Module

1

Introduction to Digital Communications and Information Theory

Lesson 2

Signals and Sampling Theory

After reading this lesson, you will learn about

- ***Need for careful representation of signals***
- ***Types of signals***
- ***Nyquist's sampling theorem and its practical implications***
- ***Band pass representation of narrow band signals***

'Signal' in the present context means electrical manifestation of a physical process. Usually an essence of 'information' will be associated with a signal. Mathematical representation or abstraction should also be possible for a signal such that a signal and its features can be classified and analyzed.

Examples of a few signals

(a) Electrical equivalent of speech/voice as obtained at the output of a microphone.
(b) Electrical output of a transducer used to sense the temperature of a furnace.
(c) Stream of electrical pulses (digital) generated by a computer.
(d) Electrical output of a TV camera (video signal).
(e) Electrical waves received by the antenna of a radio/TV/communication receiver.
(f) ECG signal.

When a 'signal' is viewed as electrical manifestation of a process, the signal is a function of one or more independent variables. For all the examples cited above, the respective signals may commonly be considered as function of 'time'. So, a notation like the following may be used to represent a signal:

s(a, b, c, t,..), where 'a', 'b',… are the independent variables.

However, observe that a mere notation of a signal, say s(t), does not reveal all its features and behavior and hence it may not be possible to analyze the signal effectively. Further, processing and analyses of many signals may become easy if we can associate them, in some cases even approximately, with mathematical functions that may be analyzed by well-developed mathematical tools and techniques. The approximate representations, wherever adopted, are usually justified by their ease of analysis or tractability or some other evidently rewarding reason. For example, the familiar mathematical function $s(t) = A \cos(\omega t + \theta)$ is extensively used in the study, analysis and testing of several principles of communication theory such as carrier modulation, signal sampling etc. However, one can very well contest the fact that $s(t) = A \cos(\omega t + \theta)$ hardly implies a physical process because of the following reasons:

(i) no range of 't' is specified and hence, mathematically the range may be from $-\infty$ to $+\infty$. This implies that the innocent looking function s(t) should exist over the infinite range of 't', which is not true for any physical source if 't' represents time. So, some range for 't' should be specified.

(ii) s(t), over a specified range of 't', is a known signal in the sense that, over the range of 't', if we know the value of s(t) at say $t = t_0$, and the values of A, ω and θ we know the value of s(t) at any other time instant 't'. We say the signal s(t) is deterministic. In a sense, such a mathematical function does not carry information.

While point (i) implies the need for rigorous and precise expression for a signal, point (ii) underlines the usage of theories of mathematics for signals deterministic or non-deterministic (random).

To illustrate this second point further, let us consider the description of s(t) = A cos ωt, where 't' indicates time and ω = 2πf implies angular frequency:

(a) Note that s(t) = A cos ωt , -∞ < t < ∞ is a periodic function and hence can be expressed by its exponential (complex) Fourier series. However, this signal has infinite energy E, $E = \int_{-\infty}^{+\infty} s^2(t)dt$ and hence, theoretically, can not be expressed by Fourier Transformation.

(b) Let us now consider the following modified expression for s(t) which may be a closer representation of a physical signal:

$s(t) = A \cos \omega t,\ 0 \le t < \infty$

$= A.u(t). \cos \omega t$ where u(t) is the unit step function, u(t) = 0, t < 0 and

$u(t) = 1, t \ge 0$

If we further put an upper limit to 't', say, s(t) = A cos ωt , $t_1 \le t \le t_2$, such a signal can be easily generated by a physical source, but the frequency spectrum of s(t) will now be different compared to the earlier forms. For simplicity in notation, depiction and understanding, we will, at times, follow mathematical models for describing and understanding physical signals and processes. We will, though, remember that such mathematical descriptions, while being elegant, may show up some deviation from the actual behavior of a physical process. Henceforth, we will mean the mathematical description itself as the signal, unless explicitly stated otherwise.

Now, we briefly introduce the major classes of signals that are of frequent interest in the study of digital communications. There are several ways of classifying a signal and a few types are named below.

Energy signal: If, for a signal s(t), $\int_{0}^{+\infty} s^2(t)dt < \infty$ i.e. the energy of the signal is finite, the signal is called an energy signal. However, the same signal may have large power. The voltage generated by lightning (which is of short duration) is a close example of physical equivalent of a signal with finite energy but very large power.

Power signal: A power signal, on the contrary, will have a finite power but may have finite or infinite energy. Mathematically,

$$\lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{+T/2} s^2(t)\,dt < \infty$$

Note: While electrical signals, derived from physical processes are mostly energy signals, several mathematical functions, usually deterministic, represent power signals.

Deterministic and random signals: If a signal s(t), described at $t = t_1$ is sufficient for determining the signal at $t = t_2$ at which the signal also exists, then s(t) represents a deterministic signal.

Example: $s(t) = A \cos \omega t$, $T_1 \le t \le T_2$

There are many signals that can best be described in terms of a probability and one may not determine the signal exactly.

Example: (from real process) Noise voltage/current generated by a resistor.

Such signals are labeled as non-deterministic or random signals.

Continuous time signal: Assuming the independent variable 't' to represent time, if s(t) is defined for all possible values of t between its interval of definition (or existence), $T_1 \le t \le T_2$. Then the signal s(t) is a continuous time signal.

If a signal s(t) is defined only for certain values of t over an interval $T_1 \le t \le T_2$, it is a discrete-time signal. A set of sample values represent a discrete time signal.

Periodic signal: If s(t) = s(t + T), for entire range of t over which the signal s(t) is defined and T is a constant, s(t) is said to be periodic or repetitive. 'T' indicates the period of the signal and $1/T$ is its frequency of repetition.

Example: $s(t) = A \cos \omega t$, $-\infty \le t \le \infty$, where $T = 2\pi/\omega$.

Analog: If the magnitudes of a real signal s(t) over its range of definition, $T_1 \le t \le T_2$, are real numbers (there are infinite such values) within a finite range, say, $S_{min} \le S(t) \le S_{max}$, the signal is analog.

A digital signal s(t), on the contrary, can assume only any of a finite number of values. Usually, a digital signal implies a discrete-time, discrete-amplitude signal.

The mathematical theories of signals have different flavours depending on the character of a signal. This helps in easier understanding and smarter analyses. There may be considerable similarities among the mathematical techniques and procedures. We assume that the reader has some familiarity with the basic techniques of Fourier series expansion and Fourier Transform. In the following, we present a brief treatise on sampling theory and its implications in digital communications.

Sampling Theorem

The concepts and techniques of sampling a continuous-time signal have important roles in baseband signal processing like digitization, multiplexing, filtering and also in carrier modulations in wireless digital communication systems. A common use of sampling theorem is for converting a continuous-time signal to an equivalent discrete-time signal and vice versa.

Generally, a modulated signal in a wireless system is to be transmitted within an allocated frequency band. To accomplish this, the modulating message signal is filtered before it is modulated and transmitted. When the message signal is already available in digital form, an appropriate pulse shaping operation may be performed on the digital stream before it modulates a carrier. We will see later in this section how the basic concepts of sampling may be adapted to shape the digital pulses. In case the message is available as a continuous-time signal, it is first band-limited and then sampled to generate approximately equivalent set of pulses. A set of pulses may be called equivalent to the original analog signal if a technique exists for reconstructing the filtered signal uniquely from these pulses. Nyquist's famous theorems form the basis of such preprocessing of continuous-time signals.

Nyquist's Uniform Sampling Theorem for Lowpass Signals

Part - I If a signal x(t) does not contain any frequency component beyond W Hz, then the signal is completely described by its instantaneous uniform samples with sampling interval (or period) of $T_s < 1/(2W)$ sec.

Part – II The signal x(t) can be accurately reconstructed (recovered) from the set of uniform instantaneous samples by passing the samples sequentially through an ideal (brick-wall) lowpass filter with bandwidth B, where $W \le B < f_s – W$ and $f_s = 1/(T_s)$.

As the samples are generated at equal (same) interval (T_s) of time, the process of sampling is called uniform sampling. Uniform sampling, as compared to any non-uniform sampling, is more extensively used in time-invariant systems as the theory of uniform sampling (either instantaneous or otherwise) is well developed and the techniques are easier to implement in practical systems.

The concept of 'instantaneous' sampling is more of a mathematical abstraction as no practical sampling device can actually generate truly instantaneous samples (a sampling pulse should have non-zero energy). However, this is not a deterrent in using the theory of instantaneous sampling, as a fairly close approximation of instantaneous sampling is sufficient for most practical systems. To contain our discussion on Nyquist's theorems, we will introduce some mathematical expressions.

If x(t) represents a continuous-time signal, the equivalent set of instantaneous uniform samples $\{x(nT_s)\}$ may be represented as,

$$\{x(nT_s)\} \equiv x_s(t) = \Sigma\, x(t).\delta(t- nT_s) \qquad 1.2.1$$

where $x(nT_s) = x(t)\big|_{t=nTs}$, $\delta(t)$ is a unit pulse singularity function and 'n' is an integer

Conceptually, one may think that the continuous-time signal x(t) is multiplied by an (ideal) impulse train to obtain $\{x(nT_s)\}$ as (1.2.1) can be rewritten as,

$$x_s(t) = x(t).\Sigma\, \delta(t- nT_s) \qquad 1.2.2$$

Now, let X(f) denote the Fourier Transform F(T) of x(t), i.e.

$$X(f) = \int_{-\infty}^{+\infty} x(t).\exp(-j2\pi ft)dt \qquad 1.2.3$$

Now, from the theory of Fourier Transform, we know that the F.T of $\Sigma\ \delta(t-nT_s)$, the impulse train in time domain, is an impulse train in frequency domain:

$$F\{\Sigma\ \delta(t-nT_s)\} = (1/T_s).\Sigma\ \delta(f-n/T_s) = f_s.\Sigma\ \delta(f-nf_s) \qquad 1.2.4$$

If $X_s(f)$ denotes the Fourier transform of the energy signal $x_s(t)$, we can write using Eq. (1.2.4) and the convolution property:

$$\begin{aligned} X_s(f) &= X(f)^* F\{\Sigma\ \delta(t-nT_s)\} \\ &= X(f)^*[f_s.\Sigma\ \delta(f-nf_s)] \\ &= f_s.X(f)^*\Sigma\ \delta(f-nf_s) \\ &= f_s.\int_{-\infty}^{+\infty} X(\lambda).\Sigma\ \delta(f-nf_s-\lambda)d\lambda = f_s.\ \Sigma \int X(\lambda).\delta(f-nf_s-\lambda)d\lambda = f_s.\Sigma\ X(f-nf_s) \end{aligned} \qquad 1.2.5$$

[By sifting property of $\delta(t)$ and considering $\delta(f)$ as an even function, i.e. $\delta(f) = \delta(-f)$]

This equation, when interpreted appropriately, gives an intuitive proof to Nyquist's theorems as stated above and also helps to appreciate their practical implications. Let us note that while writing Eq.(1.2.5), we assumed that x(t) is an energy signal so that its Fourier transform exists. Further, the impulse train in time domain may be viewed as a periodic singularity function with almost zero (but finite) energy such that its Fourier Transform [i.e. a train of impulses in frequency domain] exists. With this setting, if we assume that x(t) has no appreciable frequency component greater than W Hz and if $f_s > 2W$, then Eq.(1.2.5) implies that $X_s(f)$, the Fourier Transform of the sampled signal $x_s(t)$ consists of infinite number of replicas of X(f), centered at discrete frequencies $n.f_s$, $-\infty < n < \infty$ and scaled by a constant $f_s = 1/T_s$ (**Fig. 1.2.1**).

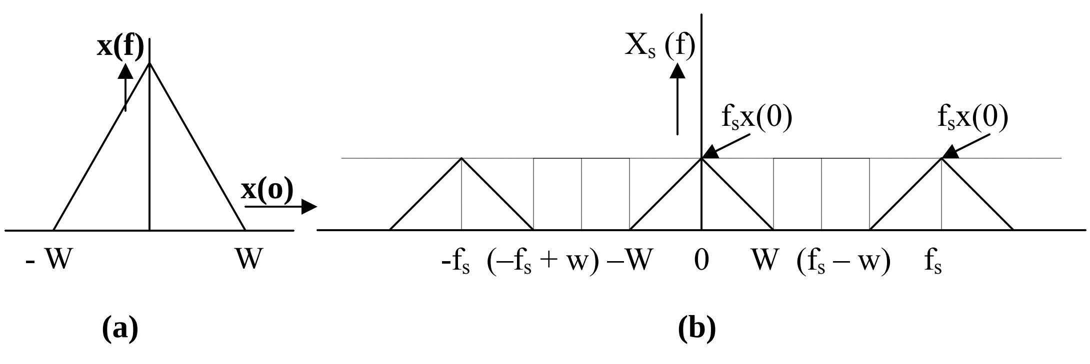

Fig. 1.2.1 *Spectra of (a) an analog signal x(t) and (b) its sampled version*

Fig. 1.2.1 indicates that the bandwidth of this instantaneously sampled wave $x_s(t)$ is infinite while the spectrum of x(t) appears in a periodic manner, centered at discrete frequency values $n.f_s$.

Now, Part – I of the sampling theorem is about the condition $f_s > 2.W$ i.e. $(f_s – W) > W$ and $(– f_s + W) < – W$. As seen from Fig. 1.2.1, when this condition is satisfied, the spectra of $x_s(t)$, centered at f = 0 and $f = \pm f_s$ do not overlap and hence, the spectrum of x(t) is present in $x_s(t)$ without any distortion. This implies that $x_s(t)$, the appropriately sampled version of x(t), contains all information about x(t) and thus represents x(t).

The second part of Nyquist's theorem suggests a method of recovering x(t) from its sampled version $x_s(t)$ by using an ideal lowpass filter. As indicated by dotted lines in Fig. 1.2.1, an ideal lowpass filter (with brick-wall type response) with a bandwidth $W \leq B < (f_s – W)$, when fed with $x_s(t)$, will allow the portion of $X_s(f)$, centered at f = 0 and will reject all its replicas at $f = n\, f_s$, for $n \neq 0$. This implies that the shape of the continuous-time signal $x_s(t)$, will be retained at the output of the ideal filter. The reader may, supposedly, have several queries at this point such as:

(a) Can a practical LPF with finite slope be used when $W \leq B < (f_s – W)$?
(b) Can we not recover x(t) if there is overlapping ?
(c) What happens to the above description when non-ideal but uniform sampling (like flat-top or natural) is used instead of instantaneous sampling?

One may observe that the above questions are related towards use of Nyquist's sampling theorems in practical systems. Instead of addressing these questions directly at this point, we wish to cast the setting a bit more realistic by incorporating the following issues:

(1) a practical uniform sampling scheme in place of instantaneous sampling (flat top sampling) and
(2) frequency response of conveniently realizable analog lowpass filters.

Flat Top Sampling

A train of pulses with narrow width (τ), rather than an impulse train, is practically realizable (**Fig.1.2.2**). The pulse width τ, though small, is considered to be significant compared to the rise and fall times of the pulses to avoid unnecessary mathematical complexity.

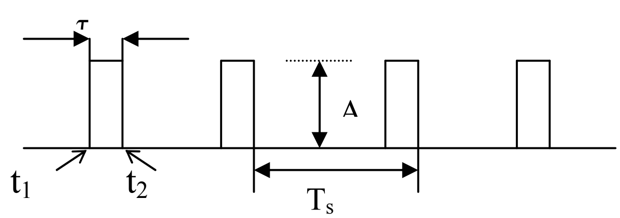

Fig. 1.2.2 *A train of narrow pulses with pulse width 'τ'*

In a flat top sampling scheme, the amplitude of a pulse after sampling is kept constant and is related to the value of signal x(t) at a pre-decided instant within the pulse duration τ. (e.g., we may arbitrarily decide that the pulse amplitude after sampling will be proportional to x(t) at the beginning instant of the pulse). The point to be noted is that, though τ is non zero, the top of a sampled pulse does not follow x(t) during τ and instead is held constant, thus retaining the flavour of instantaneous sampling to an extent. However, the spectrum of such flat-lop sampled signal for the same signal x(t), is different from the earlier spectrum.

The periodic train of sampling pulses with width τ, amplitude 'A' and sampling interval 'T', may be expressed as,

$$P(t)=\sum_{n=-\infty}^{\infty} A\pi\left(t-nT_s\tau/2\right) \qquad 1.2.6$$

Here, π(t) indicates a unit pulse of width τ, $-\tau/2 \le t < \tau/2$.

On Fourier Series expansion of p(t), it can be shown that the envelop of p(t) follows a |sinc x| function instead of a constant amplitude train of frequency pulses as earlier. Hence, the spectrum of the flat-top sampled signal $x_s(t)$ will not show up the spectrum of x(t) and its exact replicas at $f = \pm nf_s$. Instead, the spectrum of x(t) will be contoured by the $\text{sinc}(n\tau f_s)$ function. **Fig. 1.2.3** sketches this feature only for the main lobe of the sinc function.

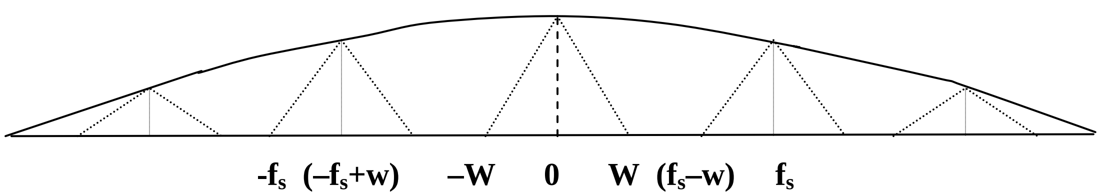

Fig. 1.2.3 *A sketch indicating the change in $X_s(f)$ due to flat-top sampling w.r.t* **Fig. 1.2.1 (b)**

So to be exact, an ideal lowpass filter, with flat amplitude response (versus frequency) will recover the signal x(t) with some distortion. The higher frequency components of x(t) will get more attenuated compared to the lower frequency components of x(t). Sometimes, 'τ' is chosen sufficiently small such that the distortion in the recovered signal x(t) is within a tolerable limit. However, one may note from **Fig. 1.2.3** that perfect recovery of x(t) may still be possible if a different and special kind of lowpass filter is used (instead of an ideal LPF) which can compensate the 'sin x/x' amplitude distortion (and an associated phase distortion which has not been highlighted for the sake of brevity).

Frequency response of conveniently realizable analog lowpass filter

A system designer more often than not searches for a convenient filter realization, which can strike a compromise amongst several conflicting properties. To be specific, let us reiterate that Nyquist's first sampling theorem gives a lower bound of the sampling rate (f_s= 2W) and from **Fig.1.2.2** we can infer that the replicas of the spectrum of x(t) can be separated more from one another by increasing the sampling rate. When the replicas are widely separated, a practical LPF of low/moderate order and with constant passband covering at least up to W H_Z (and also maintaining linear phase relationship up to W H_Z) is good enough to select the spectrum of x(t) located around f = 0. While higher and higher sampling rate eases the restriction on the design of lowpass reconstruction filter, other issues such as bit rate, multiplexing several signals, cost and complexity of the high speed sampling circuit usually come in the way. On the whole, the sampling rate f_s is chosen marginally higher than 2W samples/sec striking a balance among the contending issues.

Sampling of narrow bandpass signals: - an inefficient approach

A form of a narrow bandpass signal that is often encountered in the design and analysis of a wireless communication system is:

$$x(t) = A(t)\cos\{\omega_c t + \theta(t)\} \qquad 1.2.7$$

The spectrum of such a bandpass signal is centered at frequency f_c (= $\omega_c/2\pi$) and the bandwidth is usually small (less than 10%) compared to the centre frequency. Is it possible to represent such a continuous time signal by discrete-time samples? If possible, how is the sampling rate related to f_c and the bandwidth of the signal? How to reconstruct the original signal x(t) from its equivalent set of samples? We will try to find reasonable answers to these questions in the following introductory discussion.

Let the bandpass signal x(t), centered around 'f_c' have a band width 2B i.e. let x(t) be band limited from (f_c-B) to (f_c+B). By taking clue from Nyquist's uniform sampling theorem for lowpass signals, one may visualize the bandpass signal x(t) as a real lowpass signal, whose maximum frequency content is (f_c+B) and the lowest allowable frequency is 0 H_Z though actually there is no frequency component between 0 H_Z and (f_c-B) H_Z. Then one may observe that if x(t) is sampled at a rate greater than 2x(f_c+B), a set of valid samples is obtained which completely represents x(t). While this general approach is flawless, there exists an efficient and smarter way of looking at the problem. We will come back to this issue after having a brief look at complex low pass equivalent description of a real bandpass signal with narrow bandwidth.

Base band representation of narrow bandpass signal

Let, for a narrow band signal,

f_c: Center Frequency $= \frac{f_2 - f_1}{2} + f_1$

f_2: maximum frequency component ; f_1: minimum frequency component

So, the band width of the signal is: BW = $f_2 - f_1$

Now, we describe a real signal as a narrowband signal if BW << f_c. A rule of thumb, used to decide quickly whether a signal is a narrowband, is: $0.01 < (BW/f_c) < 0.1$. That is, the bandwidth of a narrow bandpass signal is considerably less than 10% of the centre frequency [refer **Table 1.1.4]**

Representation of narrow band signals

A general form of expressing a bandpass signal is: $x(t) = A(t)\cos[2\pi f_c t + \Phi(t)]$

Now, x(t) may be rewritten as:

$$x(t) = \{A(t)\cos\Phi(t)\}.\cos 2\pi f_c t - \{A(t)\sin\Phi(t)\}.\sin 2\pi f_c t$$
$$= u_I(t).\cos 2\pi f_c t - u_Q(t).\sin 2\pi f_c t$$
$$= \mathrm{Re}\{\tilde{u}(t).\exp(j2\pi f_c t)\}$$

Where, $\tilde{u}(t) = u_I(t) + ju_Q(t)$

$\tilde{u}(t)$: Complex Low pass Equivalent of x(t)

Note: Real band pass x(t) is completely described by complex low pass equivalent $\tilde{u}(t)$ and the centre frequency 'f_c'.

Spectra of x(t) and $\tilde{u}(t)$:

Let X(f) denote the Fourier Transform of x(t). Then,

$$X(f) = \int_{-\infty}^{\infty} x(t)\exp(-j2\pi ft)\,dt$$
$$= \int_{-\infty}^{\infty} R_e\{\tilde{u}(t)\exp(j2\pi f_c t)\}.\exp(-j2\pi ft)\,dt$$

Now, use the subtle observation that, for any complex number Z, $R_e\{\tilde{Z}\} = \frac{1}{2}\{\tilde{Z} + \tilde{Z}^*\}$

$$\therefore X(f) = \frac{1}{2}\int_{-\infty}^{\infty}\left[\tilde{u}(t)\exp(j\omega_c t) + \tilde{u}^*(t)\exp(-j\omega_c t)\right]$$
$$= \frac{1}{2}\left[U(f - f_c) + U^*(-f - f_c)\right],$$

where

$$U(f) = \int_{-\infty}^{\infty} \tilde{u}(t)\exp(-j\omega t)\,dt$$

Fig. 1.2.4 shows the power spectra of a narrowband signal and its lowpass equivalent

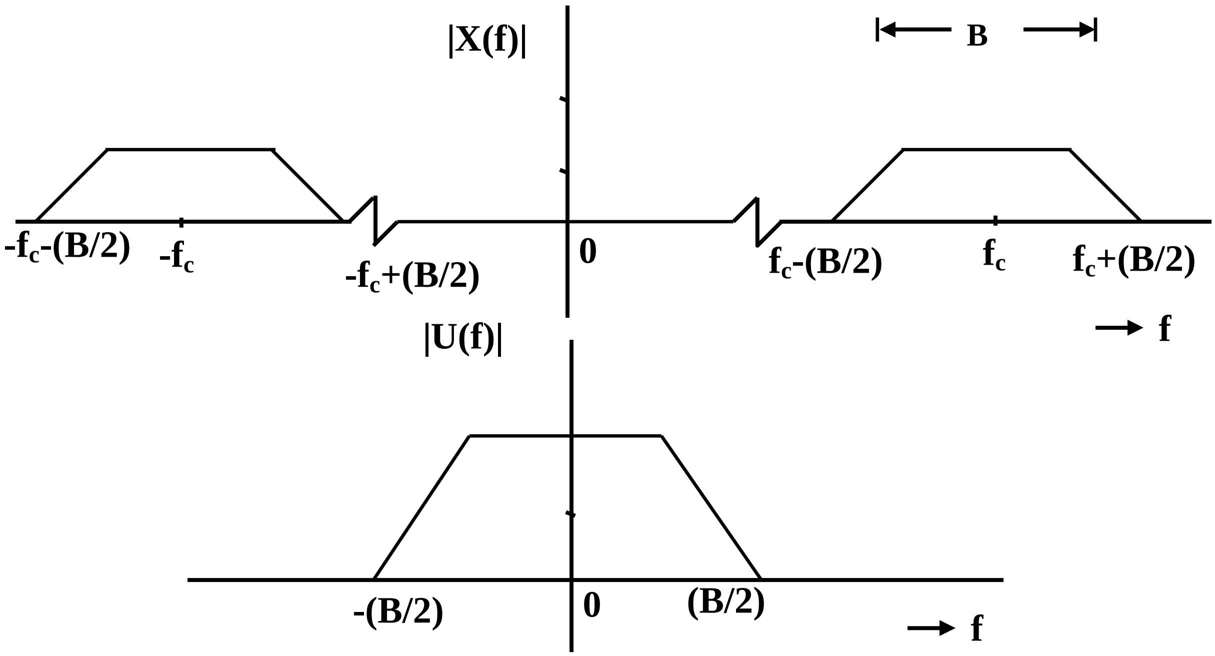

Fig. 1.2.4 *Power spectra of a narrowband signal and its lowpass equivalent*

Now, energy of $x(t) = \int_{-\infty}^{\infty} x^2(t)dt$

$$= \int_{-\infty}^{\infty} \left\{ R_e\left[\tilde{u}(t)\exp\left(j\omega_c t\right)\right]\right\}^2 dt$$

$$= \frac{1}{2}\int_{-\infty}^{\infty} \left|\tilde{u}(t)\right|^2 dt + \frac{1}{2}\int_{-\infty}^{\infty} \left|\tilde{u}(t)\right|^2 .\cos\left[4\pi f_c t + 2\varphi(t)\right] dt$$

Note: $\tilde{u}(t)$ as well as $\left|\tilde{u}(t)\right|^2$ vary slowly compared to the second component whose frequency is around $2f_c$.

This implies, energy of $x(t) \simeq \frac{1}{2}\int_{-\infty}^{\infty} \left|\tilde{u}(t)\right|^2 dt$

The sketch in **Fig. 1.2.5** helps to establish that the energy associated with a narrow band pass signal can be closely assessed by the energy of its equivalent complex lowpass representation.

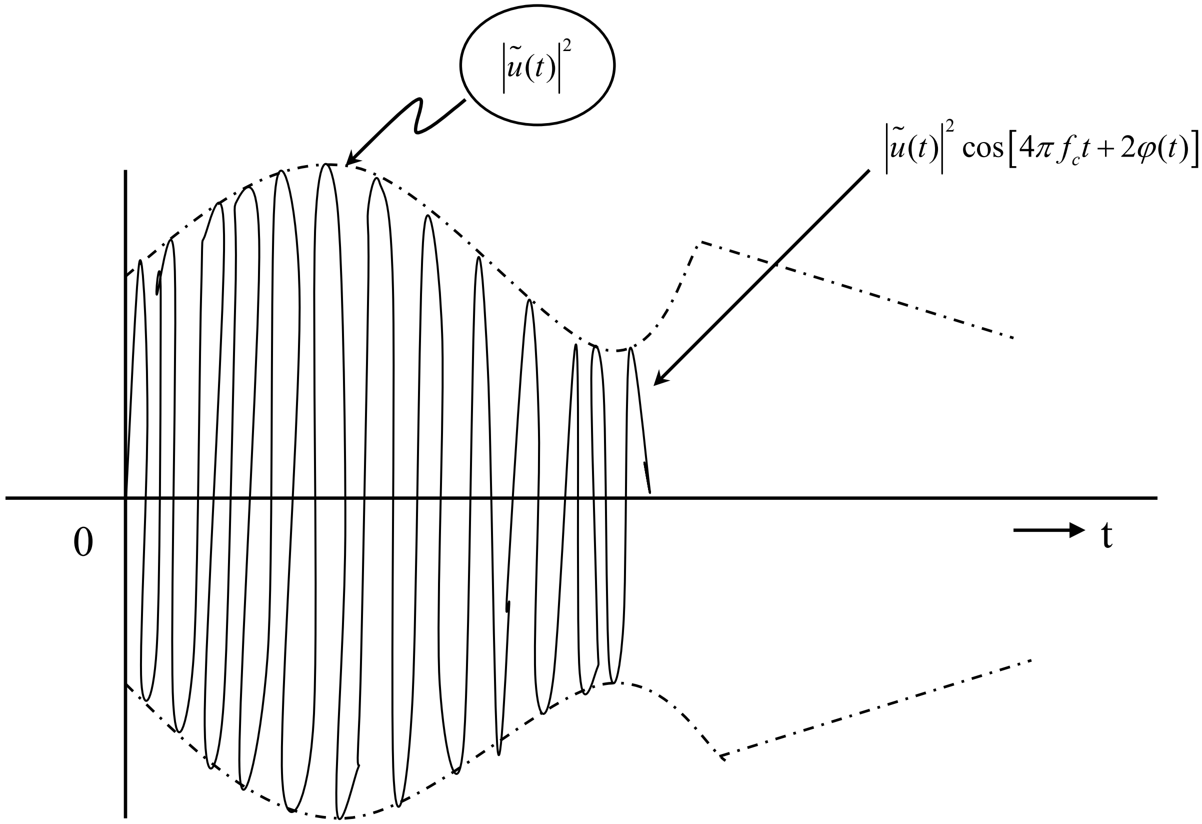

Fig. 1.2.5 *Sketch to establish that the energy associated with a narrow band pass signal can be closely assessed by the energy of its equivalent complex lowpass representation*

Sampling of narrow bandpass signals: - a better approach

Let u_I (t) and u_Q (t) represent x(t) such that each of them is band limited between 0 and B H_Z. Now, if u_I (t) and u_Q (t) are given instead of the actual signal x(t), we could say that the sampling rate for each of u_I (t) and u_Q (t) would be only 2B samples/sec [much less than $2(f_c+B)$ as f_c>>B]. Hence, the equivalent sampling rate may be as low as 2 x 2B or 4B samples/sec. **Fig.1.2.6** shows a scheme for obtaining real lowpass equivalent u_I (t) and u_Q (t) from narrow bandpass signal x(t) and their sampling.

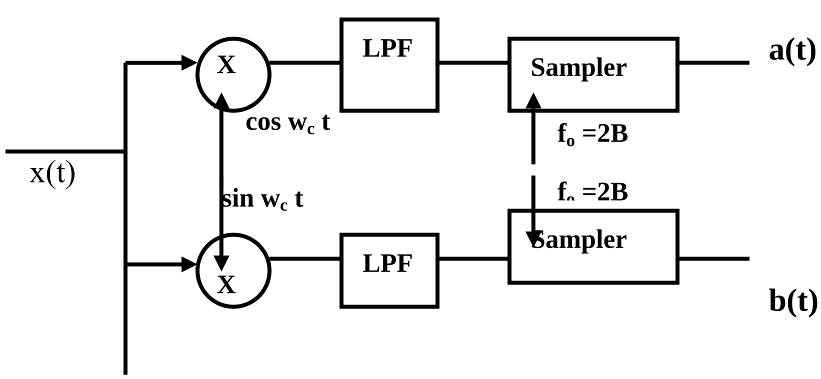

Fig.1.2.6 *A scheme for obtaining real lowpass equivalent a(t) and b(t) from narrow bandpass signal x(t) and their sampling*

Let us now have a look at the next figure (**Fig. 1.2.7**), showing the spectrum of x(t) and also the spectra of a train of sampling impulses at an arbitrary rate f_{1s}. See that, if we make $m.f_{1s} = f_c$ ('m' is an integer) and select f_{1s} such that, after sampling, there is no spectral overlap (or aliasing), then,

$$f_{1smin.} = 2.B$$

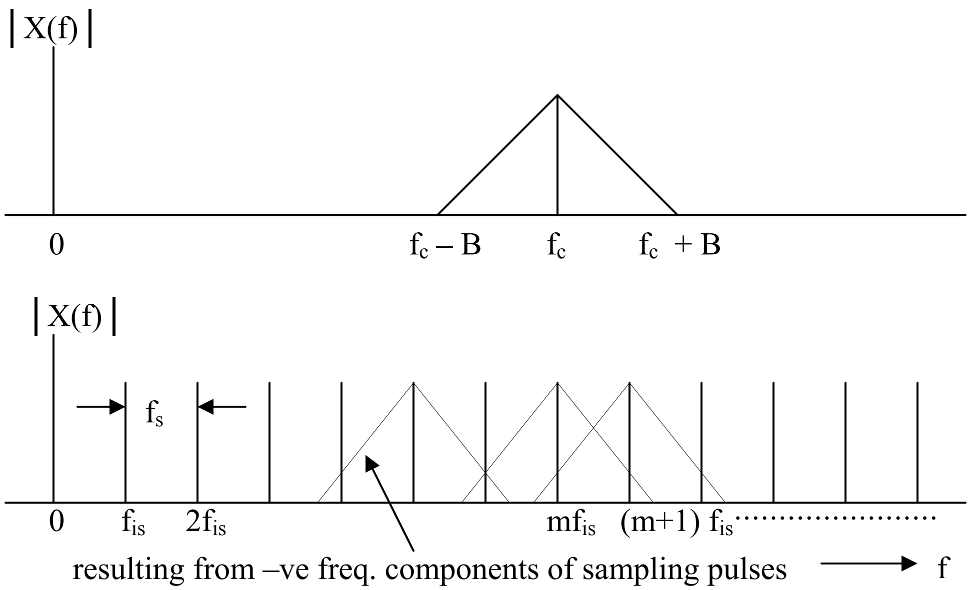

Fig. 1.2.7 *The spectrum of x(t) and the spectra of a train of sampling impulses at an arbitrary rate* f_{1s}

Further, even if we don't satisfy the above condition (i.e. $m.f_{1s} = f_c$) precisely but relax it such that, $(f_c\text{-}B) < m.f_{1s} < (f_c\text{+}B)$ for some integral m, it is possible to avoid spectral overlap for the shifted spectrum (after sampling) provided, $2 \times 2B < f_{1s} = 4 \times 2B$ [The factor of 2 comes to avoid overlapping from shifted spectra due to –ve frequency of the sampling impulses]. These basic issues on sampling will be useful in subsequent modules.

Problems

Q1.2.1) Why an electrical signal should be represented carefully and precisely?

Q1.2.2) How can a random electrical signal be represented?

Q1.2.3) Mention two reasons why a signal should be sampled for transmission through digital communication system.

Q1.2.4) Mention three benefits of representing a real narrowband signal by its equivalent complex baseband form