Measurement

an introduction

Rico A. R. Picone
Department of Mechanical Engineering
Saint Martin's University

Fall 2017*

*These notes are updated frequently. Please be sure download the most recent version.

Contents

01

Theoretical foundations

Measurement is fundamental to science and technology. It may seem at first to be simple because it is familiar. A length, a duration, a mass—these types of quantities are the building blocks of our scientific conception of the world. To know their actuality in a given instance is the primary aim of measurement.

There are several manners in which philosophers, scientists, and other thinkers have tried to understand measurement. Rather than rushing ahead with our intuitive understanding, we will slow down and consider these theoretical foundations of measurement.[1]

[1] A great summary of the theories that follow can be found in Tal (2017).

Lecture 01.01 Mathematical measurement theory

Mathematical measurement theory (sometimes just "measurement theory") concerns itself with how relations among mathematical objects (e.g. numbers) can represent relations among qualities of empirical objects (e.g. length). For instance, summing numbers is analogous to how the quality of length of a composite object is related to the quality of length of each of the object's constituents, as is illustrated in Figure 01.1.

But there's subtlety here that's best considered with precise language. Let's consider some fundamental definitions of measurement theory.

01.01.1 Quality, quantity, magnitude, and scale

When trying to describe what we are doing when we measure, there are certain terms are seemingly unavoidable. Therefore, it's worth considering some precise definitions of them. In the following, let's take the term *object* to mean the object of measurement.

object

Definition 01.01.1: quality

The *quality* of an object is the manner in which it interacts. It is the totality of its properties, which are aspects of the way the object interacts. (Spirkin, 1983)

We can think of the properties that constitute the quality of an object as sets to which a given object belongs or not, like "heavy" or "round." Immediately, however, we become suspicious that a real object can belong to such a set so completely or not. *Fuzzy set theory* allows members of a set to belong *to a certain degree* (Ross, 2010). Applying fuzzy set theory to measurement theory is beyond the scope of this text, but suffice it to say that the fuzziness of membership suggests a blurring of the boundary between quality and our next definition, quantity.

fuzzy set theory

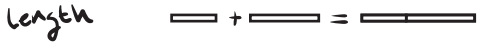


Figure 01.1: mathematical measurement theory explores the correspondence between mathematical objects like numbers and empirical qualities like lengths.

Definition 01.01.2: quantity

The *quantity* of an object is the amount of that which comprises it. The quantity of a finite collection of objects is the number of them. (Spirkin, 1983)

Note that we are already beginning to use mathematical analogy in our definition of quantity when we use a number to represent quantity. Quantity can be continuous or discrete. In the former case, it is often represented by a *real number*; in the latter, it is often represented by an *integer*.

real number

integer

We have used the term "amount" in the definition of the quantity of an object. This is a bit of a swindle, considering we have not yet defined this term. In fact, "amount" is not the usual scientific term. Rather, the term *physical magnitude* has emerged from physics (Hall, 2016). It is here that we bump against the limitation of language to define fundamental phenomena: physical magnitude is typically defined as a that which can be represented by a number about an object. Thus, we have interdependent definitions of quantity and physical magnitude when describing an object. We use both terms interchangeably.

physical magnitude

"I got a 20 on the exam." Without a measurement *scale*, there is no measurement.

Definition 01.01.3: measurement scale

A measurement *scale* is a mapping of quantities and qualities of an object to mathematical objects for representation. (Tal, 2017)

Measurement theory is hardly homogeneous, but we can think of it as being primarily comprised of considerations of (1) the nature of the objects of measurement and (2) the ways in which the correspondance between an object and its measure can be established. These are considered in the following sections (01.01.2, 01.01.3, and 01.01.4).

01.01.2 The nature of measurement objects

Every theory assumes an *ontology* (in the sense of metaphysics): a theoretical understanding of the nature of being. Unfortunately, we rarely consider ontology and instead thrash about with some assumed ontology—for there is no theory that does not have at least an implicit ontology. We will pause at ontology for just a moment before thrashing on.

ontology

There are several competing understandings of the ontological status of the objects of measurement. Tal (2017) describes them as

- concrete individual objects,
- qualitative observations of concrete individual objects,
- abstract representations of individual objets, and
- universal properties of objects.

This is especially important to realist theories of measurement, but is important to consider in all measurement theories.

01.01.3 Establishing scales

So, under mathematical measurement theory, quantities and qualities of an object are said to correspond in some way to mathematical objects. But how do we decide on the mathematical objects (scale)? What criteria are there for determining the efficacy of the mathematical objects?

This is another universal aspect of measurement theories: establishing the way in which scales can be properly established. It is a central consideration of most measurement theories.

01.01.4 Intrinsic and extrinsic quantities

intensive quantity

extensive quantity

Intensive quantities are those that represent properties of the constitutive substance of an object. Conversely, *extensive quantities* are those that are unique to each object.[2]

The quality of quantities (lol) to admit representation by a number leads to a simple manner in which to define the difference between an intrinsic and an extrinsic quantity: if an attribute of an object can be represented by the addition of numbers, it is an extrinsic quantity; otherwise, it is an intrinsic quality.[3] For instance, weight is best represented by a quantity because combining two objects with weights represented by w_1 and w_2 gives a composite object with weight $w_1 + w_2$. Similarly, the densities of two objects ρ_1 and ρ_2, if the objects are combined, are not.

[2] There's some ambiguity here, but this is approximately how the terms are used in the study of thermodynamics.

[3] The idea for this comes from (Campbell, 1920, p. 267), but his concepts of quantity and quality seem conflated with what we have called intrinsic and extrinsic quantities.

01.01.5 Fundamental and derived magnitudes

Early theories (Campbell, 1920) made the additivity and lack thereof of magnitudes the crucial aspect of a distinction between two types of physical magnitude: *fundamental* and *derived*.

fundamental magnitude
derived magnitude

Later theories distinguished between these two types of magnitude in the following way:

- *fundamental magnitudes* are those that can be measured directly and
- *derived magnitudes* are those that must be computed from a definition that depends on fundamental magnitudes.

For instance, density is measured by measuring both mass and volume of an object, then dividing them—making density a derived magnitude.

This means there is nothing intrinsic about the difference between a fundamental and a derived quantity; rather, a magnitude that is derived now may become fundamental if a method for measuring it directly is developed.

Example 01.01-1 fundamental versus derived

Of the following magnitudes, which is fundamental and which is derived?

1. resistance
2. length
3. mass
4. weight

There is an ambiguity here that I want to merely suggest and leave open. Let us take mass, for instance. We say it can be compared directly to another mass via a balance and therefore it is fundamental. However, how do we determine when a balance is balanced? By measuring, for instance, the angle of the balancing arm, which is surely never zero. It can only be "small enough." This hints at an issue with our conception of fundamental and derived quantities.

01.01.6 Classification of scales

Measurement scales have been classified by the types of transformation to which they are invariant without loss of empirical information. We will consider the following scales originated by Stevens (Tal, 2017; Robert, 1985).

nominal scales Nominal scales are those that are invariant to one-to-one substitution. Those that have no order are quintessential. For instance, gender or concave/convex (innie/outie) navels are nominal scales.

ordinal scales Ordinal scales are those that are invariant to monotonic, increasing transformations. Those that have a specific order are quintessential. For instance, one could feel *terribly*, *poorly*, or *meh*. Another example is physical hardness.

interval scales Interval scales are those that are invariant to positive linear transformation. Celcius and Farenheit scales for temperatures are related by just such a transformation

$$T_F = \frac{9}{5} \cdot T_C + 32 \tag{01.1}$$

without a loss in emperical information.

ratio scales Ratio scales are those that are invariant to multiplication by positive numbers. For instance, length can be represented in meters or kilometers via multiplication by a constant. Kelvin, unlike Celcius and Farenheit, is a ratio scale if negative Kelvin temperatures are excluded from consideration. Whenever a scale admits positive multiplication and excludes negative values (i.e. has an "absolute zero"), it is considered to be a ratio scale.

Example 01.01-2 scale classification

Classify the following measurement scales.

1. mass in kg
2. air quality index
3. numbered uniforms 1-99
4. time interval in sec
5. calendar time (e.g. 2017)

01.01.7 Representational theory

Representational Theory of Measurement

The *Representational Theory of Measurement* (RTM) is the most generally accepted mathematical measurement theory. It combines the considerations above—the nature of measurement objects and the classification of scales—to define measurement as "the construction of mappings from empirical relational structures into numerical relational structures" Tal (2017).

homomorphism

In this theory, measurement scales are *homomorphisms* (many-to-one mappings) from empirical relational structures to numerical relational structures.

Lecture 01.02 Operationalism, conventionalism, and realism

Mathematical measurement theory developed alongside another dimension of the study of measurement. This dimension is mostly concerned with the "reality" and meaning of measurement.

01.02.1 Operationalism

Most operationalists hold that the terms we apply to quantities—for instance, "duration" or "length"—depend essentially on the operations we use to measure them. In fact, one operationalist goes so far as to claim that

> *we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations. (Tal, 2017)*

Implications include that using two different instruments—say a ruler and calipers—to measure what we would typically consider to be the "length" of the same object would, in fact, need to be described as two different quantity-terms such as "length-ruler" and "length-calipers."

logical positivism

Logical positivism—a now-defunct philosophical school popular in the 1920s and 1930s in Europe, the central thesis of which is the theory that only those statements that are empirically verifiable are meaningful[4]—initially embraced this view. However, as with positivism, operationalism was found to have many issues, including (Tal, 2017):

- operationalism seems to imply that a measurement is automatically reliable,
- meaning seems to apply beyond the strict criteria of operationalism,
- operational definitions cannot be applied to some useful theoretical concepts, and
- the concept of operation itself is ambiguous.

For these reasons and others, operationalism was outpaced by the approach we turn to next.

[4] This theory is called *verificationism* and is still to be found in public discourse. This is unfortunate because philsophers have long since abandoned it along with positivism.

01.02.2 Conventionalism

A sort of operationalism-lite, *conventionalism* says that many of the quantities we define, such as temperature, are conventional. Ernst Mach, for instance, claimed that there is no possible truth or falsity to the question of which thermometric fluid expands more uniformly because temperature intervals are defined in terms of the expansion of a thermometric fluid (Tal, 2017). This is called the *principle of coordination*.

principle of coordination

Logical positivists Hans Reichenbach and Rudolf Carnap used conventionalism alongside their verificationism (unverifiable statements are neither true nor false). A *coordinative definition* of an unverifiable statement like "a meter is the length of a standard rod in Paris" (Fieser and Dowden, 2017)—this is actually how the SI system used to define a meter. (Now the SI uses the definition: "the length of the path travelled by light in vacuum during a time interval of $1/299,792,458$ of a second" (The International System of Units, 1984).) These sorts of conventional definitions were used to supplement explicit definitions.

coordinative definition

01.02.3 Realism

Most realists argue that, independent of convention or belief or measurement instrument—that is, *objectively*—objects have some real measureable properties. These properies can include those that are psychologically measureable (i.e. some subjective experiences can be measured). They are typically considered to be *estimated* by a measurement process.

estimation

We find that ordering objects by length is a very repeatable process. Similarly, we find that concatenating objects "lengthwise" yields a repeatable composite length, regardless of the ordering. Realism posits that the best explanation of these phenomena is that some objects have the property that they can relate to other objects with the relations "longer than" and "is the sum of."

Note that this means that lengths share a structure with real numbers, which can be related to each other by the relations "larger than" and "is the sum of." Some realists even go so far as to claim that we can define numbers themselves as ratios of quantities.

It is difficult to describe how the concepts of measurement *accuracy* and *error* without some form of realism (looking at you, operationalists and conventionalists). For the realist, the error is easy to define: it's the difference between the estimate and the real quantity.

Mathematical measurement theory is generally palatable to realists. However, measurement theorists have largely ignored the realists (Tal, 2017).

Example 01.02-1 which ism?

Decide which of school of thought might affirm each statement.

1. "There is no such thing as an objective property, only measurement processes."
2. "Measuring is estimating."
3. "A watch and an atomic clock measure different quantities."
4. "It is customary to define pressure as corresponding height of a column of mercury, which is precisely what pressure is."
5. "Whereof one cannot speak, thereof one must be silent."
6. "Do or do not, there is no try."

Lecture 01.03 Information-theoretic descriptions of measurement[5]

information theory

metrology

Information theory studies information quantification, storage, and communication. Liberal use of it is made by *metrology*: the study measurement and measurement applications.[6] A measurement instrument can be considered an "information machine" that takes an input *state* of an object and yields an output *reading*.

Information theory provides metrologists with a mathematical apparatus for computing the amount of information in a given information channel, typically in terms of probabilities. This same apparatus can be used for measurement, where we consider measurement to be the communication of information from the state of the measurement object to the measurement reading. The analogy between information transmission and measurement is incomplete in that information encoded and transmitted is (in theory) known by the encoder, whereas information the object "encodes" and "transmits" via the measurement instrument is typically considered unknowable any other way.

This is an active area of research. A related, recent proposal is one by Bas van Fraassen that posits two levels:

- a *physical* level at which the instrument yields a reading of the state of an object and
- an *abstract* level in which measurement reduces the region in which a mathematical representation of the physical state exists in a state space.

[5] See Tal (2017) for more information.

[6] Metrologists work for places like the US National Institute of Standards and Technology (NIST), research laboratories that develop precise measurements, and companies that develop precision instruments.

Lecture 01.04 Model-based descriptions of measurement[7]

Model-based descriptions of measurement frame measurement as

1. a concrete process of measurement object, measurement instrument, and their environment and
2. the construction of a theoretical and/or statistical model of the concrete process.

Correspondance between the measurement and the model are the primary consideration for the construction of models.

Two types of measurements are described:

instrument indications

measurement outcomes

1. *instrument indications* or *readings* are the output of the concrete process of measurement, such as the dial position on a pressure gauge;
2. *measurement outcomes* or *results* are knowlege claims about measurement objects that usually include a statement of probability or uncertainty, such as "the pressure in the tank is 50 ± 3 psi with 95% confidence."

Measurement outcomes are model-dependent and the uncertainty calculation depends on assumptions about the effectiveness of the model, including the statistical models of uncertainty.

The primary requirement of model-based descriptions of measurement is twofold:

1. coherence of the assumptions of the model and background theory and
2. consistency across instruments and environments.

[7] See Tal (2017) for much more information.

Lecture 01.05 Epistemology of measurement

epistemology

Epistemology is the study of knowledge and how it is acquired. Epistemology of measurement is the study of how knowledge is acquired through measurement. The model-based description of measurement, above, is closely associated with this study.

One of the most interesting aspects of this work is the way it both questions the validity of and offers an alternative to the intuitive realism about measurement that suggests that there is a real quantity independent of the measurement and that measurement error is the difference between that quantity and the estimate of the quantity.

01.05.1 Questioning realism

Here are four ways of questioning our intuitive realism.

01.05.1.1 The unknowable real

A first assault on realism is the following. The exact real value of a quantity cannot be known (at least for continuous scales). This is generally accepted. Therefore, the error between the measurement and the real value cannot be known. We are left to compare (inherently inaccurate) measurements with each other, using statistical techniques. Even if measurements are made with different techniques and averaged, we cannot know all these techniques don't share a common systematic error.

01.05.1.2 Circles are a thing

Another assault on realism can be made by emphasizing the fact that when measurements take place, some sort of idealizing model of the concrete situation is implicit. For instance, what would it mean to measure measure the temperature of an object without an idealized model of temperature making fluid volumes expand? Ok, instead, could we start from the theoretical model? What, then, would "temperature" mean? There is a circularity the confounds us at the bottom; however, this circularity is apparently *non-vicious* when considered from the historical point-of-view, which we will now consider.[8]

non-vicious circularity

[8]The conventionalists tried to take a "way out" of the circle by attempting to arbitrarily fix meaning with "coordinative definitions." The historical point-of-view questions the validity of this given the fact that definitions seem to require revision upon the development of sufficiently accurate measurements. That is, the definition of a quantity cannot remain

01.05.1.3 History teaching us a thing

Thomas Kuhn, among others, pointed out that the idea that the hypothesize→ measure → interpret → judge "scientific method" is rarely actual for significant scientific progress. Instead, we refine our measurements to the point that they expose a lack in our theory. The lack becomes the focus, and is measured with greater and greater precision. Finally, a new theory emerges that can explain the lack in the previous theory's explanatory power.

01.05.1.4 Bliss is ignorance

Finally, we call into question the very idea that a "real" value even makes sense without a certain amount of (useful) pretending. What does it mean to (say) measure the speed of light in a medium? We must, of course, assume (and try to control) certain aspects of the medium that we know are never *completely* the case, such as that it has uniform properties like density, temperature, and pressure. Moreover, we are assuming a completely static, repeatable measurement environment, while we all know perfectly well that one "cannot step twice into the same stream." (Tal, 2017)

These suggests a new understanding of the relationship between measurement and theory.

01.05.2 A new paradigm

These critiques of realism lead us to an emerging model-based paradigm in the epistemology of measurement. This view understands the limitations, circularity, and ideality of measurement, yet forges a new path forward. We might summarize this view as advocating something like the following attitude toward the epistemology of measurement (Tal, 2017):

accuracy The *accuracy* of a measurement has been defined as its nearness to the *unknowable* "real value." Now we define it to be its "agreement with values reasonably attributed to a quantity given available empirical data and background knowledge."

precision The *precision* of a measurement has been defined as the size of its clustering (e.g. standard deviation). Now we define it to be a type of *inaccuracy* from "uncontrolled variations" of indications.

static.

On a higher level, we develop models and measure and refine the models and measure some more. That is, *measurement refines theory* and *theory refines measurement*. Measurement processes inform theoretical models, through which measurement results are interpreted. The interpretation refines the model that inform the next generation of measurement processes. Etc.

Embrace the circle as it embraces you.

02

Signals

A *signal* "is a function that conveys information about the behavior or attributes of some phenomenon" (Priemer, 1991). **signal**

We will primarily be considering *electrical signals* (e.g. voltage or current varying in time) and *mechanical signals* (e.g. velocity or force varying in time). **electrical signal** **mechanical signal**

Lecture 02.01 Types of signals

There are many types of signals that will be encountered in engineering analysis. This lecture surveys a few of the most common. Many of our signals are functions of *time* t, and are presented as such here, but it is possible to have other independent variables for a signal.

periodic **aperiodic**

The primary distinction that we make among types of signals is between those that are *periodic* and those that are *aperiodic*. Periodic signals repeat. In maths, a signal $v(t)$ is periodic if for all t and for some $T \in \mathbb{R}$,

$$v(t+T) = v(t). \tag{02.1}$$

If a function satisfies this condition, it is periodic with period T. An aperiodic function is one that is not periodic.

02.01.1 Sinusoidal signals

The familiar sine wave is the most popular periodic signal.

Equation 02.2 sinusoidal signal

amplitude **angular frequency** **phase** **cyclic frequency**

We call A the *amplitude*, ω the *angular frequency* in rad/s, and ϕ the *phase* in rad. We define the *cyclic frequency* f in Hz to be

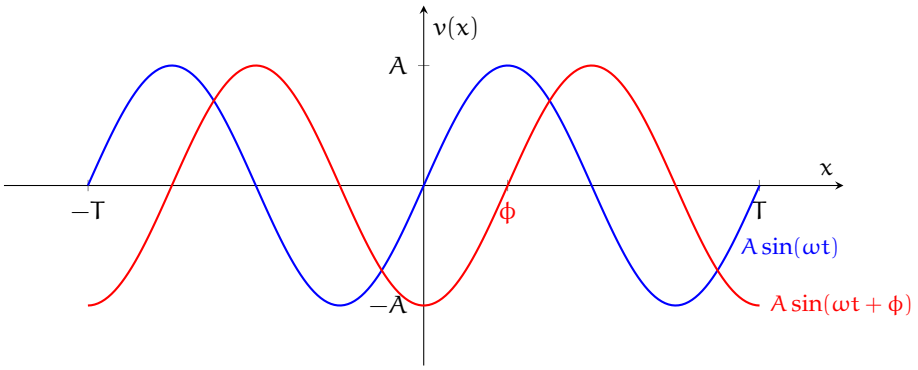
$$f = \frac{\omega}{2\pi} \tag{02.3}$$

period

and the *period* to be

$$T = \frac{1}{f}. \tag{02.4}$$

Below is a sinusoidal signal with and without a phase shift ϕ.



There are three common types of "amplitude" for sinusoids.

GOFA Good old-fashioned amplitude A is the amplitude we've described already.

P2P *Peak-to-peak amplitude* A_{pp} is twice the GOFA: $A_{pp} = 2A$. peak-to-peak amplitude

RMS *Root mean square* amplitude is conceptually like the "mean" amplitude but is defined over the time interval $[t_1, t_2]$ as root mean square amplitude

$$A_{rms} = \left(\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} v(\tau)^2 d\tau \right)^{1/2}. \tag{02.5}$$

For a sinusoid, for time intervals that are multiples of the period, this reduces to

$$A_{rms} = A/\sqrt{2}. \tag{02.6}$$

One must be careful to specify which is being used and be aware that in some instances it is assumed to be "conventional."

Finally, a sinusoid's mean value is called the *dc offset* and is defined as the integral over a period divided by a period: dc offset

$$\bar{v} = \frac{1}{T} \int_{-T/2}^{T/2} v(t) dt. \tag{02.7}$$

Many signals are approximately sinusoidal. For instance, ac electrical power is sinusoidal, as is the motion of a pendulum.

02.01.2 Decibels

Although this is not a "type" of signal, it is worth mentioning decibels alongside sinusoids. The *decibel* (dB) is a conventional logarithmic ratio of amplitudes. decibel

Equation 02.8 decibel

The "reference" amplitude A_0 is sometimes taken to be 1, but typically it is some conventional quantity like 1 V for a voltage signal. In certain cases like this, the abbreviation dB might be given some decoration, like dBV.

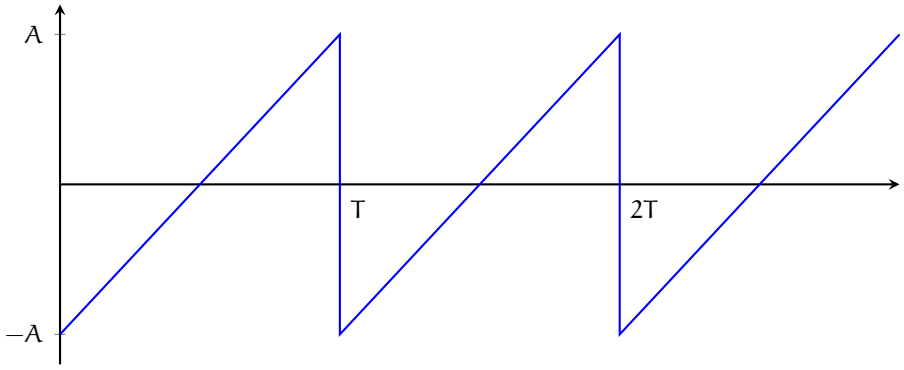
The decibel is sometimes expressed in terms of the ratio of the powers of two signals, but we aren't there yet.

02.01.3 Ramp and sawtooth

ramp
sawtooth

A *ramp* function is an aperiodic signal that increases linearly. A *sawtooth* function is a periodic signal that repeats a section of a ramp function.

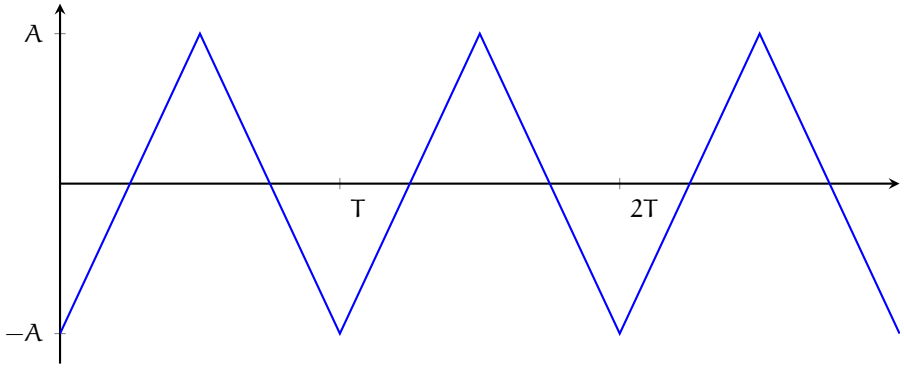
Below, a sawtooth signal with amplitude A and period T is plotted.



02.01.4 Triangle

A *triangle* function is a periodic signal with a section of a ramp followed by a negative section of the same ramp. **triangle**

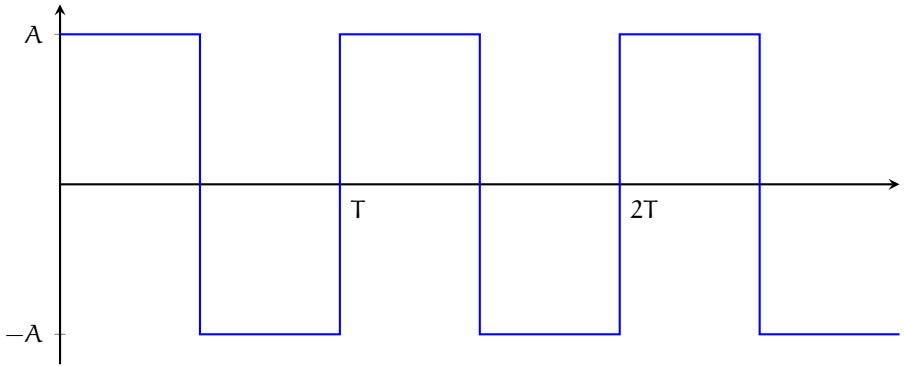
Below, a triangle signal with amplitude A and period T is plotted.



02.01.5 Square

A *square* signal is a periodic signal that switches between two constant values. **square**

Below, a square signal with amplitude A and period T is plotted.



Lecture 02.02 Fourier series

Fourier series are mathematical series that can represent a periodic signal as a sum of sinusoids at different amplitudes and frequencies. They are useful for solving for the response of a system to periodic inputs. However, they are probably most important *conceptually*: they are our gateway to thinking of signals in the *frequency domain*—that is, as functions of *frequency* (not time). To represent a function as a Fourier series is to *analyze* it as a sum of sinusoids at different frequencies ω_n and amplitudes and a_n. It's *frequency spectrum* is the functional representation of amplitudes a_n versus frequency ω_n.

frequency domain
Fourier analysis
frequency spectrum

Let's begin with the definition.

Definition 02.02.1: Fourier series: trigonometric form

The *Fourier analysis* of a periodic function $y(t)$ is, for $n \in \mathbb{N}_0$ and period T,

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} y(t)\cos(2\pi nt/T)\,dt \tag{02.9}$$

$$b_n = \frac{2}{T}\int_{-T/2}^{T/2} y(t)\sin(2\pi nt/T)\,dt. \tag{02.10}$$

The *Fourier synthesis* of a periodic function $y(t)$ with analysis components a_j and b_j corresponding to ω_j is

$$y(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n\cos(2\pi nt/T) + b_n\sin(2\pi nt/T). \tag{02.11}$$

Let's consider the complex form of the Fourier series, which is analogous to Definition 02.02.

Definition 02.02.2: Fourier series: complex form

The *Fourier analysis* of a periodic function $y(t)$ is, for $n \in \mathbb{N}_0$ and period T,

$$c_{\pm n} = \frac{1}{T}\int_{-T/2}^{T/2} y(t)e^{-j2\pi nt/T}\,dt. \tag{02.12}$$

The *Fourier synthesis* of a periodic function $y(t)$ with analysis components c_n corresponding to ω_n is

$$y(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi nt/T}. \tag{02.13}$$

We call the integer n a *harmonic* and the frequency associated with it

harmonic

$$\omega_n = 2\pi n/T \tag{02.14}$$

the *harmonic frequency*. There is a special name for the first harmonic ($n = 1$): the *fundamental frequency*. It is called this because all other frequency components are integer multiples of it.

harmonic frequency
fundamental frequency

It is also possible to convert between the two representations above.

Definition 02.02.3: Fourier series: converting between forms

The complex Fourier analysis of a periodic function $y(t)$ is, for $n \in \mathbb{N}_0$ and a_n and b_n as defined above,

$$c_{\pm n} = \frac{1}{2}\left(a_{|n|} \mp jb_{|n|}\right) \tag{02.15}$$

The sinusoidal Fourier analysis of a periodic function $y(t)$ is, for $n \in \mathbb{N}_0$ and c_n as defined above,

$$a_n = c_n + c_{-n} \text{ and} \tag{02.16}$$
$$b_n = j\left(c_n - c_{-n}\right). \tag{02.17}$$

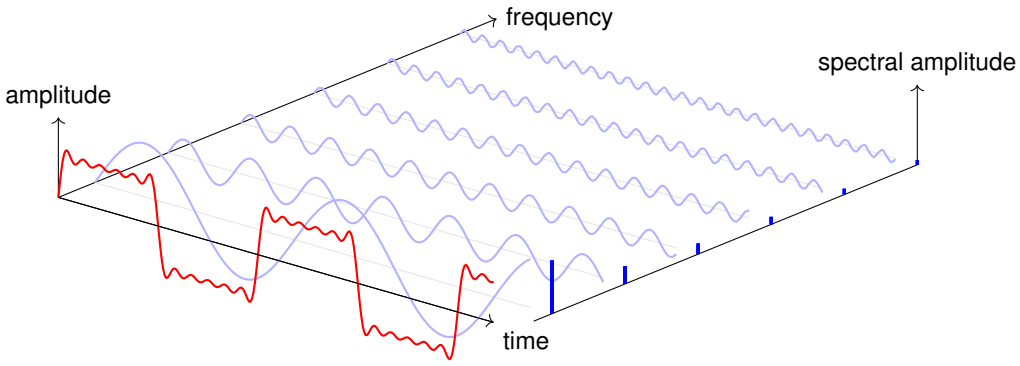
The *harmonic amplitude* is

harmonic amplitude

$$C_n = \sqrt{a_n^2 + b_n^2} \tag{02.18}$$
$$= 2\sqrt{c_n c_{-n}}. \tag{02.19}$$

line spectrum A *line spectrum* is a graph of the harmonic amplitudes as a function of the harmonic frequencies.

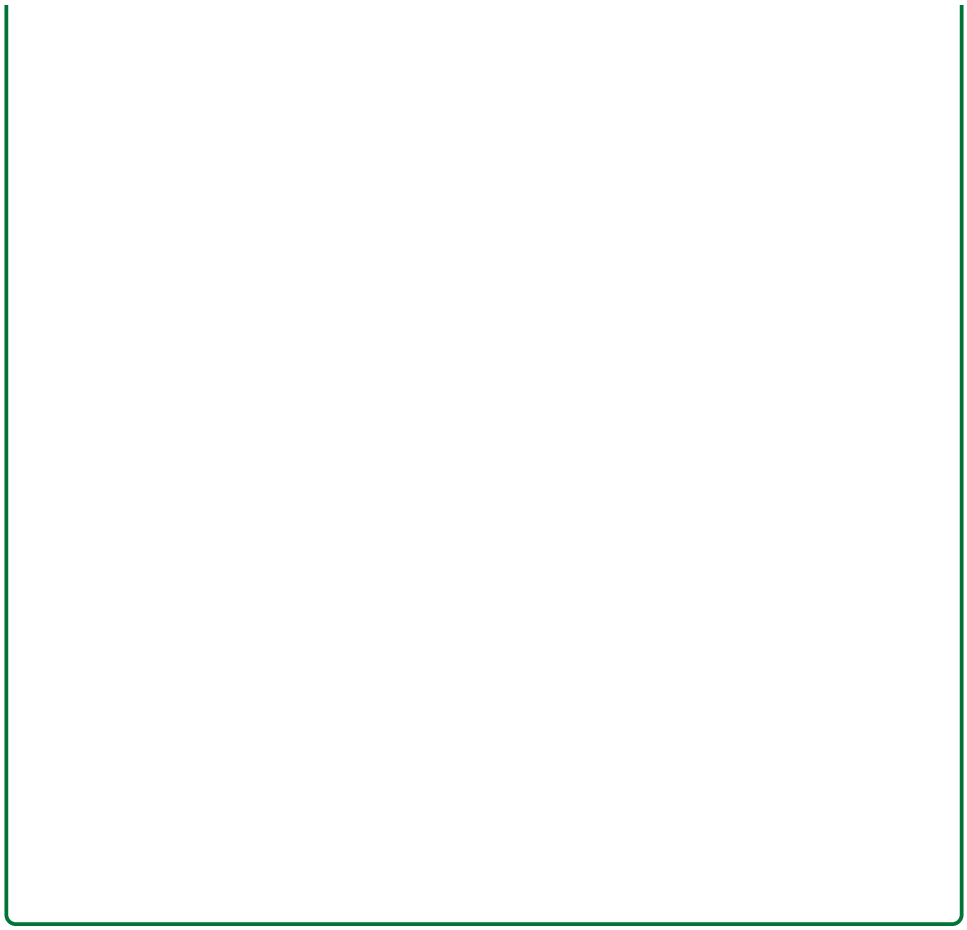
The following illustration demonstrates how sinusoidal components sum to represent a square wave. A line spectrum is also shown.



Let us compute the associated spectral components in the following example.

Example 02.02-1 Fourier series analysis: line spectrum

Compute the first five harmonic amplitudes that represent the line spectrum for a square wave in the figure above.



Lecture 02.03 Fourier transforms

The source for this lecture is in *SageMath* kernel *Jupyter* notebook. For more information, see jupyter.org and sagemath.org.

See ricopic.one/measurement/notebooks for the source code notebook. First, we import packages and all that. We use `matplotlib` for plotting and `numpy` for numerics.

Let's consider a periodic function `f` with period `T` (`T`). Each period, the function has a triangular pulse of width δ (`pulse_width`) and height $\delta/2$.

```
save_figures = False # true to save LaTeX figures
T = 35 # period
pulse_width = 2 # pulse width
f1(x) = pulse_width/2-abs(x) # first pulse
f2(x) = pulse_width/2-abs(x-T) # second pulse
omega_max = 12 # rad/s max frequency in line spectrum
n_max = round(omega_max*T/(2*pi)) # corresponding max harmonic
```

First, we plot the function `f` in the time domain. Using the *SageMath* `piecewise` function due to its Fourier Series methods (used momentarily), we define it and use `matplotlib` to plot it.

```
f = piecewise([[(-pulse_width/2,pulse_width/2),f1]]) # for FS series
fp = piecewise( # for plotting
    [
        [[-T/2,-pulse_width/2],0],
        [(-pulse_width/2,pulse_width/2),f1],
        [[pulse_width/2,T/2],0],
        [(T/2,T-pulse_width/2),0],
        [[T-pulse_width/2,T+pulse_width/2],f2],
        [(T+pulse_width/2,T+T/2),0],
        [[T+T/2,T+T/2],0]
    ]
)
N = 201 # number of points to plot
tpp = np.linspace(-T/2,3*T/2,N) # numeric array of time values
fpp = []
for i in range(0,N):
    fpp.append(fp(tpp[i])) # build array of function values
axes = plt.figure(1)
plt.plot(tpp,fpp,'b-',linewidth=2) # plot
plt.xlabel('time (s)')
```

```
plt.xlim([-T/2,3*T/2])
plt.xticks([pulse_width/2,T],['$\\frac{\delta}{2}$','$T='+str(T)+'$ s'])
plt.yticks([0,pulse_width/2],['0','$\delta/2$'])
if save_figures:
    tikz_save( # save for LaTeX's pgfplots
        'figures/fourier_series_to_transform_pulse'+
        str(T)+'.tex',
        figureheight='.5\linewidth',
        figurewidth='1\linewidth'
    )
plt.show() # display here
```

For $\delta = 2$ and $T \in [5, 15, 25]$, the left-hand column of Figure 02.1 shows two triangle pulses for each period T.

Consider the following argument. Just as a Fourier series is a frequency domain representation of a periodic signal, a Fourier transform is a frequency domain representation of an *aperiodic* signal (we will rigorously define it in a moment). The Fourier series components will have an analog, then, in the Fourier transform. Recall that they can be computed by integrating over a period of the signal. If we increase that period infinitely, the function is effectively aperiodic. The result (within a scaling factor) will be the Fourier transform analog of the Fourier series components.

Let us approach this understanding by actually computing the Fourier series components for increasing period T. *SageMath* has nice methods for its `piecewise` class, `fourier_series_cosine_coefficient(n,T/2)` and `fourier_series_sine_coefficient(n,T/2)`, that can compute the Fourier series cosine and sine components a_n and b_n for component n (`n`) and period T (`T`).

```
f_cos = [];
f_sin = [];
f_harmonic_amplitude = [];
omega = [];
for i in range(0,n_max):
    f_cos.append(f.fourier_series_cosine_coefficient(i,T/2))
    f_sin.append(f.fourier_series_sine_coefficient(i,T/2))
    f_harmonic_amplitude.append(
        T/pulse_width*sqrt(f_cos[i]**2+f_sin[i]**2)
    )
    omega.append(2*pi*i/T)
```

Furthermore, we have computed the *harmonic amplitude* (`f_harmonic_amplitude`):

$$C_n = \sqrt{a_n^2 + b_n^2} \tag{02.20}$$

which we have also scaled by a factor T/δ in order to plot it with a convenient scale.

```
axes = plt.figure(2)
markerline, stemlines, baseline = plt.stem(
    omega, f_harmonic_amplitude,
    linefmt='b-', markerfmt='bo', basefmt='r-'
)
plt.xlabel('frequency $\omega$ (rad/s)')
plt.xlim([0,omega_max])
plt.yticks([0,pulse_width/2],['0','$\delta/2$'])
if save_figures:
    tikz_save( # save for LaTeX
        'figures/fourier_series_to_transform_spectrum'+
        str(T)+'.tex',
        figureheight='.5\linewidth',
        figurewidth='1\linewidth'
    )
plt.show() # show here
```

{

} The line spectra are shown in the right-hand column of Figure 02.1. Note that with our chosen scaling, as T increases, the line spectra reveal a distinct waveform.

Let F be the continuous function of angular frequency ω

$$F(\omega) = \frac{\delta}{2} \cdot \frac{\sin^2(\omega\delta/4)}{(\omega\delta/4)^2}. \tag{02.21}$$

First, we plot it.

```
F(w) = pulse_width/2* \
    sin(w*pulse_width/(2*2))**2/ \
    (w*pulse_width/(2*2))**2
N = 201 # number of points to plot
wpp = np.linspace(0.0001,omega_max,N) # numeric array of time values
```

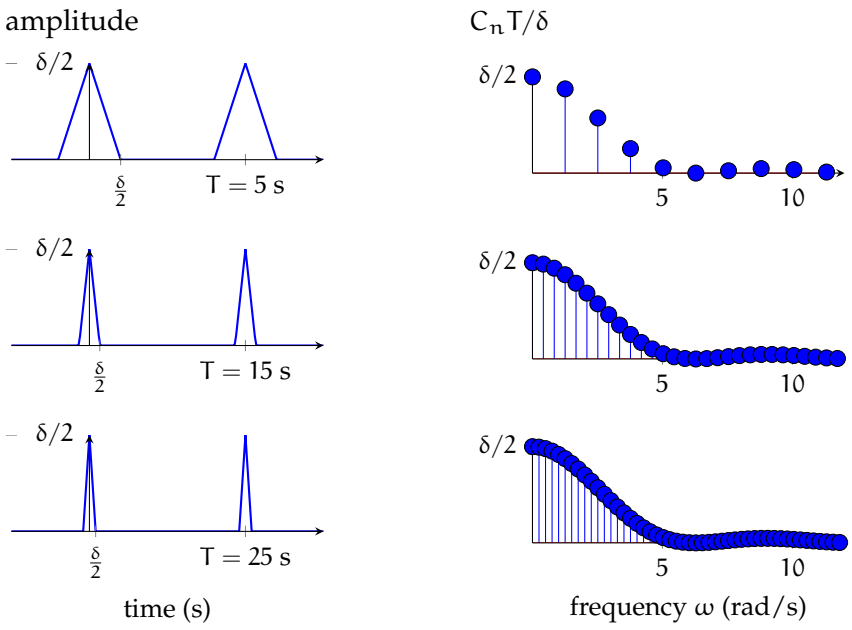


Figure 02.1: triangle pulse trains (left column) with longer periods, descending, and their corresponding line spectra (right column), scaled for convenient comparison.

```
Fpp = []
for i in range(0,N):
    Fpp.append(F(wpp[i])) # build array of function values
axes = plt.figure(3)
plt.plot(wpp,Fpp,'b-',linewidth=2) # plot
plt.xlim([0,omega_max])
plt.yticks([0,pulse_width/2],['0','$\delta/2$'])
plt.xlabel('frequency $\omega$ (rad/s)')
plt.ylabel('$F(\omega)$')
if save_figures:
    tikz_save( # save for LaTeX
        'figures/fourier_series_to_transform_transform.tex',
        figureheight='.5\linewidth',
        figurewidth='1\linewidth'
    )
plt.show()
```

Let's consider the plot in Figure 02.2 of `F`. It's obviously the function emerging in Figure 02.1 from increasing the period of our pulse train.

Now we are ready to define the Fourier transform and its inverse.

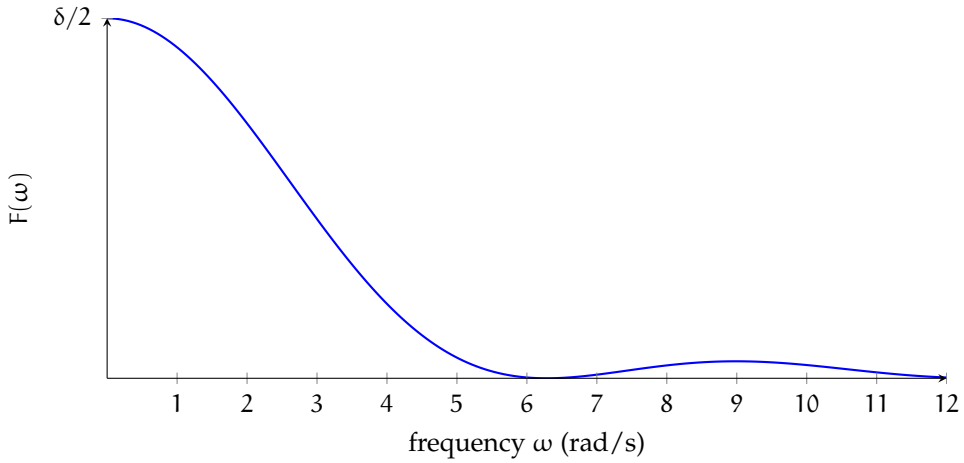


Figure 02.2: $F(\omega)$, our mysterious Fourier series amplitude analog.

Definition 02.03.1: Fourier transforms: trigonometric form

Fourier transform (analysis):

$$A(\omega) = \int_{-\infty}^{\infty} y(t)\cos(\omega t)dt \tag{02.22}$$

$$B(\omega) = \int_{-\infty}^{\infty} y(t)\sin(\omega t)dt. \tag{02.23}$$

Inverse Fourier transform (synthesis):

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} A(\omega)\cos(\omega t)d\omega + \frac{1}{2\pi}\int_{-\infty}^{\infty} B(\omega)\sin(\omega t)d\omega. \tag{02.24}$$

Definition 02.03.2: Fourier transforms: complex form

Fourier transform $\mathcal{F}$ (analysis):

$$\mathcal{F}(y(t)) = Y(\omega) = \int_{-\infty}^{\infty} y(t) e^{j\omega t} dt. \tag{02.25}$$

Inverse Fourier transform $\mathcal{F}^{-1}$ (synthesis):

$$\mathcal{F}^{-1}(Y(\omega)) = y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega) e^{-j\omega t} d\omega. \tag{02.26}$$

So now we have defined the Fourier transform. There are many applications, including solving differential equations and *frequency domain* representations—called *spectra*—of *time domain* functions.

There is a striking similarity between the Fourier transform and the Laplace transform, with which you are already acquainted. In fact, the Fourier transform is a special case of a Laplace transform with Laplace transform variable $s = j\omega$ instead of having some real component. Both transforms convert differential equations to algebraic equations, which can be solved and inversely transformed to find time-domain solutions. The Laplace transform is especially important to use when an input function to a differential equation is not absolutely integrable and the Fourier transform is undefined (for example a step or ramp function). However, the Laplace transform is also preferred for *initial value problems* due to its convenient way of handling them. The two transforms are equally useful for solving steady state problems. Although the Laplace transform has many advantages, for spectral considerations, the Fourier transform is the only game in town.

A table of Fourier transforms and their properties can be found on the course website in the "Resources" section.

Lecture 02.04 Sampling

While most quantities are continuous, making them naturally represented by continuous functions, in order to represent a signal in a computer, it must be given a *discrete* representation.[1] Constructing a discrete representation of a signal is called *sampling* it: a numerical value is assigned at discrete values of its domain. Since the quintessential signal has time as its domain, we will henceforth speak as if it is the only type.

discrete
sampling

Consider a function $f : \mathbb{R} \to \mathbb{R}$ of time t over the interval $[t_1, t_2]$ sampled at a constant interval T to form a sequence of reals $(f_n)_{n\in\mathbb{N}_0}$ of length $N = 1 + (t_2 - t_1)/T$ called the *sampled sequence*, defined as follows.

sampled sequence

Equation 02.27 sampled sequence values

We represent the sampling process as the multiplication of the continuous function f by the *Dirac comb function* $s : \mathbb{R} \to \mathbb{R}$ defined as

Dirac comb function

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT), \tag{02.28}$$

where δ is the Dirac delta function, and illustrated in Figure 02.3. This sampling representation yields the *sampled function* $f^\star : \mathbb{R} \to \mathbb{R}$ defined as follows.

sampled function

Equation 02.29 sampled function

This representation requires some interptretation. Recall that δ is zero everywhere in its domain except at $t = 0$, when it is *undefined*, but has an integral over the "pulse" of unity. Further recall that *scaling* δ by some factor λ yields the same functional value, but it is understood in the *distribution* sense to be multiplying the integral over the pulse by λ. We call this the pulse *strength*. So the sample sequence values f_n are computed from $f^\star$ by

δ strength

[1] Good references for the sampling and related topics are Rowell (2008) and Gene F. Franklin (1998).

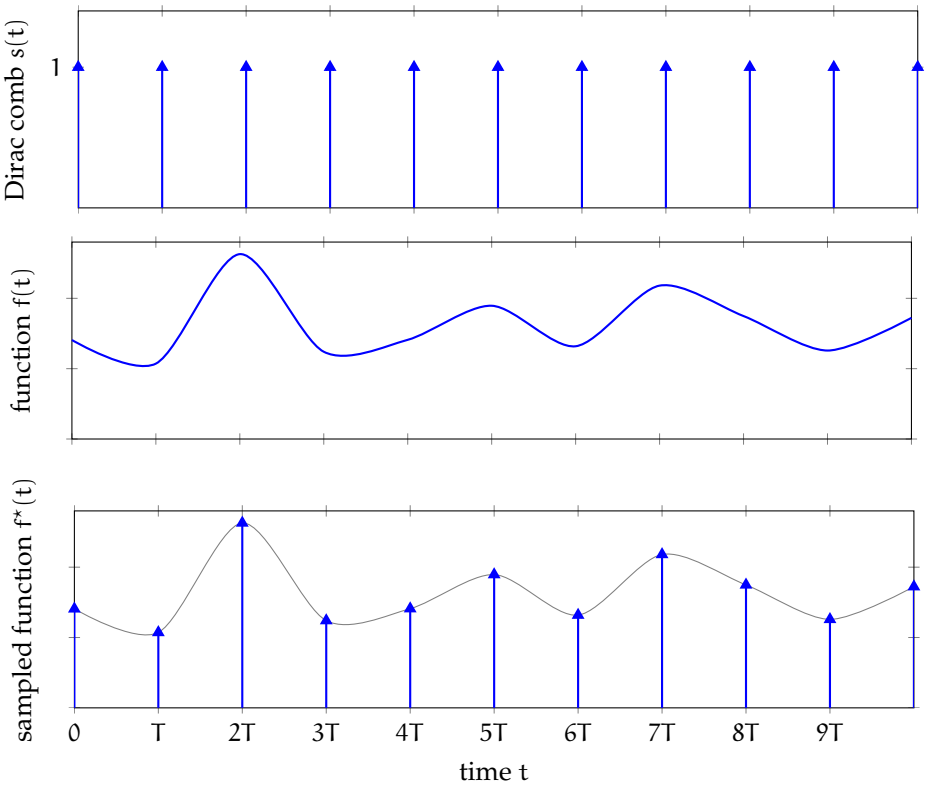


Figure 02.3: (top) the Dirac comb function s of Equation 02.28, (middle) a function f to be sampled, and (bottom) a sampled function $f^\star$. The height of each pulse is f_n and represents the strength of the pulse.

the following equation.

Equation 02.30 sampled sequence values from $f^\star$

In other words, we integrate over the n^{th} pulse to get the sample f_n.

02.04.1 Spectrum of a sampled function

The Dirac comb function s is periodic and therefore has a Fourier series representation. From Definition 02.02, we can compute the components

sifting property

The Dirac delta function has a nice property called the *sifting property* that states that for some function g, the integral $\int_I \delta(t-\tau)g(t)dt = g(\tau)$ over the interval I and 0 otherwise. This yields

We happend to choose the easiest interval over which to integrate, but the same holds for any other period, which means $c_{\pm n} = 1/T$ for all time. The Fourier series synthesis, then, is

The spectrum $F^\star(\omega)$ of a sampled function $f^\star$ can be found directly from the definition of the Fourier transform (**??**)

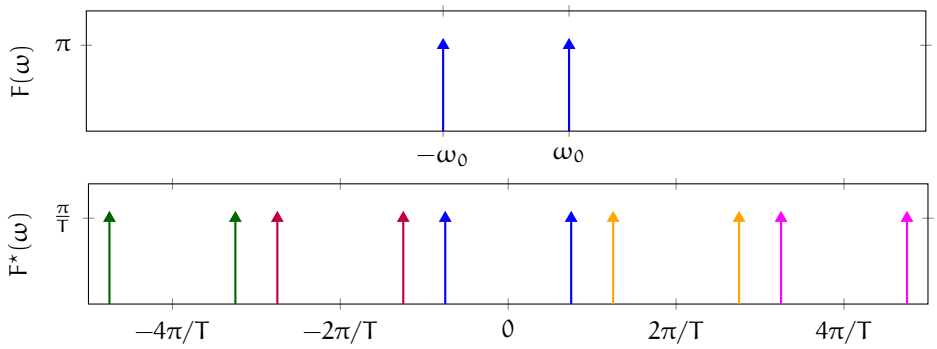


Figure 02.4: (top) the Fourier transform F of function $f(t) = \cos \omega_0 t$ and (bottom) the Fourier transform $F^\star$ of sampled function $f^\star$. T is the *sampling period*.

This result means the Fourier transform $F^\star$ of the sampled function $f^\star$ is a periodic repetition (with frequency-domain period $2\pi/T$) of the Fourier transform F of the continuous signal f, scaled by $1/T$.

For instance, although it is periodic and so has a trivial Fourier series, the cosine function $f(t) = \cos \omega_0 t$ also has Fourier transform

$$F(\omega) = \pi\delta(\omega - \omega_0) + \pi\delta(\omega + \omega_0). \tag{02.31}$$

This is illustrated in the top spectrum of Figure 02.4. Our expression above for $F^\star$ in terms of F allow us to construct the spectrum for $F^\star$, shown in the bottom spectrum of Figure 02.4.

Lecture 02.05 Nyquist sampling theorem, aliasing, and reconstruction

When we sample a signal, we lose information. However, we can reconstruct an approximation of the original signal if we make some assumptions about it.

Theorem 02.05.1: Nyquist sampling theorem

If a signal has maximum frequency component f_{max}, if it is sampled at a frequency greater than

$$f_N = 2f_{max} \tag{02.32}$$

it can be *unambiguously reconstructed* and otherwise not. We call f_N the *Nyquist frequency* of a signal.

aliasing

The "not" case is characterized by *aliasing*: when a signal appears to have frequency components that are in fact higher than they appear—an artifact of sampling.

This is the first time in our journey that Fourier analysis and frequency domain concepts will light the path. There are several ways to analyze aliasing, but perhaps considering the spectrum of a sampled sinusoid is the most accessible.

Let us consider, as in Figure 02.4, the sampling of the function $y(t) = \cos \omega_0 t$, where $\omega_0 \in \mathbb{R}$. The signal has the Fourier transform

which is just a pulse with strength π at $\omega = -\omega_0$ and one at $\omega = +\omega_0$. The Fourier transform $Y^\star$ of the sampled signal $y^\star$ simply "copies" this pair of pulses to be mirrored about integer multiples of the sampling angular frequency ω_s.

According to **??**, $\omega_s > \omega_N = 2\pi f_N$ is required to avoid aliasing and allow us to uniquely reconstruct the original signal. Let's consider the situation in which the Nyquist sampling frequency requirements are met, as shown in the upper plot of Figure 02.5. Now, consider the case for which the Nyquist sampling requirements are not met and there is aliasing, as shown in the lower plot of Figure 02.5. These two spectra are

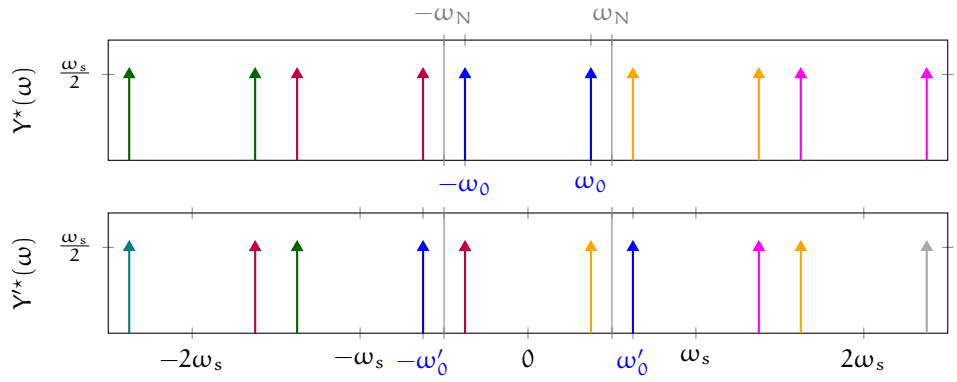


Figure 02.5: (top) the Fourier transform $Y^\star$ of sampled function $y^\star(t) = \cos\omega_0 t$ and (bottom) the Fourier transform $Y'^\star$ of sampled function $y'^\star = \cos\omega'_0 t$, where the Nyquist sampling frequency ω_N is between them: $\omega_0 < \omega_N < \omega'_0$. That is, $y^\star$ is sufficiently sampled but $y'^\star$ is *not*, and is therefore aliased. The two spectra are *indistinguishable* and therefore one must be confident *a priori* that the sampling frequency is greater than ω_N.

indistinguishable and therefore if we simply assume that the component at ω_0 is "real" (i.e. not aliased), we might be mistaken.

It is important to note that *once the signal is sampled, it's too late to do anything about it*. There are two ways to mitigate aliasing:

1. sample at a high enough rate to capture all frequency components of the signal and
2. *low-pass* or *anti-aliasing filter* the signal *before* it is sampled.

anti-aliasing filter

The former always begs the question—it is never known if the sample rate is high enough. The latter is always advisable, although it only minimizes the effects of aliasing.

Assuming there is no aliasing (which is never more than approximately true), a continuous signal y can be fully reconstructed from its sample sequence (y_n) for N samples and sample period T by the *Whittaker–Shannon interpolation formula* Rowell (2008)

Whittaker–Shannon interpolation formula

$$y(t) = \sum_{n=0}^{N-1} y_n \frac{\sin(\pi(t - nT)/T)}{\pi(t - nT)/T}. \tag{02.33}$$

Lecture 02.06 Discrete Fourier transforms

The source for this lecture is in *SageMath* kernel *Jupyter* notebook. For more information, see jupyter.org and sagemath.org.

See ricopic.one/measurement/notebooks for the source code notebook. First, we import packages and all that. We use `matplotlib` for plotting, `numpy` for numerics, and `scipy` for discrete (fast) Fourier transforms.

Modern measurement systems primarily construct spectra by sampling an analog electronic signal $y(t)$ to yield the sample sequence (y_n) and perform a *discrete Fourier transform*.

Definition 02.06.1: discrete Fourier transform

The *discrete Fourier transform* (DFT) of a sample sequence (y_n) of length N is (Y_m), where $m \in [0, 1, \cdots, N-1]$ and

$$Y_m = \sum_{n=0}^{N-1} y_n e^{-j2\pi mn/N}.$$

The *inverse discrete Fourier transform* (IDFT) reconstructs the original sequence for $n \in [0, 1, \cdots, N-1]$ and

$$y_n = \frac{1}{N} \sum_{n=0}^{N-1} Y_m e^{j2\pi mn/N}.$$

The DFT (Y_m) has a frequency interval equal to the sampling frequency ω_s/N and the IDFT (y_n) has time interval equal to the sampling time T. The first $N/2+1$ DFT (Y_m) values correspond to frequencies

and the remaining $N/2-1$ correspond to frequencies

In practice, the definitions of the DFT and IDFT are not the most efficent methods of computation. A clever algorithm called the *fast Fourier transform* (FFT) computes the DFT much more efficiently. Although it is a good exercise to roll our own FFT, in this lecture we will use `scipy`'s built-in FFT algorithm, loaded with the following command.

```
from scipy import fft
```

Now, given a time series array `y` representing (y_i), the DFT (using the FFT algorithm) can be computed with the following command.

```
fft(y)
```

In the following example, we will apply this method of computing the DFT.

02.06.0.1 A DFT/FFT example

We would like to compute the DFT of a sample sequence (y_n) generated by sampling a spaced-out sawtooth. Let's first generate the sample sequence and plot it.

We define the sampling rate `fs`, which defines the sampling interval `Ts`. Furthermore, we define the frequency of the spaced sawtooth signal `f_signal`.

```
fs = 200 # sampling rate
Ts = 1.0/fs # sampling interval
f_signal = 10 # frequency of the signal
```

We want an interval of ramp followed by an interval of "space" (zeros). The following method of generating the sampled signal `y` helps us avoid *leakage*, which we'll describe after the example.

```
arr_zeros = np.zeros(fs/f_signal/2) # half signal period worth of zeros
arr_ramp = np.arange(fs/f_signal/2) # half signal period worth of ramp
y = [] # initialize time sequence
j = 0
for i in range(fs):
    if i % (fs/f_signal/2) == 0:
        # if we are at the start of a signal period
        if j % 2 == 0:
            # every other signal period
            y = np.append(y,arr_zeros)
        else:
            y = np.append(y,arr_ramp)
        j += 1
```

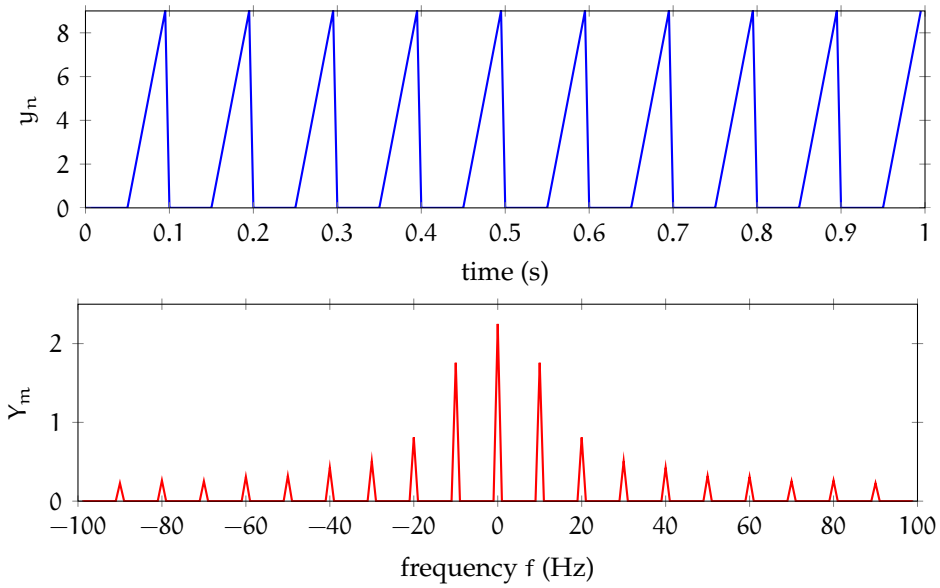


Figure 02.6: (top) a sampled sequence (y_n) plotted through time and (bottom) its discrete Fourier transform sequence (Y_m) plotted through frequency.

From this sequence, we can compute the following parameters.

```
N = len(y) # number of samples
t_a = np.arange(0,N*Ts,Ts) # time array
time_total = N*Ts # total time in series
```

Plotting this with `matplotlib` is fairly straightforward. The result is shown in the top plot of Figure 02.6.

```
plt.figure()
plt.plot(t_a,y,'b-',linewidth=2)
plt.xlabel('time (s)')
plt.ylabel('$y_n$');
```

Display the plot with the following command.

```
plt.show()
```

Now we have a nice time sequence on which we can perform our DFT. It's easy enough to compute the FFT.

```
Y = fft(y)/N # FFT with proper normalization
```

Recall that the latter values correspond to negative frequencies. In order to plot it, we want to rearrange our `Y` array such that the elements corresponding to negative frequencies are first. It's a bit annoying, but *c'est la vie*.

```
Y_positive_zero = Y[range(N/2)]
Y_negative = np.flip(
    np.delete(
        Y_positive_zero,
        0
    ),
    0
)
Y_total = np.append(Y_negative,Y_positive_zero)
```

Now all we need is a corresponding frequency array.

```
freq_total = np.arange(-N/2+1,N/2)*fs/N
```

Now, just to plot.

```
plt.figure()
plt.plot(freq_total, abs(Y_total),'r-',linewidth=2)
plt.xlabel('frequency $f$ (Hz)')
plt.ylabel('$Y_m$');
```

And now display the plot of the spectrum, shown on the bottom of Figure 02.6.

```
plt.show();
```

02.06.0.2 Leakage

The DFT assumes the sequence (y_n) is periodic with period N. An implication of this is that if any periodic components have period $N_{\text{short}} < N$, unless N is divisible by N_{short}, spurious components will appear in (Y_n).

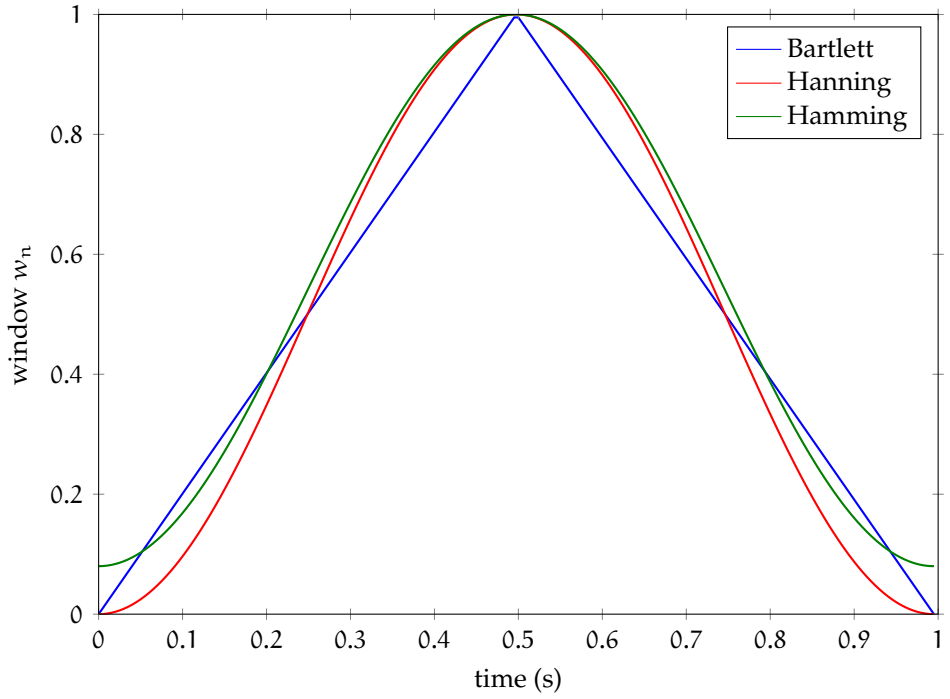


Figure 02.7: three sample window functions.

Avoiding leakage is difficult, in practice. Instead, typically we use a *window function* to mitigate its effects. Effectively, windowing functions—such as the *Bartlett*, *Hanning*, and *Hamming windows*—multiply (y_n) by a function that tapers to zero near the edges of the sample sequence.

Numpy has several window functions such as `bartlett()`, `hanning()`, and `hamming()`. For usage information on a function, the following `?` idiom is useful.

```
np.hanning?
```

Let's plot the windows to get a feel for them.

```
bartlett_window = np.bartlett(N)
hanning_window = np.hanning(N)
hamming_window = np.hamming(N)
```

```
plt.figure()
plt.plot(t_a,bartlett_window,'b-',label='Bartlett',linewidth=2)
plt.plot(t_a,hanning_window,'r-',label='Hanning',linewidth=2)
plt.plot(t_a,hamming_window,'g-',label='Hamming',linewidth=2)
plt.xlabel('time (s)')
plt.ylabel('window $w_n$')
plt.legend();
```

Show the figure Figure 02.7.

```
plt.show()
```

02.07 Problems for Chapter 02

02.07.1 Encoding and decoding with DFTs

The source for this exercise lecture is in a *Python 2* kernel *Jupyter* notebook. For more information, see python.org and jupyter.org.

See ricopic.one/measurement/notebooks for the source code notebook.

This exercise encodes a "secret word" into a sampled waveform for decoding via a *discrete Fourier transform* (DFT). The nominal goal of the exercise is to decode the secret word. Along the way, plotting and interpreting the DFT will be important.

First, load relevant packages.

We define two functions: `letter_to_number` to convert a letter into an `int`eger index of the alphabet (a becomes 1, b becomes 2, etc.) and `string_to_number_list` to convert a string to a `list` of `int`s, as shown in the example at the end.

```
def letter_to_number(letter):
    return ord(letter) - 96

def string_to_number_list(string):
    out = [] # list
    for i in range(0,len(string)):
        out.append(letter_to_number(string[i]))
    return out # list

print '"aces" = '+str(string_to_number_list('aces'))
```

```
"aces" = [1, 3, 5, 19]
```

Now, we encode a code string `code` into a signal by beginning with "white noise," which is *broadband* (appears throughout the spectrum) and adding to it `sin` functions with amplitudes corresponding to the letter assignments of the code and harmonic corresponding to the position of the letter in the string. For instance, the string `'bad'` would be represented by noise plus the signal

$$2\sin 2\pi t + 1\sin 4\pi t + 4\sin 6\pi t. \tag{02.34}$$

```
N = 2000
Tm = 30
```

```
T = float(Tm)/float(N)
fs = 1/T
x = np.linspace(0, Tm, N)
noise = 4*np.random.normal(0, 1, N)
code = 'abracadabra'
code_number_array = np.array(string_to_number_list(code)) # list into array
y = np.array(noise)
for i in range(0,len(code)):
    y = y + code_number_array[i]*np.array(np.sin(2.*np.pi*(i+1.)*x))
```

Now, we plot.

```
plt.figure()
plt.plot(x,y)
plt.xlim([0,Tm/4])
plt.xlabel('time (s)')
plt.ylabel('$y_n$')
plt.show()
```

Finally, we can save our data to a file `secrets` to distribute our message. We save it in two formats: 1. `secrets.npy` the `numpy` format that's not all that compressed and 2. `secrets.mat` the *MATLAB* format that's impressively compressed.

```
np.save('secrets',y)
scipy.io.savemat('secrets.mat',mdict={'y':y.astype('float')})
```

In order to load the `.npy` file into *Python*, we can use the following command.

```
secret_array = np.load('secrets.npy')
```

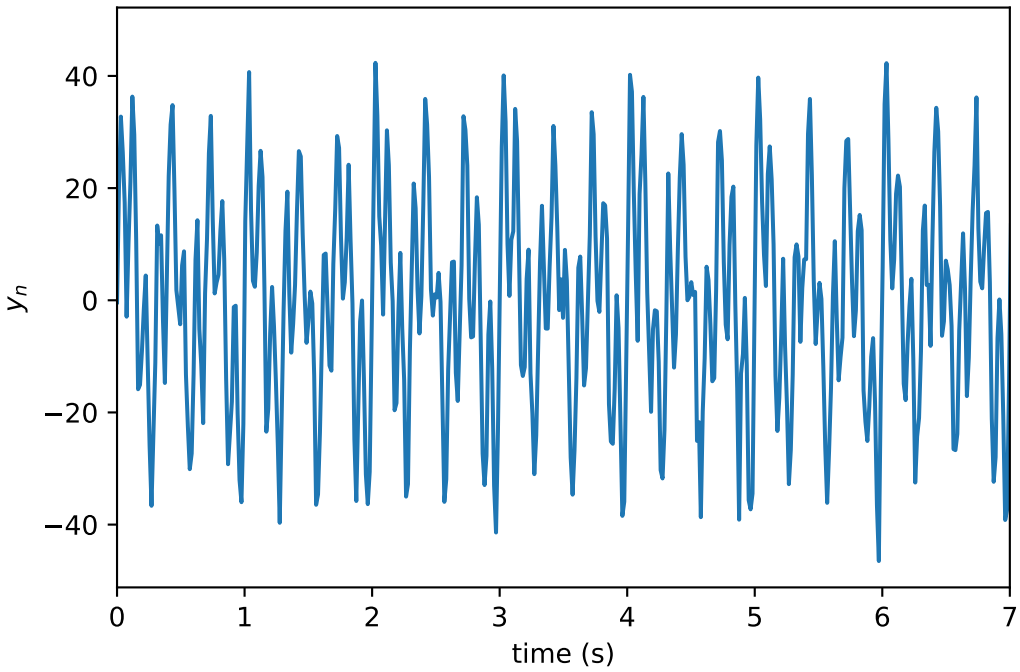


Figure 02.8: png

03

Measurement systems as dynamic systems

Measurement systems *respond* to the quantities measured. This response is crucial to our interpretation of a measurement reading. For instance, a thermometer does not give instantaneous measurement; rather, we wait for the response to "settle" before taking a reading. However, time delay is only one way in which the measurement system response affects the reading. More complicated effects require further analysis.

response

System dynamics is a field devoted to such system analysis by representing systems with *mathematical models*. Specifically, we will consider *lumped-parameter models* of measurement systems. These models give insight into the measurement system's behavior and allow us to properly design measurement systems and interpret their readings. Modeling does not replace calibration, but complements it.

system dynamics

mathematical model
lumped-parameter modeling

Lecture 03.01 Dynamic system representations

graphical representations
mathematical representations

Dynamic systems—measurement and otherwise—have several representations, mostly *graphical* or *mathematical*. For instance, schematics, linear graphs, and block diagrams are graphical representations. The foci of this lecture are three mathematical representations and a graphical representation:

1. the input-output, linear ordinary differential equation,
2. the transfer function,
3. the frequency response function, and
4. the block diagram.

Refer to Figure 03.1 for an illustration of the relations among system representations.

03.01.1 Input-output ordinary differential equations

Consider a dynamic system described by the *input-output differential equation*—with independent variable $y(t)$ (called the *output*), dependent variable time t, *input* $u(t)$, constant coefficients a_i, b_j, order n, and $m \leqslant n$ for $n \in \mathbb{N}_0$—as:

$$\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_1\frac{dy}{dt} + a_0 y = b_m\frac{d^m u}{dt^m} + b_{m-1}\frac{d^{m-1}u}{dt^{m-1}} + \cdots + b_1\frac{du}{dt} + b_0 u. \tag{03.1}$$

This might describe, for instance, an output voltage reading from an input force on a piezo-electric transducer. Mechanical, electronic, thermal, fluidic, and many other types of dynamic systems can be described with Equation 03.2. It is important to note that this describes the relationship between a *single-input* and a *single-output (SISO)*. However, a great many measurement systems behave approximately like linear SISO systems (at least in some operating regimes).

single-input, single output (SISO

03.01.2 Transfer functions

Consider a dynamic system described by the *input-output differential equation*—with variable y representing the *output*, dependent variable time t, variable u representing the *input*, constant coefficients a_i, b_j, order n, and

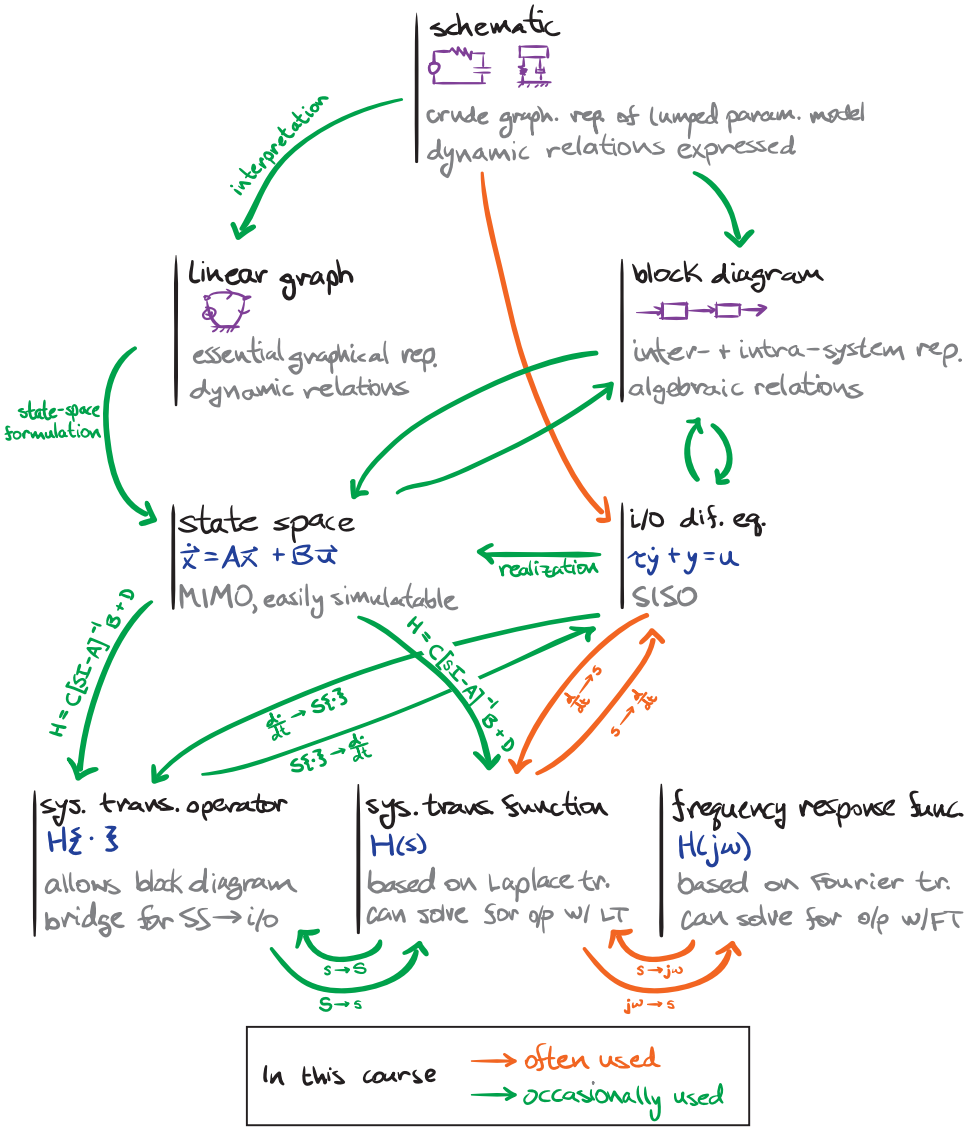


Figure 03.1: relations among system representations.

$m \leqslant n$ for $n \in \mathbb{N}_0$—as:

$$\begin{aligned} \frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_1\frac{dy}{dt} + a_0 y = \\ b_m\frac{d^m u}{dt^m} + b_{m-1}\frac{d^{m-1}u}{dt^{m-1}} + \cdots + b_1\frac{du}{dt} + b_0 u. \end{aligned} \tag{03.2}$$

Laplace transform

The *Laplace transform* $\mathcal{L}$ of Equation 03.2 yields something interesting (assuming zero initial conditions):

$$\begin{aligned} \mathcal{L}\left(\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_1\frac{dy}{dt} + a_0 y\right) = \\ \mathcal{L}\left(b_m\frac{d^m u}{dt^m} + b_{m-1}\frac{d^{m-1}u}{dt^{m-1}} + \cdots + b_1\frac{du}{dt} + b_0 u\right) \quad \Rightarrow \\ \mathcal{L}\left(\frac{d^n y}{dt^n}\right) + a_{n-1}\mathcal{L}\left(\frac{d^{n-1}y}{dt^{n-1}}\right) + \cdots + a_1\mathcal{L}\left(\frac{dy}{dt}\right) + a_0\mathcal{L}(y) = \\ b_m\mathcal{L}\left(\frac{d^m u}{dt^m}\right) + b_{m-1}\mathcal{L}\left(\frac{d^{m-1}u}{dt^{m-1}}\right) + \cdots + b_1\mathcal{L}\left(\frac{du}{dt}\right) + b_0\mathcal{L}(u) \quad \Rightarrow \\ s^n Y + a_{n-1}s^{n-1}Y + \cdots + a_1 sY + a_0 Y = \\ b_m s^m U + b_{m-1}s^{m-1}U + \cdots + b_1 sU + b_0 U. \end{aligned}$$

Solving for Y,

forced response

The inverse Laplace transform $\mathcal{L}^{-1}$ of Y is the *forced response*. However, this is not our primary concern; rather, we are interested to solve for the transfer function H as the ratio of the output transform Y to the input transform U, i.e.

$$H(s) \equiv \frac{Y(s)}{U(s)} \tag{03.3}$$

$$= \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0}. \tag{03.4}$$

block diagram

This is the foundation of another graphical technique called *block diagrams*.

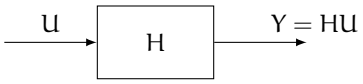


Figure 03.2: a block diagram for transfer function $H(s)$ from input $U(s)$ to output $Y(s)$.

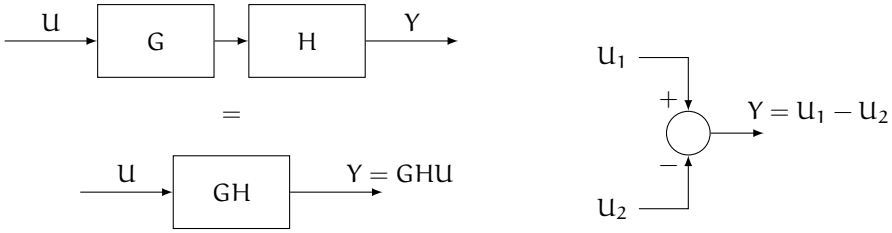


Figure 03.3: block diagrams showing (left) concatenation (i.e. multiplication) of G and H and (right) summation of inputs U_1 and U_2 (actually, subtraction).

03.01.3 Block diagrams

Block diagrams (also called *operational block diagrams*) represent inter- and intra-system relationships. A *block*, shown in Figure 03.2, represents a dynamic system with input $U(s)$ and output $Y(s)$ and (typically) transfer function $H(s)$. **block**

Each block can be considered, in the Laplace domain, to *multiply* its input by the transfer function, as illustrated in Figure 03.3 (left). Also shown in Figure 03.3 (right) is the *summation block* or *junction* that adds (or subtracts) its inputs. **multiplication** **summation block**

Block diagrams differ fundamentally from circuit diagrams and linear graphs in that do not "load" each other. That is, concatenating two systems does not affect the operation of the first. However, since we can write an input-output differential equation for a circuit and we can write a transfer function for a differential equation—block diagrams can represent a circuit or many other dynamic systems. We must be careful to make sure that subsequent connections do not load the output of another block. That is, *blocks have infinite input impedance*.

03.01.4 Frequency response functions

Let a system have input u and output y. We define its *frequency response function* $H(j\omega)$ to be the ratio of the Fourier transform its output to the Fourier transform of its input: **frequency response function**

Now, we have used a strange notation for the Fourier transform of a function, including the $j = \sqrt{-1}$ in the argument. In the context of frequency response, this is both conventional and gross. We will see why it's (barely) useful in a moment.

Typically, a system's dynamical model will first be developed as a differential equation—for instance, our input-output differential equation. We have already learned how to find a system's transfer function $H(s)$—which is the *Laplace* transform of the ratio of input to output. A shortcut to deriving the frequency response function $H(j\omega)$ from the transfer function $H(s)$ is to note that the Fourier transform of a function is equal to the Laplace transform of that function with the substitution $s \to j\omega$. For the frequency response function,

$$H(j\omega) = H(s)|_{s \to j\omega}. \tag{03.7}$$

gauche af

Note that we've used the same symbol H for both the Fourier and Laplace transforms, which is *gauche af*, but which is somewhat mitigated by our other violation of good taste: the argument s goes with the Laplace transform and the argument $j\omega$ goes with the Fourier transform.

meaning of $H(j\omega)$

So what does the frequency response function *mean*? It describes, in the frequency domain, how a system's input relates to its output. In Lecture 03.09, we will consider this interpretation in greater detail. Finally, note that the steady-state solution can be had by taking the inverse Fourier transform:

$$\begin{aligned} y(t) &= \mathcal{F}^{-1}\left(Y(j\omega)\right) \\ &= \mathcal{F}^{-1}\left(H(j\omega)U(j\omega)\right). \end{aligned} \tag{03.8}$$

Due to its greater generality (i.e. broader applicability), we typically prefer the inverse Laplace transform (and transfer function) for this task.

Example 03.01-1 converting representations

For the input-output differential equation

$$2\frac{d^2y}{dt} + 4\frac{dy}{dt} + 6y = 10\frac{du}{dt} + 2u.$$

find the transfer function $H(s)$ and the frequency response function $H(j\omega)$.

Lecture 03.02 Zeroth-order measurement systems

In special cases, we can consider measurement systems to be modeled by *zeroth*-order "differential" equations—that is, by algebraic systems.

If both input u *and* output y are static, the time-derivatives of Equation 03.2 are zero. This leaves the input-output equation, with $K \in \mathbb{R}$ in the form

$$y = Ku. \tag{03.9}$$

static sensitivity

We call K the *static sensitivity*. Technically, the input needn't be static, but one would question the efficacy of a measurement system that has a dynamic input and a static output. In fact, we have assumed neither input nor output changes. This can be approximately true if we calibrate and measure only in steady-state.

Another way to have Equation 03.9 as our model is if we assume there are no energy storage or dissipative elements. This type of measurement system does not exist, but in certain situations it can be considered approximately valid.

Example 03.02-1 zeroth or not?

Explain conditions under which the following systems can be considered zeroth-order.

1. A bulb thermometer.
2. A mass balance.
3. A speedometer.

Lecture 03.03 First-order measurement systems

First order measurement systems have input-output differential equations of the form

$$\tau \frac{dy}{dt} + y = b_1 \frac{du}{dt} + b_0 u \tag{03.10}$$

with $\tau \in \mathbb{R}$ called the *time constant* of the system. Measurement systems with a single energy storage element—such as those with electrical or thermal capacitance—can be modeled with first-order systems.

time constant

03.03.1 Step response

Commonly, a scaling of the *unit step function* $u_s(t)$, which is 0 for $t < 0$ and 1 for $t \geqslant 0$, can be considered the input to this and other measurement systems (e.g. whenever the input is changed suddenly, u_s is a good approximation). If we consider the common situation that $b_1 = 0$ and $u(t) = K u_s(t)$ for some $K \in \mathbb{R}$, the solution to Equation 03.10 is

unit step function

If we assume the steady-state solution is the proper measurement value, the transient response is *error*. Considering it never reaches zero in finite

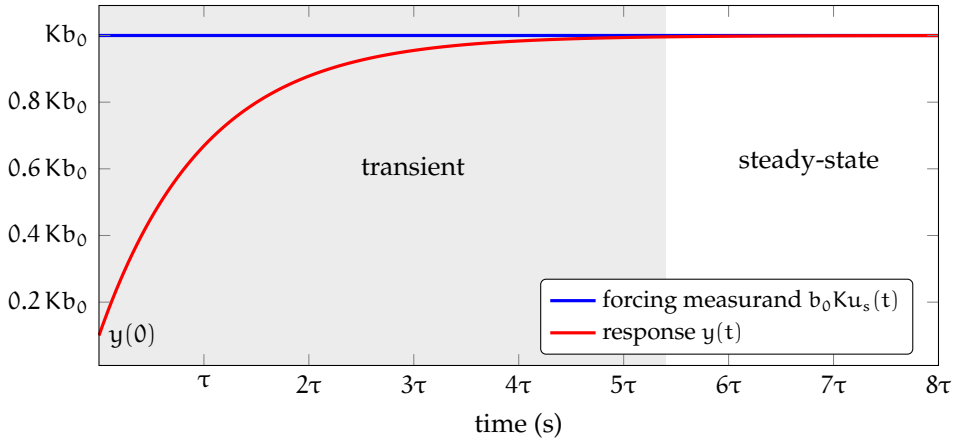


Figure 03.4: (step) response $y(t)$ of a first order system with input $u(t) = K u_s(t)$ and $b_1 = 0$.

time, this is a bummer ☹. However, it does decay exponentially, so in 5τ, the transient response is less than 1 % of difference between $y(0)$ and steady-state. A plot of the step response is shown in Figure 03.4.

03.03.2 Sinusoidal response

Another common input to first-order measurement systems modeled by Equation 03.10 is the sinusoid $u(t) = A \sin \omega t$. For $b_1 = 0$, the solution is

where κ can be found from the initial condition $y(0)$ to be

$$\kappa = y(0) + \frac{b_0 A}{\sqrt{1 + (\omega\tau)^2}} \sin(\arctan(\omega\tau)). \tag{03.11}$$

Figure 03.5 shows responses of a first-order measurement system to sinusoidal inputs (measurands) at different frequencies ω. Note the transient response decays with the same rate τ no matter the input frequency. However, there are two differences in the steady-state responses: the amplitude and phase. In fact, the steady-state amplitude and phase of an output compared to an input present a form of *error* in the measurement. Ideally, the ratio of the output and input is unity; however, for positive ω, this is never quite the case. We define this ratio, called the *magnitude ratio* $M(\omega)$, to be the steady-state output amplitude over the forcing amplitude. For first-order systems,

magnitude ratio $M(\omega)$

dynamic error

A metric for the nearness of $M(\omega)$ to unity is called the *dynamic error* $\delta(\omega)$, given by

Ideally, $\delta(\omega) = 0$, but as we can see from the expression for $M(\omega)$, this is only ever approximately true for nonzero τ and ω. However, it is small when the product $\omega\tau$ is small. So, in order to minimize dynamic

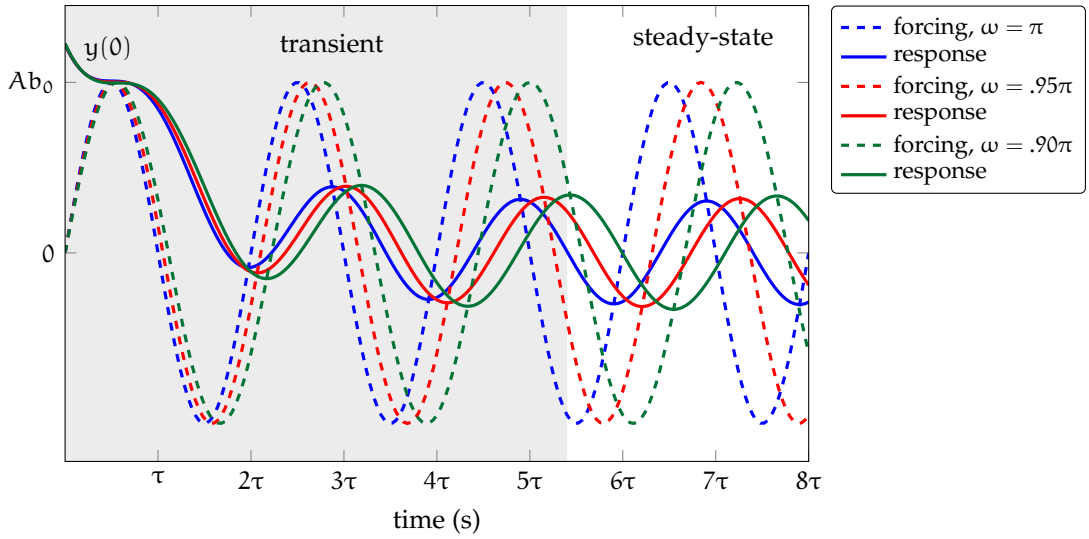


Figure 03.5: response $y(t)$ of a first order system with input $u(t) = A \sin \omega t$ and $b_1 = 0$ for three values of ω. The forcing function (measurand) is $b_0 A \sin \omega t$.

error for the measurement of a sinusoid at a given frequency, we must strive to minimize the time constant τ. It is common to call "good enough" $M(\omega) \geqslant .707$.

Similarly, the phase difference of the output relative to the input is ideally zero. Therefore, the phase shift $\phi(\omega)$ is another type of error and, for first-order systems, is given by

This corresponds to a *time-delay* $\beta_1(\omega)$ in the measurement:

time-delay $\beta_1(\omega)$

Clearly, we want to minimize $\phi(\omega)$ and $\beta_1(\omega)$. Typically, this is achieved by minimizing τ, which corresponds to the minimization of τ for the minimization of the dynamic error.

Note that the steady-state response of the measurement system to sinusoidal inputs is characterized by $M(\omega)$ and $\phi(\omega)$. In fact, a crucial identify will be observed here:

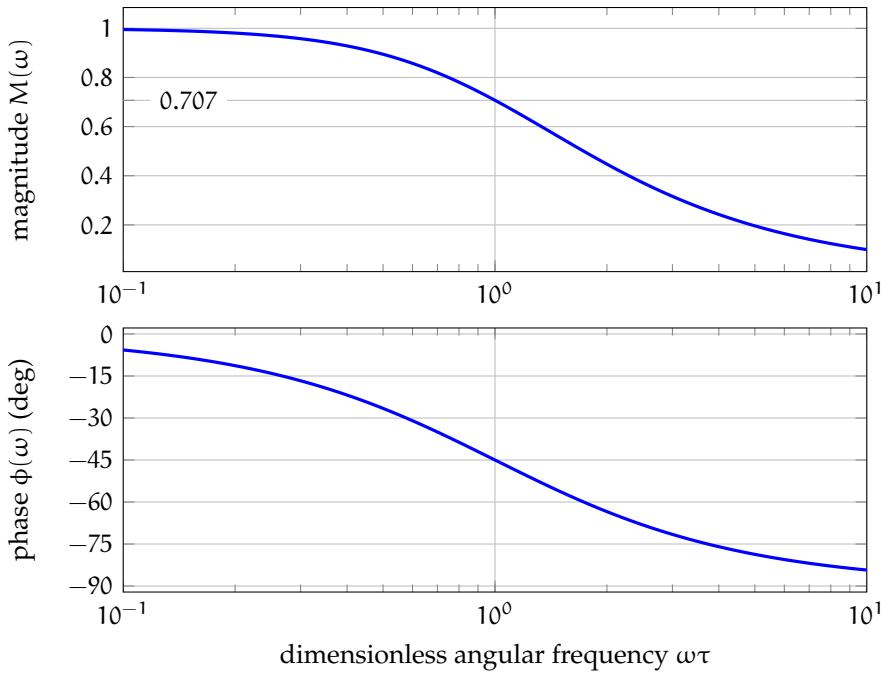


Figure 03.6: the magnitude ratio and phase.

> *the magnitude ratio* $M(\omega)$ *and phase* $\phi(\omega)$ *are equal to the magnitude* $|H(j\omega)|$ *and phase* $\angle H(j\omega)$ *of the frequency response function* $H(j\omega)$.

This is recognized as being the complex amplitude of the output over the input, which are plotted in Figure 03.6.

Lecture 03.04 Second-order measurement systems: free response

Second-order measurement systems have input-output differential equations of the form

$$\frac{d^2y}{dt^2} + 2\zeta\omega_n \frac{dy}{dt} + \omega_n^2 y = f(t) \tag{03.12}$$

where ω_n is called the *natural frequency*, ζ is called the (dimensionless) *damping ratio*, and f is a forcing function that depends the input u as

natural frequency ω_n
damping ratio ζ

$$f(t) = b_2 \frac{d^2u}{dt^2} + b_1 \frac{du}{dt} + b_0 u. \tag{03.13}$$

Measurement systems with two energy storage elements—such as those with an inertial element and a spring-like element—can be modeled with second-order systems.

For distinct roots ($\lambda_1 \neq \lambda_2$), the homogeneous solution is, for some real constants κ_1 and κ_2,

$$y_h(t) = \kappa_1 e^{\lambda_1 t} + \kappa_2 e^{\lambda_2 t} \tag{03.14}$$

where

$$\lambda_1, \lambda_2 = -\zeta\omega_n \pm \omega_n \sqrt{\zeta^2 - 1}. \tag{03.15}$$

03.04.1 Free response

The *free response* y_{fr} of a system is its response to initial conditions and no forcing ($f(t) = 0$). This is useful for two reasons:

free response y_{fr}

1. perturbations of the measurement system from equilibrium result in free response, making it critical; and
2. the free response can be added to a forced response.

The free response is found by applying initial conditions to the homogeneous solution. With initial conditions $y(0) = y_0$ and $\dot{y}(0) = 0$, the free response is

$$y_{fr}(t) = y_0 \frac{1}{\lambda_2 - \lambda_1} \left(\lambda_2 e^{\lambda_1 t} - \lambda_1 e^{\lambda_2 t} \right). \tag{03.16}$$

There are five possibilities for the location of the roots λ_1 and λ_2, all determined by the value of ζ.

$\zeta \in (-\infty, 0)$**: unstable** This case is very undesirable because it means our measurement system is unstable and, given any nonzero input or output, will *explode* ☹ to infinity. Not a good look.

boom ☹

$\zeta = 0$**: undamped** An undamped system will oscillate forever if perturbed from zero output. Once again, a bad look for a measurement device.

$\zeta \in (0, 1)$**: underdamped** Roughly speaking, underdamped systems oscillate, but not forever. Let's consider the form of the solution for initial conditions and no forcing. The roots of the characteristic equation are

$$\lambda_1, \lambda_2 = -\zeta\omega_n \pm j\omega_n\sqrt{1-\zeta^2} = -\zeta\omega_n \pm j\omega_d \tag{03.17}$$

damped natural frequency ω_d

where the *damped natural frequency* ω_d is defined as

$$\omega_d \equiv \omega_n\sqrt{1-\zeta^2}. \tag{03.18}$$

From Equation (03.16) for the free response, using Euler's formulas to write in terms of trigonometric functions, and the initial conditions $y(0) = y_0$ and $\dot{y}(0) = 0$, we have

$$y_{fr}(t) = y_0 \frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}} \cos(\omega_d t - \psi) \tag{03.19}$$

where the phase ψ is

$$\psi = \arctan\frac{\zeta}{\sqrt{1-\zeta^2}}. \tag{03.20}$$

This is an oscillation that decays to the value it oscillates about, $y(t)|_{t\to\infty} = 0$. So any perturbation of a critically damped measurement system will result in a decaying oscillation about equilibrium.

$\zeta = 1$**: critically damped** In this case, the roots of the characteristic equation are equal:

$$\lambda_1 = \lambda_2 = -\omega_n \tag{03.21}$$

So we must modify Equation 03.14 with a factor of t for the homogeneous solution. The free response for initial conditions $y(0) = y_0$ and $\dot{y}(0) = 0$, we have

$$y_{fr}(t) = y_0\,(1 + \omega_n t)\, e^{-\omega_n t}. \tag{03.22}$$

This decays without oscillation, but just barely.

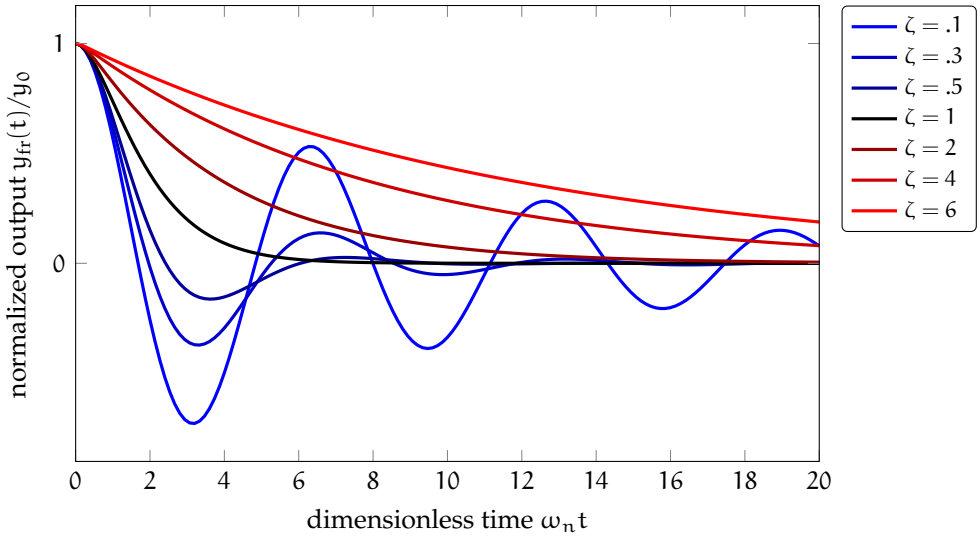


Figure 03.7: free response $y_{fr}(t)$ of a second-order system with initial conditions $y(0) = y_0$ and $\dot{y}(0) = 0$ for different values of ζ. Underdamped, critically damped, and overdamped cases are displayed.

$\zeta \in (1, \infty)$**: overdamped** Here the roots of the characteristic equation are distinct and real. From Equation (03.16) with free response to initial conditions $y(0) = y_0$ and $\dot{y}(0) = 0$, we have the sum of two decaying real exponentials. This response will neither overshoot nor oscillate—like the critically damped case—but with even lesser gusto.

Figure 03.7 displays the free response results. Note that a small damping ratio results in overshooting and oscillation about the equilibrium value. In contrast, large damping ratio results in neither overshoot nor oscillation. However, both small and large damping ratios yield responses that take longer durations to approach equilibrium than damping ratios near unity. For this reason, the damping ratio of a measurement system should be close to one. There are tradeoffs on either side of $\zeta = 1$. Slightly less than one yields faster responses that overshoot a small amount. Slightly greater than one yields slower responses less prone to oscillation.

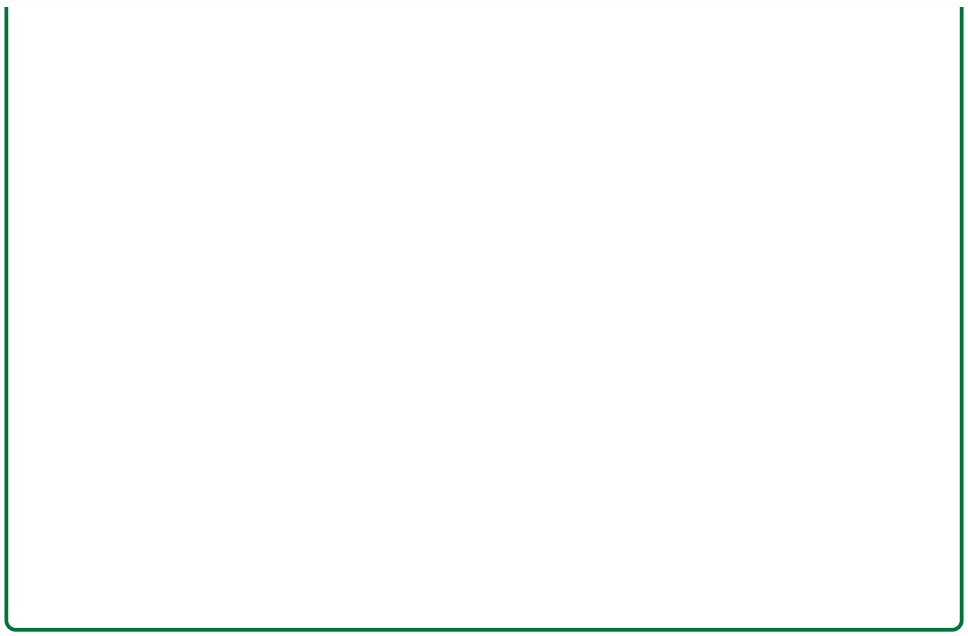
Example 03.04-1 MRFM cantilever beam detector

In magnetic resonance force microscopy (MRFM), the primary detector is a small cantilever beam with a magnetic tip. Model the beam as

a spring-mass-damper system with mass $m = 6$ pg,[a] spring constant $k = 15$ mN/m, and damping coefficient $B = 37.7 \cdot 10^{-15}$ N·s/m.

1. What is the natural frequency ω_n?
2. What is the damping ratio ζ?
3. In free response, how long before the amplitude must be less than 10% of its initial value? An upper bound is sufficient.

[a] One pg $= 10^{-12}$g $= 10^{-15}$kg.



Lecture 03.05 Second-order measurement systems: forced response

Second-order measurement systems are subjected to a variety of forcing functions f. In this lecture, we examine two common varieties: step forcing and sinusoidal forcing. In what follows, we develop *forced response* y_{fo} solutions, which are the *specific solution* responses of systems to given inputs and *zero initial conditions*: all initial conditions set to zero. In Lecture 03.08, a method is presented for combining free and forced response.

forced response y_{fo}

zero initial conditions

03.05.1 Step response

Step forcing of the form $f(t) = Ku_s(t)$, where $K \in \mathbb{R}$ and u_s is the unit step function, models abrupt changes to the input (measurand). The solution is found by applying zero initial conditions ($y(0) = 0$ and $\dot{y}(0) = 0$) to the general solution. If the roots of the characteristic equation λ_1 and λ_2 are distinct, the forced response is

$$y_{fo}(t) = \frac{K}{\omega_n^2}\left(1 - \frac{1}{\lambda_2 - \lambda_1}\left(\lambda_2 e^{\lambda_1 t} - \lambda_1 e^{\lambda_2 t}\right)\right) \tag{03.23}$$

where

$$\lambda_1, \lambda_2 = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}. \tag{03.24}$$

Once again, there are five possibilities for the location of the roots of the characteristic equation λ_1 and λ_2, all determined by the value of ζ. However, there are three important cases for measurement systems: underdamped, critically damped, and overdamped.

$\zeta \in (0, 1)$ **underdamped** In this case, the roots are distinct and complex:

$$\lambda_1, \lambda_2 = -\zeta\omega_n \pm j\omega_d. \tag{03.25}$$

From Equation 03.23, the forced step response is

$$y_{fo}(t) = \frac{K}{\omega_n^2}\left(1 - \frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}}\cos(\omega_d t + \psi)\right) \tag{03.26}$$

where the phase ψ is

$$\psi = \arctan\frac{\zeta}{\sqrt{1-\zeta^2}}. \tag{03.27}$$

This response overshoots, oscillates about, and decays to K/ω_n^2.

$\zeta = 1$ **critically damped** The roots are equal and real:

$$\lambda_1, \lambda_2 = -\omega_n \tag{03.28}$$

so the forced step of Equation 03.23 must be modified; it reduces to

$$y_{fo}(t) = \frac{K}{\omega_n^2}\left(1 - e^{-\omega_n t}(1 + \omega_n t)\right). \tag{03.29}$$

This response neither oscillates nor overshoots its steady-state of $\frac{K}{\omega_n^2}$, but just barely.

$\zeta \in (1, \infty)$ **overdamped** In this case, the roots are distinct and real, given by Equation 03.24. The forced step given by Equation 03.23 is the sum of two decaying real exponentials. These responses neither overshoot nor oscillate about their steady-state of K/ω_n^2. With increasing ζ, approach to steady-state slows.

Figure 03.8 displays the forced step response results. These responses are inverted versions of the free responses of 03.04.1. Note that a small damping ratio results in overshooting and oscillation about the steady-state value. In contrast, large damping ratio results in neither overshoot nor oscillation. However, both small and large damping ratios yield responses that take longer durations to approach equilibrium than damping ratios near unity. For this reason, the damping ratio of a measurement system should be close to $\zeta = 1$. There are tradeoffs on either side of one. Slightly less yields faster responses that overshoot a small amount. Slightly greater than one yields slower responses less prone to oscillation.

03.05.2 Sinusoidal response

Here we consider only steady-state sinusoidal response, allowing us to focus on frequency-domain considerations. The second-order system transfer function, found from the Laplace transform of Equation 03.12, from input u (generating forcing function $f(t) = Ku(t)$) to output y has the form

$$H(s) = \frac{K/\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \tag{03.30}$$

The frequency response function $H(j\omega)$ is found via the substitution $s \to j\omega$, where ω is the input sinusoidal frequency:

$$H(j\omega) = \frac{K/\omega_n^2}{(1 - (\omega/\omega_n)^2) + j(2\zeta\omega/\omega_n)}. \tag{03.31}$$

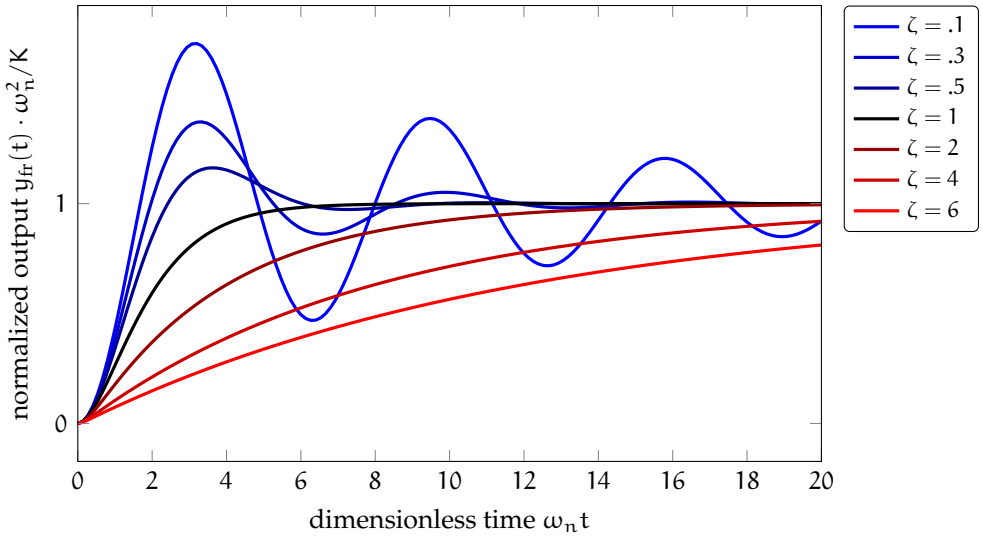


Figure 03.8: forced step response $y_{fo}(t)$ of a second-order system for different values of ζ. Underdamped, critically damped, and overdamped cases are displayed.

Writing this in terms of a magnitude and phase,

$$|H(j\omega)| = \frac{K/\omega_n^2}{\sqrt{(1-(\omega/\omega_n)^2)^2 + (2\zeta\omega/\omega_n)^2}} \quad \text{and} \tag{03.32a}$$

$$\angle H(j\omega) = \arctan \frac{-2\zeta\omega/\omega_n}{1-(\omega/\omega_n)^2}. \tag{03.32b}$$

These functions are plotted in Figure 03.9 for a range of ζ. Note especially that the magnitude $|H(j\omega)|$ is near unity for low frequency, peaks (for underdamped systems) near ω_n, and tapers to zero high frequency. This corresponds to amplitude ratios between the input sinusoidal amplitude and output sinusoidal amplitude.

The phase $\angle H(j\omega)$ is near zero for low frequency is -90 deg at ω_n, and approaches -180 deg for high frequency. This corresponds to a phase lag between the input and output sinusoids.

For input $u(t) = A\sin(\omega t + \phi)$, the steady-state response y_{ss} can be found directly from the frequency response:

$$y_{ss}(t) = A|H(j\omega)|\sin(\omega t + \phi + \angle H(j\omega)). \tag{03.33}$$

We use the same metric as before for the nearness of $|H(j\omega)|$ to unity—the *dynamic error*—

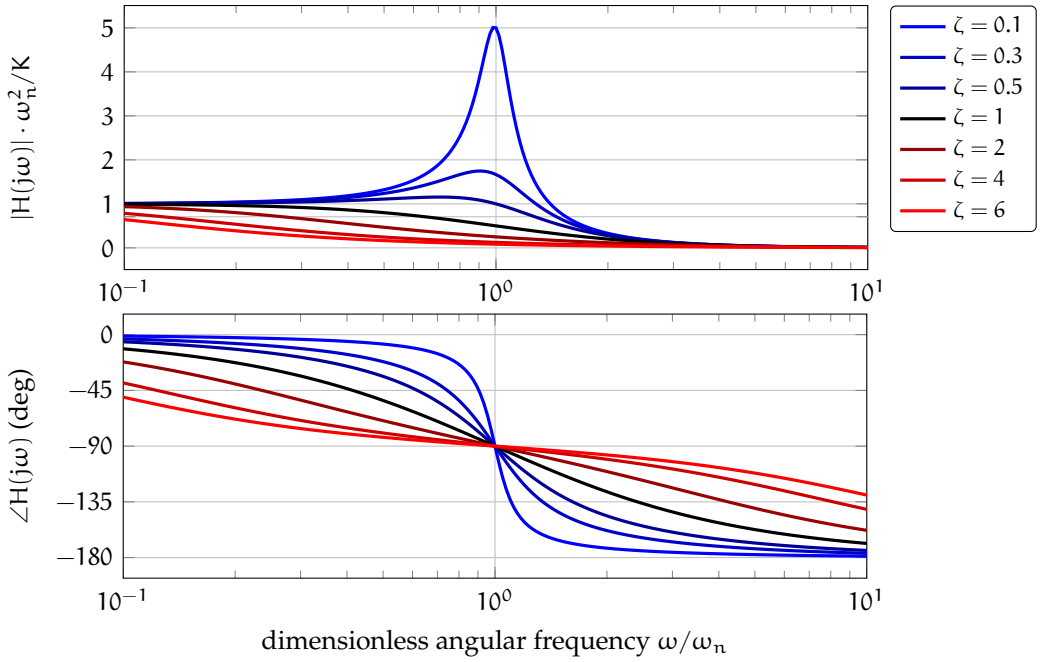


Figure 03.9: the magnitude and phase of the frequency response function $H(j\omega)$.

When $\delta(\omega) \approx 0$, the input (measurand) amplitude and output (indication) amplitude are approximately equal. Note that, according to Figure 03.9, when $\delta(\omega) \approx 0$ (i.e. $|H(j\omega)| \approx 1$), the phase lag (and therefore the time lag) is relatively small. This is ideal for second-order measurement systems.

Lecture 03.06 Transient response characteristics

step response

We define four transient response characteristics, all defined in terms of a system's *step input response*. For the following, please refer to the illustration in Figure 03.10.

rise time

1. The *rise time* T_r is the duration from the time the response reaches 10 % to the time it reaches 90 % of its final value.

peak time

2. The *peak time* T_p is the time at which the response reaches its first or maximum peak.[1]

percent overshoot

3. The *percent overshoot* %OS expresses the amount the response overshoots its steady-state value, expressed as a percentage of the steady-state value.

settling time

4. The *settling time* T_s is the time at which the response reaches, and thereafter remains within, ± 2 % of its steady-state value.[1]

[1] This definition assumes the step input occurs at $t = 0$. Otherwise, subtract the nonzero initial time.

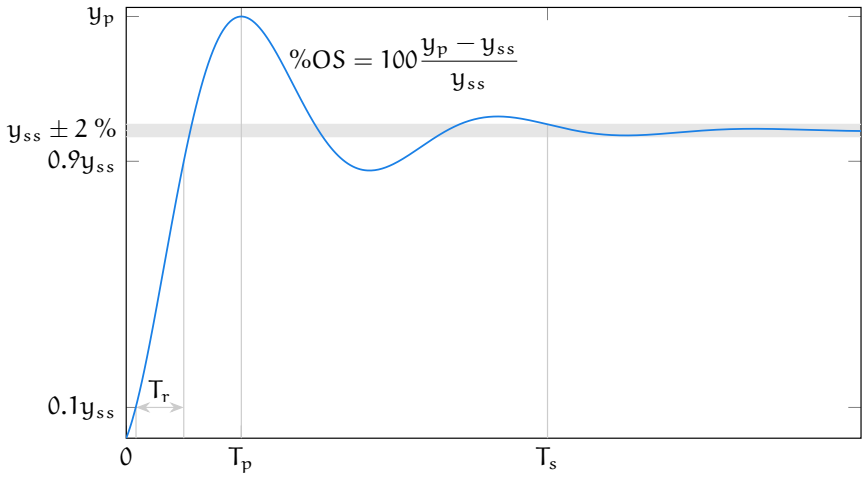


Figure 03.10: transient response characteristics rise time T_r, peak time T_p, percent overshoot %OS, and settling time T_s in terms of a response's steady-state y_{ss} and peak y_p.

Lecture 03.07 Exact analytical transient response characteristics of first- and second-order systems

03.07.1 First-order systems

First-order systems without zeros have transient responses characterized by a *time-constant* τ that appears in the total response as

time-constant

$$\_e^{-t/\tau} + \_\_. \tag{03.34}$$

The transient exponential decays such that in three time constants 3τ only 5 % of the term remains; in 5τ, less than 1 %.

There is neither peak nor overshoot for this type of response. However, the *rise time* for these systems is found from the solution to the first-order input-output ode for $y(0) = 0$:

rise time

$$y(t) = Kb_0 \left(1 - e^{-t/\tau}\right). \tag{03.35}$$

The rise time is, by definition, the duration of the time interval $[t_1, t_2]$ such that

$$y(t_1) = 0.1y_{ss} \quad \text{to} \quad y(t_2) = 0.9y_{ss}. \tag{03.36}$$

The first of these yields

$$Kb_0 \left(1 - e^{-t_1/\tau}\right) = 0.1Kb_0 \Rightarrow \tag{03.37a}$$

$$t_1 = -\tau \ln 0.9 \tag{03.37b}$$

$$\approx 0.1054\tau. \tag{03.37c}$$

Solving in an analogous fashion, we find $t_2 \approx 2.3026\tau$. The interval, then, is $t_2 - t_1 = 2.1972\tau$.

Equation 03.38 first-order system rise time

Finally, the *settling time* can be derived in a fashion similar to the rise time.

settling time

Equation 03.39 first-order system settling time

03.07.2 Second-order systems

Second-order system transient responses are characterized by a natural frequency ω_n and damping ratio ζ. Following a procedure very similar to that for first-order systems, the following relationships can be derived.

rise time

The *rise time* T_r does not have an analytical solution in terms of ω_n and ζ. However, Figure 03.11 can be developed, numerically.

peak time

The *peak time* T_p has the following, simple expression

$$T_p = \frac{\pi}{\omega_d}, \tag{03.40}$$

where $\omega_d = \omega_n\sqrt{1-\zeta^2}$ is the damped natural frequency.

percent overshoot

The *percent overshoot* %OS is related directly to ζ as follows

$$\%OS = 100\exp\frac{-\zeta\pi}{\sqrt{1-\zeta^2}} \quad \Leftrightarrow \quad \zeta = \frac{-\ln(\%OS/100)}{\sqrt{\pi^2+\ln^2(\%OS/100)}}. \tag{03.41}$$

settling time

Finally, the *settling time* T_s is expressed as

$$T_s = \frac{4}{\zeta\omega_n}. \tag{03.42}$$

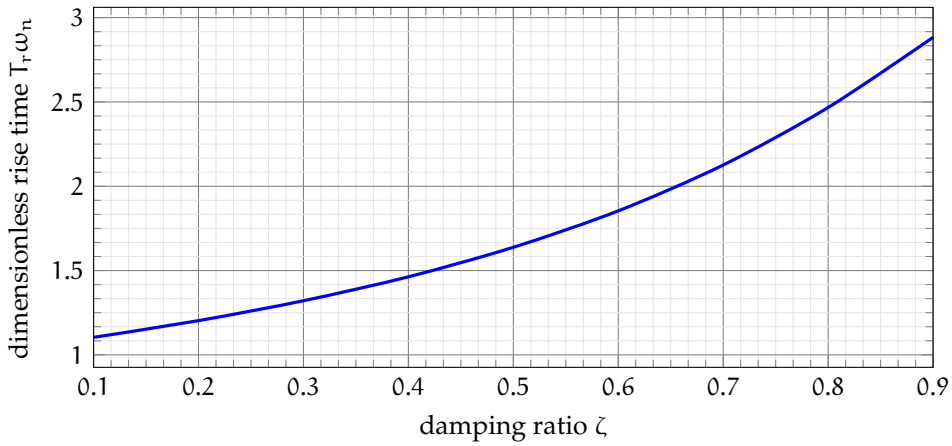


Figure 03.11: the relationship between rise time and damping ratio.

Lecture 03.08 Properties of linear, time-invariant systems

superposition
linear, time-invariant (LTI) systems

From the principle of *superposition*, *linear, time invariant* (LTI) system responses to both initial conditions and nonzero forcing can be obtained by summing the free and forced responses:

$$y(t) = y_{fr}(t) + y_{fo}(t).$$

Moreover, superposition says that any linear combination of inputs yields a corresponding linear combination of outputs. That is, we can find the response of a system to each input, separately, then linearly combine (scale and sum) the results according to the original linear combination. That is, for inputs u_1 and u_2 and constants $a_1, a_2 \in \mathbb{R}$, a forcing function

would yield output

where y_1 and y_2 are the outputs for inputs u_1 and u_2, respectively.

This powerful principle allows us to construct solutions to complex forcing functions by decomposing the problem. It also allows us to make extensive use of existing solutions to common inputs.

There is one more LTI system property worth noting here. Let a system have input u_1 and corresponding output y_1. If the system is then given input $u_2(t) = \dot{u}_1(t)$, the corresponding output is

Similarly, if the same system is then given input $u_3(t) = \int_0^t u_1(\tau)d\tau$, the corresponding output is

Lecture 03.09 Response to periodic inputs

We have already considered the response of first- and second-order measurement systems to sinusoidal inputs (measurands). These are not the only periodic inputs encountered by measurement systems; in fact, we frequently encounter non-sinusoidal periodic inputs.

Fortunately, we already have the mathematical apparatus to deal with these inputs. Recall that a periodic signal u with period T has a *Fourier series* representation, for $n \in \mathbb{N}_0$ and $\omega_n \equiv 2\pi n/T$ is the angular frequency of component n,

Fourier series

$$u(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} C_n \sin(\omega_n t + \phi_n). \tag{03.43}$$

where we have written the sum in terms of harmonic amplitudes C_n and phases ϕ_n defined as via A.02.11:

$$C_n = \sqrt{a_n^2 + b_n^2} \text{ and} \tag{03.44}$$

$$\phi_n = \arctan \frac{b_n}{a_n} \tag{03.45}$$

and where a_n and b_n are found from the trigonometric Fourier series analysis

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} u(t) \cos(\omega_n t)\, dt \tag{03.46}$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} u(t) \sin(\omega_n t)\, dt. \tag{03.47}$$

In other words, *a periodic signal can be represented as a sum of sinusoids*.

When we combine this with the *principle of superposition*—specifically with the fact that, for linear systems, *the linear combination of inputs yields an equivalent linear combination of outputs*—we can compute the response of a system to a periodic input by

principle of superposition

1. representing the input u with a Fourier series,
2. computing the response of the system to each term in the series, and
3. summing the result.

This is valid for transient and steady-state analysis, but, when working with periodic functions, we typically are most concerned with steady-state.

Conveniently, the steady-state response y_n of a system with frequency response function $H(j\omega)$ to sinusoidal forcing $u_n = C_n \sin(\omega_n t + \phi_n)$ has already been developed:

The special case is y_0, which is[2]

From the principle of superposition, the output to the sum of the inputs u_n is just the sum of outputs y_n:

In Example 02.02-1, we found that a square wave of amplitude one has trignometric Fourier series components

$$\begin{aligned} a_n &= 0 \text{ and} \\ b_n &= \frac{2}{n\pi}(1 - \cos(n\pi)) \\ &= \begin{cases} 0 & n \text{ even} \\ \frac{4}{n\pi} & n \text{ odd.} \end{cases} \end{aligned}$$

Therefore, from the definitions of C_n and ϕ_n, with $b_n \geqslant 0$,

$$\begin{aligned} C_n &= b_n \text{ and} \\ \phi_n &= \arctan\frac{b_n}{a_n} \\ &= \begin{cases} ¿ & \text{for } n \text{ even} \\ \pi/2 & \text{for } n \text{ odd.} \end{cases} \end{aligned}$$

If we consider the steady-state response of a system with frequency response function $H(j\omega)$ to this square wave input, we can create Figure 03.12 and Figure 03.13, showing how the system responds to this input. These figures are generated by applying the expression for y_n.

[2] This is derived by assuming an input amplitude $a_0/2$ and angular frequency 0 rad/s.

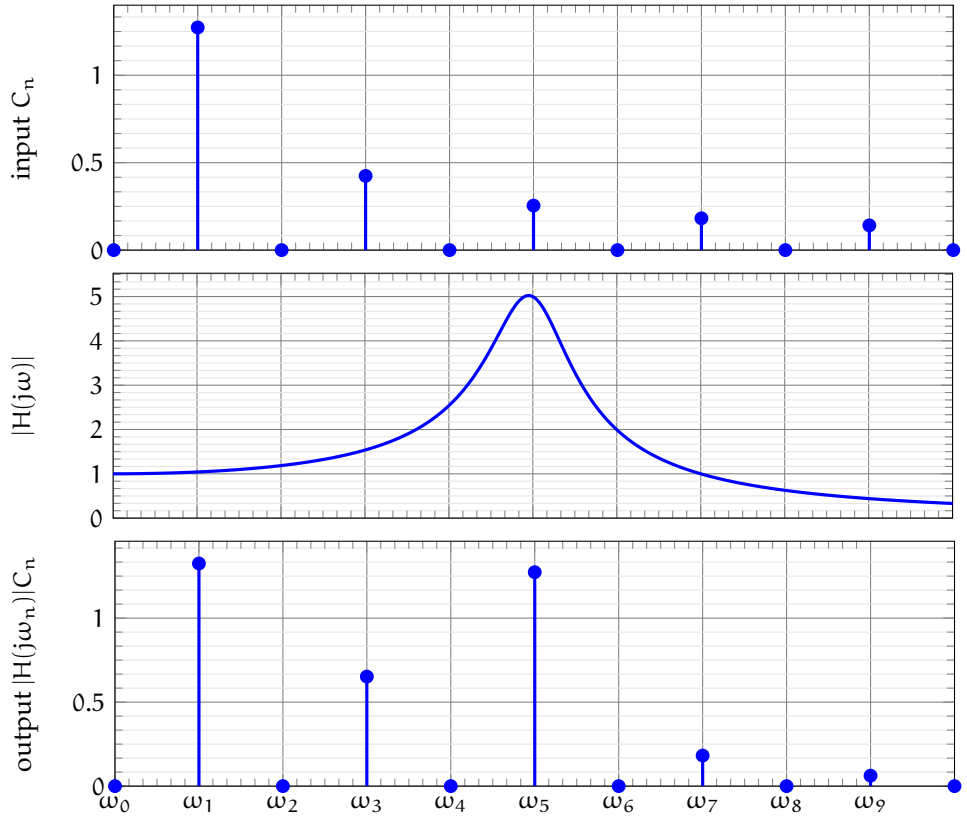


Figure 03.12: the magnitude line spectrum C_n of the input, which is operated on by the measurement system with frequency response function $H(j\omega)$ to form the output magnitude line spectrum $|H(j\omega_n)|C_n$.

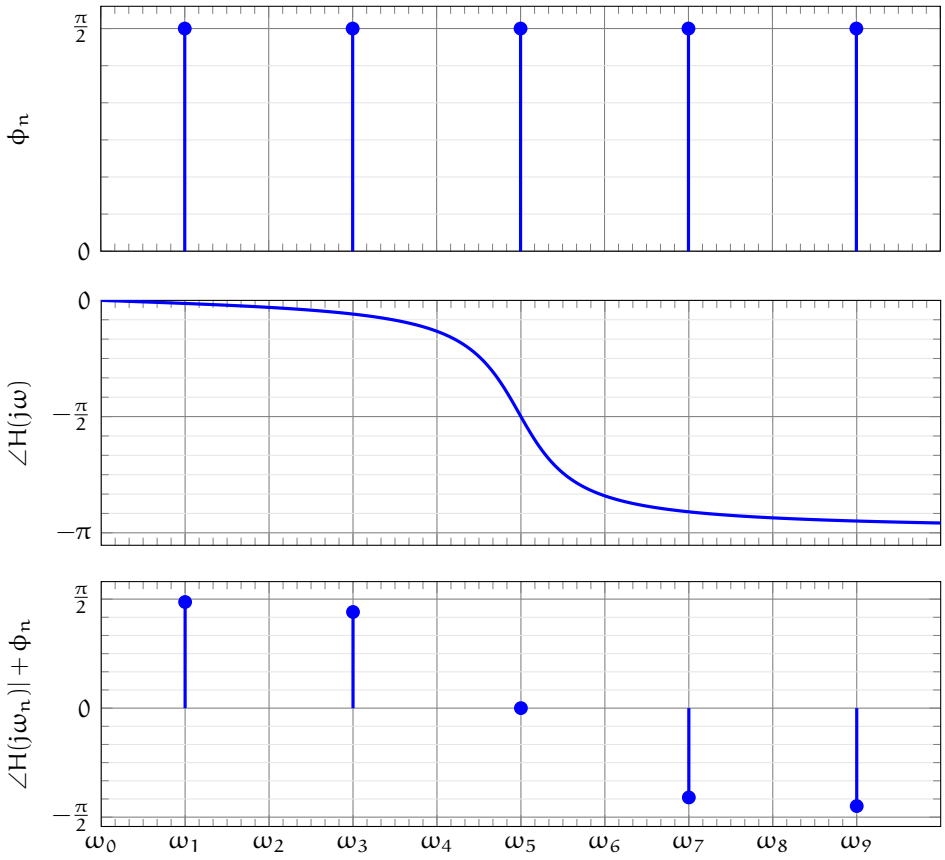


Figure 03.13: the phase line spectrum ϕ_n of the input, which is operated on by the measurement system with frequency response function $H(j\omega)$ to form the output phase line spectrum $\angle H(j\omega_n) + \phi_n$.

Lecture 03.10 Phase linearity

The source for this exercise lecture is in a *Python 2* kernel *Jupyter* notebook. For more information, see python.org and jupyter.org.

See ricopic.one/measurement/notebooks for the source code notebook.

We observed in the last lecture that significant distortion of a non-sinusoidal periodic function can occur. Due to the system's significant underdamping and confounded by the input spectral component right on the natural frequency, significant resonance distorted the signal. But for most measurement systems, we prefer damping ratios nearer unity. Can these systems also exhibit distortion? Yes, they can. Consider the Fourier series of a system with periodic input:

$$u(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} C_n \sin(n\omega_1 t + \phi_n). \tag{03.48}$$

We derived the output:

$$y(t) = \frac{a_0}{2}H(j0) + \sum_{n=1}^{\infty} C_n |H(j\omega_n)| \sin(n\omega_1 t + \phi_n + \angle H(j\omega_n)). \tag{03.49}$$

Let the system's phase shift $\angle H(j\omega)$ depend *linearly* on ω. In this case, the fundamental component has shift $\angle H(j\omega_1)$, the second component has shift $2\angle H(j\omega_1)$, the third $3\angle H(j\omega_1)$, and each subsequent component $n\angle H(j\omega_1)$. The nth sinusoid in the sum becomes

$$\sin(n\omega_1 t + n\angle H(j\omega_1) + \phi_n) \tag{03.50}$$

$$= \sin(n(\omega_1 t + \angle H(j\omega_1)) + \phi_n). \tag{03.51}$$

Letting $\theta = t + \angle H(j\omega_1)/\omega_1$, this becomes

$$\sin(n\omega_1 \theta + \phi_n) \tag{03.52}$$

Within a time-shift, this is exactly the same as the input! Therefore, for an input with dominant frequency components below some frequency ω_c, if the magnitude ratio below ω_c is approximately 1 and if the measurement system's phase shift below ω_c is approximately linear in frequency, the

output should have the same amplitude and shape as the input, with a time lag.

This is an important result. Let ω_c be the *cutoff frequency*, above which the magnitude ratio declines and phase shift becomes significant. Any periodic input with significant frequency components above ω_c will exhibit distortion. Therefore, it is not simply the resonance that distorted the signal from the previous lecture. We explore this in further detail by applying the square wave to the same system with a single difference: now the damping ratio is $\zeta = 0.707$, which has a frequency response that is optimally "flat" and "wide."

First, we generate and plot the input square wave at a few frequencies. The motivation for multiple frequencies will become clear, later. The angular frequency of the `n`th harmonic is defined in terms of the square wave period `T`.

```
def w(n,T): # angular frequency
    return 2*pi*n/T

# three periods to explore
T1 = 2*pi
T2 = 2*T1
T3 = 3*T1

# plot the square waves
t = linspace(0,6*pi,201)
u_1 = signal.square(w(1,T1)*t)
u_2 = signal.square(w(1,T2)*t)
u_3 = signal.square(w(1,T3)*t)
f,ax=plt.subplots(3,1)
f.tight_layout()
ax=plt.subplot(311)
ax.plot(t,u_1)
ax=plt.subplot(312)
ax.plot(t,u_2)
ax=plt.subplot(313)
ax.plot(t,u_3)
plt.xlabel("time (s)")
plt.show()
```

The system will be second-order and defined via its magnitude and phase functions, derived previously. We choose a natural frequency of `5`.

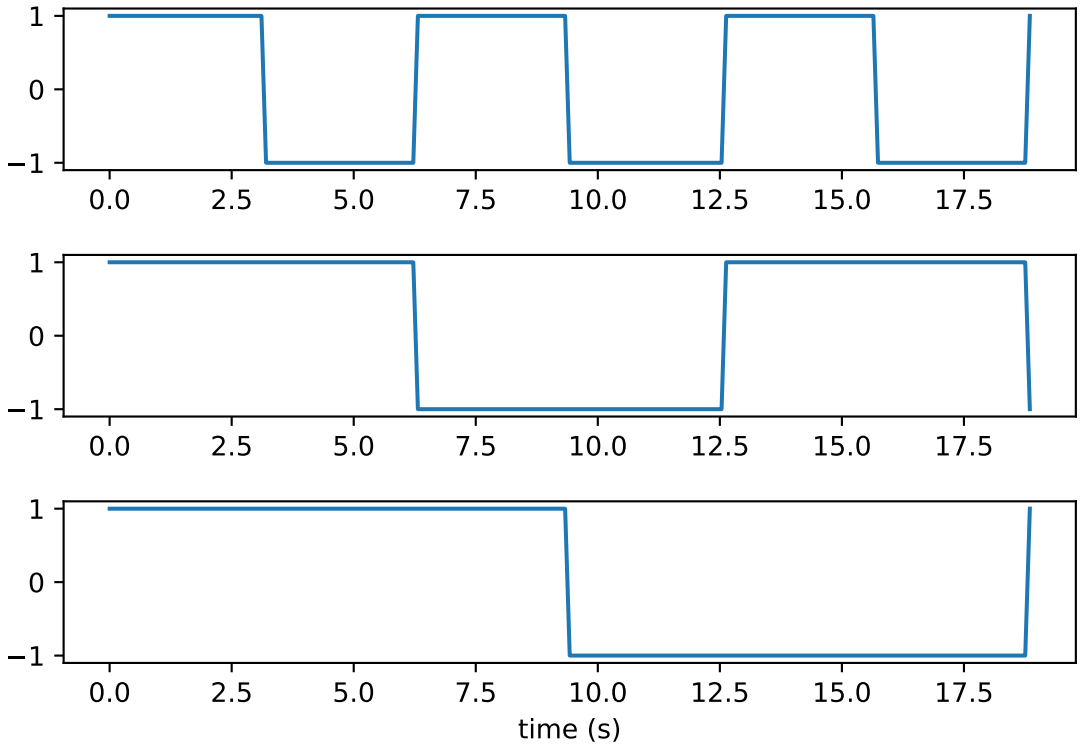


Figure 03.14: png

$$|H(j\omega)| = \frac{1}{\sqrt{(1-(\omega/5)^2)^2 + (2\zeta\omega/5)^2}} \quad \text{and} \tag{03.53}$$

$$\angle H(j\omega) = \arctan\frac{-2\zeta\omega/5}{1-(\omega/5)^2}. \tag{03.54}$$

```
def Hm(w,z):
    return 1/sqrt((1-(w/5)**2)**2 + (2*z*w/5)**2)

def Hp(w,z):
    return arctan2(-2*z*w/5,1-(w/5)**2)
```

Let's plot this for two values of ζ. The results are shown in Figure 03.15.

```
wl = logspace(-1,2,201)
plt.semilogx(wl,Hm(wl,.1),wl,Hm(wl,.7))
plt.ylabel('$|H(j\omega)|$')
plt.legend(['$\zeta = 0.1$','$\zeta = 0.7$'])
if save_figures:
    tikz_save( # save for LaTeX
        'figures/phase_linearity_mag.tex',
        figureheight='.5\linewidth',
        figurewidth='1\linewidth'
    )
plt.show()
plt.semilogx(wl,Hp(wl,.1),wl,Hp(wl,.7))
plt.ylabel('phase of $H(j\omega)$')
if save_figures:
    tikz_save( # save for LaTeX
        'figures/phase_linearity_phase.tex',
        figureheight='.5\linewidth',
        figurewidth='1\linewidth'
    )
plt.show()
```

Now, we must write the Fourier series of the output. Several functions are defined below, culminating in the partial sum `psum` of the series that can be used to plot the response.

```
def a(n):
    return 0

def b(n):
    return 2/(n*pi)*(1-cos(n*pi))

def c(n,a,b):
    return sqrt(a(n)**2+b(n)**2)

def phi(n,a,b):
    return arctan2(a(n),b(n))

def yn(n,a,b,c,phi,z,T,t):
    return c(n,a,b)*Hm(w(n,T),z)*sin(w(n,T)*t+phi(n,a,b)+Hp(w(n,T),z))

def psum(yn,a,b,c,phi,z,T,t,N):
    s = 0
    for n in range(1,N+1):
```

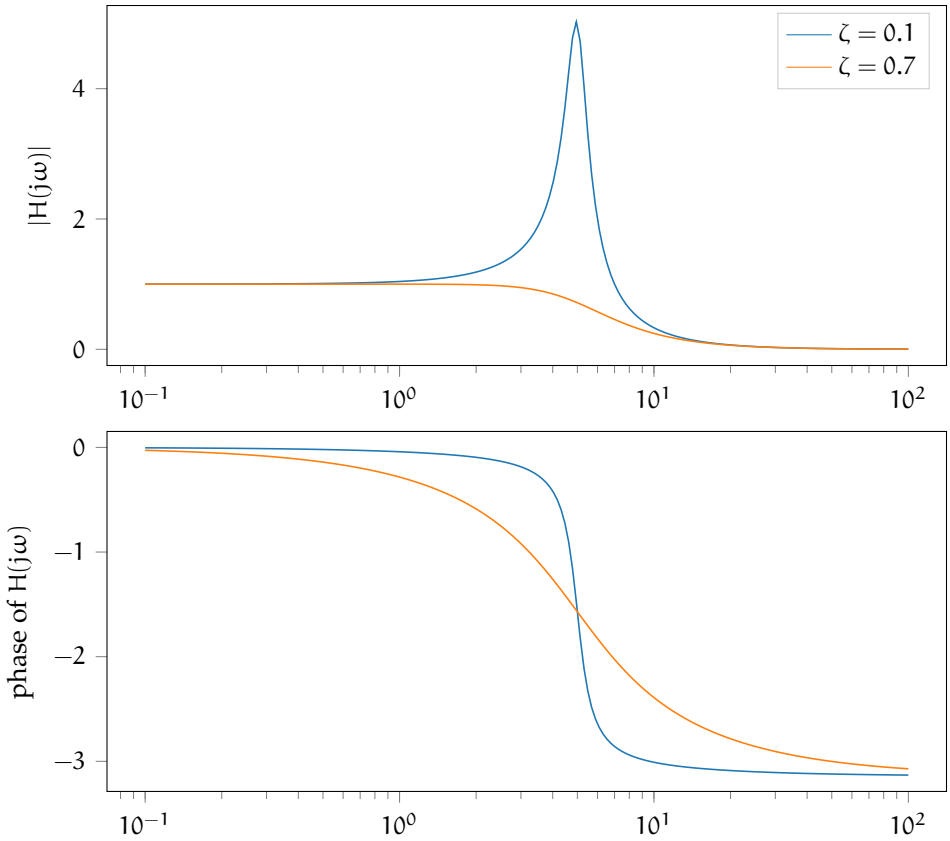


Figure 03.15: magnitude and phase of the frequency response function for two damping ratios.

```
        s = s+yn(n,a,b,c,phi,z,T,t)
    return s
```

Now we turn to three plots: one for each input period `T`. We expect our output will more closely follow the input for larger `T` because fewer significant components will be greater than the cutoff frequency (in this case, the natural frequency `5` rad/s). We also expect our output to more closely follow the input when the damping is around `0.707` rather than when it is around `0.1` because the resonance effects should be absent ("wide" and "flat").

```
N = 200
znow = [.1,.707]
```

```
Tnow = T1
plt.plot(t,u_1)
for zi in znow:
    ps = []
    for ti in t:
        ps.append(psum(yn,a,b,c,phi,zi,Tnow,ti,N))
    plt.plot(t,ps)
if save_figures:
    tikz_save( # save for LaTeX
        'figures/phase_linearity_square_T1.tex',
        figureheight='.4\linewidth',
        figurewidth='1\linewidth'
    )
plt.show()
```

```
Tnow = T2
plt.plot(t,u_2)
for zi in znow:
    ps = []
    for ti in t:
        ps.append(psum(yn,a,b,c,phi,zi,Tnow,ti,N))
    plt.plot(t,ps)
if save_figures:
    tikz_save( # save for LaTeX
        'figures/phase_linearity_square_T2.tex',
        figureheight='.4\linewidth',
        figurewidth='1\linewidth'
    )
plt.show()
```

```
Tnow = T3
plt.plot(t,u_3)
for zi in znow:
    ps = []
    for ti in t:
        ps.append(psum(yn,a,b,c,phi,zi,Tnow,ti,N))
    plt.plot(t,ps)
if save_figures:
    tikz_save( # save for LaTeX
        'figures/phase_linearity_square_T3.tex',
        figureheight='.4\linewidth',
        figurewidth='1\linewidth'
```

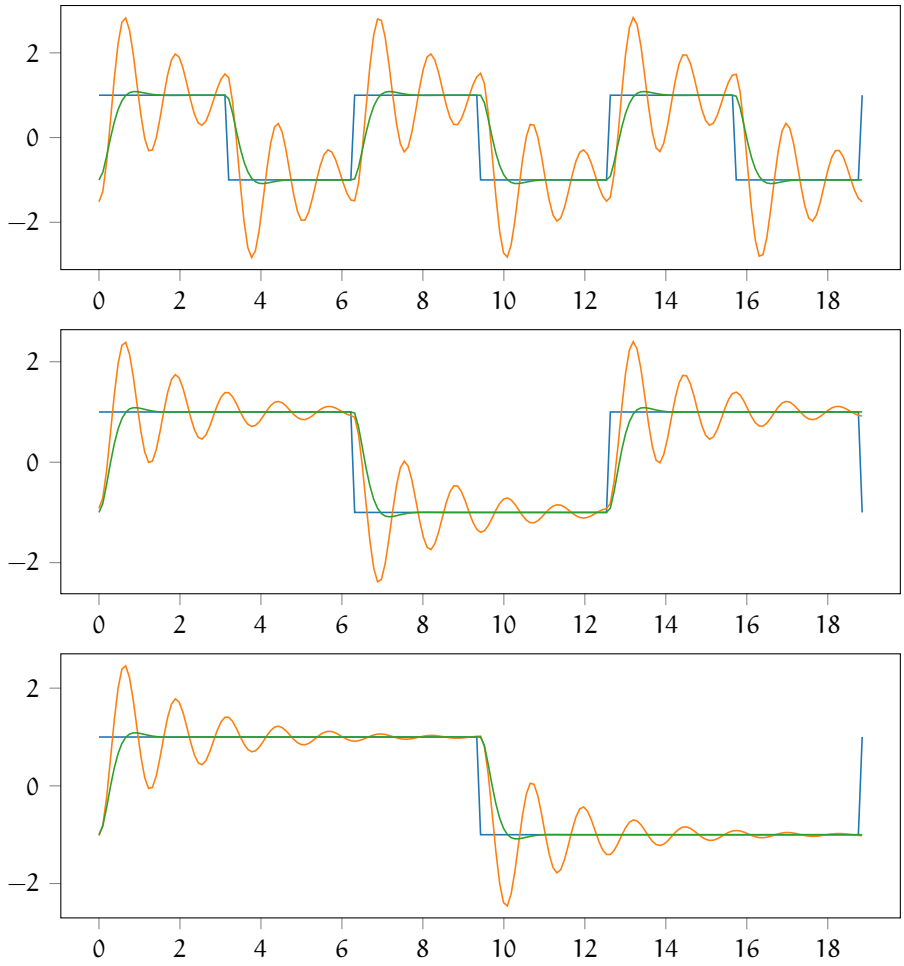


Figure 03.16: square wave inputs with corresponding outputs for $\zeta = 0.1$ and $\zeta = 0.707$.

```
    )
plt.show()
```

03.11 Problems for Chapter 03

03.11.1 Design problem

piezoelectric load sensor
pzt
load cells

Piezoelectric load sensors use the piezoelectric effect of certain materials, such as *pzt*, to transduce force applied to them into voltage across them. These sensors, sometimes called *load cells*, can be used to measure both compression and tension, and have characteristically "high frequency response," meaning they respond quickly to input forces, even those that change quickly.

A typical configuration is shown in Figure 03.17. A force is applid to a chassis that sandwiches two plates made of piezoelectric material, which sandwich an electrode.

Figure 03.18 shows a composite model for the devices. It is best to model the entire system with a single input-output differential equation or transfer function because it is not known *a priori* that the cable and charge amplifier systems will not load the load cell.

Let the load cell produce a source voltage V_p proportional to the input force f as

$$V_p = \alpha f \tag{03.55}$$

where the constant of proportionality is $\alpha = 20$ mV/kN.

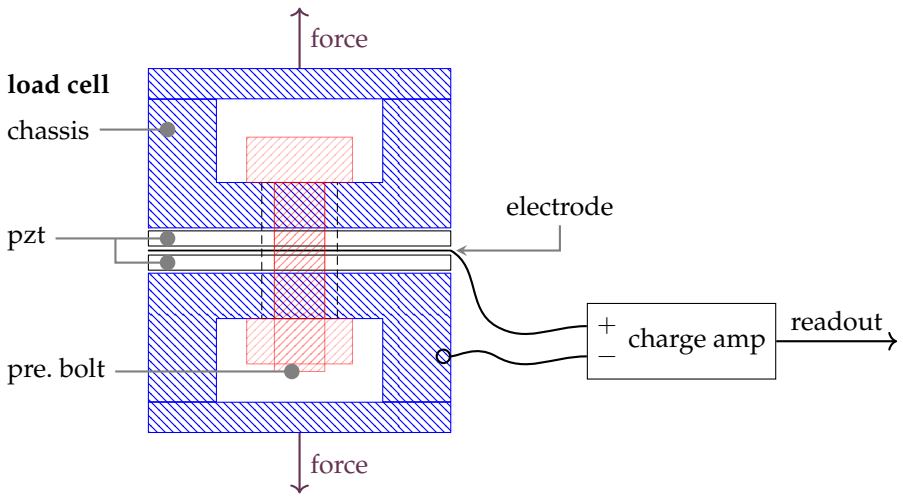


Figure 03.17

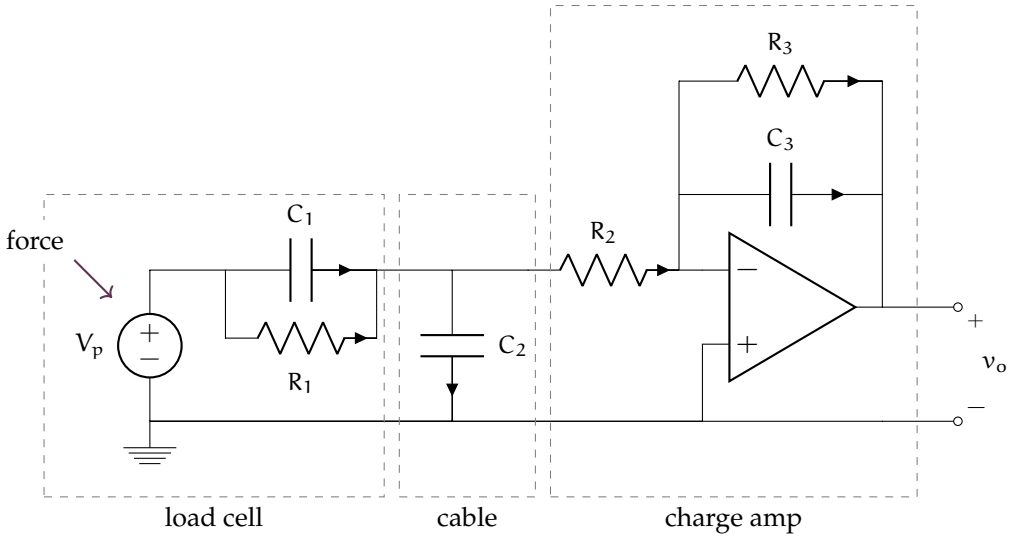


Figure 03.18

Let the piezoelectric material have capacitane $C_1 = 100$ μF and resistance $R_1 = 1$ MΩ. Let the cable have capacitance $C_2 = 30$ pF.

1. Perform a circuit analysis and express the result as a single input-output differential equation relating the input force f to the output voltage v_o.
2. Define the transfer function from $F(s)$ to $V_o(s)$.
3. Define a MATLAB (or similar) model and plot the magnitude and phase of the frequency response function for the designs you develop, below (choose arbitrary R_2, R_3, and C_3 to get started).
4. Design the charge amp by tuning R_2, R_3, and C_3 such that the system responds to a force step input of 50 N with a settling time of less than 1/2 s. Plot the step response of the system. Hint: use MATLAB's `step` function to generate the step response. Note that this system has two time constants. Our primary concern in this problem is the "fast" response to a step and its initial settling time. The feedback capacitor C_3 will slowly discharge via R_3, so be careful not to be fooled by the "slow" decay. It helps to specify the simulation interval manually. For instance, a one-second time interval should be about right to simulate your system response. MATLAB's `step` function has an option for specifying the time interval or array.
5. Design the charge amp by tuning R_2, R_3, and C_3 such that the system

responds to a force sinusoidal input of 50 N at 1 Hz and at 1 kHz with a dynamic error of $< 10\%$. Consider the "flat" interval of the frequency response to be the $|H(j\omega)| = 1$ value for the purposes of the dynamic error formula. Sometimes a system like this would be characterized by a gain equal to the magnitude ratio of this flat interval.

04

Probability, statistics, and estimation

Lecture 04.01 Probability and measurement

Probability theory

Probability theory is a well-defined branch of mathematics. Andrey Kolmogorov described a set of axioms in 1933 that are still in use today as the foundation of probability theory.[1]

interpretation of probability

event

We will implicitly use these axioms in our analysis. The *interpretation* of probability is a contentious matter. Some believe probability quantifies the frequency of the occurrence of some *event* that is repeated in a large number of trials. Others believe it quantifies the state of our knowledge or belief that some event will occur.

In experiments, our measurements are tightly coupled to probability. This is apparent in the questions we ask. Here are some examples.

1. How common is a given event?
2. What is the probability we will reject a good theory based on experimental results?
3. How repeatable are the results?
4. How confident are we in the results?
5. What is the character of the fluctuations and drift in the data?
6. How much data do we need?

[1] For a good introduction to probability theory, see Ash (2008).

Lecture 04.02 Introduction to set theory

set theory

Set theory is a very useful branch of mathematics for engineers. In probability theory, we use the language of set theory. For this reason, we review basic set theory.

set

A *set* is a collection of objects. Set theory gives us a way to describe these collections. Often, the objects in a set are numbers or sets of numbers. However, a set could represent collections of zebras and trees and hairballs. For instance, here are some sets:

field
addition
multiplication
subtraction
division
real numbers

A *field* is a set with special structure. This structure is provided by the *addition* ($+$) and *multiplication* ($\times$) operators and their inverses *subtraction* ($-$) and *division* ($\div$). The quintessential example of a field is the set of *real numbers* $\mathbb{R}$, which admits these operators, making it a field. The reals $\mathbb{R}$, the complex numbers $\mathbb{C}$, the integers $\mathbb{Z}$, and the natural numbers[2] $\mathbb{N}$ are the fields we typically consider.

set membership

Set membership is the belonging of an object to a set. It is denoted with the symbol $\in$, which can be read "is an element of," for element x and set X:

For instance, we might say $7 \in \{1, 7, 2\}$ or $4 \notin \{1, 7, 2\}$. Or, we might declare that a is a real number by stating: $x \in \mathbb{R}$.

set operations

Set operations can be used to construct new sets from established sets. We consider a few common set operations, now.

union

The *union* $\cup$ of sets is the set containing all the elements of the original sets (no repetition allowed). The union of sets A and B is denoted $A \cup B$. For instance, let $A = \{1, 2, 3\}$ and $B = \{-1, 3\}$; then

intersection

The *intersection* $\cap$ of sets is a set containing the elements common to all the original sets. The intersection of sets A and B is denoted $A \cap B$. For instance, let $A = \{1, 2, 3\}$ and $B = \{2, 3, 4\}$; then

[2] When the natural numbers include zero, we write $\mathbb{N}_0$.

empty set

If two sets have no elements in common, the intersection is the *empty set* $\varnothing = \{\}$, the unique set with no elements.

set difference

The *set difference* of two sets A and B is the set of elements in A that aren't also in B. It is denoted $A \setminus B$. For instance, let $A = \{1, 2, 3\}$ and $B = \{2, 3, 4\}$. Then

subset

A *subset* $\subseteq$ of a set is a set, the elements of which are contained in the original set. If the two sets are equal, one is still considered a subset of the other. We call a subset that is not equal to the other set a *proper subset* $\subset$. For instance, let $A = \{1, 2, 3\}$ and $B = \{1, 2\}$. Then

proper subset

complement

The *complement* of a subset is a set of elements of the original set that aren't in the subset. For instance, if $B \subseteq A$, then the complement of B, denoted $\overline{B}$ is

Lecture 04.03 Basic probability theory

The *sample space* Ω of an experiment is the set representing all possible outcomes of the experiment. If a coin is flipped, the sample space is $\Omega = \{H, T\}$, where H is *heads* and T is *tails*. If a coin is flipped twice, the sample space could be

sample space

However, *the same experiment can have different sample spaces*. For instance, for two coin flips, we could also choose

We base our choice of Ω on the problem at hand.

An *event* is a subset of the sample space. That is, an event corresponds to a yes-or-no question about the experiment. For instance, event A (remember: $A \subseteq \Omega$) in the coin flipping experiment (two flips) might be $A = \{HT, TH\}$. A is an event that corresponds to the question, "Is the second flip different than the first?" A is the event for which the answer is "yes."

event

04.03.1 Algebra of events

Because events are sets, we can perform the usual set operations with them.

Example 04.03-1 set operations with events

Consider a toss of a single die. We choose the sample space to be $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let the following define events.

$$A \equiv \{\text{the result is even}\} = \{2, 4, 6\}$$
$$B \equiv \{\text{the result is greater than } 2\} = \{3, 4, 5, 6\}.$$

Find the following event combinations:

$$A \cup B \qquad A \cap B \qquad A \setminus B \qquad B \setminus A \qquad \overline{A} \setminus B.$$

event class

The *event class* $\mathbb{F}$ is often defined as the set of all subsets of Ω. (It's actually more complicated, but we'll ignore that.) So $\mathcal{F}$ is the set of all possible events given a sample sample space Ω. When referring to an event, we often state that it is an element of $\mathcal{F}$. For instance, we might say an event $A \in \mathcal{F}$.

probability measure

We're finally ready to assign probabilities to events. We define the *probability measure* $P : \mathcal{F} \to [0, 1]$ to be a function satisfying the following conditions.

1. For every event $A \in \mathcal{F}$, the probability measure of A is greater than zero—i.e. $P(A) \geqslant 0$.
2. If an event is the entire sample space, its probability measure is unity—i.e. if $A = \Omega$, $P(A) = 1$.
3. If events $A_1, A_2, \cdots$ are disjoint sets (no elements in common), then $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$.

probability space

The three structures we've defined thus far—Ω (sample space), $\mathcal{F}$ (event class), and P (probability measure)—are called the *probability space* $(\Omega, \mathcal{F}, P)$.

We conclude with the basics by observing four facts that can be proven from the definitions above.

1.

2.

3.

4.

Lecture 04.04 Independence and conditional probability

Two events A and B are *independent* if and only if **independent**

$$P(A \cap B) = P(A)P(B).$$

If an experimenter must make a judgment without data about the independence of events, she bases it on her knowledge of the events, as discussed in the following example.

Example 04.04-1 independence

Answer the following questions and imperatives.

1. Consider a single fair die rolled twice. What is the probability that both rolls are 6?
2. What changes if the die is biased by a weight such that $P(\{6\}) = 1/7$?
3. What changes if the die is biased by a magnet, rolled on a magnetic dice-rolling tray such that $P(\{6\}) = 1/7$?
4. What changes if there are two dice, biased by weights such that for each $P(\{6\}) = 1/7$, rolled once, both resulting in 6?
5. What changes if there are two dice, biased by magnets such that for each $P(\{6\}) = 1/7$, rolled once, both resulting in 6?

04.04.1 Conditional probability

dependent
conditional probability

If events A and B are somehow *dependent*, we need a way to compute the probability of B occurring given that A occurs. This is called the *conditional probability* of B given A, and is denoted $P(B|A)$. For $P(A) > 0$, it is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{04.1}$$

We can interpret this as a restriction of the sample space Ω to A; i.e. the new sample space $\Omega' = A \subseteq \Omega$. Note that if A and B are independent, we obtain the obvious result:

Example 04.04-2 dependence

Consider two unbiased dice rolled once. Let events $A = \{\text{sum of faces} = 8\}$ and $B = \{\text{faces are equal}\}$. What is the probability the faces are equal given that their sum is 8?

Lecture 04.05 Bayes' theorem

Given two events A and B, *Bayes' theorem* (aka Bayes' rule) states that **Bayes' theorem**

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}. \tag{04.2}$$

Sometimes this is written

This is a useful theorem for determining a test's effectiveness. If a test is performed to determine whether an event has occurred, we might as questions like "if the test indicates that the event has occurred, what is the probability it has actually occurred?" Bayes' theorem can help compute an answer.

The test can be either *positive* or *negative* and this result can be either *true* ☺ or *false* ☹.

There are four options, then. Consider an event A and an event that is a test result B indicating that event A has occurred. Table 04.1 shows these four possible test outcomes. Clearly, the desirable result for any test is that it is *true*. However, no test is true 100 percent of the time. So sometimes it is desirable to err on the side of the false positive, as in the case of a medical diagnostic. Other times, it is more desirable to err on the side of a false negative, as in the case of testing for defects in manufactured balloons (when a false negative isn't a big deal).

	A	$\overline{A}$
positive B	true ☺	false ☹
negative $\overline{B}$	false ☹	true ☺

Table 04.1: test outcome B for event A.

Some interesting results can be found from this. For instance, we can plot, as in Figure 04.1 the relationship between the probability of a positive test result given that the event actually occurs $P(B|A)$ and the probability of the event occurring given that the test is positive $P(A|B)$. (Note that, in both cases, it is the conditional probability of a true positive given some condition.)

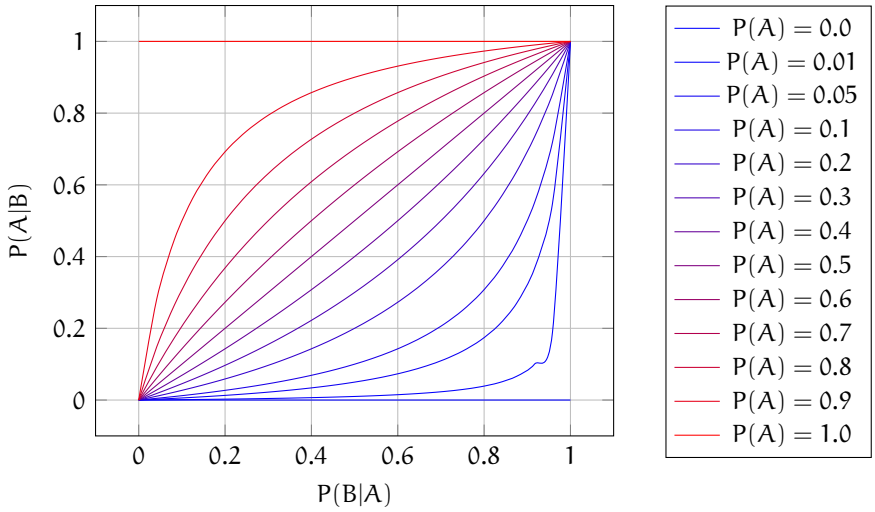


Figure 04.1: the probability that an event A occurred given that a test for it is positive B for different probabilities that the event A occurs.

Example 04.05-1 Bayes' theorem

Suppose 0.1% of springs manufactured at a given plant are defective. Suppose you need to design a test that has probability of 99% that a part, given that it is indicated defective, is actually so. What probability should your test have, given that a part is defective, of indicating that it is actually so?

Lecture 04.06 Populations, samples, and machine learning

An experiment's *population* is a complete collection of objects that we would like to study. These objects can be people, machines, processes, or anything else we would like to understand experimentally. **population**

Of course, we typically can't measure *all* of the population. Instead, we take a subset of the population—called a *sample*—and infer the characteristics of the entire population from this sample. **sample**

However, this inference that the sample is somehow representative of the population assumes the sample size is sufficiently large and that the sampling is *random*. This means selection of the sample should be such that no one group within a population are systematically over- or under-represented in the sample. **random**

machine learning

Machine learning is a field that makes extensive use of measurements and statistical inference. In it, an algorithm is *trained* by exposure to sample data, which is called a *training set*. The variables measured are called *features*. Typically, a *predictive model* is developed that can be used to extrapolate from the data to a new situation. The methods of statistical analysis we introduce in this chapter are the foundation of most machine learning methods. **training** **training set** **features** **predictive model**

Example 04.06-1 combat boots

Consider a robot, Pierre, with a particular gravitas and sense of style. He seeks just the right-looking pair of combat boots for wearing in the autumn rains. Pierre is to purchase the boots online via image recognition, and decides to gather data by visiting a hipster hangout one evening to train his style. For contrast, he also watches footage of a White Nationalist rally, focusing special attention on the boots of wearers of khakis and polos. Comment on Pierre's methods.

Lecture 04.07 Random variables

Probabilities are useful even when they do not deal strictly with events. It often occurs that we measure something that has randomness associated with it. We use random variables to represent these measurements.

random variable

A *random variable* $X : \Omega \to \mathbb{R}$ is a function that maps an outcome ω from the sample space Ω to a real number $x \in \mathbb{R}$. A random variable will be denoted with a capital letter (e.g. X and K) and a specific value that it maps to (the *image*) will be denoted with a lowercase letter (e.g. x and k).

discrete random variable
continuous random variable

A *discrete random variable* K is one that takes on discrete values. A *continuous random variable* X is one that takes on continuous values.

Example 04.07-1 dice again

Roll two unbiased dice. Let K be a random variable representing the sum of the two. Let $P(k)$ be the probability of the result $k \in K$. Plot and interpret $P(k)$.

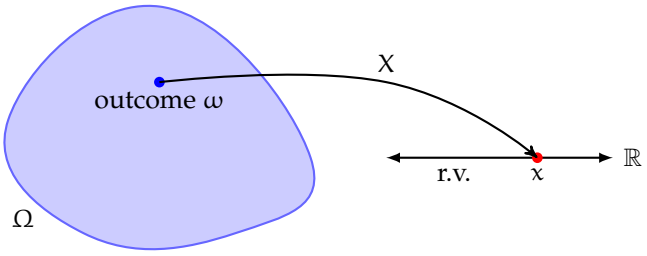


Figure 04.2: a random variable X maps an outcome $\omega \in \Omega$ to a $x \in \mathbb{R}$.

Example 04.07-2 Johnson-Nyquist noise

A resistor at nonzero temperature without any applied voltage exhibits an interesting phenomenon: its voltage randomly fluctuates. This is called *Johnson-Nyquist noise* and is a result of *thermal excitation* of charge carriers (electrons, typically). For a given resistor and measurement system, let the *probability density function* f_V of the voltage V across an unrealistically hot resistor be

$$f_V(V) = \frac{1}{\sqrt{\pi}} e^{-V^2}.$$

Plot and interpret the meaning of this function.

Lecture 04.08 Probability density and mass functions

frequency distribution
probability mass function

Consider an experiment that measures a random variable. We can plot the relative frequency of the measurand landing in different "bins" (ranges of values). This is called a *frequency distribution* or a *probability mass function* (PMF).

Consider, for instance, a probability mass function as plotted in Figure 04.3, where a frequency a_i can be interpreted as an estimate of the probability of the measurand being in the ith interval. The sum of the frequencies must be unity:

with k being the number of bins.

frequency density distribution
probability density function

The *frequency density distribution* is similar to the frequency distribution, but with $a_i \mapsto a_i/\Delta x$, where Δx is the bin width.

If we let the bin width approach zero, we derive the *probability density function* (PDF)

$$f(x) = \lim_{\substack{k\to\infty \\ \Delta x\to 0}} \sum_{j=1}^{k} a_j/\Delta x. \tag{04.3}$$

We typically think of a probability density function f, like the one in Figure 04.4 as a function that can be integrated over to find the probability of the random variable (measurand) being in an interval $[a, b]$:

$$P(x \in [a, b]) = \int_a^b f(\chi)\, d\chi. \tag{04.4}$$

Of course,

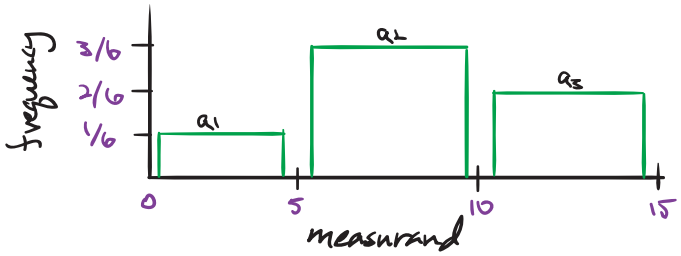


Figure 04.3: plot of a probability mass function.

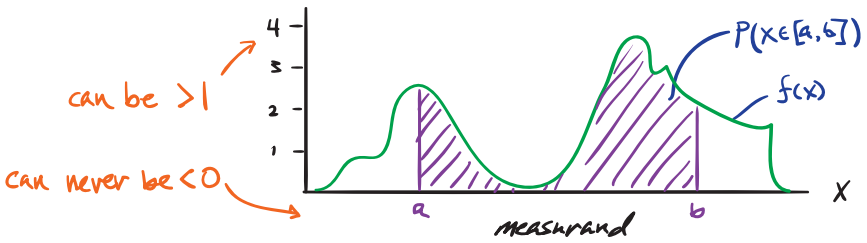


Figure 04.4: plot of a probability density function.

We now consider a common PMF and a common PDF.

04.08.1 Binomial PMF

Consider a random binary sequence of length n such that each element is a random 0 or 1, generated independently, like

$$(1, 0, 1, 1, 0, \cdots, 1, 1). \tag{04.5}$$

Let events $\{1\}$ and $\{0\}$ be mutually exclusive and exhaustive and $P(\{1\}) = p$. The probability of the sequence above occurring is

There are n choose k,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{04.6}$$

possible combinations of k ones for n bits. Therefore, the probability of any combination of k ones in a series is

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{04.7}$$

We call Equation 04.7 the *binomial distribution PDF*.

binomial distribution PDF

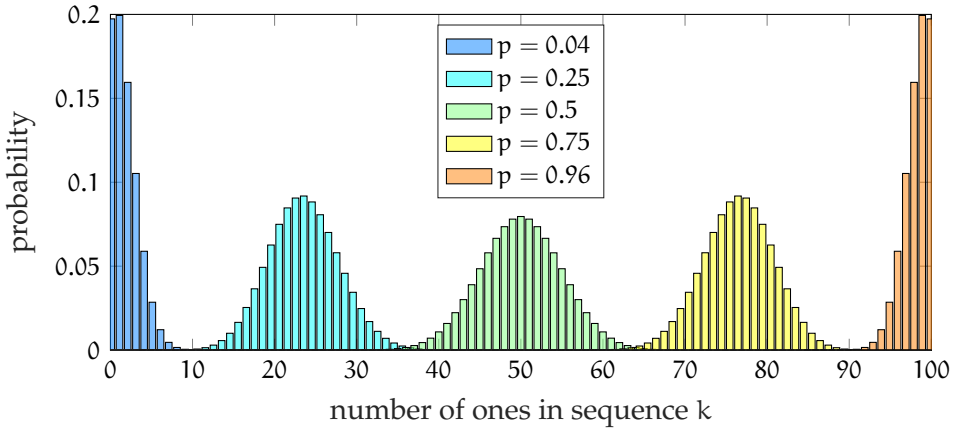


Figure 04.5: binomial PDF for $n = 100$ measurements and different values of $P(\{1\}) = p$, the probability of a measurement error. The plot is generated by the *Matlab* code of Figure 04.6.

Example 04.08-1 Binomial PMF

Consider a field sensor that fails for a given measurement with probability p. Given n measurements, plot the binomial PMF as a function of k failed measurements for a few different probabilities of failure $p \in [0.04, 0.2, 0.5]$.

Figure 04.6 shows *Matlab* code for constructing the PDFs plotted in Figure 04.5. Note that the symmetry is due to the fact that events $\{1\}$ and $\{0\}$ are mutually exclusive and exhaustive.

04.08.2 Gaussian PDF

Gaussian or normal random variable

The *Gaussian* or *normal random variable* x has PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}. \tag{04.8}$$

Although we're not quite ready to understand these quantities in detail, it can be shown that the parameters μ and σ have the following meanings:

mean
standard deviation
variance

- μ is the *mean* of x,
- σ is the *standard deviation* of x, and
- σ^2 is the *variance* of x.

```
%% parameters
n = 100;
k_a = linspace(1,n,n);
p_a = [.04,.25,.5,.75,.96];

%% binomial function
f = @(n,k,p) nchoosek(n,k)*p^k*(1-p)^(n-k);

% loop through to construct an array
f_a = NaN*ones(length(k_a),length(p_a));
for i = 1:length(k_a)
    for j = 1:length(p_a)
        f_a(i,j) = f(n,k_a(i),p_a(j));
    end
end

%% plot
figure
colors = jet(length(p_a));
for j = 1:length(p_a)
    bar(...
        k_a,f_a(:,j),...
        'facecolor',colors(j,:),...
        'facealpha',0.5,...
        'displayname', ['$p = ',num2str(p_a(j)),'$']...
    ); hold on
end
leg = legend('show','location','north');
set(leg,'interpreter','latex')
hold off
xlabel('number of ones in sequence k')
ylabel('probability')
xlim([0,100])
```

Figure 04.6: a *Matlab* script for generating binomial PMFs.

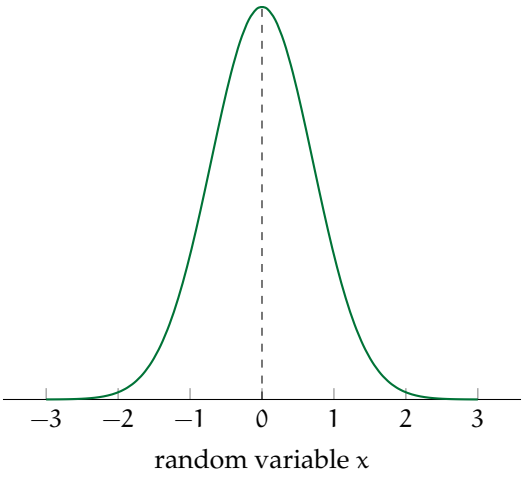


Figure 04.7: PDF for Gaussian random variable x, mean $\mu = 0$, and standard deviation $\sigma = 1/\sqrt{2}$.

Consider the "bell-shaped" Gaussian PDF in Figure 04.7. It is always symmetric. The mean μ is its central value and the standard deviation σ is directly related to its width. We will continue to explore the Gaussian distribution in the following lectures, especially in Lecture 04.12.

Lecture 04.09 Expectation

Recall that a random variable is a function $X : \Omega \to \mathbb{R}$ that maps from the sample space to the reals. Random variables are the arguments of probability mass functions (PMFs) and probability density functions (PDFs).

The *expected value* (or *expectation*) of a random variable is akin to its "average value" and depends on its PMF or PDF. The expected value of a random variable X is denoted $\langle x \rangle$ or $E(X)$. There are two definitions of the expectation, one for a discrete random variable, the other for a continuous random variable. Before we define, them, however, it is useful to predefine the most fundamental property of a random variable, its *mean*.

expected value
expectation

mean

Definition 04.09.1: mean

The *mean* of a random variable X is defined as

$$m_X = E(X).$$

Let's begin with a discrete random variable.

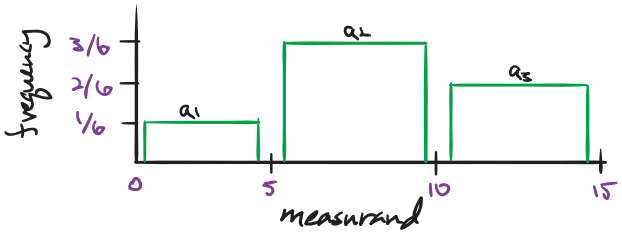
Definition 04.09.2: expectation of a discrete random variable

Let K be a discrete random variable and f its PMF. The *expected value* of K is defined as

$$E(K) = \sum_{\forall k} k f(k).$$

Example 04.09-1 expectation of a discrete random variable

Given a discrete random variable K with PMF shown below, what is its mean μ_K?



Let us now turn to the expectation of a continuous random variable.

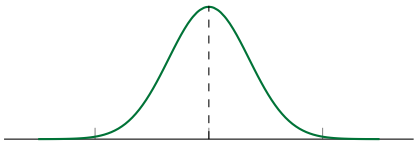
Definition 04.09.3: expectation of a continuous random variable

Let X be a continuous random variable and f its PDF. The *expected value* of X is defined as

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

Example 04.09-2 expectation of a continuous random variable

Given a continuous random variable X with Gaussian PDF f, what is the expected value of X?



random variable x

Due to its sum or integral form, the expected value $\mathsf{E}(\cdot)$ has some familiar properties for random variables X and Y and reals a and b.

$$\mathsf{E}(a) = a \tag{04.9a}$$
$$\mathsf{E}(X + a) = \mathsf{E}(X) + a \tag{04.9b}$$
$$\mathsf{E}(aX) = a\mathsf{E}(X) \tag{04.9c}$$
$$\mathsf{E}(\mathsf{E}(X)) = \mathsf{E}(X) \tag{04.9d}$$
$$\mathsf{E}(aX + bY) = a\mathsf{E}(X) + b\mathsf{E}(Y). \tag{04.9e}$$

Lecture 04.10 Central moments

central moments

Given a probability mass function (PMF) or probability density function (PDF) of a random variable, several useful parameters of the random variable can be computed. These are called *central moments*, which quantify parameters relative to its mean.

Definition 04.10.1: central moments

The nth central moment of random variable X, with PDF f, is defined as

$$E((X-\mu_X)^n) = \int_{-\infty}^{\infty} (x-\mu_X)^n f(x)\,dx.$$

For discrete random variable K with PMF f,

$$E((K-\mu_K)^n) = \sum_{\forall k}^{\infty} (k-\mu_K)^n f(k).$$

Example 04.10-1 first moment

Prove that the first moment of random variable X is zero.

variance

The second central moment of random variable X is called the *variance* and is denoted

$$\sigma_X^2 \quad \text{or} \quad \text{Var}(X) \quad \text{or} \quad E((X-\mu_X)^2). \tag{04.10}$$

The variance is a measure of the *width* or *spread* of the PMF or PDF. We usually compute the variance with the formula

Other properties of variance include, for real constant c,

$$\operatorname{Var}(c) = 0 \tag{04.11}$$
$$\operatorname{Var}(X + c) = \operatorname{Var}(X) \tag{04.12}$$
$$\operatorname{Var}(cX) = c^2 \operatorname{Var}(X). \tag{04.13}$$

The *standard deviation* is defined as **standard deviation**

Although the variance is mathematically more convenient, the standard deviation has the same physical units as X, so it is often the more physically meaningful quantity. Due to its meaning as the width or spread of the probability distribution, and its sharing of physical units, it is a convenient choice for error bars on plots of a random variable.

The *skewness* $\operatorname{Skew}(X)$ is a normalized third central moment: **skewness**

$$\operatorname{Skew}(X) = \frac{E((X - \mu_X)^3)}{\sigma_X^3}. \tag{04.14}$$

Skewness is a measure of *asymmetry* of a random variable's PDF or PMF. For a symmetric PMF or PDF, such as the Gaussian PDF, $\operatorname{Skew}(X) = 0$. **asymmetry**

The *kurtosis* $\operatorname{Kurt}(X)$ is a normalized fourth central moment: **kurtosis**

$$\operatorname{Kurt}(X) = \frac{E((X - \mu_X)^4)}{\sigma_X^4}. \tag{04.15}$$

Skewness is a measure of the *tailedness* of a random variable's PDF or PMF. "Heavier" tails yield higher kurtosis. **tailedness**

A Gaussian random variable has PDF with kurtosis 3. Given that for Gaussians both skewness and kurtosis have nice values (0 and 3), we can think of skewness and and kurtosis as measures of similarity to the Gaussian PDF.

Lecture 04.11 Estimation of sample mean and variance

Ahem.[3]

04.11.0.1 Estimation and sample statistics

The mean and variance definitions, above, apply only to a random variable for which we have a theoretical probability distribution. Typically, it is not until after having performed many measurements of a random variable that we can assign a good distribution model. Until then, measurements can help us *estimate* aspects of the data. We usually start by estimating basic parameters such as *mean* and *variance* before estimating a probability distribution.

There are two key aspects to randomness in the measurement of a random variable. First, of course, there is the underlying randomness with its probability distribution, mean, standard deviation, etc., which we call the *population statistics*. Second, there is the *statistical variability* that is due to the fact that we are *estimating* the random variable's statistics—called its *sample statistics*—from some sample. Statistical variability are decreased with greater sample size and number of samples, whereas the underlying randomness of the random variable does not decrease. Instead, our estimates of its probability distribution and statistics improve.

04.11.0.2 Sample mean, variance, and standard deviation

The *arithmetic mean* or *sample mean* of a measurand with sample size N, represented by random variable X, is defined as

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{04.16}$$

If the sample size is large, $\overline{x} \to m_X$ (the sample mean approaches the mean). The *population mean* is another term for the mean μ_X, which is equal to

[3]The source for this exercise lecture is in a *Matlab* kernel Jupyter notebook. For more information, see jupyter.org. See ricopic.one/measurement/notebooks for the source code notebook. Note, however, that running the *Matlab* code in the usual m-file environment is much easier.

$$m_X = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{04.17}$$

Recall that the *definition* of the mean is $m_X = E(x)$.

The *sample variance* of a measurand represented by random variable X is defined as

$$S_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2. \tag{04.18}$$

If the sample size is large, $S_X^2 \to \sigma_X^2$ (the sample variance approaches the variance). The *population variance* is another term for the variance σ_X^2, and can be expressed as

$$\sigma_X^2 = \lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2. \tag{04.19}$$

Recall that the *definition* of the variance is $\sigma_X^2 = E((X - m_X)^2)$.

The *sample standard deviation* of a measurand represented by random variable X is defined as

$$S_X = \sqrt{S_X^2}. \tag{04.20}$$

If the sample size is large, $S_X \to \sigma_X$ (the sample standard deviation approaches the standard deviation). The *population standard deviation* is another term for the standard deviation σ_X, and can be expressed as

$$\sigma_X = \lim_{N\to\infty} \sqrt{S_X^2}. \tag{04.21}$$

Recall that the *definition* of the standard deviation is $\sigma_X = \sqrt{\sigma_X^2}$.

04.11.0.3 Sample statistics as random variables

There is an ambiguity in our usage of the term "sample." It can mean just one measurement or it can mean a collection of measurements gathered together. Hopefully, it is clear from context.

In the latter sense, often we collect multiple samples, each of which has its own sample mean $\overline{X}_i$ and standard deviation S_{X_i}. In this situation, $\overline{X}_i$ and S_{X_i} are themselves random variables (meta af, I know). This means they have their own sample means $\overline{\overline{X}_i}$ and $\overline{S_{X_i}}$ and standard deviations $S_{\overline{X}_i}$ and $S_{S_{X_i}}$.

The mean of means $\overline{\overline{X}_i}$ is equivalent to a mean with a larger sample size and is therefore our best estimate of the mean of the underlying random process. The mean of standard deviations $\overline{S_{X_i}}$ is our best estimate of the standard deviation of the underlying random process. The standard deviation of means $S_{\overline{X}_i}$ is a measure of the spread in our estimates of the mean. It is our best estimate of the standard deviation of the statistical variation and should therefore tend to zero as sample size and number of samples increases. The standard deviation of standard deviations $S_{S_{X_i}}$ is a measure of the spread in our estimates of the standard deviation of the underlying process. It should also tend to zero as sample size and number of samples increases.

Let N be the size of each sample. It can be shown that the standard deviation of the means $S_{\overline{X}_i}$ can be estimated from a single sample standard deviation:

$$S_{\overline{X}_i} \approx \frac{S_{X_i}}{\sqrt{N}}. \tag{04.22}$$

This shows that as the sample size N increases, the statistical variability of the mean decreases (and in the limit approaches zero).

04.11.0.4 Nonstationary signal statistics

The sample mean, variance, and standard deviation definitions, above, assume the random process is *stationary*—that is, its population mean does not vary with time. However, a great many measurement signals have populations that *do* vary with time, i.e. they are *nonstationary*. Sometimes the nonstationarity arises from a "drift" in the dc value of a signal or some other slowly changing variable. But dynamic signals can also change in a recognizable and predictable manner, as when, say, the temperature of a room changes when a window is opened or when a water level changes with the tide.

Typically, we would like to minimize the effect of nonstationarity on the signal statistics. In certain cases, such as drift, the variation is a nuissance only, but other times it is the point of the measurement.

Two common techniques are used, depending on the overall type of nonstationarity. If it is periodic with some known or estimated period, the measurement data series can be "folded" or "reshaped" such that the ith measurement of each period corresponds to the ith measurement of all other periods. In this case, somewhat counterintuitively, we can consider the ith measurements to correspond to a sample of size N, where N is the number of periods over which measurements are made.

When the signal is aperiodic, we often simply divide it into "small" (relative to its overall trend) intervals over which statistics are computed, separately.

Note that in this discussion, we have assumed that the nonstationarity of the signal is due to a variable that is deterministic (not random).

04.11.0.5 Example: measurement of Gaussian noise on nonstationary signal

Consider the measurement of the temperature inside a desktop computer chassis via an inexpensive *thermistor*, a resistor that changes resistance with temperature. The processor and power supply heat the chassis in a manner that depends on processing demand. For the test protocol, the processors are cycled sinusoidally through processing power levels at a frequency of 50 mHz for $n_T = 12$ periods and sampled at 1 Hz. Assume a temperature fluctuation between about 20 and 50 C and Gaussian noise with standard deviation 4 C. Consider a *sample* to be the multiple measurements of a certain instant in the period.

1. Generate and plot simulated temperature data as a time series and as a histogram or frequency distribution. Comment on why the frequency distribution sucks.
2. Compute the sample mean and standard deviation *for each sample in the cycle*.
3. Subtract the mean from each sample in the period such that each sample distribution is centered at zero. Plot the composite frequency distribution of all samples, together. This represents our best estimate of the frequency distribution of the underlying process.
4. Plot a comparison of the theoretical mean, which is 35, and the sample mean of means with an error bar. Vary the number of samples n_T and comment on its effect on the estimate.
5. Plot a comparison of the theoretical standard deviation and the sample mean of sample standard deviations with an error bar. Vary the number of samples n_T and comment on its effect on the estimate.

6. Plot the sample means over a single period with error bars of $\pm$ one sample standard deviation of the means. This represents our best estimate of the sinusoidal heating temperature. Vary the number of samples n_T and comment on the estimate.

```
clear; close all; % clear kernel
```

Generate the temperature data The temperature data can be generated by constructing an array that is passed to a sinusoid, then "randomized" by Gaussian random numbers. Note that we add `1` to `np` and `n` to avoid the sneaky fencepost error.

```
f = 50e-3; % Hz ... sinusoid frequency
a = 15; % C ... amplitude of oscillation
dc = 35; % C ... dc offset of oscillation
fs = 1; % Hz ... sampling frequency
nT = 12; % number of sinusoid periods
s = 4; % C ... standard deviation
np = fs/f+1; % number of samples per period
n = nT*np+1; % total number of samples

t_a = linspace(0,nT/f,n); % time array
sin_a = dc + a*sin(2*pi*f*t_a); % sinusoidal array
rng(43); % seed the random number generator
noise_a = s*randn(size(t_a)); % Gaussian noise
signal_a = sin_a + noise_a; % sinusoid + noise
```

Now that we have an array of "data," we're ready to plot.

```
h = figure;
p = plot(t_a,signal_a,'o-',...
    'Color',[.8,.8,.8],...
    'MarkerFaceColor','b',...
    'MarkerEdgeColor','none',...
    'MarkerSize',3);
xlabel('time (s)');
ylabel('temperature (C)');
hgsave(h,'figures/temp');
```

This is something like what we might see for continuous measurement data. Now, the histogram.

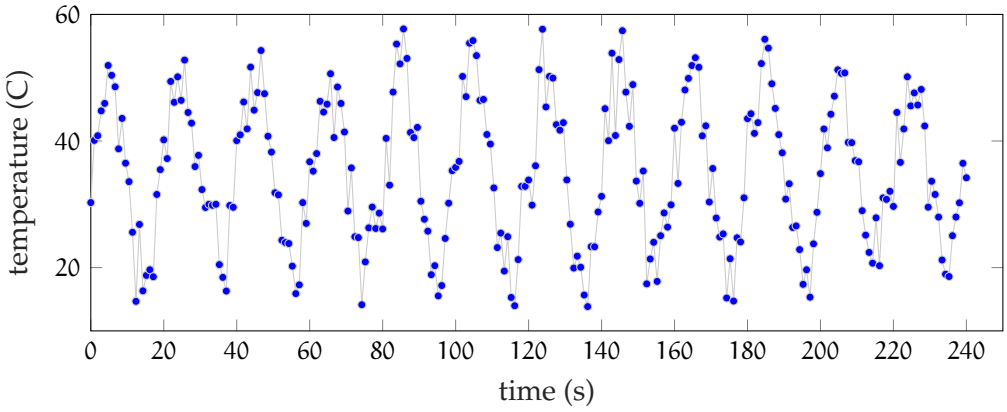


Figure 04.8: temperature over time

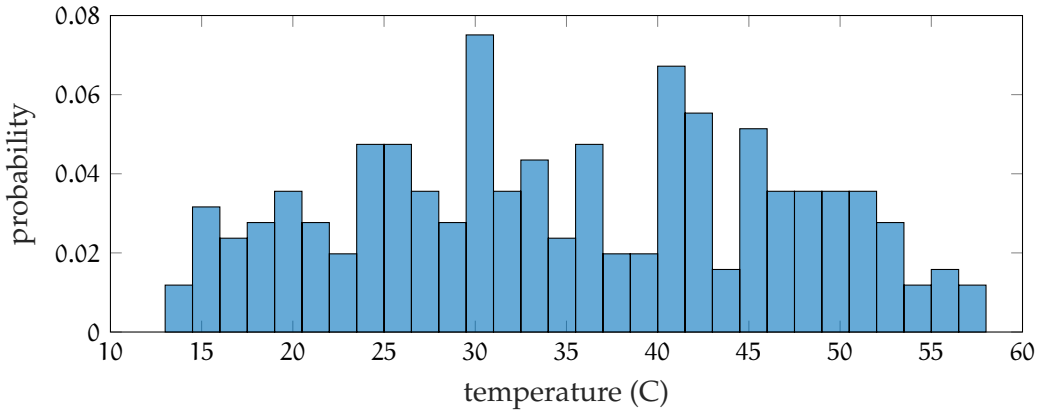


Figure 04.9: a poor histogram due to unstationarity of the signal.

```
h = figure;
histogram(signal_a,...
    30, ... % number of bins
    'normalization','probability'... % for PMF
);
xlabel('temperature (C)')
ylabel('probability')
hgsave(h,'figures/temp');
```

This sucks because we plot a frequency distribution to tell us about the random variation, but this data includes the sinusoid.

Sample mean, variance, and standard deviation To compute the sample mean μ and standard deviation s for each sample in the period, we must "pick out" the `nT` data points that correspond to each other. Currently, they're in one long $1 \times n$ array `signal_a`. It is helpful to *reshape* the data so it is in an $nT \times np$ array, which each row corresponding to a new period. This leaves the correct points aligned in columns. It is important to note that we can do this "folding" operation only when we know rather precisely the period of the underlying sinusoid. It is given in the problem that it is a controlled experiment variable. If we did not know it, we would have to estimate it, too, from the data.

```
signal_ar = reshape(signal_a(1:end-1)',[np,nT])'; % reshape
size(signal_ar) % check size
signal_ar(1:3,1:4) % print first three rows of first four columns
```

```
ans =

    12    21


ans =

   30.2718   40.0946   40.8341   44.7662
   40.1836   37.2245   49.4076   46.1137
   40.0571   40.9718   46.1627   41.9145
```

Define the mean, variance, and standard deviation functions as "anonmymous functions." We "roll our own." These are *not* as efficient or flexible as the built-in *Matlab* functions `mean`, `var`, and `std`, which should typically be used.

```
my_mean = @(vec) sum(vec)/length(vec);
my_var = @(vec) sum((vec-my_mean(vec)).^2)/(length(vec)-1);
my_std = @(vec) sqrt(my_var(vec));
```

Now the sample mean, variance, and standard deviations can be computed. We proceed by looping through each column of the reshaped signal array.

```
mu_a = NaN*ones(1,np); % initialize mean array
var_a = NaN*ones(1,np); % initialize var array
```

```
s_a = NaN*ones(1,np); % initialize std array

for i = 1:np % for each column
    mu_a(i) = my_mean(signal_ar(:,i));
    var_a(i) = my_var(signal_ar(:,i));
    s_a(i) = sqrt(var_a(i)); % touch of speed
end
```

Composite frequency distribution The columns represent samples. We want to subtract the mean from each column. We use `repmat` to reproduce `mu_a` in `nT` rows so it can be easily subtracted.

```
signal_arz = signal_ar - repmat(mu_a,[nT,1]);
size(signal_arz) % check size
signal_arz(1:3,1:4) % print first three rows of first four columns
```

```
ans =

    12    21


ans =

   -5.0881    0.9525   -0.2909   -1.5700
    4.8237   -1.9176    8.2826   -0.2225
    4.6972    1.8297    5.0377   -4.4216
```

Now that all samples have the same mean, we can lump them into one big bin for the frequency distribution. There are some nice built-in functions to do a quick reshape and fit.

```
% resize
signal_arzr = reshape(signal_arz,[1,nT*np]);
size(signal_arzr) % check size
% fit
pdfit_model = fitdist(signal_arzr','normal'); % do a fit
x_a = linspace(-15,15,100);
pdfit_a = pdf(pdfit_model,x_a);
pdf_a = normpdf(x_a,0,s); % theoretical pdf
```

```
ans =

     1   252
```

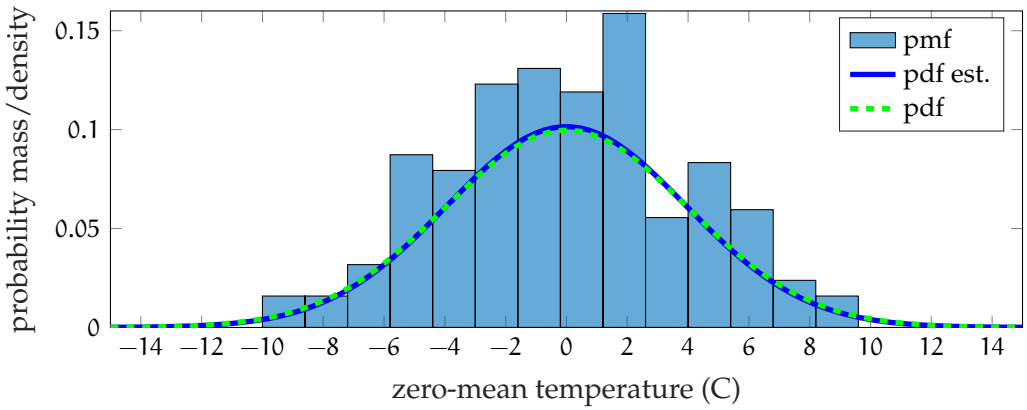


Figure 04.10: PMF and estimated and theoretical PDFs.

Plot!

```
h = figure;
histogram(signal_arzr,...
    round(s*sqrt(nT)), ... % number of bins
    'normalization','probability'... % for PMF
);
hold on
plot(x_a,pdfit_a,'b-','linewidth',2); hold on
plot(x_a,pdf_a,'g--','linewidth',2);
legend('pmf','pdf est.','pdf')
xlabel('zero-mean temperature (C)')
ylabel('probability mass/density')
hgsave(h,'figures/temp');
```

Means comparison The sample mean of means is simply the following.

```
mu_mu = my_mean(mu_a)
```

```
mu_mu =

   35.1175
```

The standard deviation that works as an error bar, which should reflect how well we can estimate the point plotted, is the standard deviation of the

means. It is difficult to compute this directly for a nonstationary process. We use the estimate given above and improve upon it by using the mean of standard deviations instead of a single sample's standard deviation.

```
s_mu = mean(s_a)/sqrt(nT)
```

```
s_mu =

    1.1580
```

Now, for the simple plot.

```
h = figure;
bar(mu_mu); hold on % gives bar
errorbar(mu_mu,s_mu,'r','linewidth',2) % gives error bar
ax = gca; % current axis
ax.XTickLabels = {'$\overline{\overline{X}}$'};
ax.TickLabelInterpreter = 'latex';
hgsave(h,'figures/temp');
```

Standard deviations comparison The sample mean of standard deviations is simply the following.

```
mu_s = my_mean(s_a)
```

```
mu_s =

    4.0114
```

The standard deviation that works as an error bar, which should reflect how well we can estimate the point plotted, is the standard deviation of the standard deviations. We can compute this directly.

```
s_s = my_std(s_a)
```

```
s_s =

    0.8495
```

Now, for the simple plot.

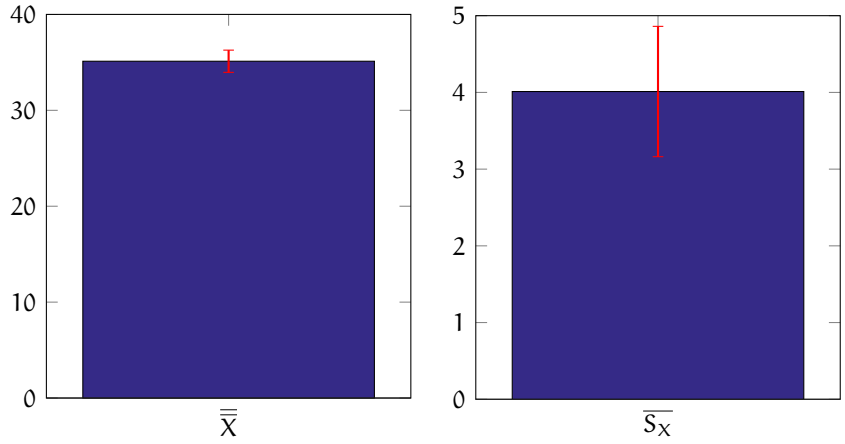


Figure 04.11: (left) sample mean of sample means and (right) sample standard deviation of sample means.

```
h = figure;
bar(mu_s); hold on % gives bar
errorbar(mu_s,s_s,'r','linewidth',2) % gives error bar
ax = gca; % current axis
ax.XTickLabels = {'$\overline{S_X}$'};
ax.TickLabelInterpreter = 'latex';
hgsave(h,'figures/temp');
```

Plot a period with error bars Plotting the data with error bars is fairly straightforward with the built-in `errorbar` function. The main question is "which standard deviation?" Since we're plotting the means, it makes sense to plot the error bars as a single sample standard deviation of the means.

```
h = figure;
e1 = errorbar(t_a(1:np),mu_a,s_mu*ones(1,np),'b'); hold on
t_a2 = linspace(0,1/f,101);
e2 = plot(t_a2,dc + a*sin(2*pi*f*t_a2),'r-');
xlim([t_a(1),t_a(np)])
grid on
xlabel('folded time (s)')
ylabel('temperature (C)')
```

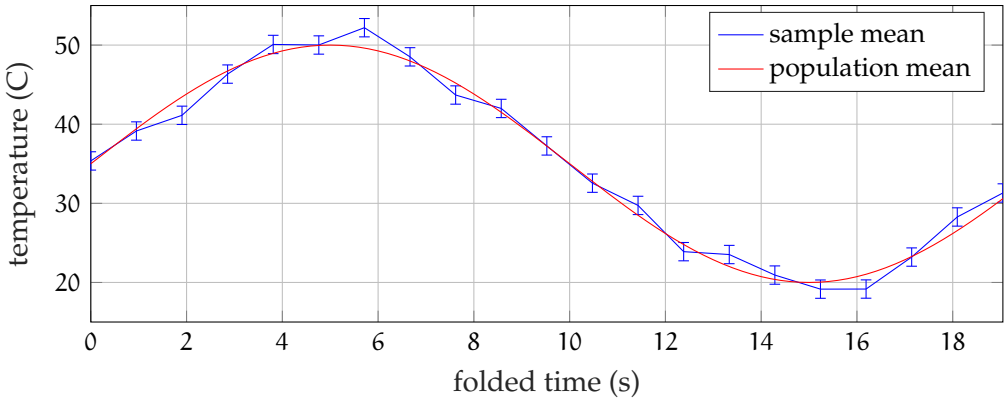


Figure 04.12: sample means over a period.

```
legend([e1 e2],'sample mean','population mean','Location','NorthEast')
hgsave(h,'figures/temp');
```

Lecture 04.12 Confidence

Ahem.[4]

One must have it to give a lecture named it. Confidence is used in the common sense, although we do endow it with a mathematical definition to scare business majors, who aren't actually impressed, but indifferent. Approximately: if, under some reasonable assumptions (probabilistic model), we estimate the probability of some event to be P%, we say we have P% confidence in it. I mean, business majors are all, "Supply and demand? Let's call that a 'law,'" so I think we're even.

So we're back to computing probability from distributions—probability density functions (PDFs) and probability mass functions (PMFs). Usually we care most about estimating the mean of our distribution. Recall from the previous lecture that when several samples are taken, each with its own mean, the mean is itself a random variable—with a mean, of course. Meanception.

But, more importantly (just kidding—equally so), the mean has a probability distribution of its own. The *central limit theorem* has as one of its implications that, as the sample size N gets large, *regardless of the sample distributions*, *this distribution of means approaches the Gaussian distribution*.

But sometimes I always worry I'm being lied to, so let's check.

```
clear; close all; % clear kernel
```

04.12.1 Generate some data to test the central limit theorem

Data can be generated by constructing an array using a (seeded for consistency) random number generator. Let's use a uniformly distributed PDF between 0 and 1.

```
N = 150; % sample size (number of measurements per sample)
M = 120; % number of samples
n = N*M; % total number of measurements
mu_pop = 0.5; % because it's a uniform PDF between 0 and 1

rng(11); % seed the random number generator
```

[4]The source for this exercise lecture is in a *Matlab* kernel *Jupyter* notebook. For more information, see jupyter.org. See ricopic.one/measurement/code for the source code in both *Jupyter* notebook and *Matlab* m-file form. Note, however, that simply running the *Matlab* code in the usual m-file environment is easier to get started.

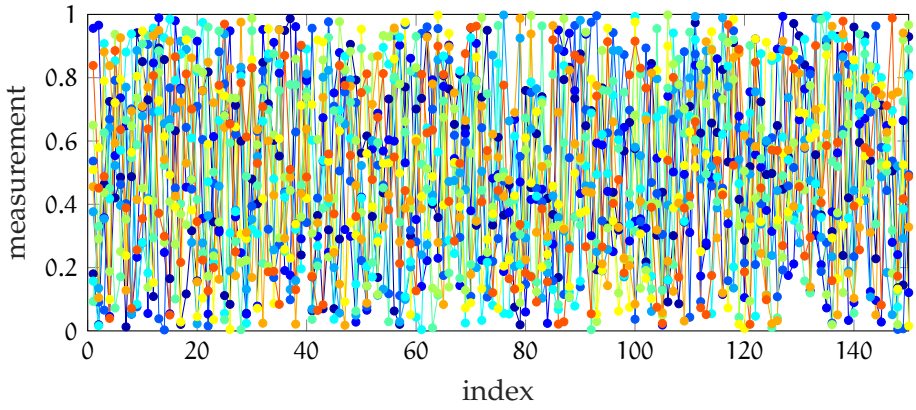


Figure 04.13: raw data with colors corresponding to samples.

```
signal_a = rand(N,M); % uniform PDF
size(signal_a) % check the size
```

```
ans =

   150   120
```

Let's take a look at the data by plotting the first ten samples (columns), as shown in Figure 04.13.

This is something like what we might see for continuous measurement data. Now, the histogram.

This isn't a great plot, but it shows roughly that each sample is fairly uniformly distributed.

04.12.2 Sample statistics

Now let's check out the sample statistics. We want the sample mean and standard deviation of each column. Let's use the built-in functions `mean` and `std`.

```
mu_a = mean(signal_a,1); % mean of each column
s_a = std(signal_a,1); % std of each column
```

Now we can compute the mean statistics, both the mean of the mean $\overline{\overline{X}}$ and the standard deviation of the mean $s_{\overline{X}}$, which we don't strictly need for

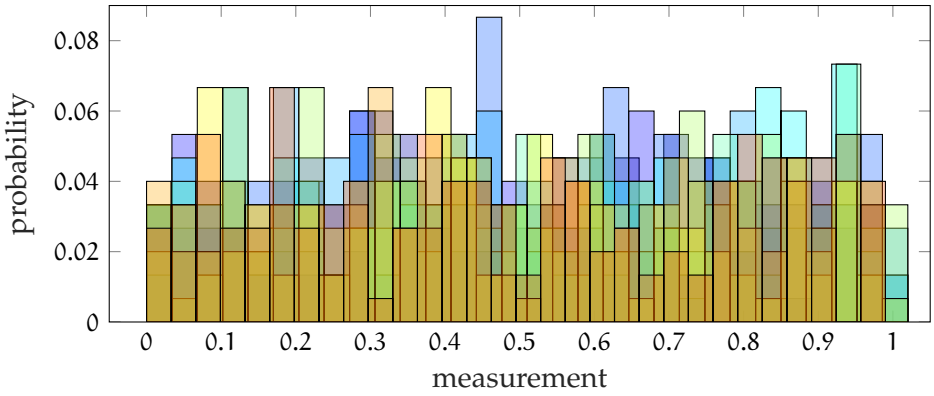


Figure 04.14: a histogram showing the approximately uniform distribution of each sample (color).

this part, but we're curious. We choose to use the direct estimate instead of the $s_X/\sqrt{N}$ formula, but they should be close.

```
mu_mu = mean(mu_a)
s_mu = std(mu_a)
```

```
mu_mu =

    0.4987


s_mu =

    0.0236
```

04.12.3 The truth about sample means

It's the moment of truth. Let's look at the distribution.

This looks like a Gaussian distribution about the mean of means, so I guess the central limit theorem is legit.

04.12.4 Gaussian and probability

We already know how to compute the probability P a value of a random variable X lies in a certain interval from a PMF or PDF (the sum or the integral, respectively). This means that, for sufficiently large sample size

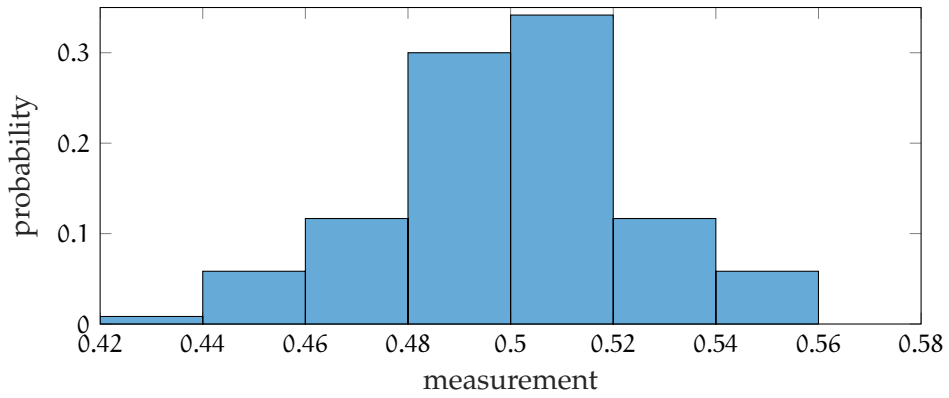


Figure 04.15: a histogram showing the approximately normal distribution of the means.

N such that we can assume from the central limit theorem that the sample means $\overline{x_i}$ are normally distributed, *the probability a sample mean value* $\overline{x_i}$ *is in a certain interval* is given by integrating the Gaussian PDF. The Gaussian PDF for random variable Y representing the sample means is

where μ is the population mean and σ is the population standard deviation.

The integral of f over some interval is the probability a value will be in that interval. Unfortunately, that integral is uncool. It gives rise to the definition of the *error function*, which, for the Gaussian random variable Y, is

This expresses the probability a sample mean being in the interval $[-y_b, y_b]$ if Y has mean 0 and variance $1/2$.

Matlab has this built-in as `erf`, shown in Figure 04.16.

We could deal directly with the error function, but most people don't and we're weird enough, as it is. Instead, people use the *Gaussian cumulative distribution function* $\Phi : \mathbb{R} \to \mathbb{R}$, which is defined as

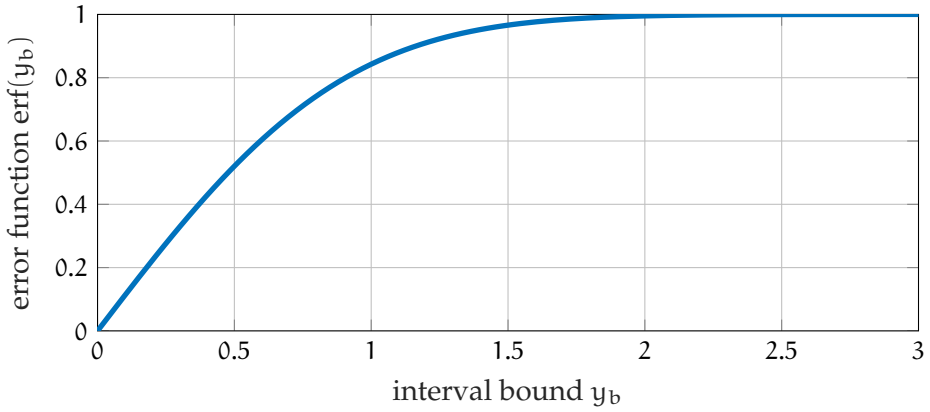


Figure 04.16: the error function.

and which expresses the probability of a Gaussian random variable Z with mean 0 and standard deviation 1 taking on a value in the interval $(-\infty, z]$.

That's great and all, but occasionally we have Gaussian random variables with nonzero means and nonunity standard deviations. It turns out we can shift any Gaussian random variable by its mean and scale it by its standard deviation to make it have zero mean and standard deviation. We can then use Φ and interpret the results as being relative to the mean and standard deviation, using phrases like "the probability it is within two standard deviations of its mean." The transformed random variable Z and its values z are sometimes called the *z-score*. For a particular value x of a random variable X, we can compute its z-score (or value z of random variable Z) with the formula

and compute the probability of X taking on a value within the interval, say, $x \in [x_{b-}, x_{b+}]$ from the table. (Sample statistics $\overline{X}$ and S_X are appropriate when population statistics are unknown.)

For instance, compute the probability a Gaussian random variable X with $\mu_X = 5$ and $\sigma_X = 2.34$ takes on a value within the interval $x \in [3, 6]$.

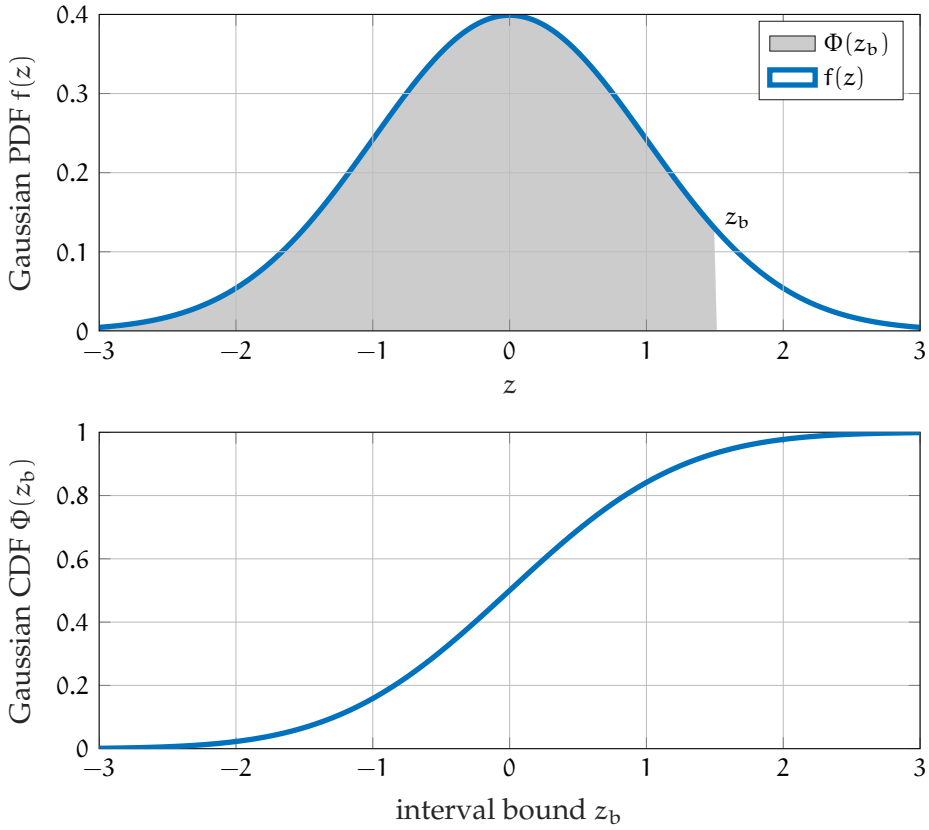


Figure 04.17: the Gaussian PDF and CDF for z-scores.

1. Compute the z-score of each endpoint of the interval:

2. Look up the CDF values for z_3 and z_6, which are $\Phi(z_3) = 0.1977$ and $\Phi(z_6) = 0.6664$.
3. The CDF values correspond to the probabilities $x < 3$ and $x < 6$. Therefore, to find the probability x lies in that interval, we subtract the lower bound probability:

So there is a 46.89% probability, and therefore we have 46.89% confidence, that $x \in [3, 6]$.

Often we want to go the other way, estimating the symmetric interval $[x_{b-}, x_{b+}]$ for which there is a given probability. In this case, we first look up the z-score corresponding to a certain probability. For concreteness, given the same population statistics above, let's find the symmetric interval $[x_{b-}, x_{b+}]$ over which we have 90% confidence. From the table, we want two, symmetric z-scores that have CDF-value difference 0.9. Or, in maths,

Due to the latter relation and the additional fact that the Gaussian CDF has antisymmetry,

Adding the two Φ equations,

and $\Phi(z_{b-}) = 0.05$. From the table, these correspond (with a linear interpolation) to $z_b = z_{b+} = -z_{b-} \approx 1.645$. All that remains is to solve the z-score formula for x:

From this,

and X has a 90% confidence interval $[1.151, 8.849]$.

04.12.5 Example: Gaussian confidence for a mean

Consider the data set generated above. What is our 95% confidence interval in our estimate of the mean?

Assuming we have a sufficiently large data set, the distribution of means is approximately Gaussian. Following the same logic as above, we need z-score that gives an upper CDF value of . From the table, we obtain the $z_b = z_{b+} = -z_{b-}$, below.

```
z_b = 1.96;
```

Now we can estimate the mean using our sample and mean statistics,

```
mu_x_95 = mu_mu + [-z_b,z_b]*s_mu
```

```
mu_x_95 =

    0.4526    0.5449
```

This is our 95% confidence interval in our estimate of the mean.

Lecture 04.13 Student confidence

Student's t-distribution

The central limit theorem tells us that, for large sample size N, the distribution of the means is Gaussian. However, for small sample size, the Gaussian isn't as good of an estimate. *Student's t-distribution* is superior for lower sample size and equivalent at higher sample size. Technically, if the population standard deviation σ_X is known, even for low sample size we should use the Gaussian distribution. However, this rarely arises in practice, so we can usually get away with an "always t" approach.

degrees of freedom

A way that the t-distribution accounts for low-N is by having an entirely different distribution for each N (seems a bit of a cheat, to me). Actually, instead of N, it uses the *degrees of freedom* ν, which is N minus the number of parameters required to compute the statistic. Since the standard deviation requires only the mean, for most of our cases, $\nu = N - 1$.

As with the Gaussian distribution, the t-distribution's integral is difficult to calculate. Typically, we will use a t-table, such as the one given here. There are three points of note.

1. Since we are primarily concerned with going from probability/confidence values (e.g. P% probability/confidence) to intervals, typically there is a column for each probability.
2. The extra parameter ν takes over one of the dimensions of the table because three-dimensional tables are illegal.
3. Many of these tables are "two-sided," meaning their t-scores and probabilities assume you want the symmetric probability about the mean over the interval $[-t_b, t_b]$, where t_b is your t-score bound.

Consider the following example.

04.13.1 Example: confidence interval

Write a *Matlab* script to generate a data set with 200 samples and sample sizes $N \in \{10, 20, 100\}$ using any old distribution. Compare the distribution of the means for the different N. Use the sample distributions and a t-table to compute 99% confidence intervals.

Generate the data set.

```
M = 200; % # of samples
N_a = [10,20,100]; % sample sizes

mu = 27; % population mean
```

```
sigma = 9; % population std

rng(1) % seed random number generator
data_a = mu + sigma*randn(N_a(end),M); % normally distributed
size(data_a) % check size
data_a(1:10,1:5) % check first 10 rows and first five columns
```

```
ans =

   100   200


ans =

   21.1589   30.2894   27.8705   30.7835   28.3662
   37.6305   17.1264   28.2973   24.0811   34.3486
   20.1739   44.3719   43.7059   39.0699   32.2002
   17.0135   32.6064   36.9030   37.9230   36.5747
   19.3900   32.9156   23.7230   22.4749   19.7709
   21.8460   13.8295   31.2479   16.9527   34.1876
   21.9719   34.6854   19.4480   18.7014   24.1642
   28.6054   32.2244   22.2873   26.9906   37.6746
   25.2282   18.7326   14.5011   28.3814   27.7645
   32.2780   34.1538   27.0382   18.8643   14.1752
```

Compute the means for different sample sizes.

```
mu_a = NaN*ones(length(N_a),M);
for i = 1:length(N_a)
    mu_a(i,:) = mean(data_a(1:N_a(i),1:M),1);
end
```

Plotting the distribution of the means yields Figure 04.18.

It makes sense that the larger the sample size, the smaller the spread. A quantitative metric for the spread is, of course, the standard deviation of the means for each sample size.

```
S_mu = std(mu_a,0,2)
```

```
S_mu =

    2.8365
```

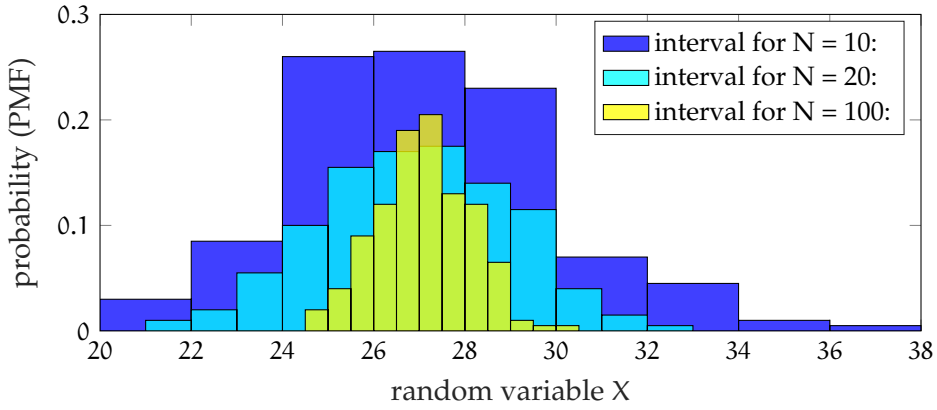


Figure 04.18: a histogram showing the distribution of the means for each sample size.

```
    2.0918
    1.0097
```

Look up t-table values for different sample sizes and 99% confidence. Use these, the mean of means, and the standard deviation of means to compute the 99% confidence interval for each N.

```
t_a = [3.25,2.861,2.626];
for i = 1:length(N_a)
    interval = mean(mu_a(i,:)) + [-1,1]*t_a(i)*S_mu(i);
    disp(sprintf('interval for N = %i: ',N_a(i)))
    disp(interval)
end
```

```
interval for N = 10:
   17.8786   36.3156

interval for N = 20:
   20.9567   32.9261

interval for N = 100:
   24.4397   29.7426
```

As expected, the larger the sample size, the smaller the interval over which we have 99% confidence in the estimate.

Lecture 04.14 Multivariate probability and correlation

Thus far, we have considered probability density and mass functions (PDFs and PMFs) of only one random variable. But, of course, often we measure multiple random variables $X_1, X_2, \ldots, X_n$ during a single event, meaning a measurement consists of determining values $x_1, x_2, \ldots, x_n$ of these random variables.

We can consider an n-tuple of continuous random variables to form a sample space $\Omega = \mathbb{R}^n$ on which a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$, called the *joint PDF* assigns a probability density to each outcome $\mathbf{x} \in \mathbb{R}^n$. The joint PDF must be greater than or equal to zero for all $\mathbf{x} \in \mathbb{R}^n$, the multiple integral over Ω must be unity, and the multiple integral over a subset of the sample space $A \subset \Omega$ is the probability of the event A.

joint PDF

We can consider an n-tuple of discrete random variables to form a sample space $\mathbb{N}_0^n$ on which a multivariate function $f : \mathbb{N}_0^n \to \mathbb{R}$, called the *joint PMF* assigns a probability to each outcome $\mathbf{x} \in \mathbb{N}_0^n$. The joint PMF must be greater than or equal to zero for all $\mathbf{x} \in \mathbb{N}_0^n$, the multiple sum over Ω must be unity, and the multiple sum over a subset of the sample space $A \subset \Omega$ is the probability of the event A.

joint PMF

Let's visualize this by plotting a bivariate Gaussian using *Matlab*'s `mvnpdf` function. The result is Figure 04.19. Note how the means and standard deviations affect the distribution.

```
mu = [10,20]; % means
Sigma = [1,0;0,.2]; % cov ... we'll talk about this
x1_a = linspace(...
    mu(1)-5*sqrt(Sigma(1,1)),...
    mu(1)+5*sqrt(Sigma(1,1)),...
    50);
x2_a = linspace(...
    mu(2)-5*sqrt(Sigma(2,2)),...
    mu(2)+5*sqrt(Sigma(2,2)),...
    50);
[X1,X2] = meshgrid(x1_a,x2_a);
f = mvnpdf([X1(:) X2(:)],mu,Sigma);
f = reshape(f,length(x2_a),length(x1_a));

h = figure;
p = surf(x1_a,x2_a,f);
xlabel('$x_1$','interpreter','latex')
ylabel('$x_2$','interpreter','latex')
zlabel('$f(x_1,x_2)$','interpreter','latex')
```

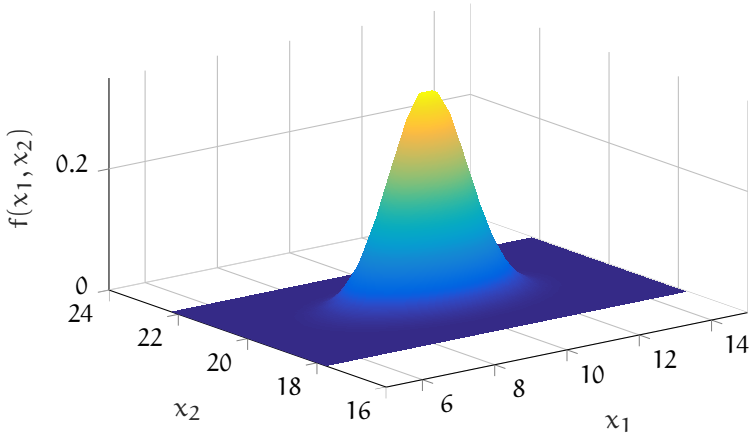


Figure 04.19: two-variable Gaussian PDF.

```
shading interp
hgsave(h,'figures/temp');
```

04.14.1 Marginal probability

marginal PDF

The *marginal PDF* of a multivariate PDF is the PDF of some subspace of Ω after one or more variables have been "integrated out," such that a fewer number of random variables remain. Of course, these marginal PDFs must have the same properties of any PDF, such as integrating to unity.

Let's demonstrate this by numerically integrating over x_2 in the bivariate Gaussian, above.

```
f1 = trapz(x2_a,f',2); % trapezoidal integration
```

Plotting this yields Figure 04.20.

We should probably verify that this integrates to one.

```
disp(['integral over x_1 = ',sprintf('%0.7f',trapz(x1_a,f1))])
```

```
integral over x_1 = 0.9999988
```

Not bad.

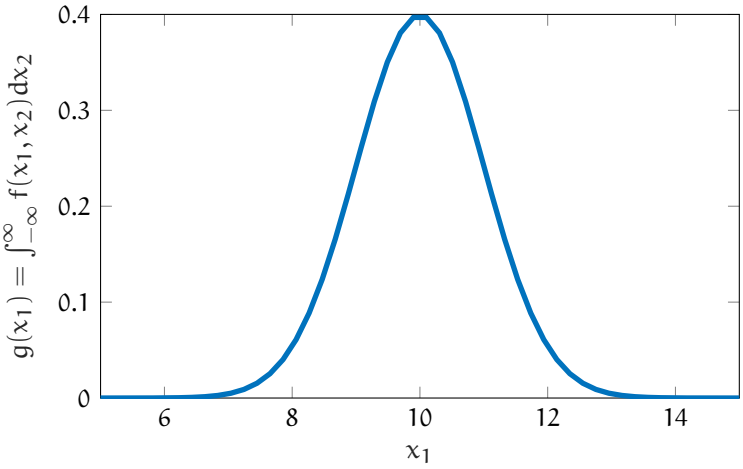


Figure 04.20: marginal Gaussian PDF $g(x_1)$.

04.14.2 Covariance

Very often, especially in *machine learning* or *artificial intelligence* applications, the question about two random variables X and Y is: how do they co-vary? That is what is their *covariance*, defined as

machine learning
artificial intelligence
covariance

Note that when $X = Y$, the covariance is just the variance. When a covariance is large and positive, it is an indication that the random variables are *strongly correlated*. When it is large and negative, they are *strongly anti-correlated*. Zero covariance means the variables are *uncorrelated*. In fact, *correlation* is defined as

correlation

This is essentially the covariance "normalized" to the interval $[-1, 1]$.

04.14.2.1 Sample covariance

As with the other statistics we've considered, covariance can be estimated from measurement. The estimate, called the *sample covariance* q_{XY}, of

sample covariance

random variables X and Y with sample size N is given by

04.14.2.2 Multivariate covariance

With n random variables X_i, one can compute the covariance of each pair. It is common practice to define an $n \times n$ matrix of covariances called the *covariance matrix* Σ such that each pair's covariance

covariance matrix

$$\operatorname{cov}(X_i, X_j) \tag{04.23}$$

appears in its row-column combination (making it symmetric), as shown below.

The multivariate *sample covariance matrix* Q is the same as above, but with sample covariances $q_{X_i X_j}$.

sample covariance matrix

Both covariance matrices have correlation analogs.

04.14.2.3 Example: car data

Let's use a built-in multivariate data set that describes different features of cars, listed below.

```
d = load('carsmall.mat') % this is a "struct"
```

```
d =

           Model: [100x33 char]
          Origin: [100x7 char]
             MPG: [100x1 double]
       Cylinders: [100x1 double]
```

```
    Displacement: [100x1 double]
      Horsepower: [100x1 double]
          Weight: [100x1 double]
    Acceleration: [100x1 double]
      Model_Year: [100x1 double]
             Mfg: [100x13 char]
```

Let's compute the sample covariance and correlation matrices.

```
variables = {...
    'MPG','Cylinders',...
    'Displacement','Horsepower',...
    'Weight','Acceleration',...
    'Model_Year'};
n = length(variables);
m = length(d.MPG);

data = NaN*ones(m,n); % preallocate
for i = 1:n
    data(:,i) = d.(variables{i});
end

cov_d = nancov(data); % sample covariance
cor_d = corrcov(cov_d) % sample correlation
```

```
cor_d =

    1.0000   -0.8367   -0.8048   -0.8028   -0.8591    0.4631    0.7112
   -0.8367    1.0000    0.9486    0.8588    0.8886   -0.6052   -0.5844
   -0.8048    0.9486    1.0000    0.9102    0.8860   -0.6719   -0.5557
   -0.8028    0.8588    0.9102    1.0000    0.8656   -0.6836   -0.5843
   -0.8591    0.8886    0.8860    0.8656    1.0000   -0.4642   -0.4673
    0.4631   -0.6052   -0.6719   -0.6836   -0.4642    1.0000    0.4404
    0.7112   -0.5844   -0.5557   -0.5843   -0.4673    0.4404    1.0000
```

This is highly correlated/anticorrelated data! Let's plot each variable versus each other variable to see the correlations of each. We use a red-to-blue colormap to contrast anticorrelation and correlation. Purple, then, is uncorrelated.

The following builds the red-to-blue colormap.

```
n_colors = 10;
cmap_rb = NaN*ones(n_colors,3);
```

```
for i = 1:n_colors
    a = i/n_colors;
    cmap_rb(i,:) = (1-a)*[1,0,0]+a*[0,0,1];
end
```

```
h = figure;
for i = 1:n
    for j = 1:n
        subplot(n,n,sub2ind([n,n],j,i))
        p = plot(d.(variables{i}),d.(variables{j}),'.'); hold on
        this_color = cmap_rb(round((cor_d(i,j)+1)*(n_colors-1)/2),:);
        p.MarkerFaceColor = this_color;
        p.MarkerEdgeColor = this_color;
    end
end
hgsave(h,'figures/temp');
```

04.14.3 Conditional probability and dependence

Independent variables are uncorrelated. However, *uncorrelated variables may or may not be independent*. Therefore, we cannot use correlation alone as a test for independence. For instance, for random variables X and Y, where X has some even distribution and $Y = X^2$, clearly the variables are *dependent*. However, the are also *uncorrelated* (due to symmetry).

Using a uniform distribution $\mathcal{U}(-1, 1)$, let's show this with some sampling. We compute the correlation for different sample sizes.

```
N_a = round(linspace(10,500,100));
qc_a = NaN*ones(size(N_a));
rng(6)
x_a = -1 + 2.*rand(max(N_a),1);
y_a = x_a.^2;
for i = 1:length(N_a)
    % should write incremental algorithm
    % but lazy
    q = cov(x_a(1:N_a(i)),y_a(1:N_a(i)));
    qc = corrcov(q);
    qc_a(i) = qc(2,1); % "cross" correlation
end
```

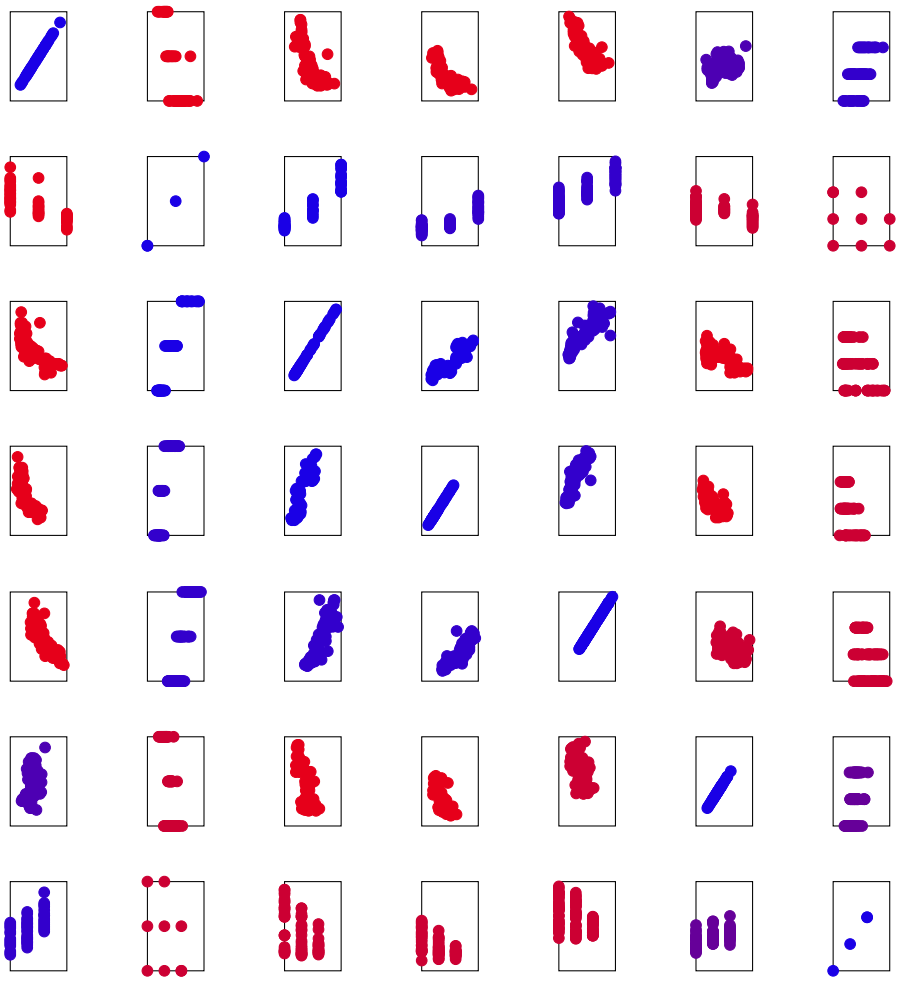


Figure 04.21: car data correlation.

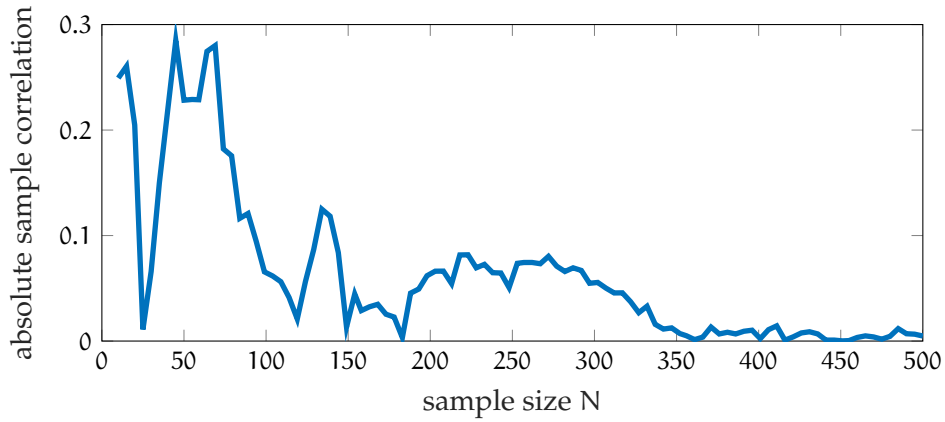


Figure 04.22: absolute value of the sample correlation between $X \sim \mathcal{U}(-1,1)$ and $Y = X^2$ for different sample size N. In the limit, the population correlation should approach zero *and yet* X *and* Y *are not independent*.

The absolute values of the correlations are shown in Figure 04.22. Note that we should probably average several such curves to estimate how the correlation would drop off with N, but the single curve describes our understanding that the correlation, in fact, approaches zero in the large-sample limit.

Lecture 04.15 Regression

Suppose we have a sample with two measurands: (1) the force F through a spring and (2) its displacement X (not from equilibrium). We would like to determine an analytic function that relates the variables, perhaps for prediction of the force given another displacement.

There is some variation in the measurement. Let's say the following is the sample.

```
X_a = 1e-3*[10,21,30,41,49,50,61,71,80,92,100]'; % m
F_a = [50.1,50.4,53.2,55.9,57.2,59.9,61.0,63.9,67.0,67.9,70.3]'; % N
```

Let's take a look at the data. The result is Figure 04.23.

```
h = figure;
p = plot(X_a*1e3,F_a,'.b','MarkerSize',15);
xlabel('$X$ (mm)','interpreter','latex')
ylabel('$F$ (N)','interpreter','latex')
xlim([0,max(X_a*1e3)])
grid on
hgsave(h,'figures/temp');
```

How might we find an analytic function that agrees with the data? Broadly, our strategy will be to assume a general form of a function and use the data to set the parameters in the function such that the difference between the data and the function is minimal.

Let y be the analytic function that we would like to fit to the data. Let y_i denote the value of $y(x_i)$, where x_i is the ith value of the random variable X from the sample. Then we want to minimize the differences between the force measurements F_i and y_i.

From calculus, recall that we can minimize a function by differentiating it and solving for the zero-crossings (which correspond to local maxima or minima).

First, we need such a function to minimize. Perhaps the simplest, effective function D is constructed by squaring and summing the differences we want to minimize, for sample size N:

(recall that $y_i = y(x_i)$, which makes D a function of x).

Now the form of y must be chosen. We consider only mth-order polynomial functions y, but others can be used in a similar manner:

$$y(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m. \tag{04.24}$$

If we treat D as a function of the polynomial coefficients a_j, i.e.

$$D(a_0, a_1, \cdots, a_m), \tag{04.25}$$

and minimize D for each value of x_i, we must take the partial derivatives of D with respect to each a_j and set each equal to zero:

This gives us N equations and $m + 1$ unknowns a_j. Writing the system in matrix form,

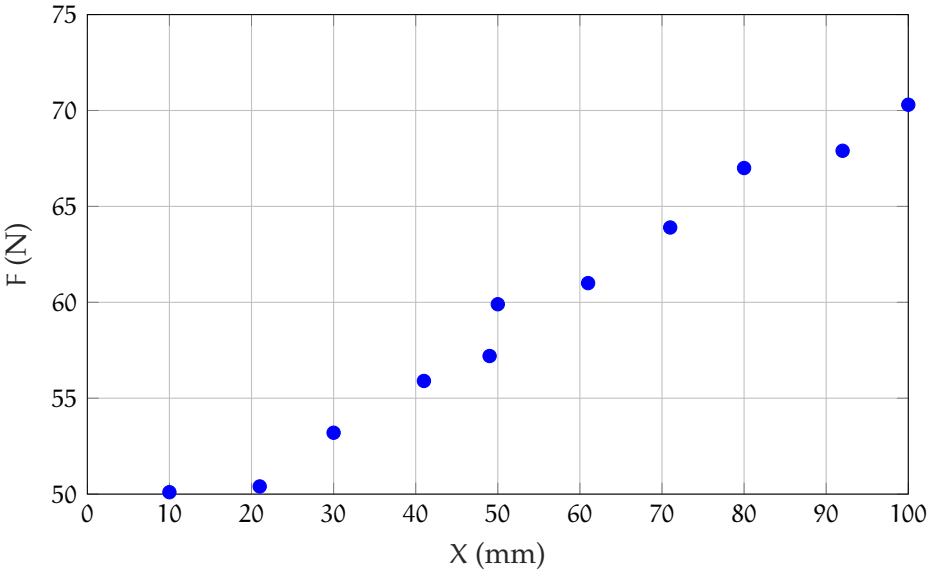


Figure 04.23: force-displacement data.

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^m \end{bmatrix}}_{A_{N\times(m+1)}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}}_{\mathbf{a}_{(m+1)\times 1}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{\mathbf{b}_{(m+1)\times 1}} . \tag{04.26}$$

Typically $N > m$ and this is an *overdetermined system*. Therefore, we usually can't solve by taking A^{-1} because A doesn't have an inverse!

Instead, we either find the *Moore-Penrose pseudo-inverse* $A^\dagger$ and have $\mathbf{a} = A^\dagger \mathbf{b}$ as the solution, which is *inefficient*—or we can approximate $\mathbf{b}$ with an algorithm such as that used by *Matlab*'s \ operator. In the latter case, `a_a = A\b_a`.

Let's use *Matlab*'s \ operator to find a good fit for the sample. Now, there's the sometimes-difficult question "what order should we fit?" Let's try out several and see what the squared-differences function D gives.

```
N = length(X_a); % sample size
m_a = 2:N; % all the order up to N

A = NaN*ones(length(m_a),max(m_a),N);
for k = 1:length(m_a) % each order
    for j = 1:N % each measurement
        for i = 1:( m_a(k) + 1 ) % each coef
            A(k,j,i) = X_a(j)^(i-1);
        end
    end
end
disp(squeeze(A(2,:,1:5)))
```

```
    1.0000    0.0100    0.0001    0.0000       NaN
    1.0000    0.0210    0.0004    0.0000       NaN
    1.0000    0.0300    0.0009    0.0000       NaN
    1.0000    0.0410    0.0017    0.0001       NaN
    1.0000    0.0490    0.0024    0.0001       NaN
    1.0000    0.0500    0.0025    0.0001       NaN
    1.0000    0.0610    0.0037    0.0002       NaN
    1.0000    0.0710    0.0050    0.0004       NaN
    1.0000    0.0800    0.0064    0.0005       NaN
    1.0000    0.0920    0.0085    0.0008       NaN
    1.0000    0.1000    0.0100    0.0010       NaN
```

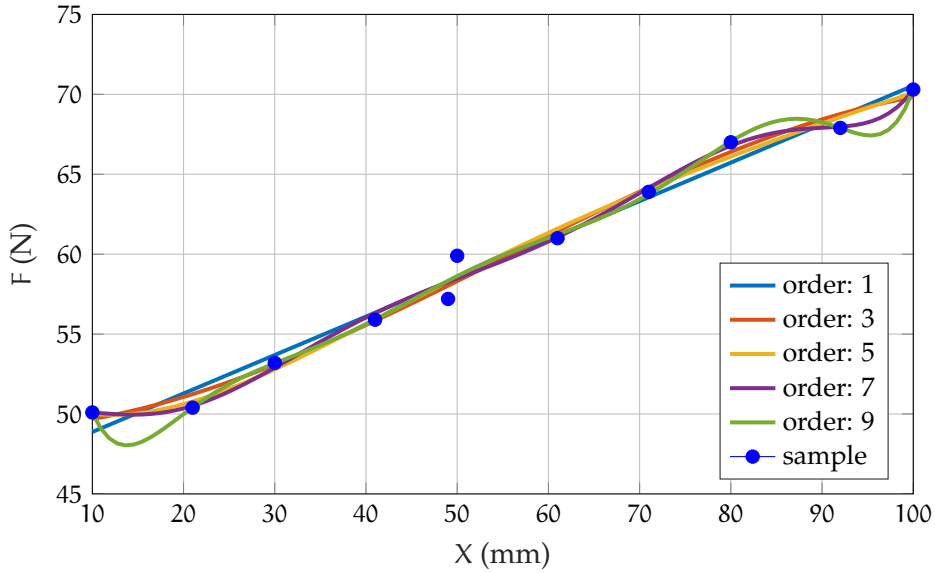


Figure 04.24: force-displacement data with curve fits.

We've printed the first five columns of the third-order matrix, which only has four columns, so `NaN`s fill in the rest.

Now we can use the \ operator to solve for the coefficients.

```
a = NaN*ones(length(m_a),max(m_a));

warning('off','all')
for i = 1:length(m_a)
    A_now = squeeze(A(i,:,1:m_a(i)));
    a(i,1:m_a(i)) = (A_now(:,1:m_a(i))\F_a)';
end
warning('on','all')
```

```
n_plot = 100;
x_plot = linspace(min(X_a),max(X_a),n_plot);
y = NaN*ones(n_plot,length(m_a)); % preallocate
for i = 1:length(m_a)
    y(:,i) = polyval(fliplr(a(i,1:m_a(i))),x_plot);
end
```

05

Uncertainty analysis

In this chapter, we explore different ways to quantify uncertainty, beginning with a *design-stage uncertainty analysis* at which point we estimate uncertainty based on minimal information with little or no data and working up to a *rigorous uncertainty analysis* at which we estimate uncertainty from the information used before and include that of multiple measurements.

design-stage uncertainty analysis

rigorous uncertainty analysis

Lecture 05.01 Design-stage uncertainty analysis

Design stage uncertainty analysis is the type of analysis done when designing a measurement system, before actual measurements are made, to predict the uncertainty in measurements made by the system being designed. It comes from different sources in a measurement system which will now be considered. Often, at this stage, we do not classify each uncertainty as systematic or random.

Zero-order uncertainty

Zero-order uncertainty u_0 is uncertainty due to instrument resolution, which is "arbitrarily" considered to be

instrument uncertainty
elemental error

Some instruments have error estimates in their manuals; this is called *instrument uncertainty* u_c. Occasionally, this is given as a single value, but more often several contributing *elemental errors* are given, such as linearity error (due to nonlinearities) and hysteresis error (due to a lack of symmetry in a measurement's increase versus its decrease).

root-sum-squares (RSS) method

The *root-sum-squares (RSS) method* allows us to estimate the total instrument uncertainty u_c due to the elemental uncertainties u_k (with number of elements K) as

The RSS method can be used to combine design-state uncertainties of concatenated (series) instruments for a measurement, as well (e.g. a force measurement with a force-to-voltage transducer and a multimeter).

The confidence/probability level $P\%$ depends on the confidence of each error estimate (ideally, they all have the same confidence). If none is given, it is common to use 95%.

design-stage uncertainty

The *design-stage uncertainty* u_d for an instrument is defined as

$$u_d = \sqrt{u_0^2 + u_c^2} \qquad (P\%). \tag{05.1}$$

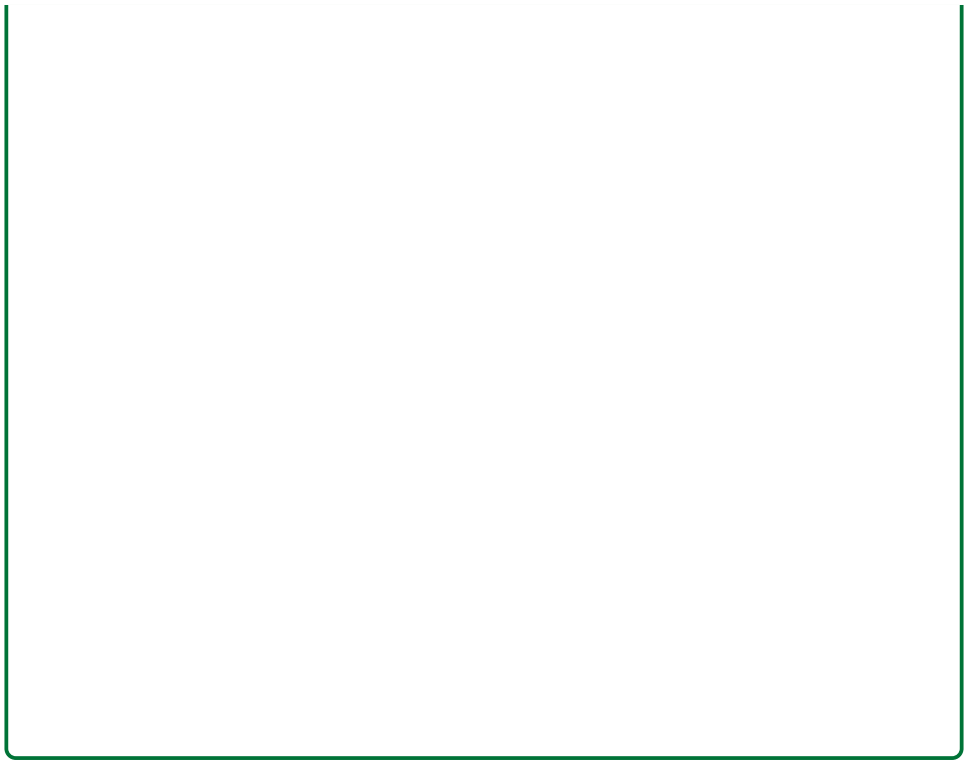
This is an estimate of our uncertainty based solely on information about the instruments. This should be considered an estimate of our *minimum* uncertainty. Factors we will later consider will add uncertainty.

Example 05.01-1 force measurement with a load cell and a digital voltage measurement

Estimate the design-stage uncertainty for the measurement of a force with a load cell (including amplifier) that transduces force to voltage and a digital voltage measurement via an analog input of a microcontroller. The following tables include specifications from each instrument's manual.

load cell	
range	$[0, 60]$ N
sensitivity	0.1 V/N
linearity error	0.15 mV/N
sensitivity error	0.25 mV/N

μC AI	
range	$[0, 5]$ V
ADC resolution	12 bits
absolute accuracy	50 mV



Lecture 05.02 Functional propagation of uncertainty

Often, we use a measurement to estimate a quantity that is functionally dependent on the measurand. For instance, perhaps we would like to estimate the volume V of an object that—would you look at that—happens to be cubic with side length ℓ, so its volume could be reasonably estimated to be $V(\ell) = \ell^3$. Your measurement of ℓ has some associated uncertainty u_ℓ, certainly. How does that propagate to an uncertainty u_V in V?

Recall that uncertainty is half of a *symmetric* interval centered at the best estimate of the value. When you drop an interval symmetric about some value $\tilde{x}$ into a nonlinear function f, that interval comes out (usually) *asymmetric* about $\tilde{x}$.

asymmetry

Let's demonstrate this with our cubic volume. Let the 95% uncertainty in $\bar{\ell}$ be u_ℓ, such that there is a 95% probability that a volume measurement value

Now, run that interval $\bar{\ell} \pm u_\ell$ through the volume function V:

…this isn't symmetric about the mean $V(\bar{\ell})$ so we linearize …which is what we also do for a multivariate function, too, and multiply each independent variable's partial derivative slope (evaluated at the mean) by the uncertainty of that variable's measurement. Then combine with RSS.

Lecture 05.03 Rigorous uncertainty analysis

confidence

We have learned about *confidence* from a statistical point-of-view. Based on sample variability, we might have $P\%$ confidence in a given measurement. As we have learned, an estimate of the variability of the random variable X—in this case, the measurand—is given by its *sample standard deviation* S_X. In the case of multiple samples, its best estimate was the *sample mean of standard deviations* $\overline{S_X}$. The best estimate of the value of X (that is, it's mean μ_X) was the *sample mean* $\overline{X}$. In the case of multiple samples, it was the sample mean of sample means $\overline{\overline{X}}$. Finally, the best estimate of the variability of the mean was the *sample standard deviation of sample means* $S_{\overline{X}}$. Further recall the nice estimate of $S_{\overline{X}}$ from a *single sample* with size N:

sample standard deviation
sample mean of standard deviations
sample mean
sample standard deviation of sample means

As we can see, as sample size N increases, $S_{\overline{X}}$ decreases. This type of error in a measurement is called *random error* and gives rise to *random uncertainty* u_r, related to what we have called *confidence intervals* about the best estimate of the mean, such as, for a single sample of size N,

random error
random uncertainty
confidence intervals

$$x \in \overline{X} \pm t_{\nu,P} S_{\overline{X}} \qquad (P\%) \tag{05.2}$$

where $\nu = N-1$ is the degree of freedom and $P\%$ is our confidence based on the probability $P\%$ of a Student random variable X taking a value x within $t_{\nu,P}$ standard deviations of the mean. The random uncertainty is a half-interval

$$u_r = t_{\nu,P}\, s \qquad (P\%) \tag{05.3}$$

where $s = S_{\overline{X}}\,(68\%)$ is the *standard random uncertainty*, which is simply one standard deviation of the means.

standard random uncertainty

However, error can arise from more than randomness. Other sources arise that *bias* the measured values—say, up or down—from the mean. This is called *systematic error* and generates *systematic uncertainty* u_b (because *bias*) and *systematic standard uncertainty* b that has 68% confidence. Let a measurement instrument's manual list an elemental error B, which (unless otherwise stated in the manual) is assigned a 95% confidence; the systematic standard uncertainty is $b = B/2$. Assuming a large sample was used to estimate B, we might report an uncertainty associated with that

bias
systematic error
systematic uncertainty
systematic standard uncertainty

error to be $u_b = 2b$ with 95% confidence (we are assuming a Gaussian distribution, but the distribution shape has little effect).

Let's consider systematic error a bit more, through an example. A scale might be systematically reading high (I know my scale does, especially around the holidays). This can be identified and mitigated by *calibration* to a standard. The National Institute of Standards and Technology (NIST) calibrates weights. Let's say you have a 10 kg NIST calibrated weight (such an object is called a *standard*) with one-normal-standard deviation confidence $\pm 200 \cdot 10^{-9}$ kg. Let's say I weigh it $N = 10$ times on my scale and the sample mean $\bar{x} = 10.5$ kg and sample standard deviation $S_x = 0.3$ kg. The calibration allows me to adjust the bias on my scale by $10 - 10.5 = -0.5$ kg.

calibration

standard

However, there remain two systematic uncertainties associated with my scale's bias: (1) NIST's standard uncertainty $b_{std} = 200 \cdot 10^{-9}$ kg due to NIST's measurement of the standard 10 kg weight and (2) our calibration standard uncertainty

The systematic standard uncertainties are combined in the usual RSS way (although the calibration uncertainty clearly dominates):

A measurement sample of size $M = 23$ of an object with unknown mass m is then performed with the calibrated scale. The sample mean is $\bar{m} = 9.04$ kg and sample standard deviation $S_m = 1$ kg. How confident can we be in the result? Certainly both the systematic and random certainties must contribute. Before we can consider their combined effect, let's compute the standard random uncertainty:

We combine the systematic and random uncertainties in the usual RSS way:

Since these are *standard* uncertainties, we must find its effective degree of freedom ν before assigning it a confidence with a Student t-score. How can we estimate this from these different sources of uncertainty, each with their own degree of freedom? A method of estimating the effective degree of freedom is given by (Figliola and Beasley, 2015) and is presented in an equivalent form here. Let a measurement have J random standard uncertainties s_j with corresponding degrees of freedom ν_{s_j}; further, let it have K systematic standard uncertainties b_k with corresponding degrees of freedom ν_{b_k}; then the effective degree of freedom is

$$\nu = \frac{\left(\sum_{j=1}^{J} s_j^2 + \sum_{k=1}^{K} b_k^2\right)^2}{\sum_{j=1}^{J} s_j^4/\nu_{s_j} + \sum_{k=1}^{K} b_k^4/\nu_{b_k}}. \tag{05.4}$$

From above, we have $J = 1$ and $K = 2$ and standard uncertainties given in Table 05.1. This gives $\nu = 29.0$. That's close enough to 30 to call it "large" and assign a 95% confidence uncertainty

So, using our 95% confidence uncertainty for our interval, our best estimate for the mass is

Table 05.1: summary of standard uncertainties for a fictional mass measurements.

uncertainty	value s or b	deg. of freedom
s	0.209 kg	22
b_{std}	$200 \cdot 10^{-9}$ kg	∞
b_{cal}	0.0949 kg	9

05.03.1 An extensive example

Consider a temperature measurement made with a linear calibrated temperature-voltage transducer. The calibration data is given as `t_cal` (units C) and `v_cal` (units V). The measurement voltage sample is given as a time series `v_a` (units V) versus `time_a` (units s), where we can assume relatively constant measurement conditions and a stationary process.

The voltmeter (used for calibration and for data) and the thermometer (used for calibration) have the systematic uncertainties defined below.

```
bv_1 = .1; % V ... voltmeter absolute uncertainty
bv_2 = .05; % V ... voltmeter linearity uncertainty
bt_1 = .05; % C ... thermometer absolute uncertainty
```

05.03.1.1 Calibration curve and its uncertainty

Let's first consider the calibration data.

```
disp('sample data (time,voltage)')
disp([time_a;v_a]')
```

```
sample data (time,voltage)
         0    5.7959
    1.8182    5.5286
    3.6364    5.2110
    5.4545    5.4191
```

```
    7.2727    5.6164
    9.0909    5.6746
   10.9091    5.4349
   12.7273    5.8535
   14.5455    5.6782
   16.3636    5.7058
   18.1818    5.5820
   20.0000    6.3077
```

Let's perform a linear regression analysis on the calibration data to find a *calibration curve*. The standard uncertainty of a polynomial regression of order $m-1$ and data with values $\tilde{y}_i$ approximating calibration curve values y_i with sample size N is (Figliola and Beasley, 2015, Equation 4.35)

$$s_{\text{fit}} = \sqrt{\frac{\sum_{i=1}^{N} (\tilde{y}_i - y_i)^2}{\nu}} \tag{05.5}$$

where the degree of freedom $\nu = N - (m+1)$. For a linear fit, $m = 2$.

```
pf_cal = polyfit(v_cal,t_cal,1)
k_trans = pf_cal(1); % this is the transducer gain
p_cal = polyval(pf_cal,v_cal);
d_cal = p_cal - t_cal;
nu_cal = length(d_cal)-(2+1)
s_cal = sqrt((sum(d_cal.^2))/nu_cal)
```

```
pf_cal =
    4.9994   -0.2028
nu_cal =
    22
s_cal =
    0.4712
```

```
h = figure;
p = plot(v_cal,t_cal,'x'); hold on
p2 = plot(v_cal,p_cal,'r-');
grid on;
xlabel('voltage (V)')
ylabel('temperature (C)')
hgsave(h,'figures/temp');
```

05.03.1.2 Random uncertainty

Another source of random error is the finite sample size. It can be computed, in the usual way, as the sample standard deviation of the sample means. And first, of course, the sample `v_a` must be passed through the calibration curve `pf_cal`.

```
t_a = polyval(pf_cal,v_a);
mu_t = mean(t_a)
s_mu_t = std(t_a)/length(t_a)
nu_a = length(t_a)-1
```

```
mu_t =
   28.0473
s_mu_t =
    0.1132
nu_a =
    11
```

The total random uncertainty is the root-sum-square (RSS) combination of the calibration and finite sample size uncertainties.

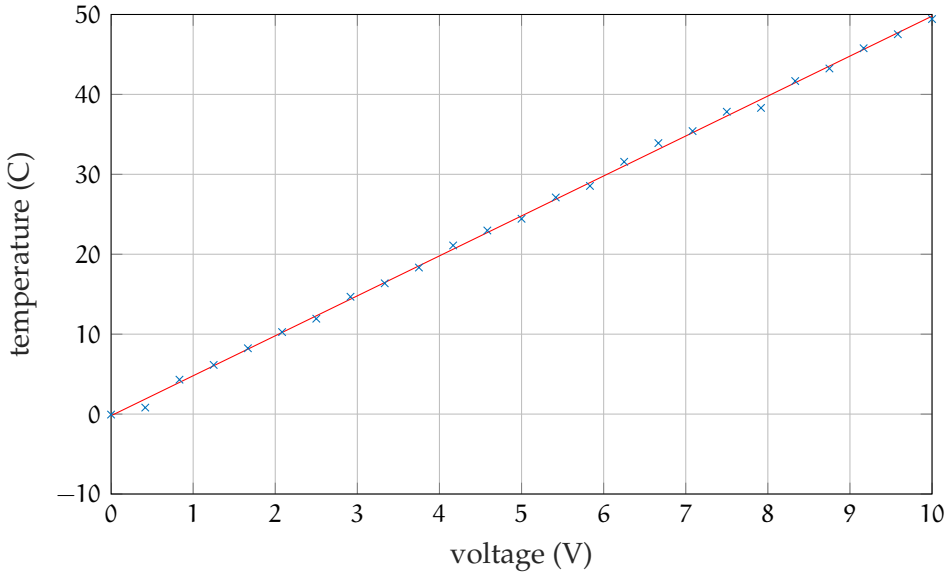


Figure 05.1: voltage-temperature transducer calibration data with its linear fit.

```
s = sqrt(s_cal^2+s_mu_t^2)
```

```
s =
    0.4846
```

05.03.1.3 Systematic uncertainty

The total systematic uncertainty is an RSS combination of those systematic uncertainties described in the problem statement. The transducer gain, found from the calibration curve, can be used to convert voltage uncertainties to temperature uncertainties.

```
b = sqrt(2*(k_trans*bv_1)^2+2*(k_trans*bv_2)^2+bt_1^2)
```

```
b =
    0.7921
```

Note the factors of two. These are due to the voltmeter's use in the calibration and in the sample.

05.03.1.4 Total uncertainty

The total *standard* uncertainty is the RSS combination of the standard random and systematic uncertainties.

```
u_t = sqrt(s^2+b^2)
```

```
u_t =
    0.9286
```

In order to assign a confidence interval via a t-score, we can use Equation 05.4 to compute the effective degree of freedom of the standard uncertainty. Given no information to the contrary, we assume the degree of freedom for each systematic uncertainty is high.

```
nu_t = (u_t^2)^2/(s_cal^4/nu_cal+s_mu_t^4/nu_a)
```

```
nu_t =
  329.5059
```

This is much greater than 30, so we can assume the distribution is Gaussian and use a z-score. Let's assign a 95% confidence uncertainty.

```
u_t_95 = 2*u_t
```

```
u_t_95 =
    1.8571
```

So a confidence interval for the estimate of the temperature is as follows.

```
mu_t_int = mu_t + [-1,1]*u_t_95
```

This is the result of our full uncertainty analysis.

06

Electricity measurement

Lecture 06.01 Instrumentation for electricity measurement

At a high level, let's survey the measurement of electricity. We begin with electricity because nearly every modern measurement device has an electronic stage, so electronic measurement is fundamental for measuring most quantities. Although many modern measurement devices are digital—that is, sampled—we first consider analog measurement techniques, the principles of which still apply to digital measurement, and which are still in practical use in some cases.

The fundamental quantities to be measured are current, voltage, and resistance. Quick, one-off measurements of these quantities can be performed with a handheld *multimeter*, which can be either analog or digital. A multimeter has different modes for measuring a current, voltage, or resistance. To measure voltage or resistance between two nodes in a circuit, the multimeter's two probes are simultaneously contacted with them. To measure the current through a circuit element, the multimeter itself must be placed in the circuit such that current flows through it. In the former case, it is best for the multimeter to have *high* input resistance such that it draws as little as possible current through itself (and thereby affecting the measurement). In the latter case of current measurement, it is preferable for the multimeter to have *low* input resistance such that it drops the voltage across itself as little as possible (and thereby affecting the measurement).

multimeter

Figure 06.1: a Fluke multimeter from the SMU Robotics lab.

Precision, (typically) benchtop multimeters are available that can reduce the uncertainty in one-off measurements.

Specific aspects of an AC electronic signal can be measured with a multimeter; most commonly, just the root-mean-square (RMS) voltage or current can be measured. However, these measurements have significant limitations, including their effective frequency bandwidth, (typical) inability to indicate the signal frequency, and lack of information about the signal's

Figure 06.2: a Tektronix oscilloscope from the SMU Robotics lab.

noise. The multimeter's wheelhouse is the DC signal and the standard 60 Hz AC power transmission signal.

Time-varying analog and digital voltage signals—including, DC, AC, and, to some extent, "other"—can be viewed effectively on an *oscilloscope* (or "scope"), a photo of which is shown in Figure 06.2 . These devices effectively show a trace of a signal across a time-window. If the window is properly "triggered," such that the window starts at the same point in a periodic signal, it will trace approximately the same path across the window. This gives the illusion of "zooming in" in time and viewing the signal across a viewing window of, say a couple milliseconds. Oscilloscopes are mostly practical for debugging and one-off measurements. But they are super fun.

oscilloscope

For repeated, continuous, stored measurements of signals—DC, AC, and "other"—it is now standard practice to use a *digital data acquisition* device that includes *analog-to-digital conversion* (ADC). A plethora of *microcontrollers* (μCs) are now available for such measurements, ranging from inexpensive and inaccurate to accurate and expensive, as usual. Dedicated data acquisition boards can be very expensive (six figures), but highly flexible and accurate.

digital data acquisition
analog-to-digital conversion
microcontrollers

It is notable that most data acquisition boards have analog inputs configured to measure *voltage only*. Therefore, if one wishes to measure current or resistance, a separate *sensor* is required. The crudest current sensor is simply a resistor with a known resistance placed in series with the

sensor

element through which one would like to measure the current. Measuring the voltage across it and hitting the result with your autOhmatic reveals the current.

This resistor-as-current-sensor has two distinct disadvantages: (1) like with the multimeter, one must "break" the circuit in order to flow current through the resistor/sensor and (2) the resistor's inclusion in the circuit will affect it by (assuming its not fighting a controlled current source) reducing the current flow and dropping the voltage. Much better current sensors exist, such as the *Hall effect sensor*. This is typically an integrated circuit (microchip) with a current pass-through that sees less than a mΩ of resistance! It outputs an analog voltage approximately linearly proportional to the current, ready for a data acquisition board analog input. Other current sensors exist that can be clamped around a wire to measure the current through it.

Hall effect sensor

There are several ways to measure resistance with a data acquisition board. Probably the easiest way is to put the unknown resistance in a voltage-divider with a known resistance and backing-out the unknown resistance value from the known input and output voltages and the known resistance. Another is to measure the voltage across and the current through (using, say, a Hall effect sensor) the unknown resistance, then letting the Ohm-g regulate. However, a much better way—with a *Wheatstone bridge circuit*—is described in detail in Lecture 06.02.

Wheatstone bridge circuit

There are also these special devices, typically benchtop and expensive, that are like the alter-ego of oscilloscopes: *spectrum analyzers*. These show "real-time" (quickly-updating) fast (discrete) Fourier transforms of signals on a screen. More band-limited spectrum analyzing functionality has relatively recently become available in higher-end oscilloscopes. I think it's reasonable to assume this label coinage will stick: *spectroscilloscopes*.[1] Ohmg.

spectrum analyzers

spectroscilloscopes

[1] I googled it 5 December 2017 and there were no results. Watch it grow.

Lecture 06.02 Measuring resistance well

resistive sensors

Many sensors are *resistive*, meaning the physical quantity to which they are sensitive affect the electrical resistance of the sensor, the accurate measurement of which is necessary for an accurate measurement of the physical quantity. In Lecture 06.01, we learned that we can measure unknown resistance R_u by applying a known voltage to it V_s, measuring the current i_{R_u} through it, and using Ohm's law:

Furthermore, we learned an alternative is to place the unknown resistor in a voltage-divider circuit with a known resistor R_k, apply a known voltage V_s, measure the output voltage v_{R_k}, and use the voltage-divider equation

to solve for the unknown resistance

The sensitivity of these methods to measured quantities i_{R_u} and v_{R_k} are:

For small i_{R_u} or v_{R_k}, which correspond to large R_u, these are *very sensitive*. This means a small uncertainty in our measurements would propagate with large (and therefore unwanted) multiplicative factors.

Wheatstone bridge circuit

We now explore the *Wheatstone bridge circuit* for measuring an unknown resistance.

06.02.1 Wheatstone bridge circuit

A Wheatstone bridge circuit for measuring unknown resistance R_u from measured (known) V_s, v_o, R_1, R_2, and R_3 is shown in Figure 06.3. We would first like to derive the relationship between V_s and v_o. The first observation we make is that the two "arms" of the bridge, R_1–R_2 and R_3–R_u, are each just voltage dividers of V_s. That is,

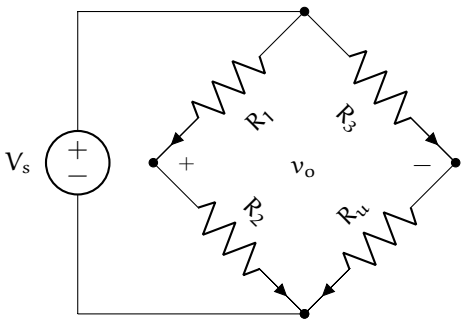


Figure 06.3: a Wheatstone bridge circuit.

By Kirchhoff's voltage law, $v_o = v_{R_2} - v_{R_u}$. These yield the desired relationship

$$v_o = \left(\frac{R_2}{R_1 + R_2} - \frac{R_u}{R_3 + R_u} \right) V_s. \tag{06.2}$$

Solving this for the unknown resistance, we obtain

$$R_u = \frac{R_3(R_2 V_s - (R_1 + R_2)v_o)}{R_1 V_s + (R_1 + R_2)v_o}. \tag{06.3}$$

It is typical common to have all resistors nearly equal to a single resistance R. Under this condition, the sensitivities of the measurement can be found to be

$$\begin{aligned} \frac{\partial R_u}{\partial v_o} &= -\frac{4RV_s}{(V_s + 2v_o)^2}, \\ \frac{\partial R_u}{\partial V_s} &= \frac{4RV_s}{(V_s + 2v_o)^2}, \text{ and} \\ \frac{\partial R_u}{\partial R} &= \frac{V_s - 2v_o}{V_s + 2v_o}. \end{aligned}$$

In all these expressions, we can control our sensitivity with the input voltage V_s.

06.02.2 Null method

The bridge is said to be *balanced* when $V_s \neq 0$ and $v_o = 0$. That is, when **balanced bridge**

From Equation 06.3, $v_o = 0$ greatly simplifies the expression for the unknown resistance

This is completely independent of V_s. Of course, if R_u is a resistive sensor, and its resistance changes such that the bridge is no longer balanced, the bridge must be re-balanced via changing another resistance some known amount. Often, R_2 is a *potentiometer* (variable resistor) that can be adjusted to balance the bridge. Sometimes feedback control is used to maintain a balanced bridge. **potentiometer**

This is called the *null method* because it requires a balanced bridge (zero output voltage). It is difficult to measure a signal that is time-varying (unless it is slow) with this method, due to the required constant balancing of the bridge. **null method**

06.02.3 Deflection method

The *deflection method* simply lets the bridge become unbalanced, logs the data, and applies Equation 06.3 to compute R_u. This is preferred for time-varying measurements, since it doesn't require a much faster bridge-balancing process. It does require that V_s is measured, which can, in some instances, lend a slight advantage to the null method in the case of stationary measurements. **deflection method**

07

Digital measurement

08

Temperature measurement

09

Pressure and velocity measurement

10

Flow measurement

11

Strain measurement

12

Sensors, actuators, and control

A

Algebra and trigonometry reference

Lecture A.01 Quadratic forms

The solution to equations of the form $ax^2 + bx + c = 0$ is

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{A.1}$$

A.01.1 Completing the square

This is accomplished by re-writing the quadratic formula in the form of the left-hand-side (LHS) of this equality, which describes factorization

$$x^2 + 2xh + h^2 = (x + h)^2. \tag{A.2}$$

Lecture A.02 Trigonometry

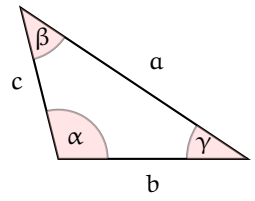
A.02.1 Triangle identities

With reference to the below figure, the *law of sines* is

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c} \tag{A.3}$$

and the *law of cosines* is

$$c^2 = a^2 + b^2 - 2ab\cos\gamma \tag{A.4a}$$
$$b^2 = a^2 + c^2 - 2ac\cos\beta \tag{A.4b}$$
$$a^2 = c^2 + b^2 - 2cb\cos\alpha \tag{A.4c}$$



A.02.2 Reciprocal identities

$$\csc u = \frac{1}{\sin u} \tag{A.5a}$$
$$\sec u = \frac{1}{\cos u} \tag{A.5b}$$
$$\cot u = \frac{1}{\tan u} \tag{A.5c}$$

A.02.3 Pythagorean identities

$$1 = \sin^2 u + \cos^2 u \tag{A.6a}$$
$$\sec^2 u = 1 + \tan^2 u \tag{A.6b}$$
$$\csc^2 u = 1 + \cot^2 u \tag{A.6c}$$

A.02.4 Co-function identities

$$\sin\left(\frac{\pi}{2}-u\right)=\cos u \tag{A.7a}$$
$$\cos\left(\frac{\pi}{2}-u\right)=\sin u \tag{A.7b}$$
$$\tan\left(\frac{\pi}{2}-u\right)=\cot u \tag{A.7c}$$
$$\csc\left(\frac{\pi}{2}-u\right)=\sec u \tag{A.7d}$$
$$\sec\left(\frac{\pi}{2}-u\right)=\csc u \tag{A.7e}$$
$$\cot\left(\frac{\pi}{2}-u\right)=\tan u \tag{A.7f}$$

A.02.5 Even-odd identities

$$\sin(-u)=-\sin u \tag{A.8a}$$
$$\cos(-u)=\cos u \tag{A.8b}$$
$$\tan(-u)=-\tan u \tag{A.8c}$$

A.02.6 Sum-difference formulas (AM or lock-in)

$$\sin(u\pm v)=\sin u\cos v\pm\cos u\sin v \tag{A.9a}$$
$$\cos(u\pm v)=\cos u\cos v\mp\sin u\sin v \tag{A.9b}$$
$$\tan(u\pm v)=\frac{\tan u\pm\tan v}{1\mp\tan u\tan v} \tag{A.9c}$$

A.02.7 Double angle formulas

$$\sin(2u)=2\sin u\cos u \tag{A.10a}$$
$$\cos(2u)=\cos^2 u-\sin^2 u \tag{A.10b}$$
$$=2\cos^2 u-1 \tag{A.10c}$$
$$=1-2\sin^2 u \tag{A.10d}$$
$$\tan(2u)=\frac{2\tan u}{1-\tan^2 u} \tag{A.10e}$$

A.02.8 Power-reducing or half-angle formulas

$$\sin^2 u = \frac{1-\cos(2u)}{2} \tag{A.11a}$$

$$\cos^2 u = \frac{1+\cos(2u)}{2} \tag{A.11b}$$

$$\tan^2 u = \frac{1-\cos(2u)}{1+\cos(2u)} \tag{A.11c}$$

A.02.9 Sum-to-product formulas

$$\sin u + \sin v = 2\sin\frac{u+v}{2}\cos\frac{u-v}{2} \tag{A.12a}$$

$$\sin u - \sin v = 2\cos\frac{u+v}{2}\sin\frac{u-v}{2} \tag{A.12b}$$

$$\cos u + \cos v = 2\cos\frac{u+v}{2}\cos\frac{u-v}{2} \tag{A.12c}$$

$$\cos u - \cos v = -2\sin\frac{u+v}{2}\sin\frac{u-v}{2} \tag{A.12d}$$

A.02.10 Product-to-sum formulas

$$\sin u \sin v = \frac{1}{2}\left[\cos(u-v) - \cos(u+v)\right] \tag{A.13a}$$

$$\cos u \cos v = \frac{1}{2}\left[\cos(u-v) + \cos(u+v)\right] \tag{A.13b}$$

$$\sin u \cos v = \frac{1}{2}\left[\sin(u+v) + \sin(u-v)\right] \tag{A.13c}$$

$$\cos u \sin v = \frac{1}{2}\left[\sin(u+v) - \sin(u-v)\right] \tag{A.13d}$$

A.02.11 Two-to-one formulas

$$A\sin u + B\cos u = C\sin(u+\phi) \tag{A.14a}$$

$$= C\cos(u+\psi) \text{ where} \tag{A.14b}$$

$$C = \sqrt{A^2+B^2} \tag{A.14c}$$

$$\phi = \quad \arctan\frac{B}{A} \tag{A.14d}$$

$$\psi = -\arctan\frac{A}{B} \tag{A.14e}$$

B

Distribution tables

Lecture B.01 Gaussian distribution table

Below are plots of the Gaussian probability density function (PDF) f and cumulative distribution function (CDF) Φ. Below them is Table B.1 of CDF values.

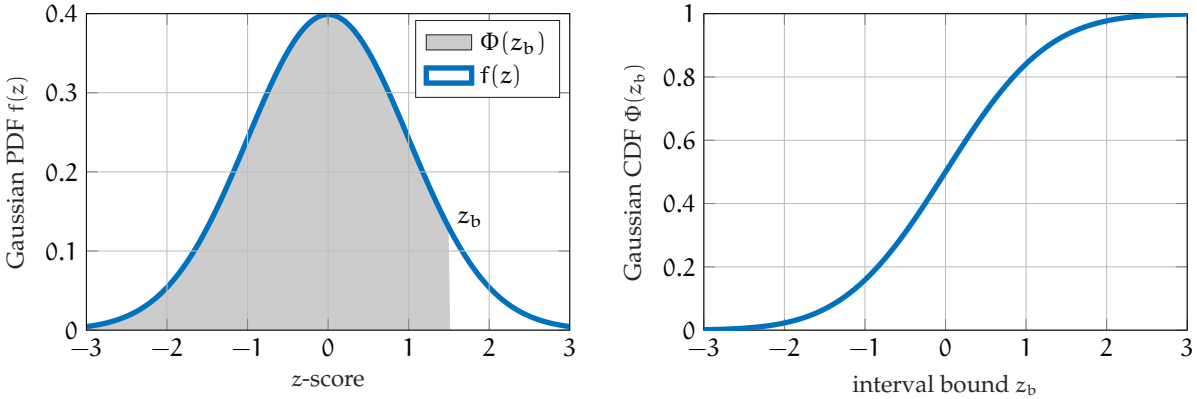


Table B.1: z-score table (columns are hundredths), $\Phi(z_b) = P(z \in (-\infty, z_b])$.

z_b	.␣0	.␣1	.␣2	.␣3	.␣4	.␣5	.␣6	.␣7	.␣8	.␣9
-3.4␣	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3␣	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2␣	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1␣	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0␣	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9␣	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8␣	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7␣	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6␣	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5␣	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4␣	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3␣	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2␣	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1␣	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0␣	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9␣	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8␣	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7␣	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6␣	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5␣	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4␣	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3␣	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2␣	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1␣	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0␣	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

Table B.1: z-score table (columns are hundredths), $\Phi(z_b) = P(z \in (-\infty, z_b])$.

z_b	·␣**0**	·␣**1**	·␣**2**	·␣**3**	·␣**4**	·␣**5**	·␣**6**	·␣**7**	·␣**8**	·␣**9**
-0.9␣	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8␣	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7␣	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6␣	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5␣	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4␣	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3␣	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2␣	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1␣	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0␣	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0␣	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1␣	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2␣	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3␣	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4␣	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5␣	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6␣	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7␣	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8␣	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9␣	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0␣	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1␣	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2␣	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3␣	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4␣	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5␣	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6␣	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7␣	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8␣	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9␣	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0␣	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1␣	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2␣	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3␣	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4␣	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5␣	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6␣	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7␣	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8␣	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9␣	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0␣	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1␣	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2␣	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3␣	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4␣	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

C

Bibliography

Robert B. Ash. *Basic Probability Theory*. Dover Publications, Inc., 2008.

Norman Robert Campbell. *Physics: The Elements*. Cambridge at the University Press, 1920. URL https://archive.org/details/physicstheelemen029733mbp.

James. Fieser and Bradley Harris. Dowden. Hans reichenbach (1891–1953). *The internet encyclopedia of philosophy*, 2017. ISSN 2161-0002. URL http://www.iep.utm.edu/reichenb/.

R.S. Figliola and D.E. Beasley. *Theory and Design for Mechanical Measurements, 6th Edition*. Wiley, 2015. ISBN 9781119031673. URL https://books.google.com/books?id=oQbYBgAAQBAJ.

Gene F. Franklin. *Digital Control Of Dynamic Systems*. Addison-Wesley, 1998.

Ned Hall. David Lewis's Metaphysics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

R. Priemer. *Introductory Signal Processing*. Advanced series in electrical and computer engineering. World Scientific, 1991. ISBN 9789971509194. URL https://books.google.com/books?id=QBT7nP7zTLgC.

Fred S. Robert. *Measurement Theory*. 1985. ISBN 9780521102438.

Timothy J. Ross. *Fuzzy Logic with Engineering Applications*. 2010. ISBN 9780470743768. doi: 10.1002/9781119994374. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84891564455{&}partnerID=tZOtx3y1.

Derek Rowell. Signal processing: Continuous and discrete. MIT OpenCourseWare Lecture Notes, 2008.

A.G. Spirkin. *Dialectical Materialism*. Progress Publishers, 1983. URL https://www.marxists.org/reference/archive/spirkin/works/dialectical-materialism/index.html.

Eran Tal. Measurement in Science. In Edward N Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 201 edition, 2017.

The International System of Units. Resolution 1 of the 17th CGPM: Definition of the metre. In *The 17th Conférence Générale des Poids et Mesures (CGPM)*, volume 20, pages 25–30, Jan 1984.