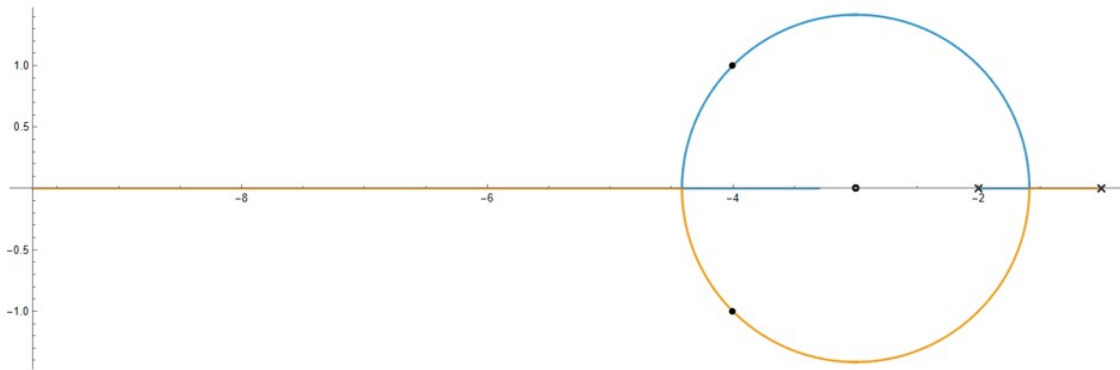


Figure Chapter12.11: Rule 5

Design Objective and Desired Pole Region

Design objective: Place the poles such that:

- The system is stable (poles in LHP)
- Damping ratio $\zeta \geq \zeta_{\min}$
- Natural frequency $\omega_n \leq \omega_{\max}$
- Settling time $t_s \leq t_{s,\max}$

This corresponds to selecting a desired pole region, often a wedge defined by lines of constant damping and a vertical boundary for maximum settling time.

We then check whether the root locus enters this region for any gain K . If it does, then proportional control is sufficient. If not, we must redesign the controller to shift the root locus.

Controller Design via Root Locus

Step 1: Try Proportional Control

$$C(s) = K \quad (\text{Chapter 12.37})$$

If the root locus of $KP(s)$ does not intersect the desired region for any value of K , then P control is insufficient.

Let us consider the open-loop transfer function:

$$L(s) = \frac{1}{s(s+1)} \quad (\text{Chapter 12.38})$$

This is a Type 1, second-order system with:

- Poles at $s = 0$ and $s = -1$
- No finite zeros
- System order: $n = 2$, number of zeros $m = 0$
- $n - m = 2$ branches of the root locus

Closed-loop characteristic equation (with P controller):

$$1 + K \cdot \frac{1}{s(s+1)} = 0 \Rightarrow s(s+1) + K = 0 \Rightarrow s^2 + s + K = 0 \quad (\text{Chapter 12.39})$$

Closed-loop poles:

$$s = \frac{-1 \pm \sqrt{1 - 4K}}{2} \quad (\text{Chapter 12.40})$$

Behavior of the poles as K varies:

- For $K < \frac{1}{4}$, poles are real and negative
- For $K = \frac{1}{4}$, repeated real pole (critical damping)
- For $K > \frac{1}{4}$, poles become complex conjugates
- Real part of poles approaches $-\frac{1}{2}$ as $K \rightarrow \infty$

Desired Closed-Loop Pole Region

The design specifications require:

- Settling time $t_s < 2$ seconds
- Maximum overshoot $< 10\%$

These constraints define the following pole region in the complex s -plane:

- Minimum damping ratio: $\zeta > 0.528$
- Minimum exponential decay rate: $\sigma > 2.3$
- Minimum natural frequency: $\omega_n > 4.35$

This defines a shaded “safe” region in the left-half plane where:

- The poles lie to the left of $\text{Re}(s) = -2.3$
- The poles lie within a cone defined by $\theta = \cos^{-1}(0.528) \approx 58.1^\circ$
- The poles lie outside a circle of radius $\omega_n = 4.35$

Compare with Root Locus of P-Controlled System

$$L(s) = \frac{K}{s(s+1)} \Rightarrow \text{Characteristic equation: } s^2 + s + K = 0 \quad (\text{Chapter 12.41})$$

This gives the closed-loop poles:

$$s = \frac{-1 \pm \sqrt{1 - 4K}}{2} \quad (\text{Chapter 12.42})$$

As $K \rightarrow \infty$, the real part of the poles approaches:

$$\text{Re}(s) \rightarrow -\frac{1}{2} \quad (\text{Chapter 12.43})$$

This does not satisfy the requirement $\text{Re}(s) < -2.3$. The root locus stays too close to the imaginary axis, and overshoot and settling time requirements are violated.

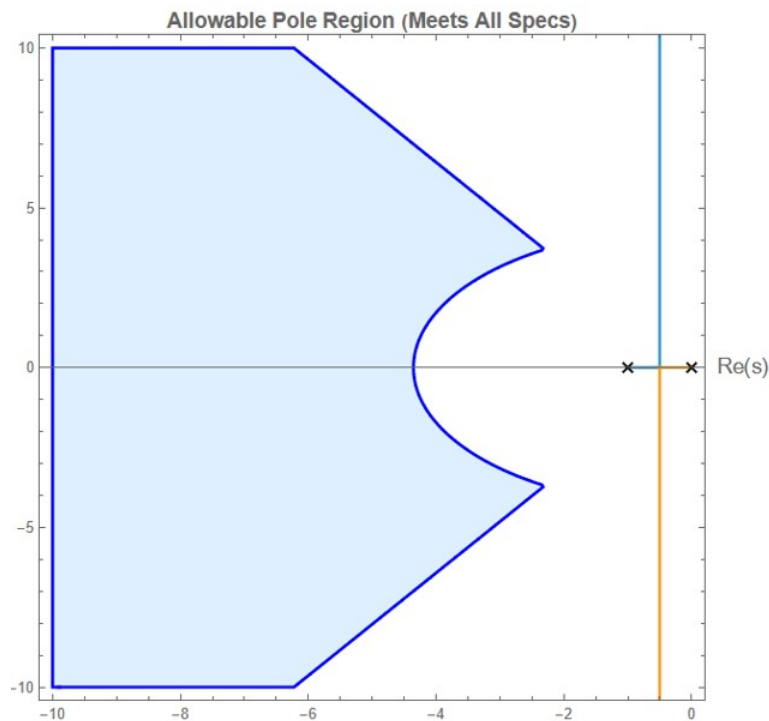


Figure Chapter12.12: Controller doesn't meet requirements

Conclusion: P Control is Insufficient

Proportional control alone cannot enter the desired region.

(Chapter 12.44)

The closed-loop poles never reach the region defined by $\zeta > 0.528$, $\sigma > 2.3$, and $\omega_n > 4.35$. Therefore, additional control (e.g., K_d , K_i) is needed to meet specs.

Therefore, we must increase the complexity of the controller.

In the next step, we can add a zero via lead control:

$$C(s) = K \cdot (s + z)$$

(Chapter 12.45)

This will bend the root locus toward the desired pole region by pulling branches leftward, enabling faster and more stable response.

Step 2: Use Lead Compensation

To move the root locus into the desired region, add a zero via lead compensation:

$$C(s) = K \cdot \frac{s + z}{s + p}, \quad z < p$$

(Chapter 12.46)

This zero attracts the locus toward the desired region. For effective placement:

- Place the zero near the desired pole location
- Position the pole further left to ensure lead (phase advancement)

Controller Design via Root Locus: Lead Compensation

When proportional control is insufficient to meet system performance requirements (e.g., damping, settling time, or overshoot), we introduce a lead compensator. The lead controller introduces a zero that attracts the root locus into the desired region of the complex plane.

Example System with Lead Compensation

Suppose we have the open-loop transfer function:

$$L(s) = \frac{k(s + 1.5)}{s(s + 1)(s + 8)(s + 6)} \quad (\text{Chapter 12.47})$$

This corresponds to a unity-feedback system with:

- One compensator zero at $s = -1.5$
- Four finite poles:
- $s = 0$ (integrator)
- $s = -1$
- $s = -6$ (lead compensator pole)
- $s = -8$
- System order $n = 4$, number of zeros $m = 1$
- $n - m = 3$ root locus branches go to infinity

Root Locus Geometry

Centroid of asymptotes:

$$\alpha = \frac{0 + (-1) + (-6) + (-8) - (-1.5)}{3} = \frac{-13.5}{3} = -4.5 \quad (\text{Chapter 12.48})$$

Asymptote angles:

$$\theta_k = \frac{(2k + 1)\pi}{n - m}, \quad k = 0, 1, 2 \Rightarrow \theta = 60^\circ, 180^\circ, 300^\circ \quad (\text{Chapter 12.49})$$

The branches to infinity depart from $s = -4.5$ at those angles.

Controller Purpose and Interpretation

The compensator adds a zero at $s = -1.5$, which attracts the root locus into the desired pole region, while the pole at $s = -6$ keeps the phase lead localized.

This lead compensator bends the locus so that for some $k > 0$, we can achieve:

- Damping ratio $\zeta > 0.6$
- Settling time $t_s < 1.5$ seconds

The corresponding pole region has:

$$\text{Re}(s) < -2, \quad \angle(\theta) > \cos^{-1}(0.6) \approx 53.1^\circ \quad (\text{Chapter 12.50})$$

Result:

By applying this compensator, we confirmed via root locus and pole calculations that the system achieves poles such as:

$$s = -2.67 \pm 4.47j \quad (\text{Chapter 12.51})$$

Which gives:

- $\zeta = \cos(\theta) \approx 0.51$
- $\omega_n = \sqrt{(-2.67)^2 + (4.47)^2} \approx 5.22$
- $t_s \approx \frac{4.6}{\sigma} = \frac{4.6}{2.67} \approx 1.72$ seconds

So this compensator brings the poles close to the desired region. With further tuning, we can fully satisfy the damping and settling time requirements.

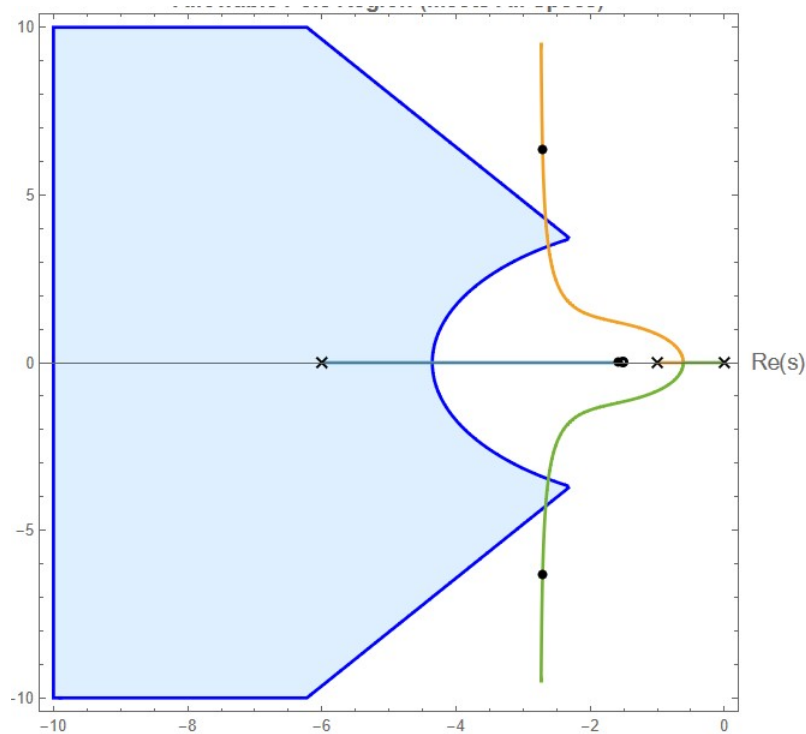


Figure Chapter12.13: Lead Controller Root Locus

Conclusion:

This example shows how lead compensation:

- Attracts the root locus into more desirable locations in the complex plane
- Enables design for speed and damping not achievable by P control alone
- Can be implemented by adding a zero near the origin and a pole further left

Step 3: Use Integral Control to Improve Steady-State

If tracking or steady-state error is unsatisfactory, add an integrator:

$$C(s) = \frac{K(s+z)}{s} \quad (\text{Chapter 12.52})$$

This makes the system Type 1 and improves steady-state accuracy, but introduces a pole at the origin and a new third-order dynamic, which must be carefully analyzed (e.g., using Routh-Hurwitz).

Controller Design via Root Locus: Integral Control

When proportional or lead compensation is not sufficient to meet steady-state performance (e.g., eliminating steady-state error to a step input), we can introduce integral control. Integral control improves tracking and regulation by increasing system type — but introduces a pole at the origin and increases system order, requiring caution in stability analysis.

Example System with Integral Control

Let the plant be:

$$P(s) = \frac{s+2}{s^2+s+3} \quad (\text{Chapter 12.53})$$

We define the controller as pure integral control:

$$C(s) = \frac{K}{s} \quad (\text{Chapter 12.54})$$

So the open-loop transfer function becomes:

$$L(s) = C(s)P(s) = \frac{K(s+2)}{s(s^2+s+3)} \quad (\text{Chapter 12.55})$$

This system has:

- Poles at:
- $s = 0$ (integrator)
- $s = -\frac{1}{2} \pm j\frac{\sqrt{11}}{2}$
- One zero at $s = -2$
- System order $n = 3$, number of finite zeros $m = 1$
- Number of asymptotes: $n - m = 2$

Asymptotes and Geometry

Centroid of asymptotes:

$$\alpha = \frac{0 + (-\frac{1}{2} + j\frac{\sqrt{11}}{2}) + (-\frac{1}{2} - j\frac{\sqrt{11}}{2}) - (-2)}{2} = \frac{-1+2}{2} = \boxed{\frac{1}{2}} \quad (\text{Chapter 12.56})$$

Asymptote angles:

$$\theta_k = \frac{(2k+1)\pi}{2}, \quad k = 0, 1 \Rightarrow \theta_0 = 90^\circ, \quad \theta_1 = 270^\circ \quad (\text{Chapter 12.57})$$

So the root locus will have branches going vertically to infinity from $\alpha = 0.5$.

Closed-Loop Characteristic Equation

We form the unity-feedback closed-loop system:

$$\begin{aligned} 1 + L(s) = 0 &\Rightarrow s(s^2 + s + 3) + K(s + 2) = 0 \\ &\Rightarrow s^3 + s^2 + 3s + K(s + 2) = 0 \Rightarrow \boxed{s^3 + s^2 + (3 + K)s + 2K = 0} \end{aligned} \quad (\text{Chapter 12.58})$$

This is a cubic polynomial. To determine stability, we apply the Routh-Hurwitz criterion.

Routh Table

Let:

$$a_3 = 1, \quad a_2 = 1, \quad a_1 = 3 + K, \quad a_0 = 2K \quad (\text{Chapter 12.59})$$

Stability Conditions (All First Column Elements > 0)

- $1 > 0$
- $1 > 0$
- $3 - K > 0 \Rightarrow K < 3$
- $2K > 0 \Rightarrow K > 0$

So for stability:

$$\boxed{0 < K < 3} \quad (\text{Chapter 12.60})$$

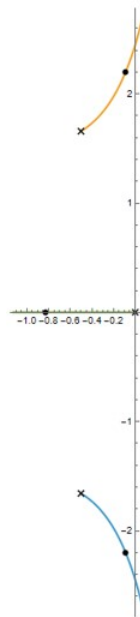


Figure Chapter12.14: Integral Controller Root Locus

Interpretation

- Adding the integrator improved steady-state tracking (Type 1 system)
- The new pole at $s = 0$ increased the system order to 3
- Stability is limited to a bounded range of K
- Beyond $K = 3$, the system becomes unstable due to a sign change in the Routh table

Why the Integrator Improves Tracking

When we add an integrator via $C(s) = \frac{K}{s}$, the open-loop system becomes Type 1:

$$L(s) = \frac{K}{s(s+1)(s+2)} \quad (\text{Chapter 12.61})$$

This guarantees:

- Zero steady-state error to step inputs:

$$e_{ss} = \lim_{s \rightarrow 0} \frac{1}{1 + L(s)} = 0 \quad (\text{Chapter 12.62})$$

- Perfect tracking of step references
- Automatic error correction over time due to integral accumulation

However, adding the integrator increases system order and can introduce instability if K is too large. Stability must be ensured via Routh-Hurwitz or root locus analysis.

Critical Gain for Stability

Suppose $s = j\omega$ lies on the imaginary axis — this is the boundary of marginal stability. Then we can find the critical gain K_{crit} such that:

$$1 + K_{\text{crit}}L(j\omega) = 0 \Rightarrow K_{\text{crit}} = \frac{1}{|L(j\omega)|} \quad (\text{Chapter 12.63})$$

This tells us the maximum gain before the system becomes unstable.

Example:

Let the critical frequency be ω_{crit} , found by solving:

$$\angle L(j\omega) = \pm 180^\circ \quad (\text{Chapter 12.64})$$

Then compute $|L(j\omega_{\text{crit}})|$, and solve:

$$K_{\text{crit}} = \frac{1}{|L(j\omega_{\text{crit}})|} \quad (\text{Chapter 12.65})$$

This can be visualized and calculated on a Bode plot as well.

Design Rule:

For stability, choose $K < K_{\text{crit}}$

Summary

- Root locus traces how closed-loop poles move as K varies
- Provides powerful visual tool for controller design
- Stability, damping, and response time all depend on pole location
- If root locus does not enter desired region, increase controller complexity (lead, lag, PID)
- Critical gain analysis gives safety margins for proportional control

This page titled [Chapter 12: Root Locus](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 13: Frequency Response, Nyquist Stability Criterion, and Bode Plots

Frequency Response and the Nyquist Stability Criterion

Frequency response is a powerful method to analyze and design control systems based on how they react to sinusoidal inputs. Unlike time-domain methods, which involve solving differential equations, frequency-domain analysis characterizes system behavior using magnitude and phase across frequencies. This method underpins many of the most robust and scalable modern control strategies.

Sinusoidal Steady-State Behavior

For any linear time-invariant (LTI) system, a sinusoidal input of frequency ω will produce a sinusoidal output of the same frequency, but scaled in amplitude and phase-shifted. Mathematically:

$$u(t) = A \cos(\omega t) \Rightarrow y(t) = |G(j\omega)| A \cos(\omega t + \angle G(j\omega)) \quad (\text{Chapter 13.1})$$

This output structure allows us to define the frequency response function:

$$G(j\omega) = |G(j\omega)| e^{j\angle G(j\omega)} \quad (\text{Chapter 13.2})$$

Here:

- $|G(j\omega)|$: gain or magnitude response
- $\angle G(j\omega)$: phase response (typically in degrees or radians)

This substitution arises from evaluating the Laplace-domain transfer function at $s = j\omega$, transforming the problem into an algebraic computation.

Physical Insight: Mass-on-Sheet Example

Imagine a small coin (mass) resting on a sheet of paper that moves sinusoidally side to side.

- At low frequencies: the coin moves nearly identically with the paper. The system tracks the input well.

$$|G(j\omega)| \approx 1, \quad \angle G(j\omega) \approx 0^\circ \quad (\text{Chapter 13.3})$$

- At high frequencies: inertia dominates. The coin barely moves while the paper vibrates beneath it.

$$|G(j\omega)| \ll 1, \quad \angle G(j\omega) \rightarrow -90^\circ \text{ (or more)} \quad (\text{Chapter 13.4})$$

This setup shows that the frequency response captures not only gain attenuation but also the phase lag introduced by system dynamics.

Frequency Response from Transfer Functions

Given a transfer function:

$$G(s) = \frac{N(s)}{D(s)} \quad (\text{Chapter 13.5})$$

its frequency response is:

$$G(j\omega) = \frac{N(j\omega)}{D(j\omega)} = \frac{|N(j\omega)| e^{j\angle N(j\omega)}}{|D(j\omega)| e^{j\angle D(j\omega)}} = |G(j\omega)| e^{j\angle G(j\omega)} \quad (\text{Chapter 13.6})$$

In this form, we can easily compute the gain and phase at each frequency. This is the basis for constructing Bode and Nyquist plots.

Bode Plots

The Bode plot consists of two separate plots versus $\log_{10}(\omega)$:

1. Magnitude Plot:

$$20 \log_{10} |G(j\omega)| \quad (\text{Chapter 13.7})$$

2. Phase Plot:

$$\angle G(j\omega) \quad (\text{in degrees}) \quad (\text{Chapter 13.8})$$

Each elementary component of $G(s)$ —poles, zeros, gain, integrators—has a well-known log-magnitude and phase contribution. The product of transfer function terms becomes the sum in log scale:

$$\log_{10}(ab) = \log_{10}(a) + \log_{10}(b) \quad (\text{Chapter 13.9})$$

This is why Bode plots use logarithmic magnitude and linear phase: they simplify system composition to visual addition.

Seven Elementary Transfer Functions

All rational transfer functions can be constructed from combinations of these elementary blocks:

1. Constant gain K
2. Pure integrator: $\frac{1}{s}$
3. First-order pole: $\frac{1}{s+a}$
4. First-order zero: $s+a$
5. Second-order term:

$$\frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (\text{Chapter 13.10})$$

6. Differentiator: s
7. Time delay: e^{-sT} (phase-only effect)

Bode plots for complex systems are built by superposing the Bode plots of these building blocks.

Nyquist Plot

The Nyquist plot is a parametric plot of $G(j\omega)$ in the complex plane as $\omega \in [0, \infty)$. It displays the real and imaginary parts of the frequency response:

$$G(j\omega) = \text{Re}(\omega) + j \cdot \text{Im}(\omega) \quad (\text{Chapter 13.11})$$

This plot encodes both gain and phase simultaneously and plays a central role in stability analysis.

Construction methods:

- Method 1: Sketch from Bode plot
- Sweep frequency from $\omega = 0^+ \rightarrow \infty$
- Record gain and phase at sample points
- Mirror the plot across the real axis for $\omega < 0$
- Indicate arrows (direction of increasing frequency)
- Method 2: Parametric plot of $G(j\omega)$

$$\omega \mapsto (\text{Re}[G(j\omega)], \text{Im}[G(j\omega)]) \quad (\text{Chapter 13.12})$$

This method is often implemented numerically or using symbolic tools like Mathematica.

Worked Bode and Nyquist Plot Examples

Example 1: First-Order Low-Pass Filter

Let:

$$G(s) = \frac{1}{s+1} \quad (\text{Chapter 13.13})$$

Frequency Response:

$$G(j\omega) = \frac{1}{j\omega + 1} = \frac{1}{\sqrt{\omega^2 + 1}} \angle -\tan^{-1}(\omega)$$

(Chapter 13.14)

Bode Plot:

- Magnitude: $20 \log_{10} \left(\frac{1}{\sqrt{\omega^2 + 1}} \right)$
- Phase: $-\tan^{-1}(\omega)$
- At $\omega = 1$: -3 dB, phase = -45°

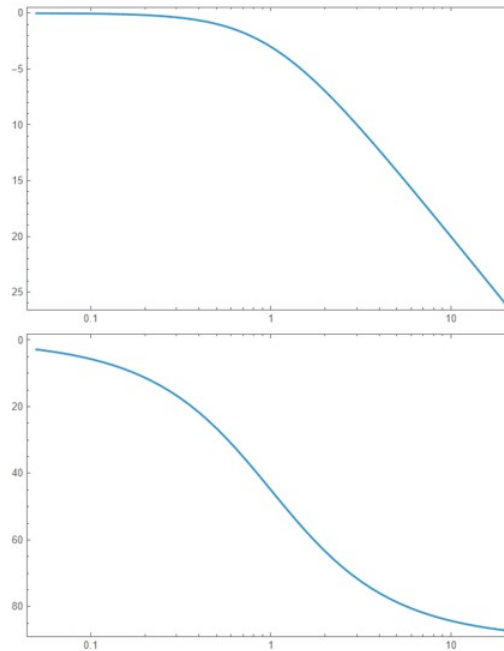


Figure Chapter13.1: Bode Plot First Order System

Nyquist Plot:

- Starts at $G(0) = 1$
- Arcs toward the origin in the fourth quadrant
- Never crosses or encircles -1

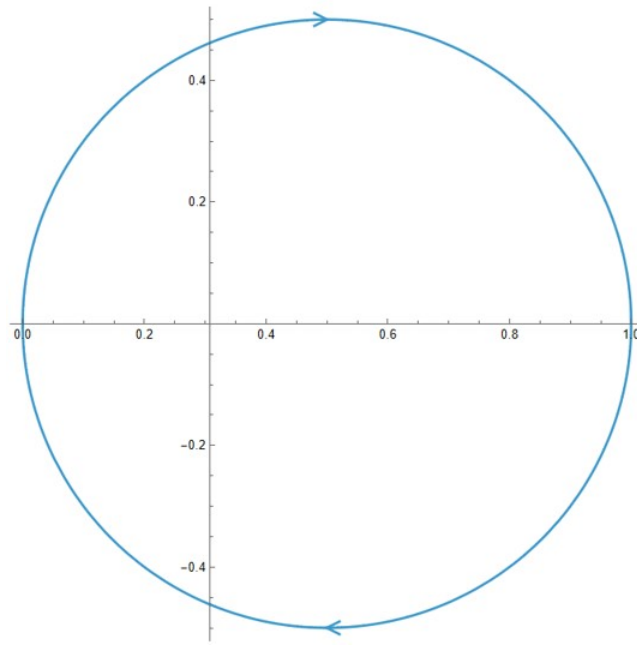


Figure Chapter13.2: Nyquist Plot First Order System

Example 2: Second-Order System with Resonance

Let:

$$G(s) = \frac{s(s+3)}{s^2 + s + 9} \quad (\text{Chapter 13.15})$$

Frequency Response:

$$G(j\omega) = \frac{j\omega(j\omega+3)}{(j\omega)^2 + j\omega + 9} \quad (\text{Chapter 13.16})$$

Use symbolic tools or software to evaluate magnitude and phase across frequencies.

Key Observations:

- Resonant peak near $\omega = 3$
- Magnitude plot shows a dip followed by a peak
- Phase starts near $+90^\circ$, decreases, and drops toward -90°

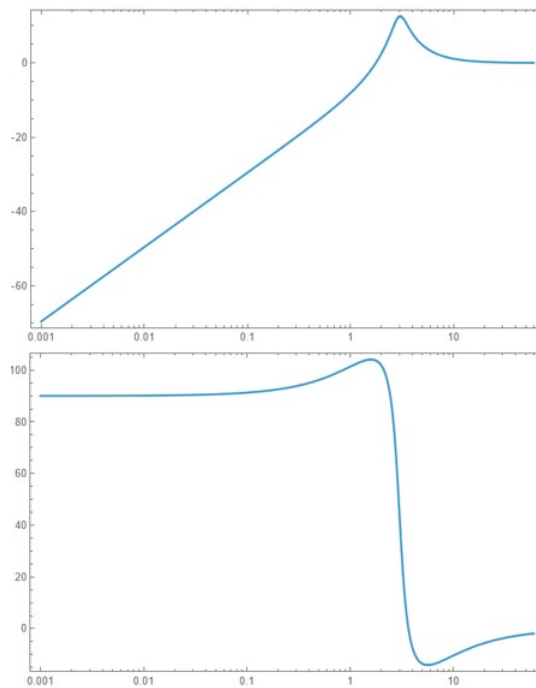


Figure Chapter13.3: Bode Plot Second Order System

Nyquist Plot:

- Begins at origin
- Swirls outward due to the numerator zero at $s = -3$
- Shows high-frequency behavior consistent with a second-order system

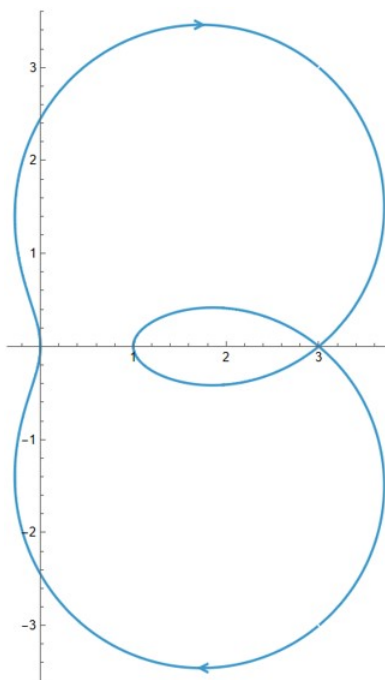


Figure Chapter13.4: Nyquist Plot Second Order System

Example 3: Lead Controller Behavior

Let:

$$G(s) = \frac{s+1}{s+10}$$

(Chapter 13.17)

Frequency Response:

$$G(j\omega) = \frac{j\omega + 1}{j\omega + 10} \quad (\text{Chapter 13.18})$$

Magnitude:

$$|G(j\omega)| = \frac{\sqrt{\omega^2 + 1}}{\sqrt{\omega^2 + 100}} \Rightarrow \text{High-frequency gain} \rightarrow 1/10 = -20 \text{ dB} \quad (\text{Chapter 13.19})$$

Phase:

$$\angle G(j\omega) = \tan^{-1}(\omega) - \tan^{-1}(\omega/10) \Rightarrow \text{Positive phase bump — typical of lead compensation} \quad (\text{Chapter 13.20})$$

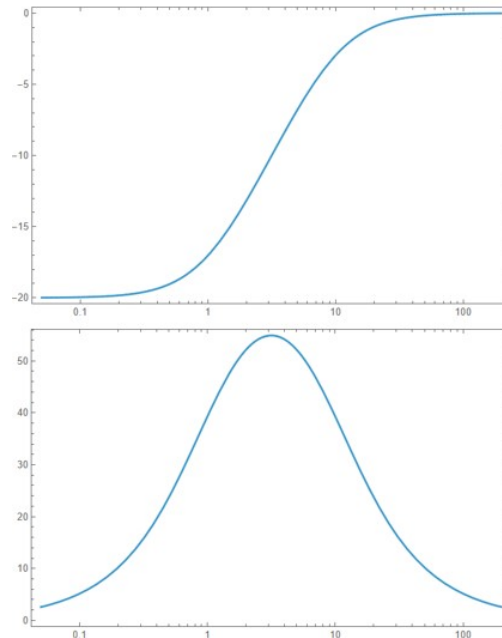


Figure Chapter13.5: Bode Plot Lead Controller System

Nyquist Plot:

- Arc in the right-half plane (phase lead)
- Ends closer to the real axis
- Useful for bending root locus toward the left

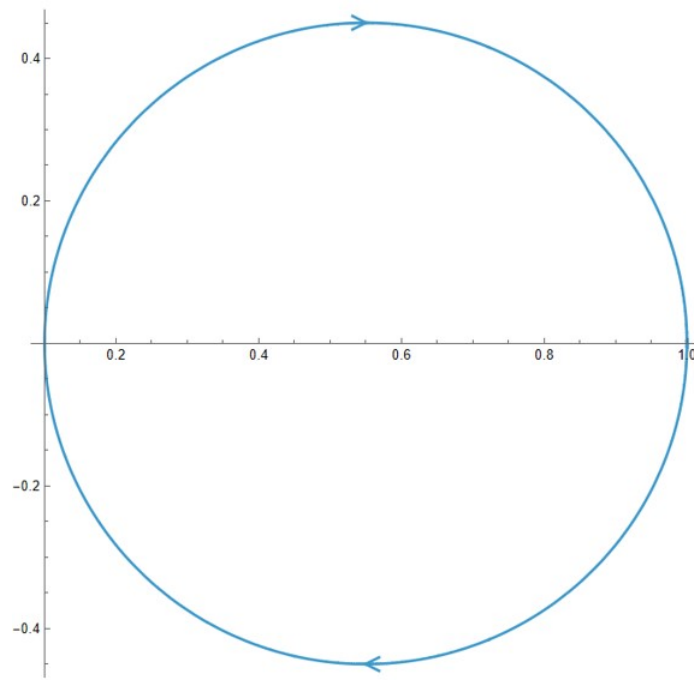


Figure Chapter13.6: Nyquist Plot Lead Controller System

Nyquist Stability Criterion

The Nyquist criterion is a frequency-domain test that links the stability of a closed-loop system to the open-loop transfer function.

Let:

$$L(s) = C(s)P(s) \Rightarrow T(s) = \frac{L(s)}{1 + L(s)} \quad (\text{Chapter 13.21})$$

The closed-loop poles are the roots of $1 + L(s) = 0$.

Nyquist Criterion:

$$Z = P - N \quad (\text{Chapter 13.22})$$

Where:

- Z : number of closed-loop poles in the right-half plane (unstable)
- P : number of open-loop poles in the RHP
- N : number of counter-clockwise encirclements of the point $-1 + j0$ by the Nyquist plot of $L(j\omega)$

Interpretation:

- For stability, we want $Z = 0 \Rightarrow N = P$
- If $L(s)$ has no RHP poles ($P = 0$), then stability requires no encirclements of -1

This criterion is a consequence of the argument principle in complex analysis, mapping the number of encirclements of a critical point to the number of zeros of a function inside a contour.

Important Considerations: Poles at the Origin

When $L(s)$ has poles on the imaginary axis (e.g., integrators at $s = 0$), the Nyquist contour must be indented to avoid singularities. In practice:

- The contour detours around the origin in the complex plane
- We analyze $\lim_{\omega \rightarrow 0^+} L(j\omega)$
- The mapping of the contour near the origin contributes a large arc in the complex plane

If the pole at the origin is stable (i.e., not causing exponential growth), then it does not increase P , but it does affect the Nyquist plot's shape and must be handled carefully.

Worked Example 1: Nyquist Stability Analysis

Let:

$$L(s) = \frac{k(s+2)}{s(s^2+s+3)} \quad (\text{Chapter 13.23})$$

Step 1: Open-loop Pole and Zero Locations

- One integrator pole at $s = 0$
- Two complex poles from $s^2 + s + 3$:

$$s = \frac{-1 \pm \sqrt{1-12}}{2} = \frac{-1 \pm j\sqrt{11}}{2} \quad (\text{Chapter 13.24})$$

- One zero at $s = -2$

This means:

$$P = 0 \quad (\text{no open-loop RHP poles}) \quad (\text{Chapter 13.25})$$

Step 2: Nyquist Stability Criterion

We check:

$$Z = P - N \Rightarrow Z = -N \quad (\text{Chapter 13.26})$$

To ensure stability, we want $Z = 0 \Rightarrow N = 0$, i.e., no encirclements of -1 .

Step 3: Find Critical Frequency

We solve numerically for the frequency ω_{crit} where:

$$\angle L(j\omega) = -180^\circ \quad (\text{Chapter 13.27})$$

This occurs at:

$$\boxed{\omega_{\text{crit}} \approx 2.44 \text{ rad/s}} \quad (\text{Chapter 13.28})$$

Step 4: Compute Gain at That Frequency

Evaluate:

$$|L(j\omega_{\text{crit}})| \approx 0.337 \Rightarrow \boxed{k_{\text{crit}} \approx \frac{1}{0.337} = 2.97} \quad (\text{Chapter 13.29})$$

Conclusion:

- For stability: $\boxed{k < 2.97}$
- For $k > 2.97$, the Nyquist plot encircles $-1 \Rightarrow Z = 1 \Rightarrow$ closed-loop unstable
- For $k = 2.97$, the system is marginally stable

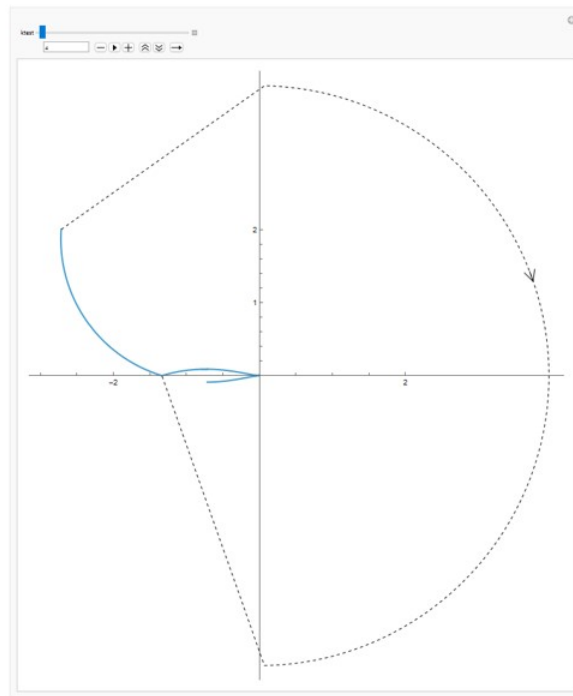


Figure Chapter13.7: Nyquist Plot Unstable

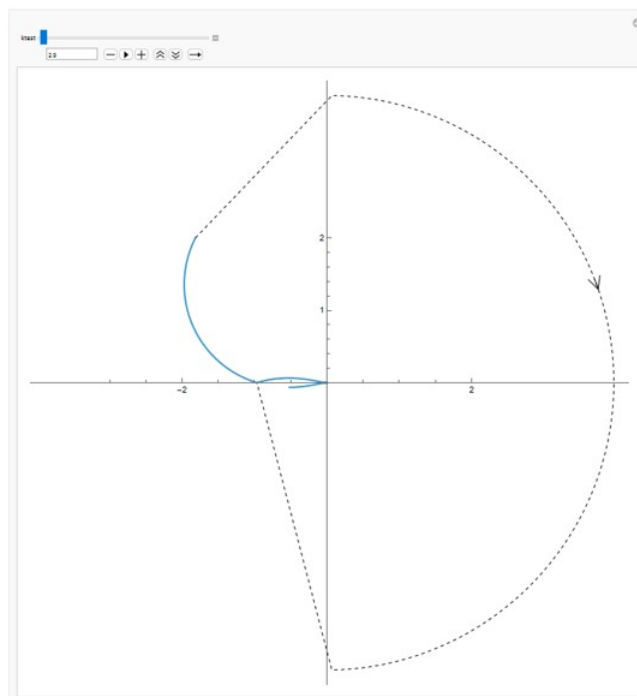


Figure Chapter13.8: Nyquist Plot Stable

Worked Example 2: Nyquist Analysis with RHP Pole

Let:

$$L(s) = \frac{k(s+1)}{s-2} \quad (\text{Chapter 13.30})$$

Step 1: Pole-Zero Structure

- One zero at $s = -1$
- One pole at $s = +2 \Rightarrow \boxed{P = 1}$

This system has an open-loop unstable pole at $s = +2$, so we must ensure:

$$Z = P - N = 1 - N \Rightarrow \boxed{N = 1} \text{ (for closed-loop stability)} \quad (\text{Chapter 13.31})$$

In other words, we require the Nyquist plot of $L(j\omega)$ to make one counterclockwise encirclement of -1 for the closed-loop system to be stable.

Step 2: Frequency Response

$$L(j\omega) = \frac{k(j\omega + 1)}{j\omega - 2} \quad (\text{Chapter 13.32})$$

Compute the magnitude and phase:

$$|L(j\omega)| = \frac{k \cdot \sqrt{\omega^2 + 1}}{\sqrt{\omega^2 + 4}} \quad (\text{Chapter 13.33})$$

$$\angle L(j\omega) = \tan^{-1}(\omega) - \pi - \tan^{-1}\left(\frac{\omega}{-2}\right) \quad (\text{Chapter 13.34})$$

Step 3: Critical Frequency for Instability

At $\omega = 0$, the phase is exactly -180° , and:

$$L(j0) = \frac{k(1)}{-2} = -\frac{k}{2} \Rightarrow |L(j0)| = \frac{k}{2} \quad (\text{Chapter 13.35})$$

$$\text{Set } |L(j0)| = 1 \Rightarrow \boxed{k_{\text{crit}} = 2}$$

Step 4: Interpretation

Since the system has a RHP pole at $s = 2$, the Nyquist criterion requires:

$$\boxed{\text{One CCW encirclement of } -1} \quad (\text{Chapter 13.36})$$

That means:

- For $k < 2$, the Nyquist plot does not encircle $-1 \Rightarrow$ closed-loop unstable
- For $k = 2$, the Nyquist plot touches $-1 \Rightarrow$ marginally stable
- For $k > 2$, the plot encircles -1 once $\Rightarrow$ closed-loop stable

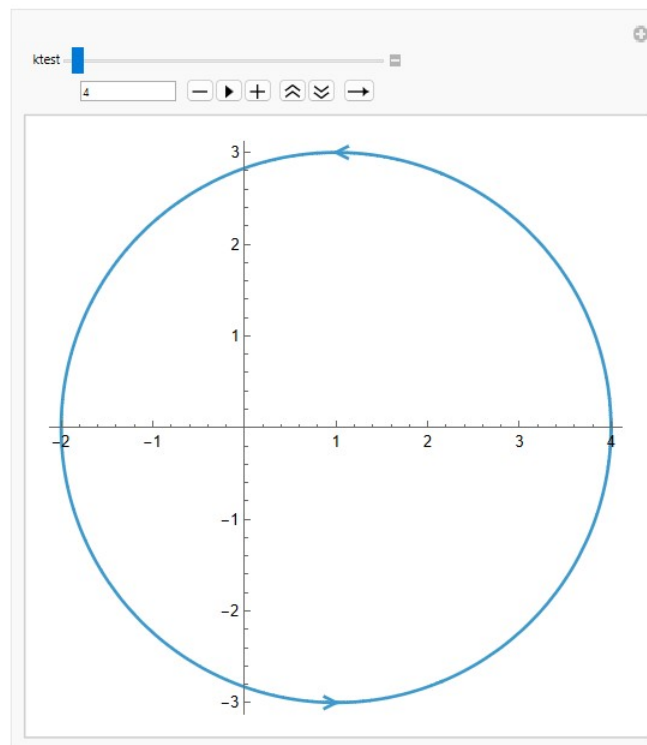


Figure Chapter13.9: Nyquist Plot Unstable

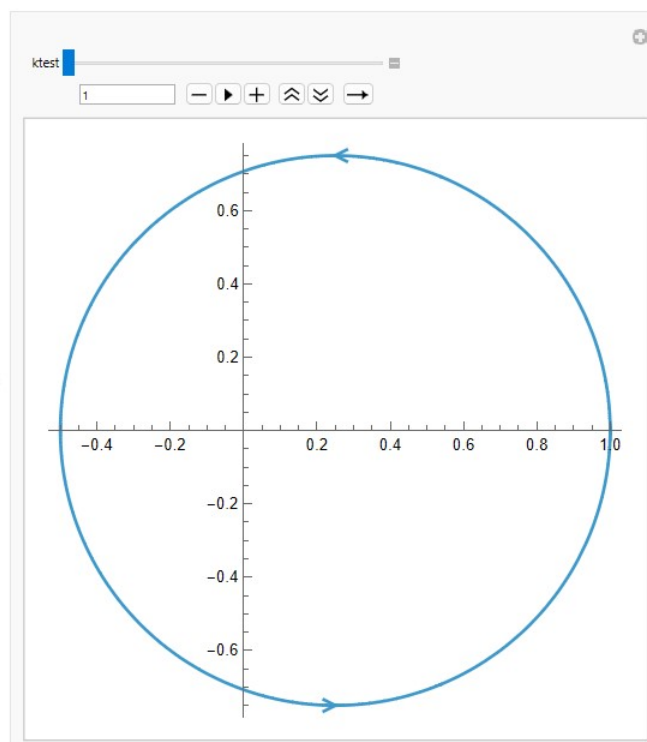


Figure Chapter13.10: Nyquist Plot Stable

Conclusion:

- Open-loop instability ($P = 1$) requires one CCW encirclement of -1
- Closed-loop stability is possible, but only if gain is tuned: $k > 2$
- Nyquist and Bode gain/phase margin analysis helps determine safe bounds on k

Conclusion

- Root locus and frequency response tools like Nyquist plots and Bode plots are critical to understanding system stability and performance.
- Bode plots provide insight into gain and phase margins, bandwidth, and resonance behavior.
- Nyquist plots visualize the response in the complex plane and help determine closed-loop stability even for systems with open-loop unstable poles.
- Nyquist stability criterion states:

$$Z = P - N \quad (\text{Chapter 13.37})$$

where:

- Z : number of unstable closed-loop poles
- P : number of open-loop poles in the right-half plane
- N : number of counter-clockwise encirclements of -1 by the Nyquist plot
- For a system to be stable, we want $Z = 0$, which implies:
 - If $P = 0$, then $N = 0$ (no encirclements of -1)
 - If $P = 1$, then $N = 1$ (one counter-clockwise encirclement of -1)

Design Guidelines

- Use Bode plots to evaluate gain and phase margins.
- Use Nyquist plots to assess closed-loop stability and validate against the Nyquist criterion.
- Tune controller gain to achieve desired margins while ensuring system stability:
- Avoid crossing the -1 point in the Nyquist plot without satisfying the Nyquist condition.
- Use frequency response plots to choose gains that avoid excessive phase lag or resonant peaking.

This page titled [Chapter 13: Frequency Response, Nyquist Stability Criterion, and Bode Plots](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 2: T-tests, P-values, and Hypothesis Tests

Central Limit Theorem

We just examined the dispersion of sample values around the mean value of the sample, $\bar{x}$. But **we did not obtain an estimate for the uncertainty in $\bar{x}$ approximated to the true mean μ .** To do this we would have to **repeat the experiment test and compare the $\bar{x}$ of each sample test.** Thus we would also obtain a set of samples for the mean.

One of the most important and profound theories in statistics deals with this set of sample means and their distribution and is referred to as the **Central Limit Theorem**. So far we have been working on the assumption that the population that we are working with is normally distributed. While many times that may be true it is not always the case. Let's think for instance if we were working with a uniform sample distribution like the one below.

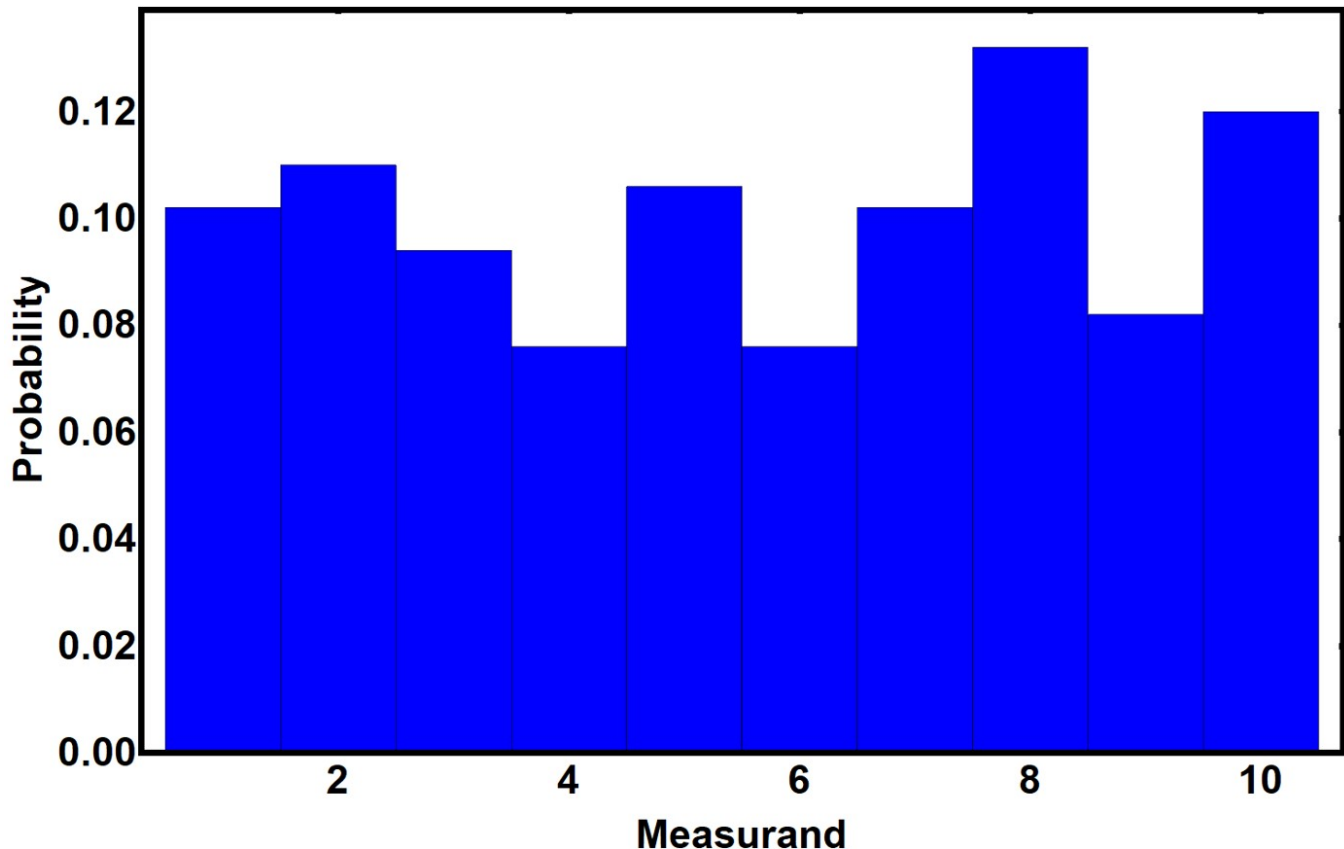


Figure Chapter2.1: Uniform Sample Distribution

At this point one might think we must give up and we cannot perform any data analysis... if only I were so kind.

No instead the **Central Limit Theorem** postulates a very powerful idea: regardless of the shape of the population distribution, the distribution of the mean values of a sample will be normally distributed as long as you obtain a large number of means, $n > 20$. We can see this visually with an example in the Mathematica Notebook for this lecture. If we randomly pick 6 samples from this distribution and average—if we look at 2 or 10 averages as seen below—the distributions do not look Gaussian. But if we select 30 or more (sometimes 20 is enough but I would obtain 30 just to be safe) then we see the distribution of the mean measurand values becomes Gaussian!

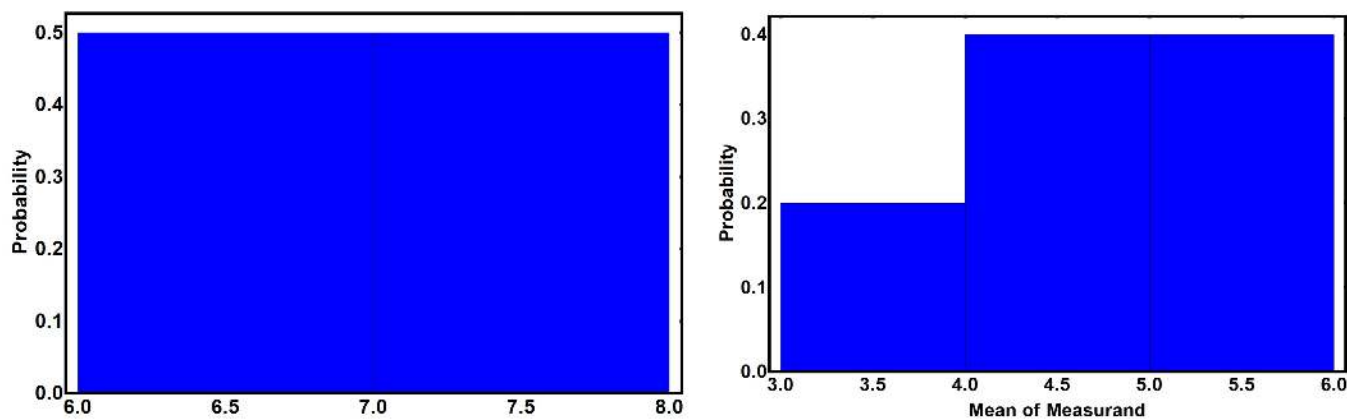


Figure Chapter2.2: Uniform Distrubution of Mean Sample Distribution of 2 and 10, left and right respectively.

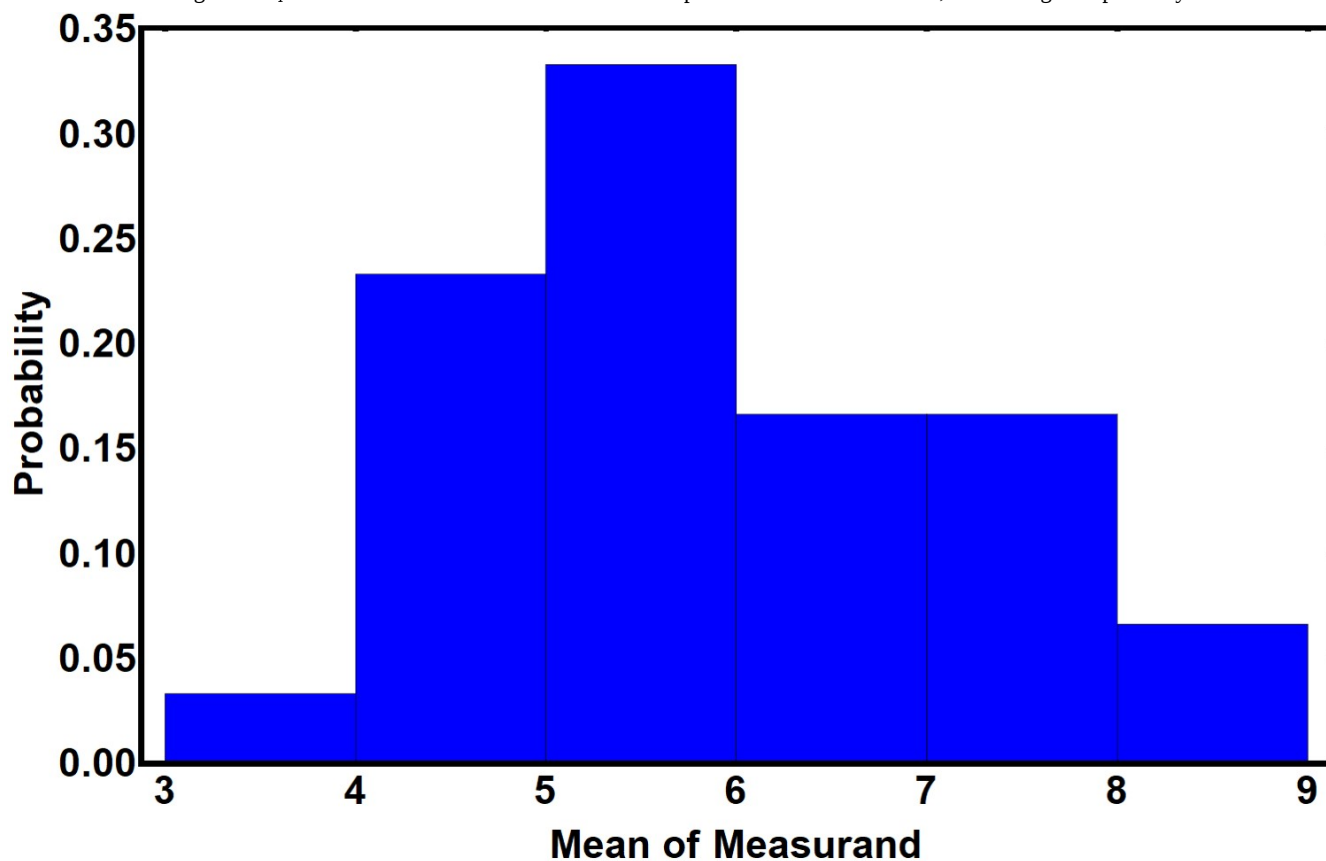


Figure Chapter2.3: Uniform Distrubiton Mean Sample Distribution of 30 Samples.

This has implications for the analysis we can perform, but let's prove this and show that it works for other types of distributions like an Exponential, χ^2 , and β . As you can see below, the distribution of the means are clearly normal or Gaussian distributions.

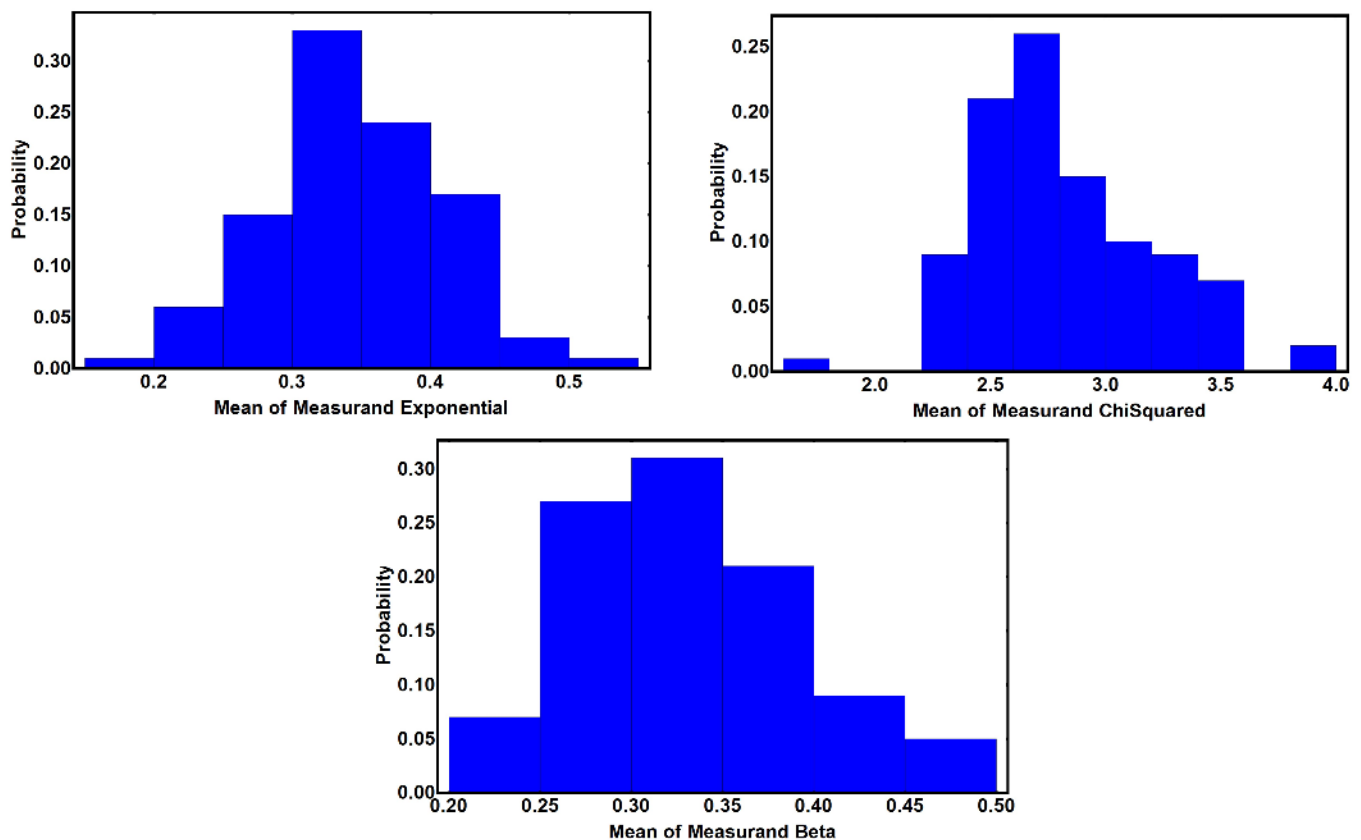


Figure Chapter2.4: Mean Distribution of 30 Samples or More for Uniform, Chi-Squared, and Beta Distributions

The practical application that is of interest to us as experimentalists is that we do not need to think about or worry about the shape of the population distribution. Since we know that the sample means are normally distributed, we can simply use the normal distribution of the means to calculate confidence intervals and do more complex analysis and comparisons moving forward... you are so lucky—t-test comparisons and hypothesis testing are coming soon!

So when we invoke the central limit theorem, we can make the following definitions where the standard deviation of the distribution of the means is:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

We can use the standard deviation of the means to now re-write our PDF or Gaussian function so z now becomes, when describing the distribution of the means:

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

This new definition of z allows us to re-write our confidence interval equation as follows, but now we can explicitly determine the confidence level or uncertainty in the population mean!

$$\mu = \overline{x} \pm z_{\frac{c}{2}} \frac{\sigma}{\sqrt{n}}$$

We can also use the approximation that $\sigma \approx S_x$ for $n > 30$, and we can then define the sample standard deviation of the sample means to be:

$$S_{\overline{x}} = \frac{S_x}{\sqrt{n}}$$

Now we can express the confidence interval of the population mean as:

$$\mu = \overline{x} \pm z_{\frac{c}{2}} \frac{S_x}{\sqrt{n}}$$

Whew!! That was a lot of work. Now let's go back to our rolling particles and calculate the 95% confidence interval for the population mean pressure.

Well that will simply be:

$$\mu = \overline{x} \pm z_{\frac{c}{2}} \frac{S_x}{\sqrt{n}} = 2.94 \pm 0.063$$

As you can see this is a much, much, much smaller range than our previous calculation due to the $\sqrt{\frac{1}{n}}$ factor. This is the key difference when trying to estimate the uncertainty in the population mean or using $\bar{x}$ as an estimate of the population mean. Previously we were just looking at the likelihood of observing a value that deviates from the population mean by a particular value.

In summary, the $c\%$ interval for the mean value is narrower than the data by a factor of $\frac{1}{\sqrt{n}}$. This is because n observations have been used to average out the random deviations of individual measurements.

Confidence Intervals for Small Samples

We have been working with large samples thus far, which is generally considered to be when $n \geq 30$, but there are many experiments when n will often be less than 30. For these situations, we will utilize the Student's t-distribution, developed by an amateur statistician writing under the pseudonym Student:

$$t = \frac{\overline{x} - \mu}{\frac{S_x}{\sqrt{n}}}$$

Student was actually William Gosset, a Guinness brewer who was applying statistical analysis to the brewing process. Guinness was against him publishing the data (and thus giving away the Guinness secrets), so he published under the pseudonym.

The assumption here is that the underlying population satisfies the Gaussian distribution. This distribution also depends on the degrees of freedom, $\nu = n - 1$.

The t-distribution is similar to the PDF—it is symmetric and the total area is unity. Moreover, the t-distribution approaches the standard Gaussian PDF as ν or n becomes large, and for $n > 30$, the two distributions are essentially identical, as we will see in the t-table soon.

The area, α , is between t and $t \rightarrow \infty$ for a given sample size. This is very different from our Z-table, where we were looking at the area between $z = 0$ and z . The value of α corresponds to the probability that for a given sample size, t will have a value greater than that given in the table. We can assert with confidence level $c = 1 - \alpha$ that the actual value of t does not fall in the shaded area.

A two-sided confidence interval is then:

$$\overline{x} - t_{\frac{\alpha}{2}, \nu} \frac{S_x}{\sqrt{n}} < \mu < \overline{x} + t_{\frac{\alpha}{2}, \nu} \frac{S_x}{\sqrt{n}}$$

Sometimes α in literature will be referred to as the level of significance. The precision uncertainty in the value of $\bar{x}$ is:

$$P_x = t_{\frac{\alpha}{2}, \nu} \frac{S_x}{\sqrt{n}} \quad (\text{Chapter 2.1})$$

Hypothesis Testing for a Single Mean for a Small Sample Size

The reason we are using these statistical tools is to make decisions regarding a measurand. One of the most common methods for doing this is hypothesis testing.

Typically, we deal with two hypotheses.

Null Hypothesis:

This is the first step in hypothesis testing.

The null hypothesis is written as

$$H_0 : \mu = \mu_0,$$

where μ_0 is some constant specific value.

Alternative Hypothesis:

This is the second step. The choice should reflect what we are attempting to show.

There are three common forms of alternative hypotheses:

- Two-tailed test: concerned with whether a population mean μ is different from a specific value μ_0

That is, $H_a : \mu \neq \mu_0$

- Left-tailed test: concerned with whether a population mean is less than a specific value

That is, $H_a : \mu < \mu_0$

- Right-tailed test: concerned with whether a population mean is greater than a specific value

That is, $H_a : \mu > \mu_0$

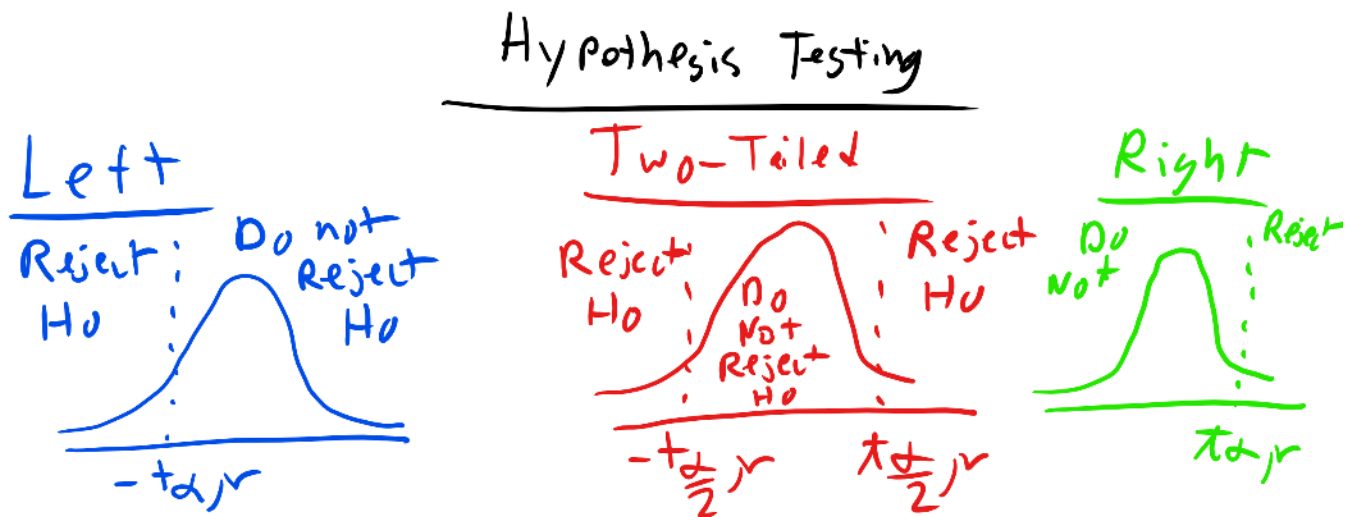


Figure Chapter2.5: Rejecting null hypothesis for two, right, and left tailed test

Procedure for Hypothesis Testing

1. Define the null hypothesis, H_0
2. Define the alternative hypothesis, H_a
3. Define the confidence level (for example, 95 percent)

4. Calculate the value of t_{exp} from the sample data
5. Determine the critical value from the t-distribution: $t_{\alpha, \nu}$ or $t_{\frac{\alpha}{2}, \nu}$, where ν is degrees of freedom
6. If t_{exp} falls in the rejection region, reject H_0 and accept H_a
7. If t_{exp} falls in the do-not-reject region, conclude there is not enough evidence to reject H_0 at the given confidence level

Let's apply this to an example. Suppose we want to test if the PCM data has a mean of 2.00 mg at 95 percent confidence:

\$\$

$H_0: \mu = 2.00 \text{ mg}$

\]

\$\$

$H_a: \mu \neq 2.00 \text{ mg}$

\]

\$\$

$t_{exp} = \frac{\overline{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} = 0.99011$

\]

\$\$

$t_{0.025, 17} = 2.11$

\]

Since $t_{exp} < t_{crit}$, we fail to reject H_0 . The data is consistent with a mean of 2.00 mg at 95 percent confidence.

P-Values

In scientific literature and statistics courses, you may have encountered the term P-value. This term is very common and often used more frequently than t-values or z-values in reporting results.

The P-value is the probability of getting a result that is more extreme than the value actually observed, assuming the null hypothesis is true.

Let's revisit the hypothesis test example we just discussed. We performed a two-tailed test with a confidence level of 95 percent, meaning each tail cutoff is 2.5 percent, or 0.025.

If the P-value is greater than 0.025, the result falls in the do-not-reject region. If the P-value is less than 0.025, it falls in the reject region.

In our case, the test yielded:

P-value = 0.336

Since 0.336 is greater than 0.025, we do not reject the null hypothesis. This agrees with the result obtained using the critical t-value.

The interpretation of the P-value in this case is that there is a 33.6 percent chance of obtaining a result as extreme or more extreme than the one observed, if the true mean were 2.00 mg.

This tells us that our sample data is consistent with the null hypothesis at the 95 percent confidence level.

Another PCM Application

Let's now apply hypothesis testing to a new scenario. Suppose we want to know:

Does the sample come from a population whose true mean weight is greater than 1.99 mg, assuming a confidence level of 99 percent?

We will use a right-tailed test.

The hypotheses are:

\$\$

$$H_0: \mu = 1.99 \text{ mg}$$

\]

\$\$

$$H_a: \mu > 1.99 \text{ mg}$$

\]

From our data, we calculate:

\$\$

$$t_{\text{exp}} = \frac{\overline{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} = 2.025$$

\]

The critical value from the t-distribution table is:

\$\$

$$t_{0.01,17} = 2.567$$

\]

Because $t_{\text{exp}} = 2.025$ is less than $t_{0.01,17} = 2.567$, it does not fall in the rejection region.

Therefore, we conclude that we do not have sufficient evidence to reject the null hypothesis at the 99 percent confidence level.

This means the population mean is not significantly greater than 1.99 mg.

Note that this does not prove the population mean is 1.99 mg. We are simply saying we do not have strong enough evidence to conclude otherwise.

Now, let's look at the P-value.

The calculated P-value is:

$$P\text{-value} = 0.02943$$

For a right-tailed test at 99 percent confidence, we compare against $\alpha = 0.01$. Since 0.02943 is greater than 0.01, the result also falls in the do-not-reject region. This agrees with the conclusion from the t-test.

Yet Another PCM Application

Let's consider a new hypothesis test.

Does the sample come from a population whose true mean weight is less than 2.01 mg, assuming a confidence level of 90 percent?

This will be a left-tailed test.

The hypotheses are:

\$\$

$$H_0: \mu = 2.01 \text{ mg}$$

\]

\$\$

$$H_a: \mu < 2.01 \text{ mg}$$

\]

From our data, we calculate:

\$\$

$$t_{\text{exp}} = \frac{\overline{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} = -0.0448035$$

\]

The critical value from the t-distribution table for a 90 percent confidence level and 17 degrees of freedom is:

$$t_{\{0.1, 17\}} = -1.33$$

Because $t_{exp} = -0.0448$ is greater than -1.33 , it does not fall in the rejection region.

Therefore, we do not reject the null hypothesis. The data does not provide sufficient evidence to conclude that the population mean is less than 2.01 mg at 90 percent confidence.

Now, let's examine the P-value.

$$P\text{-value} = 0.4824$$

For a 90 percent confidence level, the cutoff is $\alpha = 0.10$. Since the P-value is greater than 0.10, we are again in the do-not-reject region. This agrees with our earlier conclusion.

Hypothesis Testing for a Large Sample Size

When the sample size is large, we follow the same hypothesis testing procedure as with small samples. However, we use the z-distribution instead of the t-distribution.

Specifically, for large n , we replace $t_{\alpha, \nu}$ with z_{α} , and we compute the test statistic using:

$$z_{\{exp\}} = \frac{\overline{x} - \mu_0}{\frac{S_x}{\sqrt{n}}}$$

Note that when $n > 30$, we can either use the z-distribution or use the t-distribution with $\nu \approx \infty$, which produces essentially the same result.

This approach is useful when working with larger data sets, as the normal (Gaussian) approximation becomes more accurate due to the Central Limit Theorem.

Looking Back at Rolling Velocity

Let's apply hypothesis testing to a real experimental question:

Does the sample of rolling magnetic beads come from a population with a velocity less than 3.1 micrometers per second, at a confidence level of 99 percent?

This is a left-tailed test.

The hypotheses are:

$$H_0: \mu = 3.1 \, \frac{\mu\text{m}}{\text{s}}$$

$$H_a: \mu < 3.1 \, \frac{\mu\text{m}}{\text{s}}$$

From the data, we calculate:

$$t_{\{exp\}} = \frac{\overline{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} = -1.3383$$

Since this is a large sample size case, we can use the z-distribution. The critical value for a one-tailed 99 percent confidence test is:

$$z_{\{0.01\}} = -2.362$$

Now compare:

$$-t_{exp} = -1.3383$$

$$-z_{crit} = -2.362$$

Because $t_{exp} > z_{crit}$, it does not fall in the rejection region.

Conclusion: We do not reject the null hypothesis. The data does not provide sufficient evidence to conclude that the velocity is less than 3.1 micrometers per second at 99 percent confidence.

Now look at the P-value:

$$P\text{-value} = 0.09151$$

Since the P-value is greater than 0.01, this again confirms that we are in the do-not-reject region.

t-Test Comparison of Sample Means

We can also compare two samples based solely on their means using the t-test for independent samples.

The test statistic is:

\$\$

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)}}$$

Where:

- $\overline{x}_1, S_1, n_1$ are the sample mean, standard deviation, and size of sample 1

- $\overline{x}_2, S_2, n_2$ are the sample mean, standard deviation, and size of sample 2

The degrees of freedom ν can be approximated using:

\$\$

$$\nu = \frac{\left[\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

After computing this value, round ν to the nearest integer.

If the value of t falls within the interval $\pm t_{\frac{\alpha}{2}, \nu}$, then the two means are not significantly different at the specified confidence level.

This test is very versatile — it works for comparing large samples, small samples, or a mix of both.

Are These Materials Significantly Stiffer?

Let's compare two samples to determine if Material A is significantly stiffer than Material B.

Here is the sample data:

Material A:

Mean $\overline{x}_A = 302.6$ GPa

Standard deviation $S_A = 1.27$ GPa

Sample size $n_A = 12$

Material B:

Mean $\overline{x}_B = 302.3$ GPa

Standard deviation $S_B = 1.7$ GPa

Sample size $n_B = 15$

Hypotheses:

\$\$

$$H_0: \overline{x}_A = \overline{x}_B$$

\]

\$\$

$H_a: \overline{x}_A \neq \overline{x}_B$

\]

Approximated degrees of freedom:

\$\$

$\nu \approx 22$

\]

Calculated test statistic:

\$\$

$t_{\text{exp}} = 1.44$

\]

Critical value for a two-tailed 95 percent confidence level:

\$\$

$t_{\{0.025, 22\}} = 2.074$

\]

Since $t_{\text{exp}} = 1.44$ is less than 2.074, the result falls in the do-not-reject region.

Conclusion: There is no statistically significant difference in the stiffness of Material A and Material B at the 95 percent confidence level.

From a simulation, suppose the P-value was calculated to be:

P-value = 0.162

This also falls in the do-not-reject region, since $0.162 > 0.05$. Both the t-test and P-value support the same conclusion.

This page titled [Chapter 2: T-tests, P-values, and Hypothesis Tests](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 3: Propagation of Error and ANOVA

Bias and Single-Sample Uncertainty

Precision error can be calculated using statistical methods for repeat-sampled data. However, bias error cannot be uncovered using statistical methods. The only direct method is via comparison with measurements made using another more accurate apparatus. However, this is practically impossible. It is much too costly and time consuming. Instead, we will estimate the bias error using experience and knowledge of the device. Then we will obtain an estimate of the precision and accuracy of our experimental apparatus and determine if we must redesign our apparatus to measure a particular quantity of interest.

Bias Error: Measuring Volume in a Cylinder

What measurements do we need to measure the volume of a cylinder?

\$\$

$$V = \pi r^2 l$$

where r is the radius and l is the length. Uncertainty originates from the independent variables r and l . Precision error in these measurements may come from temperature fluctuation, thermal expansion, etc. Bias error comes from the type of measurement device (yard stick vs. micrometer vs. quantum laser).

Propagating Uncertainty

Usually we measure several quantities and use those measurements to calculate a desired quantity. Each measurement has some associated uncertainty which creates uncertainty in the calculated result.

We propagate the uncertainty.

Let y be a function of independent variables x_i with standard deviation σ_i . Then:

\$\$

$$\sigma_y = \sqrt{\left(\frac{\partial y}{\partial x_1} \sigma_1\right)^2 + \left(\frac{\partial y}{\partial x_2} \sigma_2\right)^2 + \dots + \left(\frac{\partial y}{\partial x_n} \sigma_n\right)^2}$$

Assuming small uncertainties and linearization via Taylor expansion:

\$\$

$$y(x_1 + u_1, \dots, x_n + u_n) \approx y(x_1, \dots, x_n) + \frac{\partial y}{\partial x_1} u_1 + \dots + \frac{\partial y}{\partial x_n} u_n$$

So:

\$\$

$$u_y = \sqrt{\left(\frac{\partial y}{\partial x_1} u_1\right)^2 + \dots + \left(\frac{\partial y}{\partial x_n} u_n\right)^2}$$

Assumes all uncertainties are independent.

Example: Volume of a Cylinder

Given:

Length = 50 m $\pm$ 0.5 m

Radius = 15 m $\pm$ 0.1 m

No precision error information, so only bias error applies.

The equation:

\$\$

$$V = \pi r^2 l$$

\]

Then:

\$\$

$$B_V = \sqrt{\left(\frac{\partial V}{\partial r} u_r\right)^2 + \left(\frac{\partial V}{\partial l} u_l\right)^2}$$

\]

Where:

\$\$

$$\frac{\partial V}{\partial r} = 2\pi r l, \quad \frac{\partial V}{\partial l} = \pi r^2$$

\]

Plug in:

\$\$

$$B_V = 589.05 \text{ m}^3$$

\]

Since no precision uncertainty:

\$\$

$$U_V = \sqrt{B_V^2 + 0^2 + 0^2} = B_V$$

\]

Convert to fractional uncertainty:

\$\$

$$\frac{B_V}{V} = \sqrt{\left(\frac{2 u_r}{r}\right)^2 + \left(\frac{u_l}{l}\right)^2} = 0.01667$$

\]

Percent = 1.667%.

Nominal volume = 35343 m³

Check: 35343 × 0.01667 = 589.05 m³

Example: Uniaxial Tension Test

Given:

- Force: mean 48.9 N, std dev 3.24, bias ±1 N
- Width: mean 0.1 m, std dev 0.01
- Length: mean 0.5 m, std dev 0.02
- Micrometer used for L and W
- 14 measurements of each

Equation:

\$\$

$$\sigma = \frac{F}{l \cdot w}$$

\]

Precision errors:

\$\$

$$P_F = t_{\{0.025,13\}} \frac{S_F}{\sqrt{n}}, \quad \frac{P_F}{F} = 0.04$$

$$P_l = t_{\{0.025,13\}} \frac{S_l}{\sqrt{n}}, \quad \frac{P_l}{l} = 0.023$$

$$P_w = t_{\{0.025,13\}} \frac{S_w}{\sqrt{n}}, \quad \frac{P_w}{w} = 0.06$$

\]

Bias uncertainty in fractional form:

\$\$

$$\frac{B_{\sigma}}{\sigma} = \sqrt{\left(\frac{u_F}{F}\right)^2 + \left(\frac{u_l}{l}\right)^2 + \left(\frac{u_w}{w}\right)^2} = 0.02$$

\]

Total uncertainty in percent:

$$U_{\sigma} = \sqrt{0.02^2 + 0.04^2 + 0.023^2 + 0.06^2} \cdot 100 = 7.93\% \quad (\text{Chapter 3.1})$$

Suggested Procedure for Calculating Total Uncertainty

1. Determine if bias, precision, or both are involved
2. Write the equation and identify independent variables
3. List known values: means, uncertainties, sample size
4. Start with precision error and convert to percent/fraction
5. Take partial derivatives and compute bias error
6. Combine to compute total uncertainty

Minimizing Error in Designing Experiments

Best time to minimize uncertainty is during design:

- Avoid subtracting large similar numbers
- Amplify signal strength
- Use null designs
- Avoid large correction factors
- Minimize sensor influence
- Calibrate the system

When in doubt about a data point, temporarily exclude it and refit. If the fit improves, the point may be an outlier — or it could indicate new behavior.

ANOVA

You may have heard of ANOVA analysis in some engineering or even more likely in the biological context. ANOVA is analysis of variance and it is an extremely useful statistical analytical tool.

ANOVA is particularly useful in experimental designs where there are multiple variables being varied simultaneously to see if they will have an effect on the measured outcome of the experiment. For example, perhaps I am creating a new high strength alloy and I want to design the new alloy material to maximize the ultimate tensile strength. I can vary a number of parameters when processing this material such as annealing time, solute concentration, amount of work hardening, and many more parameters. One question you may want to answer is: do any of these variables have a statistically significant effect on the ultimate tensile strength? Which parameter has the largest effect? Are there factors that have a combined effect?

Some of these questions we can answer using the tools we know about but others are not as easy to answer. So, before we jump into the deep end so to speak, let's start with developing a solid foundation and build a framework to analyze a single variable and perform a One-Way ANOVA.

One-Way ANOVA

Let us consider that we run an experiment with k independent random samples from k different populations and there will be a number of observations of measurands y , and we can classify them as y_{ij} where i varies from 1 to k , and j indicates the number of observations. Therefore, we will have k sample means $\bar{y}_k$.

Now in this analysis, there are several critical values that we will need to calculate. The first of which is the total sum of all observations:

\$\$

$$T = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$$

where n_i is the number of observations in group i . The total sample size is:

\$\$

$$N = \sum_{i=1}^k n_i$$

The overall sample mean is:

\$\$

$$\overline{y} = \frac{T}{N}$$

The treatment sum of squares is:

\$\$

$$SS_a = \sum_{i=1}^k n_i (\overline{y}_i - \overline{y})^2$$

The error sum of squares is:

\$\$

$$ESS_a = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

The degrees of freedom are:

\$\$

$$df_a = \sum_{i=1}^k n_i - 1 = k - 1 + \sum_{i=1}^k n_i - k$$

The mean square is defined as:

\$\$

$$\frac{SS_a}{df_a}$$

The F-ratio for treatments is:

\$\$

$$F_a = \frac{\frac{SS_a}{k-1}}{\frac{ESS_a}{N-k}}$$

The F-ratio is a key parameter in performing ANOVA analysis. A larger F value indicates large differences between sample means and thus large differences between groups.

We will also obtain a P-value from this analysis.

All of these values appear in the standard ANOVA table. Let's examine an example analyzing crop yields as a function of fertilizer type.

ANOVA[dat[[All, {3, 4}]][, 2, -1]]

		DF	SumOfSq	MeanSq	FRatio	PValue	All	177.015
{ ANOVA →	Model	2	6.06805	3.03402	7.86275	0.000699916	Model[1]	176.757
	Error	93	35.8862	0.385873			Model[2]	176.933
	Total	95	41.9542				Model[3]	177.356

Figure Chapter3.1: ANOVA Table Results

Here we can see the P-value is less than 0.05. Therefore, we can conclude that fertilizer type has a statistically significant effect on the crop yield.

Often we want to go beyond the ANOVA table and perform a post-hoc test. This lets us identify differences between specific group pairs. One common method is the Tukey post-hoc test, but it assumes equal variances. To be more general, we will use a Games-Howell post-hoc test.

```
f11 = Extract[dat[[All, {3, 4}]][, 2, -1], Partition[Position[dat[[All, {3, 4}]][, 2, -1], 1][[All, 1], 1]][[All, 2]]
{177.229, 177.55, 176.408, 177.704, 177.125, 176.778, 176.746, 177.061, 176.275, 177.967, 176.601, 177.031, 177.48, 176.874, 176.114,
176.008, 176.108, 178.357, 177.262, 176.919, 176.239, 176.573, 176.039, 176.818, 176.161, 177.226, 175.939, 177.165, 175.361, 177.277, 175.945, 175.883}

f12 = Extract[dat[[All, {3, 4}]][, 2, -1], Partition[Position[dat[[All, {3, 4}]][, 2, -1], 2][[All, 1], 1]][[All, 2]]
{176.479, 176.044, 177.412, 177.361, 177.385, 176.976, 177.38, 177.998, 176.435, 176.933, 175.983, 177.034, 176.437, 176.068, 177.121,
177.198, 176.604, 177.208, 177.149, 176.819, 176.999, 178.135, 176.429, 176.668, 176.896, 177.779, 176.414, 176.879, 177.581, 176.957, 175.748, 177.353}

f13 = Extract[dat[[All, {3, 4}]][, 2, -1], Partition[Position[dat[[All, {3, 4}]][, 2, -1], 3][[All, 1], 1]][[All, 2]]
{177.104, 178.08, 176.903, 177.54, 177.033, 178.286, 176.405, 176.431, 177.396, 176.926, 177.055, 177.344, 177.128, 177.168, 176.354,
179.061, 176.301, 177.593, 177.115, 177.794, 177.004, 178.037, 177.701, 177.633, 177.652, 177.1, 177.188, 177.405, 178.142, 177.711, 177.687, 177.118}

f11 = {f11, f12, f13}

{{177.229, 177.55, 176.408, 177.704, 177.125, 176.778, 176.746, 177.061, 176.275, 177.967, 176.601, 177.031, 177.48, 176.874, 176.114,
176.008, 176.108, 178.357, 177.262, 176.919, 176.239, 176.573, 176.039, 176.818, 176.161, 177.226, 175.939, 177.165, 175.361, 177.277, 175.945, 175.883},
{176.479, 176.044, 177.412, 177.361, 177.385, 176.976, 177.38, 177.998, 176.435, 176.933, 175.983, 177.034, 176.437, 176.068, 177.121, 177.198, 176.604,
177.208, 177.149, 176.819, 176.999, 178.135, 176.429, 176.668, 176.896, 177.779, 176.414, 176.879, 177.581, 176.957, 175.748, 177.353},
{177.104, 178.08, 176.903, 177.54, 177.033, 178.286, 176.405, 176.431, 177.396, 176.926, 177.055, 177.344, 177.128, 177.168, 176.354, 179.061, 176.301,
177.593, 177.115, 177.794, 177.004, 178.037, 177.701, 177.633, 177.652, 177.1, 177.188, 177.405, 178.142, 177.711, 177.687, 177.118}}

Table[TableForm[{{f11[i], f11[j]}], {i, 1, Length[f11]}, {j, 1, Length[f11]}] // TableForm
TableForm=
1. 0.269105 0.00042492
0.269105 1. 0.00540005
0.00042492 0.00540005 1.
```

Figure Chapter3.2: Games Howell Post-Hoc Test

Looking at the P-values, there is no statistically significant difference between fertilizer types 1 and 2. However, there are significant differences between types 1 and 3 and between types 2 and 3. This insight can guide further experiments and optimization.

Another valuable metric is the effect size. We can quantify this using η^2 , calculated as:

$$\eta^2 = \frac{SS_a}{SS_a + SS_{\text{error}}}$$

where SS_a is the treatment sum of squares. Interpretation:

- $\eta^2 \geq 0.14$: large effect
- $0.06 \leq \eta^2 < 0.14$: medium effect
- $\eta^2 < 0.01$: small effect

$$\frac{6.068046647800315}{(6.068046647800315 + 35.8861861412617)} = 0.144635$$

Figure Chapter3.3: Effect Size Eta Squared

This confirms that fertilizer has a large effect on crop yield.

Two-Way ANOVA

We can also run a Two-Way ANOVA. The total sum of squares can be expressed as:

\$\$

$$\sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \overline{y}_{..})^2 = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2 + b \sum_{i=1}^a (\overline{y}_{i.} - \overline{y}_{..})^2 + a \sum_{j=1}^b (\overline{y}_{.j} - \overline{y}_{..})^2$$

\]

We define:

- Treatment sum of squares:

\$\$

$$SS = b \sum_{i=1}^a (\overline{y}_{i.} - \overline{y}_{..})^2$$

\]

- Block sum of squares:

\$\$

$$BSS = a \sum_{j=1}^b (\overline{y}_{.j} - \overline{y}_{..})^2$$

\]

- Error sum of squares:

\$\$

$$ESS = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2$$

\]

- Total sum of squares:

\$\$

$$TSS = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \overline{y}_{..})^2$$

\]

Degrees of freedom:

\$\$

$$df = ab - a = (a - 1) + (b - 1) + (a - 1)(b - 1)$$

\]

F-ratios:

\$\$

$$F_a = \frac{\frac{TSS}{a - 1}}{\frac{ESS}{(a - 1)(b - 1)}}, \quad$$

$$F_b = \frac{\frac{BSS}{b - 1}}{\frac{ESS}{(a - 1)(b - 1)}}$$

\]

Now let's look at a real example.

							All	177.015
							fert[1]	176.856
							fert[2]	177.317
							fert[3]	176.713
							fert[4]	177.176
							dens[1]	176.757
							dens[2]	176.933
							dens[3]	177.356
ANOVA →	fert	3	5.60782	1.86927	5.42134	0.00185646	dens[1] × fert[1]	176.497
	dens	2	6.06805	3.03402	8.79939	0.000339167	dens[1] × fert[2]	177.325
	dens fert	6	1.31522	0.219203	0.635741	0.701266	dens[1] × fert[3]	176.382
	Error	84	28.9631	0.344799			dens[1] × fert[4]	176.824
	Total	95	41.9542				dens[2] × fert[1]	176.852
							dens[2] × fert[2]	177.013
							dens[2] × fert[3]	176.705
							dens[2] × fert[4]	177.164
							dens[3] × fert[1]	177.22
							dens[3] × fert[2]	177.613
							dens[3] × fert[3]	177.051
							dens[3] × fert[4]	177.541

Figure Chapter3.4: Two-Way ANOVA Table

Here we see fertilizer and crop density both have statistically significant effects on yield. The interaction between fertilizer and density does not.

Effect sizes:

- Fertilizer: large
- Density: medium
- Interaction: small

A post-hoc Games-Howell test confirms the fertilizer differences and shows that density types 1 and 2 are also significantly different.

Three-Way ANOVA

We can extend this to three independent variables.

							All	177.015
							dens[1]	176.784
							dens[2]	177.246
							block[1]	176.856
							block[2]	177.317
							block[3]	176.713
							block[4]	177.176
							fert[1]	176.757
							fert[2]	176.933
							fert[3]	177.356
							block[1] × dens[1]	176.856
							block[2] × dens[2]	177.317
							block[3] × dens[1]	176.713
							block[4] × dens[2]	177.176
							dens[1] × fert[1]	176.44
							dens[1] × fert[2]	176.778
							dens[1] × fert[3]	177.136
							dens[2] × fert[1]	177.074
							dens[2] × fert[2]	177.088
							dens[2] × fert[3]	177.577
ANOVA →	dens	1	5.12168	5.12168	14.8541	0.000226487	block[1] × fert[1]	176.497
	block	2	0.486139	0.243069	0.704959	0.497028	block[1] × fert[2]	176.852
	fert	2	6.06805	3.03402	8.79939	0.000339167	block[1] × fert[3]	177.22
	dens fert	2	0.427818	0.213909	0.620387	0.540182	block[2] × fert[1]	177.325
	block fert	4	0.887401	0.22185	0.643418	0.633043	block[2] × fert[2]	177.013
							block[2] × fert[3]	177.613
							block[3] × fert[1]	176.382
							block[3] × fert[2]	176.705
							block[3] × fert[3]	177.051
							block[4] × fert[1]	176.824
							block[4] × fert[2]	177.164
							block[4] × fert[3]	177.541
							block[1] × dens[1] × fert[1]	176.497
							block[1] × dens[1] × fert[2]	176.852
							block[1] × dens[1] × fert[3]	177.22
							block[2] × dens[2] × fert[1]	177.325
							block[2] × dens[2] × fert[2]	177.013
							block[2] × dens[2] × fert[3]	177.613
							block[3] × dens[1] × fert[1]	176.382
							block[3] × dens[1] × fert[2]	176.705
							block[3] × dens[1] × fert[3]	177.051
							block[4] × dens[2] × fert[1]	176.824
							block[4] × dens[2] × fert[2]	177.164
							block[4] × dens[2] × fert[3]	177.541

Figure Chapter3.5: Three-Way ANOVA Table

This shows:

- Fertilizer and density have statistically significant effects
- Block location does not
- Interaction between density and fertilizer is not significant

- Interaction between density and block is significant
- Three-way interaction is significant

Effect sizes:

- Fertilizer: large
- Density: medium
- Block: small
- Density and fertilizer: small
- Fertilizer and block: small

Only blocks 2 and 3 show a statistically significant difference.

ANOVA is an incredibly useful tool.

This page titled [Chapter 3: Propagation of Error and ANOVA](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 4: Fourier Analysis and Continuous Fourier Transforms

Fourier Analysis Applications

Before we get into talking about Fourier Analysis and getting in the weeds of continuous fourier transforms, discrete, spectrum analysis, signal aliasing, and spectral leakage let's take a step back and see why this analysis is useful or helpful.

Fourier analysis is a critical tool common to engineers and experimentalists, it is extremely ubiquitous and interdisciplinary. Fourier Analysis generally involves measuring some signal that varies as a function of time, one then performs a Fourier transform which allows one to examine how that signal varies in the frequency space or domain instead of the time domain. Because we know that time and frequency are inversely related you are essentially looking at a signal in one space, whether that be time, distance, etc and then you transform the signal to examine the behavior in some reciprocal space or domain.

Time to Frequency Space or Reciprocal Space

I understand that sounds very very confusing so let's look at a couple of practical applications of Fourier Analysis. One of the most common types of Fourier Analysis involves looking at a signal whether it be electrical, mechanical, optical, etc. and performing a Fourier Transform to look at the signal in frequency space via a Fourier Spectrum graph, much more on this later. We can then examine these Frequency Spectrum curves which show the amplitude of our harmonic coefficients as a function of frequency and we can identify many things about the signal we are measuring including:

- Fundamental or Lowest Frequency
- Maximum Frequency
- Dominant Frequency

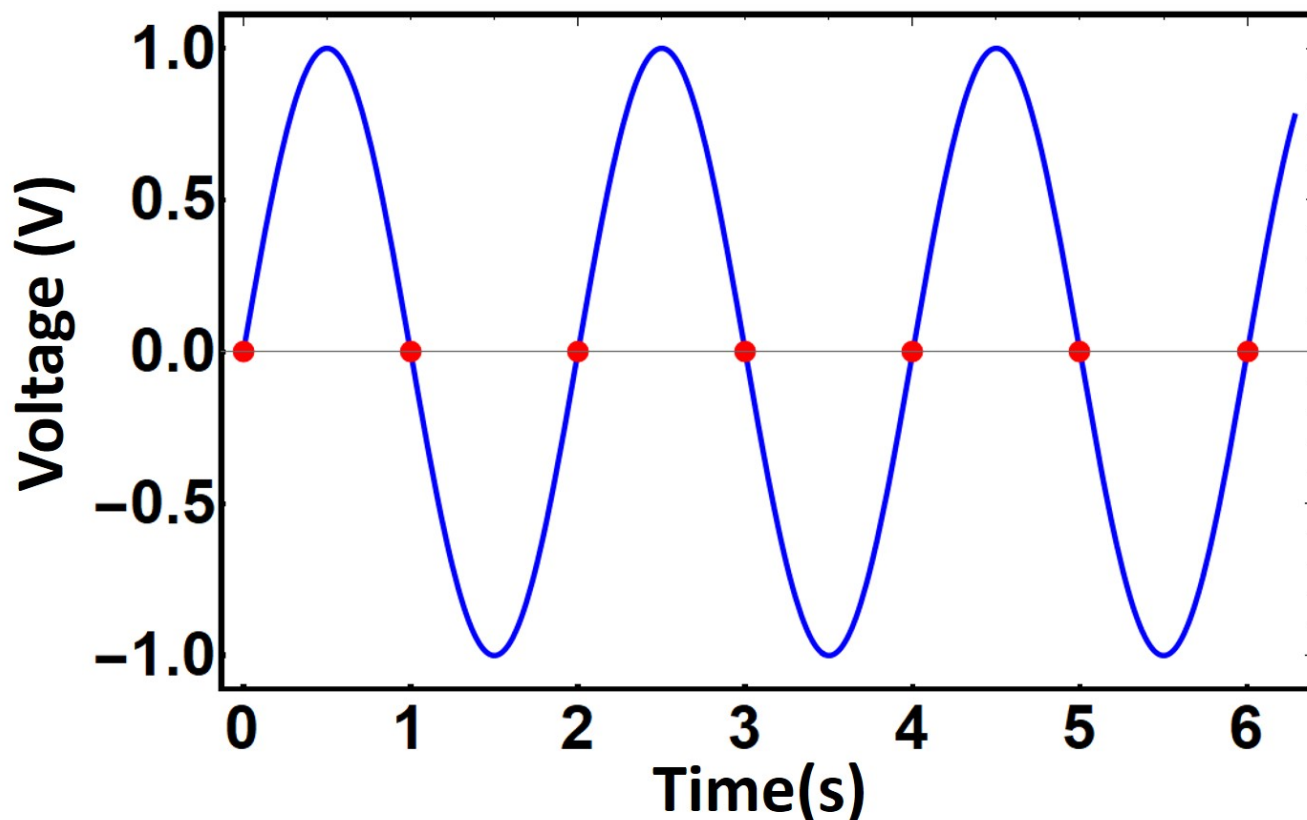


Figure Chapter4.1: Analyzing a Signal and Transforming the Signal Into Frequency Space.

These are critical parameters that will allows us to make decisions about the design of our application or experimental apparatus. As you can see in the example above these two signals are distinct in terms of frequencies and we see in the resultant Frequency

Spectrum curves there are also distinct changes in these curves as well.

ECG Measurements

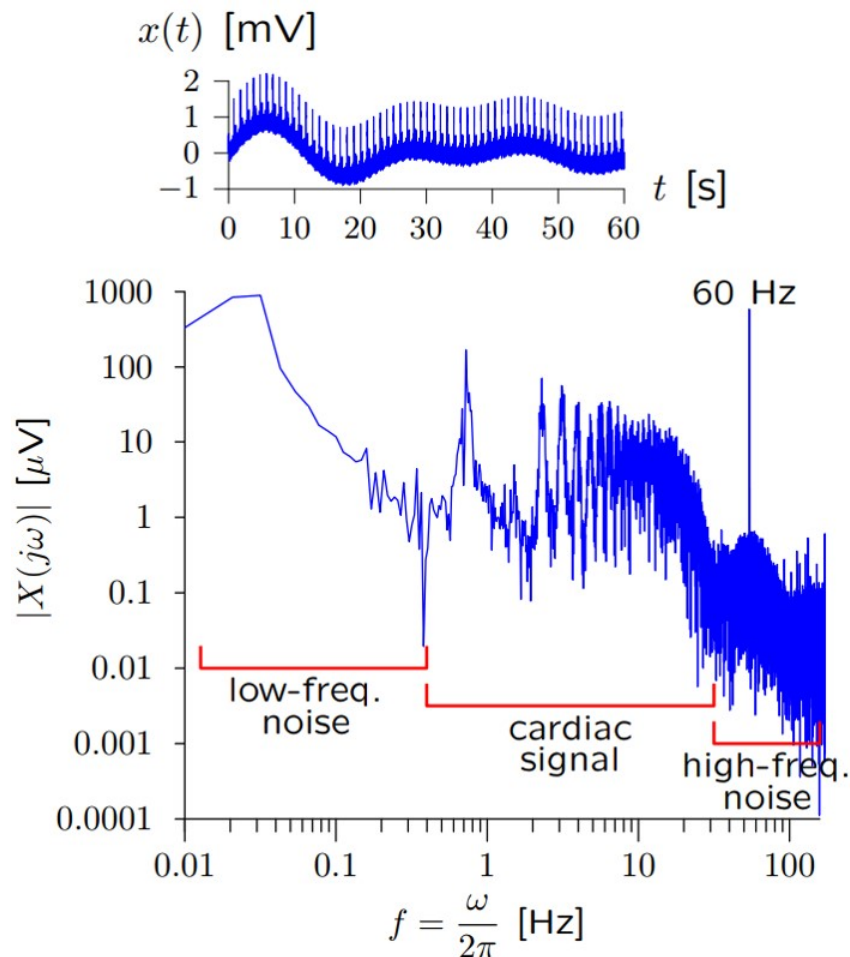
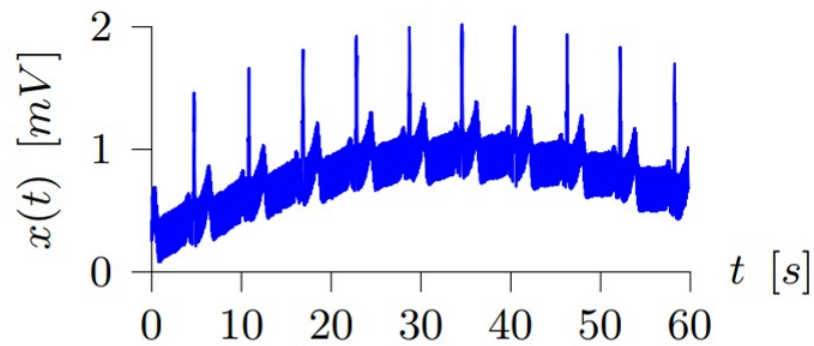


Figure Chapter 4.2: Converting Voltage Measurements to DFT transform and Frequency Spectrum.

An electrocardiogram (ECG) is a truly amazing measurement as we are effectively seeing in the curve above the bio-potential, the voltage generated from heart cells. Now typically cells can produce a potential of around 100mV. As you can see on the curve above we are measuring around 1mV but this is amazing!! Through all the tissue, blood, fluids, collagen, bone, etc we are able to measure the potential of a heart cell with electrodes placed on our skin. Really amazing! We can then take this signal and perform a Fourier transform and identify anomalous frequencies to make a diagnosis for irregular heartbeats and other issues.

Unfiltered ECG



Filtered ECG

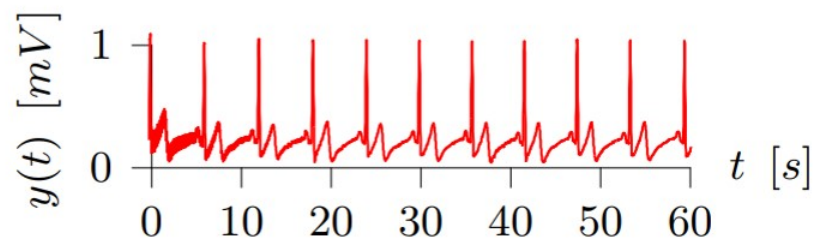


Figure Chapter4.3: Low Pass and High Pass Filtration of ECG Signal Determined from ECG analysis.

We can also see some high and low frequency noise in our signal and who can blame us we are measuring the voltage a heart cell at distances on the meter length scale away. The beauty of Fourier analysis is we can identify the high and low frequency noise typically expected for ECG measurements and then we can build high and low pass frequency filters that filter out this noise and improve our original signal of interest.

Differences in Musical Instruments

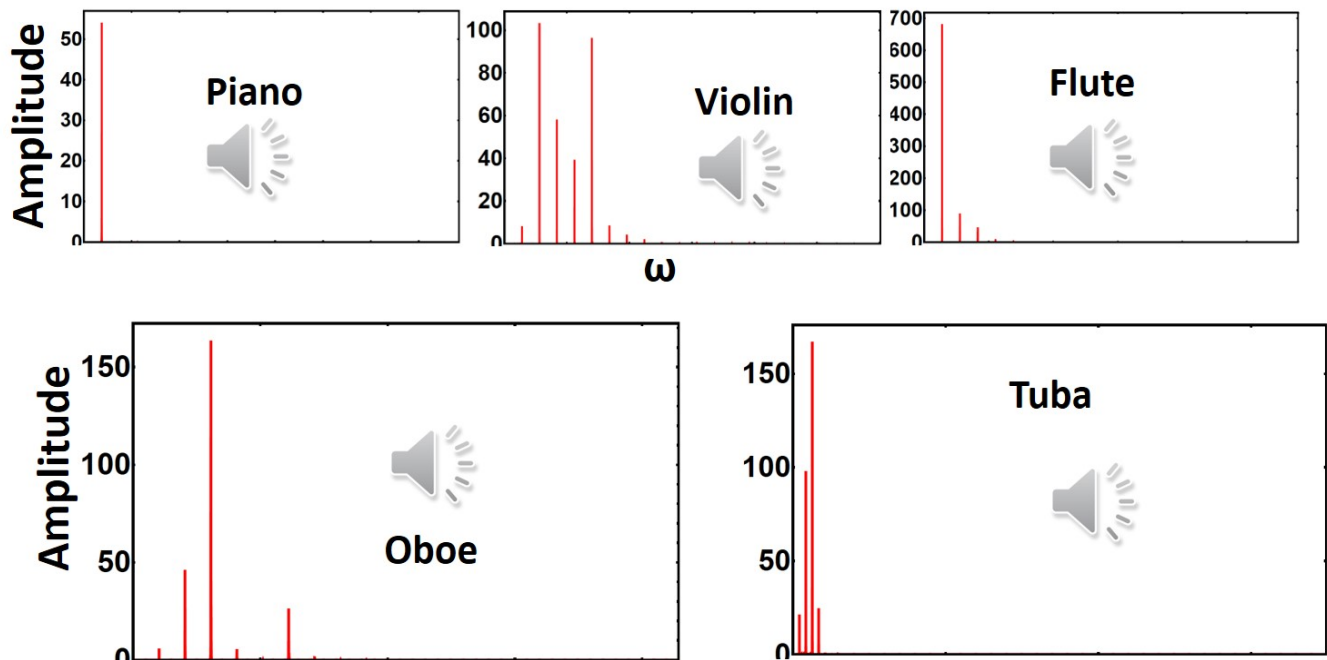


Figure Chapter4.4: Instrument Sounds and Frequency Spectrum.

We can also utilize Fourier Transforms to distinguish between the same note being played by different instruments. We can record the sound being played and measure the voltage registered by the recording device and then perform a Fourier transform on the voltage vs time signal to convert to our Frequency Spectrum curve where the frequency is on the x-axis.

Distinguish Between Vocal Sounds, Ahh, Eee, Ooo

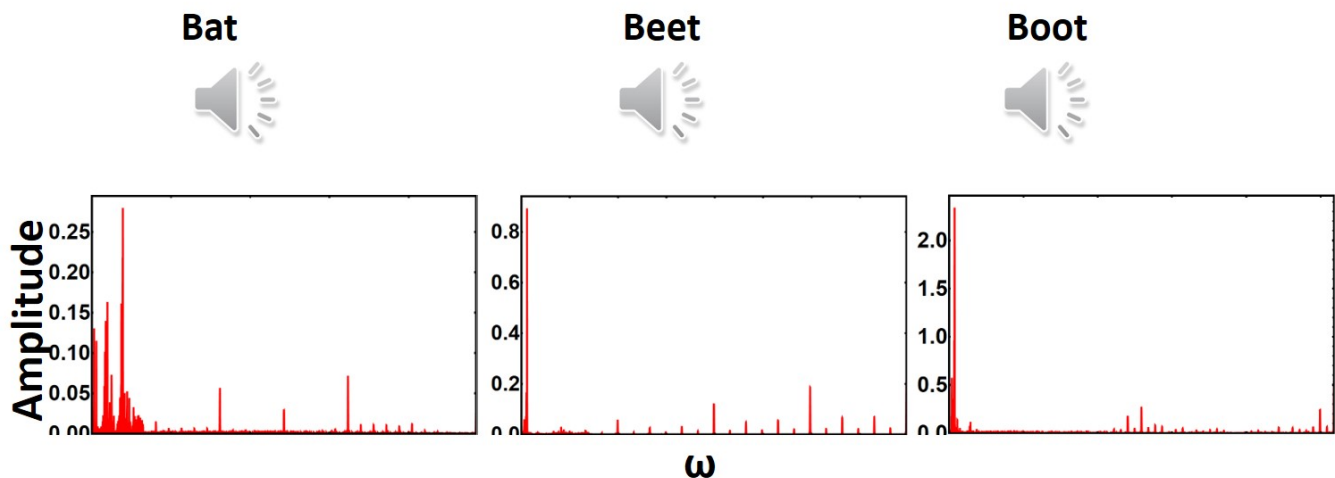


Figure Chapter4.5: Bat, Beet, and Boot Sounds.

Fourier analysis can similarly be used to distinguish between vowel sounds. You must appreciate my effort in this course as I recorded myself saying these words over and over and looked very foolish indeed. However, I was able to acquire data as a function of time and voltage and performed my Fourier Analysis.

XRD Reciprocal Space and Bragg's Law

As I have stated many times, my background is in Materials Science and Engineering, and one of the most critical tools to analyze the structure of materials is via X-Ray Diffraction (XRD). In XRD, you shoot an incident beam of radiation — typically Cu_α radiation — and that wavelength of radiation at certain angles, called Bragg Angles, will constructively interfere with planes of atoms.

We can then use simple trigonometric identities to calculate the interplanar distance d_{hkl} between these planes (or more generally any distance between periodic arrays of atoms), as described by Bragg's Law:

$$n\lambda = 2d_{hkl} \sin(\theta)$$

Where:

- n is a positive integer
- λ is the wavelength of the radiation, typically Cu_{α} , which is 1.56 Å

The XRD output is a graph of the intensity of the diffracted radiation as a function of 2θ . Now if we look at Bragg's Law, we can see that the distance d that we are trying to measure is inversely related to θ , so the graph of intensity vs. 2θ is effectively a plot of intensity vs. **reciprocal space**, i.e., $\frac{1}{d}$. This is very similar to what we have seen previously: a plot of inverse time space or frequency space.

We can use XRD to determine characteristic distances between or within polymer chains. Remember: XRD is looking for periodic distances within polymer structures, and we can only measure distances that exhibit some degree of order — either short-range or long-range.

We can see an example of this unique type of Fourier analysis by looking at a family of methacrylate polymers — specifically PMMA, PPMA, PEMA, and PBMA — which vary in the R length. In this case, the end group is a methyl, CH_3 .

Now if we place these polymers within the XRD, we can obtain the following XRD curves:

As you can see, we observe multiple peaks in our XRD curves. A peak in the intensity correlates to when we obtain constructive interference. Thus, at that angle — or equivalently, at that inverse distance — there is some distance in the polymer that has some degree of long-range or short-range order.

You can also see that some of the peaks in the curve are conserved — i.e., the peaks at larger 2θ values appear at essentially the same location. However, the peaks at low 2θ angles appear to shift based on the polymer. This suggests there is some characteristic, repeated, periodic, or ordered structure/distance in the polymer chain.

The question then becomes: Why do some distances in the polymer chain change?

To understand this, we must look at the chemical structure of our methacrylate polymers.

We saw previously that the primary structural difference between our methacrylate polymers is the length of the pendant side chain, i.e., the number of CH_2 (R groups). From this schematic:

- Intra-chain distances will **not** change — these are fixed bond lengths (like the C–C bond ≈ 1.54 Å)
- Inter-segmental distances will **not** change
- Inter-chain and intra-segment distances **will** change as R increases

Now back to our XRD plots — which peaks correspond to which distances?

- Peaks at large 2θ : these correspond to **small** distances (like intra-chain), and they don't shift — they are conserved
- Peaks at small 2θ : these correspond to **larger** distances (like inter-chain), and they do shift to lower angles as R increases

Why do peaks shift left with increasing R?

Because as R increases, inter-chain distances increase. So the associated angle θ must decrease to satisfy Bragg's Law (since $\sin(\theta) \propto \frac{1}{d}$), and thus 2θ shifts left.

This is the power of Fourier Analysis again — XRD gives us a reciprocal-space view of periodic structures and allows us to infer real-space behavior through mathematical transformation.

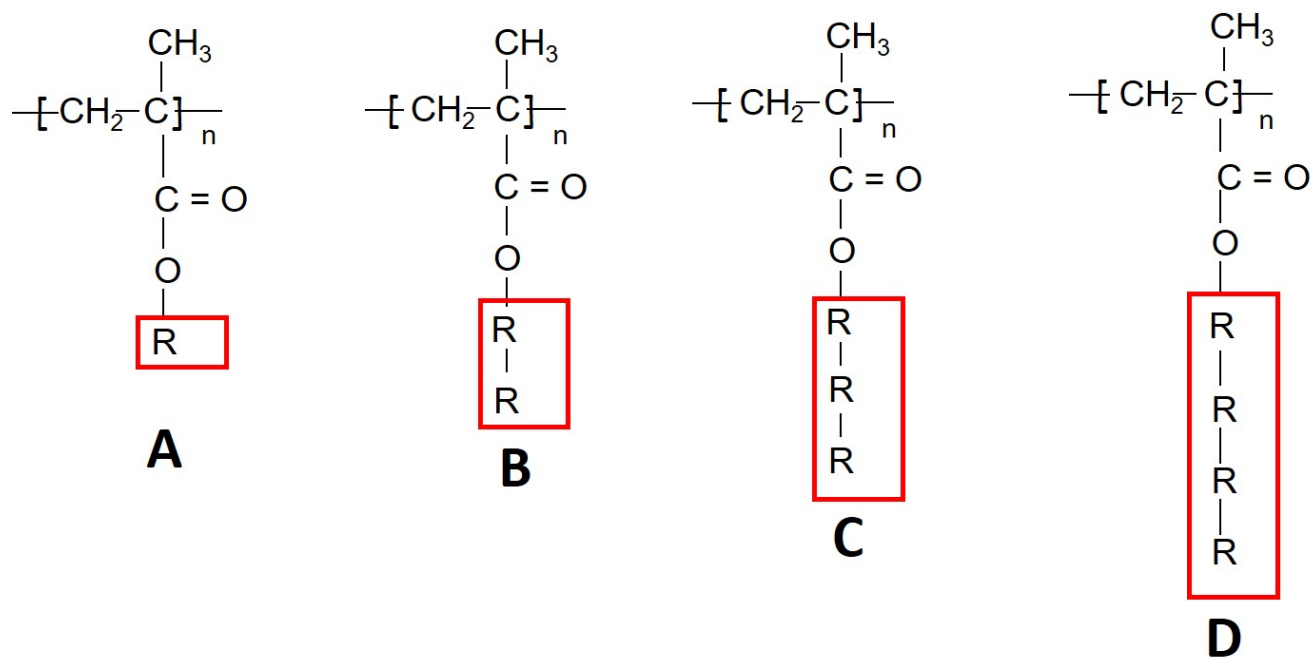


Figure Chapter4.6: PMMA, PPMA, PEMA, and PBMA structure.

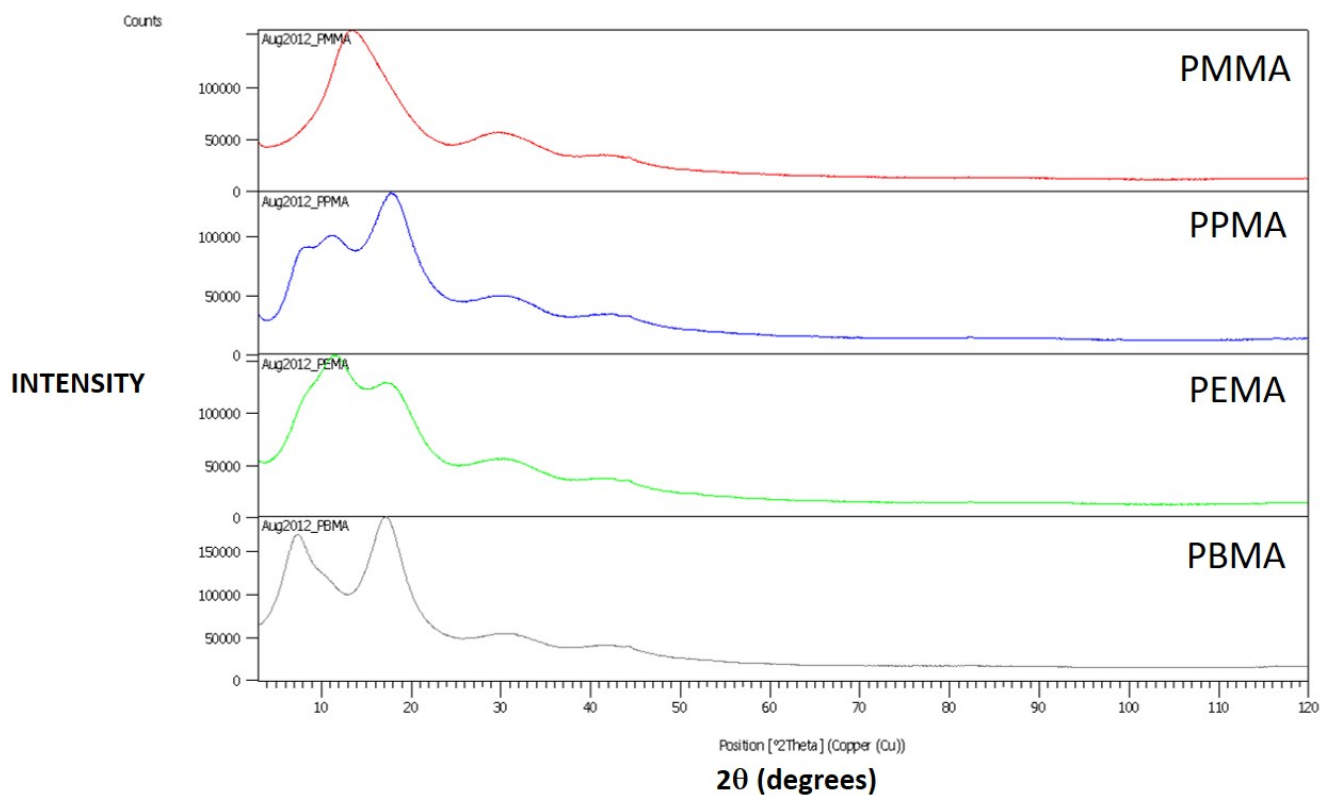


Figure Chapter4.7: XRD Plot of PMMA, PPMA, PEMA, and PBMA.

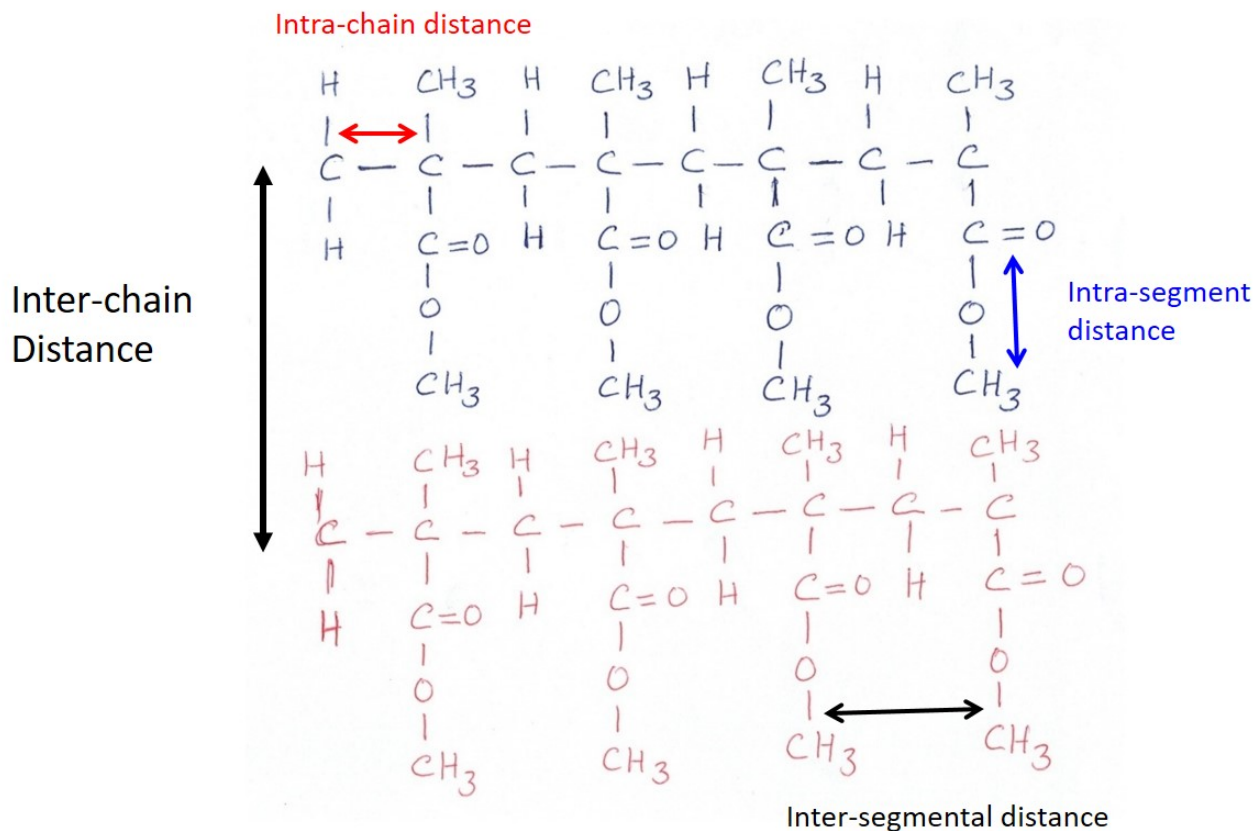


Figure Chapter4.8: Schematic of PMMA characteristic distances.

Photo 51: Rosalind Franklin

Hopefully everyone has seen this image before. What is it? Don't read the caption.

This is Rosalind Franklin's Photo 51 — an XRD picture of a DNA fiber from a calf thymus — which was then analyzed by Watson and Crick who were later credited for the discovery of DNA. Rosalind Franklin was an amazing chemist and X-ray crystallographer; however, Watson and Crick knew about Fourier.

So when they saw this picture they could clearly make out this reciprocal space image and reconstruct the structure of DNA. Rosalind Franklin was not given credit for her contribution until after her death. She died at age 37 due to ovarian cancer — a tragic story, but again showing the importance of Fourier Analysis. History books would have been written completely differently.

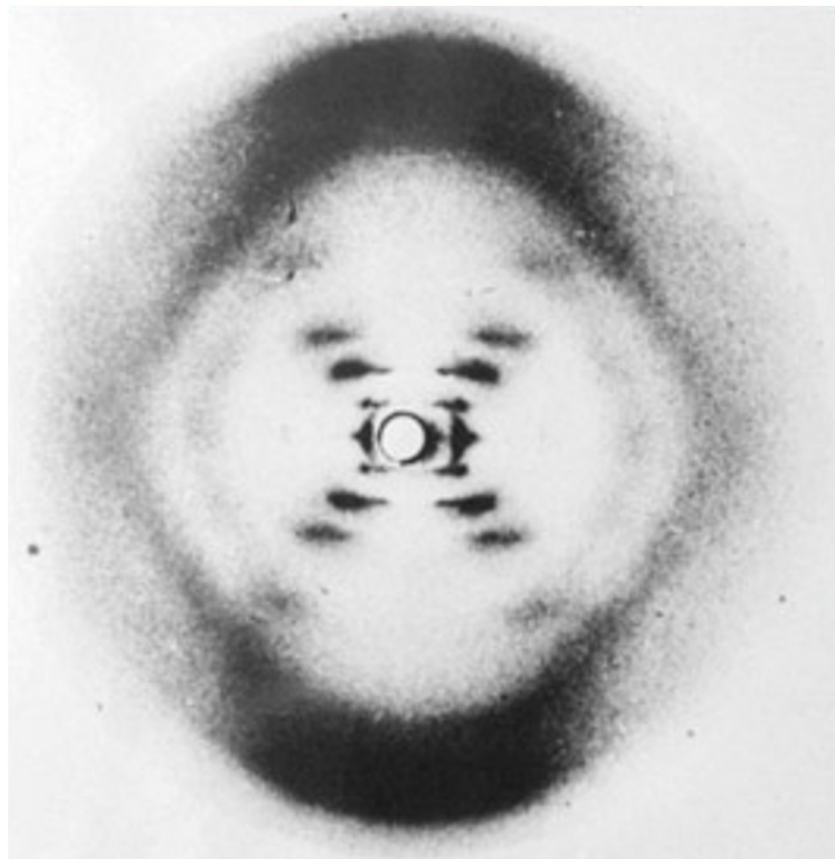
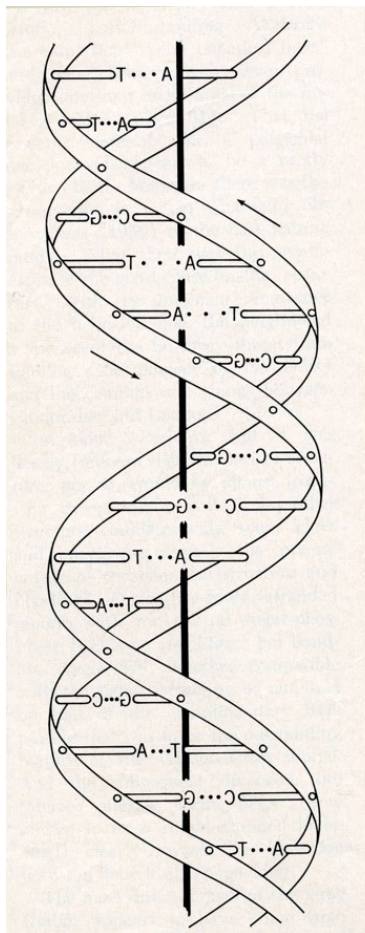


Figure Chapter4.9: Rosalind Franklin Photo 51.

Time Varying Measurements, Harmonic Functions, Cyclic and Circular Frequency:

We have previously discussed in Lecture 1 that all measurands will vary as a function of time and they can be static, dynamic, periodic, aperiodic, etc. The power and beauty of Fourier Transforms — as we will see soon — is that regardless of how the signal varies in time we will be able to analyze this using the framework of Fourier Analysis.

Now many of the signals that we will analyze will be periodic in nature and many of them will be harmonic functions. You have likely encountered many harmonic functions in your career as an engineer and hopefully have worked with sine and cosine functions. A function $y(t)$ can be classified as harmonic if:

$$\frac{\partial^2 y}{\partial t^2} = -c y(t)$$

Where c is a constant.

Prove that both cosine and sine are harmonic functions. What about $y(t) = x^2$?

These harmonic functions — and really all functions — can be treated over some period T as periodic. These functions will exhibit some frequency, and here we must distinguish:

- Circular Frequencies ω
- Cyclic Frequencies f

Circular frequency is related to cyclic frequency as:

$$\omega = 2\pi f$$

\]

Where ω is in rad/s and f is in Hz. In this course we will always use cyclic frequency f in SI units. However, you may see ω in problems or real-world systems, so be sure to convert when necessary.

Continuous Fourier Transforms: Calculating Harmonic Coefficients

Finally we are about to undertake the task of performing a Fourier Transform. Just a warning — the math will get a bit tricky here — but do not get lost in the math. Keep your high-level conceptual understanding that we developed by looking at the examples above.

So when we are analyzing a signal and we have the algebraic expression of our signal $y(t)$, i.e., we know for example $y(t) = \cos(t)$. We must transform this signal from time space t to reciprocal time space or frequency space f . To do this we must calculate the harmonic coefficients of our Fourier Series, A_n and B_n . These harmonic coefficient values at different integer values n will be critical to create our frequency spectrum. They will also allow us to represent any arbitrary signal $y(t)$ as an infinite sum of sines and cosines when we represent our original signal $y(t)$ as a Fourier Series.

That was a lot to take in — let's just focus for the moment on how to calculate our harmonic coefficients.

To calculate our harmonic coefficients we can perform the following analysis:

\$\$

$$A_n = \frac{\omega}{\pi} \int_0^{\frac{2\pi}{\omega}} y(t) \cos(\omega n t) dt \quad n = 0, 1, 2, \dots$$

\]

\$\$

$$B_n = \frac{\omega}{\pi} \int_0^{\frac{2\pi}{\omega}} y(t) \sin(\omega n t) dt \quad n = 0, 1, 2, \dots$$

\]

Where:

- A_n, B_n : harmonic coefficients
- $y(t)$: your original signal
- n : the order of the harmonic coefficient
- For example, if you are asked for the 3rd harmonic coefficient, compute A_3, B_3

You will integrate your original signal with respect to time and keep n in your solution. Then you will calculate the harmonic coefficients for increasing values of n until you sufficiently represent the signal of interest.

However, as you can see in the expression above, we have an issue — we do not want to work with our function in terms of ω ; we want to work with f . So we substitute using:

\$\$

$$\omega = 2\pi f$$

\]

Additionally, we typically analyze a function or signal for a finite period of time T . The lowest (fundamental) frequency of any signal is:

\$\$

$$f_{\text{fundamental}} = \Delta f = \frac{1}{T}$$

\]

So then:

\$\$

$$\omega = \frac{2\pi}{T}$$

\]

Plugging this into the harmonic coefficient expressions gives the form we will use in this course:

\$\$

$$A_n = \frac{2}{T} \int_0^T y(t) \cos\left(\frac{2\pi n t}{T}\right) dt \quad n = 0, 1, 2, \dots$$

\$\$

$$B_n = \frac{2}{T} \int_0^T y(t) \sin\left(\frac{2\pi n t}{T}\right) dt \quad n = 0, 1, 2, \dots$$

Once we have our harmonic coefficients, we calculate their magnitudes:

\$\$

$$C_n = \sqrt{A_n^2 + B_n^2}$$

This value C_n is what we will plot as the y-axis on the Frequency Spectrum, and the x-axis will be the corresponding frequency values — which are integer multiples of the fundamental frequency $f = \frac{1}{T}$.

That is a lot of information to take in, so let's do an example to make this clear...

Harmonic Coefficients, Frequency Spectrum, Dominant Frequency, Maximum Frequency

It's Hip to Be Square

We have the following curve of displacement and time. Calculate the Frequency Spectrum and the Fourier Series for this curve.

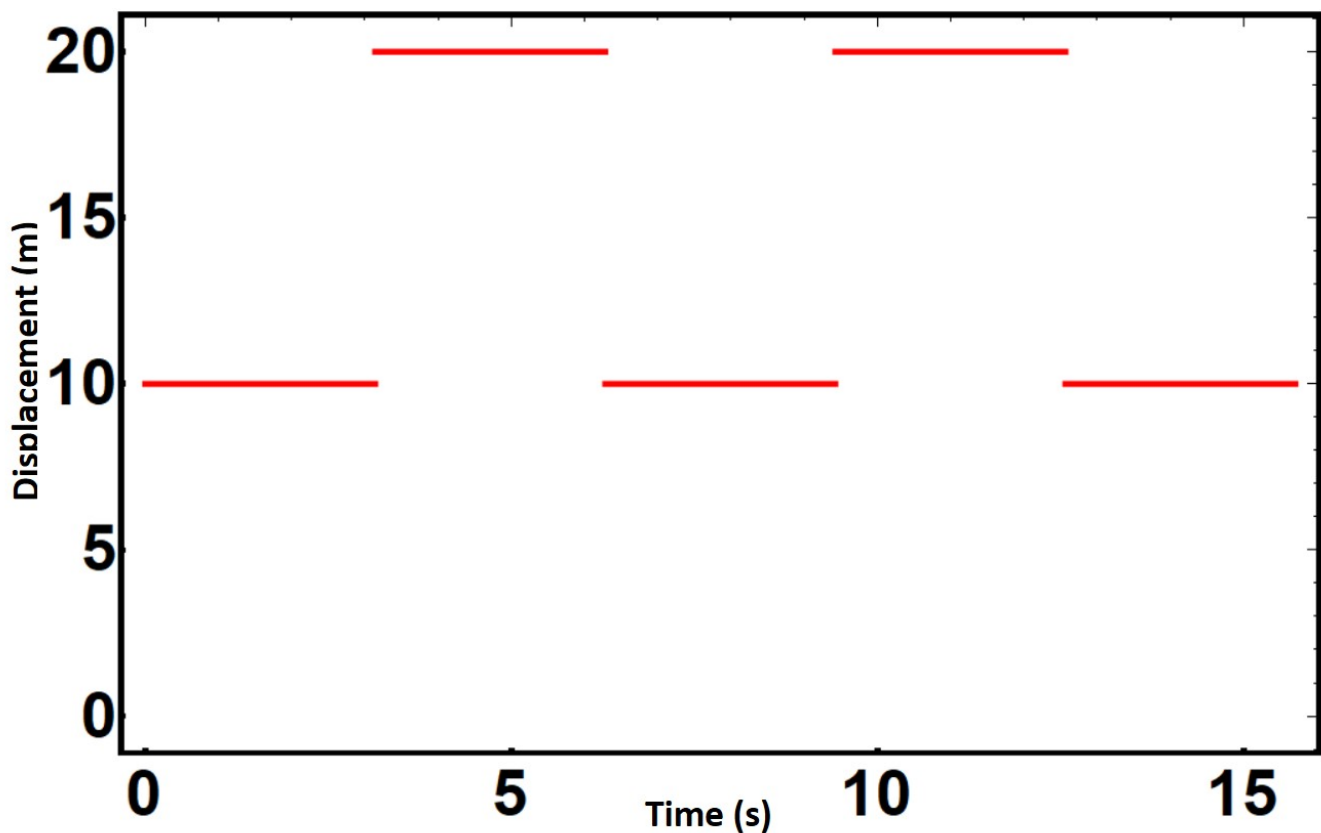


Figure Chapter4.10: Displacement vs. Time.

When looking at this curve it is a square wave, hence the Huey Lewis reference. Thus we can describe piece-wise the algebraic form of this function. Specifically we know that:

\$\$

$$y_{\{1\}}(t) = 10 \quad 0 < t < \pi \setminus$$

$$y_2(t) = 20 \quad \pi < t < 2\pi$$

\]

Now that we have our $y(t)$, we can calculate our harmonic coefficients A_n and B_n . So again we have:

\$\$

$$A_n = \frac{2}{T} \int_0^T y(t) \cos\left(\frac{2\pi n t}{T}\right) dt \quad n = 0, 1, 2, \dots$$

$$B_n = \frac{2}{T} \int_0^T y(t) \sin\left(\frac{2\pi n t}{T}\right) dt \quad n = 0, 1, 2, \dots$$

\]

Where:

- T is your period
- $y(t)$ is the signal
- n is the harmonic index

We see that the function is periodic with period $T = 2\pi$ seconds.

To calculate the Fourier coefficients, break the integration over the period T into two intervals:

\$\$

$$A_n = \frac{2}{T} \left[\int_0^\pi y_1(t) \cos\left(\frac{2\pi n t}{T}\right) dt + \int_\pi^{2\pi} y_2(t) \cos\left(\frac{2\pi n t}{T}\right) dt \right]$$

\]

Same for B_n :

\$\$

$$B_n = \frac{2}{T} \left[\int_0^\pi y_1(t) \sin\left(\frac{2\pi n t}{T}\right) dt + \int_\pi^{2\pi} y_2(t) \sin\left(\frac{2\pi n t}{T}\right) dt \right]$$

\]

Plugging in $y_1 = 10$, $y_2 = 20$, and $T = 2\pi$:

\$\$

$$A_n = \frac{1}{\pi} \left[\int_0^\pi 10 \cos(nt) dt + \int_\pi^{2\pi} 20 \cos(nt) dt \right]$$

\]

After integration:

\$\$

$$A_n = \frac{1}{\pi} \left[\frac{10 \sin(n\pi)}{n} + \frac{20 (\sin(2n\pi) - \sin(n\pi))}{n} \right]$$

\]

It turns out that $A_n = 0$ for all n .

Now solve for B_n :

\$\$

$$B_n = \frac{1}{\pi} (10(1 + \cos(n\pi)) - 2\cos(2n\pi))$$

\]

To generate the Frequency Spectrum, plot the amplitude of the harmonic coefficients:

\$\$

$$C_n = \sqrt{A_n^2 + B_n^2} = |B_n| \quad (\text{since } A_n = 0)$$

\]

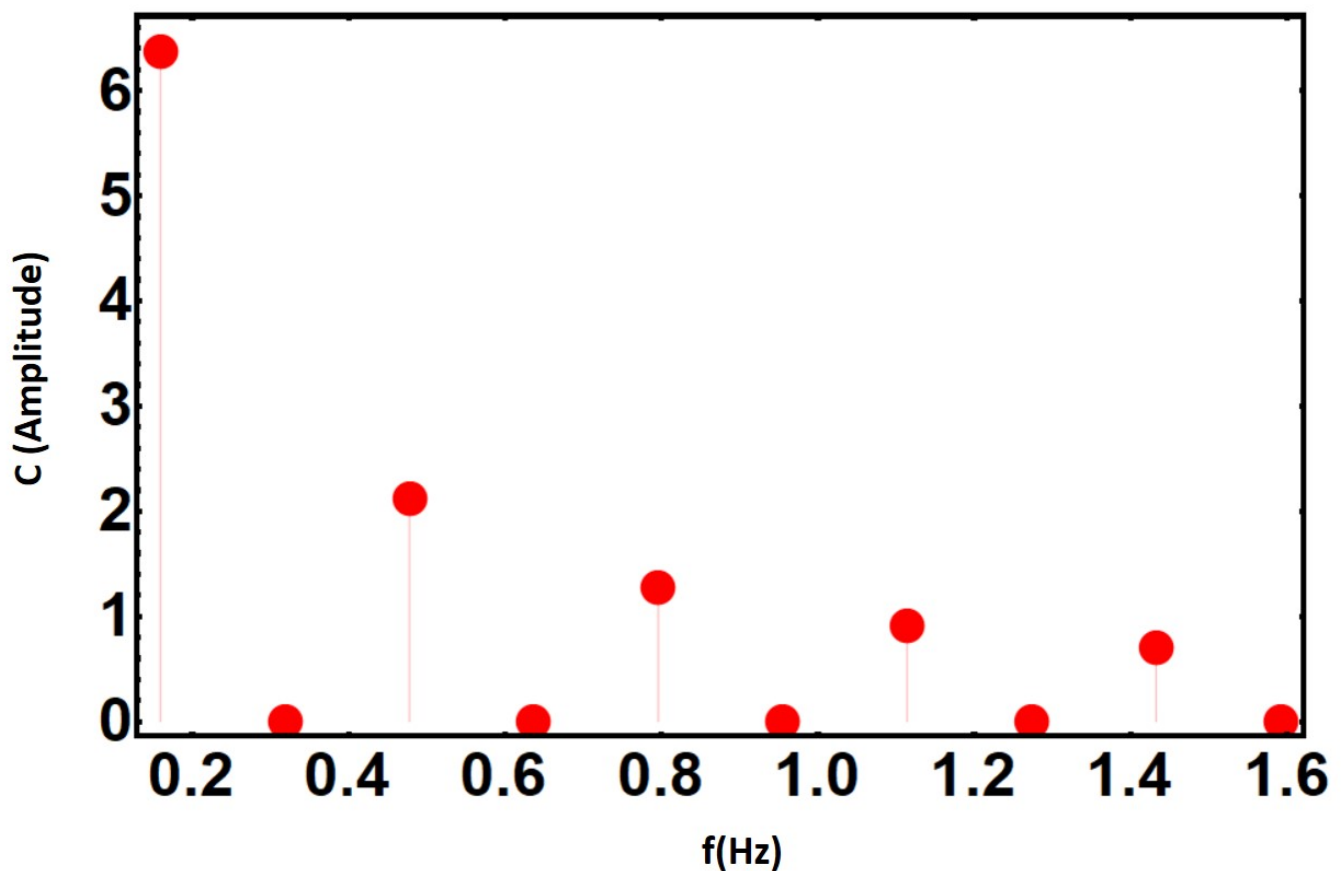


Figure Chapter4.11: Frequency Spectrum.

The x-axis is frequency: $f = \frac{1}{T} = \frac{1}{2\pi}$

So the horizontal axis will be nf for $n = 1, 2, 3, \dots$

You can also write the harmonic coefficients using this frequency format:

$$A_n = 2f \int_0^{1/f} y(t) \cos(2\pi n f t) dt$$

$$B_n = 2f \int_0^{1/f} y(t) \sin(2\pi n f t) dt$$

Although for this course, stick with:

$$A_n = \frac{2}{T} \int_0^T y(t) \cos\left(\frac{2\pi n t}{T}\right) dt$$

$$B_n = \frac{2}{T} \int_0^T y(t) \sin\left(\frac{2\pi n t}{T}\right) dt$$

Well that was awesome! We are becoming Fourier masters.

From this frequency spectrum we can determine several critical parameters moving forward. First, we can see from the graph the lowest or fundamental frequency of our signal by looking at the first non-zero frequency on our x-axis. NOTE: Sometimes people may include the 0th order harmonic on the frequency spectrum but that is not our fundamental frequency — that will be f , or the frequency associated with $n = 1$. So in this problem we can deduce that:

$$f = \frac{1}{2\pi}$$

And this value is in Hz. With the fundamental frequency in hand, just by looking at the Frequency Spectrum we can now also calculate the period T , which is just 2π seconds.

From this graph we can also calculate the maximum number of harmonics in this Fourier analysis, which will just be the number of points in our Frequency Spectrum — 10 harmonics or $n = 10$. So our maximum frequency in this signal will be $10f$.

We can also calculate our last parameter, which is a critically important parameter — the dominant frequency. This is the frequency associated with the largest magnitude of our harmonic coefficients, C_n . In this example, we can see from our Frequency Spectrum that the largest y-value in this instance corresponds to our lowest frequency. So the dominant frequency is also the fundamental frequency in this problem.

Now, why is finding the dominant frequency important? This is the frequency which dominates the behavior of our system. As we will see in lab, we can determine the natural frequency or resonant frequency of a material using this type of analysis. Furthermore, you can imagine a scenario that if you are analyzing some signal or applying some frequency to a bridge, and that frequency is the same as the resonant frequency of a material, you will get resonance — and bad, bad things will happen to that bridge.

This is the power of Fourier analysis! By analyzing our signal of interest, we can use this to make design choices to avoid catastrophic disasters and loss of life. Or we can stimulate stem cell growth or figure out the different vocal frequencies of men vs. women — whatever you want. Fourier is amazing and boundless in its applications.

Now we are not done yet... if only it were so easy... but you are at the finish line.

Aside from obtaining the Frequency Spectrum, we also want to represent our original signal $y(t)$ as an infinite sum of sines and cosines — i.e., a Fourier Series. Once we have our harmonic coefficients, we can represent our original function $y(t)$ as:

$$y_{fs}(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty} A_n \cos(n\omega t) + B_n \sin(n\omega t)$$

Where $y_{fs}(t)$ is our Fourier Series representation of our original signal.

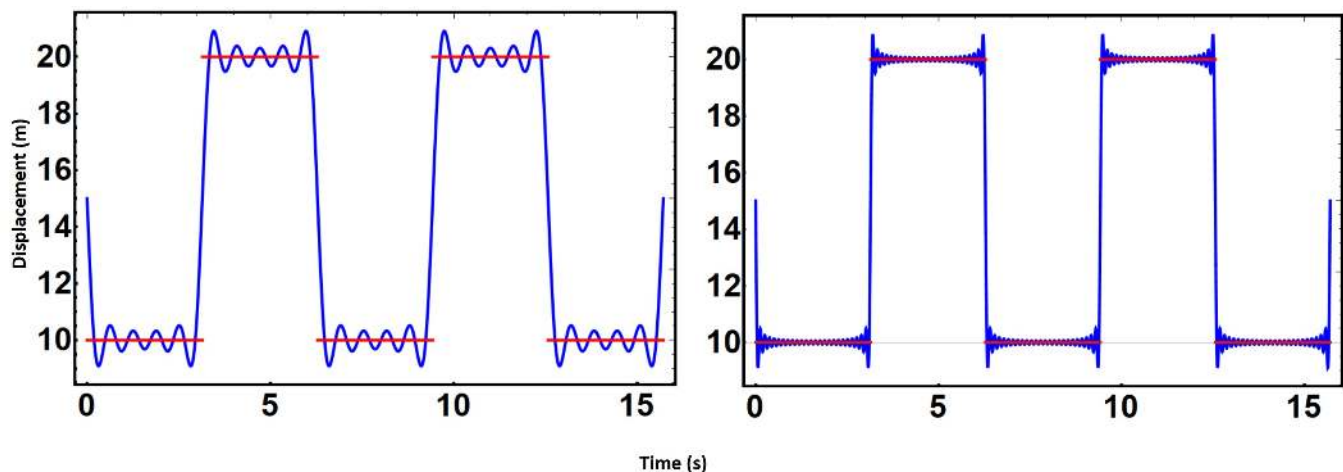


Figure Chapter4.12: Fourier Series Closer Approximates Original Signal (Red) As Harmonics are Increased from $n=10$ to $n=100$ (left to right respectively)

Now you can see the curve on the left where $n = 10$ harmonics are used is a pretty good fit to the original function. But if we increase the number of harmonics, the fit becomes much, much, much better — as seen on the right curve. The question you must ask is: for your particular application or design, is your approximation sufficient, or must you increase the number of harmonics?

Do You Remember Integration By Parts? Neither Do I — Please Help Mathematica!!

We have the following curve of displacement and time. Calculate the Frequency Spectrum and the Fourier Series for this curve. The math is a little trickier, so use Matlab or Mathematica.

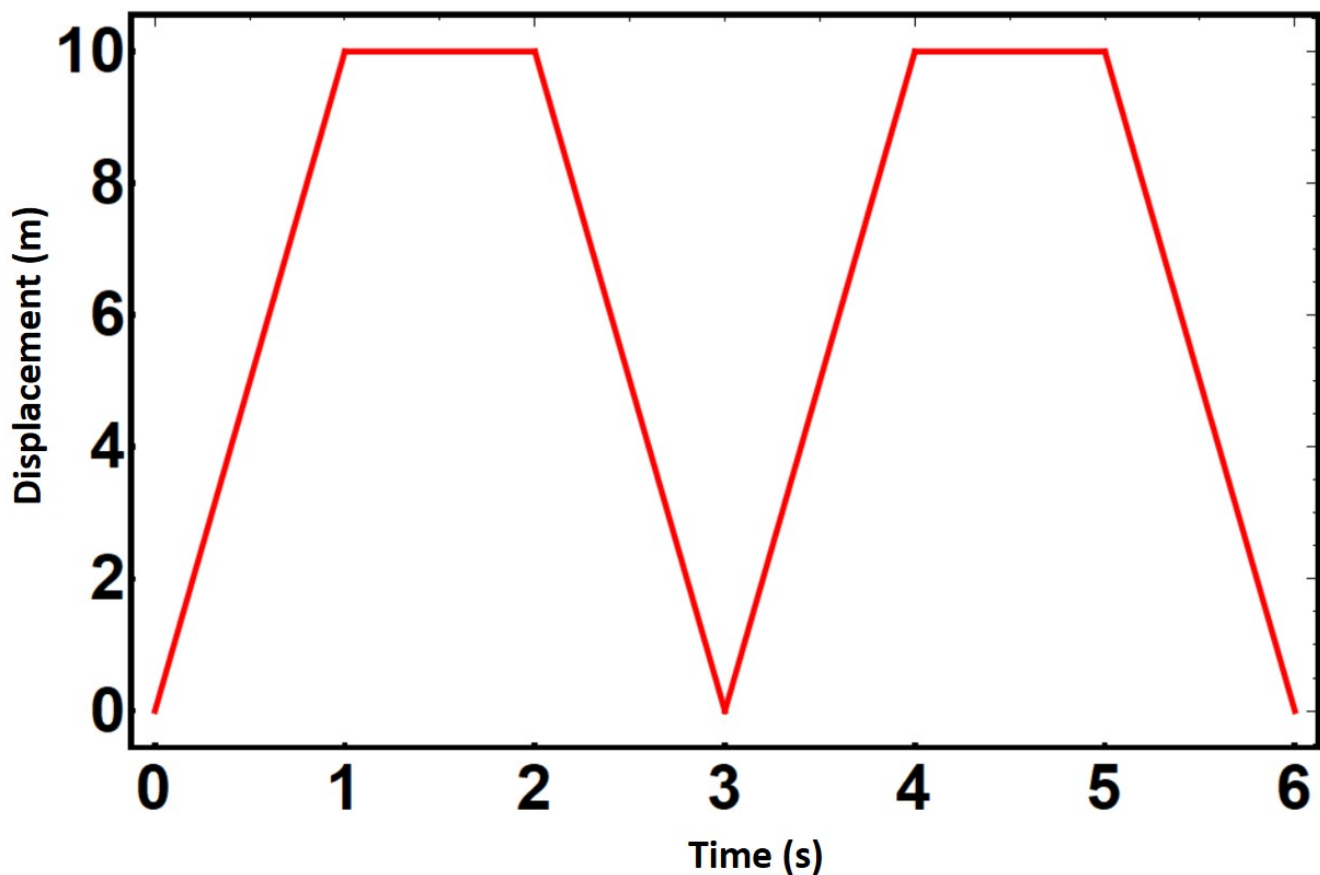


Figure Chapter4.13: Displacement vs. Time.

So the first thing you want to do is ask yourself what Fourier Case is this (i.e., Case 1, 2, 3, or 4)?

Well, looking back at your notes you should recognize that this is Case 1 — meaning that we can write out the exact algebraic equation for this curve. So if we have Case 1 and we are asked for the Frequency Spectrum, we need to calculate the harmonic coefficients A_n and B_n . For Case 1, those equations are:

\$\$

$$A_n = \frac{2}{T} \int_0^T y(t) \cos\left(\frac{2\pi n t}{T}\right) dt \quad n = 0, 1, 2, \dots$$

$$B_n = \frac{2}{T} \int_0^T y(t) \sin\left(\frac{2\pi n t}{T}\right) dt \quad n = 0, 1, 2, \dots$$

\]

Where:

- T is your period
- $y(t)$ is your signal
- n is the harmonic index

So the question becomes: how do we write the function of the displacement vs. time?

We see that the function is periodic with a period $T = 3$ seconds. Then we can write the equation for this period in three parts, which we will denote as:

\$\$

$$y_1(t) = 10t$$

$$y_2(t) = 10$$

$$y_3(t) = -10t + 30$$

\]

To calculate the Fourier coefficients, we break down our period T into three intervals:

\$\$

$$A_n = \frac{2}{T} \left[\int_0^1 y_1(t) \cos\left(\frac{2\pi n t}{T}\right) dt + \int_1^2 y_2(t) \cos\left(\frac{2\pi n t}{T}\right) dt + \int_2^3 y_3(t) \cos\left(\frac{2\pi n t}{T}\right) dt \right]$$

\]

The same structure applies to B_n :

\$\$

$$B_n = \frac{2}{T} \left[\int_0^1 y_1(t) \sin\left(\frac{2\pi n t}{T}\right) dt + \int_1^2 y_2(t) \sin\left(\frac{2\pi n t}{T}\right) dt + \int_2^3 y_3(t) \sin\left(\frac{2\pi n t}{T}\right) dt \right]$$

\]

Once we solve these integrals (by hand or using software), we can construct the Frequency Spectrum from the magnitudes:

\$\$

$$C_n = \sqrt{A_n^2 + B_n^2}$$

\]

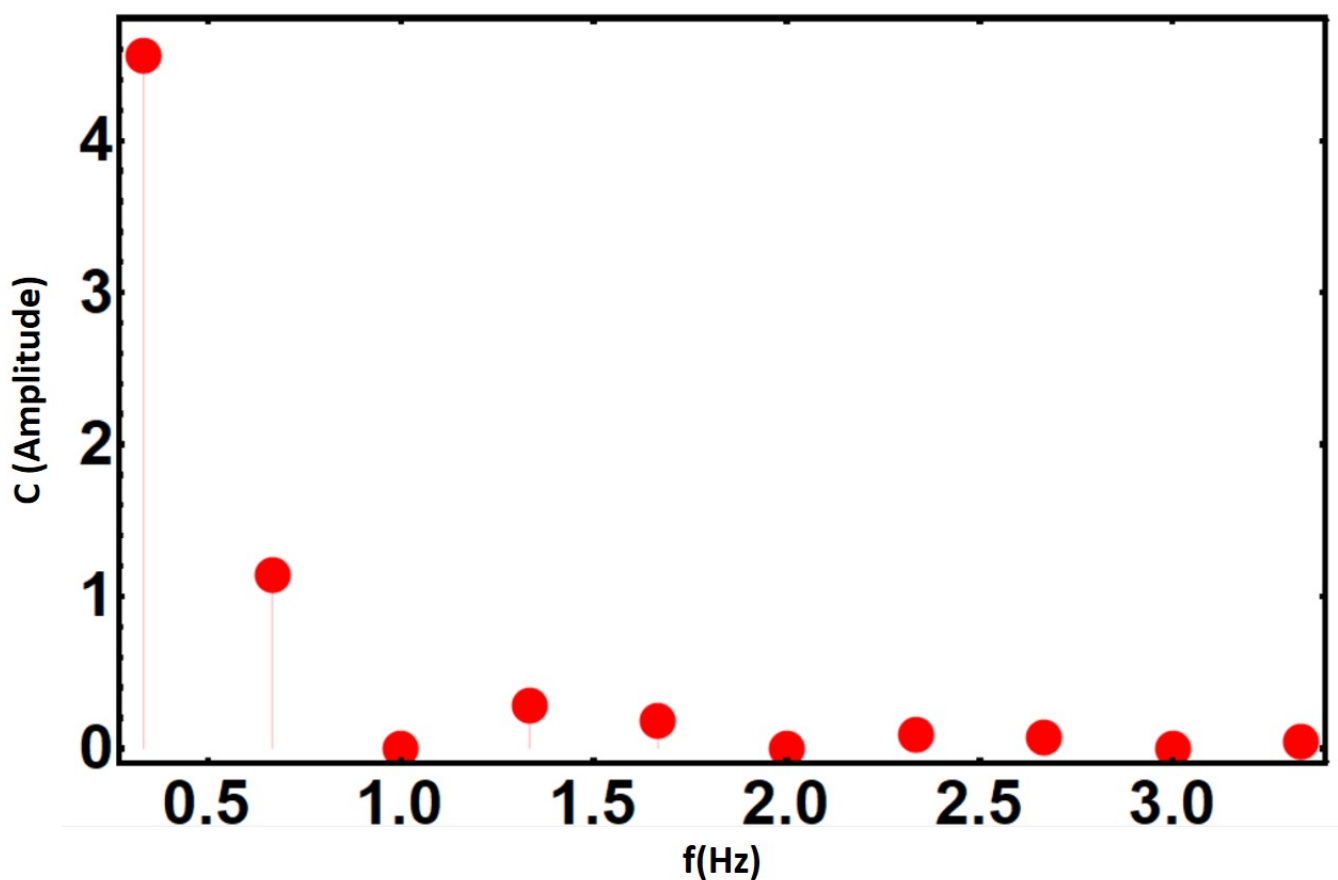


Figure Chapter4.14: Frequency Spectrum.

From this graph we can see that the dominant frequency — the frequency with the maximum amplitude — is also the fundamental frequency. In this case:

\$\$

$$f = \frac{1}{3} \text{ Hz}$$

\]

Now we can plot the Fourier Series since we have solved for A_n and B_n , by simply plugging those values into:

\$\$

$$y(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty} A_n \cos(n \omega t) + B_n \sin(n \omega t)$$

\]

And once we do that, we can overlay this reconstructed signal onto the original curve — and it's a really great fit.

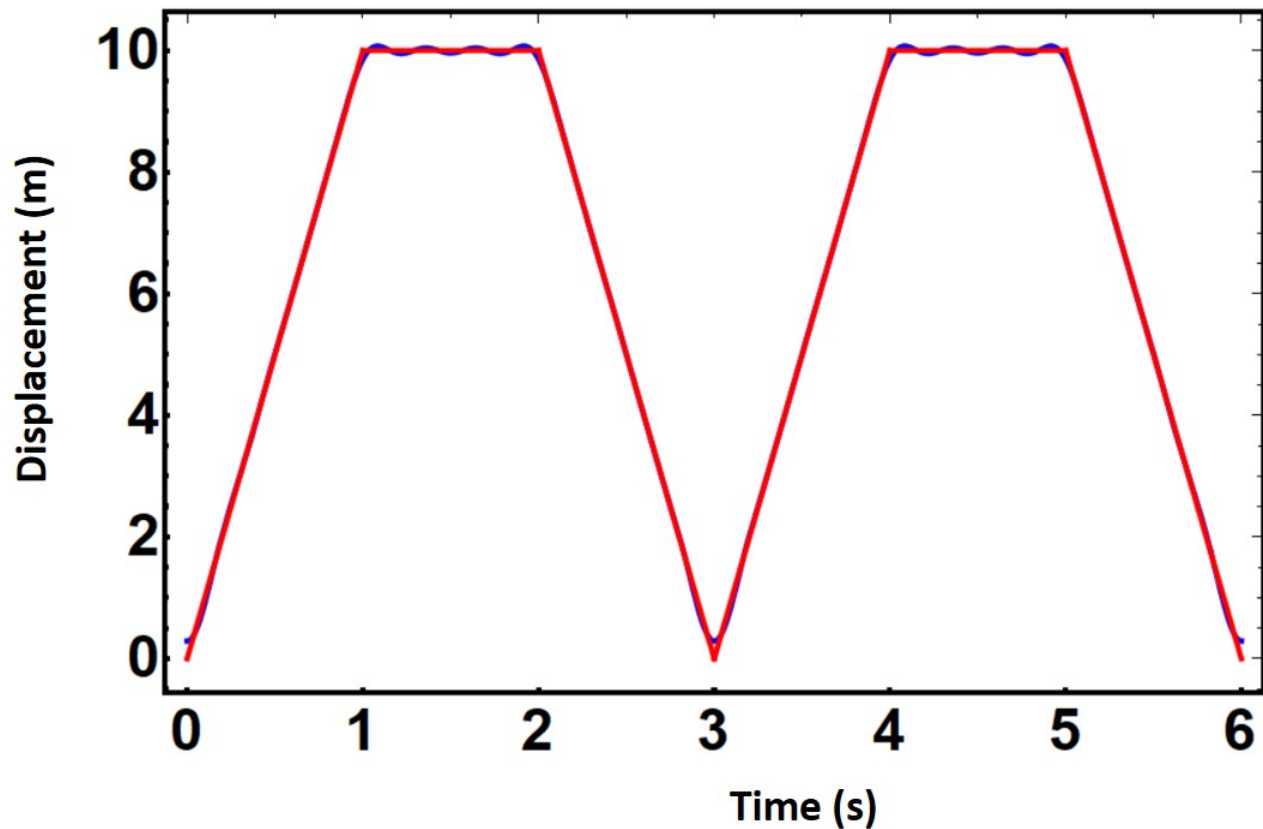


Figure Chapter4.15: Frequency Series.

The thing I want to illustrate is that our Fourier Series better approximates the original curve as we add more harmonics. However, typically we don't know the algebraic form of the equation that we are analyzing. What do we do then?

Be sure to tune in next week for another exciting episode as we tackle DFT, Aliasing, and Spectral Leakage in ENGR464.

This page titled [Chapter 4: Fourier Analysis and Continuous Fourier Transforms](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 5: Discrete Fourier Transforms, Nyquist Frequency, Signal Aliasing, and Spectral Leakage

Discrete Fourier Transform No More Integrals...Now Summations

So far you have become masters of the Continuous Fourier Transform or CFT. But we do not get too cocky just yet because this is not often the case we will deal with experimentally. I know I know I am so mean.

But in reality we are often measuring a signal and then trying to analyze the properties of the signal so we do not know the functional form of $y(t)$. Instead we will design an experimental apparatus to measure the signal — whether that be voltage, current, displacement, etc. — as a function of time and acquire, with a data acquisition device (DAQ) typically, the measurements as a function of time. Thus we will acquire a discrete data set as a function of time like seen below:

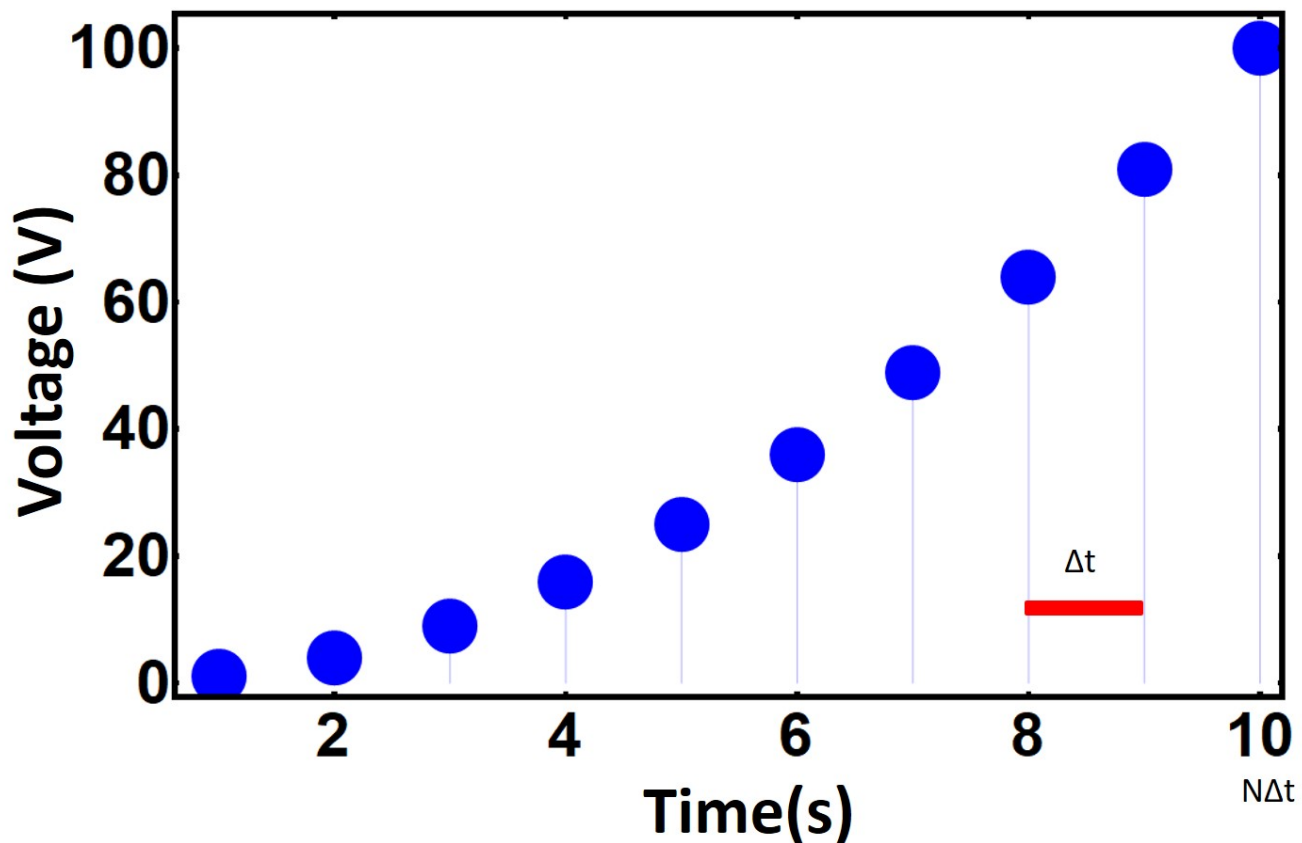


Figure Chapter5.1: Discrete Signal.

So what we will end up measuring is something like the graph above where we measure some total number of points or measurands N over some period of time T . This data will also be collected or sampled at a very specific frequency called the sampling frequency f_s . We can then determine the time intervals between each gathered sample, Δt , by the following relationship:

$$f_s = \frac{1}{\Delta t}$$

We can then also define:

$$T = N \Delta t = \frac{N}{f_s}$$

In the example above we know that $T = 10$ s, $N = 10$, $f_s = 1$ Hz, and $\Delta t = 1$ s. We can pull all that information just from looking at this curve! Amazing.

Now one question that you might have as an experimentalist is how do I choose the appropriate f_s , N , and T to capture my signal of interest? What an excellent question, and that leads us to two critical concepts in Discrete Fourier Transforms (DFT): signal aliasing and spectral leakage.

Signal Aliasing

Let's first discuss how we should choose an appropriate N and f_s in order to capture our signal.

Look at the graph below. The original signal is shown in blue, while the points I extract are shown in red. What is wrong with my experimental design in terms of my selection of N and f_s ? What are the consequences here?

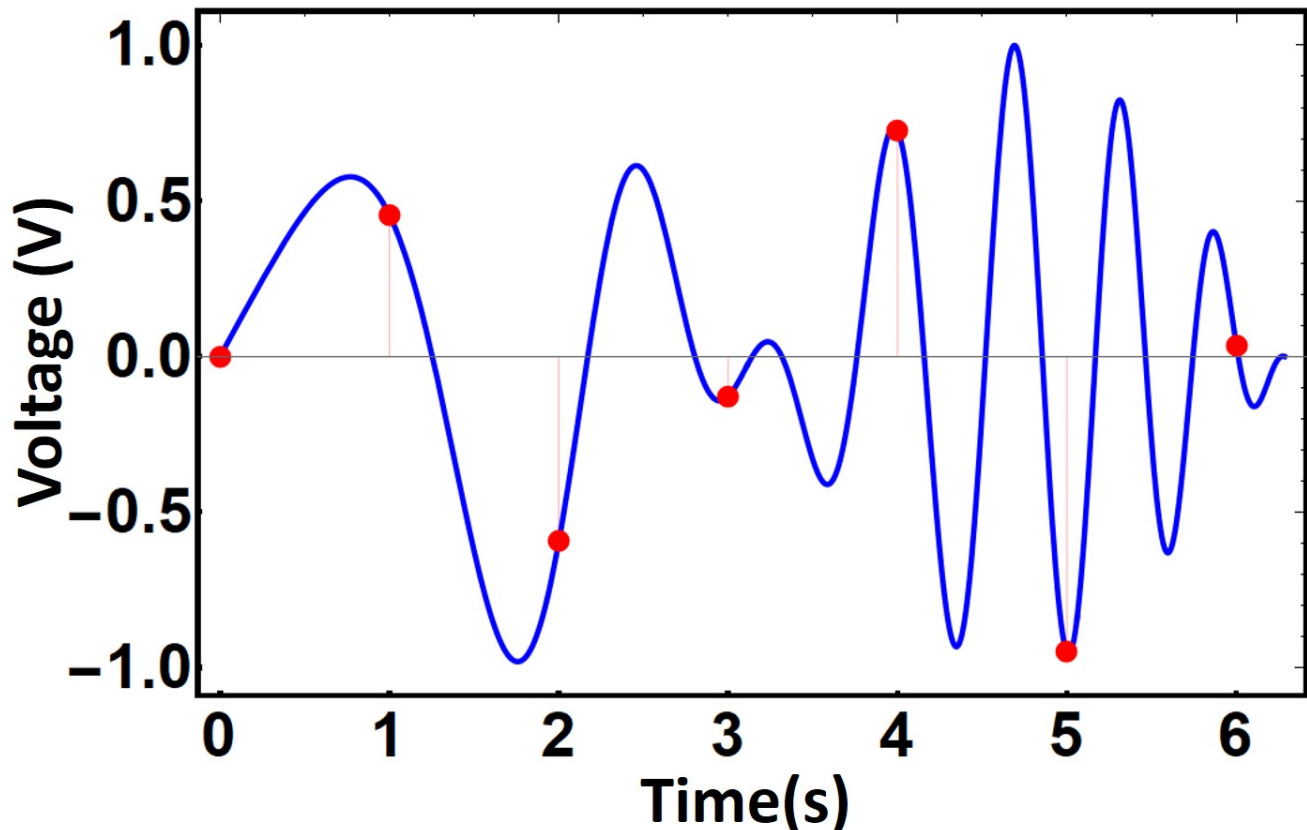


Figure Chapter5.2: f_s not sufficiently high to properly sample the signal.

Clearly, we have chosen a f_s that is not large enough to capture the properties of the original signal. What frequencies are we missing? We are missing the high-frequency components of the original signal.

The critical issue with choosing a sampling frequency that under-samples the original signal is that now, as experimentalists, we are under the mistaken impression that the signal we are measuring has a lower and different frequency than the actual signal. This is the first example of a concept that we will delve into much deeper — signal aliasing — where the sampled signal takes on a different frequency than the original signal.

You can imagine the consequences of this if you are designing a structural component in a bridge that has a particular resonant frequency which may match the vibrations present in an application. We must know the real properties of our signal.

What is happening with our sampling below here?

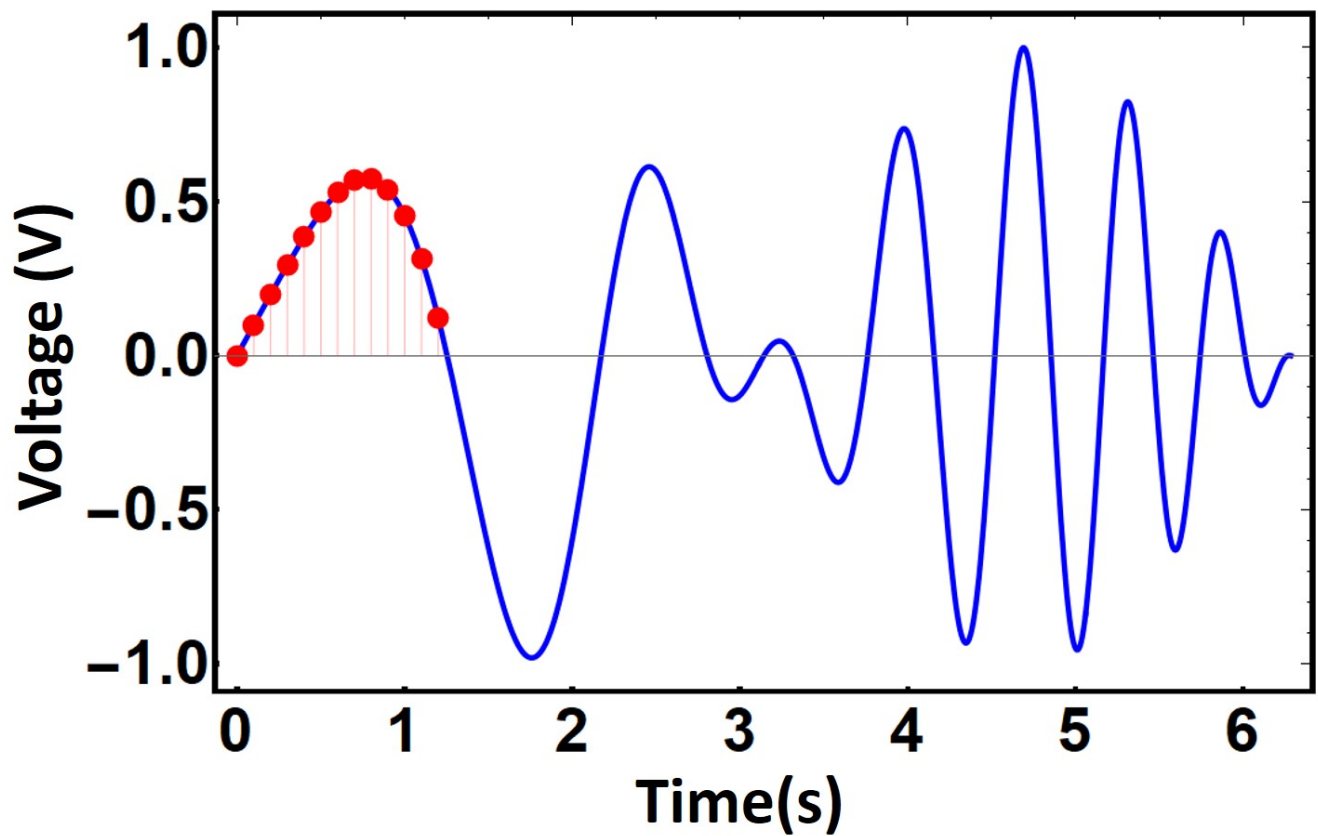


Figure Chapter5.3: N not large enough to capture entire signal.

Now we do not have a large enough N value to accurately sample our signal. Now we are missing out on the low frequencies of our signal, and again thus aliasing our signal. We must choose a large enough f_s and N in order to avoid sampling frequency issues and properly characterize our signal of interest — as shown correctly below.

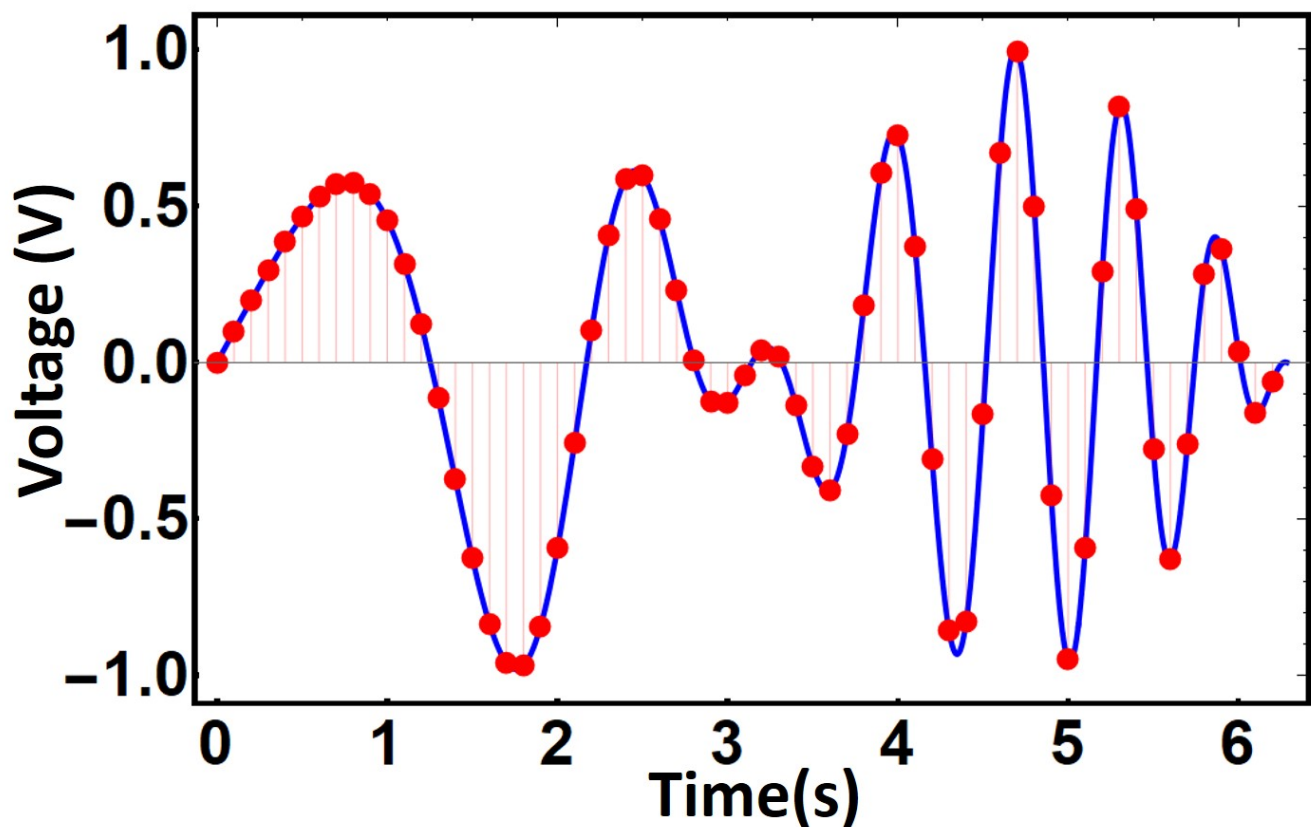


Figure Chapter5.4: Properly sampled signal f_s and N large enough.

Now you may be asking: if we are measuring an unknown signal in an experiment (which is typically the case), how can we select our f_s and N values?

That is yet another great question. The answer is: typically, you will have to do some literature review or additional experiments to try and determine some properties of the signal ahead of time. Also, you can always just run your experiment at the maximum values your DAQ can handle, i.e., largest f_s and N . If you are concerned about low- or high-frequency noise influencing your results, you can use low-pass or high-pass filters to get rid of certain frequencies.

Real experiments are hard.

Yet another question you may be asking is: how can I quantitatively avoid signal aliasing if I know the frequency of the signal I'm trying to measure?

Well, let's do this by looking at a sine function specifically:

$$y(t) = \sin(\pi t)$$

What is the frequency of my signal?

We know that $\omega = \pi$, and using the relationship $\omega = 2\pi f$, we find:

$$f_{\text{signal}} = \frac{1}{2} \text{ Hz}$$

Now let's consider what happens when we sample this signal at 0.5 Hz.

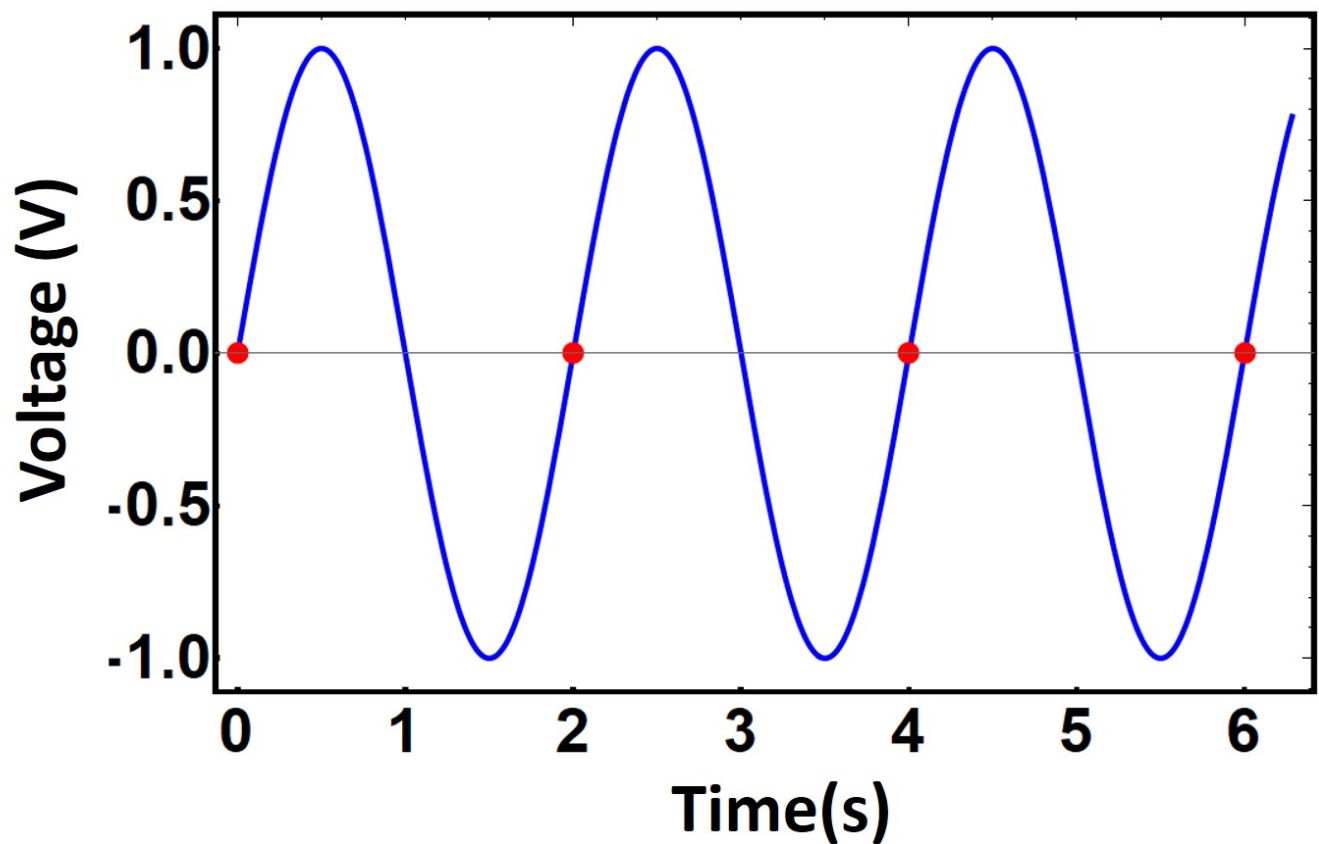


Figure Chapter5.5: Biased signal with an apparent frequency of 0Hz.

That signal is not correct — and if we look at our sampled points (red), we see the apparent frequency of our sampled signal is 0 Hz. So we have indeed aliased our signal.

We must make sure that our sampling frequency is greater than the frequency of our signal.

Let's increase the sampling frequency to $f_s = 0.75$ Hz. What happens now?

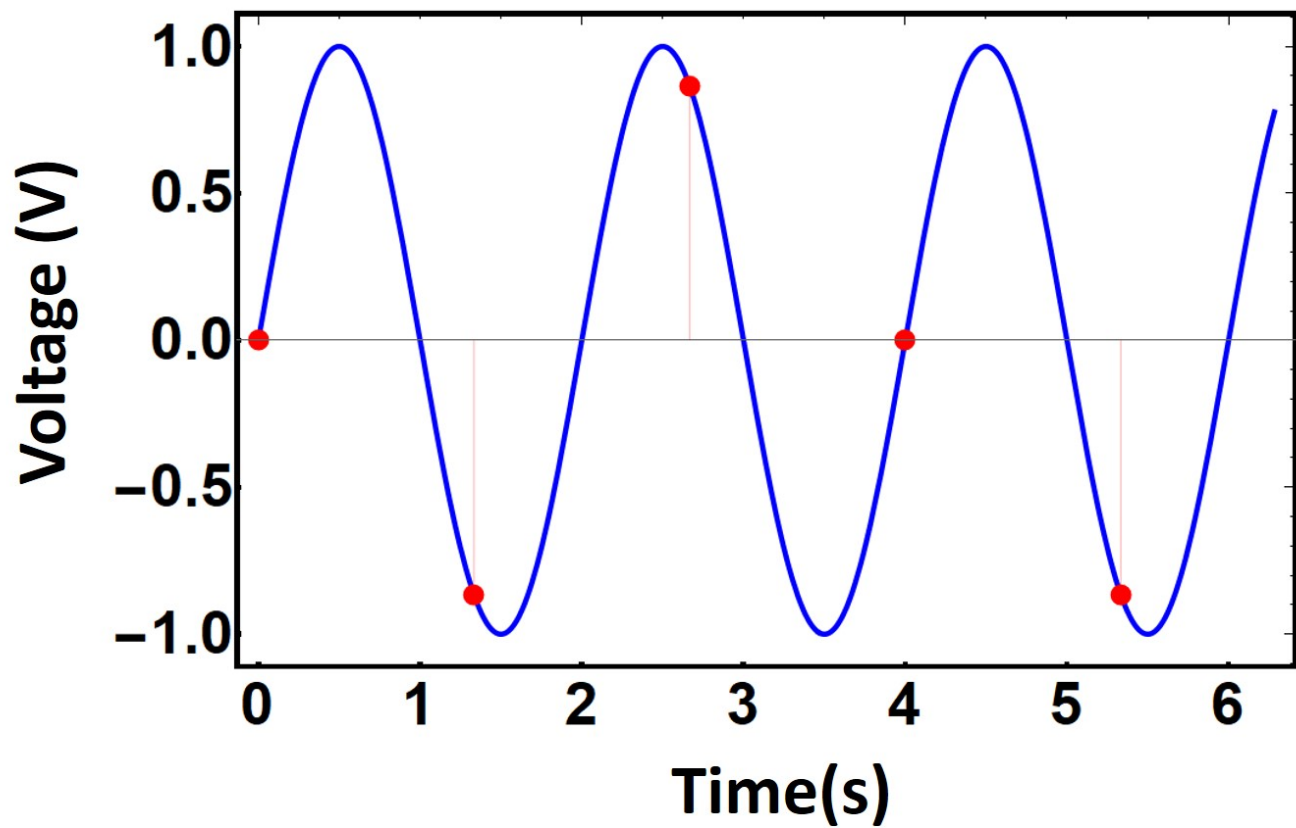


Figure Chapter5.6: Biased signal with an apparent frequency of 0.25Hz.

Nope — still haven't sampled correctly. We can clearly see that the apparent frequency is approximately 0.25 Hz, which again is not the frequency of our signal.

Faster! We must sample faster.

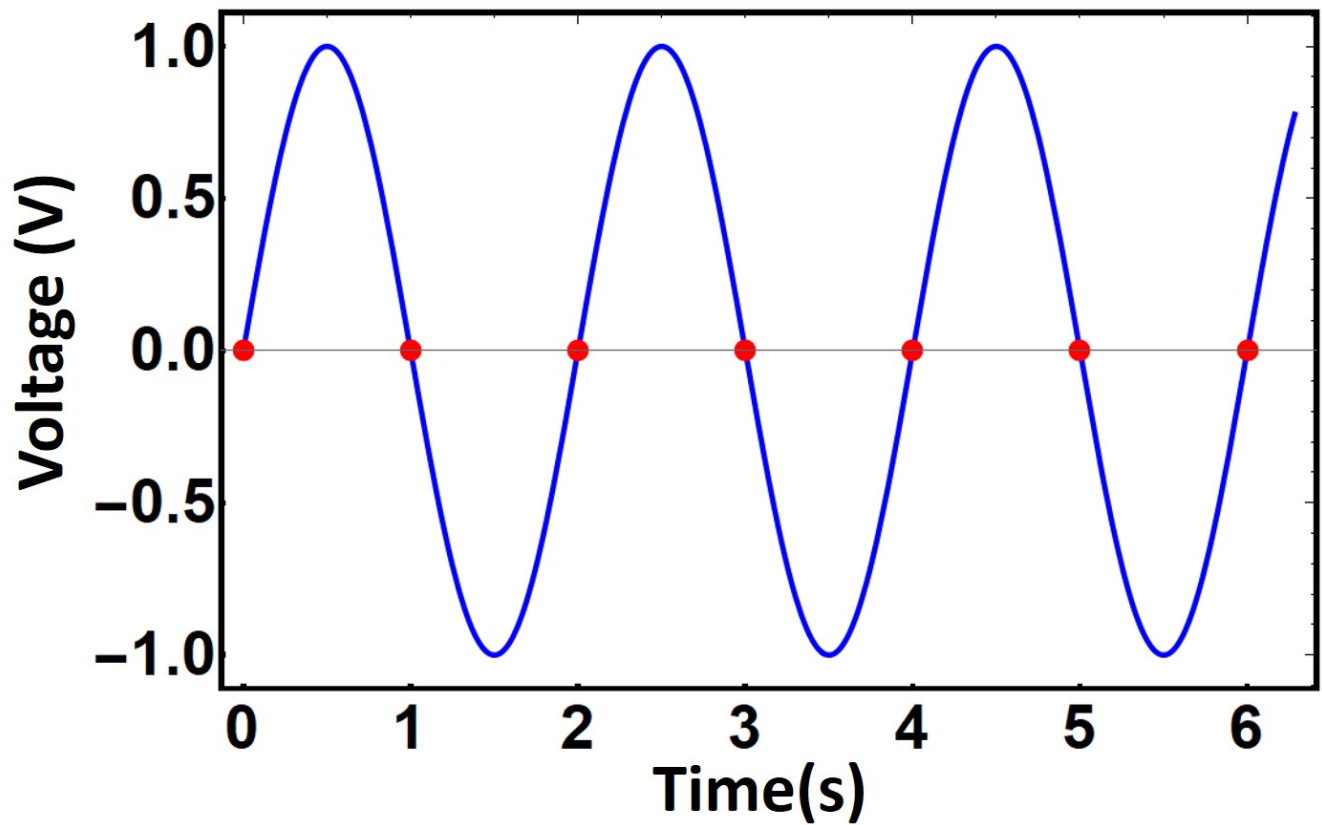


Figure Chapter5.7:Biased signal with an apparent frequency of 0Hz.

Even when we sample at exactly twice the signal frequency, we can still bias our signal — the apparent frequency here is again 0 Hz.

Can we ever properly sample our signal?

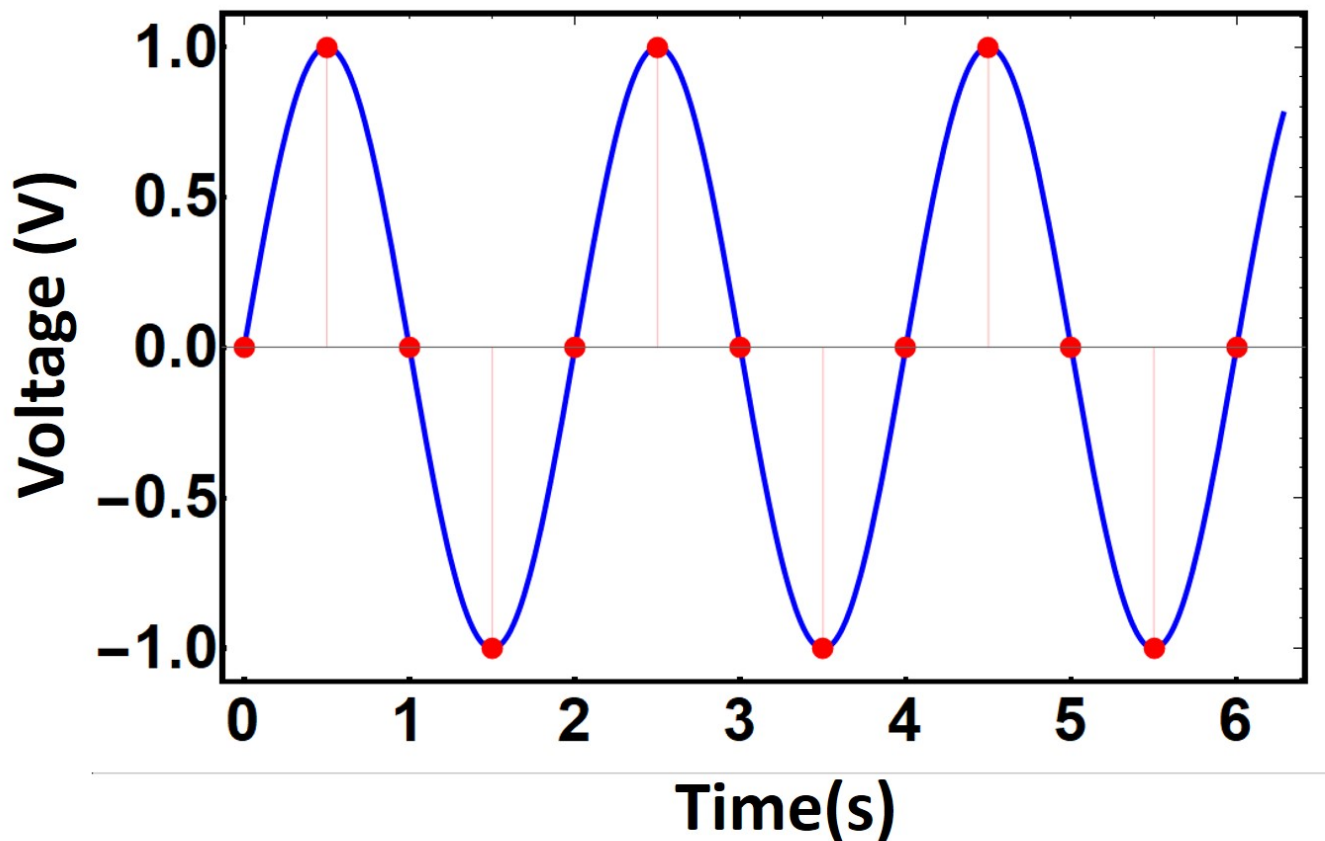


Figure Chapter5.8: Properly sampled signal with an apparent frequency of 0.5Hz.

Finally — once we choose a sampling frequency such that:

$$f_s > 2 f_{\text{max signal}}$$

Then we satisfy the Nyquist Criterion, and our signal is correctly sampled. The apparent and actual frequencies match, and we can proceed with our analysis.

Discrete Fourier Transformations: Summations and Harmonic Coefficients

Now that we are aware of the dangers of signal aliasing, we can start to perform Discrete Fourier Transforms (DFT) of signals in order to produce a Frequency Spectrum and deduce some properties of the signal of interest.

To perform a DFT, we need to convert the integrals that we have all come to love:

$$A_n = \frac{2}{T} \int_0^T y(t) \cos\left(\frac{2\pi n t}{T}\right) dt$$

$$B_n = \frac{2}{T} \int_0^T y(t) \sin\left(\frac{2\pi n t}{T}\right) dt$$

... into summations — since we are working with a discrete data set.

Let's get started.

We no longer have a continuous function $y(t)$, but instead a discrete function $y(r\Delta t)$, where r varies from 1 to N . It is just an integer counter of our discrete data points.

We know that:

\$\$

$$T = N \Delta t$$

\]

We replace the integral with a sum, and replace dt with Δt :

\$\$

$$A_n = \frac{2}{N \Delta t} \sum_{r=1}^N y(r \Delta t) \cos \left(\frac{2 \pi n r \Delta t}{N \Delta t} \right) \Delta t$$

\]

This simplifies to:

\$\$

$$A_n = \frac{2}{N} \sum_{r=1}^N y(r \Delta t) \cos \left(\frac{2 \pi r n}{N} \right)$$

\]

So the harmonic coefficients in DFT become:

\$\$

$$A_n = \frac{2}{N} \sum_{r=1}^N y(r \Delta t) \cos \left(\frac{2 \pi r n}{N} \right) \backslash$$

$$B_n = \frac{2}{N} \sum_{r=1}^N y(r \Delta t) \sin \left(\frac{2 \pi r n}{N} \right) \backslash$$

\]

Where $n = 0, 1, \dots, N/2$

Why do we stop at $\frac{N}{2}$? Because of the **Nyquist frequency**, defined as:

\$\$

$$f_{\text{Nyq}} = \frac{f_s}{2} = \frac{1}{2 \Delta t}$$

\]

This is the **maximum frequency** that can be resolved in a Frequency Spectrum graph. Frequencies above the Nyquist frequency will appear as mirror images — we'll demonstrate this next.

The lowest frequency in your spectrum is the fundamental:

\$\$

$$\Delta f = f_{\text{fundamental}} = \frac{1}{T} = \frac{1}{N \Delta t} = \frac{f_s}{N} = \frac{2 f_{\text{Nyq}}}{N}$$

\]

That gives us everything we need to start computing DFT results.

Example: DFT of Oscilloscope Signal

Let's look at this signal that I produced on my oscilloscope:

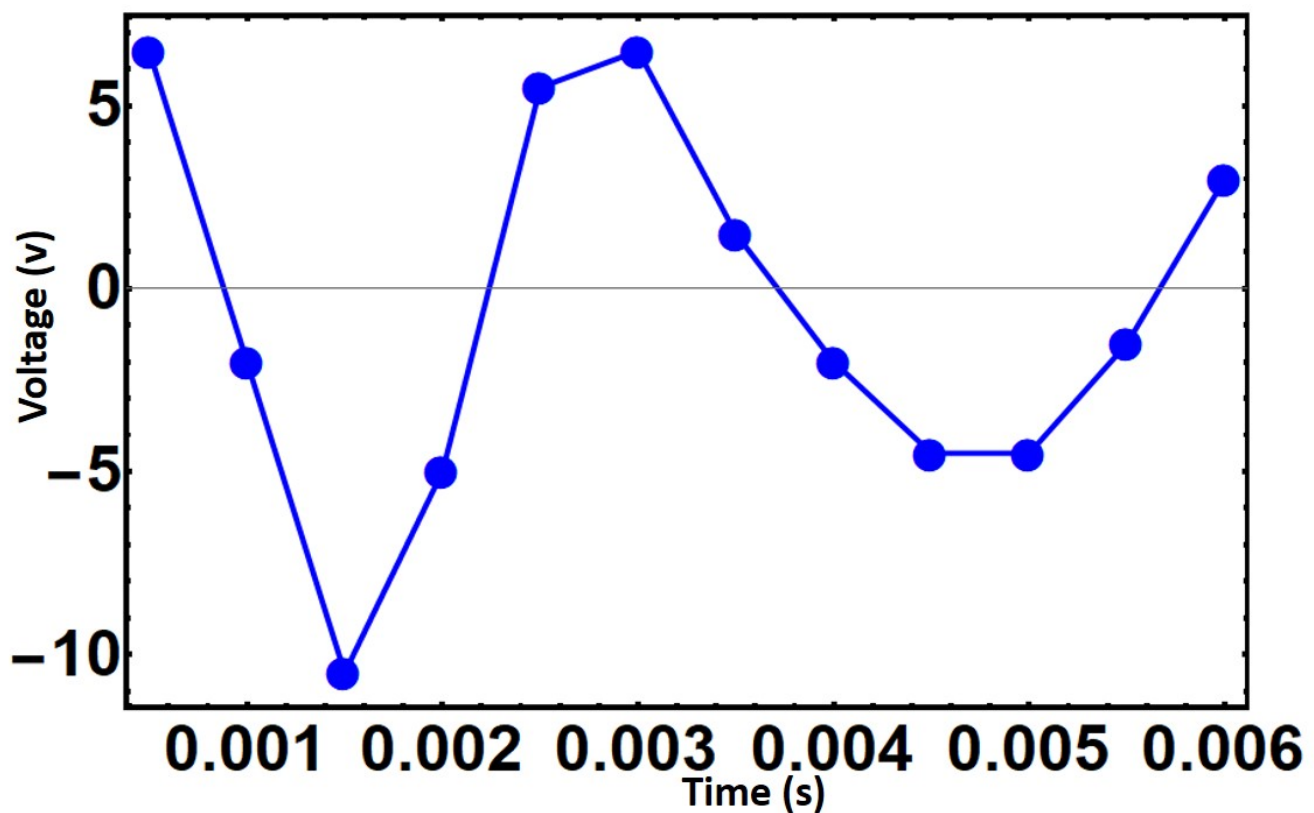


Figure Chapter5.9: Oscilloscope signal.

You can see the raw data in the Mathematica notebook, and the data is collected every 0.5 ms.

So how should we start solving this problem? What else can we identify immediately?

We see that:

- $N = 12$
- $\Delta t = 0.5 \text{ ms}$
- $T = 6 \text{ ms}$

That's a really good start.

We also know:

- $f_s = 2000 \text{ Hz}$
- $f_{\text{Nyq}} = 1000 \text{ Hz}$
- $\Delta f = 167 \text{ Hz}$

Now let's calculate the DFT for this signal. To do that, we need the harmonic coefficients:

\$\$

$$A_n = \frac{2}{N} \sum_{r=1}^N y(r \Delta t) \cos \left(\frac{2\pi r n}{N} \right) \backslash$$

$$B_n = \frac{2}{N} \sum_{r=1}^N y(r \Delta t) \sin \left(\frac{2\pi r n}{N} \right) \backslash$$

$$\backslash$$

Look at the Mathematica notebook to convince yourself how the code works — you never know if I will ask you to calculate a harmonic coefficient by hand!

We can also see the symmetry of the harmonic coefficients beyond the Nyquist frequency — the values are symmetric, so we are not seeing higher frequencies, just mirror images of the lower ones.

Here is the DFT frequency spectrum for this signal:

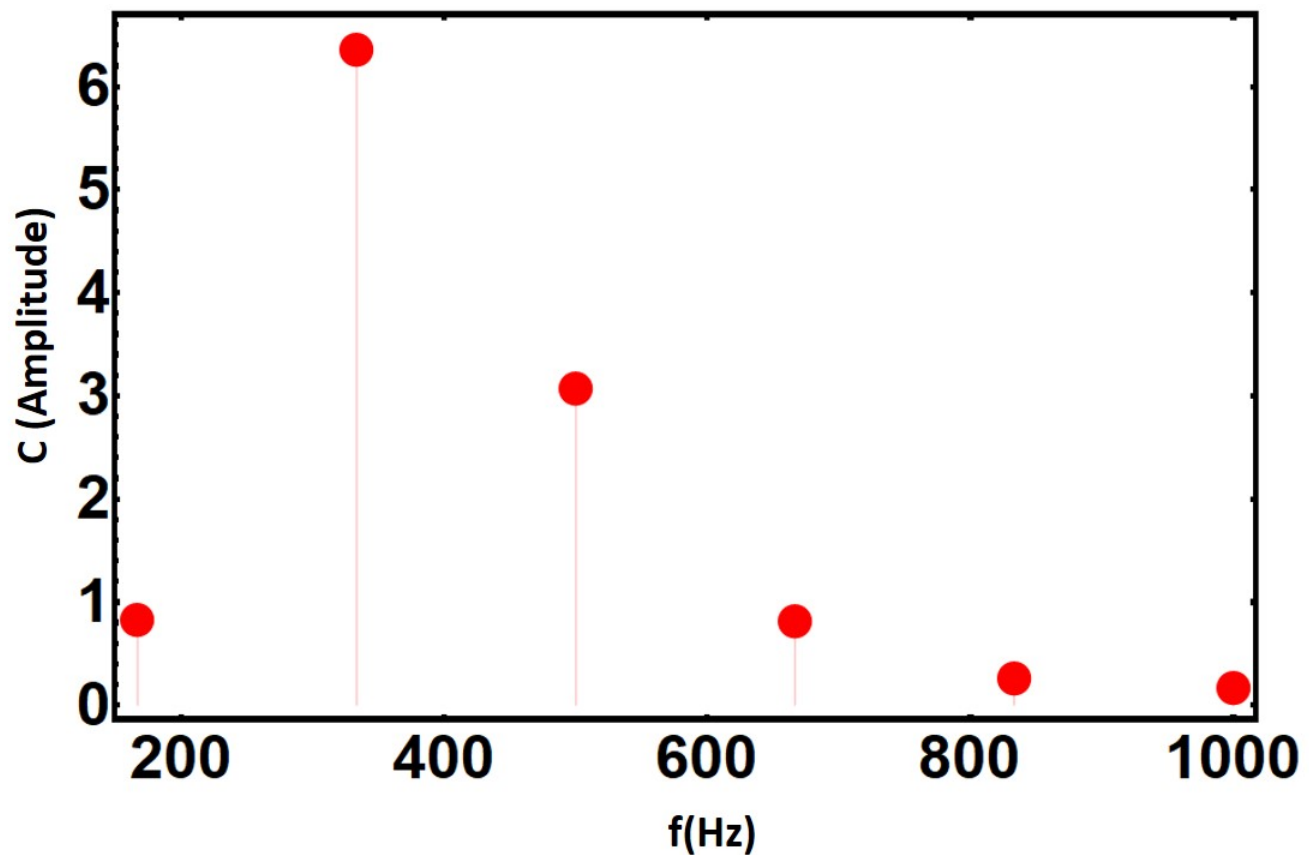


Figure Chapter5.10: DFT Fourier Spectrum.

Note again I have converted the x-axis into cyclic frequency.

How do I get these values? What is my Δf spacing?

It is simply the lowest frequency or the fundamental frequency, which we already calculated: 167 Hz.

So what is the highest frequency in this spectrum? It is $f_{\text{Nyq}} = 1000$ Hz, as confirmed here.

What is the dominant frequency? It is $2\Delta f = 333.33$ Hz

Spectral Leakage

Now let's say that I read an article and it tells me that the signal I'm measuring contains a critically important low-frequency component at 300 Hz. I need to measure this frequency — it's vital to my application.

Did I capture that in this experiment?

No.

My lowest frequency is 167 Hz, and the next frequency I measure is 333.33 Hz. That's a poor frequency resolution — another term for Δf .

Because I cannot hit 300 Hz directly, that frequency will leak into adjacent bins — the 167 Hz bin and the 333 Hz bin. This is known as spectral leakage, and it is a big problem.

From the spectrum, we see the dominant frequency is 333.33 Hz — but is that really from 333 Hz, or is it contaminated by 300 Hz?

If we want to resolve the 300 Hz frequency directly, we need better resolution.

Improving Frequency Resolution

To get a resolution of 1 Hz, we use:

\$\$

$$\Delta f = \frac{f_s}{N}$$

\]

Suppose $f_s = 2000$ Hz must stay fixed. Then:

\$\$

$$N = \frac{f_s}{\Delta f} = \frac{2000}{1} = 2000$$

\]

We would need to collect 2000 samples to get a resolution of 1 Hz. Easy as that!

DFT Fourier Series

We've done the hard part — we solved for the harmonic coefficients. Now we want to reconstruct the signal.

The formula to use is:

\$\$

$$y(t) = \frac{A_0}{2} + \sum_{n=1}^{N/2} \left[A_n \cos(2\pi n \Delta f t) + B_n \sin(2\pi n \Delta f t) \right]$$

\]

Let's plot that series using our coefficients:

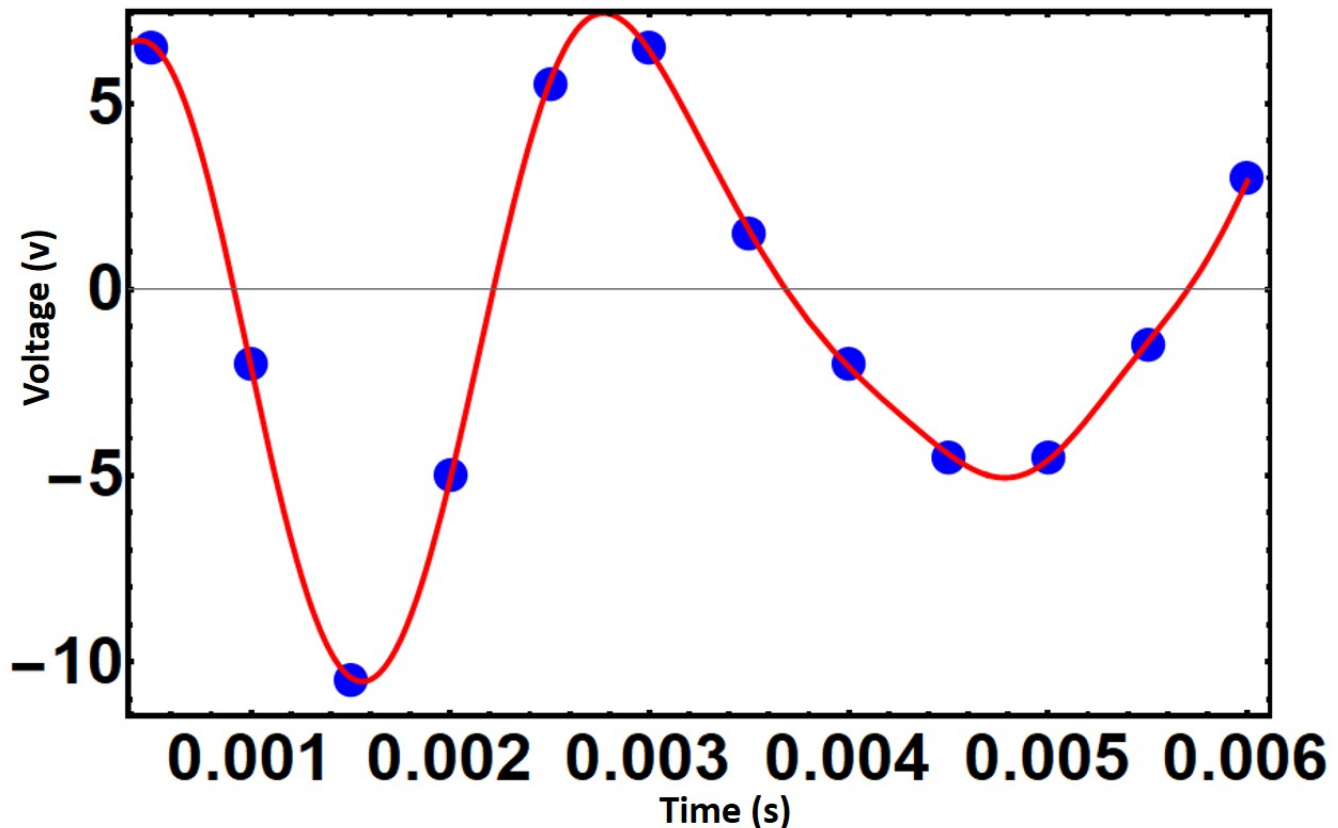


Figure Chapter5.11: DFT Fourier Series.

- Blue points: experimental data
- Red line: Fourier series reconstruction from DFT

It's a good fit! We sampled the curve well enough to reproduce the data.

If the Fourier series does not approximate the signal well, what do we do?

We must add more harmonics, which in this case means collecting more data points.

Another Example: Oscilloscope Signal

Let's look at yet another example that I recorded on my oscilloscope:

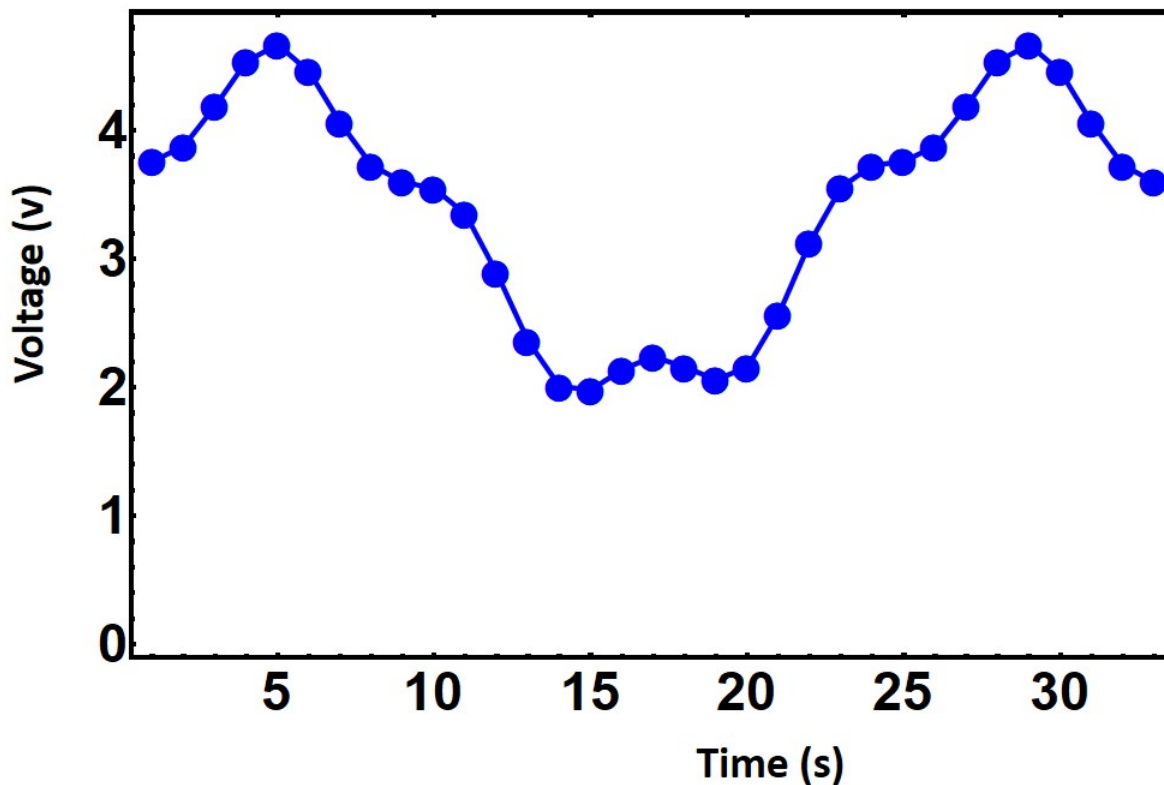


Figure Chapter5.12: Oscilloscope signal.

You can see the raw data in the mathematica notebook and the data is collected every 1ms. So how should we start solving this problem? What else can we identify immediately that is different from before? This signal is periodic!!! Specifically it repeats over a period, $T = 24 \text{ ms}$. So that will be our period and we will only select those points in that period for our DFT.

So we know as well that:

- $N = 24$
- $\Delta t = 1 \text{ ms}$
- $T = 24 \text{ ms}$

That's a really good start.

Additionally, we also know:

- Sampling frequency: $f_s = 1000 \text{ Hz}$
- Nyquist Frequency: $f_{\text{Nyq}} = 500 \text{ Hz}$
- Fundamental frequency of the DFT: $\Delta f = 42 \text{ Hz}$

Now let's calculate the DFT for this signal. To do that we need the Harmonic Coefficients, and instead of integrals we need to use summations:

\$\$

$$A_n = \frac{2}{N} \sum_{r=1}^{N/2} y(r \Delta t) \cos\left(\frac{2\pi r n}{N}\right) \quad n = 0, 1, \dots, \frac{N}{2}$$

\]

\$\$

$$B_n = \frac{2}{N} \sum_{r=1}^N y(r \Delta t) \sin\left(\frac{2\pi r n}{N}\right) \quad n = 1, 2, \dots, \frac{N}{2} - 1$$

The same math follows as the previous example. Look at the mathematica notebook to check your answers.

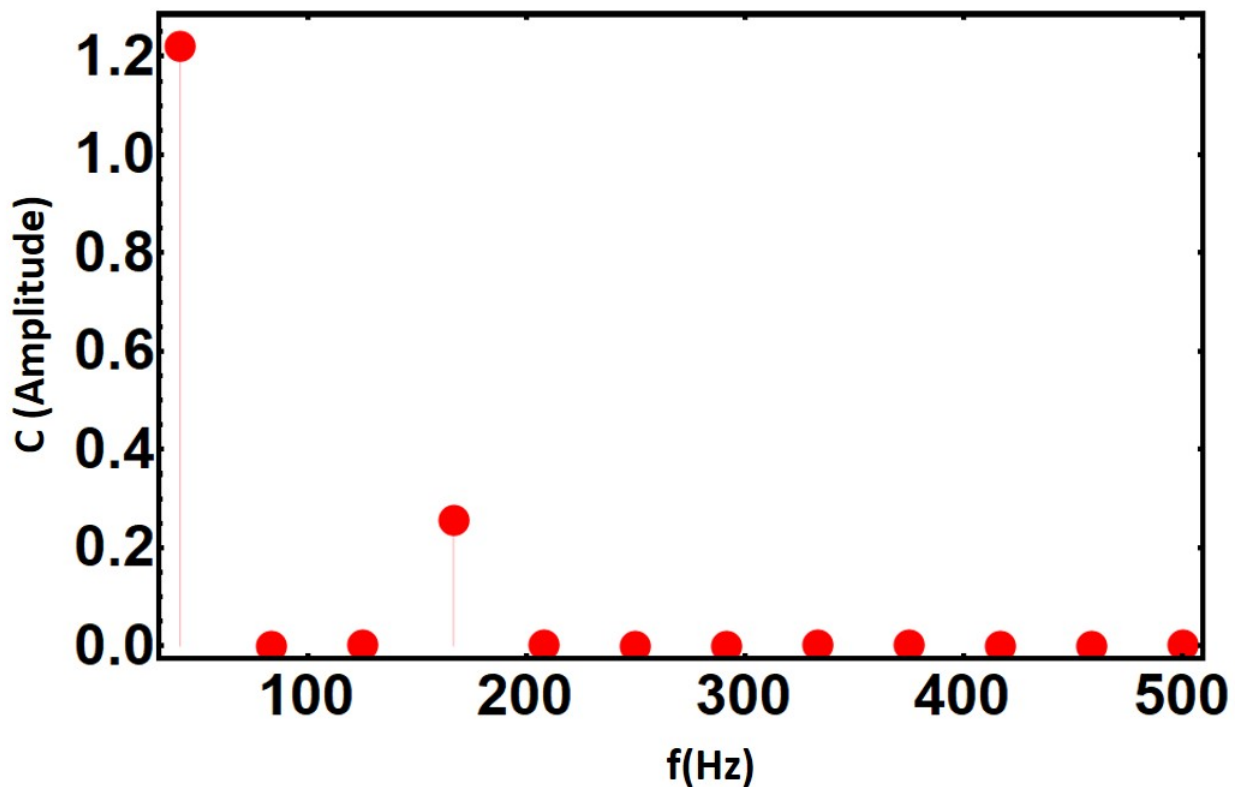


Figure Chapter5.13: DFT Fourier Spectrum.

Note again I have converted the x axis into cyclic frequency. How do I get these values? What is my Δf spacing?

Well that is simply the lowest frequency or the fundamental frequency which we have already calculated above. You should note here that the frequency resolution here is again quite poor — you should be worried about spectral leakage.

DFT Fourier Series Reconstruction

We've done the hard part, we solved for the Harmonic coefficients so now to plot the Fourier Series for this DFT what equation should we use? This one below of course,

\$\$

$$y(t) = \frac{A_0}{2} + \sum_{n=1}^{\frac{N}{2}} \left[A_n \cos(2\pi n \Delta f t) + B_n \sin(2\pi n \Delta f t) \right]$$

So we have all the values we need to solve, plug into this equation and plot as seen below:

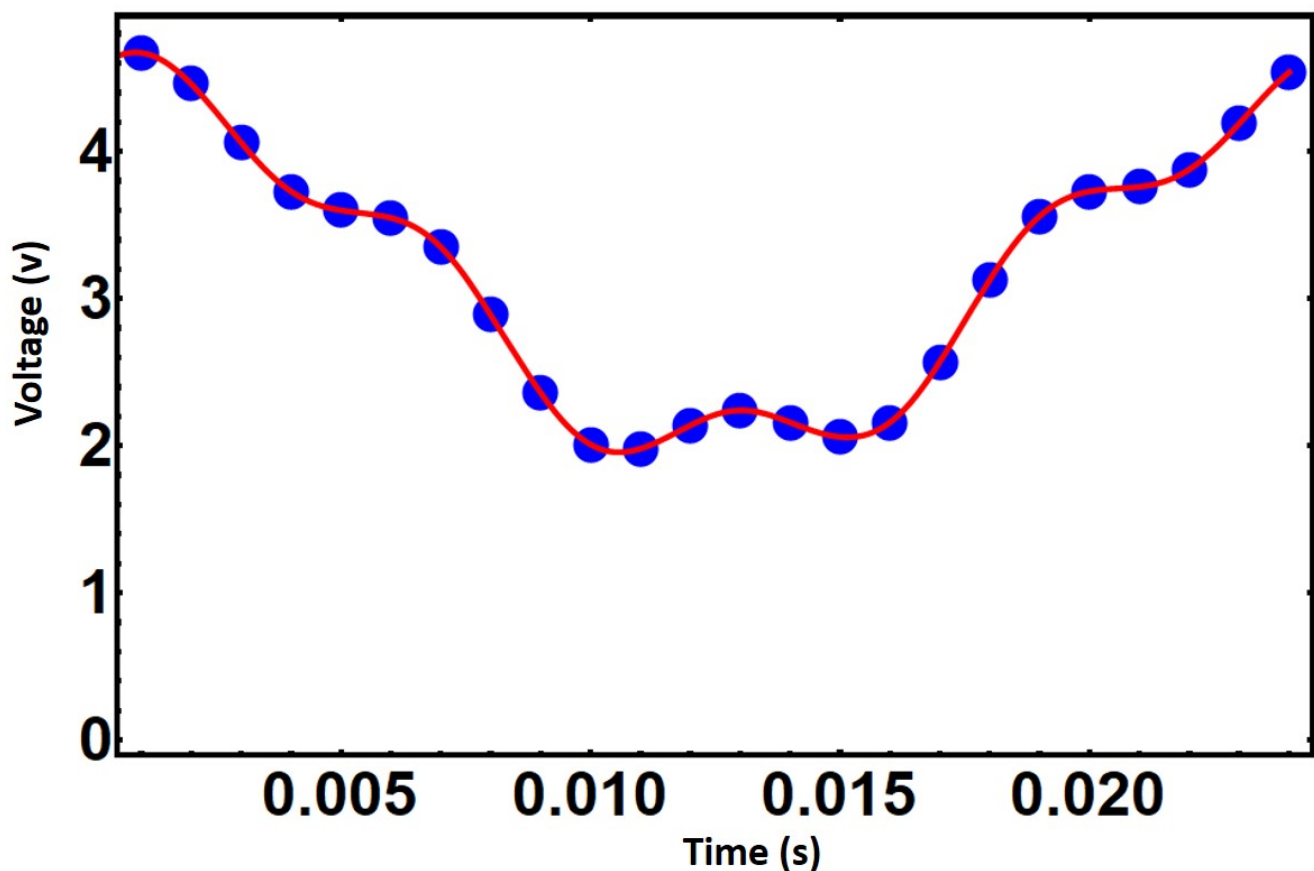


Figure Chapter5.14: DFT Fourier Series.

The blue points are our experimental data and the red line is our DFT calculated Fourier Series above. The fit looks nice so it looks like we sampled the curve sufficiently to reproduce the data.

Checklist: How to Accurately Sample a Discrete Signal

To accurately sample a discrete signal, do the following and you will be fine:

- Estimate the highest frequency in the signal and make sure that f_s is more than twice as large as that frequency.
- If you are forced to pick a Nyquist frequency less than the highest frequency in the signal, use a low-pass filter to block frequencies greater than the Nyquist frequency so those higher frequencies will not be aliased into your results.
- Estimate the lowest frequency in the signal or estimate the frequency resolution needed to accurately resolve the frequency components in the signal. Then choose the number of points in the sample N to yield the desired $\Delta f = \frac{f_s}{N}$ at the previously determined value of the sampling frequency f_s .

This page titled [Chapter 5: Discrete Fourier Transforms, Nyquist Frequency, Signal Aliasing, and Spectral Leakage](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 6: Strain Gauges and Wheastone Bridges

Everyone is now a master of linear elasticity theory but now is time to put that theory into practice and one of the experimental methods that is still being utilized to this day to measure strain is via electric resistance stain gauges.

Electrical resistive strain gauges have a long history being first discovered that the resistance of copper and iron wires will change when subjected to a mechanical load or equivalently a strain. This change in resistance was measured using a Wheatstone bridge circuit. We will analyze and utilize this Wheatstone bridge circuit configuration in order to measure the strain experimentally in our material.

While there are many other methods of measuring strain i.e. extensometer, photoelasticity, and video extensometry via tracking strain markers, strain gauges are still utilized extensively in industry and academia.

Are The Wires Elongating or Contracting? Resistance is Futile.

The resistance of a material will change when subjected to an external strain. Consider some conductive material like a wire of copper that has some length l and diameter D . If the wire is subjected to uniaxial tension along the length of the wire, as seen below, then the length will increase as diameter of the wire decreases.

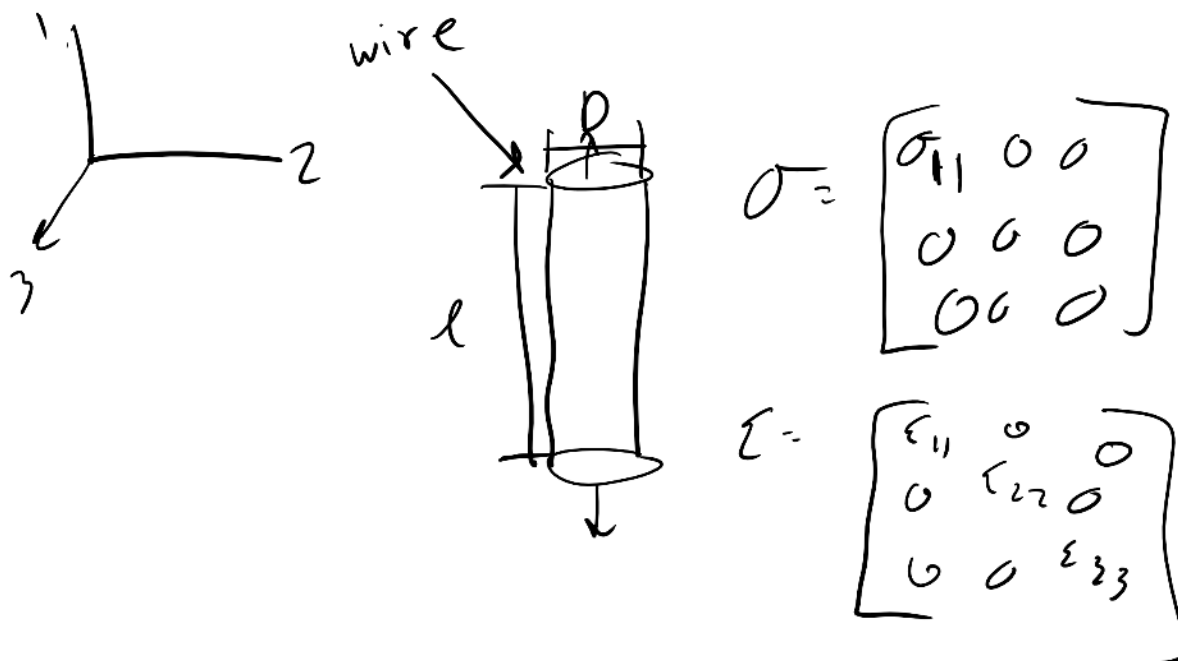


Figure Chapter6.1: Elongating Wire In Electrical Strain Gauge.

This will result in a change in the resistance of the wire and specifically the resistance will increase. We can calculate this change in resistance below where R is resistance:

$$R = \frac{\rho l}{A} = \frac{\rho l}{C D^2}$$

where ρ is resistivity and C is a constant which will describe the cross sectional area; for a wire $C = \frac{\pi}{4}$ for a circular cross section.

The change in resistance is then given by our old friend the chain rule:

$$dR = \frac{dR}{d\rho} d\rho + \frac{dR}{dl} dl + \frac{dR}{dD} dD$$

$$dR = \frac{C D^2 (1 d \rho + \rho dl) - 2 C \rho l D dD}{(C D^2)^2} \backslash$$

$$dR = \frac{1}{C D^2} \left(1 d \rho + \rho dl - 2 \rho l \frac{dD}{D} \right) \backslash$$

If we then divide by R :

$$\frac{dR}{R} = \frac{dl}{l} - 2 \frac{dD}{D} + \frac{d \rho}{\rho} \backslash$$

$$\frac{\frac{dR}{R}}{\frac{dl}{l}} = 1 - 2 \frac{\frac{dD}{D}}{\frac{dl}{l}} + \frac{\frac{d \rho}{\rho}}{\frac{dl}{l}} \backslash$$

From the equation above we can see:

$$\epsilon_{11} = \frac{dl}{l} \backslash$$

$$\epsilon_{22} = \frac{dD}{D} \backslash$$

$$\nu = -\frac{\frac{dD}{D}}{\frac{dl}{l}} \backslash$$

We can then define the gauge factor F , which is typically determined experimentally and will typically be a constant for the problems that we will work with in this class as:

$$F = \frac{\frac{dR}{R}}{\frac{dL}{L}} = \frac{\frac{dR}{R}}{\epsilon_{11}} = 1 + 2 \nu + \frac{\frac{d \rho}{\rho}}{\frac{dL}{L}} \backslash$$

This equation is a bit messy as you can see but we can re-arrange this expression and get our fundamental expression for how resistance will change with strain as seen below:

$$\epsilon_{11} = \frac{1}{F} \frac{\Delta R}{R} \backslash$$

Here we can clearly see that the change in resistance and strain are proportional — i.e. as we strain a material in tension the resistance increases. And our strain gauge is composed of these wires as seen below and they will extend or compress depending on the applied stress:

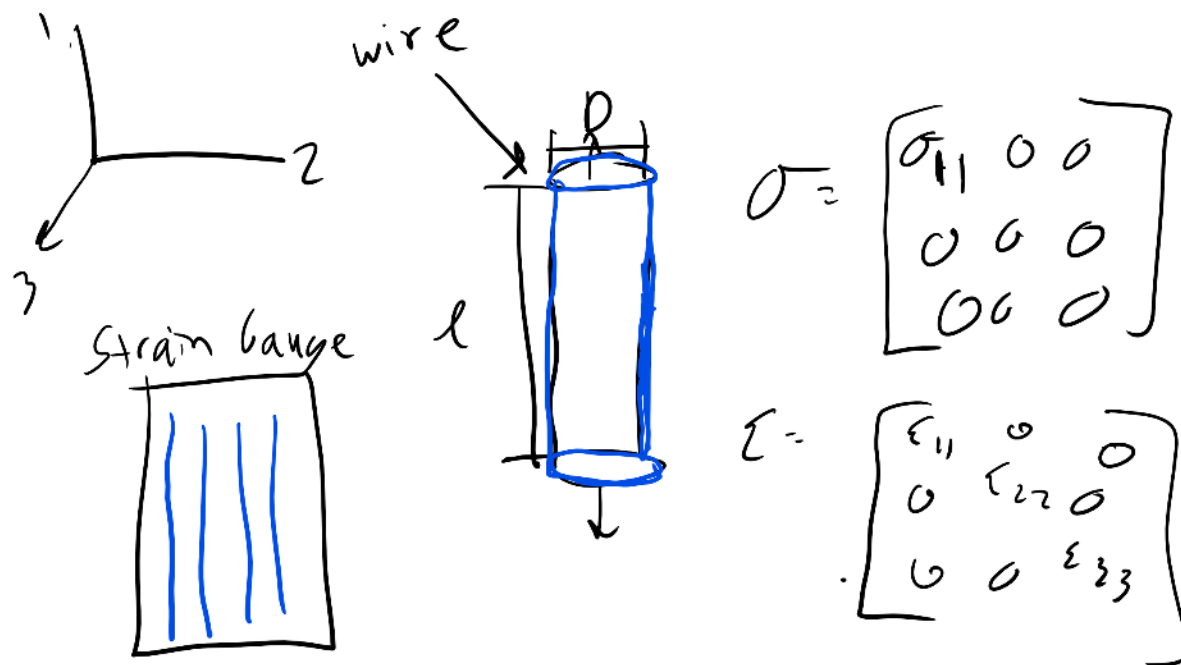


Figure Chapter6.2: Strain gauge composed of wires which change resistance of strain gauge.

Now at this point you may ask: why do we typically measure a voltage output from our strain gauge? Why don't we just measure the change in resistance directly? Why do I have to learn circuits, why are you so mean...

Well the answer lies once again in the precision and sensitivity of a measurement. If we are measuring microstrain we cannot use an ohmmeter to measure these values. Instead we will utilize circuits and specifically a Wheatstone Bridge Circuit and measure the corresponding change in voltage and correlate that to the strain in a material. However, before we jump directly into Wheatstone Bridge circuitry we should first begin with the building blocks of a Wheatstone Bridge which is a voltage divider circuit.

Voltage Divider Circuit

Below is a schematic of a voltage divider circuit. If we first imagine that initially there is a negligible amount of current taken from V_o , our output voltage, then $V_o \approx 0$ and then we find that:

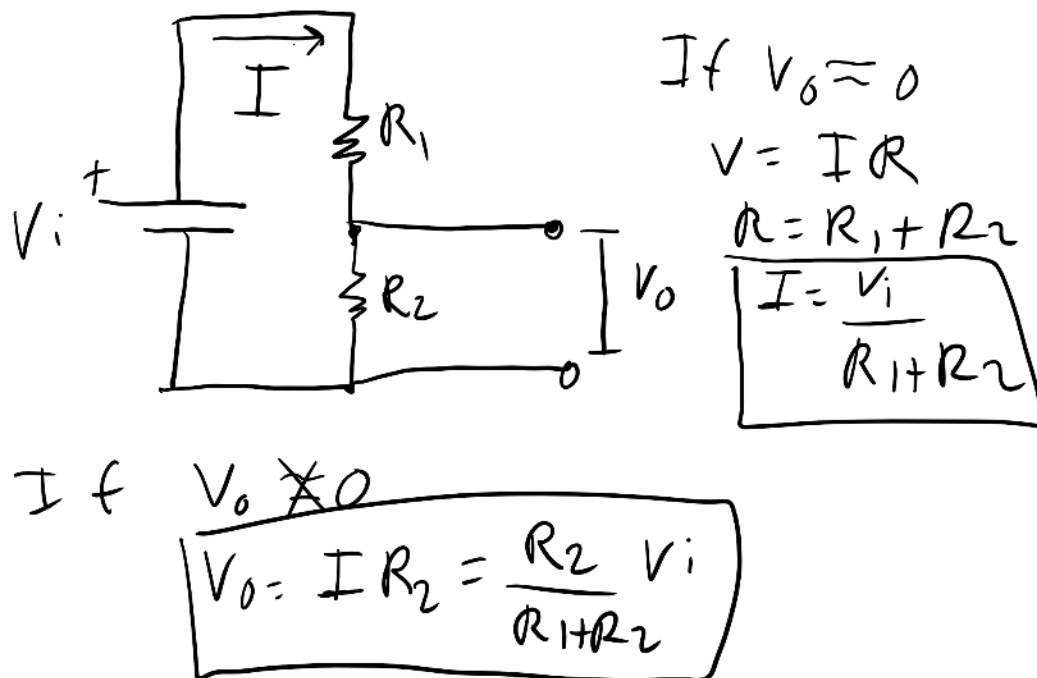


Figure Chapter6.3: Voltage divider circuit.

\$\$

$$I = \frac{V_i}{R_1 + R_2}$$

\]

where V_i is the input voltage. This gives us the current in our circuit, but if we find that there is current drawn from V_o , then we find that:

\$\$

$$V_o = i R_2 = \frac{R_2}{R_1 + R_2} V_i$$

\]

This is the output for a single voltage dividing circuit.

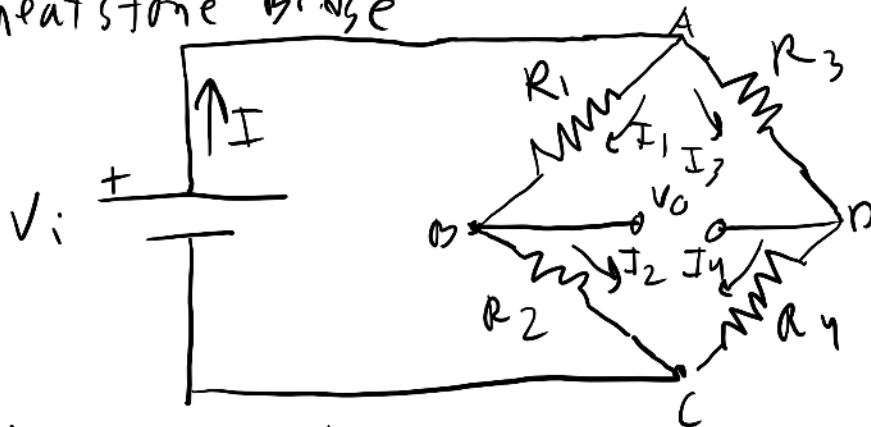
Why did we learn about a voltage dividing circuit? You said Wheatstone bridge circuits were important for strain gauges!

Yes, I agree — but as you will see our Wheatstone Bridge Circuit will look very similar to two voltage dividing circuits.

Wheatstone Bridge Circuit

We will utilize a Wheatstone Bridge circuit to measure strain in our strain gauges, and we can see an example of Wheatstone bridges and how it looks very similar to two voltage dividing circuits.

Wheatstone Bridge



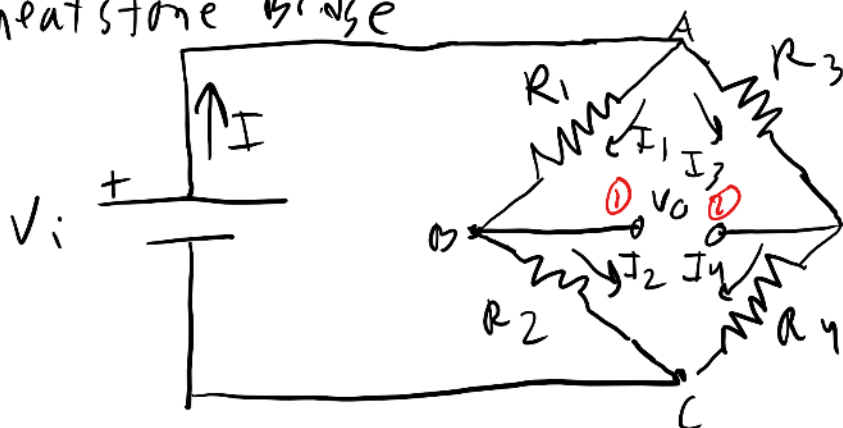
2 Voltage Dividers



Figure Chapter6.4: Wheatstone bridge circuit.

In fact, we can write out the voltage at positions 1 and 2, V_1 and V_2 , and we see that the expressions are equivalent:

Wheatstone Bridge



$$V_1 = \frac{R_2}{R_1 + R_2} V_i$$

$$V_2 = \frac{R_4}{R_3 + R_4} V_i$$

2 Voltage Dividers



$$V_1 = \frac{R_2}{R_1 + R_2} V_i$$

$$V_2 = \frac{R_4}{R_3 + R_4} V_i$$

Figure Chapter6.5: Voltage at Position 1 and 2 in Wheatstone Bridge and Voltage Dividing Circuits.

\$\$

$$V_1 = V_i \frac{R_2}{R_1 + R_2}$$

$$V_2 = V_i \frac{R_4}{R_3 + R_4}$$

]

Now at balance we should find that $V_o = 0$, and thus $V_1 = V_2$. If we set these expressions equal to one another, we find:

Wheatstone Bridge



@ Balance i.e. $V_o = 0$ $V_1 = V_2$

$$V_i \frac{R_2}{R_1 + R_2} = \frac{R_4}{R_3 + R_4} V_i$$

$$R_2 R_3 = R_1 R_4$$

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

Figure Chapter6.6: Resistance Relationship at Balance.

\$\$

$$R_2 R_3 = R_1 R_4$$

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

]

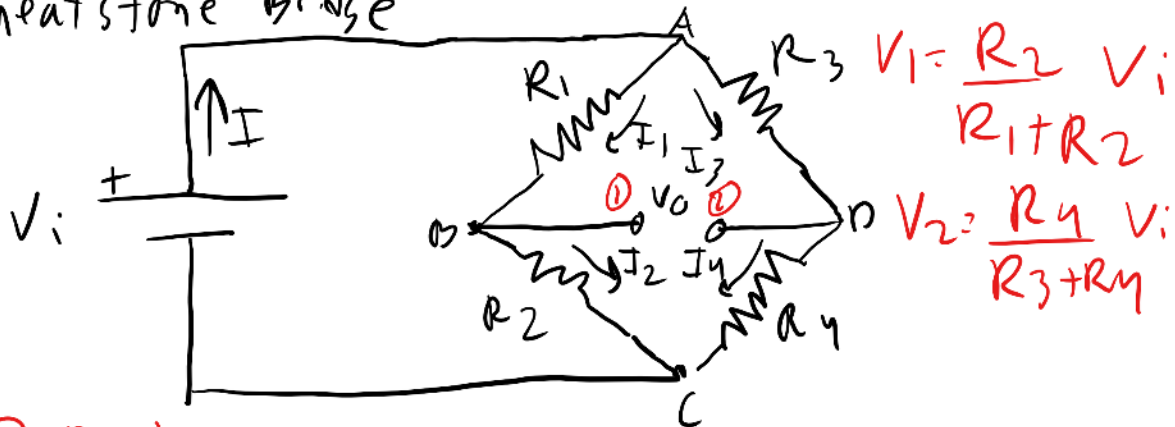
If we are not at balance, we can also calculate the output voltage of the Wheatstone Bridge configuration, similar to what we did for the voltage dividing circuit:

\$\$

$$V_o = V_1 - V_2 = V_i \left[\frac{R_2}{R_1 + R_2} - \frac{R_4}{R_3 + R_4} \right] = V_i \left[\frac{R_2 R_3 - R_1 R_4}{(R_1 + R_2)(R_3 + R_4)} \right]$$

]

Wheatstone Bridge



@ Balance i.e. $V_o = 0$ $V_1 = V_2$

$$V_o = V_1 - V_2 = V_i \left[\frac{R_2 R_3 - R_1 R_4}{(R_1 + R_2)(R_3 + R_4)} \right]$$

Figure Chapter6.7: Output voltage of Wheatstone Bridge.

You can see in this expression that if $R_2 R_3 = R_1 R_4$, then we are balanced and $V_o = 0$.

Now so far we have just looked at a Wheatstone Bridge circuit composed of only resistors, but we want to replace some of these resistors with strain gauges and then correlate the change in output voltage to the applied strain.

Let's start with the most general case where we replace all resistors with strain gauges, a configuration referred to as Full Bridge configuration.

Full Wheatstone Bridge in Standard Configuration

Let's look at the full Wheatstone bridge configuration as seen below:

Wheatstone Bridge Standard Configuration



$$V_i \parallel \rightarrow \frac{R_1}{R_2} = \frac{R_3}{R_4}$$

$$V_o = V_i \left[\frac{R_3}{R_1 + R_3} - \frac{R_4}{R_2 + R_4} \right]$$

$$dV_o = \frac{\partial V_o}{\partial R_1} dR_1 + \frac{\partial V_o}{\partial R_2} dR_2 + \frac{\partial V_o}{\partial R_3} dR_3 + \frac{\partial V_o}{\partial R_4} dR_4$$

Figure Chapter 6.8: Standard Wheatstone Bridge Configuration.

If we replace the 4 resistors with 4 strain gauges (each must have the same resistance so that at balance the output voltage is 0), we can think of the most general case where we strain a material that has 4 strain gauges attached and consider the change in resistance and corresponding change in output voltage.

So let's begin with our expression for the output voltage for the standard Wheatstone bridge configuration:

\$\$

$$V_o = V_i \left[\frac{R_3}{R_1 + R_3} - \frac{R_4}{R_2 + R_4} \right]$$

\]

Now the change in voltage for a change in resistance for each arm will be given by the following expression:

\$\$

$$dV_o = \frac{\partial V_o}{\partial R_1} dR_1 + \frac{\partial V_o}{\partial R_2} dR_2 + \frac{\partial V_o}{\partial R_3} dR_3 + \frac{\partial V_o}{\partial R_4} dR_4$$

\]

\$\$

$$\frac{dV_o}{V_i} = \frac{R_2 dR_1}{(R_1 + R_2)^2} - \frac{R_1 dR_2}{(R_1 + R_2)^2} - \frac{R_4 dR_3}{(R_3 + R_4)^2} + \frac{R_3 dR_4}{(R_3 + R_4)^2}$$

\]

For this course and the Wheatstone bridges we will work with, $R_1 = R_2 = R_3 = R_4 = R$, so we can write:

\$\$

$$\frac{dV_o}{V_i} = \frac{dR_1 - dR_2 - dR_3 + dR_4}{4R}$$

\]

And remembering back to our fundamental equation:

\$\$

$$\epsilon_n = \frac{1}{F} \frac{\Delta R}{R}$$

\]

Substituting in, we can get our final, critical, and fundamental expression for relating output voltage and strain using Wheatstone bridges as seen below:

Wheatstone Bridge Standard Configuration



$$V_i \parallel \rightarrow \frac{R_1}{R_2} = \frac{R_3}{R_4}$$

$$\frac{dV_o}{V_i} = \frac{dR_1 - dR_2 - dR_3 + dR_4}{L/R}$$

$$\epsilon = \frac{1}{F} \frac{\Delta R}{R}$$

$$\frac{dV_o}{V_i} = \frac{\Delta V_o}{V_i} = \frac{F}{L} [\epsilon_1 - \epsilon_2 - \epsilon_3 + \epsilon_4]$$

Figure Chapter6.9: Standard Wheatstone Bridge Configuration Fundamental Equation.

 alt="Standard Wheatstone Bridge Configuration Fundamental Equation." data-bbox="82 490 734 506"/>

\$\$

$$\frac{dV_o}{V_i} = \frac{F}{L} [\epsilon_1 - \epsilon_2 - \epsilon_3 + \epsilon_4]$$

We can now use this expression to describe the relationship between the output voltage, strain, and stress for a myriad of different strain gauge configurations. Let's consider several examples to illustrate how we will utilize this expression.

Strain Gauge Configuration

Uniaxial Tension Quarter Bridge Configuration

If we consider a Quarter Bridge Configuration (only 1 strain gauge in our bridge) as seen below:

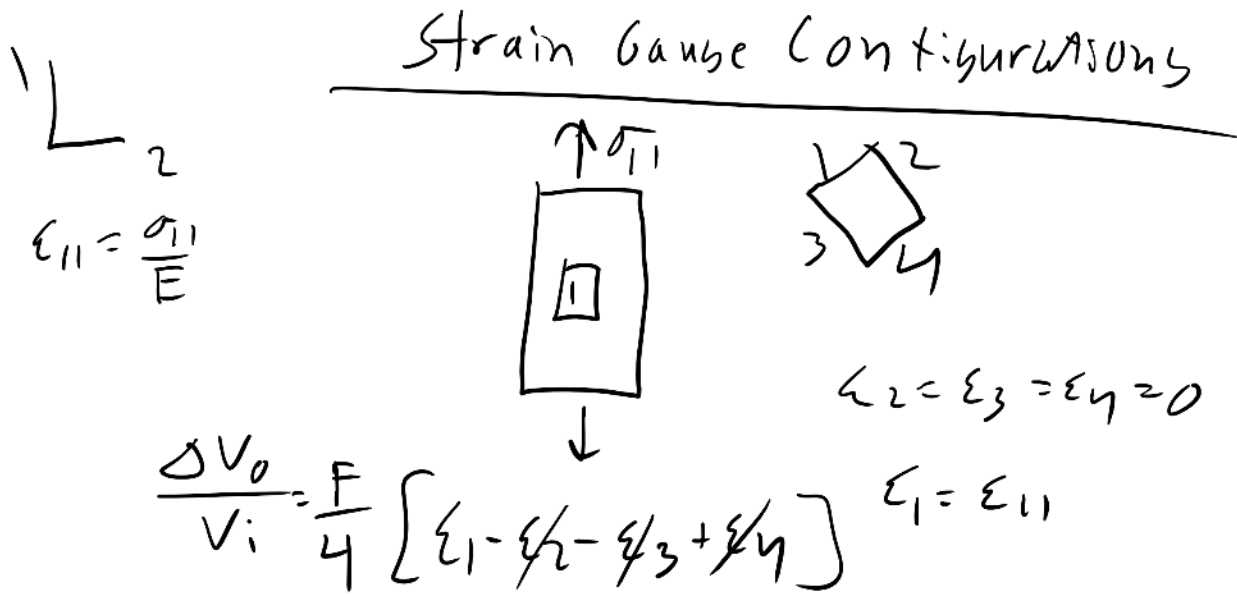


Figure Chapter6.10: Uniaxial Tension and Quarter Bridge Configuration.

So we have a strain gauge in the 1 arm of our Wheatstone bridge and it is aligned with σ_{11} . Now remember if the wires extend/elongate that correlates to a positive strain and because we only have one strain gauge the resistors in the 2, 3, and 4 arms are not being stretched so the strain there is 0. Furthermore, we know that the strain measured in the 1 arm will just be $\epsilon_1 = \epsilon_{11}$. Aren't you glad we covered mechanics first?

Let's do another example where now we have an additional stress, σ_{22} .

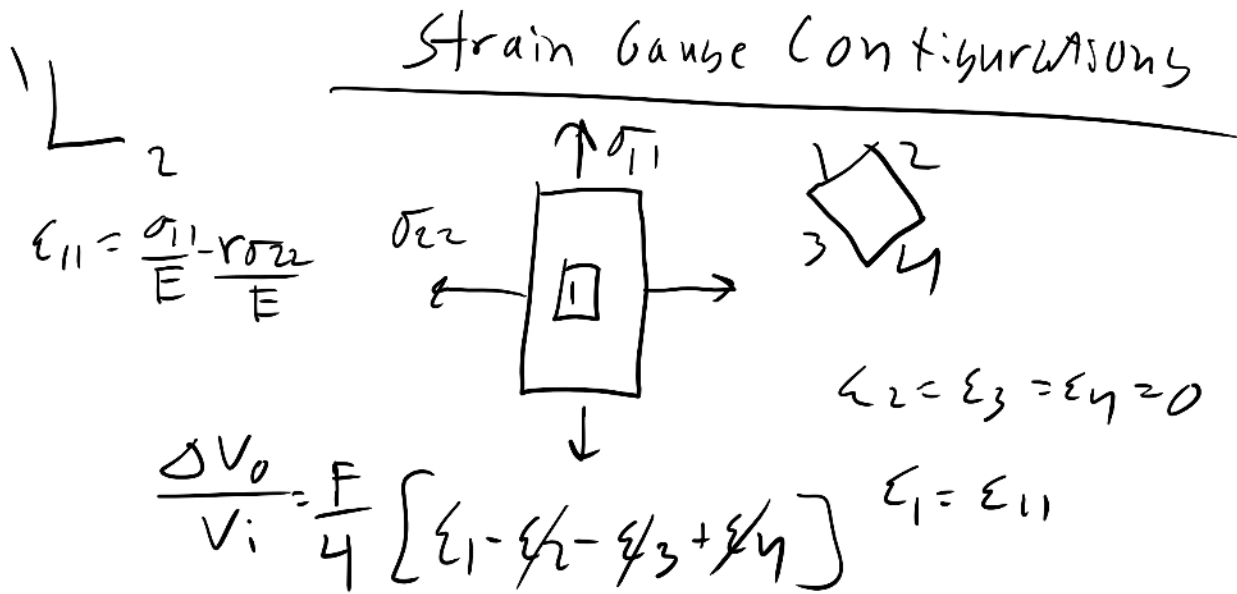


Figure Chapter6.11: Normal Plane Stress Quarter Bridge Configuration.

Well everything else holds in this example but our expression for ϵ_{11} changes and now it is:

\$\$

$$\epsilon_{11} = \frac{\sigma_{11}}{E} - \frac{\nu \sigma_{22}}{E}$$

\]

Notice the additional negative component in this strain because as the material is stretched in the 22 direction, the wires will compress and the length will decrease — thus the negative sign.

Let's look at the same stress state but add another strain gauge to make a Half-Bridge configuration as seen below:

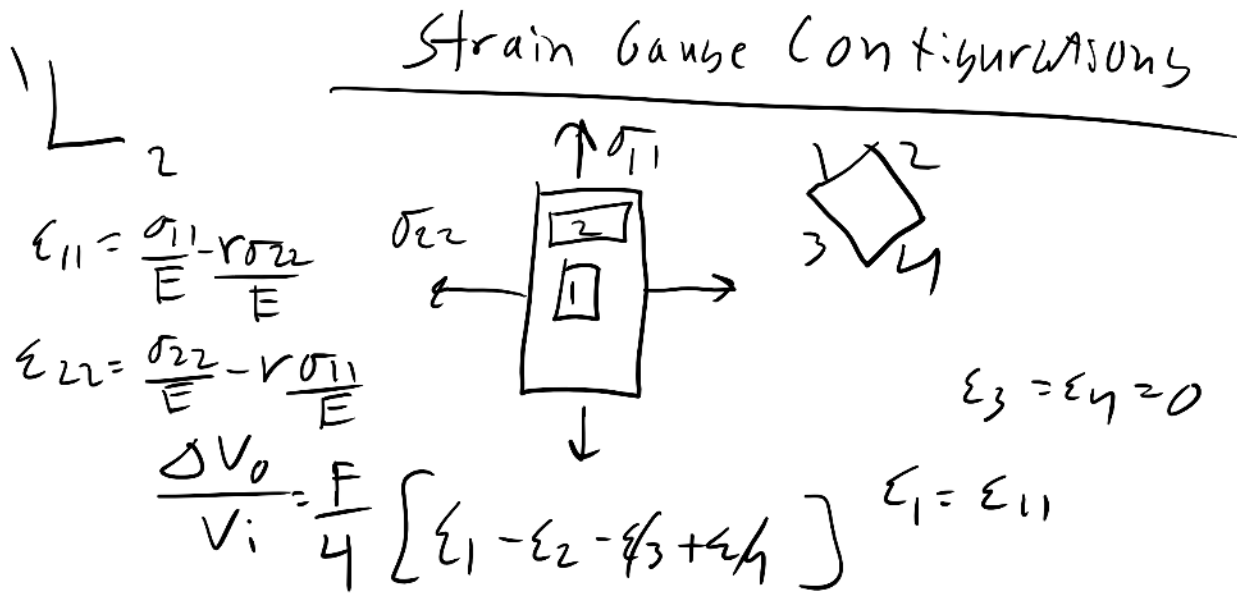


Figure Chapter6.12: Normal Plane Stress Half-Bridge Configuration.

Here now we have a strain gauge aligned with the 22 direction so it will sense ϵ_{22} , so:

\$\$

$\epsilon_2 = \epsilon_{22}$

\\

Alright, you are now masters of tension, compression, and normal stress — but what about bending?

Let's take a look at the scenario below of a cantilever beam bending:

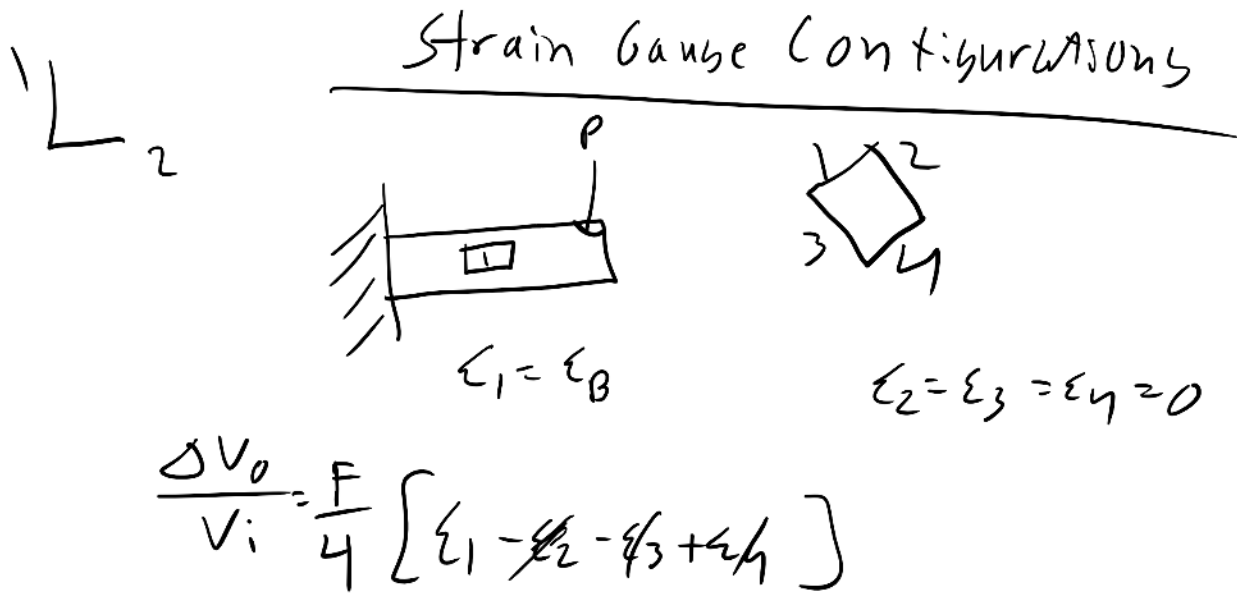


Figure Chapter6.13: Quarter Bridge Configuration Under Pure Bending

Here we again only have one strain gauge, but now it will be exposed to pure bending strain, and this strain will be positive if the strain gauge is on top of the beam (denoted by a solid outline of the strain gauge) because the wires will elongate.

However, in the figure below you can see that if we have a strain gauge placed on the bottom of the beam then the wires will compress and that will correlate to a negative bending strain. So be ready for some examples with normal and bending strain combined... I know I know so mean.

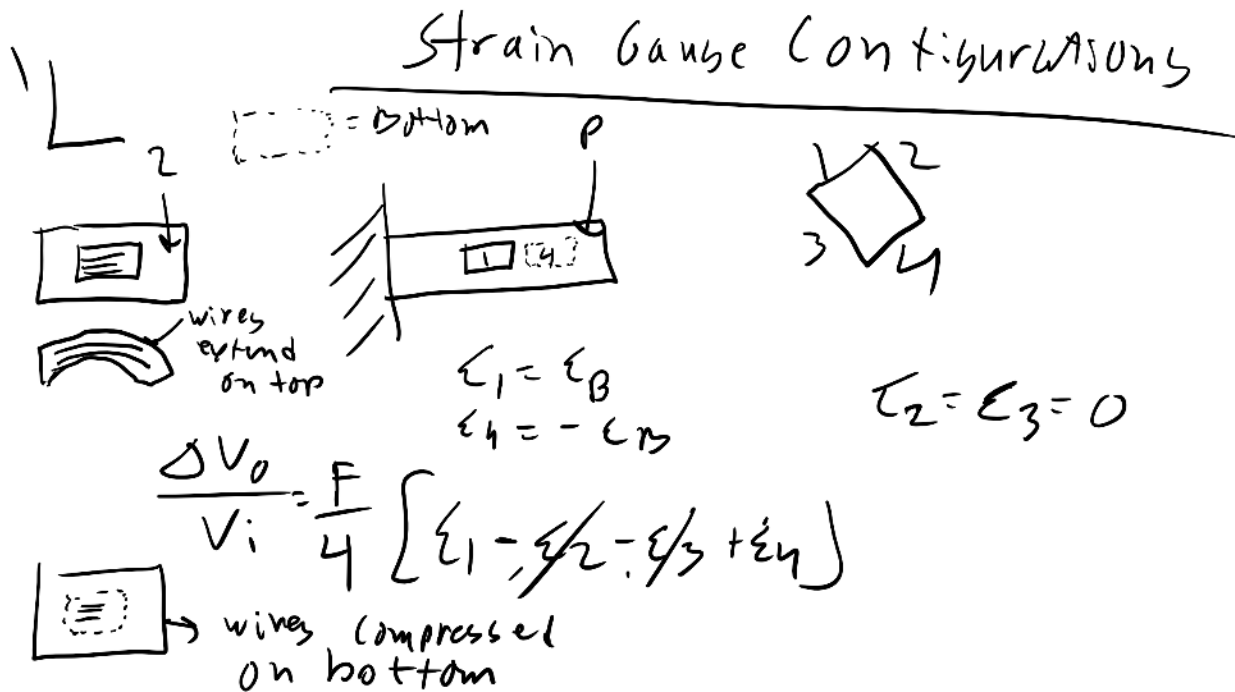


Figure Chapter6.14: Half Configuration Under Pure Bending.

Linear Thermal Expansion Contribution to Strain

Thus far we have neglected one contribution to strain and that is the contribution of thermal strain. Most materials (see exceptions like PbTiO_3 and other materials with a negative thermal expansion coefficient) will expand as temperature increases, and this behavior is reflected in a material's linear thermal expansion coefficient described below:

\$\$

$$\alpha_L = \frac{1}{L} \frac{dL}{dT}$$

\]

We can re-write this expression to obtain the thermal strain $\epsilon_{\text{thermal}}$:

\$\$

$$\epsilon_{\text{thermal}} = \frac{\Delta L}{L} = \alpha_L \Delta T$$

\]

This thermal strain will cause the wires in the strain gauge to expand, and thus this contribution will typically always be a positive contribution to strain (unless α_L is negative). You can also configure your strain gauges in such a manner to eliminate the thermal strain contribution if you do not want to measure that contribution.

Calibration of Strain Gauges

Way way back in lecture one we mentioned that when designing experiments or making an experimental apparatus, calibration was a critical step. The same calibration must be performed for strain gauges as well. However, this is a difficult proposition.

As you might imagine, calibration will involve placing a strain gauge on a material and applying a fixed strain and measuring the resulting output voltage. This is a valid way to calibrate — but as uncertainty experts think about all the potential uncertainties that can occur in this process:

- Uncertainty in the applied strain
- Defects in the material

- Variation in geometry and dimensions
- Error introduced by the mounting process itself

Instead, we will utilize an ingenious calibration technique which involves the precise application of a known change in resistance and measuring the system's response.

To do this, we add a switch S and a shunt resistor R_S to the 1-arm of our Wheatstone Bridge:

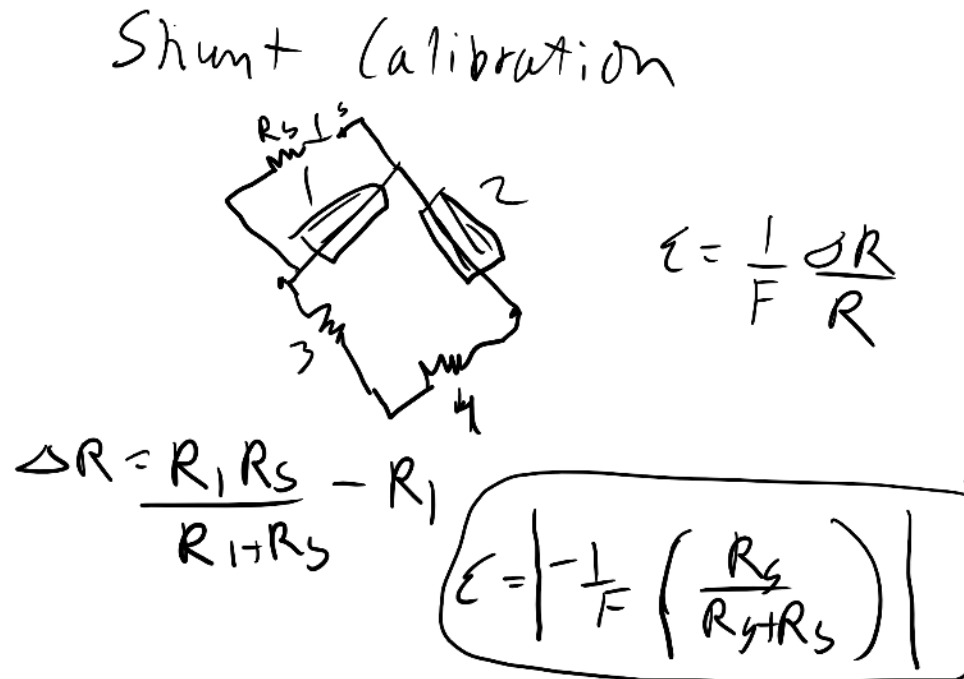


Figure Chapter6.15: Shunt Calibration of Strain Gauges.

When the switch is open, the resistance in arm 1 is just R_1 . When the switch is closed, the resistance changes to:

$$\frac{R_1 R_S}{R_1 + R_S}$$

So the change in resistance after closing the switch is:

$$\Delta R = \frac{R_1 R_S}{R_1 + R_S} - R_1 = -\frac{R_1^2}{R_1 + R_S}$$

And remembering:

$$\epsilon = \frac{1}{F} \frac{\Delta R}{R}$$

We find the corresponding strain from the change in resistance is:

$$\epsilon = \left| -\frac{1}{F} \left(\frac{R_1}{R_1 + R_S} \right) \right|$$

Full Wheatstone Cantilever Beam Bending

Consider an Al-6061 cantilever beam with full Wheatstone Bridge Configuration as seen below. The gauge factor is 2 and the excitation voltage is 5 V. The beam thickness is 0.1 m and the length is 10 m. The initial resistances in each arm of the Wheatstone Bridge are 120 Ω . What is the change in voltage after this load is applied?

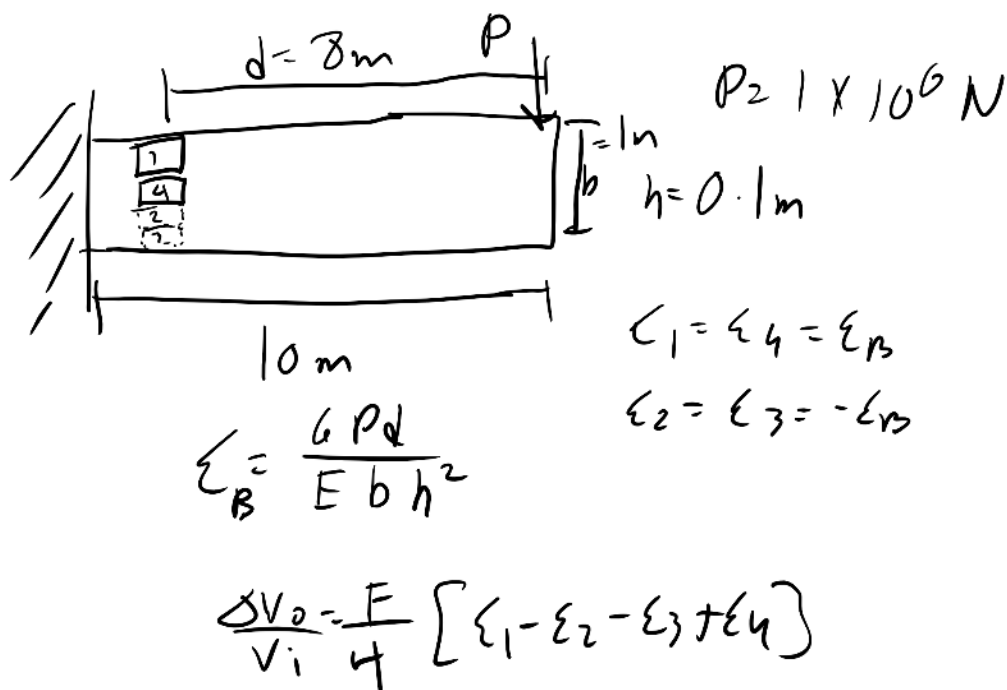


Figure Chapter6.16: Full Bridge Cantilever Beam.

Write Out the Strain

Let's look at the following material under a combination of uniaxial tension and bending:

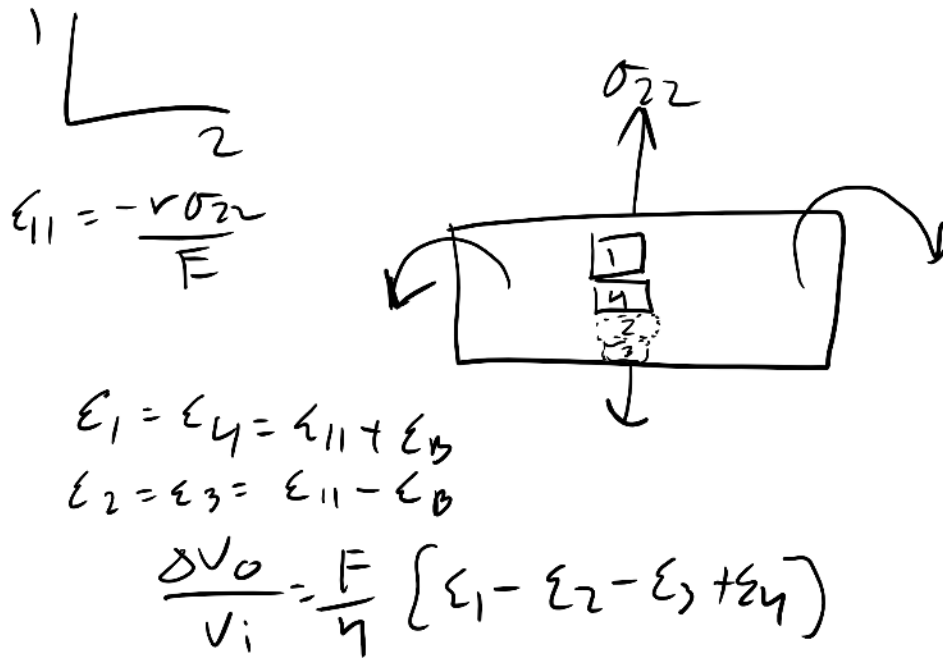


Figure Chapter6.17: Beam Under Bending and Axial Strain.

Write out the strain in each strain gauge.

Determine the configuration that eliminates bending.

Determine the configuration which eliminates axial strain components.

I Can't Stop Asking Thin Walled Pressure Vessel Problems

Take a look at the arrangement of strain gauges on the cylindrical vessel seen below:

$$\epsilon_1 = \epsilon_4 = \epsilon_{11} \quad \epsilon_2 = \epsilon_3 = \epsilon_{22}$$

$$\epsilon_{11} = \frac{1}{E} (\sigma_h - \nu \sigma_l)$$

$$\epsilon_{22} = \frac{1}{E} (\sigma_l - \nu \sigma_h)$$

$$\frac{\Delta V_0}{V_i} = \frac{F}{\eta} [\epsilon_1 - \epsilon_2 - \epsilon_3 + \epsilon_4]$$

Figure Chapter6.18: Pressure Vessel.

Consider an aluminum pressure vessel with full bridge configuration as seen below. The thickness is 0.1 m. The initial resistances of the strain gauges are 200 Ω . The gauge factor is 4. Initially a shunt is placed across gauge 1 and the output reads 20 units (your oscilloscope is so old you cannot read the units). After the cylinder is pressurized you read an output of 50.

Calculate the hoop and longitudinal stress.

Strain Rosettes: Rectangular and Delta

For the examples that we have done so far, our strain gauges have been aligned with our coordinate system or along principal directions. But if we want to measure a complex stress state — where a plane stress condition is completely unknown, like near a weld, defect, microcrack, etc. — we will need to utilize 3 strain gauges and then use Mohr's circle for our strain tensor to determine the normal and shear strains (as our strain gauges cannot directly measure shear strains — they only give us normal strains).

This configuration of 3 strain gauges is called a strain rosette. There are multiple types of rosettes (T-Delta), but in this course we will focus on rectangular and delta or equiangular rosettes as seen below:

Strain Rosettes

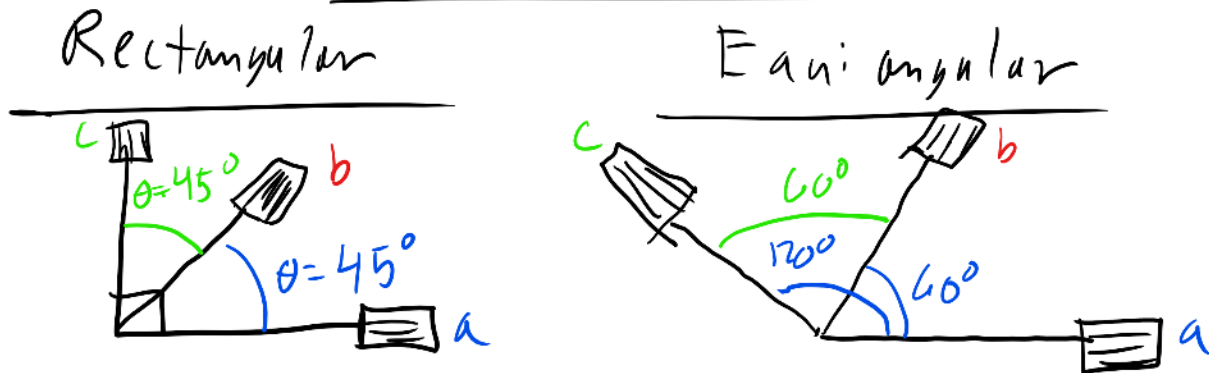


Figure Chapter6.19: Rectangular and Delta Rosette.

You can see that for the rectangular rosette the a and b arms are offset by 45° as are b and c, and the a and c arms are offset by 90° . For the delta rosette the a and b arms are offset by 60° , b and c are offset by 60° , and a and c arms are offset by 120° .

Now let's imagine that we place a rectangular rosette on a material near a weld, defect, stress concentrator, etc., i.e. near any complex stress state we may not know. If we place the rectangular rosette there and apply the load for that particular application, we will obtain strain readings from each arm of the rosette.

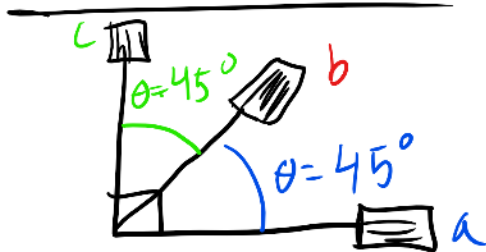
If we find that $\epsilon_a = 72$, $\epsilon_b = 120$, $\epsilon_c = 248$ in microstrain, we can then calculate the principal strains, stresses, rotation to get to the principal strain state, the center of our Mohr's circle, and the radius of our Mohr's circle.

Sounds like a lot of work right? But you are masters of Mohr's circle and rotation so this will be very similar — but with one key distinction: we need to perform our Mohr's circle in strain as seen below.

$\epsilon_a, \epsilon_b, \epsilon_c$ must be normal

Strain Rosettes

Rectangular



$$\begin{aligned}\epsilon_a &= 72 \times 10^{-6} & C? \\ \epsilon_b &= 120 \times 10^{-6} & R? \\ \epsilon_c &= 248 \times 10^{-6} & 2\theta?\end{aligned}$$

$$\begin{aligned}\epsilon_a &= c + R \cos(2\theta) \\ \epsilon_b &= c + R \cos(2\theta + 90^\circ) \\ \epsilon_c &= c + R \cos(2\theta + 180^\circ)\end{aligned}$$

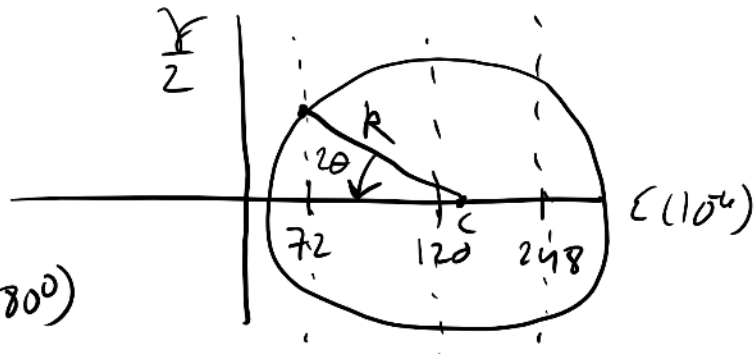


Figure Chapter6.20: Mohr's circle in strain.

Now we just mentioned that our strain gauges cannot measure shear strains directly, so thus any measurement from each arm can be considered a normal strain.

Now for our stress state in Mohr's circle, we had coordinates and drew our circle that way because we knew the normal and shear stresses. But here we only know the x-values — i.e., our normal strains.

But we know that our Mohr's circle must intersect these vertical normal strain lines at six points as seen in the figure.

We also know that we have three unknowns in our Mohr's circle:

- Center c
- Radius R
- 2θ

But we know ϵ_A , ϵ_B , and ϵ_C , so we can use our circle geometry to solve for these unknowns.

From the figure we see the relationship that relates ϵ_A to c , R , and 2θ :

$$\begin{aligned}\epsilon_A &= 72 \times 10^{-6} = c + R \cos(2\theta)\end{aligned}$$

We can also obtain a relationship for ϵ_B and ϵ_C , knowing that they are offset by a fixed angle from arm A. We know that for B in real space the angle is 45° , but in Mohr's circle that translates to 90° , so we add this to our expression:

$$\epsilon_B = 120 \times 10^{-6} = c + R \cos(2\theta + 90^\circ)$$

Similarly for arm C:

$$\epsilon_C = 248 \times 10^{-6} = c + R \cos(2\theta + 180^\circ)$$

With these three equations we can solve for our unknowns and find that:

\$\$

$$c = 1.6 \times 10^{-4} \text{ m}$$

$$R = 9.6 \times 10^{-5} \text{ m}$$

$$\epsilon_1 = 2.6 \times 10^{-4}$$

$$\epsilon_2 = 6.33 \times 10^{-5}$$

$$\theta = 12.22^\circ$$

$$\sigma_1 = 6.27 \times 10^7 \text{ Pa}$$

$$\sigma_2 = 3.19 \times 10^7 \text{ Pa}$$

\\]

We can obtain our principal stresses from our stress tensor and strain tensor relationships:

\$\$

$$\epsilon_1 = \frac{1}{E}(\sigma_1 - \nu \sigma_2)$$

$$\epsilon_2 = \frac{1}{E}(\sigma_2 - \nu \sigma_1)$$

\\]

You can use this framework to calculate many different stress and strain states. For the delta rosette, we would perform a similar procedure but the angles would be different — i.e. 60° instead of 45° . Also, you can imagine I might/will give you a different rosette on a Pset or an Exam and you might have to do even more rotations...

This page titled [Chapter 6: Strain Gauges and Wheatstone Bridges](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 7: Introduction to Feedback Control, Laplace Transforms, and Transfer Functions

Introduction to Feedback Control

Control systems are a fundamental component of modern engineering design. Applications range across disciplines, including aircraft stabilization, robotic surgery, precision manufacturing, and temperature regulation in chemical reactors. In all cases, the objective remains the same: to automatically manipulate a system's behavior so that it conforms to a prescribed performance or reference trajectory.

A control system achieves this by continuously measuring the system's output, comparing it to a desired input, and adjusting internal inputs or forces in real time to minimize the error. This process relies on feedback.

Mathematical Modeling

Designing any control system begins with developing an appropriate mathematical model of the physical system. This model typically takes the form of an ordinary differential equation (ODE), which is derived based on the system's mechanical, electrical, thermal, or fluid characteristics. The model encapsulates effects such as:

- Inertial resistance to motion
- Elastic restoring forces
- Energy dissipation via damping
- External forcing or actuation

The resulting system dynamics are then expressed in either time-domain or Laplace-domain representations, allowing for analysis and controller synthesis using established control theory frameworks.

Prerequisite Knowledge: Dynamics

Because the systems under consideration are inherently dynamic, changing in response to inputs over time, a strong foundation in engineering dynamics is required. Students must be comfortable deriving and manipulating ODEs that describe system behavior. For example, the standard second-order linear differential equation that describes a forced mass-spring-damper system is a classic system that we will analyze in depth throughout this class:

$$m\ddot{x} + c\dot{x} + kx = f(t)$$

where m is the mass, c is the damping coefficient, k is the spring constant, and $f(t)$ is the external forcing function.

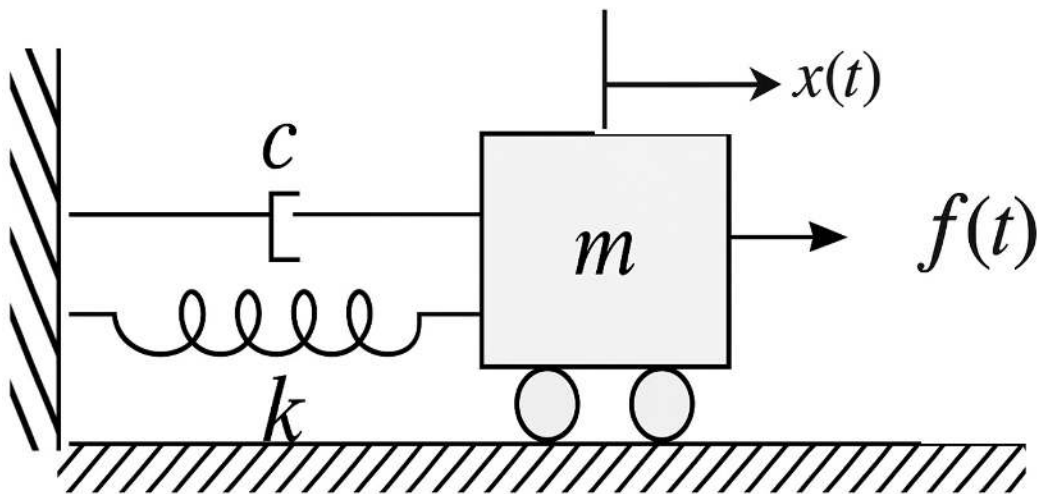


Figure Chapter7.1: Mass Spring Damper System

This system will appear frequently in mechanical system modeling and must be understood in terms of both its physical meaning and mathematical properties.

Two-Degree-of-Freedom Vertical Suspension Model

This simplified vertical suspension model captures the dynamic interaction between two vertically aligned masses: a lower mass (e.g., wheel and axle) and an upper mass (e.g., car chassis). The system is designed to model how a suspension system responds to vertical disturbances such as road irregularities.

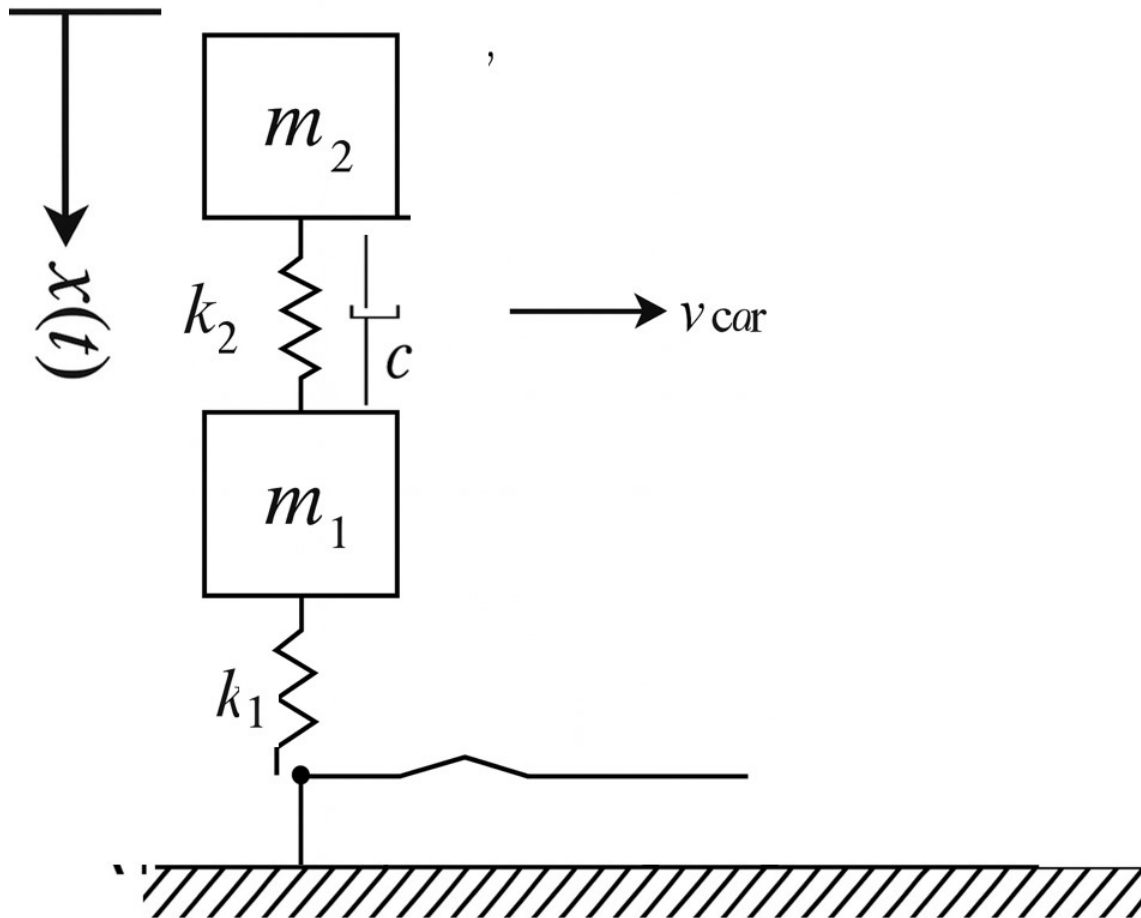


Figure Chapter7.2: Quarter-Car Suspension Model.

The model consists of:

- m_1 : the lower mass (unsprung mass), representing the wheel and axle
- m_2 : the upper mass (sprung mass), representing the chassis and vehicle body

The following elements connect the components:

- A spring of stiffness k_1 , connecting m_1 to the inertial ground at position x_0
- A spring of stiffness k_2 , connecting m_1 to m_2
- A damper with damping coefficient c , also between m_1 and m_2

Define the vertical displacements as absolute positions measured from the inertial ground:

- $x_1(t)$: absolute vertical displacement of m_1
- $x_2(t)$: absolute vertical displacement of m_2
- x_0 : fixed vertical position of the inertial ground (typically set to zero)

Equations of Motion

Applying Newton's second law in the vertical direction:

For the lower mass m_1 :

\$\$

$$m_1 \ddot{x}_1 = -k_1(x_1 - x_0) + k_2(x_2 - x_1) + c(\dot{x}_2 - \dot{x}_1)$$

This equation accounts for:

- The restoring force from the road spring: $-k_1(x_1 - x_0)$
- The coupling spring between the two masses: $+k_2(x_2 - x_1)$
- The relative damping force: $+c(\dot{x}_2 - \dot{x}_1)$

For the upper mass m_2 :

\$\$

$$m_2 \ddot{x}_2 = -k_2(x_2 - x_1) - c(\dot{x}_2 - \dot{x}_1)$$

This equation models the reaction force from the lower mass due to the spring and damper.

System Coupling and Order

These coupled second-order differential equations describe a fourth-order dynamic system. The relative motion terms $(x_2 - x_1)$ and $(\dot{x}_2 - \dot{x}_1)$ model the interaction between the sprung and unsprung masses — a key characteristic of suspension dynamics. The presence of the inertial reference x_0 allows for incorporation of road disturbances or actuator inputs in future control models.

Rotational Dynamics

Rotational systems obey Newton's second law in angular form:

\$\$

$$\sum \tau = J \alpha = J \ddot{\theta}(t)$$

where:

- τ is net torque applied to the system
- J is the moment of inertia about the axis of rotation
- $\theta(t)$ is the angular position

Analogous to mass-spring-damper translational systems, we can add:

\$\$

$$J \ddot{\theta} + b \dot{\theta} + k \theta = \tau(t)$$

Here, b is torsional damping and k is angular stiffness (e.g., from torsional springs).

This is structurally identical to the mass-spring-damper system — all the second-order analysis tools (natural frequency, damping ratio, etc.) apply directly.

Pendulum and Inverted Pendulum

A pendulum is a prototypical nonlinear system. Consider a simple pendulum of mass m , length L , with angular displacement $\theta(t)$ from vertical.

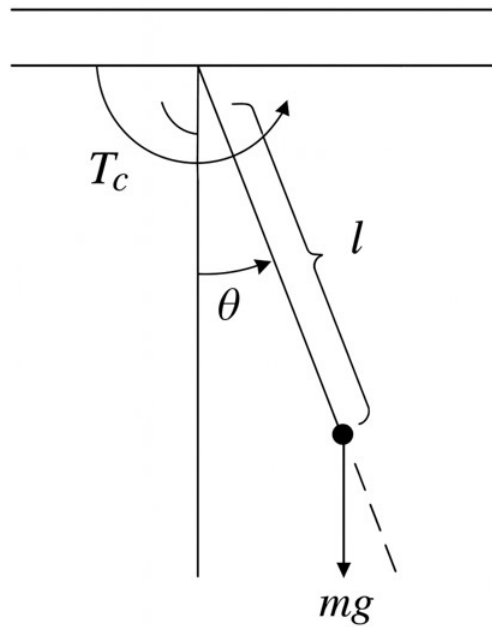


Figure Chapter7.3: Pendulum.

The torque due to gravity is:

$$\tau = -mgL \sin(\theta)$$

Using $\sum \tau = J\ddot{\theta}$ and $J = mL^2$, we get:

$$mL^2 \ddot{\theta} + mgL \sin(\theta) = 0$$

$$\ddot{\theta} + \frac{g}{L} \sin(\theta) = 0$$

This is a nonlinear second-order ODE.

Linearization around equilibrium:

- For $\theta \approx 0$: $\sin(\theta) \approx \theta \Rightarrow \ddot{\theta} + \frac{g}{L}\theta = 0$
- The solution is harmonic oscillation: $\omega_n = \sqrt{\frac{g}{L}}$

Inverted Pendulum:

The pivot is below the mass — the torque due to gravity is destabilizing.

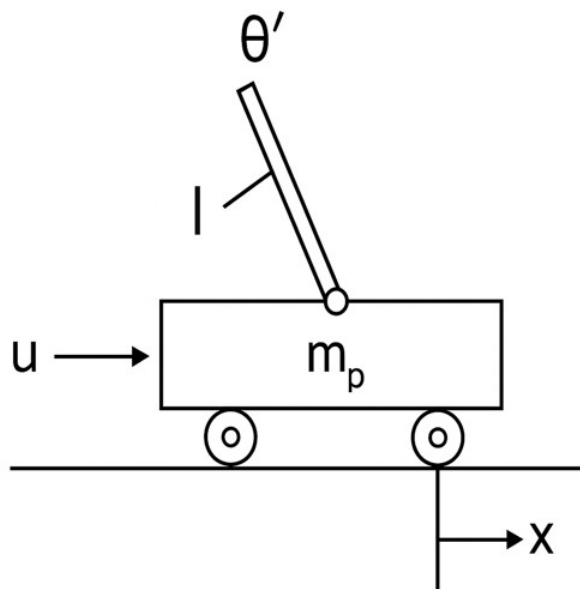


Figure Chapter7.4: Inverted Pendulum.

\$\$

$$\ddot{\theta} - \frac{g}{L} \theta = 0$$

$\Rightarrow$

$$\theta(t) = \theta_0 e^{\sqrt{g/L} t}$$

$\backslash$

This exhibits exponential divergence — instability. Such a system cannot balance itself without feedback. This is the same principle as balancing a stick on your palm (see earlier lecture section).

Inverted pendulums are central in many modern control labs: Segways, rockets, cart-pole systems, and humanoid robotics.

Gears and Gear Trains

Gears convert rotational motion and torque between shafts. For gears with radii r_1, r_2 and number of teeth N_1, N_2 , the gear ratio is:

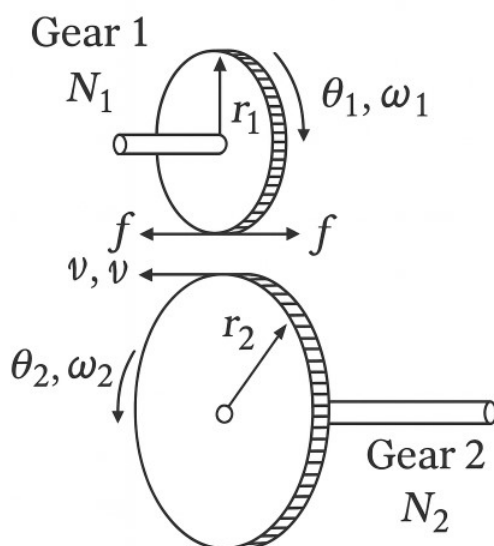


Figure Chapter7.5: Gears

$$\begin{aligned} n &= \frac{N_2}{N_1} = \frac{r_2}{r_1} = \frac{\omega_1}{\omega_2} \\ &\quad \text{and} \\ \tau_2 &= n \tau_1 \end{aligned}$$

Key principles:

- Angular displacement and velocity are inversely related to gear ratio
- Torque is multiplied by gear ratio

When modeling systems with gears, it's useful to reflect all inertias and torques to a common shaft. For instance, to reflect J_2 to shaft 1:

$$J_{\text{eq}} = J_1 + n^2 J_2$$

This simplifies the system to a single equivalent inertia. Similar transformations apply to damping and stiffness.

First-Order Thermal System (Lumped Heat Transfer)

A common thermal model is a lumped body exchanging heat with the ambient through a thermal resistance R , having thermal capacitance C . Let $T(t)$ be the body temperature, $T_a(t)$ the ambient.

$$C \frac{dT}{dt} = -\frac{1}{R}(T - T_a)$$

This is a first-order system with time constant $\tau = RC$. The system exponentially approaches ambient temperature. This same structure models RC electrical circuits, fluid tanks, and population decay.

Unity Feedback Configuration

This course begins with the unity feedback configuration, a canonical architecture in control theory. While more complex architectures exist, the unity feedback structure is sufficiently general to illustrate core principles and to control a wide range of systems.

In a unity feedback system, the output is subtracted from the reference input to form the error signal. This error is then fed to a controller, which applies a control input to the plant. The plant output is measured and looped back to be compared with the input. This closed-loop configuration is illustrated here:

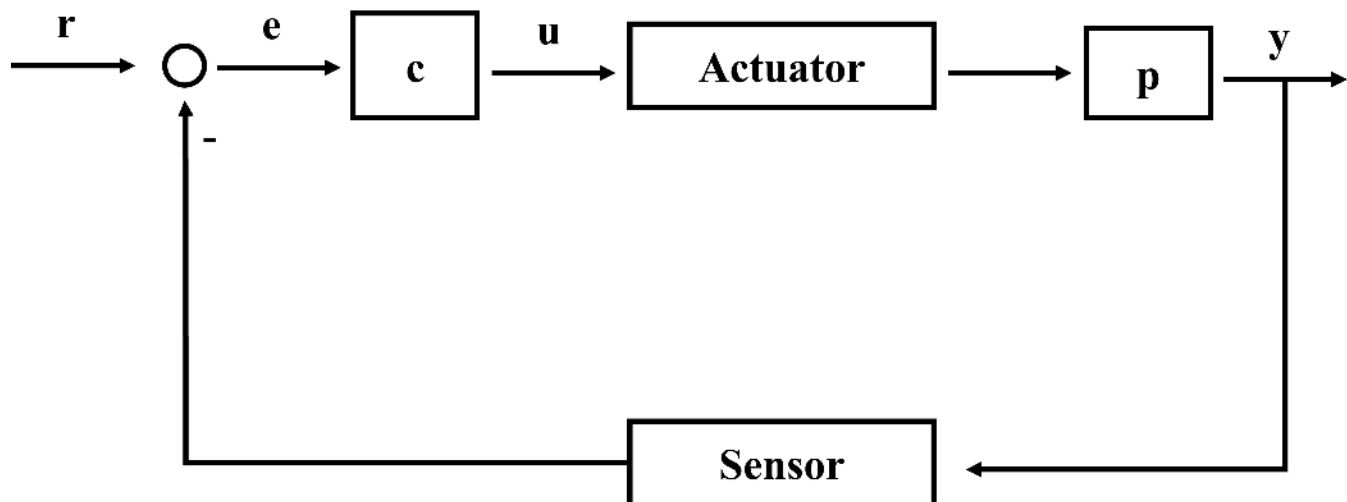


Figure Chapter7.6: Unity Feedback Closed Loop System

There are several critical components to the block diagram and inputs and outputs, much more on that later. One major component or block in this diagram is the plant, p , which is our dynamic system that we are trying to control. This could be a robot, temperature control, pendulum, etc. y is the measured output of our system and it is the variable that we are trying to control and it will be measured with a sensor. r is the reference or the signal that we want, i.e. the desired output. e is the error and is simply the difference between r and y :

\$\$

$$e = r - y$$

\]

c is the controller and we will spend a lot of the class to develop this. The output from the controller is u , the control command, and is the input to the plant via some type of actuator which can be some type of motor or other actuator.

The goals of feedback controls are:

- Stability
- Tracking: system output must track the reference signal as close as possible
- Disturbance rejection: system output must be as insensitive as possible to disturbance inputs
- Robustness

Now this may seem very abstract so let's put this in the context of an example that we will come back to over and over and over again in this class so learn to love this example — and that is cruise control in a car.

For this application, we set a particular speed, say 65 mph — that will be our reference, r . The actual speed of the car would be y and is measured via a sensor, perhaps some type of conversion of tire rotation to translational displacement. The error e is just the difference between these two signals. The plant p in this example would be the vehicle. The cruise control will take the signal from the error and then send a control command u to the actuator, which will be the accelerator or something of the like.

Let's think about another example like a Segway, which is similar to an example and lab we will conduct in this course which is an inverted pendulum where we actually use unstable dynamics to drive around. For a Segway we want to remain balanced, so our measured output y will be the angle between the Segway and the vertical line perpendicular to the surface. We want the reference r , or that angle, to be 0° and the control command will control the motor torque to the Segway wheel.

You can extend this to HVAC systems where the plant is the house, the y is the measured temperature, r is the desired temperature, and the control command will be sent to the AC damper or a valve control.

This feedback loop introduces several critical properties:

- Improved disturbance rejection
- Reduced steady-state error

- Enhanced tracking performance
- Robustness to model uncertainties

Block Diagram Analysis

Throughout this course, feedback control systems will be analyzed using block diagrams. These diagrams abstract the mathematical operations of each subsystem, allowing for algebraic manipulation of entire control loops using transfer functions. This simplifies stability analysis, error dynamics, and sensitivity functions, particularly in the frequency domain.

Feedback control is essential for the design and stabilization of dynamic systems, particularly those that are unstable in their natural configuration. To understand the role of feedback, consider the consequences of removing the feedback path in a control loop. If we eliminate the sensor and the associated feedback signal, the system becomes an open-loop configuration. This configuration lacks the capability to automatically correct deviations between the desired and actual outputs.

To illustrate this, consider the classic example of balancing a stick vertically on your palm. In the context of a unity feedback control system (so named because the feedback signal is directly subtracted from the reference), we can identify the following components:

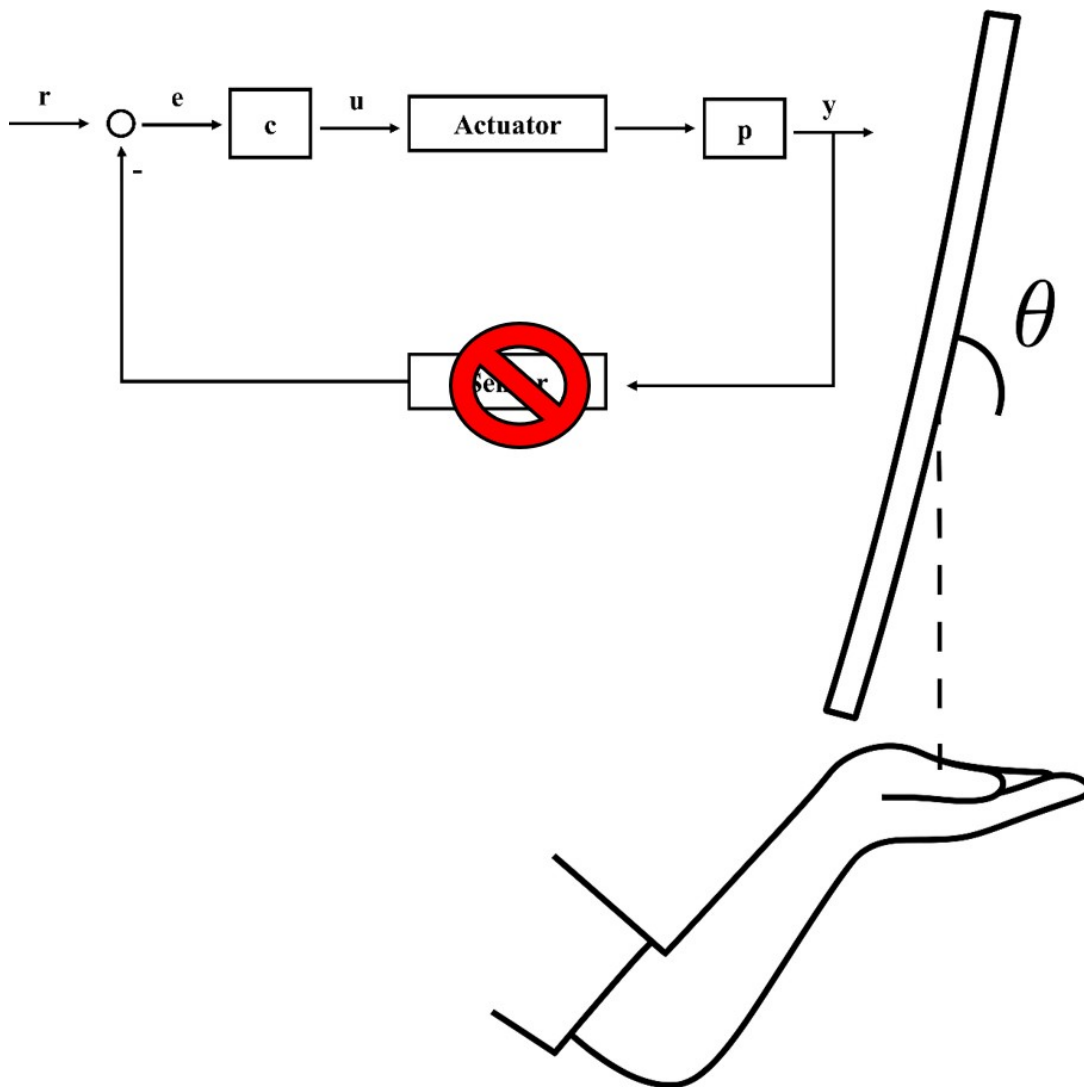


Figure Chapter7.7: Balancing a Stick on My Hand.

- Plant: the stick, which is inherently unstable when balanced upright
- Output $y(t)$: the angular deviation of the stick from the vertical axis
- Reference $r(t)$: the desired angle, ideally $r(t) = 0^\circ$
- Sensor: your eyes, providing visual feedback
- Controller C : your brain, processing the error
- Actuator: your arm or hand muscles
- Control input $u(t)$: the motion applied to keep the stick upright

In this feedback loop, you continuously observe the angle $y(t)$, compare it to the reference $r(t)$, and apply corrective motion $u(t)$ to minimize the error.

Now consider removing the sensor — closing your eyes. This breaks the feedback loop. Without sensory input, the controller has no error information and cannot generate a meaningful corrective action. As a result, the plant becomes unstable and the stick falls. This demonstrates a key point: without feedback, inherently unstable systems cannot be stabilized.

Open-Loop Control Systems

In an open-loop system, there is no feedback: the controller sends an input to the actuator, which influences the plant, and the output is observed without adjustment. The error signal $e(t) = r(t) - y(t)$ is not computed or used.

Such systems are viable only under specific conditions, primarily when the plant dynamics are well understood, repeatable, and stable. Since no corrections are made based on the output, open-loop control cannot adapt to disturbances or uncertainty.

A common example is a toaster:

- Plant: the toaster, with thermal dynamics governed by heat transfer
- Controller: the knob, regulating time or power level
- Actuator: the heating element
- Output $y(t)$: the toast produced
- Reference $r(t)$: desired doneness, e.g., golden brown

There is no feedback, no sensor measures toast color or crispness. If the bread changes (e.g., frozen vs. fresh, thin vs. thick), the system cannot adapt. Manual adjustments are required, and the output becomes inconsistent. For this reason, open-loop control is rarely suitable for complex or sensitive systems.

Closed-Loop Control Systems

In contrast, closed-loop (feedback) control systems incorporate real-time measurement and comparison between the desired output $r(t)$ and the actual output $y(t)$. The difference, the error signal $e(t) = r(t) - y(t)$, is used by the controller to determine the appropriate control input.

This architecture enables:

- Stabilization of unstable systems
- Performance improvement in already stable systems
- Robustness against disturbances and modeling uncertainty
- Tuning of speed, overshoot, and steady-state error

In feedback systems, controller design focuses on shaping system behavior to meet specific performance specifications, often involving trade-offs between speed of response, stability margins, and sensitivity to disturbances.

Understanding the dynamics of the plant is essential in control system design. The plant is the component of a feedback system that we are trying to control, typically a physical, dynamic system. To analyze such systems effectively, a solid foundation in dynamics is required.

A dynamic system is characterized by a time-dependent input-output relationship, where the output at the current time depends not only on the current input but also on past inputs and system states. In other words, dynamic systems have memory.

Consider the example of riding a bicycle. The input $u(t)$ is the torque applied by pedaling, and the output $y(t)$ is the bicycle's velocity. Even if pedaling ceases, the bicycle may continue to coast — this demonstrates that the current velocity depends on prior input. This is a classic example of a first-order dynamic system.

Car Velocity Plant Function

Let us now consider a more formal example involving a car. The control input $u(t)$ is the throttle angle, and the system output $y(t)$ is the vehicle's velocity. Clearly, the input and output are not linearly or instantaneously related — this is a dynamic system.

To study this system properly, we need to characterize how the input affects the output over time. That is the role of dynamic system modeling. Specifically, we want to develop a mathematical model that captures the behavior of the system using differential equations.

For this course, we will primarily focus on mechanical dynamic systems, and thus we will rely heavily on Newton's Second Law and free-body diagrams (FBDs) as our modeling tools. While control systems can span other physical domains (e.g., electrical systems using Kirchhoff's laws), our examples will primarily involve translational mechanical systems.

Let us model the car using a simplified, one-dimensional representation. The assumptions are:

- The vehicle has mass m
- The input is the forward thrust force from the engine $u(t)$
- The resisting force is aerodynamic drag, modeled as $F_d = bv(t)$, where b is a damping (drag) coefficient
- The output is velocity $v(t)$

We apply Newton's Second Law in the direction of motion:

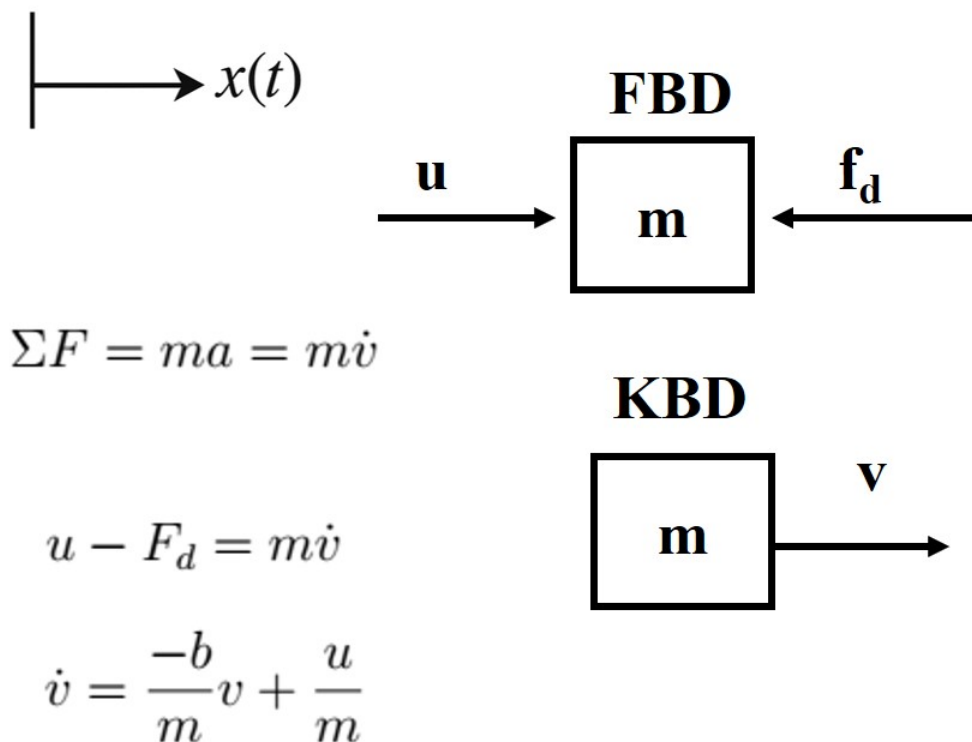


Figure Chapter7.8: Car Example.

\$\$
 $\sum F = ma = m\dot{v}(t)$
 \]

$$u(t) - F_d = m \dot{v}(t)$$

$$u(t) - b v(t) = m \dot{v}(t)$$

Rearranging, we obtain the first-order linear ordinary differential equation (ODE):

$$\dot{v}(t) = -\frac{b}{m}v(t) + \frac{1}{m}u(t)$$

This ODE captures the complete input-output dynamics of the system. It holds for arbitrary inputs, including impulse, step, or sinusoidal signals.

Let us examine the case of a constant step input $u(t) = 1$:

$$u(t) = 1 \quad \text{(unit step input)}$$

Solving the ODE with zero initial conditions yields:

$$v(t) = \frac{1}{b} \left(1 - \exp\left(-\frac{b}{m}t\right) \right)$$

This solution exhibits first-order exponential behavior, consistent with physical intuition: the velocity increases over time and asymptotically approaches a steady-state value determined by the balance between thrust and drag.

From Time Domain to Laplace (s) Domain

While time-domain modeling is fundamental, it is often advantageous to work in the Laplace domain. In this domain, differential equations become algebraic equations, and system behavior can be represented using transfer functions.

- The physical system exists in the real world.
- The time-domain model is a differential equation derived from physics.
- The Laplace-domain model is an algebraic representation via the transfer function.

These three boxes describe the same system in different but equivalent forms. The Laplace transform is the tool that lets us move between the time and frequency (s) domains. Mastery of both domains is essential for modern control analysis and design.

Laplace Transforms and Transfer Functions

To control a system, we must first understand and model its dynamic behavior. The plant refers to the system we aim to control. Through the process of modeling—usually based on Newton's Second Law—we obtain a set of ordinary differential equations (ODEs) that describe the plant's dynamics.

However, solving ODEs in the time domain can be analytically difficult and computationally expensive. To simplify the analysis, we often convert the system to the Laplace domain (s-domain) using the Laplace transform. This conversion allows us to represent dynamic relationships algebraically rather than differentially.

The Laplace transform of a time-domain function $f(t)$ is defined as:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^{\infty} f(t) e^{-st} dt$$

This operation transforms a time-domain function into an algebraic function of the complex variable s , greatly simplifying the manipulation of dynamic system models.

Transfer Function Definition

Let $u(t)$ be the input to a dynamic system and $y(t)$ be the corresponding output. The transfer function $H(s)$ is defined as:

$$H(s) = \frac{\mathcal{L}\{y(t)\}}{\mathcal{L}\{u(t)\}} = \frac{Y(s)}{U(s)}$$

This definition assumes that all initial conditions are zero.

Procedure Overview: The Four Core Applications of the Laplace Transform

1. Derive the ODE using physical laws (e.g., Newton's Second Law)
2. Take the Laplace transform of the ODE (with zero initial conditions)
3. Solve for the transfer function $H(s) = \frac{Y(s)}{U(s)}$
4. Perform inverse Laplace transforms as needed to return to the time domain

Worked Example: First-Order Car Model

Consider the simplified car model:

$$m\dot{v}(t) + b v(t) = u(t)$$

Taking the Laplace transform of both sides (assuming zero initial velocity):

$$m s V(s) + b V(s) = U(s)$$

Factor $V(s)$:

$$\begin{aligned} V(s)(ms + b) &= U(s) \quad \Rightarrow \quad \\ H(s) &= \frac{V(s)}{U(s)} = \frac{1}{ms + b} \end{aligned}$$

This is the transfer function of the system, describing how the velocity responds to the input throttle signal.

Solving an ODE via Laplace Transform

Consider the ODE:

$$\dot{v}(t) + 2v(t) = u(t), \quad v(0) = 0$$

Take Laplace transform:

$$\begin{aligned} sV(s) + 2V(s) &= U(s) \\ \Rightarrow V(s)(s+2) &= U(s) \\ \Rightarrow V(s) &= \frac{U(s)}{s+2} \end{aligned}$$

Let $u(t) = 1$ (unit step), so $U(s) = \frac{1}{s}$. Then:

\$\$

$$V(s) = \frac{1}{s(s+2)}$$

\]

Use partial fraction decomposition:

\$\$

$$\frac{1}{s(s+2)} = \frac{A}{s} + \frac{B}{s+2}$$

$$\rightarrow A = \frac{1}{2}, \quad B = -\frac{1}{2}$$

\]

\$\$

$$V(s) = \frac{1}{2} \left(\frac{1}{s} - \frac{1}{s+2} \right)$$

\]

Take inverse Laplace transform:

\$\$

$$v(t) = \frac{1}{2}(1 - e^{-2t})$$

\]

This matches the solution we obtained earlier by directly solving the differential equation.

Example: Laplace of $f(t) = te^{3t}$

$$\text{Use: } \mathcal{L}\{te^{at}\} = \frac{1}{(s-a)^2}$$

\$\$

$$\mathcal{L}\{te^{3t}\} = \frac{1}{(s-3)^2}$$

\]

The Laplace transform is a powerful tool for analyzing and solving dynamic systems. It allows us to:

- Transform complex time-domain differential equations into algebraic equations
- Define transfer functions, which relate input to output in the s-domain
- Use tables and properties to handle system responses to various inputs
- Easily move between domains using forward and inverse transforms

Mastery of this technique and its associated table of transforms is essential for efficient control system analysis and design.

Frequency Response Interpretation from Transfer Functions

Control engineers often want to understand how a system responds to sinusoidal inputs of varying frequencies. This is essential for designing controllers that reject disturbances or track desired inputs. The frequency response characterizes how each frequency component of an input signal is altered (amplified/attenuated and phase-shifted) by the system.

To analyze this, we evaluate the transfer function $H(s)$ at $s = j\omega$, where ω is a real-valued angular frequency in radians per second. This gives the frequency response of the system:

\$\$

$$H(j\omega) = |H(j\omega)| e^{j\phi(\omega)}$$

\]

This complex number encodes two key pieces of information:

- Magnitude response $|H(j\omega)|$: the amplitude gain of the output relative to the input
- Phase response $\phi(\omega)$: the phase lag or lead introduced by the system

Derivation of Frequency Response

Let the input to the system be a sinusoid:

\$\$

$$u(t) = A \cos(\omega t) = \text{Re}\{Ae^{j\omega t}\}$$

\]

For a linear time-invariant (LTI) system with transfer function $H(s)$, the Laplace transform of the input is:

\$\$

$$U(s) = \frac{A s}{s^2 + \omega^2}$$

\]

In steady-state (ignoring transients), the system's output in the frequency domain is:

\$\$

$$Y(j\omega) = H(j\omega) U(j\omega)$$

\]

Taking the inverse transform:

\$\$

$$y(t) = |H(j\omega)| A \cos(\omega t + \phi(\omega))$$

\]

This result shows that the system outputs a sinusoid at the same frequency ω , but scaled in magnitude and shifted in phase.

First-Order Low-Pass System

Let:

\$\$

$$H(s) = \frac{1}{\tau s + 1} \rightarrow H(j\omega) = \frac{1}{1 + j\omega \tau}$$

\]

Magnitude:

\$\$

$$|H(j\omega)| = \frac{1}{\sqrt{1 + (\omega \tau)^2}}$$

\]

Phase:

\$\$

$$\phi(\omega) = -\tan^{-1}(\omega \tau)$$

\]

Low-frequency behavior: $\omega \ll \frac{1}{\tau}$

\$\$

$$|H(j\omega)| \approx 1, \quad \phi(\omega) \approx 0^\circ$$

\]

High-frequency behavior: $\omega \gg \frac{1}{\tau}$

\$\$

$$|H(j\omega)| \rightarrow 0, \quad \phi(\omega) \rightarrow -90^\circ$$

\]

Underdamped Second-Order System

Consider:

\$\$

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta \omega_n s + \omega_n^2}$$

\]

Evaluating at $s = j\omega$:

\$\$

$$H(j\omega) = \frac{\omega_n^2}{(\omega_n^2 - \omega^2) + j(2\zeta \omega_n \omega)}$$

Magnitude:

\$\$

$$|H(j\omega)| = \frac{\omega_n^2}{\sqrt{(\omega_n^2 - \omega^2)^2 + (2\zeta \omega_n \omega)^2}}$$

Phase:

\$\$

$$\phi(\omega) = -\tan^{-1}\left(\frac{2\zeta \omega_n \omega}{\omega_n^2 - \omega^2}\right)$$

Key behaviors:

- Peak occurs near $\omega = \omega_n$ if $\zeta < \frac{1}{\sqrt{2}}$
- Phase shifts from 0° to -180°

Why Frequency Response Matters

- Filter Design: shaping gain to attenuate noise/disturbances
- Resonance Control: identifying and avoiding frequencies that cause large oscillations
- Stability Margins: gain and phase margins directly measurable on Bode plots
- Loop Shaping: designing feedback controllers (e.g., lead/lag compensators)

Block Diagram Algebra and Transfer Function Derivations

In a typical feedback control system, we define:

- $R(s)$: Reference input
- $C(s)$: Controller transfer function
- $P(s)$: Plant transfer function
- $Y(s)$: Output
- $E(s) = R(s) - Y(s)$: Error signal
- $U(s) = C(s)E(s)$: Control effort
- $D(s)$: Disturbance input (additive before plant)
- $N(s)$: Measurement noise (additive after plant)

We now derive transfer functions for different scenarios using block diagram algebra.

1. Open-Loop Transfer Function (No Feedback)

Assume no feedback loop. The signal path is simply:

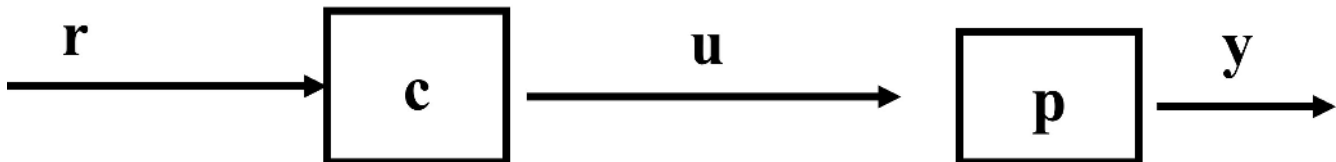


Figure Chapter7.9: Open Loop Transfer Function

Then:

\$\$

$$U(s) = C(s) R(s), \quad Y(s) = P(s) U(s) = P(s) C(s) R(s)$$

\]

Open-loop transfer function:

\$\$

$$\frac{Y(s)}{R(s)} = C(s) P(s)$$

\]

This is often denoted:

\$\$

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

\]

2. Closed-Loop Transfer Function (Unity Feedback)

Now assume unity feedback and no disturbance or noise:

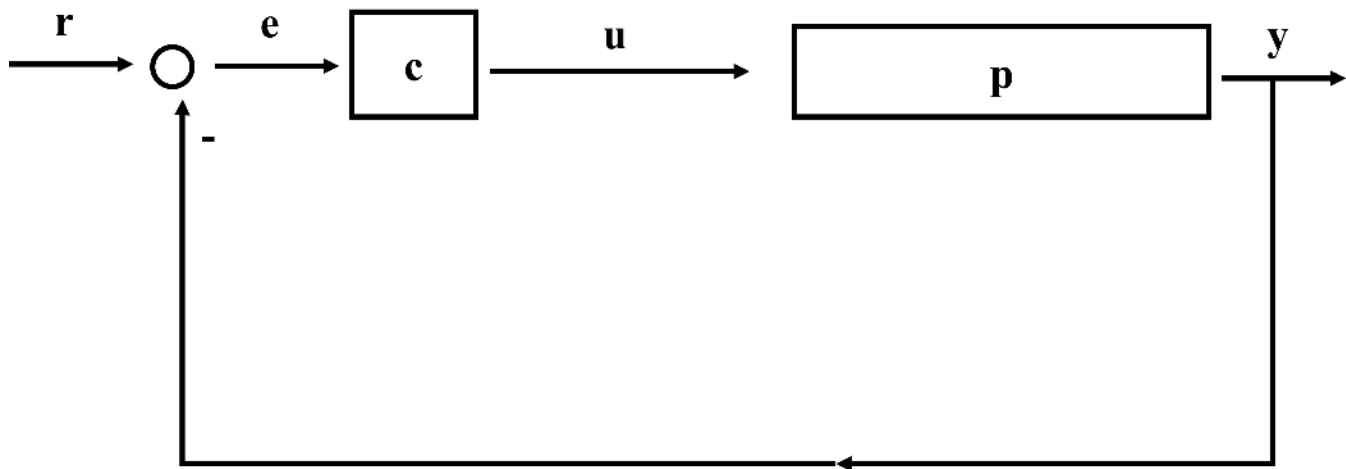


Figure Chapter7.10: Closed Loop Transfer Function

\$\$

$$E(s) = R(s) - Y(s), \quad U(s) = C(s) E(s) = C(s)[R(s) - Y(s)], \quad Y(s) = P(s) U(s)$$

\]

Substitute:

\$\$

$$Y(s) = P(s) C(s) [R(s) - Y(s)] \rightarrow Y(s) + P(s) C(s) Y(s) = P(s) C(s) R(s)$$

\]

Factor:

\$\$

$$Y(s) [1 + C(s) P(s)] = C(s) P(s) R(s)$$

\rightarrow

$$\frac{Y(s)}{R(s)} = \frac{C(s) P(s)}{1 + C(s) P(s)}$$

\]

3. Output Transfer Function with Disturbance $W(s)$

Disturbance $W(s)$ is added before the plant:

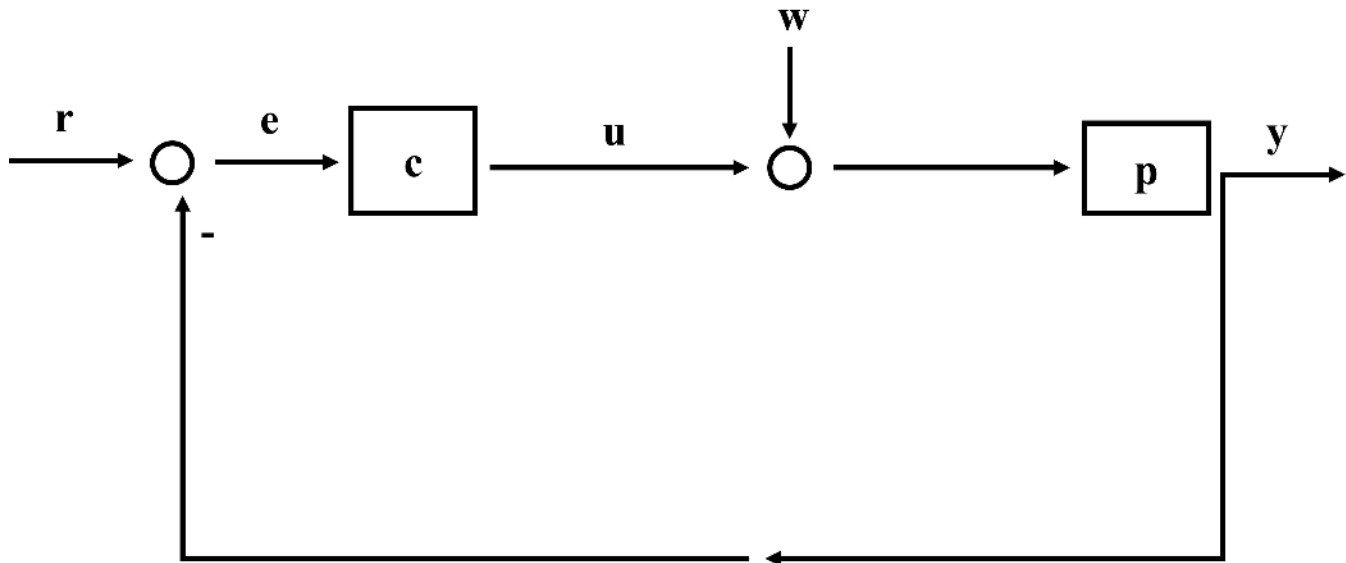


Figure Chapter7.11: Closed Loop Transfer Function with Disturbance $W(s)$

\$\$

$$U(s) = C(s) [R(s) - Y(s)]$$

$$\rightarrow Y(s) = P(s) U(s) + P(s) W(s)$$

\]

Substitute:

\$\$

$$Y(s) = P(s) C(s) [R(s) - Y(s)] + P(s) W(s)$$

$$\rightarrow Y(s)[1 + C(s) P(s)] = C(s) P(s) R(s) + P(s) W(s)$$

\]

Final expression:

\$\$

$$Y(s) = \frac{C(s) P(s)}{1 + C(s) P(s)} R(s) +$$

$$\frac{P(s)}{1 + C(s) P(s)} W(s)$$

\]

4. Output Transfer Function with Measurement Noise $V(s)$

Now let $V(s)$ be added to the measured output before it feeds back to the controller:

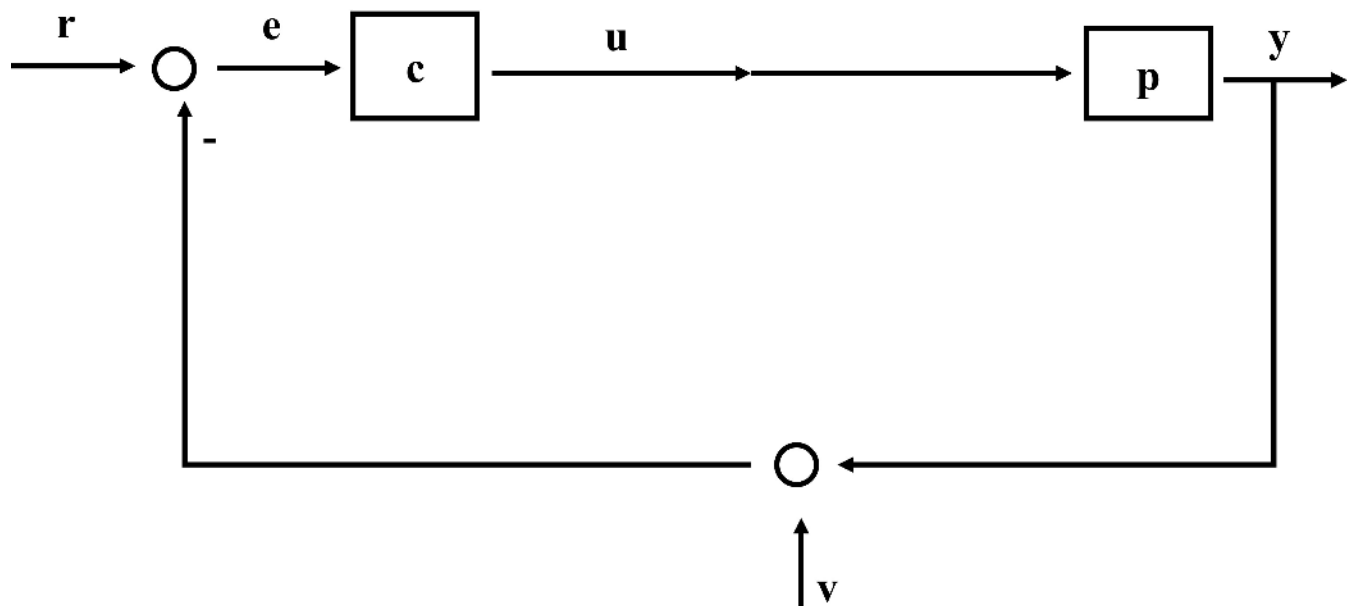


Figure Chapter7.12: Closed Loop Transfer Function with Noise $V(s)$

\$\$

$$E(s) = R(s) - [Y(s) + V(s)], \quad U(s) = C(s) E(s)$$

\]

Substitute:

\$\$

$$U(s) = C(s)[R(s) - Y(s) - V(s)], \quad Y(s) = P(s) U(s)$$

\]

\$\$

$$Y(s) = P(s) C(s)[R(s) - Y(s) - V(s)]$$

$$\Rightarrow Y(s)[1 + C(s) P(s)] = C(s) P(s) R(s) - C(s) P(s) V(s)$$

\]

Final expression:

\$\$

$$Y(s) = \frac{C(s) P(s)}{1 + C(s) P(s)} R(s) -$$

$$\frac{C(s) P(s)}{1 + C(s) P(s)} V(s)$$

\]

5. General Case with $R(s)$, $W(s)$, and $V(s)$

Now we consider the most complete block diagram, where the system is influenced by:

- $R(s)$: Reference signal
- $W(s)$: Disturbance (e.g., wind, vibration)
- $V(s)$: Measurement noise

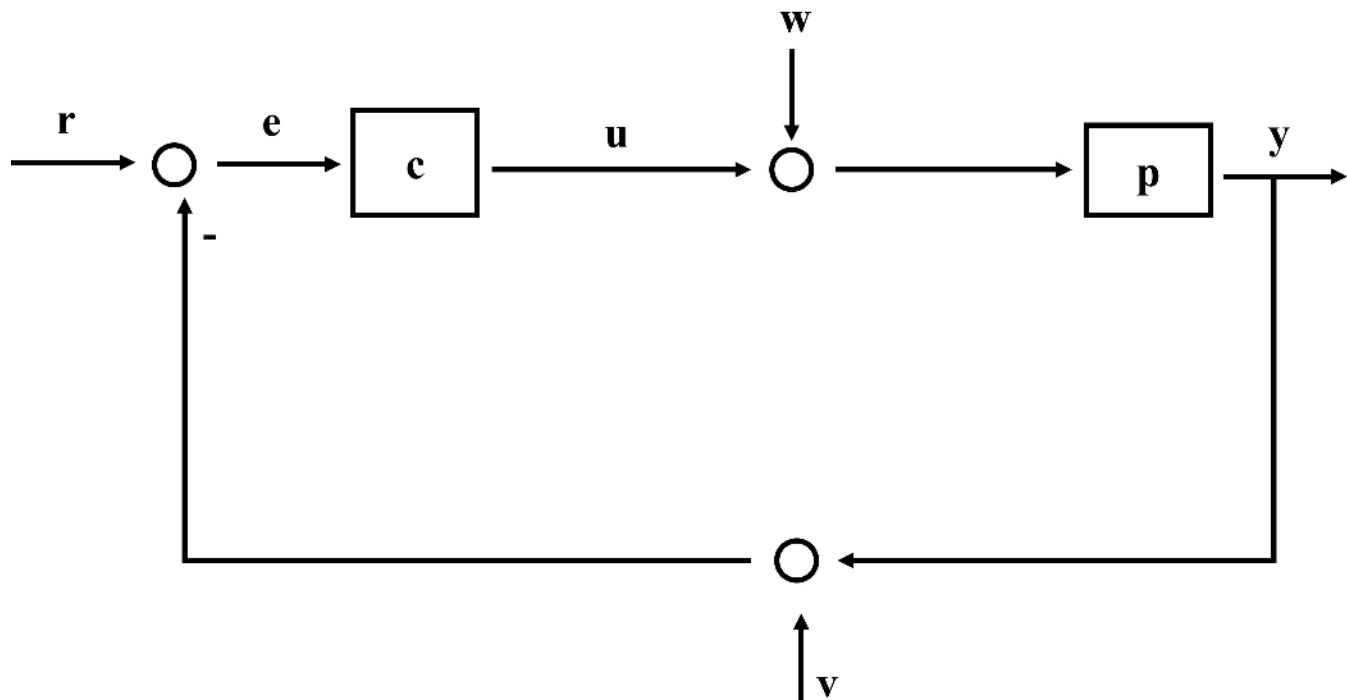


Figure Chapter7.13: Closed Loop Transfer Function with Noise $V(s)$ and Disturbance $W(s)$.

We combine everything from the previous derivations:

\$\$

$$Y(s) = \frac{C(s) P(s)}{1 + C(s) P(s)} R(s) + \frac{P(s)}{1 + C(s) P(s)} W(s) - \frac{C(s) P(s)}{1 + C(s) P(s)} V(s)$$

\]

This is the most general expression for the system output. Each term corresponds to:

- $R(s)$ driving the response via the closed-loop sensitivity function
- $W(s)$ being attenuated by the loop gain
- $V(s)$ being amplified by feedback (if $C(s)$ is too aggressive)

Key Takeaways

- $L(s) = C(s)P(s)$: the open-loop transfer function
- The term $1 + L(s)$ always appears in the denominator of feedback systems
- Feedback attenuates disturbances and noise depending on where they enter
- Closed-loop dynamics are governed entirely by the roots of $1 + C(s)P(s) = 0$
- Understanding and manipulating this structure is the core of classical control design

This page titled [Chapter 7: Introduction to Feedback Control, Laplace Transforms, and Transfer Functions](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 8: System Properties, Poles and Zeroes, System Response and Stability

System Properties, Poles and Zeroes, System Response and Stability

In this lecture, we move beyond modeling and transfer function computation to understand the underlying system properties that influence dynamic behavior. While transfer functions provide a convenient and elegant framework in the Laplace domain, they are based on certain assumptions. Understanding the limitations and implications of those assumptions is crucial for using transfer functions properly in analysis and control design.

Transfer Function Limitations

The transfer function

$$H(s) = \frac{Y(s)}{U(s)},$$

derived via Laplace transforms from a system's differential equations, assumes that the system is:

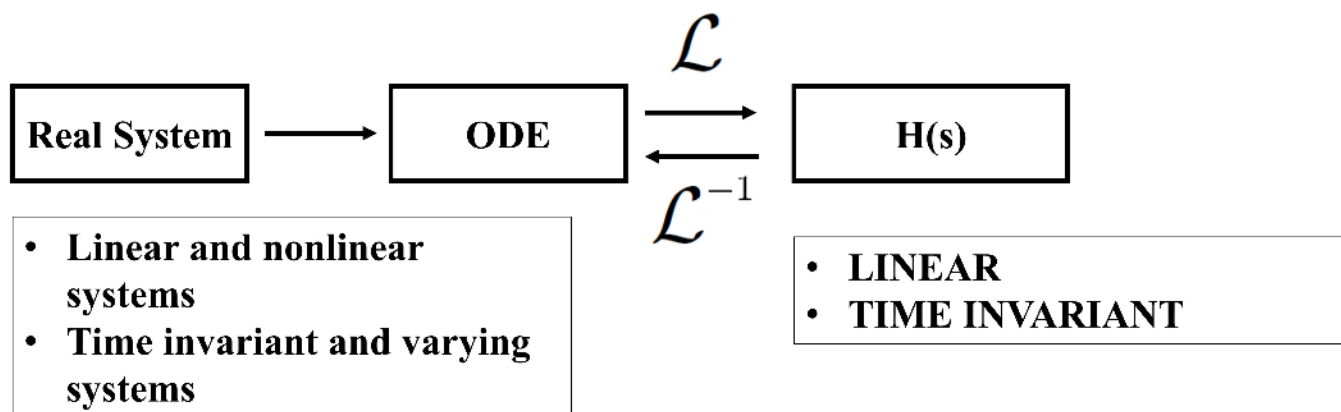


Figure Chapter8.1: System Properties for the Plant

- Linear – Follows the principle of superposition.
- Time-Invariant – System properties do not change with time.
- Causal – Output depends only on current and past inputs.

These assumptions do not hold for many real systems, which may be nonlinear, time-varying, or even non-causal in simulations. Transfer functions therefore serve as linearized approximations of the actual dynamics, valid only around an equilibrium point or within a certain input range.

Linearity

A system is linear if it satisfies both:

- Homogeneity:

$$u(t) \rightarrow y(t) \Rightarrow \alpha u(t) \rightarrow \alpha y(t)$$

- Additivity:

$$u_1(t) \rightarrow y_1(t), u_2(t) \rightarrow y_2(t) \Rightarrow u_1 + u_2 \rightarrow y_1 + y_2$$

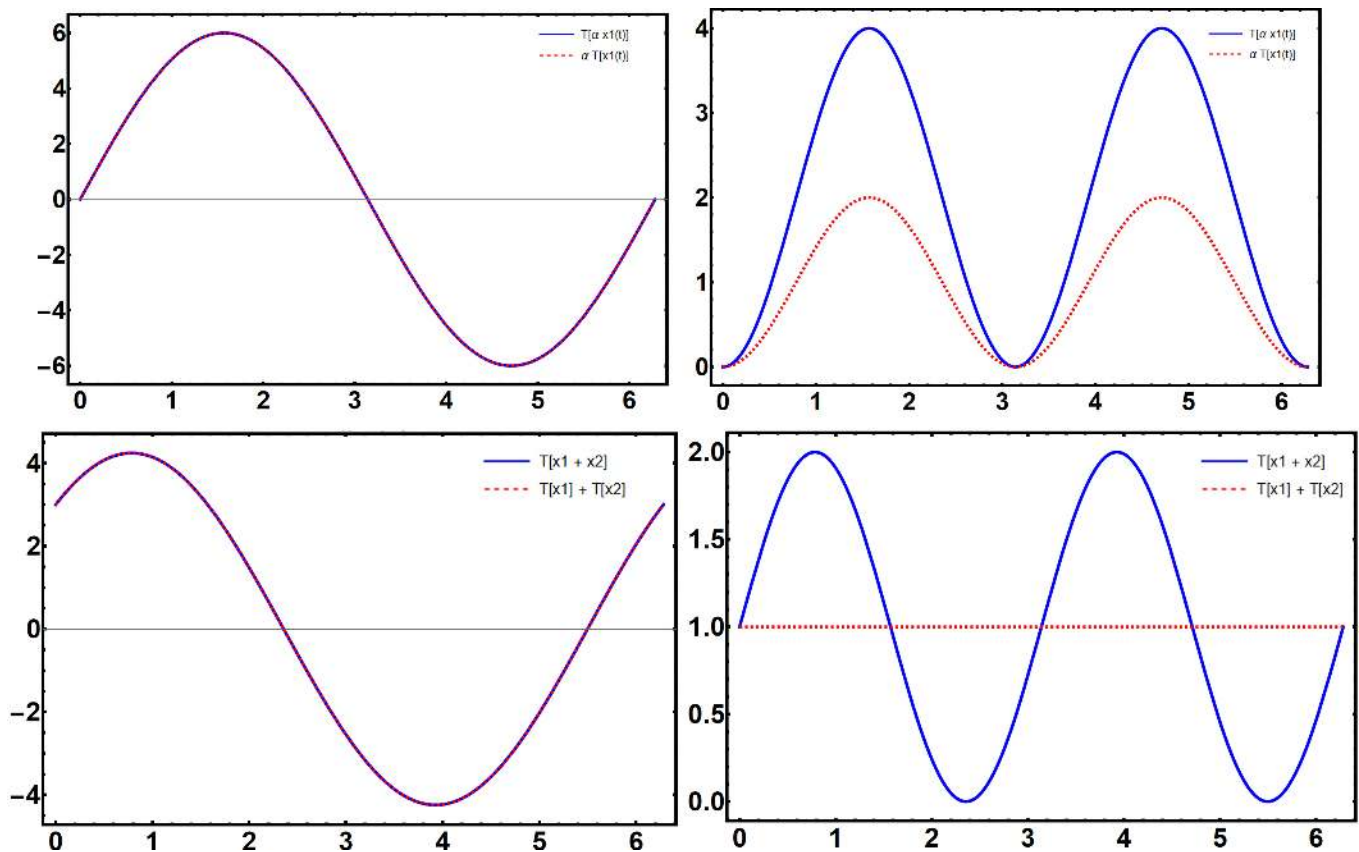


Figure Chapter8.2: Linearity Visualization

Linear System:

$$T[x(t)] = 3x(t)$$

This system satisfies both homogeneity and additivity:

- Homogeneity:

$$T[\alpha x(t)] = 3\alpha x(t) = \alpha T[x(t)]$$

- Additivity:

$$T[x_1(t) + x_2(t)] = 3(x_1(t) + x_2(t)) = 3x_1(t) + 3x_2(t) = T[x_1(t)] + T[x_2(t)]$$

Nonlinear System:

$$T[x(t)] = x(t)^2$$

This system violates both:

- Homogeneity:

$$T[\alpha x(t)] = (\alpha x(t))^2 = \alpha^2 x(t)^2 \neq \alpha x(t)^2 = \alpha T[x(t)]$$

- Additivity:

$$T[x_1(t) + x_2(t)] = (x_1(t) + x_2(t))^2 \neq x_1(t)^2 + x_2(t)^2 = T[x_1(t)] + T[x_2(t)]$$

Linear System – Mass-Spring-Damper

$$m\ddot{x} + c\dot{x} + kx = F(t)$$

This is linear since the system response scales and sums appropriately for linear combinations of inputs.

Nonlinear System – Friction or Saturation

$$m\ddot{x} + c\dot{x} + kx + \mu \sin(\dot{x}) = F(t)$$

This is nonlinear due to the discontinuous and non-scaling nature of the friction term.

Time Invariance

A system is time-invariant if a time shift in the input causes an identical shift in the output. That is:

$$u(t) \rightarrow y(t) \Rightarrow u(t - t_0) \rightarrow y(t - t_0)$$

Invariant System:

$$T[x(t)] = 3x(t)$$

$$\begin{aligned} T[x(t - t_0)] &= 3x(t - t_0) \\ y(t - t_0) &= 3x(t - t_0) \end{aligned} \quad (\text{Chapter 8.1})$$

So the time-shifted output equals the output of the time-shifted input — time invariance holds.

Time-Variant System:

$$T[x(t)] = t \cdot x(t)$$

$$\begin{aligned} T[x(t - t_0)] &= t \cdot x(t - t_0) \\ y(t - t_0) &= (t - t_0) \cdot x(t - t_0) \end{aligned} \quad (\text{Chapter 8.2})$$

These two expressions are not equal — time invariance is violated.

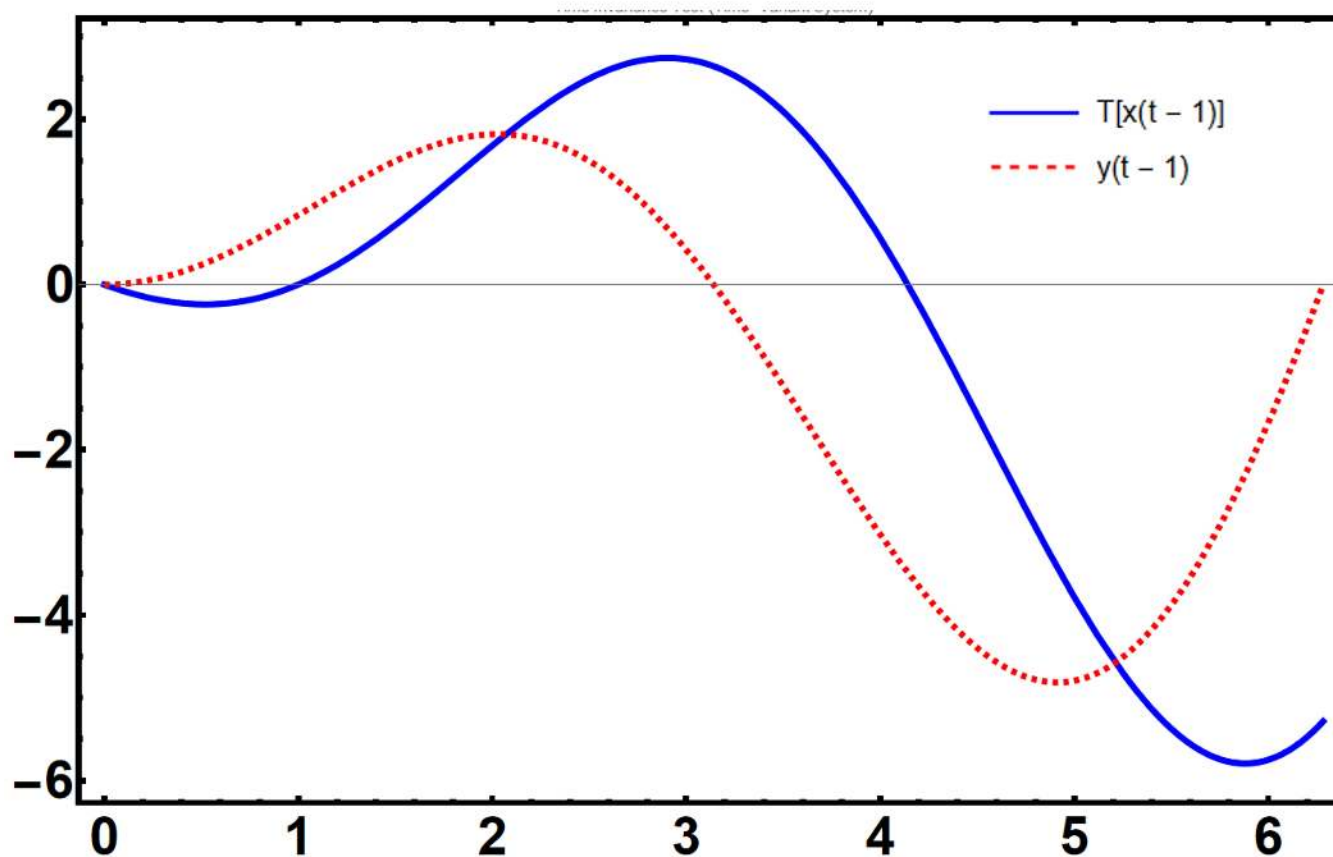
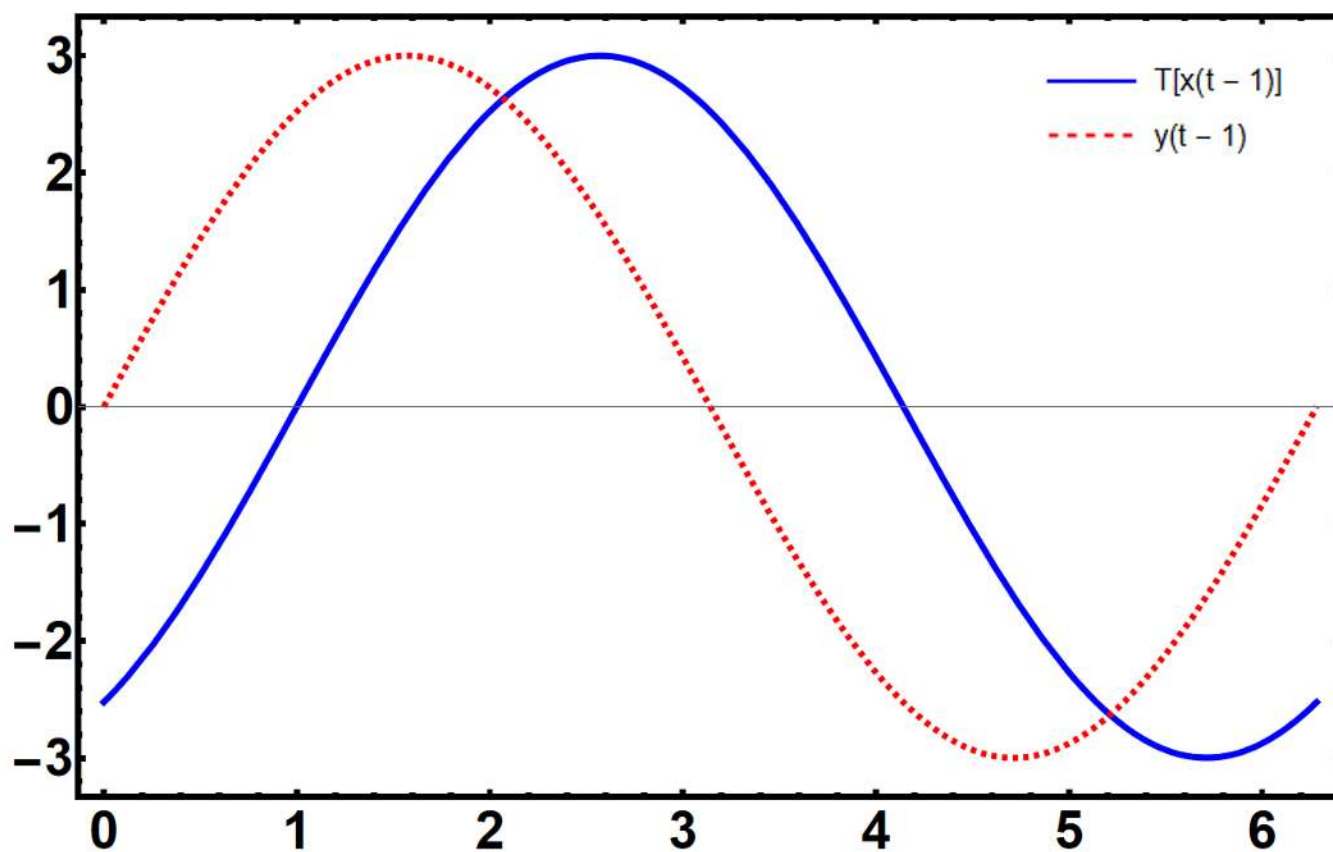


Figure Chapter8.3: Time Invariant Visualization

Real-World Example:

Driving a car – as the car consumes fuel, the total mass changes, making it a time-varying system. Transfer function models become approximate in such cases.

Causality

A causal system does not rely on future inputs.

Additionally, the degree of the denominator must be greater than or equal to the numerator:

$$H(s) = \frac{b_0 + b_1s + \dots + b_ms^m}{a_0 + a_1s + \dots + a_ns^n} \quad \text{Causal if } n \geq m \quad (\text{Chapter 8.3})$$

Example:

$$H(s) = \frac{s^2 + 3}{s^3 + 2s + 5} \quad \text{Causal} \quad H(s) = \frac{s^3 + 2}{s + 4} \quad \text{Non-Causal} \quad (\text{Chapter 8.4})$$

Poles and Zeroes of Transfer Functions

Given a transfer function:

\$\$

$$H(s) = \frac{(s - z_1)(s - z_2)\cdots(s - z_m)}{(s - p_1)(s - p_2)\cdots(s - p_n)}$$

\]

Zeroes: values of s that make $H(s) = 0$ (roots of the numerator)

Poles: values of s that make $H(s) \rightarrow \infty$ (roots of the denominator)

Poles determine:

- Stability (location in complex plane)
- Natural frequency
- Damping
- Long-term behavior

Zeroes affect:

- Response shape
- Peak amplitudes
- Transient effects

Let's consider a plant function with a delta function input. Let the plant function be:

\$\$

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

\]

The system has a pole at:

\$\$

$$s = -\sigma$$

\]

Taking the inverse Laplace transform:

\$\$

$$y(t) = \mathcal{L}^{-1}\{P(s)\} = e^{-\sigma t}$$

\]

If $\sigma > 0$, this corresponds to exponential decay. A larger σ causes faster decay.

We can visualize pole locations in the complex s -plane, which has real and imaginary axes.

Now let's return to the mass-spring-damper system:

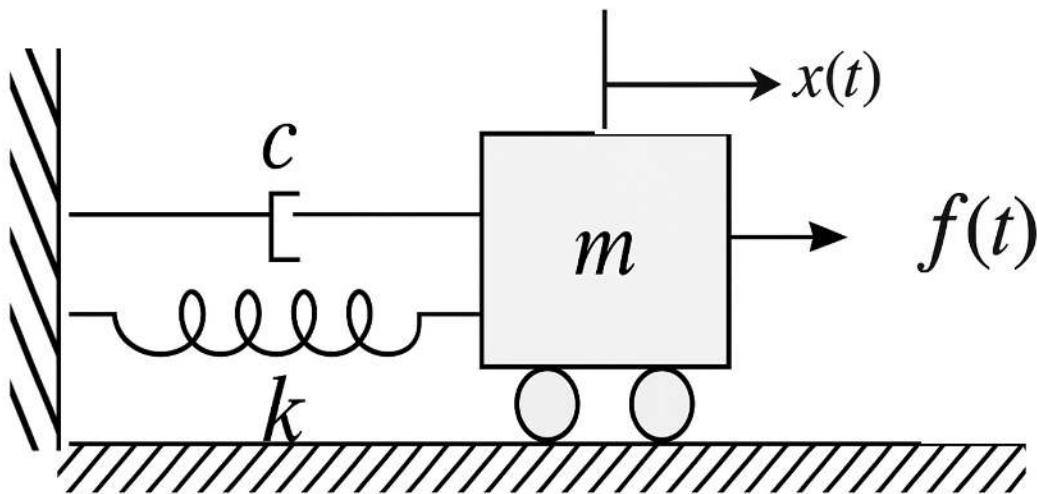


Figure Chapter8.4: Mass Spring Damper System

This system is governed by the differential equation:

$$m\ddot{x} + c\dot{x} + kx = u(t)$$

Assuming impulse input $u(t) = \delta(t)$, the plant transfer function is:

$$P(s) = \frac{X(s)}{U(s)} = \frac{1}{ms^2 + cs + k}$$

Solving the characteristic equation gives poles:

$$s_{1,2} = \frac{-c \pm \sqrt{c^2 - 4mk}}{2m}$$

Cases:

- $c^2 - 4mk > 0$: Two distinct real poles (Overdamped)
- $c^2 - 4mk = 0$: Repeated real pole (Critically damped)
- $c^2 - 4mk < 0$: Complex conjugate poles (Underdamped)

Example 1: Let $m = 1, c = 3, k = 2$. Then:

Poles: $s = -1, -2$

This yields an overdamped response:

$$x(t) = e^{-2t}(1 - e^{-t})$$

Example 2: Let $m = 1, c = 1, k = 2$. Poles are complex conjugates.

Response includes oscillations with decay envelope set by the real part, and frequency determined by the imaginary part.

Example 3: Undamped case, $c = 0$:

\$\$

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

\]

Response:

\$\$

$$x(t) = \frac{1}{\sqrt{mk}} \sin\left(\sqrt{\frac{k}{m}} t\right)$$

\]

Poles are imaginary: $s = \pm j\sqrt{\frac{k}{m}}$

Now consider a frictionless sliding mass:

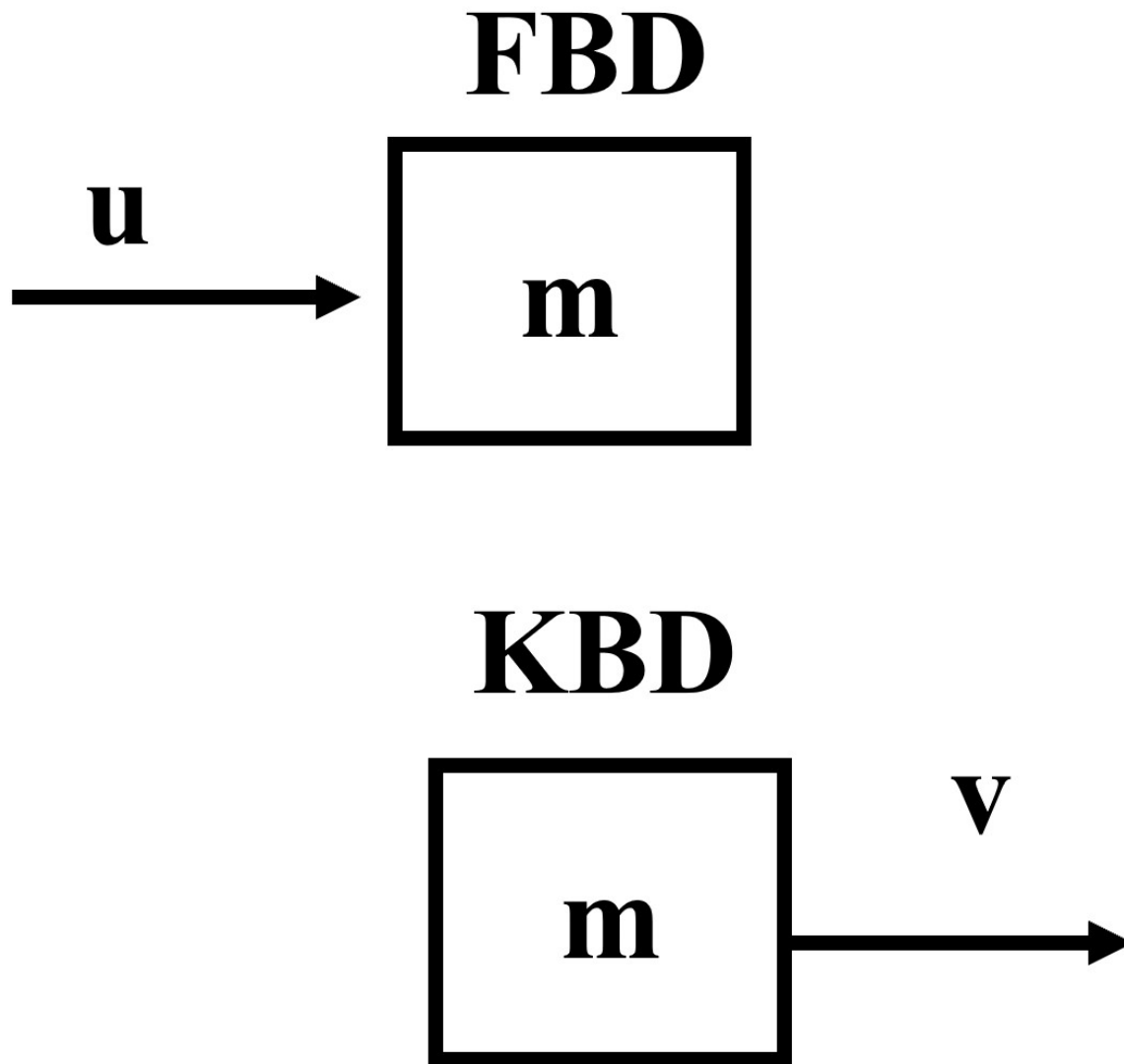


Figure Chapter8.5: Sliding Mass

Equation:

\$\$

$$u = m \dot{v}$$

\rightarrow

$$P(s) = \frac{V(s)}{U(s)} = \frac{1}{ms}$$

\]

Pole: $s = 0$

Response:

\$\$

$$v(t) = \frac{1}{m}$$

\]

This is a constant velocity response.

We can now map system responses to pole locations in the s -domain.

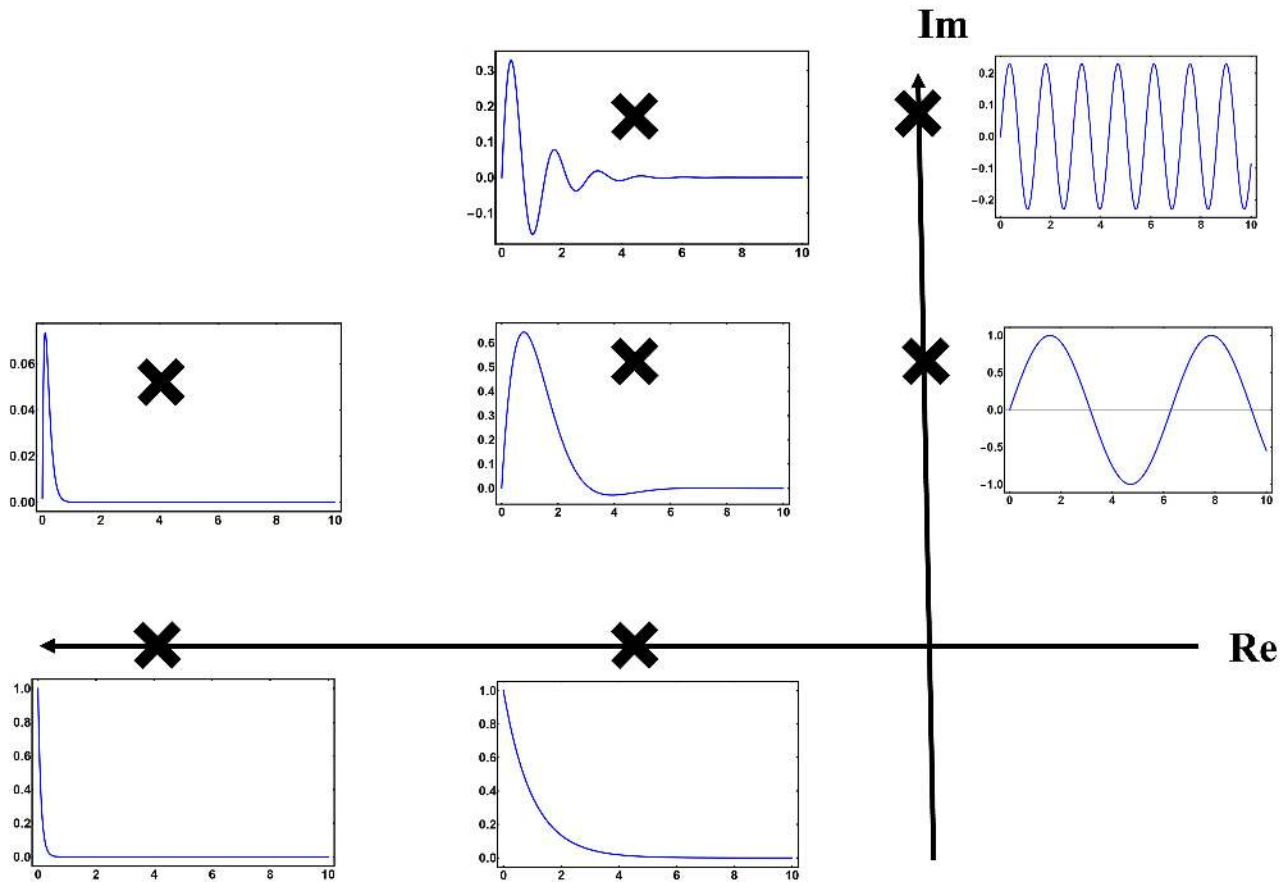


Figure Chapter8.6: System Response in S-Domain Pole Location

Summary from the figure:

- Left-half real poles → exponential decay
- Complex conjugate poles → underdamped oscillations
- Pure imaginary poles → sinusoidal oscillation
- Right-half poles → unstable growth

Zeroes Effect on the System Response

When a zero is close to or exactly cancels a pole, the pole's effect on the system response is diminished or removed.

Consider the following transfer function:

\$\$

$$P(s) = \frac{s + 2}{(s + 2)(s + 6)}$$

\]

Here we expect two real poles at $s = -2$ and $s = -6$. Ordinarily, both would contribute exponential decay terms to the time-domain response. However, because the numerator includes $(s + 2)$, the zero cancels the $s = -2$ pole. The result is:

\$\$

$$y(t) = e^{-6t}$$

\]

So the response is governed solely by the $s = -6$ pole, and the effect of the $s = -2$ pole is completely removed.

Now change the numerator slightly:

\$\$

$$P(s) = \frac{s + 2.1}{(s + 2)(s + 6)}$$

\]

Now the zero is close to the $s = -2$ pole but does not exactly cancel it. The inverse Laplace transform gives:

\$\$

$$y(t) = 0.975 e^{-6t} + 0.025 e^{-2t}$$

\]

So the zero significantly reduces the influence of the $s = -2$ pole, but does not eliminate it entirely. This is how zero placement can be used to shape system response during transient behavior.

Design Insight:

- Pole placement is critical for shaping system response.
- Zero placement can fine-tune transient performance.
- Adding zeroes near poles can help flatten or sharpen responses.

Summary of Concepts

- Transfer functions represent LTI (Linear Time-Invariant), causal systems.
- Poles determine the dominant modes of system behavior.
- Real poles lead to exponential decay.
- Complex poles lead to oscillations.
- Zeroes shape response profiles, especially during transients.
- Pole-zero maps offer intuitive visual insight into system stability and dynamics.

This page titled [Chapter 8: System Properties, Poles and Zeroes, System Response and Stability](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Chapter 9: Time Domain Specifications

Time-Domain Specifications and Response Characterization

Control engineers often describe system behavior using time-domain performance metrics. These metrics provide insight into how a system responds to changes in input, particularly in reference tracking scenarios. When designing control systems, especially those modeled as second-order systems, these specifications become critical performance targets.

Let's consider a generic 2nd order plant transfer function that can describe a number of systems, such as a spring-mass-damper system or torsional pendulum:

\$\$

$$H(s) = \frac{\sigma^2 + \omega^2}{(s + \sigma)^2 + \omega^2} = \frac{\omega_n^2}{s^2 + 2\zeta \omega_n s + \omega_n^2}$$

Here, ω_n is the undamped natural frequency and ζ is the damping ratio.

Classification of system response based on ζ :

- $\zeta > 1$: overdamped – no oscillations, slow response
- $\zeta = 1$: critically damped – fastest non-oscillatory response
- $0 < \zeta < 1$: underdamped – oscillatory, most common in mechanical systems
- $\zeta = 0$: undamped – pure sinusoidal oscillations

We are working with this general second-order transfer function:

\$\$

$$H(s) = \frac{\sigma^2 + \omega^2}{(s + \sigma)^2 + \omega^2}$$

The poles of this system are:

\$\$

$$s = -\sigma \pm j\omega$$

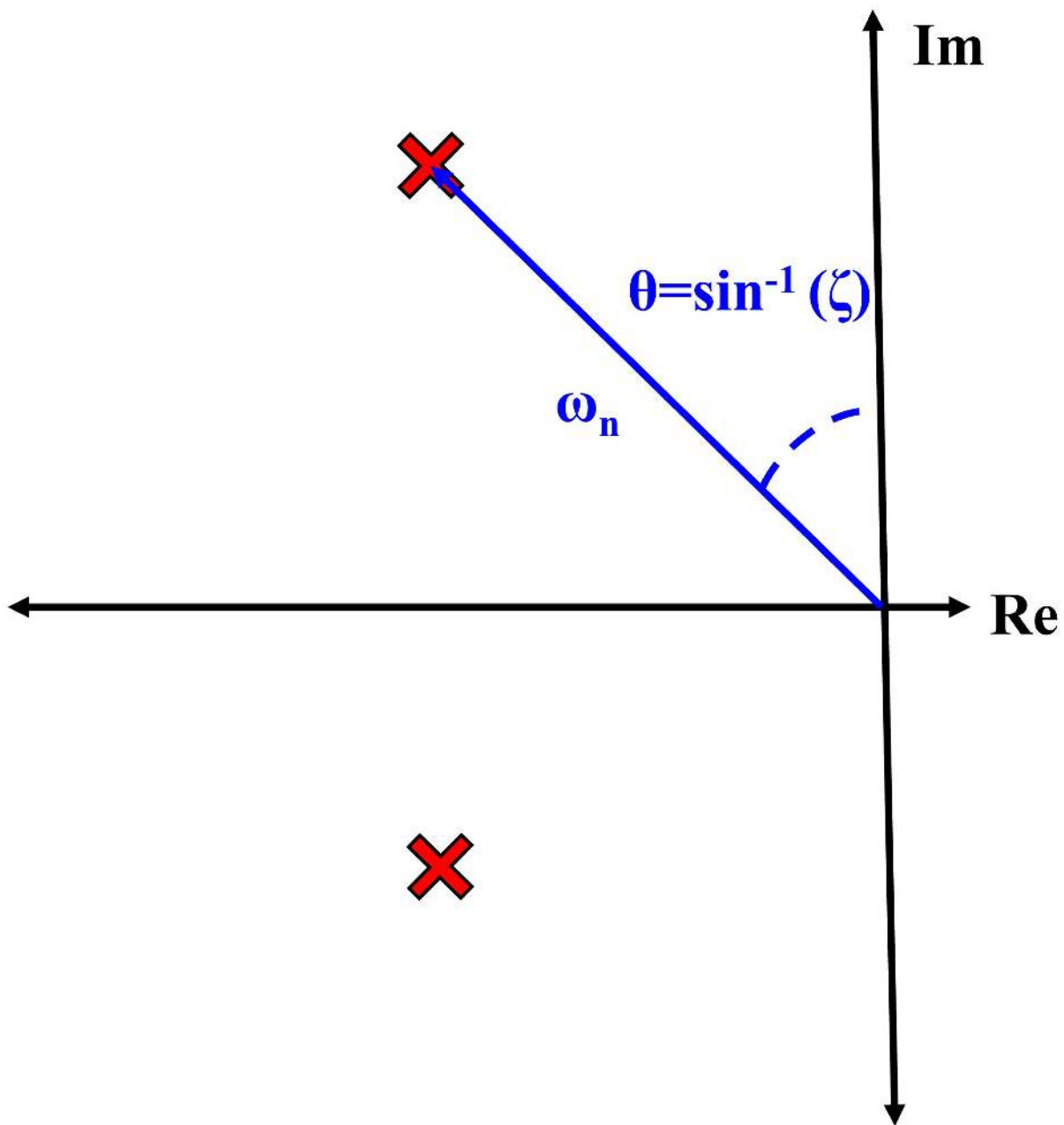


Figure Chapter9.1: Poles in S-plane

Alternatively, using the form:

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta \omega_n s + \omega_n^2}$$

Solving the denominator gives:

$$s = -\zeta \omega_n \pm j \omega_n \sqrt{1 - \zeta^2}$$

Let us define:

- $\sigma = \zeta\omega_n$ (real part)
- $\omega = \omega_n\sqrt{1-\zeta^2}$ (imaginary part = damped natural frequency)

So the poles are located at:

\$\$

$$s = -\sigma \pm j\omega$$

\]

Geometric Interpretation in the s-Plane

- The poles form a complex conjugate pair, symmetric about the real axis
- These poles lie on a semicircle of radius ω_n centered at the origin
- A right triangle can be formed with:
 - Opposite side: $\omega = \omega_n\sqrt{1-\zeta^2}$
 - Adjacent side: $\sigma = \zeta\omega_n$
 - Hypotenuse: ω_n

The angle between the pole vector and the imaginary axis is:

\$\$

$$\theta = \sin^{-1}(\zeta)$$

\]

We also use the identities:

$$\sigma = \zeta\omega_n$$

$$\omega = \omega_n\sqrt{1-\zeta^2}$$

Standard Time-Domain Performance Metrics

We define several critical time-domain performance metrics to control the shape of the system output:

- Rise Time t_r : Time for the output to rise from 10% to 90% of the steady-state value.

Approximate expression:

\$\$

$$t_r \approx \frac{1.8}{\omega_n}$$

\]

- Peak Time t_p : Time at which the output reaches its first maximum:

\$\$

$$t_p = \frac{\pi}{\omega}$$

\]

- Maximum Overshoot M_p : Percent by which output exceeds steady-state:

\$\$

$$M_p = e^{\left(-\frac{\pi\zeta}{\sqrt{1-\zeta^2}}\right)} \times 100\%$$

\]

- Settling Time t_s : Time for output to stay within 1% of the final value:

\$\$

$$t_s \approx \frac{4.6}{\zeta\omega_n}$$

\]

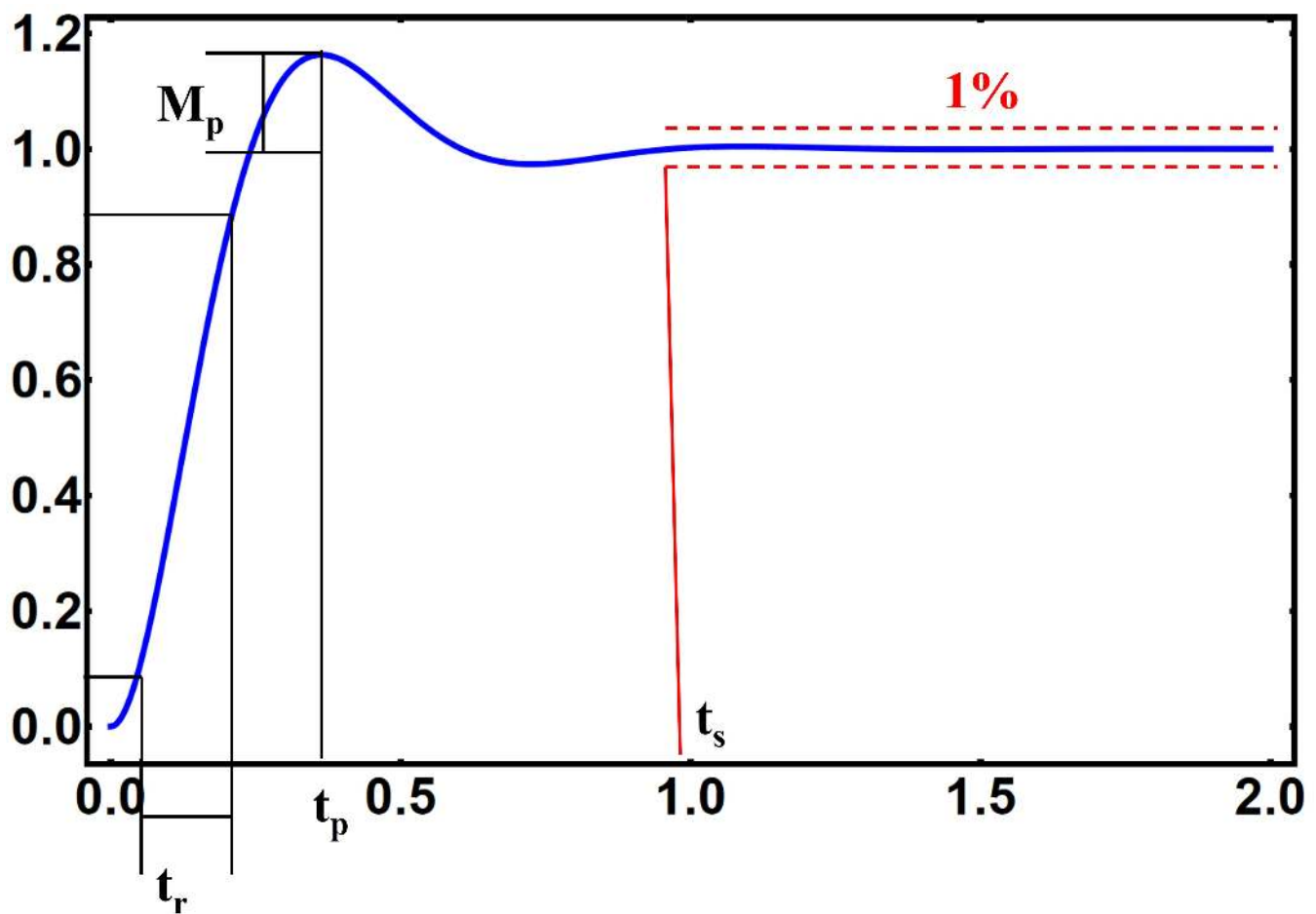


Figure Chapter9.2: Time Domain Specifications

Worked Example: Step Input Response

Let's examine a step input:

$$u(t) = 1$$

We know:

\$\$

$$P(s) = \frac{Y(s)}{U(s)}$$

\]

and

\$\$

$$U(s) = \frac{1}{s}$$

\]

We can compute the output $Y(s)$, then use the inverse Laplace transform to get the time-domain output $y(t)$.

Let's choose:

$$- \omega_n = 10 \text{ rad/s}$$

$$- \zeta = 0.5$$

We find very strong agreement between both the analytical solution and fundamental time-domain metric definitions. This can also be verified using Mathematica or other simulation tools.

It is important to recognize that some discrepancies between numerical results and metric approximations may occur. Many of these equations are empirical or derived from numerical fits. When designing control systems, we must remain aware of these approximations.

Meeting Control Specifications

In practice, specifications originate from user requirements, safety, or physical limits. For example, a temperature control system for a room might specify:

- Overshoot must be less than 10%
- Peak time must be less than 0.5 s
- Settling time must be less than 2 s

So the designer imposes:

\$\$

$$M_p < 10\%, \quad t_p < 0.5 \text{ s}, \quad t_s < 2 \text{ s}$$

The control objective becomes designing a controller that moves the system poles to meet these specifications.

Spring-Mass-Damper Response System

Let's consider a design example using a mass-spring-damper system:

- Mass: $m = 2 \text{ kg}$
- Spring constant: $k = 200 \text{ N/m}$
- Damping coefficient: $c = 8 \text{ Ns/m}$

We want to check if the system meets the following design metrics:

- Rise time $t_r < 0.8 \text{ s}$
- Maximum overshoot $M_p < 15\%$
- Settling time $t_s < 10.1 \text{ s}$

We calculate the actual system values:

\$\$

$$\begin{aligned} t_p &\approx 0.32 \text{ s} \\ M_p &\approx 52\% \\ t_r &\approx 0.18 \text{ s} \\ t_s &\approx 2.3 \text{ s} \end{aligned}$$

Conclusion:

The system fails the overshoot requirement badly. With 52% overshoot, the system might exceed design tolerances, potentially causing damage or instability. This is unacceptable.

So how can we tune the system to reduce overshoot while maintaining the other requirements?

Let's take a more fundamental approach — by analyzing pole location in the s -domain and how it relates to our time-domain metrics.

Relating Time-Domain Specs to the s -Plane

We know from earlier that:

- Settling time relates to the real part of the pole:

\$\$

$$t_s = \frac{4.6}{\sigma}$$

- Rise time relates to the natural frequency:

\$\$

$$t_r = \frac{1.8}{\omega_n}$$

\]

- Overshoot relates to damping ratio ζ , which controls the angle θ between the pole and the negative real axis:

\$\$

$$\theta = \cos^{-1}(\zeta)$$

\]

We can now construct a shaded region in the left-half of the s-plane that satisfies all of our time-domain design criteria.

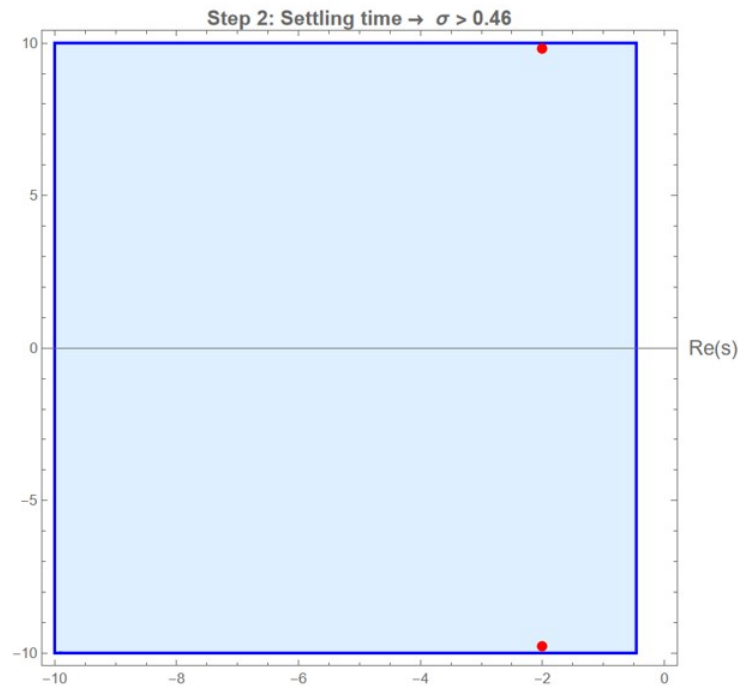


Figure Chapter9.3: Pole Domains - Settling Time Bound

This figure shows the leftmost vertical boundary determined from the settling time bound. Any poles to the left of this line will meet the settling time requirement.

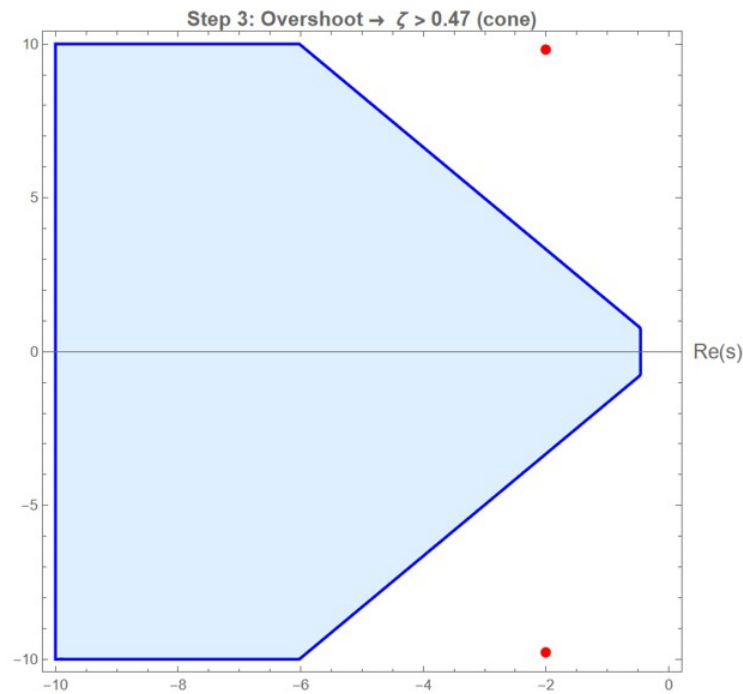


Figure Chapter9.4: Pole Domains - Rise Time Bound

This figure adds a circular arc centered at the origin representing a constant natural frequency ω_n . This corresponds to the rise time requirement.

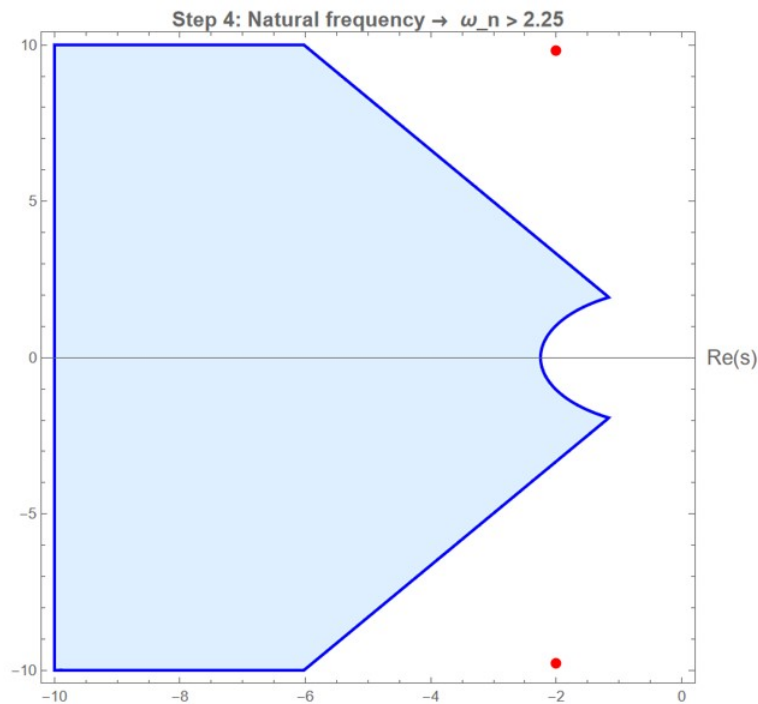


Figure Chapter9.5: Pole Domains - Overshoot Bound

This final figure overlays a sloped line indicating a fixed damping ratio ζ , or equivalently, an angle θ . This corresponds to the overshoot requirement.

Design Check for the Mass-Spring-Damper System

We are working with:

- $m = 2$
- $c = 8$
- $k = 200$

We compute:

\$\$

$$\begin{aligned}\omega_n &= \sqrt{\frac{k}{m}} = \sqrt{100} = 10 \\\zeta &= \frac{c}{2\sqrt{km}} = \frac{8}{2 \cdot 20} = 0.2\end{aligned}$$

This damping ratio is too low — it violates the overshoot constraint.

Thus, the poles fall outside the shaded "safe" region. Overshoot will be too high.

Now, let's fix that by adjusting damping.

Adjusting Damping Coefficient c to Meet Pole Placement Specifications

Consider the standard second-order transfer function for a spring-mass-damper system:

\$\$

$$H(s) = \frac{1}{m s^2 + c s + k}$$

This system has the form:

\$\$

$$\begin{aligned}H(s) &= \frac{1}{m \left(s^2 + \frac{c}{m} s + \frac{k}{m} \right)} \\&= \frac{1}{m \left(s^2 + 2\zeta \omega_n s + \omega_n^2 \right)}\end{aligned}$$

where:

\$\$

$$\begin{aligned}\omega_n &= \sqrt{\frac{k}{m}}, \quad \text{quad} \\ \zeta &= \frac{c}{2\sqrt{km}}\end{aligned}$$

Suppose we are given design specifications that include:

- Maximum overshoot less than 15% $\Rightarrow \zeta > 0.516$
- Rise time $t_r < 0.8 \text{ s}$
- Settling time $t_s < 10.1 \text{ s}$

Given our original parameters:

\$\$

$$m = 2, \quad \text{quad} \quad c = 8, \quad \text{quad} \quad k = 200$$

We find:

\$\$

$$\begin{aligned}\omega_n &= \sqrt{\frac{200}{2}} = \sqrt{100} = 10, \quad \text{quad} \\ \zeta &= \frac{8}{2 \cdot \sqrt{200 \cdot 2}} = \frac{8}{2 \cdot 20} = 0.2\end{aligned}$$

This damping ratio is insufficient ($0.2 < 0.47$) and results in excessive overshoot.

To meet the overshoot spec, we must increase c so that $\zeta > 0.47$.

Solving for the new required damping coefficient:

$$\begin{aligned} \zeta_{\text{min}} &= \frac{c_{\text{new}}}{2\sqrt{km}} \\ \Rightarrow c_{\text{new}} &= 2\zeta_{\text{min}}\sqrt{km} \end{aligned}$$

Substitute in values:

$$\begin{aligned} c_{\text{new}} &= 2 \cdot 0.47 \cdot \sqrt{200 \cdot 2} \\ &= 2 \cdot 0.47 \cdot 20 = 18.8 \end{aligned}$$

Therefore, to move the poles into the desired region, we must increase the damping coefficient from 8 to:

$$\boxed{c = 18.8}$$

This ensures $\zeta = 0.47$, satisfying the overshoot specification, and places the poles within the allowable region of the s -plane.

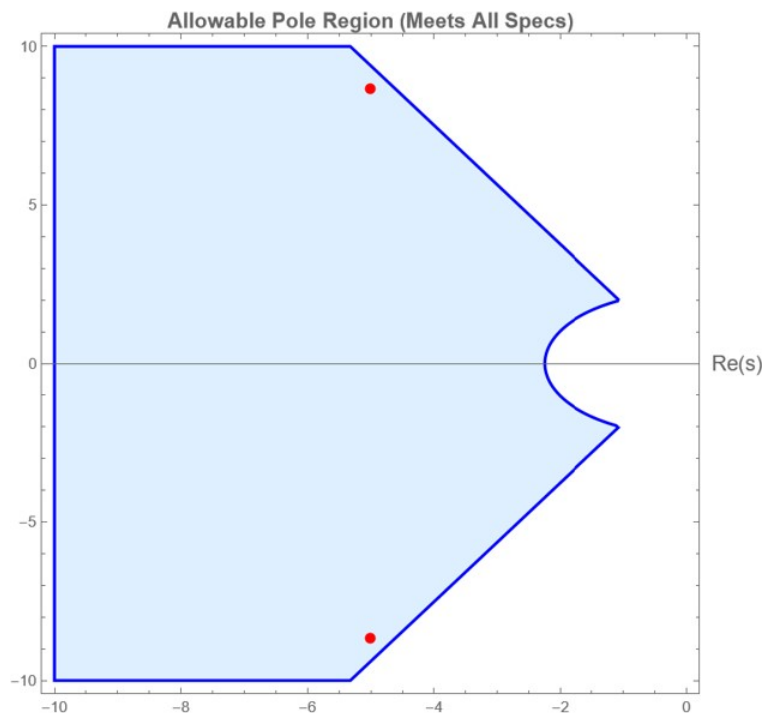


Figure Chapter9.6: Pole Domains

This should make sense conceptually because we increased the damping coefficient to damp out the overshoot so our conceptual understanding matches our numerical understanding!!

This page titled [Chapter 9: Time Domain Specifications](#) is shared under a [CC BY-NC-ND](#) license and was authored, remixed, and/or curated by [Joshua P. Steimel](#).

Index

D

dire

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Loading...