Signals and Systems: Module 1

Suggested Reading: S&S 1.0, 1.1, 1.5, 1.6.6

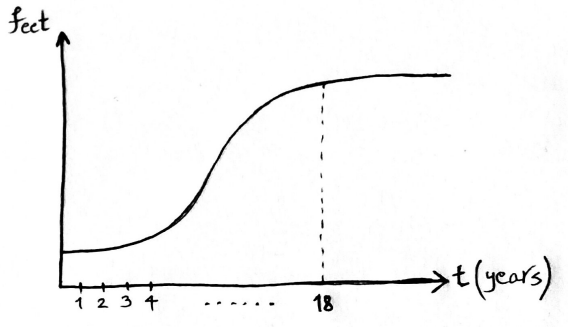
What are Signals and Systems?

The overall goal of this course is to conduct a mathematical/formal study of **signals** and **systems**. An electric current driving a circuit, a human heart beat driving a monitor, and a light driving a fiber cable each may produce very different results, but all have one fundamental feature in common: a signal driving a system. The interaction between signals and systems is pivotal to the society we know, and has been long studied by engineers, physicists, and mathematicians.

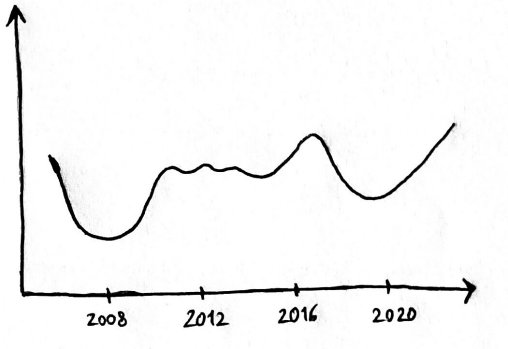
Signals

So, what are **signals**? Really, they can be anything that contains some information, whether for use by a human or a computer. A signal will encode information as some sort of a pattern of variations. Here are some examples:

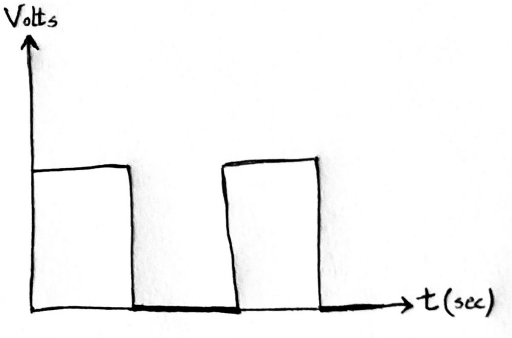
1. A person's height by time (years) since birth



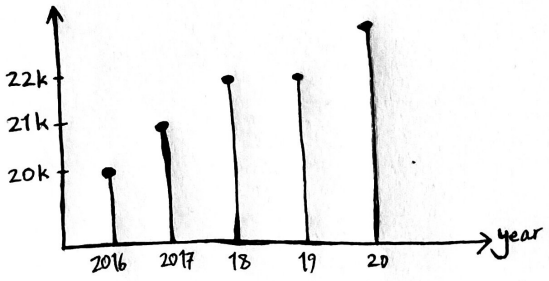
2. Dow Jones stock index



3. Voltage waveform on a circuit



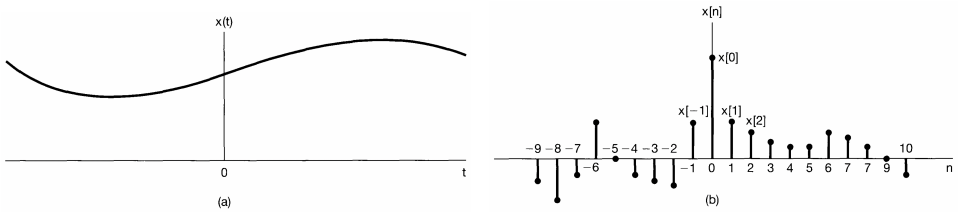
4. Purdue engineering students enrolled each year



More formally, we distinguish between **continuous-time (CT)** and **discrete-time (DT)** signals:

- CT signals have a continuous independent variable, indexed by $t \in \mathbb{R}$ (real numbers).
- DT signals have a discrete independent variable, indexed by $n \in \mathbb{N}$ (integers).

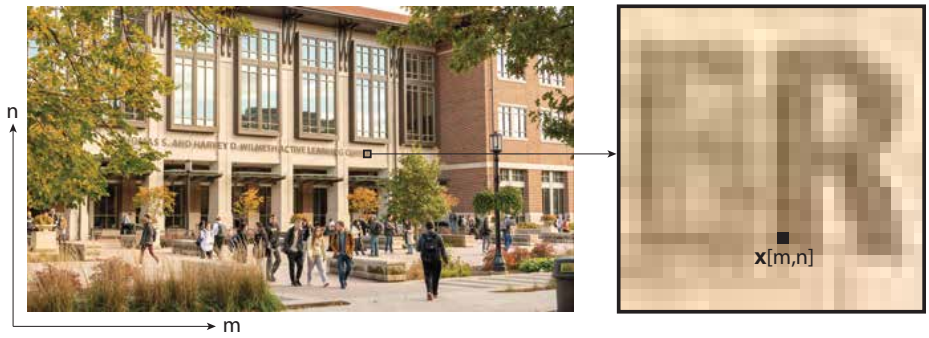
Mathematically, we use the symbol t and parentheses $(\cdot)$ to denote continuous time, whereas we use the symbol n and brackets $[\cdot]$ for discrete time. So while $x(t)$ will denote a continuous-time signal x at time t, $x[n]$ is a discrete-time signal at time n. Visually:



CT signal DT signal

It is important to note that $x[n]$ is only defined for integer values of n, e.g., $x[2]$ is defined, but $x[1.5]$ is not. To make this clear, sometimes we refer to $x[n]$ as a sequence rather than a signal.

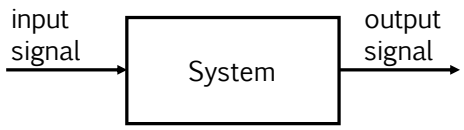
Another important notice is that although we call the signals continuous-time and discrete-time, the variable of the signals does not have to be in the time domain. Furthermore, the signals can have more than one variable. A very common example is photos. We show a photo in the picture below. A photo can be stored in many different formats in a computer. One widely used format is PNG, which stores the image as a two-dimensional DT signal $\boldsymbol{x}[m, n]$. We use the bold symbol $\boldsymbol{x}$ here because the signal value at each m and n is a 3-dimensional vector, which represents the red, green, and blue intensities of the photo at the position.



Systems

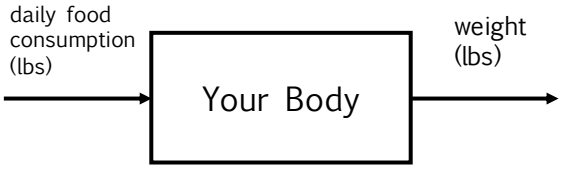
Then, what are **systems**? Broadly speaking, they are interconnections of subsystems, components, devices, or other entities. In the most general sense, a system is a process in which some input signals are transformed into some output signals.

There will be a lot of block diagrams in this course like this:



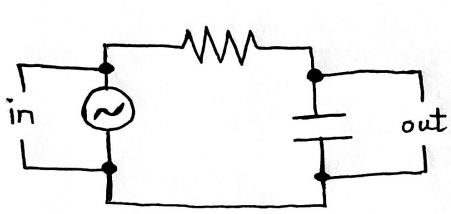
Here are some examples of systems, and potential inputs/outputs:

1. The human body

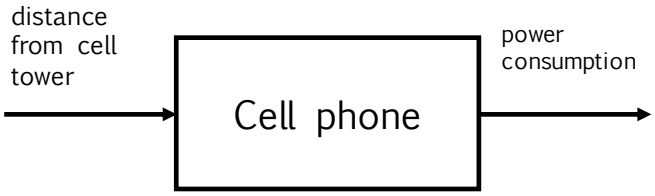


2. The stock market

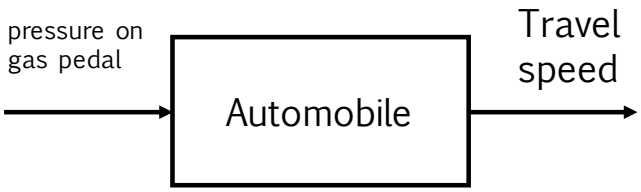
3. An RC circuit



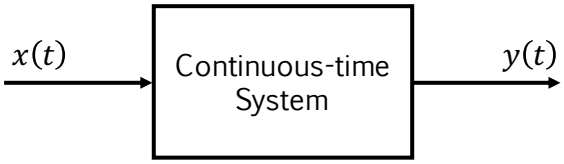
4. Cell phones



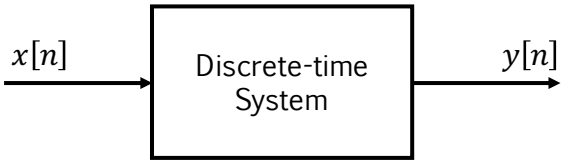
5. An automobile



A **continuous-time** system is one which has continuous-time input and output signals, e.g., a linear circuit:

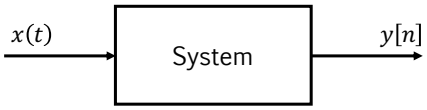


A **discrete-time** system is one which has a discrete-time input and output signals, e.g., a program on your computer:

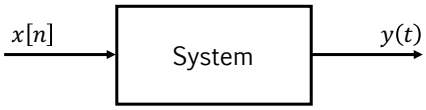


Mathematically, we often write $y(t) = \mathcal{S}(x(t))$ in CT or $y[n] = \mathcal{S}(x[n])$ in DT. $\mathcal{S}$ is used to denote the system function.

In most of this course, we will focus on CT and DT systems separately but in parallel. At the end of the course, we will see how they can be unified through the concepts of sampling and reconstruction. Indeed, we can have systems which are DT-in and CT-out, and which are CT-in and DT-out:



e.g., Audio recording

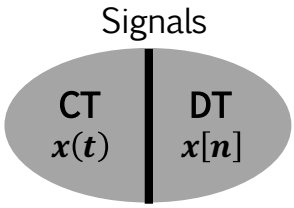


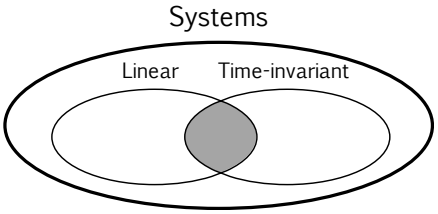
e.g., Audio playback

So, why should we study signals and systems? Is this just another math course? The answer is that the subject is fundamental to solving many engineering problems. It is a very practically-oriented math course.

- Often, we model a problem of interest as a "system" without really thinking about it: we expect some input (electrical, mechanical, biological, ...) and desire some output. This course will help write down mathematical descriptions of the input/output relationship.
- **Analyzing** a system usually involves studying various possible signals associated with the system.
- **Designing** a system requires determining a suitable system architecture as well as finding good system parameters.
- **Implementing** and **testing** the system involves checking the system under target input/output signals to see whether performance is satisfactory.

Of course, we will not be able to cover all aspects of signals and systems. We can consider the following high-level taxonomy:





In particular, while we cover both CT and DT signals, we will mostly be focused on systems that possess **linear** and **time-invariant** properties. These are practical in many settings, and emit several useful analytical tools:

- Fourier transform, Laplace transform, z-transform: Analyzing the frequency contents and stability of a system
- Convolution integral, convolution sum (as in convolutional neural networks in machine learning): Input-output relationship modeling
- Important applications: Filters, AM/FM radios, image processing, digital signal processing, quantization, ...

This is a high-cost, high-reward course. If you learn well in this course, you will develop a signal-and-system mindset that will offer a lot of valuable intuitions and insights to tackle almost all engineering problems throughout the rest of your life. But to learn well, you must spend sufficient time on the course materials. With this in mind, let's start our journey...

Linearity

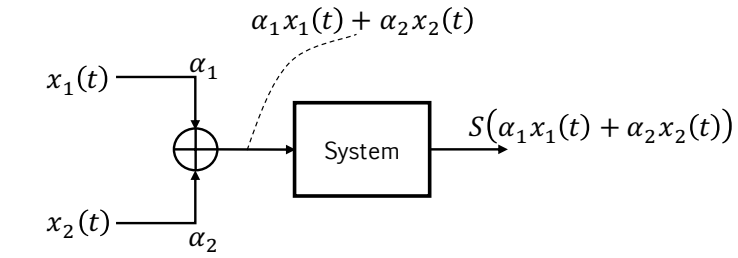
A system is **linear** if it possesses the important property **superposition**: given an input which is the weighted sum of several inputs, the system produces an output which is just the weighted sum of the individual outputs. Formally, the following are the (necessary and sufficient) conditions for linearity:

- *continuous time*: $\alpha_1 x_1(t) + \alpha_2 x_2(t) \to \alpha_1 y_1(t) + \alpha_2 y_2(t)$

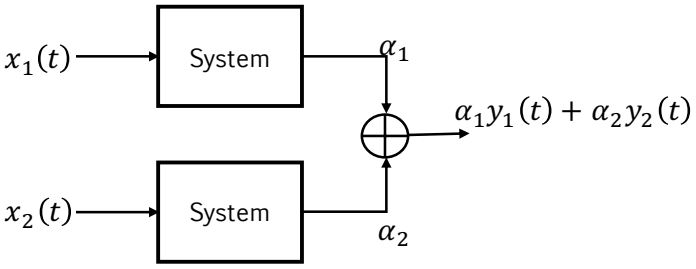
- *discrete time*: $\alpha_1 x_1[n] + \alpha_2 x_2[n] \to \alpha_1 y_1[n] + \alpha_2 y_2[n]$

which must hold for ***any and all*** possible constants α_1 and α_2.

Pictorially, linearity means the following two configurations are identical:



$=$

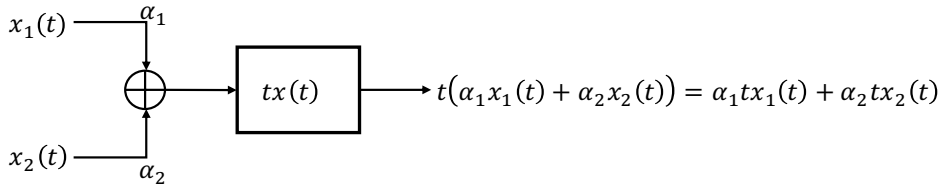


It is also easy to show that the definition of linearity generalizes to an arbitrary number of inputs.

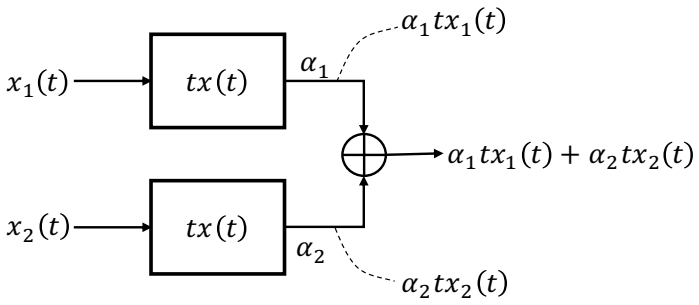
So, how do we check if a system is linear or not? Procedurally, we check whether the two configurations above give identical outputs.

Example 1. *Consider a CT system $\mathcal{S}$ given by $y(t) = tx(t)$. Is the system linear?*

Ans: For the first configuration (combining the inputs):



For the second configuration (combining the outputs):



These results are equivalent, so the system is ***linear****.*

Note: The linearity of a system is different from that of a signal. As in the above example, the output signal $y(t)$ can be nonlinear, but the system is still linear, as it follows the superposition property.

Importantly, note that when we test for linearity, we do not assign any values to x_1, x_2, α_1, or α_2. The results must hold for **all** input signals and combinations.

Example 2. *Consider a CT system $\mathcal{S}$ with $y(t) = x^2(t)$. Is the system linear?*

Ans: We have

$$x_1(t) \to y_1(t) = \mathcal{S}(x_1(t)) = x_1^2(t)$$
$$x_2(t) \to y_2(t) = \mathcal{S}(x_2(t)) = x_2^2(t)$$

and thus

$$\alpha_1 y_1(t) + \alpha_2 y_2(t) = \alpha_1 x_1^2(t) + \alpha_2 x_2^2(t).$$

On the other hand,

$$\mathcal{S}(\alpha_1 x_1(t)+\alpha_2 x_2(t)) = (\alpha_1 x_1(t) + \alpha_2 x_2(t))^2 = \alpha_1^2 x_1^2(t)+2\alpha_1\alpha_2 x_1(t)x_2(t)+\alpha_2^2 x_2^2(t)$$

Thus, $\alpha_1 y_1(t) + \alpha_2 y_2(t) \neq \mathcal{S}(\alpha_1 x_1(t) + \alpha_2 x_2(t))$. The system is ***not linear****.*

Example 3. *Consider a DT system $\mathcal{S}$ with $y[n] = ax[n] + b$. Is the system linear?*

Ans: For combining outputs, we have

$$\left.\begin{aligned} x_1[n] \to y_1[n] = ax_1[n] + b \\ x_2[n] \to y_2[n] = ax_2[n] + b \end{aligned}\right\} \alpha_1 y_1[n]+\alpha_2 y_2[n] = a(\alpha_1 x_1[n]+\alpha_2 x_2[n])+b(\alpha_1+\alpha_2)$$

On the other hand, for combining inputs,

$$\alpha_1 x_1[n] + \alpha_2 x_2[n] \to a(\alpha_1 x_1[n] + \alpha_2 x_2[n]) + b.$$

Thus, $\alpha_1 y_1[n] + \alpha_2 y_2[n] \neq \mathcal{S}(\alpha_1 x_1[n] + \alpha_2 x_2[n])$*. The system is* ***not linear****, even though its expression is a linear equation!*

By contrast, the system $y[n] = ax[n]$ *is linear. The offset term* b *creates a non-linearity. We could "linearize" the system by restricting it to only the difference between two inputs.*

Example 4. *Consider a DT system where*

$$y[n] = \begin{cases} x[n] & n \geq 1 \\ 0 & n = 0 \\ x[n+1] & n \leq -1 \end{cases}$$

Is the system linear?

Ans: We have

$$\alpha_1 x_1[n] + \alpha_2 x_2[n] \to \begin{cases} \alpha_1 x_1[n] + \alpha_2 x_2[n] & n \geq 1 \\ 0 & n = 0 \\ \alpha_1 x_1[n+1] + \alpha_2 x_2[n+1] & n \leq -1 \end{cases}$$

which can be further separated according to each interval of n*:*

$$\alpha_1 x_1[n] + \alpha_2 x_2[n] \to \alpha_1 \cdot \underbrace{\begin{cases} x_1[n] & n \geq 1 \\ 0 & n = 0 \\ x_1[n+1] & n \leq -1 \end{cases}}_{y_1[n]} + \alpha_2 \cdot \underbrace{\begin{cases} x_2[n] & n \geq 1 \\ 0 & n = 0 \\ x_2[n+1] & n \leq -1 \end{cases}}_{y_2[n]}$$

Thus, the system is ***linear****.*

Importance of linearity. Why do we care if a system is linear? Throughout this course, we will see it gives many useful properties. One is that it gives us an alternative way to compute the output.

Specifically, suppose we know some "test signals" $x_1, \cdots, x_n$ and the corresponding outputs $y_1, \cdots, y_n$. When we want to find the output for a new input signal x, we have two options:

① Direct computation, $x \to y$, i.e., passing x through the system.

② Write x as a linear combination of $x_1, \cdots, x_n$: $x = \alpha_1 x_1 + \cdots \alpha_n x_n$. If the system is linear, we can compute the output as $y = \alpha_1 y_1 + \cdots + \alpha_n y_n$.

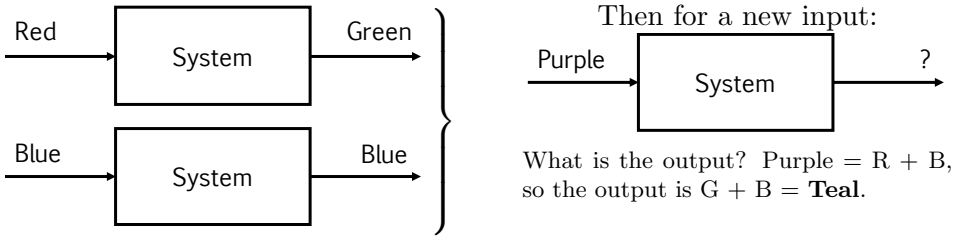
Why might ② be preferable to ①? A number of potential reasons:

- Many systems are "black boxes". Thus, it may be *hard to access* the mechanism inside the system to directly compute the output for any arbitrary x. It also may cost money to keep re-running the system. On the other hand, we may be able to readily find the weights $\alpha_1, \cdots, \alpha_n$ so that we cab "assemble" the new output y without knowing what is in the black box.

  For example, suppose we have a large electrical circuit comprised of hundreds of elements. If we know all of the elements are linear, then we may prefer computing the expected output of new inputs from combinations of test inputs rather than running through the computations across all the elements.

- These test signals can help us *understand the behavior* of a linear system even before applying the real signals of interest.

  For example, suppose we have an image processing program that we know is linear. We have "test" signals of red and blue pixels:



In summary, for linear systems, once we know the outputs of the test signals, we know how to construct the output of "any" signal. The caveat is that we have to choose good/convenient test signals that will work for all cases. R, G, B may be good choices for images, but what about more general systems? Our consideration of this question will later lead us to the **convolution integral**, which is based on a special single test input.

Open-ended Question

With the definition of linearity in mind, is the camera or the microphone in your smartphone a linear system or not? You may find the answer through some basic experiments.

Classifying Signal Types

In the next few lectures, we will introduce a few different classifications of signals. Each signal class has a unique set of properties that help us reason about their impacts on systems.

Classification #1: Discrete-time (DT) vs. continuous-time (CT).

We have seen this one already. But, we must also keep in mind that signals of both types can be either **real-valued** or **complex-valued**.

- **Discrete-time**: The signal $x[n]$ is a sequence of either real-valued or complex-valued numbers. For example:

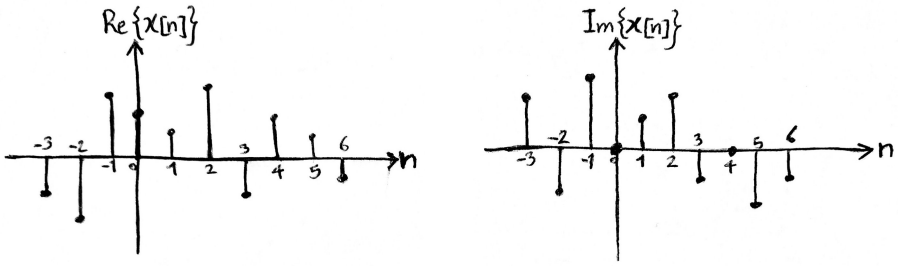
	Real-valued example	Complex-valued example
$x[n] =$	$\frac{n}{2}$	$\frac{n}{2} + (1-2n)j$
$\downarrow$	$\downarrow$	$\downarrow$
$x[0] =$	0	$0 + j$
$x[1] =$	$\frac{1}{2}$	$\frac{1}{2} - j$

  In general, we represent a complex signal in terms of its real and imaginary components: $x[n] = x_{Re}[n] + jx_{Im}[n]$. In the example above, $x_{Re}[n] = n/2$ and $x_{Im}[n] = (1-2n)$. Complex signals can also be expressed in magnitude and phase form, related to the real and imaginary components through Euler's formula:

$$x[n] = \underbrace{A[n]}_{\text{mag}} e^{j \overbrace{\phi[n]}^{\text{phase}}} = \underbrace{A[n]\cos(\phi[n])}_{\text{real}} + j \underbrace{A[n]\sin(\phi[n])}_{\text{imaginary}}$$

  See the *Math Review* supplement posted on Brightspace for a more detailed recap of complex numbers.

We often visualize a complex signal with two graphs, one for the real component and one for the imaginary component (or magnitude and phase). For example, a complex DT signal may look like this:

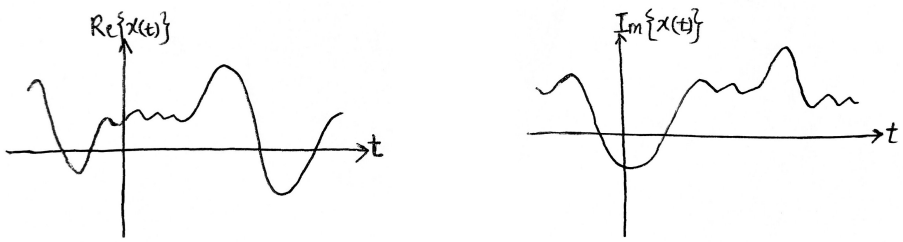


- **Continuous-time (CT)**

  Similarly, CT signals can be real or complex. For example:

Real-valued example	Complex-valued example
$x(t) = \frac{t}{2}$	$x(t) = \frac{t}{2} + (1-2t)j = x_{Re}(t) + jx_{Im}(t)$
$\downarrow$	$\downarrow$
$x(\frac{1}{3}) = \frac{1}{6}$	$x(\frac{1}{3}) = \frac{1}{6} + \frac{1}{3}j$
$x(\pi) = \frac{\pi}{2}$	$x(\pi) = \frac{\pi}{2} + (1-2\pi)j$

  As in DT, we can visualize complex CT signals with two graphs, except now they are defined over t instead of n:



Classification #2: We can also characterize signals by **energy** and **power**.

- **Energy** in signal processing is the area under the squared magnitude, which is related to how "costly" it is to store or transmit the signal.

For CT signals, the energy over time interval $t \in [t_1, t_2]$ is

$$E = \int_{t_1}^{t_2} |x(t)|^2 \, dt = \underbrace{\int_{t_1}^{t_2} \left(x_{Re}^2(t) + x_{Im}^2(t) \right) \, dt}_{\text{Follows since } |a+jb|^2 = a^2+b^2}$$

Note: As this is the first time we use integral in the course, please remember the integral interval is **closed**, i.e., the interval is $[t_1, t_2]$. Whether the interval is open or closed will not change the value of the integral as long as $x(t)$ is finite at t_1 and t_2, but it will make a difference later in the class when we talk about impulse responses.

For DT signals, the energy over time interval $n \in [n_1, n_2]$ is

$$E = \sum_{n=n_1}^{n_2} |x[n]|^2 = \sum_{n=n_1}^{n_2} \left(x_{Re}^2[n] + x_{Im}^2[n] \right)$$

For **total energy**, we consider the interval to be all time $(-\infty, \infty)$:

$$E_\infty = \int_{-\infty}^{\infty} |x(t)|^2 \, dt \text{ for CT}, \qquad E_\infty = \sum_{n=-\infty}^{\infty} |x[n]|^2 \text{ for DT}.$$

- **Power** is energy per unit time.

  For CT signals, the **average power** over $[t_1, t_2]$ is

  $$P = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} |x(t)|^2 \, dt.$$

  For DT signals, the average power over $[n_1, n_2]$ is

  $$P = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} |x[n]|^2.$$

  The **overall average power** is expressed as a limit:

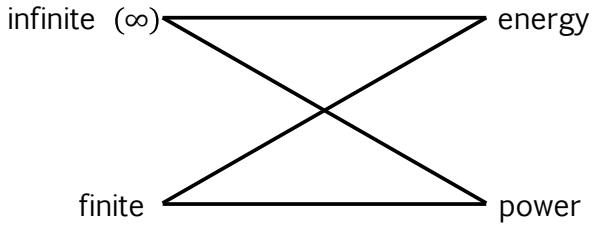
  $$P_\infty = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 \, dt \text{ for CT},$$

  $$P_\infty = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{-N}^{N} |x[n]|^2 \text{ for DT}.$$

  We can also define **instantaneous power** by

  $$|x(t)|^2 \text{ or } |x[n]|^2$$

We are often interested in characterizing signals based on E_∞ and P_∞. There are four different types (potentially): (1) infinite total energy, infinite overall average power; (2) infinite total energy, finite overall average power; (3) finite total energy, infinite overall average power; and (4) finite total energy, finite overall average power.



However, one of these combinations is not realistic. To see this, consider what happens when we have finite total energy: for example, if the total energy is 3, what is the overall average power?

$$P_\infty = \lim_{T\to\infty} \underbrace{\frac{1}{2T}}_{0} \underbrace{\int_{-T}^{T} |x(t)|^2\, dt}_{3} = 0.$$

On the other hand, consider a finite (non-zero) overall average power: if the overall average power is 3, what is the total energy? With a slight abuse of notation,

$$\frac{E_\infty}{\infty} = 3 \rightarrow E_\infty = \infty.$$

The implication is that we **do not** have finite energy, ∞-power signals. However, we can have the following three types:

1. E_∞ **finite:** Signals with finite total energy must have $P_\infty = 0$, i.e., zero overall average power, since $\lim_{T\to\infty} E_\infty/2T = \lim_{N\to\infty} E_\infty/2N = 0$.

2. P_∞ **finite:** Signals with finite (and non-zero) overall average power must have $E_\infty = \infty$, since accumulating a non-zero energy-per-time over an infinite time interval yields an infinite amount of energy.

3. **Neither** E_∞ **nor** P_∞ **finite:** It is also possible for signals to have infinite total energy AND infinite overall average power, though these signals are rarely practical. An example of this would be $x(t) = t$.

It is important to note that we are using "power" and "energy" here irrespective of whether these definitions actually correspond to physical power and energy. Even when such a relationship does exist, these equations may not have the right dimensions and scalings. For example, if $x(t)$ represents current through a resistor, P and E would need to be multiplied by resistance to obtain units of physical power and energy.

Example 5. *Consider $x(t) = cos(t) + jsin(t)$. What is the average power and energy over the interval $[-\frac{1}{2}, 1]$?*

Ans: We calculate the energy and power as

$$E = \int_{-\frac{1}{2}}^{1} |cos(t) + jsin(t)|^2 \, dt = \int_{-\frac{1}{2}}^{1} \left(cos^2(t) + sin^2(t)\right) \, dt = \int_{-\frac{1}{2}}^{1} dt = \frac{3}{2}$$

$$P = \frac{1}{t_2 - t_1} \cdot E = \frac{1}{\frac{3}{2}} \cdot \frac{3}{2} = 1$$

Since $|x(t)|^2 = 1$ for all time, the instantaneous power and average power over any time interval are also going to be 1. Thus, the total average power is $P_\infty = 1$. This means the total energy must be $E_\infty = \infty$.

Example 6. *Determine the values of P_∞ and E_∞ for the discrete-time signal*

$$x[n] = \begin{cases} \left(\frac{1}{4}\right)^n & n \geq 0 \\ 0 & n < 0 \end{cases}.$$

Ans: Starting with E_∞, we have

$$E_\infty = \sum_{n=-\infty}^{\infty} \left|x[n]\right|^2 = \sum_{n=-\infty}^{-1} |0|^2 + \sum_{n=0}^{\infty} \left| \left(\frac{1}{4}\right)^n \right|^2 = \sum_{n=0}^{\infty} \left(\frac{1}{4}\right)^{2n}$$

Noting that $\sum_{n=0}^{\infty} r^n = 1/(1-r)$ for $|r| < 1$ (see the Math Review for a recap on series summations),

$$E_\infty = \sum_{n=0}^{\infty} \left(\frac{1}{4}\right)^{2n} = \sum_{n=0}^{\infty} \left(\frac{1}{16}\right)^{n} = \frac{1}{1 - \frac{1}{16}} = \frac{16}{15}$$

Since $E_\infty < \infty$, it follows that $P_\infty = 0$. We can also see this by noting that in the formula

$$P_\infty = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=0}^{N} \left(\frac{1}{16}\right)^{n},$$

the denominator is linear in N while the numerator $\sum_{n=0}^{\infty} (1/16)^n$ converges.

Example 7. *What power-energy classification does the signal*

$$x(t) = \begin{cases} e^{-2(t-1)} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

fall into?

Ans: The total energy is

$$E_\infty = \int_{-\infty}^{0} 0dt + \int_{0}^{\infty} \left|e^{-2(t-1)}\right|^2 dt = \int_{0}^{\infty} e^{-4(t-1)} dt = \frac{1}{4} e^{-4(t-1)} \Big|_{\infty}^{0} = \frac{e^4}{4}$$

Since E_∞ is finite, it follows that $P_\infty = 0$.

Before moving on to more classifications, let us briefly digress to discuss the "algebra of signals" (which we have been using already without thinking about it). Signals are functions, and can be combined accordingly. So given two signals x_1 and x_2 (can be $x_1(t)$, $x_2(t)$, or $x_1[n]$, $x_2[n]$) we can write new signals based on them as functions:

- $y = x_1 + x_2$ means $y(t) = x_1(t) + x_2(t)$ for all t.
- $y = \alpha x_1$ means $y[n] = \alpha x[n]$ for all n.
- $y = x_1 \cdot x_2^2$ means $y(t) = (x_1(t)) \cdot (x_2(t))^2$.

These operations are used/implemented quite often in real systems. Amplifiers in a linear circuit attempt to replicate $y = \alpha x$ for $\alpha > 1$. Modulation in communications often consists of multiplying two signals: $y = x_1 \cdot x_2$.