EEEN3005J:
Communication Theory

Lecture Notes

School of Electrical, Electronic & Communications Engineering,
University College Dublin

Contents

Acronyms

A	Short for Ampere, the measurement unit for electric current
ADC	Analog to Digital Converter
Amp	Short for Ampere, the unit for electric current
C	Common letter used to denote a capacitor
dB	Decibel
F	Farad (units for capacitors)
G	Common letter used to denote a conductor
H	Henry (unit for inductors)
L	Common letter used to denote an inductor
LTI	Linear Time Invariant
PDF	Probability Density Function
PSD	Power Spectral Density
R	Common letter used to denote a resistor
RMS	Root Mean Squared
S	Siemens (units for conductance), see also $\mho$
SNR	Signal to Noise Ratio
V	Short for Volts, the measurement unit for electric Voltage
Ω	Ohm (units for resistors)
$\mho$ or Ω^{-1}	Mho (units for conductance), see also Siemens

Chapter 1

Fourier transforms

Some commonly used Fourier transforms are listed below:

time domain	frequency domain
$\delta(t)$	1
1	$\delta(f)$
$e^{j2\pi f_c t}$	$\delta(f - f_c)$
$\cos(2\pi f_c t)$	$\frac{1}{2}[\delta(f - f_c) + \delta(f + f_c)]$
$\sin(2\pi f_c t)$	$\frac{1}{2j}[\delta(f - f_c) - \delta(f + f_c)]$
$\text{rect}\left(\frac{t}{T}\right)$	$T\,\text{sinc}(fT)$
$\text{sinc}\left(\frac{t}{T}\right)$	$T\,\text{rect}(fT)$
$\sum_{n=-\infty}^{\infty}\delta(t - nT)$	$\frac{1}{T}\sum_{n=-\infty}^{\infty}\delta\left(f - \frac{n}{T}\right)$

Table 1.1: Useful Fourier Transforms.

In the above table we have the following definitions[1]:

$$\text{sinc}(x) \triangleq \frac{\sin(\pi x)}{\pi x} \text{ and}$$

$$\text{rect}(x) \triangleq \begin{cases} 1 & |x| < \frac{1}{2} \\ 0 & \text{elsewhere} \end{cases}$$

1.1 Properties of the Fourier Transform

Here we denote the FTs of $x(t)$ and $y(t)$ by $X(f)$ and $Y(f)$ respectively. Some useful properties of the FT are then as follows.

- Linearity: $\mathcal{F}\{\alpha x(t) + \beta y(t)\} = \alpha X(f) + \beta Y(f)$

[1]Note that some texts use an alternative definition $\text{sinc}(x) \triangleq \frac{\sin(x)}{x}$, but we will NOT be using that here unless otherwise stated.

- Time scaling: if $a \in \mathbb{R}$, $a \neq 0$,

$$\mathcal{F}\{x(at)\} = \frac{1}{|a|} X\left(\frac{f}{a}\right)$$

- Duality: if $\mathcal{F}\{g(t)\} = G(f)$, then $\mathcal{F}\{G(t)\} = g(-f)$

- Multiplication in the time domain corresponds to convolution in the frequency domain

$$\mathcal{F}\{x(t)\, y(t)\} = \int_{-\infty}^{\infty} X(\nu)\, Y(f-\nu)\, d\nu$$

- Multiplication in the frequency domain corresponds to convolution in the time domain

$$\mathcal{F}^{-1}\{X(f)\, Y(f)\} = \int_{-\infty}^{\infty} x(\tau)\, y(t-\tau)\, d\tau$$

- Time shift:

$$\mathcal{F}\{x(t-t_0)\} = e^{-j2\pi f t_0} X(f)$$

- Frequency shift:

$$\mathcal{F}\left\{e^{j2\pi f_c t} x(t)\right\} = X(f-f_c) \tag{1.1.1}$$

- 'Rayleigh's energy theorem' (or 'Parseval's Theorem')

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

1.2 Fourier Transform Symmetry Properties

$$
\begin{aligned}
g(t) \text{ real} &\implies G(-f) = G^*(f) \\
g(t) \text{ even} &\implies G(-f) = G(f) \\
g(t) \text{ odd} &\implies G(-f) = -G(f) \\
g(t) \text{ real and even} &\implies G(f) \text{ real and even} \\
g(t) \text{ real and odd} &\implies G(f) \text{ imaginary and odd}
\end{aligned}
$$

Note in particular that if $g(t)$ is real, then $|G(f)|$ has even symmetry, and $\angle G(f)$ has odd symmetry.

Chapter 2

Energy Signals and Power Signals

Let $x(t)$ be a real signal. The instantaneous power in $x(t)$ is given by $x^2(t)$. An *energy signal* is a real signal which has nonzero but finite energy, i.e.

$$E_x = \int_{-\infty}^{\infty} x^2(t)\, dt \qquad 0 < E_x < \infty$$

A power signal is a real signal which has nonzero but finite average power, i.e.

$$P_x = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)\, dt \qquad 0 < P_x < \infty$$

Note that power signals have infinite energy, and energy signals have zero average power. Therefore, a signal can be an energy signal or a power signal, but not both.

2.1 Real Energy Signals

The autocorrelation function of a real energy signal is given by

$$R_x(\tau) = \int_{-\infty}^{\infty} x(t)\, x(t+\tau)\, dt$$

This function provides an indication of how closely related $x(t)$ is to a time-shifted version of itself. The autocorrelation function has the properties:

1. $R_x(0) = E_x$ is the energy in the signal
2. $R_x(\tau)$ is an even function, i.e. $R_x(-\tau) = R_x(\tau)$

For a real energy signal $x(t)$, the *energy spectral density* (ESD) is given by

$$S_x(f) = |X(f)|^2$$

This function has the properties:

1. $S_x(f)$ is real
2. $S_x(f)$ is an even function, i.e. $S_x(-f) = S_x(f)$

It may be shown (proof omitted) that the autocorrelation function and the energy spectral density are a Fourier transform pair, i.e.

$$\begin{aligned} S_x(f) &= \int_{-\infty}^{\infty} R_x(\tau)\, e^{-j2\pi f\tau}\, d\tau \\ R_x(\tau) &= \int_{-\infty}^{\infty} S_x(f)\, e^{j2\pi f\tau}\, df \end{aligned}$$

This is known as the *Wiener-Khintchine theorem*.

2.2 Real Power Signals

The autocorrelation function of a real power signal is given by

$$R_x(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)\, x(t-\tau)\, dt$$

This function has the properties:

1. $R_x(0) = P_x$ is the power in the signal
2. $R_x(\tau)$ is an even function, i.e. $R_x(-\tau) = R_x(\tau)$

For a real power signal $x(t)$, the *power spectral density* (PSD) is given by

$$S_x(f) = \lim_{T\to\infty} \frac{1}{T} |X_T(f)|^2$$

where $X_T(f)$ is the FT of the truncated time signal

$$x_T(t) = \begin{cases} x(t) & \text{if } -\frac{T}{2} < t < \frac{T}{2} \\ 0 & \text{otherwise} \end{cases}$$

It may also be shown (proof omitted) that the autocorrelation function and the power spectral density are a Fourier transform pair, i.e.

$$\begin{aligned} S_x(f) &= \int_{-\infty}^{\infty} R_x(\tau)\, e^{-j2\pi f\tau}\, d\tau \\ R_x(\tau) &= \int_{-\infty}^{\infty} S_x(f)\, e^{j2\pi f\tau}\, df \end{aligned}$$

This is known as the *Wiener-Khintchine theorem*.

Chapter 3

Baseband and Bandpass Signals

3.1 Baseband signal

A *baseband* signal is one whose frequency content, or spectrum, is close to zero frequency (DC). A *bandpass* signal is one whose frequency content, or spectrum, is centered on some high frequency.

Let $x\left(t\right)$ be a baseband signal, with frequency content extending only up to $f = W$ Hz, i.e. $X\left(f\right) = 0$ for $|f| > W$.

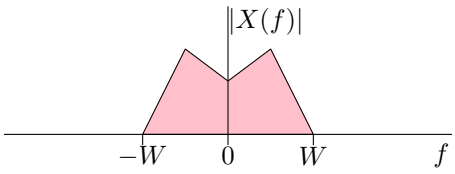


Figure 3.1.1: Spectrum of $x\left(t\right)$.

3.2 Bandpass signal

Suppose we create

$$\tilde{x}\left(t\right) = x\left(t\right)\cos\left(2\pi f_c t\right)$$

The signal is said to be frequency shifted, or *heterodyned*, up to center frequency f_c. By writing

$$\tilde{x}\left(t\right) = x\left(t\right) \cdot \frac{1}{2}\left(e^{j2\pi f_c t} + e^{-j2\pi f_c t}\right)$$

we obtain the Fourier transform of the signal $\tilde{x}(t)$ as

$$\tilde{X}(f) = \frac{1}{2}\left[X(f - f_c) + X(f + f_c)\right]$$

Therefore, $\tilde{x}(t)$ is a bandpass signal, with frequency spectrum of bandwidth $2W$ centered on f_c Hz, as shown in Figure 3.2.1

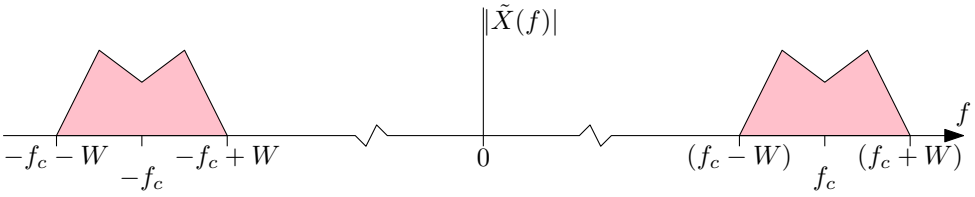


Figure 3.2.1: Spectrum of $\tilde{x}(t) = x(t)\cos(2\pi f_c t)$.

3.3 Quadrature signal

Similarly, let $y(t)$ be another baseband signal, with frequency content extending only up to $f = W$ Hz, i.e. $Y(f) = 0$ for $|f| > W$.

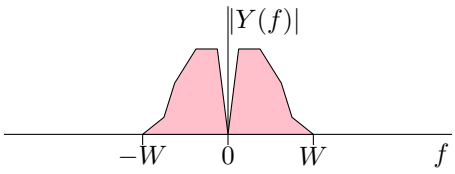


Figure 3.3.1: Spectrum of $y(t)$.

If we create

$$\tilde{y}(t) = y(t)\sin(2\pi f_c t)$$

Then, writing

$$\tilde{y}(t) = y(t) \cdot \frac{1}{2j}\left(e^{j2\pi f_c t} - e^{-j2\pi f_c t}\right)$$

we obtain the Fourier transform of $\tilde{y}(t)$ as

$$\tilde{Y}(f) = \frac{1}{2j}\left[Y(f - f_c) - Y(f + f_c)\right]$$

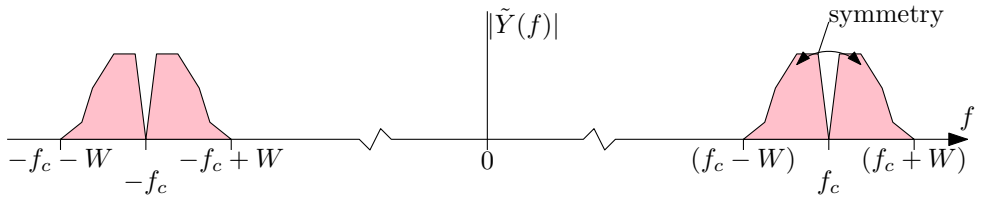


Figure 3.3.2: Spectrum of $\tilde{y}(t) = y(t)\sin(2\pi f_c t)$.

Therefore, $\tilde{y}(t)$ is also a bandpass signal, with frequency spectrum of bandwidth $2W$ centered on f_c Hz.

To create a quadrature signal we can combine $\tilde{x}(t)$ and $\tilde{y}(t)$ to obtain a more general bandpass signal $\tilde{s}(t) = \tilde{x}(t) - \tilde{y}(t)$, or

$$\tilde{s}(t) = x(t)\cos(2\pi f_c t) - y(t)\sin(2\pi f_c t) \tag{3.3.1}$$

This signal $\tilde{s}(t)$ is a bandpass signal, with frequency spectrum of bandwidth $2W$ centered on f_c Hz.

$x(t)$ and $y(t)$ are called the *quadrature components* of the signal $\tilde{s}(t)$.

The two sinusoidal signals $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$ have $90°$ phase difference and are therefore said to be in *phase quadrature* , i.e. one *quarter* of a whole cycle,with each other; hence the name “quadrature components”. The spectrum of the bandpass signal $s(t)$ is

$$\tilde{S}(f) = \frac{1}{2}\left[X(f - f_c) + X(f + f_c)\right] - \frac{1}{2j}\left[Y(f - f_c) - Y(f + f_c)\right] \tag{3.3.2}$$

This is illustrated in Figure 3.3.3.

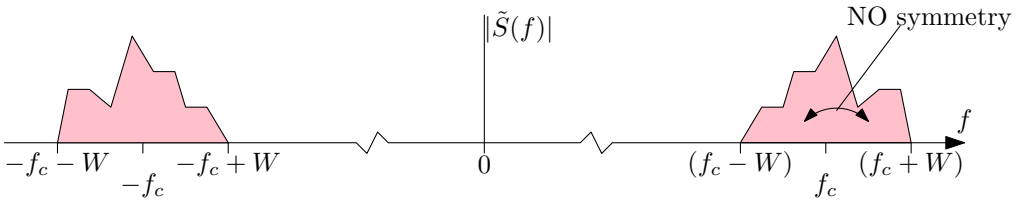


Figure 3.3.3: Spectrum of $\tilde{s}(t)$.

More generally, it may be shown that any bandpass signal with spectrum of bandwidth $2W$ centered on f_c Hz may be written in the form of equation (3.3.1) where $x(t)$ and $y(t)$ are

some pair of baseband signals of bandwidth W.

3.3.1 Why?

We claim that if we receive $\tilde{s}(t)$ then a sufficiently clever receiver can compute both $x(t)$ and $y(t)$, the two quadrature components.
The proof is a not too difficult but a little long...

Exercise:

Prove that this is true...

The importance of this is that we can transmit a signal that occupies the same bandwidth as $\tilde{x}(t)$ (or $\tilde{y}(t)$) but it contains both $\tilde{x}(t)$ and $\tilde{y}(t)$ and therefore twice the amount of information. This is illustrated in Figure 3.3.4 where two audio signals are transmitted at the same time over the same channel using the same carrier frequency, but yet the receiver can separate the two signals correctly.

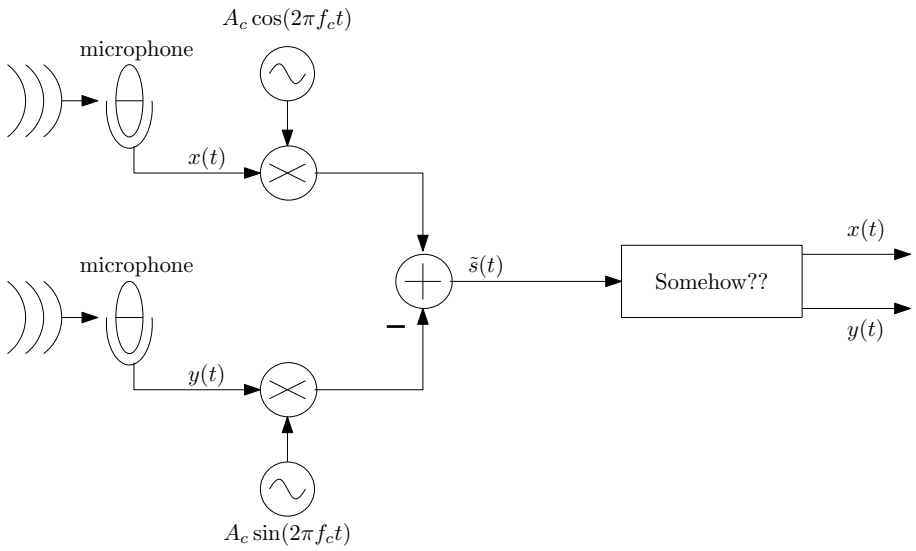


Figure 3.3.4: Two audio signals being transmitted together using the in-phase and quadrature components of a carrier.

3.3.2 Magnitude / phase notation

We may also rewrite equation (3.3.1) in the form

$$\tilde{s}(t) = A(t)\cos(2\pi f_c t + \theta(t)) \tag{3.3.3}$$

$A(t)$ is called the *envelope* of the signal, and $\theta(t)$ is called the *phase* of the signal. Equation (3.3.3) is called the *envelope-phase* representation of the bandpass signal. The envelope and phase are related to the quadrature components via

$$A(t) = \sqrt{x^2(t) + y^2(t)}; \qquad \theta(t) = \tan^{-1}\left(\frac{y(t)}{x(t)}\right)$$

or

$$x(t) = A(t)\cos\theta(t); \qquad y(t) = A(t)\sin\theta(t)$$

These relationships are illustrated graphically in figure 3.3.5.

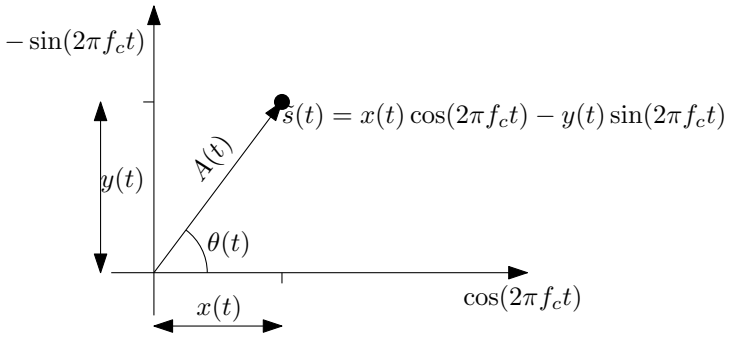


Figure 3.3.5: Illustration of the relationship between the quadrature components and the envelope and phase components of a narrowband signal.

Chapter 4

Modulation Concept

Usually the information signal $g(t)$, e.g. a voice or video signal, we wish to convey over a medium is baseband (with bandwidth W).
However many channels, most notably the wireless channel, are bandpass systems and are not directly suitable for conveying our baseband information signal (a metallic conductor is a baseband channel that would be suitable).

The engineering problem is then to transmit the base-band information signal, $g(t)$, over a bandpass channel; to achieve this we use modulation.

Modulation is the process of varying some characteristic of a sinusoidal carrier signal $c(t)$ according to the value of our information signal $g(t)$.

Consider a high-frequency sinusoid (generated by an oscillator at the transmitter), called a *carrier*, having frequency f_c that is suitable for the bandpass channel:

$$c(t) = A_c \cos(2\pi f_c t)$$

The resulting band-pass *modulated signal* is:

$$\begin{aligned} \tilde{s}(t) &= x(t)\cos(2\pi f_c t) - y(t)\sin(2\pi f_c t) \\ &\text{or} \\ \tilde{s}(t) &= A(t)\cos(2\pi f_c t + \theta(t)) \end{aligned}$$

has a narrow bandwidth centered on f_c, and therefore *can* pass through the channel.

For now we will focus on the second format.

The information signal $g(t)$ is used to generate $A(t)$ and / or $\theta(t)$.

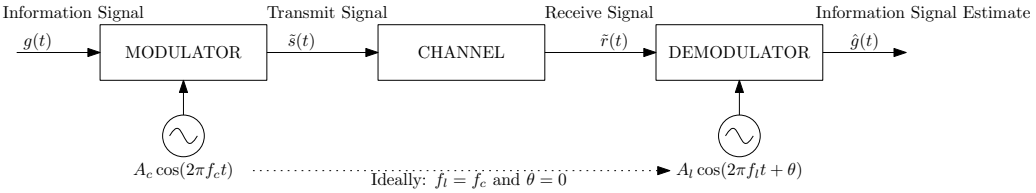


Figure 4.0.1: Bandpass Analog Communication System.

- The carrier signal, $c(t) = A_c \cos(2\pi f_c t)$, 'carries' the information at high frequency
- An *oscillator* is required to generate the carrier having frequency f_c
- There are 3 main types of modulation:
  - Amplitude modulation (AM) - $g(t)$ varies the amplitude $A(t)$
  - Angle modulation - really there are two sub-types here
    - Phase modulation (PM) - $g(t)$ directly varies the angle $\theta(t)$
    - Frequency modulation (FM) - $g(t)$ varies the frequency $\frac{d}{dt}\theta(t)$
- *Demodulation* refers to the process of undoing the modulation, i.e. mapping the received signal $\tilde{r}(t)$ back to an estimate of the information signal, $\hat{g}(t)$. As we shall see, this process often (but not always) requires an oscillator at the same frequency (f_c) as that at the transmitter - called a *local oscillator* (LO). In systems which use a LO, it is required that the transmitter and receiver oscillators be kept at the same frequency and phase as each other – this process is called *synchronisation*. This is not a trivial task since the transmitter and receiver are usually separated by a large distance.

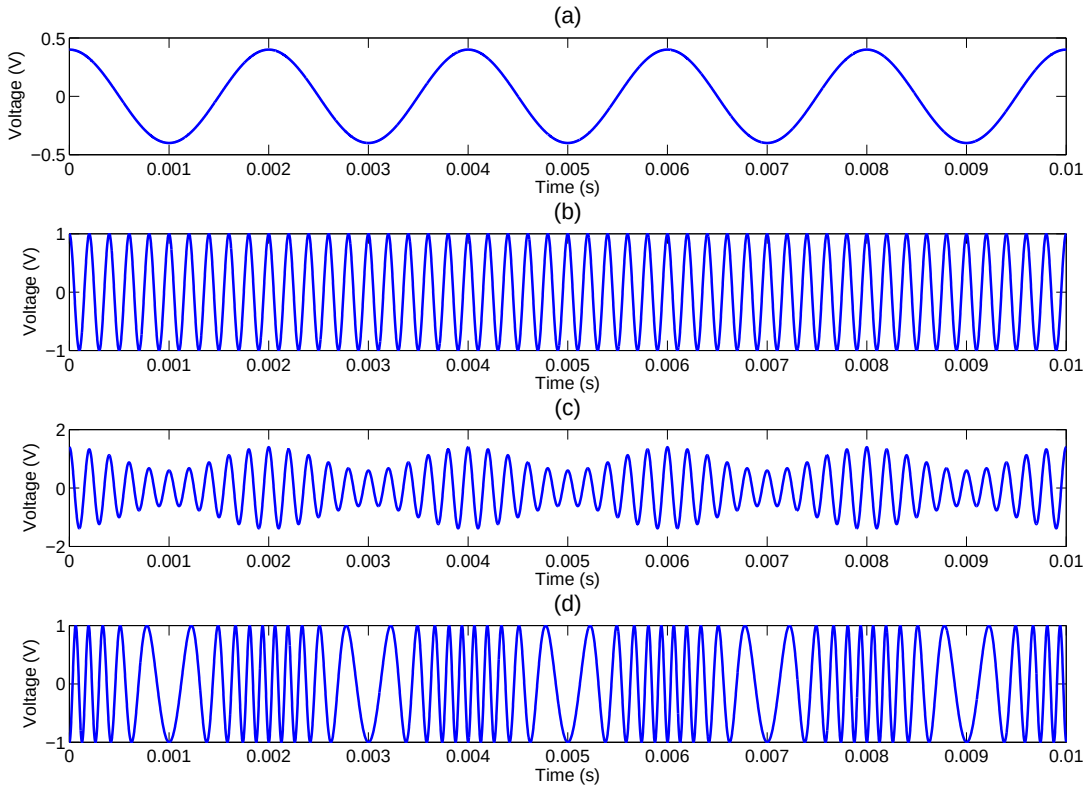


Figure 4.0.2: Illustration of amplitude and frequency modulation. (a) shows the message signal; (b) shows the carrier, a high-frequency sinusoid; (c) shows the AM signal; (d) shows the FM signal.

Modulation is performed for a number of reasons:

- To move information to a particular part of the frequency spectrum, to suit the available channel
- To put information in a better form to survive channel impairments
- To *multiplex,* i.e. to transmit many information signals on the same channel.

Chapter 5

Amplitude Modulation ('Full AM')

In amplitude modulation (AM) the amplitude of the carrier, $A(t)$, is a linear function of the information signal $g(t)$, i.e. the amplitude (not the frequency nor the phase) is modulated hence the name Amplitude Modulation (AM). Mathematically we have:

$$\begin{aligned} \tilde{s}(t) &= A(t)\cos(2\pi f_c t + \theta(t)) \\ &= A_c[1 + k_a g(t)]\cos(2\pi f_c t) \end{aligned} \tag{5.0.1}$$

where we have assumed that $\theta(t) = 0$ for all t, i.e. neither the frequency not the phase are modulated here[1]. An example is shown in Figure 5.0.1.

[1] We are also assuming that the initial phase is zero; this can always be true by appropriate choice of when you decide $t = 0$ is.

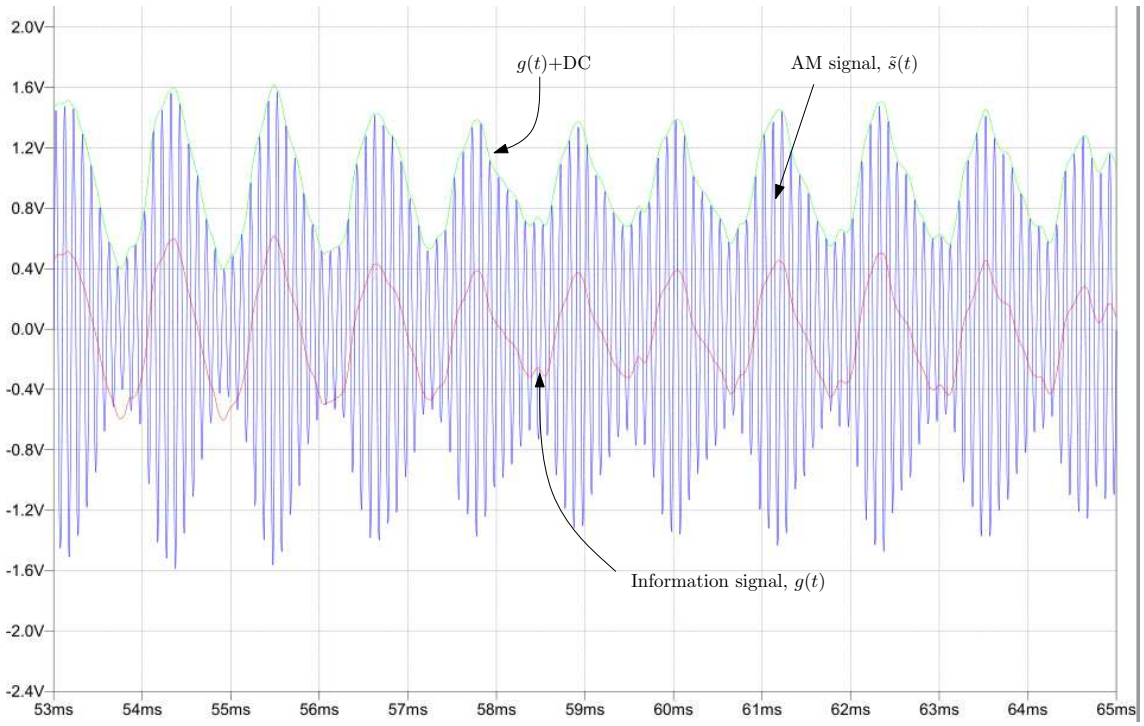


Figure 5.0.1: Illustration of an AM signal.

Definition. Envelope: Something that envelops; a wrapping (from www.thefreedictionary.com).

By looking at Figure 5.0.1 we can see that the amplitude (or envelope), $A(t)$, of the signal varies with time in proportion to $g(t)$.
It should be clear that A_c, the amplitude of the carrier before modulation, now becomes the nominal value of the amplitude after modulation, i.e. the value around which the amplitude varies.

- When $g(t) > 0$, then the envelope is $> A_c$
- When $g(t) < 0$, then the envelope is $< A_c$
- The larger $|g(t)|$ the further the envelope is from the nominal value A_c.

The positive constant k_a is called the *amplitude sensitivity*; it controls how much the envelope deviates from A_c in response to a given value of $g(t)$, i.e.:

- If k_a is small $\implies$ the envelope changes a little
- If k_a is large $\implies$ the envelope changes a lot.

Usually we choose k_a such that the envelope is always positive $\implies$ choose k_a so that $1 + k_a g(t) > 0$ for all t, or

$$k_a g(t) > -1 \quad \text{for all } t$$

We assume the information signal $g(t)$ is baseband, and therefore that the modulation is *slow*, relative to the carrier frequency. We may also write (5.0.1) as

$$\tilde{s}(t) = A(t)\cos(2\pi f_c t)$$

where $A(t) = A_c[1 + k_a g(t)]$ is called the *envelope* of the AM signal (we can think of this as an imaginary curve joining the AM signal peaks - see figure 5.0.1).

5.1 Spectrum

The information signal $g(t)$ is assumed to be a baseband signal with bandwidth W, as per Figure 5.1.1.

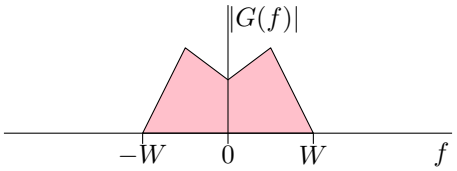


Figure 5.1.1: Information signal spectrum (baseband). Note that real signals have conjugate symmetry in the frequency domain.

Assume $W \ll f_c$, so the modulation is *slow*, relative to the carrier frequency.
The AM signal is:

$$\begin{aligned} \tilde{s}(t) &= A_c[1 + k_a g(t)]\cos(2\pi f_c t) \\ &= A_c\cos(2\pi f_c t) + A_c k_a g(t)\cos(2\pi f_c t) \end{aligned}$$

The first term is just the carrier itslef, the second term contains the information.
Let the Fourier transform of $g(t)$ be $G(f)$, thus the Fourier transform of the AM signal isL

$$\tilde{S}(f) = \frac{A_c}{2}[\delta(f - f_c) + \delta(f + f_c)] + \frac{A_c k_a}{2}[G(f - f_c) + G(f + f_c)] \tag{5.1.1}$$

This is illustrated in Figure 5.1.2.

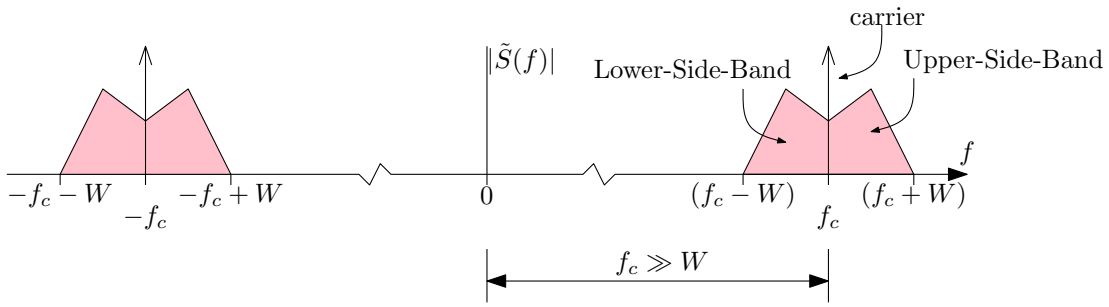


Figure 5.1.2: AM signal spectrum.

We can see the spectrum of the AM signal contains a carrier component (at f_c) and two symmetrical sidebands, called the upper sideband (USB) and lower sideband (LSB).

5.1.1 Example - Sinusoidal Modulating Signal

In analyzing and testing modulation systems, it is often convenient to use a sinusoidal test signal, in place of the information signal - it also makes the math easier!
Real information signals are never sinusoidal - we examine this for illustration only!
Let $g(t) = A_m \cos(2\pi f_m t)$, with frequency $f_m \ll f_c$. Then

$$\begin{aligned} \tilde{s}(t) &= A_c [1 + k_a A_m \cos(2\pi f_m t)] \cos(2\pi f_c t) \\ &= A_c [1 + m. \cos(2\pi f_m t)] \cos(2\pi f_c t) \end{aligned}$$

Where we define $m \triangleq k_a A_m$, the *modulation index,* or depth of modulation.
Clearly, as both k_a and A_m are positive, then $m > 0$, but what about its maximum value?
Well we said previously we should always have $1 + g(t) > 0$, i.e. $1 + m. \cos(2\pi f_m t) > 0$ for all t. But the minimum value of $\cos$ is -1, meaning that we must have $1 - m > 0$, or $m < 1$. i.e. the range of m is $0 < m < 1$ and is thus often expressed as a percentage, for example, we could say the modulation index is 50%, meaning $m = 0.5$.

The maximum value of $A(t) = A_c [1 + m. \cos(2\pi f_m t)]$ is $A_{max} = A_c (1 + m)$.
Similarly, $A_{min} = A_c (1 - m)$.
These values are easy to measure on an oscilloscope, and can be used to calculate the modulation index[2]:

$$m = \frac{A_{max} - A_{min}}{A_{max} + A_{min}} \qquad \text{for } 0 < m < 1$$

[2]This formula is easily calculated by eliminating A_c and solving for m.

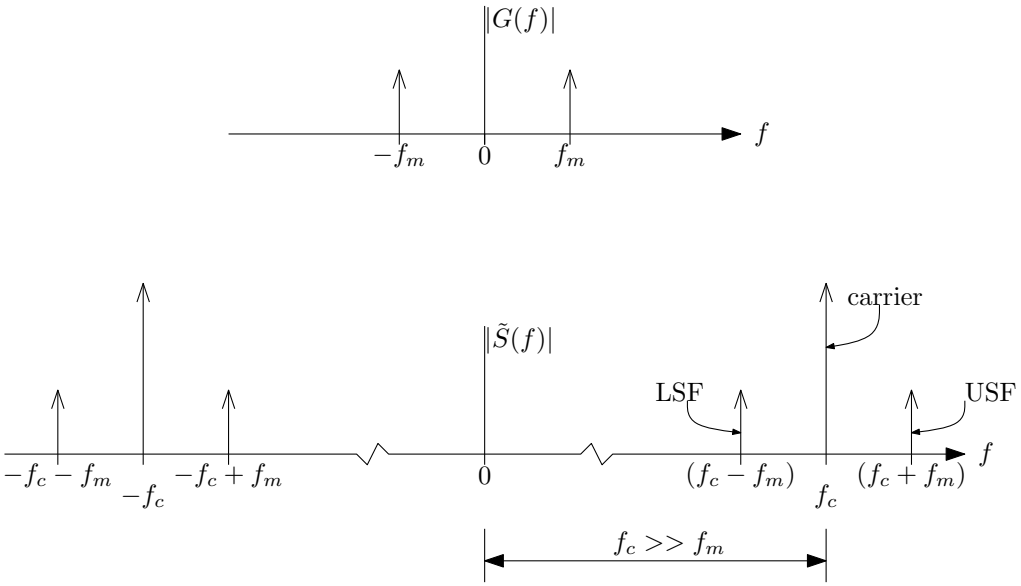


Figure 5.1.3: Spectra of original sinewave and the AM modulated sinewave.

Using the trigonometric identity

$$2\cos A\cos B = \cos\left(A+B\right) + \cos\left(A-B\right)$$

the AM signal can be broken down into three sinusoidal components:

$$\begin{aligned}
\tilde{s}\left(t\right) &= A_c\left[1 + m.\cos\left(2\pi f_m t\right)\right]\cos\left(2\pi f_c t\right) \\
&= A_c\cos\left(2\pi f_c t\right) + A_c m.\cos\left(2\pi f_m t\right)\cos\left(2\pi f_c t\right) \\
&= A_c\cos\left(2\pi f_c t\right) + \frac{A_c}{2}m.\cos\left(2\pi\left\{f_c - f_m\right\}t\right) + \frac{A_c}{2}m.\cos\left(2\pi\left\{f_c + f_m\right\}t\right)
\end{aligned}$$

These components are:

- The carrier, at frequency f_c, amplitude A_c
- A lower side-frequency (LSF), at frequency $f_c - f_m$, amplitude $\frac{A_c}{2}m$
- An upper side-frequency (USF), at frequency $f_c + f_m$, amplitude $\frac{A_c}{2}m$

5.1.2 Problem with DC

Looking at Figure 5.1.2, and also equation 5.1.1, we see that any components of $g\left(t\right)$, the modulating signal, near DC appear in the spectrum of the modulated signal near the carrier and are, from the receiver's point of view indistinguishable from the carrier.

For this reason AM is only ever used to convey signals, $g(t)$, that don't have a DC component!

This is fine for most applications, e.g. voice or music transmission is largely unaffected by this restriction as there is very little useful information close to DC.
We now redraw the AM spectra omitting the DC component from the modulating signal - this is shown in Figure 5.1.4.

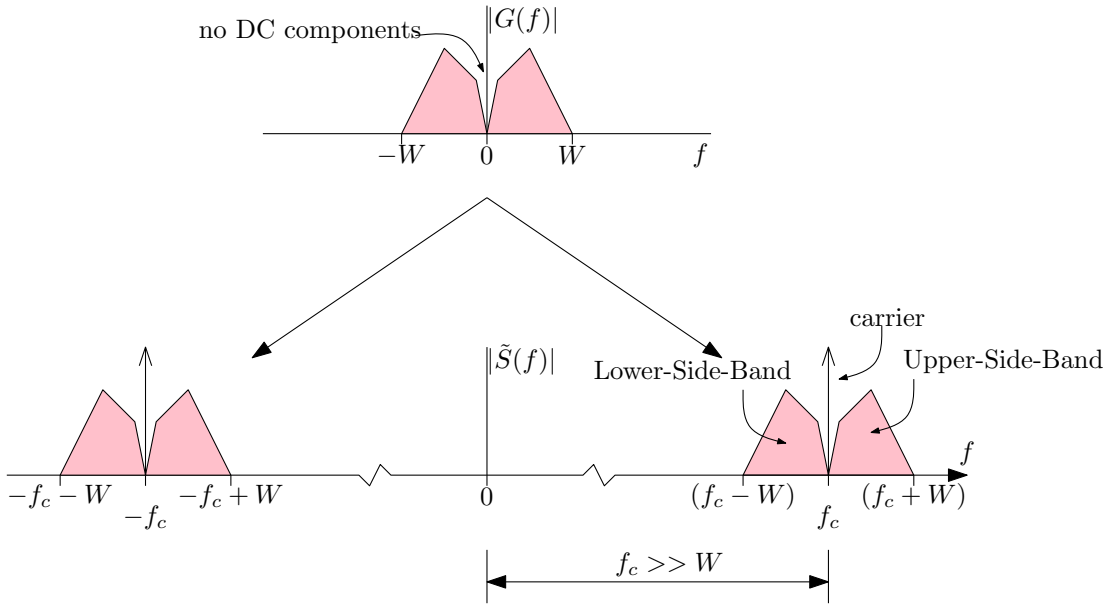


Figure 5.1.4: AM signal spectrum when (as is normal) the modulating signal has no components near DC.

5.2 Power in the AM signal

First we introduce the following defination of power in a signal $x(t)$

$$P_x \triangleq \overline{x^2(t)}$$

i.e. the mean squared.
Now we will apply this to the general case when the modulation signal is any baseband signal $g(t)$, then we will look at the specific case of sinusoid modulation.

5.2.1 General case

So for our AM signal: $\tilde{s}(t) = A(t)\cos(2\pi f_c t)$, we have:

$$\begin{aligned} P_s = \overline{\tilde{s}^2(t)} &= \overline{A^2(t)\cos^2(2\pi f_c t)} \\ &\simeq \left(\overline{A^2(t)}\right)\left(\overline{\cos^2(2\pi f_c t)}\right) \\ &= \frac{1}{2}\overline{A^2(t)} \end{aligned}$$

where the approximation above can be justified if $A(t)$ is semi-static compared to $\cos^2(2\pi f_c t)$, i.e. if $f_c \gg W$.

Now letting $A(t) = A_c[1 + k_a g(t)]$ as before, we get:

$$\begin{aligned} A^2(t) &= A_c^2\left(1 + 2k_a g(t) + k_a^2 g^2(t)\right) \\ \Rightarrow P_s &= \frac{A_c^2}{2}\left(1 + 2k_a\overline{g(t)} + k_a^2 P_g\right) \end{aligned}$$

where $P_g \triangleq \overline{g^2(t)}$ denotes the power in the modulating signal $g(t)$.

If we assume that the mean of $g(t)$ is zero (the information signal has no DC component) then[3]:

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

We have:

- the first term, $\frac{1}{2}A_c^2$, represents the power in the carrier component, and
- the second term represents the power in the sidebands, but since $|k_a g(t)| < 1$, thus $k_a^2 P_g < 1$, meaning that the power in the sidebands is always less than that of the carrier.
- In practical AM applications, we usually have $k_a^2 P_g \ll 1$. Therefore, most of the AM signal power is in the carrier component.
- $\Rightarrow$ full AM is not very power efficient.

[3]This is a common assumption and the justification is discussed in Section 5.1.2.

5.2.2 Sinusoid modulation

For the case of sinusoidal modulation, we have

$$\begin{aligned} g(t) &= A_m \cos(2\pi f_m t) \\ \Rightarrow P_g = \overline{g^2(t)} &= \frac{1}{2} A_m^2 \end{aligned}$$

and therefore:

$$\begin{aligned} P_s &= \frac{A_c^2}{2}\left(1 + k_a^2 \frac{1}{2} A_m^2\right) \\ &= \frac{A_c^2}{2}\left(1 + \frac{1}{2} m^2\right) \end{aligned}$$

From this we see that even at the maximum depth of modulation ($m = 1$), we have 66.7% of the power is in the carrier component - what a waste!

5.3 Implementation of AM

5.3.1 Modulation (the transmitter)

Two methods of implementing the AM modulator are shown in figures 5.3.1 and 5.3.2. The first is based on the formula

$$\tilde{s}(t) = [1 + k_a g(t)]\, c(t)$$

where $c(t) \triangleq A_c \cos(2\pi f_c t)$ is the carrier.
The second method is based on the formula

$$\tilde{s}(t) = c(t) + k_a g(t)\, c(t)$$

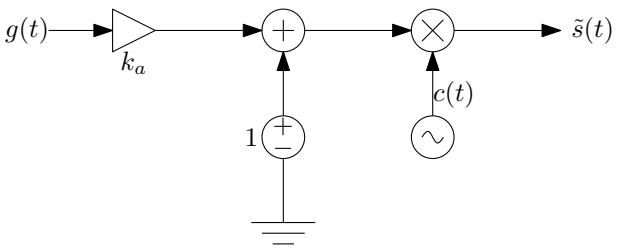


Figure 5.3.1: AM modulator - method 1.

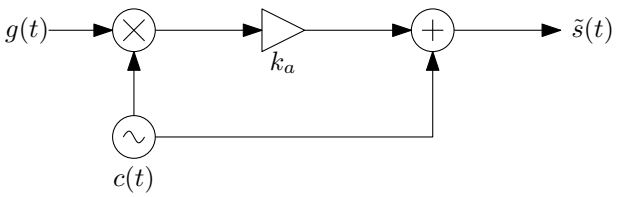


Figure 5.3.2: AM modulator - method 2.

Note that the multiplier in figure 5.3.1 needs to multiply two signals, one of which is always positive. A multiplier which performs this task is called a *two-quadrant multiplier*. On the other hand, the multiplier in figure 5.3.2 needs to multiply two signals, both of which may be positive or negative. A multiplier which performs this task is called a *four-quadrant multiplier*. Generally, a four-quadrant multiplier is harder to build than a two-quadrant multiplier.

5.3.2 Demodulation (the receiver)

Assume that the channel induces an amplitude and phase change in the signal. Let the received signal be

$$\tilde{r}(t) = A_r (1 + k_a g(t)) \cos(2\pi f_c t + \theta)$$

Where A_r is the "new" amplitude of the signal at the channel output, and θ is the phase change caused by the channel.

There are two principal methods of demodulating the AM signal, i.e. retrieving the information signal $g(t)$:

- the Synchronous demodulator
- the envelope detector

5.3.2.1 Synchronous Demodulator

This method requires a local oscillator (LO) whose frequency and phase match the received carrier, i.e.

$$l(t) = A_l \cos(2\pi f_c t + \theta)$$

If we multiply the received signal by the local oscillator as shown in Figure 5.3.3, we get

$$\begin{aligned}\tilde{r}(t)\, l(t) &= A_r A_l (1 + k_a g(t)) \cos^2(2\pi f_c t + \theta) \\ &= \frac{A_r A_l}{2} (1 + k_a g(t)) (1 + \cos(4\pi f_c t + 2\theta)) \\ &= \frac{A_r A_l}{2} (1 + k_a g(t) + \cos(4\pi f_c t + 2\theta) + k_a g(t) \cos(4\pi f_c t + 2\theta))\end{aligned}$$

This is a sum of 4 terms:

- A DC term
- The wanted signal proportional to $g(t)$
- A tone at $2f_c$ Hertz
- $g(t)$ modulated to $2f_c$ Hertz

A suitable lowpass filter (LPF) can remove the two terms around $2f_c$ and the DC term can be removed by a DC block circuit, leaving just the term proportional to $g(t)$ as shown in Figure 5.3.3.

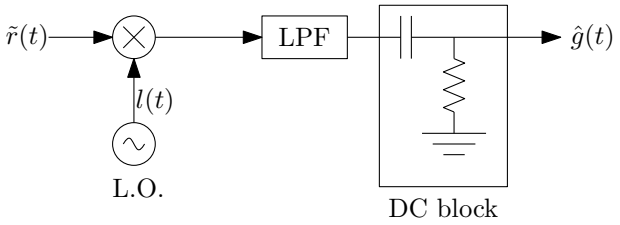


Figure 5.3.3: Synchronous demodulator for AM signal.

The output is:

$$\begin{aligned}\hat{g}(t) &= \left(\frac{A_r A_l k_a}{2}\right) g(t) \\ &\propto g(t)\end{aligned}$$

and the information signal is retrieved intact.

Note that it is not feasible to keep an independent local oscillator in phase with the carrier for any reasonable time – it would require impossible frequency accuracy. However, the AM signal contains a strong component at the carrier frequency. This can be extracted by a high-Q bandpass filter and used in place of the local oscillator. This is shown in figure 5.3.4.

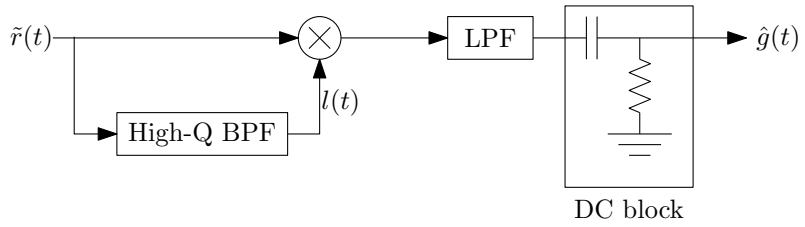


Figure 5.3.4: Derivation of the local oscillator from the received AM signal.

5.3.2.2 Envelope Detector

The envelope detector, as shown in Figure 5.3.5, is designed to follow the envelope (or the carrier peaks) of the AM signal. This is a simple method of demodulation, and is used in most AM radio receivers. It does not require a local oscillator, and is not sensitive to carrier frequency or phase.

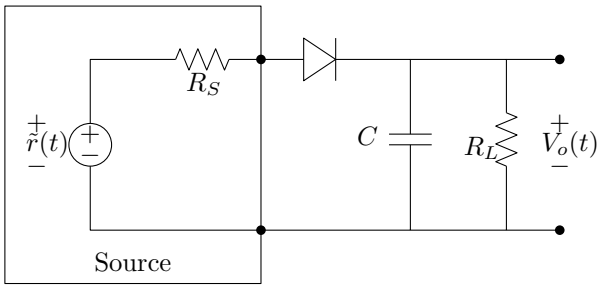


Figure 5.3.5: Envelope detector.

The resistance R_S represents the internal resistance of the signal source, and also includes the effective resistance of the diode, which is otherwise assumed ideal.

When designing such a circuit we need to have rules to help, we will look at a few (see Figure 5.3.6 to help your understanding of these points):

1. When the input voltage is greater than the capacitor voltage, the diode is forward biased, and the capacitor charges towards the input voltage, with time constant $\tau_1 = (R_S \parallel R_L)\, C \approx R_S C$.

   τ_1 must be *small*, relative to the carrier period, so that the capacitor can charge to each positive peak of the carrier.

$$R_S C \;\ll\; \frac{1}{f_c}$$

2. When the input voltage is smaller than the capacitor voltage, the diode is reverse biased, and the capacitor discharges through R_L, with time constant $\tau_2 = R_L C$.

τ_2 must be *large*, relative to the carrier period, so that the capacitor voltage only falls slightly before it is re-charged on the next positive peak.

$$R_L C \gg \frac{1}{f_c}$$

3. When the source signal amplitude increases, due to the modulation, the capacitor voltage follows. It must also be able to follow decreases in amplitude due to the modulation.

   This requires the discharge time constant τ_2 to be *small*, relative to the modulating signal $g(t)$.

$$R_L C \ll \frac{1}{W}$$

   where W is the highest frequency component in the modulating signal $g(t)$.

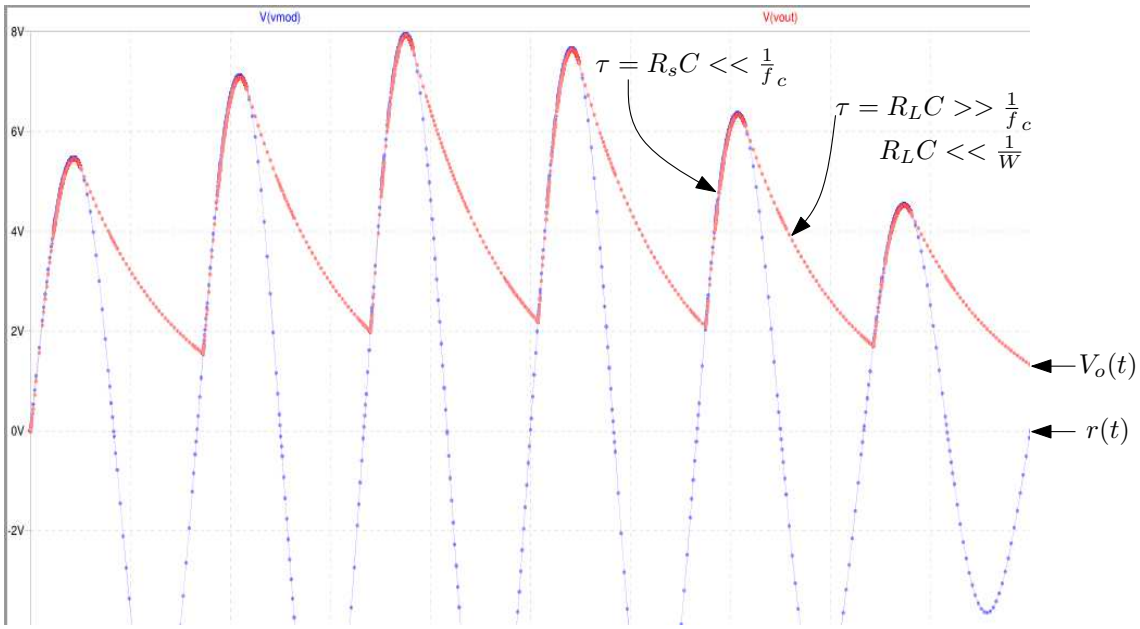


Figure 5.3.6: Output of envelope detector. The scale is exaggerated for illustration purposes.

In summary, we require:

$$R_S C \ll \frac{1}{f_c} \ll R_L C \ll \frac{1}{W}$$

where W represents the highest frequency in the information signal $g(t)$.

Chapter 6

Double-Sideband Suppressed Carrier (DSB-SC)

As we have seen, a full AM signal, given by

$$\begin{aligned} \tilde{s}(t) &= [1 + k_a g(t)]\, c(t) \\ &= A_c [1 + k_a g(t)] \cos(2\pi f_c t) \end{aligned} \tag{6.0.1}$$

consists of a carrier component and two sidebands.

Most of the power is in an unmodulated carrier component that carries no information.

Double-Sideband Suppressed-Carrier (DSB-SC) is a form of amplitude modulation where the two sidebands are transmitted, but with no carrier component.

This gives a large saving in power, compared to full AM, but uses the same bandwidth and conveys the same information. It is therefore said to be more a more power efficient modulation scheme compared to full AM.

Subtracting the carrier part from 6.0.1, we have the DSB-SC modulated signal:

$$\begin{aligned} \tilde{s}(t) &= [1 + k_a g(t)]\, c(t) - c(t) \\ &= k_a g(t)\, c(t) \end{aligned}$$

The constant k_a is now redundant (can just be merged into A_c, the amplitude of the carrier to be modulated $c(t)$), so we usually write

$$\begin{aligned}\tilde{s}(t) &= g(t)\,c(t)\\ &= A_c g(t)\cos(2\pi f_c t)\end{aligned}$$

A simplistic DSB-SC modulator is shown in Figure 6.0.1.

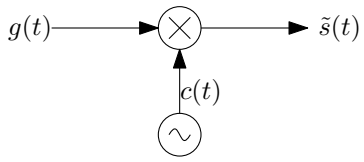


Figure 6.0.1: A DSB-SC modulator is simply a four-quadrant multiplier.

DSB-SC AM is just a multiplication of the carrier and the information signal.

Unlike full AM, the envelope can now become negative, causing a phase inversion in the modulated signal as can be seen in Figure 6.0.2.

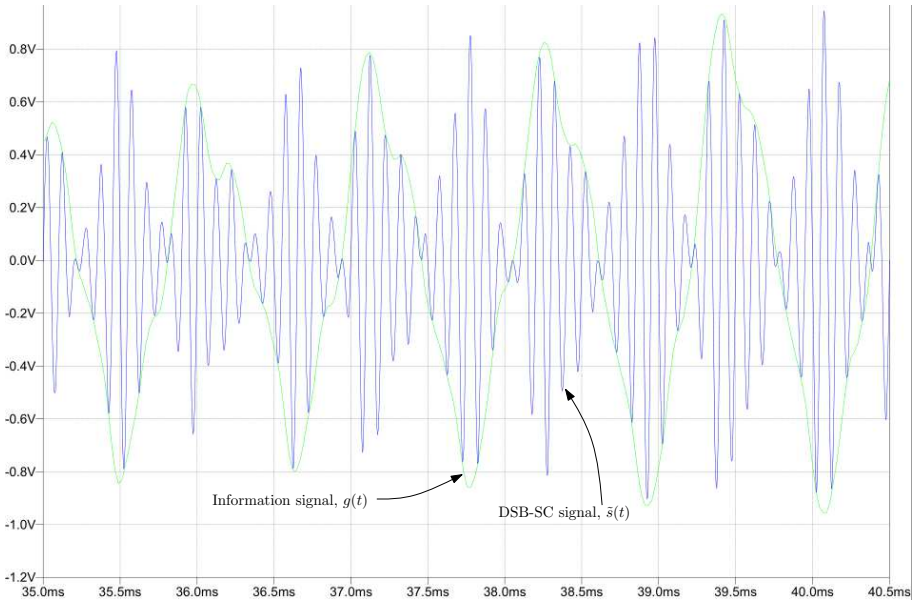


Figure 6.0.2: Illustration of a DSB-SC signal.

6.0.3 Spectrum of DSB-SC

Removing the carrier terms from equation 5.1.1[1] we get:

$$\tilde{S}(f) = \frac{A_c}{2}\left[G(f-f_c)+G(f+f_c)\right]$$

This is shown in Figure 6.0.3. Again we've drawn the spectrum assuming that $g(t)$ has no frequency components near DC, although this is strictly not needed anymore.

6.1 Implementation of DSB-SC

6.1.1 Modulation

The modulator for DSB-SC consists simply of a four-quadrant multiplier.

A popular method of realizing such a multiplier, called a "ring modulator", is shown in Figure 6.1.1.

[1] Just remove the two delta functions, and note that we've merged $k_a A_c$ into a single constant A_c.

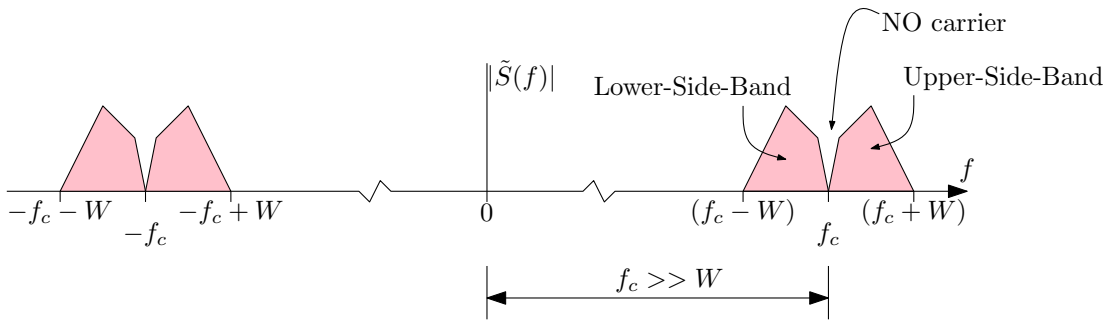


Figure 6.0.3: DSB-SC signal spectrum.

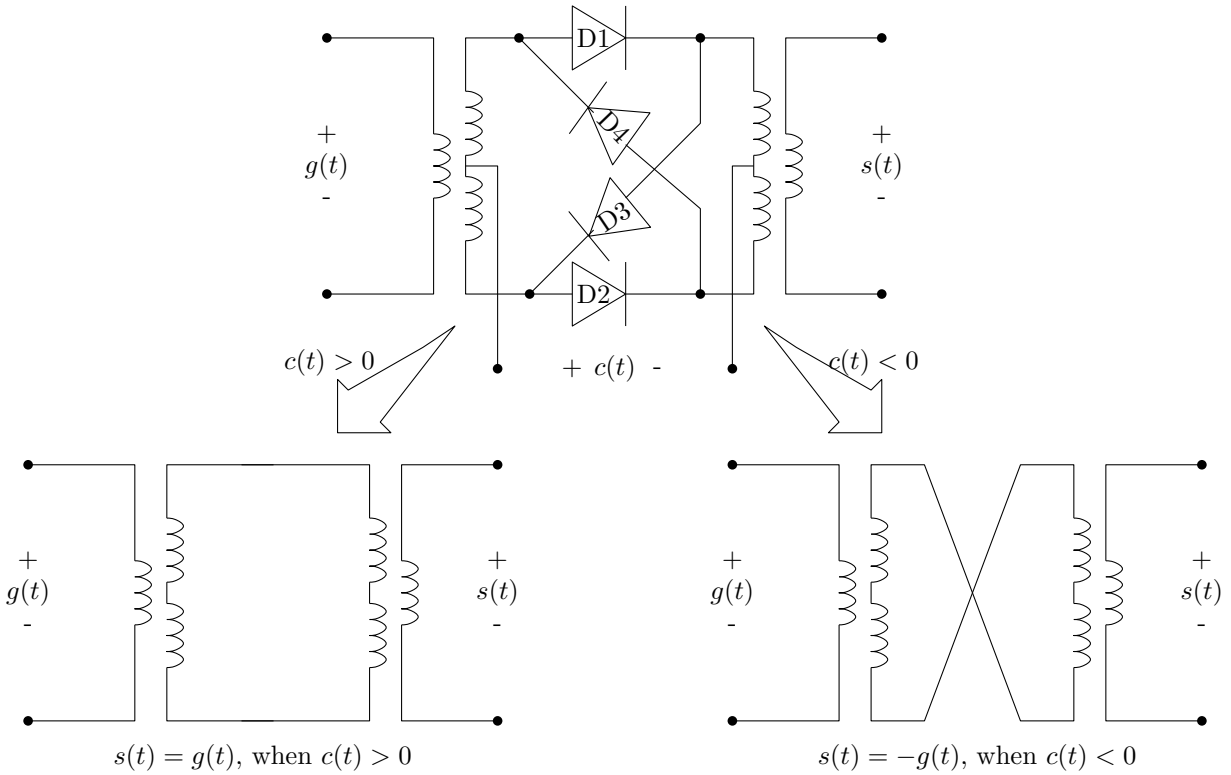


Figure 6.1.1: A "ring modulator", which performs approximate four-quadrant multiplication.

To show that this circuit indeed functions as a four-quadrant multiplier, we shall assume for ease of analysis that the diodes D1, D2, D3 and D4 are ideal.

Ignore the the presence of the carrier input, $c\,(t)$, for the moment and consider the two case:

1. D1 and D2 are ON, D3 and D4 are OFF:
   In this case the input is coupled in on the left, and coupled straight back out on the right, i.e. the output $kg\,(t)$, where k is a constant.

2. D3 and D4 are ON, D1 and D2 are OFF:
   In this case the input is coupled in on the left, and an inverted version is coupled out on the right, i.e. the output $= -kg\,(t)$, where k is a constant

Now consider $c\,(t)$ by itself.
Firstly note that applying a signal, e.g. $c\,(t)$, to the center point of an inductors causes signal cancellation between the upper and lower half resulting in no signal propagation through the inductor, i.e. $c\,(t)$ does not appear at the input nor the output. It does however get presented across the network of 4 diodes. We now note that:

1. When $c\,(t) > 0$, diodes D1 and D2 are ON and diodes D3 and D4 are OFF, and

2. When $c(t) > 0$, diodes D3 and D4 are ON and diodes D1 and D2 are OFF.

But these are exactly the two cases we previously considered, thus:

1. When $c(t) > 0$ we have output $= kg(t)$
2. When $c(t) < 0$ we have output $= -kg(t)$

or more compactly if we define the sign (or signum) function:

$$\operatorname{sgn}(x) = \begin{cases} -1 & \text{for } x < 0 \\ 0 & \text{for } x = 0 \\ +1 & \text{for } x > 0 \end{cases}$$

we have the output expression:

$$\begin{aligned} \tilde{s}(t) &= kg(t)\operatorname{sgn}(c(t)) \\ &= g(t)x(t) \end{aligned}$$

where $x(t)$ is

$$x(t) \triangleq k.\operatorname{sgn}(c(t))$$

This a square wave with frequency f_c, and therefore has Fourier Series:

$$x(t) = A_1\cos(2\pi f_c t) + A_3\cos(6\pi f_c t) + A_5\cos(10\pi f_c t) + \cdots$$

so

$$\tilde{s}(t) = g(t)x(t) = A_1 g(t)\cos(2\pi f_c t) + A_3 g(t)\cos(6\pi f_c t) + A_5 g(t)\cos(10\pi f_c t) + \cdots$$

A bandpass filter is used to select the sidebands around f_c, yielding $A_1 g(t)\cos(2\pi f_c t)$ as required.

6.1.2 Demodulation

As before assume that the channel induces an amplitude and phase change in the signal. Let the received signal be

$$\tilde{r}(t) = A_r g(t) \cos(2\pi f_c t + \theta)$$

Where A_r is the "new" amplitude of the signal at the channel output, and θ is the phase change caused by the channel.

There are two principal methods of demodulating the DCB-SC signal, i.e. retrieving the information signal $g(t)$:

- the Synchronous demodulator
- the envelope detector

6.1.2.1 Synchronous demodulator

The synchronous demodulator, as used for full AM, will work for DSB-SC also. Let the received signal be

$$\tilde{r}(t) = A_r g(t) \cos(2\pi f_c t + \theta)$$

Multiply by a local oscillator $l(t) = A_l \cos(2\pi f_c t + \theta)$. The product is

$$\begin{aligned} \tilde{r}(t) l(t) &= A_r A_l g(t) \cos^2(2\pi f_c t + \theta) \\ &= \frac{A_r A_l}{2} g(t) \{1 + \cos(4\pi f_c t + 2\theta)\} \end{aligned}$$

The LPF output is $\frac{A_r A_l}{2} g(t)$, as required.
Compared to full AM the we see that there is no extra DC component present here meaning that it would be ok for $g(t)$ to have some components close to DC.

However all is not well!

In full AM we were able to generate a local oscillator by extracting the carrier from the received signal using a high-Q BPF - this is not possible anymore! So how do we generate a local oscillator with the correct frequency and phase?, and what exactly is the affect of not getting this exactly right? This is the study of receiver synchronization - and it is a very difficult (and under appreciated) skill.

Phase error:

Consider a phase error in the local oscillator, i.e. $l(t) = A_l \cos(2\pi f_c t + \theta + \phi)$. The product is

$$\begin{aligned} \tilde{r}(t)\, l(t) &= A_r A_l g(t) \cos(2\pi f_c t + \theta) \cos(2\pi f_c t + \theta + \phi) \\ &= \frac{A_r A_l}{2} g(t) \{\cos\phi + \cos(4\pi f_c t + 2\theta + \phi)\} \end{aligned}$$

and the LPF output is $\frac{A_r A_l}{2} g(t) \cos\phi$, i.e. the ideal output multiplied by $\cos\phi$.
This is only acceptable if the phase error ϕ is constant, and not near $\pm\frac{\pi}{2}$.

Frequency error:

Consider a frequency error in the LO: $l(t) = A_l \cos(2\pi (f_c + \Delta f) t + \theta)$. The product is then

$$\begin{aligned} \tilde{r}(t)\, l(t) &= A_r A_l g(t) \cos(2\pi (f_c + \Delta f) t + \theta) \cos(2\pi f_c t + \theta) \\ &= \frac{A_r A_l}{2} g(t) \{\cos(2\pi \Delta f t) + \cos(4\pi f_c t + 2\pi \Delta f t + 2\theta)\} \end{aligned}$$

and the LPF output is $\frac{A_r A_l}{2} g(t) \cos(2\pi \Delta f t)$, i.e. the ideal output multiplied by the time varying (oscillating) term $\cos(2\pi \Delta f t)$.
This is big a problem! and we can not live with this.

6.1.2.2 Envelope detector

The envelope detector may also be used to demodulate the DSB-SC signal.
For this to work, we need the envelope of the signal to be always positive; to arrange this we can add back in a local oscillator signal within the receiver - this acts as an artificial carrier signal allowing the envelope detector to demodulate as normal.
The envelope detector input is (the received signal plus a local oscillator)

$$\begin{aligned} \tilde{r}(t) + l(t) &= A_r g(t) \cos(2\pi f_c t + \theta) + A_l \cos(2\pi f_c t + \theta) \\ &= A_l \left(1 + \frac{A_r}{A_l} g(t)\right) \cos(2\pi f_c t + \theta) \end{aligned}$$

This is a full AM signal.
For a positive envelope, we require $\left|\frac{A_r}{A_l} g(t)\right| < 1$, or $A_l > A_r |g(t)|_{MAX}$.

This gives us a condition on the receiver LO signal amplitude in order to correctly demodulate the DSB-SC signal via the envelope detector.

Chapter 7

SSB and QAM

7.1 Single-Sideband AM

We already seen:

- Full AM: Spectrum contains two side bands and the carrier
- DSB-SC: Spectrum contains two side bands but the carrier is removed

Now the question is why send both side bands? surely they contain the some information? aren't we unnecsarily using valuable spectrum?

Single-Sideband Amplitude Modulation (SSB-AM), as its name suggests, is a form of amplitude modulation where only one sideband is transmitted.
Since each sideband in a DSB-SC (or full AM) signal contains all of the information, either sideband can be used. This gives two forms of SSB:

- USB, where only the Upper SideBand is transmitted
- LSB, where only the Lower SideBand is transmitted.

These are shown in Figure 7.1.1:

Some worthwhile notes:

- The power in the SSB signal is half that of the DSB-SC signal (for the same sideband amplitude)
- This is much less than is needed for full AM.

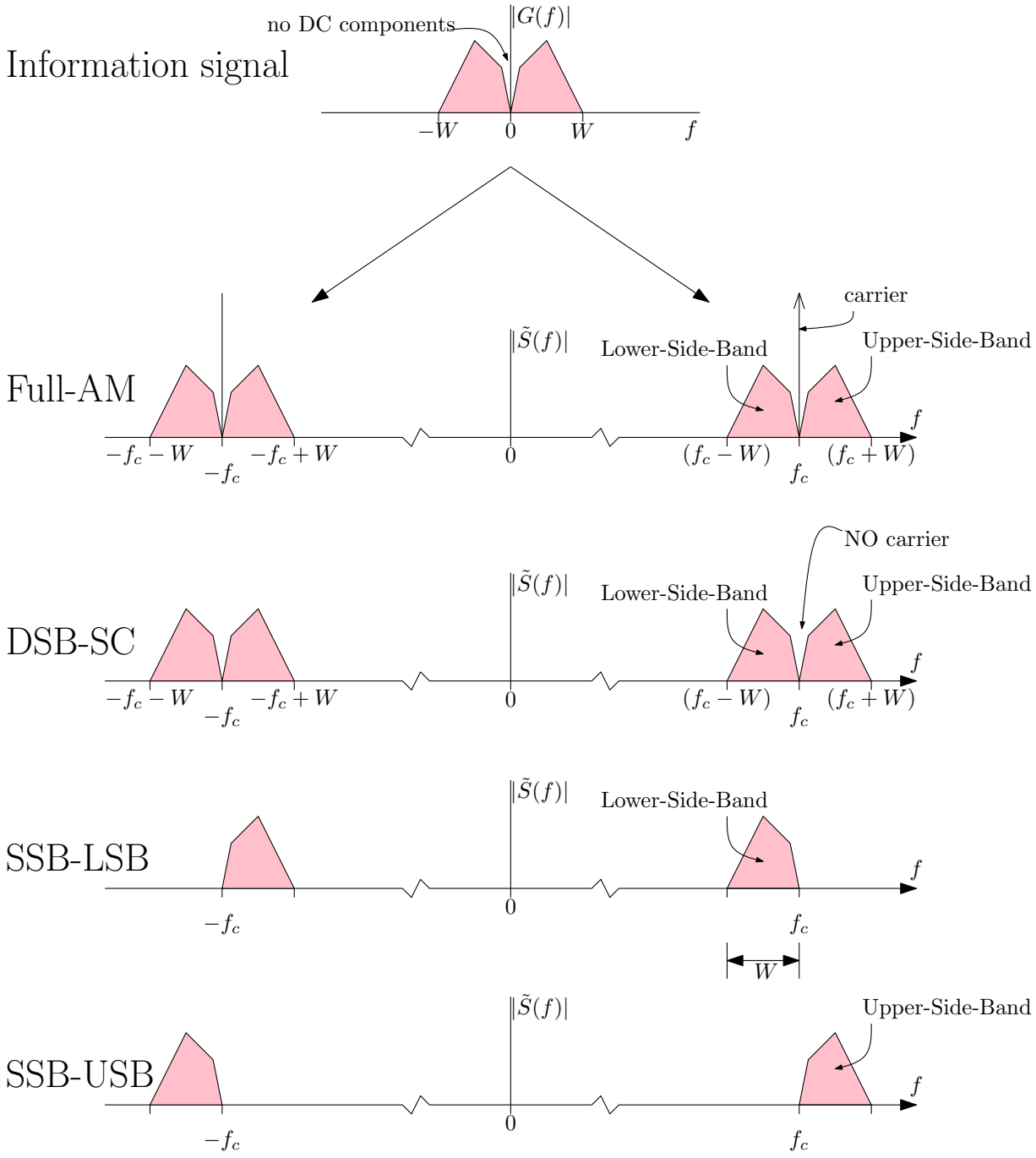


Figure 7.1.1: Comparison of the different types of Amplitude modulation.

- The bandwidth of the SSB signal is equal to the bandwidth of the information signal
- This is half that of the DSB-SC or full AM signal.

Therefore, SSB is the most bandwidth efficient form of amplitude modulation.

7.1.1 Modulation

The simplest method is to generate a DSB-SC signal, then use a filter to remove the unwanted sideband. With a real filter, this method needs a reasonable gap between the sidebands $\implies$ information signal cannot extend to DC.

7.1.2 Demodulation

A synchronous demodulator can be used. It is very difficult to synchronise the local oscillator from the received SSB signal alone, and therefore some systems transmit a small carrier component, called a *pilot signal*, to aid this process.

7.2 Analog Quadrature Amplitude Modulation (QAM)

Recap - see Chapter 3, specifically equation 3.3.2.

Consider two baseband information signals $x(t)$ and $y(t)$, each of bandwidth W.
These may be *jointly* modulated to a Quadrature Amplitude Modulation (QAM) signal via

$$\tilde{s}(t) = A_c\left(x(t)\cos(2\pi f_c t) - y(t)\sin(2\pi f_c t)\right)$$

This QAM signal has bandwidth $2W$, which is equal to that of the information signals combined.
Therefore the bandwidth efficiency of QAM is equal to that of SSB modulation.
The drawbacks of QAM are that it requires *exactly* $90°$ phase difference between the two carriers, and that it also requires *exact* phase synchronisation of the local oscillator.

7.2.1 Modulation

A QAM modulator is shown in figure 7.2.1.
The quadrature carrier is generated from the LO by putting it through a phase shift of $90°$.

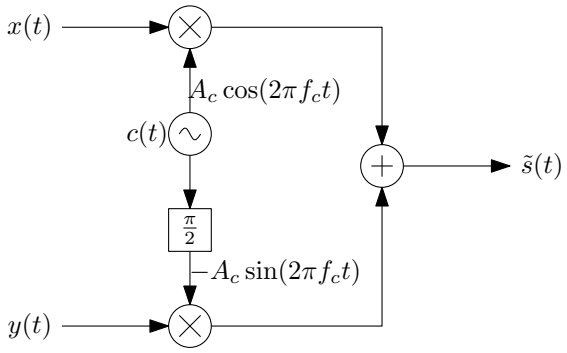


Figure 7.2.1: Analog QAM modulator.

7.2.2 Demodulation

QAM uses a synchronous demodulator, which is shown in Figure 7.2.2.

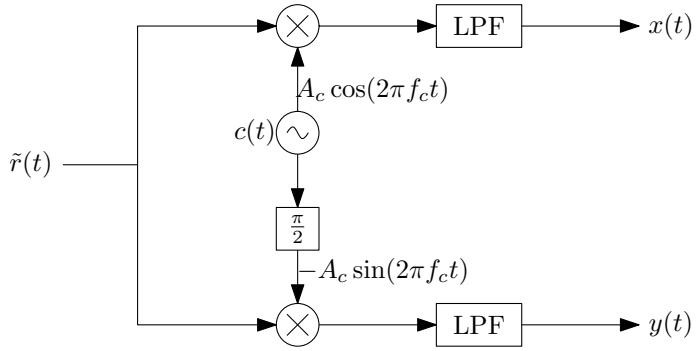


Figure 7.2.2: Synchronous QAM demodulator.

To verify that this circuit works correctly, we must consider the demodulation of $x\left(t\right)$ and $y\left(t\right)$ separately. Let the received signal be

$$\tilde{r}\left(t\right)=A_r\left(x\left(t\right)\cos\left(2\pi f_c t+\theta\right)-y\left(t\right)\sin\left(2\pi f_c t+\theta\right)\right)$$

and assume that the local oscillator (LO) is perfectly synchronized, i.e.

$$l\left(t\right)=A_l\cos\left(2\pi f_c t+\theta\right)$$

Then

$$\begin{aligned}\tilde{r}\left(t\right)l\left(t\right) &= A_r A_l\left[x\left(t\right)\cos^2\left(2\pi f_c t+\theta\right)-y\left(t\right)\cos\left(2\pi f_c t+\theta\right)\sin\left(2\pi f_c t+\theta\right)\right]\\ &= \frac{A_r A_l}{2}\left[x\left(t\right)\left(1+\cos\left(4\pi f_c t+2\theta\right)\right)-y\left(t\right)\sin\left(4\pi f_c t+2\theta\right)\right]\end{aligned}$$

The LPF removes the terms around $2f_c$, giving $x(t)$ as desired.

Exercise: Show that the demodulation of $y(t)$ works correctly, again assuming perfect synchronization of the LO.

Exercise: Examine the effect of a phase error ϕ in the local oscillator.

PM: $A_c \cos(2\pi f_c t + k_p g(t))$

FM: $A_c \cos(2\pi f_c t + 2\pi k_p \int g(t)\,dt)$

Chapter 8

Angle Modulation

Recall that *modulation* is a process where the modulating signal is used to vary some characteristic of a sinusoidal carrier signal. So far, we have only considered amplitude modulation, where the amplitude of the carrier is varied[1].

In angle modulation, the amplitude of the carrier remains constant, while the phase angle of the carrier is varied to convey the information. The modulated signal is of the form

$$\tilde{s}(t) = A_c \cos(2\pi f_c t + \theta(t))$$

where $\theta(t)$ varies to convey the information signal $g(t)$.

As previously, $g(t)$ is assumed to be a baseband signal with bandwidth W.

We can write the angle modulated signal as $\tilde{s}(t) = A_c \cos(\phi(t))$, where $\phi(t)$ is the *phase angle* of the signal.

Then $\phi(t) = 2\pi f_c t + \theta(t)$ is the sum of a linear term $2\pi f_c t$ not containing any information (i.e. not depending on $g(t)$), and $\theta(t)$ which, by construction is a function of $g(t)$.

The linear term, $2\pi f_c t$, can be thought of as the nominal value of the phase of a sinusoidal carrier.

The $\theta(t)$ term is the *phase deviation* $\theta(t)$ from this nominal value.

[1] We previously considered QAM which, as we will see later, modulates the angle also although not directly.

8.1 Phase Modulation (PM)

In phase modulation the phase of the carrier is varied about its nominal value, by an amount proportional to the value of the modulating signal, i.e.

$$\theta(t) = k_p g(t)$$

where the constant k_p is called the *phase sensitivity* of the modulator, and it determines the change in phase angle for a given change in the information signal. The phase modulated signal is then

$$\tilde{s}_p(t) = A_c \cos(2\pi f_c t + k_p g(t))$$

8.2 Frequency Modulation (FM)

This is a process where the *instantaneous frequency* of a sinusoidal carrier is varied about its nominal value, by an amount proportional to the value of the modulating signal.
Before proceeding, we need to define what is meant by the instantaneous frequency of a constant-amplitude signal.
Definition: The *instantaneous frequency* of a constant-amplitude signal $\tilde{s}(t) = A_c \cos(\phi(t))$ is

$$f_i(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{8.2.1}$$

The intuition for this equation is as follows:
The phase of a sinusoid is the rate of change of the angle, e.g. a sinusoid with a constant frequency has a phase that increases linearly with time. The higher the frequency, the higher the slope of the phase V's time graph.
So we have $f_i(t) \propto \frac{d\phi(t)}{dt}$. But this has units of $^{\text{rads}}/_{\text{sec}}$ and we want $f_i(t)$ to be in units of Hertz, a.k.a. $^{\text{cycles}}/_{\text{sec}}$. As there are 2π radians in 1 cycle, we require a $\frac{1}{2\pi}$ scaling factor thus yielding Equation 8.2.1.

We can invert equation (8.2.1) to get

$$\phi(t) = 2\pi \int f_i(t)\, dt$$

For FM we said that we would vary the instantaneous frequency about its nominal value, in

accordance with the modulating signal, i.e.

$$f_i(t) = f_c + k_f g(t)$$

The constant k_f is called the *frequency sensitivity* of the modulator, and it determines the change in instantaneous frequency for a given change in the information signal.

$$\begin{aligned}\tilde{s}(t) = A_c \cos(2\pi f_c t + \theta(t)) &= A_c \cos\left(2\pi \int f_i(t)\ dt\right) \\ &= A_c \cos\left(2\pi \int [f_c + k_f g(t)]\ dt\right) \\ &= A_c \cos\left(2\pi f_c t + 2\pi k_f \int g(t)\ dt\right)\end{aligned}$$

The frequency modulated signal is

$$\tilde{s}_f(t) = A_c \cos\left(2\pi f_c t + 2\pi k_f \int g(t)\ dt\right)$$

So we can think of this as phase modulation with a phase deviation $\theta(t) = 2\pi k_f \int g(t)\ dt$ instead of $\theta(t) = k_p g(t)$.

This observation allows us to make a frequency modulator from a phase modulator as we will now see.

If we define

$$f(t) = 2\pi \int g(t)\ dt$$

then we may express the FM signal as

$$\tilde{s}_f(t) = A_c \cos(2\pi f_c t + k_f f(t))$$

This is the FM signal, given a modulating signal $g(t)$, but it is also a PM signal, given a modulating signal $f(t)$.

This shows that there is a relationship between phase modulation and frequency modulation. This relationship is illustrated in Figure 8.2.1.

Note that $f(t)$ is derived from $g(t)$ by integration, and therefore is also a baseband signal with bandwidth W (integration is a linear operation, which does not introduce new frequency components).

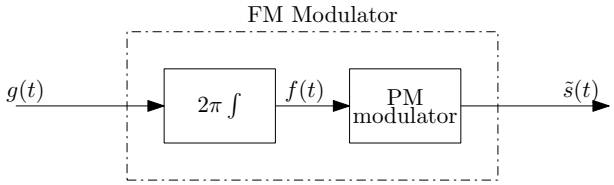


Figure 8.2.1: Illustration of the relationship between phase modulation (PM) and frequency modulation (FM). The diagram shows that a PM modulator and an integrator may be used to perform frequency modulation. The output signal is FM modulated by $g(t)$.

The converse is also true, i.e. a PM modulator can be implemented by the cascade of a differentiator and a FM modulator as is illustrated in Figure 8.2.2

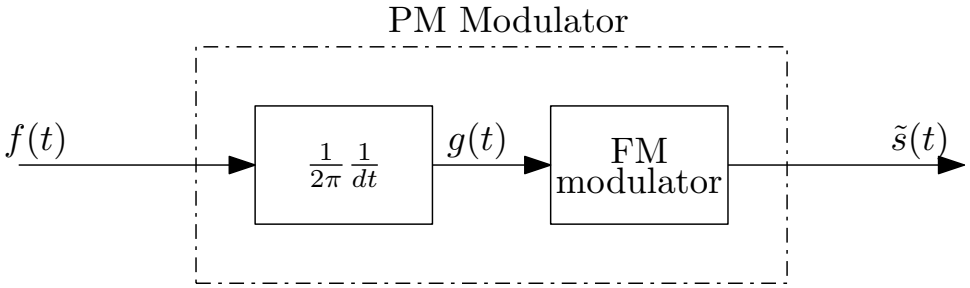


Figure 8.2.2: Illustration of the relationship between phase modulation (PM) and frequency modulation (FM). The diagram shows that an FM modulator and an differentiator may be used to perform phase modulation. The output signal is PM modulated by $f(t)$.

Chapter 9

PM and FM Spectra

The approach taken here is to firstly consider the spectra for PM and FM for the special case of a sinusoidal modulating signal which helps with understanding - then the more general result of modulating with a lowpass signal is considered.

9.1 Sinusoidal Modulation

Let the modulating signal be $g(t) = A_m \cos(2\pi f_m t)$.

9.1.1 Phase Modulation

The PM signal is

$$\tilde{s}_p(t) = A_c \cos(2\pi f_c t + k_p A_m \cos(2\pi f_m t))$$

The *peak phase deviation* is $\Delta\theta = k_p A_m$. This is also defined to be the *modulation index*, β, for the PM signal. We may thus write the PM signal as

$$\tilde{s}_p(t) = A_c \cos(2\pi f_c t + \beta \cos(2\pi f_m t)) \tag{9.1.1}$$

9.1.2 Frequency Modulation

The instantaneous frequency of the FM signal is

$$\begin{aligned} f_i(t) &= f_c + k_f g(t) \\ &= f_c + k_f A_m \cos(2\pi f_m t) \end{aligned}$$

The *peak frequency deviation* is $\Delta f = k_f A_m$. The FM signal is

$$\begin{aligned}\tilde{s}_f(t) &= A_c \cos\left(2\pi f_c t + 2\pi k_f \int g(t)\ dt\right) \\ &= A_c \cos\left(2\pi f_c t + 2\pi k_f A_m \int \cos(2\pi f_m t)\ dt\right) \\ &= A_c \cos\left(2\pi f_c t + \frac{k_f A_m}{f_m} \sin(2\pi f_m t)\right)\end{aligned}$$

The *peak phase deviation* is $\Delta\theta = k_f A_m / f_m = \Delta f / f_m$. This is also defined to be the *modulation index*, β, for the FM signal. We may thus write the FM signal as

$$\tilde{s}_f(t) = A_c \cos(2\pi f_c t + \beta \sin(2\pi f_m t)) \tag{9.1.2}$$

9.1.3 Frequency Spectrum (of FM)

Note that the expressions for the PM and FM signals (equations (9.1.1) and (9.1.2)) are very similar. This is due to the fact that integration of a sinusoidal signal yields another sinusoidal signal. We consider the FM signal:

$$\begin{aligned}\tilde{s}_f(t) &= A_c \cos(2\pi f_c t + \beta \sin(2\pi f_m t)) \\ &= A_c \left[\cos(\beta \sin(2\pi f_m t)) \cos(2\pi f_c t) - \sin(\beta \sin(2\pi f_m t)) \sin(2\pi f_c t)\right] \\ &= A_c \left[\cos(\beta \sin(x)) \cos(2\pi f_c t) - \sin(\beta \sin(x)) \sin(2\pi f_c t)\right]\end{aligned}$$

where we let $x \triangleq 2\pi f_m t$.

Note that $\cos(\beta \sin(x))$ and $\sin(\beta \sin(x))$ are periodic, and have Fourier series whose coefficients are given by Bessel functions:

$$\begin{aligned}\cos(\beta \sin(x)) &= J_0(\beta) + 2\sum_{n=1}^{\infty} J_{2n}(\beta) \cos(2nx) \\ &\text{and} \\ \sin(\beta \sin(x)) &= 2\sum_{n=1}^{\infty} J_{2n-1}(\beta) \sin((2n-1)x)\end{aligned}$$

where $J_n(\beta)$ is the n^{th} order Bessel function of the first kind.

A plot of some of these functions is shown in figure 9.1.1 for some small values of n.

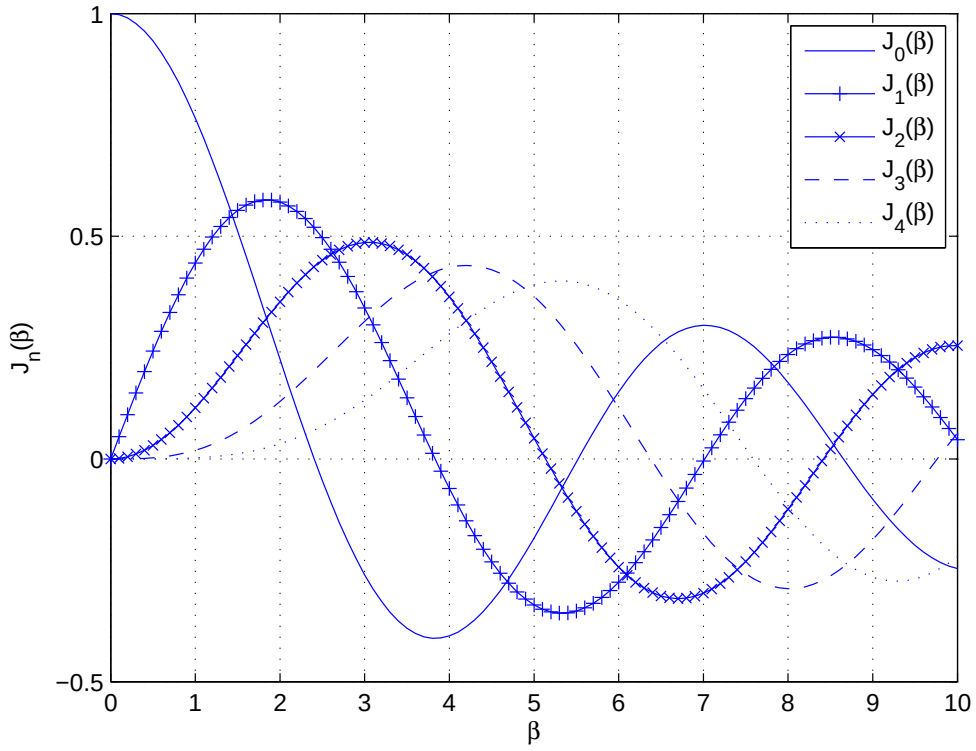


Figure 9.1.1: Some Bessel functions of the first kind, $J_n(\beta)$, plotted against β. The plots are shown for $n = 0, 1, 2, 3, 4$.

Substituting these Fourier series into the expression for the FM signal gives

$$\begin{aligned}
\tilde{s}_f(t) &= A_c J_0(\beta)\cos(2\pi f_c t) + 2A_c\sum_{n=1}^{\infty} J_{2n}(\beta)\cos(2\pi(2n)f_m t)\cos(2\pi f_c t) \\
&\quad -2A_c\sum_{n=1}^{\infty} J_{2n-1}(\beta)\sin(2\pi(2n-1)f_m t)\sin(2\pi f_c t) \\
&= A_c J_0(\beta)\cos(2\pi f_c t) + A_c\sum_{n=1}^{\infty} J_{2n}(\beta)\left[\cos\{2\pi(f_c+2nf_m)t\} + \cos\{2\pi(f_c-2nf_m)t\}\right] \\
&\quad +A_c\sum_{n=1}^{\infty} J_{2n-1}(\beta)\left[\cos\{2\pi(f_c+(2n-1)f_m)t\} - \cos\{2\pi(f_c-(2n-1)f_m)t\}\right] \\
&= A_c J_0(\beta)\cos(2\pi f_c t) \\
&\quad +A_c J_1(\beta)\left[\cos\{2\pi(f_c+f_m)t\} - \cos\{2\pi(f_c-f_m)t\}\right] \\
&\quad +A_c J_2(\beta)\left[\cos\{2\pi(f_c+2f_m)t\} + \cos\{2\pi(f_c-2f_m)t\}\right] \\
&\quad +A_c J_3(\beta)\left[\cos\{2\pi(f_c+3f_m)t\} - \cos\{2\pi(f_c-3f_m)t\}\right] \\
&\quad +A_c J_4(\beta)\left[\cos\{2\pi(f_c+4f_m)t\} + \cos\{2\pi(f_c-4f_m)t\}\right] \\
&\quad +\cdots
\end{aligned} \tag{9.1.3}$$

Here we have used the trigonometric identities

$$\begin{aligned}\cos(A+B)+\cos(A-B) &= 2\cos A\cos B\\ \cos(A+B)-\cos(A-B) &= -2\sin A\sin B\end{aligned}$$

The frequency spectrum is seen to be a line spectrum, i.e. it is composed of a carrier component, and discrete frequency components spaced at intervals of f_m around the carrier. Some sample spectra are illustrated in figure 9.1.2, for different values of modulation index β. Note that the relative strengths of the frequency components is highly dependent on β.

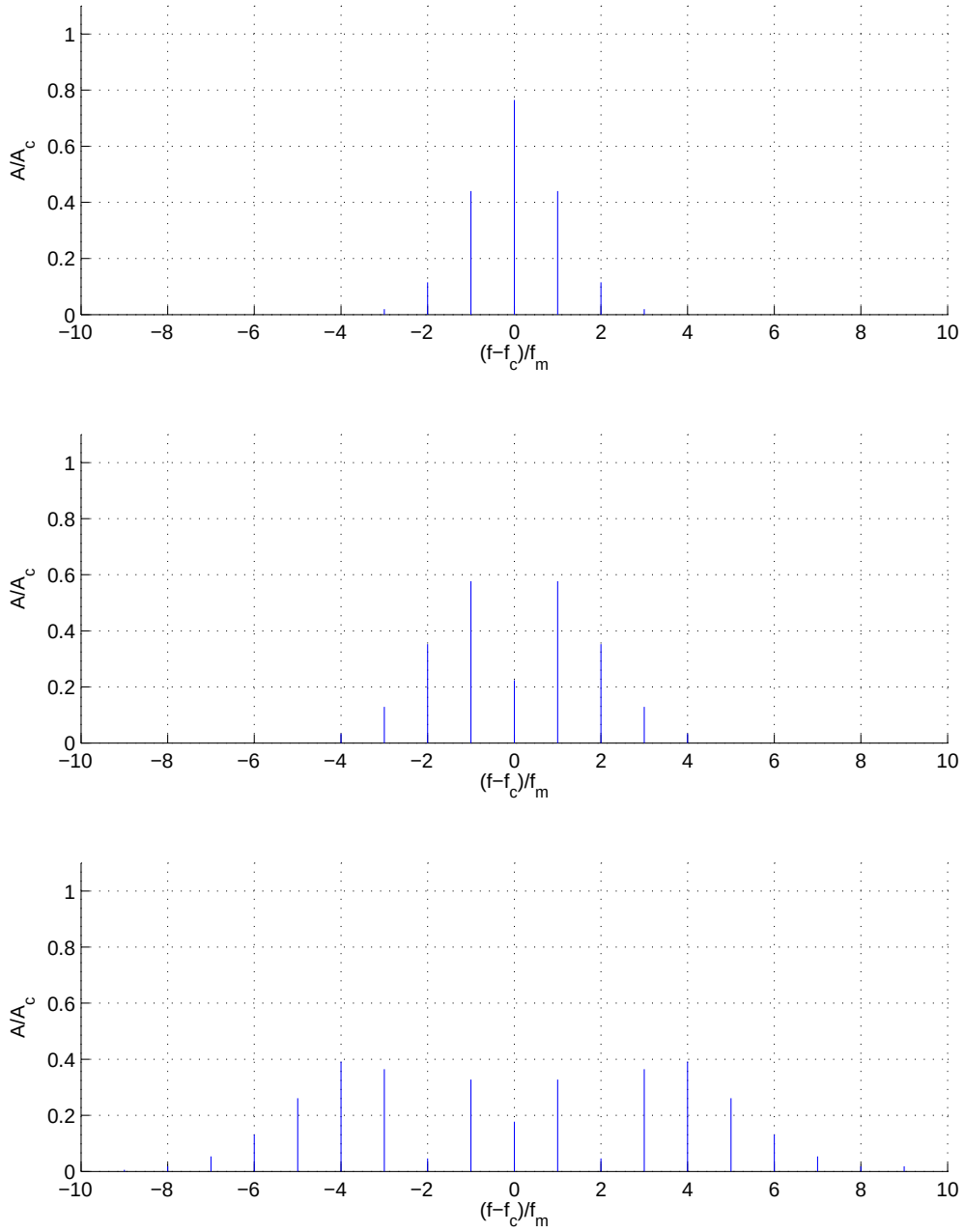


Figure 9.1.2: Frequency spectrum of an FM signal, for the case of a sinusoidal modulating signal. The spectrum is plotted for three different values of β, and in each case is centred on f_c. The upper plot is for $\beta = 1$, the middle plot is for $\beta = 2$, and the lower plot is for $\beta = 5$. The amplitude is normalised with respect to the carrier amplitude. Only the spectra for positive frequencies are shown in the figure.

9.2 General Case

Let $g(t)$ be a baseband information signal, with maximum frequency W.

9.2.1 Phase Modulation

The PM signal is

$$\begin{aligned}\tilde{s}_p(t) &= A_c \cos(2\pi f_c t + k_p g(t)) \\ &= A_c \cos(k_p g(t)) \cos(2\pi f_c t) - A_c \sin(k_p g(t)) \sin(2\pi f_c t)\end{aligned}$$

The *peak phase deviation* is $\Delta\theta = k_p \max\{g(t)\}$. If this is small, then $\cos(k_p g(t)) \approx 1$ and $\sin(k_p g(t)) \approx k_p g(t)$, giving

$$\tilde{s}_p(t) \approx A_c \cos(2\pi f_c t) - A_c k_p g(t) \sin(2\pi f_c t) \tag{9.2.1}$$

This is called 'narrowband PM', because assuming that $\Delta\theta$ is small is equivalent to assuming that the modulation index β is small and we saw in the previous section that the smaller β is the smaller the resulting modulated spectrum.

The first term is a carrier component, and the second term is a DSB-SC signal, giving sidebands around the carrier. The overall signal bandwidth is $2W$.
Note that the narrowband PM signal $s_p(t)$ is similar to a full AM signal, but there is a difference: the carrier is phase shifted by 90° relative to the sidebands.

9.2.2 Frequency Modulation

Here we have $f_i(t) = f_c + k_f g(t)$, giving peak frequency deviation $\Delta f = k_f \max\{g(t)\}$. We may write the FM signal as

$$\tilde{s}_f(t) = A_c \cos(2\pi f_c t + k_f f(t))$$

where $f(t) = 2\pi \int g(t)\, dt$. Recall that $f(t)$ is, like $g(t)$, a baseband signal, with highest frequency W. We can regard $\tilde{s}_f(t)$ as a PM signal, with modulating signal $f(t)$. This signal has peak phase deviation $\Delta\theta = k_f \max\{f(t)\}$. If this is small, we may approximate

$\cos(k_f f(t)) \approx 1$ and $\sin(k_f f(t)) \approx k_f f(t)$, giving

$$\tilde{s}_f(t) = A_c \cos(2\pi f_c t) - A_c k_f f(t) \sin(2\pi f_c t) \tag{9.2.2}$$

This is called 'narrowband FM', or NBFM. The first term is a carrier component, and the second term is a DSB-SC signal, giving sidebands around the carrier. The overall signal bandwidth is $2W$. The FM signal $\tilde{s}_f(t)$ is similar to full AM, except that the carrier is phase shifted by $90°$ relative to the sidebands. The sidebands contain $f(t)$ (and hence $g(t)$). This signal will also have small peak frequency deviation Δf, in general.

Exercise: Show that for the case of sinusoidal modulation, we may arrive at equation (9.2.2) by taking equation (9.1.3) and making the approximations that for small β, $J_0(\beta) \approx 1$, $J_1(\beta) \approx \beta/2$ and $J_n(\beta) \approx 0$ for $n \geq 2$. Also, justify these approximations by looking at figure 9.1.1.

For *wideband* FM (or WBFM), with *large* peak frequency deviation Δf, we define the *deviation ratio*

$$D = \frac{\Delta f}{W} = \frac{\text{peak frequency deviation}}{\text{bandwidth of the modulating signal}}$$

The bandwidth B of the FM signal is theoretically infinite. However, as an approximate measure of the FM signal bandwidth B, we may use *Carson's Rule*:

$$B \approx 2(D+1)W$$

For the case of a sinusoidal modulating signal, this reduces to

$$B \approx 2(\beta+1) f_m$$

A sketch of the frequency spectra of the different FM signals discussed in this section in shown in figure 9.2.1. A similar diagram may be drawn for the case of PM modulation.

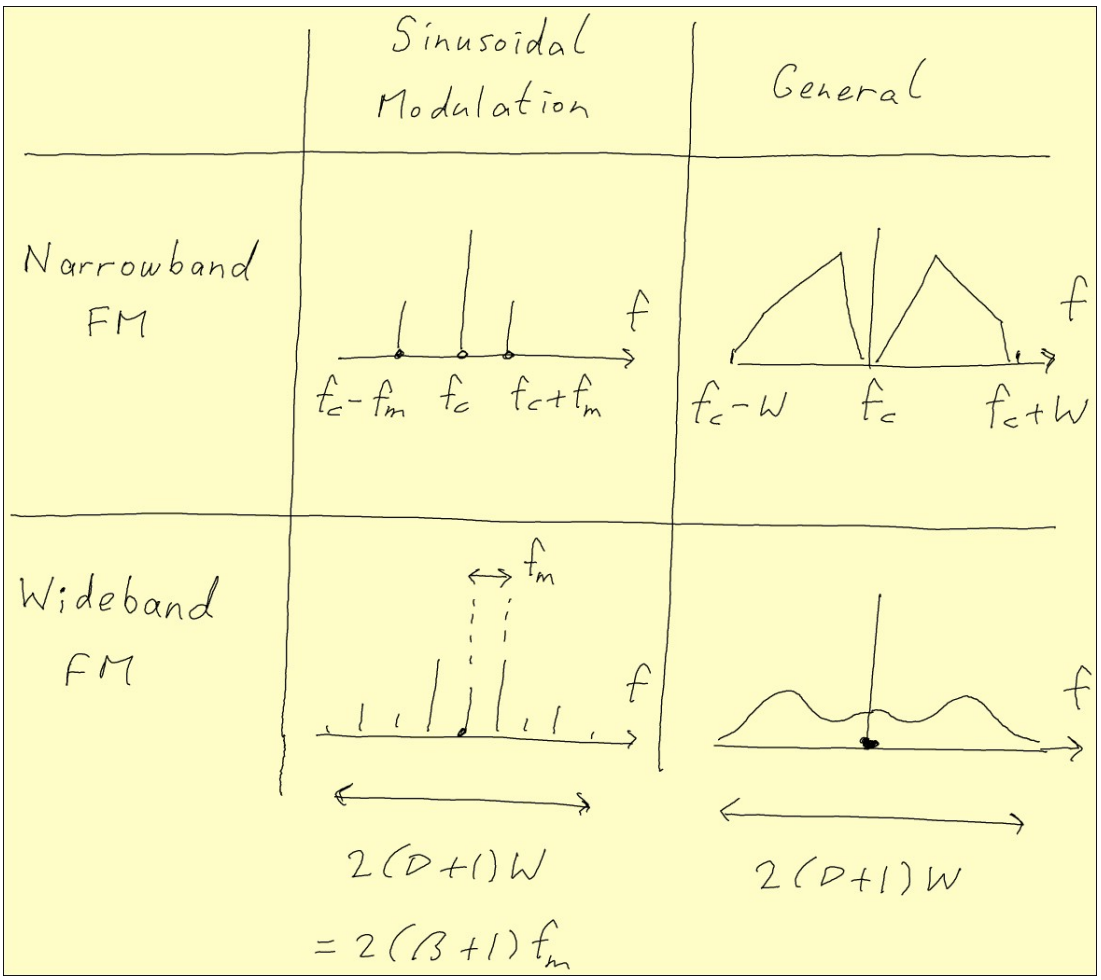


Figure 9.2.1: A sketch of the frequency spectra of FM modulated signals.

Chapter 10

FM & PM modulator implementation

From figures 8.2.1 and 8.2.2, we can see that a PM modulator may be derived quite easily from an FM modulator, and vice versa. Therefore we need only consider one form of angle modulation. We shall focus on FM modulation.

10.1 Narrowband FM (NBFM) Modulation

A circuit to implement narrowband FM modulation is shown in figure 10.1.1.

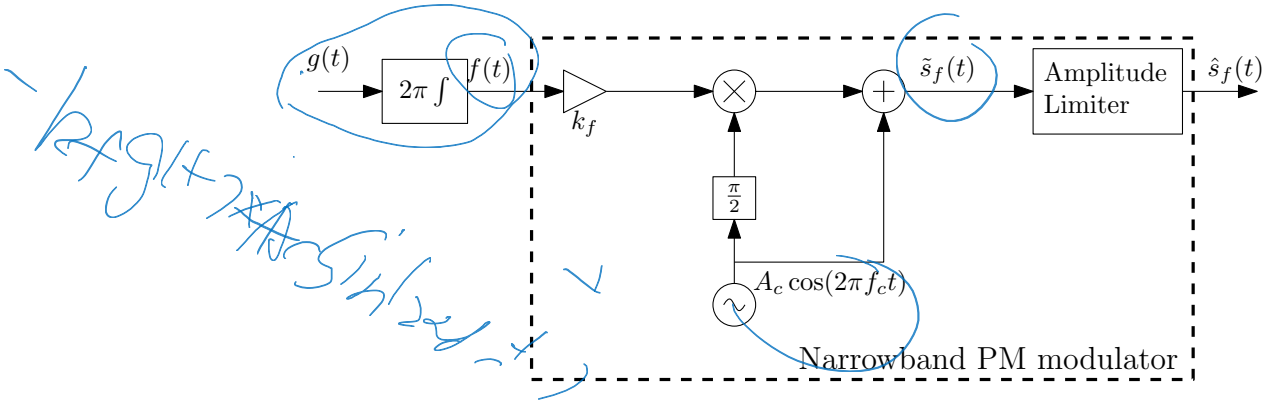


Figure 10.1.1: A narrowband FM modulator. Note the similarity of this diagram to figure 8.2.1.

This circuit simply implements equation 9.2.2 directly, followed by an amplitude limiter.

The reason for the amplitude limiter is as follows:

- From equation (9.2.2), the output signal $\tilde{s}_f(t)$ has an envelope given by

$$A(t) = \sqrt{A_c^2 + (A_c k_f f(t))^2}$$

- This is slightly time-varying.

- FM signals should have constant amplitude
- The amplitude limiter removes these amplitude variations, giving a constant-amplitude output signal $\hat{s}_f(t)$.

The reason why the amplitude was time varying is because equation 9.2.2 is an approximation and is not truly FM.

10.2 Wideband FM Modulation - Indirect Method

A NBFM modulator, such as described in the previous section, can be combined with a frequency multiplier to make a WBFM modulator.
To see how this may be accomplished, we must first understand the operation of a frequency multiplier.

10.2.1 Frequency Multiplier

A frequency multiplier consists of a nonlinear circuit (or device) followed by a BPF as shown in figure 10.2.1.

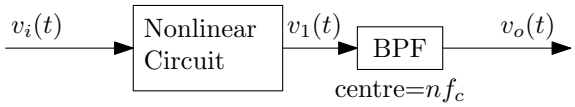


Figure 10.2.1: Frequency multiplier.

Suppose that the input signal is a bandpass signal, with frequency spectrum centred on f_c:

$$v_i(t) = A_c \cos(2\pi f_c t + \theta(t)) \tag{10.2.1}$$

The nonlinear circuit produces harmonics. The BPF has centre frequency nf_c, and therefore selects the n^{th} harmonic.
Example: $n = 5$. $v_1(t)$ contains

$$v_i^5(t) = A^5(t) \cos^5(2\pi f_c t + \theta(t))$$

Now,

$$\cos^5(x) = \frac{5}{8}\cos(x) + \frac{5}{16}\cos(3x) + \frac{1}{16}\cos(5x)$$

The BPF selects frequency components around $5f_c$; therefore the output signal is

$$v_o(t) = \frac{A^5(t)}{16}\cos\left[5\left(2\pi f_c t + \theta(t)\right)\right]$$

We see that the output frequency has increased by a factor of n $(= 5)$ with respect to the input signal frequency; hence the name 'frequency multiplier'. Note that the phase deviation has also increased by a factor of n $(= 5)$.

WBFM - Indirect Method continued

The indirect method of producing a WBFM signal is now shown in figure 10.2.2.
Its operation is as follows:

- The input to the frequency multiplier, $v_i(t)$, is a NBFM signal,
- Therefore $\theta(t)$ in equation 10.2.1 corresponds to the phase deviation of a NBFM signal.
- The effect of the frequency multiplier is to multiply this phase deviation $\theta(t)$ by a factor of n (5 in the example above).
- Therefore, $d\theta(t)/dt$ is multiplied by a factor of n, and so the frequency deviation is multiplied by a factor of n.
- Therefore, the output $v_o(t)$ is a WBFM signal.
- As before the output signal is finally passed through an amplitude limiter to remove amplitude variations.

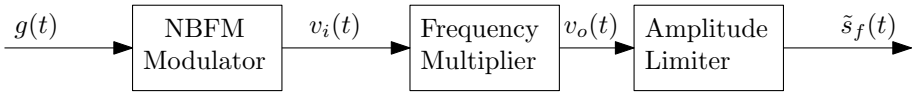


Figure 10.2.2: Indirect method of WBFM modulation.

10.2.2 Wideband FM Modulation - Direct Method

The direct method of producing an WBFM signal involves using a voltage-controlled oscillator (VCO). The output of a voltage-controlled oscillator is a sinusoidal signal with instantaneous frequency equal to

$$f_o(t) = f_c + K_o v_i(t) \tag{10.2.2}$$

Where

- $v_i\,(t)$ is the voltage at the VCO input.
- The value f_c is called the *free-running frequency* and is equal to the VCO output frequency when the input voltage is zero.

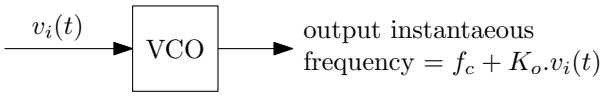


Figure 10.2.3: Voltage-controlled oscillator.

In this application, the free-running frequency is chosen to match the carrier frequency. Since the VCO output frequency deviation is proportional to the applied input signal, it is easily seen that the VCO implements FM modulation directly.

A problem with the VCO is poor frequency stability, i.e. variable frequency oscillators are more difficult to make compared to fixed-frequency oscillator. It may be stabilised to some extent by using a feedback control mechanism, and phase Locked Loops but these are beyond the scope of this module.

Chapter 11

FM demodulator implementation

We will consider demodulation of FM signals. There are two principal methods of demodulating an FM signal:

11.1 Frequency Discriminator

A frequency discriminator is shown in figure 11.1.1. This consists of a differentiator followed by an envelope detector. The differentiator is also called a *slope circuit*.

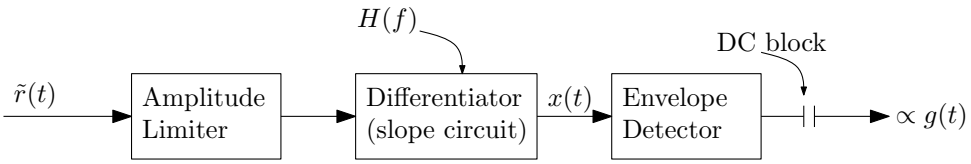


Figure 11.1.1: Frequency discriminator.

The received signal is given by

$$\tilde{r}\left(t\right) = A\left(t\right)\cos\left(2\pi f_c t + \theta\left(t\right)\right)$$

This is passed through an amplitude limiter, which gives $A_c \cos\left(2\pi f_c t + \theta\left(t\right)\right)$; this signal is applied as input to the differentiator. The differentiator gives output

$$x\left(t\right) = -A_c \sin\left(2\pi f_c t + \theta\left(t\right)\right) \cdot \left[2\pi f_c + \frac{d\theta\left(t\right)}{dt}\right]$$

Now the phase deviation $\theta(t)$ of an FM signal is given by:

$$\begin{aligned} \theta(t) &= 2\pi k_f \int g(t)\, dt \\ \Rightarrow \frac{d\theta(t)}{dt} &= 2\pi k_f g(t) \end{aligned}$$

Therefore, the input to the envelope detector is:

$$x(t) = -A_c \left[2\pi f_c + 2\pi k_f g(t)\right] \sin\left(2\pi f_c t + \theta(t)\right)$$

But this is just a full sinusoid, with envelope:

$$2\pi A_c \left(f_c + k_f g(t)\right)$$

A capacitor at the envelope detector output blocks DC, giving an output signal proportional to $g(t)$.

11.1.1 differentiator

In practice, to build the differentiator (or slope circuit) we use some circuit which has a transfer function $H(f)$ which is approximately linear near $f = f_c$, i.e. $H(f) \approx K_1 f + K_2$ in the vicinity of f_c as illustrated in Figure 11.1.2.

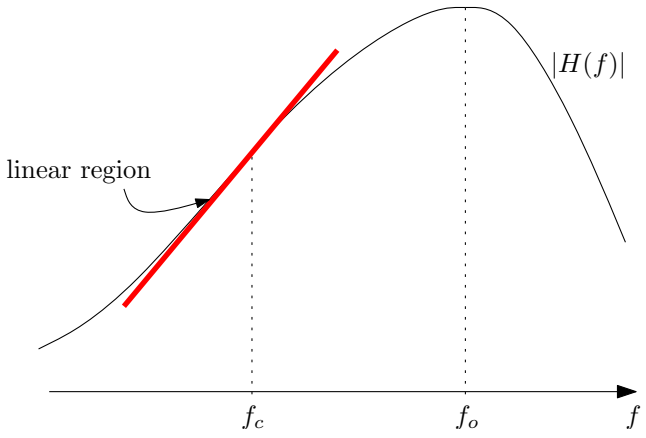


Figure 11.1.2: Transfer function of slope circuit. Here an oscillator circuit is used as slope circuit, with frequency of oscillation f_o.

This circuit thus performs linear conversion of frequency changes to amplitude changes.

Any filter circuit will have this ability, over some range of frequency.

For good FM demodulation, we require good linearity (to avoid distortion), and a steep slope in the linear region (as frequency changes are relatively small).

11.2 Phase-Locked Loop (PLL)

A diagram of a phase-locked loop is shown in figure 11.2.1.

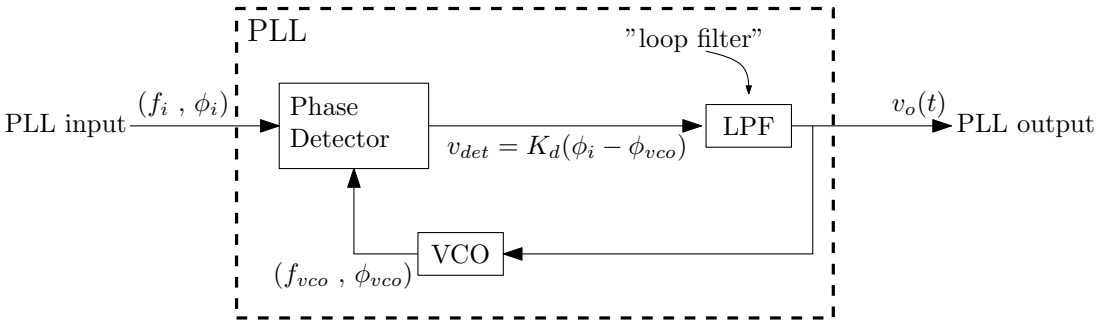


Figure 11.2.1: Phase-locked loop (PLL) circuit.

It consists of a voltage-controlled oscillator (VCO), a phase detector and a loop filter (low pass filter).

When operating correctly output from the VCO is *locked* to the PLL's input $v_i(t)$; meaning they have approximately same frequency, i.e. $f_{vco} = f_i$. How does this work?

To understand this firstly imagine a steady-state situation:

11.2.1 Steady-state

This is where the input $v_i(t)$ has a fixed frequency f_i Hertz and phase offset of ϕ_i.
The claim is that the system *locks* to a point where the VCO is also oscillating at f_i with some fixed phase offset ϕ_{vco}. For the VCO to be a fixed frequency it's control voltage, and consequently the PLL output $v_o(t)$, are constant.

As the output from the phase detector is a constant, the LPF has no function in this simplified steady-state scenario, so we will ignore it in our analysis:

- When the $f_{vco} = f_i$ then the phase error, $\Delta\phi = \phi_i - \phi_{vco}$ is a constant (but not necessarily zero)

- If the f_{vco} *increases* (due to noise) by some small amount $f_{vco} \mapsto f_{vco} + \delta f$, then the polarity of the phase detector is such that the phase error $\Delta\phi$ would *decrease* by some small amount (i.e. negative feedback) $\Longrightarrow$ the VCO control voltage *decreases* $\Longrightarrow$ the f_{vco} *decreases* toward f_i again.
- The same can be argued changes in frequency in the other direction.
- Thus the system is stable with $f_o \approx f_i$ Hertz.

Ok, but how does this help in demodulation?

The frequency of oscillation, f_{vco}, of the VCO is (from equation 10.2.2)

$$\begin{aligned} f_{vco} &= f_{zero} + K_{vco} v_o \\ \Rightarrow v_o &= \frac{1}{K_{vco}} \left(f_{vco} - f_{zero} \right) \\ &\approx \frac{1}{K_{vco}} \left(f_i - f_{zero} \right) \end{aligned}$$

i.e. the output voltage is directly proportional difference between the input frequency and the natural oscillation frequency of the VCO - this is key, as we will now see!

11.2.2 Slowly varying frequency

Now we argue that if the input frequency f_i varies slowly (compared to nominal) then the PLL loop will remain locked to this slowly varying frequency, i.e. the relationship:

$$v_o \approx \frac{1}{K_{vco}} \left(f_i - f_{zero} \right)$$

still holds true.
In FM, the frequency f_i is given by:

$$\begin{aligned} f_i &= f_c + k_f g\left(t\right) \\ \Rightarrow v_o\left(t\right) &\approx \frac{1}{K_{vco}} \left(f_c + k_f g\left(t\right) - f_{zero} \right) \\ &= \text{Linear function}\left(g\left(t\right)\right) \end{aligned}$$

So the output voltage is a linear function of the modulating function, and the parameters of said function are know, so demodulation is achieved.

11.2.3 low pass filter

What is the purpose of the Low Pass Filter (LPF) in Figure 11.2.1?
Well thus far we've not considered noise.
If there is any noise present at the output of the PLL (and there will be) then, without a LPF, this noise would propagate straight out onto $v_o(t)$, causing two effects

1. The demodulated output would be noisy, and
2. The VCO control signal would be noisy causing the PLL to not lock as well as it should - which in extreme case can cause the system to fail to track the modualtion.

Both these issues can be solved by including a LPF within the loop, as shown in Figure 11.2.1? The bandwidth of the the LPF should be:

1. Greater than W, the bandwidth of the modulating signal $g(t)$. This is needed to ensure that the output $v_o(t) = \text{Linear function}(g(t))$, and
2. As small as possible to limit the amount of noise in the feedback path.

Ideally it would be a brick-wall filter with cut-off frequency W, typically however these circuits are implemented with simple RC filters with bandwidths $> W$.

11.2.4 Phase detector

In practice, the phase detector is often implemented by an analog multiplier - this is shown in Figure 11.2.2.

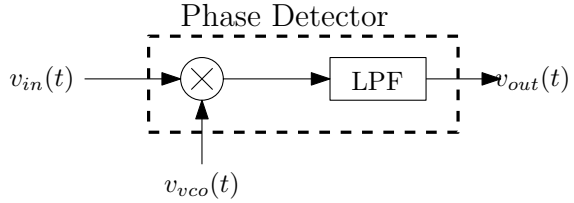


Figure 11.2.2: Phase detector implemented using an analog multiplier.

When the PLL is locked, the frequency of the two inputs to the phase detector are the same, but they have different phases, i.e.:

$$\begin{aligned} v_{in}(t) &\propto \sin(\phi_i(t)) = \sin(2\pi f_i t + \theta_i(t)), \quad \text{and} \\ v_{vco}(t) &\propto \sin(\phi_{vco}(t)) = \sin(2\pi f_i t + \theta_{vco}(t)) \end{aligned}$$

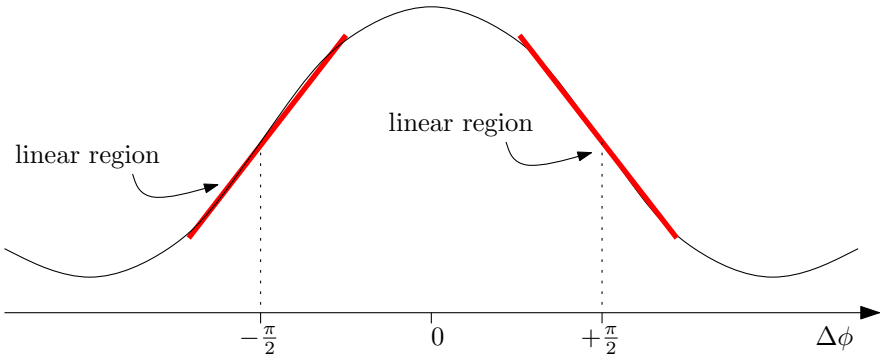


Figure 11.2.3: v_{out} verses $\Delta\phi$ for a multiplier based phase detector

the output from the multiplier is:

$$\begin{aligned} &\propto \sin\left(2\pi f_i t + \theta_i\left(t\right)\right)\sin\left(2\pi f_i t + \theta_{vco}\left(t\right)\right) \\ &= \cos\left(\theta_i\left(t\right) - \theta_{vco}\left(t\right)\right) - \cos\left(4\pi f_i t + \theta_i\left(t\right) + \theta_{vco}\left(t\right)\right) \end{aligned}$$

The inclusion of the LPF removes the high double frequency leaving the output:

$$\begin{aligned} v_{out}\left(t\right) &\propto \cos\left(\theta_i\left(t\right) - \theta_{vco}\left(t\right)\right) = \cos\left(\Delta\phi\right) \\ &= K_d \cos\left(\Delta\phi\right) \end{aligned}$$

where K_d is the constant of proportionality which we call the *gain* of the detector.

Refer to Figure 11.2.3 for a graph of v_{out} verses $\Delta\phi$.

Chapter 12

Random Signals

Before looking at random signal let's review some basic probability theory.

12.1 Random Variables (RVs)

A *random variable* X is a number associated with the outcome of an experiment. If we repeat the experiment, we may get a different number, in an unpredictable way - each repetition of the experiment yields a *sample value*.

12.1.1 Discrete RVs

A *discrete* random variable is one which may only take on values from a finite set of possible values, examples include:

- The throw of a die
- Picking a card from a deck
- The value on a roulette wheel.

We characterize a discrete RV by its *Probability Mass Function* (PMF) $f_X(x)$ defined as follows:

$$f_X(x) = \Pr(X = x)$$

i.e. the pmf tells us the probability for each possible outcome. Example:

- Throwing a die: $f_X(x) = \frac{1}{6} \quad \forall x$

- Picking a card: $f_X(x) = \frac{1}{52} \quad \forall x$
- Roulette wheel: $f_X(x) = \begin{cases} \frac{18}{38} & \text{for red} \\ \frac{18}{38} & \text{for black} \\ \frac{2}{38} & \text{for green ('0' or '00')} \end{cases}$

Notice how the sum of all the probabilities $= 1$.

This is because, of course, it is guaranteed (i.e. probability$= 1$) that one of the outcomes will happen each time. We can put this in math:

$$\sum_{\forall x} f_X(x) \quad = \quad 1$$

where the summation is over all possible values of x.

12.1.2 Continuous RVs

A *continuous* random variable is one which can take on any value in a continuous range, e.g.

- The height of a randomly selected person.
- The voltage amplitude across a noisy resistor

Lets look further at the first example, and ask the following questions:

1. What is the probability a randomly selected person has height 1.7m?
2. What is the probability a randomly selected person has height 2.7m?

Well in Beijing (2010) the average person's height was 1.747m, so we might think that the probability of a person being 1.7m is far higher than 2.7m - right? Well maybe not. What if we rephrased the questions as follows:

1. What is the probability a randomly selected person has height 1.70000000'm?
2. What is the probability a randomly selected person has height 2.70000000'm?

(where the ' on the end means repeating forever).
We haven't changed the question - but now what's the answer?

Zero!

We can put this in math, and say that for a continuous RV

$$\Pr(X = x) \;=\; 0 \;\; \forall x$$

So the PMF of a continuous RV is zero[1] for every possible value of x; i.e. the PMF is a useless measure of probability for continuous RVs, we need another tool...

12.1.2.1 Probability Density Function (PDF)

Based on the above discussion it only makes sense to talk about the probability of X on a range of values. Therefore we define the Probability Density Function (PDF) which has the

[1] If it helps, you can think of discrete RVs having mass and are well described by a probability mass function, whereas a continuous RV has zero mass for every possible values and so the PMF is useless.

following property:

$$\int_a^b f_X(x)\,dx \quad \triangleq \quad \Pr(a \leq x \leq b)$$

i.e. its integral over a range $[a, b]$ (or area under the curve from $x = a$ to b) give the probability that x obtains a value in that range (inclusive).
So now we may ask the questions:

1. What is the probability a randomly selected person has height between 1.69 and 1.71m?
2. What is the probability a randomly selected person has height between 2.69 and 2.71m?

Now these questions will have non-zero answers, as illustrated in Figure 12.1.1.

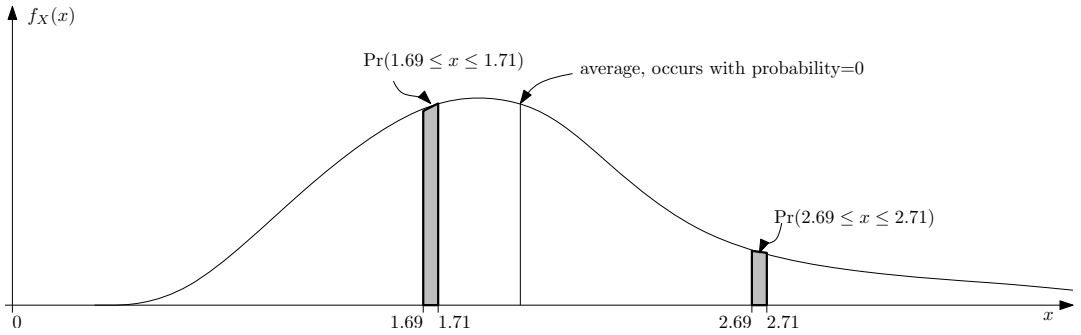


Figure 12.1.1: Illustration of PDF of people's heights (not drawn to scale)

The PDF has the following properties:

- This function is non-negative[2], i.e. $f_X(x) \geq 0$ for all x.
- It integrates to unity, i.e. $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$

12.1.2.2 Cumulative Distribution Function (CDF)

Sometimes (actually fairly often) it is easier to consider the probability that a RV is less than (or equal to) some value, and so we define the *Cumulative Distribution Function* (CDF) of a continuous random variable X as:

$$F_X(x) = \Pr(X \leq x)$$

[2] – as this would imply that there are some ranges when the area under the curve is negative, and hence negative probabilities would result which of course is not possible

Some properties

- It contains all the same information as the PDF
- This is a non-decreasing function, i.e. $F_X(x_1) \leq F_X(x_2)$ if $x_1 < x_2$
- $F_X(-\infty) = 0$, $F_X(\infty) = 1$

12.1.2.3 Relationship between CDF and PDF

From the definition of the CDF we have

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) = \Pr(-\infty \leq X \leq x) \\ &= \int_{-\infty}^{x} f_X(\nu)\, d\nu \end{aligned}$$

where f_X is the PDF of X.
This can also be re-written as:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

From these two expressions it is clear that either f_X or F_X is sufficient to completely characterize an independent RV as we can use either to compute the probability of any outcome. From them we can calculate the value of a function of a RV, and importantly some basic statistics e.g. the mean and variance.

12.1.3 Expectation

Q. If you repeat an experiment many times what would the average result value be?
A. Add them all up, and divide by the number of experiments.

Lets consider the discrete example of throwing a die N times. We would expect to get a 'one' $\frac{1}{6}N$ times, a 'two' $\frac{1}{6}N$ times, etc... Putting these in a table we have:

outcome	expected number of occurrences	contribution to sum
1	$\frac{1}{6}N$	$1 \times \frac{1}{6}N$
2	$\frac{1}{6}N$	$2 \times \frac{1}{6}N$
3	$\frac{1}{6}N$	$3 \times \frac{1}{6}N$
4	$\frac{1}{6}N$	$4 \times \frac{1}{6}N$
5	$\frac{1}{6}N$	$5 \times \frac{1}{6}N$
6	$\frac{1}{6}N$	$6 \times \frac{1}{6}N$
	expected sum of all outcomes =	$(1+2+3+4+5+6) \times \frac{1}{6}N = \frac{21}{6}N$
	average = Expected value =	$\frac{21}{6} = 3.5$

You can check this in Matlab easily:

```
>> X = randi(6,[1e6,1]); % 1 million random die throws
>> mean(X)
ans =
    3.5019
```

We can generalize and say the the expected value (or mean, or true average) of a discrete RV is:

$$\begin{aligned} E[X] &= \sum x.\Pr(X = x) \\ &= \sum x.f_X(x) \end{aligned}$$

where the summation is done over all possible values of x, and f_X is the PMF for X.

By taking the limiting case of this, it can be shown that for continuous variables we have the following expected value:

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

where f_X is now the PDF of the RV.

The expectation represents the 'true' mean, taking into account all possible (infinite in number) outcomes and their probabilities. It is in contrast to $\bar{X}$, the *sample mean*, which is the

average of N (a finite number) sample values:

$$\bar{X} \;=\; \frac{1}{N}\sum_{n=1}^{N} x_n$$

We would, of course, like to believe that $\lim_{N\to\infty} \bar{X} = m_X$, in which case we'd say that $\bar{X}$ is an un-biased estimator of the mean.

12.1.3.1 Expected value of a function

For any function $g(\cdot)$, the expectation of $g(X)$, i.e. a function of a random variable, is

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

From this it is easy to see that expectation is *linear,* i.e.

$$E[\alpha g(X) + \beta h(Y)] = \alpha E[g(X)] + \beta E[h(Y)]$$

Thus it follows that the expectation function is linear:

$$E[\alpha X + \beta Y] \;=\; \alpha E[X] + \beta E[Y]$$

12.1.3.2 Moments

The n^{th} *moment* of X is defined by

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x)\; dx$$

Example: $n = 1$; $E[X^1] = m_X$ is the *mean value* of X.

Example: $n = 2$; $E[X^2]$ is the *mean-square value* of X. The mean-square value of X gives us an indication of how 'large' x is.

12.1.3.3 Central moments (and variance)

The n^{th} *central moment* of X is given by

$$E[(X - m_X)^n] = \int_{-\infty}^{\infty} (x - m_X)^n\, f_X(x)\; dx$$

Example: $n = 2$; $E\left[(X - m_X)^2\right] = \sigma_X^2 = Var\left[X\right]$ is the *variance* of X.

The variance of X gives us an indication of the extent to which X varies about its mean value.
Variance has the following properties:

$$\begin{aligned} Var\left[X + \alpha\right] &= Var\left[X\right] \\ Var\left[\alpha X\right] &= \alpha^2 Var\left[X\right] \end{aligned}$$

The square root of the variance, σ_X, is called the *standard deviation* of X.

Note that

$$\begin{aligned} \sigma_X^2 &= E\left[(X - m_X)^2\right] = E\left[X^2 - 2X.m_X + m_X^2\right] = E\left[X^2\right] - 2m_X.E\left[X\right] + m_X^2 \\ &= E\left[X^2\right] - m_X^2 \end{aligned}$$

12.1.3.4 Gaussian distribution

Example: A Gaussian-distributed random variable (also known as normal-distributed random variable) is described by the PDF

$$f_X\left(x\right) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{\left(x - m_X\right)^2}{2\sigma_X^2}\right)$$

Note that there are only two parameters to this distribution, namely m_X and σ_X^2

Exercise: show that the mean and variance of this RV are indeed m_X and σ_X^2.

Based on this it is sufficient to say that the variable is Gaussian distributed with some specified mean and variance, and we often use the shorthand notation:

$$X \sim N\left(m_X, \sigma_X^2\right)$$

where the $\sim$ is often used in probability theory to indicate how a RV is distributed (it does not mean "equals to").

12.1.3.5 Normalized Gaussian, and the Q function

The *normalized* Gaussian random variable has $m_X = 0$, variance $\sigma_X^2 = 1$, i.e. $X \sim N(0,1)$. It has a PDF given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

The area under the tail of the normalized Gaussian PDF is used so often in communication theory that we give it a special name, the Q-function:

$$\begin{aligned} Q(x) &\triangleq P(X \geq x) \\ &= \int_x^\infty f_X(t)\, dt \\ &= \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{t^2}{2}\right) dt \end{aligned}$$

Mathematicians like the *complementary error function* is given by

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp\left(-t^2\right) dt$$

These are related by

$$Q(x) = \frac{1}{2}\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right)$$

This relationship is illustrated in Figure 12.1.2.

As we shall see later on, the so-called "Q-function" $Q(x)$ is useful in analyzing the performance of *digital* communication systems.

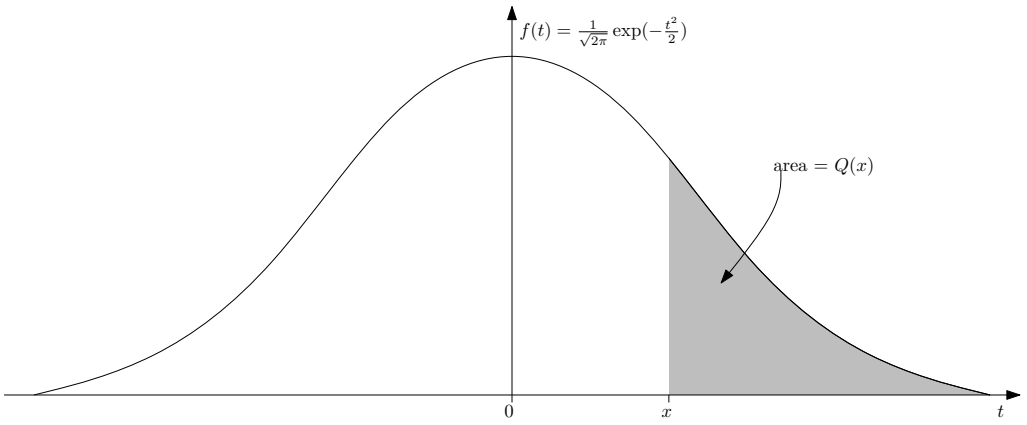


Figure 12.1.2: The Q function and its relation to the normalized Gaussian PDF.

12.1.3.6 Central limit theorem

The central limit theorem (not proven here) says that if you add together enough i.i.d.[3] RVs the result has a Gaussian distribution irrespective of the original distribution, in fact:

$$\lim_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}Y_n\right) \sim N\left(m_Y,\sigma_Y^2\right)$$

this is why we find the Gaussian RVs everywhere - electronic engineering is no exception, i.e. in a large circuit there are many independent noise sources which can (with care) all be lumped together into one Gaussian contribution resulting in dramatically reduced complexity analysis.

12.1.4 Correlated RVs

Thus far we've only considered independent RVs, i.e. ones where the outcome is independent of anything else. Before considering random processes we'd take a brief look at correlated RVs. Consider the circuit of Figure 12.1.3.

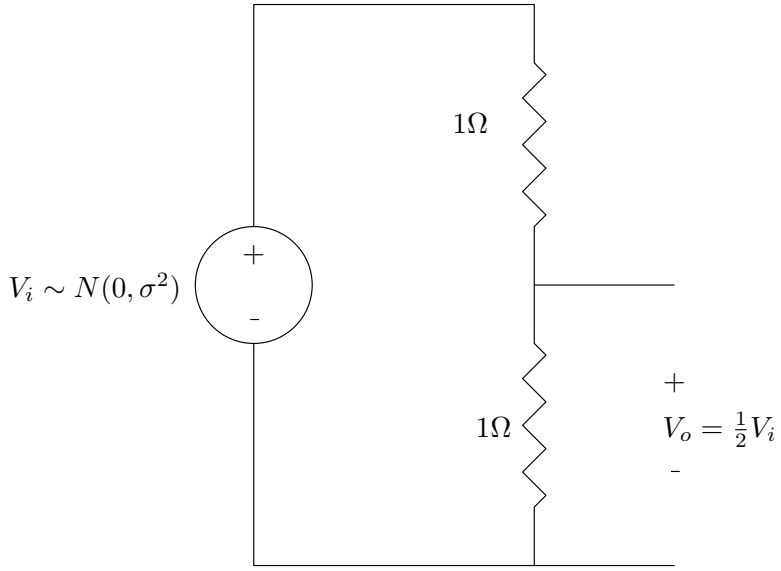


Figure 12.1.3: Division of a random voltage produces a correlated random variable

Clearly we have:

$$\begin{aligned} V_i &\sim N\left(0,\sigma^2\right) \text{ and} \\ V_o &\sim N\left(0,\frac{1}{4}\sigma^2\right) \end{aligned}$$

[3] independent and identical distributed.

However THEY ARE NOT INDEPENDENT.

By that we mean that if we know one of them then we know something about the other, in this case we know *everything* about the other; we say the two RVs are correlated.

We measure correlation between two RVs X and Y by the *correlation coefficient* $\rho_{X,Y}$, defined as:

$$\rho_{X,Y} = \frac{E\left[(X - m_X)(Y - m_Y)\right]}{\sigma_X \sigma_Y}$$

Exercise: Shown that for our voltage divider circuit $\rho_{v_i,v_o} = 1$..

In general we have:

- $\rho = 0$ $\Rightarrow$the variables are independent
- $0 < \rho < 1$ $\Rightarrow$the variables are somewhat correlated, i.e. knowing one tells us something (but not everything) about the other.
- $\rho = 1$ $\Rightarrow$the variables are 100% dependent, i.e. knowing one tells us everything about the other.

12.2 Stationary Random Processes

A *random process*, or *a stochastic process*, is some process, $X(t)$, that produces a output, $x(t)$, a function of time, that can only be described statistically, examples include:

- A machine that automatically flips a coin - this is discrete random process as the possible values are either "heads" or "tails".
- The number of customers arriving per unit time at a shop - this is discrete random process as the possible values are $1, 2, 3 \cdots$, etc. Often this is modeled by with a single "rate of arrival" parameter.
- A resistor; the voltage across it always contains a random component that varies with time - the noise is often zero-mean Gaussian distributed $\sim N(0, \sigma^2)$

In theory all of these example processes are themselves time varying, i.e. their statistics vary with time, i.e.

- After many years of operation the coin flipping machine will have statistics that differ from its original setting
- We'd expect the rate of arrival of customers in a shop to be different at night compared to day time.
- The noise across a resistor varies with temperature, so there will likely be variations over the day and over the year.

Therefore these are examples of *non-stationary random processes*.

However!! non-stationary random processes are difficult to deal with - so in engineering we usually assume the processes are stationary over the time-span of our analysis.
We will not consider non-stationary processes any further!

A *stationary random process* is one whose statistical parameters do not vary with time.

In communication theory, the study of random processes allows us to do three things:

1. The noise in the communication channel is a random process - characterizing this process allows us to work out the effect of the channel noise on the performance of the communication system.

2. Having done this, we may go further and design the communication system (transmitter and receiver) in order to minimize this effect.

3. The information signal may also be regarded as a random process, from the receiver's point of view. If the receiver had knowledge of the information signal, there would be no need for the communication. Therefore the information signal also needs to be characterized statistically.

Let X be a stationary random process having an output, $x(t)$, then we can consider this as a random variable - no different to the random variables in the previous session, except we need to include t, so we use slightly different notation: It's CDF is:

$$F_X(x;t) = P(X(t) \leq x(t))$$

and its PDF:

$$f_X(x;t) = \frac{d}{dx} F_X(x;t)$$

For example, a very common process called a *Gaussian process*, $f_X(x;t)$ has a Gaussian distribution for all t:

$$f_X(x;t) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{(x(t) - m_X)^2}{2\sigma_X^2}\right)$$

The mean of X is:

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

Which is a constant (over time) as we assumed this is a stationary process.

12.2.1 Auto-correlation function

So for, with the exception of the t parameter in some of the above definitions the aspects of stationary random processes considered are no different from regular independent random variables, i.e. the PDF and CDF allow use to describe the statistics of the output from $X(t)$ at some time t_o without reference to any other time t_1. To examine how the output from a random process varies over time we need some additional statistical tools.

12.2.1.1 Motivational example

Consider the two signals $x(t)$ and $y(t)$ shown in Figure 12.2.1. These were generated in Matlab.

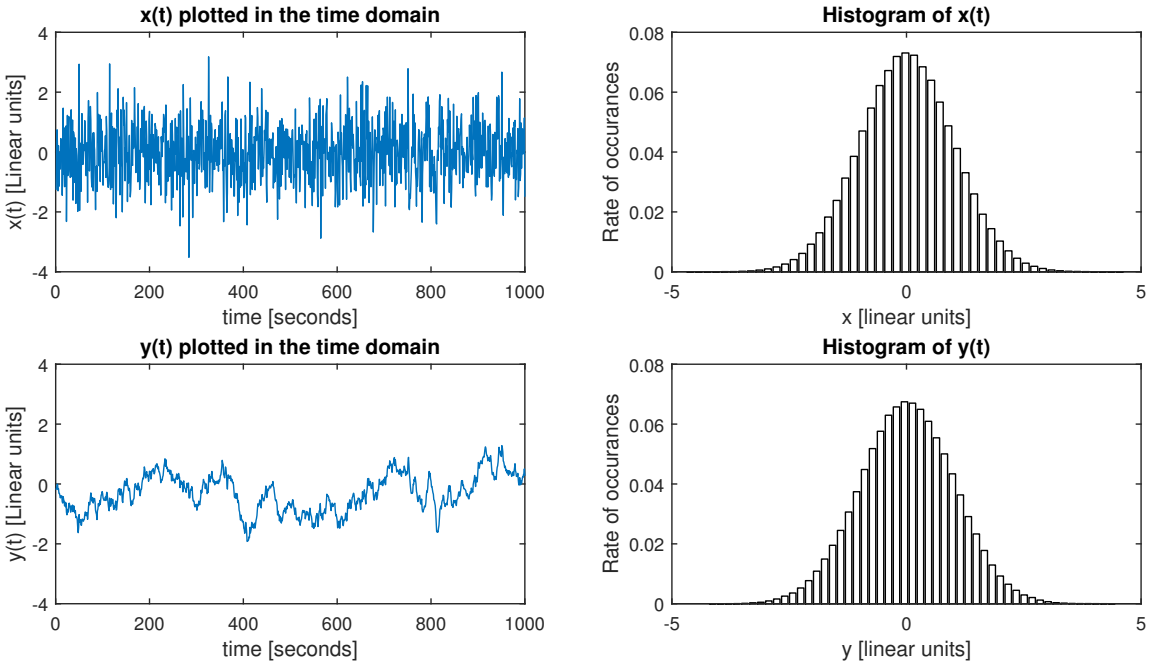


Figure 12.2.1: Two signal with same PDF, but different auto-correlation functions.

What do we notice?

Based on the histograms they both appear to be Gaussian distributed - that is correct, in fact they are both $\sim N(0,1)$.

But clearly they "look" very different in the time domain - how can this be?

I know (because I generated them) that each sample of $x(t)$ is independent of any other - however the samples of $y(t)$ are not - they are correlated with each other. If I know $y(t_1)$ then I know something (but not everything) about $y(t_2)$, we could say they have a auto-correlation coefficient of $\rho_{Y(t_1),Y(t_2)}$.
Of course this is different depending on the exact time instances t_1 and t_2 so we define the *auto-correlation* function of $Y(t)$ is defined as

$$R_Y(t_1,t_2) = E[Y(t_1)Y(t_2)]$$

which differs from the definition of the auto-correlation coefficient by the absence of the nor-

malization factors.[4]

This measures the correlation between two samples of the process at two different times t_1, t_2.

12.2.2 Wide Sense Stationary (WSS) Random Processes

A random process is *strict sense stationary* if all its statistics do not vary with time. A looser but widely used definition is that of Wide Sense Stationarity. A random process is *Wide Sense Stationary* (WSS) if just the following two conditions are satisfied:

- The mean is independent of time, i.e.

$$E\left[X\left(t\right)\right] = m_X \quad \forall t$$

- The auto-correlation function depends only on the time difference between samples (leading to the follow abuse notation):

$$R_X\left(\tau\right) \;=\; R_X\left(t, t-\tau\right) \quad \forall t$$

  where τ is a time difference. The point here is that the auto-correlation function only depends on this time shift, i.e. the time between t_1 and t_2 , but not on t_1 nor t_2 persa.

12.2.3 Ergodic Random Processes

The time average of the output from a stationary random process $X\left(t\right)$ over an interval of length T is given by

$$\bar{m}_X\left(T\right) = \frac{1}{T}\int_{-T/2}^{T/2} X\left(t\right)\, dt$$

If $\bar{m}_X\left(T\right) \to m_X$ as $T \to \infty$, then X is *ergodic in the mean*.

The auto-correlation of the output from a stationary random process $X\left(t\right)$ over an interval of length T is given by

$$\bar{R}_X\left(\tau, T\right) = \frac{1}{T}\int_{-T/2}^{T/2} X\left(t\right) X\left(t+\tau\right)\, dt$$

[4]The mean is not removed which makes no difference for zero mean signal as is often the case. The normalization scaling factor is generally omitted in signal processing - this convention allows a compact notation for the *Wiener-Khintchine* theorem which we will see very soon.

If $\bar{R}_X(\tau, T) \to R_X(\tau)$ as $T \to \infty$, then X is *ergodic in the auto-correlation function*.

Ergodicity is a nice property for a random process to have, as it means that we can estimate the 'true' mean and auto-correlation by substituting measured time averages, and these estimates become accurate as $T \to \infty$.

Also note that an ergodic process must be stationary, but a stationary process need not necessarily be ergodic.

We will assume that all random processes are WSS, and ergodic in the mean and auto-correlation.

Chapter 13

Power Spectral Density (PSD)

The auto-correlation function of a zero-mean wide-sense stationary random process $x(t)$ is given by.

$$R_x(\tau) = E[x(t)\, x(t+\tau)]$$

This function has the properties:

1. $R_x(0) = E[x^2(t)] = P_X$ is the *average power* of the process.
2. $R_x(\tau)$ is an even function, i.e. $R_x(-\tau) = R_x(\tau)$

The *power spectral density* (PSD) is given by

$$S_x(f) = \lim_{T\to\infty} \frac{1}{T} E\left[|X_T(f)|^2\right] \tag{13.0.1}$$

where $X_T(f)$ is the Fourier transform of the truncated signal

$$x_T(t) = \begin{cases} x(t) & \text{if } -\frac{T}{2} < t < \frac{T}{2} \\ 0 & \text{otherwise} \end{cases}$$

It may be shown (proof omitted) that

$$P_x = E\left[x^2(t)\right] = \int_{-\infty}^{+\infty} S_x(f)\, df = 2\int_0^{+\infty} S_x(f)\, df$$

This gives two methods of evaluating the average power in the process - one in the time domain, and one in the frequency domain.

Since $S_x(f)$ contains contributions from both positive and negative frequencies, it is called the *two-sided* PSD.

We may define a *one-sided* PSD as $S'_x(f) \triangleq 2S_x(f)$ for $f > 0$.
It can be shown that $\int_{f_1}^{f_2} S'_x(f)\, df$ represents the power in the process lying in the band of frequencies between f_1 and f_2.
In particular, the power contained in a very small band of width Δf centered on f_c is approximately $S'_x(f_c)\,\Delta f$; hence the name "power spectral density".

It may also be shown that the auto-correlation function and the power spectral density are a Fourier transform pair, i.e.

$$\begin{aligned} S_x(f) &= \int_{\infty}^{\infty} R_x(\tau)\, e^{-j2\pi f\tau}\, d\tau \\ R_x(\tau) &= \int_{\infty}^{\infty} S_x(f)\, e^{j2\pi f\tau}\, df \end{aligned}$$

This is known as the *Wiener-Khintchine theorem*.

If $S_x(f)$ is constant for all f, then the process has equal power contribution from all frequencies, hence the name "white noise".

Example: Find the PSD of a train of randomly weighted impulses

$$x(t) = \sum_{n=-\infty}^{\infty} a_n \delta(t - nT)$$

where

$$E[a_n a_{n+k}] = \begin{cases} \sigma^2 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases}.$$

The solution (not given here is):

$$S_X(f) = \frac{1}{T}\sigma^2$$

So the PSD is white! Happy days.

Notes:

- that we didn't assume anything about the distribution of the a_n, we simply insisted that they were independent - so this works for Gaussian, uniform, or indeed any distribution...
- Also note that this signal looks very different to AWGN noise, but they both have the same PSD!! so the PSD by no means tells us everything about a random process.

13.1 Response of a linear system

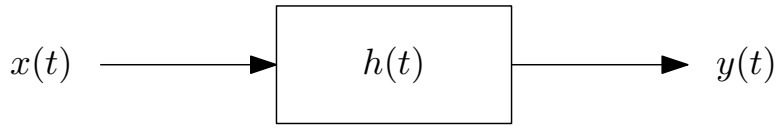


Figure 13.1.1: The random process $x(t)$, passed through an LTI system with impulse response $h(t)$, results in another random process $y(t)$.

If a random process $x(t)$ is applied as input to an LTI system with transfer function $H(f)$, then the output $y(t)$ will also be a random process. It may also be shown (proof omitted) that

$$S_y(f) = S_x(f)\,|H(f)|^2 \tag{13.1.1}$$

This is illustrated in Figure 13.1.2 where the system, $h(t)$, is a low-pass filter.

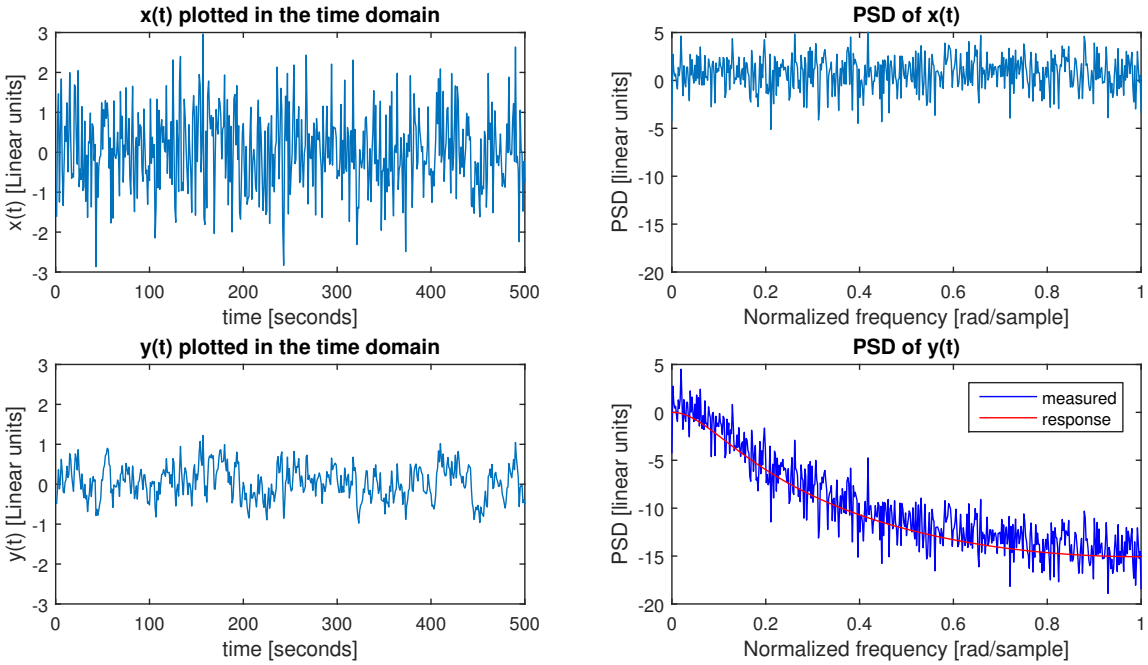


Figure 13.1.2: The PSD of input, $S_x(f)$, is white, but the output PSD, $S_y(f)$, has being shaped by the frequency response of the system $|H(f)|$.

Chapter 14

Noise in Communication Receivers

14.1 Thermal Noise

Consider a resistor R, open circuit, at some temperature T (in Kelvin).
An ideal resistor would have zero voltage between its terminals, i.e. $v(t) = 0$ for all t.
A real resistor contains a large number of electrons, all in random motion. On average, electrons are uniformly distributed throughout the resistor - there should be no tendency for a higher electron density at one end than at the other, so the average terminal voltage is zero:

$$E[v(t)] = 0 \quad \forall t$$

However, there are local, short-term variations in the electron density. These cause small variations in the terminal voltage $v(t)$.
Thermodynamic analysis gives the mean-square voltage as

$$E\left[v^2(t)\right] = 4kTRB \tag{14.1.1}$$

where

- $k =$ Boltzmann's constant, $1.38 \times 10^{-23} J/C$
- T is the absolute temperature, typically $290K$
- B is the bandwidth of the system or measurement (in Hz).

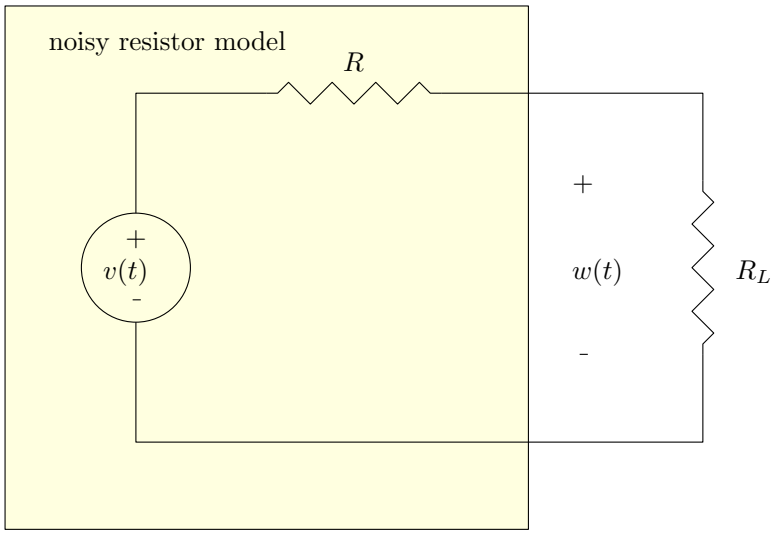


Figure 14.1.1: A noisy resistor viewed as a noise voltage source in series with an ideal resistor.

We can model this noisy resistor as a noise voltage source $v(t)$, in series with an ideal resistance R (see figure 14.1.1).

In a communication receiver, the load is usually matched[1], which means that $R_L = R$. Hence $w(t) = v(t)/2$, and so the noise power delivered to the load is $w^2(t)/R = v^2(t)/4R$. The mean (or expected value) of this dissipated power is therefore $P = E[v^2(t)/4R] = kTB$, from equation (14.1.1).
This maximum dissipated power is called the *available noise power*, and is independent of the value of the resistance R.
Note that the available power is proportional to the bandwidth B, hence we say the noise has power spectral density $N_0 = kT$ Watts/Hz.

A thermal noise voltage $w(t)$ is a random process with the following properties:

- Wide-sense stationary - mean and auto-correlation function do not change with time
- Ergodic - time averages tend to the mean and auto-correlation function
- Gaussian - samples of the process have Gaussian probability density function (this is due to the central limit theorem)
- Zero mean - there is no long-term drift of electrons in one direction

[1] From electronics we know that the maximum amount of signal power is delivered to the load when the impedances are matched, so this is usually arranged.

- Constant power spectral density $S_w(f) = N_0/2$, where $N_0 = kT$ - noise power is spread uniformly across the frequency spectrum - called *white noise*.

Note that in reality, the PSD is not constant over all frequencies, but drops at high frequency. This is because real resistors have some stray (parasitic) inductance and capacitance, which forms a low-pass filter, limiting the bandwidth at high frequency. However, the white noise model still serves as a good model in communication theory.

By superposition, the noise is added to the received signal, and is called Additive White Gaussian Noise (AWGN); this is shown in Figure 14.1.2.

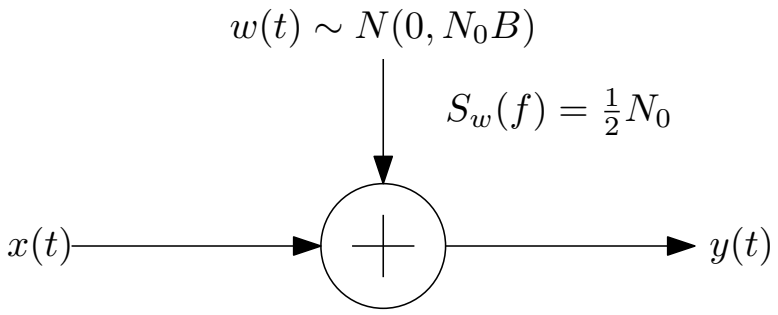


Figure 14.1.2: A simple AWGN model. The input, $x(t)$, has White Gaussian Noise, $w(t)$, added to it to form the output $y(t) = x(t) + w(t)$

14.2 Filtered Noise

Communication systems operate using signals of finite bandwidth.

The receiver normally includes a bandpass filter, which will only pass frequencies within the signal bandwidth $f_c \pm B$ - let us call this filter $H(f)$.

Thus whatever noise is added in the channel will be filtered before it reaches the demodulator.

Denoting the wideband noise by $w(t)$ as before, and its power spectral density by $S_w(f)$, we have

$$S_n(f) = |H(f)|^2 S_w(f)$$

This gives us the expression for the PSD of the noise at the filter output.

In receivers for modulated signals, the signal bandwidth is usually very small relative to the center (carrier) frequency f_c. Thus the filtered noise at the demodulator input also has small bandwidth $2B$, relative to its center frequency, and is called *narrow-band noise*.

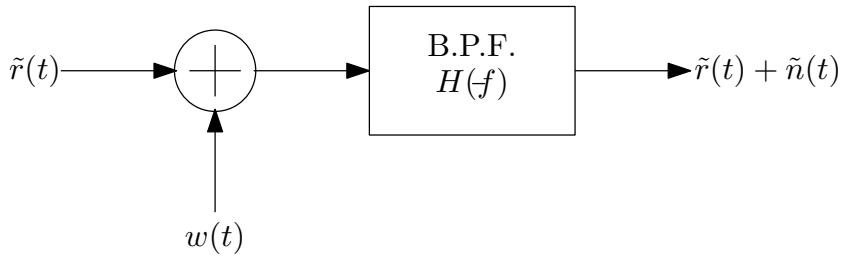


Figure 14.2.1: The receive signal is filtered by the bandpass filter $H(f)$ prior to demodulation. This is in order to reduce the effect of noise. The information-bearing signal $\tilde{r}(t)$ passes through unaltered, but the noise is changed from the wideband noise process $w(t)$ to a narrowband noise process $\tilde{n}(t)$.

If the noise at the filter input is white, with two-sided PSD $N_0/2$,

$$S_n(f) = \frac{N_0}{2} |H(f)|^2$$

so the power in the narrowband noise is

$$P_n = N_0 \int_0^\infty |H(f)|^2 \, df$$

Suppose that, instead of a real filter, the receive filter was an *ideal* bandpass filter with unity gain over a passband of width B_N. Then the narrowband noise power would be

$$P_n = N_0 \int_0^\infty |H(f)|^2 \, df = N_0 B_N$$

Based on this comparison, we define the *noise equivalent bandwidth* of a filter $H(f)$ by

$$B_N = \int_0^\infty |H(f)|^2 \, df$$

Therefore, an ideal bandpass filter over the bandwidth B_N gives the same noise power at its output as the actual filter with transfer function $H(f)$.

14.3 Quadrature Components of Narrowband Noise

Since the bandpass filtered white noise $\tilde{n}(t)$ is a narrowband noise signal, it may be represented in the form

$$\tilde{n}(t) = n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)$$

where $n_c(t)$ and $n_c(t)$ are the quadrature components of $\tilde{n}(t)$. The quadrature components have the following properties:

- $n_c(t)$ and $n_s(t)$ are statistically identical baseband random processes, with bandwidth B
- $n_c(t)$ and $n_s(t)$ are zero-mean and Gaussian
- $n_c(t)$ and $n_s(t)$ are wide sense stationary
- The power is each of the quadrature components is the same as that in the original noise process, i.e.
$$E\left\{n_c^2(t)\right\} = E\left\{n_s^2(t)\right\} = E\left\{\tilde{n}^2(t)\right\}$$
- The PSD of each of the quadrature components is the same, and is related to that the PSD of the original noise process by
$$S_{n_c}(f) = S_{n_s}(f) = \begin{cases} S_{\tilde{n}}(f - f_c) + S_{\tilde{n}}(f + f_c) & \text{if } |f| < B \\ 0 & \text{otherwise} \end{cases}$$

Chapter 15

Analog demodulation with noise

In this section of the course we are interested in analysing the performance of analog modulation schemes and their associated receivers. But how do we put a measure on the quality of reception? This question is answered by considering the receiver *output*, i.e. the point at which the information signal is reconstructed. We take as our performance measure the *output signal-to-noise ratio*, or *output SNR*. This is the ratio of signal power to noise power at the receiver output:

$$\left(\frac{S}{N}\right)_o = \frac{P_{so}}{P_{no}} = \frac{\text{Signal Power at Receiver Output}}{\text{Noise Power at Receiver Output}}$$

15.1 AM in the Presence of Noise

Consider the communication receiver for a full AM signal, shown in figure 15.1.1. We have already analysed the receiver in the absence of noise; our goal now is to characterise performance in the presence of additive white Gaussian noise (AWGN).
Recall:

- The information signal $g(t)$ is baseband, of bandwidth W. It has average power $P_g = E\{g^2(t)\}$.
- The received AM signal is bandpass, of bandwidth $2W$, and is given by

$$\tilde{r}(t) = A_r\{1 + k_a g(t)\}\cos(2\pi f_c t)$$

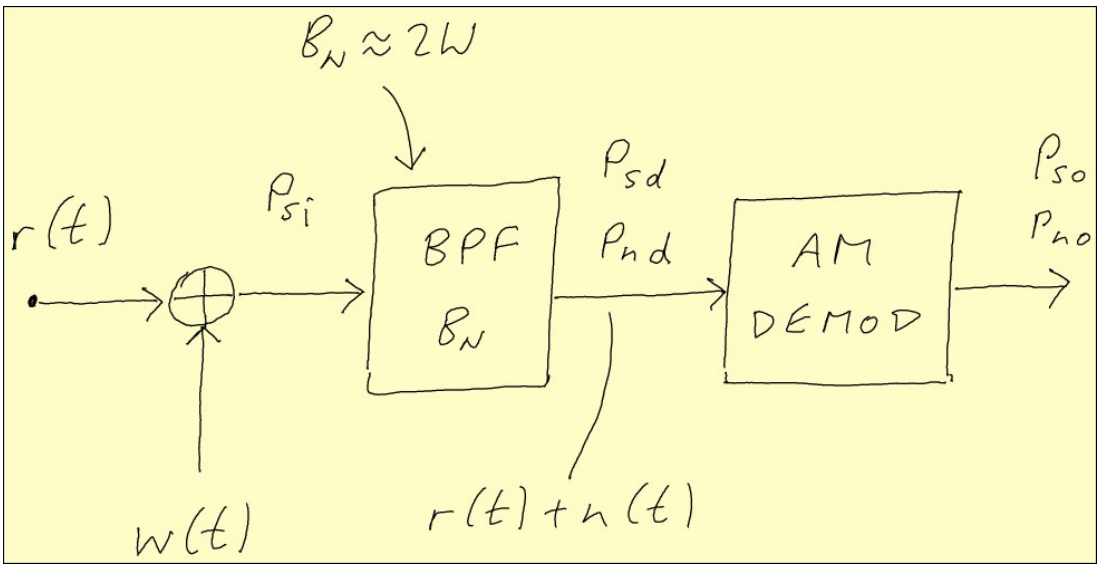


Figure 15.1.1: AM receiver.

Therefore, the signal power at the receiver input is

$$P_{si} = \frac{A_r^2}{2}\left\{1 + k_a^2 P_g\right\}$$

Assume white Gaussian noise $w(t)$ is added in the channel, of PSD $S_w(f) = N_0/2$. The received signal $\tilde{r}(t) + w(t)$ is then bandpass filtered to reduce the effect of noise. We assume that the signal component $\tilde{r}(t)$ passes unaltered, but the wideband noise $w(t)$ is filtered to produce narrowband noise $\tilde{n}(t) = n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)$. Therefore, the signal power at the demodulator input is $P_{sd} = P_{si}$. The noise equivalent bandwidth of the BPF is B_N, so the noise power at the demodulator input is $P_{nd} = N_0 B_N$. We define the signal-to-noise ratio (SNR) at the demodulator input as

$$\left(\frac{S}{N}\right)_d = \frac{P_{sd}}{P_{nd}} = \frac{A_r^2\left\{1 + k_a^2 P_g\right\}}{2N_0 B_N}$$

The demodulator input is

$$\tilde{r}(t) + \tilde{n}(t) = \left[A_r\left\{1 + k_a g(t)\right\} + n_c(t)\right]\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)$$

15.1.1 Synchronous Demodulator

In the case of synchronous demodulation, we multiply by a local oscillator $l(t) = A_l\cos(2\pi f_c t)$ to get

$$A_l\left[A_r\left\{1 + k_a g(t)\right\} + n_c(t)\right]\cos^2(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)\cos(2\pi f_c t)$$

then LPF to get

$$\frac{1}{2}A_l\left[A_r + A_r k_a g\left(t\right) + n_c\left(t\right)\right]$$

Therefore the demodulator output consists of a DC component, an information signal component and an additive noise component. To simplify, let $A_l = 2$ (it will scale the signal and noise equally, and therefore will not affect the SNR). Then:

- Signal power at demodulator output:

$$P_{so} = E\left\{\left(A_r k_a g\left(t\right)\right)^2\right\} = A_r^2 k_a^2 P_g$$

- Noise power at demodulator output:

$$P_{no} = E\left\{n_c\left(t\right)^2\right\} = E\left\{n\left(t\right)^2\right\} = P_{nd} = N_o B_N$$

- Signal-to-noise ratio at demodulator output:

$$\left(\frac{S}{N}\right)_o = \frac{P_{so}}{P_{no}} = \frac{A_r^2 k_a^2 P_g}{N_0 B_N}$$

The ratio of SNRs at the input and output of the demodulator is then

$$\left(\frac{S}{N}\right)_o \Big/ \left(\frac{S}{N}\right)_d = \frac{2k_a^2 P_g}{1 + k_a^2 P_g} \tag{15.1.1}$$

Recall that a practical AM system has $k_a^2 P_g \ll 1$; therefore this ratio is small due to the transmission of the carrier signal. Therefore, full AM gives poor SNR performance**.**

15.1.2 Envelope Detector

The envelope detector input is

$$\tilde{r}\left(t\right) + \tilde{n}\left(t\right) = \left[A_r\left\{1 + k_a g\left(t\right)\right\} + n_c\left(t\right)\right]\cos\left(2\pi f_c t\right) - n_s\left(t\right)\sin\left(2\pi f_c t\right)$$

Assuming an ideal envelope detector, the output $m\left(t\right)$ is equal to the envelope of the input, i.e.

$$m\left(t\right) = \sqrt{\left[A_r\left\{1 + k_a g\left(t\right)\right\} + n_c\left(t\right)\right]^2 + n_s^2\left(t\right)}$$

This can be seen easily from Figure 15.1.2.

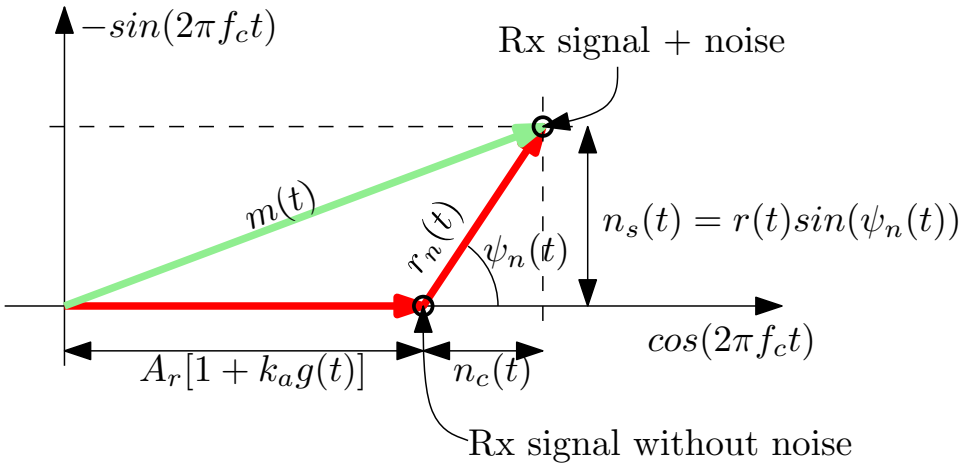


Figure 15.1.2: Additive noise in AM can be considered as adding a vector with magnitude $r_n(t)$ and angle $\psi_n(t)$ to the received signal given by $A_r\{1 + k_a g(t)\}$ when plotted on the axis defined by $\cos(2\pi f_c t)$ and $-\sin(2\pi f_c t)$. The output from an envelope detector is just the magnitude of the resulting vector, i.e. $m(t)$ in this figure.

Full analysis here is difficult due to the nonlinearity. We consider two important modes of operation:

15.1.2.1 Normal operation (High input SNR):

$$\left(\frac{S}{N}\right)_d \gg 1 \implies A_r^2 \gg n_s^2(t) \implies m(t) \approx A_r\{1 + k_a g(t)\} + n_c(t)$$

So for normal operation, we get the same result as for synchronous demodulation.

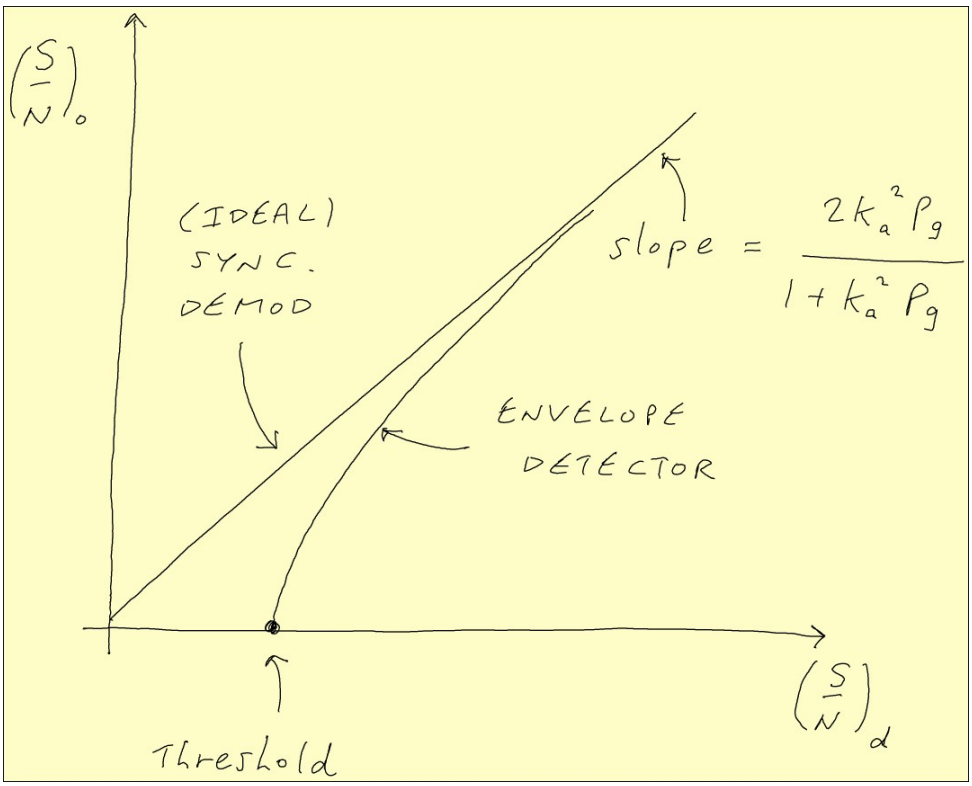


Figure 15.1.3: Demodulator output SNR versus demodulator input SNR for an AM receiver. The envelope detector exhibits a threshold effect, i.e. $(S/N)_d$ needs to be above the threshold value indicated in order to produce any useful output. The ideal synchronous demodulator maintains a linear relationship between $(S/N)_o$ and $(S/N)_d$. In practice, this relationship will fail at low $(S/N)_d$ due to loss of LO synchronization.

15.1.2.2 Low input SNR:

For low input SNR, the analysis is made easier by using the envelope-phase representation of narrowband noise:

$$n_c(t) = r_n(t)\cos\psi_n(t)$$

$$n_s(t) = r_n(t)\sin\psi_n(t)$$

Applying the cosine rule, and letting $A(t) = A_r\{1 + k_a g(t)\}$, we have:

$$\begin{aligned} m^2(t) &= A^2(t) + r_n^2(t) - 2A(t)\,r_n(t)\cos(\pi - \psi_n(t)) \\ &= A^2(t) + r_n^2(t) + 2A(t)\,r_n(t)\cos(\psi_n(t)) \end{aligned}$$

Now, at VERY low input SNR, then $A^2(t) \ll r_n^2(t)$, and

$$\begin{aligned} m^2(t) &\approx r_n^2(t) + 2A(t) r_n(t) \cos(\psi_n(t)) \\ \Rightarrow m(t) &\approx r_n(t) \sqrt{1 + 2\frac{A(t)}{r_n(t)} \cos\psi_n(t)} \end{aligned}$$

Here we may use the binomial approximation $(1+x)^{1/2} \approx 1 + \frac{1}{2}x$, which is valid for $x \ll 1$ (which it will be under the low SNR assumption). Therefore

$$\begin{aligned} m(t) &\approx r_n(t) \left(1 + \frac{A(t)}{r_n(t)} \cos\psi_n(t)\right) \\ &= r_n(t) + A(t) \cos\psi_n(t) \end{aligned}$$

Now putting back $A(t) = A_r \{1 + k_a g(t)\}$, we get:

$$m(t) \approx r_n(t) + A_r \cos\psi_n(t) + A_r k_a g(t) \cos\psi_n(t)$$

The information signal is now *multiplied* by a random process, $\cos\psi_n(t)$. Therefore $g(t)$ cannot be recovered, and all information is lost!!!

The phenomenon of no useful output below a certain input SNR is called the *threshold effect*, and is found in all *non-linear* demodulators. The ideal synchronous demodulator has output SNR proportional to input SNR, even as the input SNR goes to zero (by equation (15.1.1)).
In reality, synchronization of the local oscillator will at some point fail.
These considerations are illustrated in figure 15.1.3.

15.2 FM in the Presence of Noise

Our goal in this section is to characterise the performance of an FM receiver, shown in figure 15.2.1, in the presence of AWGN. We will consider wideband FM only, for reasons which will become clear after the analysis is complete. The receiver diagram is similar to that of the previous section, with the AM demodulator replaced by an FM demodulator. However, there are two other differences. First, the noise equivalent bandwidth, B_N, of the BPF is approximately $2(D+1)W$, which is, in general, a good deal larger than $2W$. Note also that the demodulator output is passed through a further LPF to obtain the receiver output; the

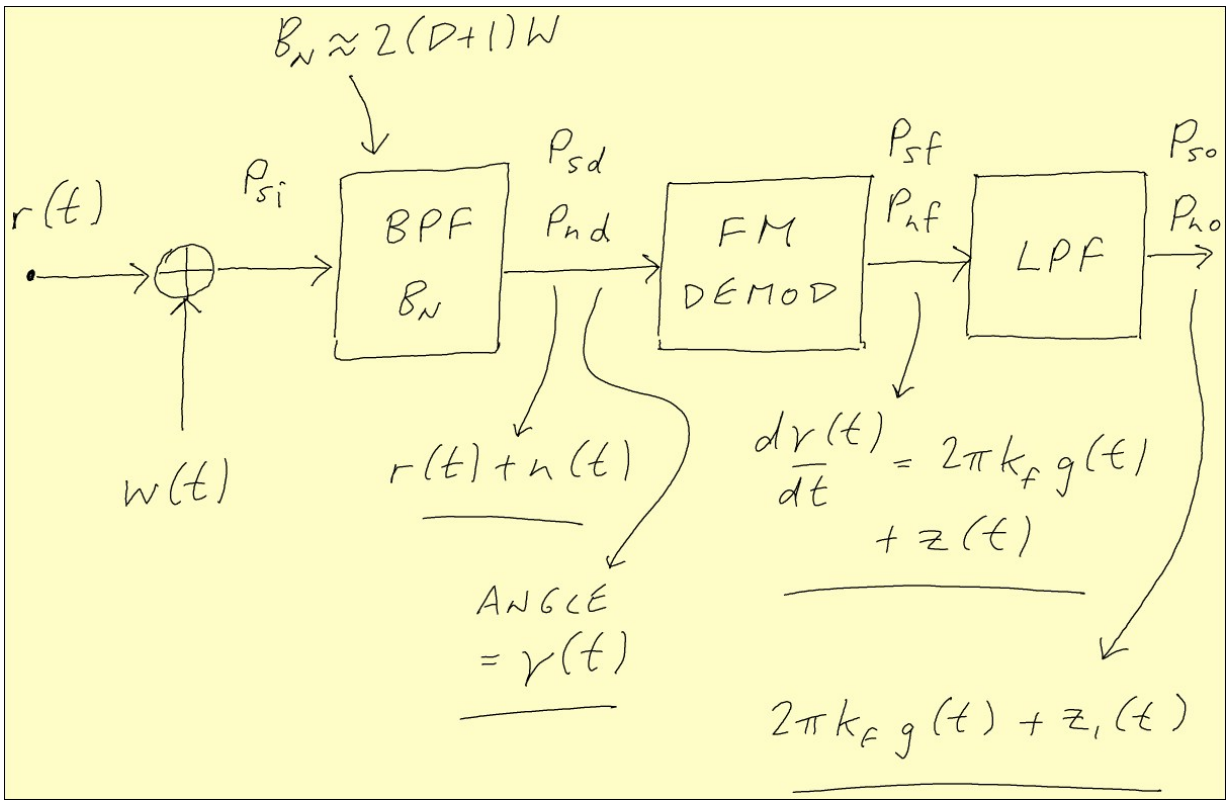


Figure 15.2.1: FM receiver.

reason for this final LPF will become clear as we perform the analysis of this receiver.

Recall:

- The information signal $g(t)$ is baseband, of bandwidth W. It has average power $P_g = E\{g^2(t)\}$.
- The received FM signal is $r(t) = A_r \cos(2\pi f_c t + \theta(t))$, with

$$\frac{d\theta(t)}{dt} = 2\pi k_f g(t)$$

Therefore, the signal power at receiver input is

$$P_{si} = \frac{A_r^2}{2}$$

As in the previous section, the signal at the output of the BPF is $\tilde{r}(t) + \tilde{n}(t)$, where $\tilde{n}(t) = n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)$. Therefore, the signal power at the demodulator input is $P_{sd} = P_{si}$. The noise equivalent bandwidth of the BPF is B_N, so the noise power at the demodulator input is $P_{nd} = N_0 B_N$. The SNR at the demodulator input is

$$\left(\frac{S}{N}\right)_d = \frac{P_{sd}}{P_{nd}} = \frac{A_r^2}{2N_0 B_N}$$

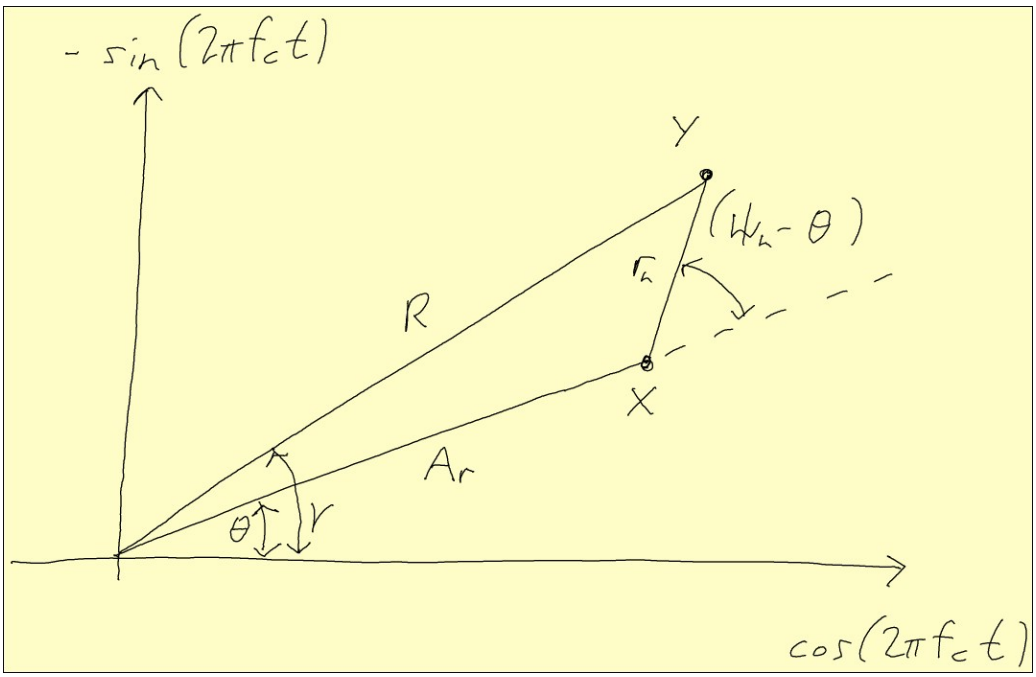


Figure 15.2.2: Phasor diagram illustrating the effect of adding the noise signal $\tilde{n}(t)$ to the information-bearing FM signal $\tilde{r}(t)$. In the absence of noise, the receive signal corresponds to the point X in the diagram; with additive noise $n(t)$, the receive signal corresponds to the point Y. Note that the angle of the received signal is corrupted by the noise from $\theta(t)$ to $\gamma(t)$.

We use the envelope-phase representation for the narrowband noise,

$$\tilde{n}(t) = r_n(t)\cos(2\pi f_c t + \psi_n(t))$$

Thus the demodulator input is

$$\tilde{r}(t) + \tilde{n}(t) = A_r\cos(2\pi f_c t + \theta(t)) + r_n(t)\cos(2\pi f_c t + \psi_n(t))$$

The phasor diagram of figure 15.2.2 illustrates the effect of the noise $n(t)$ on the received signal. The angle of the resultant signal is $\gamma(t)$, which has signal component, $\theta(t)$, plus a component due to noise.

From figure 15.2.2 we can see that

$$\gamma(t) = \tan^{-1}\left[\frac{r_n(t)\sin(\psi_n(t) - \theta(t))}{A_r + r_n(t)\cos(\psi_n(t) - \theta(t))}\right]$$

Consider *normal* operation, with large $(S/N)_d$, so $A_r \gg r_n(t)$. Then we may use the approx-

imation $\tan x \approx x$ for small x, giving

$$\gamma(t) \approx \theta(t) + \frac{r_n(t)}{A_r} \sin(\psi_n(t) - \theta(t)) \tag{15.2.1}$$

From this equation we may note:

- The signal component, $\theta(t)$, is independent of the received signal amplitude, A_r.
- The noise component *reduces* as A_r increases. This is called the *noise quieting effect*.

Assuming an ideal FM demodulator, the demodulator output is proportional to $d\gamma(t)/dt$ (in the absence of noise, this would be $d\theta(t)/dt = 2\pi k_f g(t)$). From equation (15.2.1) we get

$$\begin{aligned} \frac{d\gamma(t)}{dt} &= \frac{d\theta(t)}{dt} + \frac{1}{A_R}\frac{d}{dt}\left[r_n(t)\sin(\psi_n(t) - \theta(t))\right] \\ &= 2\pi k_f g(t) + \frac{1}{A_R}\frac{d}{dt}\left[r_n(t)\sin(\psi_n(t) - \theta(t))\right] \end{aligned}$$

Therefore, the signal power at the demodulator output is

$$P_{sf} = E\left\{(2\pi k_f g(t))^2\right\} = (2\pi k_f)^2 P_g$$

The noise power at the demodulator output is easier to calculate if we consider an unmodulated carrier ($g(t) = \theta(t) = 0$):

$$\frac{d\gamma(t)}{dt} = \frac{1}{A_R}\frac{d}{dt}\left[r_n(t)\sin\psi_n(t)\right] = \frac{1}{A_R}\frac{dn_s(t)}{dt}$$

So we may write

$$\frac{d\gamma(t)}{dt} = 2\pi k_f g(t) + z(t)$$

where the noise signal $z(t)$ is given by

$$z(t) = \frac{1}{A_R}\frac{dn_s(t)}{dt}$$

To simplify the analysis, we will now assume that the BPF is *ideal*. This means that $n_s(t)$ has two-sided PSD

$$S_{n_s}(f) = \begin{cases} N_0 & \text{if } |f| < B_N/2 \\ 0 & \text{otherwise} \end{cases}$$

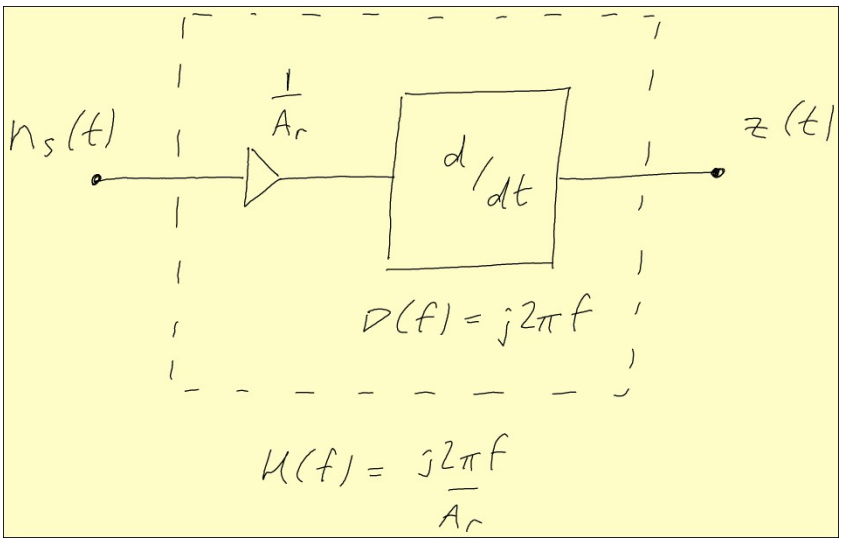


Figure 15.2.3: Illustration of the relationship between $n_s(t)$ and $z(t)$. The overall transfer function of the gain-differentiator combination is $H(f) = j2\pi f/A_r$.

The one-sided PSD of $n_s(t)$ is then

$$S'_{n_s}(f) = \begin{cases} 2N_0 & \text{if } 0 < f < B_N/2 \\ 0 & \text{otherwise} \end{cases} \tag{15.2.2}$$

Now consider the relationship between $n_s(t)$ and $z(t)$, illustrated in figure 15.2.3. The transfer function of the differentiator is $D(f) = j2\pi f$ and so the transfer function from $n_s(t)$ to $z(t)$ is $H(f) = j2\pi f/A_r$. It follows that the PSD of $z(t)$ is given by $S'_z(f) = S'_{n_s}(f)|H(f)|^2$. Substituting for $S'_{n_s}(f)$ from equation (15.2.2) gives

$$S'_z(f) = \begin{cases} 2N_0(2\pi f)^2/A_r^2 & \text{if } 0 < f < B_N/2 \\ 0 & \text{otherwise} \end{cases}$$

This PSD is illustrated in figure 15.2.4. We may evaluate the power in $z(t)$ by integrating $S'_z(f)$ from 0 to $B_N/2$. However, observe that the signal bandwidth at this point is $W \ll B_N/2$, so it makes sense to pass the demodulator output through a low-pass filter. Assume an ideal LPF; this will pass the information signal but remove most of the noise. So

$$P_{so} = P_{sf} = (2\pi k_f)^2 P_g$$

$$P_{no} = \int_0^W S'_z(f)\, df = \frac{8\pi^2 N_0}{A_R^2} \int_0^W f^2\, df = \frac{8\pi^2 N_0 W^3}{3A_R^2}$$

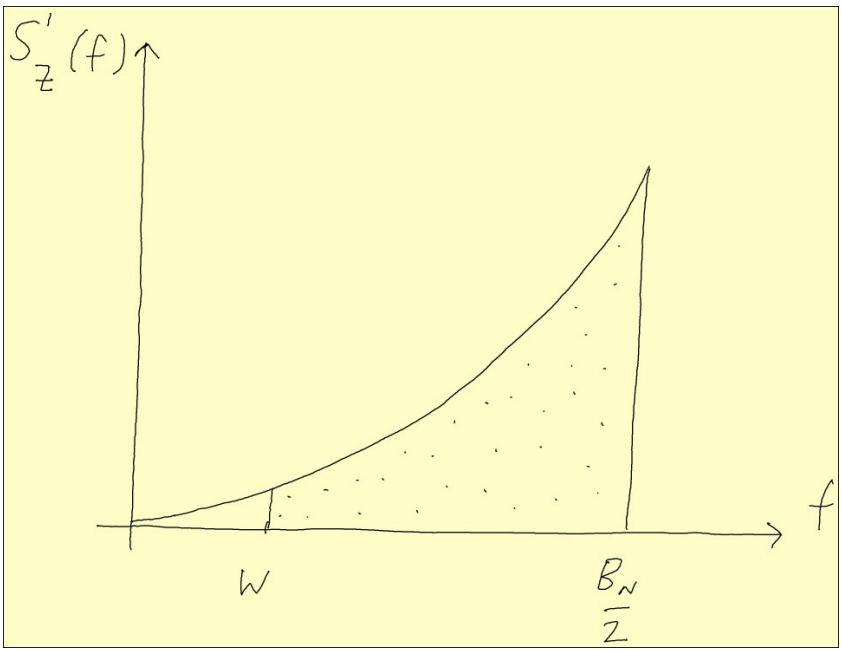


Figure 15.2.4: PSD of the noise at the FM demodulator output. The noise PSD increases as f^2, up to the point $f = B_N/2$. The information-bearing signal occupies the frequency band $0 < f < W \ll B_N/2$. Therefore, a LPF following the demodulator will get rid of most of the noise (dotted region).

Therefore, the SNR at the receiver output is

$$\left(\frac{S}{N}\right)_o = \frac{P_{so}}{P_{no}} = \frac{3A_R^2 k_f^2 P_g}{2N_0 W^3}$$

There are two important points to note about this equation:

- We can increase this SNR as required, simply by increasing k_f
- The penalty for this increase in SNR is a larger peak frequency deviation, implying a larger bandwidth for the FM signal.

From these two points we conclude that FM allows us to *trade bandwidth for SNR*. This is a major advantage of FM over all forms of AM.

Note: It may be shown that the FM demodulator exhibits a threshold effect, i.e. for sufficiently low $(S/N)_d$, $\gamma(t)$ contains no component proportional to $\theta(t)$.

15.2.1 Pre-emphasis and De-emphasis in FM Radio

Note that the noise PSD at the output of the FM demodulator increases as f^2 (see figure 15.2.4). Typical information signals (e.g. audio, video) tend to have less power at the

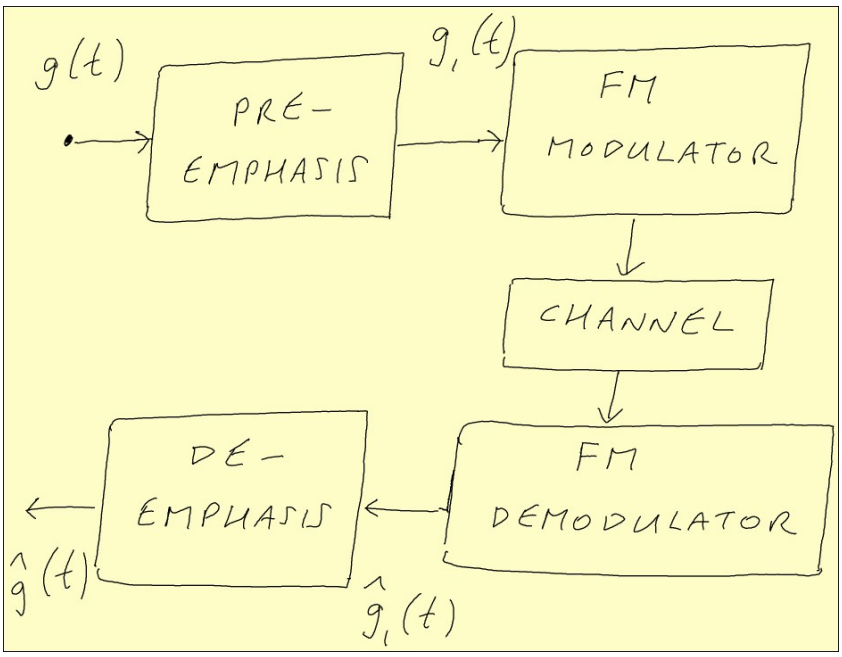


Figure 15.2.5: An FM system using pre-emphasis and de-emphasis.

higher frequencies (near W). These frequency components are therefore highly vulnerable to noise. To overcome this, we amplify higher-frequency components of the information signal (relative to the low-frequency components) before modulation. This is called *pre-emphasis,* and the amplifying filter is called a *pre-emphasis filter*. After demodulation at the receiver, a *de-emphasis filter* restores the original signal spectrum. This process is illustrated in figure 15.2.5. The result is that the high-frequency components in the information signal are better protected against noise, and thus the SNR performance of the system is improved.

Chapter 16

Analog Pulse Modulation

16.1 Types of modulation

Thus far we've only considered analog modulation schemes where the information signal $m(t)$ is continuous in both amplitude and time. We are now ready to consider other possibilities, examples of which are given in Table 16.1.

		Amplitude	
		Continuous	Discrete
Time	Continuous	e.g. AM, FM, PM	not used
	Discrete	Analog Pulse Modulation	pure digital communication systems, e.g. Pulse Coded Modulation (PCM)

Table 16.1: Broad categories of modulation based on nature of information source $g(t)$

Our main goal is to study pure digital communication systems, where information source is discrete in both time and amplitude, but it is necessary to firstly study an intermediate system, that of Analog Pulse Coded Modulation (PCM) systems.

16.2 Analog Pulse Modulation

In Analog Pulse Modulation we perform two steps:

1. Ideal sampling
2. These samples are used to modulate some aspect of a train of pulses.

These are now considered.

16.2.1 Ideal sampling

Ideal sampling on the information signal $g(t)$ at a rate of $f_s = \frac{1}{T_s}$, to create a collection of discrete time samples $\{g_n\}$ for $-\infty < n < +\infty$.
Point to note:

- The signal is converted to be discrete in time
- It is still however continuous in amplitude (there is an infinity of possible amplitudes)

16.2.2 Pulse modulation

Here we take a train of pulses each having a shape $h(\tau)$ and each separated by T_s (the sampling interval). Thus unmodulated train of pulses is given by:

$$\sum_{n=-\infty}^{+\infty} h(t - nT_s)$$

Then we modulate (change) some aspect of this train in accordance with the signal samples $\{g_n\}$.
There are several obvious options:

- PAM: Pulse Amplitude Modulation
  This is where the amplitude of the n^{th} pulse is proportional to the n^{th} sample g_n
- PWM: Pulse Width Modulation
  This is where the width of the n^{th} pulse is proportional to the n^{th} sample g_n
- PPM: Pulse position Modulation
  This is where the position (in time) of the n^{th} pulse is proportional to the n^{th} sample g_n

These three schemes are illustrated in Figure 16.2.1 for the case of using a rectangular pulse given by:

$$h(\tau) = \begin{cases} 1 & \text{for } |\tau| < \frac{1}{2}T \\ 0 & \text{elsewhere} \end{cases}$$

Note that many options are possible for $h(\tau)$, this particular example is a baseband pulse. Radio frequency pulses are also possible.

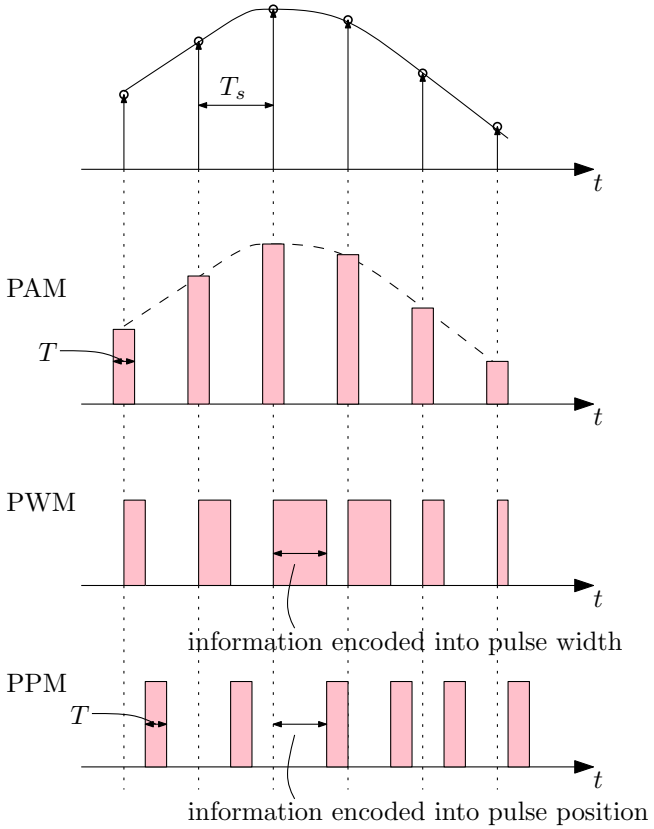


Figure 16.2.1: The three types of analog pulse modulation, namely PAM, PWM, and PPM. This example assumes a rectangular pulse shape.

Lets look closer at Pulse Amplitude Modulation (PAM), we won't consider the other schemes in this course as they are not widely used in communication systems (PWM signals are used to drive some actuators, e.g. motors and lights in many applications but not for communications per-sa. PPM also has some non-communication applications).

16.3 PAM

16.3.1 Spectrum of PAM

We can model the time domain process of sampling and the subsequent mapping to a pulse amplitudes as an ideal impulse sampling process followed by a linear *pulse shaping filter* with impulse response $h(\tau)$, as shown in Figure 16.3.1 below.

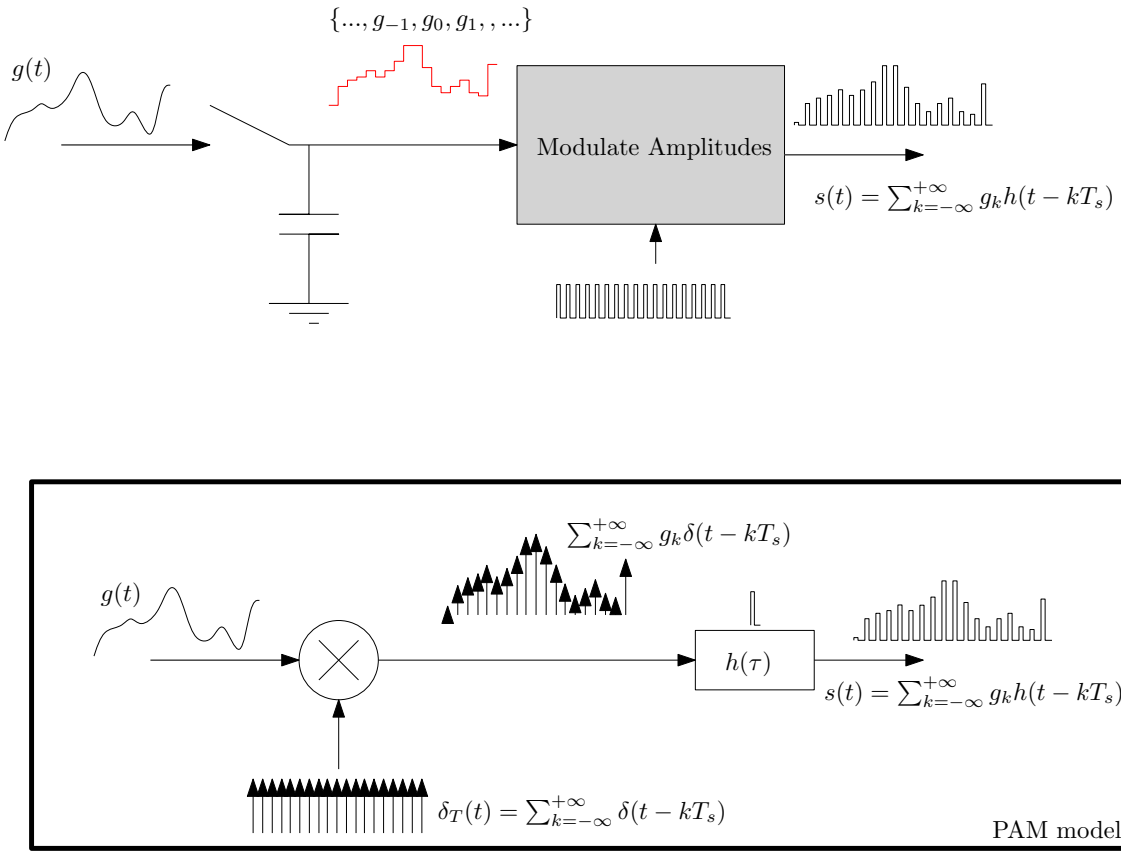


Figure 16.3.1: Sampling followed by pulse modulation is equivalent to ideal impulse sampling followed by a *pulse shaping filter* with impulse response $= h(\tau)$, the shape of the pulses to be modulated.

Equipped with this model we can now develop simple frequency domain model by taking the Fourier transforms and noting that filtering is a convolution in the time domain which corresponds to multiplication in the frequency domain.

We know from Digital Signal Processing (DSP) that the Fourier transform on an impulse sampled signal is the Discrete Time Fourier Transform (DTFT[1]), $\bar{G}(f)$. We also know from DSP that the DTFT is just the scaled spectrum of $g(t)$ repeated every integer multiple of F_s, i.e.:

$$\bar{G}(f) = F_s \sum_{k=-\infty}^{+\infty} G(f - kF_s)$$

Let the Fourier transform of $h(\tau)$ be $H(f)$, thus the spectrum of the modulated signal $s(t)$ is:

$$S(f) = F_s H(f) \sum_{k=-\infty}^{+\infty} G(f - kF_s)$$

[1] In DSP we use the notation $\bar{G}(j\omega)$, here we use the slightly different notation $\bar{G}(f)$, where f is assumed to be real and $\omega = 2\pi f$

16.3.1.1 Example: rectangular pulse

When $h(\tau)$ is a rectangular pulse with unit magnitude and width T seconds we have:

$$h(\tau) = \begin{cases} 1 & \text{for } |\tau| < \frac{1}{2}T \\ 0 & \text{elsewhere} \end{cases}$$

and the Fourier transform is:

$$H(f) = T\text{sinc}(\pi f T)$$

where $\text{sinc}(x) \triangleq \frac{\sin(x)}{x}$.

So $H(f)$ has nulls when $\pi f T$ is a multiple of π, i.e. when fT is an integer, i.e. when f is a multiple of $\frac{1}{T}$.

Note that yet again we see the time / frequency domain duality where a narrow pulse in the time domain (small T), results in a wide pulse in the frequency domain (the nulls are separated by $\frac{1}{T}$).

Putting these results together we have the spectrum of the analog PAM signal:

$$S(f) = \left(\frac{T}{T_s}\right)\text{sinc}(\pi f T)\sum_{k=-\infty}^{+\infty} G(f - kF_s)$$

This is shown in Figure 16.3.2.

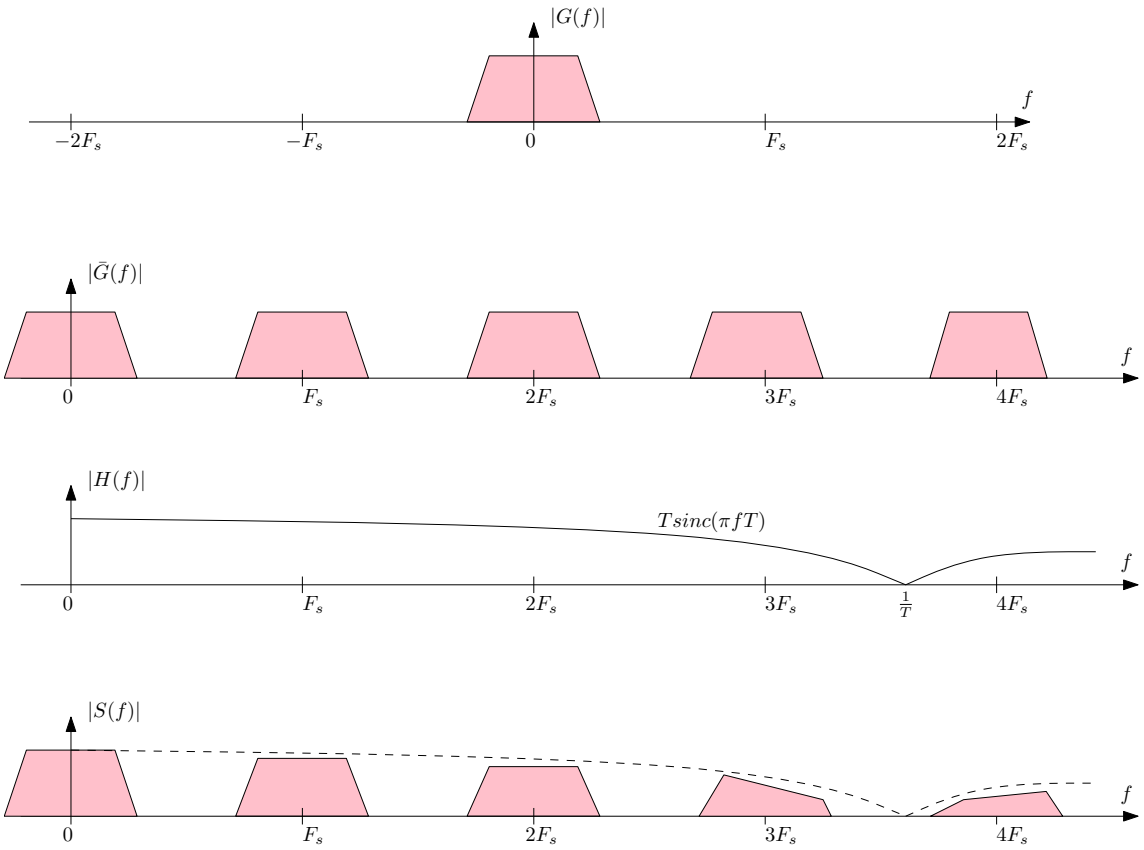


Figure 16.3.2: The spectrum of an analog PAM signal when a rectangular pulse having a width $T \approx \frac{1}{3.5} T_s$ is used.

16.3.1.2 Example: rectangular pulse, $T = T_s$

When $h(\tau)$ is as above, but the width of the pulse is equal to the sampling interval, then the null in the pulse shaping filter's response occurs every F_s exactly where each of the repeated spectra in the DTFT of $g(t)$ occurs, as shown in Figure 16.3.3.

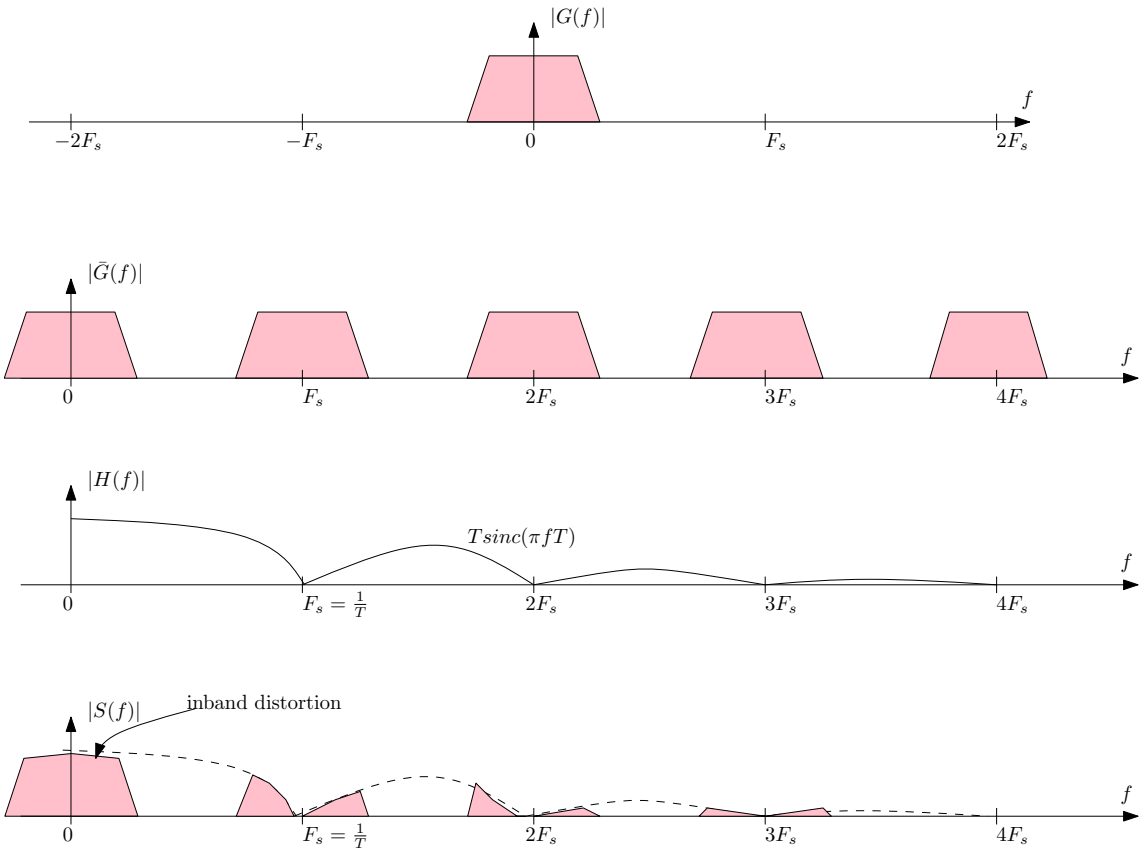


Figure 16.3.3: The spectrum of an analog PAM signal when a rectangular pulse having a width $T = T_s$ is used.

16.3.1.3 Terminology

In common usage are the following two terms in connection with PAM.

1. Return-to-Zero (RZ): This is when a short pulse is used, i.e. $T < T_s$, so the signal returns to zero before the next pulse.

2. Non-Return-to-Zero (NRZ): This is when pulse duration equals the sampling interval, i.e. $T = T_s$, so the signal does not zero before the next pulse.

In truth, these terms are more commonly used in connection with pure digital transmission schemes but they are interoduced here as they have a slightly more general meaning.

16.3.1.4 Signal reconstruction

When $s(t)$ as above (i.e. rectangular pulse with $T = T_s$) is received how do we recover the original signal?
An ideal reconstruction filter that selects the main lobe of $S(f)$ will almost work, but not fully as there is some inband distortion due to the sinc response of $H(f)$. Thus signal

reconstruction can be achieved by an ideal reconstruction filter (ideal low pass filter with a bandwidth of $\frac{1}{2}F_s$) followed by a compensating filter with a response equal to $\frac{1}{H(f)}$ on the range $0 < f < \frac{1}{2}F_s$. This compensating filter is sometimes also called an equalizer, and is illustrated in Figure 16.3.4.

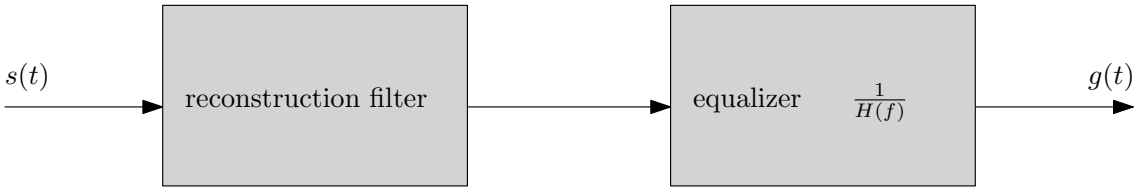


Figure 16.3.4: The reception of a PAM signal.

16.3.2 Why?

In of itself analog PAM is fairly useless - it has a larger spectrum than its continuous time equivalent, i.e. compared to just transmitting the original analog signal along a wire! So why is it ever used?

There is one very important application, namely time division multiplexing of signals from many sources onto the one wire. Say, for example, we want to send 4 analog signals down the one (telephone) wire; we could sample each one and use analog PAM to represent each by a pulse of length $\frac{1}{4}T_s$ and each just delayed in time so that they can all be added together as illustrated in Figure 16.3.5.

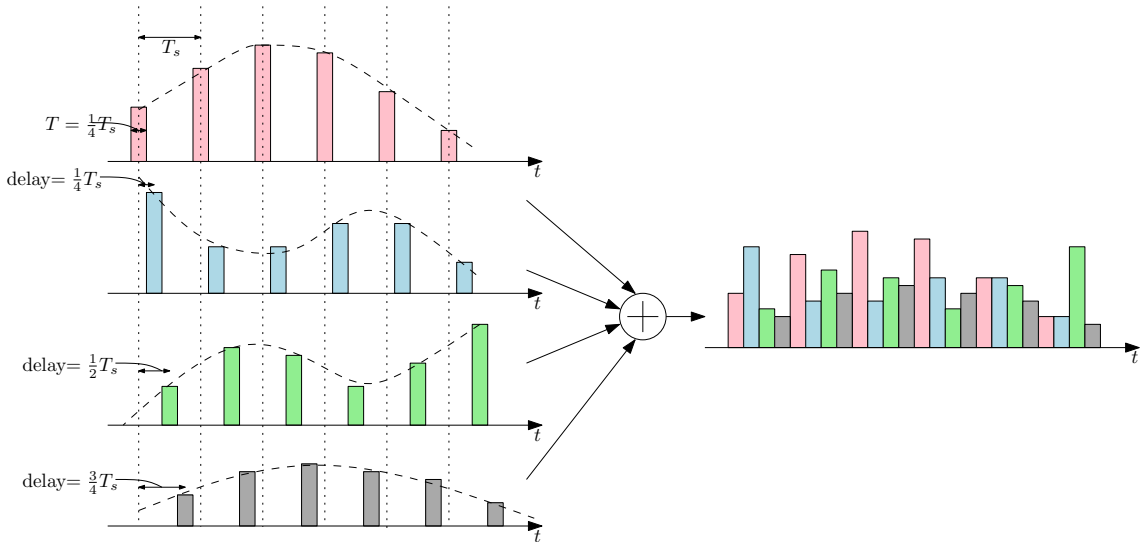


Figure 16.3.5: Using analog PAM to multiplex 4 signals into 1 signal using time division multiplexing.

This system is used in the Plain Old Telephone System (POTS).

Chapter 17

Pure Digital Communications

17.1 Introduction

- In a digital system, we communicate by means of a *finite set of signals*. This is markedly different to an analog communication system, where the transmitted signal is chosen from an infinite number of possible signals.
- Each signal represents a discrete symbol, chosen from a set of M possible symbols.
- Examples of symbol streams:
  - Chinese symbols
  - Binary stream $[1\,1\,0\,1\,\cdots]$ - a stream of bits.
  - A stream of symbols resulting from the *digitization* (sampling and analog-to-digital conversion) of an analog signal.
- There are a number of codes used for representing characters (letters of the alphabet etc.):
  - ASCII: American Standard Code for Information Interchange
  - EBCDIC: Extended binary coded decimal interchange code.
  - Example: 7-bit ASCII: A = 1000001 ; g = 1110011
- Bits can be grouped together to form *symbols*. In digital communications, each symbol is represented by a unique waveform. Usually all symbols have the same duration (called the symbol interval).

- If we group N bits together to form a symbol, the number of possible symbols is $M = 2^N$ this is called M-ary signaling. Therefore the number of bits per symbol is $N = \log_2 M$.
- The transmission rate is measured in bits per second (bps or bits/s), or in symbols per second. Sometimes one symbol per second is called *one baud*.
- Features of a digital communication system:
  - Compression (or Source Coding): Removes unwanted redundancy from the information symbol stream.
  - Channel coding: Puts in structured redundancy in order to help correct errors at the receiver.

17.1.1 Digital communication system performance measure

Suppose we've built a digital communication system, and we want to measure how good it is? For this purpose, we use:

$$\text{Bit Error Rate} = \frac{\text{Number of bits reeived in error}}{\text{Number of bits sent}}$$

17.1.2 Advantages of Digital Communications

- It is possible to do compression
- It is possible to do channel coding
- It is possible to make data private using encryption
- All information signals may be treated equally (as a sequence of bits)
- Digital circuits are inexpensive
- It is possible to perform detection and subsequent retransmission at intervals along the communication path (relaying) without degrading the signal to noise ratio.
- Well suited to digital sources of information, e.g. file transfer.

17.1.3 Advantages of Analog Communications

- Synchronization of analog communication systems is generally easier
- Analog systems exhibit ”graceful degradation” - when system performance decreases, it does so gradually.
- Digital communication systems tend to fail in a sudden manner.
- Well suited to analog signal sources, e.g. voice over a telephone line.

Chapter 18

PCM and line codes

In analog pulse modulation, discrete samples of an analog signal were used to modulate some aspect of a pulse train (e.g. pulse amplitude, pulse width, or position of the pulses). In Pulse Coded Modulation, the analog samples are converted (or mapped) to a finite set of codewords which in turn are used to create a train of electrical pulses.

Pulse Code Modulation (PCM) is used to transmit an analog signal in digital form. It has been used in telephone networks for many years.

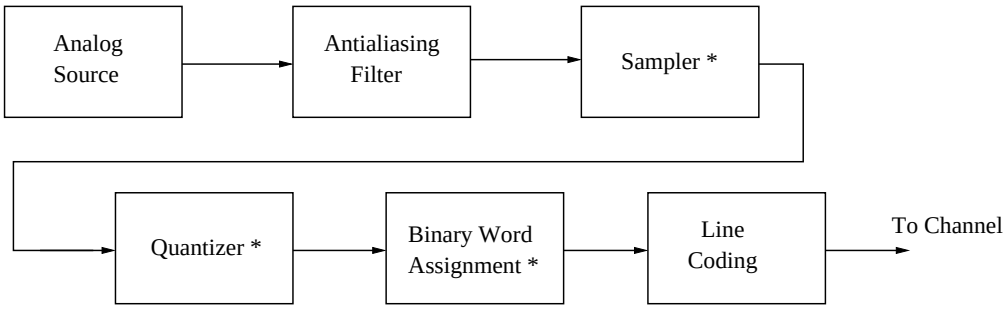


Figure 18.0.1: PCM system.

The three blocks marked with an asterisk (*) constitute the analog-to-digital conversion (A/D or ADC) process, but they are listed here separately as it is necessary to consider the effect of each.

18.1 ADC

18.1.1 Sampler

The sampler is exactly as per the analog pulse modulation system; it simply samples the analog signal at a uniform spacing in time (every T_s seconds). These samples are continuous in amplitude.

Provided the Nyquist sampling criteria is satisfied this process does not introduce any information loss.

18.1.2 Quantizer

After sampling, the signal is discrete in time, but is still a continuum in amplitude.

- The signal needs to be mapped (i.e. rounded) to discrete levels - separated by Δ volts.
- This process is called quantization.
- This process introduces a error which can't be recovered - however this is the only place in a well designed digital communication system where unrecoverable noise is introduced.
- This is known as quantization noise.
- The quantization noise is uniformly distributed between $\pm\frac{1}{2}\Delta$, and can be shown to have a variance of $\frac{1}{12}\Delta^2$ (see Appendix 4), leading to the approximate linearized model shown in Figure 18.1.1

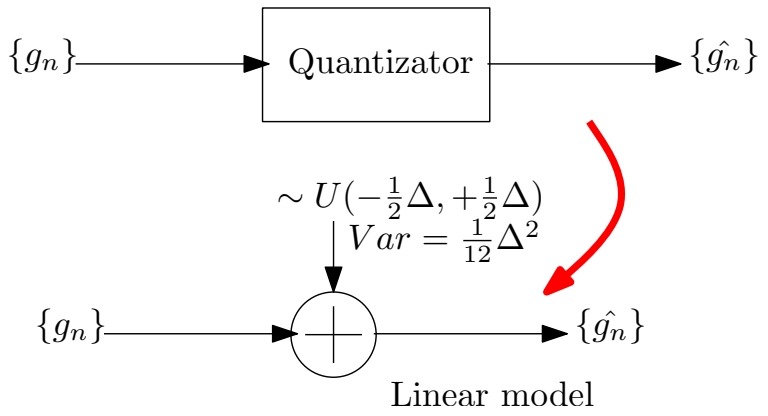


Figure 18.1.1: Approximate linear noise model for quantization noise.

If we consider a sine wave input to an ADC quantized to b-bit binary words, i.e. 2^b quantization levels the Signal-to-Quantization-Noise (SQNR) may be calculated and is presented in the following table:

#bits = b	#level =2^b	SQNR [dB]
8	256	49.92
10	1,024	61.96
12	4,096	74
14	16,384	86.04
16	65,536	98.08
18	262,144	110.12

A word on terminology:

For the 12 bit case, we might say: “The quantization noise is 74dB down on the signal”, or “The quantization is -74dB”, or “The SNR due to quantization is 74dB”.

The choice of how many bit to use really depends on the application. If for example, we are dealing with an audio signal that has been picked up by a cheap microphone and a noisy amplification circuit (as might be the case in a mobile phone for example), then it might be that the signal being presented to the ADC circuit already has -40dB of noise present. Noting that dB are on a log scale, we can see that -40dB of noise is far noisy than -74dB. Note the 24dB difference is 8×3dB, and each 3dB corresponds to a doubling of power. Therefore the analog audio signal being presented to the 12-bit ADC would already have 256 times the amount of noise present than what the ADC is going to add.

18.1.2.1 Examples

Compact Disc (CD)

- Stereo, i.e. 2 ADC channels
- $F_s = 44.1\text{kHz}$ per channel
- $b = 16\text{bits}$ resolution per channel

Digital part of Plain Old Telephone Service (POTS)

- Single channel
- $F_s = 8\text{kHz}$
- $b = 8\text{bits}$ resolution

- Analog still used on local loop (to / from the home)

18.1.2.2 Non-Uniform Quantization

In regular uniform quantization, low amplitude signals have worse signal to quantization noise ratio (SQNR) than large amplitude signals. This is a a particular problem for speech. The situation can be significantly improved through the use of logarithmic *companding*. In μ-law companding, the signal, before quantization, is passed through a nonlinear compression circuit (*compressor*) which has the following characteristic:

$$x_2 = \frac{\log\left[1 + \mu x_1\right]}{\log\left[1 + \mu\right]} \quad \text{where } 0 \leq x_1 \leq 1 \ .$$

The value $\mu = 255$ is used for telephony in the US. Compression is followed by a uniform quantizer - see Figure 18.1.2 for an illustration

The concatenation of nonlinear circuit and uniform quantizer is equivalent to a *non-uniform quantizer*. A device in the receiver, called an *expander*, will later restore the reconstructed signal to its correct (original) level.

The combination of *compressing* and *expanding* is referred to as *companding*. In Europe, a similar scheme, called A*-law companding*, is used – here the compander has a slightly different characteristic to the μ-law.

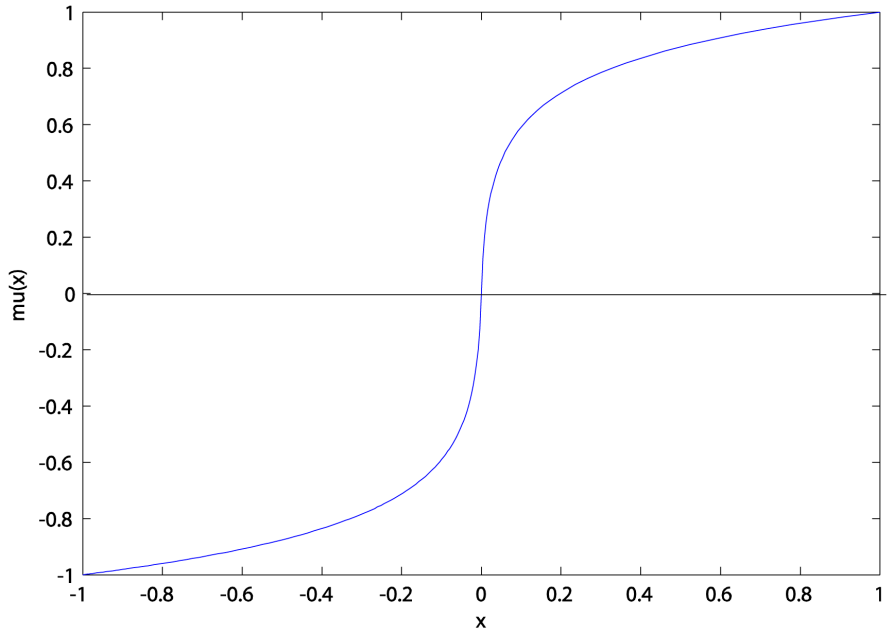


Figure 18.1.2: μ-law companding.

18.1.3 Binary word assignment

Binary word assignment can be a simple mapping to regular 2's complement binary notation, or it could be to some other format which may be more suitable to the communication task. Some examples are shown in the following table:

Level number	Binary code	Gray code
1	000	010
2	001	011
3	010	001
4	011	000
5	100	100
6	101	101
7	110	111
8	111	110

The leftmost bit is a sign bit (in both cases).
In the Gray code, the binary words representing adjacent levels differ by exactly one bit; hence, a single bit error in the received code will result in minimal errors in the reconstructed analog levels.

18.2 Line coding

For the moment we'll just consider baseband modulation, i.e. encoding data onto a low-pass signal suitable for transmission over a wire line (hence the name *line code*). Many of the ideas and concepts can and have been extended to RF transmission also.

18.2.1 Motivation

In a binary digital communication system we could represent 0, and 1 as follows:

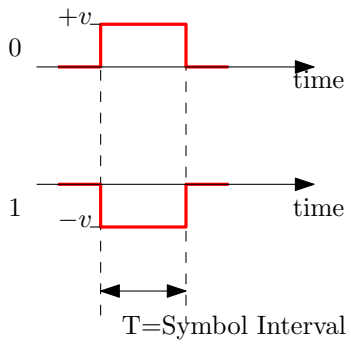


Figure 18.2.1: Binary pulse example

If we have a larger alphabet, then we can use a large set of "pulses":

Example: crazy pulses!

Consider the message "THINK!"

Lets encode it using 6-bit binary number, and split each of these into 2 x 3-bit number each having $8 = 2^3$ possibilities, see Table 18.1.

message	T		H		I		N		K		!	
6-bit ASCII	001010		000100		100100		011100		110100		100010	
2 x 3-bit	001	010	000	100	100	100	011	100	110	100	100	010
pulse index	1	2	0	4	4	4	3	4	6	4	4	2

Table 18.1: THINK! example

Thus to send each letter, we need to send 2 pulses, each drawn from a size-8 alphabet. In theory all we require is that each pulse shape used is unique, see Figure 18.2.2 for an example of what this *might* look like.

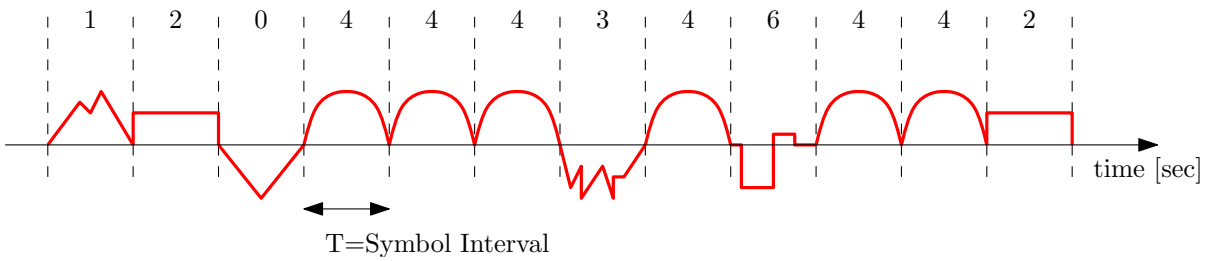


Figure 18.2.2: "THINK!" example

In practice pulse shape selection can be a complex business, and the illustration shown in Figure 18.2.2 would never be used! - it is just presented here to demonstrate a general principle.

18.2.2 Pulse Coded Modulation

A more practical approach is to do the same as analog pulse modulation but now the amplitude is be discrete; in the example above there would be 8 possible amplitudes.
Consider a 4 amplitude example using rectangular pulses as illustrated in Figure 18.2.3, we have

1. RZ: Return to zero with using pulses with amplitudes $\{\pm 3, \pm 1\}$. Symmetric (about zero) choice of amplitudes is usually preferred as it can be shown that it usually guarantees best noise performance[1].
2. Clearly, NRZ is also possible.
3. PWM: 4 different pulse widths are used.
4. Pulse position modulation: 4 different pulse positions are used.

[1] In the presence of AWGN noise

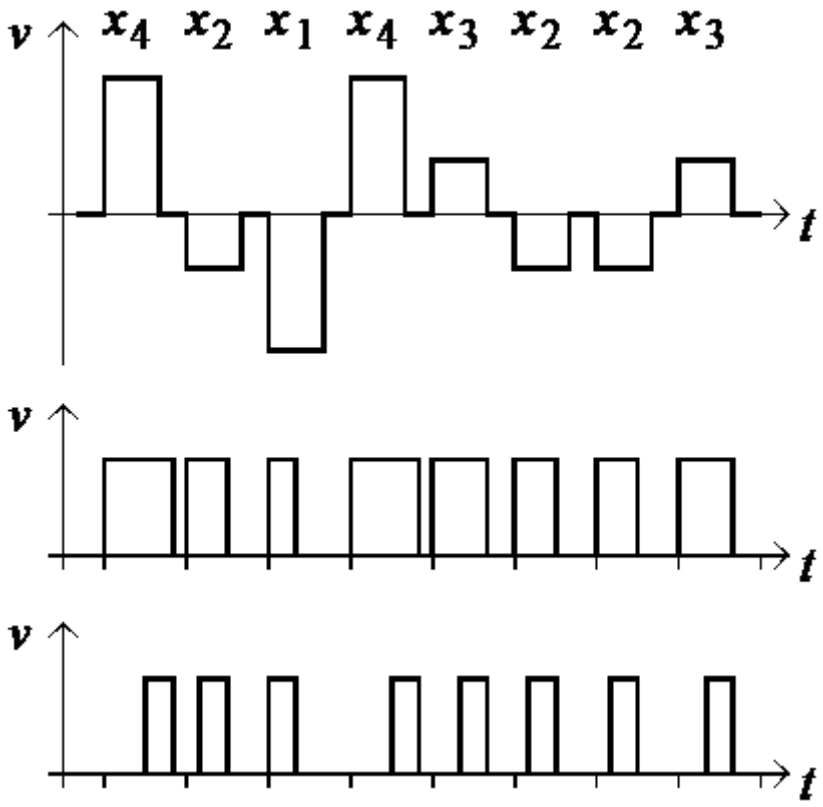


Figure 18.2.3: Illustration of several types of binary PCM. (top) Return-to-Zero (RZ), Pulse-Width-Modulation (PWM), and (bottom) Pulse-Position-Modulation.

18.2.3 Binary line codes

Often we are concerned with the transmission of binary data, be that a binary source, e.g. a file, or the bits coming from an ADC.

We only need to define two electrical signals, with Figure 18.2.4 showing some typical examples all of which use rectangular pulses, they are:

1. Return-to-Zero (RZ): This is when a short pulse positive pulse is used to represent a 1 and a negative version of the same pulse represents a 0. Here the pulses are rectangular and have a duration less than the bit interval meaning that the signal returns to zero before the start of the next bit.

2. Non-Return-to-Zero (NRZ): This is the same as RZ, but the pulse duration equals the bit interval.

3. Pulse-Width-Modulation (PWM): Rather than modifying the amplitude of the pulses (by either ± 1) to convey the desired information, we can change the width of the pulse, e.g.

narrow pulse for a 0, and a wide pulse for a 1.

4. Pulse-Position-Modulation: Yet another option is to keep the pulse completely fixed, but just vary their position in time, e.g. delay for a 1, don't delay for a 0.

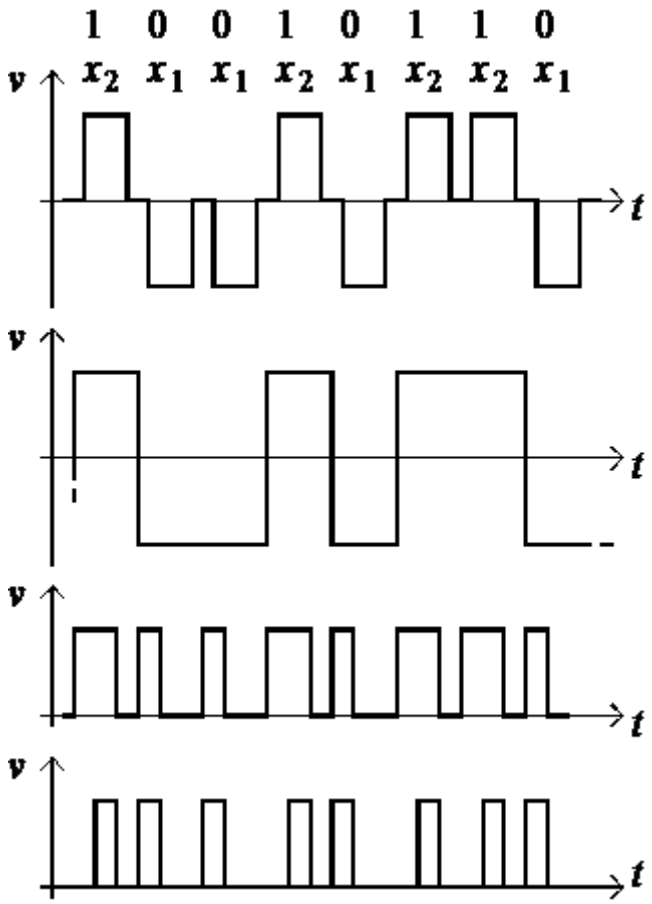


Figure 18.2.4: Illustration of several types of binary PCM. (top) Return-to-Zero (RZ), Non-Return-to-Zero (NRZ), Pulse-Width-Modulation (PWM), and (bottom) Pulse-Position-Modulation.

18.2.3.1 More advanced line codes

There are problems with RZ or NRZ

- Synchronization can be difficult
- If there is an unequal number of 1's to 0's in a binary data stream, then there can be a DC component in the signal

A scheme called **Alternate Mark Inversion** (AMI) can resolve these. In this scheme we do the following:

- use 0Volts to represent a logic 0

- alternate between a positive and negative pulse to represent a logic 1.

see Figure 18.2.5.

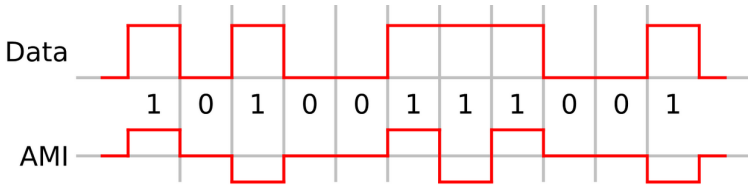


Figure 18.2.5: Illustration of AMI coding. Note how the logic "1" alternates between + and -.

Another common scheme is **Manchester coding**. There are several ways to explain the operation of Manchester coding, but in keeping with the approach taken thus far, we can simply define it as regular NRZ coding but the pulse shape is one cycle of a bipolar clock pulse as illustrated in Figure 18.2.6.

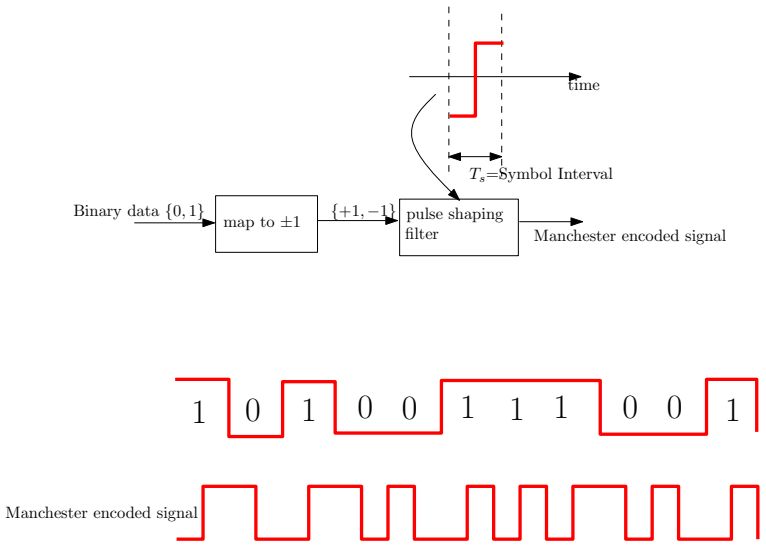


Figure 18.2.6: Illustration of Manchester coding

The advantage of Manchester coding is that there is no DC component and, due to the large number of transitions, it is easy to synchronize to at the receiver.

Chapter 19

Inter-Symbol-Interference (ISI)

The concept of Inter-Symbol-Interference (ISI) is most easily explained with reference to a binary Pulse Amplitude Modulated (PAM) system but the effect if not limited to said system.

19.1 binary Pulse Amplitude Modulated (PAM)

In an PAM system the k^{th} pulse has an amplitude b_k which is chosen from a finite alphabet (usually equally spaced values symmetric about zero); each modulated pulse is called a symbol and each one is separated in time by a symbol period T seconds (the symbol interval), i.e. the k^{th} symbol is sent at time $t = kT$.

Terminology

- The symbol rate is $R_s = \frac{1}{T}$ symbols/sec.
- If there are B bits per symbol, the bit rate $R_b = \frac{B}{T}$ bits/sec.

In a **binary** PAM system, such as the one depicted in Figure 19.1.1, each symbol is selected from a binary alphabet $= \{-1, +1\}$.

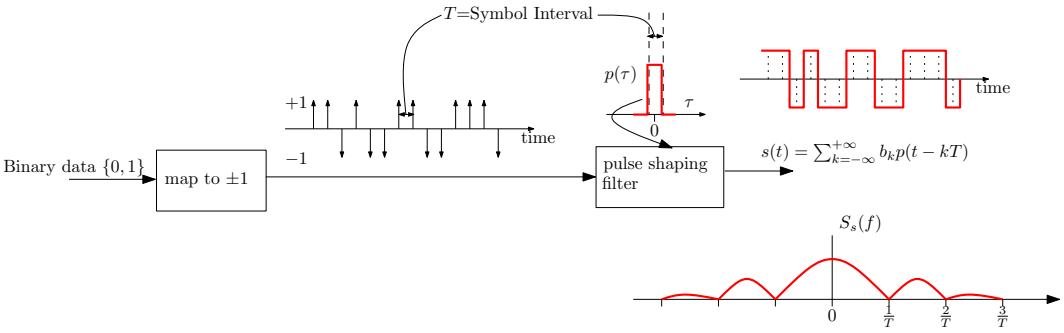


Figure 19.1.1: An example binary Pulse Amplitude Modulated (PAM) with source symbols $= \{-1, +1\}$

In this example we've taken the pulse to be a rectangle having width equal to the symbol interval:

$$p(\tau) = \begin{cases} 1 & \text{when } |\tau| < \frac{1}{2}T \\ 0 & \text{elsewhere} \end{cases}$$

Note that in general any pulse shape is possible, even ones longer than T, in fact even infinitely long ones are also possible. For mathematically simplicity (and without loss in generality) we always take the middle of symmetric pulses to occur at $\tau = 0$ as in this case.

19.1.1 A possible reception process

Suppose some Additive White Gaussian noise is added to the signal and we are tasked with making a receiver that can accurately estimate the transmitted symbols.

Q. How would we do it?

A. See Figure 19.1.2.

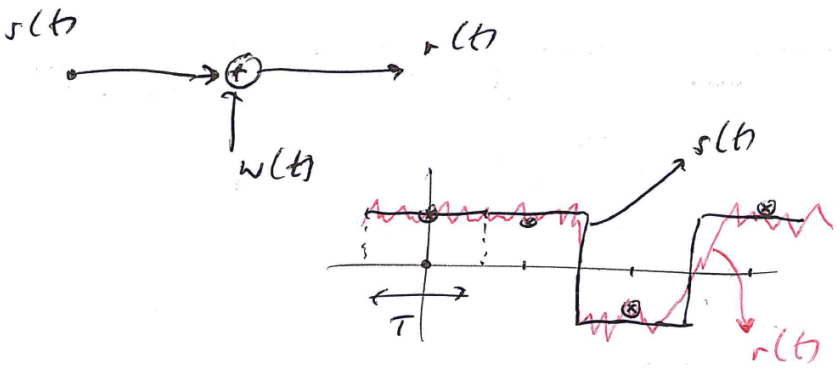


Figure 19.1.2: Binary PAM currupted by noise. Possible receiver sampling points are shown in thr middle of each symbol.

To detect the symbols $\{b_k\}$ we *could* sample the received signal $r(t) = s(t) + w(t)$ every T seconds in the middle of each symbol, i.e. at times $t = kT$.
We *could* apply the following simple rule:

- Let $\hat{b}_k = +1$, if $r(kT) > 0$,
- Let $\hat{b}_k = -1$, if $r(kT) \leq 0$,

Where the equality case was arbitrarily chosen.
This technique will work when the amount of noise is small, but (as we will see later) we can do a lot better.

19.2 Different pulse shapes

A nice property of the rectangular pulse shape used in Figure 19.1.1 is that the simple reception process described above works because at the selected sampling points ($t = kT$) the value of the signal depends only on the k^{th} symbol and the noise at that moment in time;

$$\begin{aligned} r(kT) &= s(kT) + w(kT) \\ &= b_k + w(kT) \end{aligned}$$

i.e. it does not depend on other symbols.
We say that there is no Inter-Symbol-Interference (ISI) - happy days!
However, all is not well, as we will now see.

19.2.1 Spectrum of PAM

We saw in Chapter 13 that the PSD of a train of modulated delta function is:

$$S_X(f) = \frac{1}{T}\sigma^2$$

where σ^2 is the variance of the modulating values, which is this case are ± 1 with equal probability, so $\sigma^2 = 1$, and $S_X(f) = \frac{1}{T}$. This is then filtered by $p(\tau)$ so the PSD of the resulting signal is:

$$S_s(f) = \frac{1}{T}|P(f)|^2$$

where $P(f)$ is the Fourier transform of $p(t)$.

For rectangular PAM, $P(f) = T\text{sinc}(\pi f T)$, and so the PSD is:

$$S_s(f) = T\text{sinc}^2(\pi f T)$$

This is illustrated in Figure 19.1.1.

Comment:

The rectangular time domain pulse is nice as there is no ISI, but it has infinite bandwidth. Can we have a finite bandwidth pulse? Yes!, but what does it look like in the time domain - will there be ISI - Let's find out.

19.2.2 Sinc pulse

If we use a sinc pulse in the time domain:

$$p(\tau) = \text{sinc}\left(\pi\frac{\tau}{T}\right)$$

This has a rectangular spectrum between $\pm\frac{1}{2T}$Hertz, i.e, this will have a finite bandwidth $= \frac{1}{T}$Hertz.

But the time domain pulse now has an infinite duration!

In theory every symbol overlaps with every other symbol! as can be seen in the time domain expression for our PAM signal:

$$\begin{aligned} s(t) &= \sum_{k=-\infty}^{+\infty} b_k p(t-kT) \\ &= \sum_{k=-\infty}^{+\infty} b_k \text{sinc}\left(\pi\frac{t-kT}{T}\right) \end{aligned}$$

i.e. the sample at any time instant t_0 contains weighted contributions from EVERY b_k - thus we have a lot of Inter-Symbol-Interference (ISI) !!!!

But is this a problem???

19.2.2.1 Simple receiver

The simple receiver presented in Section 19.1.1 can be applied to this signal, i.e. to decide on b_k sample the received signal at time $t = kT$,

$$\begin{aligned} r(kT) &= s(kT) + w(kT) \\ &= \sum_{n=-\infty}^{+\infty} b_n \text{sinc}\left(\pi \frac{kT - nT}{T}\right) + w(kT) \\ &= \sum_{n=-\infty}^{+\infty} b_k \text{sinc}(\pi(k-n)) + w(kT) \end{aligned}$$

(be careful above, note how I changed the summation index from k to n to avoid confusion with the sampling instant kT).

The argument to the sinc function is an integer multiple of π, so we have

$$\text{sinc}(\pi(k-n)) = \begin{cases} 1 & \text{for } k = n \\ 0 & \text{for } k \neq n \end{cases}$$

Thus we have:

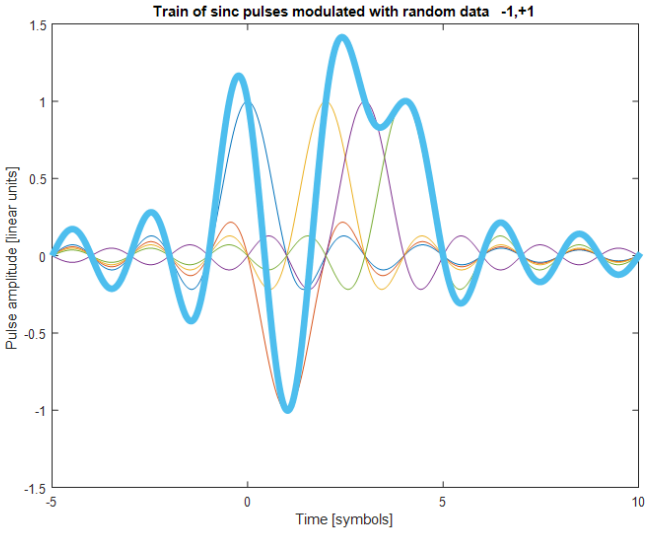
$$r(kT) = b_k + w(kT)$$

This is amazing! The samples taken at times $t = kT$ only depend on b_k and the noise at that instant in time, and not on any of the other symbols!

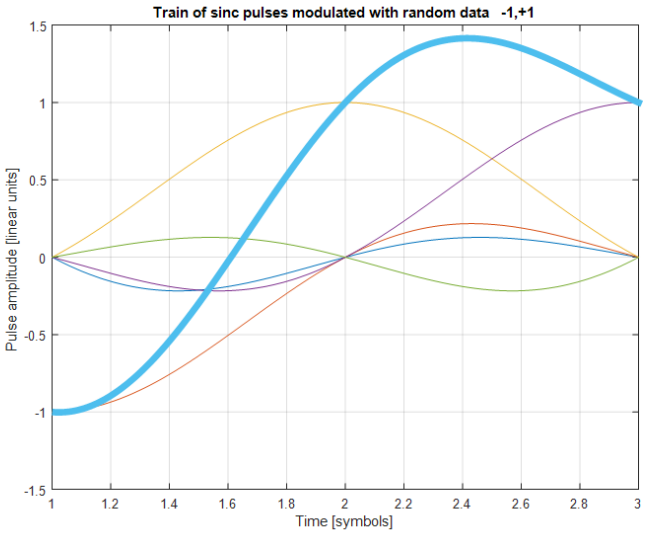
Thus, despite first impressions, if we take the samples of the received signal correctly then the samples have *no* ISI!!!

What has happened here?

Well the pulse shape used has a value of 1 in the middle, but is 0 at the middle point of every other symbol. So samples taken in the middle of each symbol only contain contributions from that symbol. This is illustrated in Figure 19.2.1 below.



(a)



(b)

Figure 19.2.1: (a) Train of 5 modulated sinc pulses, (b) zoomed in version. The transmitted signal is the sum of them and is shown as the thick line on each plot. Note how the transmit signal either passes through ± 1 at the middle of each symbol, and the contributions from all other sinc pulses are zero at those time instances.

19.2.3 Time domain rule for ISI-free operation

From the above discussion, we have the following condition on the pulse shaping filter for ISI free operation:

$$p(kT) = \begin{cases} 1 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases} \tag{19.2.1}$$

19.3 Frequency domain rule

We have a time domain rule for designing pulse shaping filters to guarantee ISI free operation. We would also like a frequency domain rule to help use design pulse shaping filters that are both compact in the frequency domain and are ISI free.

19.3.1 Nyquist's First Criteria (NFC)

Most books take a couple of pages to derive this, but we will use some result from our DSP course to compress the derivation down to just a few lines.
If we sample $p(\tau)$ with a train of delta functions each separated by T seconds then, according to Equation 19.2.1, the result will be zeros everywhere except for $t = 0$, i.e.:

$$\begin{aligned} p(t) \sum_{k=-\infty}^{+\infty} \delta(t - kT) &= 1.\delta(0) + \sum_{\substack{k=-\infty \\ k \neq 0}}^{+\infty} 0.\delta(t - kT) \\ &= \delta(0) \end{aligned}$$

Taking the Fourier transform of both sides and using the following facts:

- The left hand side is just the sampled version of $p(t)$ and so its Fourier transform is the DTFT, which can be written as the infinite sum of repeated spectra (see DSP course)
- The Fourier transform of the right hand side is just 1

we get:

$$F_s \sum_{k=-\infty}^{+\infty} P(f - kF_s) = 1 \tag{19.3.1}$$

where $F_s = \frac{1}{T}$ is the symbol rate.

This is known as Nyquist's First Criteria (NFC) for ISI-free baseband pulse transmission.

19.3.2 Interpretation

What the NFC rule says is that $P(f)$ can be any shape you like, so long as the repeated spectrum is unity. This leads immediately to some interesting results:

Pulse bandwidth must be $\geq \frac{1}{2}F_s$

If the bandwidth of the pulse is $< \frac{1}{2}F_s$ then it is not possible to have ISI free operation. This is obvious from Figure 19.3.1.

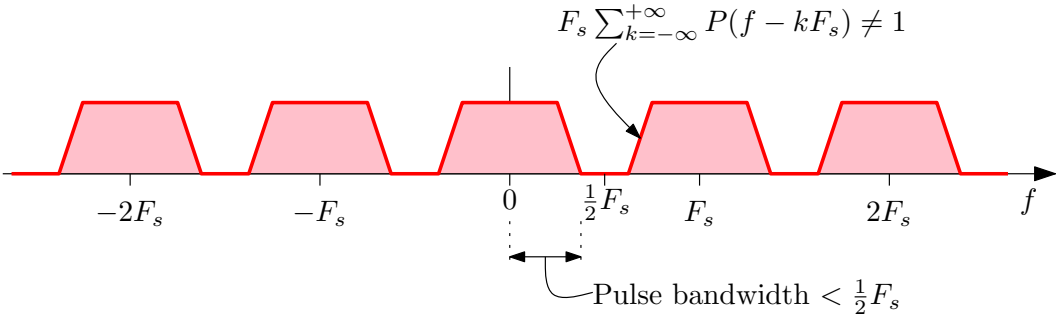


Figure 19.3.1: An pulse with bandwidth $< \frac{1}{2}F_s$ can't have ISI free operation.

If pulse bandwidth $= \frac{1}{2}F_s$, then it must be a rectangle

If the bandwidth of the pulse is exactly $\frac{1}{2}F_s$ Hertz, then there is no overlap. So to satisfy the NFC the pulse spectrum must be a rectangle as per Figure 19.3.2.

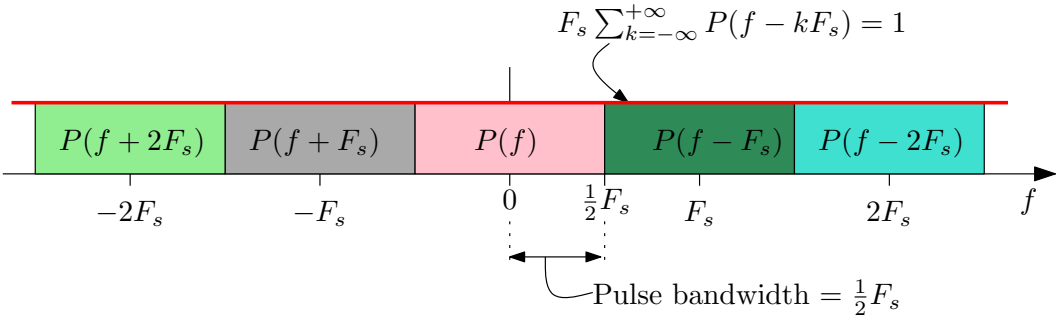


Figure 19.3.2: An ISI free pulse with bandwidth $= \frac{1}{2}F_s$ must be a rectangle.

This corresponds to a sinc time domain pulse.

Corollary. *Of all possible ISI free pulse, the sinc pulse has the smallest bandwidth.*

Bandwidth $> \frac{1}{2}F_s$ is possible

All that we require is that there is a kind of odd symmetry about $\frac{1}{2}F_s$ as shown in the example in Figure 19.3.3

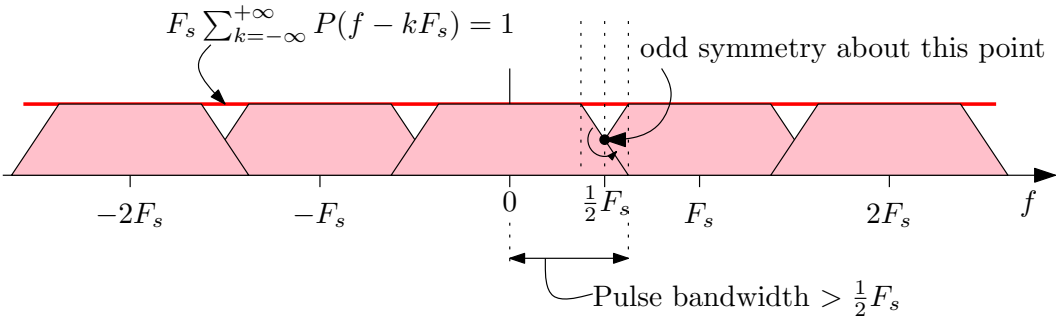


Figure 19.3.3: An ISI free pulse with bandwidth $> \frac{1}{2}F_s$ must have the odd symmetry shown about $\frac{1}{2}F_s$.

Excess bandwidth & roll-off factor

We've already established that the smallest bandwidth and ISI-free pulse can occupy is $\frac{1}{2}F_s$, but in practice real world systems use pulses that have a larger bandwidth than that.
Thus we define a parameter called the *excess bandwidth* as the amount of extra bandwidth used compared to the minimum possible
We also define the roll-off factor (usually denoted β) as the ratio of the excess bandwidth to the minimum possible; it is usually quoted as a percentage.
Example:
Let the symbol rate is 1MHz, and the pulse bandwidth be 750kHz:

- The minimum ISI-free bandwidth is 500kHz
- The excess bandwidth is 250kHz, i.e. the pulse bandwidth of 750kHz is 250kHz higher than the minimum ISI-free bandwidth of 500kHz
- The roll-off factor $\beta = \frac{250}{500} = 0.5$, or 50%.

19.3.3 Raised Cosine Pulse

A more practical and realizable ISI-free pulse is the Raised-Cosine filter. This filter has a frequency domain response that is flat for the central part of the spectrum, and is a raised cosine shape during the overlap regions as illustrated in Figure 19.3.4.
It is parametrized by it's roll-off factor β, as defined above, and in this case $0 \leq \beta \leq 1$.

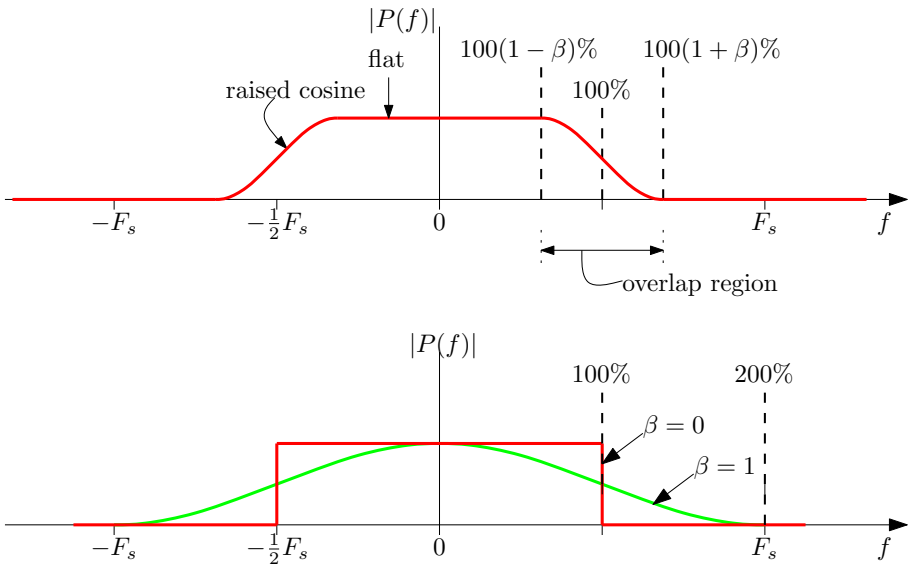


Figure 19.3.4: A raise cosine response is designed so that it complies with NFC.

The mathematical definition of a raised cosine pulse is:

$$P(f) = \begin{cases} T & \text{for} \quad |f| \leq (1-\beta)\,\frac{1}{2}F_s \\ \frac{1}{2}T\left[1+\cos\left(\frac{\pi T}{\beta}\left(|f|-(1-\beta)\,\frac{1}{2}F_s\right)\right)\right] & \text{for} \quad (1-\beta)\,\frac{1}{2}F_s < |f| < (1+\beta)\,\frac{1}{2}F_s \\ 0 & \text{for} \quad |f| \geq (1+\beta)\,\frac{1}{2}F_s \end{cases}$$

Taking the inverse Fourier transform the time domain pulse can be computed:

$$p(t) = \operatorname{sinc}(\pi F_s t)\,\frac{\cos(\pi\beta F_s t)}{1-(2\beta F_s t)^2}$$

This is illustrated in Figure 19.3.5.

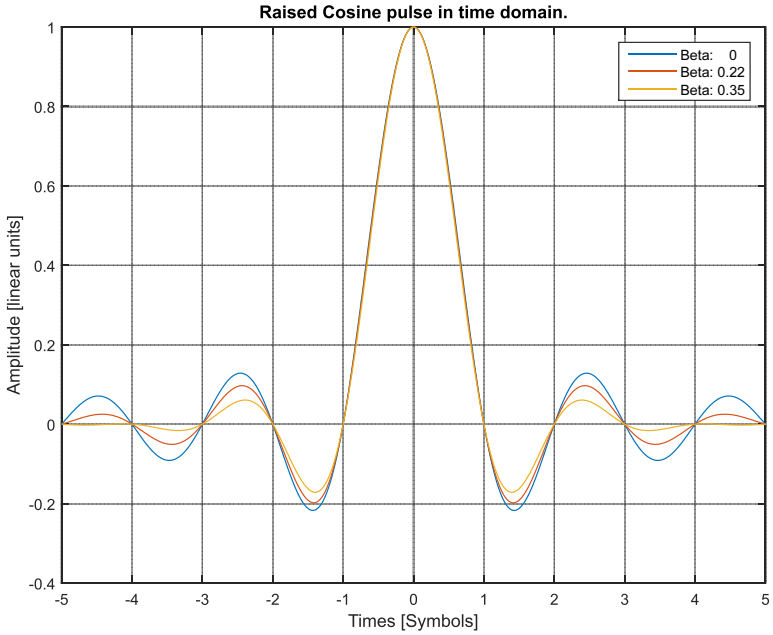


Figure 19.3.5: Example raise cosine pulses in the time domain.

Note how the pulse (by design) passes through 0 at every symbol interval except at time $t = 0$; this is the ISI-free condition.

Note also how $\beta = 0$ corresponds to a sinc pulse which decays much slower compared to the $\beta = 0.35$, a commonly used roll-off factor. The Raised Cosine pulse decays at a rate $\propto \frac{1}{t^3}$, whereas the since pulse decays only at a rate $\propto \frac{1}{t}$. This, in fact, is why raised cosine pulses are used, i.e. they have finite bandwidth and decay quickly in the time domain (but in theory they are, of course, infinity long).

A comparison of square and RC filtered signal is shown in Figure 19.3.6; note how the RC signal passes through ± 1 at the symbol instances, but at other times instances it is very different to the square wave version. The bandwidth of the RC version is finite (in this case it has a 22% excess bandwidth), but the bandwidth of the square wave version is infinite.

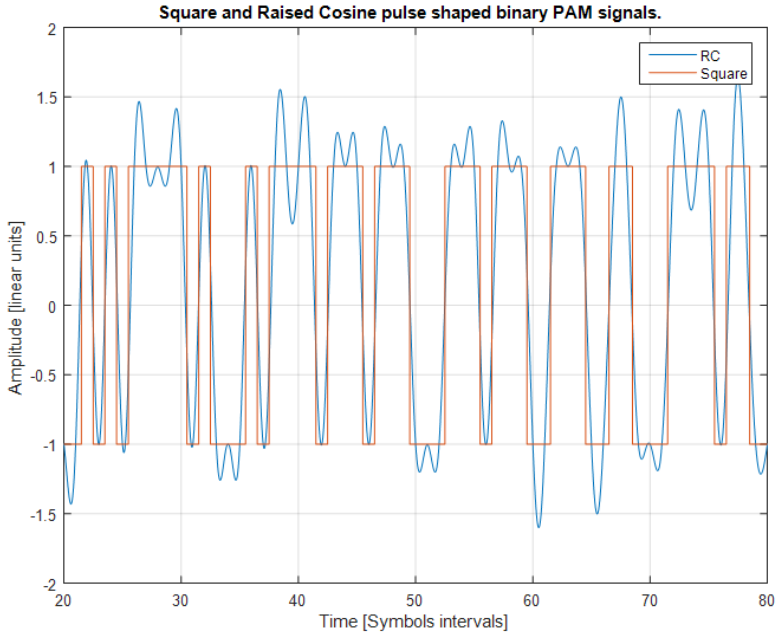


Figure 19.3.6: A comparison of square and RC filtered binary PAM signals

19.4 Sampling errors

In the above discussions regarding ISI free pulses it is assumed that the timing synchronization in the receiver is perfect; i.e. we sample at time instance $t = kT$ exactly. In practice the sampling instances will have an error, i.e. sampling might occur at times $t = kT + \epsilon$ where ϵ could be a significant proportion of a symbol period, e.g. 20%. What happens in this case? Consider the ISI-free pulses shown in Figures 19.2.1(a) and (b). If we have a sampling error of, say -20%, then the value of sample number 2 would be approximately 0.5 instead of 1.0 as illustrated in Figure 19.4.1

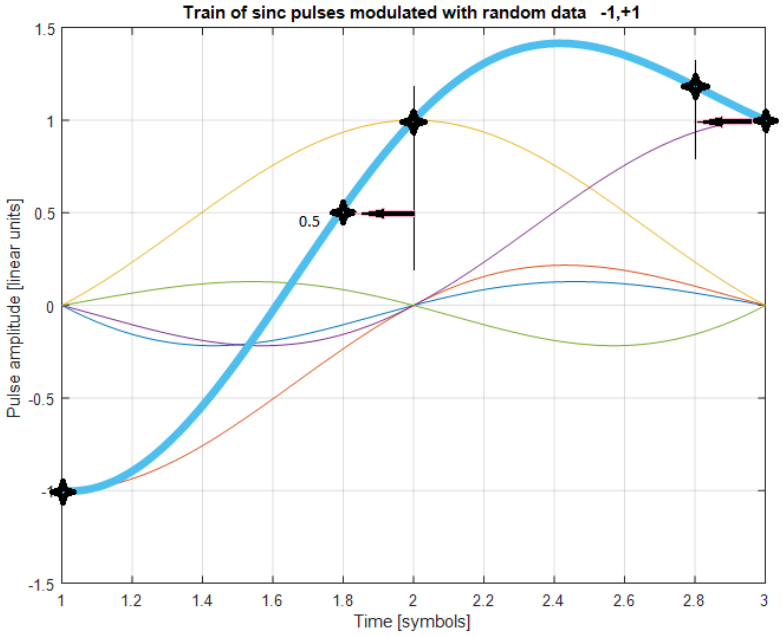


Figure 19.4.1: Pulses from Figure 19.2.1(b) with a timing error of -20%.

This will have a huge effect on the error performance of the receiver.
Note: The reason for the change in amplitude is due to two factors:

- The amplitude of the "wanted" pulse is reduced, and
- There are now contributions from other "unwanted" pulses, i.e. we have ISI back again.

So the point here is that even if we use ISI-free pulses, if we don't sample at the correct time instances, then we will have a certain amount of ISI in our system.
Note: Observe in Figure 19.3.5 that the larger the roll-off factor the smaller the amplitudes of the "side lobes"[1] are, and so the amount of ISI introduced by timing errors will be less. This leads to a general rule:
The larger the excess bandwidth, then less sensitive the system is to timing errors.

[1]The Raised Cosine pulse decays at a rate $\propto \frac{1}{t^3}$, whereas the since pulse decays only at a rate $\propto \frac{1}{t}$.

Chapter 20

Matched Filtering

20.1 Motivation

Lets revisit our example binary PAM receiver discussed in Section 19.1.1, but this lets add a bit more noise as illustrated in Figure 20.1.1. If we apply our simple decision rule:

$$\hat{b}_k = \begin{cases} +1 & \text{if } r(kT) > 0 \\ -1 & \text{if } r(kT) \leq 0 \end{cases}$$

then we get a bit error at the position shown ($\hat{b}_2 = -1$, where it should be $+1$).

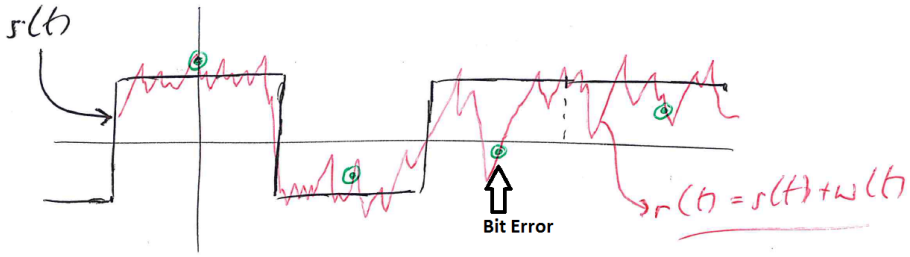


Figure 20.1.1: Binary PAM currupted by noise. Possible receiver sampling points are shown in thr middle of each symbol.

Surely we could have done better as it is (fairly) clear that had we taken an average of the signal over the symbol interval and used that in our decision rule instead then we would have got the correct result - this suggests an alternate receiver structure, shown in Figure 20.1.2, which we might expect to work better. The decision rule now becomes:

$$\hat{b}_k = \begin{cases} +1 & \text{if } y(kT) > 0 \\ -1 & \text{if } y(kT) \leq 0 \end{cases}$$

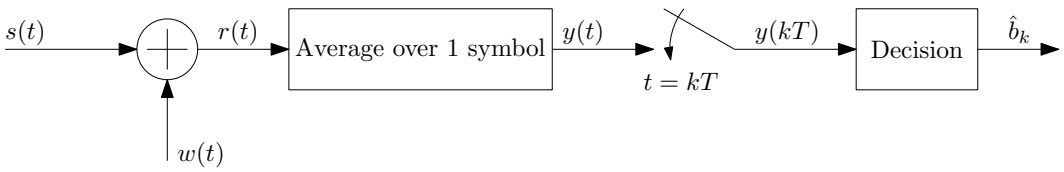


Figure 20.1.2: PAM receiver using an averaging process followed by a sampler and a simple decision rule.

20.1.1 Linear model

A device that averages its input over T seconds, like the one in Figure 20.1.2, can be written mathematically as:

$$y(t) = \int_{t-\frac{1}{2}T}^{t+\frac{1}{2}T} r(\tau)\, d\tau$$

(ignoring causality for the moment).

We can change the limits to be $\pm\infty$ by introducing a new function $h(\tau)$ as follows:

$$y(t) = \int_{-\infty}^{+\infty} h(t-\tau)\, r(\tau)\, d\tau \tag{20.1.1}$$

where

$$h(\tau) = \begin{cases} 1 & |\tau| < \frac{1}{2}T \\ 0 & \text{elsewhere} \end{cases}$$

But Equation 20.1.1 is just the output from a filter with input $r(t)$ and impulse response $h(t)$, a rectangular pulse of length T. So we have the linear model shown in Figure 20.1.3.

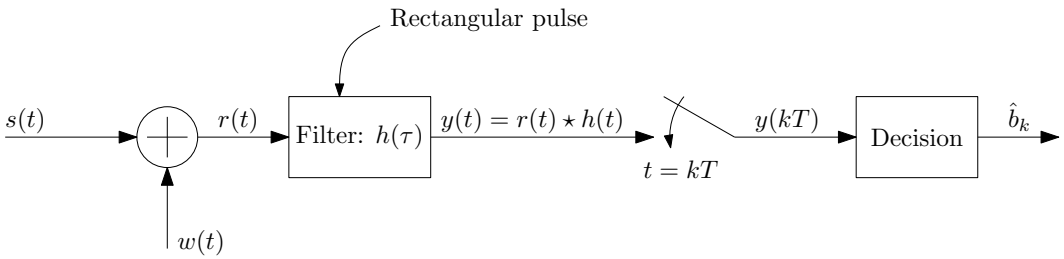


Figure 20.1.3: Linear model of PAM receiver shown in Figure 20.1.2

20.1.2 Questions

Ok, so we think we have a linear receiver that should work better than our previous attempt, but we still have the following questions:

1. Is it actually better?
2. Is this rectangular filter the best choice?
3. The example considered above is a PAM signal with a rectangular pulse shaping filter, $p(t)$, but what if a different $p(t)$ were used instead, e.g. a raised cosine pulse?
4. What if the noise is not white?

We now attempt to answer these.

20.2 The Matched Filter

20.2.1 Preliminaries; the Cauchy-Schwarz inequality

The Cauchy-Schwartz inequality is a very general inequality involving inner-products with many application from simple geometric proofs (the triangle inequality) to more exotic domains. We will use one particular version of it where it is applied to square-integrable complex-valued functions which we now state:

Theorem. *Given any two complex functions of a single real variable, call them $A(f)$ and $B(f)$ we have the following inequality:*

$$\left|\int_{-\infty}^{+\infty} A(f)\,B(f)\,df\right|^2 \;\leq\; \left(\int_{-\infty}^{+\infty} |A(f)|^2\,df\right)\left(\int_{-\infty}^{+\infty} |B(f)|^2\,df\right) \tag{20.2.1}$$

And, importantly, the equality is true when $A(f) \propto B^*(f)$.

Proof. See Theorem 5 in Appendix A. □

Essentially the inequality states that the integral on the LHS has a maximum value given by the product on the RHS, and it tells us under what conditions said function maximized.
It is necessary in many optimization problems where we wish to maximize some quantity, e.g. we will what to select a receive filter response that maximizes the receiver signal to noise ratio. The inequality not only tells us what the maximum SNR is, but also how to get it!

20.2.2 SNR of a single modulated pulse

Imagine just 1 symbol of information, call it b_0, was transmitted over an noisy channel using a pulse shaping filter $p(t)$, i.e. the received signal is:

$$r(t) = b_0 p(t) + w(t)$$

where $w(t)$ is the additive wideband noise.
Using the linear receiver proposed in the previous Section and illustrated in Figure 20.1.3, we can write the output from the receive filter as:

$$\begin{aligned} y(t) &= r(t) \star h(t) \\ &= b_0 p(t) \star h(t) + w(t) \star h(t) \\ &\triangleq b_0 g(t) + n(t) \end{aligned}$$

where $g(t) \triangleq p(t) \star h(t)$ and $n(t) \triangleq w(t) \star h(t)$ are the signal and (narrowband) noise components of the signal post-filtering retrospectively. We can write these in the frequency domain[1]:

$$\begin{aligned} \text{Signal: } G(f) &= P(f) H(f) \\ \text{PSD of Noise: } S_n(f) &= S_w(f) |H(f)|^2 \end{aligned}$$

Where $S_n(f)$ and $S_w(f)$ are the Power Spectral Densities (PSD) of the narrowband and wideband noise components respectively.
In our proposed receiver we would like to take a sample of $y(t)$ at time $t = 0$, i.e. obtain $y(0)$, and to base our decision just on this value. Thus we would like to choose the filter $h(t)$ (or equivalently $H(f)$) so as to maximize the Signal to Noise Ratio (SNR) at that time instant. To do this we will follow this procedure:

1. Derive an expression for the power of the wanted signal part $|b_0 g(0)|^2$ in terms of $P(f)$ and $H(f)$
2. Derive an expression for the noise power at time $t = 0$, i.e. $E[n^2(0)]$.
3. Compute the SNR, and use the Cauchy-Schwarz inequality to maximize it.

[1] We simply take the Fourier transform of the deterministic signal part, but for the noise part it makes more sense to characterize it's PSD.

4. Derive the conditions under which this maximization occurs - this will yield an expression for the optimum receive filter.

Let's begin:

Power of the wanted signal

We know that $G(f) = P(f) H(f)$. Applying the inverse Fourier Transform we get:

$$\begin{aligned} g(t) &= \int_{-\infty}^{+\infty} P(f) H(f) e^{j2\pi ft} df \\ \Rightarrow |b_0 g(0)|^2 &= b_0^2 \left| \int_{-\infty}^{+\infty} P(f) H(f) df \right|^2 \end{aligned}$$

Note how this already "looks" like the LHS of the Cauchy-Schwarz inequality...

Noise power at time $t = 0$

Lets just find $E[n^2(t)]$ first. Well this is just the power of the noise signal, which is obtained from it's PSD by integrating over the entire frequency range:

$$\begin{aligned} E[n^2(t)] &= \int_{-\infty}^{+\infty} S_n(f) df \\ &= \int_{-\infty}^{+\infty} S_w(f) |H(f)|^2 df \end{aligned}$$

which is independent of t, so it is also true at time $t = 0$.

SNR at time $t = 0$

The SNR we seek is given by:

$$\text{SNR} = b_0^2 \frac{\left| \int_{-\infty}^{+\infty} P(f) H(f) df \right|^2}{\int_{-\infty}^{+\infty} S_w(f) |H(f)|^2 df}$$

We wish to apply the Cauchy-Schwarz inequality to the numerator in such a way that one of the two resulting product terms, (see Equation 20.2.1), cancels with the denominator, so let:

$$\begin{aligned} \int_{-\infty}^{+\infty} |B(f)|^2 df &= \text{denominator} = \int_{-\infty}^{+\infty} S_w(f) |H(f)|^2 df \\ \Rightarrow \text{Let } B(f) &= \sqrt{S_w(f)} H(f) \end{aligned}$$

So we can now rewrite the SNR as:

$$\begin{aligned}\mathrm{SNR} &= b_0^2 \frac{\left|\int_{-\infty}^{+\infty}\left(\frac{P(f)}{\sqrt{S_w(f)}}\right)\left(\sqrt{S_w(f)}H(f)\right)df\right|^2}{\int_{-\infty}^{+\infty} S_w(f)\left|H(f)\right|^2 df} \\ &\leq b_0^2 \int_{-\infty}^{+\infty}\left|\frac{P(f)}{\sqrt{S_w(f)}}\right|^2 df\end{aligned}$$

Where we applied the Cauchy-Schwarz inequality in the 2^{nd} line and, by design, part of the numerator canceled with the denominator. This yields a simple expression for the maximum possible SNR

$$\mathrm{SNR}_{\max} = b_0^2 \int_{-\infty}^{+\infty} \frac{\left|P(f)\right|^2}{S_w(f)} df$$

Note that this maximum SNR only depends on the transmission system and the PSD of the additive noise (but is only achieved if the correct receive filter is used).

Condition for SNR maximization

The Cauchy-Schwarz inequality also tells us when the maximization occurs, namely when $A(f) \propto B^*(f)$ which in our case is:

$$\begin{aligned}\frac{P(f)}{\sqrt{S_w(f)}} &\propto \left(\sqrt{S_w(f)}H(f)\right)^* \\ \Rightarrow H(f) &\propto \frac{P^*(f)}{S_w(f)}\end{aligned}$$

This an incredible simple result!
It states that if you know the transmit pulse shape, and the PSD of the noise, you can uniquely (to within a scaling factor) compute the optimum receive filter.

This result underpins all digital communications and is not appreciated nor fully understood by many practitioners.

20.3 AWGN channel

The AWGN channel is so important it gets it's own section!
The only difference from the general derivation provided above is that the wideband noise

PSD is

$$S_w(f) = \frac{N_o}{2} \quad \forall f$$

Thus the optimum filter is:

$$H(f) \propto P^*(f)$$

where the proportionality absorbs the constant noise factor.
By taking the inverse Fourier transform of both sides we can derive the following time domain relationship:

$$h(t) \propto p^*(-t)$$

So the remarkably simple result is that the receive filter is just the time reversed conjugate filter of the transmit one!, i.e. the Rx filter is MATCHED to the Tx filter.

20.3.1 Rx = Tx filter

As a further special case consider, as is usually the case, if the transmit filter's impulse response is real and symmetric about $t = 0$ (remember the sinc, and the Raised Cosine filters were both examples where the impulse response was symmetric about $t = 0$, and it is true of any linear phase filter).
In this case the Rx filter, being the time reversed conjugate of the Tx filter, is actually the same (to within a constant scaling factor) as the Tx filter!

20.3.2 Questions answered

We are now in a position to answer the question posed in Section 20.1.2:

1. Is [this linear receiver] actually better?
   ANS: Yes, as the previous method didn't use a matched filter and so wasn't optimum, and so this Rx is better.

2. Is this rectangular filter the best choice?
   ANS: In that example, the Tx pulse was rectangular and so if the noise is AWGN then Yes, a rectangular Rx filter is best.

3. The example considered above is a PAM signal with a rectangular pulse shaping filter, $p(t)$, but what if a different $p(t)$ were used instead, e.g. a raised cosine pulse?
   ANS: If noise is AWGN then let : $h(t) \propto p^*(-t)$.

4. What if the noise is not white?
   ANS: If the PSD of the noise is $S_w(f)$, then let $H(f) \propto \frac{P^*(f)}{S_w(f)}$ and use the inverse Fourier transform to compute the Rx filter's impulse response.

20.3.3 Causality

In the above proof, causality was completely ignored, i.e. the Rx and Tx filter's response could begin at time $t < 0$ which in practice is not possible. To overcome this we simply add sufficient delays in the transmitter and receiver filters and sample the output of the receive filter with a delay equal to the sum of the Tx and Tx filter delays; problem solved!, but the math is much easier if we ignore causality.

20.3.4 Root-Raised-Cosine (RRC) filters

Lets consider our complete system, comprising both a Tx and an Rx filter as illustrated in Figure 20.3.1.

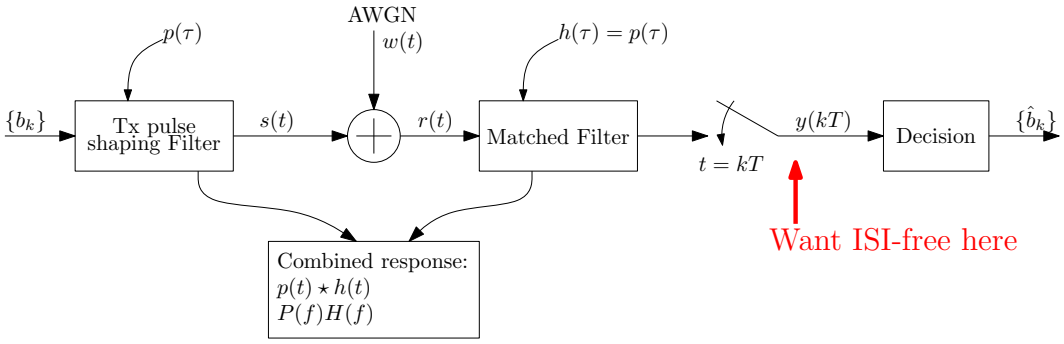


Figure 20.3.1: Full PAM system with linear Rx matched filter.

Assuming AWGN, and a real valued symmetric Tx pulse shaping filter, then the Rx filter $H(f) = P(f)$, i.e. they are the same. Notice in Figure 20.3.1 that we require ISI-free operation after the sampling device by which time each transmitted symbol has not just been filtered by the Tx pulse shaping filter, but also by the Rx matched filter - a combined response of $P(f) H(f) = P^2(f)$.
But Nyquist First Criteria (NFC) requires that a series of pulses are ISI-free iff they satisfy

some frequency domain requirement[2] which we previously applied only to the Tx filter, but in truth we should apply it to the combined Rx and Tx filter, which in this case is $P^2(f)$. Thus we should design $P(f)$ such that $P^2(f)$ is ISI free, not $P(f)$ itself, so that post matched filtering there will be not ISI at the ideal sampling points.

Perhaps the most common (2G, 3G, 4G, cellular systems etc....) choice is to make $P^2(f)$ a Raised Cosine as defined in Section 19.3.3, and so $P(f)$ is a square *Root-Raised-Cosine (RRC)*.

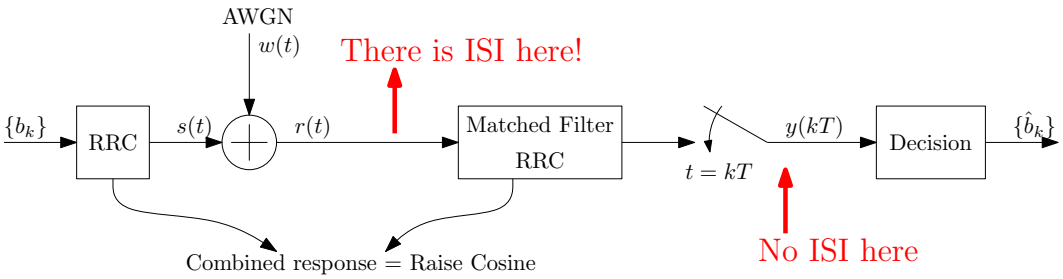


Figure 20.3.2: Full PAM system with RRC filters.

Note, an RRC filter is NOT and ISI free filter - it does not obey Nyquist's First Criteria for ISI-free transmission, so there is ISI present in the signal in the channel, but after matched filtering the combined response is ISI free. So the Rx RRC filter removes ISI and behaves as a matched filter maximizing the SNR.

[2]The repeated spectra is constant

Chapter 21

Bit Error Rate (BER) analysis

21.1 Introduction

Communication engineers spend a lot of time analyzing system under various channel conditions. This can be useful for a number of reasons:

1. If designing a system, this analysis can allow different parameter / system choices to be quantified as an aid to selection, e.g. the 3GPP had 100's of people working for years before finalizing the details of the various generations of cellular protocols.

2. For a given transmission system, the performance and structure of optimum receiver can be found - this serves as a bound beyond which any real world receiver can never better.

3. For a given receiver the performance can be computed and compared against the optimum or the minimum user requirements.

4. This allows a trade off between complexity and performance to be done.

The above tasks can be achieved analytically for many systems.
As an example will will consider the simplest of all systems, that of binary PAM transmission over an AWGN channel where the transmit pulse shaping filter is the square root of a ISI-free filter, e.g. a rectangular, or RRC pulse.
Before doing so, we'll first define a signal quality measure and the performance metrics we'll like to derive as a function of this signal quality.

21.2 Signal quality and performance metrics

21.2.1 SNR is not good enough

In analog modulation we considered the signal to noise power ratio (SNR) at the input to the receiver as a suitable measurement of the signal quality which worked well and allowed analysis of these systems in a semi-comparable way - this is the traditional way of analyzing analog communication systems and indeed in the early days of digital communication the same concept was also used with some success.
However as many different types of digital modulation schemes emerged it became necessary to define a new signal quality measurement that allowed easy comparison of diverse systems. Specifically two areas needed to be addressed:

- The noise power in the SNR measurement depends on the bandwidth, and so the meaning of SNR is different for different bandwidth systems. For this reason modern analysis simply uses N_o, the noise spectral density, instead of absolute noise power ($= N_o B_n$). This allows easy comparison of system with different bandwidths.
- Many digital modulation schemes operate at various bits rates. Moreover some bits may require more energy to transmit than other, e.g. consider multilevel PAM, and so to facilitate like-for-like system comparison the average Energy per bit, denoted E_b, is now generally used as a fairer measure.

Together these two points has lead to the definition of a new signal quality measure called *Energy per bit to noise spectral density ratio*, sometime called pronounced "Eb-No" :

$$\left(\frac{E_b}{N_o}\right)$$

It is measured at the input to the receiver, and if used correctly allows modulation schemes which hugely different characteristics to be effective compared in a like-for-like manner and has now become the de-facto standard.

21.2.2 BER and SER

Again in analog system we consider the relationship between the SNR at the output of a receiver compared to that of the input which made perfect sense for analog signals. However in digital systems we transmit symbols which usually represent bits of information, and

the "user" doesn't care about the SNR at the receiver output, he/she only cares about the probability of receiving the symbols or bits correctly.
We define the following performance metrics:

- BER: Bit Error Rate. This is the average number of bits error per every bit transmitted. For wireless system this can be as high as 10^{-3}, but for wired systems it is typically much smaller, e.g. 10^{-12}.
- SER: Symbol Error rate. This is the average number of symbol errors per every symbol transmitted.

In truth the "user" is primarily interested in the Bit Error Rate (BER) not the SER, but when it comes to analysis it is often mathematically intractable to accurately compute the BER, whereas close solutions for SER are often possible. So one common approach is to derive the SER, and then use some approximations to derive a less accurate BER.

21.2.3 Summary

We will attempt to compute the BER of our system in terms of $\frac{E_b}{N_o}$ as per Figure 21.2.1.

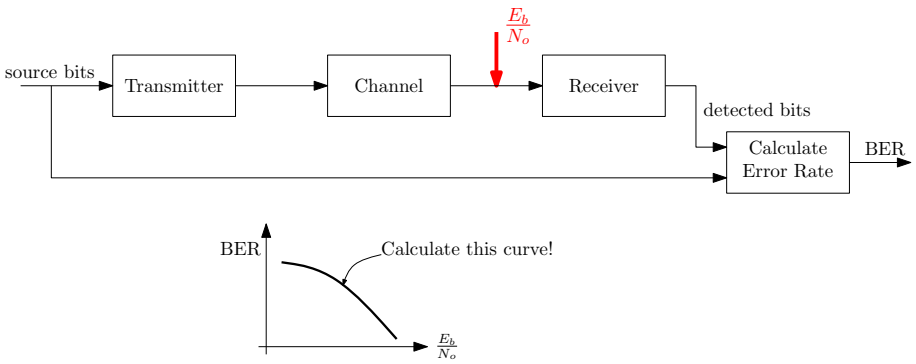


Figure 21.2.1: Assessment of digital communication systems

21.3 Binary PAM over AWGN

Consider the system illustrated in Figure 21.3.1. It is a binary PAM transmission system over an AWGN channel where the transmit pulse shaping filter is the square root of a ISI-free filter, e.g. a rectangular, or RRC pulse.

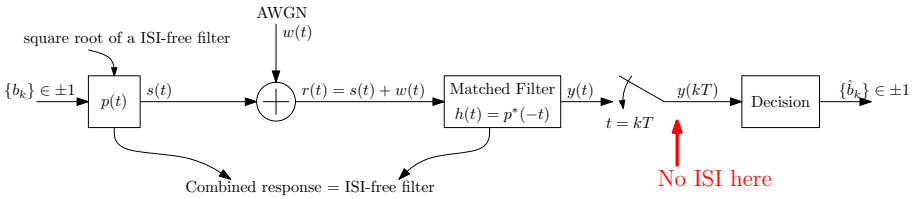


Figure 21.3.1: Binary PAM transmission over an AWGN channel with an optimum matched filter receiver.

Before considering the performance of the receiver, we should derive an expression for E_b, the Energy per bit at the input to the receiver (as this is independent of the receiver itself).

21.3.1 Energy per bit E_b

Each received pulse is either $b_k p(t)$, where $b_k = \pm 1$.
As the sign of the pulse doesn't effect the pulse's energy, it is clear that each pulse has the same energy, E_b, and it is given by:

$$E_b = \int_{-\infty}^{+\infty} p^2(t)\, dt$$

21.3.2 Post-Matched filter SNR

In the derivation of the matched filter we not only found the optimum filter response, but also the SNR achieved when said filter is used, the result was:

$$\text{SNR} = b_0^2 \int_{-\infty}^{+\infty} \frac{|P(f)|^2}{S_w(f)} df$$

But in our case the wideband noise is white, so $S_w(f) = \frac{N_o}{2}$ and the symbols (corresponding to b_0 in the above equation) are ± 1, so we have:

$$\text{SNR} = \frac{2}{N_o} \int_{-\infty}^{+\infty} |P(f)|^2 df$$

But the integral of $|P(f)|^2$ over all f is, by Parseval's theorem, is the same as integral of $p^2(t)$ over all time, thus:

$$\begin{aligned} \text{SNR} &= \frac{2}{N_o} \int_{-\infty}^{+\infty} p^2(t)\, dt \\ &= 2\left(\frac{E_b}{N_o}\right) \end{aligned}$$

21.3.3 Simplify the model

The combination of the transmit pulse shaping filter and the receive matched filter is assumed to be ISI free. Furthermore we can set the gain of the matched filter to be any value without affecting the resulting SNR, thus we can set the gain to a value that results in the signal part of $y(kT) = \pm 1$, i.e. the signal power at the sample instances can be set to unity, and thus the amount of noise power is $= \frac{1}{\text{SNR}} = \frac{N_o}{2E_b}$.

Putting these together allows use to develop a simplified sampled model as illustrated in Figure 21.3.2.

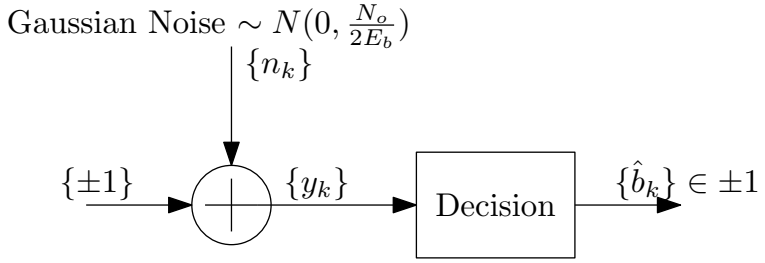


Figure 21.3.2: Simplified sampled model of binary PAM system.

21.3.4 BER computation

Two possible types of bit errors can occur:

- Type A: $b_k = -1$, but $\hat{b}_k = +1$
- Type B: $b_k = +1$, but $\hat{b}_k = -1$

Lets start by computing the probability of Type A:

21.3.4.1 Type A: $b_k = -1$, but $\hat{b}_k = +1$

When $b_k = -1$ we have:

$$y(kT) = -1 + n(kT)$$

Where $n(kT)$ is a sample of the zero-mean Gaussian noise having variance (=average power) of $\sigma^2 = \frac{N_o}{2E_b}$.

Thus $y\left(kT\right)$ has mean $=-1$ and $\sigma^2=\frac{N_o}{2E_b}$, and has probability density function:

$$f_Y\left(y\right) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y+1\right)^2}{2\sigma^2}\right)$$

and is illustrated in Figure 21.3.3.

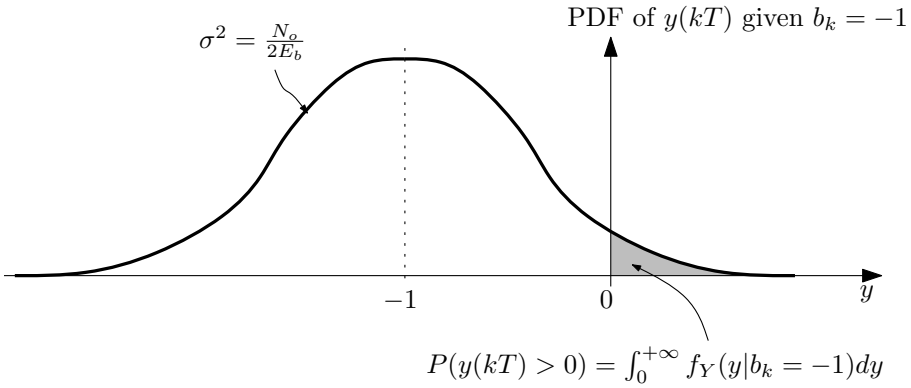


Figure 21.3.3: PDF of the sampled signal $y\left(kT\right)$ when $b_k=-1$

If we apply our simple decision rule:

- Let $\hat{b}_k=+1$, if $y\left(kT\right)>0$,
- Let $\hat{b}_k=-1$, if $y\left(kT\right)\leq 0$,

Then we will have an error if $y\left(kT\right)>0$ (as we assumed that $b_k=-1$), so the probability this type of error is:

$$\begin{aligned}
P\left(\text{error}\middle|\, b_k=-1\right) &= \int_0^{\infty} f_Y\left(y\,|\,b_k=-1\right)dy \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_0^{\infty}\exp\left(-\frac{\left(y+1\right)^2}{2\sigma^2}\right)dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{\frac{1}{\sigma}}^{\infty}\exp\left(-\frac{1}{2}x^2\right)dx \\
&= Q\left(\frac{1}{\sigma}\right)=Q\left(\sqrt{2\frac{E_b}{N_o}}\right)
\end{aligned}$$

Where the $Q\left(\cdot\right)$ function was defined in Section 12.1.3.5:

$$Q\left(\alpha\right) \;\triangleq\; \frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\infty}\exp\left(-\frac{x^2}{2}\right)dx$$

21.3.4.2 Type B: $b_k=+1$, but $\hat{b}_k=-1$

By symmetry it is easy to prove that type A errors have the same probability as Type A ones.

21.3.5 Overall BER

The overall BER is:

$$\begin{aligned} P_e &= P(b_k = -1)\, P(\text{error}|\, b_k = -1) + P(b_k = +1)\, P(\text{error}|\, b_k = +1) \\ &= 0.5Q\left(\sqrt{2\frac{E_b}{N_o}}\right) + 0.5Q\left(\sqrt{2\frac{E_b}{N_o}}\right) \\ &= Q\left(\sqrt{2\frac{E_b}{N_o}}\right) \end{aligned}$$

Normally when this is plotted both the BER and the $\frac{E_b}{N_o}$ are drawn on a log scale as shown in Figure 21.3.4.

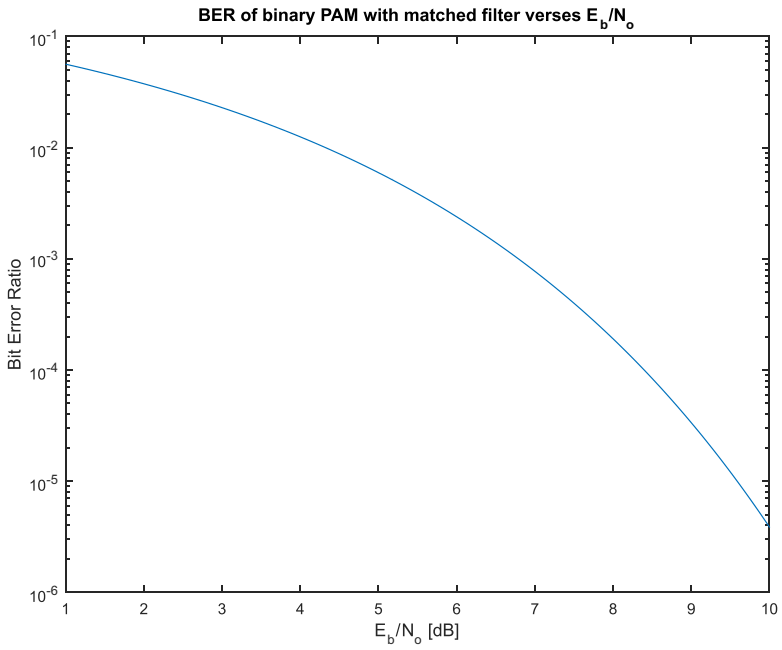


Figure 21.3.4: The BER of a binary PAM system using an optimum matched filter.

Note: The BER does not depend on the type of pulse used!, it just depends on the Energy of that pulse at the input to the receiver, and of course the amount of noise.

Chapter 22

High order modulation

22.1 Introduction

Thus far we've only really considered binary baseband (i.e. low frequency) digital transmission schemes such as binary PAM with arbitrary pulse shaping filters. However many of these schemes can be generalized to:

- Have more bits per symbol, i.e. non-binary systems, and
- To utilize a high frequency channel, e.g. a wireless channel, using what is called passband (or bandpass) transmission.

We will now consider some of these generalizations.

22.2 Multi-level PAM

22.2.1 Recap: Binary PAM

In binary PAM, the binary data to be transmitted (e.g. a collection of 1's and 0's) are mapped to a binary alphabet of symbols $b_k \in \{\pm 1\}$, and these are used to drive a pulse shaping filter, $p(\tau)$, generating a train of either positive of negatively modulated pulses.

$$s(t) = \sum_{k=-\infty}^{+\infty} b_k p(t - kT)$$

This can be represented as points on a constellation diagram as illustrated in Figure 22.2.1.

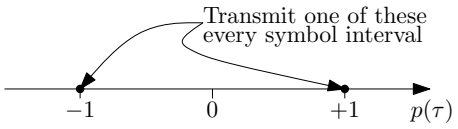


Figure 22.2.1: Binary PAM constellation diagram (x-axis is amount of $p(\tau)$ to send each time).

Assuming ISI-free reception in the presence of noise the PDF of the sampled signal, $y(kT)$, at the output of the receiver's matched filter will be something like that illustrated in Figure 22.2.2:

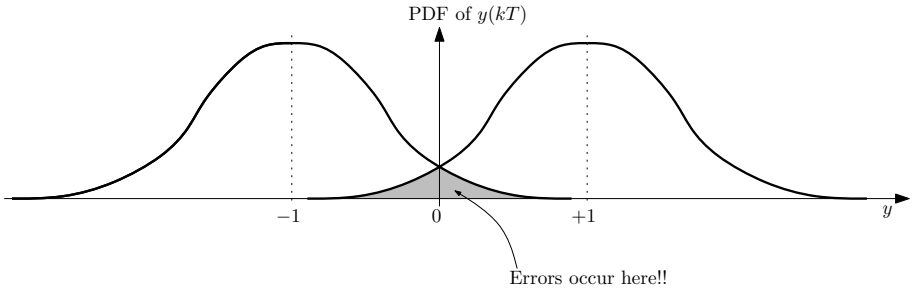


Figure 22.2.2: PDF of the sampled signal at output from receiver's matched filter in an ISI Free binary PAM system.

22.2.2 Extend to multilevel PAM

All the same principals of ISI free transmission and matched filtering reception can be applied to multilevel PAM modulation schemes. Usually the number of constellation point is a power-of-2 so that we can have a simple bits-to-symbol mapping process. For example we could group the incoming bit stream into pairs (groups of 2) resulting in $2^2 = 4$ possible symbols which could be $\{\pm 1, \pm 3\}$ as shown in Figure 22.2.3.

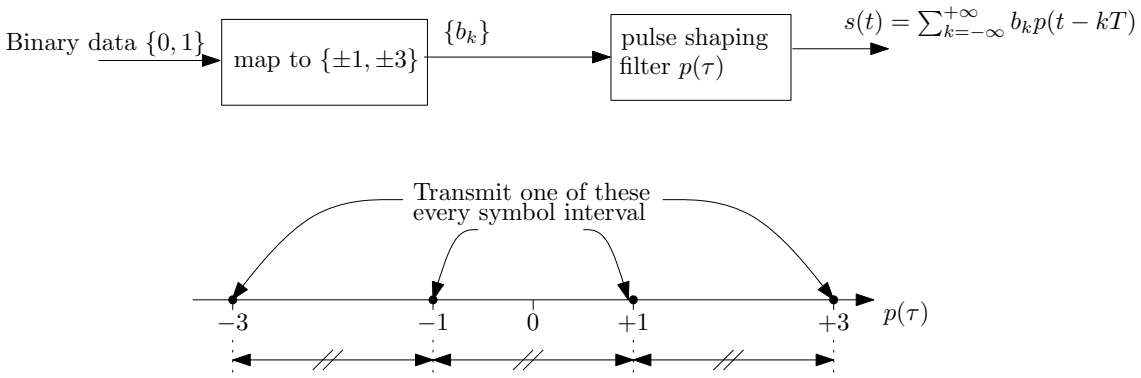


Figure 22.2.3: Multilevel PAM constellation diagram (x-axis is amount of $p(\tau)$ to send each time).

Again assuming ISI-free reception in the presence of noise the PDF of the sampled signal,

$y(kT)$, at the output of the receiver's matched filter will be something like that illustrated in Figure 22.2.4:

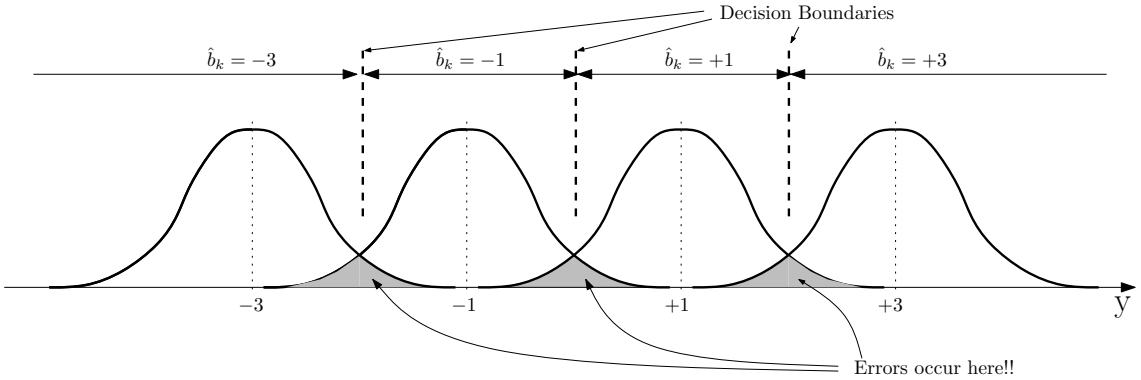


Figure 22.2.4: PDF of the sampled signal at output from receiver's matched filter in an ISI Free multilevel PAM system.

Note the position of the decision boundaries, and how symbols $b_k = \pm 3$ are most likely to be received correctly as there are "at the edge", whereas symbols $b_k = \pm 1$, can easily be mistaken for their neightbours on either side. This is known as a constellation *edge effect*.
This signal will have twice the data rate compared to binary PAM, and will have the same spectrum, as it only depends on the spectrum of the pulse shaping filter which can be the same in both cases - so, as if by magic, we can get twice the amount of information into the same bandwidth! It has higher number of bits per Hz.
However you don't something for nothing in this world, as we will now see
Assuming the symbols are all equally likely, i.e. probability $= 0.25$, then the average energy per symbol is:

$$\begin{aligned} E_s &= \left[0.25 \times (-3)^2 + 0.25 \times (-1)^2 + 0.25 \times (+1)^2 + 0.25 \times (+3)^2\right] E_p \\ &= 5E_p \end{aligned}$$

where E_p is the Energy in a single pulse $p(t)$.
As there two bits per symbol, the average energy per bit is:

$$E_b = 2.5E_p$$

This compares to binary PAM, where the average energy per bit is:

$$\begin{aligned} E_b &= \left[0.5 \times (-1)^2 + 0.5 \times (+1)^2\right] E_p \\ &= E_p \end{aligned}$$

i.e. the E_b used in the multilevel PAM scheme illustrated in Figure 22.2.3 is $\times 2.5$ the E_b used in 22.2.1. This will "shift" the BER curve to the right by $10\log_{10}(2.5) = 4\text{dB}$.
However the symbol error rate will be approximately the same (as can be seen by the proportion of the PDFs shown in Figures 22.2.2 and 22.2.4 where errors occur)[1], and therefore the BER is approximately halved, i.e. the BER curve "shifts" down by an amount. Examination of the BER curve shown in Figure 21.3.4 demonstrated that this corresponds to no more than about 1dB.
Conclusion is that the resulting BER curve (plotted against E_b/N_o) is shifted to the right by about 3dB - this is an approximate calculation, more accurate analysis can be done, but are beyond the scope of this course.

22.2.3 In summary

We can easily modify the binary PAM scheme discussed in previous chapters to permit multilevel signal without increasing the bandwidth thereby improving the spectrum utilization however there is a downside - the BER, for a given E_b/N_o gets worse, which may or may not be acceptable.

22.3 Passband PAM system

Here the channel is a high frequency, or passband channel, so just as was the case in AM chapters it is necessary to "shift" the baseband signal up in frequency to match that of the channel permitting transmission.

22.3.1 Transmitter

This can be done by a simple multiplication by a local oscillator in the transmitter and subsequent frequency down shifting in the receiver, as was done in the AM synchronous demodulator. This is shown in Figure 22.3.1.

[1] This is only an approximate calculation as we haven't taken into account constellation edge effects where ± 3 symbols perform better than ± 1 - this can easily be done.

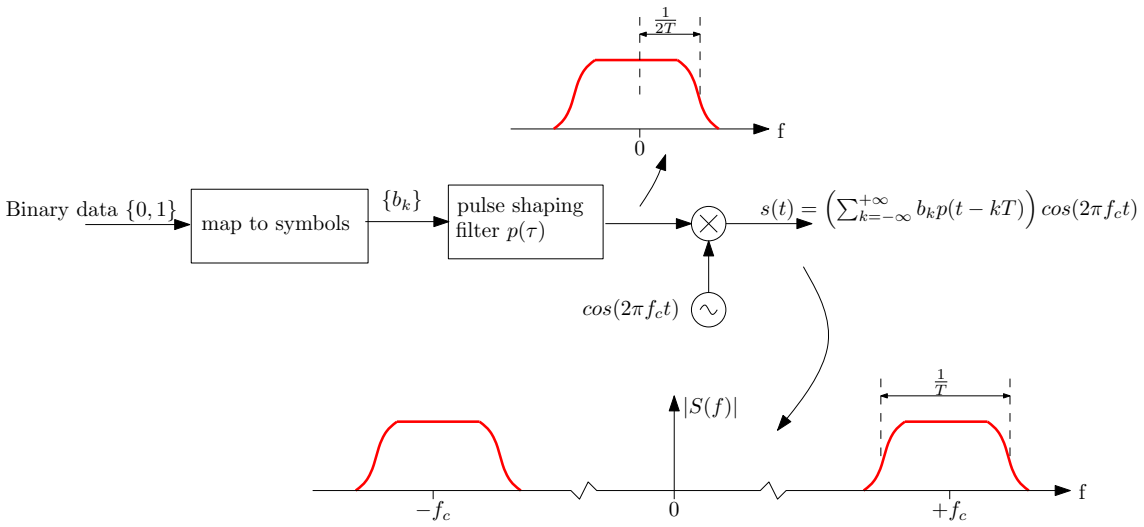


Figure 22.3.1: Passband multilevel PAM transmitter.

The resulting transmitted signal is given by:

$$s(t) = \left(\sum_{k=-\infty}^{+\infty} b_k p(t - kT) \right) \cos(2\pi f_c t)$$

and has a spectrum similar to that also shown in Figure 22.3.1.

The minimum bandwidth is $\frac{1}{T}$Hertz, which occurs when we use a sinc pulse shaping filter, but in reality there is an amount of excess bandwidth as previously defined resulting in an occupied bandwidth of $\frac{1}{T}(1+\beta)$ Hertz, where β is the roll-off factor.

22.3.2 Receiver

Just as in AM, we can implement a synchronous demodulator where we multiply the received signal by a local cosine reference signal as shown in Figure 22.3.2.

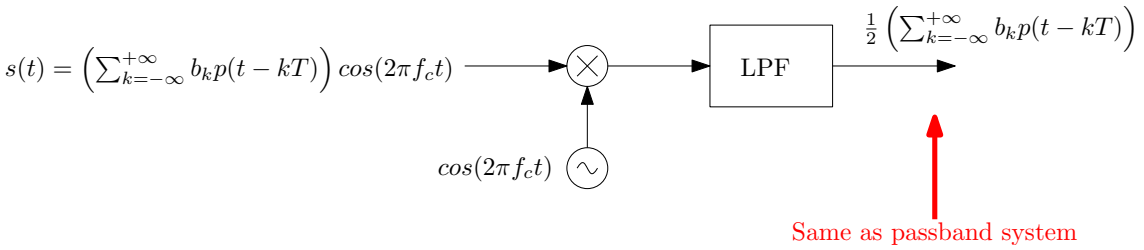


Figure 22.3.2: A down conversion process used in a passband receiver to produce a signal suitable for normal baseband reception, e.g. matched filtering etc...

Assuming no noise nor any other channel effects, the output from the multiplier is:

$$\begin{aligned} s(t)\cos(2\pi f_c t) &= \left(\sum_{k=-\infty}^{+\infty} b_k p(t-kT)\right)\cos^2(2\pi f_c t) \\ &= \frac{1}{2}\left(\sum_{k=-\infty}^{+\infty} b_k p(t-kT)\right)(1+\cos(4\pi f_c t)) \end{aligned}$$

After the Low pass filter, we have a signal proportional to:

$$\sum_{k=-\infty}^{+\infty} b_k p(t-kT)$$

, as per a baseband system!.

Therefore we can apply all the same matched filtering and ISI free transmission theory developed for passband transmission to this system instead.

22.4 QAM

As per Section 7.2 it is possible to modulate two signals, one with cosine, the other with sine, and to simply add them together prior to transmission over the same passband channel - essentially we can transmit two signals at the same time on the same channel without interference.

22.4.1 QAM transmission

Using this technique we can double the spectral efficiency of the passband PAM scheme (Section 22.3) presented above by sending two different PAM signals at the same time - one modulated using cosine, the other using sine. This is illustrated in Figure 22.4.1

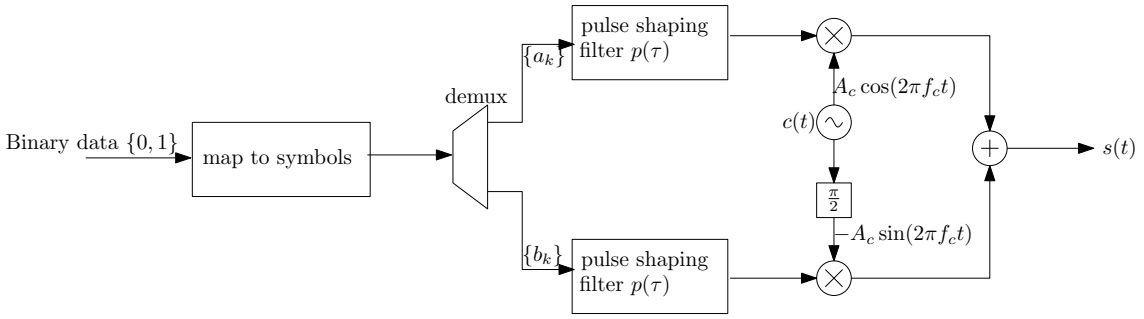


Figure 22.4.1: An example QAM transmitter. This can be viewed as two PAM transmitters (90^o apart). The symbol stream is divided between the two independent paths such that the overall data rate is doubled, i.e. different data is sent on each path, the symbols are NOT duplicated.

22.4.2 Constellation diagram

The constellation diagram for this scheme is now a 2-D plot, one axis showing the amount of the pulse sent each symbol interval on cosine, the other axis corresponds to sine. This is shown in Figure 22.4.2 assuming two 4-level PAM paths, corresponding to a 16-point constellations; hence this modulation scheme is called 16-QAM.

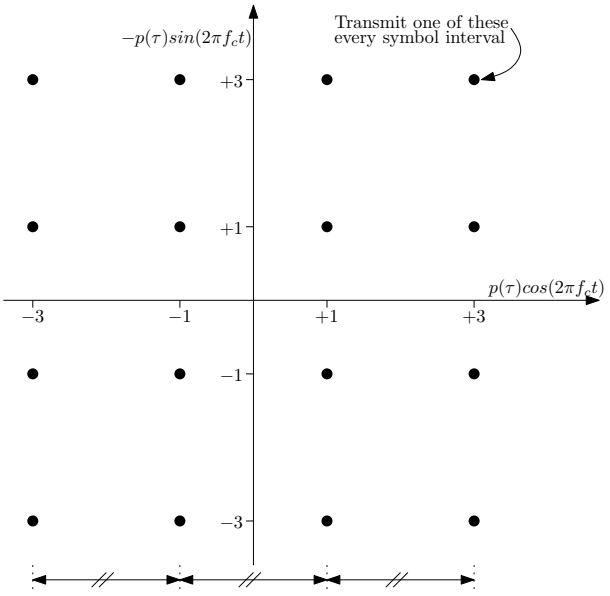


Figure 22.4.2: The constellation diagram for a 16-QAM system, i.e. $a_k, b_k \in \{\pm 1, \pm 3\}$.

22.4.3 QAM receiver

Again, borrowing ideas from Section 7.2 we have the synchronous receiver which can now be combined with the theory of matched filtering to produce the receiver architecture illustrated in Figure 22.4.3.

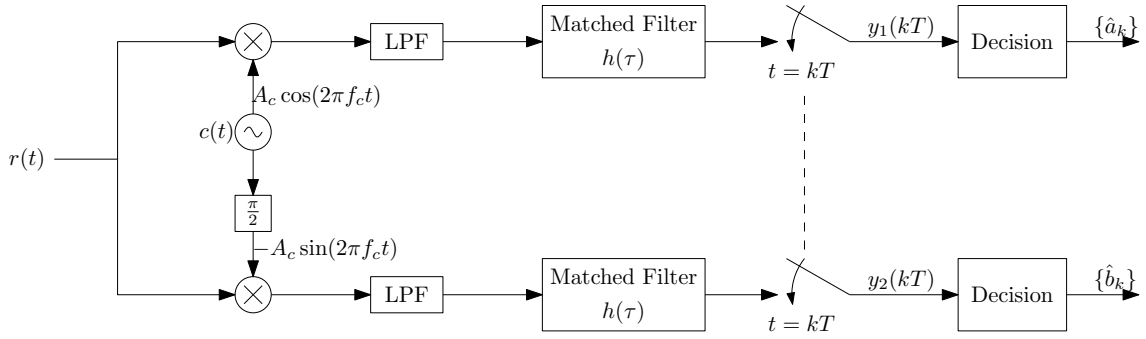


Figure 22.4.3: An example passband QAM receiver; really its just two PAM receivers.

Provided the noise on each of the two paths is independent (which it is in the case of AWGN), then decisions can be made on each of the two paths independently as shown in Figure 22.4.3.

Appendix A

Various theorems

A.1 Fourier Transform theorems

Theorem 1. *Fourier transform of complex exponential*

Claim. The Fourier transform of $x(t) = e^{j2\pi f_o t}$ is $\delta(f - f_o)$

Proof. The sifting property of the delta function tells us how it behaves in an integral:

$$\int_{-\infty}^{+\infty} \delta(t - t_o)\, g(t)\, dt \quad = \quad g(t_o)$$

of course renaming t to f doesn't change a thing:

$$\int_{-\infty}^{+\infty} \delta(f - f_o)\, g(f)\, df \quad = \quad g(f_o)$$

Thus we can write the function $e^{j2\pi f_o t}$ as:

$$\begin{aligned} e^{j2\pi f_o t} &= \int_{-\infty}^{+\infty} \delta(f - f_o)\, e^{j2\pi f t} df \\ &= \mathcal{F}^{-1}[\delta(f - f_o)] \\ \Rightarrow \mathcal{F}\left[e^{jf_o t}\right] &= \delta(f - f_o) \end{aligned}$$

□

Based on this we have the following two theorems:

Theorem 2. *Fourier transform of* $\cos(2\pi f_o t)$ *is* $\frac{1}{2}(\delta(f - f_o) + \delta(f + f_o))$.

Proof. Euler's equation gives us:

$$\begin{aligned} e^{j2\pi f_o t} &= \cos(2\pi f_o t) + j\sin(2\pi f_o t) \\ e^{-j2\pi f_o t} &= \cos(2\pi f_o t) - j\sin(2\pi f_o t) \end{aligned}$$

Adding (both sides of) these two equations yields

$$\begin{aligned} \cos(2\pi f_o t) &= \frac{1}{2}\left(e^{j2\pi f_o t} + e^{-j2\pi f_o t}\right) \\ \Rightarrow \mathcal{F}\left[\cos(2\pi f_o t)\right] &= \frac{1}{2}\left(\mathcal{F}\left[e^{j2\pi f_o t}\right] + \mathcal{F}\left[e^{-j2\pi f_o t}\right]\right) \\ &= \frac{1}{2}\left(\delta(f - f_o) + \delta(f + f_o)\right) \end{aligned}$$

□

And

Theorem 3. *Fourier transform of* $\sin(2\pi f_o t)$ *is* $\frac{1}{2j}\left(\delta(f - f_o) - \delta(f + f_o)\right)$.

Proof. Subtracting (both sides of) the two Euler equations in the previous proof we get:

$$\begin{aligned} \sin(2\pi f_o t) &= \frac{1}{2j}\left(e^{j2\pi f_o t} - e^{-j2\pi f_o t}\right) \\ \Rightarrow \mathcal{F}\left[\cos(2\pi f_o t)\right] &= \frac{1}{2j}\left(\mathcal{F}\left[e^{j2\pi f_o t}\right] - \mathcal{F}\left[e^{-j2\pi f_o t}\right]\right) \\ &= \frac{1}{2j}\left(\delta(f - f_o) - \delta(f + f_o)\right) \end{aligned}$$

□

A.1.1 Fourier transform pair summary

	$x(t)$	$(\mathcal{F}[x(t)])(f)$
1	$e^{j2\pi f_o t}$	$\delta(f - f_o)$
2	$\cos(2\pi f_o t)$	$\frac{1}{2}\left(\delta(f - f_o) + \delta(f + f_o)\right)$
3	$\sin(2\pi f_o t)$	$\frac{1}{2j}\left(\delta(f - f_o) - \delta(f + f_o)\right)$

Table A.1: Some common Fourier Transform pairs.

A.2 Statistical proofs

Theorem 4. *The variance of a zero uniform RV on* $\pm\frac{1}{2}\Delta$ *is* $\frac{\Delta^2}{12}$

Proof. First compute the non-central second moment:

$$\begin{aligned} E\left[X^2\right] &= \int_{-\infty}^{+\infty} x^2 f\left(x\right) dx \\ &= \int_{-\Delta/2}^{+\Delta/2} \frac{x^2}{\Delta} dx = \frac{1}{\Delta}\left.\frac{x^3}{3}\right|_{-\Delta/2}^{+\Delta/2} \\ &= \frac{1}{3\Delta}\left[\left(\Delta/2\right)^3 - \left(-\Delta/2\right)^3\right] \\ &= \frac{\Delta^2}{12} \end{aligned}$$

The variance is given by:

$$\begin{aligned} Var\left[X\right] &= E\left[X^2\right] - \left(E\left[X\right]\right)^2 \\ &= \frac{\Delta^2}{12} \end{aligned}$$

because the mean of a symmetric RV is zero. □

A.3 Cauchy-Schwarz

Theorem 5. *The Cauchy-Schwarz inequality:*
If $\phi_1\left(x\right)$ *and* $\phi_2\left(x\right)$ *are both square integrable complex functions of a real variable* x*, i.e.:*

$$0 < \int_{-\infty}^{+\infty} \left|\phi_i\left(x\right)\right|^2 dx \;\; < \;\; \infty$$

Then the Cauchy-Schwarz inequality states that:

$$\left|\int_{-\infty}^{+\infty} \phi_1\left(x\right)\phi_2\left(x\right) dx\right|^2 \;\; \leq \;\; \int_{-\infty}^{+\infty} \left|\phi_1\left(x\right)\right|^2 dx \int_{-\infty}^{+\infty} \left|\phi_2\left(x\right)\right|^2 dx$$

and that equality occurs when $\phi_1\left(x\right) = \lambda\phi_2^*\left(x\right)$ *where the proportional constant,* λ*, is real valued.*

Proof. Firstly, the equality case can be easily proven by substituting $\phi_1(x) = \lambda\phi_2^*(x)$:

$$\begin{aligned}
\left|\int_{-\infty}^{+\infty} \lambda\phi_2^*(x)\,\phi_2(x)\,dx\right|^2 &\stackrel{?}{=} \int_{-\infty}^{+\infty} |\lambda\phi_2^*(x)|^2\,dx \int_{-\infty}^{+\infty} |\phi_2(x)|^2\,dx \\
\Rightarrow \left|\int_{-\infty}^{+\infty} |\phi_2(x)|^2\,dx\right|^2 &\stackrel{?}{=} \int_{-\infty}^{+\infty} |\phi_2^*(x)|^2\,dx \int_{-\infty}^{+\infty} |\phi_2(x)|^2\,dx \\
\Rightarrow \left(\int_{-\infty}^{+\infty} |\phi_2(x)|^2\,dx\right)^2 &\stackrel{?}{=} \left(\int_{-\infty}^{+\infty} |\phi_2(x)|^2\,dx\right)^2
\end{aligned}$$

Which, of course, satisfies equality. So all we need to do now is prove that the inequality holds for all $\phi_1(x)$ and $\phi_2(x)$; this is the difficult part!

Start be converting the problem statement into one involving real valued functions.

Consider the LHS of the inequality and note that it itself is guaranteed to be$\leq \left(\int_{-\infty}^{+\infty} |\phi_1(x)|\,|\phi_2(x)|\,dx\right)^2$ and so *if* the following can be proven:

$$\left(\int_{-\infty}^{+\infty} |\phi_1(x)|\,|\phi_2(x)|\,dx\right)^2 \leq \int_{-\infty}^{+\infty} |\phi_1(x)|^2\,dx \int_{-\infty}^{+\infty} |\phi_2(x)|^2\,dx$$

then the Cauchy-Schwarz is also true. But now the above inequality is one involving real valued non-negative functions which, for mathematical convenience, we define as:

$$\psi_1(x) = |\phi_1(x)| \quad \text{and} \quad \psi_2(x) = |\phi_2(x)|$$

and we try to prove the following:

$$\left(\int_{-\infty}^{+\infty} \psi_1(x)\,\psi_2(x)\,dx\right)^2 \leq \int_{-\infty}^{+\infty} \psi_1^2(x)\,dx \int_{-\infty}^{+\infty} \psi_2^2(x)\,dx \tag{A.3.1}$$

To do this consider the following integral:

$$\int_{-\infty}^{+\infty} (\lambda\psi_1(x) + \psi_2(x))^2\,dx$$

where λ is real valued. Clearly this is a non-negative real valued function. It is also a quadratic in λ as can be seen as follows:

$$\begin{aligned}
\int_{-\infty}^{+\infty} (\lambda\psi_1(x) + \psi_2(x))^2\,dx &= \left(\int_{-\infty}^{+\infty} \psi_1^2(x)\,dx\right)\lambda^2 + \left(2\int_{-\infty}^{+\infty} \psi_1(x)\,\psi_2(x)\,dx\right)\lambda + \left(\int_{-\infty}^{+\infty} \psi_2^2(x)\,dx\right) \\
&= a\lambda^2 + b\lambda + c
\end{aligned}$$

where we define the three real valued functions (of x) as:

$$\begin{aligned} a(x) &\triangleq \int_{-\infty}^{+\infty} \psi_1^2(x)\,dx \geq 0 \\ b(x) &\triangleq 2\int_{-\infty}^{+\infty} \psi_1(x)\,\psi_2(x)\,dx \geq 0 \\ c(x) &\triangleq \int_{-\infty}^{+\infty} \psi_2^2(x)\,dx \geq 0 \end{aligned}$$

As we know the quadratic to be real and non-negative it must either have complex roots, or a double real root. For this to be true we require (from the famous quadratic solution equation[1]) $b^2 \leq 4ac$, so we have:

$$\left(\int_{-\infty}^{+\infty} \psi_1(x)\,\psi_2(x)\,dx\right)^2 \leq \int_{-\infty}^{+\infty} \psi_1^2(x)\,dx \int_{-\infty}^{+\infty} \psi_2^2(x)\,dx$$

But this is exactly Equation A.3.1 and thus the Cauchy-Schwarz inequality is proven. □

[1] Roots of a quadratic are at: $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$