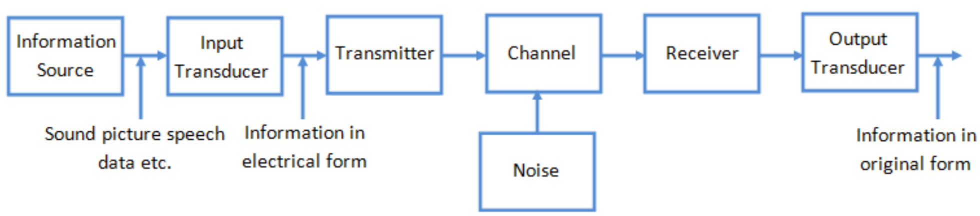
Communication Systems

Dhruva Hegde

1 Introduction to Communication

Communication is the process of exchanging (sending and receiving) data or information.

1.1 Block Diagram of Communication System



Information source generates information in different forms such as text, audio, picture, video, etc.These are non-electrical signals, which need to be converted to electrical (or electromagnetic) signals which is done by input transducer.

Transmitter is a device (or group of devices) that is used to modulated the input electrical signal and send it over a channel.

Channel is the physical medium in which the signal carrying the information travels. It can be wired (like optical fibre or copper cable) or wire-less.

Receiver is a device (or group of devices) that is used to recover back the original information from the signal obtained through the channel.

The signal at the receiver end will be in electrical form, which again needs to be converted back to it's original form. This is accomplished using output transducer, which outputs the data to the destination in it's original form.

Noise is any distortion or loss or data or addition of unwanted data on the original signal. This can happen due to various practical reasons in any of the blocks, but for convenience, noise is modelled to be added during the transportation of the signal in the channel.

1.2 Modulation

Modulation is the process of varying some feature of a high frequency carrier signal with respect to a low frequency message signal.
In simpler words, modulation is implanting the low frequency message on a high frequency carrier due for transmission and reception.

1.2.1 Need for Modulation

Wired communication generally does not required modulation. However, wireless communication is more extensively used because wired communication requires man-made connections to exist between the source and the destination.

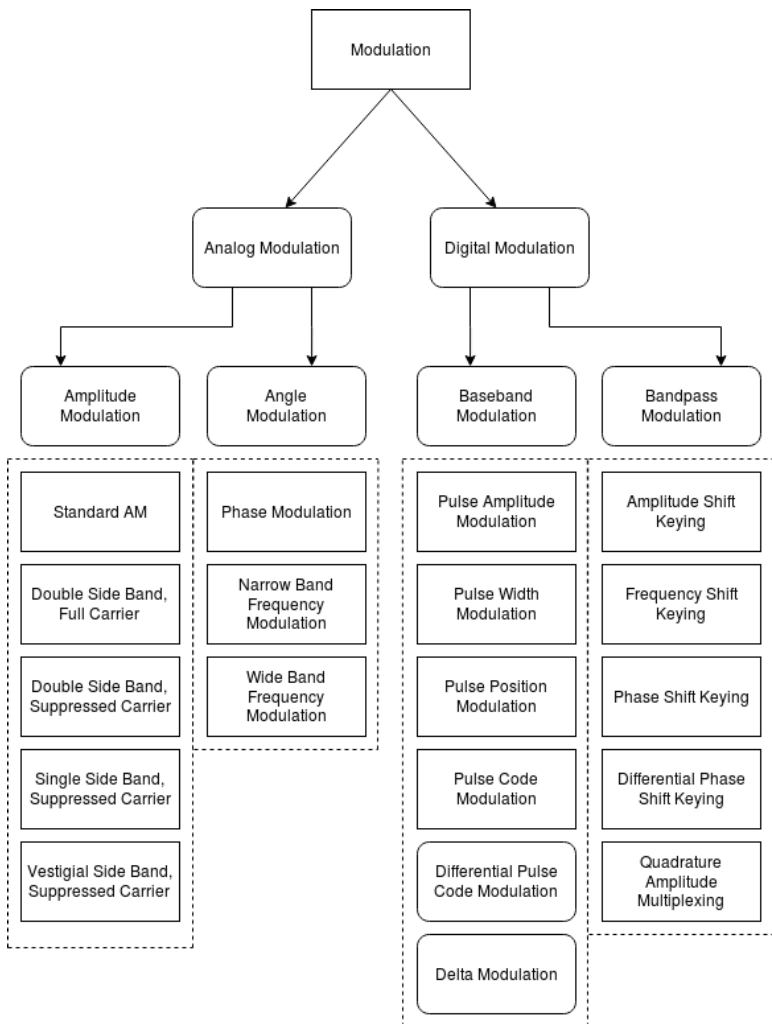
In wireless communication, the information is converted to electromagnetic waves. This is accomplished with the help of a transmitting antenna, which is transmitted through air. And the receiving antenna converts the received electromagnetic wave to an electrical signal which is processed by the receiver.

For the antenna to be able to transmit or receive an electromagnetic signal, it's length should be of the order of the wavelength of the electromagnetic signal. Typically, if λ is the wavelength of the EM wave, then the antenna length should be around $\lambda/4$.
Now audio signals are usually in the frequency range of 300-3000 Hz, meaning their wavelengths will be of the order 100-1000 km ($\lambda = c/f$), for which the antenna lengths are impossible to realize practically.
Hence, to reduce the length of antenna, a high frequency carrier (in the order of GHz) is modulated with respect to the message, which can be transmitted and received using antennae whose lengths would be of the order cm or mm.

Modulation also have other advantages such as allowing for multiplexing, increasing signal to noise ratio, etc..
Note that the received signal can't directly be used, it has to be demodulated. Meaning, the low frequency message has to be re-obtained from the modulated signal before sending to the output destination.

1.2.2 Types of Modulation

The various types of modulation schemes are shown in the flowchart.



1.3 Hilbert Transform

Hilbert Transform a time domain transformation, i.e it transforms a time domain signal to another time domain signal.
Hence, a Hilbert Transformer can be thought of as an LTI system of impulse response $h_H(t)$ which operates on a signal $x(t)$ to give Hilbert transform of $x(t)$ which is denoted by $\hat{x}(t)$.
$\implies \hat{x}(t) = h_H(t) * x(t)$

Taking Fourier Transform on both sides for the above equation,
$\hat{X}(f) = H_H(f)\ X(f)$.

$$H_H(f) = -j\ \text{sgn}(f) = \begin{cases} -j & : f > 0 \\ 0 & : f = 0 \\ j & : f < 0 \end{cases}$$

The magnitude and phase of the above function will give more clarity on the nature of behaviour of Hilbert transformer in the frequency domain.

$$|H_H(f)| = \begin{cases} 1 & : f \neq 0 \\ 0 & : f = 0 \end{cases}$$

The magnitude is unaffected except for the zero frequency (i.e DC) component on $x(t)$, which is completely attenuated. Hence, Hilbert transform kills DC input.

$$\angle H_H(f) = \begin{cases} -\pi/2 & : f > 0 \\ 0 & : f = 0 \\ \pi/2 & : f < 0 \end{cases}$$

The phase is shifted by -90^o for all positive frequencies and by 90^o for all negative frequencies. Hence, Hilbert transform is basically a 90^o phase shifter.

From the expression of $H_H(f)$, using derivative and duality properties of Fourier Transform, it can be deduced that the impulse response of the Hilbert transformer is given by, $h_H(t) = \frac{1}{\pi t}$.

$$\implies \hat{x}(t) = \frac{1}{\pi t} * x(t)$$

Properties of Hilbert Transform

- $ax(t) + by(t) \leftrightarrow a\hat{x}(t) + b\hat{y}(t)$
- $ax(t - t_0) \leftrightarrow a\hat{x}(t - t_0)$
- $x(at) \leftrightarrow sgn(a)\hat{x}(t)$
- $\frac{d[x(t)]}{dt} \leftrightarrow \frac{d[\hat{x}(t)]}{dt}$
- $x(t) * y(t) \leftrightarrow \hat{x}(t) * y(t) = x(t) * \hat{y}(t)$

If $X(f)$ is a complex quantity i.e $X(f) = a + jb$, then $\hat{X}(f) = b - ja$. Hence, Hilbert transform swaps the real and imaginary components in the frequency domain.

1.3.1 Complex Envelope

Let $x(t)$ be a pass-band signal centered around frequency f_c.
The signal $x(t)+j\hat{x}(t)$ is called the "Complex pre-envelope" of the signal $x(t)$.

The spectrum of the complex pre-envelope is $X(f) + j\hat{X}(f)$. Upon spectral analysis, it can be observed that this will result in,

$$X(f) + j\hat{X}(f) = \begin{cases} 2X(f) & : f > 0 \\ 0 & : f < 0 \end{cases}$$

If the above spectrum is shifted such that it is centered around the 0 frequency, the obtained signal is called the "Complex envelope".
Shifting by f_c in the frequency domain corresponds to multiplying the signal by $e^{-2j\pi f_c}$ in time domain. Hence, the complex envelope in time domain is given by,

$$\tilde{x}(t) = [x(t) + j\hat{x}(t)]e^{-2j\pi f_c}$$

The Complex envelope represents the complex base-band equivalent signal of the band-pass signal $x(t)$.
Observe that $x(t) =$Re$[\tilde{x}(t)e^{2j\pi f_c}]$.

1.4 Correlation Functions

Correlation functions compare two signals and measure the extent of similarity between them with respect to a certain lag. Note that a correlation function is not a function of time, but a function of time delay.

1.4.1 Cross-correlation Function

Consider 2 signals $x(t)$ and $y(t)$.
The cross-correlation function between these signals is defined as,

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)\bar{y}(t-\tau)dt$$

This function characterizes the extent of similarity between $x(t)$ and $y(t)$ for a lag τ.
Note that the cross-correlation operation is same as the convolution between $x(\tau)$ and $\bar{y}(-\tau)$.

$$\implies R_{xy}(\tau) = x(\tau) * \bar{y}(-\tau)$$

Taking FT on both sides of the above equation,

$$S_{xy}(f) = X(f)\ \bar{Y}(f)$$

The frequency domain representation of the cross-correlation function is represented using $S_{xy}(f)$.

1.4.2 Auto-correlation Function

The Auto-correlation function is a special case of cross-correlation in which both the signals being considered are the same.
Meaning, this function characterizes the extent of similarity of a signal with itself for a lag τ.
Hence, the auto-correlation function of a signal $x(t)$ is given by,

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)\bar{x}(t-\tau)dt = x(\tau) * \bar{x}(-\tau)$$

Note that the auto-correlation function is an even function i.e $R_{xx}(\tau) = R_{xx}(-\tau)$ since the both signals are same so lag being positive or negative changes nothing.

The Fourier transform of auto-correlation function is represented as $S_{xx}(f)$. Taking FT on both sides of the above equation,

$$S_{xx}(f) = X(f)\ \bar{X}(f) = |X(f)|^2$$

Evaluating $R_{xx}(\tau)$ at $\tau = 0$ i.e no lag, it is found that

$$R_{xx}(0) = \int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df$$

Note that the above equation shows Parseval's Theorem.
$R_{xx}(0)$ is the same as integrating the square of the signal for all time, and hence it is equal to the energy of the signal $x(t)$.

Also, since $\tau = 0$ means the function is comparing a signal with itself, the similarity has to be maximum. Hence, $R_{xx}(0)$ is the maximum value that $R_{xx}(\tau)$ can take.

The function $S_{xx}(f)$ represents the spread or distribution of energy of the signal in the frequency domain, and hence it is called **Energy Spectral Density** of the signal $x(t)$.

Energy of the signal $x(t)$ over a band of frequencies $[-W, W]$ is given by,

$$E = \int_{-W}^{W} S_{xx}(f)\ df$$

Note that ESD is a non-negative real quantity i.e $S_{xx}(f) > 0\ \forall\ f$ and it is an even function i.e $S_{xx}(f) = S_{xx}(-f)$.

1.5 Random Process

A random process is a time indexed random variable.
This means, a random process is a process which at every instant of time, takes values from the distribution corresponding to a random variable.
Hence, a random process $X(t)$ is characterized by a probability distribution function $f_X(x, t)$.

The mean of the random process $X(t)$ is given by,

$$\mu_x(t) = E[X(t)] = \int_{-\infty}^{\infty} X(t) f_X(x,t) dx$$

Note that here, x is the dummy variable in the PDF.

The variance of a random process is given by,

$$\sigma^2(t) = E[X^2(t)] - [\mu_x(t)]^2$$

The auto-correlation function can be defined for a random process as,

$$R_{xx}(t_1, t_2) = E[X(t_1)X(t_2)]$$

Hence, the ACF of a random process gives the relation between the same process at two different instances of time.

1.5.1 Wide Sense Stationary Random Process

A WSS random process is a special type of random process whose mean is a constant value (i.e mean does not depend on t) and whose auto-correlation function is only a function of time delay and not of time instances (i.e ACF is a function of τ only and not t).

$$\implies \mu_x(t) = k$$

If $t_1 = t$ and $t_2 = t + \tau$, then the ACF of a WWS random process is given by,

$$R_{xx}(t_1, t_2) = R_{xx}(t, t+\tau) = E[X(t), X(t+\tau)] = R_{xx}(\tau)$$

Hence for a WSS random process, the ACF will be stationary for a given time delay and mean will always be stationary.

If the time delay $\tau = 0$, then the ACF will be,

$$R_{xx}(0) = E[X(t)X(t)] = E[X^2(t)]$$

The value of ACF at $\tau = 0$ gives the average power of the random process.

Note that for a zero mean WSS process ($\mu_x(t) = 0$), the variance and the ACF at $\tau = 0$ are the same i.e mean of square of $X(t)$. Hence, for a WSS process with zero mean, the power of the process is also equal to its variance.

Power Spectral Density of a WSS random process is given by the Fourier transform of the auto-correlation function.

$$\implies S_{xx}(f) = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{-j2\pi f\tau} d\tau$$

PSD of a WSS random process has same properties as ESD of a signal i.e it is always positive, real and even (symmetric).

It can be noted that $R_{xx}(0)$ is equal to the area under PSD. Hence, area under PSD also gives the average power of the WSS random process.

Power of a random process in a band of frequencies f_1 to f_2 is given by,

$$P = \int_{-f_2}^{-f_1} S_{xx}(f)df + \int_{f_1}^{f_2} S_{xx}(f)df = 2\int_{f_1}^{f_2} S_{xx}(f)df$$

Few observations regarding applications of ACF of a WSS random process:

- $R_{xx}(\tau)$ as $\tau \to \infty$ gives the DC power of $X(t)$ if $X(t)$ is an ergodic random process.
  $\implies R_{xx}(\infty) = [E[X(t)]]^2$
  The square root of this is simply the mean $\mu_x(t)$, which gives the DC value of $X(t)$.

- $R_{xx}(\tau)$ at $\tau = 0$ gives the total average power of $X(t)$.
  $\implies R_{xx}(0) = E[X^2(t)]$.
  The square root of this gives the RMS value of $X(t)$.

- AC power can be found as $R_{xx}(0) - R_{xx}(\infty)$.
  $\implies R_{xx}(0) - R_{xx}(\infty) = E[X^2(t)] - [E[X(t)]]^2$.
  Note that this is same as variance of $X(t)$.

Note that for a non WSS random process, the ACF will be a function of both t and τ, so it has to be denoted as $R_{xx}(t, \tau)$.

1.5.2 Random Process' & LTI Systems

Consider a random process $X(t)$ being transmitted through an LTI system of impulse response $h(t)$.
The output $Y(t)$ is given by,

$$Y(t) = X(t) * h(t) = \int_{-\infty}^{\infty} X(t-k)h(k)dk$$

This indicates that $Y(t)$ is also a random process.

If $X(t)$ is a WSS process, then a few interesting properties can be obtained. Since the mean of $X(t)$ is constant, the following result can be derived.

$$E[Y(t)] = \mu_x \int_{-\infty}^{\infty} h(k)dk = \mu_y$$

The resulting random process $Y(t)$ will also have mean that is constant with time.

Similarly, the auto-correlation function of $Y(t)$ is obtained as,

$$R_{yy}(t, t+\tau) = R_{xx}(\tau) * h(\tau) * h(-\tau)$$

$Y(t)$ has ACF that depends only on the delay τ and not on time instances, meaning the ACF is constant of a given time delay.

Therefore, if the input to an LTI system is a WSS random process, then the output of the system will also be a WSS random process.

Since $h(\tau) \leftrightarrow H(f)$ and $h(-\tau) \leftrightarrow \bar{H}(f)$, the PSD of the output random process $Y(t)$ is given by,

$$S_{yy}(f) = S_{xx}(f)H(f)\bar{H}(f) = S_{xx}(f)|H(f)|^2$$

This result is very useful in finding the power of a random signal which is output from LTI systems (such as filters).

2 Analog Communication

As mentioned earlier, modulation and demodulation of message signals are essential for communication systems. In analog communication, a high frequency carrier is modulated to embed a low frequency message.

Consider a sinusoidal carrier $c(t) = A_c \cos(2\pi f_c t + \phi)$ where A_c is its amplitude, f_c is its frequency and ϕ is its phase where $2\pi f_c t+\phi$ represents the angle.

Obviously, analog modulation can be done in 3 ways i.e by changing one of the 3 parameters mentioned above with respect to the message signal.

2.1 Amplitude Modulation

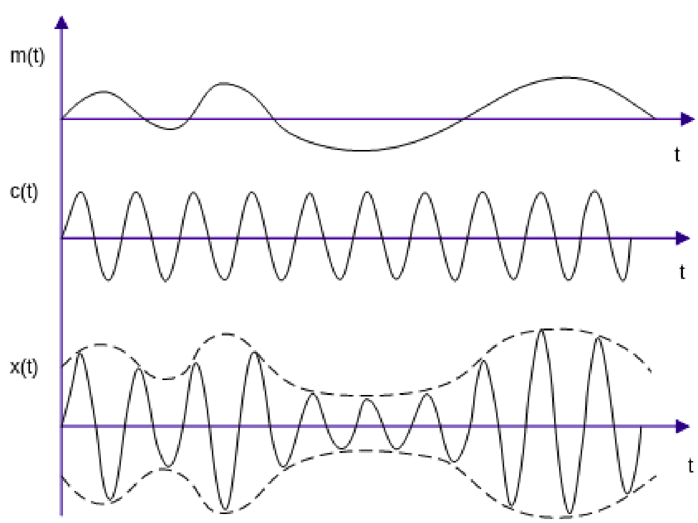
For a high frequency carrier sinusoid of the form $c(t) = A_c \cos(2\pi f_c t + \phi)$, if its amplitude A_c is varied according to the message signal $m(t)$, then the carrier is said to be "amplitude modulated".

2.1.1 Standard Amplitude Modulation

The amplitude modulated signal is given by,

$$x(t) = A_c\ [1 + k_a\ m(t)]\ \cos(2\pi f_c t)$$

where k_a is the "sensitivity" of the modulated signal



Observe that the message signal is encoded in the envelope of the modulated carrier signal.

Consider a single-tone message signal $m(t) = A_m \cos(2\pi f_m t)$.
Then the modulated signal will be,

$$x(t) = A_c\ [1 + \mu\ \ \cos(2\pi f_m t)]\ \ \cos(2\pi f_c t)$$

where $\mu = k_a\ A_m$ which is called "Modulation Index"

The above equation can be further modified as

$$x(t) = A_c \cos(2\pi f_c t) + \frac{\mu A_c}{2} \cos(2\pi(f_c - f_m)t) + \frac{\mu A_c}{2} \cos(2\pi(f_c + f_m)t)$$

where the first term is simply the carrier while the second and third terms represent the lower and upper side bands respectively.

Hence, standard amplitude modulation is also called Double side band full carrier (DSB-FC) modulation.

The value of μ should be such that envelope distortion should not occur while demodulating the modulated signal, which will happen if zero amplitude crossing occurs in the modulated signal.

If $\mu > 1$, then the carrier is said to be over-modulated and it will result in envelope distortion. Hence, for proper recovery of the modulated signal, $\mu \leq 1$.

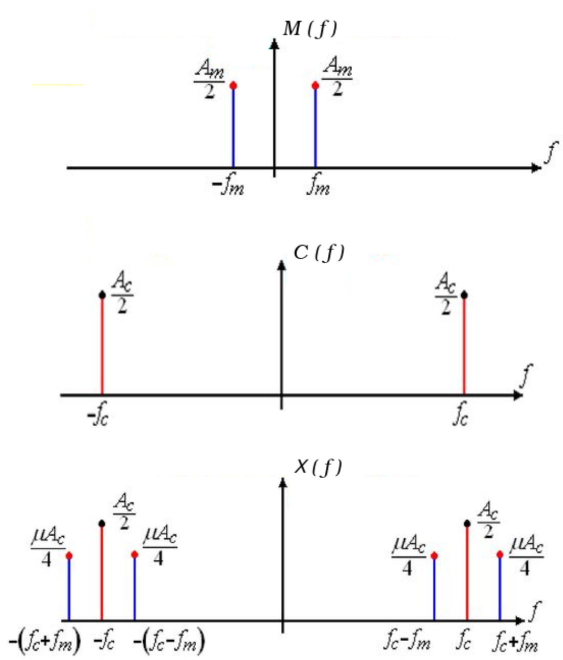
For any arbitrary message signal $m(t)$, the modulation index is given by $\mu = k_a\ |\min[m(t)]|$.

Spectrum of Standard AM Signal
Consider that the message signal $m(t)$ is band-limited to frequency f_c, and $M(f)$ is the spectrum of $m(t)$.
The spectrum of the carrier is found using Fourier transform of a sinusoidal signal which will be two shifted impulses.
Then using modulation property of Fourier transform, the spectrum of the modulating signal $x(t)$ i.e $X(f)$ can be obtained as follows.

$$X(f) = \frac{A_c}{2}\big[\delta(f - f_c) + k_a M(f - f_c) + \delta(f + f_c) + k_a M(f + f_c)\big]$$

It can be noted that the bandwidth of an amplitude modulated signal is twice the maximum frequency component of the message signal.
$BW = (f_c + f_m) - (f_c - f_m) = 2f_m$

For a sinusoidal message signal, the spectrum of the modulated signal will be as illustrated.



Power in AM signal
An AM modulated signal $x(t) = A_c\,[1 + k_a\, m(t)]\ \cos(2\pi f_c t)$ can be split into two components i.e the carrier itself and the modulated message.

The power of the carrier is obviously $P_c = A_c^2/2$.
The power of the message can be approximated by treating it as a sinusoid of amplitude $A_c k_a m(t)$ and hence, the power is given by, $P_{mod} \approx A_c^2 k_a^2 P_m$ where P_m is the power of the message signal $m(t)$.

The total power of $x(t)$ will be

$$P_t = \frac{A_c^2}{2}\left(1 + k_a^2 P_m\right)$$

Since only the message part of the modulated signal carrier information, the efficiency of AM is given by the ratio of modulated message power and the total power.

$$\eta = \frac{k_a^2 P_m}{1 + k_a^2 P_m}$$

Considering the specific case of a sinusoidal message signal $m(t) = A_m \cos(2\pi f_m t)$, the power of the message signal will be $P_m = A_m^2/2$.
Hence, the total power is given by,

$$P_t = \frac{A_c^2}{2}\left(1 + k_a^2 \frac{A_m^2}{2}\right) = \frac{A_c^2}{2}\left(1 + \frac{\mu^2}{2}\right)$$

Note that the power is LSB and USB will be $\mu^2 A_c^2/8$ and hence the side band power (i.e power of the components that carry the message) is $\mu^2 A_c^2/4$.

And the efficiency will be,

$$\eta = \frac{k_a^2 A_m^2}{2 + k_a^2 A_m^2} = \frac{\mu^2}{2 + \mu^2}$$

Since to avoid envelope distortion the value of $\mu \leq 1$, maximum efficiency is obtained when $\mu = 1$.
$\implies \quad \eta_{max} = \frac{1}{3}$ i.e 33.33%.

If I_c is the antenna current before modulation and I_t is the antenna current after modulation, the relation between them is,

$$I_t = I_c \sqrt{1 + \frac{\mu^2}{2}}$$

Some useful results for standard AM calculations:

- Maximum amplitude of modulated signal, $V_{max} = A_c(1 + \mu)$
- Minimum amplitude of modulated signal, $V_{min} = A_c(1 - \mu)$
- $\implies \mu = \frac{V_{max} - V_{min}}{V_{max} + V_{min}}$
  In case of single tone modulation, $\mu = \frac{A_m}{A_c}$

For a multi-tone modulated signal, the effective Modulation Index is given by, $\mu_T = \sqrt{\mu_1^2 + \mu_2^2 + ...}$ Hence, the total power will be,

$$P_t = P_c \left(1 + \frac{\mu_1^2 + \mu_2^2 + ...}{2}\right)$$

Square Law Modulator

Standard AM signal is generated using the square law modulator. The message and carrier are added to obtain $v_1(t) = m(t) + c(t)$, which is passed through a non-linear system which generates $v_2(t) = a\ v_1(t) + b\ v_1^2(t)$.

$$v_2(t) = a\ m(t) + a\ c(t) + b\ m^2(t) + b\ c^2(t) + 2b\ m(t)c(t)$$

The above signal is then passed through a band pass filter of cut-off frequencies $f_c - f_m$ and $f_c + f_m$.

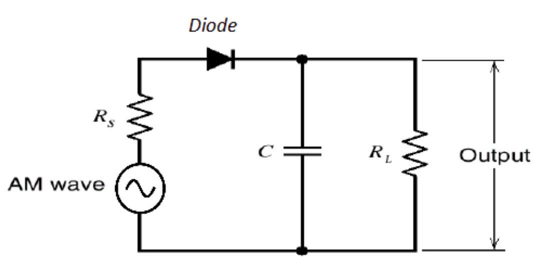
$$\implies x(t) = a\ c(t) + 2b\ m(t)c(t) = aA_c \cos(2\pi f_c t) + 2bA_c m(t) \cos(2\pi f_c t)$$

The same can be expressed in standard AM form as,

$$x(t) = aA_c\big(1 + \frac{2b}{a} m(t)\big) \cos(2\pi f_c t)$$

Envelope Detector

The envelope detector is a simple circuit that is used to recover back the message signal from a modulated signal by tracking the envelope of the modulated signal.
This is done using a diode and an RC circuit with appropriate time constant.



When the modulated signal amplitude is increasing, the diode will be forward biased and hence the capacitor charges. When the modulated signal amplitude is decreases, the diode is reverse biased so the capacitor can't discharge

back through the same path, but it has to discharge through the load.
If the charging time constant ($\tau_c = R_s C$) is low and discharging time constant ($\tau_d = R_L C$) is appropriately high, then the message signal will be tracked by the envelope detected.

For enveloped detected to work properly, $\tau_c << \frac{1}{f_c}$ and $\frac{1}{f_c} << \tau_d << \frac{1}{f_m}$.

- If τ_d is too close to $\frac{1}{f_c}$, then it will start tracking the carrier itself instead of the envelope, which is not desirable.

- If τ_d is too close to $\frac{1}{f_m}$, then it the capacitor discharges too slow and hence will not track any fast changes in the message signal. This is called "diagonal clipping".

Note that envelope detector will not work if the value of Modulation Index is greater than 1.
In general, the output of a properly designed envelope detector will be $|m(t)|$.

Ideal value of discharging time constant is given by,

$$\tau_d \; (= R_L C) = \frac{1}{2\pi f_m} \; \frac{\sqrt{1-\mu^2}}{\mu}$$

2.1.2 Double Side Band, Suppressed Carrier

In standard amplitude modulation, the carrier is also sent along with the modulated signal. This caused low efficiency.
The DSB-SC scheme only sends the modulated signal and does not send the carrier (hence the name).

$$s(t) = m(t) \; A_c \cos(2\pi f_c t)$$

The spectrum of a DSB-SC signal is given by,

$$S(f) = \frac{A_c}{2}\big[M(f-f_c) + M(f+f_c)\big]$$

The bandwidth of DSB-SC signal is same as that of standard AM signal.
However, since the carrier is completely suppressed, there is no component

in the signal which does not contain the message.

Meaning, the total power is same as the modulated message signal power which is given by $P_t = A_c^2 P_m/2$ and hence the efficiency of DSB-SC is 100%.

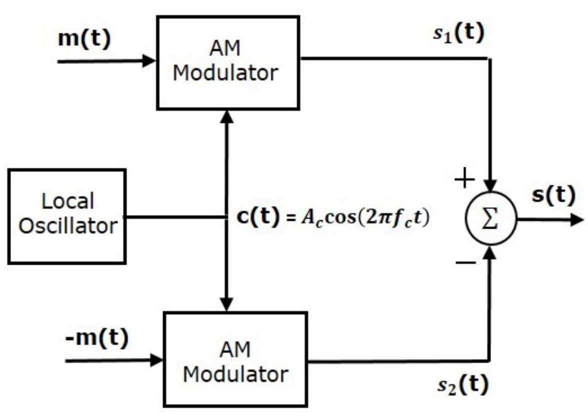
Consider single tone modulation i.e the message is $m(t) = A_m \cos(2\pi f_m t)$. Then the power of the modulated signal is given by $P_t = A_c^2 A_m^2/4$.

Similarly for multi-tone modulation i.e the message is $m(t) = A_{m1} \cos(2\pi f_{m1} t) + A_{m2} \cos(2\pi f_{m2} t) + ...$ Then the power of the modulated signal is given by $P_t = A_c^2 (A_{m1}^2 + A_{m2}^2 + ...)/4$.

Balanced Modulator

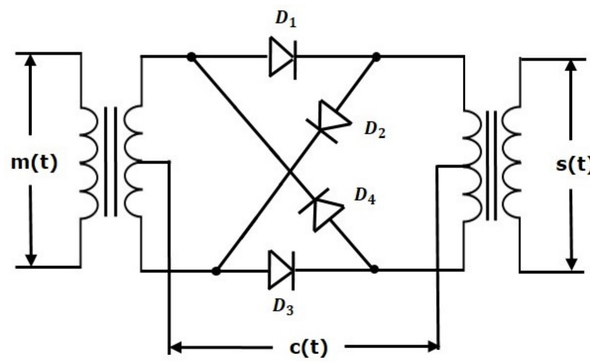
DSB-SC signal can be generated using a balanced modulator which is done by generated two standard AM signals and subtracting them to remove the carrier component.

Here, $s_1(t) = A_c\,[1 + k_a m(t)] \cos(2\pi f_c t)$ and $s_2(t) = A_c\,[1 - k_a m(t)] \cos(2\pi f_c t)$ and hence the final DSB-SC signal is $s_1(t) = 2A_c k_a m(t)\ \cos(2\pi f_c t)$.



Ring Modulator

A more efficient way to generate a DSB-SC signal is by using ring modulator.

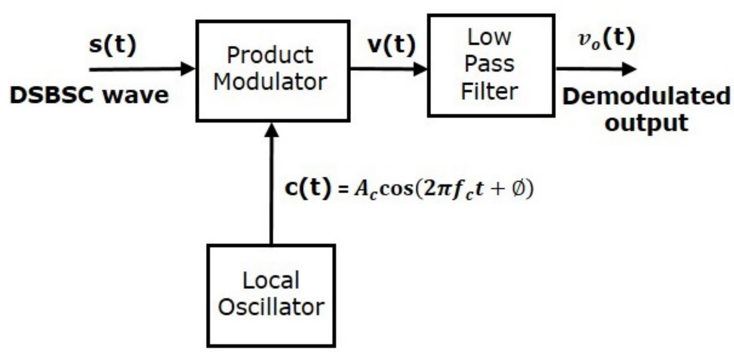


- When the carrier is positive, diodes D_1 and D_3 are on, while D_2 and D_4 are off. Hence, $s(t) = m(t)$.
- When the carrier is negative, diodes D_2 and D_4 are on, while D_1 and D_3 are off. Hence, $s(t) = -m(t)$.

The signal $s(t)$ changes back and forth between $m(t)$ and $-m(t)$ at the high frequency of the carrier, and hence DSB-SC signal is generated.

Synchronous Demodulator

It is obvious that enveloped detector can't be used to demodulate a DSB-SC signal. Hence, the synchronous (or coherent) detector is employed.



The output of the product modulator is given by,

$$v(t) = \frac{A_c^2}{2} m(t) \cos\phi + \frac{A_c^2}{2} \cos(4\pi f_c t + \phi) m(t)$$

If the cut-off frequency of the low-pass filter is chosen to be less than $2f_c - f_m$, then the demodulated output will be a scaled version of the message signal multiplied by the phase difference between the original carrier and then locally generated carrier.

$$m_d(t) = \frac{A_c^2}{2} m(t) \cos\phi$$

Ideally, the phase difference $\phi = 0$ and hence the demodulated signal is simply scaled version of the message.
But due to mobile receivers, when the distance between the transmitter and sender keeps changing, a phase difference is noticed in the modulated signal when it arrives as the received signal.
Note that if $\phi = \pi/2$, then the output will be 0. This effect is called "Quadrature nulling" and it must be avoided.

In coherent demodulation of DSB-SC signal, in-phase component of noise will be retained whereas quadrature-phase component of noise will be rejected.

2.1.3 Quadrature Carrier Multiplexing

QCM is a technique where two message signals are transmitted over the same channel at the same carrier frequency using orthogonal carrier signals.

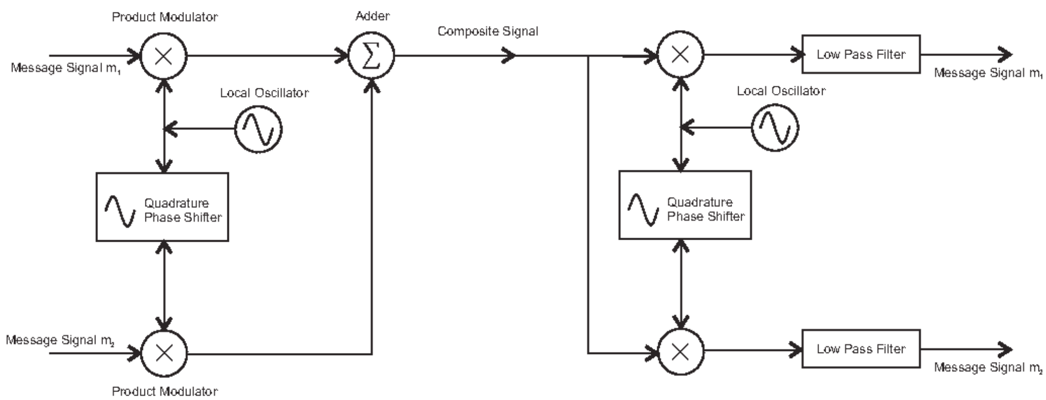
Orthogonal signals are those signals that are independent of each other i.e they have 90^o phase shift.
Example - $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$

Condition for two 2 real value signals $f_1(t)$ and $f_2(t)$ to be orthogonal is,

$$\int_{-\infty}^{\infty} f_1(t) f_2(t) dt = 0$$

If $m_1(t)$ and $m_2(t)$ are the two message signals, then the QAM signal will be $s(t) = m_1(t)\cos(2\pi f_c t) + m_2(t)\sin(2\pi f_c t)$

Usually, demodulation is done using coherent demodulation.



This technique of using orthogonal carriers of same frequency will reduce the bandwidth necessary to transmit signals.

The message signals can be referred to as $m_I(t)$ i.e the in-phase message and $-m_Q(t)$ i.e the quadrature-phase message.
$\implies s(t) = m_I(t)\cos(2\pi f_c t) - m_Q(t)\sin(2\pi f_c t)$

The complex pre-envelope of the band-pass QCM signal will be,

$$s_p(t) = [m_I(t) + m_Q(t)]e^{2\pi f_c t}$$

Now by shifting it to base-band (i.e centered around 0 frequency), the complex envelope is given by,

$$\tilde{s}(t) = s_p(t)\ e^{-2\pi f_c t} = m_I(t) + j\ m_Q(t)$$

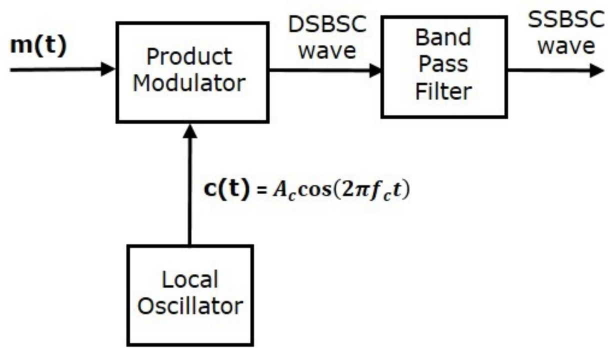
Hence, the base-band equivalent of the QCM signal is simply the in-phase and quadrature-phase messages that are phase-shifted by 90^o i.e are orthogonal.

2.1.4 Single Side Band, Suppressed Carrier

Since signal spectrum is always symmetric, there is no need to transmit both LSB and USB. The entire information in the message can be recovered by transmitting either one of the side bands, which will reduce the power consumption and bandwidth.

Frequency Discrimination Method
Ideally, SSB-SC signal can be generated by first generating DSB-SC signal

and then passing it through ideal band pass filter with appropriate cut-off frequencies to get either only USB or only LSB (and hence obtaining SSB-SC signal).



The disadvantage with this method is that ideal filters can't actually be designed. There will always be a transition band in practical filters which will allow some parts of unwanted frequency components to pass as well.

Phase Discrimination Method
An alternative, more practical way to generate SSB-SC signal uses the concept of Hilbert transform.

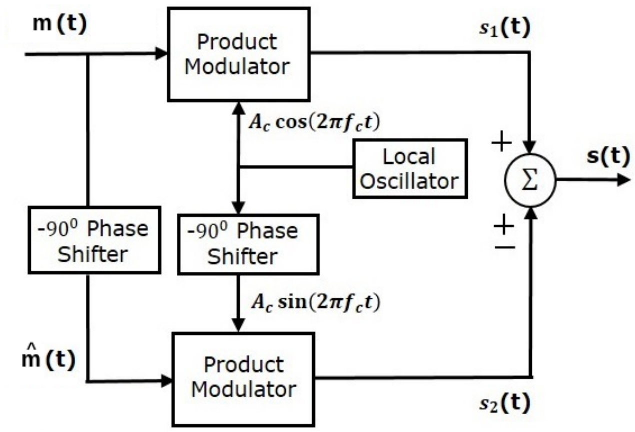
The SSB-SC signal in time domain will be as follows.
USB only transmission:

$$s(t) = \frac{A_c}{2}\left[m(t)\cos(2\pi f_c t) - \hat{m}(t)\sin(2\pi f_c t)\right]$$

LSB only transmission:

$$s(t) = \frac{A_c}{2}\left[m(t)\cos(2\pi f_c t) + \hat{m}(t)\sin(2\pi f_c t)\right]$$

The above equations indicate that either USB or LSB signal can be generated by adding or subtracting the message and its phase shifted version that are DSB-SC modulated using orthogonal carriers.



Demodulation of SSB-SC
Envelope detector can't be used since the output will be constant DC ($A_c/2$). Synchronous detector must be used.

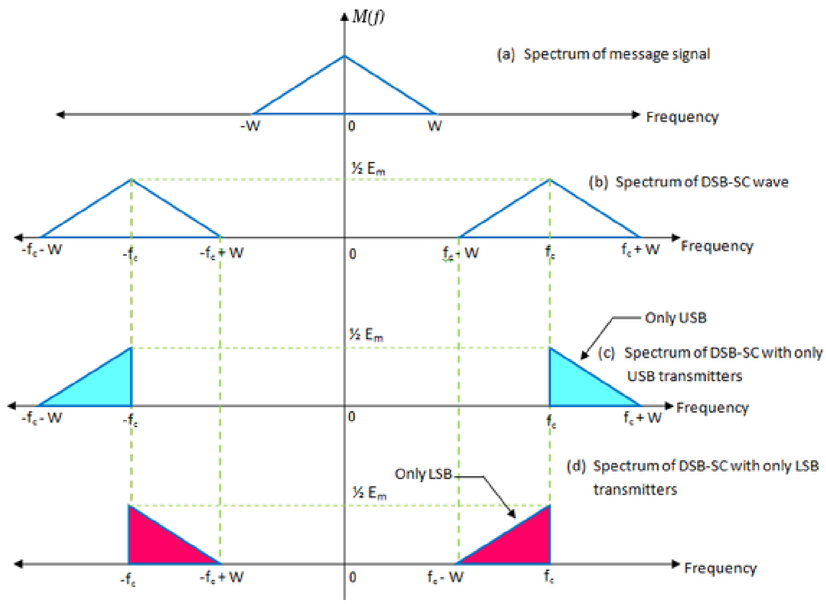
SSB-SC signal demodulated using carrier $A_c \cos(2\pi f_c t + \phi)$ and necessary BPF is given by,

$$m'(t) = \frac{A_c^2}{2}\ m(t)\ \cos\phi \pm \frac{A_c^2}{2}\ \hat{m}(t)\ \sin\phi$$

- If $\phi = 0$, then $m'(t) \propto m(t)$
- If $\phi = \pi/2$, then $m'(t) \propto \hat{m}(t)$

Hence, quadrature nulling will not occur in case of synchronous demodulation of SSB-SC signal.

The spectra of both DSB-SC and SSB-SC are illustrated.



2.2 Angle Modulation

For a high frequency carrier sinusoid of the form $c(t) = A_c \cos(2\pi f_c t + \phi)$, if its angle $\theta(t) = 2\pi f_c t + \phi$ is varied according to the message signal $m(t)$, then the carrier is said to be "angle modulated".

Note that there are two variable components in the angle, which are phase ϕ and frequency f_c. Meaning, there are broadly two types of angle modulation schemes.

2.2.1 Phase Modulation

In phase modulation, the phase ϕ of a high frequency carrier $c(t) = A_c \cos(2\pi f_c t + \phi)$ is varied according to message signal $m(t)$.

Assume the unmodulated carrier is $c(t) = A_c \cos(2\pi f_c t)$ where the angle $\theta(t) = 2\pi f_c t$.
In Phase modulation, the angle is modulated along with the message such that instantaneous phase varies proportionally with the message.
$\implies \theta_i(t) = 2\pi f_c t + k_p m(t)$.

Hence, the phase modulated signal will be,

$$x(t) = A_c \cos(\theta_i(t)) = A_c \cos(2\pi f_c t + k_p m(t))$$

The maximum phase deviation is obviously $\Delta\phi_{max} = k_p \text{max}[m(t)]$.

Instantaneous frequency is calculated as follows.

$$f_i(t) = \frac{1}{2\pi}\frac{d[\theta_i(t)]}{dt} = f_c + \frac{k_p}{2\pi}m'(t)$$

where $m'(t)$ is the derivative of the message signal $m(t)$

Hence, the maximum frequency deviation is given by,
$\Delta f_{max} = \text{max}(f_i(t) - f_c) = \frac{k_p}{2\pi}\text{max}[m'(t)]$.

For a single tone sinusoidal message signal, $m(t) = A_m \cos(2\pi f_m t)$, the phase modulated signal will be,

$$x(t) = A_c \cos(\theta_i(t)) = A_c \cos(2\pi f_c t + k_p A_m \cos(2\pi f_m t))$$

Hence, the modulation index is $\beta = k_p A_m$.

- $\Delta\phi_{max} = k_p A_m = \beta$
- $\Delta f_{max} = \frac{k_p A_m}{f_m} = \frac{\beta}{f_m}$

2.2.2 Frequency Modulation

Assume the unmodulated carrier is $c(t) = A_c \cos(2\pi f_c t)$ where the angle $\theta(t) = 2\pi f_c t$.
In Frequency modulation, the frequency is modulated along with the message such that instantaneous frequency varies proportionally with the message.
$\implies f_i(t) = f_c + k_f m(t)$.

The instantaneous angle is given by,

$$\theta_i(t) = 2\pi \int_0^t f_i(\tau)d\tau = 2\pi f_c t + 2\pi k_f \int_0^t m(\tau)d\tau$$

And hence, the frequency modulated signal will be,

$$x(t) = A_c \cos(\theta_i(t)) = A_c \cos(2\pi f_c t + 2\pi k_f \int_0^t m(\tau)d\tau)$$

The maximum frequency deviation is obviously $\Delta f_{max} = k_f \text{max}[m(t)]$.

The maximum phase deviation is given by $\Delta\phi_{max} = 2\pi k_f \text{max}[\int m(\tau)d\tau]$.

For a single tone sinusoidal message signal, $m(t) = A_m \cos(2\pi f_m t)$, the frequency modulated signal will be,

$$x(t) = A_c \cos(\theta_i(t)) = A_c \cos(2\pi f_c t + \frac{k_f A_m}{f_m} \sin(2\pi f_m t))$$

Hence, the modulation index is $\beta = \frac{k_f A_m}{f_m}$.

- $\Delta\phi_{max} = \frac{k_f A_m}{f_m} = \beta$
- $\Delta f_{max} = k_f A_m = \beta \ f_m$

If the modulating signal consists of both sin and cos terms such as,
$m(t) = A_1 \sin(2\pi f_m t) + A_2 \cos(2\pi f_m t)$, then the maximum value $m(t)$ can take is $|m(t)|_{max} = \sqrt{A_1^2 + A_2^2}$.
This result will be useful while calculating maximum phase and frequency deviations.

Narrow Band FM

Narrow band FM signal is a frequency modulated signal where the modulation index is significantly smaller than 1. $\beta << 1 \quad \implies \Delta f_{max} << f_m$.

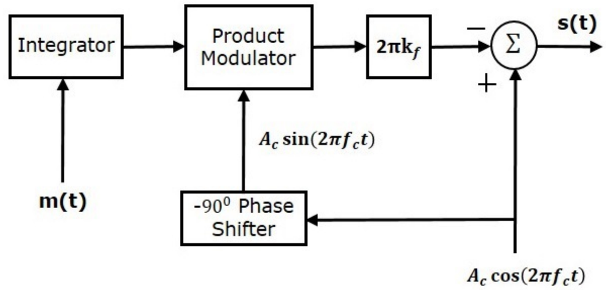
Consider an FM signal (with single tone sinusoidal message).

$$s(t) = A_c \cos(2\pi f_c t + \beta \sin(2\pi f_m t))$$

Upon expanding and substituting since $\beta << 1$, $\cos\beta \approx 1$ and $\sin\beta \approx \beta$, the equation for narrow band FM can be obtained as follows.

$$s(t) = A_c \cos(2\pi f_c t) - A_c \beta \sin(2\pi f_c t) \sin(2\pi f_m t)$$

Note that the narrow band FM signal is similar to standard AM signal and it can be generated using product modulator.

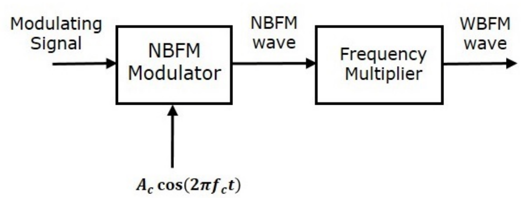


So if $\beta << 1$, the FM signal behaves somewhat similar to a standard AM signal, which is not very useful.

Wide Band FM

Wide band FM signal is a frequency modulated signal where the modulation index is a larger quantity. In general, $\beta > 0.5$ for it to be considered as a wide band FM signal.

Wide band FM signal is usually generated by passing a narrow band FM signal through a frequency multiplier.
A frequency multiplier consists of a non-linear device of order n followed by band-pass filter to remove unwanted frequency components. This is called "Indirect method".



If the instantaneous frequency of NBFM signal is $f_i(t) = f_c + k_f m(t)$, the output of frequency multiplier (of factor n) will be $f_o(t) = nf_i(t) = nf_c + nk_f m(t)$.
Hence, both the carrier frequency and the modulation index (β) are multiplied by a factor of n, which increases the value of β making it a WBFM signal.

Spectral analysis of FM signal

A frequency modulated signal can be expressed as,

$$s(t) = Re[A_c\ e^{j(2\pi f_c t + \beta \sin(2\pi f_m t))}] = Re[A_c\ e^{j\beta \sin(2\pi f_m t)}\ e^{j2\pi f_c t}]$$

From the above expression, it can be noted that $A_c e^{j\beta \sin(2\pi f_m t)}$ is the complex base-band equivalent of the FM signal $s(t)$.
Hence, the spectrum of FM signal can be found by first finding the spectrum of its complex envelope and then frequency translating it.

Let $\tilde{s}(t) = A_c e^{j\beta \sin(2\pi f_m t)}$. This signal is periodic with period $T_0 = 1/f_m$. So the spectrum will be given by complex Fourier series coefficients.

$$\tilde{s}(t) = \sum_{k=-\infty}^{\infty} C_k e^{j2\pi f_m k t}$$

$$C_k = \frac{1}{T_0} \int_0^{T_0} A_c e^{\beta \sin(2\pi f_m t)} e^{-j2\pi k f_m t} dt$$

Upon solving the above integral, it is found that the complex Fourier series coefficients can be represented using Bessel's Function of 1^{st} kind and k^{th} order.

$$C_k = A_c\ \frac{1}{2\pi} \int_0^{2\pi} e^{\beta \sin x - kx} dx = A_c J_k(\beta)$$

$$\implies \tilde{s}(t) = \sum_{k=-\infty}^{\infty} A_c J_k(\beta) e^{j2\pi f_m k t}$$

Now the expression for original FM signal is obtained as,

$$s(t) = Re\Big[\sum_{k=-\infty}^{\infty} A_c J_k(\beta) e^{j2\pi f_m k t}\ e^{2\pi f_c t}\Big] = A_c \sum_{k=-\infty}^{\infty} \big[J_k(\beta) \cos(2\pi(f_c + k f_m)t)\big]$$

Finally, the spectrum of FM signal is obtained by using Fourier transform.

$$S(f) = \frac{A_c}{2} \sum_{k=-\infty}^{\infty} \Big[J_k(\beta)\big[\delta(f - f_c - k f_m) + \delta(f + f_c + k f_m)\big]\Big]$$

Properties of Bessel's Function:

- $J_n(x) = (-1)^n J_{-n}(x)$
- $\sum J_n^2(x) = 1$

Magnitude of carrier in the FM signal depends on the value of β. For $\beta = 2.4, 5.5, 8.6, 11.8, ...$, the carrier magnitude will turn out to be zero and hence, 100% modulation efficiency is obtained.

Power of FM signal is same as the power of the carrier. $\implies P_t = P_c = \frac{A_c^2}{2}$.

From the spectrum, it can be observed that the bandwidth of the FM signal is infinite since k ranges from $-\infty$ to ∞. However, practically the Bessel function values decay rapidly with increasing k and hence most of the power is contained in the first few frequency components.

The method to calculate practical bandwidth of an FM signal is given by **Carson's Rule**.

$$BW = 2\left(1 + \frac{1}{\beta}\right)\Delta f_{max} = 2(1+\beta)f_m$$

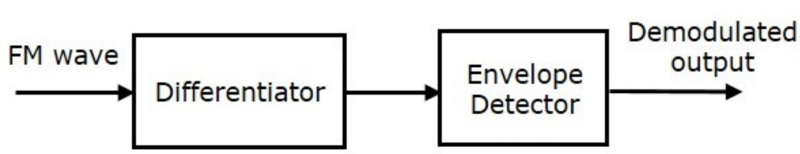
Note that $\beta = \frac{\Delta f_{max}}{f_m}$

- If $\beta >> 1$, then $BW \approx 2\Delta f_{max}$
- If $\beta << 1$, then $BW \approx 2f_m$

It is observed that approximately 98% of the power of an FM signal is contained within the bandwidth calculated using Carson's rule.

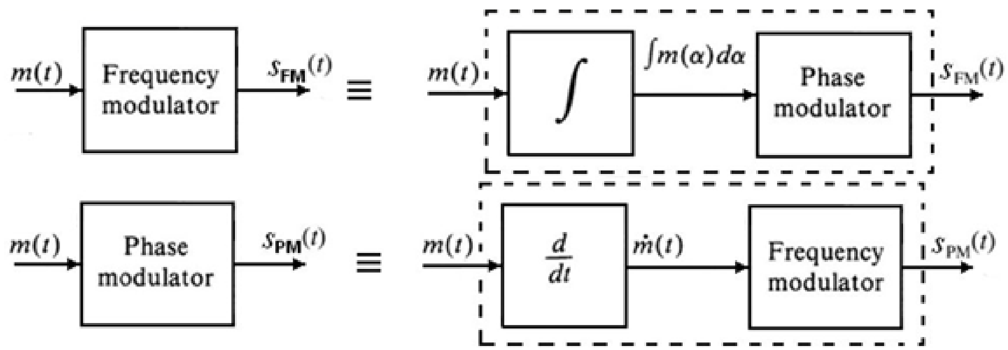
Demodulation of FM signal

FM signal is demodulated by first passing it through a differentiator which will convert it to a signal in standard AM form, and then using envelope detector.



Note that the condition for envelope detector to work without distortion is $f_c > \Delta f_{max}$, which will pretty much always be naturally satisfied.

Relation between PM and FM
From the equations of PM and FM signals, it is obvious that one can be realized by modifying the other as follows.



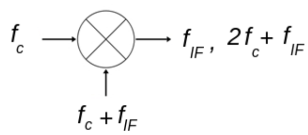
2.3 Frequency Translation

In signal processing and communication systems, it will be necessary to shift a particular signal centred around frequency f_c to a different centre frequency f_{IF}. This process is called frequency translation or "Heterodyning".
The new frequency at which the original signal is shifted to is usually referred to as "intermediate frequency".

For example, all AM radio signals that are in the range of 540 kHz to 1600 kHz are shifted around the intermediate frequency of 455 kHz.

Frequency translation is done by using a mixer and a band pass filter. A mixer is simply a product modulator which multiplies the modulated signal with another sinusoid of different frequency.

If the modulated signal $s(t) = A_c\ m(t)\ \cos(2\pi f_c t)$ is multiplied with signal generated by local oscillator $l(t) = \cos(2\pi(f_c + f_{IF})t)$, then the output will consist of, $s_m(t) = 0.5A_c\ m(t)\ [\cos(2\pi f_{IF} t) + \cos(2\pi(2f_c + f_{IF})t)$.



The signal $s_m(t)$ can be passed through BPF to obtain $s_t(t) = 0.5A_c\, m(t)\ \cos(2\pi f_{IF} t)$ which is the same modulated signal centered around f_{IF} rather than f_c.
(note that $l(t) = \cos(2\pi(f_c - f_{IF})t)$ can also be used for the same purpose)

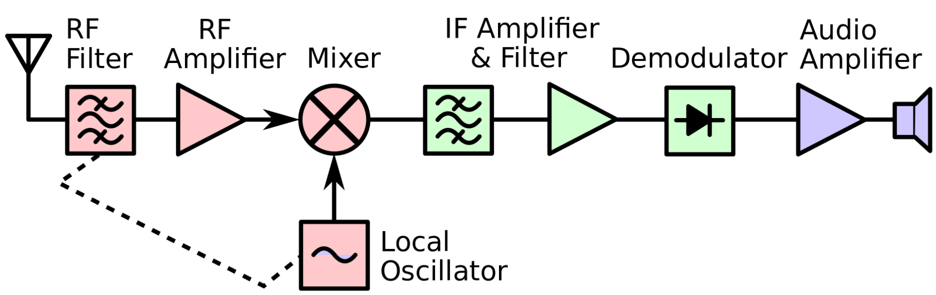
Note that if there is another modulated signal being transmitted, say $r(t) = A_c\ m(t)\ \cos(2\pi(f_c + 2f_{IF})t)$, then after frequency translation, $r_m(t) = 0.5A_c\ m(t)\ [\cos(2\pi f_{IF} t) + \cos(2\pi(2f_c + 3f_{IF})t)$ is obtained which upon filtering will provide $r_t(t) = 0.5A_c\ m(t)\ \cos(2\pi f_{IF} t)$.

Note that in this case, $s(t) = r_t(t)$. Meaning, modulated signals of two different frequencies can be translated to the same intermediate frequency, which will cause interference between the two. This effect is not desirable.
For a modulated signal of frequency f_c, the frequencies of $f_c \pm 2f_{IF}$ can cause interference. These are called ”image frequencies”.

In order to properly receive the required message signal without interference from image frequencies, the image frequency components have to be attenuated by a pre-filter.

2.3.1 Super-heterodyne Receiver

The super-heterodyne receiver is a system used to receive the modulated signal using the concepts discussed above.



As mentioned, to suppress image frequency components before mixing with a local oscillator, a filter is used (tuned RF filter).

The image frequency rejection ratio of the tuned filter is given by,

$$\alpha = \sqrt{1 + P^2Q^2}; \qquad P = \frac{f_{img}}{f_c} - \frac{f_c}{f_{img}}$$

where Q is the quality factor of the antenna coupling circuit

The local oscillator frequency is adjusted using capacitors.
The capacitance ratio is given by,

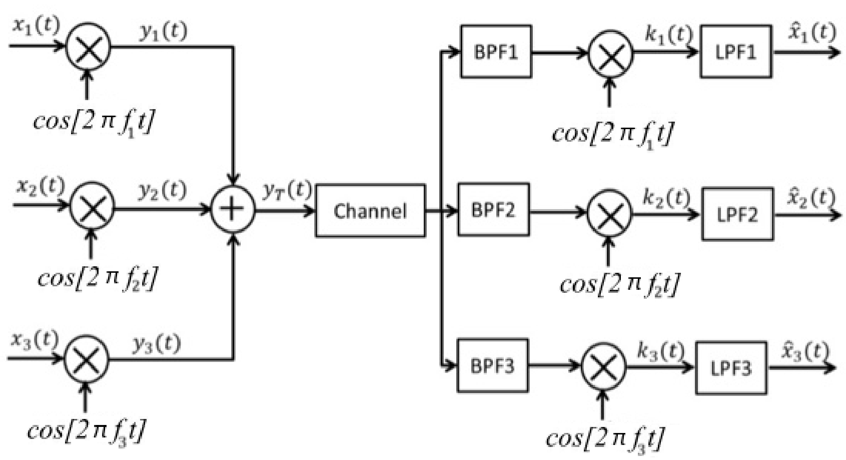
$$R = \frac{C_{max}}{C_{min}} = \left(\frac{f_{max}}{f_{min}}\right)^2$$

where $f_{max} = f_{c_{max}} + f_{IF}$ and $f_{min} = f_{c_{min}} + f_{IF}$.

Multiple such filters can be cascaded to obtain overall rejection ratio as product of individual ratios.

2.4 Frequency Division Multiplexing

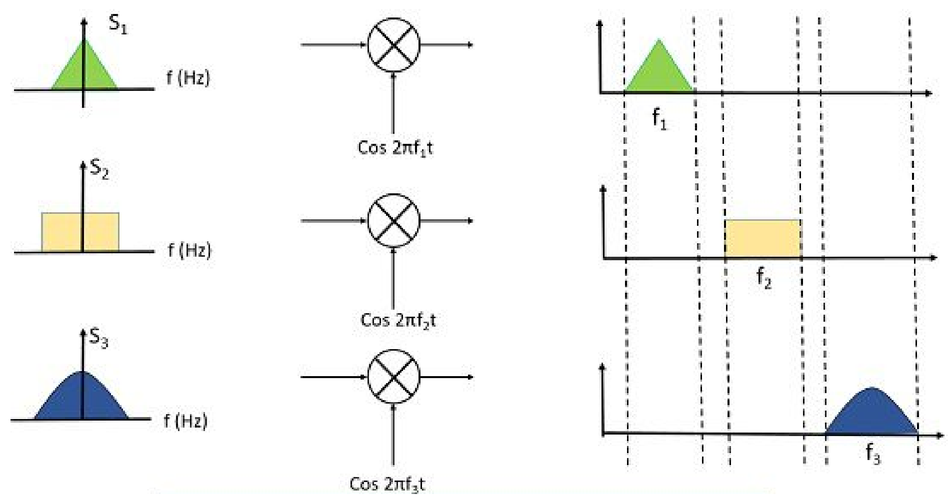
Frequency Division Multiplexing (FDM) is a process of combining multiple signals centred at different carrier frequencies and transmitting them through the same channel.



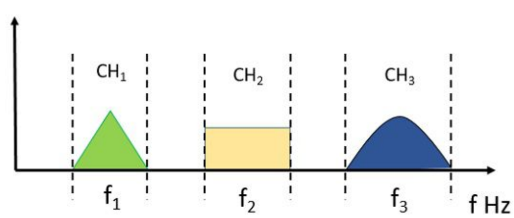
If $x_1(t), x_2(t)$ and $x_3(t)$ are band-limited to frequency W, they are translated to different frequencies using different carriers as shown.
Hence, $y_1(t), y_2(t)$ and $y_3(t)$ will be centred at f_1, f_2 and f_3 respectively. These signals are added and transmitted.

At the receiving end, band pass filters of relevant cut-off frequencies are used to receive only the required signal (for example BPF of cut-off frequencies $f_1 - W$ and $f_1 + W$ for signal $y_1(t)$ and so on), which are then demodulated to recover the original signals.

The figure below illustrates how base-band signals are frequency translated to different center frequencies.

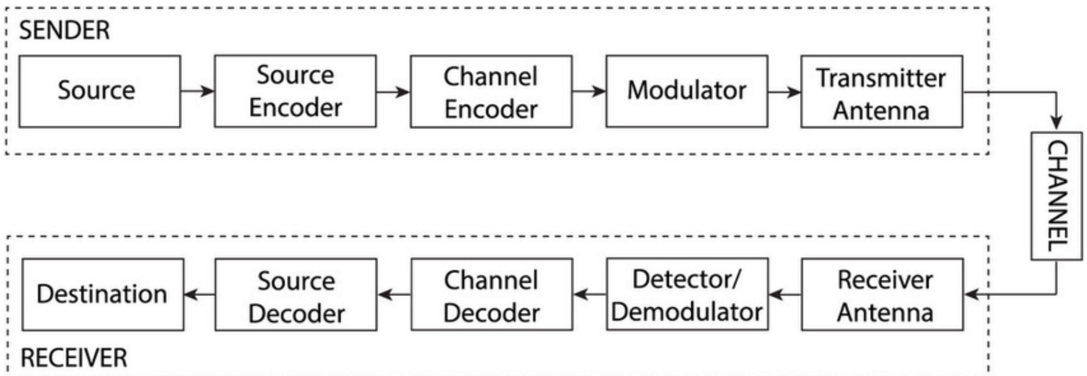


The translated signals are then added to obtain the FDM signal as shown.



3 Digital Communication

Noise immunity of digital signals is significantly higher than that of analog signals. Also, designing systems that deal with digital systems is more memory efficient and more structured when compared to analog systems.
Hence, digital communication schemes are more widely used. The block diagram shown illustrates a general digital communication system.



3.1 Basics of Digital Communication Systems

Before going into how digital signals are modulated, transmitted and received, some basic concepts regarding digital pulses and digital channel need to be covered.

3.1.1 Digital Pulses

Digitally modulated symbols are transmitted using pulses. A generic pulse will have amplitude 1 for a duration T and is denoted by $p(t)$.
A digital signal will be a combination of several shifted and scaled pulses. Hence the structure of a typical digitally transmitted signal is,

$$X(t) = \sum_{k=-\infty}^{\infty} a_k \, p(t - kT)$$

The two most generic coding schemes are polar and unipolar.

$$\text{Polar} \implies a_k = \begin{cases} +A : 1 \\ -A : 0 \end{cases} \qquad \text{Unipolar} \implies a_k = \begin{cases} +A : 1 \\ \;\;0 : 0 \end{cases}$$

Since a_k is a discrete random variable, $x(t)$ is a random process. Assuming "0" and "1" are equally likely to occur, it can be calculated that the expected value of a_k, and hence $X(t)$ as well as $X(f)$ are all 0.

Hence, the relevant way to measure the spectral components of a random process is by using Power Spectral Density (for which Auto-correlation function needs to be computed).

A more general form of a binary random sequence is,

$$X(t) \sum_{k=-\infty}^{\infty} a_k \; p(t - kT + T_d)$$

where T_d is the starting delay due to absence of time synchronization. The value of T_d is not deterministic, it can take any value between 0 to T with equal probability.

The auto-correlation function (ACF) of $X(t)$ is derived to be,

$$R_x(\tau) = \frac{1}{T} \sum_{-\infty}^{\infty} R_a(n) r_p(\tau + nT)$$

where $R_a(n)$ is the ACF of the discrete random variable a_k and $r_P(\tau)$ is the ACF of the deterministic pulse signal $p(t)$.

The above expression can be simplified for different schemes.

For polar coding, it can be found that

$$R_a(n) = \begin{cases} A^2 : n = 0 \\ 0 \quad : n \neq 0 \end{cases}$$

Hence, the ACF expression reduces to:

$$R_x(\tau) = E[x(t)x(t + \tau)] = \frac{A^2}{T} \; R_p(\tau)$$

And the power spectral density is given by,

$$S_x(f) = \frac{A^2}{T} \; |P(f)|^2$$

The energy spectral density of the pulse is, $|P(f)|^2 = T^2 \text{ sinc}^2(fT)$.
Hence, the final expression for PSD of the digital signal is,

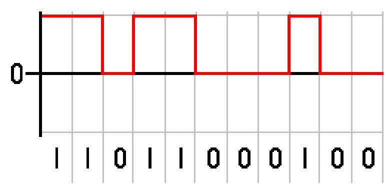
$$S_x(f) = A^2 T \text{ sinc}^2(fT)$$

Note that the polar scheme is the most common and simplest scheme for random binary pulse transmission. More complicated schemes are illustrated in the following section.

3.1.2 Line Codes

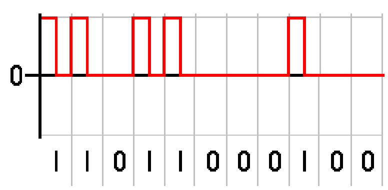
A line code is a pattern of voltage, current, or photons used to represent digital data.
There are 4 types of line coding, which are further sub-divided as follows.

- Unipolar:
  - Non Return to Zero (U-NRZ)



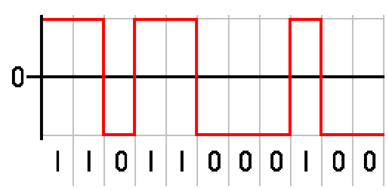
$BW = R_b$

  - Return to Zero (U-RZ)



$BW = 2R_b$

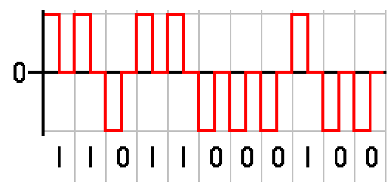
- Polar:
  - Non Return to Zero (P-NRZ)



$BW = R_b$

Polar Non Return to Zero is the line code which was considered earlier for the derivation of ACF and PSD of pulse of a random binary sequence.

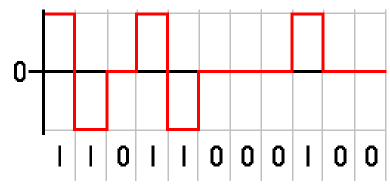
– Return to Zero (P-RZ)



$BW = 2R_b$

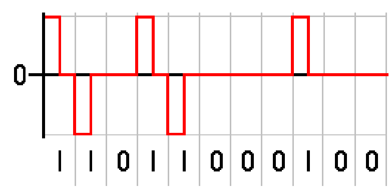
- Bipolar:

  – Non Return to Zero (B-NRZ)



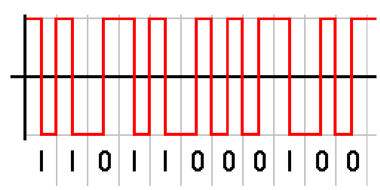
$BW = R_b$

  – Return to Zero (B-RZ)



$BW = R_b$

- Manchester:



$BW = 2R_b$

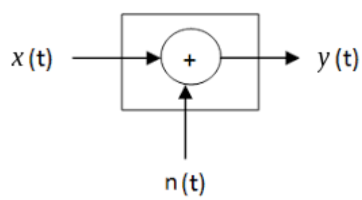
3.1.3 Digital Channel

The most commonly used digital channel is called **AWGN Channel** i.e the Additive White Gaussian Noise Channel.

If a digital signal $x(t)$ is transmitted through an AWGN channel, then the received signal $y(t)$ is given by,

$$y(t) = x(t) + n(t)$$

Here, $n(t)$ represents the noise signal.



In an AWGN channel, the noise is characterized by the following properties.

- $n(t)$ is additive in nature i.e it adds to the input signal.
- $n(t)$ is a Gaussian random process.
- $n(t)$ adds equally to all frequencies of the input.

The auto-correlation of $n(t)$ is given by,

$$R_n(\tau) = \frac{N_0}{2}\delta(\tau)$$

The above equation indicates that the correlation exists only at one instant and hence, the noise signal is highly uncorrelated.

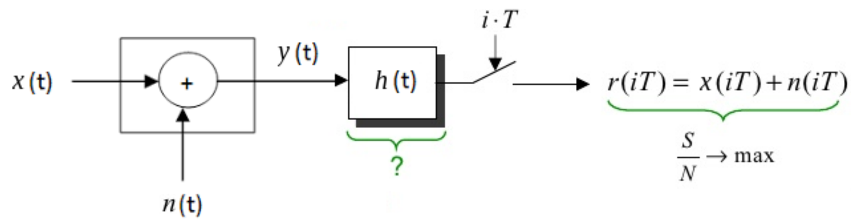
Taking Fourier transform, the power spectral density of $n(t)$ is found.

$$S_n(f) = \frac{N_0}{2} \quad \forall\ f$$

Since the PSD is flat, noise adds equally to all frequencies.

3.1.4 Digital Receivers

The transmitted signal, after being sent through a channel (AWGN) has to be received such that the noise is minimized and signal component is maximized.



As shown above, received signal $y(t)$ is passed through a filter (which is an LTI system), and then sampled at the duration of pulse to obtain the information.
The filter $h(t)$ is to be designed such that the Signal to Noise ratio (SNR) is maximum.

By calculating the signal power and the noise power, the following inequality can be obtained.

$$\frac{S}{N} \leq \frac{\int_{-\infty}^{\infty} |X(f)|^2 df}{N_0/2}$$

The numerator term is simply the energy of the input signal (E) and the denominator term is obtained by the PSD of the Gaussian Noise.

$$\implies \left.\frac{S}{N}\right|_{max} = \frac{2E_{max}}{N_0}$$

The energy of the input signal will be maximum if the transfer function of the filter satisfies,

$$H(f) = \bar{X}(f)\ e^{-j2\pi fT}$$

Therefore, the impulse response of the filter must be

$$\boxed{h(t) = \bar{x}(T-t)}$$

For real signals, $h(t) = x(T-t)$.

Such a filter is called **Matched Filter** and the receiver which uses this is called "Matched Filter Receiver".

3.1.5 Probability of Error

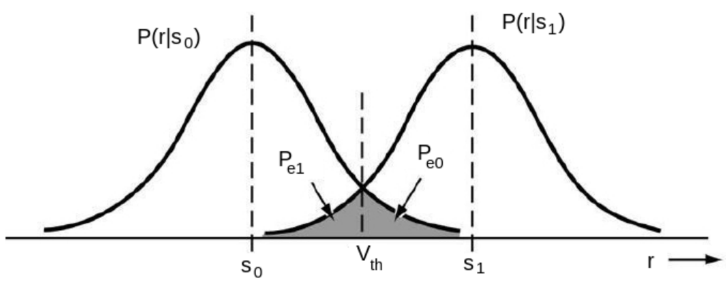
Due to the addition of noise, there is always the chance for error, even if the signal to noise ratio is maximized.

The output of the matched filter is sampled at every T duration. Since the input signal is a random process, the output will be a random variable.

For a binary sample space $S = [0, 1]$, let s_0 correspond to "0" and s_1 correspond to "1". $n_0(t)$ represents that noise component in the output of the filter.

If "0" is transmitted, then the output random variable will be $r = s_0 + n_0(T)$.
If "1" is transmitted, then the output random variable will be $r = s_1 + n_0(T)$.

Hence, depending on the channel and the noise model, the probability density function of the output random variable is obtained.
If the noised model is Gaussian (as it will be for an AWGN channel), then the PDFs of the output random variables will be as illustrated.



$P(r|s_0)$ is the PDF of the received symbol if 0 was transmitted and $P(r|s_1)$ is the PDF of the received symbol of 1 was transmitted.

$V_{th} = \frac{s_1 - s_0}{2}$ is the threshold value which makes the decision in the output, i.e if the received value is greater than V_{th}, it is assumed that 1 was sent and if the received value is lesser than V_{th}, it is assumed that 0 was sent.

If the two PDFs coincide/overlap, it means there is a chance for error. This probability is found by integrating the PDFs over the region of the graph beyond V_{th}.

Note that if the noise that adds to both the symbols equally (as does AWGN), then the probability of errors for both 1 and 0 will be the same.

For a system with AWGN Channel which uses Matched Filter Receiver, the probability of error is given by the complementary error function or the Q-function:

$$\boxed{P_e = Q\left(\sqrt{\frac{E_d}{2N_0}}\right)}$$

E_d is the difference of signal energy i.e

$$E_d = \int_0^T (s_1(t) - s_0(t))^2 dt = \int_0^T s_1^2(t)\ dt + \int_0^T s_0^2(t)\ dt - 2\int_0^T s_1(t)\ s_0(t)\ dt$$

$$\implies \boxed{P_e = Q\left(\sqrt{\frac{E_{s1} + E_{s0} - 2E_{s1s0}}{2N_0}}\right)}$$

The Q-function expression is:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{z=x}^{\infty} e^{-z^2/2}\ dz$$

Using the given expression, probability of error for each of the modulation schemes can be found (which will be introduced soon).

In case of there is a phase shift of ϕ between the modulation and demodulation carriers, then the bit energy at the receiver end will be multiplied by $\cos^2(\phi)$, and so will the parameter inside the Q-function in probability of error.

In case of polar signaling, since $s_1 = -s_0$, the energies are given by $E_{s1} = E_{s0} = E_s$ and $E_{s1s0} = -E_s$.
Average energy per bit, $E_b = 0.5E_p + 0.5E_p = E_p$.

$$\implies P_e = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$$

For any orthogonal binary signaling (such as on-off), $E_{s1s0} = 0$.
Average energy per bit, $E_b = 0.5(E_p + E_q)$.

$$\implies P_e = Q\left(\sqrt{\frac{E_b}{N_0}}\right)$$

The above results indicate that any orthogonal binary signaling is inferior to polar signaling by $3dB$.

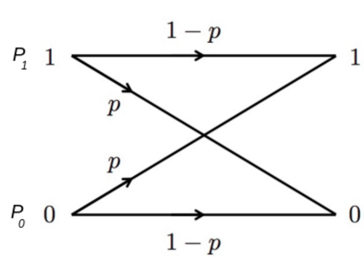
It is known that using orthogonal pulses will save bandwidth. Hence, there is a trade-off between probability of error and bandwidth required for transmission between the two different schemes.

- Polar signaling gives lower probability of error and power consumption
- Orthogonal signaling gives lower bandwidth consumption

3.1.6 Maximum Aposteriori Probability (MAP) Rule

Consider a binary sample space $S = [0, 1]$ where s_0 corresponds to "0" and s_1 corresponds to "1".
Note that s_0 and s_1 are mutually exclusive and exhaustive.

Probability of generating s_1 is given by $P(s_1) = P_1$ and probability of generating s_0 is given by $P(s_0) = P_0$. Hence, P_1 and P_2 are the prior probabilities. Assume the channel is binary symmetric channel as illustrated as below with p being probability of error or probability of flipping.



If r is the event corresponding to receiving "1" and $\tilde{r}$ is the event corresponding to receiving "0", then the likelihoods are given by,

$$P(r|s_0) = P(\tilde{r}|s_1) = 1 - p \qquad\qquad P(r|s_1) = P(\tilde{r}|s_0) = p$$

Using Bayes theorem, the aposteriori probabilities of the transmitted symbol given the received symbol can be calculated.

Consider the aposteriori probabilities given that 1 was received i.e r has occurred.

$$P(s_0|r) = \frac{P(s_0)P(r|s_0)}{P(s_0)P(r|s_0) + P(s_1)P(r|s_1)} = \frac{P_0\ (1-p)}{P_0\ (1-p) + P_1\ p}$$

$$P(s_1|r) = \frac{P(s_1)P(r|s_1)}{P(s_0)P(r|s_0) + P(s_1)P(r|s_1)} = \frac{P_1\ p}{P_0\ (1-p) + P_1\ p}$$

MAP rule states that, choose the symbol with maximum aposteriori probability. Meaning in the above case, if $P(s_0|r) > P(s_1|r)$, then choose s_0 and if $P(s_1|r) > P(s_0|r)$, then choose s_1.

Similarly, consider the aposteriori probabilities given that 0 was received i.e $\tilde{r}$ has occurred.

$$P(s_0|\tilde{r}) = \frac{P(s_0)P(\tilde{r}|s_0)}{P(s_0)P(\tilde{r}|s_0) + P(s_1)P(\tilde{r}|s_1)} = \frac{P_0\ p}{P_0\ p + P_1\ (1-p)}$$

$$P(s_0|\tilde{r}) = \frac{P(s_1)P(\tilde{r}|s_1)}{P(s_0)P(\tilde{r}|s_0) + P(s_1)P(\tilde{r}|s_1)} = \frac{P_1\ (1-p)}{P_0\ p + P_1\ (1-p)}$$

Here if $P(s_0|\tilde{r}) > P(s_1|\tilde{r})$, then choose s_0 and if $P(s_1|\tilde{r}) > P(s_0|\tilde{r})$, then choose s_1.

Therefore if a receiver operates on MAP rule, the decision is taken by comparing the aposteriori probabilities.

The **optimal receiver** is the one that operates on MAP Rule since it minimizes the probability of error.

3.1.7 Maximum Likelyhood (ML) Rule

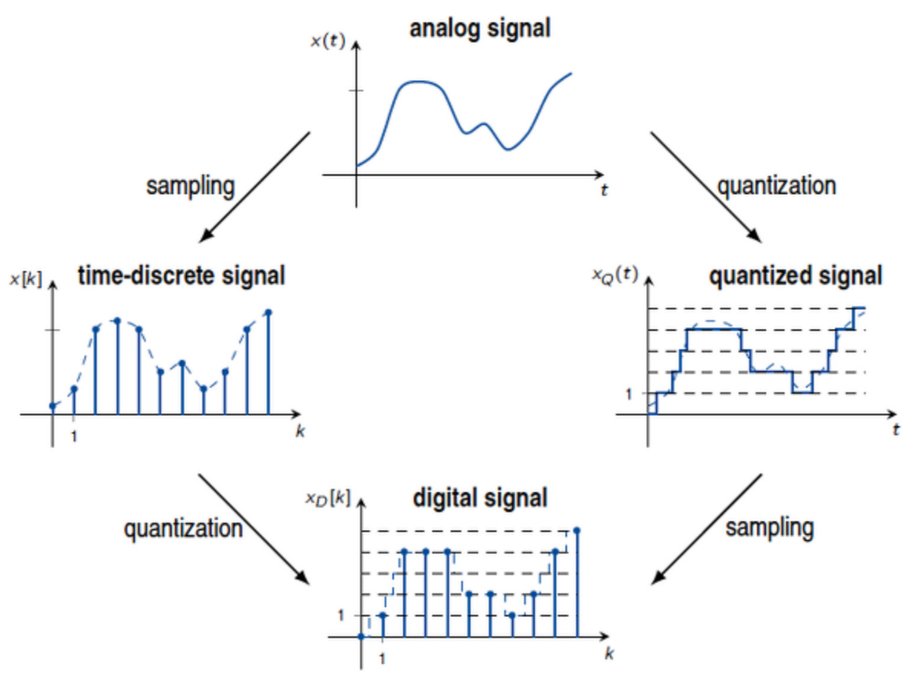
If the decision is taken based on likelihoods rather than aposteriori probabilities, then the receiver is said to operate on maximum likelihood rule.
ML rule states that, choose the symbol with maximum likelihood.

Note that the MAP rule reduces to ML rule if the symbols are equiprobable i.e if $P(s_1) = P(s_0) = P$.
Therefore, if the probability of generation of 0 and 1 are the same (0.5 since they are mutually exclusive and exhaustive), then the decision taken by ML rule and MAP rule will be exactly the same.

- Received symbol is r:
  - $P(r|s_0) > P(r|s_1)$ or $1 - p > p$; choose s_0
  - $P(r|s_1) > P(r|s_0)$ or $p > 1 - p$; choose s_1
- Received symbol is $\tilde{r}$:
  - $P(\tilde{r}|s_0) > P(\tilde{r}|s_1)$ or $p > 1 - p$; choose s_0
  - $P(\tilde{r}|s_1) > P(\tilde{r}|s_0)$ or $1 - p > p$; choose s_1

3.2 Base Band Modulation

Message signals that occur naturally are analog in nature. In a digital communication system, the analog message signals need to be digitized using some modulation techniques before being transmitted.



3.2.1 Sampling

Sampling is the process of converting an analog signal to discrete signal.

The process of sampling has been discussed from time domain point of view in Signals. Here, sampling from the frequency domain perspective is elaborated in detail.

Sampling period, T_s is the time duration at which samples of the continuous signal are taken. Sampling frequency is given by $f_s = 1/T_s$.

Let $x(t)$ be a continuous time signal and $X(f)$ be its Fourier Transform.
A band-limited signal is a signal which is zero outside a specific range of frequencies.

- Low pass signal : $X(f) = 0 : f > |W|$
- Band pass signal : $X(f) = 0 : |W_1| < f < |W_2|$

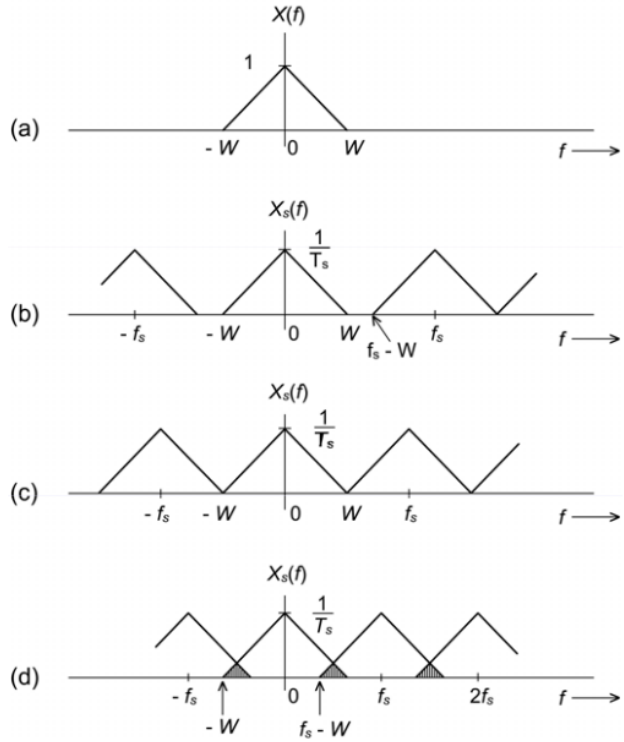
Consider $x(t)$ to be low pass signal for now.
Nyquist Rate is the sampling rate/frequency which is equal to twice the maximum frequency of the signal. $\implies f_s = 2W$.

The figure shown below represent the original spectrum (a) and the spectra obtained after sampling above Nyquist rate (b), at Nyquist rate (c) and below Nyquist rate (d).
It can be observed that there will be loss of higher frequency components if sampled at a rate lower than the Nyquist rate. This phenomenon is called "Aliasing".

In the other two cases, there is no loss of any information as the original spectrum has just repeated itself. And hence, it can be recovered by simply using a filter with cut-off frequency W and gain T_s.

The expression for spectrum of the sampled signal is obtained as,

$$X_s(f) = X(f) * f_s \sum_{k=-\infty}^{\infty} \delta(f - kf_s) = f_s \sum_{k=-\infty}^{\infty} X(f - kf_s)$$



The transfer function of the required ideal filter is given by,

$$H(f) = \begin{cases} T_s & : f \leq W \\ 0 & : f > W \end{cases}$$

Obviously this is a rectangular function, and its inverse Fourier transform will give the impulse response of the required filter.

$$\implies h(t) = \text{sinc}(f_s t) \quad \left(= \frac{\sin(\pi f_s t)}{\pi f_s t} \right)$$

The recovered signal $x_r(t)$ is given by,

$$x_r(t) = x_s(t) * h(t) = \sum_{k=-\infty}^{\infty} x(kT_s)\delta(t - kT_s) * \text{sinc}(f_s t) = \sum_{k=-\infty}^{\infty} x(kT_s)\,\text{sinc}(f_s(t - kT_s))$$

Practically, sampling through impulses can't be done since impulse function does not actually exist. Hence, flat top sampling is generally employed.

Here, a train of periodic pulses of duration T and period T_s is multiplied with the message signal. This is generally done using a "sample and hold circuit".

This process is also called "Pulse Amplitude Modulation" and the signal obtained is the PAM signal.

$$x_p(t) = \sum_{k=-\infty}^{\infty} x(kT_s)p(t - kT_s)$$

where $p(t)$ is a simple pulse signal given by,

$$p(t) = \begin{cases} 1 & : 0 < t < T \\ 0 & : T < t < T_s \end{cases}$$

3.2.2 Quantization

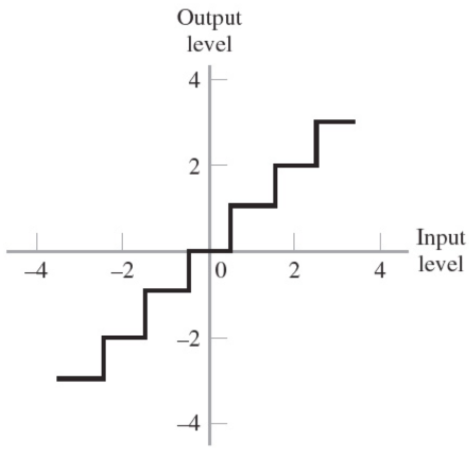
Quantization is the process of converting a discrete signal to quantized signal.

Meaning, a few discrete levels are defined and the amplitudes of the discrete samples are mapped to the nearest levels, so that the discrete samples can take only finite number of predefined values which can be encoded in a finite number of bits.

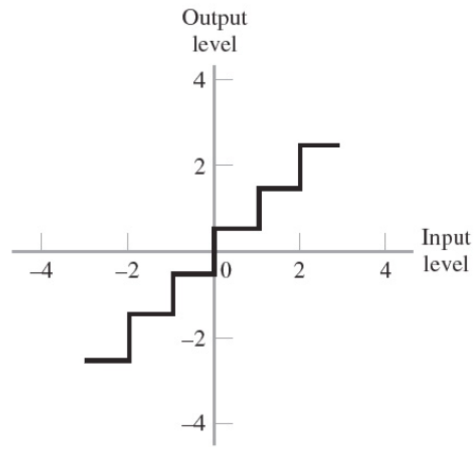
Uniform Quantization is the simplest form of quantization where the difference between two consecutive levels is constant and is denoted by Δ.
Quantization levels are the midpoints of the intervals. Hence, a sample value which lies in a particular interval is mapped to its midpoint.

Mid-Tread Quantization:
The number of quantization levels are odd. Hence, the midpoint of the input-output characteristics will be horizontal.

Mid-Rise Quantization:
The number of quantization levels are even. Hence, the midpoint of the input-output characteristics will be vertical.



Mid-Tread Quantizer



Mid-Rise Quantizer

The dynamic range of a discrete signal is the range of values/amplitudes the signal can take.
Hence, the dynamic range is given by $[-V_{min}, V_{max}]$.
Usually this is assumed to be symmetric, so in that case, the range will be $[-V_m, V_m]$ and the length of the dynamic range is $V_m - (-V_m) = 2V_m$.

Hence, the number of quantization levels is given by,

$$L = \frac{V_{max} - (-V_{min})}{\Delta} = \frac{2V_m}{\Delta}.$$

Since quantization is a process of approximating the actual amplitude to the nearest fixed amplitude level, there is bound to be loss of information or error.

Quantization error/noise, Q_e is the difference between the actual value of a sample and the quantized value.
For a uniform quantizer, the maximum value that Q_e can take is $\Delta/2$.

Note that all quantization need not be uniform. Quantization can be customized to have any number of levels, any step size and any number in each step can be the mapped value (instead of midpoint).

3.2.3 Pulse Code Modulation

The process of converting a quantized signal to encoded digital signal is called "Pulse Code Modulation".

If L is the number of quantization levels and n is the number of bits necessary to encode the quantized signal (also termed as the "resolution"), n can be found using the relation $L = 2^n$ (where n is an integer).

This means n bits can be used to encode upto $L = 2^n$ levels and hence, n bits can naturally be used to encode any number of levels less than 2^n, so it is not necessary that all combinations of bits have to be used.

The Quantization noise can be considered as a uniform distribution in the range $(-\Delta/2, \Delta/2)$.
Hence, Quantization noise power (Q) is given by expectation of square of the uniform distribution. $\implies Q = \Delta^2/12$

The signal power can be calculated based on the nature of input signal.
If the input signal is sinusoidal, then the signal power (P) is given by $A_m^2/2$ where A_m is peak amplitude. In this case, the dynamic range of the quantizer can be approximated as $2A_m$.
$\implies P = A_m^2/2$ and $\Delta = 2A_m/2^n$.

The Signal to Noise Ratio (SNR) is the ratio of signal power to quantization noise power. $SNR = P/Q \implies SNR = 6A_m^2/\Delta^2 = 3(2^{2n-1})$

Expressing the same in dB,

$$\boxed{SNR = 1.8 + 6n \;\; dB}$$

Note that this will vary if the input signal is not sinusoidal.

Another important observation is that the signal to noise ratio increases by 6 dB for 1 bit increase in the resolution.

Since there are n bits for every period T_s, the duration of each bit is given by $T_b = T_s/n$.
Hence, the bit rate is $R_b = 1/T_b$.

Maximum bandwidth required is equal to the bit rate.
Minimum bandwidth required is equal to half the bit rate.

$\Longrightarrow$ Increasing quantization levels for more accuracy will reduce error but will increase bandwidth.

Note that the above expressions and methods for calculating SNR are valid if the input signal is deterministic and uniform quantization is used.

Random Signal and Arbitrary Quantization Scheme:

Consider the case where input signal is a random signal $X(t)$ which follows some distribution $f_X(x)$.
The signal power is calculated using:

$$S = E[X^2(t)] = \int_{-\infty}^{\infty} X^2(t) f_X(x) dx$$

Consider the case where quantization is non-uniform, with mapped values given by x_q for different intervals. The quantization error in general will be $X(t) - x_q$.
The quantization noise is calculated using:

$$N = E[(X(t) - x_q)^2] = \int_{-\infty}^{\infty} (X(t) - x_q)^2 f_X(x) dx$$

For illustration, assume that 3 quantization intervals are defined between intervals l_1 to l_2, l_2 to l_3 and l_3 to l_4 and these intervals are mapped to levels x_{q1}, x_{q2}, x_{q3} respectively.

$$N = \int_{l_1}^{l^2} (X(t) - x_{q1})^2 f_X(x) dx + \int_{l_2}^{l^3} (X(t) - x_{q2})^2 f_X(x) dx + \int_{l_3}^{l^4} (X(t) - x_{q3})^2 f_X(x) dx$$

This method works for uniform quantization as well, so it is the general procedure to find the quantization noise given any quantization scheme.

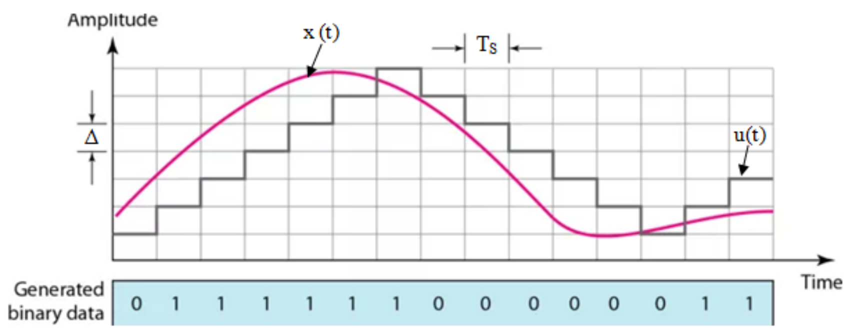
Differential Pulse Code Modulation

DPCM is a modified version of PCM where instead of directly encoding the quantized samples, the difference between consecutive samples is quantized. This can potentially decrease the dynamic range, which will decrease the quantization error without changing the bandwidth.

3.2.4 Delta Modulation

A modified version of DPCM where only 2 quantization levels exist. Hence, it is also called 1-bit DPCM.

- If difference between current and previous samples is positive, then Δ is added.
- If difference between current and previous samples is negative, then $-\Delta$ is added.



Since $n = 1$, $T_b = T_s$ and $R_b = 1/T_s = f_s$.
Therefore, the bandwidth has decreased from previous schemes.

Delta Modulation works very well if the correlation between samples is high i.e the samples are not changing rapidly.

There are however, 2 problems that are encountered in Delta modulation.

Granular Noise Distortion occurs if Δ value is too large and hence slight variations in the output is causing unnecessary ups and downs in modulated signal.

Slope Overload Distortion occurs if Δ value is too small and hence the modulator will be unable to track rapid changes in the original signal.

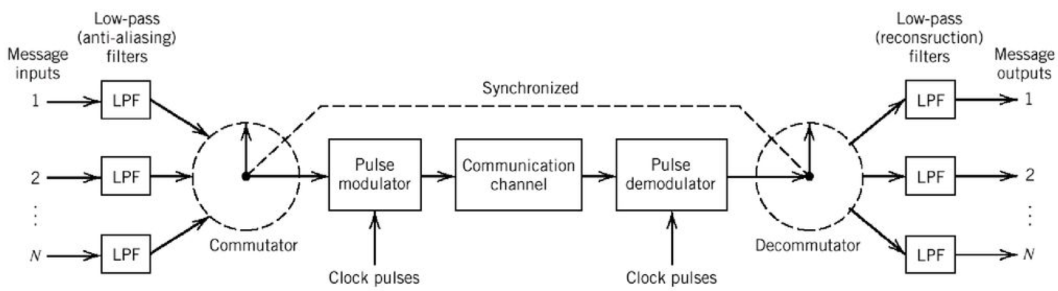
It can be observed that there will be a trade-off in performance with respect to the above two problems. Hence, an optimal value for Δ has to be found.

$$\boxed{\Delta_{opt}\ f_s = \left|\frac{dm(t)}{dt}\right|_{max}}$$

- $\Delta\ f_s << |m'(t)|$ Will result in slope overload
- $\Delta\ f_s >> |m'(t)|$ Will result in granular noise

3.3 Time Division Multiplexing

Time division multiplexing (TDM) is a method of transmitting and receiving independent signals over a common signal path by means of synchronized switches at each end so that each signal appears in an alternating pattern.



If M is the number of message signals, n is the number of encoding bits, f_s is sampling frequency and $T_s(= 1/f_s)$ is the sampling period, the bit rate necessary for synchronization is given by:

$$R_b = \frac{M\ n}{T_s} = M\ n\ f_s$$

Overhead bits are usually added for marking/indication. If a extra bits are added, then

$$R_b = \frac{M\ n + a}{T_s} = (M\ n + a)\ f_s$$

The above formulae work only when the multiple signals are band-limited to the same frequency.
If the signals are band-limited to different frequencies, then TDM can be achieved using 2 different methods. For example, consider message signals are band-limited to f_m, $2f_m$ and $3f_m$.

- Choose sampling rate to be twice the maximum frequency among the messages i.e $f_s = 2f_{max}$.
  Ex: $f_s = 6f_m$. $\implies R_b = 6f_m \times 3 \times n$

- Divide the higher frequency signal into multiple lower frequency signals by feeding the same signal into the commutator multiple times. This will work if the frequencies present are integral multiples of the lowest frequency.
  Ex: $f_s = 3f_m$; feed the second signals twice, the third signal thrice. So total number of signals will be 6. $\implies R_b = 3f_m \times 6 \times n$

3.4 Band Pass Data Transmission

Digital signals (like PCM or DM signals) need to be modulation with high frequency carriers before transmission.
This is done using Digital Carrier Modulation techniques, which converts the base band digital signals to band pass digital signals of high frequency.

3.4.1 Signal Space

The signal space is a framework to better understand and design band-pass transmission of digital signals.

Consider two pulses $p_1(t)$ and $p_2(t)$ that are orthogonal to each other i.e

$$\int_{-\infty}^{\infty} p_1(t)\ p_2(t)\ dt = 0$$

Also, let the energy of both the pulses be equal to unity i.e

$$\int_{-\infty}^{\infty} |p_1(t)|^2\ dt = \int_{-\infty}^{\infty} |p_2(t)|^2\ dt = 1$$

Here, $p_1(t)$ and $p_2(t)$ are called "Orthonormal basis functions".
A signal space consists of all linear combinations of the pulses $p_1(t)$ and $p_2(t)$.

Example for orthonormal basis functions of a signal space:

$$p_1(t) = \sqrt{\frac{2}{T}} \cos(2\pi f_1 t) \quad : \quad 0 < t < T$$

$$p_2(t) = \sqrt{\frac{2}{T}} \sin(2\pi f_2 t) \quad : \quad 0 < t < T$$

where f_1 and f_2 are integral multiples of same fundamental frequency f_0

T is the symbol duration such that integer number of cycles of f_1 and f_2 occur in the duration i.e. $T = k_1/f_1 = k_2/f_2$ where k_1 and k_2 are integers.

The signal space is mainly used to draw the **constellation diagram**, which is a plot of the symbols to be transmitted using the orthonormal basis functions as the axes.

The probability of error will vary inversely with the change is distance between the symbols as represented in the signal space constellation diagram.

3.4.2 Amplitude Shift Keying

Binary data is represented using on-off (unipolar NRZ) signaling. Hence it is also called "on-off keying" (OOK).

The digital message signal a which is $+A$ for "1" or 0 for "0" is simply multiplied with the carrier pulse $p(t) = \sqrt{\frac{2}{T}} \cos(2\pi f_c t)$ which lasts for the duration $0 < t < T$ such that $T = k/f_c$ (k is an integer).

Note that the carrier is normalized to unit energy (i.e $E_p = 1$).

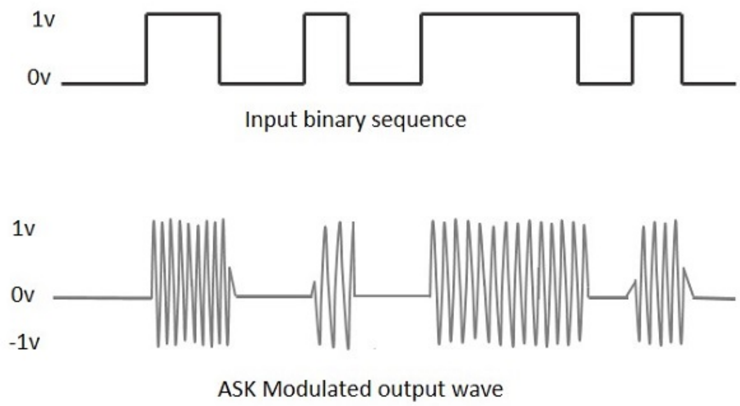
$$s(t) = \begin{cases} s_1 = A\sqrt{\frac{2}{T}} \cos(2\pi f_c t) & : a = 1 \\ s_2 = 0 & : a = 0 \end{cases}$$

Energy of bit if s_1 occurs is A^2 and energy of bit if s_2 occurs is 0.
If s_1 and s_2 are equally likely to occur, then the average energy per bit is give by,

$$E_b = \frac{A^2}{2} \; ; \qquad A = \sqrt{2E_b}$$

Hence the modified expression of ASK modulated signal is given by,

$$s(t) = \begin{cases} s_1 = \sqrt{\frac{4E_b}{T}} \cos(2\pi f_c t) & : a = 1 \\ s_2 = 0 & : a = 0 \end{cases}$$



The bandwidth of the base-band polar NRZ pulse is $1/T$ but now since it is modulated to be a pass-band signal, the bandwidth of ASK modulated signal will be $2/T$.

Difference energy is given by,

$$E_d = E_b + 0 - 2(0) = E_b$$

Assuming AWGN channel, the received symbol will be $r(t) = s(t) + n(t)$, which will be sampled at $t = T$. In detection of ASK modulated signal using matched filter receiver and threshold decision device, the optimal threshold will be $A/2$.

- $r(T) > A/2 \implies : a = 1$
- $r(T) < A/2 \implies : a = 0$

The probability of error in ASK detection is given by,

$$\boxed{P_e = Q\left(\sqrt{\frac{E_b}{2N_0}}\right)}$$

3.4.3 Frequency Shift Keying

Binary data is represented using polar non-return to zero signaling.
Two carriers of different frequencies are generated.

$f_1 = f_c + K_f V \;\rightarrow$ Mark Frequency
$f_2 = f_c - K_f V \;\rightarrow$ Space Frequency

Note that here, $f_1 = k_1/T$ and $f_2 = k_2/T$ where T is the bit duration and k_1, k_2 are integers.

The binary data a represented in polar NRZ form is multiplied to one of the two carriers depending on its value, which is done using Voltage Controlled Oscillator (VCO).

$$s(t) = \begin{cases} s_1 = A\sqrt{\frac{2}{T}}\cos(2\pi f_1 t) & : a = 1 \\ s_2 = A\sqrt{\frac{2}{T}}\cos(2\pi f_2 t) & : a = 0 \end{cases}$$

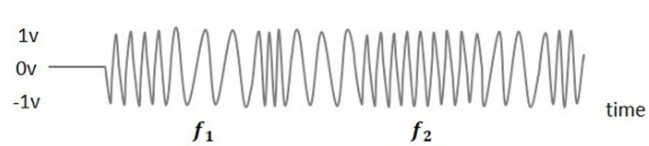
Energy of bit if s_1 occurs is A^2 and energy of bit if s_2 occurs is also A^2.
If s_1 and s_2 are equally likely to occur, then the average energy per bit is give by,

$$E_b = \frac{A^2}{2} + \frac{A^2}{2} = A^2 \; ; \qquad A = \sqrt{2E_b}$$

Hence the modified expression of FSK modulated signal is given by,

$$s(t) = \begin{cases} s_1 = \sqrt{\frac{2E_b}{T}}\cos(2\pi f_1 t) & : a = 1 \\ s_2 = \sqrt{\frac{2E_b}{T}}\cos(2\pi f_2 t) & : a = 0 \end{cases}$$

Input binary sequence



FSK Modulated output wave

Bandwidth of FSK modulated signal is $f_1 - f_2 + 2/T$ or $(k_1 - k_2 + 2)/T$.

Difference energy is given by,

$$E_d = E_b + E_b - 2(0) = 2E_b$$

Assuming AWGN channel, the received symbol will be $r(t) = s(t) + n(t)$, which will be sampled at $t = T$. In detection of FSK modulated signal using matched filter receiver and threshold decision device, the decision has to be taking by measuring the frequency instead of amplitude.

- $r(T) > f_c \implies : a = 1$
- $r(T) < f_c \implies : a = 0$

The probability of error in FSK detection is given by,

$$\boxed{P_e = Q\left(\sqrt{\frac{2E_b}{2N_0}}\right) = Q\left(\sqrt{\frac{E_b}{N_0}}\right)}$$

Note that the bandwidth needed for FSK is higher when compared to ASK but since the symbols are place further apart, the probability of error will be reduced.

3.4.4 Binary Phase Shift Keying

Binary data is represented using polar non-return to zero signaling.
The digital message signal a which is $+A$ for "1" or $-A$ for "0" is simply multiplied with the carrier pulse $p(t) = \sqrt{\frac{2}{T}}\cos(2\pi f_c t)$ which lasts for the duration $0 < t < T$ such that $T = k/f_c$ (k is an integer).

Note that the carrier is normalized to unit energy (i.e $E_p = 1$).

$$s(t) = \begin{cases} s_1 = A\sqrt{\frac{2}{T}}\cos(2\pi f_c t) & : b(t) = 1 \\ s_2 = -A\sqrt{\frac{2}{T}}\cos(2\pi f_c t) & : b(t) = 0 \end{cases}$$

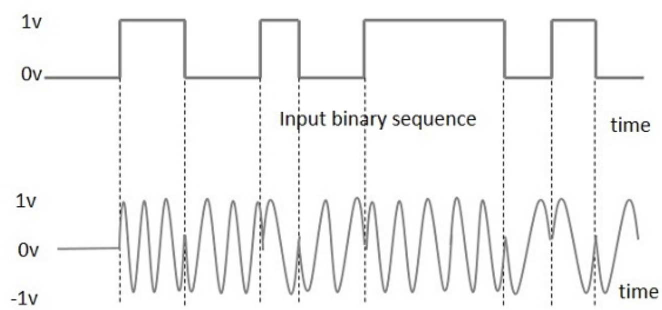
Energy of bit if s_1 occurs is A^2 and energy of bit if s_2 occurs is also A^2.
If s_1 and s_2 are equally likely to occur, then the average energy per bit is give by,

$$E_b = \frac{A^2}{2} + \frac{A^2}{2} = A^2 \ ; \qquad A = \sqrt{E_b}$$

Hence, the modified expression of BPSK modulated signal is give by,

$$\implies s(t) = \begin{cases} s_1 = \sqrt{\frac{2E_b}{T}} \cos\left(2\pi f_c t\right) & : a = 1 \\ s_2 = -\sqrt{\frac{2E_b}{T}} \cos\left(2\pi f_c t\right) & : a = 0 \end{cases}$$

Since $\cos(x+\pi) = -\cos(\pi)$, it can be noted that the transmitted wave-forms for "1" and "0" are phase shifted by 180^o.



BPSK Modulated output wave

The bandwidth of the base-band polar NRZ pulse is $1/T$ but now since it is modulated to be a pass-band signal, the bandwidth of BPSK modulated signal will be $2/T$.

Difference energy is given by,

$$E_d = E_b + E_b - (-2E_b) = 4E_b$$

Assuming AWGN channel, the received symbol will be $r(t) = s(t) + n(t)$, which will be sampled at $t = T$. In detection of BPSK modulated signal using matched filter receiver and threshold decision device, the optimal threshold amplitude will be 0.

- $r(T) > 0 \implies : a = 1$
- $r(T) < 0 \implies : a = 0$

The probability of error in BPSK detection is given by,

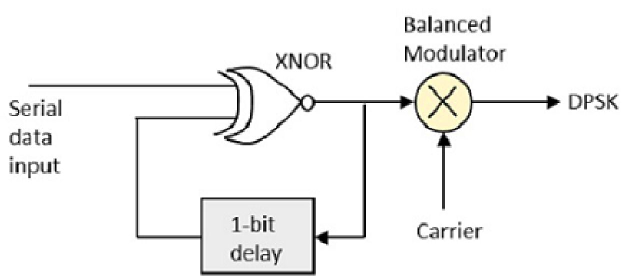
$$\boxed{P_e = Q\left(\sqrt{\frac{4E_b}{2N_0}}\right) = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)}$$

Note that the bandwidth required for BPSK is same ASK (and hence lower than FSK) and since the symbols are place further apart, the probability of error will less than both the previous schemes.

3.4.5 Differential Phase Shift Keying

The disadvantage of BPSK is that it requires the carrier used in demodulation need to be phase synchronized with the carrier used in demodulation.
If the carriers are out of phase, then quadrature nulling can occur and distort the original signal.
DPSK is a scheme designed so that no carrier is necessary at the receiver end, hence quadrature nulling effect is eliminated entirely.

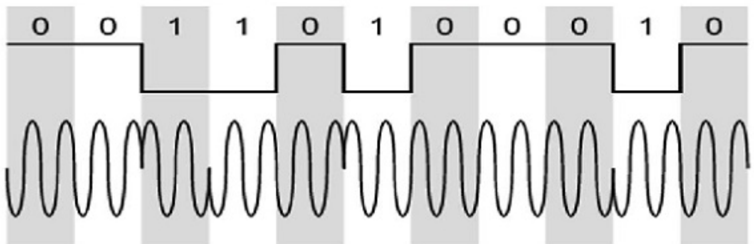
DPSK modulator is shown in the figure below.



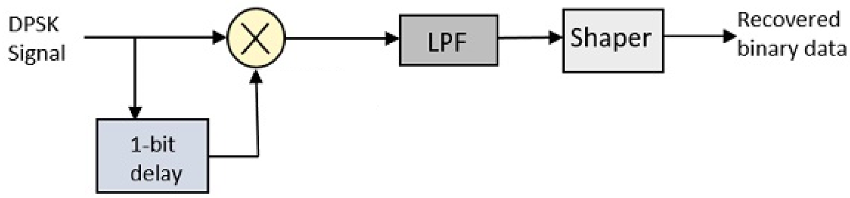
Reference signal for the first XNOR operation is generally taken as 1.

Hence, the basic working is as follows:

- Previous bit is 0 $\implies$ Complement current bit
- Previous bit is 1 $\implies$ Retain current bit



DPSK demodulator is shown in the figure below.



3.4.6 Quadrature Phase Shift Keying

QPSK is a simple case of using quadrature carrier multiplexing technique. The binary data is represented using 2 bits ($n = 2$) and 4 levels ($M = 4$).

Let the two orthogonal carrier pulses be $p_1(t) = \sqrt{\frac{2}{T}}\cos(2\pi f_c t)$ and $p_2(t) = \sqrt{\frac{2}{T}}\sin(2\pi f_c t)$. Note that these two are normalized to unit energy, hence they form an orthonormal basis for a signal space.

A linear combination of the two carriers is given by $s(t) = a_1 p_1(t) + a_2 p_2(t)$. If binary data (i.e the messages a_1 and a_2) can take values A or $-A$, then the possible combinations of transmitted signals are,

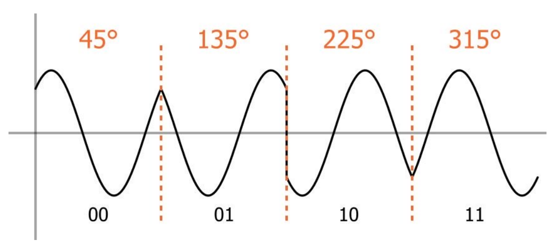
$$s(t) = \begin{cases} s_1 = A\sqrt{\frac{2}{T}}\cos(2\pi f_c t) + A\sqrt{\frac{2}{T}}\sin(2\pi f_c t) \\ s_2 = A\sqrt{\frac{2}{T}}\cos(2\pi f_c t) - A\sqrt{\frac{2}{T}}\sin(2\pi f_c t) \\ s_3 = -A\sqrt{\frac{2}{T}}\cos(2\pi f_c t) + A\sqrt{\frac{2}{T}}\sin(2\pi f_c t) \\ s_4 = -A\sqrt{\frac{2}{T}}\cos(2\pi f_c t) - A\sqrt{\frac{2}{T}}\sin(2\pi f_c t) \end{cases}$$

Hence, there will be 4 distinct symbols to transmit using only 2 carriers. Upon simplifying the above expressions, the modified QPSK wave-forms are obtained as:

$$s(t) = \begin{cases} s_1 = \frac{2A}{\sqrt{T}} \cos\left(2\pi f_c t - 45^o\right) & : a_1 a_2 = 11 \\ s_2 = \frac{2A}{\sqrt{T}} \cos\left(2\pi f_c t + 45^o\right) & : a_1 a_2 = 10 \\ s_3 = \frac{2A}{\sqrt{T}} \cos\left(2\pi f_c t + 135^o\right) & : a_1 a_2 = 01 \\ s_4 = \frac{2A}{\sqrt{T}} \cos\left(2\pi f_c t + 225^o\right) & : a_1 a_2 = 00 \end{cases}$$

The phase shift between each of the successive symbols is 90^o.

The waveform illustrates the phase shifted cosine signal as per the QPSK scheme given.



For receiving the QPSK signal, two matched filters have to be used i.e one corresponding to each carrier.

$h_1(t)$ is used for $s_1(t)$ and optimal decision boundary is 0 to determine is the message sent with $p_1(t)$ i.e a_1 is $+A$ or $-A$. Similarly, $h_2(t)$ is used for $s_2(t)$ and optimal decision boundary is 0 to determine is the message sent with $p_2(t)$ i.e a_2 is $+A$ or $-A$.

Now the received symbol has error if error has occurred in detection of either a_1 or a_2 or both.

Since the scheme is phase shift keying, the probability of error in detection of a_1 or a_2 is,

$$P_{e1} = P_{e2} = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$$

Overall probability of error is,
$P_e = P_{e1}(1 - P_{e2}) + P_{e2}(1 - P_{e1}) + (1 - P_{e1})(1 - P_{e2}) = 1 - (1 - P_{e1})^2$

$$\implies P_e = 2Q\left(\sqrt{\frac{2E_b}{N_0}}\right) - Q^2\left(\sqrt{\frac{2E_b}{N_0}}\right)$$

Since $P_{e1} << 1$, the square term will be even smaller and can usually be neglected.

$$\boxed{\therefore P_e \approx 2Q\left(\sqrt{\frac{2E_b}{N_0}}\right)}$$

3.4.7 M-ary Pulse Amplitude Modulation

So far, all the schemes studied were binary i.e a pulse could take one of two values (except for QPSK).
In M-ary schemes, a pulse can take one of M distinct values. Hence, the number of bits per symbol (level) if M levels are present is $n = \log_2 M$.

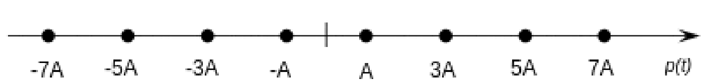
If the transmitted symbol is generalized as $s(t) = ap(t)$, then $a = \pm(2i+1)A$ where $i = 0, 1, 2, ...\frac{M}{2} - 1$ and $2A$ is the spacing between two consecutive levels.

Average symbol energy is given by,

$$E_s = \frac{2}{M} \sum_{i=0}^{\frac{M}{2}-1} (2i+1)^2 A^2 = \frac{A^2}{3}(M^2 - 1)$$

$$\implies A = \sqrt{\frac{3E_s}{M^2 - 1}}$$

The constellation diagram of M-ary PAM scheme with $M = 8$ is illustrated below.



At the receiver end, decision device follows nearest neighbor decision rule. For example in the above scenario:

- $r(T) < -6A \ : \quad \Longrightarrow \ a = -7A$
- $-6A < r(T) < -4A \ : \quad \Longrightarrow \ a = -5A$
- $-4A < r(T) < -2A \ : \quad \Longrightarrow \ a = -3A$
- $-2A < r(T) < 0 \ : \quad \Longrightarrow \ a = -A$
- $0 < r(T) < 2A \ : \quad \Longrightarrow \ a = A$
- $2A < r(T) < 4A \ : \quad \Longrightarrow \ a = 3A$
- $4A < r(T) < 6A \ : \quad \Longrightarrow \ a = 5A$
- $r(T) > 6A \ : \quad \Longrightarrow \ a = 7A$

Assuming AWGN channel and matched filter receiver, the probability can be calculated as follows.
Consider the message signal $a = A$ (or any intermediate level). Then the received signal will be $r(T) = s + \tilde{n}$. Now if $r(T)$ is less than 0 or greater than $2A$, there will be an error in detection.

$$P_{e1} = P(r(T) > 2A) + P(r(T) < 0) = P(\tilde{n} > A) + P(\tilde{n} < -A) = 2P(\tilde{n} < A)$$

Hence, probability of error in receiving an intermediate level is give by,

$$\Longrightarrow \ P_{e1} = 2Q\left(\sqrt{\frac{A^2}{N_0/2}}\right) = 2Q\left(\frac{A}{\sigma}\right)$$

Now consider the message signal $a = 7A$ (or $a = -7A$ i.e an end point). Then the received signal will be $r(T) = s + \tilde{n}$. Now if $r(T)$ greater than $7A$ only there will be an error in detection.

$$P_{e2} = P(r(T) > 7A) = P(\tilde{n} > A)$$

Hence, probability of error in receiving the first or last level is give by,

$$\Longrightarrow \ P_{e2} = Q\left(\sqrt{\frac{A^2}{N_0/2}}\right) = Q\left(\frac{A}{\sigma}\right)$$

Assuming all M levels occur at equal probabilities of $1/M$, since there are $M-2$ intermediate levels and 2 end levels, the average probability of error will be,

$$P_e = \frac{M-2}{M}\, 2Q\left(\frac{A}{\sigma}\right) + \frac{2}{M}\, Q\left(\frac{A}{\sigma}\right)$$

Upon simplifying and substituting for A and σ,

$$\boxed{\therefore P_e = 2\ (1 - \frac{1}{M})\ Q\left(\sqrt{\frac{6E_s}{N_0(M^2-1)}}\right)}$$

3.4.8 M-ary Quadrature Amplitude Modulation

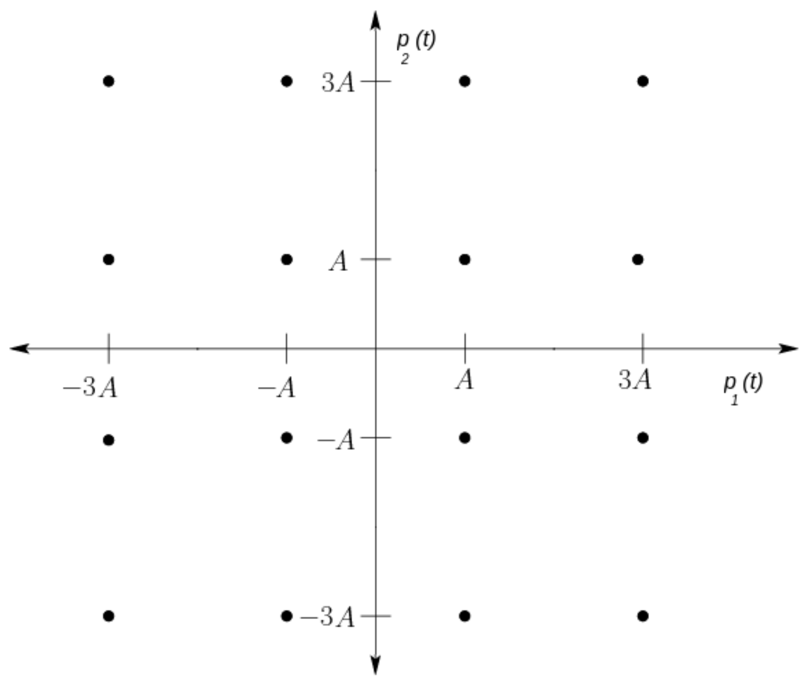
M-ary QAM is a scheme which uses two $\sqrt{\text{M}}$-ary PAM signals with two orthogonal carrier pulses.
The number of bits required transmitted per symbol is again, $n = \log_2 M$.

The transmitted symbol is generalized as $s(t) = a_1p_1(t) + a_2p_2(t)$, where both a_1 and a_2 can take $\sqrt{M}$ levels.
$a_1 = \pm(2i+1)A$ where $i = 0, 1, 2 ... \frac{\sqrt{M}}{2} - 1$
$a_2 = \pm(2j+1)A$ where $j = 0, 1, 2 ... \frac{\sqrt{M}}{2} - 1$
$2A$ is the spacing between two consecutive levels.

Average symbol energy of M-ary QAM is E_s and since it consists of two independent $\sqrt{\text{M}}$-ary PAM signals, the average symbol energy of each PAM signal is $E_s/2$.

$$\frac{E_s}{2} = \frac{A^2}{3}\,(M-1)\ ; \qquad \Longrightarrow A = \sqrt{\frac{3E_s}{2(M-1)}}$$

Since M-QAM is orthogonal combination of $\sqrt{\text{M}}$-PAM, the constellation diagram will have a square structure in the signal space formed by the two orthonormal pulses.



The M-QAM signal has to be received using two matched filters (similar to QPSK) since there are two different pulses.

$h_1(t)$ is used for receiving $a_1 p_1(t)$, which is sampled and nearest neighbor decision rule is applied on $r_2(T)$ to get a_1.
Similarly, $h_2(t)$ is used for receiving $a_2 p_2(t)$, which is sampled and nearest neighbor decision rule is applied on $r_2(T)$ to get a_2.

Now the received symbol has error if error has occurred in detection of either a_1 or a_2 or both.
Since the individual scheme is $\sqrt{\text{M}}$-PAM, the probability of error in detection of a_1 or a_2 is,

$$P_{e1} = P_{e2} = 2\left(1 - \frac{1}{\sqrt{M}}\right) Q\left(\sqrt{\frac{3E_s}{N_0(M-1)}}\right)$$

Overall probability of error is, $P_e = 1 - (1 - P_{e1})^2 = 2P_{e1} - P_{e1}^2$
Since $P_{e1} << 1$, the square term will be even smaller and can usually be neglected.

$$\boxed{\therefore P_e \approx 4\ \big(1 - \frac{1}{\sqrt{M}}\big)\ Q\Bigg(\sqrt{\frac{3E_s}{N_0(M-1)}}\Bigg)}$$

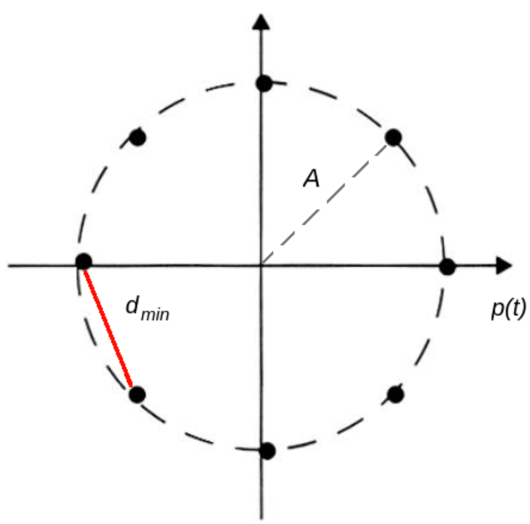
3.4.9 M-ary Phase Shift Keying

M-ary PSK is a more general scheme of BPSK which modulates M symbols by differentiating each of them using M phases (as opposed to just 2 in BPSK).
The number of bits per symbol is obviously $n = \log_2 M$.

Each symbol is modulated at a phase of $2\pi i/M$ where $i = 0, 1..M - 1$.

For example, consider $8 - PSK$ i.e $M = 8$.
The phases at which each of the symbols are modulated are $0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi, \frac{5\pi}{4}, \frac{3\pi}{2}, \frac{7\pi}{4}, 2\pi$.

It can easily be observed that in this scheme, the signal constellation diagram takes a circular shape since all the symbols are separated only by phase difference and hence they are equidistant from the origin of the signal space.



Optimal decision rule is selecting the symbol in the constellation diagram with the least distance from the received symbol.

Average energy per symbol is $E_s = A^2$. Hence, $A = \sqrt{E_s}$.

Probability of error is calculated by using an approximate formula which is obtained using number of nearest neighbors and minimum distance.
Here, each signal in the signal space diagram has $\alpha = 2$ nearest neighbors separated by d_{min}.
The minimum distance can be calculated using basic geometry as,

$$d_{min} = 2\sqrt{E_s}\sin\left(\frac{\pi}{M}\right)$$

The probability of error is given by,

$$\boxed{P_e = \alpha\ Q\left(\frac{d_{min}}{\sqrt{2N_0}}\right) = 2\ Q\left(\sqrt{\frac{2E_s\ \sin^2\left(\frac{\pi}{M}\right)}{N_0}}\ \right)}$$

3.5 Inter Symbol Interference

If a pulse is limited to a finite time interval, then in reality the bandwidth occupied by it will be infinite (though 95% of the power will be present within major lobe).
Similarly, if a pulse is band-limited to a certain range of frequencies, it can't be time limited i.e its time response will be non-zero for all time.

Spreading of a pulse beyond its allocated time interval will cause it to interfere with neighbouring pulses. This phenomenon is called "Inter symbol interference" (ISI).

ISI can't be avoided completely since time-limited pulse are not band-limited and band-limited pulse are not time-limited.
However, pulse amplitude can be detected correctly despite this by properly shaping the pulse.

3.5.1 Nyquist Pulse Shaping Criteria

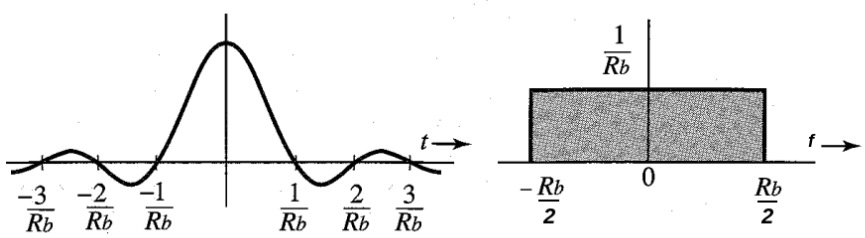
Zero ISI can be achieved by using a pulse shape that has a non-zero amplitude at its center ($t = 0$) and zero amplitudes at $t = nT_b$ where T_b is the separation between successive pulses and $n = 1, 2, 3....$

$$p(t) = \begin{cases} 1 & : t = 0 \\ 0 & : t = nT_b \end{cases}$$

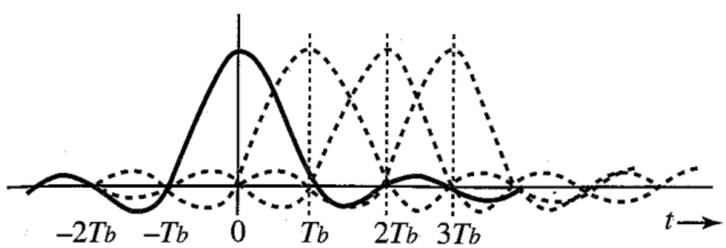
If the duration is T_b, then the bit-rate will be $R_b = 1/T_b$, which means minimum transmission bandwidth is $R_b/2$.

The only pulse $p(t)$ that will satisfy the Nyquist criterion and has bandwidth $R_b/2$ is the sinc function i.e,

$$p(t) = \text{sinc}(\pi R_b t) \longleftrightarrow P(f) = \frac{1}{R_b}\text{rect}\left(\frac{f}{R_b}\right)$$



The reason why it works is because addition of time shifted versions of the sinc pulse will note interfere with itself at integral multiples of T_b as illustrated below.



Since the sinc function decays as $1/t$, it is too slow and hence is not very practical to use because even slight jitters might cause ISI.
In order to obtain a pulse that decays much faster the sharp transition edge of the rectangular spectrum has to be smoothed.

Hence, making the signal decay faster comes at the expense of extra bandwidth.

If the excess bandwidth required is f_x and α is the roll-off factor of the smoothed spectrum (beyond $0.5R_b$), the relation between them is given by,

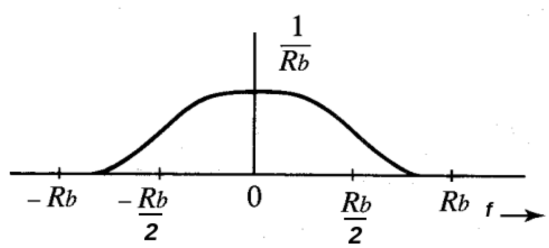
$$\alpha = \frac{f_x}{0.5R_b} = 2f_xTb$$

f_x has to be less than $0.5R_b$ and hence $0 < \alpha < 1$.

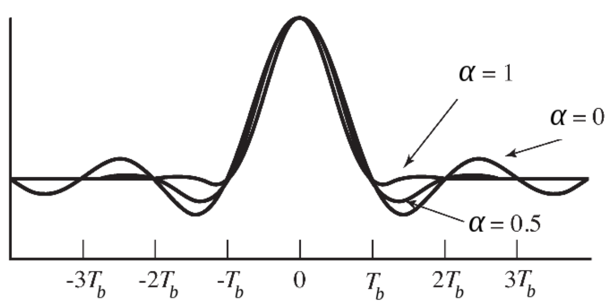
The new overall bandwidth required to transmit the reshaped pulse will be

$$BW = 0.5\ R_b + f_x = \frac{(1+\alpha)R_b}{2}$$

The spectrum of the shaped signal is called **Raised Cosine Spectrum**.



The sinc pulse obtained after smoothening the transition band in the spectrum for few standard roll-off rates are illustrated.



Note that for maximum value of α i.e 1, the sinc pulse decays almost instantaneously and the signal is pretty much non-existent outside the major lobe, which makes it desirable.

3.5.2 Raised Cosine Filter

The raised cosine filter is the filter used to shape a pulse in such a way that the raised cosine spectrum i.e one which satisfies the Nyquist criterion is obtained.
One such practical filter that realizes this has frequency response as follows.

$$H_{rc}(f) = \begin{cases} 1 & : f < 0.5R_b - f_x \\ 0.5\left[1 - \sin\pi\left(\frac{f-0.5R_b)}{2f_x}\right)\right] & : 0.5R_b - f_x < f < 0.5R_b + f_x \\ 0 & : f > 0.5R_b + f_x \end{cases}$$

If the filter has such a frequency response, it will shape the signal such that Nyquist criterion is obeyed and hence zero ISI transmission is achieved.

However practical communications systems since a matched filter is used in the receiver, for zero ISI, it is the net response of the transmit and receive filters that must give the frequency response of the raised cosine filter $H_{rc}(f)$.
$\implies H_{rc}(f) = H_R(f)\ H_T(f)$
$\therefore |H_R(f)| = |H_T(f)| = \sqrt{|H_{rc}(f)|}$

These filters with the response $\sqrt{|H_{rc}(f)|}$ are called "Root Raised Cosine Filters".

4 Information Theory

Information theory studies the quantification, storage, and communication of digital information.

As noted before, a message signal produced is encoded into a stream of bits hence producing a binary signal, which is used in digital communication systems.
Information theory deals with various schemes of encoding a message signal, quantifying different schemes in order to figure out how to maximize the information content per bit or minimize probability of error or minimize bandwidth usage, etc.

4.1 Characterizing Information

An information source can either be digital or analog. But usually analog information is converted to digital information anyway.

A digital source sequentially produces symbols from a given finite set of symbols. For example, S set of M symbols: $S = [s_0, s_1, s_2, ...s_{M-1}]$.

The source will generate symbols based on a probability distribution function. Meaning, each of the discrete symbols is generated with some probability.
Hence, s_0 is generated with probability p_0, s_1 is generated with probability p_1 and so on.
(note that $p_0 + p_1 + ... + p_{M-1} = 1$)

It is generally assumed that the generation of symbols are uncorrelated and generation of a symbol is independent of what is going to be generated next or what has already been generated.

The information content in a symbol is characterized as follows:

$$I_{s_i} = \log_2 \left(\frac{1}{p_i} \right)$$

where I_{s_i} is the information of symbol s_i and p_i is the probability of occurrence of the symbol

Information content is measured in bits.

The above quantification of information says that symbols (or events) that have high probability of occurrence carry less information and symbols (or events) that have low probability of occurrence carry high information.

4.1.1 Entropy

Entropy of a digital source S is defined as the expected value of information or the average information content present in the source. $H(S) = E[I(S)]$

$$\implies \boxed{H(S) = \sum_{i=0}^{M-1} p_i \log_2 \Big(\frac{1}{p_i}\Big)}$$

Note that the entropy is a sum of non-negative terms, hence $H(S) \geq 0$ and it is measured in bits/symbol.

If $p_i = 1$, then $\log_2\left(\frac{1}{p_i}\right) = 0$
If $p_i = 0$, then $\log_2\left(\frac{1}{p_i}\right) = 0$ (use L-Hospital's rule to evaluate limit)

Another property of entropy is : $H(S^n) = nH(S)$.

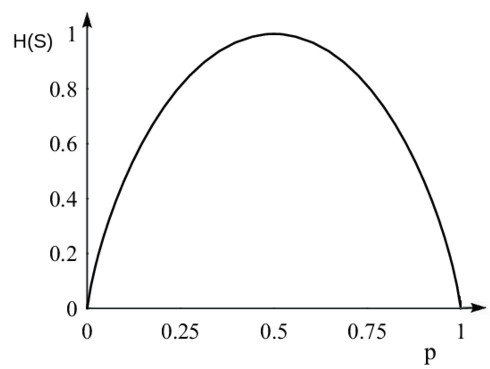
Consider a binary information source $S = [s_0, s_1]$ where probability of occurrence of s_0 is p and probability of occurrence of s_1 is $1 - p$.
Entropy of the binary source is given by,

$$H(S) = p\log_2\Big(\frac{1}{p}\Big) + (1-p)\log_2\Big(\frac{1}{1-p}\Big)$$

Note at as $p \to 0$ or as $p \to 1$, the value of $H(s) \to 0$.

This is because if $p = 0$, then the source will always generate s_0 and if $p = 1$, the source will always generate s_1 and hence in either case there is no average information content.

The maximum value of entropy $H(S)$ can be obtained by differentiating it w.r.t p and equating to 0. This will show that $H(S)_{max} = 1$ when $p = 1 - p$ i.e $p = 0.5$.



Therefore, the entropy is maximum when all the symbols are equally likely to occur (this result can be extended to more than just binary sources).

In case of an analog information source S, the source has to be modelled such that it takes values based on some probability distribution function. Hence, S can be considered a random process and the value it takes is a random variable that follows some distribution $f_S(x)$.

In this case, the entropy of the source is also called "differential entropy" and can be calculated as follows.

$$H(S) = \int_{-\infty}^{\infty} f_S(x) \log_2 \left(\frac{1}{f_S(x)} \right) dx$$

4.1.2 Information Rate & Symbol Rate

The information rate of a source is the number of bits transmitted per second and is denoted by R.

The symbol rate of a source is the number of symbols transmitted per second and is denoted by r. Note that the number of symbols can also be the number of quantization levels.

Since the entropy of the source (H) is the average information content, it gives the average number of bits present in each symbol. Hence, the relation between the average information, information rate and symbol rate is given by $R = H\ r$.

4.1.3 Joint Entropy

Consider two sources $X = [s_0, s_1, ..., s_{M-1}]$ and $Y = [r_0, r_1, ...r_{N-1}]$.
A joint probability distribution can be obtained from the two sources.
$p_{ij} = P(X = s_i \cap Y = r_j) = P(s_i, r_j)$

The joint entropy of the two sources X and Y is given by,

$$H(X,Y) = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} p_{ij}\ \log_2\left(\frac{1}{p_{ij}}\right) = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} P(s_i, r_j)\ \log_2\left(\frac{1}{P(s_i, r_j)}\right)$$

Some useful relations in the joint distribution:

$$\rightarrow \quad \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} P(s_i, r_j) = 1$$

$$\rightarrow \quad \sum_{i=0}^{M-1} P(X = s_i \cap Y = r_a) = P(Y = r_a)$$

$$\rightarrow \quad \sum_{j=0}^{N-1} P(X = s_b \cap Y = r_j) = P(X = s_b)$$

Probabilities $P(X = s_b)$ and $P(Y = r_a)$ are called "Marginal Probabilities".

The entropies of individual sources i.e $H(X)$ and $H(Y)$ can be calculated from the usual entropy formula using the marginal probabilities.

Properties of joint entropy:

- The joint entropy of two sources is always greater than or equal to the individual entropies of the two sources.
  $\implies \quad H(X,Y) \geq H(X); \quad H(X,Y) \geq H(Y)$
- The joint entropy of two sources is always lesser than or equal to the sum of individual entropies of the two sources.
  $\implies \quad H(X,Y) \leq H(X) + H(Y)$

The second property gives rise to Jensen's inequality that says,
$H(X,Y) - H(X) - H(Y) \leq 0$.

4.1.4 Conditional Entropy

Conditional entropy is the average information content present in one source, given another source.
Consider two sources X having M symbols $[s_0, s_1, ...s_{M-1}]$ and Y having N symbols $[r_0, r_1, ...r_{N-1}]$. Then the conditional entropy on source Y given source X is defined as,

$$H(Y|X) = \sum_{i=0}^{M-1} P(s_i) \sum_{j=0}^{N-1} P(r_j|s_i) \log_2 \left(\frac{1}{P(r_j|s_i)} \right)$$

By taking $P(s_i)$ inside the second summation and using $P(s_i, r_j) = P(s_i)P(r_j|s_i)$, the equation becomes

$$H(Y|X) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} P(r_j, s_i) \log_2 \left(\frac{1}{P(r_j|s_i)} \right)$$

Relation between Joint entropy and Conditional entropy:

- $H(X, Y) = H(Y|X) + H(X)$
- $H(X, Y) = H(X|Y) + H(Y)$

From the above relations, it can be intuitively concluded that the conditional entropy is the information content left in one source when the other source has already been observed.

4.1.5 Mutual Information

Mutual information gives the common information content contained in two different sources.

Consider two sources X and Y. The mutual information between the two sources is given by,

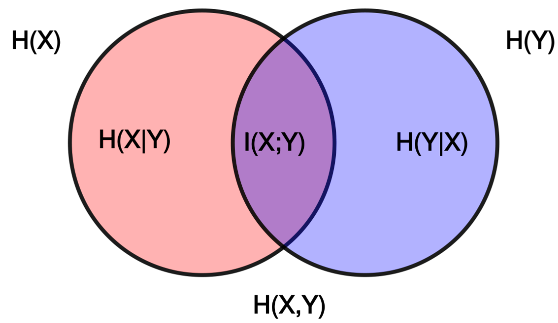
$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Using the relations between joint and conditional entropies, mutual information can also be expressed as, $I(X;Y) = H(X) + H(Y) - H(X, Y)$.

If X denotes the set of symbols transmitted through a channel and Y denotes the set of symbols received from a channel, then mutual information can be interpreted as the amount of information in X obtained after observing Y.

Ideally the value of $I(X, Y)$ has to be as large as possible since maximum information has to obtained regarding X by observing Y.

Note that $I(X; Y) = I(Y; X)$ and $I(X; Y) \geq 0$.



Consider a binary source X with $s_0 = 0$ and $s_1 = 1$ and let Y be the received symbol with $r_0 = 0$ and $r_1 = 1$.
Let the probability of generation of s_0 be α and hence the probability of generation of s_1 will be $1 - \alpha$.
Assume binary symmetric channel with flipping probabilities of p.

$$P(r_0|s_0) = P(r_1|s_1) = 1 - p \qquad\qquad P(r_0|s_1) = P(r_1|s_0) = p$$

The following notation will be used for convenience.

$$\left[H(k) = k \log_2 \frac{1}{k} + (1 - k) \log_2 \frac{1}{(1 - k)} \right]$$

Now the conditional entropy of Y given X can be calculated as,

$$H(Y|X) = P(s_0)H(Y|s_0) + P(s_1)H(Y|s_1) = \alpha H(p) + (1 - \alpha)H(p) = H(p)$$

To calculate mutual information, $H(Y)$ is necessary. It can be computed as, $H(Y) = H(p + \alpha - 2p\alpha)$.

$$\implies I(X; Y) = H(p + \alpha - 2p\alpha) - H(p)$$

4.1.6 Probability matrices

While performing calculations in information theory, it is convenient to represent the systems in matrix form.

Consider a system with source X having M symbols $[s_0, s_1, ...s_{M-1}]$ and received symbols Y having N symbols $[r_0, r_1, ...r_{N-1}]$.

- The matrix $[P(X)]$ is an $M \times 1$ matrix consisting of probabilities of occurrence of each of the source symbols in X.
- The matrix $[P(Y)]$ is an $N \times 1$ matrix consisting of probabilities of occurrence of each of the received symbols in Y.
- The matrix $[P(Y|X)]$ is an $M \times N$ matrix consisting of probabilities of receiving a symbol of Y, given that a symbol in X is transmitted. This is called the "channel matrix" or the "conditional probability matrix".
  - The sum of each row in a channel matrix is equal to 1.
- The matrix $[P(X, Y)]$ is an $M \times N$ matrix consisting of probabilities of a symbol transmitted from X being received as a symbol in Y. This is called the "joint probability matrix".
  - The individual probabilities of the source symbols i.e elements of $[P(X)]$ can be obtained by adding the corresponding row elements of $[P(X, Y)]$.
  - The individual probabilities of the received symbols i.e elements of $[P(Y)]$ can be obtained by adding the corresponding column elements of $[P(X, Y)]$.
- The matrix $[P(X|Y)]$ is an $M \times N$ matrix consisting of probabilities of the transmitted symbol being a symbol in X provided that a symbol in Y has been received. This is called the "posterior probability matrix".
  - The sum of each column in a channel matrix is equal to 1.

The elements of joint matrix $[P(X, Y)]$ are usually obtained by multiplying corresponding elements of $[P(X)]$ and $[P(Y|X)]$ (because source probabilities $[P(X)]$ and channel matrix $[P(Y|X)]$ are usually available data).

The elements of posterior probability matrix $[P(X|Y)]$ (which are helpful in MAP decoding) are obtained by dividing the elements of joint matrix $[P(Y|X)]$ with the corresponding received symbol probabilities from $[P(Y)]$.

The entorpies can be obtained more clearly from these probability matrices. $H(X)$ is calculated from $[P(X)]$. $H(Y)$ is calculated from $[P(Y)]$.

$H(X,Y)$ is calculated from $P(X,Y)$ i.e,

$$H(X,Y) = \sum P(X,Y) \log_2 \left(\frac{1}{P(X,Y)} \right)$$

$H(Y|X)$ is calculated as:

$$H(Y|X) = \sum P(X,Y) \log_2 \left(\frac{1}{P(Y|X)} \right)$$

$H(X|Y)$ is calculated as:

$$H(X|Y) = \sum P(X,Y) \log_2 \left(\frac{1}{P(X|Y)} \right)$$

From the calculated entropies, mutual information $I(X;Y)$ can be obtained using all the 3 formulae and verified. They must turn out to be equal.

4.1.7 Types of Channels

A binary channel is a channel that has 2 input symbols (say s_0 and s_1) and 2 output symbols (say r_0 and r_1).

- A binary channel is said to be symmetric if the flipping probabilities (or probabilities of error) are the same i.e $P(r_1|s_0) = P(r_0|s_1)$.
- A binary channel is said to be asymmetric if the flipping probabilities (or probabilities of error) are the different i.e $P(r_1|s_0) \neq P(r_0|s_1)$.

In general, there can be some special types of channels. They are:

1. Lossless channel:
   A lossless channel is a channel such that for every received symbol, there is no uncertainty in which symbol is transmitted i.e $H(X|Y) = 0$ and hence $I(X;Y) = H(X)$.
   In a lossless channel matrix $P(Y|X)$, every column has only one non-zero element.

2. Deterministic channel:
   A deterministic channel is a channel such that for every transmitted symbol, there is no uncertainty in which symbol is received i.e $H(Y|X) = 0$ and hence $I(X;Y) = H(Y)$.
   In a deterministic channel matrix $P(Y|X)$, every row has only one non-zero element, which will be equal to 1.

3. Noiseless channel:
   A noiseless channel is a channel that is both lossless and deterministic. This means there will be certain one to one mapping between source symbols and received symbols i.e $H(Y|X) = H(X|Y) = 0$ and hence $I(X;Y) = H(X) = H(Y)$.
   The noiseless channel matrix $P(Y|X)$ is an identity matrix.

4. Useless channel:
   A useless channel is a channel which has 0 mutual information and hence 0 channel capacity (more on channel capacity to be explained).

4.2 Channel Capacity

Channel capacity is defined as the maximum rate at which information can be transmitted over a channel.
It is given by the maximum of all values of mutual information calculated for a channel for different probability distributions of the source.

This is also obvious since a good channel must have high capacity meaning mutual information should be maximized.
$\implies C = \max[I(X;Y)]$

Capacity of Binary Symmetric Channel
For a binary symmetric channel considered earlier, the channel capacity can be calculated by maximizing the expression for mutual information.
$I(X;Y)$ is maximum if $H(p + \alpha - 2p\alpha)$ is maximum since $H(p)$ is constant for a given channel.
$H(p+\alpha-2p\alpha)$ will be equal to 1 when $p+\alpha-2p\alpha = 0.5$ and hence $\alpha = 0.5$.

$\implies C = 1 - H(p)$
Since $H(p) = H(1 - p)$, it is obvious that $C = 1 - H(1 - p)$.

The channel capacity will be maximum if $p = 1$ or $p = 0$ i.e it will be 1 *bit* per channel use.
The channel capacity will be minimum if $p = 0.5$ i.e it will be 0 *bits* per channel use.

Capacity of Gaussian Channel
For a Gaussian channel (i.e a channel which adds white Gaussian noise to the input), the channel capacity can be calculated from Shannon-Hartley theorem.
The Shannon–Hartley theorem states the channel capacity C, i.e the upper bound on the information rate of data that can be communicated at an arbitrarily low error rate using an average received signal power S through an analog communication channel subject to additive white Gaussian noise (AWGN) of power N is,

$$\boxed{C = BW \log_2 \left(1 + \frac{S}{N}\right)} \qquad bits/s$$

where BW is the bandwidth of the channel and hence N is the average power of the noise and interference over the bandwidth.
(note that the SNR is expressed as a linear power ratio, not as logarithmic decibels)

The above expression is obtained by considering the fact that the maximum number of distinguishable pulse levels that can be transmitted and received reliably over a communications channel is limited by the dynamic range of the signal amplitude and the precision with which the receiver can distinguish amplitude levels.
The number of levels M is given by,

$$M = \sqrt{\frac{N+S}{N}} = \sqrt{1 + \frac{S}{N}}$$

It can be observed that the channel capacity can also be expressed as,

$$C = 2BW \log_2 M$$

For a channel of infinite bandwidth, the channel capacity can be derived using the same expression ($BW \rightarrow \infty$).

$$\implies C = 1.44\frac{S}{N_0}$$

N_0 is the amplitude of the noise (AWGN) power spectral density.

4.3 Source Coding

It has been established that in digital communication, a digital source sequentially generates symbols. Usually, the source will be a discrete memoryless source and hence the probabilities of occurrence of any symbol at any instant of time is independent of previous symbols.
Therefore, the symbols are independent and identically distributed.

Source coding refers to the process of encoding (or mapping) these symbols into a sequence of digits like bits which is called the "code".
The goal of source coding would be to minimize the average number of bits used to represent the symbols in a source.

For a source that generates M symbols $s_0, s_1, ...s_{M-1}$ with probabilities $p_0, p_1, ...p_{M-1}$ respectively, the average code-word length (obtained by using any coding technique) is,

$$L = \sum_{i=0}^{M-1} p_i \; l_i$$

where $l_0, l_1, ...l_{M-1}$ are the lengths of the corresponding code-words.

There are two types of source coding techniques.

1. Fixed Length Coding
2. Variable Length Coding

Fixed Length Coding
In fixed length coding, the length of the code-word for each of the symbols is equal and fixed. For example, consider a source that generates 4 symbols s_0, s_1, s_2 and s_3 and the code-words corresponding to the 4 symbols are

00,01,10 and 11 respectively. Here, each code-word is of length 2.

A fixed length code is always "Uniquely Decodable". Meaning, the original sequence of symbols can be extracted from the sequence of code-words. Even though the decoding complexity is low, fixed length coding does not minimize the average length of the code-words and hence is not the most efficient technique.

Variable Length Coding
In variable length coding, the length of the code-word for different symbols is different. There are various schemes in which variable length codes can be generated, and all aim to minimize the average length of the code-words.

A variable length code if not constructed properly will not be uniquely decodable. For example, for the same source illustrated before, if the code-words are 0,1,00 and 11 respectively, it can be noted that given an arbitrary sequence of code-words, the original sequence of symbols can not be obtained.

Therefore, source coding mainly involves in finding uniquely decodable variable length codes that minimize the average length of the code-words.

The most convenient way to obtain uniquely decodable codes is to ensure that the code-words are prefix-free. Meaning, no code-word is a prefix of another code-word. Such codes are called "Prefix-Free Codes" or "Instantaneous Codes" (since they can be decoded instantaneously without the necessity of waiting for arrival of next code-word).

For example, for the same source illustrated before, if the code-words are 1,01,001 and 000 respectively, it can be noted that given an arbitrary sequence of code-words, the original sequence of symbols can be obtained.

To generate a prefix-free code, the following inequality called "Kraft Inequality" has to be followed. It is given by,

$$\sum_{i=0}^{M-1} 2^{-l_i} \leq 1$$

The coding efficiency is defined as the ratio of the entropy of the source to the average code-word length.

$$\eta = \frac{H(s)}{L} = -\frac{\sum_{i=0}^{M-1} p_i \log_2 p_i}{\sum_{i=0}^{M-1} p_i \ l_i}$$

The maximum efficiency that can be obtained is 1, which is possible using prefix-free coding.

4.4 Channel Coding

It is quite obvious that there is finite non-zero probability of error in the received information/signal when compared to the transmitted information/signal. Therefore, it is necessary for the receiver to have mechanisms that can detect and correct the potential errors in the received bits of information.

The basic idea used is to transmit extra bits called "Parity bits" along with the data bits. These parity bits will contain some data about the data bits that can help detect and correct errors. This technique is called "Channel Coding" or "Error Correction Coding".
$\Longrightarrow$ Code-word = Data (message) bits + Parity bits

If there are k data bits and m parity bits, then the length of the code word will consist of $n = k + m$ bits. Such a code is represented as (n, k) code.

The Code Rate of an (n, k) code is the ratio k/n.
Hence, the code rate denotes the fraction of a code-word that actually consists of the data (message).
Note that code rate is always less than 1 (a code rate of 1 means no parity bits have been added and hence no coding has been done).

There are mainly two types of channel coding techniques.

1. Linear Block Codes
2. Convolution Codes

Linear block codes will be covered in detail, convolution codes are left to more advanced study. A sub-division in linear block codes called "Cyclic codes" will also be introduced.

Linear Block Code properties:

- Sum of two code-words belonging to a code is also a code-word belonging to the code
- The all zero code-word is always a valid code-word
- The minimum distance of the code will be equal to the minimum weight among all code-words of the code (excluding the all zero code-word).

Hamming Distance:
The Hamming distance between two code-words is the number of elements (or bits) that are different between them. Hamming distance of any two arbitrary code-words c_1 and c_2 (both of same length) is denoted as $d(c_1, c_2)$.
Example: $c_1 = 01101$, $c_2 = 11100$; $\implies d(c_1, c_2) = 2$

The minimum distance of a code is the minimum Hamming distance between any two code-words in the code.
Consider a code $C = (c_0, c_1, ...c_{M-1})$, then the minimum distance is given by,
$d^* = \min\ [d(c_i, c_j)] \quad i \neq j$.

This is denoted as d_{min} and is an important parameter in error correction codes as it gives the error detection and correction capacity of a code.

- An LBC can detect up-to t errors provided $d_{min} \geq t + 1$
- An LBC can detect up-to t errors provided $d_{min} \geq 2t + 1$

Hamming Weight:
The Hamming weight of any code-word is the number of non-zero elements (or bits) present in the code-word. Hamming weight of any code-word c is denoted as $w(c)$.
Example: $c = 11010$; $\implies w(c) = 3$

Note that the Hamming distance of two code-words is equal to the Hamming weight of the difference between the code-words i.e $d(c_1, c_2) = w(c_1 - c_2)$.

The minimum weight of a code is the minimum Hamming weight of any non-zero code-word in the code and is denoted by w^*.

Generator matrix G is the matrix that converts a data sequence of length k to an encoded sequence of length n.

Consider a message block (data sequence) given by (d_1,d_2,...d_k) which needs to be encoded to a code-word given by (c_1,c_2,...c_k, c_{k+1},...c_{k+m}).
Hence, the generator matrix takes input $d_{k\times 1}$ and gives output $c_{n\times 1}$.
$\implies \quad c = dG$. This means that the generator matrix G is of size $k \times n$.

The first k bits in the code-word will be same as the message bits. The last m bits i.e the parity bits will be linear combinations of the k message bits.
$c_{k+1} = p_{11}d_1 + p_{12}d_2 + ...p_{1k}d_k$, ... , $c_{k+m} = p_{m1}d_1 + p_{m2}d_2 + ...p_{mk}d_k$.
This means that the generator matrix G has rank k.

The coefficients of the data bits that result in the parity bits can be expressed in matrix form, called the "Coefficient Matrix" P.

$$\implies P_{m\times k} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mk} \end{pmatrix}$$

The generator matrix can be expressed by combining the above observations. It will consist of an identity matrix of dimensions k followed by the transpose of the coefficient matrix.
$\implies G_{k\times n} = [I_k : P^T]$.

Parity Check matrix H is the matrix that decodes the encoded sequence back to the original data sequence.
H is obtained such that $GH^T = 0$. Hence, it can be deduced that the parity check matrix should be $H = [P : I_{n-k}]$.
This implies that the parity check matrix H is of size $(n - k) \times n$.

Since $GH^T = 0$ and $c = dG$, it can be concluded that $cH^T = 0$.
Meaning, any valid code-word generated from the generator matrix, when multiplied with the transpose of the parity check matrix, will give [0].
This property is used for detecting and correcting errors.

Syndrome Decoding: The algorithm used to detect and correct errors in the received code-word using the Parity check matrix.

If the transmitted code-word is c and the received code-word is r (both will be rows of n elements), then the error pattern $e = c + r$, which can be modified as $r = c + e$.
The error pattern e will have 0 if there is no error in the corresponding bit position i.e r and c are same and will have 1 if there is an error in the corresponding bit position i.e r and c are different.

Consider $rH^T = cH^T + eH^T = eH^T$. This matrix is called the Syndrome (S).

- If $S = 0$, then no error has occurred.
- If $S \neq 0$, then error has occurred.

$S = eH^T$ will give the row of the parity check matrix which corresponds to the position of the error. So the algorithm simply compares the obtained syndrome S with each row of the matrix H^T and the index of the row which matches with S indicates the position of the error in r, which has to be inverted to obtain back c.

If none of the rows of H^T matches with a non-zero S, then it means error has occurred in more than one position and the algorithm has to check for linear combinations of rows of H^T which match with S to identify error positions.

An (n, k) LBC will have 2^{n-k} distinct syndromes and $2^{n-k} - 1$ distinct non-zero syndromes.

4.4.1 Hamming Code

Hamming coding is the most fundamental error correction code. It can be used correct 1 error.
Hamming codes are (n, k) codes with $n = 2^m - 1$ and $k = 2^m - 1 - m$ and minimum distance $m = 2t + 1$ where t is the number of errors the scheme is capable of correcting (which is 1 here and hence $m = 3$).
The most popular Hamming code is $(7, 4)$ code.

$(7, 4)$ Hamming code consists of 7 bits overall with 4 message bits and 3 parity bits. The parity bits occupy positions that are powers of 2 i.e positions

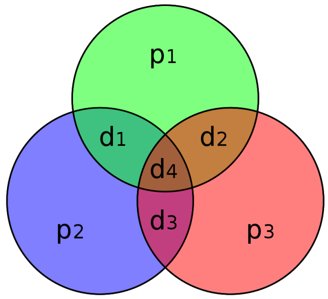
1,2 and 4. The remaining positions are occupied by the data bits.
Hence, in general the Hamming code can be represented as $(d_7, d_6, d_5, p_4, d_3, p_2, p_1)$.

A general algorithm can be deduced from the following description:

- Number the bits starting from 1: bit 1, 2, 3, 4, 5, 6, 7, etc.
- Write the bit numbers in binary: 1, 10, 11, 100, 101, 110, 111, etc.
- All bit positions that are powers of two (have a single 1 bit in the binary form of their position) are parity bits: 1, 2, 4, 8, etc. (1, 10, 100, 1000)
- All other bit positions, with two or more 1 bits in the binary form of their position, are data bits.
- Each data bit is included in a unique set of 2 or more parity bits, as determined by the binary form of its bit position.
  - Parity bit 1 covers all bit positions which have the least significant bit set: bit 1 (the parity bit itself), 3, 5, 7, 9, etc.
  - Parity bit 2 covers all bit positions which have the second least significant bit set: bit 2 (the parity bit itself), 3, 6, 7, 10, 11, etc.
  - Parity bit 4 covers all bit positions which have the third least significant bit set: bits 4–7, 12–15, 20–23, etc.
  - Parity bit 8 covers all bit positions which have the fourth least significant bit set: bits 8–15, 24–31, 40–47, etc.

(In general each parity bit covers all bits where the AND of the parity position and the bit position is non-zero)
The parity bits are calculated by finding the bits necessary to be placed in their respective positions described earlier so as to obtain even parity for their corresponding positions.
In (7,4) Hamming code:

- p_1 is responsible for setting even parity to p_1, d_3, d_5, d_7
- p_2 is responsible for setting even parity to p_2, d_3, d_6, d_7
- p_4 is responsible for setting even parity to p_4, d_5, d_6, d_7



After Hamming code is generated by adding the parity bits to the data bits and transmitted, the received bit sequence is tested for errors.
If due to error any bit is flipped, then the corresponding parity will be found to be odd instead of even and hence it can be concluded that error has been detected.
Process of error correction:

- From the error detection process, a binary number has to be found whose decimal equivalent gives the position of the error.
- For all parity bits that satisfy even parity, set the corresponding binary number bit to 0.
- For all parity bits that do not satisfy even parity (i.e error is detected), set the corresponding binary number bit to 1.
- Write the binary number in order of highest parity to lowest parity bit and convert this binary number to decimal equivalent to get the position of the error.
- Flip the bit in this position to correct the error.

Example: Data bits are (0,1,1,0). $\implies$ (0,1,1,p_4,0,p_2,p_1).

$p_1 = 1$, $p_2 = 1$ and $p_4 = 0$ to set even parities.
The Hamming code sequence is (0,1,1,0,0,1,1) which is transmitted.

If the received sequence is (1,1,1,0,0,1,1), then error is detected because on finding the parity bits again for the data, p_1 should be 0 but it is 1; p_2 should be 0 but it is 1; p_4 should be 1 but it is 0. Here all parity bits are wrong but in general if any one is wrong, it still means error has occurred.

For error correction, note that all 3 parity bits are wrong meaning the binary value of the error location is 111 which corresponds to d_7. Therefore, d_7 has error and it should be 0 instead of 1.

(Note that if two errors have occurred, then Hamming code does not work)

4.4.2 Cyclic Redundancy Check (CRC)

Cyclic redundancy check (CRC) is an error-detecting code based on the theory of cyclic error-correcting codes.

A CRC for a sequence of data bits (message) is generated by the sender using a particular algorithm and another sequence which is known by the receiver as well called the "divisor".

CRC Generation:

- Find length of divisor, L.
- Append $L - 1$ zeros (0s) at the end of the data bit sequence.
- Perform binary division i.e the message with appended zeros divided by the divisor.
  - Subtracting or adding is same as performing XOR operation.
  - If MSB of divisor is 1 and MSB of dividend is 1, then the corresponding quotient bit is 1.
  - If MSB of divisor is 1 and MSB of dividend is 0, then the corresponding quotient bit is 0.
  - If MSB of divisor is 0, then the corresponding quotient bit is always 0.
- The obtained remainder is the CRC.
  Note that the length of the CRC will also be $L - 1$ bits.
- The encoded message to be transmitted will be the original sequence appended by the CRC bits.

Error Detection using CRC:
Once the encoded data is received, the receiver performs the same binary division operation on the received data using the same divisor (note that the common divisor is predetermined by the protocol).

- If the remainder obtained by the division performed by the receiver is a sequence of $L-1$ zeros (0s), then there is no error in the transmitted data.
- If the remainder obtained by the division performed by the receiver is any other sequence, then there is error in the transmitted data and hence error detection is achieved.