InnoSpaceTool Unit 7 complementary material - Signals and Channels

Endurosat Team

This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 768049

In this unit we will generalize our developments from unit 2 to non-periodic signals and review one of the most important results in digital signal processing - the Nyquist-Shannon theorem. To that end, we will take a closer look at the differences between continuous (analog) and discrete (digital) signals, which we outlined in unit 1.

1 A general (non-periodic) signal - Fourier Transform

In unit 2 we showed that a general periodic signal can be represented as an infinite superposition of sine and cosine signals of different frequencies through Fourier series. Our task now is to find a similar process for representing non-periodic signals as well. We will once again only outline the important results and give a justification without diving into formal mathematics too deeply.
To begin with, consider the Fourier series again, but this time using complex exponential functions instead of sine and cosine functions, and in the region $[-\frac{T}{2}, \frac{T}{2}]$ instead of the previously considered $[0, T]$. We will show that these series are equivalent to those in unit 2. Using Euler's formula ($e^{i\alpha} = \cos\alpha + i\sin\alpha$), it is easy to rewrite the sine and cosine functions in terms of exponentials:

$$\sin(\omega t) = \frac{e^{i\omega t} - e^{-i\omega t}}{2i} \tag{1.1}$$

$$\cos(\omega t) = \frac{e^{i\omega t} + e^{-i\omega t}}{2} \tag{1.2}$$

Then demanding that $c_{-m} = \frac{1}{2}(a_m + ib_m)$, $c_m = \frac{1}{2}(a_m - ib_m = c^*_{-m})$ and $c_0 = \frac{a_0}{2}$, we can easily see that:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{i\omega_n t} = \frac{a_0}{2} + \sum_{n=1}^{\infty}[a_n\cos(\omega_n t) + b_n\sin(\omega_n t)] \tag{1.3}$$

To see this, note first that due to our identification for the angular frequencies ($\omega_n = 2\pi n/T$), we have $\omega_{-m} = -\omega_m$ and so:

$$\sum_{n=-\infty}^{\infty} c_n e^{i\omega_n t} = ...c_{-m}e^{-i\omega_m t} + ... + c_0 + ... + c_m e^{i\omega_m t} + ... =$$

$$= \frac{a_0}{2} + \sum_{n=1}^{\infty} [c_n e^{i\omega_n t} + c_{-n} e^{-i\omega_n t}] = \frac{a_0}{2} + \sum_{n=1}^{\infty} [\frac{1}{2}(a_n - ib_n) e^{i\omega_n t} + \frac{1}{2}(a_n + ib_n) e^{-i\omega_n t}] =$$

$$= \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \frac{1}{2}(e^{i\omega_n t} + e^{-i\omega_n t}) + b_n \frac{1}{2i}(e^{i\omega_n t} - e^{-i\omega_n t})] = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(\omega_n t) + b_n \sin(\omega_n t)]$$

It follows that the Fourier Series for a function $f(t)$ can be given as a complex exponential series of the form 1.3. To switch the interval of the series so that it is symmetric about the origin, we simply need to translate the functions by $T/2$ to the left, the only difference is that the integrals for calculating a_n, b_n and so c_n will be given by:

$$a_0 = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) dt \tag{1.4}$$

$$a_n = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \cos\left(\frac{2\pi n t}{T}\right) dt \tag{1.5}$$

$$b_n = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \sin\left(\frac{2\pi n t}{T}\right) dt \tag{1.6}$$

Comparing with the formulas from unit 2, we see that only the periods of integration have changed. Furthermore, since $c_n = 1/2(a_n - ib_n)$, it follows that we can calculate the complex constants directly:

$$c_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-i\omega_n t} dt \tag{1.7}$$

Now let's take a look at a general non-periodic function $f(t)$. We can again form a periodic function from any "portion" of it (using different intervals), but can we somehow represent all of it, without making it periodic? The answer for a specific class of functions is yes, and the process comes to the following idea - a non-periodic function which falls off quickly enough is just a periodic function with an infinite period $T \to \infty$! We will not clarify what "quickly enough" means mathematically, but any function which is zero everywhere apart from a compact subset of $\mathbb{R}$ is in this class (these are known as compact support functions). Consider such a function $f(t)$, which is equivalently 0 outside some region around the origin $[-L, L]$. We can readily obtain a periodic function by taking the usual integrals 1.4 to 1.6 for $T = 2L$, but instead we can form a new function in the following way:

$$g(t) = \begin{cases} f(t) & \text{for } -L < t < L \\ 0 & \text{for } t < -L \cup t > L \end{cases}$$

Now if we take a larger period for $g(t)$, say $10L$, since it is simply zero outside of $[-L, L]$, the result will be a periodic function with the non-zero repetitions of $f(t)$ "spaced out" (there will be $8L$ distance between the closest non-zero values of two different periods). Obviously, taking $T \to \infty$ will space them out infinitely and so we will obtain an effectively non-periodic function, which is simply $f(t)$ in the region $[-L, L]$ and zero outside - but this was exactly the

original function. Carrying this process out formally is relatively difficult, but the following will give us an idea of how it works:
First notice that the frequencies are evenly spaced: $\Delta\omega_n = \omega_{n+1} - \omega_n = \dfrac{2\pi}{T} = \omega$. Notice further that making the period larger makes $\Delta\omega_n$ smaller, and in the case we're interested (making T infinite), we have:

$$\lim_{T\to\infty} \Delta\omega_n = 0$$

Using the facts above:

$$1 = \frac{T}{2\pi}\Delta\omega_n$$

And so we can rewrite 1.3 as:

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{T}{2\pi} c_n e^{i\omega_n t}\Delta\omega_n$$

Since the constants c_n contain division by T, they will all tend to zero as we take the limit, and so we will use new constants $F_n = Tc_n$. These are given simply by the integrals 1.7 (no division by T). Using them, we obtain:

$$f(t) = \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} F_n e^{i\omega_n t}\Delta\omega_n$$

We can now take the limit. In it, the separation between the frequencies becomes infinitely small $\Delta\omega_n \to d\omega$, the sum becomes an integral and the sequences $\{\omega_n\}$ and $\{F_n\}$ become continuous variable and continuous function respectively. This is the part where we will not be very formal, but it is a mathematical result that:

$$\lim_{T\to\infty} \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} F_n e^{i\omega_n t}\Delta\omega_n = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega \tag{1.8}$$

Then we can finally describe the function $f(t)$ as an infinite (and continuous) superposition of waves, without making it periodic! The function $F(\omega)$ can be found by taking the same limit $T \to \infty$ in 1.7 after multiplying it by T (since $F_n = Tc_n$). The two functions $F(\omega)$ and $f(t)$ are connected and their connection is called a Fourier (or inverse Fourier) transform. Given one, we can find the other in the following way:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt \tag{1.9}$$

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega \tag{1.10}$$

Given a function $f(t)$, the function $F(\omega)$ is called its (forward) Fourier transform. Given a function $F(\omega)$, the function $f(t)$ is called its inverse Fourier transform. Note that f is a function of time, while F is a function of angular frequency! We can interpret F as a "weight", telling us how much each frequency ω contributes to the signal $f(t)$. This means that if we have a given non-periodic signal in time, there is a single distribution of frequencies corresponding to it and vice versa! We only hinted of this duality between time and frequency in unit 2, but we will now fully explore it with a very popular example in the next section!

Remark: The definitions 1.9 and 1.10 may be given differently in different sources. For example some norm both the forward and inverse transforms by $1/\sqrt{2\pi}$ as opposed to norming just one with its square. Also, some sources may be using the opposite convention for the sign of the exponentials (and so for them positive sign would be forward transform and negative would be inverse transform). These conventions are all equivalent and give equivalent results, as long as one does not switch between them while in the process of Fourier analysis! We will be using the derived formulas 1.9 and 1.10.

2 The frequency and time domains - an example

Let us consider a simple rectangular impulse as a signal:

$$U(x) = \begin{cases} c & \text{for } -1 < x < 1 \\ 0 & \text{for all other values of x} \end{cases}$$

By finding the Fourier series for this signal for different periods T, we can compose periodic functions with different spacings of the impulses. As long as $T > 1$ (which is required if we don't want the impulses to overlap), the angular frequencies will be given as before and the constants will be given by 1.7, which yields:

$$c_n = \frac{1}{T}\int_{-1}^{1} e^{-i\omega_n t}dt = \frac{1}{i\omega_n T}(e^{i\omega_n} - e^{-i\omega_n}) = \frac{2}{\omega_n T}\sin\omega_n$$

Note that the integration is carried out only over $[-1, 1]$, since our function is zero outside of this region. Furthermore, we have taken $c = 1$ to simplify things. Noting the relationship between T and ω_n, we can write this as:

$$c_n = \frac{2}{\omega_n T}\sin\omega_n = \frac{1}{n\pi}\sin\left(\frac{2\pi n}{T}\right) \tag{2.1}$$

Now let us find the Fourier transform of this signal (the case when we represent it as non-periodic as derived in the previous section). Denoting $U(t)$'s transform by $\mu(\omega)$ and using 1.9, we easily obtain:

$$\mu(\omega) = \int_{-\infty}^{+\infty} U(t)e^{-i\omega t}dt = \int_{-1}^{+1} e^{-i\omega t}dt = \frac{2}{\omega}\sin(\omega) \tag{2.2}$$

Comparing this result with 2.1, we can see an obvious similarity! The two functional forms are identical (apart from a division by T in the first), except for the fact that 2.2 is a continuous distribution of frequencies (all frequencies are included with a weight given by $\mu(\omega)$) and 2.1 consists of only a discrete number of frequencies (each included with the same weight as it would be if it was continuous). This can be interpreted in the following way -

representing a periodic function with sine waves requires an enumerable number of frequencies (even though they may be infinitely many, they are "spaced out"), while representing a non-periodic function generally requires a continuum of frequencies (not just infinitely many, but also not enumerable)!
To see this, let's represent $U(t)$ with three different periods - $T_1 = 3$, $T_2 = 6$, $T_3 = \infty$ (two periodic and one non-periodic representation).
The function $\mu(\omega)$ in the region $[-15, 15]$ of the frequency domain is plotted on Fig.7.1. As it can be seen, larger frequencies contribute less for the formation of the signal (we will define a quantity which tells us how much energy each "mode" of the signal carries soon). The discrete constants c'_n are plotted alongside $\mu(\omega)$ on Fig.7.2 for the two periods listed above. It immediately becomes apparent that the larger the period is, the more "dense" the required frequencies become, culminating in a truly continuous function (as $T \to \infty$), which is the Fourier transform.

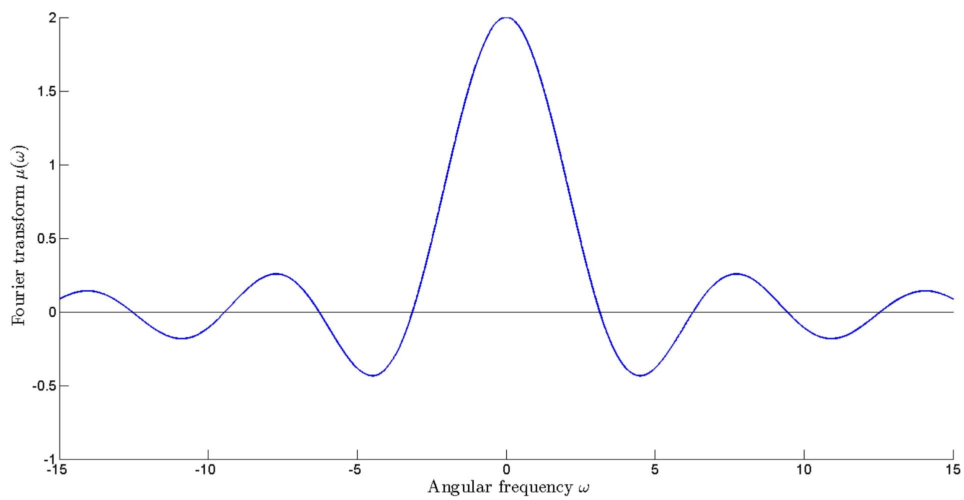
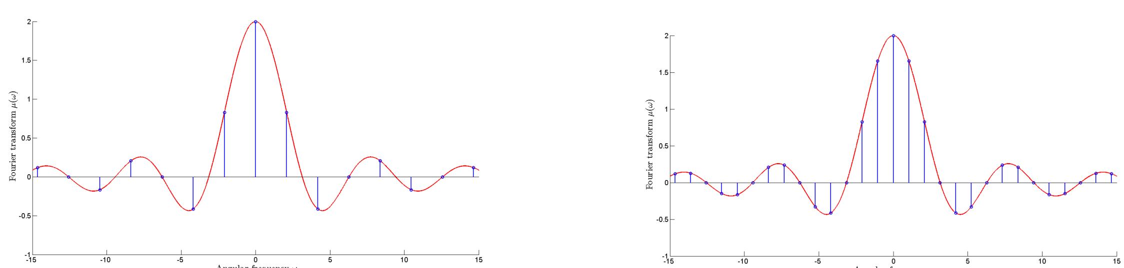


Fig.7.1 - The Fourier transform of a rectangular impulse in the region $[-1, 1]$ with height 1.



(a) Period T=3 (b) Period T=6

Fig.7.2 - The Fourier transform (red) alongside the Fourier series constants c'_n (blue) in the cases T_1 and T_2 in the region $[-15, 15]$ of the frequency domain.

The implications are as follows - if we are to recover the spectre of a signal and then try to reproduce it, since our machines cannot truly describe continuous functions (only discrete

points), the reproduced signal will never be a perfect non-periodic copy of the original! Its period, however, can become sufficiently large (and so the discretization step of the function $\mu(\omega)$ will be sufficiently small) so that this will not have any practical consequences! As an example, (Fig7.3) on the next page shows the same sequence of frequencies 2.1, but this time for a period $T = 50$. One can continue increasing the period further and realize as close description as one needs.
There is, however, another restriction when reconstructing signals from their spectre due to our equipment, and it is far more serious. First note that finding the inverse Fourier transform (using 1.10) of $\mu(\omega)$ given by 2.2, will give us exactly our original rectangular impulse. In practice, however, no synthesizer can do this, since there are infinitely large frequencies included in this square impulse! For example $|\mu(\omega = 3.10^{15})| \simeq 10^{15}$ tells us that there would be some ionizing radiation included (however small it may be) if the impulse was composed of EMWs.

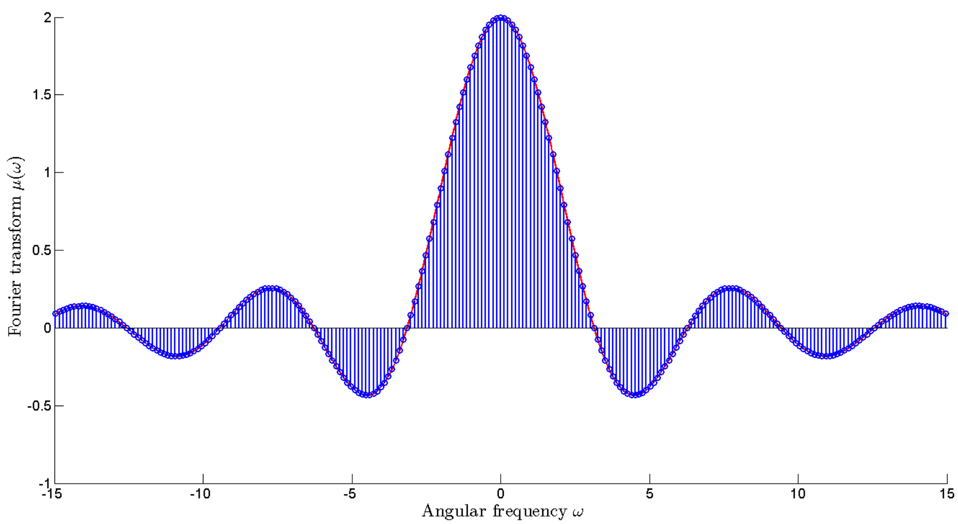


Fig.7.3 - The Fourier transform (red) alongside the Fourier series constants c'_n (blue) in the case $T = 50$ in the region $[-15, 15]$ of the frequency domain.

Now is the time to define a very important characteristic, which will tell us how important each frequency is for forming the final signal.
Def: *For signal $U(t)$ confined in time (compact carrier signal), the total energy carried by it is given by the expression:*

$$E = \int_{-\infty}^{+\infty} |U(t)|^2 dt \tag{2.3}$$

Note that the term "energy" is a general term, not necessarily the physical energy for a physical process. For example, if $U(t)$ was the magnitude of the electric field, then this quantity would indeed be proportional to the total energy carried by the impulse due to our electric energy density definition. With this term clear, the following important result (known as Parseval's theorem) tells us how each "mode" of frequency contributes to the total energy of an impulse:

Theorem: *For a signal $U(t)$, which is confined in time (compact carrier signal) and its Fourier transform $\mu(\omega)$:*

$$\int_{-\infty}^{+\infty} |U(t)|^2 dt = \int_{-\infty}^{+\infty} |\mu(\omega)|^2 d\omega \tag{2.4}$$

This tells us that we can interpret $|\mu(\omega)|^2$ as spectral energy density - a function which tells us how much energy each angular frequency ω carries! We will denote this by $S(\omega) = |\mu(\omega)|^2$. Note that some sources define the spectral energy density as a function of frequency ν instead of angular frequency ω, but since $\omega = 2\pi\nu$, the two are equivalent up to some constant factor and the result 2.4 will hold up to a constant factor again.

To see how the concept of spectral energy density helps us, consider our example again, for which:

$$S(\omega) = |\mu(\omega)|^2 = \frac{4}{\omega^2} \sin^2 \omega$$

It can be shown that more than 95% of the energy for this signal is carried by the frequencies in the region $[-6.5, 6.5]$ of the frequency domain. In practice, one cannot use all frequencies which compose a general non-periodic signal, and so one "cuts" the higher frequencies off by setting the Fourier transform $\mu(\omega)$ identically equal to zero above (bellow) some $\pm\omega_{max}$.

To see what effects this brings into the reconstructed signal, let us consider the signal resulting from cutting all frequencies of our $U(t)$ above $\omega = 6$ Hz. This is equivalent to setting $\mu(\omega > 6 \cup \omega < -6) = 0$. The new signal obtained is then given by:

$$U_{cut}(t) = \frac{1}{2\pi} \int_{-6}^{6} \frac{2}{\omega} \sin \omega e^{i\omega t} d\omega$$

This integral cannot be evaluated in terms of elementary functions and so the shape of $U_{cut}(t)$ must be found numerically. The resulting impulse is shown bellow.

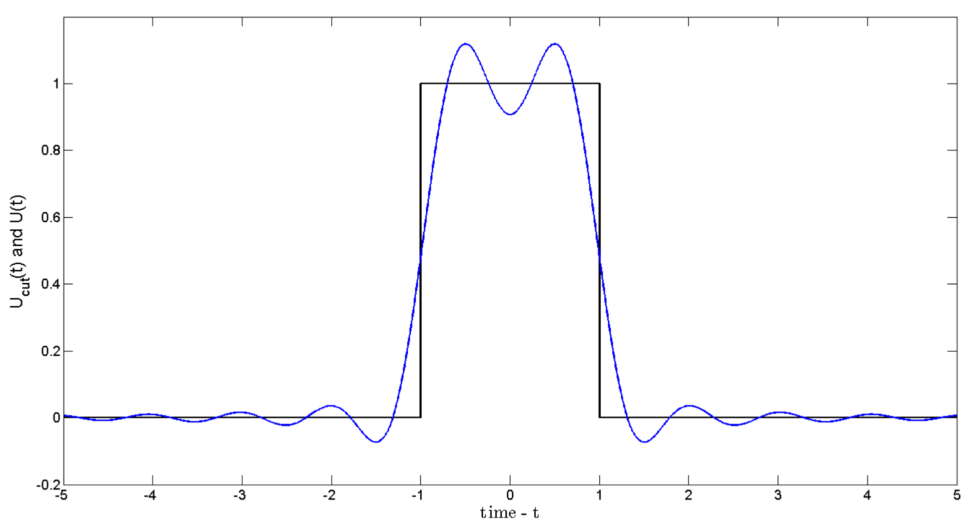


Fig.7.4 - The signal resulting from cutting off frequencies higher than $\omega = 6$ (blue) and the original signal (black).

As expected, we can see deviations from the true rectangular impulse we started with, which tells us that in practice, impulses are not reconstructed perfectly even in the absence of any outside noise, due to the limitations of our tools and their inability to operate with all possible frequencies. Just as it was with the Fourier series, adding more frequencies (so expanding the limits of integration in the frequency domain) will make the reconstruction better and above some frequency interval the signal will be practically indistinguishable from the original!

3 Digitalization, sampling and the Nyquist - Shannon theorem

Now is the time to formally distinguish between discrete (digital) and continuous (analog) signals. To see the difference, consider the microphone example again. We stated that a microphone will convert sound wave impulses into continuous impulses of current. These can then be fed into a speaker (which is the opposite of a microphone) and be reproduced by a similar process. On the other hand, if we wanted to keep a sound signal in the memory of a computer, it would have to be "digitalized". The reason is that computers do not use continuous signals. Information in them is stored in bits and each bit can either be 1 or 0 ("on" or "off"). So how do we represent a continuous signal in terms of elements which have only two states (seemingly infinitely less than the range of continuous signals). To see this, let us consider the continuous impulse $F(t)$ shown bellow as the signal to be digitalized.

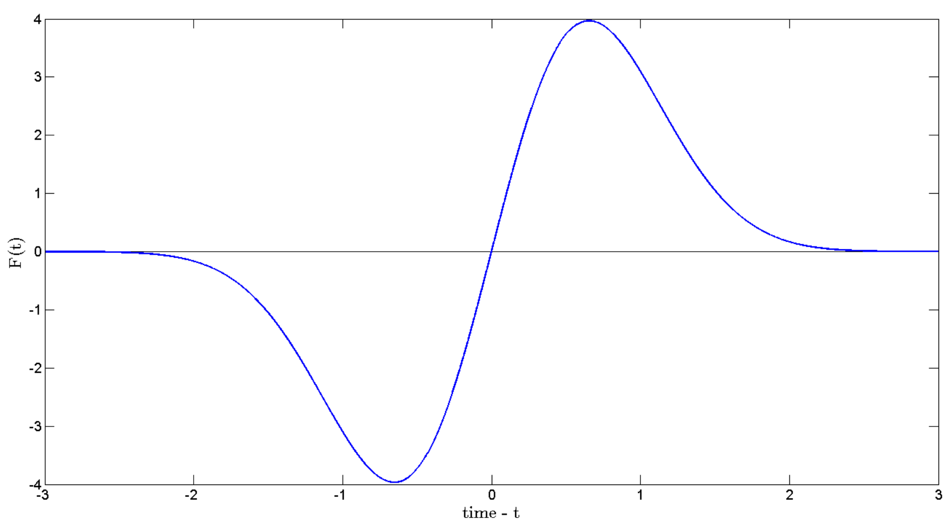


Fig.7.5 - The signal to be digitalized.

Now each individual bit has just two states (1 or 0), but combining them, we can form more states. Since each is independent, two bits will give us 4 states (01, 10, 00 and 11), three bits will give us twice as many (since we will have 4 states with the third being 0 and 4 more with the third being 1) and so on. In general, N bits will give us 2^N possible states.
Considering only 4 bits for our example, we have 16 possible states. Since our signal is in

the range $[-4, 4]$, we can break up the y axis into 16 different discrete regions, so that each corresponds to a bit. Leaving the first bit for a $+$ or $-$ indicator, we are left with 8 states above zero and 8 states bellow 0, each of which corresponds to an interval of length 0.5. The division into discrete intervals for the y's of our example is given in the table bellow:

Interval range (+)	Designation	Interval range (-)	Designation
$[0, 0.5)$	1000	$[-0.5, 0)$	0000
$[0.5, 1)$	1001	$[-1, -0.5)$	0001
$[1, 1.5)$	1010	$[-1.5, -1)$	0010
$[1.5, 2)$	1011	$[-2, -1.5)$	0011
$[2, 2.5)$	1100	$[-2.5, -2)$	0100
$[2.5, 3)$	1101	$[-3, -2.5)$	0101
$[3, 3.5)$	1110	$[-3.5, -3)$	0110
$[3.5, 4)$	1111	$[-4, -3.5)$	0111

Table 7.1 Discretization of the interval $[-4, 4]$ into bits.

Now if we break up t into intervals of the same length (this is not required, but it will make our discrete grid square), we can approximate our continuous signal $F(t)$ in the range of $t \in [-2.5, 2.5]$ as a 10 component vector, which can be given fully as a sequence of bits if we associate one of the 16 combinations to each "time step" ($t = -2.5$, $t = -2$, ... $t = 2.5$). This is shown in the graph bellow:

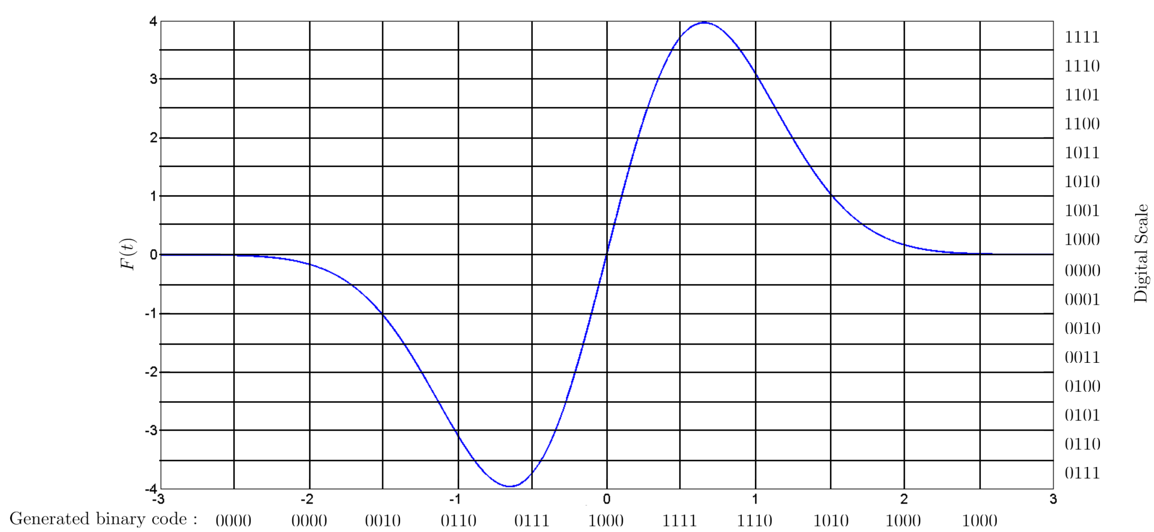


Fig.7.6 - Discretization of the interval and range with steps 0.5 for the signal $F(t)$.

So the signal $F(t)$ from Fig.7.5 can be represented digitally (using 0.5 as both time and amplitude discretization step) by:

00000000001001100111100011111110101010001000

It is obvious that perfect reconstruction in our example is not possible, since to each time step, there is a whole range of possible values. What is worse, in some regions (for example around $t = 0$) our function changes a lot for just one time step. These problems cannot be avoided and in general converting a digital signal back to an analog is not ideal - the deviations in each timestep introduce "digitalization noise". It is obvious, however, that making the discrete levels more dense (by using more bits) and making the time step smaller will improve the digitalization process and reduce the noises introduced by it. The problem is that the more the levels are and the smaller the time step is, the longer the binary sequence which represents the signal is, and so the more memory it will require. It would be nice if we knew the largest time step we can use to reconstruct a signal "well enough" so that we don't waste resources for better reconstruction, all the while obtaining good results. Such a time step turns out to exist for a specific class of signals, as we will shortly see and when it does, it depends on the signal itself (as can be expected). In general, there are infinitely many possible analog signals corresponding to a single digital signal, but Harry Nyquist and Claude Shannon showed (as we will now) that for signals with limited spectre, unique reconstruction is possible under certain conditions. To give their result, let us first get a few definitions out of the way.

Def: *Given a signal $F(t)$, the spectre of which is zero outside some interval $[-L, L]$, the total length of the interval ($2L$) will be called the signal bandwidth.*

Def: *The process of discretizing a continuous signal and associating a given interval to each time step of will be called sampling. The time step will be called sampling time (t_S) and its inverse will be called sampling rate f_S.*

$$f_S = \frac{1}{t_S}$$

We are now ready to give one of the most important results in digital signal processing:

Theorem: *If a continuous signal F(t) contains no frequencies higher than B Hertz, it is completely determined by digitalizing it at a sampling rate higher than $2B$ (sampling time less than $1/2B$ seconds.)*

Note that the theorem in its statement above uses the usual frequency ν instead of the angular frequency ω. Since the two are connected by $\omega = 2\pi\nu$, a signal containing no frequency higher than B Hertz is equivalent to one containing no angular frequency higher than $2\pi B$ Hertz. All of our results in the Fourier series and transforms can be easily given in terms of the usual frequency in the same way. We will take a brief look at the proof Shannon gave, since it gives us some more insight on the connection between Fourier series and transform of a function, and how they affect the required sampling rate:

Proof: Let $\Phi(\omega)$ be the spectre of $F(t)$. Since it contains no frequencies higher than B (no angular frequencies higher than $2\pi B$, we can give $F(t)$ as:

$$F(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi(\omega) e^{i\omega t} d\omega = \frac{1}{2\pi} \int_{-2\pi B}^{+2\pi B} \Phi(\omega) e^{i\omega t} d\omega$$

Using discretization step $1/2B$, we have $t = n/2B$ for n any positive or negative integer n. Then we obtain for the values of F at the sampling points:

$$F\left(\frac{n}{2B}\right) = \frac{1}{2\pi} \int_{-2\pi B}^{+2\pi B} \Phi(\omega) e^{i\omega\left(\frac{n}{2B}\right)} d\omega$$

Comparing the integral on the right with 1.7, we see that it is proportional to the n^{th} Fourier series coefficient in the expansion of the function $\Phi(\omega)$ for a period $2B$. This means that the Fourier series of $\Phi(\omega)$ are completely determined by the sampling points values $F(\frac{n}{2B})$. Since $\Phi(\omega)$ is zero outside of its fundamental period $[-B, B]$, it follows that it is completely determined. But since a function is completely determined by its spectrum, this means that $F(t)$ is also completely determined! Therefore the sample rate $2B$ is sufficient to reconstruct the original function completely as promised!

The sample rate $2B$ for a given signal with compact spectrum (zero for frequencies higher than B) is called the Nyquist rate, and the Nyquist criterion $f_S > 2B$ provides us with a sufficient condition for reconstructing the signal after digitalization! Undersampling a signal (sampling it with rate lower than the Nyquist rate) results in deviations after its reconstruction, called aliasing. Since each sample gives a Fourier coefficient (as we saw in the proof), aliasing is similar to the effect which we observed with the reconstruction of our rectangular impulse (Fig.7.4).

For completeness, we will comment further that the discretization levels of the signal can be picked dynamically (not equally spaced), so that they are more "dense" near zero and less dense further away from it. This can improve the process further and help reduce the sampling requirements.

Now that we know how digital signals are obtained from analog such, we can consider the transmission speed. For n bits and sampling rate f_S, since each sample consists of n bits, the effective transmission speed (in bits per second) is:

$$R_b = f_S n \tag{3.1}$$

It can be shown that in order to reconstruct the digital signal properly in a receiver, it is sufficient to transmit the first harmonic to its highest frequency spectral component. It can be seen as a sequence of rectangular impulses with repetition frequency $f_T/2$ and k harmonics ($k = 1, 2, 3...$) with frequencies $f_k = (2k-1)f_T/2$ and so the effective bandwidth of a digital signal can be given by the so called Nyquist bandwidth (B_N) by:

$$B_N = \frac{f_T}{2} = \frac{1}{T_b} \tag{3.2}$$

where T_b is the duration of a single bit.

Example: Considering telephone signals as an example again (which we said have analog bandwidth of about 3 kHz), the sampling rate for them would have to be higher than 6 kHz. A sampling rate of 8 kHz is used in practice and 8 bits are used for the discretization (7 for the signal level and 1 for the sign). This means that the transmission speed is 64 kbit/sec and the effective bandwidth is 32 kHz (so 10 times larger than the bandwidth for analog transmission).