Digital Communication Systems

S Le Goff

October 2023

School of Engineering

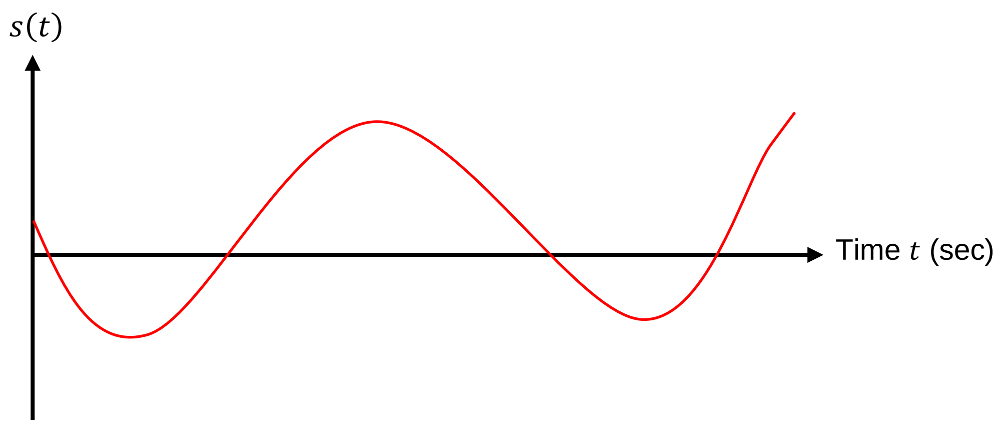
Newcastle University

Contents

1. Basic Definitions

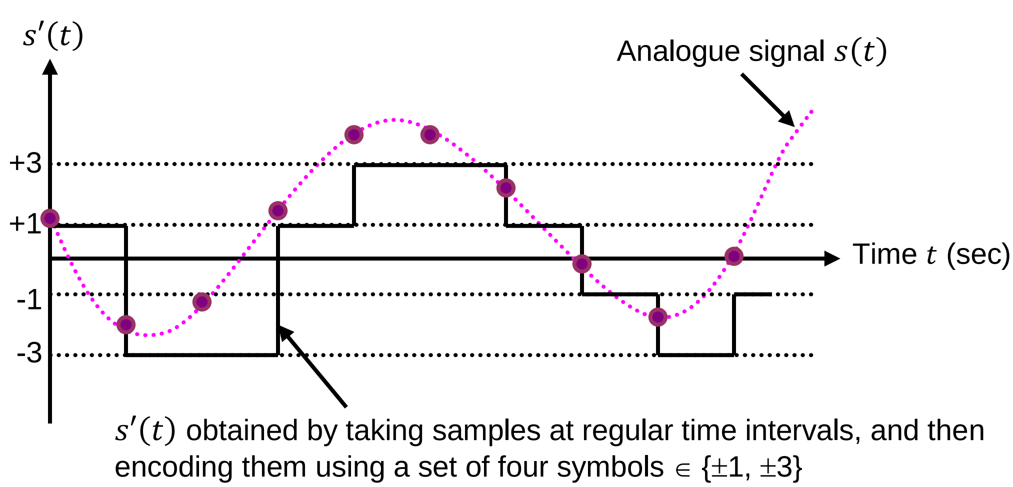
• Analogue signal

An analogue signal $s(t)$ is defined as a physical time-varying quantity and is usually smooth and continuous, e.g., acoustic pressure variation when speaking.



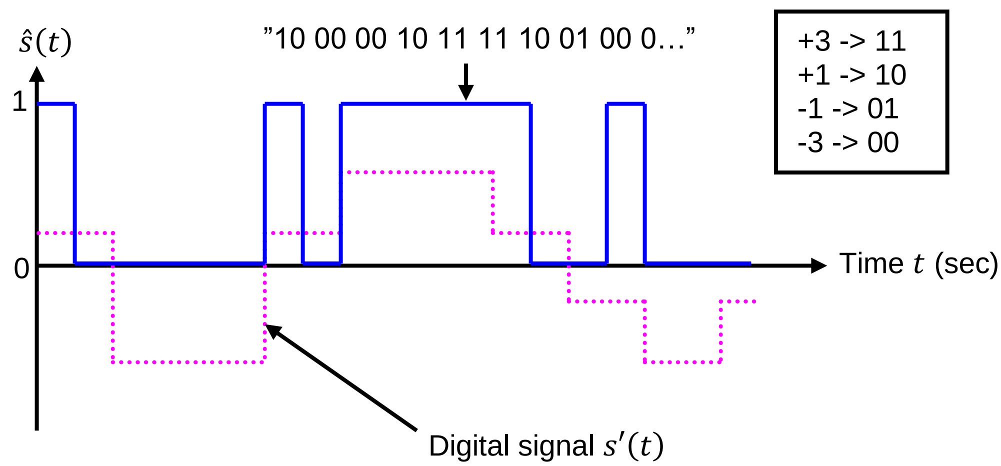
• Digital signal

A digital signal, on the other hand, is made up of discrete symbols selected from a finite set, e.g., letters from the alphabet or binary data. For example, the previous analogue signal $s(t)$ can be converted into a four-level digital signal $s'(t)$ as shown in the figure below.



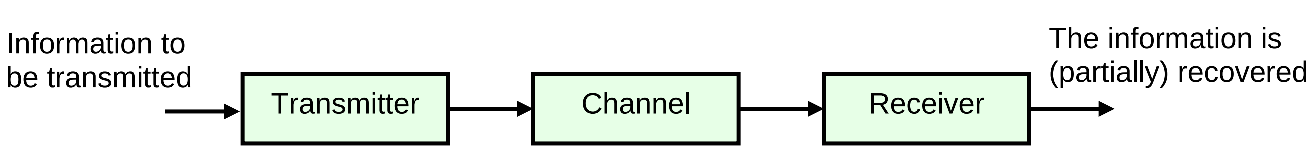
The digital signal $s'(t)$ can be seen as a vector of 4-ary real symbols $s_n \in \{-3, -1, +1, +3\}$, $n \in \{0, 1, 2, 3, \ldots\}$. A particular symbol s_n is obtained by quantizing the sample $s'(nT)$, where T denotes the sampling period. Since the number of values taken by symbols s_n is finite, these symbols are said to be discrete.

In many systems, the digital signal is a binary signal in the sense that it can take only two values: 0 or 1. For example, the digital signal $s'(t)$ depicted above can be converted into a binary digital signal $\hat{s}(t)$ as shown below.



The main advantage of digital systems over analogue systems is that they are far more resistant to noise. In addition, digital systems allow for the use of powerful digital processing techniques such as error correction/detection, data compression, easier data multiplexing, etc.

• General structure of a communication scheme



The transmitter element in a communication system processes the message signal in order to produce a signal most likely to pass reliably and efficiently through the channel. This usually involves modulation of a carrier signal by the message signal, coding of the signal to help correct for transmission errors, filtering of the message or modulated signal to limit the occupied bandwidth, and power amplification to overcome channel losses.

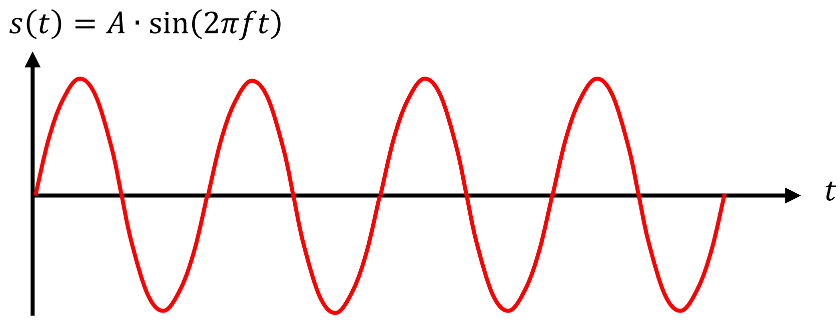
The channel is defined as the electrical medium between source and destination, e.g. cable, optical fiber, free space, etc. Any channel is characterized by its loss/attenuation, bandwidth, noise/interference and distortion.

The receiver function in a communication system is principally to reverse the modulation processing of the transmitter in order to recover the message signal, attempting to compensate for any signal degradation introduced by the channel. This will normally involve amplification, filtering, demodulation and decoding, and, in general, is a more complex task than the transmit processing.

2. Deterministic Signals

A signal is defined as any sign, gesture, token, etc. that serves to communicate information.

A signal is said to be deterministic if its future values can be predicted. Therefore, deterministic signals do not carry information, but they are used by transmitters to carry information. An example of deterministic signal is $s(t) = A \cdot \sin(2\pi f t)$.



- **Periodic signals**

A deterministic signal s(t) is said to be periodic if $s(t) = s(t + nT)$, where n is an integer and T (in sec) is the period of the signal. Periodic signals are eternal.

The frequency f of a periodic signal is given by $f = \frac{1}{T}$ and is expressed in Hertz (Hz).

The mean value, m, of a periodic signal can be computed using

$$m = \frac{1}{T} \cdot \int_{t_0}^{t_0+T} s(t)dt,$$

while its power P is given by

$$P = \frac{1}{T} \cdot \int_{t_0}^{t_0+T} [s(t)]^2 dt.$$

Note that the parameter t_0 can be chosen arbitrarily in these expressions.

The total energy of a periodic signal is infinite. Periodic signals are sometimes called *power signals*.

Example

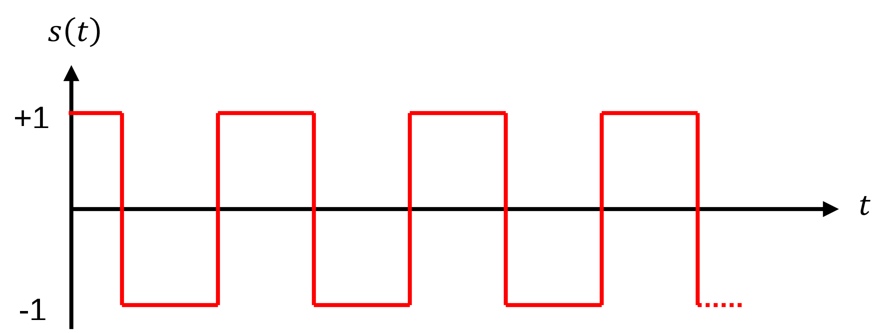
The power of the periodic signal $s(t) = A \cdot \cos(2\pi f t)$ is given by

$$P = \frac{1}{T} \cdot \int_0^T [A \cdot \cos(2\pi f t)]^2 dt = \frac{A^2}{T} \cdot \int_0^T \cos^2(2\pi f t)\, dt = \frac{A^2}{2T} \cdot \left[\int_0^T dt + \int_0^T \cos(4\pi f t)\, dt\right] = \frac{A^2}{2T} \cdot \left[[\ t\ ]_0^T + \left[\frac{\sin(4\pi f t)}{4\pi f}\right]_0^T\right] = \frac{A^2}{2T} \cdot \left[T + \left(\frac{\sin(4\pi f T) - \sin(4\pi f 0)}{4\pi f}\right)\right] = \frac{A^2}{2}.$$

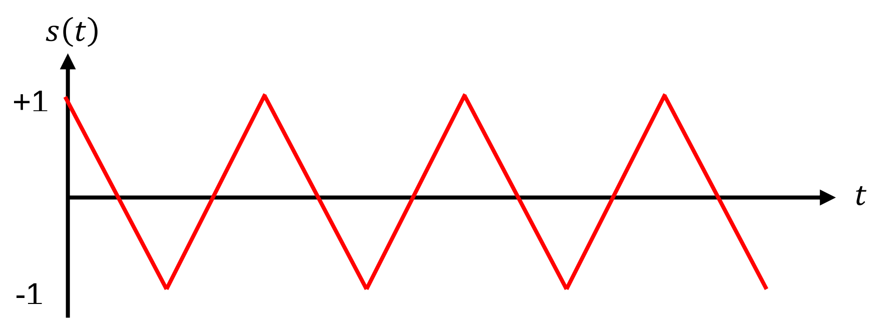
Note that the power of a sine wave is proportional to the square of the signal amplitude.

Examples of periodic signals

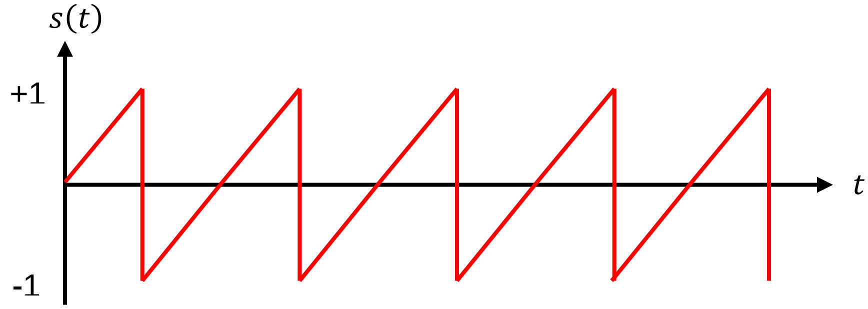
Square signal:



Triangular signal:



Saw-tooth signal:



- **Orthogonal periodic signals**

Two periodic signals $s_1(t)$ and $s_2(t)$ are said to be orthogonal if $\int_{t_0}^{t_0+T} s_1(t) \cdot s_2(t) dt = 0$.

Orthogonal signals are useful in many applications. For instance, by *attaching* different signals to orthogonal periodic signals, it is possible to *mix* these signals and still be able to separate them later.

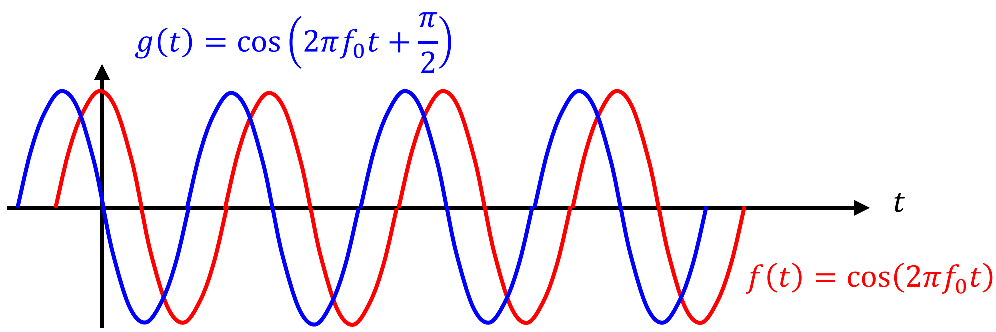
Example: Digital communication systems ---

Digital communication systems allow for the transmission of binary digits (bits) from a transmitter to a receiver. In these communication schemes, the information that is sent by the transmitter actually consists of a stream of discrete complex symbols in which these bits are embedded.

Let $s_n = s_{r,n} + js_{i,n}$ be the complex symbol transmitted between times $n \cdot T$ and $(n+1) \cdot T$, where T denotes the duration of a symbol and n is an integer. Each symbol s_n can take M different values and thus carry $\log_2(M)$ bits.

To optimize the use of the frequency spectrum, the real and imaginary parts of s_n are transmitted over the channel at the same time by using the same radio wave. This is done by embedding both symbols $s_{r,n}$ and $s_{i,n}$ in the same carrier signal during transmission and separating them afterwards at the receiver side.

To this end, we use two orthogonal periodic signals as follows: the two symbols $s_{r,n}$ and $s_{i,n}$ are used to modulate the amplitudes of the sinusoidal functions $f(t) = \cos(2\pi f_0 t)$ and $g(t) = \cos\left(2\pi f_0 t + \frac{\pi}{2}\right)$, respectively. Both functions $f(t)$ and $g(t)$ represent the same radio carrier signal of frequency f_0. Note that the carrier frequency f_0 is a multiple of $\frac{1}{T}$, with $f_0 \gg \frac{1}{T}$.

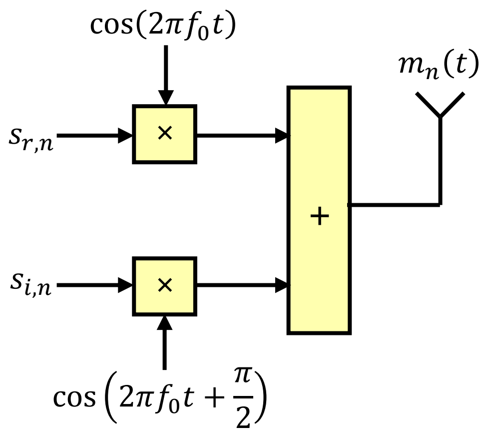


The functions $f(t)$ and $g(t)$ are orthogonal since $\int_{t_0}^{t_0+T} f(t) \cdot g(t) dt = \int_{t_0}^{t_0+T} \cos(2\pi f_0 t) \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right) dt = \frac{1}{2}\int_{t_0}^{t_0+T} \cos\left(4\pi f_0 t + \frac{\pi}{2}\right) dt + \frac{1}{2}\int_{t_0}^{t_0+T} \cos\left(\frac{\pi}{2}\right) dt = 0$.

Over the time period of an individual symbol $s_n = s_{r,n} + j s_{i,n}$, the radio signal radiated by the transmit antenna can be written as

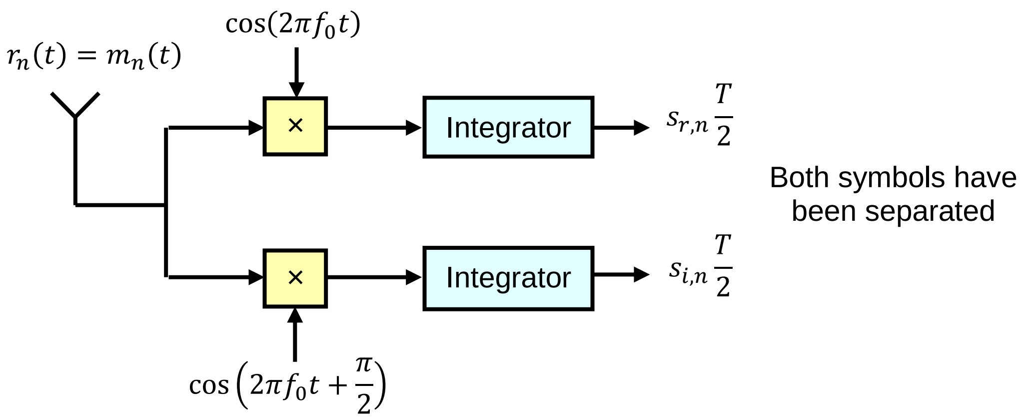
$$m_n(t) = s_{r,n} \cdot \cos(2\pi f_0 t) + s_{i,n} \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right).$$

The radio signal $m_n(t)$ is thus the sum of an "in-phase" component, $s_{r,n} \cdot \cos(2\pi f_0 t)$, and a "quadrature" component, $s_{i,n} \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right)$. These two components are thus "mixed" prior to transmission. The reason for doing so is that it allows for an efficient use of the frequency spectrum. In fact, the bandwidth required to transmit the combined signal $m_n(t)$ is identical to that needed to transmit a single component. In other words, by combining the in-phase and quadrature components in a single radio wave, we can transmit two real symbols instead of one by using the same amount of bandwidth.



At the receiver side, we can separate the symbols $s_{r,n}$ and $s_{i,n}$ by exploiting the orthogonality of the functions $f(t) = \cos(2\pi f_0 t)$ and $g(t) = \cos\left(2\pi f_0 t + \frac{\pi}{2}\right)$.

Assume that the received signal is $r_n(t) = m_n(t)$, thus implying that no noise was added to $m_n(t)$ by the channel. The recovery of $s_{r,n}$ is achieved by multiplying $r_n(t) = m_n(t)$ by $f(t) = \cos(2\pi f_0 t)$ and then integrating the result as follows: $\int_{n\cdot T}^{(n+1)\cdot T} m_n(t) \cdot f(t) dt = s_{r,n} \cdot \int_{n\cdot T}^{(n+1)\cdot T} \cos^2(2\pi f_0 t) dt + s_{i,n} \cdot \int_{n\cdot T}^{(n+1)\cdot T} \cos\left(2\pi f_0 t + \frac{\pi}{2}\right) \cdot \cos(2\pi f_0 t)\, dt = s_{r,n} \frac{T}{2}$.



In a similar fashion, to recover the symbol $s_{i,n}$, we multiply $r_n(t) = m_n(t)$ by $g(t) = \cos\left(2\pi f_0 t + \frac{\pi}{2}\right)$ and then integrate the result as follows: $\int_{n\cdot T}^{(n+1)\cdot T} m_n(t) \cdot g(t) dt = s_{r,n} \cdot \int_{n\cdot T}^{(n+1)\cdot T} \cos(2\pi f_0 t) \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right) dt + s_{i,n} \cdot \int_{n\cdot T}^{(n+1)\cdot T} \cos^2\left(2\pi f_0 t + \frac{\pi}{2}\right) dt = s_{i,n} \frac{T}{2}$.

In practice, the whole process is a bit more complicated because of the presence of noise in the received signal as well as the need for "attaching" both symbols to a suitable pulse signal, $h(t)$, before modulation in order to optimize the use of the frequency spectrum. The task of the receiver is therefore not only to separate symbols $s_{r,n}$ and $s_{i,n}$, but also to process the noisy received signal in order to minimize the probability of symbol detection failure. A consequence of this is that the integrator is actually replaced by a low-pass filter whose impulse response is matched to the pulse shape $h(t)$.

Note that a realistic expression for the received signal, between times $n \cdot T$ and $(n+1) \cdot T$, is, in the typical case of an additive white Gaussian noise (AWGN) channel, given by

$$r_n(t) = m_n(t) + n(t) = s_{r,n} \cdot h(t) \cdot \cos(2\pi f_0 t) + s_{i,n} \cdot h(t) \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right) + n(t),$$

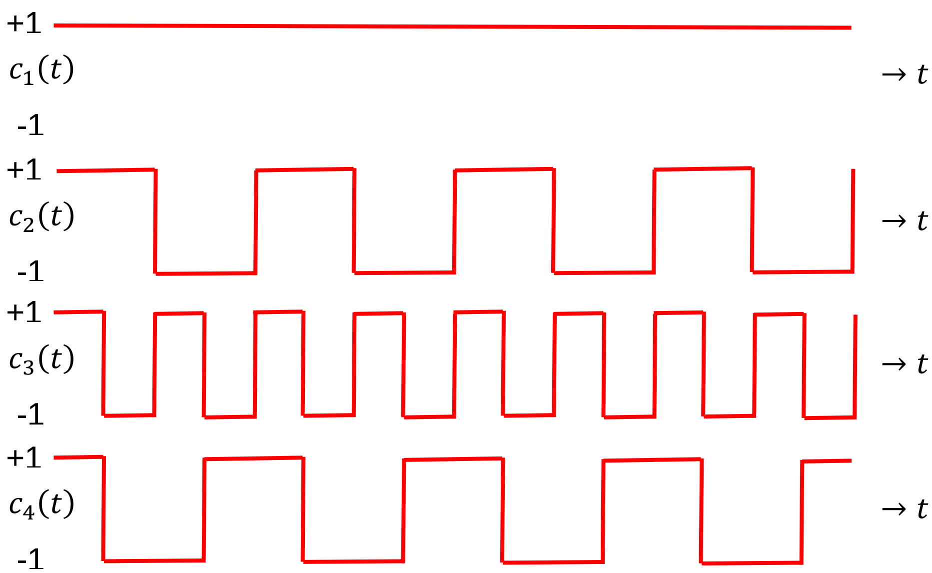
where $n(t)$ denotes a white Gaussian noise process.

Example: Code division multiple access (CDMA) ---

Multiple access refers to the techniques employed in wireless communication systems to enable several users to share the same channel without interfering. Code-division multiple access (CDMA), proposed for wireless telephony by the San Diego-based telecommunication company Qualcomm in the 1980s, is one of these techniques. In CDMA, each user is given, by the network, a different binary code chosen from a set of N orthogonal periodic codes.

Let us consider the simple example of a CDMA wireless system with N = 4 users. Let $s_{n,i}$ be the M-ary complex symbol transmitted by user $i \in \{1, 2, 3, 4\}$ between times $n \cdot T$ and $(n+1) \cdot T$, where T denotes the duration of a symbol and n is an integer.

Since there are four users, the network needs four different orthogonal codes. They are as follows: $c_1(t) = (+1, +1, +1, +1)$, $c_2(t) = (+1, +1, -1, -1)$, $c_3(t) = (+1, -1, +1, -1)$, and $c_4(t) = (+1, -1, -1, +1)$.

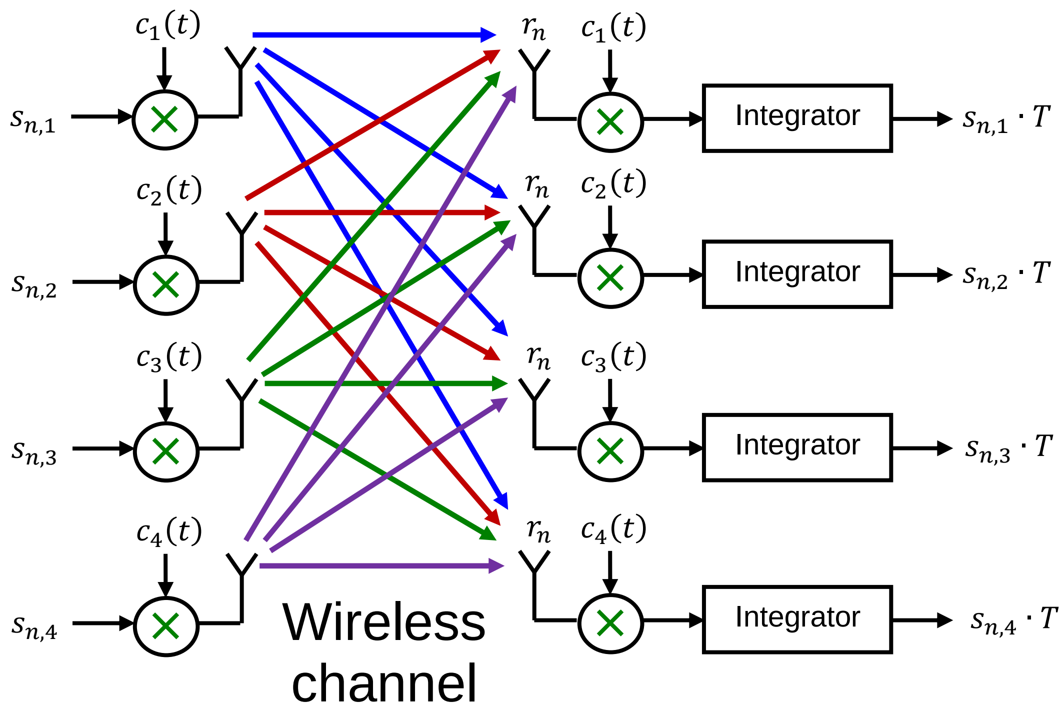


For instance, the notation $c_1(t) = (+1, +1, +1, +1)$ means that the pattern $(+1, +1, +1, +1)$ is repeated every T second, i.e., for each successive symbol to be transmitted, which implies that the codes are indeed periodic and the duration of one bit in the code is given by $\frac{T}{N} = \frac{T}{4}$ seconds.

We can see that $\int_{n \cdot T}^{(n+1) \cdot T} c_i(t) \cdot c_j(t) dt = 0$ for $i \neq j$ and $\int_{n \cdot T}^{(n+1) \cdot T} c_i^2(t) dt = T$ for any $i \in \{1, 2, 3, 4\}$.

We assume hereafter that the wireless network allocates the code $c_i(t)$, $i \in \{1, 2, 3, 4\}$, to a particular user i so that the latter can transmit its symbol from transmitter i to receiver i.

Between $n \cdot T$ and $(n + 1) \cdot T$, the information generated by a particular transmitter $i \in \{1, 2, 3, 4\}$ is therefore a four-symbol vector $s_{n,i} \cdot c_i(t)$, whereas the receivers of all users process the same channel observation given by $r_n = \sum_{j=1}^{N} s_{n,j} \cdot c_j(t)$. Here, we assume, for simplicity's sake, a noiseless channel with equal receive power for all users. The symbols generated by all transmitters are clearly "mixed" during transmission over the radio channel and the task of each receiver will be to separate the desired symbol from the three unwanted ones.



A particular receiver $i \in \{1, 2, 3, 4\}$ is able to retrieve symbol $s_{n,i}$ if and only if it has been given (by the network) the knowledge of the code $c_i(t)$ employed at the transmitter side. This receiver i multiplies the channel observation r_n by its allocated code $c_i(t)$ and then integrates the result in order to recover symbol $s_{n,i}$: $\int_{n\cdot T}^{(n+1)\cdot T} c_i(t) \cdot r_n dt = \int_{n\cdot T}^{(n+1)\cdot T} c_i(t) \cdot \sum_{j=1}^{N} s_{n,j} \cdot c_j(t)\, dt = \sum_{j=1}^{N} s_{n,j} \cdot \int_{n\cdot T}^{(n+1)\cdot T} c_i(t) \cdot c_j(t) dt = s_{n,i} \cdot T$.

To clarify this derivation, let us consider the example of the first receiver ($i = 1$) for which we we can write

$$\int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot r_n dt = \int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot \sum_{j=1}^{N} s_{n,j} \cdot c_j(t)\, dt$$

$$= \int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot \left[s_{n,1} \cdot c_1(t) + s_{n,2} \cdot c_2(t) + s_{n,3} \cdot c_3(t) + s_{n,4} \cdot c_4(t)\right] dt$$

$$= \int_{n\cdot T}^{(n+1)\cdot T} \left[s_{n,1} \cdot c_1(t) \cdot c_1(t) + s_{n,2} \cdot c_1(t) \cdot c_2(t) + s_{n,3} \cdot c_1(t) \cdot c_3(t) + s_{n,4} \cdot c_1(t) \cdot c_4(t)\right] dt$$

$$= \int_{n\cdot T}^{(n+1)\cdot T} s_{n,1} \cdot c_1(t) \cdot c_1(t)dt + \int_{n\cdot T}^{(n+1)\cdot T} s_{n,2} \cdot c_1(t) \cdot c_2(t)dt + \int_{n\cdot T}^{(n+1)\cdot T} s_{n,3} \cdot c_1(t) \cdot c_3(t)dt + \int_{n\cdot T}^{(n+1)\cdot T} s_{n,4} \cdot c_1(t) \cdot c_4(t)dt$$

$$= s_{n,1} \cdot \int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot c_1(t)dt + s_{n,2} \cdot \int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot c_2(t)dt + s_{n,3} \cdot \int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot c_3(t)dt + s_{n,4} \cdot \int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot c_4(t)dt$$

$$= s_{n,1} \cdot \int_{n\cdot T}^{(n+1)\cdot T} c_1(t) \cdot c_1(t)dt = s_{n,1} \cdot \int_{n\cdot T}^{(n+1)\cdot T} 1 \cdot dt = s_{n,1} \cdot T.$$

A receiver is able to recover symbols generated by a transmitter if and only if it has the knowledge of the code used by this particular transmitter. This simple example has shown that CDMA relies on the use of orthogonal periodic codes to allow for the sharing of a common channel by many users.

- **Finite-energy (or energy) signals**

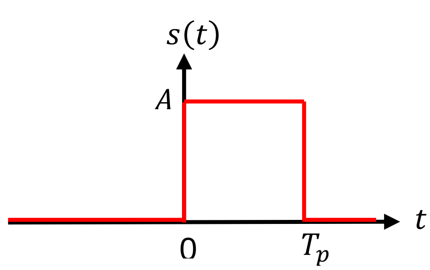
A deterministic signal is said to be finite-energy if it is time-limited. Pulses are a good example of such signals. The energy E of a finite-energy signal $s(t)$ is given by

$$E = \int_{-\infty}^{+\infty} [s(t)]^{2} dt.$$

Example ---

The energy of a square pulse $s(t)$ of amplitude A and duration T_p is given by

$$E = \int_0^{T_p} [A]^2 dt = A^2 \cdot T_p.$$



Note that this energy is proportional to the duration of the pulse and the square of its amplitude.

3. Time/Frequency Representations of Deterministic Signals

To understand the operation of any communication link, it is essential to have a good grounding in the relationship between the shape of a waveform in the *time domain* and its corresponding spectral content in the *frequency domain*.

Any signal $s(t)$ represented in the time domain has an equivalent representation $S(f)$ in the frequency domain. $S(f)$ is often referred to as the *spectrum* of $s(t)$. The spectrum shows the distribution of the signal energy/power as a function of frequency.

• Spectrum of a periodic signal: Fourier series

Assume that $s(t)$ is a periodic signal with period T. Fourier showed that such signal can be seen as an infinite sum of sine and cosine terms:

$$s(t) = a_0 + 2 \cdot \sum_{n=1}^{+\infty} \left[a_n \cdot \cos\left(\frac{2\pi n}{T} t\right) + b_n \cdot \sin\left(\frac{2\pi n}{T} t\right)\right],$$

where

- $a_0 = \frac{1}{T} \cdot \int_{-T/2}^{T/2} s(t) dt$ is the mean value of the periodic signal,
- $a_n = \frac{1}{T} \cdot \int_{-T/2}^{T/2} s(t) \cdot \cos\left(\frac{2\pi n}{T} t\right) dt$,
- $b_n = \frac{1}{T} \cdot \int_{-T/2}^{T/2} s(t) \cdot \sin\left(\frac{2\pi n}{T} t\right) dt$.

The term $f_1 = \frac{1}{T}$ is the fundamental frequency of the periodic signal, and the quantity $f_n = \frac{n}{T} = n \cdot f_1$ represents its $(n-1)$-th harmonic.

It is possible to re-write the equations in a more compact way as follows:

$$s(t) = a_0 + 2 \cdot \sum_{n=1}^{+\infty} \Re\left[c_n \cdot e^{-j2\pi f_n t}\right],$$

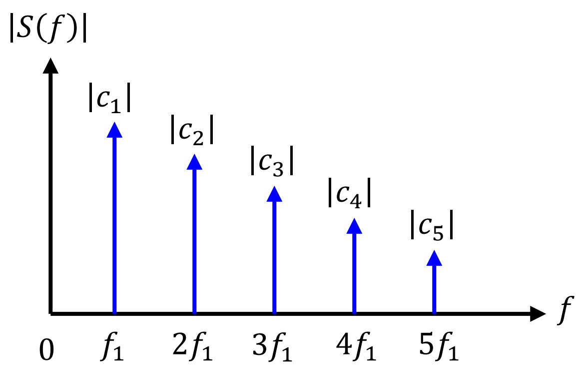
where $c_n = \frac{1}{T} \cdot \int_{-T/2}^{T/2} s(t) \cdot e^{+j2\pi f_n t} \, dt$.

The complex coefficient $c_n = a_n + jb_n$ is the complex amplitude of the $(n-1)$-th harmonic.

The representation of a periodic signal by a Fourier series is equivalent to the resolution of the signal into its various harmonic components. We see that a signal $s(t)$ with period T is composed of sine waves with frequencies 0, f_1, $2f_1$, $3f_1$ and so on. The frequency-domain description, i.e., the spectrum, of the signal therefore consists of components at frequencies 0, f_1, $2f_1$, $3f_1$ and so on. The spectrum of periodic signals is discrete.

If we specify the spectrum $S(f)$, we can determine the corresponding time-domain periodic signal $s(t)$, and vice-versa. It is important to understand that $s(t)$ and $S(f)$ are two different representations of a same signal.

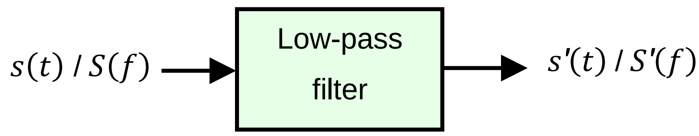
It is common practice to represent only the magnitude of each component c_n in the spectrum. However, one must not forget that the terms c_n in the Fourier series expansion are, generally, complex numbers, and hence both magnitude and phase of c_n are required for a complete representation of the signal.



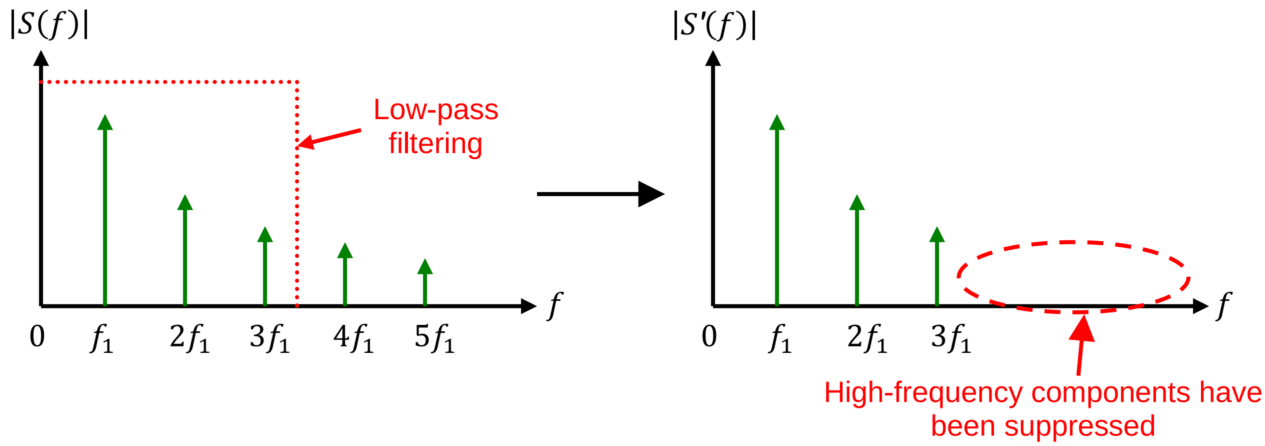
This discrete spectrum shows the distribution of the signal energy on the frequency axis.

If the signal is transmitted through a filter, some frequency components in the spectrum will be suppressed. This will introduce some distortion in the time-domain signal.

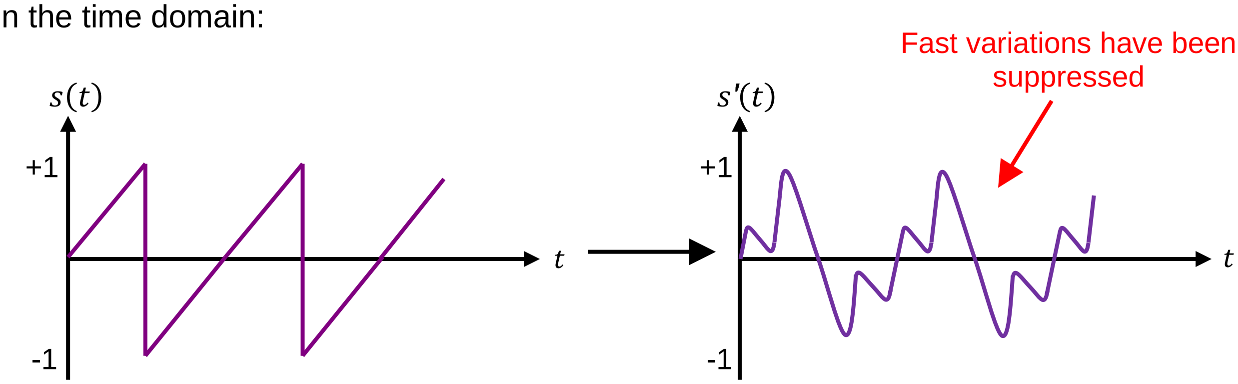
Example: Effect of low-pass filtering on a saw-tooth wave



In the frequency domain:



In the time domain:



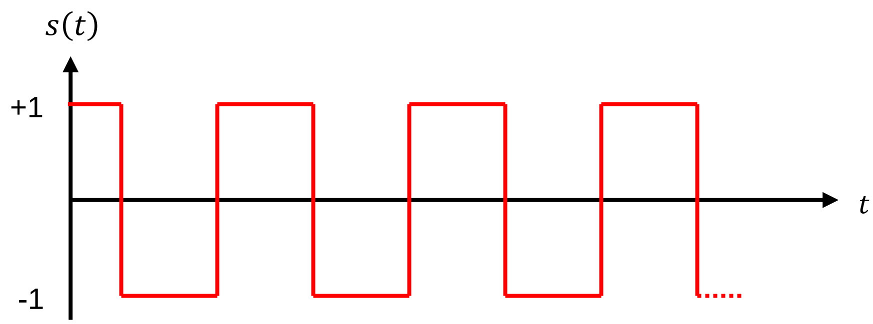
The bandwidth of a signal is defined as the frequency band occupied by its spectrum. Depending on their shape in the time domain, certain types of signals have a finite bandwidth (typically, signals with smooth variations, e.g., sine waves), while some others have an infinite bandwidth (typically, signals with infinitely sharp variations, e.g., square waves). Bandwidth is a scarce and valuable resource in communication engineering because the whole available spectrum must be shared among many users.

Examples of periodic signals and their Fourier series

$$s(t) = a_0 + 2 \cdot \sum_{n=1}^{+\infty} [a_n \cdot \cos(n\omega t) + b_n \cdot \sin(n\omega t)]$$

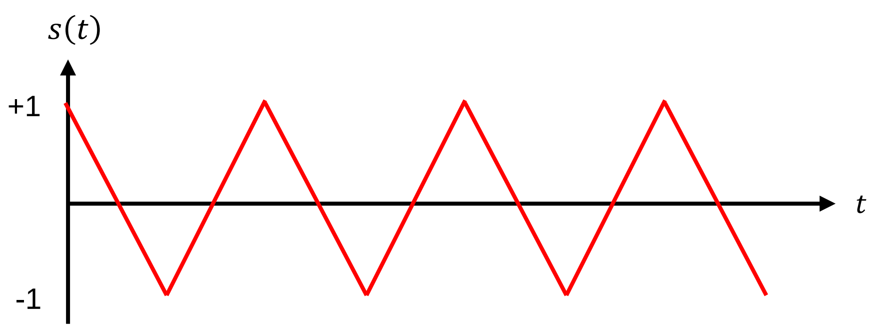
$$\omega = 2\pi f_1 = \frac{2\pi}{T}$$

Square signal:



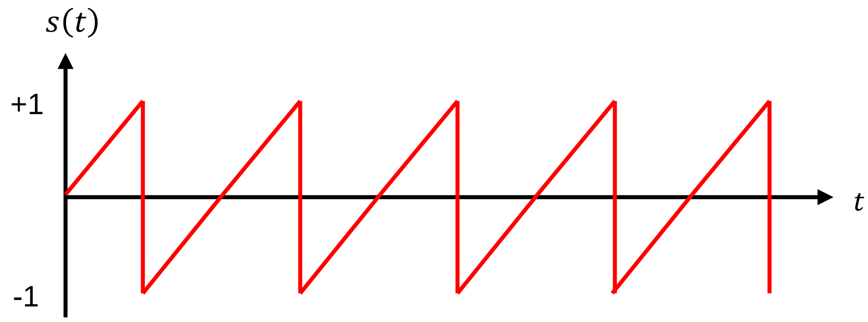
$$s(t) = \frac{4}{\pi} \cdot \left(\cos(\omega t) - \frac{1}{3}\cos(3\omega t) + \frac{1}{5}\cos(5\omega t) - \frac{1}{7}\cos(7\omega t) + \cdots\right)$$

Triangular signal:



$$s(t) = \frac{8}{\pi^2} \cdot \left(\cos(\omega t) + \frac{1}{9}\cos(3\omega t) + \frac{1}{25}\cos(5\omega t) + \frac{1}{49}\cos(7\omega t) + \cdots\right)$$

Saw-tooth signal:



$$s(t) = \frac{2}{\pi} \cdot \left(\sin(\omega t) - \frac{1}{2}\sin(2\omega t) + \frac{1}{3}\sin(3\omega t) - \frac{1}{4}\sin(4\omega t) + \cdots \right)$$

It is worthwhile noting that the spectrum of a constant (DC) signal is composed of a single component located at the zero frequency.

• Spectrum of a finite-energy signal: Fourier transform

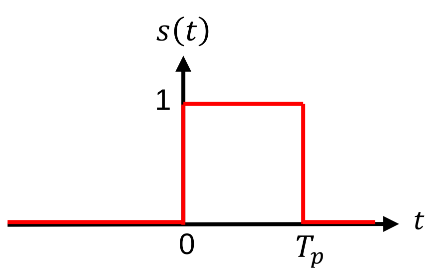
The frequency content of a deterministic finite-energy signal cannot be obtained by evaluating its Fourier series expansion because such signal is not periodic. The mechanism for obtaining the spectrum of a finite-energy signal is to calculate the Fourier transform of its time-domain representation as follows:

$$S(f) = \int_{-\infty}^{+\infty} s(t) \cdot e^{-j2\pi f t}\, dt.$$

This equation cannot be evaluated for infinite-energy signals, such as periodic signals for example. Unlike the spectrum obtained for periodic signals, the function $S(f)$ is now continuous.

Example ---

Square pulse



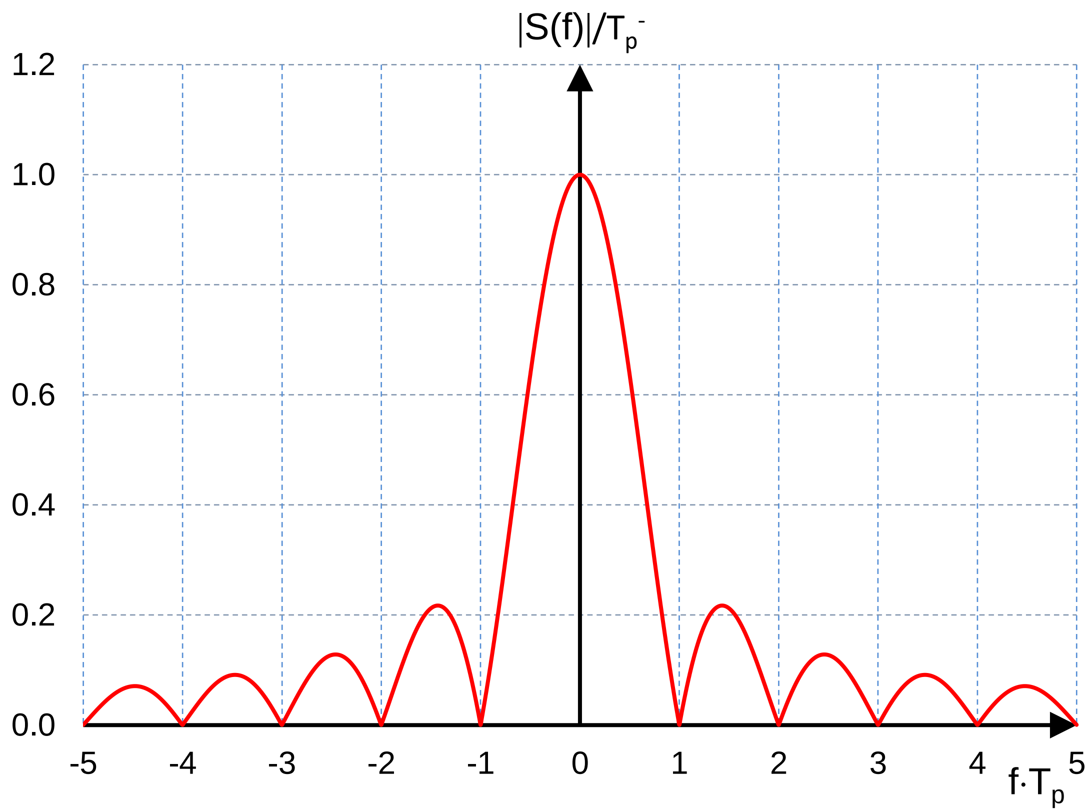
The Fourier transform of $s(t)$ is given by

$$S(f) = \int_{-\infty}^{+\infty} s(t) \cdot e^{-j2\pi ft}\, dt = \int_0^{T_p} e^{-j2\pi ft}\, dt = \left[\frac{e^{-j2\pi ft}}{-j2\pi f}\right]_0^{T_p} = -\frac{1}{j2\pi f} \cdot \left(e^{-j2\pi f T_p} - e^{-j2\pi f 0}\right) =$$

$$-\frac{1}{j2\pi f} \cdot \left(e^{-j\pi f T_p} - e^{+j\pi f T_p}\right) \cdot e^{-j\pi f T_p} = \frac{\sin(\pi f T_p)}{\pi f} \cdot e^{-j\pi f T_p} = T_p \cdot \mathrm{sinc}\left(\pi f T_p\right) \cdot e^{-j\pi f T_p}.$$

We are more particularly interested in the magnitude of the complex function $S(f)$ since it contains the valuable information related to the energy distribution versus frequency, while its argument only represents a delay in the time domain. Here, the magnitude of $S(f)$ is given by

$$S(f) = T_p \cdot \left|\mathrm{sinc}\left(\pi f T_p\right)\right|.$$



Strictly speaking, the bandwidth of the square pulse is infinite. This means that, if one attempts to transmit such pulse through a practical communication system, some (high) frequency components will be filtered out, which will result in distortion in the time-domain.

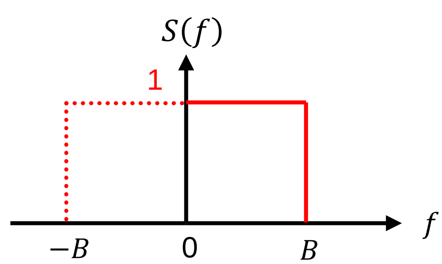
The factors affecting the bandwidth of a signal are both the shape and duration T_p of its time domain version. A smooth, slow-varying signal $s(t)$ tends to have a small bandwidth, while a sharp, fast-varying signal has a larger bandwidth. Also, the bandwidth of this pulse signal is reduced when its duration T_p is increased. Short pulses occupy a larger frequency band than longer ones.

If we know the spectrum of a signal, it is possible to obtain its time-domain representation by computing the inverse Fourier transform as follows:

$$s(t) = \int_{-\infty}^{+\infty} S(f) \cdot e^{+j2\pi ft}\, df.$$

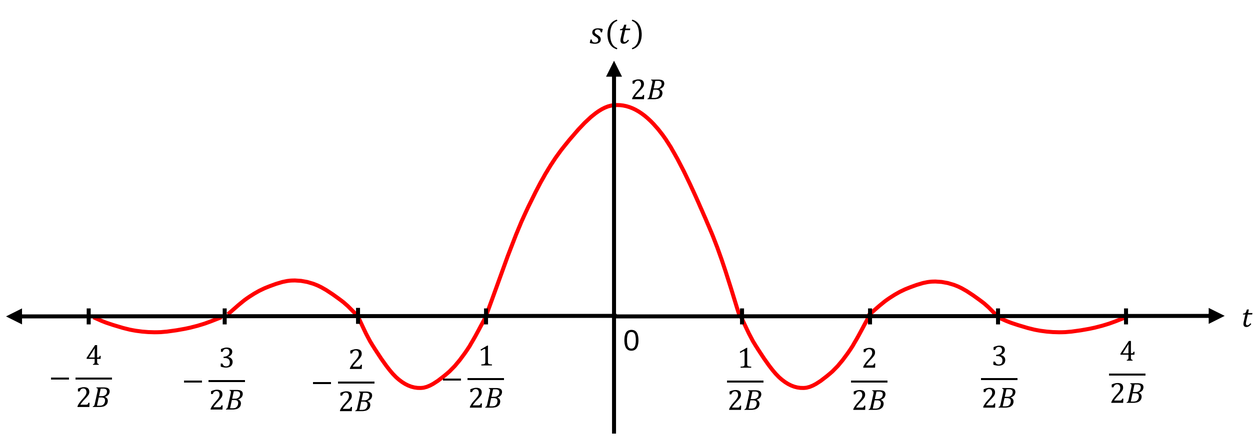
Example --

Signal with bandwidth B

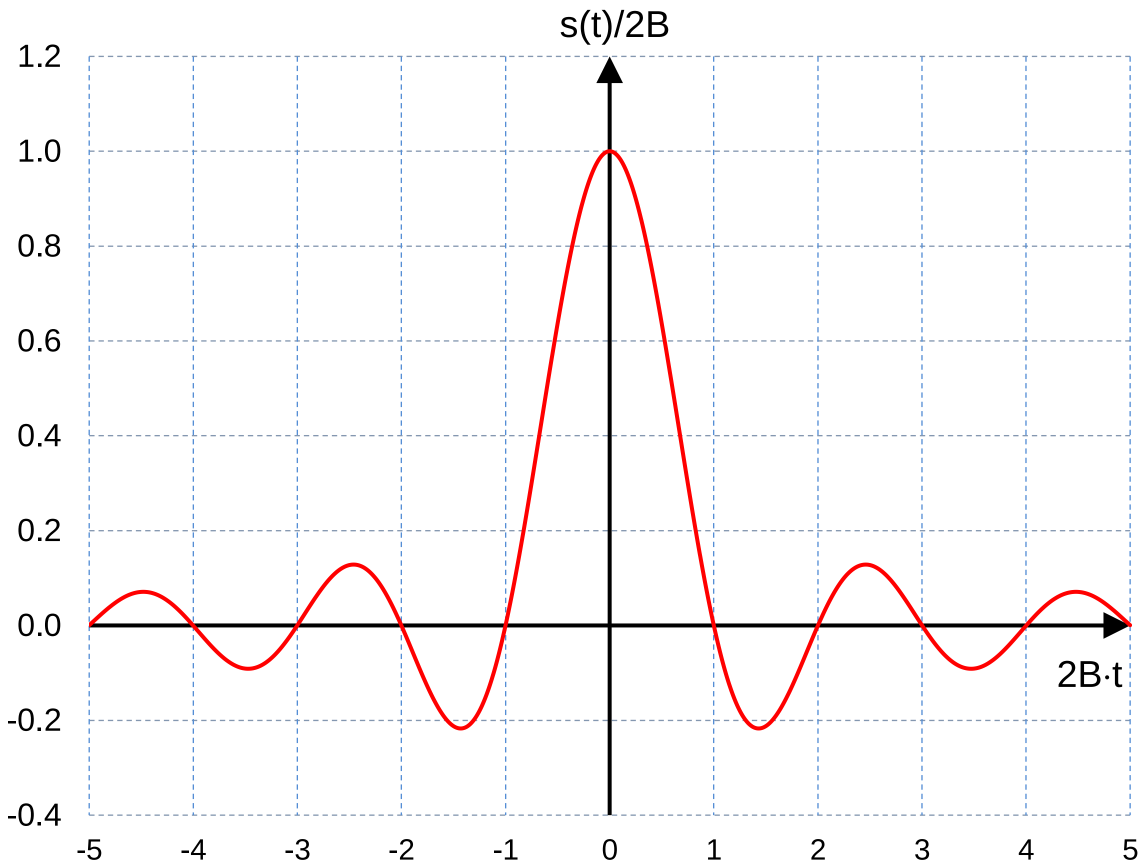


We define the spectrum over the interval $[-B, +B]$ to ensure that the time-domain signal $s(t)$ is going to be a real function. The inverse Fourier transform of $S(f)$ is given by

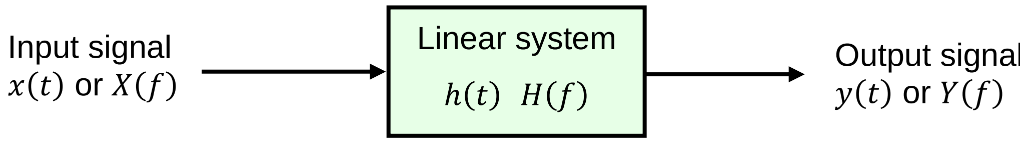
$$s(t) = \int_{-\infty}^{+\infty} S(f) \cdot e^{+j2\pi ft}\, df = \int_{-B}^{+B} e^{+j2\pi ft}\, df = \left[\frac{e^{+j2\pi ft}}{j2\pi t}\right]_{-B}^{+B} = \frac{1}{j2\pi t} \cdot \left(e^{+j2\pi Bt} - e^{-j2\pi Bt}\right) = \frac{\sin(2\pi Bt)}{\pi t} = 2B \cdot \mathrm{sinc}(2\pi Bt).$$



The shape of this sine cardinal pulse is very smooth, which explains why its bandwidth is not infinite, unlike the square pulse studied previously. It can be seen that, when the bandwidth B is decreased, the sine cardinal pulse "spreads out", i.e., becomes even smoother.



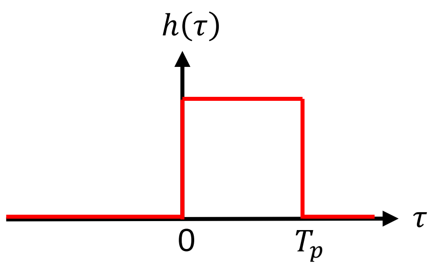
The Fourier transform is often used to evaluate the response of a linear system (also sometimes referred to as linear filter) to an input signal. To illustrate this, let us consider a linear system characterized by its *impulse response* $h(t)$ or, equivalently, its *transfer function* $H(f)$. The transfer function $H(f)$ is the Fourier transform of the impulse response $h(t)$.

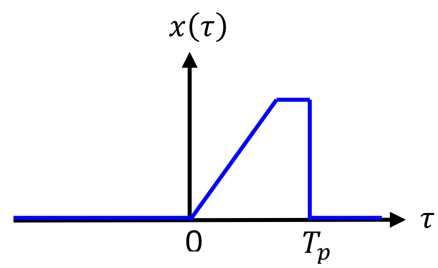


In the time domain, the output signal $y(t)$ is obtained by determining the convolution product of the input signal and the impulse response:

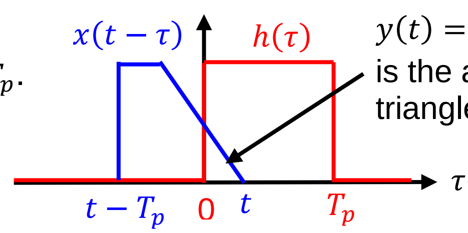
$$y(t) = h(t) * x(t) = \int_{-\infty}^{+\infty} h(\tau) \cdot x(t-\tau)\, d\tau.$$

To ensure causality of a linear system, we must define its impulse response so that $h(t) = 0, \forall t < 0$. A causal system is a system whose output depends on past and current inputs, but not future inputs, i.e., an output $y(t_0)$ only depends on the input $x(t)$ for values of $t \leq t_0$. Systems must be causal to process signals in real time.



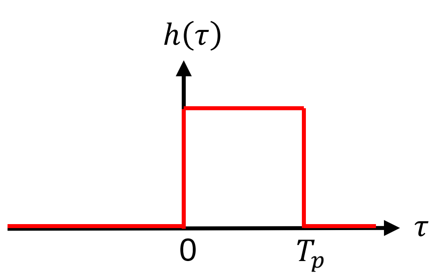


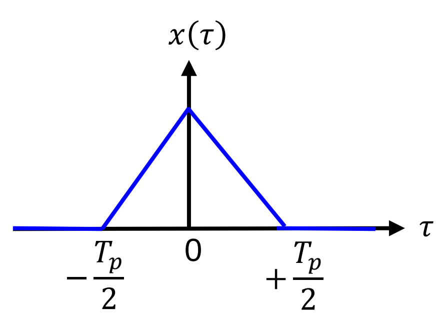
In this example,
$y(t) = 0$ for $t < 0$ or $t > 2T_p$.
$y(t)$ is maximal for $t = T_p$.



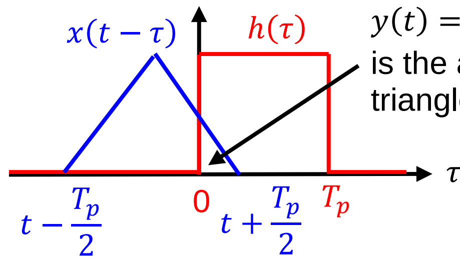
$y(t) = \int_{-\infty}^{+\infty} h(\tau) \cdot x(t-\tau)\, d\tau$ is the area contained in this triangle.

We can be more flexible regarding other types of pulse signals in the sense that they can also be defined as starting at any time $t < 0$.





In this example,
$y(t) = 0$ for $t < -\frac{T_p}{2}$ or $t > \frac{3T_p}{2}$.
$y(t)$ is maximal for $t = \frac{T_p}{2}$.



$y(t) = \int_{-\infty}^{+\infty} h(\tau) \cdot x(t-\tau)\, d\tau$ is the area contained in this triangle.

In the frequency domain, the output signal $Y(f)$ is obtained by determining the product of the input signal $X(f)$ and the transfer function $H(f)$:

$$Y(f) = H(f) \cdot X(f).$$

A product of two functions is generally much easier to evaluate than the convolution product of these functions. This is why it is often preferable to process signals in the frequency domain rather than in the time domain.

Example --

Show that the Fourier transform of a real signal $s(-t)$ is $S^*(f)$ which is the complex conjugate of $S(f)$.

The Fourier transform of $s(-t)$ is given by

$$\int_{-\infty}^{+\infty} s(-t) \cdot e^{-j2\pi ft}\, dt = \int_{+\infty}^{-\infty} s(u) \cdot e^{+j2\pi fu}\, (-du) = \int_{-\infty}^{+\infty} s(u) \cdot e^{+j2\pi fu}\, du$$

$$= \int_{-\infty}^{+\infty} \left[s(u) \cdot e^{-j2\pi fu}\right]^* du = \left[\int_{-\infty}^{+\infty} s(u) \cdot e^{-j2\pi fu}\, du\right]^* = S^*(f).$$

Example --

Show that the Fourier transform of $s(t-T)$, where T is a constant, is $S(f) \cdot e^{-j2\pi fT}$.

The Fourier transform of $s(t-T)$ is given by

$\int_{-\infty}^{+\infty} s(t-T)\cdot e^{-j2\pi ft}\,dt = \int_{-\infty}^{+\infty} s(u)\cdot e^{-j2\pi fu}\ e^{-j2\pi fT}\,du = e^{-j2\pi fT}\cdot\int_{-\infty}^{+\infty} s(u)\cdot e^{-j2\pi fu}\,du = S(f)\cdot e^{-j2\pi fT}$.

Example --

Consider two signals $g(t)$ and $h(t)$ defined so that $g(t) = h(t) * h(-t)$. Show that the energy of $h(t)$ is equal to $g(0)$.

The energy of $h(t)$ is given by

$$E = \int_{-\infty}^{+\infty} [h(\tau)]^2\ d\tau = \int_{-\infty}^{+\infty} h(\tau)\cdot h(\tau-0)\ d\tau = \int_{-\infty}^{+\infty} h(\tau)\cdot h\big(-(0-\tau)\big)\ d\tau = g(0).$$

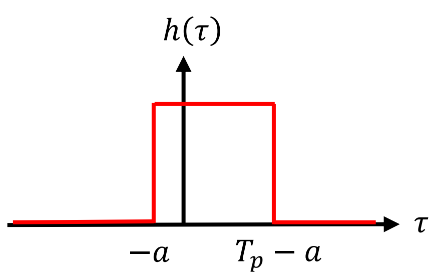
Example --

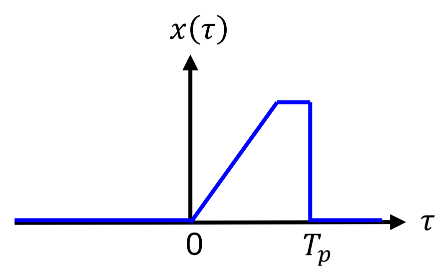
Show that $s(0) = \int_{-\infty}^{+\infty} S(f)\,df$ and $S(0) = \int_{-\infty}^{+\infty} s(t)\,dt$.

We have $\int_{-\infty}^{+\infty} S(f)\,df = \int_{-\infty}^{+\infty} S(f)\,e^{+j2\pi f0}df = s(0)$ and $\int_{-\infty}^{+\infty} s(t)\,dt = \int_{-\infty}^{+\infty} s(t)\ e^{-j2\pi 0t}dt = S(0)$.

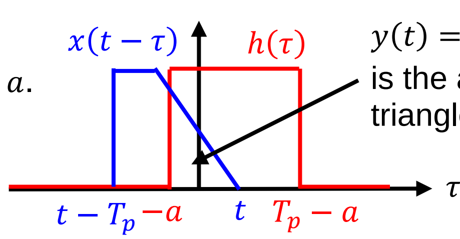
Example --

Show using a simple example that the principle of causality implies that an impulse response must always be defined so that $h(t) = 0, \forall t < 0$.





In this example (with $a > 0$),
$y(t) = 0$ for $t < -a$ or $t > 2T_p - a$.
$y(t)$ is maximal for $t = T_p - a$.



$y(t) = \int_{-\infty}^{+\infty} h(\tau)\cdot x(t-\tau)\ d\tau$
is the area contained in this triangle.

In this example, we see that the output signal starts at time $t = -a < 0$, although the input signal starts at $t = 0$. In other words, the linear system starts responding to the input signal before the latter has even started. This does not make any sense.

• **Digital filters**

In practice, a linear system is not always implemented using an analogue filter with a continuous impulse response $h(t)$. Instead, it is implemented in a digital fashion. The digital filter is fed with successive (quantized) samples $x_n = x(nT),\ n \in \{0, 1, 2, 3, \ldots\}$, of the input signal $x(t)$. Based on these samples and the filter coefficients, the digital filter generates samples $y_n,\ n \in \{0, 1, 2, 3, \ldots\}$, that represent the output signal $y(t)$.

In the case of a finite impulse response (FIR) filter, a sample y_n is computed using

$$y_n = \sum_{i=0}^{L} h_i \cdot x_{n-i} = h_0 \cdot x_n + h_1 \cdot x_{n-1} + h_2 \cdot x_{n-2} + \cdots + h_L \cdot x_{n-L},$$

where the quantities $h_i,\ i \in \{0, 1, \ldots, L\}$, are the filter coefficients (representing the discrete impulse response of the filter) and L is the filter order. This computation is also known as discrete convolution.

Example --

Consider a FIR filter with $L + 1 = 5$ coefficients given by $h_0 = h_1 = h_2 = h_3 = h_4 = 0.2$. Hence, we can write $y_n = 0.2 \times (x_n + x_{n-1} + x_{n-2} + x_{n-3} + x_{n-4})$. What type of filter is it?

To answer, simply consider the response of the filter to a step function, i.e., find the sequence $\{y_n\}$ of output samples when the sequence $\{x_n\}$ of input samples is given by

$$\{x_n\} = \{\ldots, 0, 0, 0, 0, +1, +1, +1, +1, +1, +1, +1, +1, \ldots\}.$$

We obtain $\{y_n\} = \{\dots, 0, 0, 0, 0, +0.2, +0.4, +0.6, +0.8, +1, +1, +1, +1 \dots\}$. The effect of the filter has been to remove the sharp transition, i.e., high frequencies, from the input signal. This filter is therefore a low-pass filter.

Example --

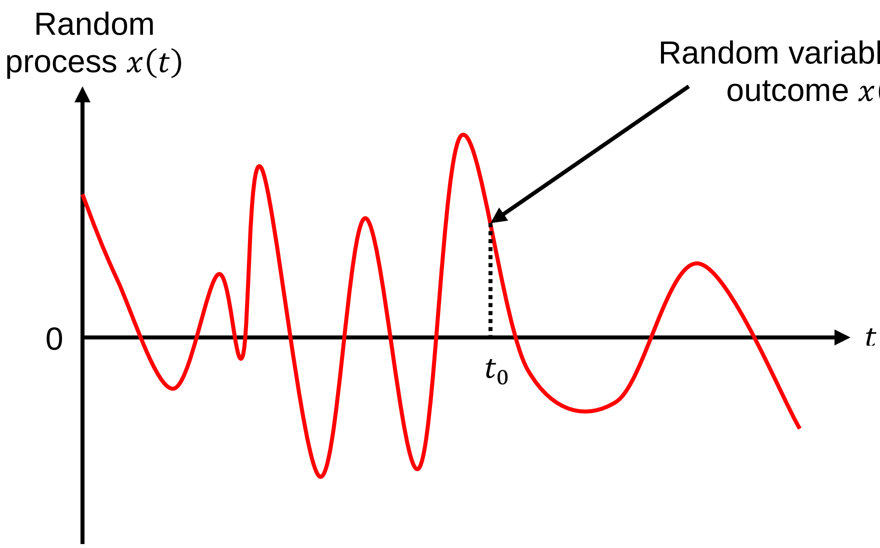
Consider a FIR filter with $L + 1 = 4$ coefficients given by $h_0 = h_1 = +1$ and $h_2 = h_3 = -1$. Hence, we can write $y_n = x_n + x_{n-1} - x_{n-2} - x_{n-3}$. What type of filter is it?

Once again, you can find the answer by determining the response of the filter to a step function, i.e., find the sequence $\{y_n\}$ of output samples when the sequence $\{x_n\}$ of input samples is given by $\{x_n\} = \{\dots, 0, 0, 0, 0, +1, +1, +1, +1, +1, +1, +1, +1, \dots\}$.

We obtain $\{y_n\} = \{\dots, 0, 0, 0, 0, +1, +2, +1, 0, 0, 0, 0, 0 \dots\}$. The effect of the filter has been to remove the constant parts, i.e., the zero frequency, and keep the sharp transition, where the high frequencies reside. This filter is therefore a high-pass filter.

4. Random Processes and Variables

In communication engineering, one must often deal with *random processes*, i.e., time-varying random "signals" (voice, noise…). A random process is not deterministic, meaning that it is impossible to predict what values such "signal" will take in the future. One can just have an idea about its behavior. In general, communication engineers have the knowledge of some of the statistical properties of the random process.



If we sample the random process $x(t)$ at time $t = t_0$, we obtain a *random variable* X which has some statistical properties, such as a *mean*, defined as the average value of the outcomes of X, and a *variance* that measures the variability of the outcomes of X around their mean value. By sampling $x(t)$ at a particular time $t = t_0$, we obtain a sample $x(t_0)$ that can also be viewed as an outcome of the random variable X associated with $t = t_0$.

If the statistical characteristics of $x(t)$ do not change over time, $x(t)$ is said to be stationary. Throughout these notes, we will only consider stationary random processes.

Communication engineers often make no distinction between the concepts of random process, random variable, and outcome of this random variable. So, it can become confusing at times. We will have to be a bit flexible regarding notations.

For instance, one can encounter the three following notations that are generally considered as being equivalent:

1. The notation $E\{x(t)\}$ refers to the expected value, i.e., the mean value, of a random process $x(t)$ at time t. If $x(t)$ is stationary, the value of $E\{x(t)\}$ does not depend on t.
2. The notation $E\{X\}$ refers to the expected value of a random variable X which is generated by sampling a random process $x(t)$ at a given time t. The fact that the notation X does not contain any reference to time implies that the statistical properties of X do not depend on t, i.e., $x(t)$ is stationary.
3. The notation $E\{x_0\}$ refers to the expected value of a sample, here called x_0, of a random process $x(t)$ taken at a particular time t_0. The quantity x_0 can also be viewed as an outcome of the random variable X generated by sampling $x(t)$ at time t_0.

If X can only take a finite number of values, X is said to be *discrete*.

If X can take an infinite number of values, X is said to be *continuous*.

• Discrete random variables

Consider a random variable X that can take m values a_i with probabilities p_i. We must have

$$\sum_{i=1}^{m} p_i = 1.$$

If we plot the probability p_i as a function of values a_i, we obtain a discrete distribution (histogram, for example).

The mean m of X, also called expected value, $E\{X\}$, of X, is given by

$$m = E\{X\} = \sum_{i=1}^{m} a_i \cdot p_i.$$

The mean m of X is simply the average value of the outcomes of X.

The variance σ^2 of X is given by

$$\sigma^2 = E\{(X-m)^2\} = [\sum_{i=1}^{m}(a_i - m)^2 \cdot p_i].$$

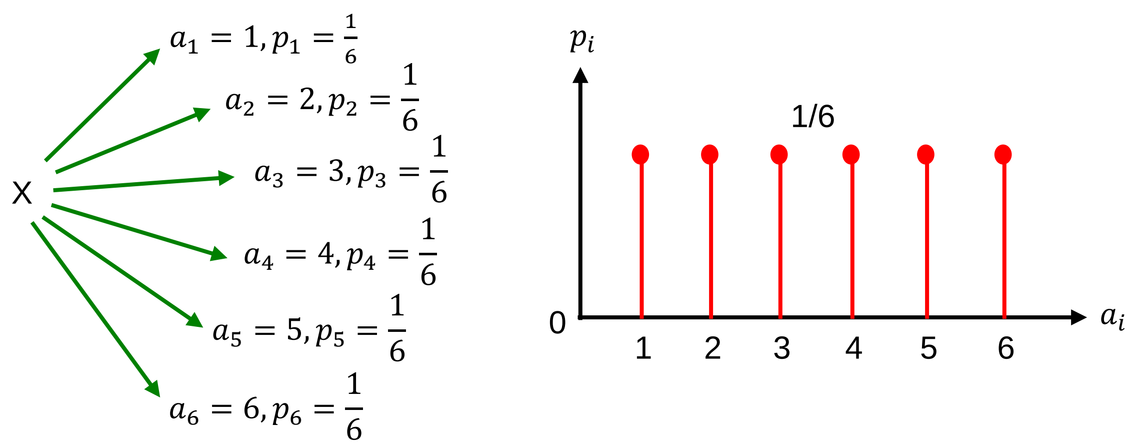
The variance measures the variability of the outcomes of X around their mean value.

Another useful expression of the variance can be derived as follows:

$\sigma^2 = E\{(X-m)^2\} = E\{X^2 - 2Xm + m^2\} = E\{X^2\} - E\{2Xm\} + E\{m^2\} = E\{X^2\} - 2mE\{X\} + m^2 = E\{X^2\} - 2m^2 + m^2 = E\{X^2\} - m^2 = [\sum_{i=1}^{m}(a_i)^2 \cdot p_i] - m^2$.

Example: cast of a dice --

Consider a random variable X corresponding to the cast of a dice. It can be modelled using a uniform discrete distribution.



Show that its mean m is equal to 3.5, and its variance σ^2 is approximately equal to 2.92.

- **Continuous random variables**

Since a continuous random variable X can take an infinite number of values, we need to use a mathematical function to specify the distribution of X.

This function is called the probability density function (PDF), $P_X(x)$, of X.

The mean m of X is then given by

$$m = E\{X\} = \int_{-\infty}^{+\infty} x \cdot P_X(x)\, dx,$$

while its variance σ^2 is expressed as

$$\sigma^2 = E\{X^2\} - m^2 = \int_{-\infty}^{+\infty} x^2 \cdot P_X(x) dx - m^2.$$

Using the function $P_X(x)$, it is possible to evaluate the probability for an outcome x_0 of X to be in a certain range by using

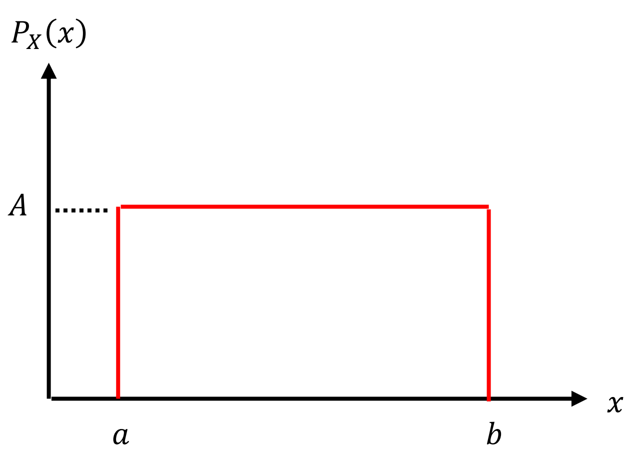
$$\Pr\{\, A < x_0 < B \,\} = \int_A^B P_X(x)\; dx.$$

This is a very important equation that is often used in practice to compute the probabilities of certain events. As we must have $\Pr\{-\infty < x_0 < +\infty\} = 1$, the expression above implies that a PDF must always be normalized so that

$$\int_{-\infty}^{+\infty} P_X(x) dx = 1.$$

Example: uniform PDF --

Consider a random variable with a uniform distribution over an interval $[a, b]$.



Show that

(1) The amplitude A of $P_X(x)$ must be equal to $\frac{1}{b-a}$;

(2) The mean of X is given by $m = \frac{a+b}{2}$;

(3) The variance of X is given by $\sigma^2 = \frac{(b-a)^2}{12}$;

(4) We can write $\Pr\{a < x_0 < c\} = \frac{c-a}{b-a}$ for $a < c < b$.

• **Addition and multiplication of random variables**

Consider two random variables X_1 and X_2.

We can always write $\mathrm{E}\{X_1 + X_2\} = \mathrm{E}\{X_1\} + \mathrm{E}\{X_2\}$.

We can also always write $\mathrm{E}\{a \cdot X_1 + b \cdot X_2\} = a \cdot \mathrm{E}\{X_1\} + b \cdot \mathrm{E}\{X_2\}$, where a and b are two constant quantities.

But, in general, we have $\mathrm{E}\{X_1 \cdot X_2\} \neq \mathrm{E}\{X_1\} \cdot \mathrm{E}\{X_2\}$.

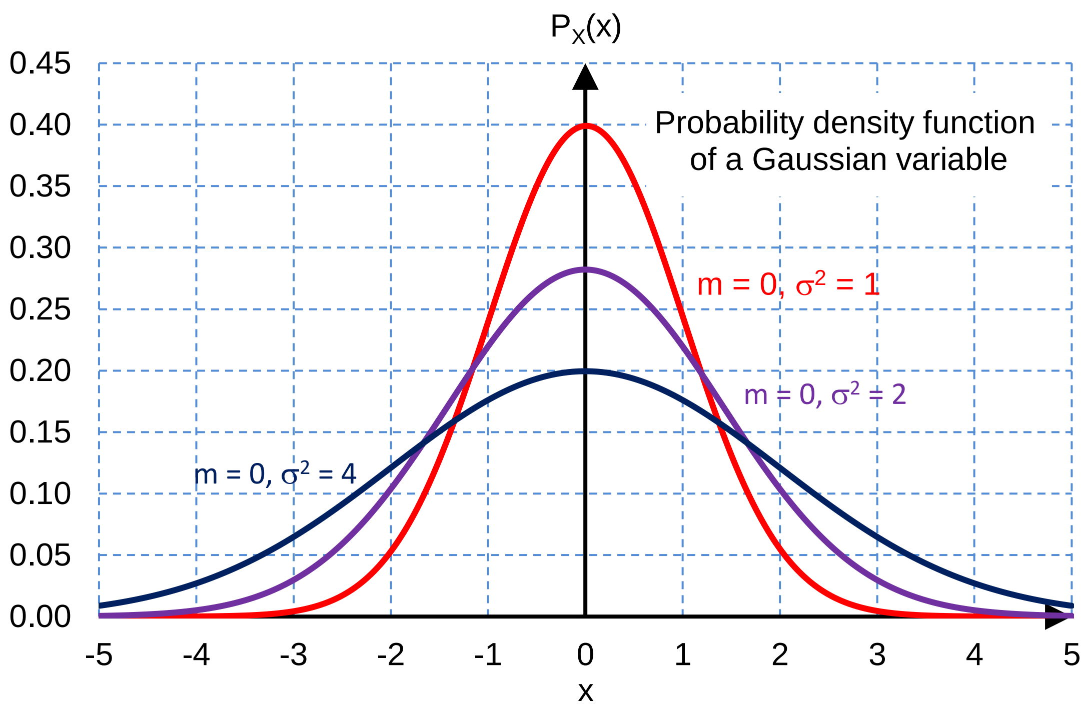
However, in the special case where X_1 and X_2 are independent, then we can write $\mathrm{E}\{X_1 \cdot X_2\} = \mathrm{E}\{X_1\} \cdot \mathrm{E}\{X_2\}$.

Two random variables X_1 and X_2 are said to be independent if the knowledge of the outcome of X_1 does not tell us anything about the outcome of X_2, and vice versa.

• **Gaussian random variable**

A random variable X is said to be Gaussian if it has a Gaussian PDF given by

$$P_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}}.$$



Gaussian variables are very common in all fields of science (engineering, physics, biology, medicine, and so on). If a random variable is generated by the sum of the individual contributions of many independent random variables, which happens very often in science, then the distribution of this random variable is going to be Gaussian. This is true even in cases where the distribution of the individual variables is not Gaussian (see the central limit theorem).

The Gaussian distribution is also known as the normal distribution. We often use the following abbreviated notation to designate a Gaussian random variable X with mean m and variance σ^2: $X \sim \mathcal{N}(m, \sigma^2)$.

Example ---

Consider a Gaussian random variable X with a mean m and a variance σ^2**.** What are the characteristics of the random variable aX generated by multiplying X by a real constant a?

Multiplying a random variable by a constant does not change the shape of its probability density function. Hence, the random variable aX is Gaussian with a mean

$$\mathrm{E}\{aX\} = a \cdot \mathrm{E}\{X\} = a \cdot m,$$

and a variance

$$\mathrm{E}\{a^2X^2\} - (a \cdot m)^2 = a^2 \cdot \mathrm{E}\{X^2\} - (a \cdot m)^2 = a^2\sigma^2 + a^2m^2 - (a \cdot m)^2 = a^2\sigma^2.$$

As a conclusion, $X \sim \mathcal{N}(m, \sigma^2) \rightarrow aX \sim \mathcal{N}(am, a^2\sigma^2)$.

Example --

Consider two independent Gaussian random variables $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(m_2, \sigma_2^2)$**.** What are the characteristics of the random variable $X = X_1 + X_2$?

Summing two independent Gaussian random variables generates another Gaussian variable with a mean given by

$$\mathrm{E}\{X_1 + X_2\} = \mathrm{E}\{X_1\} + \mathrm{E}\{X_2\} = m_1 + m_2,$$

and a variance given by

$$\mathrm{E}\{(X_1 + X_2)^2\} - (m_1 + m_2)^2 = \mathrm{E}\{X_1^2\} + 2\mathrm{E}\{X_1\} \cdot \mathrm{E}\{X_2\} + \mathrm{E}\{X_2^2\} - m_1^2 - 2m_1 \cdot m_2 - m_2^2$$
$$= \sigma_1^2 + m_1^2 + 2m_1 \cdot m_2 + \sigma_2^2 + m_2^2 - m_1^2 - 2m_1 \cdot m_2 - m_2^2 = \sigma_1^2 + \sigma_2^2.$$

As a conclusion, $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(m_2, \sigma_2^2) \rightarrow X_1 + X_2 \sim \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$ if X_1 and X_2 are independent.

It is important to mention that the sum of two Gaussian variables that are not independent is NOT necessarily Gaussian.

Example --

Consider two independent Gaussian random variables $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(m_2, \sigma_2^2)$**.** What are the characteristics of the random variable $X = X_1 - X_2$?

Subtracting two independent Gaussian random generates another Gaussian variable with a mean

$$\mathrm{E}\{X_1 - X_2\} = \mathrm{E}\{X_1\} - \mathrm{E}\{X_2\} = m_1 - m_2,$$

and a variance

$$\mathrm{E}\{(X_1 - X_2)^2\} - (m_1 - m_2)^2 = \mathrm{E}\{X_1^2\} - 2\mathrm{E}\{X_1\} \cdot \mathrm{E}\{X_2\} + \mathrm{E}\{X_2^2\} - m_1^2 + 2m_1 \cdot m_2 - m_2^2$$
$$= \sigma_1^2 + m_1^2 - 2m_1 \cdot m_2 + \sigma_2^2 + m_2^2 - m_1^2 + 2m_1 \cdot m_2 - m_2^2 = \sigma_1^2 + \sigma_2^2.$$

As a conclusion, $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(m_2, \sigma_2^2) \rightarrow X_1 - X_2 \sim \mathcal{N}(m_1 - m_2, \sigma_1^2 + \sigma_2^2)$ if X_1 and X_2 are independent.

Example: Detection of a binary symbol corrupted by Gaussian noise ------------------------

Assume that we were sent a binary symbol $s \in \{-1, +1\}$. However, we only have access to an estimate r of it. This estimate can be written as $r = s + n$, where n is a Gaussian noise sample with zero mean and variance σ^2.

We assume that the symbol s can take the values -1 and $+1$ with equal probabilities: $\Pr\{s = -1\} = \Pr\{s = +1\} = \frac{1}{2}$.

Our goal is to recover the value of the symbol s that was sent to us by solely relying on what we have access to, i.e., the observation r. Of course, we also know that it can be equal to either -1 and $+1$ with equal probabilities.

This optimal technique to recover the value of a symbol s based on the observation r is known as the *maximum likelihood (ML) detection* method. It is optimal in the sense that it minimizes the probability of failure to detect the correct value of symbol s.

The universal ML detection rule is as follows: we decide that symbol s was equal to -1 rather than $+1$ if $\Pr\{s = -1; r\} > \Pr\{s = +1; r\}$. If this not the case, we decide that it was equal to $+1$.

So, how to apply ML detection in this example to recover the value of symbol s?

We start our calculation by using Bayes' rule:

$$\Pr\{s=-1;r\} > \Pr\{s=+1;r\} \leftrightarrow \Pr\{r|s=-1\}\cdot\Pr\{s=-1\} > \Pr\{r|s=+1\}\cdot\Pr\{s=+1\}.$$

As $\Pr\{s=-1\} = \Pr\{s=+1\} = \frac{1}{2}$, the above inequality is equivalent to $\Pr\{r|s=-1\} > \Pr\{r|s=+1\}$. Now, we can replace the probabilities with mathematical expressions by exploiting the PDF of the observations $r = s + n$ given the knowledge of s:

$$\Pr\{r|s=-1\} = \frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(r-(-1))^2}{2\sigma^2}},$$

and

$$\Pr\{r|s=+1\} = \frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(r-1)^2}{2\sigma^2}}.$$

The inequality becomes

$$\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(r+1)^2}{2\sigma^2}} > \frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(r-1)^2}{2\sigma^2}} \rightarrow (r+1)^2 < (r-1)^2 \rightarrow r < 0.$$

We thus decide that symbol s was equal to -1 if the observation r is negative. If $r > 0$, we decide that $s = +1$ instead. ML detection consists of selecting the possible symbol value -1 or $+1$ that is at minimum Euclidean distance from the observation r.

If we employ ML detection to recover the value of s, what is then the probability, P_{wd}, of wrong detection?

To answer this question, we can start with

$$P_{wd} = \Pr\{(+1\text{ TX}; +1\text{ not DX}) \cup (-1\text{ TX}; -1\text{ not DX})\},$$

where $(s\text{ TX}; s\text{ not DX})$, $s \in \{-1, +1\}$, denotes the event where s is the transmitted symbol and s is not the detected symbol. Such event clearly represents an error event.

As both events $(s\text{ TX}; s\text{ not DX})$ are mutually exclusive, we can write

$$P_{wd} = \Pr\{+1\text{ TX}; +1\text{ not DX}\} + \Pr\{-1\text{ TX}; -1\text{ not DX}\}.$$

By using Bayes' rule, we obtain

$$\Pr\{s\text{ TX}; s\text{ not DX}\} = \Pr\{s\text{ not DX}|s\text{ TX}\}\cdot\Pr\{s\text{ TX}\} = \frac{1}{2}\cdot\Pr\{s\text{ not DX}|s\text{ TX}\}.$$

So, we can write

$P_{wd} = \frac{1}{2} \cdot (\Pr\{+1 \text{ not DX}|+1 \text{ TX}\} + \Pr\{-1 \text{ not DX}|-1 \text{ TX}\}) = \frac{1}{2} \cdot (\Pr\{r < 0|+1 \text{ TX}\} + \Pr\{r > 0|-1 \text{ TX}\}) = \frac{1}{2} \cdot (\Pr\{n + 1 < 0\} + \Pr\{n - 1 > 0\}) = \frac{1}{2} \cdot (\Pr\{n < -1\} + \Pr\{n > +1\})$.

The Gaussian noise has a mean equal to zero, which implies that its probability density function is an even function. Therefore, we have $\Pr\{n < -1\} = \Pr\{n > +1\}$. This leads to

$$P_{wd} = \Pr\{n > +1\} = \int_{+1}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{x^2}{2\sigma^2}} dx.$$

This integral does not have a closed-form expression. The way forward is to introduce the complementary error function, $\text{erfc}(x)$, into the expression of P_{wd}. This function is defined as

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \cdot \int_{x}^{+\infty} e^{-u^2} du.$$

The complementary error function does not have an exact closed-form expression, but it can be evaluated using tables or software packages such as Matlab and Excel.

By performing the change of variable $u = \frac{x}{\sqrt{2\sigma^2}}$, the expression of the probability of wrong detection becomes

$$P_{wd} = \frac{\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} \cdot \int_{\sqrt{\frac{1}{2\sigma^2}}}^{+\infty} e^{-u^2} du = \frac{1}{\sqrt{\pi}} \cdot \int_{\sqrt{\frac{1}{2\sigma^2}}}^{+\infty} e^{-u^2} du = \frac{1}{2} \cdot \text{erfc}\left(\sqrt{\frac{1}{2\sigma^2}}\right).$$

As $\text{erfc}(x)$ is a monotonically decreasing function, the lower the noise variance σ^2, the lower the probability P_{wd} of wrong detection of a symbol.

This result does make sense because a lower noise variance corresponds to a noise process that has lesser power as it tends to vary less around its mean value.

5. Power Spectral Density and Correlation Functions

A random process is a non-periodic infinite-energy *signal*. Hence, both its Fourier series expansion and Fourier transform cannot be evaluated.

• Power spectral density and autocorrelation function

The spectral characteristic of a random process is obtained by computing its *power spectral density* (PSD), $\Phi(f)$, which represents the distribution of power with frequency.

The PSD of a random process can be used to evaluate its power P in a frequency band ranging from f_1 to f_2 by computing the area under $\Phi(f)$: $P = \int_{f_1}^{f_2} \Phi(f) df$.

The total power of the random process is thus given by

$$P_{total} = \int_{-\infty}^{+\infty} \Phi(f) df.$$

The function $\Phi(f)$ of a random process $x(t)$ is the Fourier transform of its autocorrelation function $\Gamma(t)$ which is defined as follows:

$$\Gamma(t) = E_\tau\{x(\tau) \cdot x(\tau - t)\}.$$

We can therefore write $\Phi(f) = \int_{-\infty}^{+\infty} \Gamma(t) \cdot e^{-j2\pi f t} dt$ as well as $\Gamma(t) = \int_{-\infty}^{+\infty} \Phi(f) \cdot e^{+j2\pi f t} df$.

The autocorrelation function informs us about the degree of correlation between successive samples of a random process.

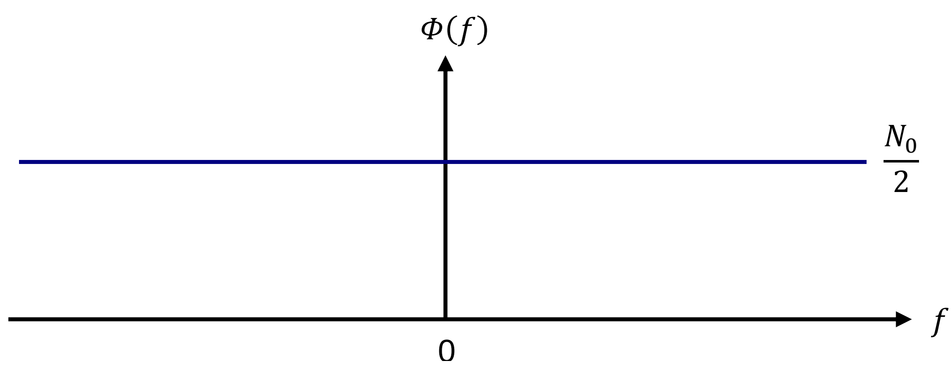
The maximum value of $\Gamma(t)$ is obtained for $t = 0$ as the correlation is maximal between a sample taken at any time and itself. Also, we have $\Gamma(t) \to m^2$ as $t \to +\infty$ and $t \to -\infty$. In fact, two samples $x(\tau)$ and $x(\tau - t)$ become independent as $t \to +\infty$ and $t \to -\infty$. This independence leads to $E_\tau\{x(\tau) \cdot x(\tau - t)\} \to E_\tau\{x(\tau)\} \cdot E_\tau\{x(\tau - t)\} = m^2$.

In a slow-varying random process, there is plenty of memory between successive samples. Consequently, the autocorrelation function of such process tends to reach the value m^2 slowly as $t \to +\infty$ and $t \to -\infty$, which means that its Fourier transform, i.e., the power spectral density $\Phi(f)$, tends to occupy a narrow frequency band. This is consistent with the general principle of time-frequency representations stating that signals that vary slowly in the time domain occupy less bandwidth in the frequency domain than fast-varying signals.

Note that $P_{total} = \int_{-\infty}^{+\infty} \Phi(f)\ df = \Gamma(0) = E_\tau\{[x(\tau)]^2\} = \sigma^2 + m^2$. This result means that the total power P_{total} of a random process is equal to the maximum value of the autocorrelation function $\Gamma(t)$, obtained for $t = 0$. It is also the sum of its variance σ^2 and the square of its mean m.

• **White noise process**

A *white noise* process $n(t)$ is defined to have a flat PSD over the entire frequency range. This type of noise is of particular importance in electronic and communication engineering because it corresponds to the noise generated by the thermal agitation of the charge carriers (mostly electrons) inside electronic components. This thermal noise can have a very detrimental effect on the operation of sensitive systems, such as radio receivers.



A pure white noise model is, however, unrealistic since it implies that such random process has a total power P_{total} that is infinite because we have $P_{total} = \int_{-\infty}^{+\infty} \Phi(f)df = \int_{-\infty}^{+\infty} \frac{N_0}{2} df = +\infty$.

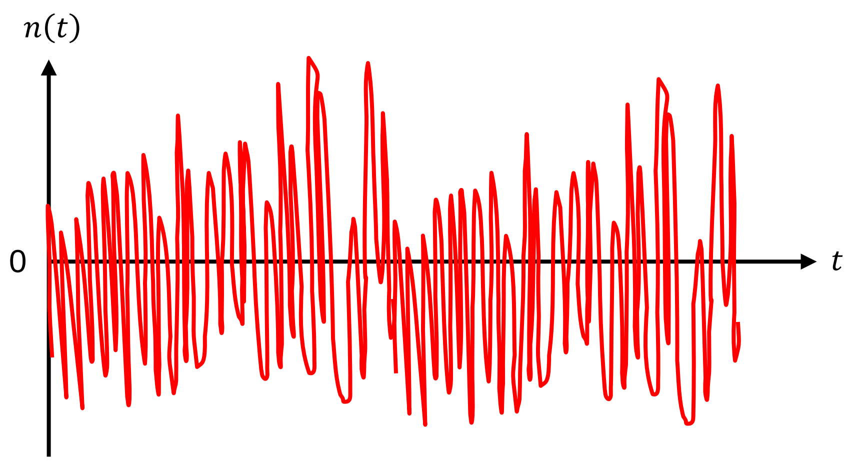
The fact that a white noise process has an infinite total power P_{total} implies that it also has an infinite variance. Recall that we can write $P_{total} = \sigma^2 + m^2$.

Since we have $\Phi(f) = \frac{N_0}{2}$, $\forall f$, the autocorrelation function of $n(t)$ is given by

$$\Gamma(t) = \frac{N_0}{2} \cdot \int_{-\infty}^{+\infty} e^{+j2\pi ft} df = \frac{N_0}{2} \cdot \delta(t),$$

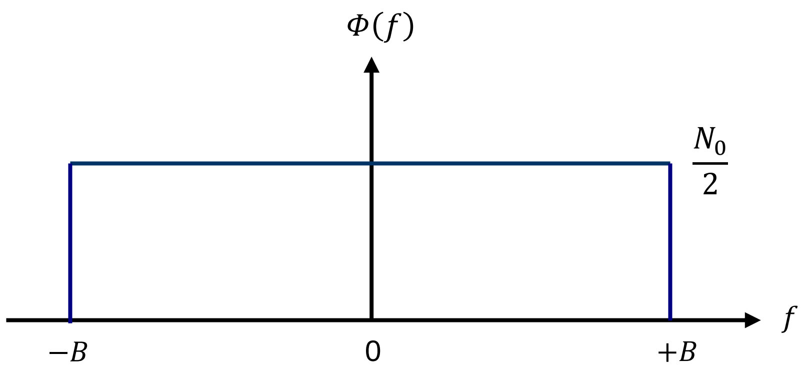
where $\delta(t)$ is called the Dirac delta function and defined as being equal to 0 for any $t \neq 0$ and having unit area, i.e., $\int_{-\infty}^{+\infty} \delta(t)dt = 1$. This result implies that two successive white noise samples are not correlated, no matter how close they are. This is, once again, something that is unrealistic.

A white noise process would typically display many *infinitely sharp* variations of unlimited amplitude because its power is infinite and there is no correlation between successive samples $n(\tau)$ and $n(\tau - t)$, $\forall t \neq 0$. A more realistic, i.e., almost-white, noise process typically looks like the random process plotted below.



A pure white noise process cannot exist in the real world, but a noise process can still be considered as white in practice if its PSD is constant in the bandwidth of interest. In other words, if no signal in our communication system contains frequency components outside a band

$[-B,+B]$, then any noise process that has a flat PSD between $f=-B$ and $f=+B$.can be considered as white.

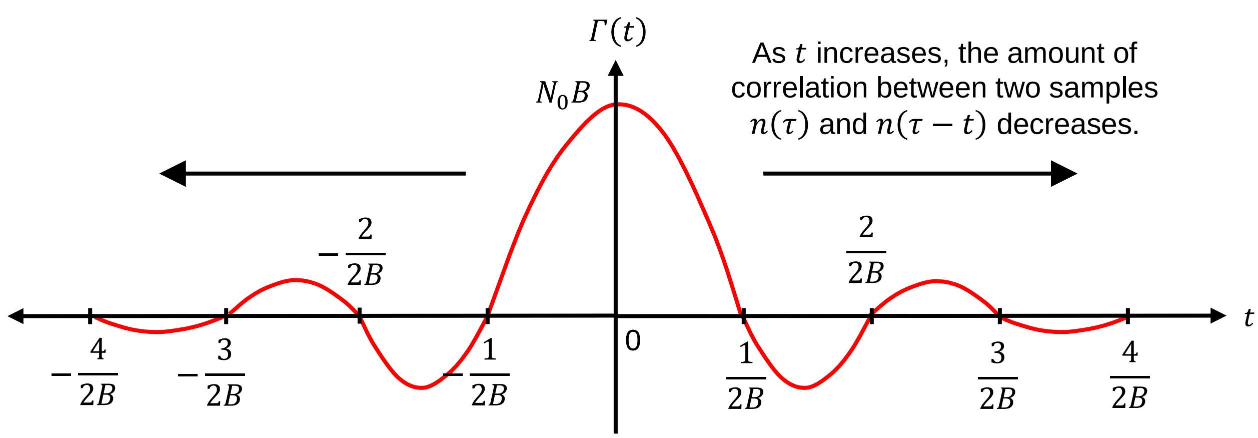


In this case, the total power of the noise process is given by $P_{total} = \int_{-B}^{+B} \Phi(f)df = N_0 B$.

The fact that a noise process cannot be perfectly white in the real world means that we must expect to see some correlation between two noise samples $n(\tau)$ and $n(\tau - t)$ for $t \neq 0$, especially for small values of t. However, if we make sure to never take noise samples that are too close to each other, then these samples are not going to be correlated, and the noise process is going to appear as if it was perfectly white.

As an illustration, let us consider once again a noise process whose PSD is flat inside a $[-B,+B]$ frequency range and equal to zero elsewhere. Its autocorrelation function is thus given by

$$\Gamma(t) = N_0 B \cdot \mathrm{sinc}(2\pi B t).$$

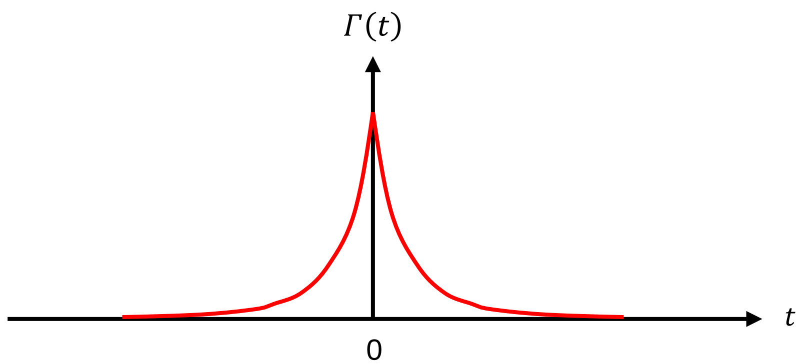


Remarkably, this expression indicates that, if the time t between two samples $n(\tau)$ and $n(\tau - t)$ is a multiple of $\frac{1}{2B}$, then these samples are not correlated. Therefore, if we were to sample such noise process every $\frac{1}{2B}$ seconds, then this noise would appear as if it was perfectly white.

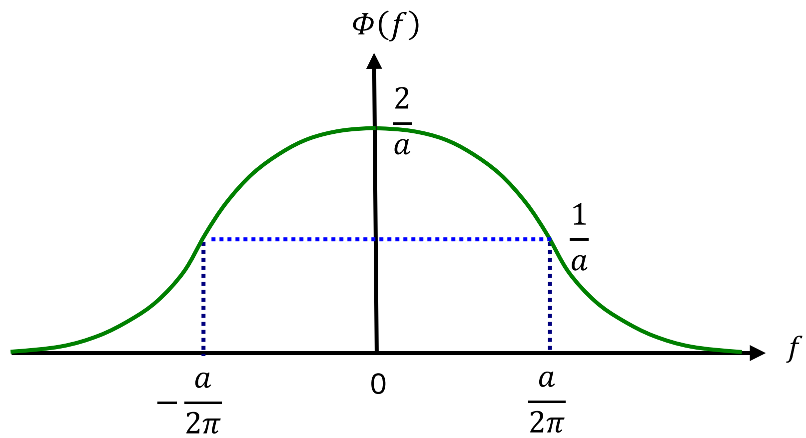
It is worth mentioning that the white noise is also often *Gaussian*. However, it is important to note that white and Gaussian are two adjectives that have nothing to do with each other. As an example, a noise process can be Gaussian without being white, and vice-versa. For instance, if a white Gaussian noise process is filtered, the output of the filter will be another noise process which is still Gaussian, but no longer white.

Example ---

Assume $\Gamma(t) = e^{-a\cdot|t|}$. Determine the PSD of this random process.



$$\Phi(f) = \int_{-\infty}^{+\infty} \Gamma(t)\cdot e^{-j2\pi ft}dt = \int_{-\infty}^{+\infty} e^{-a|t|}\cdot e^{-j2\pi ft}dt = \int_{-\infty}^{0} e^{+at}\cdot e^{-j2\pi ft}dt + \int_{0}^{+\infty} e^{-at}\cdot e^{-j2\pi ft}dt = \int_{-\infty}^{0} e^{+(a-j2\pi f)t}dt + \int_{0}^{+\infty} e^{-(a+j2\pi f)t}dt = \left[\frac{e^{+(a-j2\pi f)t}}{a-j2\pi f}\right]_{-\infty}^{0} + \left[-\frac{e^{-(a+j2\pi f)t}}{a+j2\pi f}\right]_{0}^{+\infty} = \frac{e^{+(a-j2\pi f)0}-0}{a-j2\pi f} - \frac{0-e^{-(a+j2\pi f)0}}{a+j2\pi f} = \frac{1}{a-j2\pi f} + \frac{1}{a+j2\pi f} = \frac{a+j2\pi f+a-j2\pi f}{(a-j2\pi f)\cdot(a+j2\pi f)} = \frac{2\cdot a}{a^2+(2\pi f)^2}.$$



• Cross-correlation

The cross-correlation function, $\Gamma_{x,y}(t)$, informs us about the degree of correlation between two random processes $x(t)$ and $y(t)$. It is defined as follows:

$$\Gamma_{x,y}(t) = E_\tau\{x(\tau) \cdot y(\tau - t)\}.$$

If we have $\Gamma_{x,y}(t) = m_x m_y$, $\forall t$, then the two processes $x(t)$ and $y(t)$ are said to be uncorrelated. The fact that they are uncorrelated does not mean that they are independent.

Note that it is also possible to define a cross-correlation for two random variables X and Y. The cross-correlation function $\Gamma_{X,Y}$, which informs us about the degree of correlation between X and Y, is defined as $\Gamma_{X,Y} = E\{X \cdot Y\}$.

If we have $\Gamma_{X,Y} = m_x m_y$, then X and Y are said to be uncorrelated. The fact that they are uncorrelated does not mean that they are independent.

• **Independence implies non-correlation. Non-correlation does NOT imply independence.**

If two random processes or variables are found to be uncorrelated, it means that there is no linear relationship between them. It does not necessarily mean that they are independent.

However, when two random processes or variables are independent, then they are also uncorrelated. It can be easily demonstrated as follows:
- For random processes, the independence of $x(\tau)$ and $y(\tau - t)$ leads to $E_\tau\{x(\tau) \cdot y(\tau - t)\} = E_\tau\{x(\tau)\} \cdot E_\tau\{y(\tau - t)\} = m_x m_y$.
- For random variables, the independence of X and Y leads to $E\{X \cdot Y\} = E\{X\} \cdot E\{Y\} = m_x m_y$.

In the context of Gaussian processes, two random processes $x(t)$ and $y(t)$ can be proven to be independent if they satisfy the following conditions:
1. They are uncorrelated, i.e., we have $\Gamma_{x,y}(t) = m_x m_y$, $\forall t$.
2. They are jointly Gaussian. In other words, any linear combination of them, $\alpha \cdot x(t) + \beta \cdot y(t)$, where α and β designate two arbitrary real numbers, is Gaussian.

In the context of Gaussian variables, two random variables X and Y can be proven to be independent if they satisfy the following conditions:
1. They are uncorrelated, i.e., we have $\Gamma_{X,Y} = m_x m_y$;
2. They are jointly Gaussian, i.e., any linear combination of them, $\alpha \cdot X + \beta \cdot Y$, is Gaussian.

Example: Cross-correlation and independence of random variables ----------------------------
Consider a random variable $Y = X \cdot Z$ obtained by multiplying a zero-mean Gaussian variable X by a binary variable Z that takes the value -1 with probability $1/2$ and the value $+1$ with probability $1/2$. The random variables X and Z are assumed to be independent.

We have $m_x = E\{X\} = 0$ and $m_z = E\{Z\} = 0$. Hence, we can write $m_y = E\{Y\} = E\{X \cdot Z\} = E\{X\} \cdot E\{Z\} = 0$.

Are X and Y independent?

To prove the independence of X and Y, we must show that they satisfy the following conditions:
1. They are uncorrelated, i.e., we have $\Gamma_{X,Y} = 0$;
2. They are jointly Gaussian, i.e., any random variable in the form $\alpha \cdot X + \beta \cdot Y$ is Gaussian.

The first condition is satisfied because we have $\Gamma_{X,Y} = E\{X \cdot Y\} = E\{X^2 \cdot Z\} = E\{X^2\} \cdot E\{Z\} = 0$.

However, the random variables X and Y are not jointly Gaussian because, for instance, the sum $X + Y$ is not a Gaussian variable. This can be understood by noticing that $X + Y = X + X \cdot Z = X \cdot (1 + Z)$.

Clearly, the random variable $X + Y = X \cdot (1 + Z)$ has a Gaussian distribution when the outcome of Z is $+1$, but is always equal to zero when the outcome of Z is -1. Hence, the distribution of $X + Y$ is not Gaussian. This result is remarkable because we could show that Y is actually Gaussian, which means that X and Y are both Gaussian, but their sum is not.

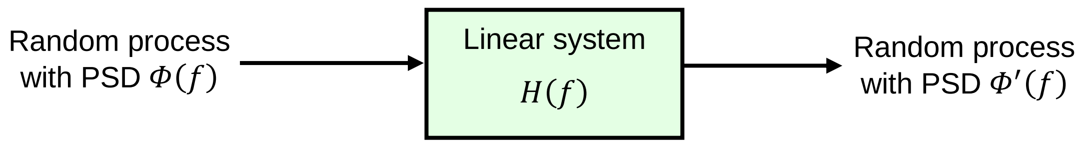
So, we conclude that X and Y are not independent despite being uncorrelated.

We could have reached the same conclusion using a more intuitive reasoning: X and Y are not independent because the knowledge of the outcome of X provides us with some knowledge of the outcome of Y. For instance, if the outcome of X is a number x_0, we then know that the outcome of Y can only be equal to x_0 or $-x_0$.

We therefore conclude from this example that uncorrelation and independence are not the same thing. Remarkably, we have seen that the sum of two random variables is not necessarily Gaussian. This happens when these random variables are not independent, which is indeed the case throughout this example.

• Linear filtering of a random process

We can finally have a quick look at how a PSD is modified by a linear system characterized by its transfer function $H(f)$.



It can be shown that the output PSD is simply obtained by applying $\Phi'(f) = \Phi(f) \cdot |H(f)|^2$.