ECE 306: Signals and Systems

Bryan Van Scoy

November 21, 2024

Contents

Part I

Introduction to Signals and Systems

1

Introduction

The goal of this course is to give students a general understanding of the fundamental principles of signals and systems. To start out, let's define what we mean by signals and systems.

A *signal* is. . .

- a detectable physical quantity, such as voltage or current, by which information may be transmitted
- a set of information or data

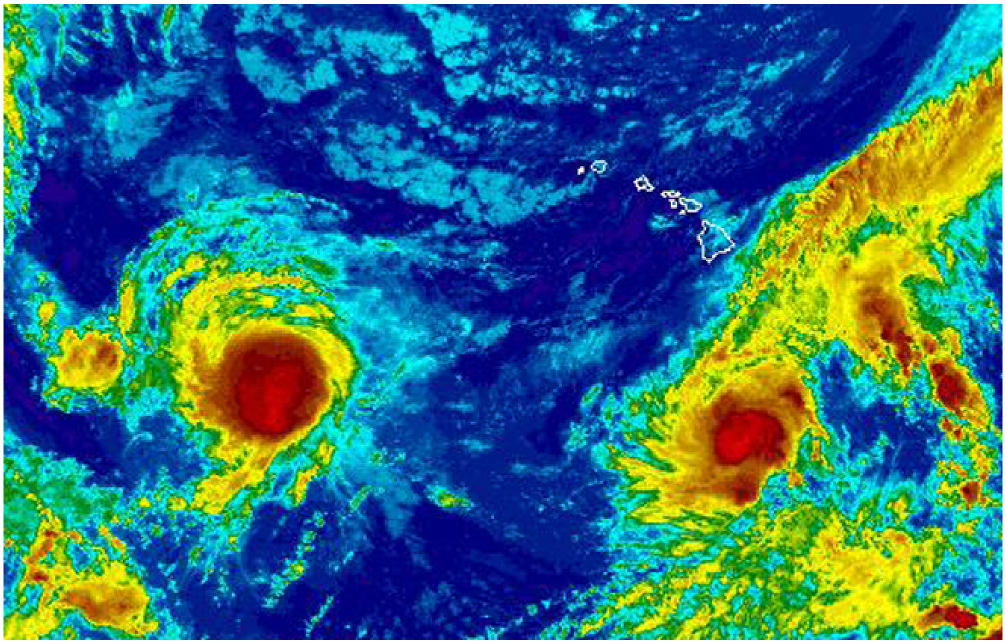
A *system* is. . .

- any interacting or interdependent group of items forming a unified whole
- any collection of objects for which some set of cause-and-effect relations exists
- an entity that processes a set of signals (inputs, excitations) to yield another set of signals (outputs, response)

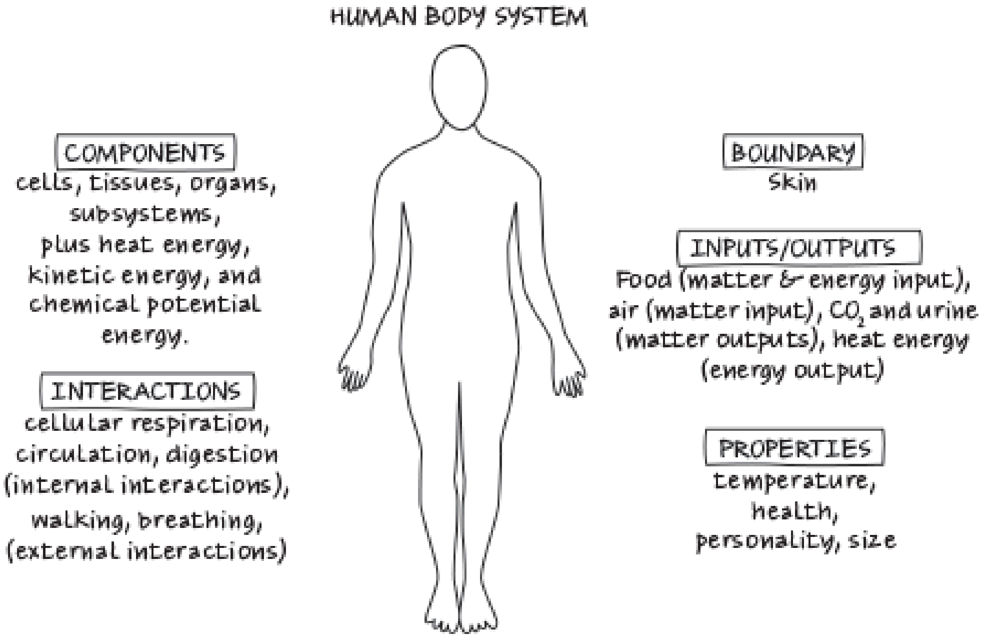
1.1 Examples

We now provide some examples of signals and systems.

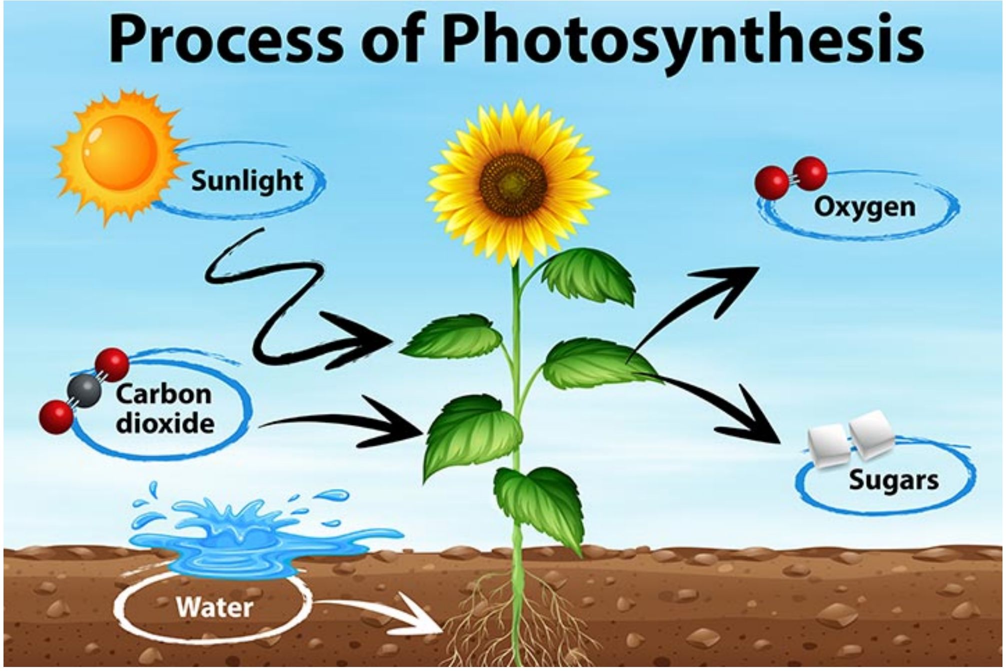
- **Weather.** The weather is a complex system that relates signals such as temperature, pressure, density, and air velocity (wind) at each point in time. Meteorologists use models to predict weather based on current and past conditions, such as tracking the location of a hurricane. These models may be based on physics or constructed from historical data.



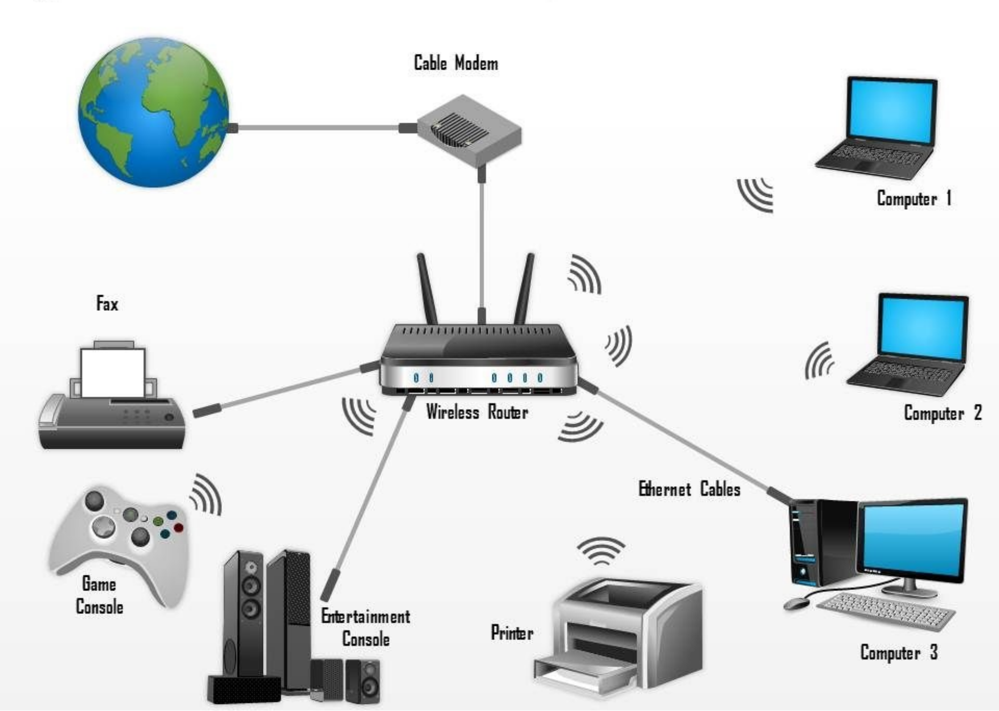
- **Human body.** The human body is a complex system that relates signals such as height, weight, temperature, blood flow, and electrical signals between neurons. Doctors use medications to influence (control) these signals in order to improve bodily function, such as fighting off a disease.


HUMAN BODY SYSTEM
COMPONENTS
cells, tissues, organs, subsystems, plus heat energy, kinetic energy, and chemical potential energy.
INTERACTIONS
cellular respiration, circulation, digestion (internal interactions), walking, breathing, (external interactions)
BOUNDARY
Skin
INPUTS/OUTPUTS
Food (matter & energy input), air (matter input), CO_2 and urine (matter outputs), heat energy (energy output)
PROPERTIES
temperature, health, personality, size

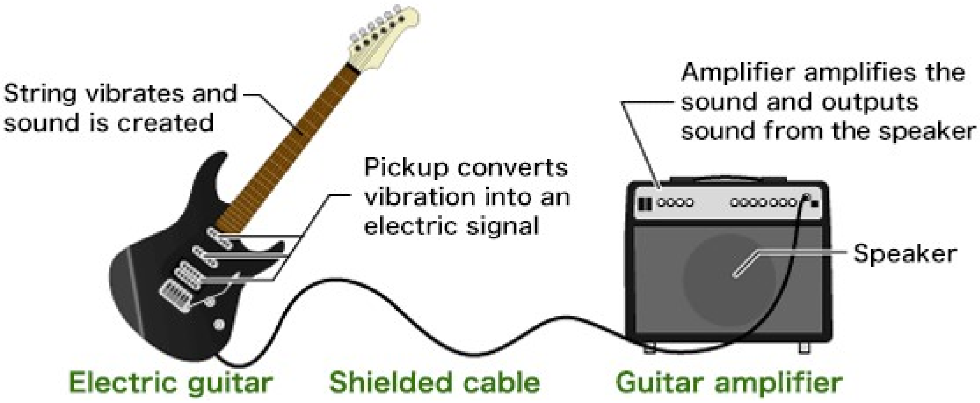
- **Chemical process.** The transformation of interacting chemicals forms a system, where the signals are the quantities and concentrations of the chemicals. For example, plants use photosynthesis to convert sunlight, carbon dioxide, and water into sugars for food while releasing oxygen as a byproduct.



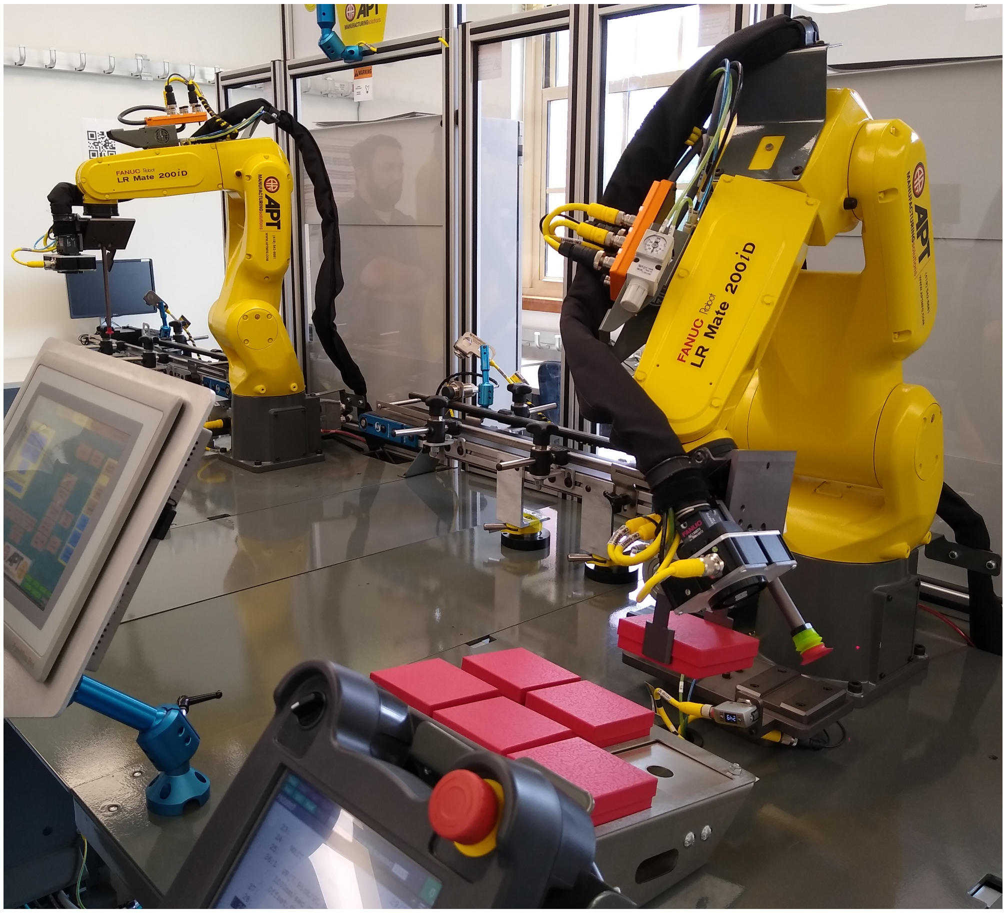
- **Internet.** The internet is a complex interconnection of routers, modems, servers, and computers that enables signals to be transmitted between devices, either electrically through wires or wirelessly via electromagnetic waves. This can be viewed as a single (very large!) system, or as the connection of many small subsystems.



- **Audio system.** An audio system may be used to produce sound waves from a musical instrument or microphone. In an electric guitar, for example, the sound is created by the musician plucking a string to make it vibrate, which is then converted to an electric signal by the pickup, which is then amplified and converted to a sound wave by the speaker.


String vibrates and sound is created
Pickup converts vibration into an electric signal
Amplifier amplifies the sound and outputs sound from the speaker
Speaker
Electric guitar
Shielded cable
Guitar amplifier

- **Robot.** A robot is an electromechanical system, meaning that it is composed of both electrical and mechanical components. Signals include the force, position, and velocity of each mechanical component, as well as voltages and currents in the circuits. The robot below is used in the course ECE 317: Industrial Robotics at Miami University.

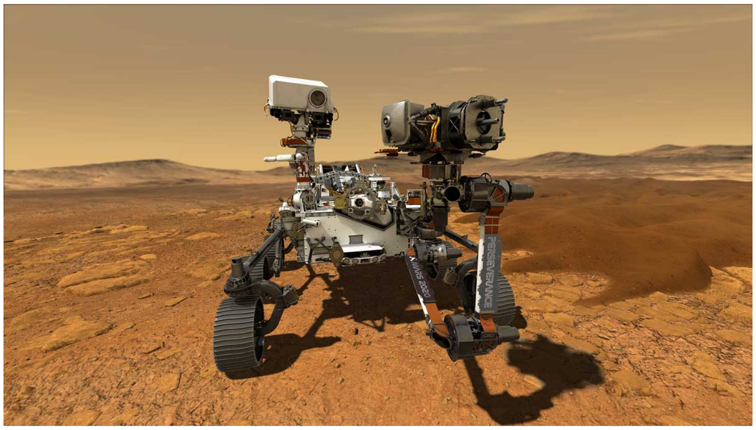


This is a very incomplete list of some signals and systems; the main point is that signals and systems are everywhere! As you go about your day, take a moment to observe some of the signals and systems that you interact with.

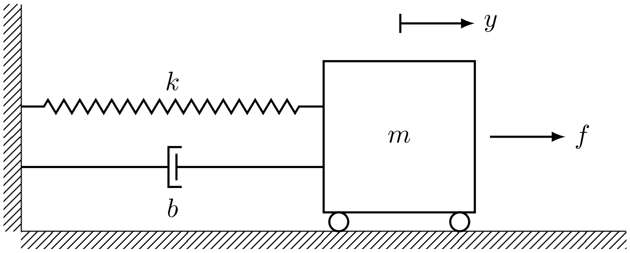
1.2 Modeling

Now that we have seen some examples of signals and systems, let's discuss how we can describe them. We will start from the most natural and realistic description, and then abstract parts of the system to move towards a mathematical model that can be used to analyze the system.

- **Physical.** The most natural and realistic description of a system is its physical description. A description of the Mars rover, for instance, would include its dimensions, materials, and components.

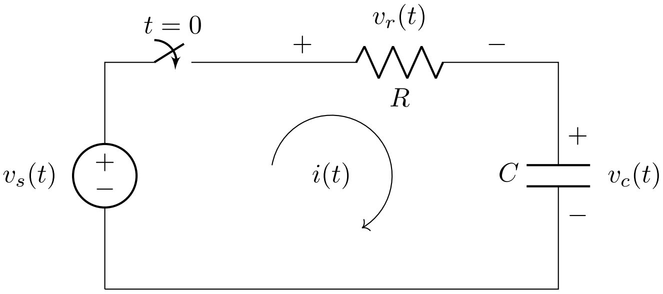


- **Diagrams.** Another way to describe a system is by schematic diagrams. Continuing with our example, we can describe the mechanical components of the Mars rover using a mechanical diagram:



Here, m is the mass of the robot, k is its spring constant, b is its damping coefficient, y is its position, and f is the force applied to it, such as through an electric motor. Note that the mass, spring constant, and damping coefficient are *parameters* that typically do not change with time, while force and position are *signals* that do depend on time.

Likewise, the electrical components of the system can be described using a circuit diagram:



Here, v_s is the source voltage applied to the circuit, R is the resistance, C is the capacitance, i is the current, v_r is the voltage across the resistor, and v_c is the voltage across the capacitor. In this case, the resistance and capacitance are parameters of the circuit, while the voltages and current are signals.

- **Equations.** From the schematic diagrams, we can write down differential equations that describe the relationship between the signals as a function of time. For instance, Newton's laws (force equals mass times acceleration) say that the force $f(t)$ and position $x(t)$ of the mass in our mechanical diagram are related by the second-order differential equation

  $$m\,\ddot{y}(t) + c\,\dot{y}(t) + k\,y(t) = f(t)$$

  Similarly, Kirchhoff's laws allow us to relate the source voltage $v_s(t)$ and the capacitor voltage $v_c(t)$ in the circuit diagram by the first-order differential equation

  $$RC\,\dot{v}_c(t) + v_c(t) = v_s(t)$$

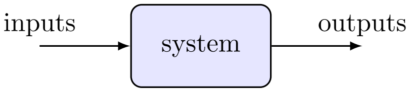
  The dot notation means a derivative with respect to time, such as

  $$\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t) \qquad \text{and} \qquad \ddot{x}(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}x(t)$$

  The derivation of such equations that describe a system is called *modeling* and requires knowledge of the particular application domain.

1.3 Overview

As we have seen in each of the examples, a system describes the relationship between the various signals in a system. We typically think of the signals that we have control over as *inputs* and those that we can measure as *outputs*. The general depiction is as follows:



In this class, we. . .

- abstract signals and systems to study their fundamental properties
- focus on systems that are well-defined mathematically

In other words, we assume that a mathematical model of the system has already been constructed, and we study this model of the system. Keep in mind, however, that models do not always describe the system perfectly, and there is a trade-off between the complexity of the model and the difficulty in analyzing it!

For a given system, there are several types of problems that we may want solve:

- **Analysis.** The analysis problem is to quantify how the system reponds to various inputs. We may want to characterize qualitative properties of the system (such as whether or not it is stable) or quantitative properties (such as how much the system amplifies a sinusoidal input at a certain frequency).
- **Design.** The design problem is to construct a system with some desired behavior. In robotics, for instance, engineers may design the electrical and mechanical components so that the robot is able to walk or run (as in a humanoid robot).
- **Control.** In some cases, we do not have the ability to design the system (such as the weather or human body examples). Such systems can still be manipulated by choosing the inputs so that the system has some desired behavior. The main approach in control is to choose the input signals as a function of the outputs, which is called feedback.

- **Identification.** For some systems, such as the weather example, it is too complicated to write down governing equations from first principles. In such cases, an alternative approach is to collect input/output data and use that to construct a mathematical model that approximately describes the behavior of the system. This is called system identification.

In this class, you will learn how to

- classify signals and systems based on their properties
- convert between various representations of signals and systems
- compute and interpret the response of linear time-invariant systems
- interpret signals and systems in the frequency domain
- model a linear time-invariant system from input/output data

Discussion: What are some other examples of systems?

- What are the input signals?
- What are the output signals?
- Why do you want to understand the behavior of the system?

1.4 Fundamental concepts

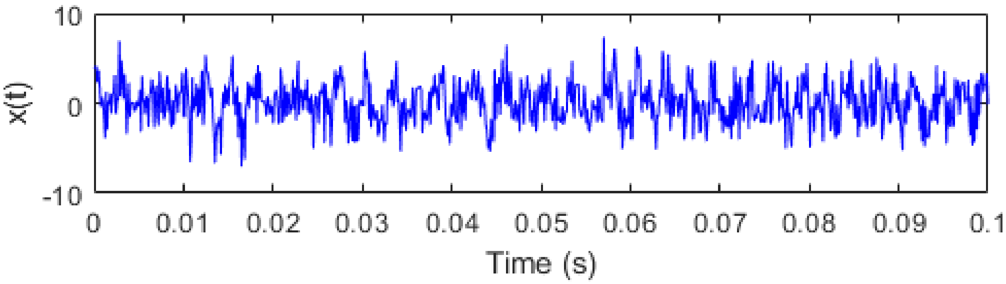
There are several fundamental concepts in signals and systems that we will be studying throughout the course.

Signals can be viewed as functions of *time* or *frequency*

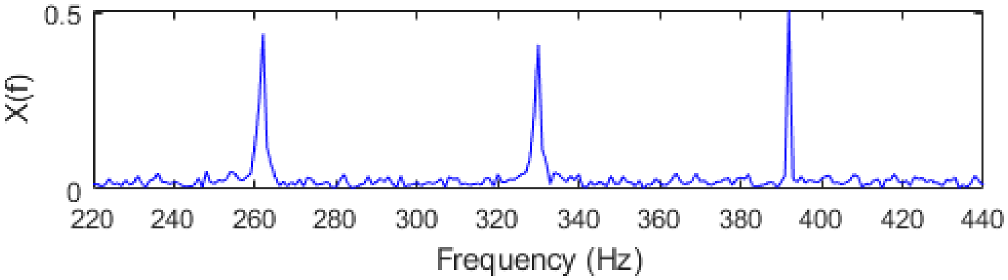
While we typically think of signals as functions of time (such as the voltage in a circuit), it is often useful to interpret signals in terms of their *frequencies*.

As an illustrative example, consider a musical signal such as an audio clip of a song. The signal consists of sound waves that apply pressure to your eardrums at each point in time. A *tone* is an audio signal that consists of a single frequency. For instance, the tone A_4 is a single sinusoidal signal at a frequency of 440 Hz (the tone often used to calibrate instruments). Songs consist of multiple tones happening at once to create music.

Consider a C major chord, which consists of the frequencies: 261.63 Hz (C_4), 329.23 (E_4), and 392.00 (G_4). As a function of time, this signal looks as follows:



From this time-domain representation of the signal, it is difficult to see that the signal is actually just three tones. If we instead represent this signal as a function of its frequencies (its frequency-domain representation), we get three “spikes” at the individual tones. There are small amounts of the other frequencies due to noise in the signal, but the amount of noise is small compared to the size of the signal.



Throughout the course, we will learn to interpret both signals and systems in the time and frequency domains. Some reasons for this are the following:

- We may want to know what frequencies are in a signal. For instance, we are only able to hear frequencies between 20 Hz to 20,000 Hz, so any frequency content outside of this range is inperceptible (so we may want to remove it to save space when storing the signal on a computer).
- Many systems are much easier to understand in the frequency domain. In particular, systems that are *linear* and *time invariant* will play a large role in this course, and we will see that they are best understood in the frequency domain.

Applications often involve continuous- and discrete-time signals and systems

Continuous-time signals are defined at all points in time (like the voltage in a circuit or the force on an object), while discrete-time signals are defined only at discrete points in time. Many applications involve both continuous and discrete time due to the fact that signals are often *continuous* (voltage, force, sound, image), while computers process and store *discrete* information.

As an example, consider an audio signal that is processed (or stored) by a computer and then played back through speakers. The audio signal begins as a continuous-time signal of sound waves. A microphone converts the pressure waves to an electric signal (still in continuous time). The computer then uses an analog-to-digital converter to convert the electric signal to a series of bits that it can process and store. To play the signal back, those bits are passed through a digital-to-analog converter to produce an electric signal that is then put through the speaker which converts it to an audio signal.

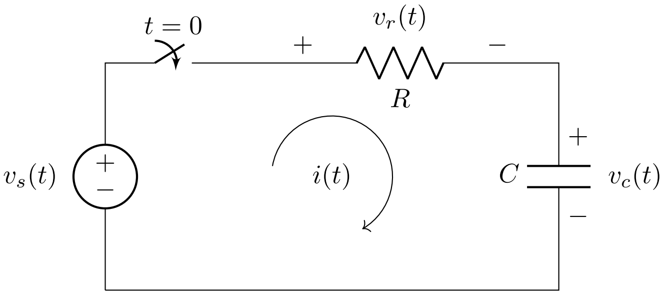
audio → microphone → computer → speaker → sound

This basic idea of converting a continuous-time signal to discrete time, processing it using a computer, and then converting back to continuous time occurs in numerous applications. For instance, robots involve continuous-time signals such as forces, positions, velocities, voltages, and currents, while they are programmed using microcontrollers that operate on discrete-time signals.

We will only scratch the surface of the relationship between continuous and discrete time in this course. For those interested in learning more, this is one of the main topics in digital signal processing (DSP).

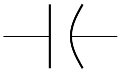
1.5 Example: RC circuit

To illustrate several more concepts that we will see throughout the course, consider a simple RC circuit.



- the input signal $v_s(t)$ is the supplied voltage
- the output signal $v_c(t)$ is the capacitor voltage
- the initial condition $v_c(0) = a$ is the capacitor voltage at time $t = 0$
- the time constant is RC

A circuit is a system that relates voltages and currents. The following table summarizes the relationship between voltage and current across a resistor, capacitor, and inductor.

component	**voltage-current**	**impedence**
capacitor 	$v(t) = \dfrac{1}{C} \displaystyle\int_{-\infty}^{t} i(\tau)\,\mathrm{d}\tau$	$\dfrac{1}{Cs}$
resistor 	$v(t) = R\,i(t)$	R
inductor 	$v(t) = L\,\dfrac{\mathrm{d}i(t)}{\mathrm{d}t}$	Ls

We can analyze the circuit using these relationships along with Kirchhoff's voltage and current laws.

- **Kirchhoff's voltage law (KVL):** the sum of voltages around any closed loop is zero
- **Kirchhoff's current law (KCL):** the sum of currents entering and exiting a node must be equal

From KVL, the supplied voltage is equal to the sum of the voltage across the resistor and capacitor:

$$v_s(t) = v_R(t) + v_c(t)$$

From KCL, the same current flows through the resistor and capacitor. Using the relationship between the voltage and current across each component, we can write this in terms of voltages as

$$i(t) = \frac{1}{R}\,v_R(t) = C\,\dot{v}_c(t)$$

Combining these equations, we find that the system is described by the differential equation

$$v_c(t) + RC\,\dot{v}_c(t) = v_s(t)$$

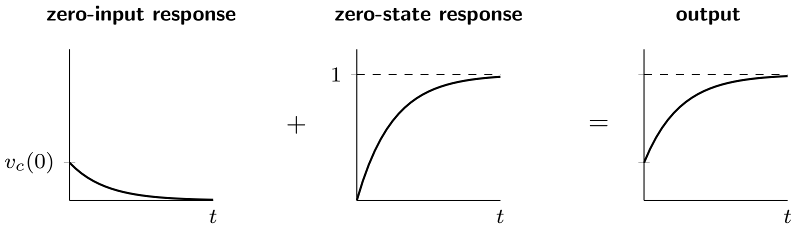
This is a differential equation because it involves both signals (such as $v_s(t)$ and $v_c(t)$) and their derivatives ($\dot{v}_c(t)$). Given the source signal $v_s(t)$, the solution of the differential equation is the entire signal $v_c(t)$ as a function of time t.

From this representation, we can make several observations:

- Continuous-time systems can be modeled using *differential equations*. Discrete-time systems have a similar representation as *difference equations*. This is only one way to represent systems, and we will see multiple other ways of describing a system.
- Dynamical systems have *memory*. The solution depends on the initial condition $v_c(0)$, which is the capacitor voltage before the source voltage is applied. The response (or output) of a system depends on both the input to the system and what state it was in before the input was applied.
- Consider what happens to the circuit if the source voltage is zero. If the capacitor has an initial charge, then the charge will dissipate through the resistor when the circuit is connected until it has no charge left. This is called the *zero-input* response of the system.

  Now consider what happens if the capacitor has no initial charge but we apply a constant source voltage. This constant voltage will charge the capacitor until it is fully charged. This is called the *zero-state* response of the system because the system was initially "at rest".

  When there is both an initial state (the charge on the capacitor) and an input signal (the source voltage), the system response is the superposition of the zero-input and zero-state responses.



1.6 Course mechanics

- **pre-requisites**
  - circuits
- **co-requisites**
  - differential equations
- **post-requisites**
  - signal processing
  - communication systems
  - robotics
  - control

2

Signals

A *signal* is a function of time. Some examples of signals are the force on a mass, the voltage in a circuit, and the acoustic pressure at a point in space.

Notation. We use lowercase letters such as u, x, and y to denote signals. The symbol u refers to the *entire* signal, while $u(t)$ refers to the value of the signal at time t. ■

While it may not seem obvious at first, a signal is actually a vector. We typically think of a vector as a finite set of numbers, such as

$$v = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

while a signal is a function of time, such as

$$u(t) = \cos(t)$$

But everything we can do with vectors, we can also do with signals. For instance, we will see how to scale a signal by a scalar, add two signals, take the inner product between two signals, determine whether or not two signals are orthogonal, and compute the size (or length) of a signal. While linear algebra is the study of *finite-dimensional* vectors, signals have *infinite* dimensions.

2.1 Properties

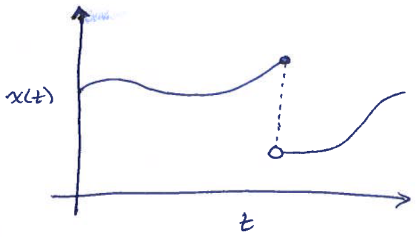
We will first study how to categorize signals based on their properties.

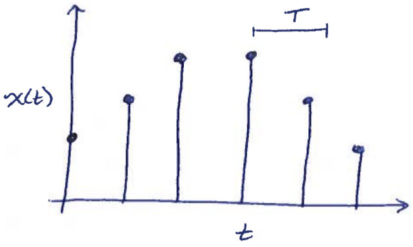
Domain

The *domain* of a signal is the set of times for which it is defined. There are two common domains.

- *Continuous-time signals* are defined on an entire interval of time, such as the set of all real numbers, the set of nonnegative real numbers, or the interval $[a, b]$. For continuous-time signals, we often use the symbols t or τ for time to indicate that it is a real number.
- *Discrete-time signals* are only defined on a discrete set of times, such as the set of integers, the set of nonnegative integers, or the finite set $\{0, 1, 2, \ldots, n\}$. For discrete-time signals, we often use the symbols k or m for time to indicate that it is an integer.

Notation. We often denote continuous-time variables by t or τ, while we use k or n to denote discrete-time variables to emphasize that time is an integer. For instance, a discrete-time signal may be denoted by $x(k)$, $x[k]$, or x_k, while a continuous-time signal is typically denoted by $x(t)$. ■





Example. Since time itself is continuous, all physical signals are in continuous time. Some examples of continuous-time signals are voltages and currents in a circuit, forces and positions in a mechanical system, and radio transmissions. Discrete-time signals, however, are often used to represent quantities at a discrete set of times, such as the monthly balance of a bank account, the iterates of a numerical algorithm, a sampled continuous-time signal, or a sequence of numbers stored on a computer (such as an image or audio signal).

The values of a discrete-time signal are often evenly spaced in time. We can think of a discrete-time signal as a sampled version of a continuous-time signal, where the relationship between continuous time t and discrete time k is

$$t = kT$$

where T is the sampling period. We often write $x(k)$ instead of $x(kT)$, where the sampling period T is implicitly assumed.

Codomain

The *codomain* of a signal u is the set of values that it can have at any given time.

- *Scalar signal:* $u(t)$ is a scalar (number)
- *Vector signal:* $u(t)$ is a vector
- *Binary signal:* $u(t)$ is either 0 or 1

Example. The continuous-time vector signal

$$v(t) = \begin{bmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \end{bmatrix}$$

might give the voltage at three places on an antenna.

Example. The discrete-time binary signal $u(k) = \{1, 1, 0, 0, 1, 0, 1, \ldots\}$ may give the value (high or low) of a transistor in a circuit at each clock cycle.

The values of a signal often have *units* associated with them, such as volts, amps, meters, etc. A signal may also be unitless, or the units may be unspecified.

Size

The size of a signal is a single number that describes how "big" the signal is. Size is a quantitative property of the signal. If we think of a signal as a vector, then the size of a signal is the same thing as the length of the vector (in general, this is called a *norm*).

There are many ways to measure the size of a signal. For example, consider a continuous-time scalar signal u with domain $t \geq 0$.

- The *energy* of the signal is the integral of its amplitude squared,

  $$\int_0^\infty u(t)^2 \, \mathrm{d}t$$

  An *energy signal* is a signal with finite energy. For the integral to converge to a finite number, the values of the signal $u(t)$ must converge to zero as $t \to \infty$.

- The *power* of the signal is the time average of the amplitude squared,

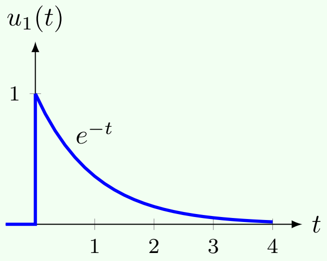
  $$\lim_{T \to \infty} \frac{1}{T} \int_0^T u(t)^2 \, \mathrm{d}t$$

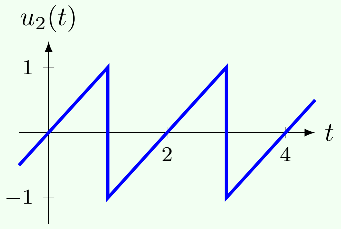
  A *power signal* is a signal with finite power. The *root-mean-square* (RMS) value is the square root of the power.

- The *peak* of the signal is its maximum absolute value,

  $$\max_{t \geq 0} \; |u(t)|$$

Example. To illustrate the various ways of measuring the size of a signal, consider the following two continuous-time scalar signals defined on the time interval $[0, \infty)$.





- **Energy.** The energy of the signal $u_1(t)$ is

$$\text{energy}(u_1) = \int_0^\infty \left(e^{-t}\right)^2 \, \mathrm{d}t = \int_0^\infty e^{-2t} \, \mathrm{d}t = -\frac{1}{2} e^{-2t} \bigg|_{t=0}^{\infty} = -\frac{1}{2}(0-1) = \frac{1}{2}$$

Since the energy is finite, this is an energy signal. The signal $u_2(t)$ does not converge to zero, so it has infinite energy,

$$\text{energy}(u_2) = \int_0^\infty u_2(t)^2 = \infty$$

- **Power.** Since the signal u_1 converges to zero, its power is zero,

$$\text{power}(u_1) = 0$$

Since the signal $u_2(t)$ repeats, we can calculate the power by integrating over a single period as

$$\text{power}(u_2) = \lim_{T\to\infty} \frac{1}{T} \int_0^T u_2(t)^2 \, \mathrm{d}t = \frac{1}{T} \int_T u_2(t)^2 \, \mathrm{d}t = \frac{1}{2} \int_{-1}^{1} t^2 \, \mathrm{d}t = \frac{1}{6} t^3 \bigg|_{t=-1}^{1} = \frac{1}{6}(1-(-1)) = \frac{1}{3}$$

The power is finite, so this is a power signal.

- **Peak.** The peak of both signals is one.

$$\text{peak}(u_1) = 1 = \text{peak}(u_2)$$

The energy, pwer, and peak are various ways of measuring the size of a signal. For these two examples, the signals have the same peak values while u_1 has finite energy and u_2 has finite power. These measures are all quite different!

We can also use the size of a signal to measure how similar two signals are to each other. Given two signals u_1 and u_2, we can measure their similarity by taking the size of their difference. For example, some measures of similarity are $\text{energy}(u_1 - u_2)$, $\text{power}(u_1 - u_2)$, and $\text{peak}(u_1 - u_2)$.

Remark. The terms *energy* and *power* do not indicate these quantities in the conventional sense, but instead refer to a generalization to arbitrary signals. For a voltage v across a 1-ohm resistor, the above integral is equal to the physical energy dissipated in the resistor. ■

Qualitative properties

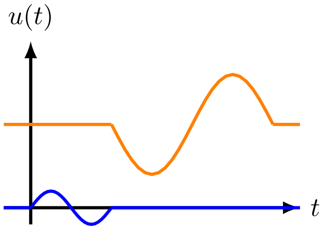
There are many qualitative properties that a signal may or may not have.

- u *converges* if $u(t)$ converges to a constant as $t \to \infty$
- u *decays* if $u(t)$ converges to zero as $t \to \infty$
- u is *bounded* if its peak is finite, meaning that there exists some $M > 0$ such that $|u(t)| < M$ for all t
- u *blows up* if its magnitude diverges to infinity as $t \to \infty$
- u is *periodic* if there exists $T > 0$ such that $u(t+T) = u(t)$ for all t, and the smallest such T is the *period*

2.2 Transformations

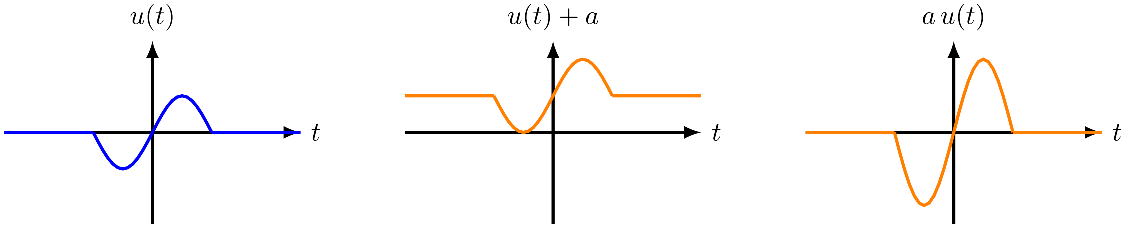
Given a signal, we can perform various transformations to obtain other related signals. We now describe the basic transformations of shifting and scaling of both the magnitude of the signal and its time axis. When plotting a signal, time is plotted on the horizontal axis while the magnitude of the signal is plotted on the vertical axis. Therefore, transforming the magnitude shifts and stretches the signal vertically, while transforming the time variable shifts and stretches it horizontally.

The following figure illustrates these transformations. The original signal (blue) is both shifted and stretched in time, and the magnitude of the signal is both scaled and shifted to produce the transformed signal (orange).



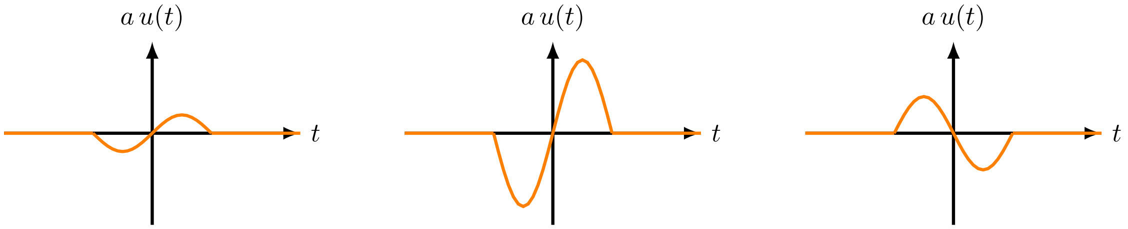
Shifting and scaling the magnitude

Scaling the magnitude of a signal stretches or compresses the graph vertically, while shifting the magnitude shifts the graph of the signal up or down. These transformations are illustrated in the following figure.



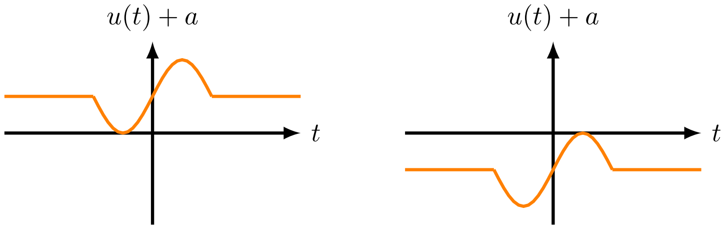
Original signal Shifted magnitude Scaled magnitude

Scaling the magnitude. Scaling the magnitude of a signal $u(t)$ by an amount a produces the signal $a\,u(t)$. If a is larger than one, then this stretches the graph of the signal vertically, while it compresses the graph if $0 < a < 1$. When a is negative, this also flips the graph vertically about the horizontal axis.



$0 < a < 1$ $a > 1$ $a = -1$

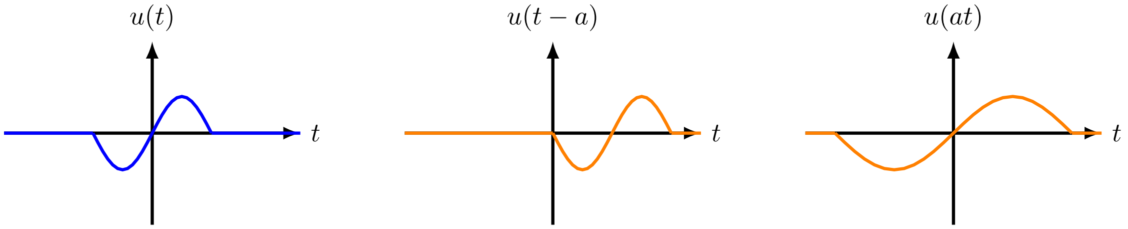
Shifting the magnitude. Shifting the magnitude of a signal $u(t)$ by an amount a produces the signal $u(t) + a$. This shifts the graph of the signal up if a is positive and down if a is negative.



$a > 0$ $a < 0$

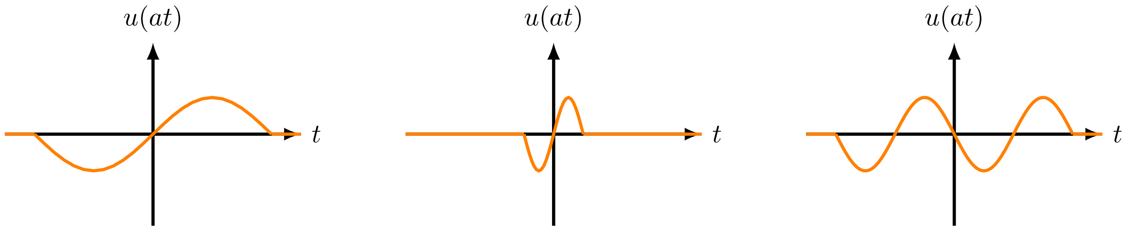
Shifting and scaling time

Scaling the independent time variable of a signal stretches or compresses the graph horizontally, while shifting time moves the graph of the signal left or right. These transformations are illustrated in the following figure.



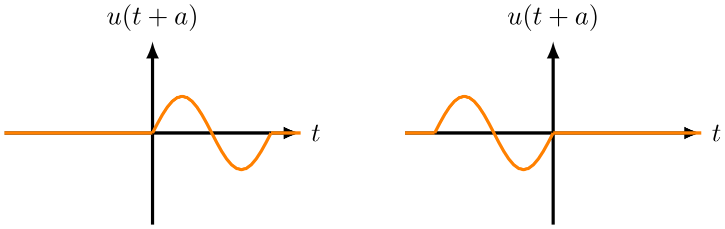
Original signal Shifted time Scaled time

Scaling time. Scaling the time of a signal $u(t)$ by an amount a produces the signal $u(at)$. If a is larger than one, then this stretches the graph of the signal horizontally, while it compresses the graph if $0 < a < 1$. When a is negative, this also flips the graph horizontally about the vertical axis.



$0 < a < 1$ (expand time) $a > 1$ (compress time) $a = -1$ (reverse time)

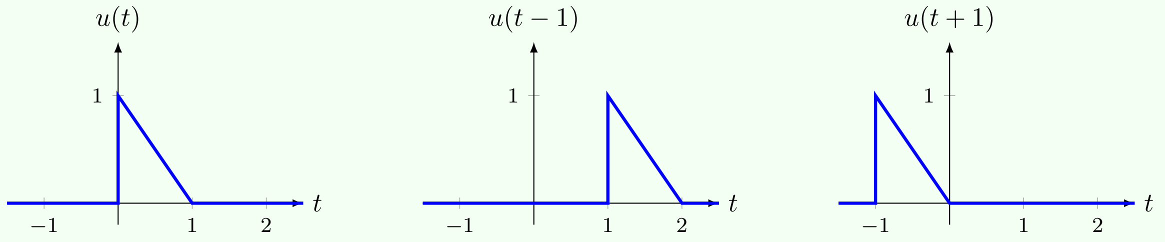
Shifting time. Shifting the time of a signal $u(t)$ by an amount a produces the signal $u(t+a)$. This shifts the graph of the signal left if a is positive (a time *delay*) and right if a is negative (a time *advance*).



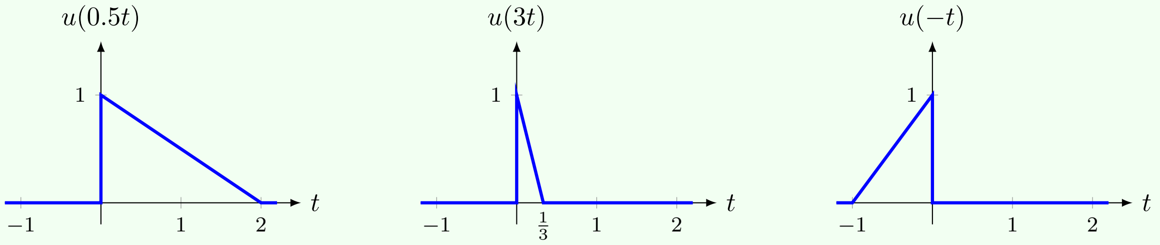
$a < 0$ (delay time) $a > 0$ (advance time)

Remark. Time shifting and scaling may be the opposite of what you intuitively think. Subtracting from time shifts to the right, while adding to time shifts to the left. Scaling time by a number larger than one compresses time, while scaling by a number smaller than one expands time. If you ever forget or get confused, just substitute in a couple values for t to figure out the correct direction to shift and scale. ∎

Example (Time shifting and scaling)**.** Consider the continuous-time signal $u(t) = 1 - t$ for $0 \le t \le 1$ and zero otherwise. Then $u(t-1)$ is the signal shifted to the right by one. While the signal $u(t)$ starts at time $t = 0$, the delayed signal $u(t-1)$ starts at time $t = 1$. Similarly, $u(t+1)$ is the signal shifted to the left by one, which advances the signal to start at time $t = -1$. These signals are illustrated below.



Also, $u(0.5t)$ is the signal expanded by a factor of two, $u(3t)$ is the signal compressed by a factor of three, and $u(-t)$ is the signal flipped about the vertical axis. These signals are illustrated below.



Combined shifting and scaling

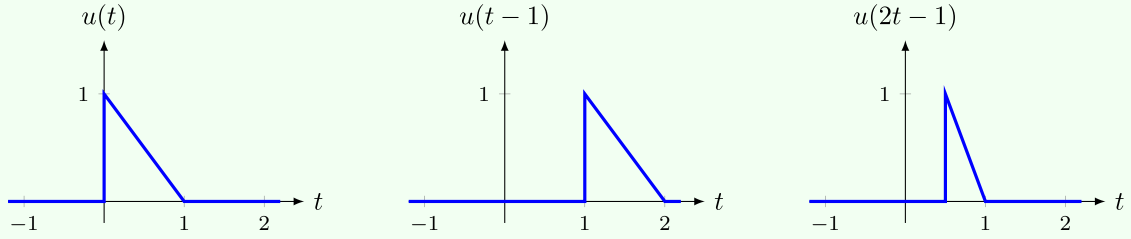
When we both shift and scale a signal (in either time or magnitude), we must be careful about the order in which the transformations are applied, since the order affects the resulting signal. There are two cases.

- **Shift then scale.** Shifting the signal $u(t)$ in time by b produces the signal $u(t+b)$, and then scaling time by a produces $u(at+b)$.
- **Scale then shift.** Scaling the signal $u(t)$ in time by a produces the signal $u(at)$, and then shifting time by b produces $u(a(t+b))$.

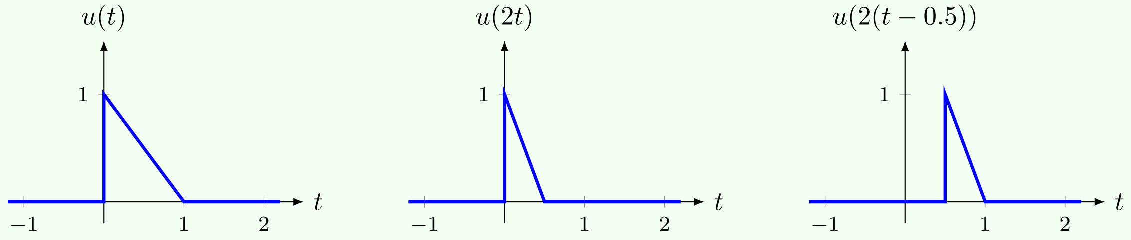
Remark. Shifting and then scaling is *not* the same as scaling and then shifting (by the same amounts). The order of shifting and scaling matters! ■

Example. Consider the continuous-time signal $u(t) = 1 - t$ for $0 \le t \le 1$ and zero otherwise. We will form the shifted and scaled signal $u(2t-1)$ using both methods.

- Using the first method, we need to shift the signal to the right by one and then scale by a half.



- Alternatively, we can find the same signal using the second method by writing it as $u(2(t-0.5))$. To form this signal, we need to first scale by a half and then shift to the right by a half.



Both methods produce the same signal, though we had to shift by the appropriate amount in each case.

2.3 Common signals

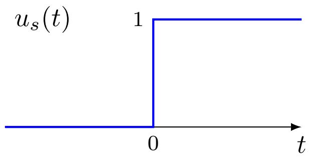
We now introduce some common signals in both continuous and discrete time. These signals can be used as building blocks to construct more complex signals.

Unit step signal

A unit step signal is a signal that is zero for negative time and "steps" to a value of one at time zero. We denote the step signal in continuous time by $u_s(t)$ and in discrete time by $u_s(k)$, where the subscript s stands for "step". This is also sometimes called the Heaviside step function.

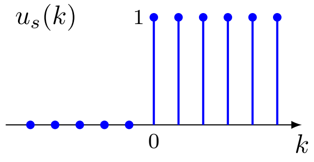
Continuous time

$$u_s(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$



Discrete time

$$u_s(k) = \begin{cases} 1 & \text{if } k \geq 0 \\ 0 & \text{if } k < 0 \end{cases}$$



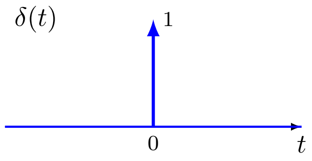
Unit impulse signal

A unit impulse signal is a signal that represents an impulse at time zero and a value of zero at all other times. We denote the impulse signal in continuous time by $\delta(t)$ and in discrete time by $\delta(k)$.

Remark (continuous-time impulse)**.** While the discrete-time impulse signal is rather simple, the continuous-time impulse signal is actually quite complicated. In fact, it is not even a function! There is no explicit expression for $\delta(t)$. Instead, the continuous-time impulse signal is defined by how it acts under integration. You can think of a continuous-time impulse as a signal that is zero everywhere, but is infinite right at time zero in such a way that its integral is one. For this reason, we draw an impulse as an arrow with a height of one at time zero. ■

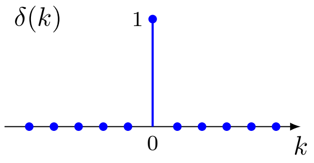
Continuous time

$\delta(t)$ has no explicit formula



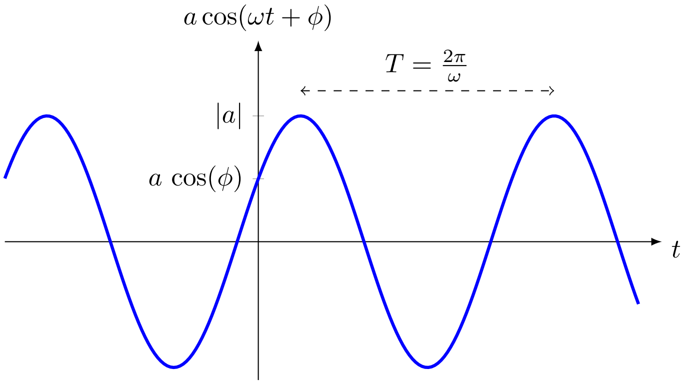
Discrete time

$$\delta(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases}$$



Sinusoidal signal

A continuous-time sinusoid is a scalar signal of the following form.



A sinusoid is parameterized by the following three numbers:

- a is the magnitude
- ω is the angular frequency (rad/s)
- ϕ is the phase (rad)

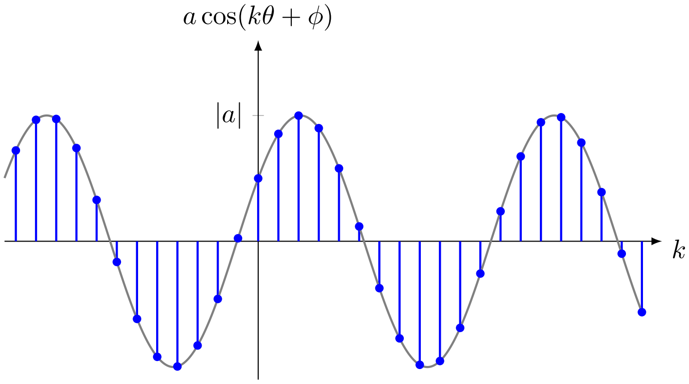
The magnitude a scales the height of the sinusoid, the angular frequency ω scales the sinusoid in time, and the phase ϕ shifts the sinusoid in time.

There are three related quantities that describe how quickly a sinusoid oscillates. The period T has units of seconds and is the period of the signal, which is the amount of time before the signal repeats itself. The frequency f has units of Hertz (or inverse seconds) and describes how quickly the signal oscillates, and the angular frequency ω has units of radians per second which also describes the rate of oscillation. These three quantities are related by

$$T = \frac{1}{f} = \frac{2\pi}{\omega}$$

The frequencies f and ω are proportional to each other and inversely proportional to the period T, so higher frequencies correspond to smaller periods and vice-versa.

Discrete-time sinusoids are obtained by sampling a continuous-time sinusoid. Let T_s denote the sampling period. Then the continuous time t is related to the discrete time k by $t = kT_s$. Substituting this into the formula for a continuous-time sinusoid, we obtain the discrete-time sinusoid $a\cos(\omega kT_s + \phi)$. To simplify the expression, define the angle $\theta = \omega T_s$ which has units of radians. A discrete-time sinusoid then has the following form.

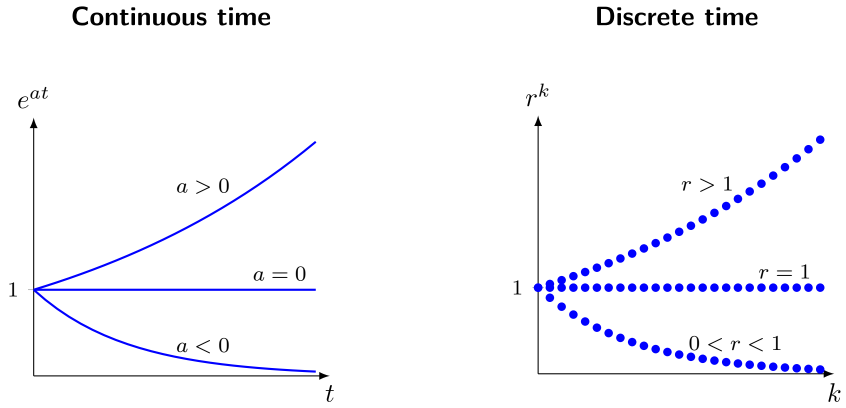


One cycle of the discrete-time sinusoid is completed when $k\theta = 2\pi$, so we can find the number of samples per cycle by solving for k to obtain

$$\text{number of samples per cycle} = \frac{2\pi}{\theta}$$

Exponential signal

An exponential signal is a signal in which time is in the exponent of a constant. A continuous-time exponential signal has the form e^{at} and a discrete-time exponential signal has the form r^k where r is a constant parameter.



Complex exponential signal

Complex exponential signals generalize all of the signals that we have seen so far (step, impulse, sinusoidal, and exponential). These signals will be important when we study systems.

A continuous-time complex exponential signal has the form e^{st} where s is a complex number. The shape of the signal depends on the value of the complex number s. To understand the shape of the signal, write the complex number in rectangular form as $s = a + jb$. Then the complex signal is

$$e^{st} = e^{(a+jb)t} = e^{at}e^{jbt} = e^{at}\left(\cos(bt) + j\sin(bt)\right)$$

where we used properties of the exponential and Euler's formula for complex numbers. This reveals that a complex number is the product of a (real) exponential signal with the summation of real and imaginary sinusoidal signals.

$$e^{st} \quad = \quad \underbrace{e^{at}}_{\text{exponential}} \quad \underbrace{\left(\cos(bt) + j\sin(bt)\right)}_{\text{sinusoid}}$$

The parameter a is the real part of the complex number s, which determines whether the exponential is growing, decaying, or constant. The parameter b is the imaginary part of the complex number s, which determines how fast the real and imaginary parts of the signal are oscillating.

A discrete-time complex exponential signal has the form z^j where z is a complex number. The shape of the signal depends on the value of the complex number z. To understand the shape of the signal, write the complex number in polar form as $z = r\,e^{j\theta}$. Then the complex signal is

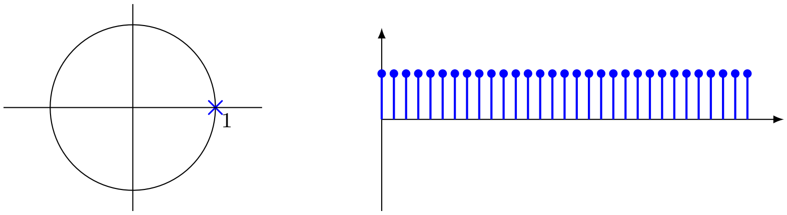
$$z^k = \left(r\,e^{j\theta}\right)^k = r^k\,e^{jk\theta} = r^k\left(\cos(k\theta) + j\sin(k\theta)\right)$$

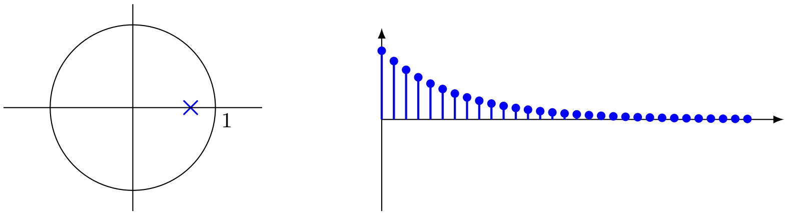
where we used properties of the exponential and Euler's formula for complex numbers. This reveals that a complex number is the product of a (real) exponential signal with the summation of real and imaginary sinusoidal signals.

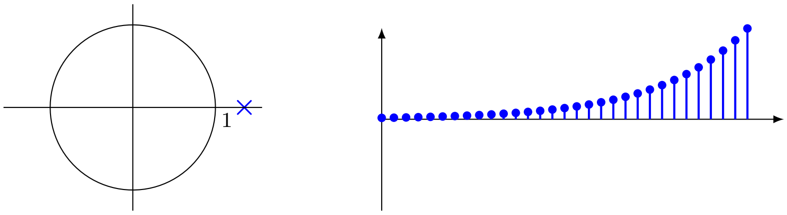
$$z^k = \underbrace{r^k}_{\text{exponential}} \underbrace{\big(\cos(k\theta) + j\sin(k\theta)\big)}_{\text{sinusoid}}$$

The parameter r is the magnitude of the complex number z, which determines whether the exponential is growing, decaying, or constant. The parameter θ is the angle of the complex number z, which determines how fast the real and imaginary parts of the signal are oscillating.

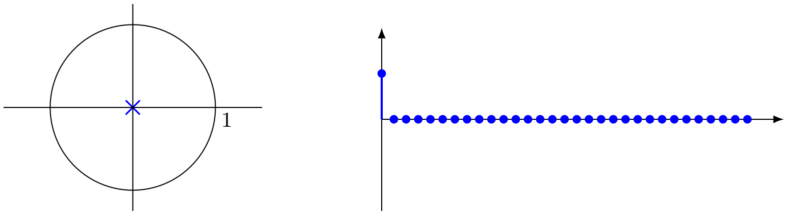
We now visualize each of the various types of discrete-time complex exponential signals along with the corresponding location of the complex number z in the complex plane. First, notice that a complex exponential reduces to a real exponential signal if z is a real number. In this case, the real part of z determines whether the signal grows, decays, or stays constant.



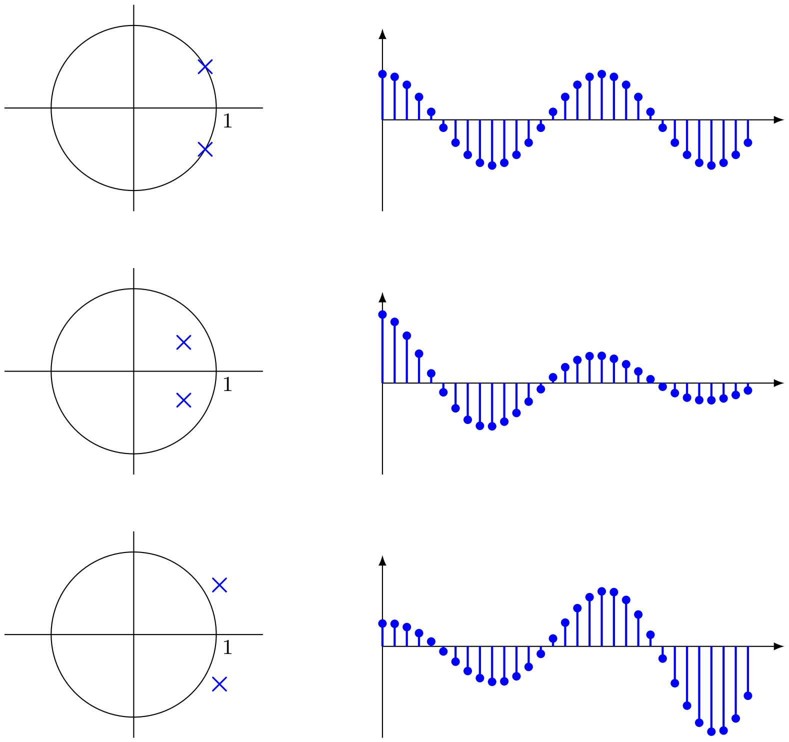




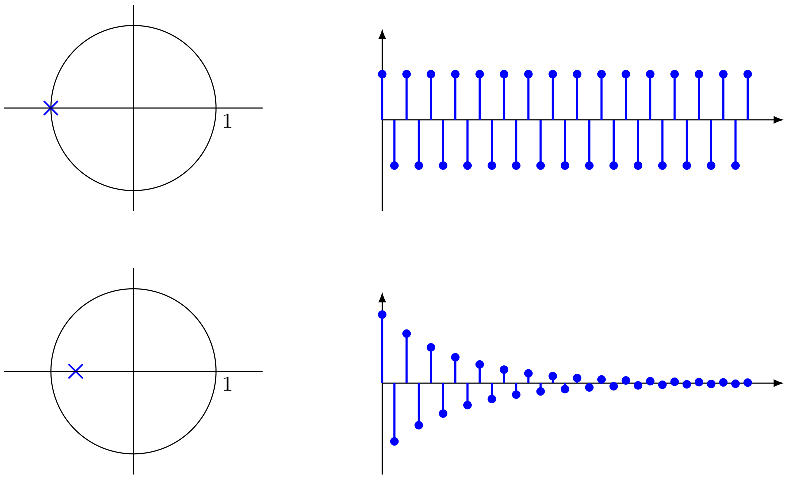
We recover a discrete-time impulse signal when $z = 0$ since $0^0 = 1$ and zero raised to any other power is zero.

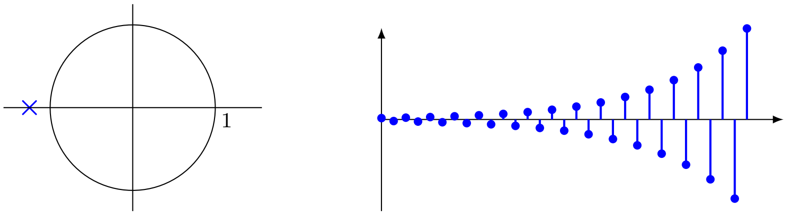


When the magnitude of z is one (so z is on the unit circle in the complex plane), then the real exponential is constant, so the complex exponential signal oscillates without growing or decaying. Since the angle of the complex number determines the frequency of oscillation, moving z inside the unit circle at the same angle results in decaying oscillations, while moving it outside the unit circle at the same angle results in growing oscillations.

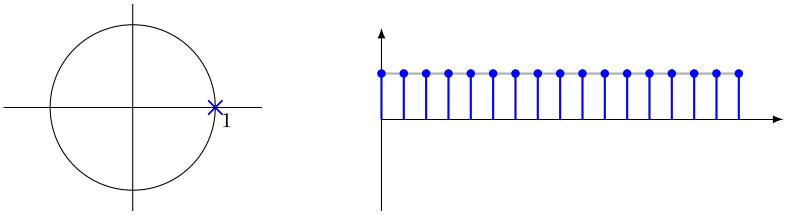


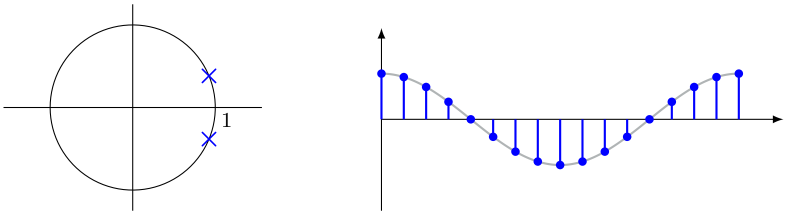
For real negative values of z, the complex exponential is a real signal that oscillates between positive and negative numbers. Once again, if the real part of z is inside the unit circle then the oscillations decay over time, and if the real part of z is outside of the unit circle then they grow over time.

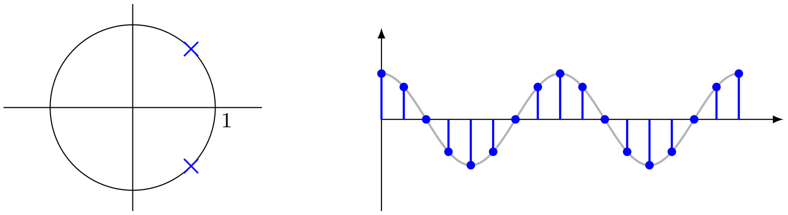


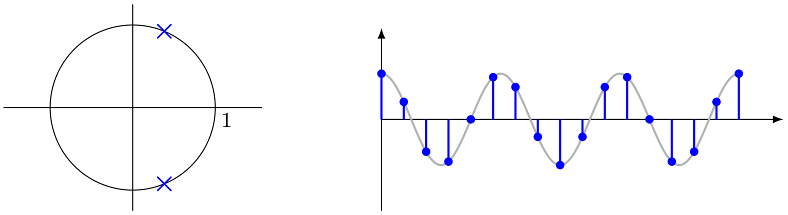


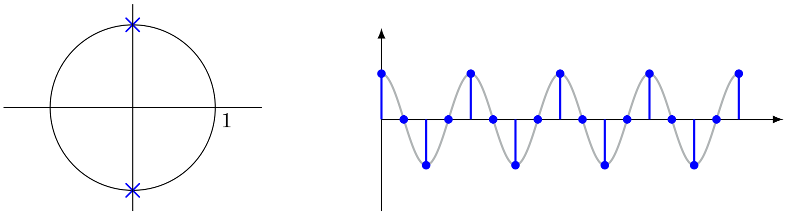
And finally, let's observe how the frequency of oscillation depends on the angle of the complex number z. As z moves around the unit circle, its angle gets larger so the signal oscillates faster.

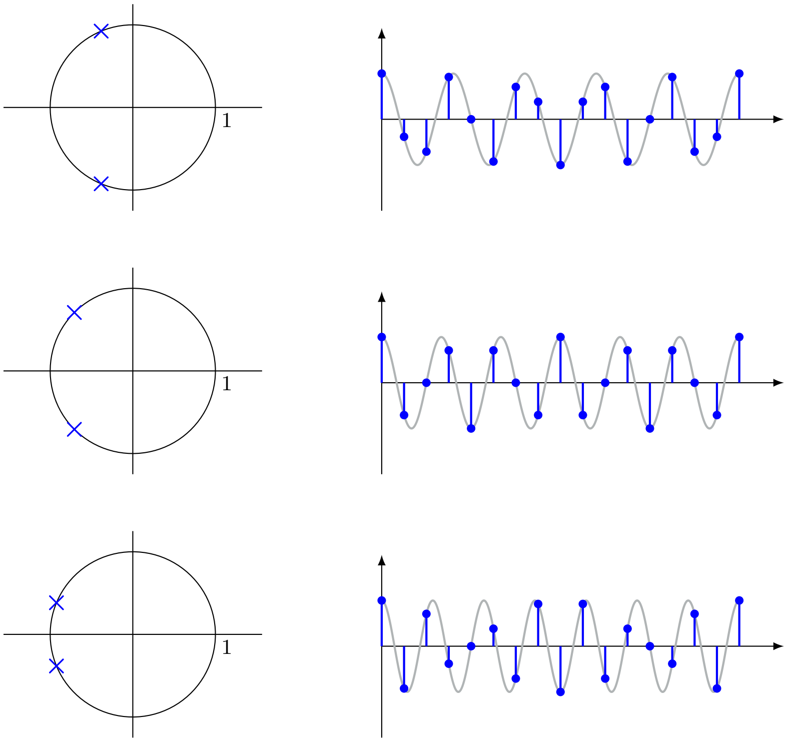




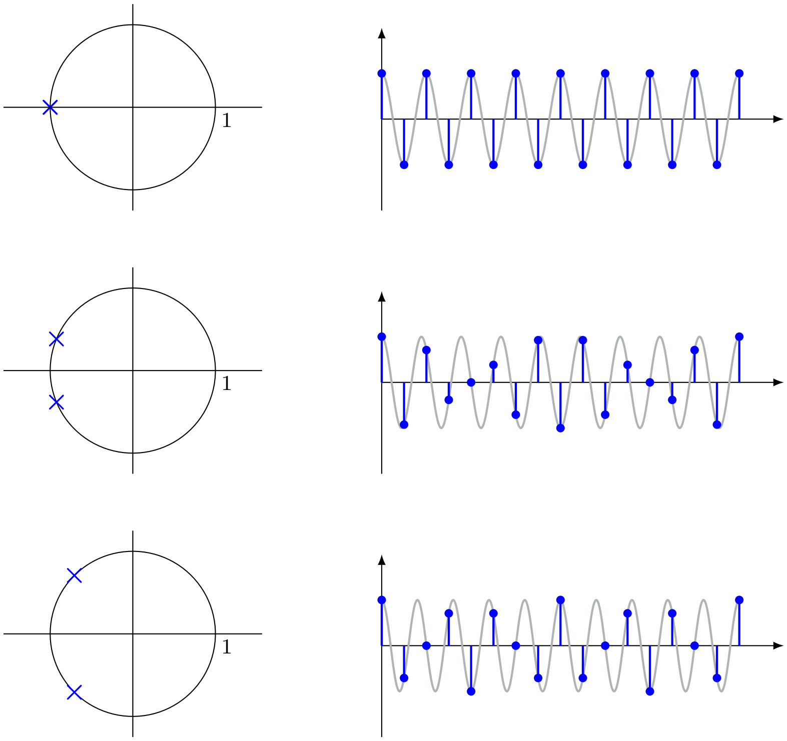


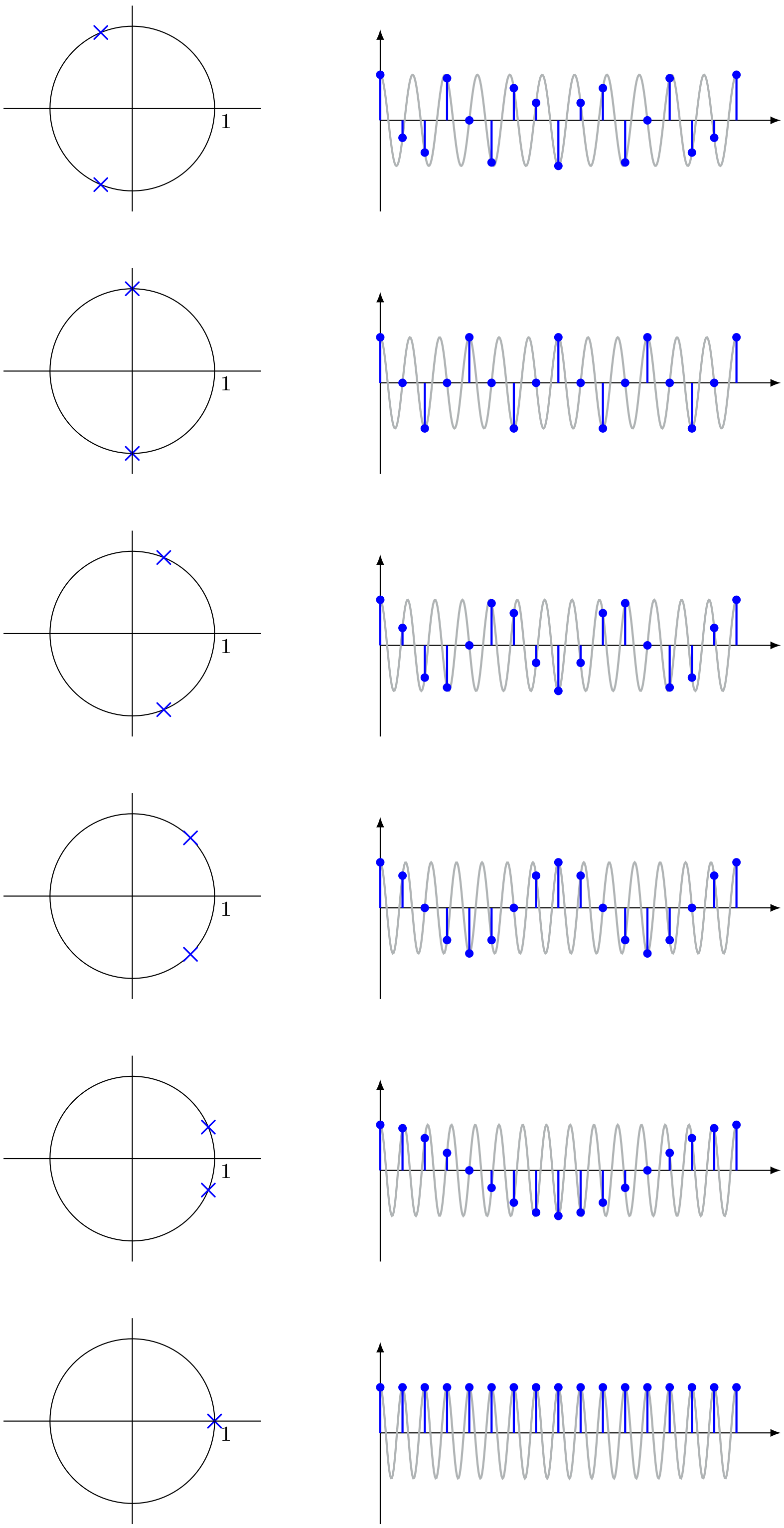






The fastest discrete-time frequency is $\theta = \pi$. After this frequency, faster sinusoids look like lower-frequency sinusoids, which is known as *aliasing*.




1
1
1
1
1
1

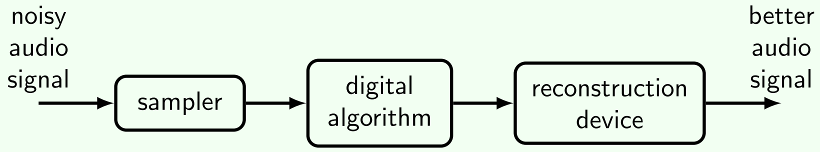
2.4 Converting between continuous and discrete time

Many applications involve both continuous-time and discrete-time signals and systems. The main reason for this is due to the fact that

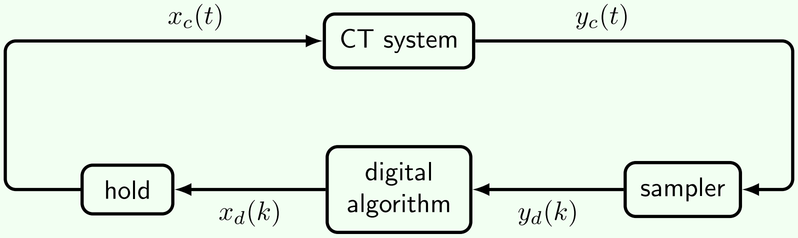
- physical signals are continuous (voltage, force, sound, position), while
- computers process and store discrete information.

Below are a couple prominent examples that involve both continuous-time and discrete-time signals and systems. There are entire courses dedicated to each of these problems.

Example (digital signal processing)**.** Digital signal processing (DSP) studies the use of computers to process signals. Since computers inherently process and store discrete information, DSP focuses on discrete-time systems. An example application is processing a noisy audio signal to produce a better (less noisy) signal. The high-level procedure is shown below, were the continuous-time audio signal is sampled to produce a discrete-time signal, which is then processed on a computer by a digital algorithm, and a continuous-time signal is reconstructed from the processed signal.



Example (digital control)**.** The field of control studies how to choose the inputs to a system to make it behave in a desired manner. Control is often applied to physical systems, such as a mobile robot. In this case, the inputs to the system are continuous-time signals such as voltages applied to motors, while the outputs are continuous-time signals such as the position and velocity of the robot. Robots are often controlled using computers, so these continuous-time signals must first be converted to discrete-time signals to be processed, and then converted back to be applied to the system.



In the remainder of this section, we will look at how to convert signals between continuous and discrete time.

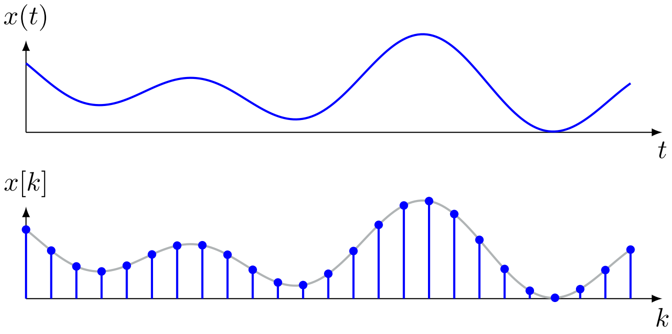
Notation. In this section, we will use parentheses to denote continuous-time signals (such as $x(t)$) and square brackets to denote discrete-time signals (such as $x[k]$). This is just to help the reader distinguish between the two types of signals. ■

Continuous to discrete

The process of converting a continuous-time signal to a discrete-time signal is called *sampling*. In the most basic form of sampling, the discrete-time signal is formed by taking the value of the continuous-time signal at evenly-spaced intervals. The time between samples is called the *sampling period*, which we denote by T_s. Given a continuous-time signal $x(t)$, the sampled discrete-time signal is then

$$x[k] \;=\; x(kT_s)$$

The sampling process is illustrated below.



There are three related quantities typically used to describe how fast a continuous-time signal is sampled.

- the sampling period T_s which has units of seconds (s)
- the sampling frequency f_s which has units of Hertz (Hz, or inverse seconds)
- the (angular) sampling frequency ω_s which has units of radians per second (rad/s)

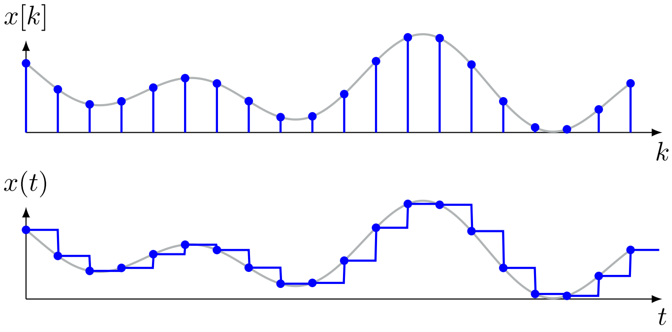
$$T_s \;=\; \frac{1}{f_s} \;=\; \frac{2\pi}{\omega_s}$$

Discrete to continuous

Given a discrete-time signal, there are many ways to "fill in the gaps" to construct a continuous-time signal. The simplest conversion method is called a *zero-order hold*. Here, the value of the discrete-time signal is "held" until the time of the next sample. This constructs a piecewise constant continuous-time signal as follows:

$$x(t) = x[k] \quad \text{for} \quad kT_s \le t < (k+1)T_s$$

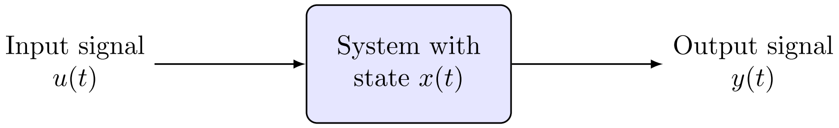
Applying this procedure to the discrete-time signal from above, we obtain the following continuous-time signal (the original signal is shown in gray). There are other (more complicated) methods to reconstruct the signal.



3

Systems

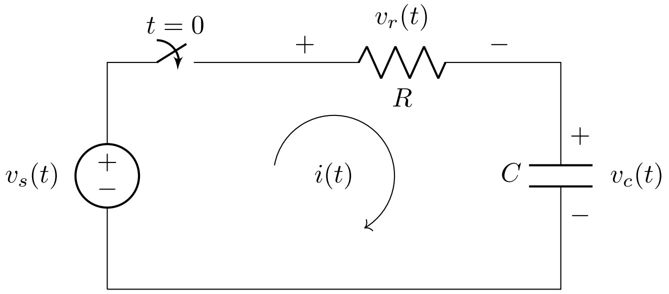
A *system* is a relationship between signals. We typically split the signals into two groups: the *input signals* are those that we can use to influence the behavior of the system, and the *output signals* are those that we can observe or want to control. Systems are often *dynamic*, meaning that the output of the system at any given time depends not only on the current value of the input, but also on previous values of the input. In other words, dynamic systems have *memory*, so the behavior now depends on what has happened to the system in the past. This memory is captured by the *state* of the system, which is a set of signals internal to the system that completely describe its current state of memory. A *continuous-time system* has input, output, and state that are continuous-time signals, while a *discrete-time system* involves discrete signals. The following diagram illustrates the main components of a general system.



3.1 Examples

Circuit

Circuits are electrical systems, where the signals that the system relates to each other are the voltages and currents in the circuit. For instance, consider the following circuit consisting of a voltage source, switch, resistor, and capacitor.



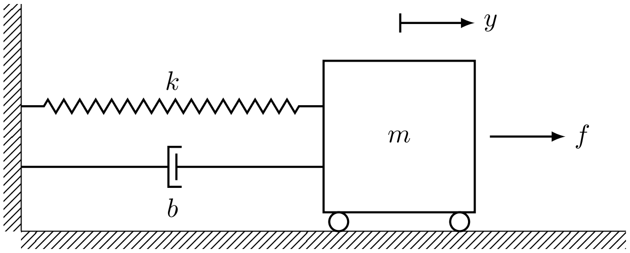
The input signal to this system is the source voltage $v_s(t)$, and the output signal is the voltage across the capacitor $v_c(t)$. Using Kirchhoff's laws, the relationship between the source and capacitor voltage is described by the following first-order differential equation:

$$v_s(t) = RC\,\dot{v}_c(t) + v_c(t)$$

The solution to this equation gives the voltage across the capacitor $v_c(t)$, which depends on both the source voltage $v_s(t)$ and the initial capacitor voltage $v_c(0)$. The memory of this system is captured by the voltage across the capacitor, so the state is $x(t) = v_c(t)$. Together, the initial state and the input signal uniquely determine the output.

Mechanical system

Mechanical systems relate signals like the position and velocity of an object with the forces that act on it. For instance, consider the following mechanical system consisting of a mass attached to the wall by a spring and a damper.



The input signal to this system is the external force $f(t)$ applied to the mass, and the output signal is the horizontal position $y(t)$ of the mass. Using Newton's laws, the relationship between the applied force and the position of the mass is described by the following second-order differential equation:

$$f(t) = m\ddot{y}(t) + b\dot{y}(t) + ky(t)$$

Since this is a second-order differential equation, two initial conditions are needed to solve for the position. For instance, we could use the initial position $y(0)$ and the initial velocity $\dot{y}(0)$ as the initial conditions. Since these are both needed to specify the output, the state consists of both the position *and* velocity, that is, $x = (y, \dot{y})$.

Savings account

A savings account can be modeled as a discrete-time system, where the input signal is the amount $A(k)$ deposited into the account in month k and the output signal is the principal (or balance) $P(k)$ of the account. Here, the time index k corresponds to the month and takes nonnegative integer values $k = 0, 1, 2, \ldots$, where month zero corresponds to when the account was established. Given that the account earns interest i, the dynamics relating the amount deposited with the principal is

$$\underbrace{P(k+1)}_{\text{balance next month}} = \underbrace{P(k)}_{\text{balance this month}} + \underbrace{\frac{i}{12}P(k)}_{\text{interest}} + \underbrace{A(k)}_{\text{deposits}}$$

While continuous-time systems are described by *differential* equations, discrete-time systems are described by *difference* equations that involve the signals at shifted times (such as k and $k+1$). For this system, the state is the principal $P(k)$.

Numerical algorithms

An iterative numerical algorithm is a sequence of instructions that may be used to solve a mathematical problem. We can interpret such algorithms as discrete-time systems. There are a numerous algorithms that can be applied to solve a variety of problems. We now discuss a few such applications.

Square root calculation

As a simple example, we can use Newton's method to compute the square root $\sqrt{a}$ of a positive number $a > 0$. To do so, let $x(0) > 0$ be the initial estimate of the square root. Then for each iteration $k = 0, 1, 2, \ldots$, update the estimate as

$$x(k+1) = \frac{1}{2}\left(x(k) + \frac{a}{x(k)}\right)$$

As this system is iterated, the iterates $x(k)$ converge to the square root $\sqrt{a}$ in the limit as $k \to \infty$.

Fixed-point iterations

We can use numerical algorithms to solve nonlinear equations using fixed-point iterations. Consider the nonlinear equation

$$2x - e^{-x} = 1$$

To solve this equation, we first isolate one instance of x. For example, solving for the first instance of x gives

$$x = \frac{1}{2}\big(1 + e^{-x}\big)$$

To turn this into a discrete-time system, we replace x on the left-hand side by $x(k+1)$ and x on the right-hand side by $x(k)$. This gives the discrete-time system

$$x(k+1) = \frac{1}{2}\big(1 + e^{-x(k)}\big)$$

We can start at some estimate $x(0)$ of the solution and iterate for $k = 0, 1, 2, \ldots$. If the iterates of the system converge, then this must be a solution to the original nonlinear equation.

Instead of isolating the first instance of x in the equation, we could have isolated the second instance to obtain the iteration

$$x(k+1) = -\ln\big(2x(k) - 1\big)$$

But this iteration does not converge! In general, there is no guarantee that a fixed-point iteration will converge to a solution of the nonlinear equation, so you may need to try isolating a different instance of x, or start the iteration from a different initial condition $x(0)$.

Optimization

Consider the problem of finding the value x that minimizes a function $f(x)$. Some examples of optimization problems are the following:

- x may be the number of widgets that a compancy produces and $f(x)$ the associated production costs.
- x may be the route that you take to campus and $f(x)$ the amount of time that you spend commuting.

Such optimization problems are typically described mathematically as

$$\underset{x}{\text{minimize}} \;\; f(x)$$

If the objective function $f(x)$ is differentiable (meaning that its derivative exists), we can use the gradient descent algorithm to find the minimizer:

$$x(k+1) = x(k) - \alpha \frac{\mathrm{d}}{\mathrm{d}x} f(x(k))$$

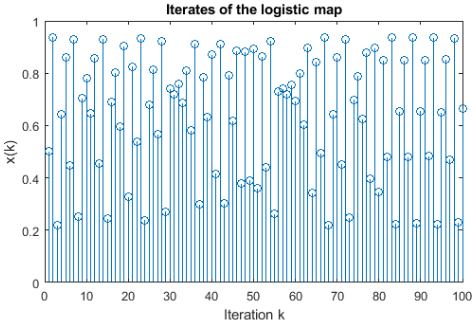
where $\alpha > 0$ is the stepsize. Whether this algorithm converges to the optimal solution or not depends on the properties of the objective function $f(x)$ and the choice of stepsize α.

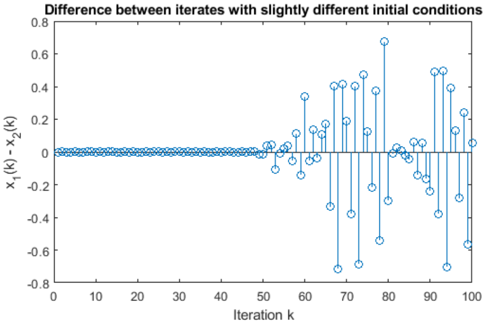
Logistic map

An interesting discrete-time system is the *logistic map*, which is defined by the iterations

$$x(k+1) = r\, x(k)\, (1 - x(k))$$

where $r \in [0, 4]$ is a constant parameter. The dynamics of this system highly depend on the parameter r. For instance, the iterates with $r = 3.75$ and initial condition $x(0) = 0.5$ are shown on the left.





One of the interesting properties of this system is that it is *chaotic*, meaning trajectories starting arbitrarily closed together eventually diverge far apart. To illustrate this, suppose we instead initialize the system with $x(0) = 0.500001$. The difference between the iterates using this initialization and the initialization $x(0) = 0.5$ from before is shown on the right. While the trajectories start out close (so that their difference is approximately zero), they eventually become very far apart. This illustrates some of the interesting behavior that can occur when the dynamics are *nonlinear*.

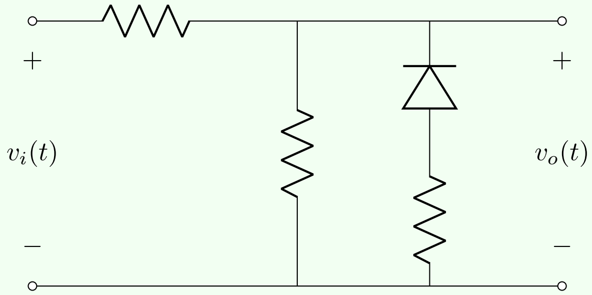
3.2 Properties

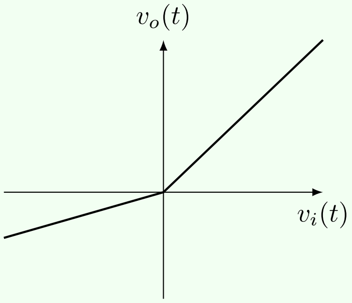
It is often useful to characterize systems based on their properties. We now define some of the basic properties that a system may have.

Memory

A system is *static* if the value of the output signal at any given time depends only on the value of the input signal *at the same time*. In other words, $y(t)$ is a function of $u(t)$, but not $u(\tau)$ for any $\tau \neq t$. Static systems have no state.

Example (Static system). An example of a static system is a circuit without any energy storage elements (such as capacitors or inductors). For instance, the output voltage $v_o(t)$ depends only on the input voltage $v_i(t)$ at the current time in the following circuit:





The plot of the output voltage as a function of the input voltage is shown on the right. Since the system has no memory, the output at time t is a function of the input at time t only.

A system is *dynamic* if the value of the output signal depends on the input signal at past (or future) times. In other words, $y(t)$ depends on $u(\tau)$ for some $\tau \neq t$. Dynamic systems have a state, and the response of the system depends not only on the input signal, but on the initial state as well. All systems that have memory are dynamic, so essential all “interesting” systems are dynamic.

An RC circuit is an example of a dynamic system. The voltage across the capacitor depends on what currents have been applied to it in the past, and the output voltage depends not only on the supply voltage but the initial capacitor voltage as well. Other examples of dynamic systems are the spring–mass–damper system, savings account, fixed-point iterations, and logistic map.

Linearity

A system is *linear* if a weighted sum of input signals produces the same weighted sum of the corresponding output signals. In particular, a system is linear if the input $au_1 + bu_2$ produces the output $ay_1 + by_2$ for all scalars a and b and all input signals u_1 and u_2 and their corresponding output signals y_1 and y_2.

To show that a system is nonlinear, we only need to find specific input-output pairs $u_1 \mapsto y_1$ and $u_2 \mapsto y_2$ and constants a and b such that the weighted sum of inputs $a\,u_1 + b\,u_2$ does *not* produce the weighted sum of outputs $a\,y_1 + b\,y_2$.

Example (Linear vs nonlinear). Examples of linear systems are the RC circuit, savings account, and the system

$$y(t) = u(t) + 3\int_{-\infty}^{t} u(\tau)\,\mathrm{d}\tau + 2\,\dot{u}(t)$$

Examples of nonlinear systems include the following:

$$\begin{aligned}
y(t) &= \tfrac{1}{2}(u(t)+3)^2 \\
y(t) &= \int_0^t \sqrt{u(\tau)}\,\mathrm{d}\tau \\
\dot{y}(t) &= y(t)\,u(t),\ \ y(0) = 1 \\
y(k+1) &= \frac{1}{2}\left[y(k) + \frac{u(k)}{y(k)}\right] \\
y(k+1) &= [y(k)]^2 + u(k)
\end{aligned}$$

The previous example suggests a general principle that we may use to quickly detect nonlinearity: if any of the input or output signals are raised to a power, multiplied, or divided, the system is nonlinear.

Time invariance

A system is *time invariant* if shifting the input signal in time shifts the output signal by the same amount in time. Suppose the input signal $u_1(t)$ produces the output signal $y_1(t)$. Then a system is time invariant if the shifted input signal $u_2(t) = u_1(t-\tau)$ produces the shifted output signal $y_2(t) = y_1(t-\tau)$ for all scalars τ and all input signals u_1 and its corresponding output signal y_1.

A system is *time varying* if shifting the input signal in time does *not* shift the output signal by the same amount in time. In other words, the dynamics of the system change over time.

Example (Time-varying parameter). Time-varying systems often arise from systems whose parameters vary with time. For instance, the RC circuit is time invariant if the resistance and capacitance are constant. But if the resistance $R(t)$ of the resistor changes over time, then the dynamics become time varying:

$$v_s(t) = R(t)\, C\, \dot{v}_c(t) + v_c(t)$$

Likewise, the savings account system is time invariant if the interest rate parameter is constant, but it is time varying if the interest rate changes over time (which it does!):

$$P(k+1) = \left(1 + \frac{i(k)}{12}\right) P(k) + A(k)$$

Causality

A system is *causal* if the output at any given time depends only on the input at previous times. In particular, a system is causal if $y(t)$ only depends on the input $u(\tau)$ at past times $\tau \leq t$ and *not* future times $\tau > t$.

Example. All physical systems are causal, as a physical signal cannot predict the values of future input signals. However, a noncausal system can be implemented in several circumstances:

- We can implement a noncausal system if the independent variable does not actually represent time. In image processing, for example, signals are images and the independent variables correspond to dimensions in space along the image.
- Even when the independent variable does represent time, we can implement a noncausal system when the entire input signal is available before processing. For example, consider processing an audio signal that is stored on a computer. Since the entire signal is available, the algorithm need not process the audio signal in a causal manner.

Dimensionality

A system is *finite-dimensional* if the statespace (the space in which the state takes its values) is a finite-dimensional vector space such as $\mathbb{R}^n$. All of the systems that we will study are finite dimensional.

Example (Continuous-time delay)**.** A relevant example of an infinite-dimensional (linear time-invariant) system is a delay in continuous time. For a delay of time T, the output is $y(t) = u(t - T)$. The state of this system is the set of all inputs over the past T seconds,

$$x(t) = \{u(\tau) \mid t - T \leq \tau \leq t\}$$

This is an infinite-dimensional space, since it is a function over a continuous interval.

- **Time.** A system evolves in *continuous time* if time t takes values in a continuous interval, such as all real numbers $(-\infty, \infty)$ or nonnegative real numbers $[0, \infty)$. In contrast, a system evolves in *discrete time* if time t takes values in a discrete set, such as all integers $\{\dots, -2, -1, 0, 1, 2, \dots\}$ or natural numbers $\{0, 1, 2, \dots\}$.
- **Linearity.** A system is *linear* if a weighted sum of input signals produces the same weighted sum of the corresponding output signals. In particular, a system is linear if the input $a\,u_1 + b\,u_2$ produces the output $a\,y_1 + b\,y_2$ for all scalars a and b and all input signals u_1 and u_2 and their corresponding output signals y_1 and y_2.
- **Time invariance.** A system is *time invariant* if shifting the input signal in time shifts the output signal by the same amount in time. In particular, a system is time invariant if the shifted input signal $t \mapsto u(t - \tau)$ produces the shifted output signal $t \mapsto y(t - \tau)$ for all scalars τ and all input signals u and its corresponding output signal y.
- **Causality.** A system is *causal* if at each time the output at that time only depends on the input at previous times. In particular, a system is causal if $y(t)$ only depends on the $u(\tau)$ for $\tau \leq t$. All physical systems are causal.
- **Dimensionality.** A system is *finite-dimensional* if the state $x(t)$ at each time t is in a finite-dimensional vector space such as $\mathbb{R}^n$.

3.3 LTI systems

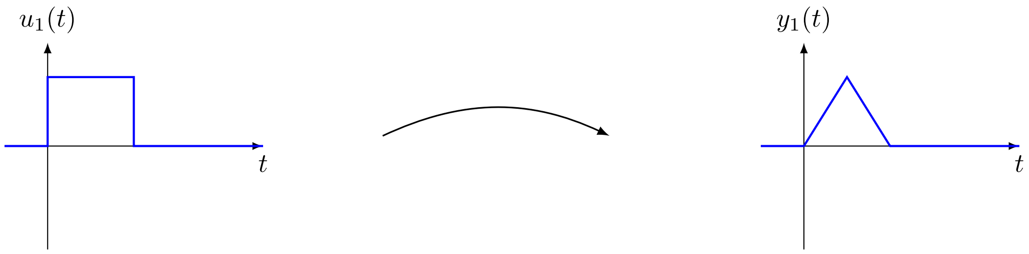
An important class of systems are those that are *linear* and *time invariant*, or LTI systems. There are several reasons to study LTI systems:

- Some systems are inherently linear and time invariant, such as RLC circuits, spring–mass–damper systems, and many mathematical operations like differentiation and integration.
- Many systems can be adequately approximated by LTI systems. Even if a system is not LTI, we can often approximate it by one that is.
- Understanding LTI systems forms a foundational basis for understanding more complex systems.
- LTI systems have a rich and elegant theory. They strike a balance between being simple enough to thoroughly understand and yet are rich enough to be broadly applicable.

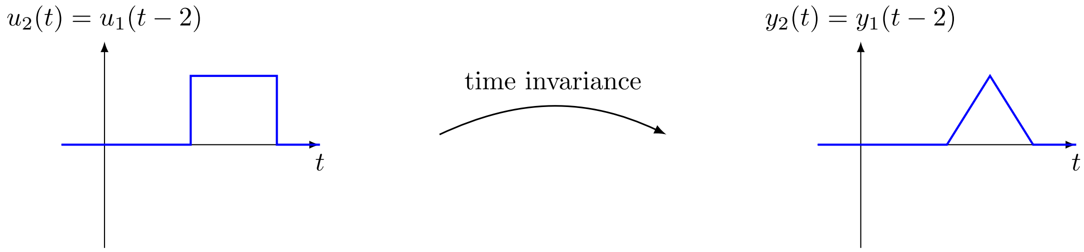
LTI systems are a rich class of systems for which we have a complete and thorough understanding of how they work, and the tools that we learn will help us in understanding more complicated systems. LTI systems are the types of systems that you study in introductory courses in circuits, mechanics, and mathematics.

Motivation

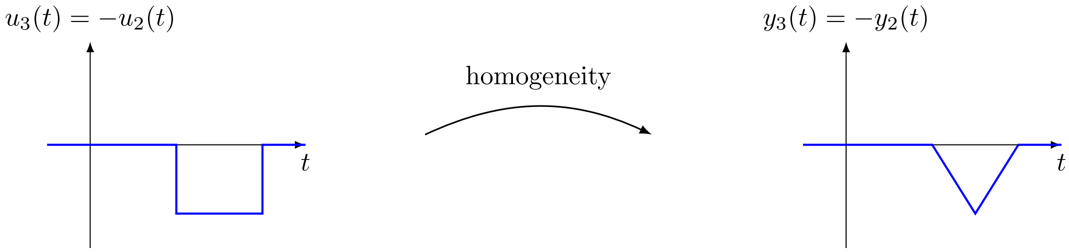
Consider an LTI system, and suppose we know a single input/output pair for the system as shown below.



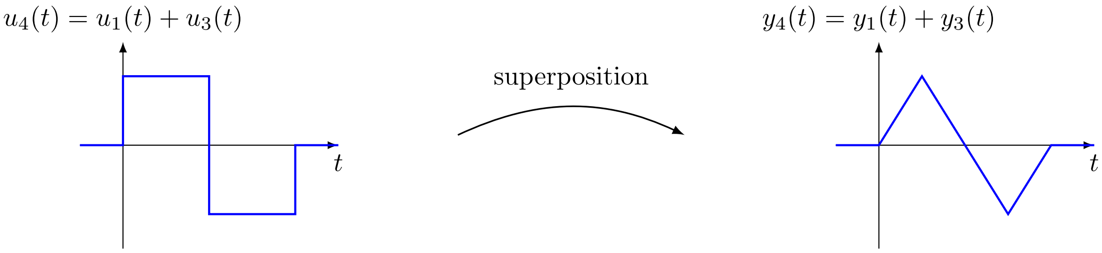
Because the system is time invariant, we can shift the input in time and the output gets shifted in time by the same amount.



Since the system is linear (and therefore also homogeneous), we can scale this shifted input and the output gets scaled by the same amount.



Also from linearity (this time using superposition), we can sum the first and third inputs and the output will be the sum of the corresponding outputs.



We could continue using the linearity and time invariance properties to construct new input/output pairs, all from knowing a single input/output pair and that the system is LTI!

Generalization

Let's now generalize the motivating example to see all the ways in which we can exploit that a system is linear and time invariant. As before, suppose we know a single input-output pair of an LTI system (this time we will work with a discrete-time system for simplicity, although the procedure is similar in continuous time).

$$u(k) \quad \mapsto \quad y(k)$$

Since the system is time invariant, then shifting the input signal in time shifts the output signal by the same amount in time.

$$u(k-m) \quad \mapsto \quad y(k-m)$$

This produces a different input-output pair for each time shift (corresponding to the value of m). Now from linearity, taking a weighted sum of inputs produces an identical weighted sum of outputs. Applying this to the above input-output pair for all values of m gives

$$\sum_m w(m)\, u(k-m) \quad \mapsto \quad \sum_m w(m)\, y(k-m)$$

where there is a weight $w(m)$ for each value of m.

Later, we will see that this expression is general enough to represent *any* input-output pair of the system. In other words, knowing a single input-output pair is enough to completely characterize any LTI system!

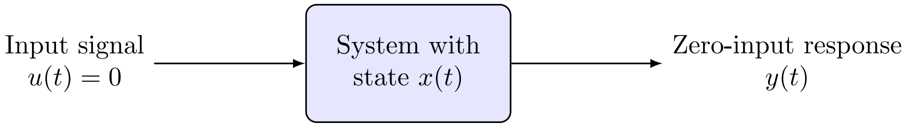
3.4 System response

The most fundamental question regarding the analysis of a system is: *how does the system respond to various excitations?* The output of a system, known as the *response*, depends on both the input signal and the initial conditions (or initial state). In an RC circuit, for instance, the voltage across the capacitor depends on both the source voltage and the initial capacitor voltage.

There are various types of responses, based on the nature of the input signal and initial condition. We now describe several important responses of a system.

Zero-input response

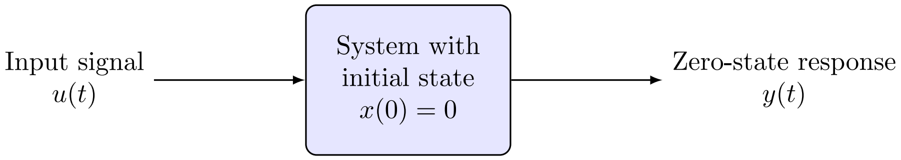
The zero-input response (ZIR) is the output of the system due to the initial conditions when the input signal is zero (at all times). The zero-input response depends only on the initial state of the system.



Example (ZIR of spring–mass-damper system). Going back to our example of a spring–mass–damper mechanical system, the zero-input response is the position of the mass when no force is applied. In this case, the response depends on only the initial position and velocity of the mass.

Zero-state response

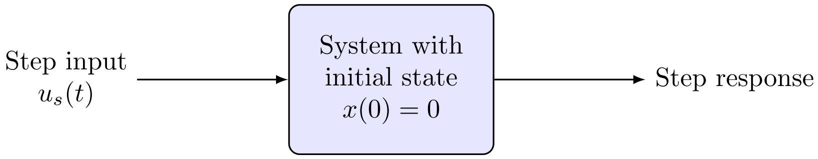
The *zero-state response* is the output of the system due to the input signal when the initial state is zero. By "initial state", we typically mean the state at time zero, in which case it is assumed that the input signal is zero for times before the initial time.



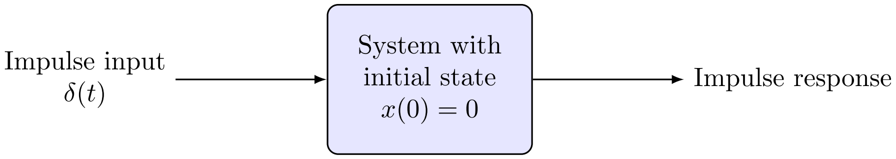
Example (ZSR of spring–mass-damper system)**.** Continuing our example, the zero-state response of the mechanical system is the position of the mass when it starts at rest. In this case, the response depends on only the force applied to the mass.

The zero-state response depends on the particular input signal. Two common zero-state responses are the following.

- The **step response** is the zero-state response due to a unit step input signal.



- The **impulse response** is the zero-state response due to a unit impulse input signal.



Part II

Signal Transforms

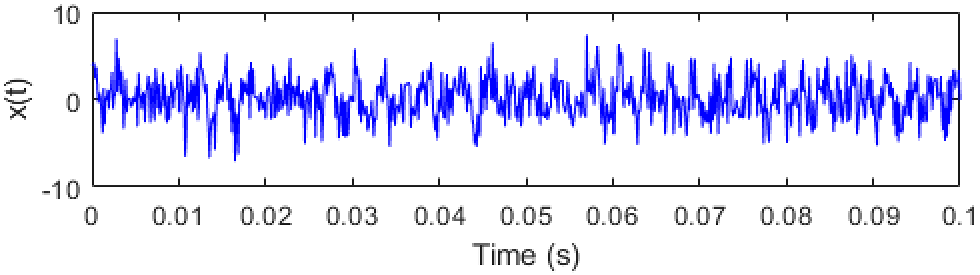
4

Fourier Series

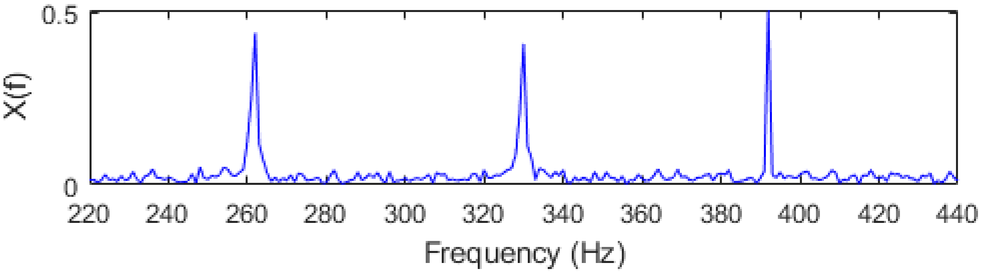
Signals are functions of time. Another interpretation of a signal, however, is as a function of *frequency*. This frequency-domain representation of a signal describes how much of each frequency is contained in the signal. In this chapter, we study the frequency-domain representation of periodic signals. Later, we will study frequency-domain representations of more general signals.

4.1 Motivation

Recall that a signal is a function of time. Fourier analysis is used to express a signal as a function of *frequency*. As an example, consider the following signal $x(t)$.



We can express the same signal in terms of its frequency components as follows.



It is now clear that the signal consists of the three distinct frequencies 262 Hz, 330 Hz, and 392 Hz which correspond to the notes C4, E4, and G4 in a C-major chord. The small values at other frequencies correspond to noise that is corrupting the signal. This representation of the signal is *equivalent* to its description as a function of time. There is a unique mapping that allows us to convert back and forth between the time-domain representation $x(t)$ and the frequency-domain representation $X(f)$ of the signal.

The time-domain representation of a signal can be continuous or discrete depending on the values that the time variable can have (its domain). In the same way, the frequency-domain representation can also be continuous or discrete depending on the values that the frequency variable can have. Signals that can be represented by a discrete set of frequencies are *periodic* in time, while non-periodic signals consist of a continuous set of frequencies. Therefore, the type of transformation between the time-domain and frequency-domain representations of a signal depend on the time domain of the signal (continuous or discrete) and its periodicity (periodic or not). The representation of periodic signals by their frequencies is called the *Fourier series*, and the representation of nonperiodic signals by their frequencies is called the *Fourier transform*. The following table summarizes the transformations from the time-domain representation of a signal to its frequency-domain representation based on the type of signal.

	Continuous time	**Discrete time**
Periodic (power) signals	Fourier Series	Discrete Fourier Transform
Energy signals	Fourier Transform	Discrete-Time Fourier Transform
Arbitrary signals	Laplace Transform	z-Transform

4.2 Vectors

The Fourier series represents a periodic signal in terms of a discrete set of frequencies. Each individual frequency is a sinusoidal signal. We will represent periodic signals as a weighted sum of distinct frequencies. In continuous time, for example, we will represent a periodic time-domain signal $x(t)$ as a summation of a set of sinusoids with (angular) frequencies $\omega_1, \omega_2, \omega_3, \ldots$, where each frequency has a corresponding weight $w_1, w_2, w_3, \ldots$ that describes how much of that frequency is in the signal:

$$x(t) = w_1 \sin(\omega_1 t) + w_2 \sin(\omega_2 t) + w_3 \sin(\omega_3 t) + \ldots$$

In this section, we will learn how to determine the appropriate frequencies and their corresponding weights such that this equation holds for all time. The expression on the right is then the frequency-domain representation of the signal.

Orthogonal decomposition of a vector

To find the Fourier series representation of a signal, we need to find "how much" of a given sinusoid is contained in the signal (which corresponds to the weight of that sinusoid in the summation). Before discussing (infinite-dimensional) signals, let's consider how to do this with finite-dimensional vectors. In the discussion that follows, we will consider arbitrary vectors, so the formulas for signals will be the same!

Given a vector u, how much of the vector is in some given direction v? In other words, let's approximate the vector u by the scaled vector cv, where c is a scalar. We will find c such that the error between the vector u and its approximation cv is minimized. We can measure the similarity between two vectors by taking the norm of their difference. Recall that the norm of a vector is the square root of its inner product with itself. So its squared norm is just the inner product with itself. In particular, the squared-norm of the error is

$$\|u - cv\|^2 = \langle u - cv, u - cv \rangle = \|u\|^2 - 2c \langle u, v \rangle + c^2 \|v\|^2$$

This is a quadratic function of c. The value of c that minimizes the error is

$$c = \frac{\langle u, v \rangle}{\|v\|^2}$$

Therefore, the best approximation of the vector u in terms of the vector v is

$$u \approx \frac{\langle u, v\rangle}{\|v\|^2}\, v$$

Example (vectors). As a simple example, consider the two-dimensional vectors

$$u = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad \text{and} \qquad v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Computing the inner product and norm gives

$$\langle u, v\rangle = u^{\mathsf{T}} v = \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \qquad \text{and} \qquad \|v\|^2 = v^{\mathsf{T}} v = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1$$

Therefore, the coefficient is $c = 2$ as expected.

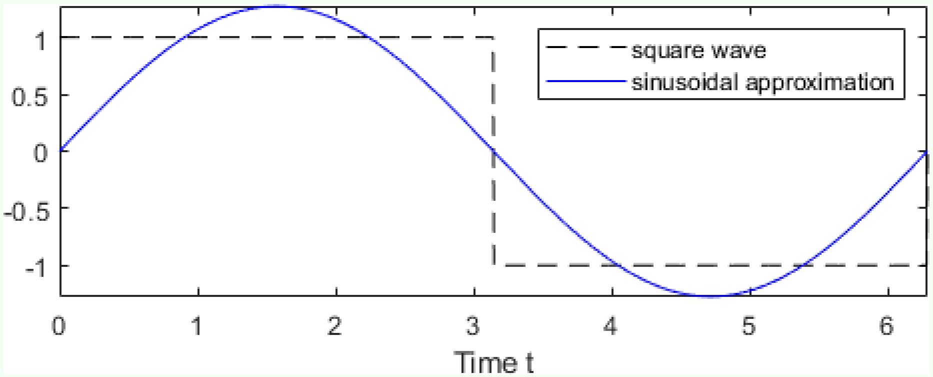
Example (signals). As a more interesting example, consider the square wave $u(t) = 1$ for $t \in [0, \pi)$ and $u(t) = -1$ for $t \in [\pi, 2\pi)$. Suppose that we want to approximate this signal by a sinusoid of the form $v(t) = \sin(t)$. In this case, the inner product between the two signals is the integral of their product,

$$\langle u, v\rangle = \int_0^{2\pi} u(t)\, v(t)\, \mathrm{d}t = \int_0^{\pi} \sin(t)\, \mathrm{d}t + \int_{\pi}^{2\pi} -\sin(t)\, \mathrm{d}t = 4$$

The squared-norm is the inner product of the signal with itself, which is also the power of the signal,

$$\|v\|^2 = \langle v, v\rangle = \int_0^{2\pi} v(t)^2\, \mathrm{d}t = \int_0^{2\pi} \sin^2(t)\, \mathrm{d}t = \pi$$

Therefore, the amount of the signal u in the direction v is $c = 4/\pi$. The original signal u and its sinusoidal approximation cv are shown below.



So far, we have represented a vector by another single vector. But if we want a better approximation, we need to use more vectors. Now suppose that we want to approximate a vector u using a set of vectors $v_1, v_2, v_3, \ldots$,

$$u \approx c_1\, v_1 + c_2\, v_2 + c_3\, v_3 + \ldots$$

To obtain a good approximation of u, we need the set of vectors $v_1, v_2, v_3, \ldots$ to be "different" enough (for instance, in the extreme case that they are all the same vector, this would not be a very good approximation!). In particular, we will assume that the vectors are orthogonal, meaning that the inner product between any two different vectors is zero:

$$\langle v_n, v_m \rangle = 0 \quad \text{for } n \neq m.$$

As before, we will consider the squared-norm of the error between the signal and its approximation. Let's first do the computation when there are only two vectors v_1 and v_2. In this case, the squared norm of the error is

$$\|u - (c_1\, v_1 + c_2\, v_2)\|^2 = \langle u - c_1\, v_1 - c_2\, v_2, u - c_1\, v_1 - c_2\, v_2 \rangle$$

Using that the properties of the inner product, we can expand this to obtain

$$\|u - (c_1\, v_1 + c_2\, v_2)\|^2 = \langle u, u \rangle - 2c_1 \langle u, v_1 \rangle - 2c_2 \langle u, v_2 \rangle - 2c_1 c_2 \langle v_1, v_2 \rangle - c_1^2 \langle v_1, v_1 \rangle - c_2^2 \langle v_2, v_2 \rangle$$

Since v_1 and v_2 are orthogonal (by assumption), their inner product is zero. For a general set of orthogonal vectors, the squared norm of the error is

$$\left\| u - \sum_n c_n\, v_n \right\|^2 = \|u\|^2 - \sum_n 2c_n\, \langle u, v_n \rangle + \sum_n c_n^2\, \|v_n\|^2$$

This is a quadratic function of the coefficients c_n. Moreover, because the vectors are orthogonal, each coefficient c_n only affects terms involving the corresponding vector v_n. We can therefore minimize the error with respect to each coefficient individually to obtain

$$c_n = \frac{\langle u, v_n \rangle}{\|v_n\|^2}$$

Therefore, the best approximation of a vector u in terms of an orthogonal set of vectors v_n is

$$u \approx \sum_n \frac{\langle u, v_n \rangle}{\|v_n\|^2}\, v_n$$

In general, this is only an approximation of the original signal (the error is not zero). But if we use a rich enough set of vectors v_n, then the approximation is exact and the above relation holds with equality. In particular, the set of vectors must form a *complete basis*, meaning that any vector u can be written as a weighted sum of the vectors v_n in the set.

Example (vectors). For finite-dimensional vectors, it is trivial to write them as a weighted sum of basis vectors. For example,

$$\begin{bmatrix} 3 \\ 5 \\ -1 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 5 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + (-1) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

This is called the *standard basis* and is how we typically interpret a vector. When we say u_2, for example, what we really mean is the coefficient c_2 of the vector u using the standard basis vector $v_2 = (0, 1, 0)$. But we can also represent the vector using a different set of basis vectors, such as

$$\begin{bmatrix} 3 \\ 5 \\ -1 \end{bmatrix} = (1) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + (-1) \begin{bmatrix} -2 \\ 1 \\ 4 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}$$

Note that these vectors are *not* orthogonal, so the above formula for the coefficients does not apply.

The *Fourier series* represents a periodic signal using a complete orthogonal set of basis vectors that consist of sinusoidal signals at evenly-spaced frequencies.

Parseval's theorem

Parseval's theorem states that the "size" of a vector is the same as the "size" of its coefficients in a complete orthogonal basis. That is,

$$\|u\|^2 \quad = \quad \sum_n \frac{|\langle u, v_n\rangle|^2}{\|v_n\|^2}$$

The left-hand side is the squared norm (or energy) of the vector u, while the right-hand side is the sum of all of the coefficients of its representation in terms of the basis vectors $v_1, v_2, v_3, \ldots$. Therefore, the n^{th} term in the summation is proportional to the amount the vector v_n that is in the vector u.

$$\frac{|\langle u, v_n\rangle|^2}{\|v_n\|^2} \quad \propto \quad \text{amount of } v_n \text{ in } u$$

4.3 Continuous time

The continuous-time Fourier series represents a periodic signal in terms of sinusoids at evenly-spaced frequencies. There are several forms of the Fourier series representation for a periodic signal, depending on what signals are used for the basis vectors. We will first use both sines and cosines, then use a more compact representation using only cosines, and then another representation using complex exponentials (which generalize sinusoids).

Inner product

For periodic continuous-time signals with period T, the inner product between two periodic signals is

$$\langle u, v\rangle = \int_T u(t)^* \, v(t) \, \mathrm{d}t$$

where the notation $\int_T$ denotes an integral over any interval of length T (which is one period), and $(\cdot)^*$ denotes the complex conjugate (for complex signals). The squared-norm of a periodic signal is the inner product of the signal with itself,

$$\|u\|^2 = \langle u, u\rangle = \int_T |u(t)|^2 \, \mathrm{d}t$$

which is the signal power (or average energy).

Harmonics

For continuous-time perioidic signals with period $T > 0$, the *fundamental frequency* is

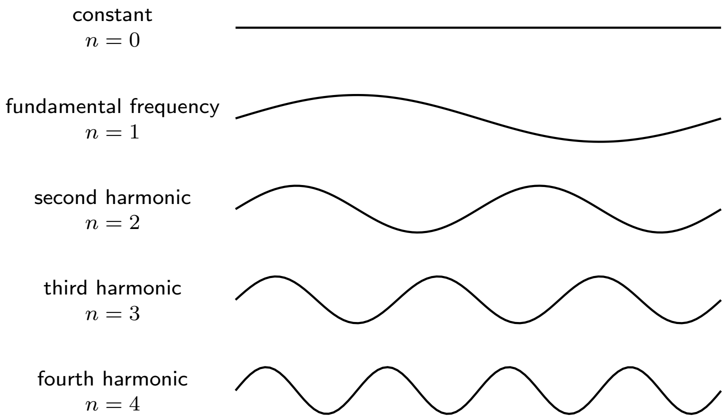
$$\omega_0 = \frac{2\pi}{T}$$

which has units of radians per second. A sinusoid at the fundamental frequency has period T and therefore oscillates once per period of the signal. This is the "slowest" frequency that is useful in representing a periodic signal with period T.

The Fourier series representation of a signal will use a sinusoid at the fundamental frequency along with its *harmonics*, which are sinusoids whose frequencies are multiplies of the fundamental frequency. The set of harmonics is the infinite set of signals

$$\{1, \ \cos(\omega_0 t), \ \cos(2\omega_0 t), \ldots, \sin(\omega_0 t), \ \sin(2\omega_0 t), \ldots\}$$

Each harmonic is periodic with period T as illustrated below.



Trigonometric form

The set of harmonically related sinusoids are both orthogonal and complete, so any periodic signal $x(t)$ with period T can be written as the weighted sum

$$x(t) = a_0 + 2\sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)$$

This is the *trigonometric form* of the Fourier series representation of the signal. The coefficient a_0 is the constant portion of the signal, and the coefficients a_n and b_n are the amount of the sinusoids $\cos(n\omega_0)$ and $\sin(n\omega_0 t)$ in the signal. (The factor of two in front of the summation is just for convenience.)

For sinusoidal signals, the power is

$$\|\cos(n\omega_0 t)\|^2 = \int_T \cos^2(n\omega_0 t)\,\mathrm{d}t = \begin{cases} T/2 & \text{if } n \neq 0 \\ T & \text{if } n = 0 \end{cases}$$

We can use these expressions for the signal power in our general expression for the Fourier coefficients. For instance, the coefficient for the constant term is

$$a_0 = \frac{\langle x(t), 1\rangle}{\|1\|^2} = \frac{1}{T}\int_T x(t)\,\mathrm{d}t$$

where 1 denotes the constant signal with unit value. Note that a_0 is simply the average value of the signal. Similarly, a general coefficient a_n for a cosine term is

$$a_n = \frac{\langle x(t), 2\cos(n\omega_0 t)\rangle}{\|2\cos(n\omega_0 t)\|^2} = \frac{1}{T}\int_T x(t)\cos(n\omega_0 t)\,\mathrm{d}t.$$

Note that this expression is also valid when $n = 0$ since $\cos(0) = 1$.

The trigonometric Fourier series representation of a periodic signal $x(t)$ with period T is the weighted sum of sinuosoids

$$x(t) = a_0 + 2\sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)$$

where the Fourier series coefficients are

$$a_n = \frac{1}{T}\int_T x(t)\, \cos(n\omega_0 t)\,\mathrm{d}t \qquad n = 0, 1, 2, 3, \ldots$$

$$b_n = \frac{1}{T}\int_T x(t)\, \sin(n\omega_0 t)\,\mathrm{d}t \qquad n = 1, 2, 3, \ldots$$

and $\omega_0 = 2\pi/T$ is the fundamental frequency.

Compact trigonometric form

The trigonometric form contains both sine and cosine terms of the same frequency. We can combine these terms to obtain a single shifted sinusoid using the trigonometric identity

$$a\cos\theta + b\sin\theta = c\cos(\theta+\phi) \qquad \text{where} \qquad c = \sqrt{a^2+b^2} \quad \text{and} \quad \phi = \tan^{-1}\left(-\frac{b_n}{a_n}\right).$$

Applying this identity to the trigonometric Fourier series, we obtain the following.

The compact trigonometric Fourier series representation of a periodic signal $x(t)$ with period T is the weighted sum of shifted sinuosoids

$$x(t) = a_0 + 2\sum_{n=1}^{\infty} c_n \cos(n\omega_0 t + \theta_n)$$

where the Fourier series coefficients c_n and θ_n can be obtained from the coefficients a_n and b_n as

$$c_n = \sqrt{a_n^2 + b_n^2} \qquad \text{and} \qquad \theta_n = \tan^{-1}\left(-\frac{b_n}{a_n}\right).$$

Exponential form

Instead of sinusoids, the Fourier series can also be expressed in terms of complex exponential signals. Here, the basis vectors are signals of the form $e^{jn\omega_0 t}$ for $n = 0, \pm 1, \pm 2, \ldots$. These vectors also form a complete orthogonal basis, where the inner product between two complex exponentials is

$$\langle e^{jn\omega_0 t}, e^{jm\omega_0 t}\rangle = \int_T \left(e^{jn\omega_0 t}\right)^* \left(e^{jm\omega_0 t}\right)\mathrm{d}t = \int_T e^{j(m-n)\omega_0 t}\,\mathrm{d}t = \begin{cases} 0 & \text{if } m \neq n \\ T & \text{if } m = n \end{cases}$$

Using these expressions with our general formula for the Fourier series coefficients, we obtain the following.

The exponential Fourier series representation of a periodic signal $x(t)$ with period T is the weighted sum of complex exponentials

$$x(t) = \sum_{n=-\infty}^{\infty} d_n \, e^{jn\omega_0 t}$$

where the Fourier series coefficients d_n are

$$d_n = \frac{1}{T} \int_T x(t) \, e^{-jn\omega_0 t} \, \mathrm{d}t.$$

In both trigonometric Fourier series representations, the coefficients a_n, b_n, c_n, and θ_n are all real numbers defined for nonnegative integers n. In contrast, the coefficients d_n are *complex* numbers (in general) that are defined for *all* integers n.

To find the relationship between the trigonometric and exponential Fourier series representations, we can rewrite the trigonometric Fourier series as follows:

$$\begin{aligned} x(t) &= a_0 + 2 \sum_{n=1}^{\infty} c_n \cos(n\omega_0 t + \theta_n) \\ &= a_0 + 2 \sum_{n=1}^{\infty} c_n \frac{e^{j(n\omega_0 t + \theta_n)} + e^{-j(n\omega_0 t + \theta_n)}}{2} \\ &= a_0 + \sum_{n=1}^{\infty} c_n \, e^{j\theta_n} e^{jn\omega_0 t} + c_n \, e^{-j\theta_n} e^{-jn\omega_0 t} \end{aligned}$$

Comparing this to the exponential Fourier series representation and using the relationship between (a_n, b_n) and (c_n, θ_n), we obtain the following relationships.

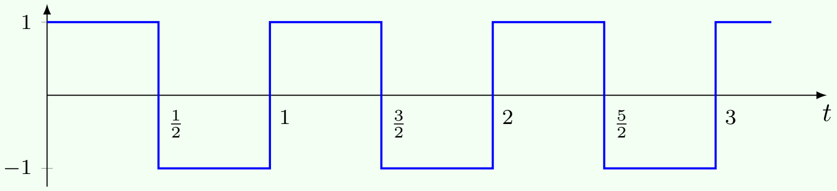
The exponential Fourier series coefficients are related to the trigonometric Fourier coefficients as follows:

$$d_n = \begin{cases} a_0 & n = 0 \\ c_n \, e^{j\theta_n} = a_n - jb_n & n > 0 \\ c_n \, e^{-j\theta_n} = a_n + jb_n & n < 0 \end{cases}$$

We can also solve for the trigonometric Fourier series coefficients in terms of the exponential coefficients:

$$a_n = \mathrm{Re}(d_n) \qquad b_n = -\mathrm{Im}(d_n) \qquad c_n = |d_n| \qquad \theta_n = \angle d_n$$

Example (square wave). Find the Fourier series representation of the following square wave $x(t)$.



The signal is periodic with fundamental period $T = 1$, so the fundamental frequency is $\omega_0 = 2\pi$. We will first compute the exponential Fourier series coefficients, which we can then use to find the trigonometric coefficients. The n^{th} coefficient is

$$d_n = \frac{1}{T}\int_0^1 x(t)\, e^{-jn\omega_0 t}\, \mathrm{d}t = \int_0^{1/2} e^{-j2\pi nt}\, \mathrm{d}t + \int_{1/2}^{1} (-1)\, e^{-j2\pi nt}\, \mathrm{d}t$$

where we split the integral up into two pieces since $x(t) = 1$ on the interval $[0, \frac{1}{2})$ and $x(t) = -1$ on the interval $[\frac{1}{2}, 1)$. Both integrals are of the form

$$\int_a^b e^{-j2\pi nt}\, \mathrm{d}t = \frac{1}{-j2\pi n} e^{-j2\pi nt}\Big|_{t=a}^{b} = \frac{e^{-j2\pi na} - e^{-j2\pi nb}}{j2\pi n}$$

Substituting this into the expression for the coefficients, we obtain

$$d_n = \frac{1 - e^{-j\pi n}}{j2\pi n} - \frac{e^{-j\pi n} - e^{-j2\pi n}}{j2\pi n}$$

We can simplify the expression using that $e^{-j2\pi n} = 1$ for any integer n to obtain

$$d_n = \frac{1 - e^{-j\pi n}}{j\pi n}$$

This expression is valid for $n \neq 0$, but we must then compute the coefficient for the constant term separately since setting $n = 0$ makes the denominator zero. Using the general expression for d_0, we have

$$d_0 = \frac{1}{T}\int_T x(t)\, \mathrm{d}t = 0$$

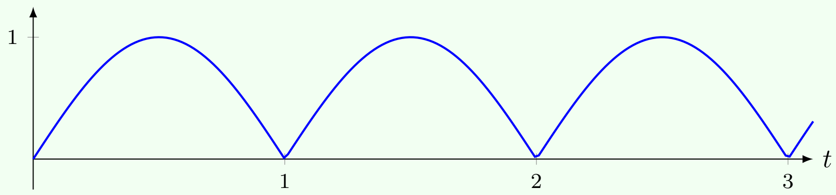
since the average value of the signal over one period is zero. To find the trigonometric coefficients, we need to express d_n is rectangular and polar form. To do so, note that $e^{-j\pi n}$ is positive one if n is even and negative one if n is odd. So all of the even coefficients are actually zero, and the odd coefficients are simply $d_n = \frac{2}{j\pi n}$ for n odd. Using the relationship between the exponential and trigonometric Fourier series coefficients, each of the coefficients are

$$a_n = 0 \qquad d_n = \begin{cases} 0 & n \text{ even} \\ \frac{2}{j\pi n} & n \text{ odd} \end{cases} \qquad b_n = c_n = \begin{cases} 0 & n \text{ even} \\ \frac{2}{\pi n} & n \text{ odd} \end{cases} \qquad \theta_n = \begin{cases} 0 & n \text{ even} \\ -\frac{\pi}{2} & n \text{ odd} \end{cases}$$

The Fourier series representation of the square wave in trigonometric form is

$$x(t) = \sum_{\substack{n \geq 1 \\ n \text{ odd}}} \frac{4}{\pi n} \sin(2\pi nt)$$

Example (rectified sinusoid)**.** Find the Fourier series representation of the rectified sinusoid $x(t) = |\sin(\pi t)|$.



The signal is periodic with period $T = 1$, so the fundamental frequency is $\omega_0 = 2\pi$. The exponential Fourier series coefficients are

$$d_n = \frac{1}{T}\int_0^1 x(t)\, e^{-jn\omega_0 t}\, \mathrm{d}t = \int_0^1 \sin(\pi t)\, e^{-j2\pi n t}\, \mathrm{d}t$$

To compute the integral, we will use the trigonometric identity $\sin\theta = \frac{1}{j2}(e^{j\theta} - e^{-j\theta})$. Applying this to the integrand,

$$d_n = \frac{1}{j2}\int_0^1 \left(e^{j\pi t} - e^{-j\pi t}\right) e^{-j2\pi n t}\, \mathrm{d}t = \frac{1}{j2}\int_0^1 \left(e^{j\pi(1-2n)t} - e^{-j\pi(1+2n)t}\right) \mathrm{d}t$$

Evaluating each of the integrals of an exponential gives

$$d_n = \frac{1}{j2}\left(\frac{1}{j\pi(1-2n)} e^{j\pi(1-2n)t}\bigg|_{t=0}^{1} - \frac{1}{-j\pi(1+2n)} e^{-j\pi(1+2n)t}\bigg|_{t=0}^{1}\right)$$

Applying the limits and using the fact that $j^2 = -1$, we obtain

$$d_n = \frac{e^{j\pi(1-2n)} - 1}{-2\pi(1-2n)} - \frac{e^{-j\pi(1+2n)} - 1}{2\pi(1+2n)}$$

Using that $e^{j2\pi n} = 1$ for all integers n and $e^{\pm j\pi} = -1$, this simplifies to

$$d_n = \frac{1}{\pi\,(1-2n)} + \frac{1}{\pi\,(1+2n)}$$

We can now get a common denominator to obtain the simple expression

$$d_n = \frac{2}{\pi\,(1-4n^2)}$$

While the coefficients d_n may be complex in general, they are real numbers in this case. The compact trigonometric Fourier series coefficients are

$$a_n = \mathrm{Re}(d_n) = \frac{2}{\pi\,(1-4n^2)} \qquad \text{and} \qquad b_n = -\mathrm{Im}(d_n) = 0.$$

(We could also use the standard trigonometric form, but then we would have $c_0 = a_n$ and $\theta_0 = 0$, but $c_n = -a_n$ and $\theta_n = \pi$ for $n \geq 1$, since in this case d_n is negative. In this case, it is simpler to use the compact trigonometric form since d_n is real.) Therefore, the Fourier series representation of the rectified sinusoid in compact trigonometric form is

$$|\sin(\pi t)| \quad = \quad \frac{2}{\pi} + \sum_{n=1}^{\infty} \frac{4}{\pi\,(1-4n^2)} \cos(2\pi n t)$$

The total power in the signal is

$$\|\sin(\pi t)\|^2 = \int_0^1 \sin(\pi t)^2\, \mathrm{d}t = \frac{1}{2}.$$

4.4 Discrete time

The discrete-time Fourier series, also known as the *discrete Fourier transform (DFT)*, represents a periodic discrete-time signal as a weighted summation of complex exponentials. While we could also use sinusoids as in the continuous-time case, complex exponentials are easier to work with and are much more common in the discrete-time case.

Consider a discrete-time signal $x(k)$ that is periodic with period $N > 0$. Define the *fundamental frequency* as

$$\theta_0 = \frac{2\pi}{N}$$

which has units of radians. The set of *harmonics* corresponding to a fundamental frequency θ_0 is the set of signals

$$\{1,\ e^{\pm j\theta_0 k},\ e^{\pm j2\theta_0 k},\ e^{\pm j3\theta_0 k}, \ldots\}$$

In the continuous-time case there were an infinite number of harmonics. In discrete time, however, there are only a finite number of distinct harmonics. This is due to the fact that

$$e^{j(\theta_0 \pm 2\pi)k} = e^{j\theta_0 k}\, e^{\pm j2\pi k} = e^{j\theta_0 k}$$

The consequence of this is that the N^{th} harmonic is equal to the zero harmonic, the $(N+1)^{\text{th}}$ harmonic is equal to the first harmonic, and so on. In general, the $(N+r)^{\text{th}}$ harmonic is equal to the r^{th} harmonic. Therefore, the set of harmonics in discrete time is the finite set of N complex exponential signals

$$\{1,\ e^{j\theta_0 k},\ e^{j2\theta_0 k}, \ldots, e^{j(N-1)\theta_0 k}\}$$

The discrete-time Fourier series represents a periodic discrete-time signal as a weighted summation of these complex exponential basis functions.

$$x(k) = \sum_{n=\langle N\rangle} c_n\, e^{jn\theta_0 k}$$

where the notation $\sum_{k=\langle N\rangle}$ denotes a summation over any sequence of length N (that is, one period), c_n are the (complex) Fourier series coefficients, and $\theta_0 = 2\pi/N$ is the fundamental frequency. To find the coefficients, we can use the general expression for the coefficients in this basis. Here, the inner product between two periodic signals is

$$\langle u, v\rangle = \sum_{k=\langle N\rangle} u(k)^*\, v(k)$$

where $(\cdot)^*$ denotes the complex conjugate (for complex signals). The squared-norm of a periodic signal is

$$\|u\|^2 = \langle u, u\rangle = \sum_{k=\langle N\rangle} |u(k)|^2$$

which is the signal power. Just as in the continuous-time case, any two distinct discrete-time complex exponentials are orthogonal, meaning that their inner product is zero.

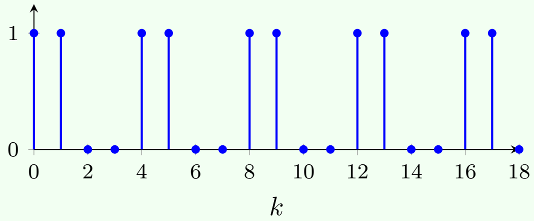
$$\langle e^{jn\theta_0 k}, e^{jm\theta_0 k}\rangle = \sum_{k=\langle N\rangle} \left(e^{jn\theta_0 k}\right)^* \left(e^{jm\theta_0 k}\right) = \sum_{k=\langle N\rangle} e^{j(m-n)\theta_0 k} = \begin{cases} 0 & \text{if } m \neq n \\ N & \text{if } m = n \end{cases}$$

Therefore, the discrete-time Fourier series coefficients are

$$c_n = \frac{1}{N} \sum_{k=\langle N\rangle} x(k)\, e^{-jn\theta_0 k}$$

The coefficients c_n are complex numbers in general and are called the *spectrum* of the signal.

Example (square wave)**.** Find the Fourier series representation of the discrete-time square wave $x(k)$ shown below.



The signal is periodic with period $N = 4$, so the fundamental frequency is $\theta_0 = \pi/2$. When computing the Fourier series coefficients, we can sum over any sequence of length $N = 4$. Let us choose the sequence $\{0, 1, 2, 3\}$. The coefficients are then

$$c_n = \frac{1}{4}\sum_{k=0}^{3} x(k)\, e^{-jnk\pi/2}$$

Since $x(k)$ is zero for $k = 2$ and $k = 3$, the summation simplifies to

$$c_n = \frac{1}{4}\big(1 + e^{-jn\pi/2}\big)$$

Specifically, the coefficients are

$$c_0 = \frac{1}{2} \qquad c_1 = \frac{1-j}{2} \qquad c_2 = 0 \qquad c_3 = \frac{1+j}{2}$$

Therefore, the Fourier series representation of the discrete-time square wave is

$$x(k) = \sum_{n=0}^{3} c_n\, e^{jnk\pi/2} = \frac{1}{2} + \frac{1-j}{2} e^{jk\pi/2} + \frac{1+j}{2} e^{jk3\pi/2}$$

While this expression makes the signal appear to be complex, it is actually real. We can see this by simplifying the expression. Using that $e^{jk3\pi/2} = e^{-jk\pi/2}$, we can rewrite this as

$$x(k) = \frac{1}{2} + \frac{1}{2}\Big(e^{jk\pi/2} + e^{-jk\pi/2}\Big) + \frac{1}{j2}\Big(e^{jk\pi/2} - e^{-jk\pi/2}\Big)$$

Now using the relationships $\cos\theta = \frac{1}{2}(e^{j\theta} + e^{-j\theta})$ and $\sin\theta = \frac{1}{j2}(e^{j\theta} - e^{-j\theta})$, we have that

$$x(k) = \tfrac{1}{2} + \cos\big(\tfrac{k\pi}{2}\big) + \sin\big(\tfrac{k\pi}{2}\big)$$

which emphasizes that the signal is in fact real.

Discrete-time Fourier series as matrix multiplication. The discrete-time Fourier series transforms a sequence of N points (the values of the periodic signal) to another sequence of N points (the Fourier series coefficients) via the formula

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} x(k)\, e^{-jn\theta_0 k}$$

We can equivalently express this relationship as the multiplication of a matrix with a vector. In particular, the Fourier series coefficients are given by

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{N-1} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & e^{-j\theta_0} & \dots & e^{-j(N-1)\theta_0} \\ 1 & e^{-j2\theta_0} & \dots & e^{-j2(N-1)\theta_0} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-j(N-1)\theta_0} & \dots & e^{-j(N-1)(N-1)\theta_0} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(N-1) \end{bmatrix}$$

Similarly, the discrete-time Fourier series represents the signal as a weighted sum of complex exponentials via the formula

$$x(k) = \sum_{n=0}^{N-1} c_n \, e^{jn\theta_0 k}$$

We can express this relationship as the following matrix-vector multiplication:

$$\begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(N-1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & e^{j\theta_0} & \dots & e^{j(N-1)\theta_0} \\ 1 & e^{j2\theta_0} & \dots & e^{j2(N-1)\theta_0} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j(N-1)\theta_0} & \dots & e^{j(N-1)(N-1)\theta_0} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{N-1} \end{bmatrix}$$

Example (square wave — continued)**.** Continuing our previous example, the Fourier series coefficients for the square wave can be represented using matrix multiplication as

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix}$$

and the values of the discrete-time signal can be recovered from the Fourier series coefficients using

$$\begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

Note that the two matrices are very similar. In fact, they are *conjugate transposes* of each other. Also, the eigenvalues of both matrices are $\{2, -2, -j2, 2\}$, which all have magnitude two.

5

Fourier Transform

5.1 Continuous time

The frequency response is the transfer function evaluated on the imaginary axis, where the transfer function is the Laplace transform of the impulse response. This is called the *Fourier tranform*.

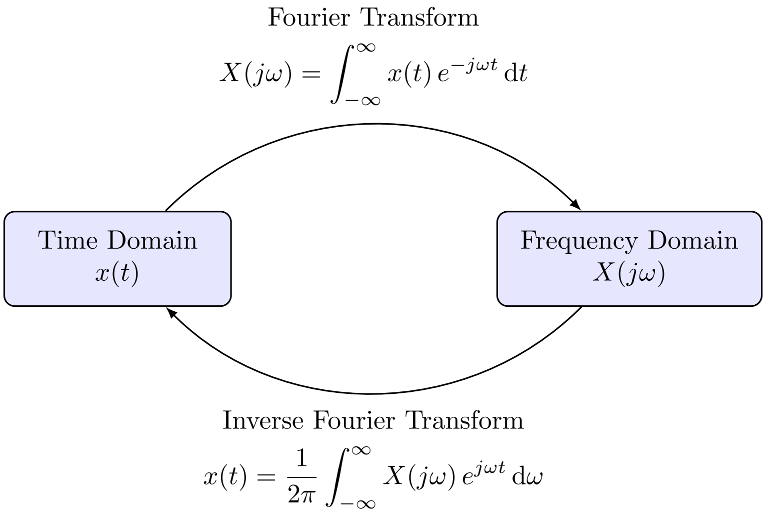
Definition (Continuous-time Fourier transform)**.** The Fourier transform of a continuous-time signal $x(t)$ is the complex function

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, \mathrm{d}t$$

if the integral exists. The inverse Fourier transform is

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega)\, e^{j\omega t}\, \mathrm{d}\omega.$$

For a time-domain signal $x(t)$, its Fourier transform $X(j\omega)$ is a frequency-domain representation of the signal. Since the Fourier transform is invertible, this frequency-domain representation of the signal is equivalent to its time-domain representation. This relationship is illustrated below.


Fourier Transform
$X(j\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, \mathrm{d}t$
Time Domain
$x(t)$
Frequency Domain
$X(j\omega)$
Inverse Fourier Transform
$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega)\, e^{j\omega t}\, \mathrm{d}\omega$

Comments

- The signal $x(t)$ need not be periodic.
- The Fourier transform $X(j\omega)$ is in general complex.
- The magnitude of $X(j\omega)$ is the amount of energy of $x(t)$ at frequency ω.
- The factor of $1/(2\pi)$ is sometimes in first equation instead.
- For signals that start at time zero ($x(t) = 0$ for $t < 0$), the Fourier transform is the Laplace transform evaluated on the imaginary axis.
$$X(j\omega) = X(s)\Big|_{s=j\omega}$$
- The Fourier transform exists if the imaginary axis is in the region of convergence of the Laplace transform.
- While the formula for the inverse Laplace transform requires complex integration, the inverse Fourier transform is a much simpler integral that is quite similar to the Fourier transform itself.
- Parseval's theorem states that the energy of the signal is the same in both the time domain and frequency domain.
$$\int_{-\infty}^{\infty} |x(t)|^2 \,\mathrm{d}t \quad = \quad \frac{1}{2\pi}\int_{-\infty}^{\infty} |X(j\omega)|^2 \,\mathrm{d}\omega$$
- For a stable continuous-time LTI system, the frequency response is the Fourier transform of the impulse response.

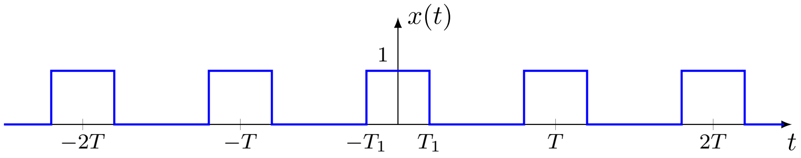
Motivation

We now motivate the definition of the Fourier transform by deriving it from the Fourier series of a periodic signal in the limit as the period goes to infinity.

Recall that any signal that is represented by a Fourier series is periodic since

$$\begin{aligned} x(t+T) &= \sum_{k=-\infty}^{\infty} a_k\, e^{jk\omega_0(t+T)} \\ &= \sum_{k=-\infty}^{\infty} a_k\, e^{jk\omega_0 t}\, e^{j2\pi k} && \text{since } T = 2\pi/\omega_0 \\ &= \sum_{k=-\infty}^{\infty} a_k\, e^{jk\omega_0 t} \\ &= x(t) \end{aligned}$$

Therefore, the Fourier series cannot represent *non-periodic* signals. To construct a frequency-domain representation of non-periodic signals, let's first consider the following periodic square wave.



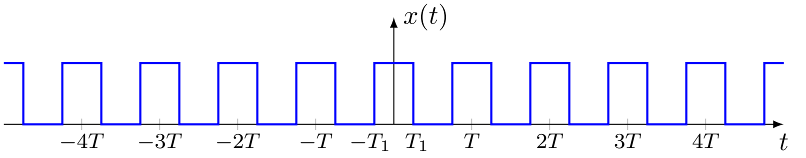
The complex exponential Fourier series coefficients of this signal are

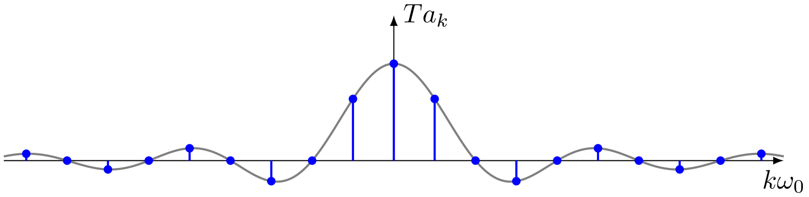
$$a_k = \frac{\sin(k\omega_0 T_1)}{k\pi}.$$

We can interpret the (scaled) Fourier series coefficients as samples of an envelope function:

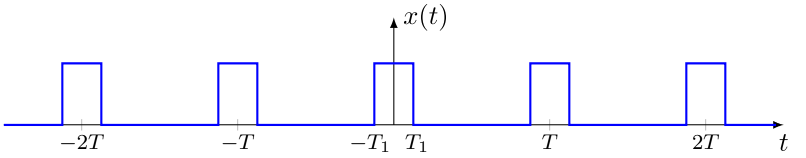
$$Ta_k = \left.\frac{2\sin(\omega T_1)}{\omega}\right|_{\omega=k\omega_0}$$

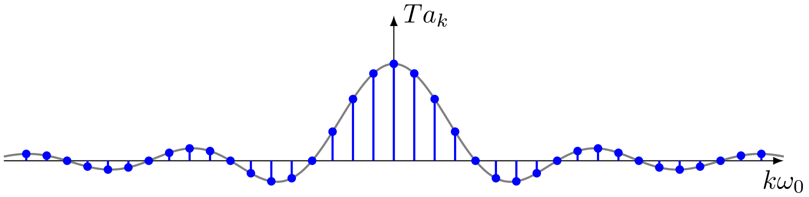
The scaled Fourier series coefficients Ta_k are equally spaced samples of the envelope $2\sin(\omega T_1)/\omega$. Moreover, for fixed T_1, the envelope is independent of the period T.



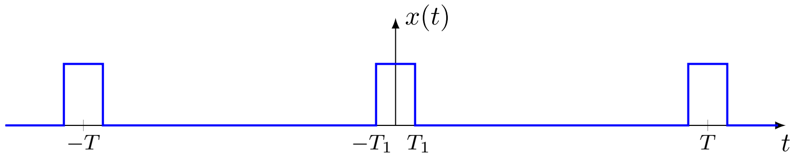


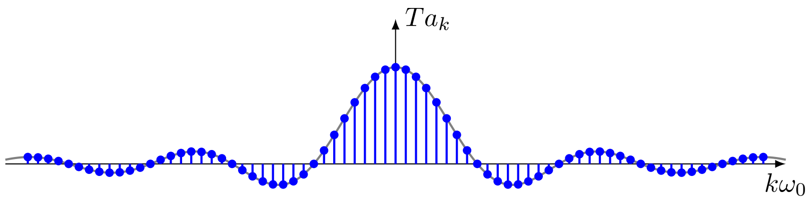
Now consider how the envelope changes as we double the period. Since the (scaled) Fourier series coefficients are independent of the period T, the envelope remains the same. The only thing that changes is the interval at which the coefficients are sampled from the envelope.



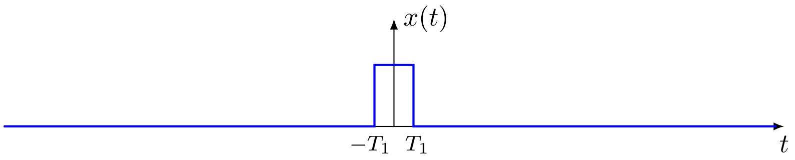


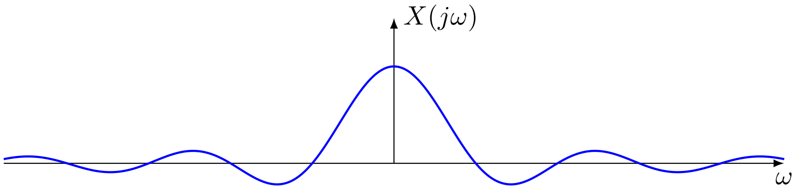
Doubling the period again, the Fourier series coefficients are sampled twice as fast from the envelope.





In the limit as the period T goes to infinity, the signal $x(t)$ becomes non-periodic and we obtain the smooth envelope. The limiting envelope of Ta_k is the *Fourier transform* of $x(t)$, which is denoted by $X(j\omega)$ where the continuous-time frequency is $\omega = k\omega_0$.



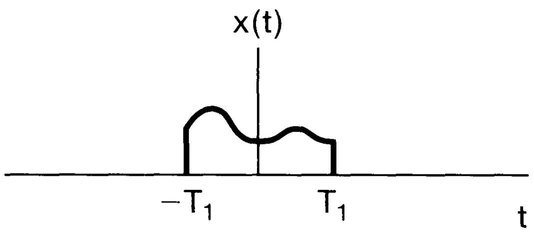


Therefore, the Fourier transform of the pulse that has a value of one on the interval $[-T_1, T_1]$ and is zero otherwise is

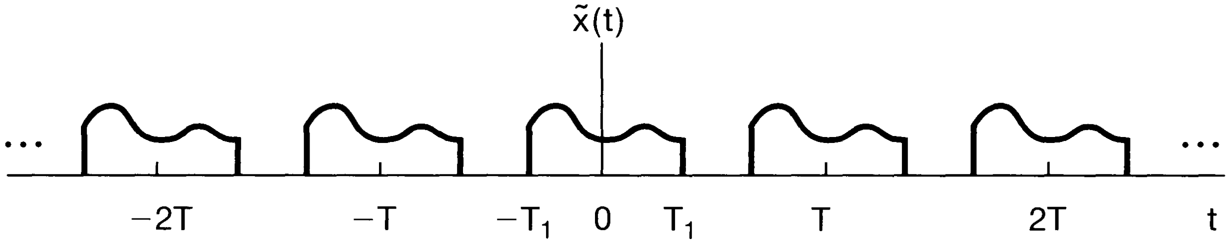
$$X(j\omega) = \frac{2\sin(\omega T_1)}{\omega}.$$

This is called a *sinc* function and is commonly used in signal processing.

Now consider a general non-periodic signal $x(t)$ that is zero outside an interval $[-T_1, T_1]$. An illustrative example of such a signal is as follows.



Let $\tilde{x}(t)$ be this signal repeated with period $T > 2T_1$ (so that there is no overlap). Note that $\tilde{x}(t)$ is periodic with period T, and $\tilde{x}(t)$ converges to $x(t)$ in the limit as $T \to \infty$.



The signal $\tilde{x}(t)$ is periodic, and therefore has the (exponential) Fourier series

$$\tilde{x}(t) = \sum_{k=-\infty}^{\infty} a_k\, e^{jk\omega_0 t} \qquad \xleftrightarrow{\text{FS}} \qquad a_k = \frac{1}{T}\int_T \tilde{x}(t)\, e^{-jk\omega_0 t}\, \mathrm{d}t.$$

Similar to the square wave, define the continuous-time frequency as $\omega = k\omega_0$, and then define the Fourier transform as the envelope of the (scaled) Fourier series coefficients,

$$X(j\omega) = Ta_k = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, \mathrm{d}t$$

(since $x(t) = \tilde{x}(t)$ for $|t| < T/2$ and is zero outside this interval).

To obtain the inverse Fourier transform, substitute this into the Fourier series representation for $\tilde{x}(t)$,

$$\begin{aligned}\tilde{x}(t) &= \sum_{k=-\infty}^{\infty} a_k\, e^{jk\omega_0 t} \\ &= \sum_{k=-\infty}^{\infty} \frac{1}{T} X(j\omega)\, e^{j\omega t} \\ &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} X(j\omega)\, e^{j\omega t}\, \omega_0\end{aligned}$$

Taking the limit as $\omega_0 \to 0$, the sum becomes an integral and $\tilde{x}(t)$ becomes $x(t)$:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(j\omega)\, e^{j\omega t}\, \mathrm{d}\omega.$$

5.2 Discrete time

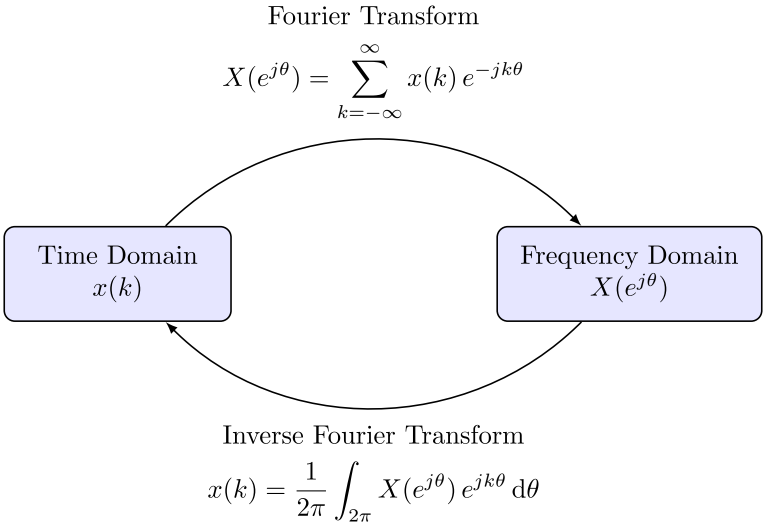
Definition (Discrete-time Fourier transform). The Fourier transform of a discrete-time signal $x(k)$ is the complex function

$$X(e^{j\theta}) = \sum_{k=-\infty}^{\infty} x(k)\, e^{-j\theta k}$$

if the summation exists. The inverse Fourier transform is

$$x(k) = \frac{1}{2\pi}\int_{2\pi} X(e^{j\theta})\, e^{j\theta k}\, \mathrm{d}\theta.$$

For a time-domain signal $x(k)$, its Fourier transform $X(e^{j\theta})$ is a frequency-domain representation of the signal. Since the Fourier transform is invertible, this frequency-domain representation of the signal is equivalent to its time-domain representation. This relationship is illustrated below.



Fourier Transform

$$X(e^{j\theta}) = \sum_{k=-\infty}^{\infty} x(k)\, e^{-jk\theta}$$

Inverse Fourier Transform

$$x(k) = \frac{1}{2\pi} \int_{2\pi} X(e^{j\theta})\, e^{jk\theta}\, \mathrm{d}\theta$$

Comments

- The signal $x(k)$ need not be periodic.
- The Fourier transform $X(e^{j\theta})$ is in general complex.
- The magnitude of $X(e^{j\theta})$ is the amount of energy of $x(k)$ at frequency θ.
- The factor of $1/(2\pi)$ is sometimes in first equation instead.
- The magnitude of $X(e^{j\theta})$ is the amount of frequency θ that is contained in the signal $x(k)$.
- For signals that start at time zero ($x(k) = 0$ for $k < 0$), the Fourier transform is the z-transform evaluated on the unit circle.
$$X(e^{j\theta}) = X(z)\Big|_{z=e^{j\theta}}$$
- The Fourier transform exists if the unit circle is in the region of convergence of the z-transform.
- While the formula for the inverse z-transform requires complex integration, the inverse Fourier transform is a much simpler integral that is quite similar to the Fourier transform itself.
- Parseval's theorem states that the energy of the signal is the same in both the time domain and frequency domain.
$$\sum_{k=-\infty}^{\infty} |x(k)|^2 \quad = \quad \frac{1}{2\pi} \int_{2\pi} |X(e^{j\theta})|^2\, \mathrm{d}\theta$$
- For a stable discrete-time LTI system, the frequency response is the Fourier transform of the impulse response.

6

z-Transform

The z-transform is a useful tool in understanding LTI systems that interprets signals in terms of their *frequencies*. The z-transform is a generalization of the discrete Fourier series, and it is the discrete-time equivalent of the Laplace transform for continuous-time signals.

6.1 Definition

Definition. The z-transform of a discrete-time signal $x(k)$ is the complex function

$$X(z) = \sum_{k=0}^{\infty} x(k)\, z^{-k}$$

where z is any complex number for which the summation converges. The values of z for which the summation converges is the region of convergence (ROC).

Notation. We use lowercase letters for a time-domain signal such as $x(k)$, while we use the corresponding uppercase letter for its corresponding z-transform $X(z)$. We denote a signal and its z-transform as the pair

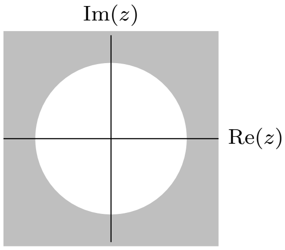
$$x(k) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad X(z)$$

■

Remark (Alternative definitions)**.** This definition of the z-transform ignores the signal for negative times: $X(z)$ does not depend on $x(k)$ for $k < 0$. This is sometimes called the *unilateral* z-transform. There is also a version where the summation starts at negative infinity, which is called the *bilateral* z-transform. Both versions of the z-transform are equivalent for signals that are zero for $k < 0$. Moreover, the unilateral z-transform will be useful in studying how the system response depends on the initial conditions. ■

Remark (Connection with Laplace transform)**.** You may have seen the Laplace transform when studying circuits or differential equations. The Laplace transform represents a *continuous-time* signal in terms of complex exponentials, while the z-transform represents a *discrete-time* signal in terms of complex exponentials. Almost all of the results for the z-transform have similar results for the Laplace transform. ■

The definition of the z-transform is in terms of an infinite summation, which may not converge for all values of the complex number z. It always converges, however, when the magnitude of z is sufficiently large. The set of complex numbers z for which the summation converges is called the *region of convergence*, or *ROC*. This is illustrated below, where the gray region is the region of convergence.



6.2 Computation

There are several ways to compute the z-transform of a discrete-time signal. The most straightforward way is to directly apply the definition, which involves computing an infinite summation. Alternatively, we can use a table of z-transform pairs along with properties of the z-transform to find the transforms of many signals. We now show how to compute the z-transform using each of these techniques.

Apply the definition

We now compute the z-transform for several simple signals by directly applying the definition. This requires computing an infinite summation in general. We will see a few cases in which this is quite simple.

Example (z-transform of an impulse signal)**.** The z-transform of an impulse $x(k) = \delta(k)$ is

$$X(z) = \sum_{k=0}^{\infty} \delta(k)\, z^{-k} = \delta(0)\, z^{-0} = 1$$

Therefore, the z-transform pair for an impulse signal is

$$\delta(k) \quad \xleftrightarrow{\mathcal{Z}} \quad 1$$

Example (z-transform of a finite-duration signal)**.** A finite-duration signal has the form

$$x(k) = \{x_0, x_1, x_2, \ldots, x_m, 0, 0, 0, \ldots\}$$

The signal is zero after some time m. The z-transform is the finite summation

$$X(z) = \sum_{k=0}^{\infty} x(k)\, z^{-k} = x_0 + x_1\, z^{-1} + x_2\, z^{-2} + \ldots + x_m\, z^{-m}$$

Therefore, the z-transform pair for a finite-duration signal is

$$\{x_0, x_1, x_2, \ldots, x_m, 0, 0, 0, \ldots\} \quad \xleftrightarrow{\mathcal{Z}} \quad \frac{x_0\, z^m + x_1\, z^{m-1} + \ldots + x_m}{z^m}$$

Example (z-transform of an exponential signal)**.** The z-transform of an exponential $x(k) = a^k\, u_s(k)$ is

$$X(z) = \sum_{k=0}^{\infty} a^k\, z^{-k} = \sum_{k=0}^{\infty} \left(\frac{a}{z}\right)^k = \frac{1}{1 - a/z} = \frac{z}{z-a}$$

where the summation converges only if $|a/z| < 1$. Therefore, the z-transform of an exponential signal is

$$a^k\, u_s(k) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad \frac{z}{z-a}$$

The region of convergence in this case is the set of z such that $|z| > |a|$.

6.3 Poles and zeros

The z-transform is usually a rational function of the complex number z, so it has the form

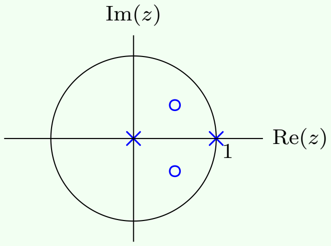
$$X(z) = \frac{N(z)}{D(z)}$$

where $N(z)$ and $D(z)$ are polynomials in the complex variable z.

- The *zeros* of $X(z)$ are the roots of the numerator polynomial $N(z)$.
- The *poles* of $X(z)$ are the roots of the denominator polynomial $D(z)$.

We can visualize the z-transform by plotting its poles and zeros in the complex plane. We use the symbol $\times$ to represent poles and the symbol $\circ$ to represent zeros.

Example (Pole-zero plot)**.** The pole-zero plot of a rational complex function $X(z)$ with zeros at $z = 0.5 \pm j0.5$ and poles at $z = 0$ and $z = 1$ is as follows.



Remark (Complex conjugates)**.** Since the coefficients of the z-transform are real numbers, the poles and zeros always appear in complex-conjugate pairs, meaning that if $a + jb$ is a pole or zero, then $a - jb$ is also a pole or zero. The pole-zero plot is therefore symmetric about the real axis. ■

Example. Find the z-transform and its associated poles, zeros, and region of convergence of the signal

$$x(k) = 1 + (0.5)^k \quad \text{for } k \geq 0.$$

Using the definition of the z-transform,

$$X(z) = \sum_{k=0}^{\infty} \left(1 + (0.5)^k\right) z^{-k}$$

Splitting this into two sums,

$$X(z) = \sum_{k=0}^{\infty} \left(\frac{1}{z}\right)^k + \left(\frac{0.5}{z}\right)^k$$

We can use the formula for a geometric series to evaluate the sums to obtain

$$X(z) = \frac{1}{1 - 1/z} + \frac{1}{1 - 0.5/z}$$

where the region of convergence of the first term is $|z| > 1$, and the region of convergence of the second term is $|z| > 0.5$. For $X(z)$ to exist, *both* summations must converge, which means the combined region of convergence is the intersection of the two individual ROCs, which is $|z| > 1$.

Multiplying the top and bottom by z in both terms, we get

$$X(z) = \frac{z}{z-1} + \frac{z}{z-0.5}$$

To find the poles and zeros, we must combine these terms into a single rational expression. Getting a common demoninator, we have

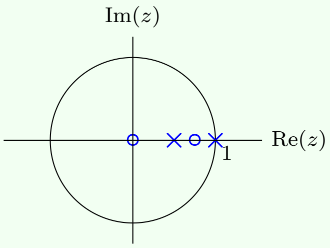
$$X(z) = \frac{z\,(z-0.5) + z\,(z-1)}{(z-1)(z-0.5)} = \frac{z\,(2z-1.5)}{(z-1)(z-0.5)}$$

The numerator and denominator polynomials are

$$N(z) = z\,(2z - 1.5) \qquad \text{and} \qquad D(z) = (z-1)(z-0.5).$$

The zeros are the values of z that make the numerator zero, which are $z = 0$ and $z = 0.75$. The poles are the values of z that make the denominator zero, which are $z = 1$ and $z = 0.5$.

The pole-zero plot for $X(z)$ is as follows.



6.4 Properties

The z-transform has many properties that are useful in simplifying calculations of the z-transform and its inverse.

Linearity

The z-transform maps a discrete-time signal $x(k)$ to its corresponding z-transform $X(z)$. This map is *linear*, meaning that the z-transform of a weighted sum of signals is the same weighted sum of their individual z-transforms.

Specifically, suppose we have two signals $x_1(k)$ and $x_2(k)$ and their corresponding z-transforms $X_1(z)$ and $X_2(z)$. Then for any scalars a and b,

$$a\,x_1(k) + b\,x_2(k) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad a\,X_1(z) + b\,X_2(z)$$

Example. As we have seen, the z-transform of an impulse is one, and the z-transform of a unit step is $\frac{z}{z-1}$. We can then use the linearity property to obtain the z-transform pair

$$2\,\delta(k) - u_s(k) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad 2 - \frac{z}{z-1}$$

Time shifting

Advancing a signal in time multiplies its z-transform by z.

$$x(k+1) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad z\,X(z) - z\,x(0)$$

Delaying a signal in time divides its z-transform by z.

$$x(k-1) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad \frac{1}{z}X(z) + x(-1)$$

Notice that the second term on the right in both cases involves the time-domain signal.

Example. We know that the z-transform of the exponential signal $x(k) = a^k\,u_s(k)$ is $X(z) = \frac{z}{z-a}$. We can then use the time shifting properties of the z-transform to obtain the additional pairs

$$a^{k+1}\,u_s(k+1) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad z \cdot \frac{z}{z-a} - z\,(1) = \frac{az}{z-a}$$

and

$$a^{k-1}\,u_s(k-1) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad \frac{1}{z} \cdot \frac{z}{z-a} + 0 = \frac{1}{z-a}$$

Multiplication by time

Multiplying a signal by k in time differentiates the z-transform.

$$k\,x(k) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad -z\,\frac{\mathrm{d}}{\mathrm{d}z}\Big(X(z)\Big)$$

Example. Using that the z-transform of a unit step is $\frac{z}{z-1}$, the z-transform of $k\,u_s(k)$ is

$$-z\,\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{z}{z-1}\right) = -z\,\frac{(1)(z-1)-z\,(1)}{(z-1)^2} = \frac{z}{(z-1)^2}$$

where we used the quotient rule to compute the derivative.

Initial-value theorem

In terms of its z-transform, the initial value of a signal is

$$x(0) \quad = \quad \lim_{z\to\infty} X(z) \quad \text{if the limit exists.}$$

Example.

$$\lim_{z\to\infty} \frac{z}{z-a} \quad = \quad 1 \quad = \quad a^k\,u_s(k)\Big|_{k=0}$$

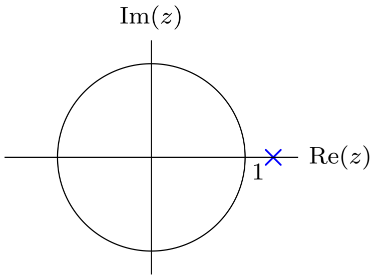
Final-value theorem

In terms of its z-transform, the final value of a signal is

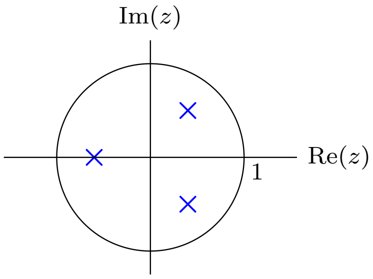
$$\lim_{k\to\infty} x(k) \quad = \quad \lim_{z\to 1}\,(z-1)\,X(z)$$

if both limits exist.

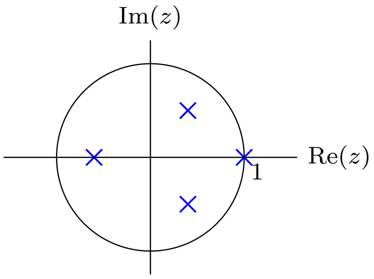
- The final value theorem only exists when all poles of $X(z)$ are strictly inside the unit circle except for possibly a single pole at $z=1$.
- If all poles of $X(z)$ are strictly inside the unit circle, then the limit is trivially zero.
- Therefore, the final value theorem only exists and is useful when all poles are strictly inside the unit circle except for a single pole at $z=1$.



The limit of $x(k)$ as $k\to\infty$ does not exist, so the FVT does not apply.



The limit of $x(k)$ as $k\to\infty$ is zero; the FVT applies but is not useful.



The FVT applies and is useful.

Example. The limit of $(0.5)^k\, u_s(k)$ is

$$\lim_{z\to 1}\,(z-1)\cdot\frac{z}{z-0.5}=0$$

The limit of $u_s(k)$ is

$$\lim_{z\to 1}\,(z-1)\cdot\frac{z}{z-1}=1$$

6.5 Inverse z-transform

The z-transform maps a discrete-time signal $x(k)$ to the complex function $X(z)$. The z-transform does not lose any information about the signal in that $X(z)$ contains all the same information as $x(k)$, just in a different form. Mathematically, this means that the z-transform is invertible, so we can always go back and forth between a discrete-time signal $x(k)$ and its z-transform $X(z)$. We now see how to compute the inverse.

There is a formula for the inverse z-transform, but we will never use it! For those that are interested, the formula is

$$x(k)=\frac{1}{j2\pi}\oint X(z)\,z^{k-1}\,\mathrm{d}z$$

where the contour integral is over a counterclockwise closed path in the complex plane that encircles the origin and is entirely in the region of convergence. While we could use this formula, it involves complex integration. Instead, we will use a method called partial fraction expansion that allows us to simply look up the z-transform from a table.

Partial fraction expansion

Partial fraction expansion (PFE) is a method for computing the inverse z-transform. The idea is to express $X(z)$ as a sum of simple terms and then use linearity to find $x(k)$. Here, *simple terms* are those that we either have in our table or can easily obtain using properties of the z-transform.

$$\begin{array}{ccccccc} X(z) & = & X_1(z) & + & X_2(z) & + & X_3(z) \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ x(k) & = & x_1(k) & + & x_2(k) & + & x_3(k) \end{array}$$

Example. Find the inverse z-transform of $X(z) = \dfrac{z-2}{z-1}$.

Let's try writing $X(z)$ as the sum

$$X(z) = \frac{z-2}{z-1} = A\frac{z}{z-1} + B\,(1)$$

where $\frac{z}{z-1}$ and 1 (a constant) are z-transforms in our table. If we can find coefficients A and B such that this equation holds for all z, then the inverse z-transform is

$$x(k) = Au_s(k) + B\delta(k)$$

To see if this is possible, multiply both sides by $z-1$ to obtain the polynomial equation

$$z - 2 = Az + B\,(z-1)$$

This is a polynomial equation in z. For the equation to hold for all z, the coefficients of each power of z much match. The constant terms give the equation $-2 = -B$, and the z terms give the equation $1 = A + B$. Solving this system of equations gives the coefficients $B = 2$ and $A = -1$. Therefore, the inverse z-transform is

$$x(k) = 2\delta(k) - u_s(k)$$

For the previous example, the partial fraction expansion consisted of the term $\frac{z}{z-1}$ due to the pole of $X(z)$ at $z = 1$ and a constant. In the next few examples we will see what form the partial fraction expansion should have in more complicated scenarios.

Example. Find the inverse z-transform of $X(z) = \dfrac{z^2+3z+5}{(z-1)(z-2)}$.

In this case $X(z)$ has two poles: $z = 1$ and $z = 2$. Therefore, the form of the partial fraction expansion is

$$X(z) = \frac{z^2+3z+5}{(z-1)(z-2)} = A\frac{z}{z-1} + B\frac{z}{z-2} + C$$

Now that we have the form of the partial fraction expansion, we can apply our general method. Multiplying both sides by the denominator of $X(z)$ gives the polynomial equation

$$z^2 + 3z + 5 = Az(z-2) + Bz(z-1) + C\,(z-1)(z-2)$$

Equating the coefficients of the powers of z produces the linear system of three equations:

$$\begin{aligned} 1 &= A + B + C && z^2 \text{ terms} \\ 3 &= -2A - B - 3C && z^1 \text{ terms} \\ 5 &= 2C && z^0 \text{ terms} \end{aligned}$$

The solution to this linear system of equations is $A = -9$, $B = 15/2$, and $C = 5/2$. From the table of z-transform pairs, the inverse z-transform of $z/(z-a)$ is $a^k\,u_s(k)$, and the inverse z-transform of the constant 1 is an impulse $\delta(k)$. Therefore, the inverse z-transform is

$$x(k) = -9u_s(k) + \frac{15}{2}\,(2)^k u_s(k) + \frac{5}{2}\delta(k)$$

Example. Find the inverse z-transform of $X(z) = \dfrac{z^3}{(z+1)(z-2)(z-1)}$.

$$x(k) = \left[-\frac{1}{6}(-1)^k + \frac{4}{3}(2)^k - \frac{1}{2}\right] u_s(k)$$

Example (Repeated roots)**.** Find the inverse z-transform of $X(z) = \dfrac{5z\,(z+1)}{(z+2)^2(z+3)}$.

Since the root at $z = 2$ is repeated twice, the partial fraction expansion has a term of the form $z/(z+2)^2$. This corresponds to the following entry in the table,

$$k\,2^{k-1}\,u_s(k) \quad \xleftrightarrow{\;z\;} \quad \frac{z}{(z+2)^2}$$

The form of the partial fraction expansion is then

$$X(z) = \frac{5z\,(z+1)}{(z+2)^2(z+3)} = \frac{Az}{z+2} + \frac{Bz}{(z+2)^2} + \frac{Cz}{z+3} + D$$

Multiplying both sides by the denominator of $X(z)$ yields the polynomial equation

$$5z\,(z+1) = Az(z+2)(z+3) + Bz(z+3) + Cz(z+2)^2 + D\,(z+2)^2(z+3)$$

Equating the coefficients of powers of z, we obtain the linear system of equations

$$\begin{aligned} 0 &= A + C + D && z^3 \text{ terms} \\ 5 &= 5A + B + 4C + 7D && z^2 \text{ terms} \\ 5 &= 6A + 3B + 4C + 16D && z^1 \text{ terms} \\ 0 &= 12D && z^0 \text{ terms} \end{aligned}$$

The coefficients are $A = 10$, $B = -5$, $C = -10$, and $D = 0$. Therefore, the inverse z-transform is

$$x(k) = \left[10\,(-2)^k - 5k\,(-2)^{k-1} - 10\,(-3)^k\right] u_s(k)$$

Example (Complex roots)**.** Find the inverse z-transform of $X(z) = \dfrac{z+1}{z^2+z+1}$.

The denominator of $X(z)$ does not factor, so we need to use the quadratic formula to find the poles:

$$z = -\frac{1}{2} \pm j\,\frac{\sqrt{3}}{2}$$

The magnitude of the poles is

$$|z| = \sqrt{\left(-\tfrac{1}{2}\right)^2 + \left(\tfrac{\sqrt{3}}{2}\right)^2} = 1$$

so the poles are on the unit circle in the complex plane. The angle of the poles are $\pm\theta$ where

$$\theta = \operatorname{atan2}\!\left(\tfrac{\sqrt{3}}{2}, -\tfrac{1}{2}\right) = \tfrac{2\pi}{3}$$

For complex roots on the unit circle at an angle θ, the relevant entries in the z-transform table are

$$\begin{aligned} \cos(k\theta)\,u_s(k) \quad &\overset{z}{\longleftrightarrow} \quad \frac{z\,(z-\cos\theta)}{z^2 - 2\cos\theta z + 1} = \frac{z\,(z+0.5)}{z^2+z+1} \\ \sin(k\theta)\,u_s(k) \quad &\overset{z}{\longleftrightarrow} \quad \frac{\sin\theta z}{z^2 - 2\cos\theta z + 1} = \frac{0.5\sqrt{3}z}{z^2+z+1} \end{aligned}$$

The form of the partial fraction expansion is then

$$X(z) = \frac{z+1}{z^2+z+1} = A\frac{z\,(z+0.5)}{z^2+z+1} + B\,\frac{0.5\sqrt{3}z}{z^2+z+1} + C$$

Multiplying both sides by the denominator, we obtain the polynomial equation

$$z + 1 = Az(z+0.5) + B\,0.5\sqrt{3}z + C\,(z^2+z+1)$$

Equating the coefficients of powers of z, we obtain the linear system of equations

$$\begin{aligned} 0 &= A + C && z^2 \text{ terms} \\ 1 &= 0.5A + 0.5\sqrt{3}B + C && z^1 \text{ terms} \\ 1 &= C && z^0 \text{ terms} \end{aligned}$$

The coefficients are $A = -1$, $B = \frac{1}{\sqrt{3}}$, and $C = 1$. Therefore, the inverse z-transform is

$$x(k) = \left[-\cos\!\left(k\tfrac{2\pi}{3}\right) + \tfrac{1}{\sqrt{3}}\sin\!\left(k\tfrac{2\pi}{3}\right)\right] u_s(k) + \delta(k)$$

A general method for finding the inverse z-transform of a rational function $X(z)$ is as follows:

a) Write out the form of the partial fraction expansion and set it equal to $X(z)$.

- The partial fraction expansion always includes a constant term.
- A pole at $z = a$ produces a term of the form $\frac{z}{z-a}$ in the partial fraction expansion.
- A pole at $z = a$ that is repeated m times produce terms of the form

$$\frac{z}{z-a}, \qquad \frac{z}{(z-a)^2}, \qquad \dots, \qquad \frac{z}{(z-a)^m}$$

- Complex conjugate poles with magnitude r and angle $\pm\theta$ produce terms of the form

$$\frac{z\,(z - r\cos\theta)}{z^2 - 2\,r\cos(\theta)\,z + r^2} \qquad \text{and} \qquad \frac{z\,r\sin\theta}{z^2 - 2\,r\cos(\theta)\,z + r^2}$$

- A pole at $z = 0$ produces a term of the form $1/z$ in the partial fraction expansion.

b) Multiply both sides of the equation by the denominator of $X(z)$ to obtain a polynomial equation in z.

c) Equate the coefficients of each power of z to obtain a system of equations.

d) Solve the system of equations to find the coefficients of the partial fraction expansion.

e) Use a table to find the inverse z-transform of each term, and use the linearity property of the z-transform to write down the inverse z-transform $x(k)$.

Shortcuts

The above procedure can always be used to find the inverse z-transform. The process, however, requires solving a system of equations that can be rather tedious. In many cases, there are some shortcuts that we can take to simplify the calculations.

The partial fraction expansion always includes a constant term. All of the other terms, however, have a factor of z in the numerator. Setting $z = 0$ in $X(z)$ and its partial fraction expansion, we have that

$$X(0) = \text{constant term}$$

Therefore, we can solve for the constant term before solving the system of equations, which reduces the dimension of the system of equations by one.

Similarly, we can directly solve for the coefficients corresponding to terms of the form $z/(z-a)$ for simple roots without solving the system of equations. The idea is to multiply both $X(z)$ and its partial fraction expansion by $z - a$ and then set $z = a$ to cancel all other terms in the partial fraction expansion.

Example (Repeated roots, revisited). Find the inverse z-transform of $X(z) = \dfrac{5z\,(z+1)}{(z+2)^2(z+3)}$.

As before, the form of the partial fraction expansion is

$$X(z) = \frac{5z\,(z+1)}{(z+2)^2(z+3)} = \frac{Az}{z+2} + \frac{Bz}{(z+2)^2} + \frac{Cz}{z+3} + D$$

We can immediately find the constant term $D = X(0) = 0$. Similarly, we can immediately find the coefficient C as follows. First, multiply both sides by $z + 3$ to obtain

$$\frac{5z\,(z+1)}{(z+2)^2} = \frac{Az(z+3)}{z+2} + \frac{Bz(z+3)}{(z+2)^2} + Cz + D\,(z+3)$$

All terms on the right-hand side except for the C term contain a factor of $z + 3$ in the numerator, so they all become zero when we set $z = -3$. Substituting this into the equation, we have

$$\frac{5(-3)(-3+1)}{(-3+2)^2} = -3C$$

which implies that $C = -10$. We can also find the coefficient B as follows. First, multiply both sides by $(z + 2)^2$ to obtain

$$\frac{5z\,(z+1)}{z+3} = Az(z+2) + Bz + \frac{Cz(z+2)^2}{z+3} + D\,(z+2)^2$$

All terms on the right-hand side except for the B term contain a factor of $z + 2$ in the numerator, so they all become zero when we set $z = -2$. Substituting this into the equation, we have

$$\frac{5(-2)(-2+1)}{-2+3} = -2B$$

which implies that $B = -5$. We cannot do something similar to find A since multiplying by $z + 2$ gives

$$\frac{5z\,(z+1)}{(z+2)(z+3)} = Az + \frac{Bz}{z+2} + \frac{Cz(z+2)}{z+3} + D\,(z+2)$$

This still has $z+2$ in the denominator, so we cannot set $z = -2$ as before. In this case, we must procede as in the general procedure, except that we already know B, C, and D so the system of equations will only involve the single unknown A. Using the linear system of equations from before, the z^3 coefficients give that $0 = A + C + D$ which implies that $A = 10$. Therefore, the inverse z-transform is

$$x(k) = \left[10\,(-2)^k - 5k\,(-2)^{k-1} - 10\,(-3)^k\right] u_s(k)$$

This shortcut is most useful when $X(z)$ has all real distinct poles, since it can then be used to find all coefficients in the partial fraction expansion without needing to solve a system of equations.

7

Laplace Transform

The Laplace transform is a useful tool in understanding LTI systems that interprets signals in terms of their *frequencies*. The Laplace transform is a generalization of the Fourier series, and it is the continuous-time equivalent of the z-transform for discrete-time signals.

7.1 Definition

Definition. The Laplace transform of a continuous-time signal $x(t)$ is the complex function

$$X(s) = \int_0^\infty x(t)\, e^{-st}\, \mathrm{d}t$$

where s is any complex number for which the integral converges. The values of s for which the integral converges is the region of convergence (ROC).

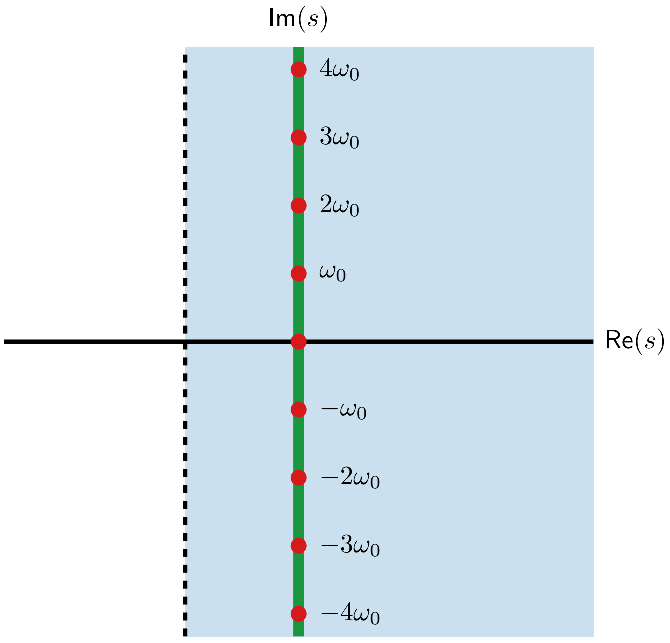
8

Transform Summary

We have seen various ways of transforming time-domain signals into a frequency-domain representation. We now summarize these various transforms in both continuous and discrete time.

8.1 Continuous time

- The most general transform in continuous time is the Laplace transform, which is a rational function of the complex number s. The zeros and poles of the Laplace transform are the roots of the numerator and denominator polynomials. The region of convergence is the set of complex numbers s for which the integral converges, which is to the right of the rightmost pole.
- The Fourier transform is the Laplace transform evaluated on the imaginary axis. The Fourier transform exists only if all poles of the Laplace transform are strictly in the left-half plane so that the imaginary axis is inside the ROC.
- For periodic signals, the Fourier series coefficients are evenly-spaced values of the Fourier transform.



Laplace transform

$$X(s) = \int_0^\infty x(t)\, e^{-st}\, \mathrm{d}t$$

Fourier transform

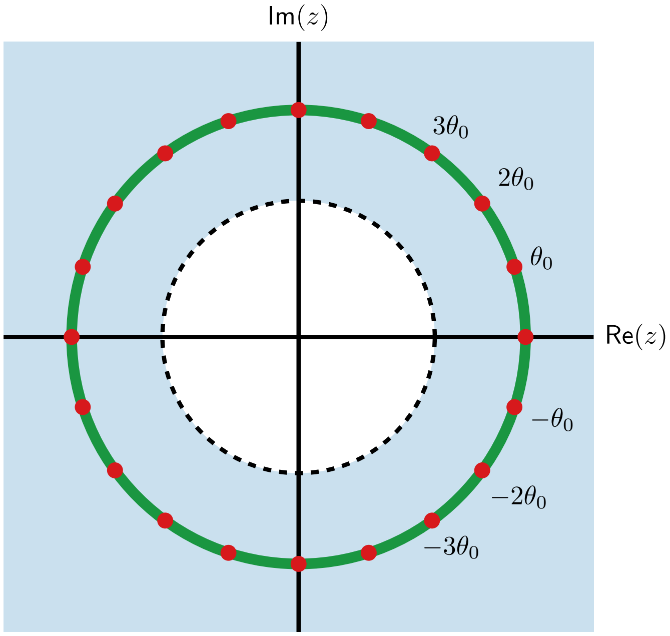
$$X(j\omega) = \int_{-\infty}^\infty x(t)\, e^{-j\omega t}\, \mathrm{d}t$$

Fourier series

$$c_k = \frac{1}{T}\int_T x(t)\, e^{-jk\omega_0 t}\, \mathrm{d}t \quad \text{where} \quad \omega_0 = \frac{2\pi}{T}$$

8.2 Discrete time

- The most general transform in discrete time is the z-transform, which is a rational function of the complex number z. The zeros and poles of the z-transform are the roots of the numerator and denominator polynomials. The region of convergence is the set of complex numbers z for which the summation converges, which is outside of the circle through the outermost pole.
- The discrete-time Fourier transform is the z-transform evaluated on the unit circle. The discrete-time Fourier transform exists only if all poles of the z-transform are strictly inside the unit circle so that the unit circle is inside the ROC.
- For periodic signals, the discrete Fourier transform coefficients are evenly-spaced values of the discrete-time Fourier transform.



z-transform

$$X(z) = \sum_{k=0}^{\infty} x(k)\, z^{-k}$$

Discrete-time Fourier transform

$$X(e^{j\theta}) = \sum_{k=-\infty}^{\infty} x(k)\, e^{-jk\theta}$$

Discrete Fourier transform

$$c_k = \frac{1}{N} \sum_{k=\langle N \rangle} x(k)\, e^{-jk\theta_0 n} \quad \text{where} \quad \theta_0 = \frac{2\pi}{N}$$

Part III

Discrete-Time LTI Systems

9

Input-Output Representations

We now focus on causal finite-dimensional discrete-time linear time-invariant systems. We can represent such systems in various ways. While the representations are all equivalent, each one will lead to a deeper understanding of the system, and some representations are more useful in certain scenarios than others. Here, we focus on *input-output* representations that do not explicitly model the internal state of the system.

9.1 Difference equation

A difference equation is an equation whose expressions consist of discrete-time signals evaluated at different times.

Example (savings account). Previously, we modeled a savings account as a discrete-time system, where we expressed the principal balance *next month* in terms of the balance *this month*. Because the equation involves a signal (the principle balance) at different points in time (this month and next month), this is a difference equation for the savings account system.

Theorem. Every causal finite-dimensional discrete-time LTI system can be represented by a difference equation of the following form:

$$y(k) + \sum_{i=1}^{n} a_i \, y(k-i) = \sum_{j=0}^{m} b_j \, u(k-j)$$

where

- $u(k)$ is the input signal
- $y(k)$ is the output signal
- n is the order of the system (assuming $a_n \neq 0$)
- a_i and b_i are constant parameters

Conversely, every difference equation of this form represents a causal finite-dimensional discrete-time LTI system.

Example. An example of a difference equation is

$$y(k) - \frac{1}{2}y(k-1) = u(k-1)$$

This fits into the general form with $n = m = 1$, $a_1 = -\frac{1}{2}$, $b_0 = 0$, and $b_1 = 1$.

An alternative way to write a difference equation is as a *recursion*, where we solve for the last output $y(k)$ in terms of the input signal and previous outputs:

$$y(k) = -\sum_{i=1}^{n} a_i\, y(k-i) + \sum_{j=0}^{m} b_j\, u(k-j)$$

This is a recursion because the current value of the output $y(k)$ depends on its previous values $y(k-1)$, $y(k-2)$, ..., $y(k-n)$. In order to compute the response (or output signal), we need n initial conditions.

Example (iterating a difference equation by hand)**.** Consider the first-order difference equation from the example above:

$$y(k) - \frac{1}{2}y(k-1) = u(k-1)$$

To set this up as a recursion, we solve for the last-most output:

$$y(k) = \frac{1}{2}y(k-1) + u(k-1)$$

Given the input signal $u(k)$ and the initial condition $y(0)$, we can use recursion to compute the output at each time k. Let's consider several cases.

- **Zero-input response.** First, suppose that the input signal is $u(k) = 0$ for all k and the initial condition is $y(0) = 2$. Then we can substitute $k = 1$ into the recursion to obtain

  $$y(1) = \frac{1}{2}y(0) + u(0) = 1$$

  Now that we know $y(1)$, we can substitute $k = 2$ into the recursion to obtain

  $$y(2) = \frac{1}{2}y(1) + u(1) = \frac{1}{2}$$

  And now that we know $y(2)$, we can substitute $k = 3$ into the recursion to obtain

  $$y(3) = \frac{1}{2}y(2) + u(2) = \frac{1}{4}$$

  and so on. Looking at the recursion, we notice that the output is being multiplied by a half at each iteration. In fact, we can guess the solution in this case! It is just

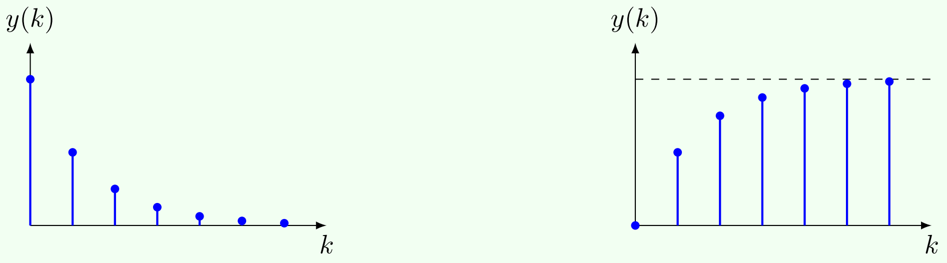
  $$y(k) = 2\left(\tfrac{1}{2}\right)^k$$

  which is a decaying exponential.

- **Step response.** Now suppose that the input signal is a unit step, $u(k) = u_s(k)$, and the initial condition is $y(0) = 0$. Iterating the recursion, we obtain the output:

  $$\begin{aligned} y(1) &= \tfrac{1}{2}y(0) + x(0) = 1 \\ y(2) &= \tfrac{1}{2}y(1) + x(1) = \tfrac{3}{2} \\ y(3) &= \tfrac{1}{2}y(2) + x(2) = \tfrac{7}{4} \\ &\vdots \\ y(k) &= 2 - 2\left(\tfrac{1}{2}\right)^k \end{aligned}$$

  where we found the general expression again by inspection. Since we put in a constant input, the output now exponentially approaches the value two. Both responses are plotted below.

  This first-order discrete-time system behaves similar to a first-order continuous-time system, such as an RC circuit. In this analogy, the first scenario corresponds to the capacitor discharging its initial charge, while the second scenario corresponds to the capacitor being charged by a constant input voltage.

Example (undamped oscillator)**.** While first-order systems have relatively simple behavior, second-order systems can be more complex. Consider the second-order system

$$y(k) - y(k-1) + y(k-2) = 2u(k-1) + u(k-2)$$

Once again, we can turn this into a recursion by solving for $y(k)$.

$$y(k) = y(k-1) - y(k-2) + 2u(k-1) + u(k-2)$$

Instead of iterating by hand, let's write a MATLAB program to do this for us. Since this is a second-order difference equation, we will need two initial conditions and the input signal in order to iterate. We also need to specify how many iterations to perform. We will put our code in a function so that we can call it with different scenarios, and we'll have it create a plot showing both the input and output signals.

```matlab
function undamped_oscillator(N,y1,y2,u,name)
%
% To use, run the following code in the command line:
%
% u1 = @(k) k-3 == 0;
% u2 = @(k) zeros(size(k));
% u3 = @(k) ( (mod(k,6)==0) - (mod(k,6)==3) );
%
% undamped_oscillator(30,0,0,u1,'undamped oscillator with impulse input');
% undamped_oscillator(30,0,1,u2,'undamped oscillator with nonzero initial
   conditions');
% undamped_oscillator(30,0,0,u3,'undamped oscillator with resonance');

% initial conditions
y(1) = y1;
y(2) = y2;

% iterate the difference equation
for k = 3:N
    y(k) = y(k-1) - y(k-2) + 2*u(k-1) + u(k-2);
end

% plot the result
figure;
stem(1:N,u(1:N)); hold on;
stem(1:N,y);
xlim([0 N]);
grid on;
legend('input signal u(k)','output signal y(k)','Location','Northwest');
xlabel('k');
title(name);

pbaspect([2 1 1]);

end
```

- **Zero-input response.** For the first scenario, let's see what happens to the system when we start it in some nonzero initial state and let it go (that is, the input is zero).

```matlab
N  = 30;                    % number of iterations
y1 = 0;                     % initial condition y(1)
y2 = 1;                     % initial condition y(2)
u  = @(k) zeros(size(k));   % input signal u(k)

undamped_oscillator(N,y1,y2,u);
```

Our program produces the following plot. Even though we didn't put any input into the system besides the initial conditions, it continues oscillating forever! This is why the system is called an undamped oscillator.

Remark (Noncausal recursions)**.** When setting up the recursion, what would happen if we solved for a different term involving the output? Let's go back to our first-order system:

$$y(k) - \frac{1}{2}y(k-1) = u(k-1)$$

Insteading of solving for the first term $y(k)$, let's solve for the second term $y(k-1)$. Doing so, we get

$$y(k-1) = 2y(k) - 2u(k)$$

But how do we iterate this? Given the input and the initial condition $y(0)$, we can solve for the *previous* output $y(-1)$. And then we can use this value to find $y(-2)$, and so on. Here, we are iterating backwards in time by viewing the difference equation as a noncausal system, as opposed to the causal recursion running forward in time that we did above. We can interpret any difference equation as representing either a causal system or a noncausal system. This is why you will sometimes see things like

$$y(k) - \frac{1}{2}y(k-1) = u(k-1), \qquad k \geq 0$$

to specify that the iteration goes forward in time, so the difference equation represents a causal system. Unless specified, you can typically assume that difference equations represent causal systems. ■

9.2 Transfer function

We previously saw how to represent a discrete-time LTI system as a difference equation. We now learn how to represent it as a transfer function, which is a powerful and compact representation of the system.

Advance and delay

In order to describe the transfer function, we need one of the most simple systems: an advance. This system, which we denote by E, takes in a signal and shifts it to the left by one iteration:

$$u(k) \qquad \xmapsto{E} \qquad u(k+1)$$

This is a discrete-time LTI system, but it is noncausal since the output at iteration $k+1$ depends on the previous input at iteration k. For example, advancing the signal

$$u(k) = (0.75)^k\, u_s(k) = \begin{cases} (0.75)^k & \text{if } k \geq 0 \\ 0 & \text{if } k < 0 \end{cases}$$

produces

$$E\, u(k) = u(k+1) = \begin{cases} (0.75)^{k+1} & \text{if } k+1 \geq 0 \\ 0 & \text{if } k+1 < 0 \end{cases}$$

What system undoes an advance? A delay! Shifting a signal to the left by one and then shifting it to the right by one are inverse operations, so we represent a delay by $1/E$ or E^{-1}.

$$u(k) \qquad \xmapsto{E^{-1}} \qquad u(k-1)$$

This is also a discrete-time LTI system, and this system is also causal! Going back to our example,

$$E^{-1}\, u(k) = u(k-1) = \begin{cases} (0.75)^{k-1} & \text{if } k-1 \geq 0 \\ 0 & \text{if } k-1 < 0 \end{cases}$$

Describing difference equations using the advance operator

We are now going to form more general systems by taking weighted summations of powers of the advance operator. In other words, we are going to do algebra with E. For example, we can use the advance to form a new system $aE + b$ where a and b are scalars. When we apply this system to an input signal, we can distribute the signal through to each term

$$(aE + b)\, u(k) = a\, E\, u(k) + b\, u(k)$$

and then replace the advance operator by its definition

$$(aE + b)\, u(k) = a\, u(k + 1) + b\, u(k)$$

Notice that the right-hand side of this equation looks like a difference equation! Let's take a look at how we can use the advance operator to rewrite a difference equation.

Example. Consider the undamped oscillator that we studied previously:

$$y(k + 2) - y(k + 1) + y(k) = 2u(k + 1) + u(k)$$

Working backwards, we will rewrite this difference equation using the advance operator. First, rewrite each side so that the signals only appear as $u(k)$ and $y(k)$.

$$E^2\, y(k) - E\, y(k) + y(k) = 2\, E\, u(k)\; + u(k)$$

Now factor out the signals $u(k)$ and $y(k)$ to obtain

$$(E^2 - E + 1)\, y(k) = (2E + 1)\, u(k)$$

Since our goal is to find the output $y(k)$ in terms of the input $u(k)$, let's solve for the output just as you would in algebra.

$$y(k) = \frac{2E + 1}{E^2 - E + 1}\, u(k)$$

This is an equivalent representation of the difference equation. We could just as easily do each of these steps in reverse to get back to the original difference equation. The system that gets applied to the input to obtain the output is the transfer function, which is often denoted $H(E)$. In this example,

$$H(E) = \frac{2E + 1}{E^2 - E + 1}$$

The transfer function is the system $H(E)$, which depends on the advance E, that produces the output signal when applied to the input:

$$y(k) = H(E)\, u(k)$$

The transfer function...

- is a rational function of the advance operator E, meaning that it is the ratio of two polynomials
- is equivalent to the difference equation (and it is straightforward to convert between the two representations)
- represents only the difference equation and provides no information about initial conditions

System properties from the transfer function

Since the transfer function is a complete representation of the system, we should be able to determine the properties of the system directly from $H(E)$. To describe these properties, denote the numerator by $N(E)$

and the denominator by $D(E)$ so that

$$H(E) = \frac{N(E)}{D(E)}$$

- **Causality.** A system is causal if and only if the degree of the numerator polynomial $N(E)$ is less than or equal to the degree of the denominator polynomial $D(E)$, that is,

  $$\deg\big(N(E)\big) < \deg\big(D(E)\big)$$

  For example, consider the system with transfer function $H(E) = \frac{E^2}{E-2}$ in which the degree of the numerator (two) is larger than that of the denominator (one). The output is

  $$y(k) = \frac{E^2}{E-2}\,u(k)$$

  Multiplying both sides by the denominator,

  $$(E-2)\,y(k) = E^2\,u(k)$$

  which represents the difference equation

  $$y(k+1) - 2\,y(k) = u(k+2)$$

  This system is noncausal since the output at iteration $k+1$ depends on the future input at time $k+2$.

- **Degree.** The degree of the system is the integer n in the difference equation representation

  $$y(k) + \sum_{i=1}^{n} a_i\,y(k-i) = \sum_{j=0}^{m} b_j\,u(k-j)$$

  which is the degree of the denominator polynomial $D(E)$.

Example (Hidden structure). One simple application of the transfer function is to find a minimal representation of a system. For example, consider the system described by the third-order difference equation

$$y(k+3) + 2\,y(k+2) - y(k+1) - 2\,y(k) = u(k+2) + 3\,u(k+1) + 2\,u(k)$$

At first observation, it is difficult to understand how this system behaves; it is quite complicated! Let's compute its transfer function. Rewriting the difference equation using the advance operator,

$$(E^3 + 2E^2 - E - 2)\,y(k) = (E^2 + 3E + 2)\,u(k)$$

Solving for the output,

$$y(k) = \frac{E^2 + 3E + 2}{E^3 + 2E^2 - E - 2}\,u(k)$$

This isn't much simpler yet. However, notice that the numerator and denominator polynomials factor.

$$y(k) = \frac{(E+1)(E+2)}{(E+1)(E+2)(E-1)}\,u(k)$$

The numerator and denominator both have factors of $(E+1)$ and $(E+2)$. Cancelling these common factors, we have the simple representation

$$y(k) = \frac{1}{E-1}\,u(k)$$

Now going back to a difference equation, we have

$$y(k+1) = y(k) + u(k)$$

It is now straightforward to interpret the behavior of the system: it sums the input signal.

Caution! The transfer function can only represent LTI systems (and any system represented by a transfer function is LTI). For nonlinear or time-varying systems, the transfer function does not exist!

10

Convolution

For discrete-time systems, we can solve for the response by iterating the difference equation that represents the system. This approach, however, does not give insight into how the response changes for various scenarios (different input signals, initial conditions, and/or system parameters). In this chapter, we focus on the zero-state response (the output of the system when it is initially at rest) and learn how to find a closed-form expression for the response of a discrete-time LTI system given the input signal. The main result is that:

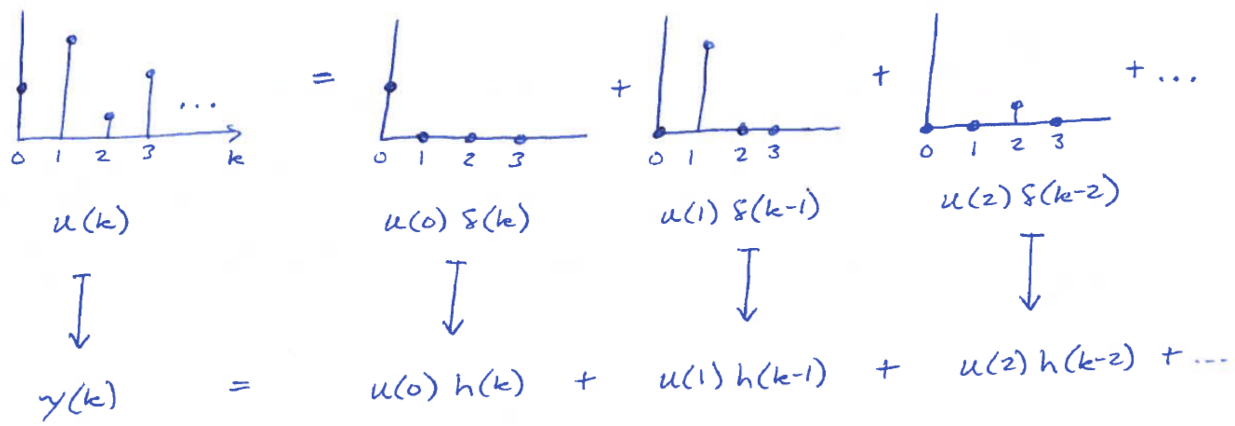
The zero-state response is the convolution of the input signal with the impulse response of the system.

10.1 Motivation

Suppose we want to find the zero-state response of a discrete-time LTI system due to an input signal $u(k)$. Let's first consider just the time $k = 0$. The value of the signal at this time is $u(0)$. We can also represent this as an entire signal with only this single value as $u(0)\,\delta(k)$. At the next time step, the value of the signal is $u(1)$, and a signal with only this value is $u(1)\,\delta(k-1)$, where the impulse is now shifted to occur at time $k = 1$. At a general point $k = m$, the point is represented by $u(m)\,\delta(k-m)$. We can keep doing so to represent each individual point in the signal as a weighted and shifted impulse. The signal is the sum of each of its individual points, so

$$u(k) = \sum_{m=-\infty}^{\infty} \delta(k-m)\,u(m)$$

This is known as the *sifting property* of the impulse signal and holds for any signal $u(k)$.



So far, we have just rewritten the input signal in a more complicated way (as an infinite sum of weighted impulses). However, suppose we know the impulse response of the system:

$$\delta(k) \quad \mapsto \quad h(k)$$

Because the system is time-invariant, we know that shifting the input signal will shift the output signal by the same amount, so

$$\delta(k-m) \quad \mapsto \quad h(k-m)$$

Recall that linearity is equivalent to homogeneity and superposition. From homogeneity, we know that multiplying the input signal by a constant will scale the output signal by the same amount, so

$$\delta(k-m)\,u(m) \quad \mapsto \quad h(k-m)\,u(m)$$

where $u(m)$ is constant with respect to the time index k. This holds for any time shift m and any constant $u(m)$. Applying superposition, we know that summing the input signals (over m) will sum the corresponding output signals, so

$$\sum_{m=-\infty}^{\infty} \delta(k-m)\,u(m) \quad \mapsto \quad \sum_{m=-\infty}^{\infty} h(k-m)\,u(m)$$

From the sifting property of the impulse signal, the left-hand side is simply $u(k)$. In other words,

$$u(k) \quad \mapsto \quad y(k) = \sum_{m=-\infty}^{\infty} h(k-m)\,u(m)$$

Recall that this holds for any signal $u(k)$. Therefore, the zero-state response due to an input signal $u(k)$ is the signal $y(k)$ given by the summation on the right-hand side. The zero-state response depends on the input signal $u(k)$ and the impulse response $h(k)$.

10.2 Definition and main result

Definition (convolution)**.** The convolution of discrete-time signals h and u is the signal $h * u$ given by

$$(h * u)(k) = \sum_{m=-\infty}^{\infty} h(k-m)\,u(m)$$

Using the sifting property of the impulse signal with our motivation that a weighted sum of shifted inputs produces the same weighted sum of shifted outputs, we obtain the following main result for convolution.

Main result (convolution): The zero-state response of an LTI system is the convolution of the input signal with the impulse response:

$$y_{\mathrm{ZSR}}(k) \quad = \quad (h * u)(k) \quad = \quad \sum_{m=-\infty}^{\infty} h(k-m)\,u(m)$$

We can often simplify the limits of the summation.

- If the system is *causal*, then the impulse response is zero for negative times, so $h(k) = 0$ for $k < 0$. Then all terms in the convolutional sum with $m > k$ are zero, so we can set the upper limit to k.
- If the input signal does not start until time zero (that is, $u(k) = 0$ for $k < 0$), then all terms in the convolutional sum with $m < 0$ are zero, so we can set the lower limit to zero.

We often work with causal systems and input signals that are initially zero, in which case the zero-state response simplifies to the finite summation

$$y_{\mathrm{ZSR}}(k) = \sum_{m=0}^{k} h(k-m)\, u(m)$$

Example (first-order system). Consider the discrete-time LTI system with impulse response

$$h(k) = (0.5)^k \, u_s(k)$$

Find the zero-state response due to the input signal $u(k) = 3^k \, u_s(k)$.

The zero-state response is the convolution of the input signal with the impulse response:

$$y(k) = (h * u)(k) = \sum_{m=-\infty}^{\infty} h(k-m)\, u(m)$$

Substituting the input signal and the impulse response,

$$y(k) = \sum_{m=-\infty}^{\infty} (0.5)^{k-m} \, u_s(k-m) \, 3^m \, u_s(m)$$

The first unit step $u_s(k-m)$ is zero whenever $k - m < 0$, or equivalently $m > k$. Therefore, we can replace the upper limit in the summation with $m = k$ since all terms after this are zero.

$$y(k) = \sum_{m=-\infty}^{k} (0.5)^{k-m} \, 3^m \, u_s(m)$$

The other unit step $u_s(m)$ is zero whenever $m < 0$, so we can replace the lower limit in the summation with $m = 0$ since all terms before this are zero.

$$y(k) = \sum_{m=0}^{k} (0.5)^{k-m} \, 3^m$$

To evaluate the summation, let's first break the first exponential into two terms,

$$y(k) = \sum_{m=0}^{k} (0.5)^k \, (0.5)^{-m} \, 3^m$$

The first exponential does not depend on m and can therefore be pulled outside of the summation, and we can combine the other two terms into a single exponential.

$$y(k) = (0.5)^k \, \sum_{m=0}^{k} \left(\frac{3}{0.5} \right)^m$$

We can now apply the geometric series formula

$$\sum_{m=0}^{k} a^k = \frac{1 - a^{k+1}}{1-a} \, u_s(k)$$

to obtain

$$y(k) = (0.5)^k \, \frac{1 - 6^{k+1}}{1 - 6} \, u_s(k)$$

Simplifying this expression, we have

$$y(k) = \tfrac{1}{5} \, (0.5)^k \left(6^{k+1} - 1 \right) u_s(k)$$

which further simplifies to

$$y(k) \quad = \quad \underbrace{\tfrac{6}{5}\,(3)^k \, u_s(k)}_{\text{particular}} \quad - \quad \underbrace{\tfrac{1}{5}\,(0.5)^k \, u_s(k)}_{\text{homogeneous}}$$

where we can identify the homogeneous and particular solutions in the response.

Example (second-order system). Consider the discrete-time LTI system with impulse response

$$h(k) = \left[-100\,(0.1)^k + 50\,(0.2)^k\right] u_s(k-1)$$

Find the step response of the system. To check your solution, the step response is

$$y(k) = \left[\tfrac{25}{18} - \tfrac{25}{2}\,(0.2)^k + \tfrac{100}{9}\,(0.1)^k\right] u_s(k-1)$$

10.3 Properties

Convolution is an operator that takes two signals and produces another signal. The convolution operator satisfies the following properties.

- **Commutative**
  Convolution is commutative, meaning that the result does not depend on the order in which two signals are convolved.
  $$u * h = h * u$$
- **Distributive**
  Convolution is distributive over addition, meaning that the convolution of a sum is equal to the sum of the convolutions.
  $$u * (h_1 + h_2) = u * h_1 + u * h_2$$
- **Associative**
  Convolution is associative, meaning that the order in which you convolve multiple signals does not change the result.
  $$u * (h_1 * h_2) = (u * h_1) * h_2$$
- **Sifting property**
  The sifting property of the impulse signal shows that convolving any signal with an impulse does not change the signal.
  $$\delta * u \quad = \quad u \quad = \quad u * \delta$$
  In other words, the impulse signal is the identity element for convolution (similar to how the number one is the identity element for multiplication).
- **Time shifting**
  Convolution also has a property that multiplication does not have, which is the time shifting property. This property states that if we shift either of the signals in the convolution in time, then the convoluted signal also gets shifted by the same amount in time. Mathematically,
  $$u_1 * u_2 = y \qquad \Longrightarrow \qquad E^n u_1 * E^m u_2 = E^{n+m} y$$
  where E is the advance operator.

Remark (analogy to scalar multiplication). As an analogy, let's first think about basic *multiplication*. Multiplication is an operator that takes two numbers and multiplies them together to produce another number. While we typically denote multiplication by juxtaposition (writing two numbers side-by-side as in ab), let's make this explicit and write $a \times b$. As we all know, multiplication satisfies several important properties. For example, it is *commutative*, meaning that the order in which you multiply two numbers does not matter, so $a \times b = b \times a$. It is also *distributive* over addition, meaning that $a \times (b + c) = a \times b + a \times c$. Multiplication is also *associative*, meaning that $a \times (b \times c) = (a \times b) \times c$. And finally, the number one is special in that multiplying any number by one does not change the number, sp $a \times 1 = a$. The number one is called the *identity element* for multiplication.

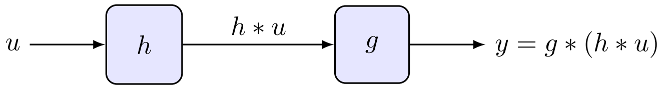
These properties of addition and multiplication are generalized in the mathematical field of abstract algebra as a *ring*. While the multiplication operator acts on *scalars*, the convolution operator acts on *signals*. Just like for multiplication, addition and convolution of signals also form a ring and therefore have all of the same properties. While the number one is the identity element for multiplication, the impulse signal is the identity element for convolution. ■

Implications for interconnected systems

The properties of convolution have important implications for interconnections of LTI systems.

- **Series connection.** *The zero-state response of two LTI systems connected in series is independent of the order of the systems.*

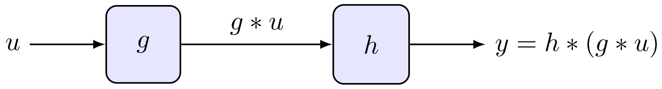
  Consider the series interconnection of two LTI systems with impulse responses h and g.

  The zero-state response of the first system is the signal $h * u$, and the response of the second system is the signal

  $$\begin{aligned} y &= g * (h * u) \\ &= (g * h) * u && \text{(associativity)} \\ &= (h * g) * u && \text{(commutativity)} \\ &= h * (g * u) && \text{(associativity)} \end{aligned}$$

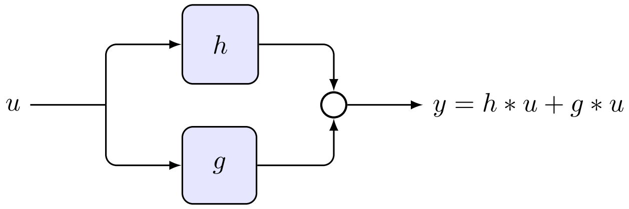
  This is the same output as the interconnected system with the order of the two subsystems switched.

  This property only holds for *single-input single-output* systems. This is because, while multiplication commutes for scalars, matrix multiplication is in general not commutative.

- **Parallel connection.** *A parallel connection of LTI systems is equivalent to an LTI system whose impulse response is the sum of the individual impulse responses.*

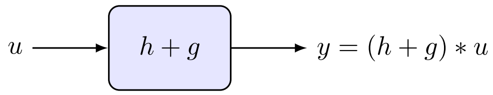
  Consider the parallel connection of two LTI systems with impulse responses h and g.

  The zero-state response of the interconnection is

  $$\begin{aligned} y &= h * u + g * u \\ &= u * h + u * g && \text{(commutativity)} \\ &= u * (h + g) && \text{(distributivity)} \\ &= (h + g) * u && \text{(commutativity)} \end{aligned}$$

  This is the same output as a single LTI system whose impulse response is the sum of the two individual impulse responses.



10.4 Application: Matched filter

An important application of convolution is the matched filter, which is an LTI system that is designed to detect a specific signal. Matched filters are used in a variety of systems.

- **Digital communication.** In communication systems, information may be encoded as a sequence of bits. To transmit these signals, each bit may be associated with a signal. To decode the message, the receiver must detect which signal has been transmitted. One way to do this is to have two filters in the receiver, each one tuned to detect one of the signals.
- **Radar.** In radar systems, an electromagnetic pulse is transmitted at a target. The target reflects the pulse, which then returns to the transmitter. The amount of time between when the signal was transmitted and received is proportional to the distance to the target. Ideally, the received signal will be identical to the original signal, except shifted in time and scaled (due to attenuation). The receiver should then be designed to detect this signal to measure the distance to the target.

Let $s(k)$ be the signal that we want the system to detect, and suppose that the signal is zero for $k < 0$ and $k > T$ for some time T. Our goal is then to design an LTI system (or filter) to detect this signal.

Definition (matched filter)**.** The matched filter designed to detect a signal $s(k)$ is the LTI system whose impulse response is the shifted and time-reversed signal, $h(k) = s(T - k)$.

Note that $h(k)$ is zero for $k < 0$ and $k > T$ due to the corresponding properties for $s(k)$. The output of the filter due to an input signal $u(k)$ is then the convolution of the input signal with the impulse response:

$$y(k) = (h * u)(k) = \sum_{m=0}^{k} h(k-m)\, u(m)$$

When the input to the matched filter is the signal $s(k)$, the output is

$$y(k) = \sum_{m=0}^{k} s(N - (k-m))\, s(m)$$

In particular, the output at time T is the energy of the signal $s(k)$,

$$y(T) = \sum_{m=0}^{T} s(m)^2 = \|s\|^2$$

The matched filter is precisely the filter that maximizes this value.

We can use the matched filter to detect a delayed version of the signal $s(k)$ as follows. Suppose we receive the signal $s(k - d)$, where d is the time delay. Then the output of the matched filter at time $T + d$ is the energy, $y(T + d) = \|s\|^2$. So given the time $d + T$ at which the output of the matched filter has its peak value, the amount by which the transmitted signal is delayed is d.

11

Time Domain

For a given system, the most fundamental analysis question is: *how does the system respond to various excitations?* The output of a system, known as the *response*, depends on both the input signal and the initial conditions. In an RC circuit, for instance, the voltage across the capacitor depends on both the source voltage and the initial capacitor voltage. For discrete-time systems, we can solve for the response by iterating the difference equation. This approach, however, does not give insight into how the response changes for various scenarios (different input signals, initial conditions, and/or system parameters). In this chapter, we learn how to find a closed-form expression for the response of a discrete-time LTI system given the input signal and initial conditions.

11.1 Overview

Consider an LTI system described by the difference equation

$$D(E)\,y(k) = N(E)\,u(k)$$

where $u(k)$ is the input signal, $y(k)$ is the output signal (or response), and $N(E)$ and $D(E)$ are polynomials in the advance operator E. As we will see, the structure of the response depends on that of the input signal and the roots of the *characteristic polynomial* $D(E)$.

To find the response of the system to a particular input signal and set of initial conditions, we apply the following general procedure that will be described in detail throughout the remainder of this chapter.

To find the response of an n^{th}-order discrete-time LTI system:

a) Find the homogeneous solution, which contains n parameters.

b) Find a particular solution.

c) Add the homogeneous and particular solutions.

d) Apply the initial conditions to solve for the n parameters.

11.2 Homogeneous solution

The *homogeneous equation* is the difference equation describing the dynamics of the system when the input signal is zero,

$$D(E)\,y(k) = 0$$

This equation is called *homogeneous* because we can scale any solution $y(k)$ and it remains a solution. The solution to the homogeneous equation is the *homogeneous solution*. For an n^{th}-order system, there are n independent solutions to the homogeneous equation, so the homogeneous solution contains n parameters. To find the homogeneous solution, we will start with a simple system and work toward the more general case.

Case: first-order system

The characteristic polynomial of a first-order system is of the form $D(E) = E - a$, so the homogeneous equation is the first-order difference equation

$$y(k+1) - a\,y(k) = 0$$

Suppose the initial condition is $y(0) = m$. Since this system is so simple, we can find the solution by just iterating until we see a pattern:

$$\begin{aligned} y(0) &= m \\ y(1) &= a\,y(0) = a\,m \\ y(2) &= a\,y(1) = a^2\,m \\ y(3) &= a\,y(2) = a^3\,m \\ &\vdots \end{aligned}$$

This leads to the homogeneous solution

$$y(k) = m\,a^k$$

The solution depends on the root a of the characteristic polynomial and has a single parameter m since this is a first-order system.

Case: second-order system with distinct real roots

The characteristic polynomial of a second-order system is quadratic. For now, let's suppose that it factors as

$$(E-b)(E-a)\,y(k) = 0$$

where a and b are not equal. From our analysis of first-order systems, one solution is $y_1(k) = m_1\,a^k$ since

$$(E-b)(E-a)\,y_1(k) = (E-b)\big((E-a)\,m_1\,a^k\big) = (E-b)\cdot 0 = 0$$

Similarly, we can switch the order of the two factors to show that another solution is $y_2(k) = m_2\,b^k$. If a and b are not equal, then these are two different solutions. Since the system is linear, we can take a linear combination of these two solutions to get the general solution

$$y(k) = m_1\,a^k + m_2\,b^k \qquad \text{if } a \neq b$$

Case: n^{th}-order system with distinct real roots

For an n^{th}-order system whose characteristic polynomial has all distinct roots, the polynomial factors and the homogeneous equation can be written as

$$(E - a_1)(E - a_2) \dots (E - a_n)\, y(k) = 0$$

where the roots a_i are all distinct. This equation has n independent solutions of the form $m_i\, a_i^k$, so the general solution is

$$y(k) = m_1\, a_1^k + m_2\, a_2^k + \dots + m_n\, a_n^k \qquad \text{if } a_i \text{ all distinct}$$

Example. Let's solve the homogeneous difference equation

$$y(k+2) - 3\, y(k+1) + 2\, y(k) = 0$$

First, write the difference equation using the advance operator E.

$$(E^2 - 3E + 2)\, y(k) = 0$$

Factor the characteristic polynomial.

$$(E - 1)(E - 2)\, y(k) = 0$$

The characteristic polynomial is quadratic with two distinct roots, so the homogeneous solution is

$$y(k) = m_1\, (1)^k + m_2\, (2)^k = m_1 + m_2\, 2^k$$

So far, we have assumed that the roots of the characteristic polynomial are *real* and *distinct*. But in general, the roots of a polynomial may be *complex* and *repeated*. We cover these two cases next.

Case: two repeated roots

Let's start with the quadratic case where the root is repeated.

$$(E - a)^2\, y(k) = 0$$

As before, one solution is $y_1(k) = m_1\, a^k$. We need two solutions for a second-order system. But since the roots are the same, we cannot take $y_2(k) = m_2\, a^k$ as before because this is the same as the first solution. Instead, let's define an intermediate signal

$$w(k) = (E - a)\, y(k)$$

This signal satisfies the *first-order* homogeneous equation

$$(E - a)\, w(k) = 0$$

which we know has the solution $w(k) = m\, a^k$. Now substituting this back into how we defined $w(k)$,

$$(E - a)\, y(k) = m\, a^k$$

This is still a first-order equation, but it is no longer homogeneous (it has a nonzero right-hand side). To find its solution, let's iterate the difference equation and look for a pattern:

$$\begin{aligned} y(1) &= m\,a^0 = m \\ y(2) &= a\,m + m\,a^1 = 2ma \\ y(3) &= a \cdot 2ma + m\,a^2 = 3ma^2 \\ &\vdots \\ y(k) &= k\,m\,a^{k-1} \end{aligned}$$

Since m was an arbitrary parameter, we can set $m_2 = m/a$ to simplify the solution to $y(k) = k\,m_2\,a^k$. The homogeneous solution is then the sum of both solutions,

$$y(k) = (m_1 + k\,m_2)\,a^k$$

Example (two repeated roots)**.** Let's solve the homogeneous recursion

$$y(k) = 2.5\,y(k-1) - 2\,y(k-2) + 0.5\,y(k-3)$$

We first shift time by three iterations and bring all the terms to the same side.

$$y(k+3) - 2.5\,y(k+2) + 2\,y(k+1) - 0.5\,y(k) = 0$$

Now rewrite the equation using the advance operator.

$$(E^3 - 2.5E^2 + 2E - 0.5)\,y(k) = 0$$

To find the homogeneous solution, we need to factor the characteristic polynomial.

$$(E-0.5)(E-1)^2\,y(k) = 0$$

The characteristic polynomial has a single real root at 0.5 and a repeated real root at 1, so the homogeneous solution is

$$y(k) = m_1\,(0.5)^k + m_2 + m_3\,k$$

where the first term is due to the real root at 0.5 and the second two terms are due to the repeated root at 1.

Case: more repeated roots

The form of the solution when a root is repeated more times is as follows:

$$\begin{aligned} (E-a)\,y(k) = 0 \quad &\Longrightarrow \quad y(k) = m_1\,a^k \\ (E-a)^2\,y(k) = 0 \quad &\Longrightarrow \quad y(k) = (m_1 + m_2\,k)\,a^k \\ (E-a)^3\,y(k) = 0 \quad &\Longrightarrow \quad y(k) = (m_1 + m_2\,k + m_3\,k^2)\,a^k \\ &\vdots \\ (E-a)^n\,y(k) = 0 \quad &\Longrightarrow \quad y(k) = (m_1 + m_2\,k + \ldots + m_n\,k^{n-1})\,a^k \end{aligned}$$

Example (more repeated roots). The homogeneous equation

$$(E-0.5)^2(E+0.2)^3\,y(k)=0$$

has the homogeneous solution

$$y(k)=(m_1+m_2\,k)\,(0.5)^k+(m_3+m_4\,k+m_5\,k^2)\,(-0.2)^k$$

Roots at zero

For repeated roots at zero, we cannot use the general formula above because the solutions all collapse to a single solution (an impulse signal). For n repeated roots at zero, the solution to the homogeneous difference equation

$$E^n\,y(k)=0$$

is a weighted sum of shifted impulses:

$$y(k)=m_1\,\delta(k)+m_2\,\delta(k-1)+\ldots+m_n\,\delta(k-n+1)$$

Example (roots at zero). The homogeneous equation

$$E^2(E-0.5)\,y(k)=0$$

has the homogeneous solution

$$y(k)=m_1\,\delta(k)+m_2\,\delta(k-1)+m_3\,(0.5)^k$$

Case: complex roots

Since the coefficients of the characteristic polynomial are real numbers, complex roots of the characteristic polynomial always come in complex conjugate pairs. Recall that, for a complex number $a+jb$, its complex conjugate is $a-jb$. For a second-order system with two complex conjugate roots, the homogeneous equation has the form

$$(E-p)(E-\bar{p})\,y(k)=0$$

where p is a complex number and $\bar{p}$ is its complex conjugate. We can still apply the formula for distinct roots in this case to get the solution

$$y(k)=m_1\,p^k+m_2\,\bar{p}^k$$

but now the coefficients m_1 and m_2 will in general be complex, even though the signal $y(k)$ is real! Instead, we prefer to use a form with only real numbers to emphasize that the signal is real.

To find a better expression, let's write p in polar form as $r\,e^{j\theta}$ where r is the magnitude and θ the angle of the complex number. Then its complex conjugate is $\bar{p}=r\,e^{-j\theta}$, so the solution is

$$y(k)=m_1\left(r\,e^{j\theta}\right)^k+m_2\left(r\,e^{-j\theta}\right)^k$$

Since the output signal $y(k)$ is real, the coefficients m_1 and m_2 must be complex conjugates of each other.

Proof. For the signal

$$y(k)=m_1\,p^k+m_2\,\bar{p}^k$$

to be real for all k, we will show that the complex coefficients m_1 and m_2 must be complex conjugates of each other. Let's write all of the complex numbers in polar form:

$$p = r\,e^{j\theta} \qquad m_1 = a\,e^{j\phi} \qquad m_2 = b\,e^{j\psi}$$

Substituting these expressions into $y(k)$, we obtain

$$y(k) = a\,e^{j\phi}\,r\,e^{jk\theta} + b\,e^{j\psi}\,r\,e^{-jk\theta}$$

Combining the complex exponentials and factoring out r,

$$y(k) = r\left[a\,e^{j(\phi+k\theta)} + b\,e^{j(\psi-k\theta)}\right]$$

Now use Euler's formula to rewrite the complex exponentials in rectangular form,

$$y(k) = r\left[a\Big(\cos(\phi+k\theta) + j\sin(\phi+k\theta)\Big) + b\Big(\cos(\psi-k\theta) + j\sin(\psi-k\theta)\Big)\right]$$

Grouping the real and imaginary parts,

$$y(k) = r\left[\Big(a\cos(\phi+k\theta) + b\cos(\psi-k\theta)\Big) + j\Big(a\sin(\phi+k\theta) + b\sin(\psi-k\theta)\Big)\right]$$

All of the parameters are real (they are the magnitudes and angles of the complex numbers p, m_1, and m_2), so we can now clearly identify the real and imaginary parts of the signal. Since we know that $y(k)$ must be a real signal, the imaginary part must be zero. That is,

$$0 = \mathrm{Im}\{y(k)\} = a\sin(\phi+k\theta) + b\sin(\psi-k\theta)$$

For this to be zero for all values of k, we must have that $\psi = -\phi$ and $a = b$ so that the two terms cancel (using that sine is an odd function). Therefore, the complex coefficients m_1 and m_2 must be complex conjugates of each other. ■

Since m_1 and m_2 are complex conjugates of each other, we can write them in polar form as

$$m_1 = \frac{m}{2}\,e^{j\phi} \qquad \text{and} \qquad m_2 = \frac{m}{2}\,e^{-j\phi}$$

(we included the factor of one half for convenience, so here m is twice the magnitude of m_1 and m_2). Substituting these expressions, the homogeneous solution is

$$y(k) = \frac{m}{2}\,r^k\left(e^{j(k\theta+\phi)} + e^{-j(k\theta+\phi)}\right)$$

Applying the trigonometric identity $\cos\theta = \frac{1}{2}(e^{j\theta} + e^{-j\theta})$ we get the form

$$y(k) = m\,r^k\cos(k\theta+\phi)$$

where the parameters are now m and ϕ, which are both real. The other parameters r and θ in the solution are the magnitude and angle of the complex roots of the characteristic polynomial. While this form emphasizes the fact that the response is a real signal, to find the parameters m and ϕ we need to solve a *nonlinear* system of equations. Instead, we can use the equivalent form

$$y(k) = r^k\left[m_1\cos(k\theta) + m_2\sin(k\theta)\right]$$

where the parameters are now m_1 and m_2 (these are both real, and are not the same as the complex ones used above). The parameters now appear linearly, so it is generally easier to solve for the parameters in this form. However, both forms are equivalent and you can use whichever you like.

When the characteristic polynomial has two complex conjugate roots with magnitude r and angle $\pm\theta$, the homogeneous solution can be written in either form

$$y(k) \quad = \quad m\, r^k \cos(k\theta + \phi) \quad = \quad r^k \left[m_1 \cos(k\theta) + m_2 \sin(k\theta)\right]$$

where the two parameters are m and ϕ in the first form and m_1 and m_2 in the second form.

Example (complex roots). Consider the difference equation

$$(E^2 - 2E + 4)\, y(k) = 0$$

The characteristic polynomial does not factor, so we have to resort to the quadratic equation. This gives the two complex roots

$$p = \frac{2 \pm \sqrt{4-16}}{2} = 1 \pm j\sqrt{3}$$

The magnitude of the roots is

$$r = |p| = \sqrt{(1)^2 + \big(\sqrt{3}\big)^2} = 2$$

and the angle is

$$\theta = \angle p = \arctan\big(\sqrt{3}/1\big) = \frac{\pi}{3} = 60^\circ$$

Therefore, the homogeneous solution is of the form

$$y(k) = m\, 2^k \cos\big(k\tfrac{\pi}{3} + \phi\big) = 2^k \left[m_1 \cos\big(k\tfrac{\pi}{3}\big) + m_2 \sin\big(k\tfrac{\pi}{3}\big)\right]$$

with real parameters m and ϕ, or equivalently, m_1 and m_2.

11.3 Particular solution

A *particular solution* is a response $y(k)$ that satisfies the difference equation (with the input), but need not satisfy the initial conditions. In general, the particular solution has the same form as the input signal. There are many particular solutions, but we only need to find one of them.

The general method to find a particular solution is to try a generalized form of the input signal. For instance, if the input signal is sinusoidal, then the particular solution should also be sinusoidal with the same frequency. Here are some examples of input signals $u(k)$ and the corresponding form of the particular solution $y(k)$.

$u(k) = 2^k$	$y(k) = A\, 2^k$
$u(k) = k\,(0.5)^k$	$y(k) = A\,(0.5)^k + Bk\,(0.5)^k$
$u(k) = 6\cos\big(k\frac{\pi}{6}\big)$	$y(k) = A\cos\big(k\frac{\pi}{6}\big) + B\sin\big(k\frac{\pi}{6}\big)$
$u(k) = 4\,(0.5)^k + 2\,(0.3)^k$	$y(k) = A\,(0.5)^k + B\,(0.3)^k$

The generalized form of the input signal contains some parameters that we must solve for. To find the parameters, substitute the particular solution into the difference equation. Unlike when solving the homogeneous equation, we can immediately solve for the parameters in the particular solution.

Example. Find a particular solution to the difference equation

$$(E - 0.3)\, y(k) = 3 \cdot 2^k$$

Using a similar form as the input signal, let's try $y(k) = A\, 2^k$. Substituting this into the difference equation gives

$$3 \cdot 2^k = (E - 0.3)(A\, 2^k) = A\, 2^{k+1} - 0.3A\, 2^k$$

While this equation depends on k, we can divide both sides by 2^k to eliminate time from the equation:

$$3 = 2A - 0.3A$$

We can now solve for the coefficient to obtain $A = 3/1.7$. Therefore, the particular solution is

$$y(k) = \frac{3}{1.7}\, 2^k$$

A troublesome example

Find the particular solution to the difference equation

$$(E - 1)(E - 2)\, y(k) = 2^k$$

As before, let's try using a generalized form of the input signal for the particular solution, so $y(k) = A\, 2^k$. Substituting this into the difference equation,

$$2^k = (E^2 - 3E + 2)\,(A\, 2^k) = A\, 2^{k+2} - 3A\, 2^{k+1} + 2A\, 2^k$$

Once again, we can divide by 2^k to eliminate k from the equation,

$$1 = 4A - 6A + 2A = 0$$

But this says that $1 = 0$, which is a contradiction! To see what happened, let's think about the solution to the homogeneous equation:

$$y_h(k) = m_1 + m_2\, 2^k$$

The form that we chose for the particular solution already appears in the homogeneous solution! So our chosen form for the particular solution of $A\, 2^k$ solves the *homogeneous* difference equation, where the right-hand side is zero instead of 2^k as desired. To fix this, we need to include an extra factor of k in the particular solution (similar to when we were solving the homogeneous equation and the characteristic polynomial had repeated roots). Setting $y(k) = Ak\, 2^k$ and substituting this into the difference equation, we now get

$$2^k = A(k+2)2^{k+2} - 3A(k+1)2^{k+1} + 2Ak2^k$$

Dividing both sides by 2^k, all of the terms with k cancel and we obtain

$$1 = 4A(k+2) - 6A(k+1) + 2Ak = 2A$$

which implies that $A = 1/2$. The particular solution is then

$$y(k) = \tfrac{k}{2}\, 2^k$$

Example. Find the particular solution to the difference equation

$$(E-1)(E-2)\,y(k) = 2^k + 3^k$$

Learning from the previous example, we should use the form of the homogeneous solution when choosing the form of the particular solution. The solution to the homogeneous equation is

$$y_h(k) = m_1 + m_2\,2^k$$

so the particular solution should have the form

$$y_p(k) = Ak\,2^k + B\,3^k$$

where we used an extra factor of k in the 2^k term since it already appears in the homogeneous solution. Substituting this into the difference equation,

$$\begin{aligned} 2^k + 3^k &= (E^2 - 3E + 2)\big(Ak\,2^k + B\,3^k\big) \\ &= A\,(k+2)\,2^{k+2} - 3A\,(k+1)\,2^{k+1} + 2Ak\,2^k + B\,3^{k+2} - 3B\,3^{k+1} + 2B\,3^k \end{aligned}$$

Collecting like terms,

$$2^k + 3^k = (4A - 6A + 2A)\,k\,2^k + (8A - 6A)\,2^k + (9B - 9B + 2B)\,3^k$$

In order for these to be equal for all k, the coefficients in front of like terms must be equal. The first term on the right-hand side is zero, which it should be since there is no $k\,2^k$ term on the left-hand side. From the 2^k terms, we get the equation $2A = 1$, and from the 3^k terms we get $2B = 1$. Therefore, the particular solution is

$$y_p(k) = \tfrac{k}{2}\,2^k + \tfrac{1}{2}\,3^k$$

Example. Find the particular solution to the difference equation

$$(E-1)^2\,y(k) = 3$$

The homogeneous solution has the form

$$y_h(k) = m_1 + m_2\,k$$

Since the form of the input (constant) already appears in the homogeneous solution, we should multiply by k. But this also appears in the homogeneous solution, so we need to multiply by k again. The form of the particular solution is then

$$y_p(k) = Ak^2$$

To find the coefficient, we can substitute the particular solution into the difference equation to obtain

$$3 = (E-1)^2\,(Ak^2) = A(k+2)^2 - 2A(k+1)^2 + Ak^2$$

Collecting like terms,

$$3 = (A - 2A + A)\,k^2 + (4A - 4A)\,k + (4A - 2A) = 2A$$

All of the k and k^2 terms on the right-hand side cancel out, which is expected since they are in the homogeneous solution. Solving gives $A = 3/2$, so the particular solution is

$$y_p(k) = \tfrac{3}{2}\,k^2$$

11.4 The total response

Now that we know how to find the particular solution, we can solve for the total response by summing the homogeneous and particular solutions and then applying the initial conditions.

Example. Solve the difference equation

$$(E - 0.3)\, y(k) = 3 \cdot 2^k, \qquad y(0) = 0$$

We already found that the homogeneous and particular solutions are

$$y_h(k) = m\,(0.3)^k \qquad \text{and} \qquad y_p(k) = \frac{3}{1.7}\,2^k$$

Summing the homogeneous and particular solutions, the total response is

$$y(k) = m\,(0.3)^k + \frac{3}{1.7}\,2^k$$

We can now apply the initial condition,

$$0 = y(0) = m + \frac{3}{1.7}$$

which implies that the parameter is $m = -3/1.7$. Therefore, the total response is

$$y(k) = \frac{3}{1.7}\Big(2^k - (0.3)^k\Big)$$

Example. Solve the difference equation

$$(E-1)(E-2)\,y(k) = 2^k + 3^k \qquad y(0)=1,\ y(1)=1$$

Since we already found the homogeneous and particular solutions, the total response is the sum

$$y(k) \quad = \quad \underbrace{m_1 + m_2\,2^k}_{\text{homogeneous}} \quad + \quad \underbrace{\tfrac{k}{2}\,2^k + \tfrac{1}{2}\,3^k}_{\text{particular}}$$

Since this is a second-order system, we have two initial conditions and two parameters. Applying the initial conditions yields the linear system of equations

$$\begin{aligned} 1 &= y(0) = m_1 + m_2 + \tfrac{1}{2} \\ 1 &= y(1) = m_1 + 2m_2 + 1 + \tfrac{3}{2} \end{aligned}$$

Let's see how to solve this system of linear equations two different ways.

- The most straightforward approach is to pick one of the equations, solve for one of the parameters, substitute the solution into the other equation, and then solve for the other parameter. It does not matter which equation we choose to start with, so let's pick the first one.

  $$1 = m_1 + m_2 + \tfrac{1}{2}$$

  It also does not matter which variable we solve for first, so let's solve for m_1 to obtain

  $$m_1 = \tfrac{1}{2} - m_2$$

  Now substituting this into the second equation, we get

  $$1 = \underbrace{\left(\tfrac{1}{2} - m_2\right)}_{m_1} + 2m_2 + 1 + \tfrac{3}{2}$$

  This equation now depends on only m_2, so we can solve it to obtain $m_2 = -2$. Now that we know m_2, we can substitute this back into our equation for m_1 to get

  $$m_1 = \tfrac{1}{2} - m_2 = \tfrac{5}{2}$$

  Therefore, the solution to the system of equation is $m_1 = -2$ and $m_2 = \frac{5}{2}$.

- Instead of solving the system of equations by hand, we can set this up in matrix form as

  $$\begin{bmatrix} \frac{1}{2} \\ -\frac{3}{2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

  Inverting this matrix (using a computer) gives the solution

  $$\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{2} \\ -\frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{5}{2} \\ -2 \end{bmatrix}$$

  which is the same as we found by directly solving by hand. This high-level approach is more easily done using a computer.

Therefore, the solution to the difference equation is

$$y(k) = \tfrac{5}{2} + \left(-2 + \tfrac{k}{2}\right) 2^k + \tfrac{1}{2}\,3^k$$

11.5 Zero-input response

Two important outputs of a system are the zero-input and zero-state responses.

- The *zero-input response* is the output of the system due to the initial conditions when the input is zero, that is, $u(k) = 0$ for all k.
- The *zero-state response* is the output of the system due to the input signal when the initial conditions are zero (meaning that the system is initially at rest), that is, $y(-1) = y(-2) = \ldots = y(-n) = 0$.

For LTI systems, the total response is the sum of the zero-input and zero-state responses.

$$\text{total response} \quad = \quad \text{zero-input response} \quad + \quad \text{zero-state response}$$

We already have all of the tools needed to find the zero-input and zero-state responses from the difference equation describing the dynamics of the system. The zero-state response is found by setting the initial conditions to zero, while the zero-input response is found by setting the input signal to zero and then solving the corresponding difference equation.

Since the zero-input response is the solution to the homogeneous difference equation, it has the same form as the homogeneous solution. To find the zero-input response, we will find the homogeneous solution and then apply the initial conditions to find the n parameters (there is no particular solution in this case since the input is zero).

Example. Find the zero-input response of the system described by the difference equation

$$y(k+2) - 3\,y(k+1) + 2\,y(k) = u(k), \qquad y(0) = 2, \quad y(1) = 1$$

Since we are finding the zero-input response, the input is $u(k) = 0$. Write the homogeneous difference equation using the advance operator E.

$$(E^2 - 3E + 2)\,y(k) = 0$$

Factor the characteristic polynomial.

$$(E-1)(E-2)\,y(k) = 0$$

Since the characteristic polynomial has two real distinct roots, the general solution is then of the form

$$y(k) = m_1\,(1)^k + m_2\,(2)^k = m_1 + m_2\,2^k$$

To find the coefficients, apply the initial conditions.

$$\begin{aligned} 2 &= y(0) = m_1 + m_2 \\ 1 &= y(1) = m_1 + 2m_2 \end{aligned}$$

This is a system of two coupled linear equations. One way to solve a system of equations is by picking one of the equations, solving for one of the coefficients, and then substituting this into the other equation. For instance, let's solve the first equation for m_1 to get

$$m_1 = 2 - m_2$$

We are not done yet since this depends on m_2, so let's substitute this expression for m_1 into the second equation:

$$1 = (2 - m_2) + 2m_2$$

This equation only depends on m_2, so we can solve it to get that $m_2 = -1$. We can now back-substitute this into our previous expression for m_1 in terms of m_2 to find that $m_1 = 3$. The zero-input response is then

$$y(k) = 3 - 2^k$$

Example. Find the zero-input response of the system described by the difference equation

$$(E^2 - 2E + 4)\,y(k) = u(k), \qquad y(0) = -1, \quad y(1) = 2$$

The characteristic polynomial does not factor in this case, so we have to resort to the quadratic equation. This gives the two complex roots

$$p = \frac{2 \pm \sqrt{4-16}}{2} = 1 \pm j\sqrt{3}$$

The magnitude and angle of the roots are

$$r = |p| = \sqrt{(1)^2 + \left(\sqrt{3}\right)^2} = 2 \qquad \text{and} \qquad \theta = \angle p = \arctan\left(\sqrt{3}/1\right) = \frac{\pi}{3} = 60^\circ$$

We saw two equivalent forms for the solution to the homogeneous equation when the roots of the characteristic polynomial are complex:

$$y(k) \quad = \quad m\,r^k\,\cos(k\theta + \phi) \quad = \quad r^k\,(m_1 \cos k\theta + m_2 \sin k\theta)$$

Using the first form, the homogeneous solution is of the form

$$y(k) = m\,2^k \cos\left(k\tfrac{\pi}{3} + \phi\right)$$

with parameters m and ϕ. Substituting the initial conditions yields the nonlinear equations

$$\begin{aligned} -1 &= y(0) = m\cos\left(\phi\right) \\ 2 &= y(1) = 2m\cos\left(\tfrac{\pi}{3} + \phi\right) \end{aligned}$$

Applying the trigonometric identity

$$\cos(a \pm b) = \cos a \cos b \mp \sin a \sin b$$

to the second equation produces the system of equations

$$\begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2\cos\left(\frac{\pi}{3}\right) & 2\sin\frac{\pi}{3} \end{bmatrix} \begin{bmatrix} m\cos\phi \\ m\sin\phi \end{bmatrix}$$

This is a linear system of equations in the unknowns $m\cos\phi$ and $m\sin\phi$, which has the solution

$$m\cos\phi = -1 \qquad \text{and} \qquad m\sin\phi = -\sqrt{3}$$

Squaring both equations and summing gives that $m = \pm 2$, where we used the trigonometric identity

$$\cos^2\phi + \sin^2\phi = 1$$

Taking the positive solution $m = 2$, the angle is $\phi = -2\pi/3$. Therefore, the zero-input response is

$$y(k) = 2^{k+1}\cos\left(k\tfrac{\pi}{3} - \tfrac{2\pi}{3}\right).$$

Example (Continued)**.** For the previous example, instead of using all of the trigonometric identities needed to solve the nonlinear system of equations, we can instead express the homogeneous solution in the second form

$$y(k) = 2^k \left[m_1 \cos\!\left(k\tfrac{\pi}{3}\right) + m_2 \sin\!\left(k\tfrac{\pi}{3}\right)\right]$$

with real parameters m_1 and m_2. Substituting in the initial conditions leads to the *linear* system of equations

$$\begin{aligned} -1 &= y(0) = m_1 \\ 2 &= y(1) = m_1 + \sqrt{3}\, m_2 \end{aligned}$$

which has the solution $m_1 = -1$ and $m_2 = \sqrt{3}$. Therefore, the zero-input response is

$$y(k) = 2^k \left[-\cos\!\left(k\tfrac{\pi}{3}\right) + \sqrt{3} \sin\!\left(k\tfrac{\pi}{3}\right)\right].$$

While these expressions are equivalent, the second form is much easier to find since it involves solving a linear system of equations.

11.6 Zero-state response

The zero-state response is the solution when the initial conditions are zero, which means that

$$0 = y(-1) = y(-2) = \cdots = y(-n)$$

One complication in applying these initial conditions is that they all apply to times *before* zero, while the expression for the solution is only valid for times greater than or equal to zero. Therefore, we cannot apply the initial conditions directly. Instead, we will need to iterate the recursion n times using these initial conditions and the input signal to find $y(0), y(1), \dots, y(n-1)$. We can then use these values to solve for the n parameters in the homogeneous solution. The form of the zero-state response will include those of the homogeneous and particular solutions.

Two common zero-state responses are:

- The **step response** is the zero-state response due to a unit step input signal, $u(k) = u_s(k)$.
- The **impulse response** is the zero-state response due to a unit impulse input signal, $u(k) = \delta(k)$. We typically denote the impulse response by $h(k)$.

Example. Find the zero-state response of the system

$$y(k+2) - 0.3\,y(k+1) + 0.02\,y(k) = u_s(k)$$

We first write the difference equation in operator form as

$$(E^2 - 0.3E + 0.02)\,y(k) = u_s(k)$$

Factoring the characteristic polynomial,

$$(E - 0.1)(E - 0.2)\,y(k) = u_s(k)$$

Since the characteristic polynomial has two distinct real roots, the homogeneous solution has the form

$$y_h(k) = m_1\,(0.1)^k + m_2\,(0.2)^k$$

The input signal is constant, which does not appear in the homogeneous solution, so the particular solution is also a constant, $y_p(k) = A$. Substituting this into the difference equation,

$$1 = A - 0.3A + 0.02A = \frac{18}{25}A$$

Therefore, the particular solution is

$$y_p(k) = \frac{25}{18}$$

The total response is the sum of the homogeneous and particular solutions,

$$y(k) = m_1\,(0.1)^k + m_2\,(0.2)^k + \frac{25}{18}$$

which is only valid for $k \geq 0$, so we cannot apply the zero-state conditions $y(-1) = y(-2) = 0$ directly. Instead, we need to iterate the recursion

$$y(k) = 0.3\,y(k-1) - 0.02\,y(k-2) + u_s(k-2)$$

to obtain $y(0) = 0$ and $y(1) = 0$. We can now apply these initial conditions to the total response, which leads to the linear system of equations

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} \frac{25}{18} \\ \frac{25}{18} \end{bmatrix}$$

which has the solution $m_1 = \frac{100}{9}$ and $m_2 = -\frac{25}{2}$. Therefore, the zero-state response due to a unit step input signal is

$$y(k) = \frac{100}{9}\,(0.1)^k - \frac{25}{2}\,(0.2)^k + \frac{25}{18} \qquad \text{for } k \geq 0$$

Example. Find the zero-state response of the system

$$y(k+2) - 0.3\, y(k+1) + 0.02\, y(k) = \delta(k)$$

This is the same system as before, only the input has changed. Therefore, the form of the homogeneous solution is the same as before. The input signal is now an impulse, which does not appear in the homogeneous solution, so the particular solution is also an impulse, $y_p(k) = A\, \delta(k)$. Substituting this into the difference equation, we have

$$A\, \delta(k+2) - 0.3A\, \delta(k+1) + 0.02A\, \delta(k) = \delta(k)$$

Keep in mind that this equation only holds for $k \geq 0$, so the first two impulses never affect the equation! If we use any $k > 0$, the equation is vacuous since all of the terms on both sides are zero. So the only non-trivial equation comes from $k = 0$, in which case we find that $A = 50$. The total response is the sum of the homogeneous and particular solutions,

$$y(k) = m_1\, (0.1)^k + m_2\, (0.2)^k + 50\, \delta(k)$$

Once again, this difference equation is only valid for $k \geq 0$, so we cannot apply the zero-state conditions $y(-1) = y(-2) = 0$ directly. Instead, we need to iterate the recursion

$$y(k) = 0.3\, y(k-1) - 0.02\, y(k-2) + \delta(k-2)$$

to obtain $y(0) = 0$ and $y(1) = 0$. We can now apply these initial conditions to the total response, which leads to the linear system of equations

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} 50 \\ 0 \end{bmatrix}$$

which has the solution $m_1 = -100$ and $m_2 = 50$. Therefore, the zero-state response due to an impulse input signal is

$$y(k) = -100\, (0.1)^k + 50\, (0.2)^k + 50\, \delta(k) \qquad \text{for } k \geq 0$$

Example. Find the step responses of the linear time-invariant system

$$y(k) = \frac{2E^2 + 3E + 4}{E^2 + 3E + 2}\, u(k)$$

First, let's multiply by the denominator to get the difference equation

$$(E^2 + 3E + 2)\, y(k) = (2E^2 + 3E + 4)\, u(k)$$

or equivalently,

$$y(k+2) + 3\, y(k+1) + 2\, y(k) = 2\, u(k+2) + 3\, u(k+1) + 4\, u(k)$$

Shifting the time index and writing this as a recursion,

$$y(k) = -3\, y(k-1) - 2\, y(k-2) + 2\, u(k) + 3\, u(k-1) + 4\, u(k-2)$$

We first find the solution to the homogeneous equation. Factoring the characteristic polynomial, the homogeneous equation is

$$(E+1)(E+2)\, y(k) = 0$$

which has the solution

$$y_h(k) = m_1\, (-1)^k + m_2\, (-2)^k$$

For the step response, the input signal is a unit step, $u(k) = u_s(k)$. The particular solution is then also a unit step, $y_p(k) = A\, u_s(k)$. Substituting this into the difference equation,

$$A + 3A + 2A = 2 + 3 + 4$$

which implies that $A = 3/2$, so the total response is

$$y(k) = m_1\, (-1)^k + m_2\, (-2)^k + \tfrac{3}{2} \qquad \text{for } k \geq 0$$

To find the parameters, we need to iterate the recursion using zero initial conditions. Doing so, we find that the output at time $k = 0$ is

$$y(0) = -3\, \underbrace{y(-1)}_{0} - 2\, \underbrace{y(-2)}_{0} + 2\, \underbrace{u_s(0)}_{1} + 3\, \underbrace{u_s(-1)}_{0} + 4\, \underbrace{u_s(-2)}_{0} = 2$$

and the output at time $k = 1$ is

$$y(1) = -3\, \underbrace{y(0)}_{2} - 2\, \underbrace{y(-1)}_{0} + 2\, \underbrace{u_s(1)}_{1} + 3\, \underbrace{u_s(0)}_{1} + 4\, \underbrace{u_s(-1)}_{0} = -1$$

Applying these initial conditions to our expression for the total response, we obtain the linear system of equations

$$\begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} 3/2 \\ 3/2 \end{bmatrix}$$

which has the solution $m_1 = -3/2$ and $m_2 = 2$. Therefore, the step response is

$$y(k) = -\tfrac{3}{2}\, (-1)^k + 2\, (-2)^k + \tfrac{3}{2} \qquad \text{for } k \geq 0$$

Example. Find the impulse response of the system in the previous example.

For the impulse response, the input signal is an impulse, $u(k) = \delta(k)$. The homogeneous solution has the same form as before, so we only need to find the particular solution. The particular solution is also an impulse, $y_p(k) = A\,\delta(k)$. Substituting this into the difference equation,

$$A\,\delta(k+2) + 3A\,\delta(k+1) + 2A\,\delta(k) = 2\,\delta(k+2) + 3\,\delta(k+1) + 4\,\delta(k) \qquad \text{for } k \geq 0$$

Note that this difference equation only holds for nonnegative times $k \geq 0$. Substituting $k = 0$, we get that $2A = 4$, so $A = 2$. The total response is then

$$y(k) = m_1\,(-1)^k + m_2\,(-2)^k + 2\,\delta(k) \qquad \text{for } k \geq 0$$

To find the parameters, we need to iterate the recursion using zero initial conditions. Doing so, we find that the output at time $k = 0$ is

$$y(0) = -3\underbrace{y(-1)}_{0} - 2\underbrace{y(-2)}_{0} + 2\underbrace{\delta(0)}_{1} + 3\underbrace{\delta(-1)}_{0} + 4\underbrace{\delta(-2)}_{0} = 2$$

and the output at time $k = 1$ is

$$y(1) = -3\underbrace{y(0)}_{2} - 2\underbrace{y(-1)}_{0} + 2\underbrace{\delta(1)}_{0} + 3\underbrace{\delta(0)}_{1} + 4\underbrace{\delta(-1)}_{0} = -3$$

Applying these initial conditions to our expression for the total response, we obtain the linear system of equations

$$\begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

which has the solution $m_1 = -3$ and $m_2 = 3$. Therefore, the impulse response is

$$y(k) = -3\,(-1)^k + 3\,(-2)^k + 2\,\delta(k) \qquad \text{for } k \geq 0$$

12

Frequency Domain

We can interpret signals as functions of either time or frequency. When representing signals as functions of time, the zero-state response of an LTI system is the convolution of the input signal with the impulse response of the system. While this characterizes the response of the system, we must recompute the convolution for each particular input signal. In this chapter, we interpret signals as functions of frequency. Instead of convolution in the time domain, the zero-state response of an LTI system in the frequency domain will be described by basic multiplication. This simple interpretation for the response of the system will provide significant insight into how the system behaves which can then be used for design.

12.1 Zero-state response due to complex exponentials

To start, let's see how a causal LTI system responds to a complex exponential input signal. Recall that we can always represent a complex number in polar form as $z = r\, e^{j\theta}$ where r is the magnitude and θ the angle of the complex number. Then using Euler's formula, a complex exponential signal has the form

$$z^k = r^k \big(\cos(k\theta) + j\sin(k\theta)\big).$$

The first term is an exponential signal and the second term is a complex sinusoid. If the input signal is $u(k) = z^k$, then the zero-state response is the convolution

$$y(k) = (h * u)(k) = \sum_{m=0}^{\infty} z^{k-m}\, h(m).$$

Splitting up z^{k-m} and pulling z^k outside of the summation (since it is constant with respect to m) gives

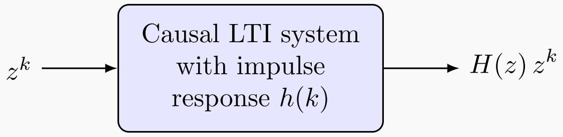
$$y(k) = z^k \sum_{m=0}^{\infty} h(m)\, z^{-m}.$$

Notice that the term inside the summation does not depend on k at all; it only depends on the impulse response $h(k)$ and the complex number z. In fact, the summation is precisely the z-transform of the impulse response,

$$H(z) = \sum_{m=0}^{\infty} h(m)\, z^{-m}.$$

Then the zero-state response is $y(k) = H(z)\, z^k$, which is the input scaled by the complex number $H(z)$.

Zero-state response due to complex exponentials. The zero-state response of a causal LTI system with impulse response $h(k)$ due to a complex exponential signal z^k is the same complex exponential signal scaled by the complex number $H(z)$.

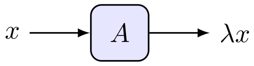


where $\quad H(z) = \sum_{k=0}^{\infty} h(k)\, z^{-k}$

Remark (Connection to linear algebra)**.** The fact that LTI systems scale complex exponential signals is connected with a fundamental concept from linear algebra: eigenvalues and eigenvectors. Recall that an eigenvalue $\lambda \neq 0$ and eigenvector x for a matrix A are such that

$$Ax = \lambda x,$$

which can also be represented by the block diagram:



The interpretation of this is that the matrix A maps the eigenvector x to itself scaled by the eigenvalue. Eigenvectors are directions in which multiplication by the matrix does not change, and the scaling of the eigenvector is the eigenvalue. We can think of signals as infinite-dimensional vectors and linear systems as infinite-dimensional matrices that map the input signal to the output signal. With this interpretation, the complex exponential signal z^k is an eigenvector of the system with eigenvalue $H(z)$.

Complex exponential signals are eigenvectors of LTI systems.

The remarkable property of LTI systems is that *every* complex exponential signal is an eigenvector for *any* LTI system. This is certainly not true in general for finite-dimensional matrices, which typically have different eigenvectors. ■

12.2 Zero-state response

We previously found the zero-state response of a causal LTI system due to a complex exponential signal. We now generalize this to find the zero-state response due to an arbitrary input signal using the z-transform.

Recall that the zero-state response of a causal LTI system with impulse response $h(k)$ is the convolution of the impulse response with the input signal,

$$y(k) = (h * u)(k) = \sum_{m=0}^{\infty} h(k-m)u(m).$$

By definition, the z-transform $Y(z)$ of the output $y(k)$ is

$$Y(z) = \sum_{k=0}^{\infty} y(k)\, z^{-k}.$$

Substituting the expression for the output as the convolution of the input with the impulse response,

$$Y(z) = \sum_{k=0}^{\infty} \left(\sum_{m=0}^{\infty} h(k-m)\, u(m) \right) z^{-k}.$$

Switching the order of the summations,

$$Y(z) = \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} h(k-m)\, u(m)\, z^{-k}$$

where we started the second summation at $k = m$ since the impulse response $h(k-m)$ is zero for $k < m$ since the system is causal. Rewriting z^{-k} as $z^{-(k-m)} z^{-m}$ and rearranging terms,

$$Y(z) = \sum_{m=0}^{\infty} \left(\sum_{k=m}^{\infty} h(k-m)\, z^{-(k-m)} \right) u(m)\, z^{-m}.$$

To simplify the term in parentheses, let $\ell = k - m$ so that

$$Y(z) = \sum_{m=0}^{\infty} \left(\sum_{\ell=0}^{\infty} h(\ell)\, z^{-\ell} \right) u(m)\, z^{-m}.$$

It is now clear that the term inside the parentheses is $H(z)$, the z-transform of the impulse response $h(k)$. Since this does not depend on m, we can pull it out of the first summation to have

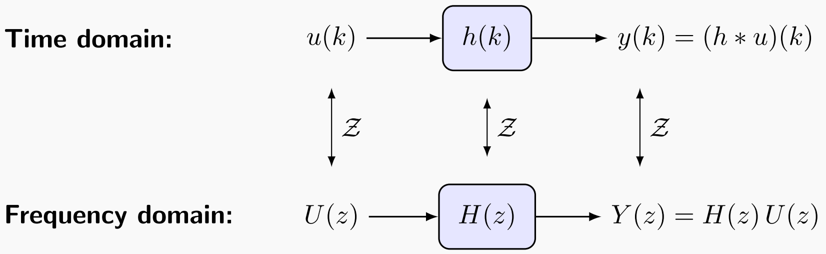
$$Y(z) = H(z) \sum_{m=0}^{\infty} u(m)\, z^{-m}.$$

The summation is $U(z)$, the z-transform of the input signal $u(k)$. Therefore, the z-transform of the zero-state response is the product of the z-transform of the impulse response with the z-transform of the input signal,

$$Y(z) = H(z)\, U(z).$$

Finding the zero-state response required convolution in the time domain, but this becomes simple multiplication in the frequency domain (in terms of z-transforms).

Zero-state response. In the time domain, the zero-state response is the convolution of the input signal with the impulse response. The corresponding relationship in the frequency domain is that the z-transform of the zero-state response is the product of the z-transform of the impulse response with the z-transform of the input signal.



Example. Consider an LTI system with finite-duration impulse response $h(k) = \{0, 2, 1, 0, 0, \ldots\}$. Find the zero-state response due to the finite-duration input signal $u(k) = \{1, 2, 3, 0, 0, \ldots\}$.

To illustrate the connection by the time domain and frequency domain, we will find the zero-state response using both convolution and the z-transform.

- **Convolution.** Since both the impulse response and input signal are finite in duration, the convolutional sum is the finite summation

$$y(k) = \sum_{m=0}^{\infty} h(k-m)\, u(m) = u(0)\, h(k) + u(1)\, h(k-1) + u(2)\, h(k-2).$$

  Substituting the values for the signals, we have that

$$\begin{aligned} y(k) &= 1 \cdot \{0, 2, 1, 0, 0, 0, 0, \ldots\} \\ &\quad + 2 \cdot \{0, 0, 2, 1, 0, 0, 0, \ldots\} \\ &\quad + 3 \cdot \{0, 0, 0, 2, 1, 0, 0, \ldots\}. \end{aligned}$$

  Summing over each time gives $y(k) = \{0, 2, 5, 8, 3, 0, 0, \ldots\}$.

- z-**Transform.** We now find the zero-state response using the z-transform. The z-transform of the input signal is

$$U(z) = u(0) + u(1)\, z^{-1} + u(2)\, z^{-2} = \frac{z^2 + 2z + 3}{z^2}$$

  and the z-transform of the impulse response is

$$H(z) = h(0) + h(1)\, z^{-1} + h(2)\, z^{-2} = \frac{2z + 1}{z^2}.$$

  The z-transform of the zero-state response is then the product of these two,

$$Y(z) = H(z)\, U(z) = \frac{2z+1}{z^2} \cdot \frac{z^2 + 2z + 3}{z^2} = \frac{2z^3 + 5z^2 + 8z + 3}{z^4} = \frac{2}{z} + \frac{5}{z^2} + \frac{8}{z^3} + \frac{3}{z^4}.$$

  We can then simply read off the coefficients of the zero-state response, $y(k) = \{0, 2, 5, 8, 3, 0, 0, \ldots\}$.

Example. Find the zero-state response of the LTI system with impulse response $h(k) = (0.5)^k\, u_s(k)$ due to the input signal $u(k) = 3^k\, u_s(k)$.

Using a table to take z-transforms of the signals, we have that

$$H(z) = \frac{z}{z-0.5} \qquad \text{and} \qquad U(z) = \frac{z}{z-3}.$$

The z-transform of the zero-state response is then

$$Y(z) = H(z)\,U(z) = \frac{z^2}{(z-0.5)(z-3)} = \frac{Az}{z-0.5} + \frac{Bz}{z-3} + C.$$

To find $y(k)$ we need to take the inverse z-transform. Using the form of the partial fraction expansion above, we can find the coefficient A by multiplying both sides by $z-0.5$ and then setting $z = 0.5$ to obtain

$$(z-0.5)\,Y(z)\big|_{z=0.5} \quad = \quad \frac{(0.5)^2}{0.5-3} \quad = \quad 0.5A$$

which implies that $A = -1/5$. Similarly,

$$(z-3)\,Y(z)\big|_{z=3} \quad = \quad \frac{3^2}{3-0.5} \quad = \quad 3B$$

which implies that $B = 6/5$. Setting $z = 0$, we find that $C = 0$. Therefore, the zero-state response is

$$y(k) = \frac{6}{5}\,(3)^k - \frac{1}{5}\,(0.5)^k \quad \text{for } k \ge 0.$$

Example. Find the step response of the LTI system with impulse response

$$h(k) = \left[50\,(0.2)^k - 100\,(0.1)^k\right] u_s(k-1)$$

To use the table of z-transform pairs, we first rewrite the impulse response so that the exponents are $k-1$ to match the unit step,

$$h(k) = \left[10\,(0.2)^{k-1} - 10\,(0.1)^{k-1}\right] u_s(k-1).$$

We can now use a table to take z-transforms of the signals to find that

$$H(z) = \frac{10}{z-0.2} - \frac{10}{z-0.1} = \frac{1}{(z-0.2)(z-0.1)}.$$

Since we are finding the step response, the input is a step signal $u(k) = u_s(k)$ whose z-transform is $U(z) = \frac{z}{z-1}$. The z-transform of the zero-state response is then

$$Y(z) = H(z)\,U(z) = \frac{z}{(z-1)(z-0.1)(z-0.2)} = \frac{Az}{z-1} + \frac{Bz}{z-0.1} + \frac{Cz}{z-0.2} + D.$$

To find $y(k)$ we need to take the inverse z-transform. Using the form of the partial fraction expansion above, we find that the coefficients are $A = 25/18$, $B = 100/9$, $C = -25/2$, and $= 0$. Therefore, the zero-state response is

$$y(k) = \frac{25}{18} + \frac{100}{9}\,(0.1)^k - \frac{25}{2}\,(0.2)^k \quad \text{for } k \ge 0.$$

12.3 Transfer function

So far, we have used the symbol H to denote two quantities:

- $H(E)$ is the transfer function of an LTI system, which is a function of the advance operator E.
- $H(z)$ is the z-transform of the impulse response $h(k)$, which is a function of the complex number z.

In this section, we describe the relationship between these two quantities. We have been using the same symbol for both (a slight abuse of notation), which hints that there is a strong relationship between them. As we will see, $H(E)$ and $H(z)$ have the same expression and are therefore *both* a representation of the transfer function of the system.

The transfer function and z-transform of the impulse response

To understand this relationship, let's consider a generic second-order difference equation,

$$y(k+2) + a\,y(k+1) + b\,y(k) = c\,u(k+2) + d\,u(k+1) + e\,u(k).$$

The transfer function relates the input to the output by $y(k) = H(E)\,u(k)$ and is given by

$$H(E) = \frac{cE^2 + dE + e}{E^2 + aE + b}.$$

Now let's find $H(z)$ for this system. While $H(z)$ is the z-transform of the impulse response, this is not easy to calculate for our generic second-order system. Instead, we can use the fact that the zero-state response is the product of the z-transforms of the impulse response and input, $Y(z) = H(z)\,U(z)$. If we can find $Y(z)$ and $U(z)$, then we can solve this equation for $H(z)$. To find the z-transforms, let's first shift the difference equation in time to obtain

$$y(k) + a\,y(k-1) + b\,y(k-2) = c\,u(k) + d\,u(k-1) + e\,u(k-2).$$

Now take the z-transform of this equation and use that the initial conditions are zero (since $y(k)$ is the zero-state response) to obtain

$$\Big(1 + \frac{a}{z} + \frac{b}{z^2}\Big)\,Y(z) = \Big(c + \frac{d}{z} + \frac{e}{z^2}\Big)\,U(z)$$

Solving for the zero-state response gives

$$Y(z) = \frac{cz^2 + dz + e}{z^2 + az + b}\,U(z)$$

which implies that the z-transform of the impulse response is

$$H(z) = \frac{cz^2 + dz + e}{z^2 + az + b}.$$

This has the same form as $H(E)$, just with the advance operator E replaced with the complex number z. For this reason, we refer to both $H(E)$ and $H(z)$ as the transfer function of the system. This is summarized as follows.

The transfer function of an LTI system has two interpretations.

- The transfer function $H(E)$ is the operator such that the zero-state response is $y(k) = H(E)\,u(k)$.
- The transfer function $H(z)$ is the z-transform of the impulse response $h(k)$.

While both $H(E)$ and $H(z)$ have the same expression, one is a function of the advance operator E while the other is a function of the complex number z.

$$h(k) \quad \xleftrightarrow{\mathcal{Z}} \quad H(z) \quad = \quad H(E)\big|_{E\to z}$$

Using this equivalence between $H(E)$ and $H(z)$, we can now find the impulse response of an LTI system by taking the inverse z-transform of its transfer function (instead of solving the difference equation).

Example. Find the impulse response of the system

$$y(k+1) - a\,y(k) = u(k+1).$$

We could find the impulse response by solving the difference equation. It is typically easier, however, to find the impulse response by computing the inverse z-transform of the transfer function. In this case, the transfer function is $H(E) = \frac{E}{E-a}$, so the z-transform of the impulse response is $H(z) = \frac{z}{z-a}$. Taking the inverse z-transform gives the impulse response $h(k) = a^k\,u_s(k)$.

Transfer function of a difference equation

From our previous discussion, we found that the difference equation

$$y(k+2) + a\,y(k+1) + b\,y(k) = c\,u(k+2) + d\,u(k+1) + e\,u(k)$$

has the transfer function

$$H(z) = \frac{cz^2 + dz + e}{z^2 + az + b}.$$

The coefficients on the input signal in the difference equation are the coefficients of the numerator polynomial, while the coefficients on the output signal in the difference equation are the coefficients of the denominator polynomial. Therefore, we can construct the transfer function directly from the coefficients of the difference equation. Conversely, given a transfer function, we can use the coefficients of its numerator and denominator polynomials to directly write down the corresponding difference equation. This relationship is summarized as follows.

A discrete-time LTI system described by the n^{th}-order difference equation

$$y(k) + \sum_{i=1}^{n} a_i\,y(k-i) = \sum_{j=0}^{m} b_j\,u(k-j)$$

has transfer function

$$H(z) = \frac{\displaystyle\sum_{j=0}^{m} b_j\,z^{-j}}{1 + \displaystyle\sum_{i=1}^{n} a_i\,z^{-i}}$$

Example. The discrete-time LII system described by the difference equation

$$y(k+1) - a\,y(k) = u(k+1)$$

has the transfer function (and impulse response)

$$h(k) = a^k\,u_s(k) \quad \xleftrightarrow{\mathcal{Z}} \quad H(z) = \frac{z}{z-a}.$$

12.4 System response

We have seen that the zero-state response is represented in the frequency domain as $Y(z) = H(z)\,U(z)$ where $H(z)$ is the transfer function. But this only describes the part of the response due to the input signal. To find the complete response (including the part due to the initial conditions), we can use the following procedure.

To find the total response of an LTI system using the z-transform,

a) take the z-transform of the difference equation using the advance/delay properties of the z-transform

b) solve for $Y(z)$

c) compute the inverse z-transform to obtain the response $y(k)$

Example. Solve the difference equation

$$y(k+1) - 0.2\,y(k) = u(k+1)$$

with initial condition $y(0) = 5$ and input signal $u(k) = u_s(k)$.

Recall the advance property of the z-transform,

$$y(k+1) \quad \overset{\mathcal{Z}}{\longleftrightarrow} \quad z\,Y(z) - z\,y(0).$$

Taking the z-transform of the difference equation and using this property twice, we have that

$$z\,Y(z) - z\,y(0) - 0.2\,Y(z) = z\,U(z) - z\,u(0).$$

Solving for the response gives

$$Y(z) \quad = \quad \underbrace{\frac{z}{z-0.2}\big(U(z) - u(0)\big)}_{\text{zero-state response}} \quad + \quad \underbrace{\frac{z\,y(0)}{z-0.2}}_{\text{zero-input response}}$$

The first term is the zero-state response, which is zero when the initial condition is zero, and the second term is the zero-input response which is zero when the input is zero. It is not necessary to separate the terms like this; we only did it here to illustrate the different components of the response. This form also highlights the transfer function which is the term that multiplies $U(z)$, so $H(z) = \frac{z}{z-0.2}$. From this, we can tell that the impulse response is $h(k) = (0.2)^k\,u_s(k)$, the system has a zero at $z = 0$ and a pole at $z = 0.2$, and the system is stable.

Returning to solving the difference equation, let's substitute in values. Since the input is a unit step, its z-transform is $U(z) = \frac{z}{z-1}$. Substituting this and the initial values $u(0) = 1$ and $y(0) = 5$,

$$Y(z) = \frac{z^2}{(z-1)(z-0.2)} + \frac{4z}{z-0.2}$$

While we could combine these, it is simpler to just find the partial fraction expansion of the first term since the second term is already of a form in the table. The partial fraction expansion is

$$Y(z) = \frac{5}{4}\,\frac{z}{z-1} - \frac{1}{4}\,\frac{z}{z-0.2} + \frac{4z}{z-0.2}.$$

Therefore, the solution to the difference equation is

$$y(k) = \frac{5}{4} + \frac{15}{4}\,(0.2)^k \quad \text{for } k \geq 0$$

Example. Solve the difference equation

$$y(k+1) - 0.3\,y(k) = u(k)$$

with initial condition $y(0) = 0$ and input signal $u(k) = 3\,(2)^k\,u_s(k)$.

Taking the z-transform of the difference equation and using the advance property, we have

$$\big[z\,Y(z) - z\,y(0)\big] - 0.3\,Y(z) = \frac{3z}{z-2}$$

Since the initial condition is zero, the output is

$$Y(z) = \frac{3z}{(z-0.3)(z-2)} = \frac{Az}{z-0.3} + \frac{Bz}{z-2} + C$$

where the right-hand side is a partial fraction expansion.

- To find the coefficient A, multiply by $z - 0.3$ and then set $z = 0.3$.

$$(z-0.3)Y(z)\Big|_{z=0.3} = 0.3A = \frac{0.9}{-1.7} \qquad \Longrightarrow \qquad A = -\frac{3}{1.7}$$

- To find the coefficient B, multiply by $z - 2$ and then set $z = 2$.

$$(z-2)Y(z)\Big|_{z=2} = 2B = \frac{6}{1.7} \qquad \Longrightarrow \qquad B = \frac{3}{1.7}$$

- To find the coefficient C, set $z = 0$ to obtain $C = 0$.

Therefore, the partial fraction expansion is

$$Y(z) = \frac{3}{1.7}\left(\frac{z}{z-2} - \frac{z}{z-0.3}\right)$$

Then using the table of z-transforms, the inverse z-transform is

$$y(k) = \frac{3}{1.7}\Big(2^k - (0.3)^k\Big) \qquad \text{for } k \geq 0$$

Example. Solve the difference equation

$$y(k+2) - 3\,y(k+1) + 2\,y(k) = u(k)$$

with initial condition $y(0) = 1$ and $y(1) = 1$ and input signal $u(k) = (2^k + 3^k)\,u_s(k)$.

Taking the z-transform of the difference equation and using the advance property, we have

$$\left[z^2\,Y(z) - z^2\,y(0) - z\,y(1)\right] - 3\left[z\,Y(z) - z\,y(0)\right] + 2\,Y(z) = \frac{z}{z-2} + \frac{z}{z-3}$$

Substituting the initial condition, we have

$$(z-1)(z-2)\,Y(z) = \frac{z}{z-2} + \frac{z}{z-3} + z\,(z-2)$$

Solving for the output,

$$Y(z) = \frac{z}{(z-1)(z-2)^2} + \frac{z}{(z-1)(z-2)(z-3)} + \frac{z}{z-1}$$

We could get a common denominator to write $Y(z)$ as a rational function of z, but this is unecessary and leads to a complicated expression. Instead, let's just use this expression to find the partial fraction expansion.

$$Y(z) = \frac{Az}{z-1} + \frac{Bz}{z-2} + \frac{Cz}{(z-2)^2} + \frac{Dz}{z-3} + E$$

- To find the coefficient A, multiply by $z-1$ and then set $z=1$.

$$(z-1)\,Y(z)\Big|_{z=1} = A = 1 + \frac{1}{2} + 1 \qquad \Longrightarrow \qquad A = \frac{5}{2}$$

- To find the coefficient C, multiply by $(z-2)^2$ and then set $z=2$.

$$(z-2)^2\,Y(z)\Big|_{z=2} = 2C = 2 \qquad \Longrightarrow \qquad C = 1$$

- To find the coefficient D, multiply by $z-3$ and then set $z=3$.

$$(z-3)\,Y(z)\Big|_{z=3} = 3D = \frac{3}{2} \qquad \Longrightarrow \qquad D = \frac{1}{2}$$

- To find the coefficient E, set $z=0$ to obtain $E=0$.
- To find the coefficient B, set $z=-1$ to obtain

$$Y(-1) = \frac{A}{2} + \frac{B}{3} - \frac{C}{9} + \frac{D}{4} + E$$

  Substituting the values for A, C, D, and E, we find that $B=-2$.

Therefore, the partial fraction expansion is

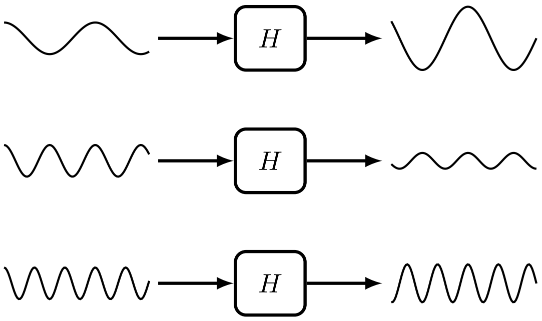
$$Y(z) = \frac{\frac{5}{2}z}{z-1} - \frac{2z}{z-2} + \frac{z}{(z-2)^2} + \frac{\frac{1}{2}z}{z-3}$$

Then using the table of z-transforms, the inverse z-transform is

$$y(k) = \frac{5}{2} - 2\,(2)^k + k\,(2)^{k+1} + \frac{1}{2}(3)^k \qquad \text{for } k \geq 0$$

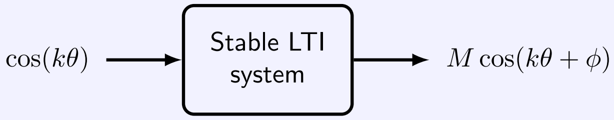
12.5 Frequency response

The frequency response of a stable LTI system characterizes how the system acts on individual frequencies. The frequency response is based on the fact that, for stable LTI systems, sinusoid inputs produce sinusoid outputs.



The output sinusoid has the same frequency, but it may be shifted in time and/or scaled in magnitude.

Definition. For a stable LTI system, the frequency response is the amount that the system scales and shifts sinusoidal input signals.



For each frequency θ, the frequency response magnitude M is the amount that the sinusoid is scaled, and the frequency response phase ϕ is the amount that it is shifted.

We can find the frequency response in terms of the transfer function $H(z)$. First, decompose the sinusoidal input as a sum of complex exponentials,

$$u(k) = \cos(k\theta) = \frac{e^{jk\theta} + e^{-jk\theta}}{2}$$

Recall that $H(z)$ is the amount that the system scales the complex exponential z^k. Since the system is linear, a weighted sum of inputs produces the same weighted sum of the corresponding outputs. Therefore, the output due to the sinusoid is

$$y(k) = \frac{H(e^{j\theta})\, e^{jk\theta} + H(e^{-j\theta})\, e^{-jk\theta}}{2}$$

The second term is the complex conjugate of the first. Since the sum of a complex number and its conjugate is twice its real part (that is, $(a + jb) + (a - jb) = 2a$), we have that

$$y(k) = \mathrm{Re}[H(e^{j\theta})\, e^{jk\theta}]$$

Let M be the magnitude and ϕ the phase of the transfer function evaluated on the unit circle at an angle θ so that $H(e^{j\theta}) = M\, e^{j\phi}$. Then

$$y(k) = \mathrm{Re}[M\, e^{j(k\theta+\phi)}]$$

Applying Euler's formula to the complex exponential and then taking the real part,

$$y(k) = M\cos(k\theta + \phi)$$

Since M and ϕ were the magnitude and phase of the transfer function evaluated on the unit circle at an angle θ, we have the following.

For a stable discrete-time LTI system, the frequency response is the transfer function evaluated on the unit circle in the complex plane:

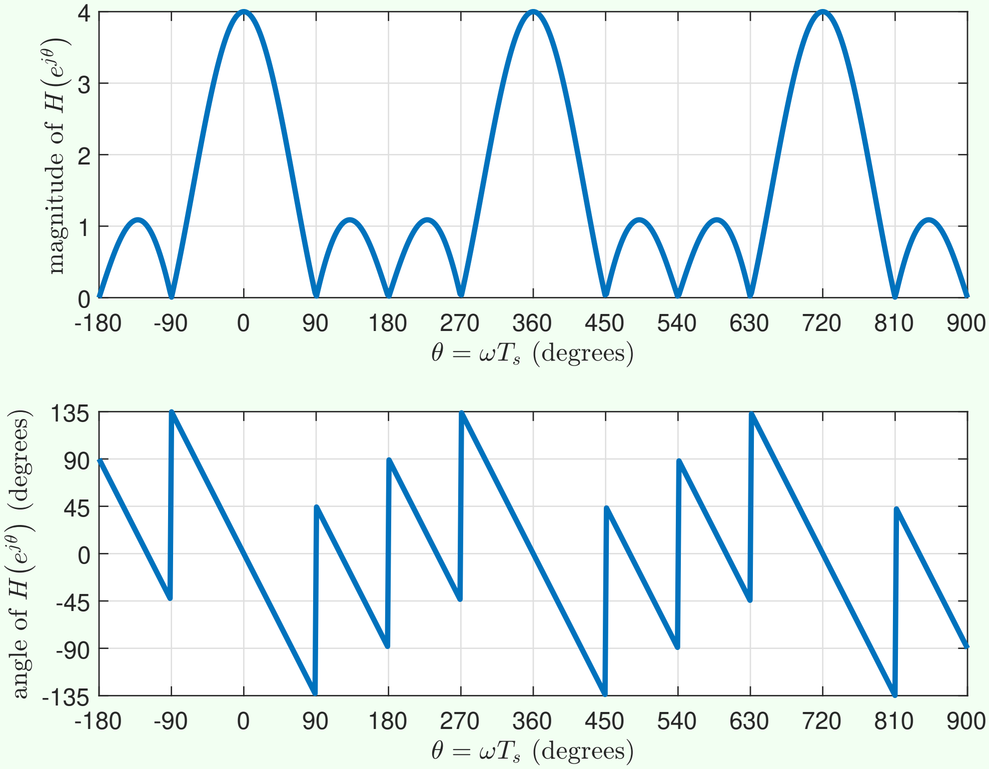
$$\text{Frequency response at } \theta \quad = \quad H(e^{j\theta})$$

The magnitude $M = |H(e^{j\theta})|$ is how much the system scales a sinusoid with frequency θ, and the phase $\phi = \angle H(e^{j\theta})$ is how much the system shifts the sinuoid.

Plotting the frequency response

We can visualize the frequency response by plotting the magnitude and phase as a function of the input frequency.

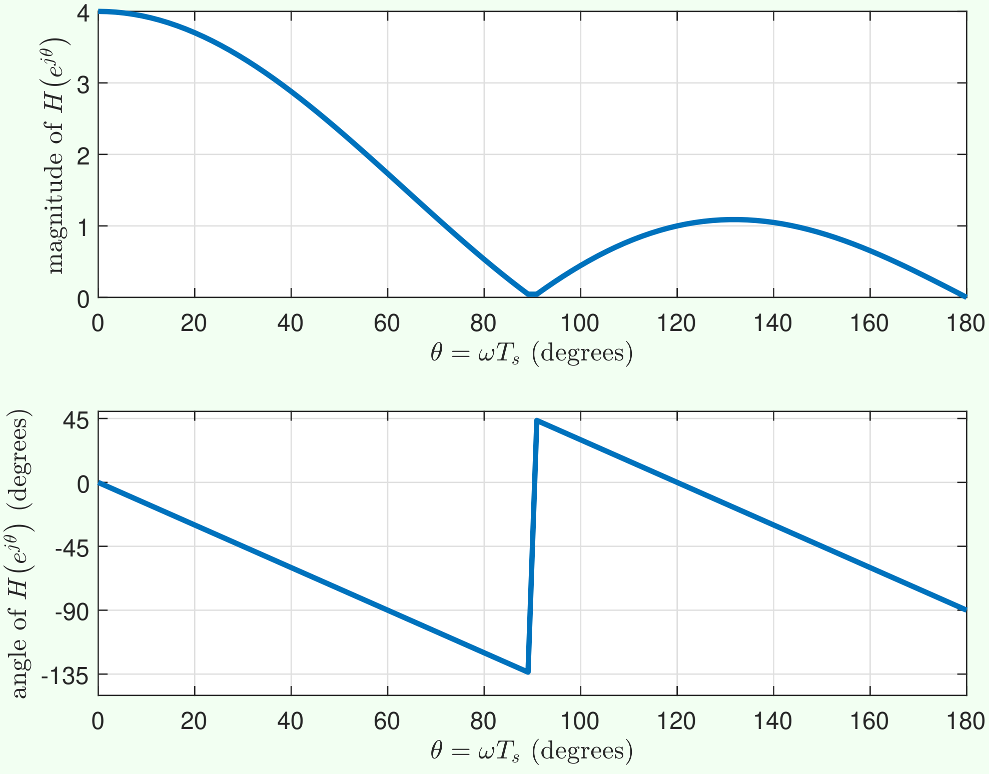
Example. The frequency response of the LTI system with transfer function $H(z) = \frac{(z+1)(z^2+1)}{z^3}$ is as follows.



Comments

- The frequency response is periodic with period 2π since
$$H(e^{j(\theta+2\pi)}) = H(e^{j\theta})$$
- The frequency response is conjugate symmetric (meaning the magnitude is even and the phase is odd) since
$$\overline{H(e^{j\theta})} = M\, e^{-j\phi} = H(e^{-j\theta})$$
- Combining these facts, the frequency response is completely specified by its magnitude and phase on the interval $[0, \pi]$. We therefore typically plot the magnitude M and phase ϕ as functions of θ in the interval $[0, \pi]$.

Example. Going back to the previous example, the frequency response over the interval $[0, \pi]$ is as follows.



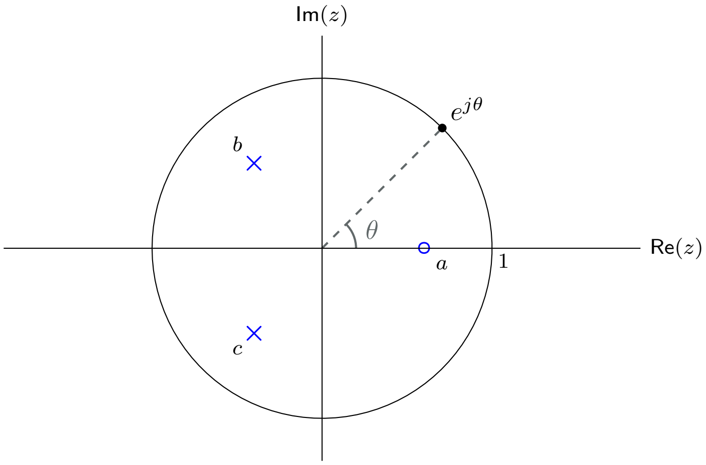
Frequency response from pole-zero plot

The frequency response can be evaluated geometrically (up to a constant) from the pole-zero plot of the transfer function. This provides insight into how the poles and zeros of the transfer function affect the frequency response and is particularly useful for studying frequency-selective filters.

Consider a discrete-time LTI system with transfer function

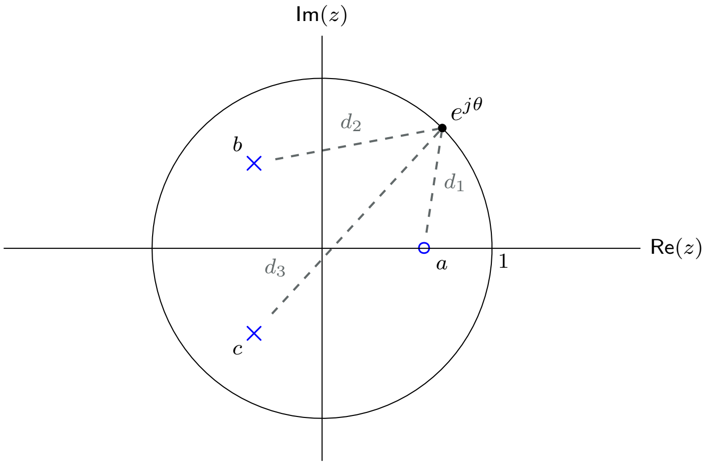
$$H(z) = \frac{z-a}{(z-b)(z-c)}$$

where a is real and b and c are complex conjugates. The pole-zero plot of the transfer function is as follows.



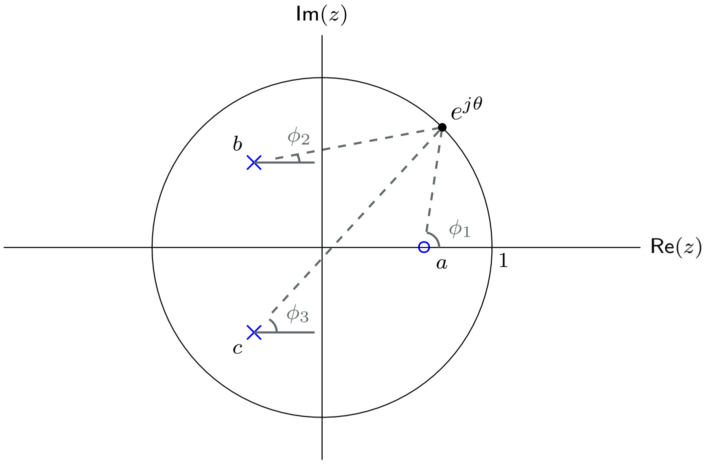
The complex exponential $e^{j\theta}$ is on the unit circle with an angle of θ with the positive real axis. Let d_1, d_2, and d_3 denote the distances from each of the poles and zeros to the complex exponential as shown below. The magnitude of the frequency response at the angle θ is then

$$|H(e^{j\theta}) = \frac{|e^{j\theta} - a|}{|e^{j\theta} - b|\,|e^{j\theta} - c|} = \frac{d_1}{d_2\,d_3}$$



Similarly, let ϕ_1, ϕ_2, and ϕ_3 denote the angles from each of the poles and zeros to the complex exponential as shown below. The phase of the frequency response at the angle θ is then

$$\angle H(e^{j\theta}) = \angle(e^{j\theta} - a) - \angle(e^{j\theta} - b) - \angle(e^{j\theta} - c) = \phi_1 - \phi_2 - \phi_3$$



The frequency response can be found (up to a constant) from the poles and zeros of the transfer function as follows.

$$|H(e^{j\theta})| = K \cdot \frac{\text{product of distances to zeros}}{\text{product of distances to poles}}$$

and

$$\angle H(e^{j\theta}) = \text{sum of angles to zeros} - \text{sum of angles to poles}$$

Based on this geometric interpretation of the frequency response, we make the following observations.

Observation 1. A zero near the unit circle produces a “low spot” (notch) in the frequency response magnitude.

The time-domain interpretation of this is that the zero near the unit circle approximately cancels inputs at that frequency, that is, $N(E)\cos(k\theta) \approx 0$ for θ near the zero, so the input produces a small output.

Example. Consider the system with transfer function $H(z) = \frac{z-1}{z-0.5}$. The input corresponding to the zero at $z = 1$ is a constant, such as $u_s(k)$ which has z-transform $U(z) = \frac{z}{z-1}$. The difference equation for this system is

$$y(k+1) = 0.5\,y(k) + u(k+1) - u(k)$$

Iterating the difference equation,

$$\begin{aligned} y(0) &= \quad 0 + 1 - 0 = 1 \\ y(1) &= \; 0.5 + 1 - 1 = 0.5 \\ y(2) &= 0.25 + 1 - 1 = 0.25 \\ &\vdots \end{aligned}$$

where the input terms cancel due to the zero at $z = 1$.

Observation 2. A pole near the unit circle produces a “high spot” (peak) in the frequency response magnitude.

The time-domain interpretation of this is that the input resonates with the pole near the unit circle to amplify the output at that frequency.

Example. Consider the system with transfer function $H(z) = \frac{z}{z+1}$. The input corresponding to the pole at $z = -1$ is $u(k) = (-1)^k\, u_s(k)$ which has z-transform $U(z) = \frac{z}{z+1}$. The difference equation for this system is

$$y(k+1) = -y(k) + u(k+1)$$

Iterating the difference equation,

$$\begin{aligned} y(0) &= 0 + 1 = 1 \\ y(1) &= -1 - 1 = -2 \\ y(2) &= 2 + 1 = 3 \\ y(3) &= -3 - 1 = -4 \\ &\vdots \end{aligned}$$

The output grows unbounded due to the pole at $z = -1$.

Bode plot

A Bode plot is a particular way of plotting the frequency response.

- The magnitude is plotted on a logarithmic scale in units of decibels.
- The angle is plotted on a linear scale.
- The frequency $\omega = \theta/T$ is plotted on a logarithmic scale, where T is the sampling period.

A *decibel* is a unit of measure that is often used to measure the intensity of sound. Decibels use a logarithmic scale that is useful in analyzing signals over a wide range of amplitudes, from very small (such as 10^{-9}) to very large (such as 10^9).

A unitless gain G is equivalent to $g = 20\log_{10} G$ decibels, or equivalently, a gain of g decibels is . Some common gains and their corresponding value in decibels are shown below.

Gain (unitless)	Decibels
100	40 dB
10	20 dB
1	0 dB
0.7079	-3 dB
0.1	-20 dB
0.01	-40 dB

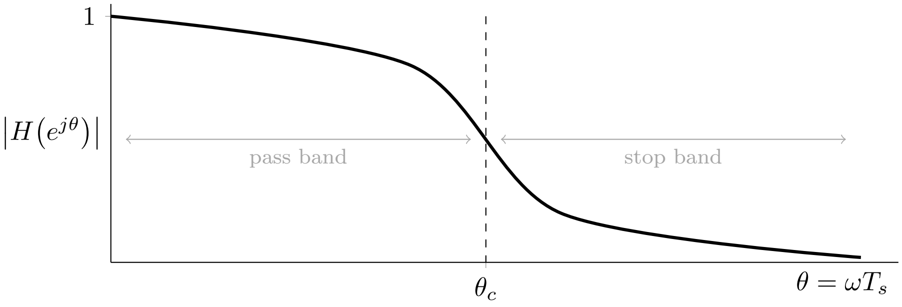
13

Filters

Filtering is the process of changing relative amplitudes of the frequency components of a signal. A common application of filtering is in audio systems, where we may want to adjust the energy at low frequencies (bass) and high frequencies (treble). When playing music through speakers, for instance, we may want to apply an equalizing filter to compensate for the frequency response of the speakers, or we may want to use a lowpass filter to remove high-frequency noise. Frequency-selective filters are designed to select some bands of frequencies and reject others. In this chapter, we study the main types of frequency-selective filters: lowpass, highpass, bandpass, and bandstop filters. We will also see why it is not possible (and often not desireable) to implement an ideal frequency-selective filter.

13.1 Lowpass filter

A lowpass filter is a frequency-selective filter that passes low frequencies and blocks high frequencies. The magnitude of the frequency response for a lowpass filter looks like the following:

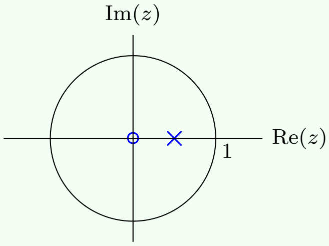


Low frequencies in the pass band are scaled by approximately one and therefore have approximately the same magnitude after filtering. High frequencies in the stop band are scaled by a number close to zero and therefore are attenuated (or reduced) after filtering. The cutoff frequency is the (approximate) frequency that separates the pass band and stop band.

Example (First-order lowpass filter). A simple first-order lowpass filter has the transfer function

$$H(z) = K\,\frac{z}{z-a}$$

which has a zero at the origin and a pole at the parameter a. The pole-zero plot of the transfer function is as follows:



From our intuition about the shape of the frequency response from the locations of the poles and zeros, we should have $0 < a < 1$ for this to be a lowpass filter. The closer a is to one the lower the cutoff frequency will be, and the magnitude of the frequency response will be the largest at $\theta = 0$. Therefore, let's choose the gain K such that the magnitude of the frequency response is one for a constant input (that is, zero frequency). The magnitude of the frequency response at $\theta = 0$ is

$$|H(e^{j0})| = |H(1)| = \frac{K}{1-a}$$

Therefore, we choose $K = 1 - a$. The lowpass filter is then

$$H(z) = \frac{(1-a)z}{z-a}$$

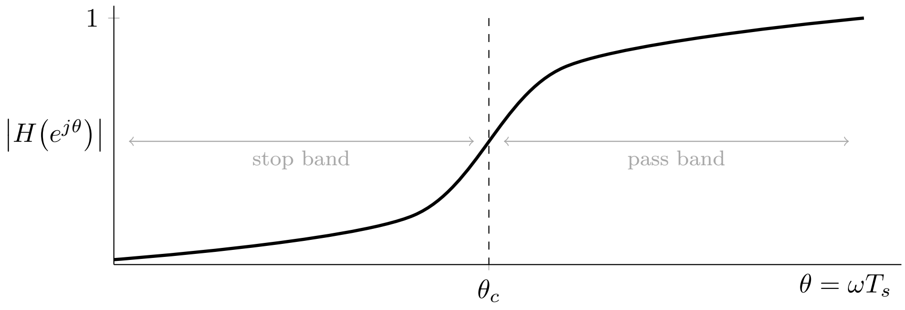
which corresponds to the recursion

$$y(k+1) = a\,y(k) + (1-a)\,u(k+1)$$

The output is a weighted difference of the current input with the previous output, where the weights are positive and sum to one. In the extreme cases, when $a = 0$ the filter passes everything (the output is equal to the input), and when $a = 1$ the filter rejects everything (the output is constant).

13.2 Highpass filter

A highpass filter is a frequency-selective filter that passes high frequencies and blocks low frequencies (the opposite of a lowpass filter). The magnitude of the frequency response for a highpass filter looks like the following:

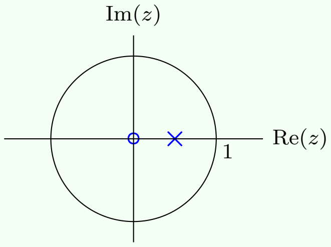


Just like a lowpass filter, a highpass filter also has a pass band, stop band, and cutoff frequency. The only difference is that the pass band consists of all frequencies *above* the cutoff frequency while the stop band consists of all frequencies *below* the cutoff frequency.

Example (First-order highpass filter). A simple first-order highpass filter has the transfer function

$$H(z) = K\,\frac{z-a}{z}$$

where the parameter a of the zero satisfies $0 < a \leq 1$. The pole-zero plot of the transfer function is as follows:



The magnitude of the frequency response is largest when $\theta = \pi$, so let's choose the gain K to be unity at this frequency. Then

$$1 = |H(e^{j\pi})| = |H(-1)| = K\,(1+a)$$

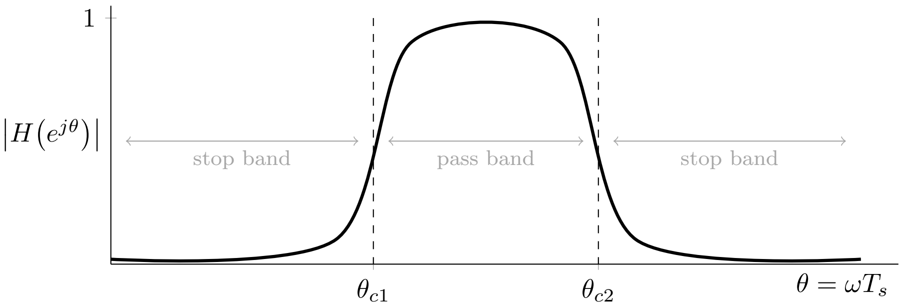
which implies that $K = \frac{1}{1+a}$. The corresponding difference equation is

$$y(k+1) = \frac{1}{1+a}\big(u(k+1) - a\,u(k)\big)$$

Note that this is an FIR filter since all poles of the transfer function are at zero and the difference equation has no feedback! The output is a weighted difference of the two previous inputs, where the weights are positive and sum to one. When $a = 1$, the output is simply $y(k+1) = \frac{1}{2}u(k+1) - \frac{1}{2}u(k)$. This is called the *backward difference* of the input signal and is an approximation to the derivative.

13.3 Bandpass filter

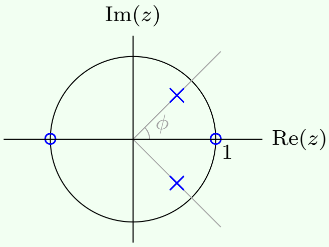
A bandpass filter is a frequency-selective filter that passes an interval of frequencies $[\theta_{c1}, \theta_{c2}]$ and blocks both higher and lower frequencies. The magnitude of the frequency response for a bandpass filter looks like the following:



Example (Second-order bandpass filter). While we have seen first-order highpass and lowpass filters, a bandpass filter must have degree at least two. A simple bandpass filter has the transfer function

$$H(z) = K \frac{z^2 - 1}{z^2 - 2r\cos(\phi)z + r^2}$$

with parameters r and ϕ. The zeros are at ± 1, and the poles are complex conjugates at $r\exp(\pm j\phi)$. The pole-zero plot is as follows:

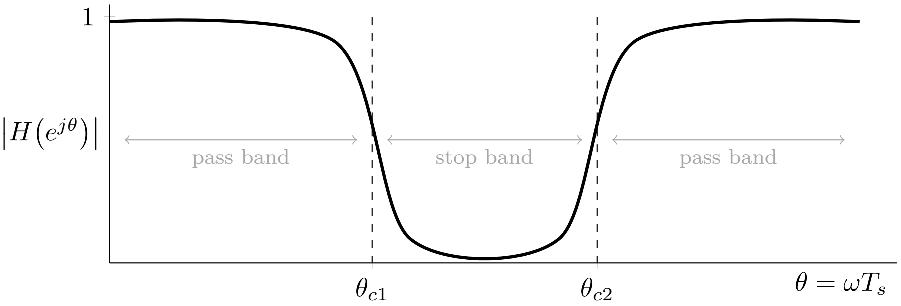


The poles near the unit circle cause a peak in the magnitude of the frequency response at that angle, so the passband of the filter is centered about ϕ. As r approaches one, the poles approach the unit circle which causes the peak in the magnitude of the frequency response to increase.

To design the gain, we may choose K such that the peak of the magnitude of the frequency response is one. Since this occurs at $\theta = e^{j\phi}$, we choose K such that $|H(e^{j\phi})| = 1$.

13.4 Bandstop filter

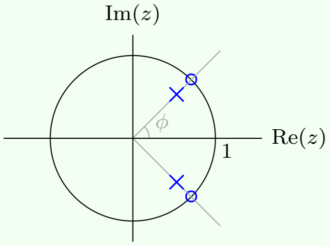
A bandstop filter is a frequency-selective filter that attenuates an interval of frequencies $[\theta_{c1}, \theta_{c2}]$ and passes both higher and lower frequencies. The magnitude of the frequency response for a bandstop filter looks like the following:



Example (Second-order bandstop filter). A simple bandstop filter has the transfer function

$$H(z) = K\,\frac{z^2 - 2\cos(\phi)z + 1}{z^2 - 2r\cos(\phi)z + r^2}$$

with parameters r and ϕ. The zeros are complex conjugates on the unit circle at $\exp(\pm j\phi)$, and the poles are complex conjugates at $r\exp(\pm j\phi)$. The pole-zero plot is as follows:



The zero on the unit circle cause a low spot in the magnitude of the frequency response at that angle, so the stopband of the filter is centered about ϕ. As r approaches one, the poles approach the zeros on the unit circle which causes the width of the stop band to narrow.

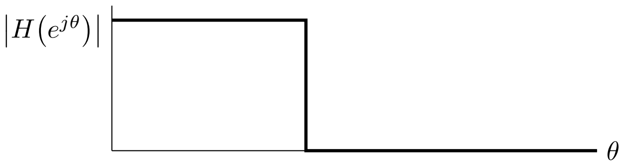
To design the gain, we may choose K such that the magnitude of the frequency response is one at zero frequency, so $|H(1)| = 1$.

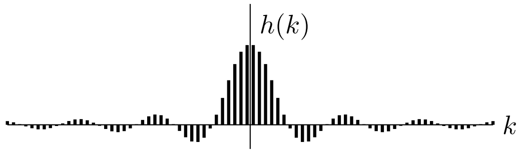
13.5 Ideal filters

You may wonder, why not use an *ideal* frequency-selective filter? An ideal lowpass filter with cutoff frequency θ_c has the frequency response

$$H(e^{j\theta}) = \begin{cases} 1 & \text{if } 0 \le \theta \le \theta_c \\ 0 & \text{otherwise} \end{cases}$$

The ideal filter exactly passes low frequencies and completely blocks high frequencies. But if we compute the inverse z-transform of the ideal filter, we get the following impulse response.





The impulse response is not zero for negative times, so the filter is noncausal and therefore cannot be implemented in real time. However, we typically do not actually want an ideal filter anyway, since the cutoff frequency is often an estimate of the separation between the signal and the noise, with some parts of the noise below this frequency and some parts of the signal above this frequency.

Example (Ideal lowpass filter). Consider the ideal lowpass filter with cutoff frequency θ_c,

$$H\left(e^{j\theta}\right) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq \theta_c, \\ 0 & \text{otherwise.} \end{cases}$$

The inverse discrete-time Fourier transform is

$$h(k) = \frac{1}{2\pi} \int_{2\pi} H\left(e^{j\theta}\right) e^{j\theta k} \, \mathrm{d}\theta = \frac{1}{2\pi} \int_{-\theta_c}^{\theta_c} e^{j\theta k} \, \mathrm{d}\theta.$$

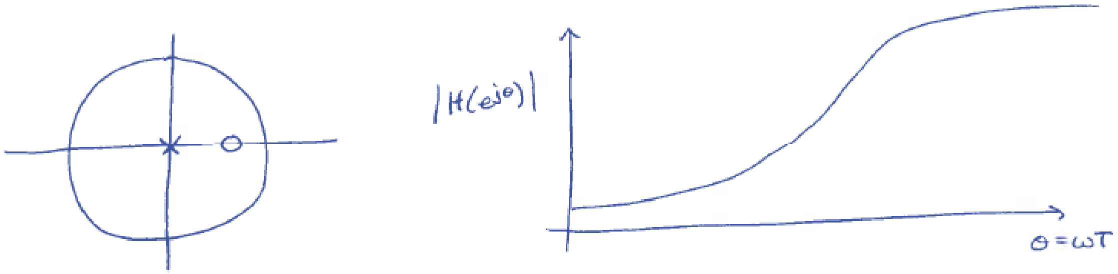
Evaluating the integral yields the impulse response

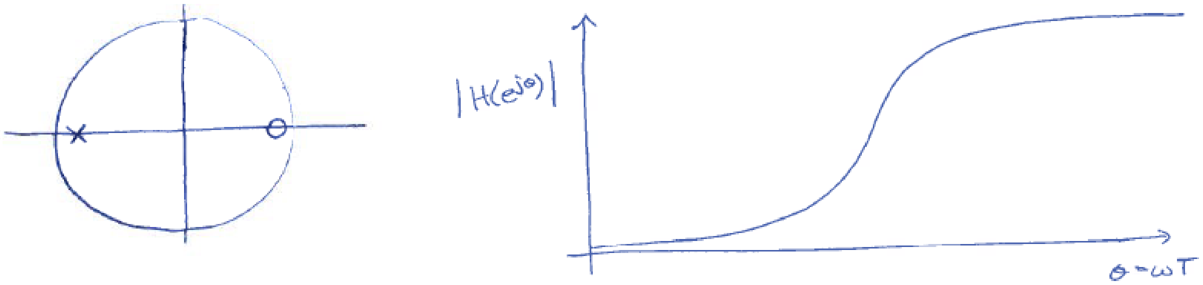
$$h(k) = \frac{1}{2\pi jk} \left(e^{jk\theta_c} - e^{-jk\theta_c}\right) = \frac{\sin(k\theta_c)}{\pi k}.$$

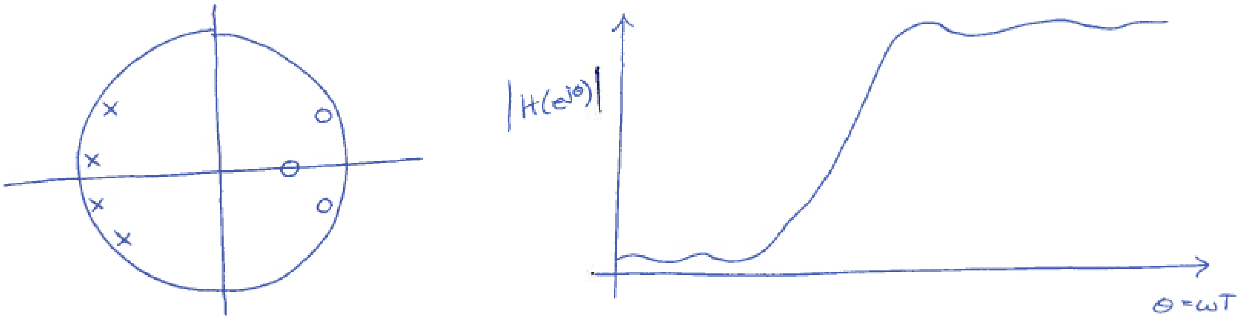
This is called a *sinc* function, which is a sinusoid that decays to zero both as $k \to +\infty$ and as $k \to -\infty$. In particular, the impulse response is nonzero for $k < 0$, so the filter is noncausal and therefore not implementable.

13.6 More advanced filters

In this brief introduction to frequency selective filters, we only considered the most basic filter configurations. Using our intuition about the shape of the frequency response in terms of the poles and zeros of the transfer function, however, we could easily construct more sophisticated higher-order filters. Some examples of highpass filters are as follows:


$|H(e^{j\theta})|$
$\theta = \omega T$


$|H(e^{j\theta})|$
$\theta = \omega T$


$|H(e^{j\theta})|$
$\theta = \omega T$

14

BIBO Stability

Stability is a fundamental concept in the study of dynamical systems that describes the behavior of the system trajectories over time. A system is said to be stable if its trajectories remain bounded or converge to a fixed point as time progresses. This means that small perturbations or disturbances to the system do not cause its behavior to change drastically. Stability is particularly important in the field of control, where it ensures that a controlled system behaves predictably and does not exhibit undesirable behavior.

14.1 Overview

There are many ways to characterize stability of a dynamical system. For instance, we may describe stability in terms of properties of the output signal in relation to the input signal, which is called *external stability*. Alternatively, we could describe stability in terms of all signals internal to the system, which is called *internal stability*.

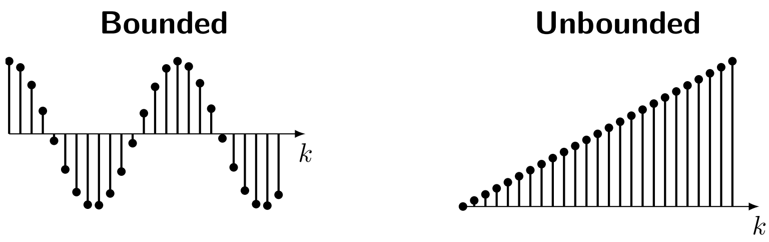
Here, we focus on external (or input-output) stability, which means that the output of a system is "small" whenever the input is "small". There are many ways of characterizing external stability depending on how we measure the size of a signal (what does it mean for a signal to be small?). For LTI systems, however, all notions of stability are equivalent.

Example (Stability examples).

- For a chair in which the inputs are applied forces and the outputs are its position, we want the system to be stable in that small forces to result in the chair falling over.
- Positive feedback in an audio system is an example of undesirable unstable system. Here, a small sound into the microphone gets amplified by the speaker, which then is picked up by the microphone and amplified, over and over again, resulting in loud piercing noises that hurt our ears.
- A system that is purposefully unstable is a fighter jet. Pilots need to be able to perform fast maneuvers, so the dynamics of the aircraft are intentionally designed to be unstable. These unstable dynamics are then stabilized by a control system so that the aircraft can fly (otherwise, the pilot would be unable to control the aircraft!).
- A passenger airline, on the other hand, is intentionally very stable so that even large disturbances (such as turbulence or losing an engine) do not result in drastic changes in flight.

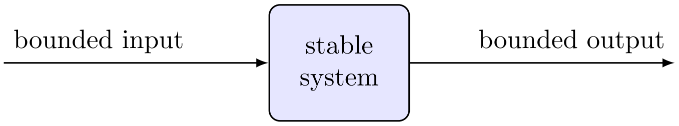
Definition (Bounded signal). A signal $y(k)$ is *bounded* if there exists a constant $M > 0$ such that

$$|y(k)| < M \quad \text{for all } k.$$



Remark. A discrete-time signal can be unbounded only if its magnitude grows without bound as $k \to \infty$, it cannot be unbounded "in the middle" (at finite times). Continuous-time signals, however, can be unbounded at finite times (if it has a vertical asymptote). ■

Definition (BIBO stability). A system is *bounded-input bounded output (BIBO) stable* if the output signal is bounded whenever the input signal is bounded.



Remark. For BIBO stability, the bounds on the input and output signals can be different. For instance, if the output signal is bounded by 100 whenever the input signal is bounded by 1, the system is still BIBO stable. ■

Example (Using the definition). One way to determine stability of a system is to directly apply the definition. Consider the system

$$y(k) = 3\,\sin(k+2)\,u(k-3)$$

If the input is bounded, then there exists a constant M such that $|u(k)| < M$ for all k, which implies that

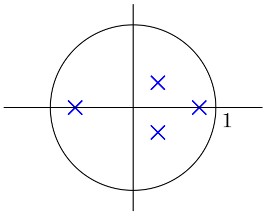
$$|y(k)| = 3\,|\sin(k+2)|\,|u(k-3)| < 3M,$$

so the output is also bounded. Therefore, the system is BIBO stable.

14.2 Characterization using the characteristic polynomial

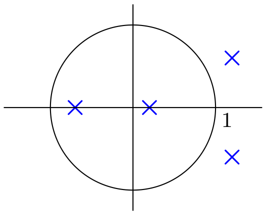
A discrete-time system is BIBO stable if and only if all roots of the characteristic polynomial $D(E)$ have magnitude strictly less than one.

- **Case 1:** if all roots of $D(E)$ lie strictly within (not on) the unit circle, then the system is stable by any definition

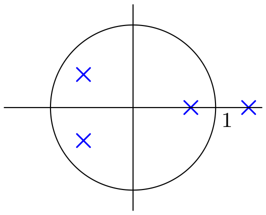
  

  - the homogeneous solution consists of decaying exponentials and/or decaying sinusoids
  - the particular solution has the same form as the input, so it is bounded if the input is bounded

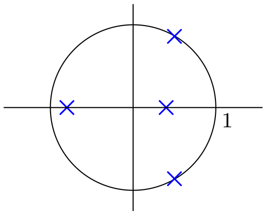
- **Case 2:** if any root of $D(E)$ is outside the unit circle, then the system is unstable by any definition

  or

- **Case 3:** if any root of $D(E)$ is on the unit circle, then the system is *not* BIBO stable (this is sometimes called *marginally stable*)

  - the homogeneous solution is constant or sinusoidal and therefore bounded
  - if the input has the same form as the homogeneous solution, then the particular solution will be multiplied by a factor of k, which will be unbounded (resonance)

Example (Using the poles)**.** Consider the system described by the difference equation

$$y(k+1) = 2.5y(k) - y(k-1) + 3u(k) - u(k-1)$$

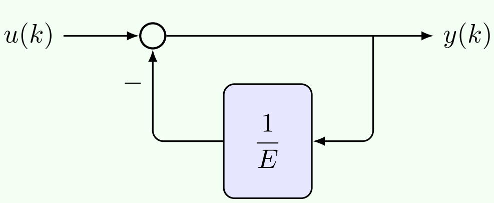
The transfer function of the system is

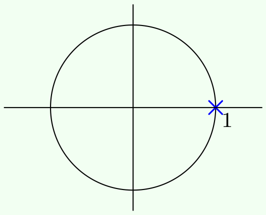
$$H(z) = \frac{3z-1}{z^2 - 2.5z + 1} = \frac{3z-1}{(z-0.5)(z-2)}$$

The transfer function has a pole at $z = 2$ which is outside the unit circle, so the system is not BIBO stable.

Example (Integrator).

$$y(k+1) = y(k) + u(k) \qquad H(z) = \frac{1}{z-1} \qquad h(k) = u_s(k-1)$$





Suppose the input signal is a (bounded) unit step. Using convolution or the z-transform, the output is

$$y(k) = k\, u_s(k)$$

which is unbounded, so the system is *not* BIBO stable.

14.3 Characterization using the impulse response

A system is BIBO stable if and only if its impulse response is absolutely summable.

$$\sum_{k=-\infty}^{\infty} |h(k)| \quad \text{finite}$$

The argument is as follows. Suppose the input is bounded. Then there exists a constant $M > 0$ such that $|u(k)| < M$ for all k. The output is then the convolution of the input signal with the impulse response. Using the bound on the input, we can bound the output as

$$\begin{aligned}
|y(k)| &= \left|\sum_{m=-\infty}^{\infty} h(k-m)\, u(m)\right| \\
&\leq \sum_{m=-\infty}^{\infty} |h(k-m)\, u(m)| \\
&= \sum_{m=-\infty}^{\infty} |h(k-m)| \cdot |u(m)| \\
&\leq M \sum_{m=-\infty}^{\infty} |h(k-m)| \\
&= M \sum_{m=-\infty}^{\infty} |h(m)|
\end{aligned}$$

So, if the impulse response is absolutely summable, then the output is also bounded and the system is BIBO stable. Moreover, it can be shown that, if the impulse response is not absolutely summable, then there is a particular bounded input signal for which the output is unbounded.

Example (Using the impulse response). Consider the discrete-time LTI system with impulse response

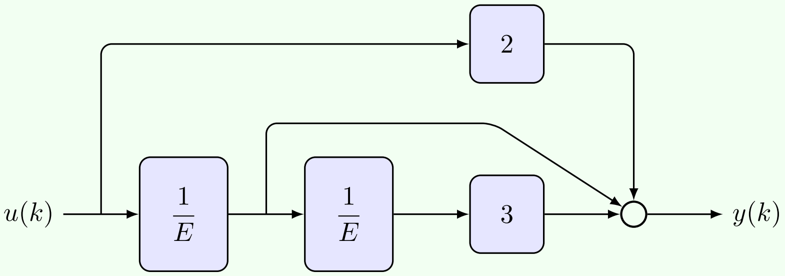
$$h(k) = (0.5)^k \, u_s(k)$$

The summation of the absolute value of the impulse response is

$$\sum_{k=-\infty}^{\infty} |h(k)| = \sum_{k=0}^{\infty} (0.5)^k = \frac{1}{1-0.5} = 2$$

which is finite, so the system is BIBO stable.

Example (FIR system). Recall that a finite-impulse response (FIR) system has an impulse response that is only nonzero for a finite amount of time. The transfer function of an FIR system has all poles at the origin ($z = 0$).



- the impulse response is absolutely summable

$$h(k) = \{2, 1, 3, 0, 0, 0, \ldots\}$$

- the characteristic polynomial has all roots at the origin

$$H(E) = 2 + \frac{1}{E} + \frac{3}{E^2} = \frac{2E^2 + E + 3}{E^2}$$

The impulse response of an FIR system is always absolutely summable (since it has finite duration), and all poles are strictly inside the unit circle in the complex plane (since they are all at $z = 0$), so FIR systems are *always* BIBO stable.

Part IV

Continuous-Time LTI Systems

15

Continuous-Time LTI Systems

We have focused our study so far on discrete-time signals and systems. Many of the same tools and concepts apply to continuous-time systems as well, with a few notable excepts. We now briefly discuss continuous-time LTI systems and emphasize the differences with discrete time.

15.1 Differential equations

Every causal finite-dimensional continuous-time LTI system can be represented by a differential equation of the following form:

$$\sum_{i=0}^{n} a_i \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^i y(t) = \sum_{j=0}^{m} b_j \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^j u(t)$$

where

- $u(t)$ is the input signal
- $y(t)$ is the output signal
- n is the order of the system (assuming $a_n \neq 0$)
- a_i and b_i are constant parameters

Conversely, every differential equation of this form represents a causal finite-dimensional continuous-time LTI system.

15.2 Impulse signal, impulse response, and convolution

Recall that we defined the discrete-time impulse signal as

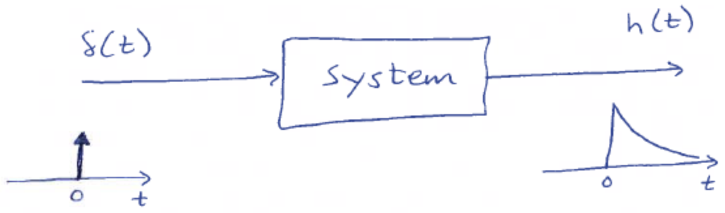
$$\delta(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases}$$

However, we did not define a continuous-time impulse signal $\delta(t)$. The reason is that, unlike in discrete time, the impulse signal in continuous time is quite complicated. In fact, it is not even a function, so we cannot write a formula for the continuous-time impulse signal as a function of time. As we will see, however, we can still define a continuous-time impulse signal that will be useful in studying continuous-time LTI systems.

To motivate the continuous-time impulse signal, let's first consider the impulse response, which is the output of the system when the input is an impulse and the initial conditions are zero (that is, the zero-state response). Then we can describe every zero-state response of the system through convolution with the impulse response just as in discrete time.

Impulse response

The impulse response is the response of the system due to an impulse input when the initial conditions are zero:



We have not yet defined what the continuous-time impulse signal is, but we draw it as an arrow at time $t = 0$. If the system is time invariant, then shifting the input signal in time also shifts the output signal by the same amount of time, so

$$\delta(t-\tau) \qquad \to \qquad h(t-\tau)$$

for any time shift τ. If the system is linear, then a weighted sum of inputs produces the same weighted sum of the corresponding outputs. For each τ, we will weight the input by $u(\tau)$ for some arbitrary signal u. Then summing over all τ results in

$$\int_{-\infty}^{\infty} u(\tau)\,\delta(t-\tau)\,\mathrm{d}\tau \qquad \to \qquad \int_{-\infty}^{\infty} u(\tau)\,h(t-\tau)\,\mathrm{d}\tau$$

This argument is similar to what we saw in discrete time, where we had

$$\sum_{m=-\infty}^{\infty} u(m)\,\delta(k-m) \qquad \to \qquad \sum_{m=-\infty}^{\infty} u(m)\,h(k-m)$$

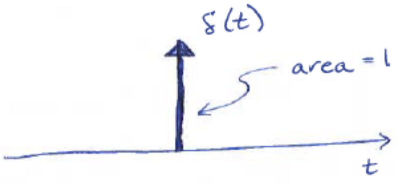
The term on the left is an infinite sum, where each term in the summation is a shifted and scaled impulse, which is nonzero at only a single point in time (that is, when $k = m$). We plotted each of these terms to conclude that the left-hand side is simply $u(k)$, which is an arbitrary input signal. We then defined the right-hand side to be the convolution of the input signal with the impulse response.

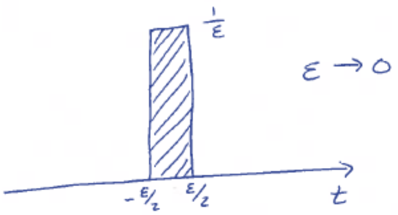
Impulse signal

We would like to make the same argument in continuous time. To do so, we need the impulse signal to satisfy

$$u(t) = \int_{-\infty}^{\infty} u(\tau)\,\delta(t-\tau)\,\mathrm{d}\tau$$

for any signal $u(t)$. The intuitive solution is to set $\delta(t)$ to be the same as the discrete-time impulse, but then the term inside the integral is zero everywhere except for at a single point which has no area so its integral is zero. In fact, there is no such function that satisfies this equation! We need a function that is zero everywhere except for a spike at $t = \tau$ that has unit area so that the integral is nonzero. While such a function does not exist, we can approximate the impulse signal as a pulse that has width ε and height $1/\varepsilon$ centered about the origin.





The continuous-time impulse signal can then be defined as the limit of this signal as $\varepsilon \to 0$.

Definition (Continuous-time impulse signal)**.** The continuous-time impulse signal, denoted $\delta(t)$, is the limit as $\varepsilon \to 0$ of the pulse centered about $t = 0$ with width ε and height $1/\varepsilon$,

$$\delta(t) \quad = \quad \lim_{\varepsilon \to 0} \; \delta_\varepsilon(t).$$

This is also known as the *Dirac delta function*.

The continuous-time impulse signal is *not* a function, so we cannot write an expression for $\delta(t)$ as a function of t (without using limits). On plots, we represent the continuous-time impulse signal as an arrow whose height is the area of the impulse. For instance, the impulse signal $2\,\delta(t-3)$ would be an arrow with height two centered at time $t = 3$.

As we have seen, the output of the system is an integral that depends on the input signal and the impulse response. We define this as the *convolution* of the two signals.

Definition (Continuous-time convolution)**.** The convolution of two continuous-time signals $h(t)$ and $u(t)$ is another signal, denoted $h * u$, whose value at time t is the integral

$$(h * u)(t) = \int_{-\infty}^{\infty} h(t-\tau)\,u(\tau)\,\mathrm{d}\tau.$$

Note that we cannot actually put an impulse into a system since it has infinite energy. We can, however, approximate the impulse response by putting its approximation into the system for small ε.

From this definition of the continuous-time impulse signal as the limit of its approximation, we have that the above equation is satisfied for any signal $u(t)$. This is known as the *sifting property* of the impulse signal.

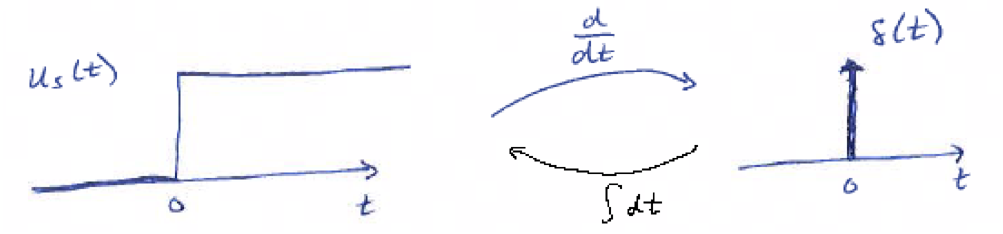
Fact (Sifting property)**.** The continuous-time impulse signal is the identity signal for convolution, meaning that convolving it with any signal does not change the signal:

$$u(t) = \int_{-\infty}^{\infty} \delta(t-\tau)\,u(\tau)\,\mathrm{d}\tau.$$

Continuous-time impulse and step signals

We can also interpret the impulse as the derivative of the unit step signal, and likewise the step signal is the integral of the impulse. While the derivative of the continuous-time step signal does not exist in the strict sense, it does when we allow for generalized signals such as the impulse.

$$\delta(t) = \frac{\mathrm{d}}{\mathrm{d}t} u_s(t) \qquad \text{and} \qquad u_s(t) = \int_{-\infty}^{t} \delta(\tau)\,\mathrm{d}\tau$$



The complexity of the continuous-time impulse signal is one of the reasons that we began by studying discrete-time signals and systems! While the continuous-time impulse signal does not actually exist, we will use it much like in the discrete-time case to study linear time-invariant systems.

Convolution

Fact (Zero-state response)**.** The zero-state response of an LTI system with impulse response $h(t)$ due to an input signal $u(t)$ is the convolution

$$y(t) = (h * u)(t) = \int_{-\infty}^{\infty} h(t-\tau)\,u(\tau)\,\mathrm{d}\tau.$$

Similar to discrete time, some properties of convolution are as follows:

- Convolution is commutative, meaning that $(h * u)(t) = (u * h)(t)$.
- The system is causal if and only if the impulse response is zero for negative times, $h(t) = 0$ for $t < 0$.
- Convolving an impulse with any signal does not affect the signal, that is, $(x * \delta)(t) = x(t)$ for any signal $x(t)$.

Example. Find the zero-state response of the continuous-time LTI system with impulse response $h(t) = e^{-t}\, u_s(t)$ due to the input signal $u(t) = e^{-3t}\, u_s(t)$.

The zero-state response is the convolution of the input signal with the impulse response,

$$y(t) = (h * u)(t) = \int_{-\infty}^{\infty} h(t-\tau)\, u(\tau)\, \mathrm{d}\tau$$

Substituting the input impulse response signals,

$$y(t) = \int_{-\infty}^{\infty} e^{-(t-\tau)}\, u_s(t-\tau)\, e^{-3\tau}\, u_s(\tau)\, \mathrm{d}\tau$$

The unit steps make the integrand zero for $\tau < 0$ and $\tau > t$, and are one for all other times. Therefore, we can change the limits of integration to obtain

$$y(t) = \int_{0}^{t} e^{-(t-\tau)}\, e^{-3\tau}\, \mathrm{d}\tau$$

Manipulating the exponentials,

$$y(t) = e^{-t} \int_{0}^{t} e^{-2\tau}\, \mathrm{d}\tau$$

Now evaluating the integral,

$$y(t) = e^{-t} \big[-\tfrac{1}{2} e^{-2\tau}\big]\Big|_{\tau=0}^{t} = -\tfrac{1}{2} e^{-t} \big[e^{-2t} - 1\big]\, u_s(t)$$

where the unit step is due to the fact that the intgral is zero if $t < 0$. Simplifying, the zero-state response is

$$y(t) = \tfrac{1}{2}\big(e^{-t} - e^{-3t}\big)\, u_s(t)$$

which contains terms similar to both the input signal and the impulse response.

15.3 Complex exponentials

A continuous-time complex exponential signal has the form e^{st} where s is a complex number. The shape of the signal depends on the value of the complex number s. To understand the shape of the signal, write the complex number in rectangular form as $s = a + jb$. Then the complex signal is

$$e^{st} = e^{(a+jb)t} = e^{at} e^{jbt} = e^{at}\, \big(\cos(bt) + j\sin(bt)\big)$$

where we used properties of the exponential and Euler's formula for complex numbers. This reveals that a complex number is the product of a (real) exponential signal with the summation of real and imaginary sinusoidal signals.

$$e^{st} \quad = \quad \underbrace{e^{at}}_{\text{exponential}} \quad \underbrace{\big(\cos(bt) + j\sin(bt)\big)}_{\text{sinusoid}}$$

The parameter a is the real part of the complex number s, which determines whether the exponential is growing, decaying, or constant. The parameter b is the imaginary part of the complex number s, which determines how fast the real and imaginary parts of the signal are oscillating.

15.4 Transfer function

Just like in discrete time, complex exponential signals are eigenvectors of LTI systems and can therefore be used to construct the transfer function, which is how much the system scales a complex exponential.

$$e^{st} \qquad \rightarrow \qquad H(s)\, e^{st}$$

To construct the transfer function, use convolution to find the response due to a complex exponential input signal. Suppose the system is causal (so the impulse is zero for negative times). If the input is the complex exponential $u(t) = e^{st}$, then the output is

$$y(t) = (h * u)(t) = \int_0^\infty h(\tau)\, e^{s\,(t-\tau)}\, \mathrm{d}\tau = e^{st} \int_0^\infty h(\tau)\, e^{-s\tau}\, \mathrm{d}\tau$$

The transfer function $H(s)$ is defined as the amount the system scales the complex exponential e^{st}.

$$H(s) = \int_0^\infty h(t)\, e^{-st}\, \mathrm{d}t$$

The transfer function $H(s)$ is the *Laplace transform* of the impulse response $h(t)$.

We can use the transfer function to find the zero-state response due to any input signal. Using convolution, the Laplace transform of the output due to the input signal $u(t)$ is

$$\begin{aligned}
Y(s) &= \int_0^\infty y(t)\, e^{-st}\, \mathrm{d}t \\
&= \int_0^\infty \left(\int_0^\infty u(\tau)\, h(t-\tau)\, \mathrm{d}\tau \right) e^{-st}\, \mathrm{d}t \\
&= \int_0^\infty \int_0^\infty h(t-\tau)\, e^{-s\,(t-\tau)}\, u(\tau)\, e^{-s\tau}\, \mathrm{d}\tau\, \mathrm{d}t \\
&= \int_0^\infty \underbrace{\left(\int_0^\infty h(t-\tau)\, e^{-s\,(t-\tau)} \mathrm{d}t \right)}_{H(s)} u(\tau)\, e^{-s\tau}\, \mathrm{d}\tau \\
&= H(s) \int_0^\infty u(\tau)\, e^{-s\tau}\, \mathrm{d}\tau \\
&= H(s)\, U(s)
\end{aligned}$$

Therefore, the zero-state response in the frequency domain is simply $Y(s) = H(s)\, U(s)$.

In the time domain, the zero-state response is the convolution of the input signal with the impulse response. In the frequency domain, the Laplace transform of the zero-state response is the product of the transfer function with the Laplace transform of the input signal.

$$\begin{tikzcd}
\textbf{Time domain:} & u(t) \arrow[r] \arrow[d, leftrightarrow, "\mathcal{L}"] & \boxed{h(t)} \arrow[r] \arrow[d, leftrightarrow, "\mathcal{L}"] & y(t) = (h * u)(t) \arrow[d, leftrightarrow, "\mathcal{L}"] \\
\textbf{Frequency domain:} & U(s) \arrow[r] & \boxed{H(s)} \arrow[r] & Y(s) = H(s)\, U(s)
\end{tikzcd}$$

15.5 Laplace transform

Definition. The Laplace transform of a continuous-time signal $x(t)$ is the complex function

$$X(s) = \int_0^\infty x(t)\, e^{-st}\,\mathrm{d}t$$

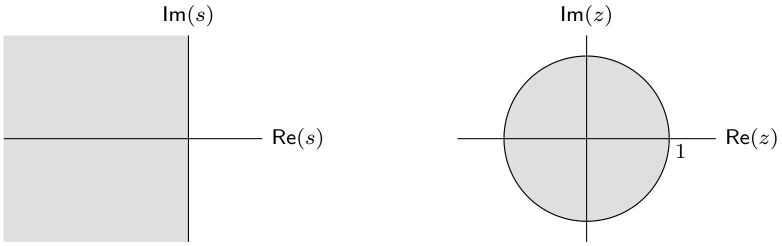
where s is any complex number for which the integral converges. The values of s for which the integral converges is the region of convergence (ROC).

- The Laplace transform is the continuous-time equivalent of the z-transform for discrete-time signals.
- We use lowercase letters for time-domain signals such as $x(t)$ and the corresponding uppercase letter for its Laplace transform such as $X(s)$.
- This is called the *unilateral* Laplace transform since the integral starts at zero; there is a bilateral version that integrates over all time.
- The ROC is all values of s to the right of the rightmost pole.

15.6 Stability

For a continuous-time LTI system, the following statements are equivalent:

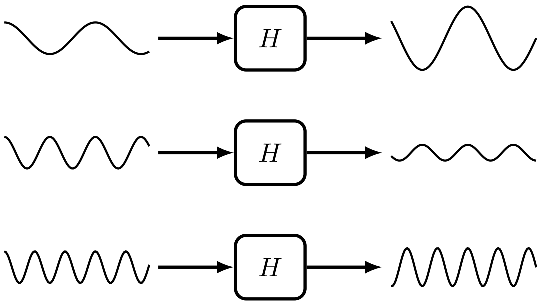
- The system is BIBO stable.
- The transfer function has all poles in the open left-half plane.
- The impulse response is absolutely integrable.



Stable discrete-time LTI systems have poles inside the unit circle, while stable continuous-time LTI systems have poles in the left-half plane. One significant difference between continuous and discrete time is that the stability region in continuous time is an unbounded set, while in discrete time the stability region is bounded. This makes it much easier to design stable systems in continuous time!

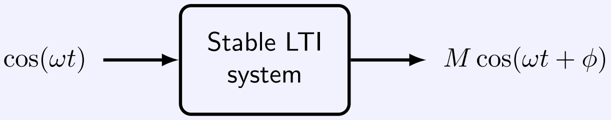
15.7 Frequency response

The frequency response of a stable LTI system characterizes how the system acts on individual frequencies. The frequency response is based on the fact that, for stable LTI systems, sinusoid inputs produce sinusoid outputs.



The output sinusoid has the same frequency, but it may be shifted and/or scaled.

Definition. For a stable LTI system, the frequency response is the amount that the system scales and shifts sinusoidal input signals.



For each frequency ω, the frequency response magnitude M is the amount that the sinusoid is scaled, and the frequency response phase ϕ is the amount that it is shifted.

We can find the frequency response in terms of the transfer function $H(s)$. First, decompose the sinusoidal input as a sum of complex exponentials,

$$u(t) = \cos(\omega t) = \frac{e^{j\omega t} + e^{-j\omega t}}{2}$$

Recall that $H(s)$ is the amount that the system scales the complex exponential e^{st}. Since the system is linear, a weighted sum of inputs produces the same weighted sum of the corresponding outputs. Therefore, the output due to the sinusoid is

$$y(t) = \frac{H(j\omega)\, e^{j\omega t} + H(-j\omega)\, e^{-j\omega t}}{2}$$

The second term is the complex conjugate of the first. Since the sum of a complex number and its conjugate is twice its real part (that is, $(a + jb) + (a - jb) = 2a$), we have that

$$y(t) = \mathrm{Re}[H(j\omega)\, e^{j\omega t}]$$

Let M be the magnitude and ϕ the phase of the complex number $H(j\omega)$. Then

$$y(t) = \mathrm{Re}[M\, e^{j(\omega t + \phi)}]$$

Applying Euler's formula to the complex exponential and then taking the real part,

$$y(t) = M\cos(\omega t + \phi)$$

We summarize this result as follows.

For a stable continuous-time LTI system, the frequency response is the transfer function evaluated on the imaginary axis in the complex plane, which is the complex function of ω given by

$$H(j\omega) = M\, e^{j\phi} \qquad \text{where} \qquad M = |H(j\omega)| \qquad \text{and} \qquad \phi = \angle H(j\omega)$$

15.8 Fourier transform

The frequency response is the transfer function evaluated on the imaginary axis, where the transfer function is the Laplace transform of the impulse response. This is called the *Fourier tranform.*

Definition. The Fourier transform of a continuous-time signal $x(t)$ is the complex function

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, \mathrm{d}t$$

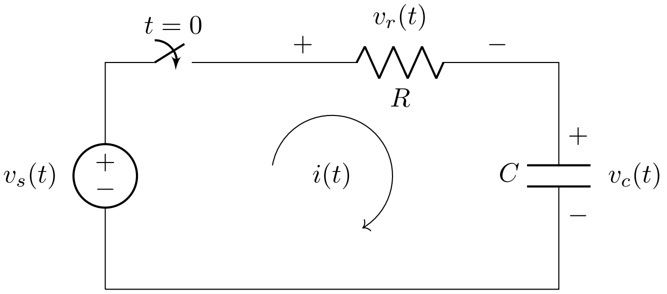
if the integral exists. The inverse Fourier transform is

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega)\, e^{j\omega t}\, \mathrm{d}t$$

- The quantity $|X(j\omega)|$ is the amount of frequency ω that is in the signal $x(t)$.
- For signals that start at time zero ($x(t) = 0$ for $t < 0$), the Fourier transform is the Laplace transform evaluated on the imaginary axis.
$$X(j\omega) = X(s)\Big|_{s=j\omega}$$
- The Fourier transform exists if the imaginary axis is in the region of convergence of the Laplace transform.
- While the formula for the inverse Laplace transform requires complex integration, the inverse Fourier transform is a much simpler integral that is quite similar to the Fourier transform itself.
- Parseval's theorem states that the energy of the signal is the same in both the time domain and frequency domain.
$$\int_{-\infty}^{\infty} |x(t)|^2\, \mathrm{d}t \quad = \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(j\omega)|^2\, \mathrm{d}\omega$$
- For a stable continuous-time LTI system, the frequency response is the Fourier transform of the impulse response.

15.9 Example: RC circuit

We now use a simple RC circuit to illustrate many of the concepts from the course.



Modeling

One way to analyze the circuit is to use Kirchhoff's laws to write down differential equations that describe the relationship between the signals in the circuit. From Kirchhoff's voltage law (KVL), the sum of the voltages around the loop is zero, so the supplied voltage is equal to the sum of the voltage across the resistor and the capacitor,

$$v_s(t) = v_r(t) + v_c(t)$$

From Kirchhoff's current law (KCL), the same current flows through the resistor and capacitor,

$$i(t) = \frac{1}{R} v_r(t) = C\,\dot{v}_c(t)$$

where we used the relationship between voltage and current for the resistor and the capacitor. Solving for the voltage across the resistor in terms of the voltage across the capacitor and substituting into the above equation yields the relationship

$$v_s(t) = RC\,\dot{v}_c(t) + v_c(t)$$

This is a first-order differential equation that relates the source voltage $v_s(t)$ and the capacitor voltage $v_c(t)$. The product $\tau = RC$ of the resistance and capacitance is called the *time constant* since it describes how quickly the circuit responds to changes in the input signal (as we will see below).

Transfer function

We can find the transfer function from the source voltage to the voltage across the capacitor by taking the Laplace transform (with zero initial conditions) of the differential equation. In particular, the Laplace transform $X(s)$ of a signal $x(t)$ satisfies

$$\dot{x}(t) \quad \xleftrightarrow{\mathcal{L}} \quad sX(s) \qquad \text{and} \qquad \ddot{x}(t) \quad \xleftrightarrow{\mathcal{L}} \quad s^2X(s)$$

when $x(0) = 0$ and $\dot{x}(0) = 0$. Taking the Laplace transform of the differential equation and using this property yields

$$V_s(s) = (\tau s + 1)V_c(s)$$

where $V_s(s)$ and $V_c(s)$ are the Laplace transforms of the source and capacitor voltages, respectively. Solving for the ratio of the output to the input yields the transfer function

$$H(s) = \frac{V_c(s)}{V_s(s)} = \frac{1}{\tau s + 1}$$

The transfer function has a single pole at $s = -1/\tau$ and no zeros. Since the pole is in the left-half of the complex plane, the system is stable.

Frequency response

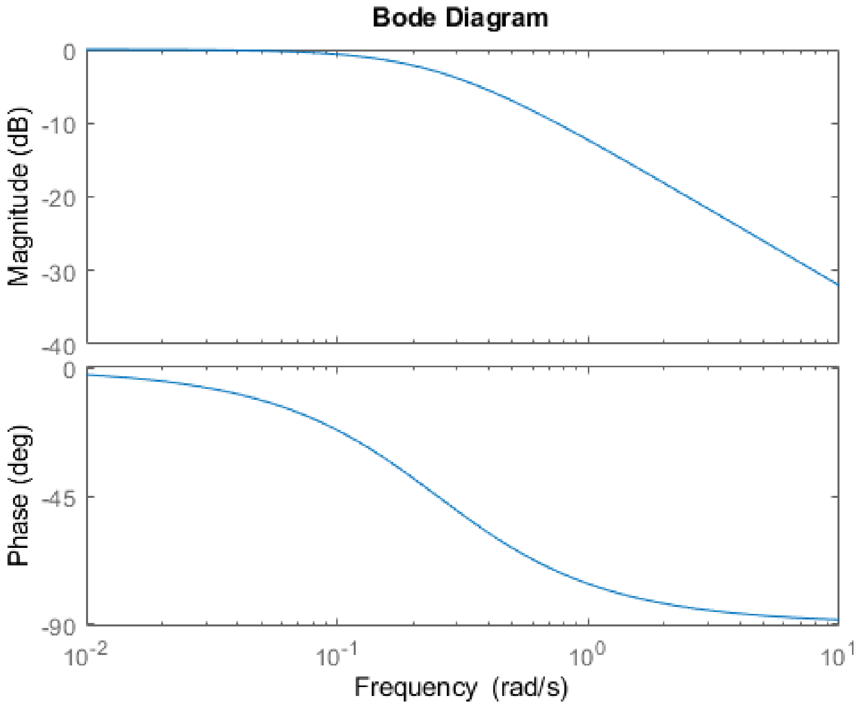
The frequency response describes how the circuit responds to various frequencies in the source voltage. For continuous-time systems, the frequency response is the transfer function evaluated on the imaginary axis (that is, $s = j\omega$), so the frequency response is

$$H(j\omega) = \frac{1}{j\omega\tau + 1}$$

The magnitude of the frequency response is

$$|H(j\omega)| = \frac{1}{\sqrt{(\omega\tau)^2 + 1}}$$

For constant input signals, the frequency is zero and the corresponding gain is one. For high frequencies, ω is large and the magnitude of the frequency response is small. Therefore, this system acts as a lowpass filter that passes low frequencies and attenuates high frequencies. The bode plot of the system is as follows.



Impulse response

Recall that the impulse response is the output of the system due to an impulse input when the system is initially at rest. Since the transfer function is the Laplace transform of the impulse response, we can take the inverse Laplace transform of $H(s)$ to obtain the impulse response $h(t)$. For this simple system, we can find the inverse Laplace transform using a table of Laplace transform pairs to obtain

$$h(t) = \frac{1}{\tau} e^{-t/\tau}\, u_s(t)$$

which is a decaying exponential. Note that the time constant τ determines how quickly the impulse response decays, with larger time constants resulting in slower decay.

Step response

Recall that the step response is the output of the system due to a step input when the system is initially at rest. Now that we have both the impulse response and transfer function, we can find the step response (which is a particular zero-state response) using either convolution in the time domain or the transfer function in the frequency domain.

In the time domain, the step response is the convolution of the impulse response with a unit step signal,

$$y(t) = (h * u_s)(t) = \frac{1}{\tau}\int_0^t e^{-r/\tau}\, \mathrm{d}r = -e^{-r/\tau}\Big|_{r=0}^{t} = \left(1 - e^{-t/\tau}\right) u_s(t)$$

which grows exponentially to one (the size of the source voltage).

We can also find the step resonse in the frequency domain. The Laplace transform of the step response is

$$Y(s) = H(s)\,U_s(s) = \frac{1}{\tau s+1}\frac{1}{s}$$

To find the inverse Laplace transform, we perform a partial fraction expansion using terms that are in the table,

$$Y(s) = \frac{A}{\tau s+1} + \frac{B}{s}$$

Unlike the partial fraction expansion for the z-transform, we do *not* include the Laplace variable in the numerator since the terms in the table are of the form $1/(s+a)$, and we do not need the constant term when using this form (you could include it, but it can always be taken as zero). We can solve for the parameters in the partial fraction expansion just like in the discrete-time case. To find A, multiply by the demoninator $\tau s+1$ and then set $s=-1/\tau$ to obtain $A=-\tau$. Similarly, multiply by s and then set $s=0$ to obtain $B=1$. Then using the entries in the table of Laplace transform pairs, the step response is

$$y(t) = \left(1 - e^{-t/\tau}\right) u_s(t)$$

which is the same as found using convolution.

Complete response

We can find the response of the system due to both the input signal and initial condition by taking the Laplace transform of the differential equation. Since the differential equation involves the derivative of the capacitor voltage, we need the following property of the Laplace transform:

$$\dot{x}(t) \quad \overset{\mathcal{L}}{\longleftrightarrow} \quad s\,X(s) - x(0)$$

Taking the Laplace transform of the differential equation and using this property, we obtain

$$V_s(s) = \tau\left[s\,V_c(s) - v_c(0)\right] + V_c(s)$$

Solving for the Laplace transform of the capacitor voltage,

$$V_c(s) = \frac{1}{\tau s+1}V_s(s) - \frac{\tau}{\tau s+1}v_c(0)$$

The first term is the zero-state response, which is the product of the transfer function with the Laplace transform of the input signal. The second term is the zero-input response, which depends on the initial capacitor voltage. Given a particular input signal and initial condition, we can substitute them into this equation and take the inverse Laplace transform (using a partial fraction expansion) to find the complete response.

MATLAB code

All of the above analysis can be done easily in MATLAB as follows.

```
tau = R*C;          % time constant
s = tf('s');        % Laplace variable
H = 1/(tau*s + 1);  % transfer function
bode(H);            % bode plot
impulse(H);         % impulse response
step(H);            % step response
lsim(H,u,t,x0);     % total response due to input u(t) with initial condition x0
```

Part V

Appendices

A

Complex Numbers

Complex numbers are are fundamental tool used to study signals and systems, and we will use them extensively throughout this course. While the term *imaginary* may make them seem less practical than other numbers, they are no less real than a negative number (such as -2), irrational number (such as π), or even a natural number (such as 3). Numbers are just mathematical objects that we use to describe the world, and we will see that they are quite useful for studying signals and systems.

A.1 Motivation

Historically, the development of complex numbers was motivated by the need to solve the cubic equation

$$x^3 + ax + b = 0$$

In 1545, the mathematician Cardano found the general solution for x given by

$$x = \sqrt[3]{-\frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}} + \sqrt[3]{-\frac{b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}$$

You can substitute this expression into the cubic to verify that it is indeed a solution. For example, $a = 6$ and $b = -20$ gives the solution $x = 2$. But when we try substituting the coefficients $a = -15$ and $b = -4$ into the formula, we get an expression that contains the square root of a negative number.

$$x = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}$$

We know how to take the square root of positive numbers: the square root of a positive number y is a number whose square is y. For instance, the square root of 16 is 4 since $4^2 = 16$. But to interpret the square root of a negative number, we need to use *complex numbers*.

Using complex numbers, we have that

$$(2 \pm j)^3 = 2 \pm j11 = 2 \pm \sqrt{-121}$$

which suggests that the cube root of $2 \pm \sqrt{-121}$ is $2 \pm j$. Substituting this into the expression for x, the solution to the cubic is simply $x = 4$, which is a real number! This example illustrates the usefulness of complex numbers: they enable us to solve *real* problems that we could not solve before, or to solve them in a much simpler way.

A.2 Definition

The *imaginary unit* is defined as the number j whose square is negative one, that is, $j^2 = -1$.

Notation. Engineers typically use j for the imaginary unit while mathematicians use i. ■

While no real number satsifies the equation $j^2 = -1$, we define it's solution to be the imaginary unit j. From this, it follows that both $\pm j$ are square roots of -1.

A.3 Representations

There are several different ways of writing a complex number z.

$$z \quad = \quad \underbrace{a + jb}_{\text{rectangular form}} \quad = \quad \underbrace{r\,e^{j\theta}}_{\text{polar form}}$$

The constants a, b, r, and θ describe various aspects of the complex number. In rectangular form, a is called the real part of the complex number while b is called the imaginary part. In polar form, r is called the magnitude and θ is called the angle (or phase). The constant e is referred to as *Euler's number* and is the irrational number

$$e = 2.7182818...$$

The rectangular and polar forms are equivalent in that any complex number can be written uniquely in either form. For a complex number z, its real and imaginary parts, magnitude, and phase are denoted as follows.

- $\text{Re}(z) = a$ is the real part
- $\text{Im}(z) = b$ is the imaginary part
- $|z| = r$ is the magnitude (which is nonnegative)
- $\angle z = \theta$ is the angle or phase

Rectangular and polar coordinates are linked through *Euler's equation*, which states that

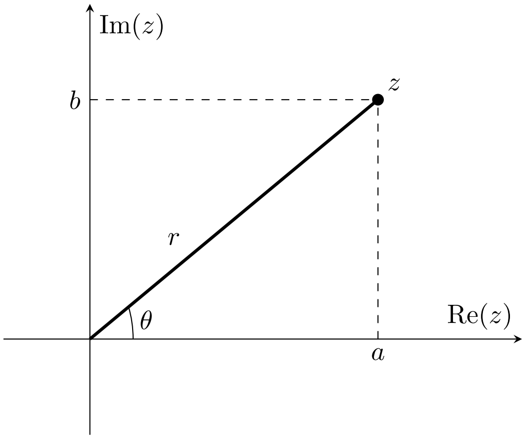
$$e^{j\theta} = \cos\theta + j\sin\theta$$

On the left hand side, $e^{j\theta}$ is a complex number in polar form, and on the right-hand side is $\cos\theta + j\sin\theta$, which is a complex number in rectangular form. Euler's equation allows us to transform complex numbers between the two forms. Multiplying both sides by r, we get that $a = r\cos\theta$ and $b = r\sin\theta$. This relates the real and imaginary parts of the complex number to its magnitude and phase. By inverting this relationship, we can solve for r and θ in terms of a and b. Doing so yields the following relationships to convert between the two forms.

Polar to rectangular	**Rectangular to polar**
$a = r\cos\theta$	$r = \sqrt{a^2 + b^2}$
$b = r\sin\theta$	$\theta = \arctan(b/a)$

A.4 Complex plane

It is often useful to visualize complex numbers in the *complex plane*. The complex plane is a two-dimensional plane in which the real part of the complex number is plotted on the x-axis and the imaginary part on the y-axis.



For a complex number in rectangular form, the coefficients a and b denote the x and y coordinates in the complex plane. In polar form, the magnitude r is the distance from the origin, and θ is the angle between the real axis and the line connecting the origin to the complex number.

From this visualization in the complex plane, we can use trigonometry to derive the relationship between the rectangular and polar forms of a complex number. For instance, since the magnitude r is the hypotenuse of the triangle with side lengths a and b, we have from Pythagorean's theorem that $r^2 = a^2 + b^2$. Similarly, the tangent of θ is the opposite side length b divided by the adjacent side length a.

A.5 Algebra with complex numbers

We can do algebra with complex numbers, just like with real numbers.

- **Addition/subtraction**
  - rectangular form

$$z_1 + z_2 = (a_1 + jb_1) + (a_2 + jb_2) = (a_1 + a_2) + j(b_1 + b_2)$$
$$z_1 - z_2 = (a_1 + jb_1) - (a_2 + jb_2) = (a_1 - a_2) + j(b_1 - b_2)$$

  - to add or subtract two complex numbers in polar form, first convert them to rectangular form

- **Multiplication/division**
  - polar form

$$z_1 z_2 = \left(r_1 e^{j\theta_1}\right)\left(r_2 e^{j\theta_2}\right) = (r_1 r_2)\, e^{j(\theta_1+\theta_2)}$$

$$\frac{z_1}{z_2} = \frac{r_1 e^{j\theta_1}}{r_2 e^{j\theta_2}} = \frac{r_1}{r_2}\, e^{j(\theta_1-\theta_2)}$$

- multiplication in rectangular form (use the fact that $j^2 = -1$)

$$(a_1 + jb_1)(a_2 + jb_2) = a_1a_2 + j(a_1b_2 + b_1a_2) + j^2b_1b_2 = (a_1a_2 - b_1b_2) + j(a_1b_2 + b_1a_2)$$

- division in rectangular form (multiply and divide by the the same complex number to make the denominator real)

$$\frac{a_1 + jb_1}{a_2 + jb_2} = \frac{a_1 + jb_1}{a_2 + jb_2} \cdot \frac{a_2 - jb_2}{a_2 - jb_2} = \frac{(a_1a_2 + b_1b_2) + j(b_1a_2 - a_1b_2)}{a_2^2 + b_2^2} = \frac{a_1a_2 + b_1b_2}{a_2^2 + b_2^2} + j\frac{b_1a_2 - a_1b_2}{a_2^2 + b_2^2}$$

Remark (Complex numbers as vectors)**.** Since we can visualize complex numbers in the complex plane, we can interpret them as two-dimensional real vectors.

$$a + jb \qquad \text{corresponds to} \qquad \begin{bmatrix} a \\ b \end{bmatrix}$$

While this is a valid interpretation, it is slightly misleading. The reason is that complex numbers have more structure than a two-dimensional real vector. In particular, we can multiply and divide any two complex numbers (as long as the denominator is nonzero for division), while there is no corresponding operation on vectors (you cannot multiply two vectors to get another vector). ■

A.6 Trigonometric formulas

Below is a list of some useful trigonometric formulas.

$$e^{j\theta} = \cos\theta + j\sin\theta \qquad \text{(Euler's formula)}$$

$$\cos\theta = \frac{e^{j\theta} + e^{-j\theta}}{2}$$

$$\sin\theta = \frac{e^{j\theta} - e^{-j\theta}}{j2}$$

$$e^{j2\pi k} = 1 \qquad k \text{ an integer}$$

$$e^{j\pi k} = (-1)^k = \begin{cases} +1 & \text{if } k \text{ even} \\ -1 & \text{if } k \text{ odd} \end{cases} \qquad k \text{ an integer}$$

B

Linear Algebra

Linear algebra is the branch of mathematics concerned with linear equations and linear maps, along with their representations in vector spaces and through matrices.

$$\begin{aligned} a_{11}\,x_1 + a_{12}\,x_2 &= b_1 \\ a_{21}\,x_1 + a_{22}\,x_2 &= b_2 \end{aligned} \qquad \Longleftrightarrow \qquad \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \qquad \Longleftrightarrow \qquad Ax = b$$

The main mathematical objects in linear algebra are scalars, vectors, and matrices.

B.1 Vector space

A vector space is a set of objects, called *vectors*, over another set of objects, called *scalars*, that satisfy certain properties. For instance, vectors can be added together, and a vector can be scaled (or multiplied) by a scalar to produce another vector.

Scalars

A scalar is a number, such as 2, $\sqrt{5}$, or π. Scalars may be real numbers or complex numbers. We can do standard algebraic operations with scalars, such as addition, subtraction, multiplication, and division.

Vectors

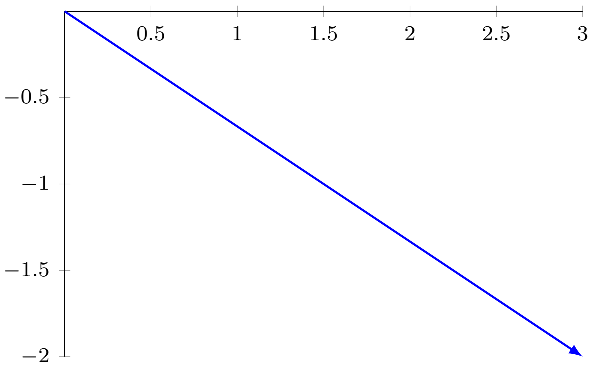
A vector is a one-dimensional array of numbers, such as

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} \sqrt{3} \\ -2 \end{bmatrix}, \qquad \text{or} \qquad \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

The dimension of a vector is its number of elements. For example,

$$v = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

is a two-dimensional vector since it has two elements. In two dimensions, we can visualize a vector as an arrow extending from the origin to its coordinates in the plane.



The inner product of two vectors with the same dimension is the scalar

$$\langle a, b \rangle = \sum_i a_i \, b_i$$

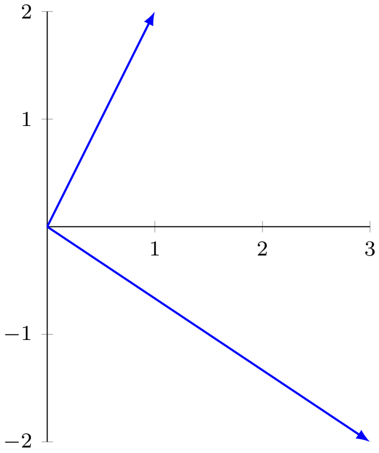
For example,

$$\left\langle \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \right\rangle = (1)(2) + (2)(0) + (3)(1) = 5$$

Two vectors of the same dimension are *orthogonal* if their inner product is zero, $\langle a, b \rangle = 0$. For example,

$$\left\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 \\ -2 \end{bmatrix} \right\rangle = (1)(4) + (2)(-2) = 0$$

Orthogonality is a generalization of two lines being perpendicular. In two dimensions, vectors are orthogonal if they are perpendicular to each other.



The *length* of a vector is the square root of its inner product with itself.

$$|a| = \sqrt{\langle a, a \rangle}$$

For example,

$$\left| \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}$$

In two dimensions, the length is the standard length of the vector from the origin (this is the Pythagorean theorem).

Matrices

A matrix is a rectangular array of numbers, such as

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \begin{bmatrix} -2 & \sqrt{3} \\ 1 & 5 \end{bmatrix}, \qquad \text{or} \qquad \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Notation. Matrices are typically denoted by uppercase letters, while their individual elements are represented by the same lowercase letter. For example, we might write $A = [a_{ij}]$ to denote a matrix A with elements a_{11}, a_{12}, and so on. ■

The *dimensions* of a matrix are the number of rows and columns. For instance, the matrix

$$\begin{bmatrix} 1 & 0 & 2 \\ 4 & 3 & 4 \end{bmatrix}$$

has dimensions 2 × 3, meaning that it has two rows and three columns.

The transpose of a matrix is denoted A^{T}, and is a matrix with the rows and columns switched. For example,

$$\begin{bmatrix} 1 & 0 & 2 \\ 4 & 3 & 4 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 1 & 4 \\ 0 & 3 \\ 2 & 4 \end{bmatrix}$$

The matrix has dimensions 2 × 3 while its transpose has dimensions 3 × 2. The first row of the matrix becomes the first column of its transpose, and so on.

The sum of two matrices of the same dimensions is

$$A + B = [a_{ij} + b_{ij}]$$

Each element of the sum is the sum of the corresponding elements of the two matrices. For example,

$$\begin{bmatrix} 1 & 0 & 2 \\ 4 & 3 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 2 \\ 4 & 4 & 4 \end{bmatrix}$$

A matrix can be multiplied by a scalar α, which multiplies every element of the matrix by the scalar.

$$\alpha A = [\alpha\, a_{ij}]$$

For example,

$$3 \begin{bmatrix} 1 & 0 & 2 \\ 4 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 6 \\ 12 & 9 & 12 \end{bmatrix}$$

Matrix multiplication

Consider a matrix A with dimensions $m \times n$ and a matrix B with dimensions $p \times q$. We can multiply A times B only if $n = p$, in which case the product is the $m \times q$ matrix

$$C = AB = [c_{ij}] \qquad \text{where} \qquad c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}$$

For example,

$$\begin{bmatrix}1 & 0 & 2\\ 4 & 3 & 4\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}1 & 1\\ 7 & 4\end{bmatrix}$$

Unlike scalar multiplication, matrix multiplication in general is not commutative! So $AB \neq BA$. For example,

$$\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix} = \begin{bmatrix}4 & 6\\ 3 & 4\end{bmatrix} \qquad \neq \qquad \begin{bmatrix}1 & 3\\ 3 & 7\end{bmatrix} = \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$$

The product may not even have the same dimensions,

$$\begin{bmatrix}1 & 0 & 2\\ 4 & 3 & 4\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}1 & 1\\ 7 & 4\end{bmatrix} \qquad \neq \qquad \begin{bmatrix}5 & 3 & 7\\ 1 & 0 & 2\\ 0 & 0 & 0\end{bmatrix} = \begin{bmatrix}1 & 1\\ 1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & 0 & 2\\ 4 & 3 & 4\end{bmatrix}$$

The product of two matrices may exist when multiplied in one order but not in the other.

$$\begin{bmatrix}1 & 1\\ 1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix} = \begin{bmatrix}4 & 6\\ 1 & 2\\ 0 & 0\end{bmatrix} \qquad \neq \qquad \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\\ 0 & 0\end{bmatrix} \text{ does not exist!}$$

Determinant

The determinant of a square matrix is a scalar, denoted $\det(A)$. While the precise formula for the determinant is complicated, for a two-dimensional matrix it is simply the product of the diagonal terms minus the product of the off-diagonal terms:

$$\det\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix} = a_{11}\,a_{22} - a_{12}\,a_{21}$$

Special matrices

The identity matrix, typically denoted I, is a square matrix with ones on the diagonal and zeros on the off-diagonal. For example, the identity matrix in three dimensions is

$$I = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

Inverse

The inverse of a square matrix A is a square matrix A^{-1} of the same dimensions such that its product with the original matrix (in any order) is the identity matrix.

$$AA^{-1} = A^{-1}A = I$$

A matrix has an inverse only if its determinant is not zero. For a two-dimensional matrix with nonzero determinant, its inverse is the two-dimensional matrix

$$\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}^{-1} = \frac{1}{a_{11}\,a_{22} - a_{12}\,a_{21}}\begin{bmatrix}a_{22} & -a_{12}\\ -a_{21} & a_{11}\end{bmatrix}$$

B.2 Eigenvalues and eigenvectors

Definition. If multiplying a square matrix A by a nonzero vector x scales the vector by a (possibly complex) scalar λ, then λ is an *eigenvalue* of A with *eigenvector* x. That is, an eigenvalue and eigenvector satisfy the relationship

$$Ax = \lambda x.$$

A pair (λ, x) of an eigenvalue and its corresponding eigenvector is an *eigenpair*.

Remark (Dimensions)**.** If A has dimensions $n \times n$, then an eigenvector x must have dimension n, and A must be square so that the vector Ax also has dimension n. ■

Example. Consider the 2×2 matrix

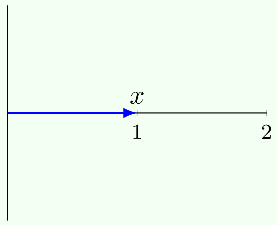
$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$$

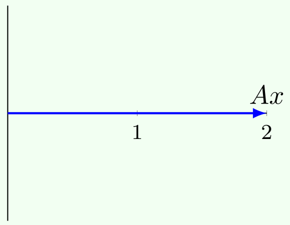
One eigenvector and eigenvalue pair is

$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \lambda = 2$$

We can directly verify that this satisfies the eigen equation

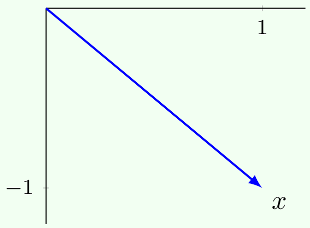
$$Ax = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \lambda x$$

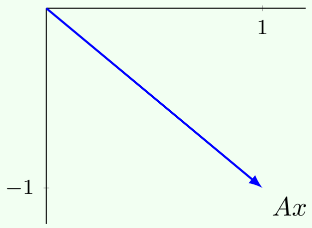




Another eigenvector and eigenvalue pair is

$$x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad \lambda = 1$$





Note that Ax points in the same direction as x and is scaled by λ.

Example (Differential). Signals are infinite-dimensional vectors, in which case the matrix A can be a general linear operator. For instance, the differential operator $\frac{\mathrm{d}}{\mathrm{d}t}$ maps a continuous-time signal to another continuous-time signal. Moreover, one of the basic results in calculus is that this operation is linear:

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(ax(t)+by(t)\big)=a\frac{\mathrm{d}}{\mathrm{d}t}x(t)+b\frac{\mathrm{d}}{\mathrm{d}t}y(t).$$

The eigenvectors (also called *eigenfunctions*) of the differential operator are the exponential signals $e^{\lambda t}$ with corresponding eigenvalue λ since they satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}e^{\lambda t}=\lambda\, e^{\lambda t}.$$

Note that this is true for *any* λ, so there are an infinite number of eigenpairs!

Computing eigenvalues

If (λ, x) is an eigenpair of A, then they satisfy $(\lambda I - A)x = 0$, so the columns of $\lambda I - A$ are linearly dependent, which implies that its determinant is zero. Therefore, the eigenvalues of A are the solutions to the polynomial equation

$$\det(\lambda I - A) = 0.$$

This is called the *characteristic equation* of A, and the polynomial $\det(\lambda I - A)$ is its *characteristic polynomial.*

Remark (Number of eigenvalues). If A is an $n \times n$ matrix, then its characteristic polynomial is a polynomial in λ of degree n, so it always has exactly n roots (that may be complex and/or repeated). ■

Example.

$$A=\begin{bmatrix}3 & 4\\ 2 & 1\end{bmatrix}$$

The characteristic equation is

$$\begin{aligned}
0=\det(\lambda I - A) &= \det\left(\lambda\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}-\begin{bmatrix}3 & 4\\ 2 & 1\end{bmatrix}\right)\\
&= \det\begin{bmatrix}\lambda-3 & -4\\ -2 & \lambda-1\end{bmatrix}\\
&= (\lambda-3)(\lambda-1)-(-4)(-2)\\
&= \lambda^2-4\lambda-5\\
&= (\lambda+1)(\lambda-5)
\end{aligned}$$

Therefore, the eigenvalues are -1 and 5.

Example (complex eigenvalues).

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

The characteristic equation is

$$0 = \det(\lambda I - A) = \det \begin{bmatrix} \lambda & -1 \\ 1 & \lambda \end{bmatrix} = \lambda^2 + 1$$

Therefore, the eigenvalues are $\pm j$. Since the eigenvalues are the roots of a polynomial, they may be complex.

Computing eigenvectors

Given an $n \times n$ matrix A and a (possibly complex) scalar eigenvalue λ, we can find the n-dimensional eigenvector x associated with the eigenvalue by solving the eigen equation

$$(\lambda I - A)x = 0$$

Example.

$$A = \begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} \qquad \Longrightarrow \qquad \lambda_1 = 5, \quad \lambda_2 = -1$$

For the eigenvalue λ_1, we can find the associated eigenvector $x_1 = \begin{bmatrix} a \\ b \end{bmatrix}$ by solving the equation

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 5-3 & -4 \\ -2 & 5-1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 & -4 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

which implies that

$$\begin{aligned} 0 &= 2a - 4b \\ 0 &= -2a + 4b \end{aligned}$$

The second equation is simply the negative of the first equation, so it provides no additional information. The first equation implies that $a = 2b$, so the eigenvector is

$$x_1 = \begin{bmatrix} 2b \\ b \end{bmatrix} \qquad \text{for any } b$$

For the second eigenvalue λ_2, the corresponding eigenvector $x_2 = \begin{bmatrix} a \\ b \end{bmatrix}$ is the solution to

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1-3 & -4 \\ -2 & -1-1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -4 & -4 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

The second equation is simply half the first equation, so again it provides no additional information. The first equation implies that $a = -b$, so the second eigenvector is

$$x_2 = \begin{bmatrix} -b \\ b \end{bmatrix} \qquad \text{for any } b$$

The eigenvectors are not unique since they can be scaled. We usually just pick a convenient scaling. For instance,

$$\lambda_1 = 5, \quad x_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \lambda_2 = -1, \quad x_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

B.3 Matrix similarity

The mapping $A \mapsto T^{-1}AT$ is called a *similarity transformation*, and two matrices A and B are *similar* if there exists an invertible matrix T such that $A = T^{-1}AT$, which is denoted $A \sim B$.

The following result provides several useful properties of similar matrices.

Proposition (Similar matrices)**.** If $A \sim B$, then

a) $A^k \sim B^k$ for $k = 0, 1, 2, \dots$

b) $e^{At} \sim e^{Bt}$

c) $(sI - A)^{-1} \sim (sI - B)^{-1}$

d) $\det(sI - A) = \det(sI - B)$

e) A and B have the same eigenvalues

■

Proof. If A and B are similar, then there exists an invertible matrix T such that $B = TAT^{-1}$.

a) $B^k = (TAT^{-1})(TAT^{-1})\cdots(TAT^{-1}) = TA^kT^{-1} \sim A^k$.

b) $e^{Bt} = \sum_{k=0}^{\infty} \frac{1}{k!} B^k t^k = \sum_{k=0}^{\infty} \frac{1}{k!} TA^kT^{-1}t^k = Te^{At}T^{-1} \sim e^{At}$

c) $(sI - B)^{-1} = (sI - TAT^{-1})^{-1} = \big(T(sI - A)T^{-1}\big)^{-1} = T(sI - A)^{-1}T^{-1} \sim (sI - A)^{-1}$

d) $\det(sI-B) = \det\big(T(sI-A)T^{-1}\big) = \det(T)\det(sI-A)\det(T^{-1}) = \det(sI-A)\det(TT^{-1}) = \det(sI-A)$

e) This follows from the fact that A and B have the same characteristic polynomials (item (d)). ■

B.4 Subspace

Definition (Subspace)**.** A *subspace* S of a vector space V is a subset of V that is itself a vector space. Equivalently, a subspace S is a subset of V that:

- contains zero: $0 \in S$
- is closed under addition: if $x, y \in S$, then $x + y \in S$
- is closed under scalar multiplication: if $x \in S$ and a is a scalar, then $ax \in S$

There are four fundamental subspaces associated with a matrix A.

Definition (Fundamental subspaces)**.** The four fundamental subspaces associated with an $m \times n$ real matrix A are the following:

$$\operatorname{col}(A) = \{Ax \mid x \in \mathbb{R}^n\} \qquad \text{(column space)}$$

$$\operatorname{row}(A) = \{A^\mathsf{T} y \mid y \in \mathbb{R}^m\} \qquad \text{(row space)}$$

$$\operatorname{null}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\} \qquad \text{((right) null space)}$$

$$\operatorname{null}(A^\mathsf{T}) = \{y \in \mathbb{R}^m \mid A^\mathsf{T} y = 0\} \qquad \text{(left null space)}$$

The row space and null space are subspaces of $\mathbb{R}^n$, while the column space and left null space are subspaces of $\mathbb{R}^m$. The interpretations of these subspaces are as follows. The column space is the set of all linear combinations of the columns of the matrix, while the row space is the set of all linear combinations of its rows. The right null space is the set of all vectors that are mapped to zero when multiplied by A on the right, while the left null space is the set of all vectors that are mapped to zero when multiplied by A on the left.

Another important subspace is the span of a set of vectors.

Definition (Span)**.** The *span* of a set of vectors is the subspace of all linear combinations of the vectors:

$$\operatorname{span}(x_1, x_2, \ldots, x_n) = \{a_1x_1 + \ldots + a_nx_n \mid a_1, \ldots, a_n \text{ scalars}\}.$$

Example. Consider the matrix

$$A = \begin{bmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 4 & 5 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The column space and left null space are

$$\operatorname{col}(A) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) \qquad \text{and} \qquad \operatorname{null}(A^\mathsf{T}) = \operatorname{span}\left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right),$$

and the row space and right null space are

$$\operatorname{row}(A) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 4 \\ 5 \end{bmatrix} \right) \qquad \text{and} \qquad \operatorname{null}(A) = \operatorname{span}\left(\begin{bmatrix} -2 \\ -4 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ -5 \\ 0 \\ 1 \end{bmatrix} \right).$$

Definition (Subspace dimension)**.** The *dimension* of a subspace S, denoted $\dim(S)$, is the number of vectors in any basis for that subspace.

If the vectors are linearly independent, then the dimension of their span is the number of vectors.

Definition (Orthogonal complement)**.** The orthogonal complement of a subspace S is the set of vectors that are orthogonal to all vectors in the subspace:

$$S^\perp = \{v \mid u^\mathsf{T} v = 0 \text{ for all } u \in S\}.$$

Fact. The orthogonal complement is an *involution*, meaning that the orthogonal complement of the orthogonal complement is the original subspace:

$$(S^\perp)^\perp = S.$$

Definition (Sum and direct sum of subspaces)**.** The sum of two subspaces U and V is the subspace

$$U + V = \{u + v \mid u \in U \text{ and } v \in V\}.$$

Moreover, this is called the *direct sum*, denoted $U \oplus V$, when the decomposition is unique.

Fact. The orthogonal complement of the four fundmantal subspaces are as follows:

$$\begin{aligned} \operatorname{row}(A)^\perp &= \operatorname{null}(A), \\ \operatorname{col}(A)^\perp &= \operatorname{null}(A^\mathsf{T}), \\ \operatorname{null}(A)^\perp &= \operatorname{row}(A), \\ \operatorname{null}(A^\mathsf{T})^\perp &= \operatorname{col}(A). \end{aligned}$$

Fact. If V is an inner product space and U is a subspace of V, then V is the direct sum of U and its orthogonal complement,

$$U \oplus U^{\perp} = V.$$

In particular, for the four fundamental spaces associated with a matrix $A \in \mathbb{R}^{m\times n}$,

$$\operatorname{row}(A) \oplus \operatorname{null}(A) = \mathbb{R}^n \qquad \text{and} \qquad \operatorname{col}(A) \oplus \operatorname{null}(A^{\mathsf{T}}) = \mathbb{R}^m.$$

B.5 Diagonalization

Suppose (λ_k, x_k) is a set of eigenpairs for $k = 1, \ldots, n$, and define the matrices

$$X = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} \qquad \text{and} \qquad \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Then, the matrix form of the eigen equation is

$$\begin{aligned} AX &= A\begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 x_1 & \lambda_2 x_n & \ldots & \lambda_n x_n \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \\ &= X\Lambda. \end{aligned}$$

If all eigenvectors are linearly independent, then the matrix X is invertible, in which case we can solve the eigen equation for the matrix of eigenvalues as

$$X^{-1}AX = \Lambda.$$

So applying a similarity transformation with the matrix of eigenvectors produces the diagonal matrix of eigenvalues. In this case, we say that the matrix A is *diagonalizable*, and the matrix X diagonalizes A.

Definition (Diagonalizable)**.** A matrix is *diagonalizable* if it is similar to a diagonal matrix.

As we saw above, if the eigenvectors are all linearly independent, then the matrix of eigenvectors is invertible and diagonalizes the matrix. The following result describes when the eigenvectors are all linearly independent.

Fact. If all eigenvalues are distinct, then all eigenvectors are linearly independent.

This implies that a matrix is diagonalizable when all of its eigenvalues are distinct. But in general, some eigenvalues may be repeated. The characteristic polynomial then has the form

$$\det(\lambda I - A) = (\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^{n_2} \cdots (\lambda - \lambda_d)^{n_d}$$

where $n_1 + \ldots + n_d = n$ (the size of A) and d is the number of distinct eigenvalues.

Definition (Eigenspace)**.** The *eigenspace* associated with a complex number λ is the subspace

$$E_\lambda = \text{null}(\lambda I - A).$$

There are several cases for the eigenspace depending on the value of λ.

- If λ is not an eigenvalues, then $\lambda I - A$ is invertible, so the eigenspace is the trivial subspace $E_\lambda = \{0\}$.
- If λ is an eigenvalue, then the eigenspace is the subspace of all the associated eigenvectors.

Definition (Eigenvalue multiplicity)**.** Each complex number λ has two associated multiplicities:

- The *algebraic multiplicity* of λ is the number of times it is repeated as a root of the characteristic polynomial (denoted n_k).
- The *geometric multiplicity* of λ is the dimension of its corresponding eigenspace, $\dim(E_\lambda)$, which is the number of linearly independent eigenvectors with eigenvalue λ.

Fact. The eigenspace satisfies the following:

a) $1 \le \dim(E_{\lambda_k}) \le n_k$

b) If $x_i \in E_{\lambda_i}$ are eigenvectors in distinct eigenspaces, then $\{x_i\}$ are linearly independent.

Fact. A matrix is diagonalizable if and only if the algebraic multiplicity of each eigenvalue is equal to its geometric multiplicity:

$$n_k = \dim(E_{\lambda_k}) \quad \text{for all } k.$$

Intuitively, this means that each eigenspace must have "enough" linearly independent eigenvectors.

Example. The matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

has eigenvalues at 3 and -1 with corresponding eigenspaces

$$E_3 = \text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \qquad \text{and} \qquad E_{-1} = \text{span}\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right).$$

Each eigenspace has dimension one (which is the same as the multiplicity of each eigenvalue), so the matrix is diagonalizable. To diagonalize the matrix, perform a similarity transformation with the matrix of eigenvectors:

$$T = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad \text{with inverse} \qquad T^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

The diagonal form is then

$$T^{-1}AT = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix},$$

which is a diagonal matrix containing all of the eigenvalues on the diagonal.

Example. The matrix $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ has two eigenvalues at zero, but the corresponding eigenspace is

$$E_0 = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$$

which has dimension one, so the matrix is *not* diagonalizable.

B.6 Jordan form

While not every matrix is diagonalizable, any square matrix is similar to a matrix in Jordan form, which is a generalization of a diagonal matrix. The Jordan form of a matrix is a decomposition of the matrix into simple blocks, which is often useful to understand properties of the matrix. If the matrix is diagonalizable, then the Jordan form is equivalent to the eigendecomposition of the matrix.

A *Jordan matrix* is a block-diagonal matrix of the form

$$J = \begin{bmatrix} J_{k_1}(\lambda_1) & & & \\ & J_{k_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{k_d}(\lambda_d) \end{bmatrix}$$

with $k_1 + k_2 + \ldots + k_d = n$, where each diagonal block $J_{k_i}(\lambda_i)$ is a *Jordan block* of the form

$$J_{k_i}(\lambda_i) \quad = \quad \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \lambda_i & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{bmatrix},$$

which is a $k_i \times k_i$ matrix where all unspecified entries are zero. Each Jordan block has the form $J_{k_i}(\lambda_i) = \lambda_i I + N$, where N is a matrix of zeros except for ones on the superdiagonal; such a matrix is called *nilpotent*. Intuitively, a Jordan matrix is "almost" diagonal.

The Jordan form of a square matrix A is

$$A = T^{-1} J T$$

where T is an invertible matrix and J is a Jordan matrix.

Example (Jordan form)**.** Some examples of Jordan forms are as follows:

$$\underbrace{\begin{bmatrix} -1 & -9 \\ 1 & 5 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 3 & 5 \\ -1 & -2 \end{bmatrix}}_{T^{-1}} \underbrace{\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}}_{J} \underbrace{\begin{bmatrix} 2 & 5 \\ -1 & -3 \end{bmatrix}}_{T}$$

$$\underbrace{\begin{bmatrix} 6 & -4 & -1 & -11 & -1 \\ -7 & 3 & 1 & 11 & 1 \\ -5 & 1 & -2 & 6 & 1 \\ 7 & -4 & -1 & -12 & -1 \\ 5 & -2 & 0 & -7 & -2 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 0 & -1 & -1 & 1 & 1 \\ 3 & 1 & 2 & -1 & -1 \\ -2 & 0 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ 1 & 0 & -2 & 0 & 1 \end{bmatrix}}_{T^{-1}} \underbrace{\begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & -2 \end{bmatrix}}_{J} \underbrace{\begin{bmatrix} 1 & 0 & 0 & -1 & 0 \\ -9 & 4 & 1 & 12 & 2 \\ -2 & 1 & 0 & 3 & 0 \\ -5 & 3 & 1 & 8 & 1 \\ -5 & 2 & 0 & 7 & 1 \end{bmatrix}}_{T}$$

Fact. The Jordan form of a matrix can be used to determine the multiplicities of its eigenvalues:

a) The algebraic multiplicity of an eigenvalue is the sum of the sizes of all Jordan blocks corresponding to that eigenvalue.

b) The geometric multiplicity of an eigenvalue is the number of Jordan blocks corresponding to that eigenvalue.

Therefore, a matrix is diagonalizable if and only if all of its Jordan blocks have dimension one.

B.7 Matrix exponential

Recall that the series expansion for the exponential of a scalar x is

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \dots$$

In the same way, we define the *matrix exponential* of a square matrix X in terms of its series expansion,

$$e^X = \sum_{n=0}^{\infty} \frac{1}{n!} X^n = I + X + \frac{1}{2}X^2 + \frac{1}{6}X^3 + \dots$$

Remark. The matrix exponential is defined by its series expansion, which is *not* the same as taking the exponential of each component of the matrix! ■

Example (diagonal matrix)**.** Consider the diagonal matrix

$$A = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}.$$

The powers of this matrix are as follows:

$$A^k = \begin{bmatrix} \lambda_1^k & 0 & 0 \\ 0 & \lambda_2^k & 0 \\ 0 & 0 & \lambda_3^k \end{bmatrix}.$$

Therefore, the matrix exponential is the diagonal matrix whose elements are the (standard) exponentials of each diagonal element:

$$e^{At} = \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 \\ 0 & e^{\lambda_2 t} & 0 \\ 0 & 0 & e^{\lambda_3 t} \end{bmatrix}.$$

Example (nilpotent matrix)**.** Consider the matrix that is all zero except with ones on the superdiagonal:

$$N = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The powers of this matrix are as follows:

$$N^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad N^3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \text{and} \qquad N^n = 0 \text{ for all } n \geq 4.$$

This type of matrix is called *nilpotent* since it is zero when raised to a high enough power (in this case, four). Since only a finite number of powers are nonzero, the series expansion is finite.

$$e^{Nt} = I + Nt + \frac{1}{2}N^2t^2 + \frac{1}{6}N^3t^3 = \begin{bmatrix} 1 & t & \frac{1}{2}t^2 & \frac{1}{6}t^3 \\ 0 & 1 & t & \frac{1}{2}t^2 \\ 0 & 0 & 1 & t \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Properties

The following properties hold for any square matrices A and B, and any scalars t, t_1, and t_2.

- e^{At} is the unique matrix satisfying $\frac{\mathrm{d}}{\mathrm{d}t}\big(e^{At}\big) = A\,e^{At}$
- $e^{A(t_1+t_2)} = e^{At_1}e^{At_2}$
- $\big(e^{At}\big)^{-1} = e^{-At}$, so the matrix exponential is always invertible
- $A\,e^{At} = e^{At}\,A$, so the matrix exponential commutes with the matrix in the exponent
- $\big(e^{At}\big)^{\mathsf{T}} = e^{A^{\mathsf{T}}t}$, so transposing the matrix exponential transposes the matrix in the exponent
- $e^{(A+B)t} = e^{At}e^{Bt}$ for all t if and only if A and B commute (that is, $AB = BA$); this always holds if A and B are scalars

Laplace transform of the matrix exponential. Recall that the Laplace transform of a scalar exponential e^{at} is $\frac{1}{s-a}$. The Laplace transform of a matrix exponential generalizes as follows:

$$\mathcal{L}(e^{At}) = (sI - A)^{-1}$$

Proof.

$$\begin{aligned}
\mathcal{L}(e^{At}) &= \mathcal{L}\left(\sum_{n=0}^{\infty}\frac{1}{n!}A^n t^n\right) && \text{(definition of matrix exponential)}\\
&= \sum_{n=0}^{\infty}\frac{1}{n!}A^n\mathcal{L}(t^n) && \text{(linearity of Laplace transform)}\\
&= \sum_{n=0}^{\infty}\frac{1}{n!}A^n\frac{n!}{s^{n+1}}\\
&= s^{-1}\sum_{n=0}^{\infty}A^n s^{-n}\\
&= s^{-1}\sum_{n=0}^{\infty}(s^{-1}A)^n\\
&= s^{-1}(I-s^{-1}A)^{-1} && \text{(geometric series)}\\
&= (sI-A)^{-1}
\end{aligned}$$

Note that the Laplace transform is only defined in its region of convergence, which are the values of s for which the geometric series converges. ■

Computation

To compute the matrix exponential, first transform the matrix to Jordan form

$$A = T^{-1}JT$$

where T is invertible and $J = \operatorname{diag}(J_1, J_2, \dots, J_m)$ is a Jordan matrix in which each J_i a Jordan block:

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 & 0\\ 0 & \lambda_i & 1 & \dots & 0 & 0\\ 0 & 0 & \lambda_i & \ddots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & 0 & \dots & \lambda_i & 1\\ 0 & 0 & 0 & \dots & 0 & \lambda_i \end{bmatrix}$$

Each Jordan block has the form $J_i = \lambda_i I + N$, where N is nilpotent. The powers of the matrix are then

$$A^n = \left(T^{-1}JT\right)^n = \underbrace{\left(T^{-1}JT\right)\left(T^{-1}JT\right)\dots\left(T^{-1}JT\right)}_{n\text{ times}} = T^{-1}J^nT$$

So we can raise the matrix A to a power simply by raising its Jordan matrix to the power. The matrix exponential is then

$$e^{At} = \sum_{n=0}^{\infty}\frac{t^n}{n!}A^n = \sum_{n=0}^{\infty}\frac{t^n}{n!}T^{-1}J^nT = T^{-1}\left(\sum_{n=0}^{\infty}\frac{t^n}{n!}J^n\right)T = T^{-1}e^{Jt}T$$

The matrix exponential of a block matrix is the block matrix exponentials:

$$\exp\left(\begin{bmatrix} A_1 & 0\\ 0 & A_2\end{bmatrix}\right) = \begin{bmatrix} e^{A_1} & 0\\ 0 & e^{A_2}\end{bmatrix}$$

And the matrix exponential of a single Jordan block is simple:

$$e^{Jt} = \begin{bmatrix} e^{\lambda t} & t\,e^{\lambda t} & \frac{1}{2}t^2 e^{\lambda t} & \dots & \frac{1}{(k-1)!}t^{k-1}e^{\lambda t} \\ 0 & e^{\lambda t} & t\,e^{\lambda t} & \dots & \frac{1}{(k-2)!}t^{k-2}e^{\lambda t} \\ 0 & 0 & e^{\lambda t} & \dots & \frac{1}{(k-3)!}t^{k-3}e^{\lambda t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda t} \end{bmatrix}$$

B.8 Quadratic forms

A quadratic form is a homogeneous quadratic function. Given a symmetric matrix A, the corresponding quadratic form is

$$f(x) = x^{\mathsf{T}} A x.$$

Minimizing quadratic forms

Ellipses

C

Polynomials

A polynomial is a function of the form

$$p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$$

The scalars $a_1, a_2, \ldots, a_n$ are the coefficients of the polynomial, and the number n is the degree of the polynomial (assuming that the coefficient a_n is nonzero).

C.1 Roots

The roots of a polynomial are the solutions to the equation $p(x) = 0$. There are several fundamental results from algebra concerning the roots of a polynomial.

- **Fundamental thereom of algebra:** every polynomial of degree n has exactly n complex roots (counting multiplicities)
- **Complex conjugate root theorem:** for polynomials with real coefficients, complex roots always appear in conjugate pairs

C.2 Quadratic polyomials

Quadratic polynomials have the form

$$p(x) = ax^2 + bx + c$$

There are several ways to find the roots of quadratic polynomials.

- **Factor:** If we can factor the polynomial as $(x - p_1)(x - p_2)$, then the roots must be p_1 and p_2 since they make each individual factor zero. For example, the quadratic

  $$x^2 - 3x + 2 = (x - 1)(x - 2)$$

  has roots 1 and 2. However, not all polynomials factor!
- **Quadratic formula:** For quadratic polynomials, we can always use the quadratic formula to find the roots. It says that the roots are

  $$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

  The ± gives two roots, and the structure of the roots is characterized by the discriminant $b^2 - 4ac$.

- if $b^2 - 4ac > 0$, then the roots are real and distinct
- if $b^2 - 4ac = 0$, then the roots are real and repeated
- if $b^2 - 4ac < 0$, then the roots are complex conjugates

D

Geometric series

$$\sum_{m=0}^{k} a^m = \frac{1-a^{k+1}}{1-a}\, u_s(k) \qquad a \neq 1$$

$$\sum_{m=0}^{\infty} a^m = \frac{1}{1-a} \qquad |a| < 1$$

$$\sum_{m=0}^{k} m\, a^m = \frac{a}{(1-a)^2}\big(1 - a^k - (1-a)\,k\,a^k\big)\, u_s(k) \qquad a \neq 1$$

$$\sum_{m=0}^{\infty} m\, a^m = \frac{a}{(1-a)^2} \qquad |a| < 1$$

$$\sum_{m=0}^{k} m = \frac{k\,(k+1)}{2}\, u_s(k)$$

$$\sum_{m=0}^{k} m^2 = \frac{k\,(k+1)(2k+1)}{6}\, u_s(k)$$

Proof. To illustrate how to construct these formulas, we will show how to derive the first formula. First note that the summation is empty if $k < 0$, in which case the value of the summation is zero due to the unit step. Now suppose $k \geq 0$ and let S_k denote the value of the summation up to k,

$$S_k = \sum_{m=0}^{k} a^m = 1 + a + a^2 + \ldots + a^k$$

Multiplying by a yields

$$aS_k = a + a^2 + a^3 + \ldots + a^{k+1}$$

We now subtract the second equation from the first one. All of the middle terms cancel and we are left with

$$S_k - aS_k = 1 - a^{k+1}$$

Solving for S_k (assuming $a \neq 1$) yields the expression in the formula. ■