Digital Modulation

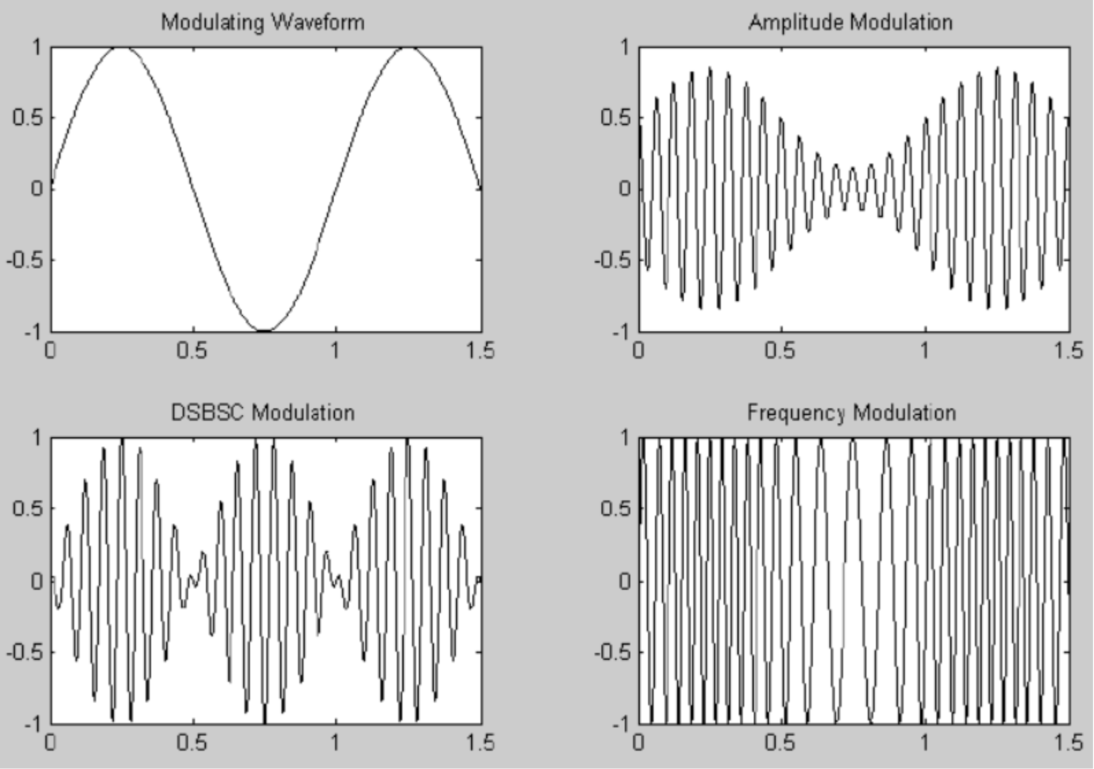
Contents

1 Analogue Modulation

Concepts from analogue modulation are often useful to help in our understanding of digital modulation techniques. Analogue Modulation techniques include

- Amplitude Modulation (AM)
  - ◇ *Where a carrier wave is modulated by a signal in terms of the amplitude.*
- Double-SideBand Suppressed Carrier (DSBSC)
  - ◇ *Similar to AM but with suppressed carrier component.*
- Single-SideBand (SSB)
  - ◇ *Similar to DSBSC but only one side-band.*
- Phase Modulation (PM)
  - ◇ *Where the modulating signal is carried by means of varying the phase of the carrier wave.*
- Frequency Modulation (FM)
  - ◇ *Where the modulating signal is carried by means of varying the frequency of the carrier wave.*

Some of these are illustrated below. In the top left is the modulating waveform, this is the signal of interest before modulation has taken place. The carrier wave is a higher frequency sine wave that can be used to communicate the signal described by the modulating waveform.



Modulation can be treated mathematically. We will not go into much detail for the analogue modulation. This is just an overview of analogue modulation to help provide a bit of background that is useful for the understanding of digital modulation techniques. Let us describe the un-modulated carrier as a sine wave:

$$x_c(t) = A\cos\left(2\pi f_c t + \phi\right).$$

The terms in this sine wave should be familiar with all of us by now and include: A – amplitude, f_c – carrier frequency, ϕ – phase offset and often $\phi = 0$. *Sometimes* $\sin()$ *used,* $\cos(2\pi f_c t) = \sin(2\pi f_c t + \pi/2)$*.*

As we have seen there are a number of different analogue modulation techniques. We can describe them with the same general formula which we will refer to as the **General Modulated Wave Formula**. But first let us look at some equations that can be used to describe the various modulation techniques.

For **Amplitude Modulation:** we can describe it with a cosine wave where the amplitude $A(t)$ is varied as a function of time t:

$$x_m(t) = A(t)\cos\left(2\pi f_c t + \phi\right).$$

For **Phase Modulation (including FM):** we can describe this type of modulation by varying the phase $\phi(t)$ as a function of time:

$$x_m(t) = A\cos\left(2\pi f_c t + \phi(t)\right).$$

As noted this equation for the description of phase modulation can also be used to describe frequency modulation. This is an interesting point which requires us to look at a simple derivation which we will do shortly.

So far we have seen that the phase and amplitude can be varied as functions of time to describe two main different types of modulation techniques. However another type of modulation technique known as Quadrature Amplitude Modulation (QAM) varies both the amplitude and phase simultaneously. We can describe QAM with the **General Modulated Wave Formula:**

$$x_m(t) = A(t)\cos\left(2\pi f_c t + \phi(t)\right),$$

QAM modulates both $A(t)$ *and* $\phi(t)$.

1.1 Amplitude Modulation

So how does the amplitude $A(t)$ in Amplitude Modulation vary as a function of time? It should depend on the information in the modulating signal waveform, of course!

$$\begin{aligned} x_m(t) &= A(t)\cos(2\pi f_c t + \phi) \\ &= E_c\left(\frac{1+m(t)}{2}\right)\cos(2\pi f_c t) \end{aligned}$$

where $A(t) = E_c\frac{1+m(t)}{2}$, $m(t) \in [-1, +1]$ is the modulating signal and E_c is the peak amplitude of the carrier wave. Here the phase is taken to be zero, *i.e.* $\phi = 0$. The constant is added to the carrier wave to enable it to take positive values only which enables it to modulate the envelope of the carrier wave.

1.2 Double-SideBand Suppressed Carrier

Double-SideBand Suppressed Carrier modulation suppresses the carrier wave which helps reduce the amount of power required to transmit a signal. Here the modulation is simply the product of the carrier wave with the modulating signal waveform $m(t)$:

$$x_m(t) = E_c \times m(t) \times \cos(2\pi f_c t)$$

where the amplitude is given by $A(t) = E_c \times m(t)$ and again there is zero phase $\phi = 0$.

1.3 Phase Modulation

Phase Modulation requires the phase to be varied:

$$x_m(t) = E_c \times \cos(2\pi f_c t + \kappa m(t))$$

where the amplitude is constant $A = E_c$, the phase varies $\phi = \kappa \times m(t)$ and κ is the peak phase deviation.

1.4 Frequency Modulation

As already noted Frequency Modulation can be described using the same equation that was used to describe Phase Modulation:

$$x_m(t) = A\cos(2\pi f_c t + \phi(t)).$$

We can account for all terms in this equation except for the phase $\phi(t)$. It is difficult to immediately see how varying the frequency will affect the phase $\phi(t)$. What we have to do is to describe $\phi(t)$ in terms of a frequency form. This can be done by taking the derivative with respect to time which changes the units to one of angle changes over time which is in effect an angular frequency formulation.

To start, we describe the argument of the cosine as the **Instantaneous angle**:

$$\theta_i = \theta_i(t) = 2\pi f_c t + \phi(t) \text{ radians}$$

Taking derivative of the phase with respect to time we get **Instantaneous angular frequency**:

$$\omega_i(t) = \frac{\mathrm{d}\theta_i}{\mathrm{d}t} = 2\pi f_c + \frac{\mathrm{d}\phi}{\mathrm{d}t}$$

which has units of *radians per second*. We can convert the angular frequency form to a conventional frequency form by dividing through out by 2π to obtain the **Instantaneous frequency**:

$$f_i(t) = f_c + \frac{1}{2\pi}\frac{\mathrm{d}\phi}{\mathrm{d}t} \text{ Hz}$$

where f_c is nominal carrier frequency. The second term is referred to as the **Instantaneous frequency modulation** or *instantaneous frequency deviation*:

$$FM(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi}{\mathrm{d}t}$$

We can also make the observation that this $FM(t)$ is somehow describing by how much the signal frequency is changing, which because we are interested in frequency modulation is a function of the modulating signal $m(t)$. Furthermore we would like to make the frequency deviate by a certain amount which we refer to as the *Maximum frequency deviation* f_d. This means that we can make the observation that the Instantaneous frequency modulation can also be made to equal:

$$\therefore FM(t) = f_d \times m(t)$$

If we equate the two we get:

$$\frac{1}{2\pi}\frac{\mathrm{d}\phi}{\mathrm{d}t} = f_d \times m(t).$$

Multiplying both sides by 2π we get:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 2\pi f_d \times m(t).$$

Remember that we started out with the argument of the modulated carrier wave $2\pi f_c t + \phi(t)$. The only term that we are interested in was $\phi(t)$. We now have it defined but in terms of the derivative only. We can reverse the derivative operation by using the antiderivative, which can be found using integration. Thus we may define the **Phase Modulation** with:

$$\begin{aligned}\phi(t) &= 2\pi \int\limits_{-\infty}^{t} FM(\tau)\mathrm{d}\tau \\ &= 2\pi f_d \int\limits_{-\infty}^{t} m(\tau)\mathrm{d}\tau.\end{aligned}$$

Therefore for Frequency Modulation the modulated carrier wave signal can be described by:

$$\begin{aligned}x_m(t) &= E_c \times \cos(2\pi f_c t + \kappa m(t)) \\ &= E_c \times \cos\left(2\pi f_c t + 2\pi f_d \int\limits_{-\infty}^{t} m(\tau)\mathrm{d}\tau\right).\end{aligned}$$

This shows how the modulated carrier wave signal will change as the modulating signal $m(t)$ changes.

1.5 Analogue QAM

Analogue Qaudrature Amplitude Modulation (QAM) is another modulation technique that involves a carrier wave being modulated in terms of both the amplitude and the phase. Analogue QAM is important for many applications however the digital equivalent, also known as QAM is also widely used. The mathematics underlying QAM uses the *compound angle formula*:

$$\cos(a + b) = \cos(a)\cos(b) - \sin(a)\sin(b)$$

to obtain two components referred to as the I/Q components. Recall the general modulated wave formula:

$$x_m(t) = A(t)\cos\left(2\pi f_c t + \phi(t)\right).$$

We can use the compound angle formula to split this into two parts. Furthermore it also enables us to describe the phase $\phi(t)$ and the nominal carrier frequency in two separate terms as well. Thus, using the compound angular formula with $a = 2\pi f_c t$ and $b = \phi(t)$ we have:

$$\begin{aligned}x_m(t) &= A(t)\cos(2\pi f_c t + \phi(t)) \\ &= A(t)[\cos(2\pi f_c t)\cos(\phi(t)) - \sin(2\pi f_c t)\sin(\phi(t))] \\ &= A(t)\cos(\phi(t))\cos(2\pi f_c t) - A(t)\sin(\phi(t))\sin(2\pi f_c t)\end{aligned}$$

Furthermore if we now set $I(t) = A(t)\cos(\phi(t))$ and $Q(t) = A(t)\sin(\phi(t))$ we get:

$$x_m(t) = I(t)\cos(2\pi f_c t) - Q(t)\sin(2\pi f_c t).$$

This also has some interesting properties. In terms of Amplitude Modulation we can say that the amplitude is described by:

$$A(t) = \sqrt{I(t)^2 + Q(t)^2}.$$

And in terms of Phase Modulation the phase is described by:

$$\phi(t) = \tan^{-1}\left(\frac{Q(t)}{I(t)}\right).$$

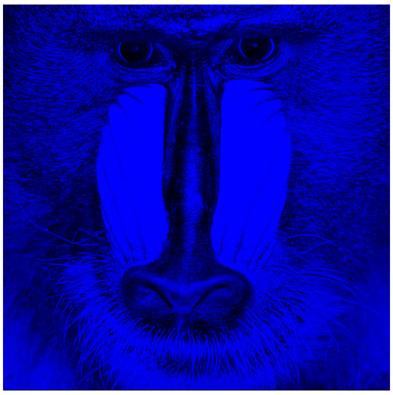
1.6 QAM: Subcarrier Chrominance Encoding

A popular example of the use of QAM in an analogue modulation scheme is for colour television. A colour picture is usually split into two chrominance parts U and V and the more important luminance or brightness information Y. However, often we may have more immediate access to an alternative representation such as where each pixel in the image is represented by an RGB colour system with three colour channels and no luminance information. For example,

Red

Green

Blue



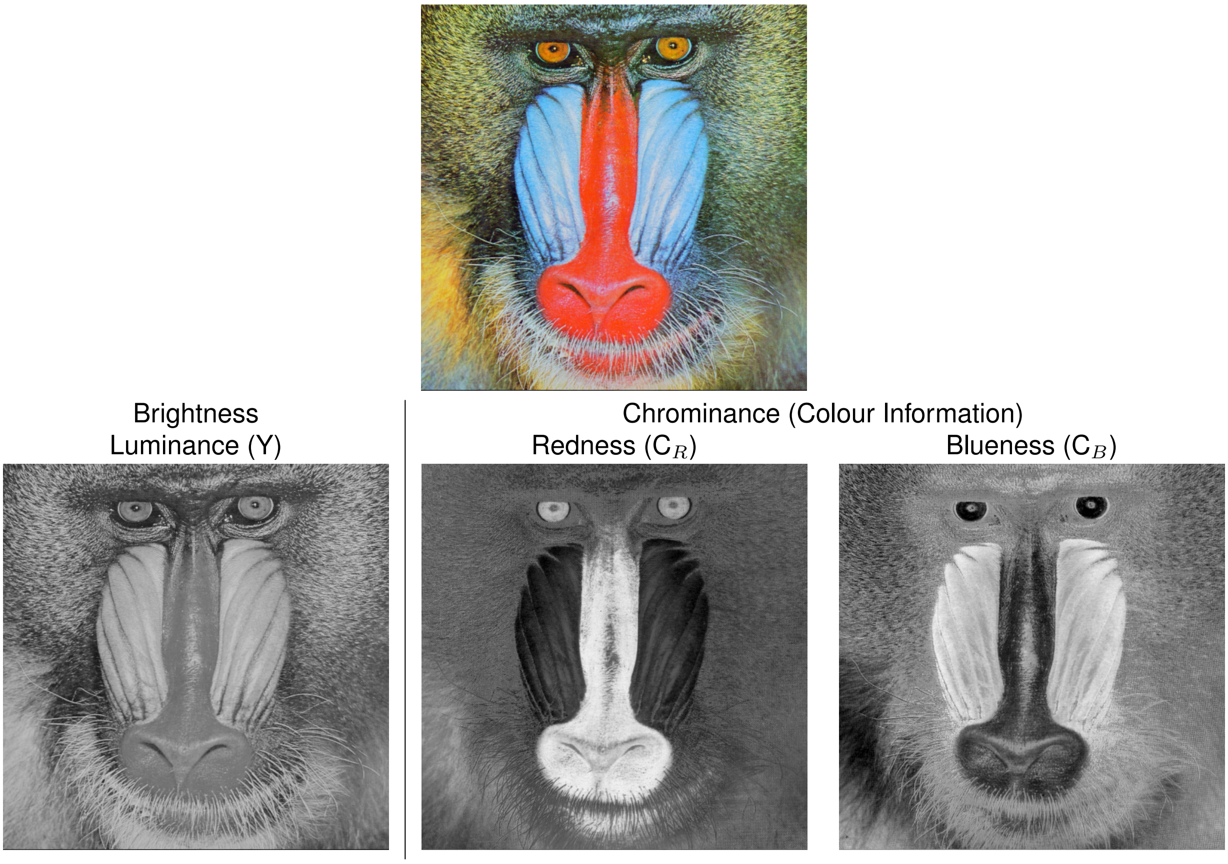
The U and V chrominance channels can be created from the R, G and B channels with:

$$V = R - Y$$

and

$$U = B - Y.$$

The luminance Y can be created with a weighted combination of all three colour components R, G and B. For the above example we might have:



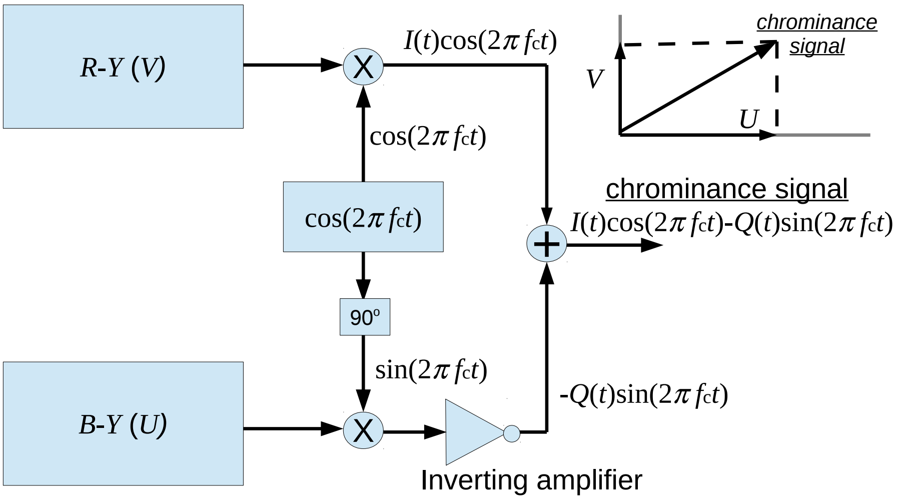
Brightness
Luminance (Y)

Chrominance (Colour Information)

Redness (C_R)

Blueness (C_B)

These chrominance channels can then be communicated simultaneously with a single modulated carrier wave using QAM, like so:



2 Digital Modulation

A famous digital modulation technique is **Morse Code**. Morse code entered by **keying** a mechanical push button.

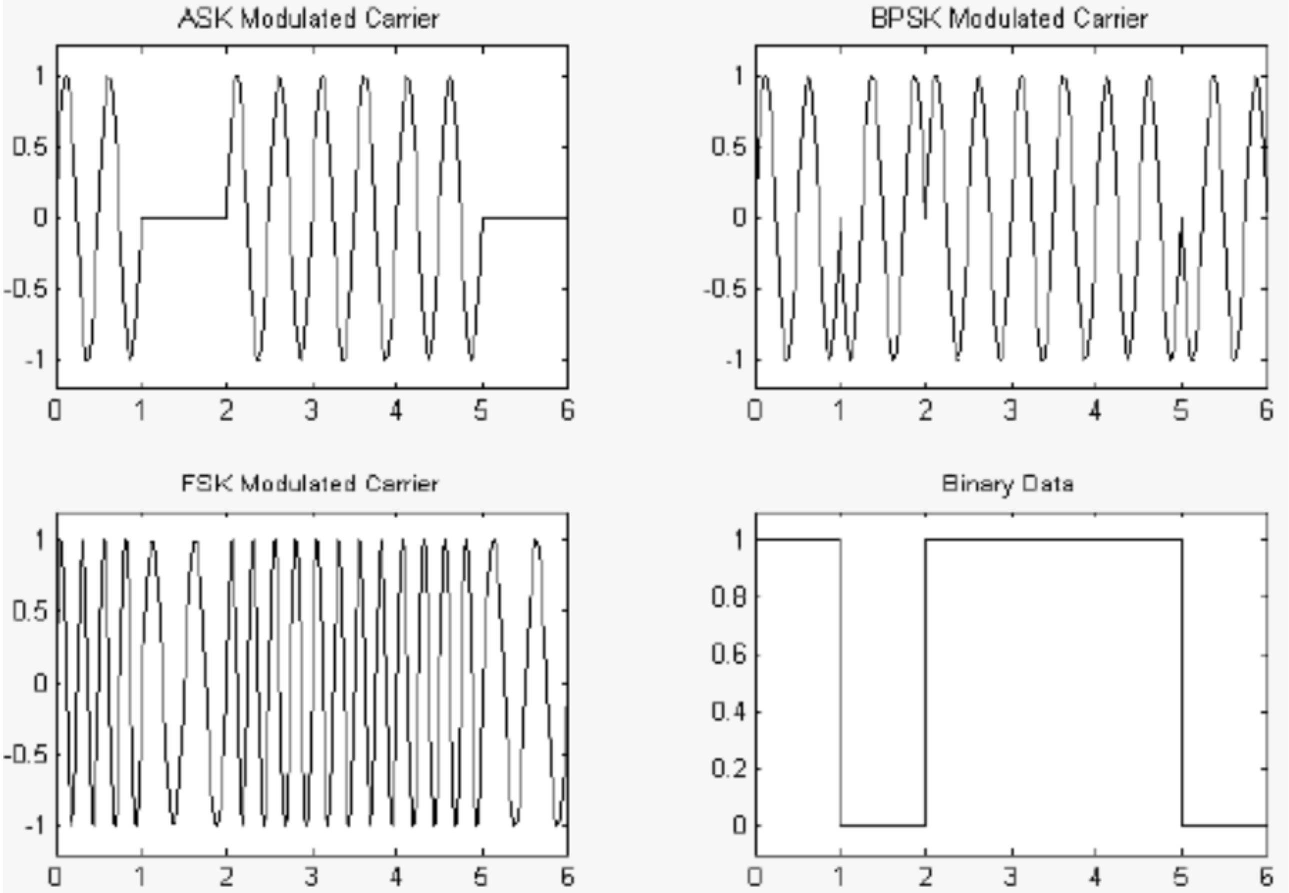
An over simplified idea of Digital modulation is that it consists of a clocked digital signal $d(t) \in \{0, 1\}$ and then the modulating signal is $m(t) = 2d(t) - 1$ which can be used to modulate a carrier wave. This idea is an oversimplification because the original signal may not be binary however it should be digital.

Digital modulation techniques include:

- Amplitude-Shift Keying (ASK) which involves varying the amplitude of a carrier wave. A special case is on-off keying or (OOK) which involves switching on the carrier wave for a binary digit 1 and completely switching it off for a binary digit 0.
- Frequency-Shift Keying (FSK) which involves varying the frequency of the transmitted signal.

- Phase-Shift Keying (PSK) which includes Binary Phase Shift Keying (BPSK), the binary case is *similar to DSBSC or digital phase modulation.* As the name suggests, this involves switching between different phase values.
- (digital) Quadrature Amplitude Modulation (QAM). This involves varying both the amplitude and phase of the carrier wave.

A few examples of the various digital modulation schemes can be seen below:



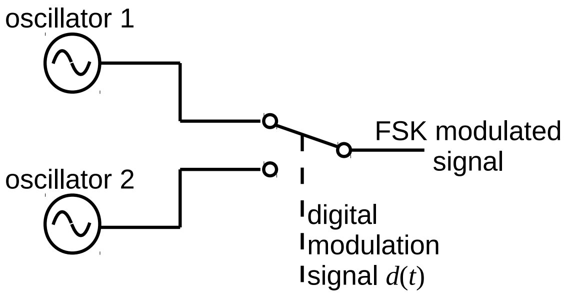
Digital Modulation Overview Here we can see an overview of some mathematical aspects of the digital modulation schemes. We will investigate each in more detail below.

Modulation	Amplitude & Phase	$I(t)$	$Q(t)$
ASK	$A(t) = E_c d(t)$ $\phi(t) = 0$	$E_c d(t)$	0
BPSK	$A(t) = E_c \zeta$ $\phi(t) = \pi(1 - d(t))$	$E_c \cos(\pi(1 - d(t)))$ $= E_c(2d(t) - 1)$ because $d(t) \in \{0, 1\}$	0 0
FSK	$A(t) = E_c \zeta$ $\phi(t) = 2\pi f_d \int\limits_{-\infty}^{t} (2d(\tau) - 1)\mathrm{d}\tau$ because $\frac{1}{2\pi}\frac{\mathrm{d}\phi}{\mathrm{d}t} = f_d m(t) = f_d(2d(t) - 1)$	too hard	too hard

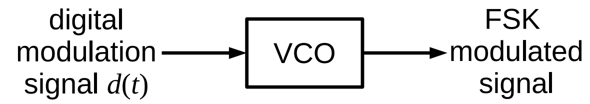
where ζ is some constant.

2.1 Frequency Shift Keying (FSK)

FSK varies an analogue carrier frequency with the digital modulating signal. How to vary frequency? Suddenly switching between frequencies can introduce relatively large sudden changes in the signal. These sudden changes mean very wide range (wideband) of undesirable frequencies are present. It is better to smoothly change between frequencies using a Voltage Controlled Oscillator (VCO) as communication channels are band limited. Furthermore large sudden discontinuities can be more difficult to linearly amplify.

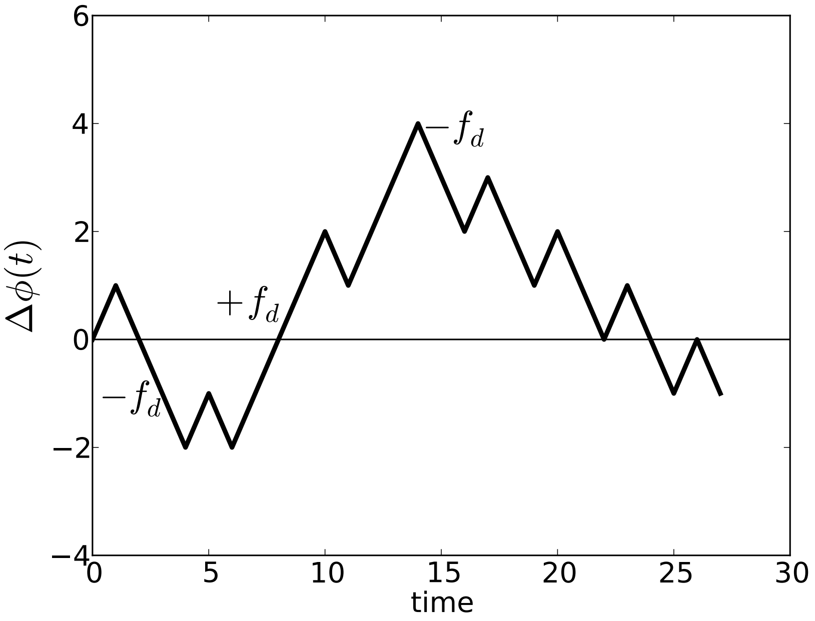


BAD
Discontinuities - high bandwidth
Only useful for low bite rate applications.
e.g. storing data on audio cassette tapes



GOOD
Continuous, smooth changes between frequencies
Also known as Continuous Phase FSK (CPFSK).

Using a VCO to switch between frequencies means that the resulting FSK scheme can be referred to as Continuous Phase FSK (CPFSK). Continuous Phase FSK can be illustrated by means of an example. For the following data sequence=100010111101110010010010010 we have the following plot of the change in phase with time $\Delta\phi(t)$:



Here the *Carrier frequency, f_c is not sent*. Instead the frequency of the modulated carrier wave signal $f_i(t)$ depends on the combination of the maximum frequency deviation f_d and the carrier frequency f_c:

$$f_i(t) = f_c + f_d m(t)$$

Also we know that the modulation signal converts a binary stream $d(t) = 0$ or $1$into positive and negative values, *i.e.* $m(t) = -1$ or $+1$ therefore:

$$f_i(t) = f_c + f_d(2d(t) - 1).$$

Recall how, for the analogue Frequency Modulation, the modulation was formulated in terms of the phase modulation $\phi(t)$ term. We found that the change in phase with respect to time is given by:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 2\pi f_d \times m(t) = 2\pi f_d(2d(t) - 1).$$

Using this, we can see that if $d(t) = 1$ then $\frac{\mathrm{d}\phi}{\mathrm{d}t} > 0$ and $f_i = f_c + f_d$ or if $d(t) = 1$ then $\frac{\mathrm{d}\phi}{\mathrm{d}t} < 0$ $f_i = f_c - f_d$. The change in phase is shown in the example above except there we have $\Delta\phi(t)$. For a discrete system, instead of a continuous derivative we have

$$\frac{\Delta\phi(t)}{\Delta t} = \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}.$$

Also, the smallest size of time that can be measured in a discrete system is a bit width T_b so that

$$\frac{\Delta\phi(t)}{T_b} = \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}.$$

Quantifying the change in phase

It is also of interest to be able to state the amount that the phase changes for each bit being communicated, *i.e.* how to describe $\Delta\phi$. Previously we found that the Instantaneous Frequency Modulation:

$$FM(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi}{\mathrm{d}t}$$

Substituting in the discrete approximation for the derivative of $\phi(t)$ we have:

$$FM(t) = \frac{1}{2\pi}\frac{\Delta\phi(t)}{T_b}.$$

We also know that the instantaneous frequency is given by $f_i(t) = f_c + FM(t)$ so that and for communicating a logic $d(t) = 1$ we have $f_i = f_c + f_d$. We can therefore say that f_d is equal to the instantaneous frequency modulation at a particular time instance resulting in

$$f_d = \frac{1}{2\pi}\frac{\Delta\phi(t)}{T_b}.$$

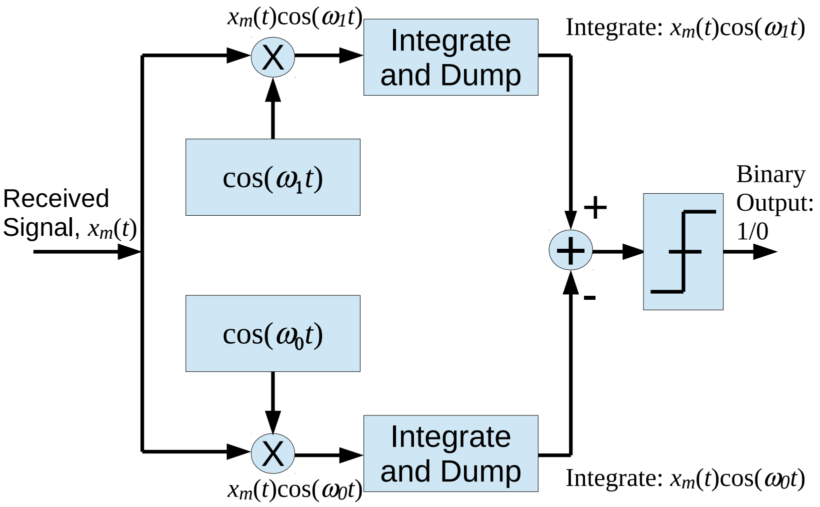
Rearranging, we can find an expression for $\Delta\phi(t)$:

$$\Delta\phi = 2\pi f_d T_b.$$

Spacing of transmitted frequencies Δf

The spacing between the frequencies is given by $\Delta f = 2f_d$.

It is beneficial for receiver design if $\Delta f \times T_b$ is an integer. This acts like a modulation index. Some special conditions arise for the following receiver circuit if this is the case.



Integrate and Dump is a standard tool to help recover a pulse that is often communicated in digital communication systems. It effectively averages signal over one symbol duration. It is used here similar to a low pass filter that would be used in a conventional analogue coherent detector circuit.

To see how this works, let us consider the two cases for when the modulated carrier wave $x_m(t)$ is carrying a pulse to represent the data $d(t) = 1$ which we will refer to as $x_m(t) = \cos(\omega_1 t)$. The other case, *i.e.* when $d(t) = 0$ we will let $x_m(t) = \cos(\omega_0 t)$.

If $d(t) = 1$ then after the upper product operation we have

$$x_{\text{upper}}(t) = x_m(t)\cos(\omega_1 t) = \cos(\omega_1 t)\cos(\omega_1 t).$$

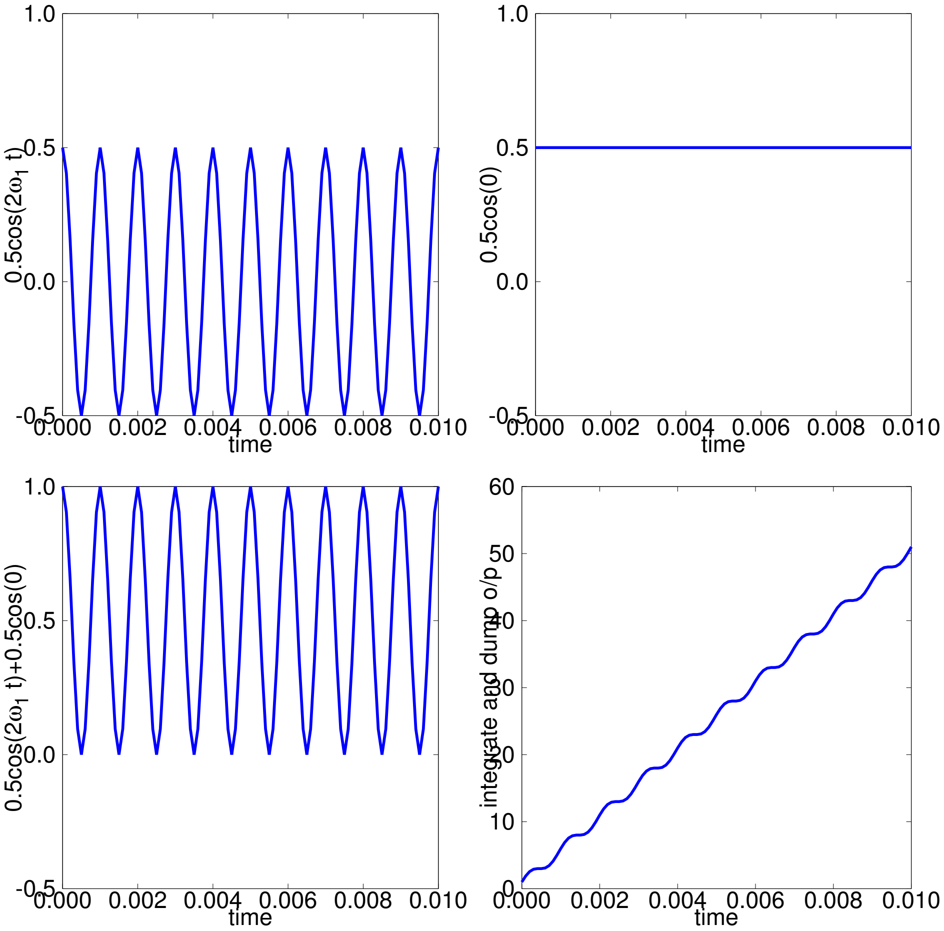
We can use the trigonometric identity

$$\cos(a)\cos(b) = \frac{1}{2}\cos(a+b) + \frac{1}{2}\cos(a-b)$$

to obtain

$$x_{\text{upper}}(t) = \frac{1}{2}\cos(2\omega_1 t) + \frac{1}{2}\cos(0) = \frac{1}{2}\cos(2\omega_1 t) + \frac{1}{2}.$$

In a coherent analogue detector this upper sum and difference result would normally be low pass filtered to remove the higher frequency $\cos(2\omega_1 t)$. However for digital modulation we can use the integrate and dump which will produce a positive voltage or signal, where the constant $\frac{1}{2}$ is combined with the higher frequency sinusoid that sits on top. This additional DC component means that the output of the integrate and dump operation is positive. This can be seen below:



Meanwhile, the output of the lower product operation is

$$x_{\text{lower}}(t) = x_m(t)\cos(\omega_0 t) = \cos(\omega_1 t)\cos(\omega_0 t).$$

Using the trigonometric identity we have

$$x_{\text{lower}}(t) = \frac{1}{2}\cos((\omega_1+\omega_0)t) + \frac{1}{2}\cos((\omega_1-\omega_0)t).$$

The first cosine term $\frac{1}{2}\cos((\omega_1+\omega_0)t)$ is high frequency and is averaged out after the integrate and dump. The second cosine term $\cos((\omega_1-\omega_0)t)$ after the integrate and dump equals zero if $\Delta f \times T_b$ is an integer or $\Delta f = \frac{\lambda}{T_b}$ where λ is an integer to enable the integrate and dump to have an integration time that is an integer multiple of the bit time. This can be understood if we manipulate the second term to show that

$$\cos((\omega_1-\omega_0)t) = \cos(2\pi(f_1-f_0)t) = \cos(4\pi f_d t) = \cos(2\pi \times 2f_d \times t) = \cos(2\pi\Delta f t).$$

And if the integration period contains a whole number of cycles of this frequency Δf then the output of the integrate and dump is zero for this part. Thus

$$\left(\underbrace{\frac{1}{2}\int_0^{\frac{N}{\Delta f}} \cos((\omega_1+\omega_0)t)\mathrm{d}t}_{\approx 0} + \underbrace{\frac{1}{2}\int_0^{\frac{N}{\Delta f}} \cos(2\pi\Delta f t)\mathrm{d}t}_{=0} \right) \approx 0.$$

The threshold operation that follows can then be used to set any positive value to 1 and any negative value to 0. The negative values occurs when the modulated carrier waveform is carrying a pulse that corresponds to a data signal value of $d(t)=0$.

2.2 Phase Shift Keying (PSK)

Information in phase, not amplitude:

$$x_m(t) = A\cos(2\pi f_c + \phi(t)).$$

Most basic PSK, **Binary PSK** (BPSK):

- Phase of 0 radians = binary 1
- Phase of π radians = binary 0

Can be encoded by:

$$\phi(t) = \begin{cases} 0 & \text{if} \quad d(t)=0 \quad \Rightarrow \quad x_m(t) = A\cos(2\pi f_c) \\ \pi & \text{if} \quad d(t)=1 \quad \Rightarrow \quad x_m(t) = A\cos(2\pi f_c+\pi) = -A\cos(2\pi f_c). \end{cases}$$

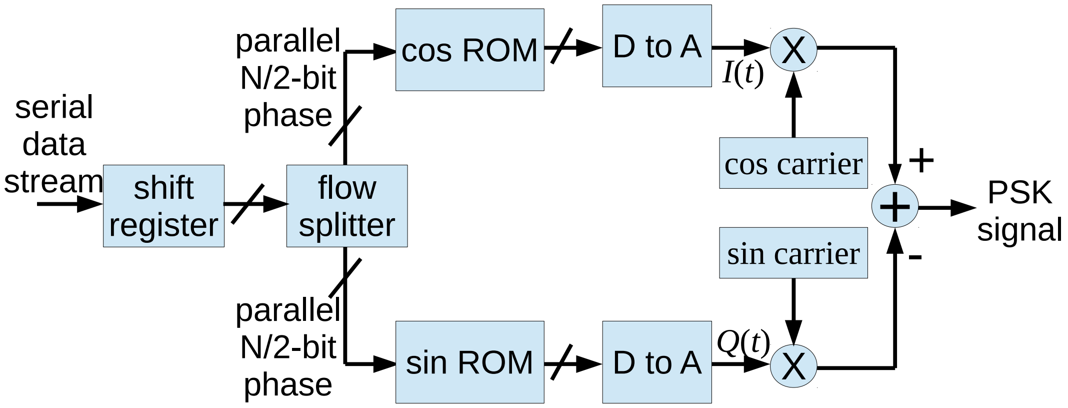
Identical to digital DSBSC.

Quadrature implementation of PSK

Can also implement using I and Q components from Quadrature formulation. If amplitude $A=1$ then:

$$I(t) = \cos(\phi(t)) \quad \text{and} \quad Q(t) = \sin(\phi(t))$$

Modulator structure is then:



The reverse of this type of structure can be used to demodulate the signal as well. Similar techniques are used for the popular “software defined radio” technology although digital signal processing might be used to extract the original transmitted signal $d(t)$.

Advantages and disadvantages of PSK

PSK is has a number of advantages and disadvantages. Some advantages include:

- Easy to implement.
- Useful for communication channels that suffer *fading* caused by *e.g.* time varying signal amplitudes (e.g. mobile radio).
- Also useful if high transmitter powers required.

A disadvantage of PSK is:

- Receivers have no phase reference...
    - ◇ "0" and "1" can become wrong way round called *Polarity Inversion*

What to do?

- Use Differential PSK (DPSK)!

2.3 Differential PSK (DPSK)

Overcomes problem of Polarity Inversion in BPSK. It provides a way for the receiver to determine when a logic 1 should occur by alternating between phases. For logic 0 the phase remains unchanged. In summary:

- To communicate Bit 0
    - ◇ **same-phase-relative-to-previous-bit**
- To communicate Bit 1
    - ◇ **opposite-phase-relative-to-previous-bit**

DPSK Example

Here a logic 1 is communicated by changing the phase by $\Delta\phi = \pi$ radians. If a logic 0 is to be communicated then the phase does not change, *i.e.* $\Delta\phi = 0$.

time =	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Data =	1	1	0	1	1	0	0	0	0	0	0	0	1	0	1
$\Delta\phi =$	π	π	0	π	π	0	0	0	0	0	0	0	π	0	π

However this type of modulation scheme can have more errors upon demodulation.

3 M-ary Communication

3.1 Digital QAM

Digital Quadrature Amplitude Modulation (QAM) relies on the same mathematical result as analogue QAM, *i.e.* (repeated here for convenience)

$$\begin{aligned} x_m(t) &= A(t)\cos(2\pi f_c t + \phi(t)) \\ &= \underbrace{A(t)\cos(\phi(t))}_{I(t)}\cos(2\pi f_c t) - \underbrace{A(t)\sin(\phi(t))}_{Q(t)}\sin(2\pi f_c t) \end{aligned}$$

So that we can express the generalized modulated wave formula in terms of two quadrature components $I(t)$ and $Q(t)$:

$$x_m(t) = I(t)\cos(2\pi f_c t) - Q(t)\sin(2\pi f_c t).$$

These quadrature components can simultaneously carry 2 analogue signals at the same instance in time. In digital modulation QAM becomes even more powerful because the amplitude and phase can be varied to enable multiple bits to be simultaneously communicated. For example, we can simultaneously transmit 2 bits of information by varying the amplitude and phase in 4 different ways (known as 4-QAM), leading to the following possible states:

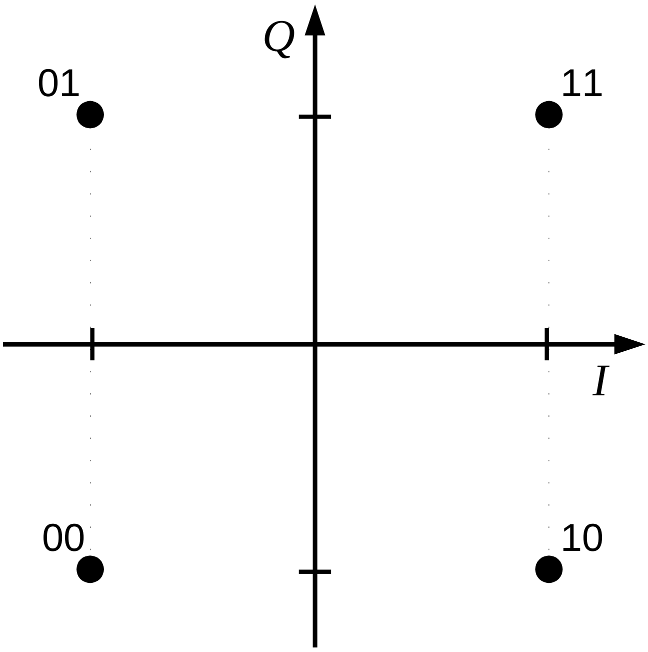
4QAM State Table

Data	I	Q
11	+1d	+1d
10	+1d	-1d
01	-1d	+1d
00	-1d	-1d

where d is some relative distance.

Furthermore these states can be drawn in a diagram known as a **constellation diagram**:

Constellation Diagram (4QAM)



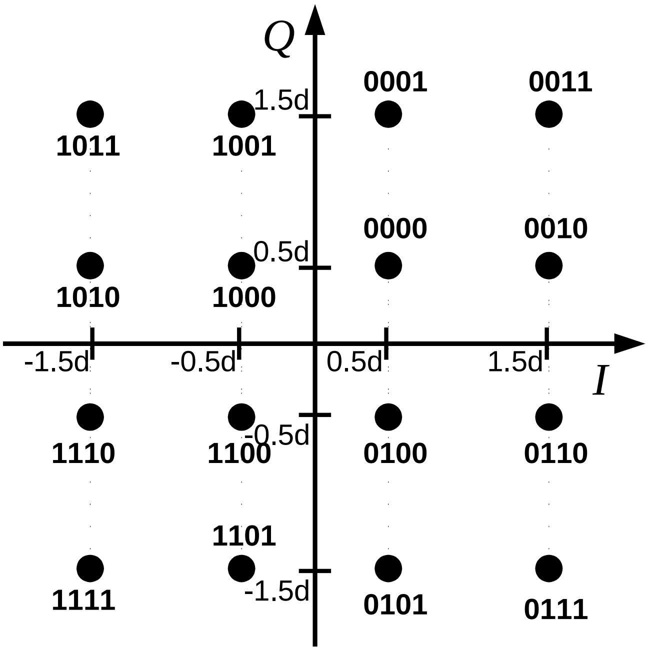
This constellation diagram helps to illustrate what is being varied for the different states. To simultaneously communicate the bits 11 we need to set $I = +1$ and $Q = +1$. To simultaneously communicate the bits 01 we need to set $I = -1$ and $Q = +1$, etc. This principle can be extended to any number of bits however there are limitations as we will shortly see.

3.2 16QAM

The ordering of the states is important to help reduce the effect of noise on a communicated signal. For 16QAM we may have the following state table:

Data	I	Q
0000	0.5d	0.5d
0001	0.5d	1.5d
0010	1.5d	0.5d
0011	1.5d	1.5d
0100	0.5d	-0.5d
0101	0.5d	-1.5d
0110	1.5d	-0.5d
0111	1.5d	-1.5d
1000	-0.5d	0.5d
1001	-0.5d	1.5d
1010	-1.5d	0.5d
1011	-1.5d	1.5d
1100	-0.5d	-0.5d
1101	-0.5d	-1.5d
1110	-1.5d	-0.5d
1111	-1.5d	-1.5d

These values are assigned to the various bit patterns so that all neighbouring states differ by at most 1 bit. This means that if a signal is affected by some noise that makes it have a modified phase or amplitude then the received signal will more likely be different by at most 1 bit. The importance of the ordering is more easily observed in terms of the constellation diagram:



Here, as you can see only 1 bit differs for any immediately neighbouring state.

Digital QAM Modulation

Digital QAM modulation requires a similar modulation scheme as outlined for PSK earlier except now the amplitude is also modulated with the relative distance d multiplied by the relevant state information. For 16 QAM we may have:

- Two bits (4 levels) to modulate I
  - ◇ Equivalent to 2bit D/A converter
- Two bits (4 levels) to modulate Q
  - ◇ Equivalent to 2bit D/A converter

3.3 M-ary Communication

M-ary communication is the simultaneous transmission of N bits of information where N is calculated according to

$$N = \log_2(M).$$

The M represents the number of states needed to communicate N bits of information and can be calculated with

$$M = 2^N.$$

For example for 4QAM then 2 bits of information can be simultaneously communicated. For 16QAM then 4 bits of information can be communicated simultaneously.

If M increases too much then more power is needed to transmit the signal to ensure that the distance between the states is as large as possible. However there are limitations on the amount of power that can be used to transmit a signal (see later). Therefore it is important to understand the power in a digitally modulated signal. The Root Mean Square (RMS) amplitude of a modulated signal is a useful term in power calculations.

3.4 Root Mean Square (RMS)

The RMS amplitude of a signal can be found with the following equation:

$$A_{\text{rms}} = \sqrt{\frac{1}{M} \sum_{i \text{ in all states}} A_i^2}$$

where A_i is the amplitude of state i. State i is one of the M states that we have already looked at, so for 16QAM there are 16 states. The process of calculating the RMS amplitude of a carrier can be found by:

- Finding A^2 for each state

  $$A_i^2 \quad \text{for all states, } i.$$

- Finding the Average A^2 for all states (mean)

  $$\frac{1}{M} \sum_{\text{for all states, } i} A_i^2$$

  where M is the number of states

- then taking the square root

  $$\sqrt{\frac{1}{M} \sum_{\text{for all states, } i} A_i^2}.$$

16QAM RMS Example

The RMS amplitude can be calculated for 16QAM using the above steps in combination with the state information. Firstly we know that there are 16 states, $M = 16$. Next we can calculate the amplitudes with the amplitude equation for QAM:

$$A = \sqrt{I^2 + Q^2}$$

Data	I	Q	A_i
0000	0.5d	0.5d	$d\sqrt{\frac{1}{2}}$
0001	0.5d	1.5d	$d\sqrt{2\frac{1}{2}}$
0010	1.5d	0.5d	$d\sqrt{2\frac{1}{2}}$
0011	1.5d	1.5d	$d\sqrt{4\frac{1}{2}}$
0100	0.5d	-0.5d	$d\sqrt{\frac{1}{2}}$
0101	0.5d	-1.5d	$d\sqrt{2\frac{1}{2}}$
0110	1.5d	-0.5d	$d\sqrt{2\frac{1}{2}}$
0111	1.5d	-1.5d	$d\sqrt{4\frac{1}{2}}$
1000	-0.5d	0.5d	$d\sqrt{\frac{1}{2}}$
1001	-0.5d	1.5d	$d\sqrt{2\frac{1}{2}}$
1010	-1.5d	0.5d	$d\sqrt{2\frac{1}{2}}$
1011	-1.5d	1.5d	$d\sqrt{4\frac{1}{2}}$
1100	-0.5d	-0.5d	$d\sqrt{\frac{1}{2}}$
1101	-0.5d	-1.5d	$d\sqrt{2\frac{1}{2}}$
1110	-1.5d	-0.5d	$d\sqrt{2\frac{1}{2}}$
1111	-1.5d	-1.5d	$d\sqrt{4\frac{1}{2}}$

Taking the mean of the square of A_i we find

$$\frac{1}{M} \sum_{\text{for all states, } i} A_i^2 = d^2 \frac{5}{2}.$$

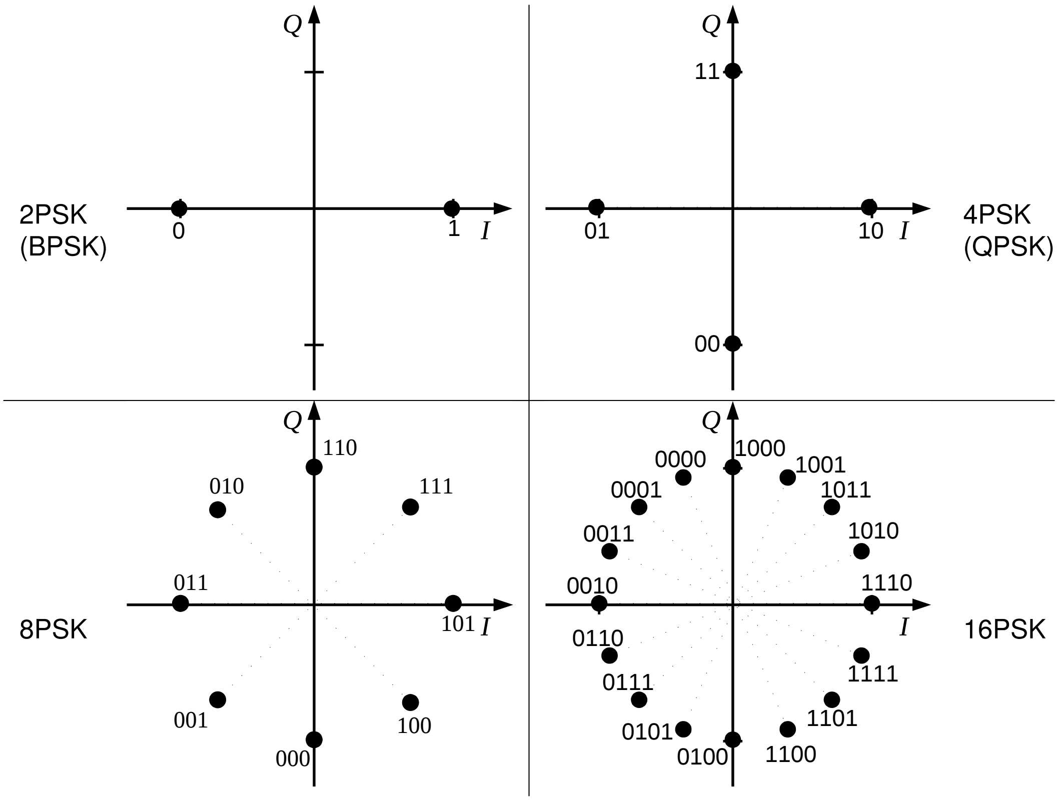
Therefore 16QAM has RMS amplitude:

$$A_{\text{rms}} = d\sqrt{\frac{5}{2}}.$$

3.5 M-ary PSK Schemes

Quadrature Amplitude Modulation is not the only technique that is able to transmit more than a single bit simultaneously. Phase Shift Keying (PSK) can utilise a range of different phase angles to simultaneously transmit multiple bits of information. Again, states and constellation diagrams are useful concepts when understanding the M-ary PSK. Recall to convey N bits, then $M = 2^N$.

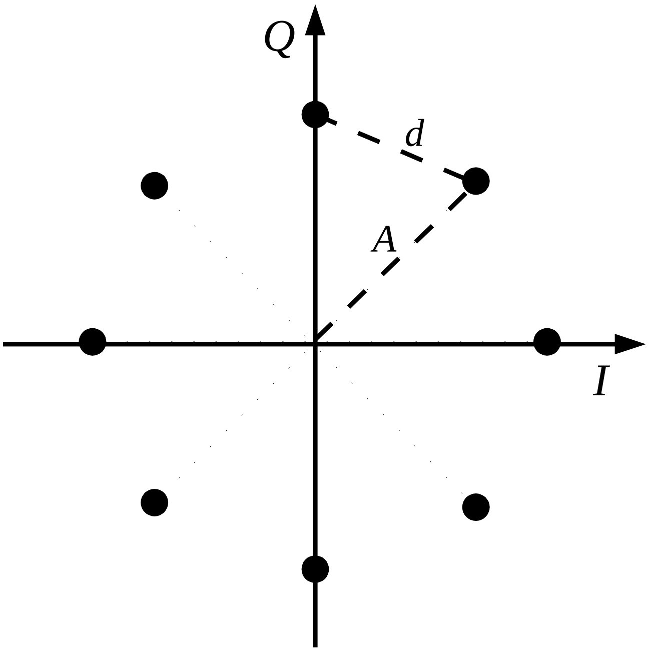
Here are some examples of different M PSK modulation:



The important thing to realise here is that PSK is very similar to QAM except the amplitude for PSK does not vary between states. Having said that, the amplitude can still be varied which can be used to control the distance between the states:

- A = amplitude or signal strength
- d = distance between states (state spacing)

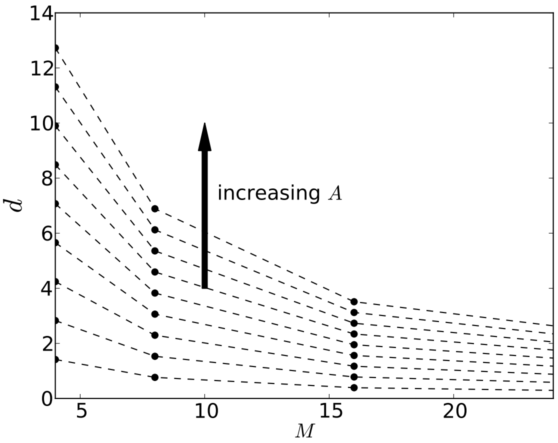
The amplitude and distance between states are illustrated below:



Distance between state points can be quantified using the following formula:

$$d = 2A \sin\left(\frac{\pi}{M}\right).$$

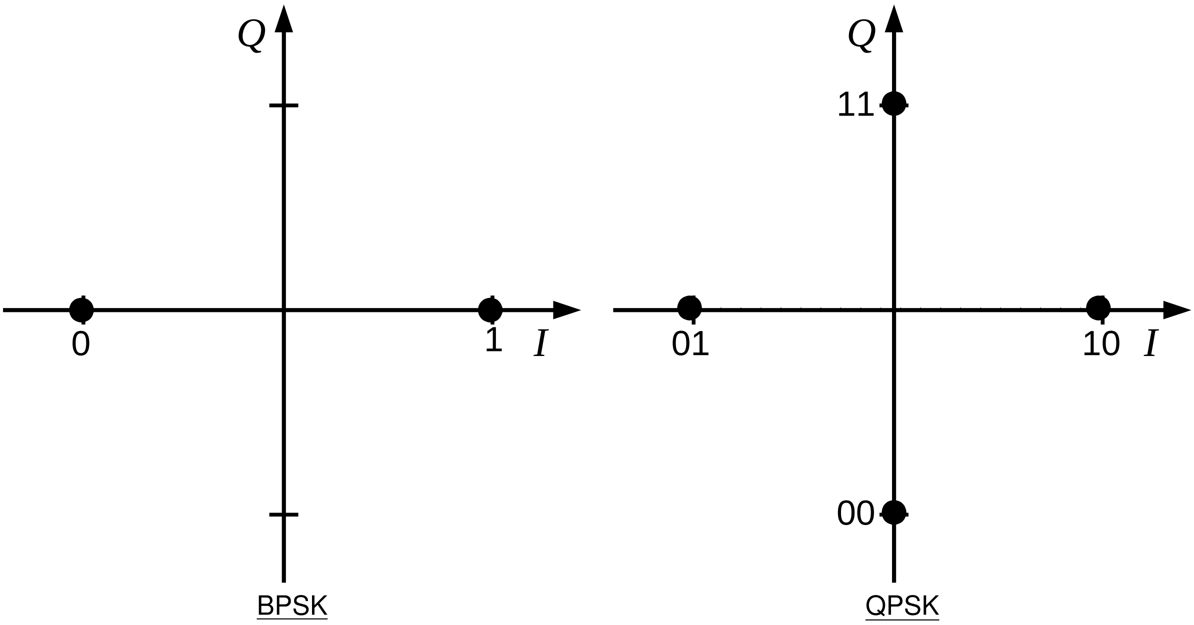
The distance between the states is important because it is a measure of the ability of the receiver to distinguish between states and therefore to be able to determine the original signal that was transmitted correctly, *i.e.* without error. As M increases it becomes more difficult for a receiver to distinguish (decide) between states which is illustrated here:



This diagram also illustrates that if M is large then d can be kept relatively constant if A is large. So in summary for greater A (signal strength) it can help to separate state points, however this act is not so useful for $M \geq 16$.

BPSK and QPSK

For the cases of $M = 2$ or Binary PSK (BPSK) and $M = 4$ for Quadrature PSK (QPSK) both these have the same probability of error on the received bit data.



This will be seen later on.

4 Comparison of Digital Modulation Schemes

We have seen that there are many different Digital Modulation techniques. Some techniques are better than others depending on the application and depending on the parameters involved. There are various advantages and disadvantages for each technique.

An obvious differentiating factor is the amount of information that can be sent simultaneously. This can be expressed in terms of a combination of bit rate and baud rate which we will now define.

- Bit rate is the number of bits being sent a second;
- Baud rate is the number of symbols being sent a second;

A symbol can represent a combination of bits being sent simultaneously and corresponds to the state information that we have seen in the M-ary schemes. Bit rate can therefore be higher than Baud rate if $M > 2$. However if we just consider schemes where only a single bit of information is sent at a time such as BPSK then the Baud rate will be equal to the bit rate.

We can also make a statement about the channel bandwidth. In general:

$$\begin{array}{c} \text{channel bandwidth} \\ \text{to transmit} \\ \text{digital signal} \end{array} \propto \begin{array}{c} \text{number of symbols} \\ \text{sent per} \\ \text{second (Baud rate)} \end{array}$$

This means that M-ary schemes need less channel bandwidth than binary schemes for same bit rate.

A simple way of comparing different schemes is in terms of relative bandwidth, B_{rel}

$$\begin{aligned} B_{\text{rel}} &= \frac{\text{Baud Rate}}{\text{Bit Rate}} = \frac{\text{symbols/sec}}{\text{bits/sec}} \\ &= \frac{1}{\text{bits/symbol}} = \frac{1}{N} \\ &= \frac{1}{\log_2(M)}. \end{aligned}$$

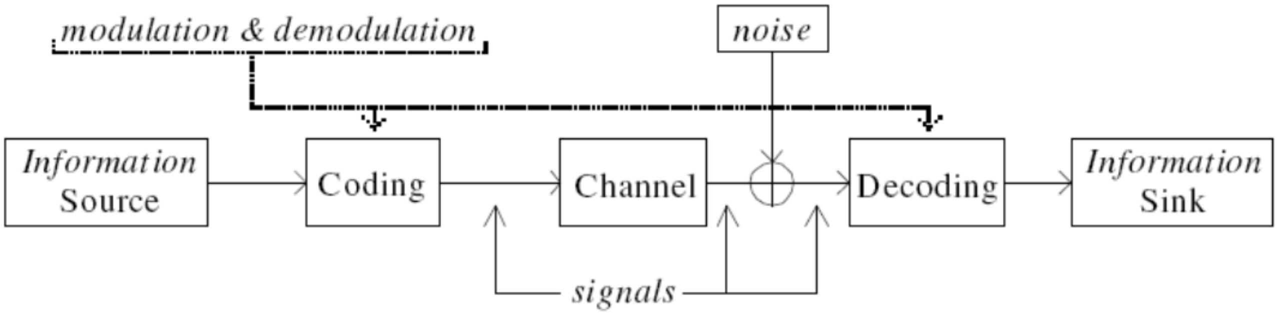
Thus, comparing various modulation schemes ability to transmit information simultaneously we have the relative bandwidth:

Scheme	M	B_{rel}
BPSK	2	1
QPSK	4	$\frac{1}{2}$
8-PSK	8	$\frac{1}{3}$
16-PSK	16	$\frac{1}{4}$
16-QAM	16	$\frac{1}{4}$

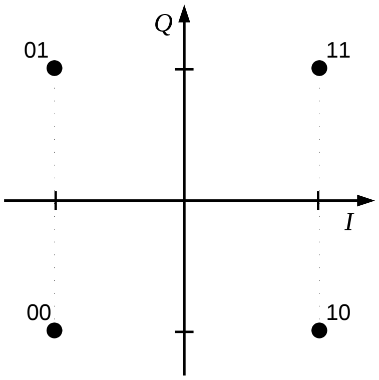
Low B_{rel} is (usually) Good - but because, as we saw earlier, this means that d becomes smaller for lower B_{rel} (because M is low) and can therefore become too sensitive to noise.

4.1 Noise Sensitivity

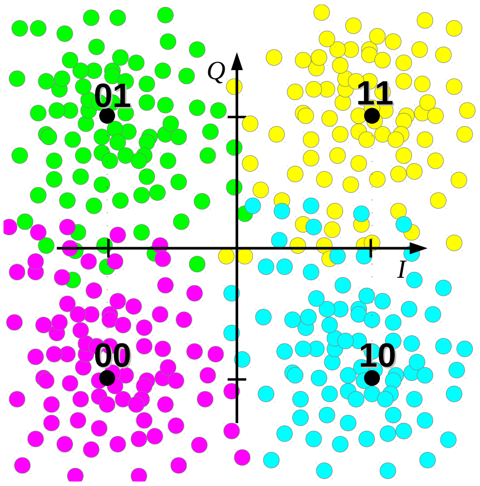
All communication systems can potentially be affected by noise. Noise is often introduced as part of the communication process and it may become more of a problem after a signal has become significantly attenuated.



In any case, we can visualise the affect of noise on a received signal by looking at the constellation diagram. The received bit information will differ from ideal state position. So for example, for 4-QAM we may have:



Transmitted or Theoretical 4QAM



Received / Noisy 4QAM

Here you can see that the received signal has points scattered over a large area of the constellation diagram. If a point lands in an incorrect quadrant then it is likely that the receiver / detector will not be able to correctly identify the bit combination corresponding to that received symbol. It is therefore useful to be able to quantify the affect of this noise on the possibility of a received signal being detected with error, *e.g.* a 10 bit sequence interpreted which should instead have been interpreted as 11.

All communication systems are potentially affected by varying levels of noise. There is a general need to minimise the number of errors which can be described in terms of the probability of bit errors or Bit-Error-Ratio (BER). As already hinted at, the likelihood of errors can be reduced if we move states as far apart as possible which will have the affect of increasing d. We can try to increase the transmitter power thus increasing A which will move the states apart.

However increasing transmitter power increases costs and there is a potential problem from running the transmitter amplifiers at high power known because of high power amplifier non-linearities which we will look at shortly.

Higher order systems are more sensitive to noise (for fixed transmitter power) because

- states are closer together.
- If A_{rms} is constant then d must get smaller.
- Leads to more sensitivity to noise...

We may consider reducing B_{rel} using lower order modulation schemes. Decreasing channel bandwidth reduces noise because noise power is proportional to bandwidth B:

$$\text{noise power} \propto B.$$

$$\text{noise rms} \propto \sqrt{B}$$

We also have relative bandwidth to express the digital communication of a number of bits simultaneously in the same amount of bandwidth, where $B \propto B_{\text{rel}}$ therefore we can say that

$$\text{noise rms} \propto \sqrt{B_{\text{rel}}}.$$

4.2 Describing Relative Signal Power

To compare modulation schemes we can assume the transmitter power for a particular modulation scheme is selected so that d is a fixed multiple of the noise RMS. This does not guarantee exactly the same BER but it does lead to an expression that can be used for the approximate equivalence in terms of required signal power:

$$P \propto B_{\text{rel}} \times \left(\frac{A}{d}\right)^2.$$

- B_{rel} depends on number of states, M
- $\left(\frac{A}{d}\right)^2$ depends on type of keying, *and* the number of states.

Transmitter power P required to obtain same level of Symbol Errors (not BER) using $P \propto B_{\mathrm{rel}} \times \left(\frac{A}{d}\right)^2$ can then be calculated for a number of different modulation schemes:

Scheme	B_{rel}	$(A/d)^2$	$(A/d)^2 B_{\mathrm{rel}}$	dB relative to BPSK
BPSK	1	0.25	0.250	0dB
4PSK	1/2	0.5	0.250	0dB
8PSK	1/3	1.7071	0.569	+3.57dB
16PSK	1/4	6.5685	1.642	+8.17dB
16QAM	1/4	2.5	0.625	+3.98dB

Two main observations can be made:

- Higher order schemes increase transmitter power need
- QAM needs less power than PSK for same channel bandwidth

However QAM is also more susceptible to non-linearities in the amplification stage.

5 High Power Amplifier Nonlinearities

All amplifiers are slightly non-linear and in particular have non-linear regions at high levels of amplification. Very high power devices can be very nonlinear e.g. satellite broadcasting transmitters using Travelling Wave Tubes (TWTs) or Solid State Power Amplifiers.

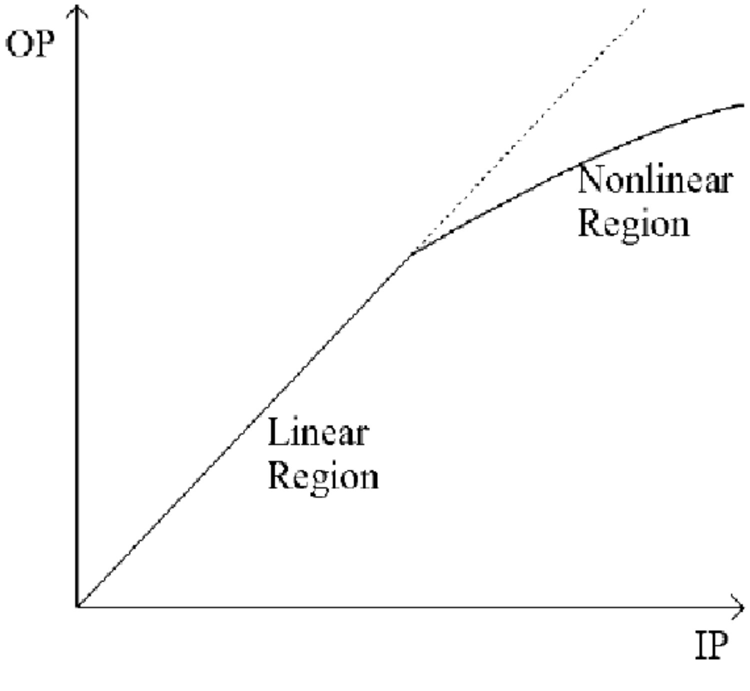
Some nonlinear effects include:

- Generation of harmonics;
- Distortion components with same frequency band as input signal which cannot be removed by filtering.

Two types of non-linear effects are: the amplitude is not linearly reproduced with a constant multiple (AM/AM conversion distortion) and the phase may also be non-linear (AM/PM conversion distortion). In other words:

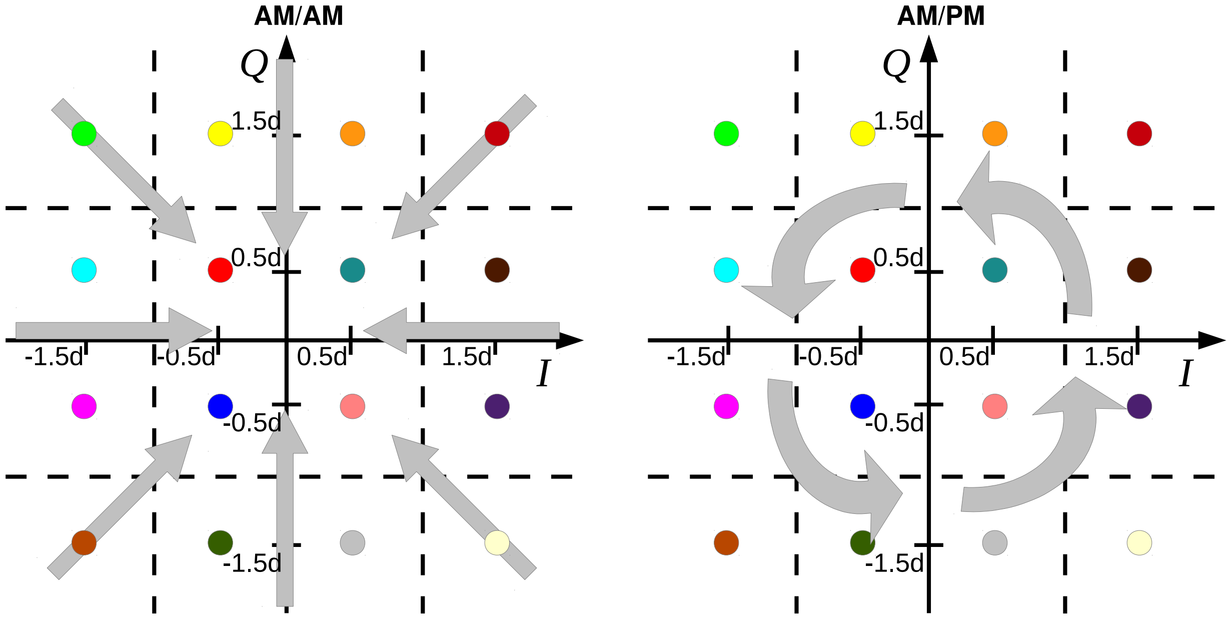
- Instantaneous amplitude amplifier output not a constant multiple of input amplitude
  - ◇ **AM/AM conversion distortion**.
- Amplifier phase shift not constant
  - ◇ **AM/PM conversion distortion**

AM/AM conversion distortion can be easily seen if the input signal is plotted as a function of the output signal:



5.1 AM/AM versus AM/PM on QAM

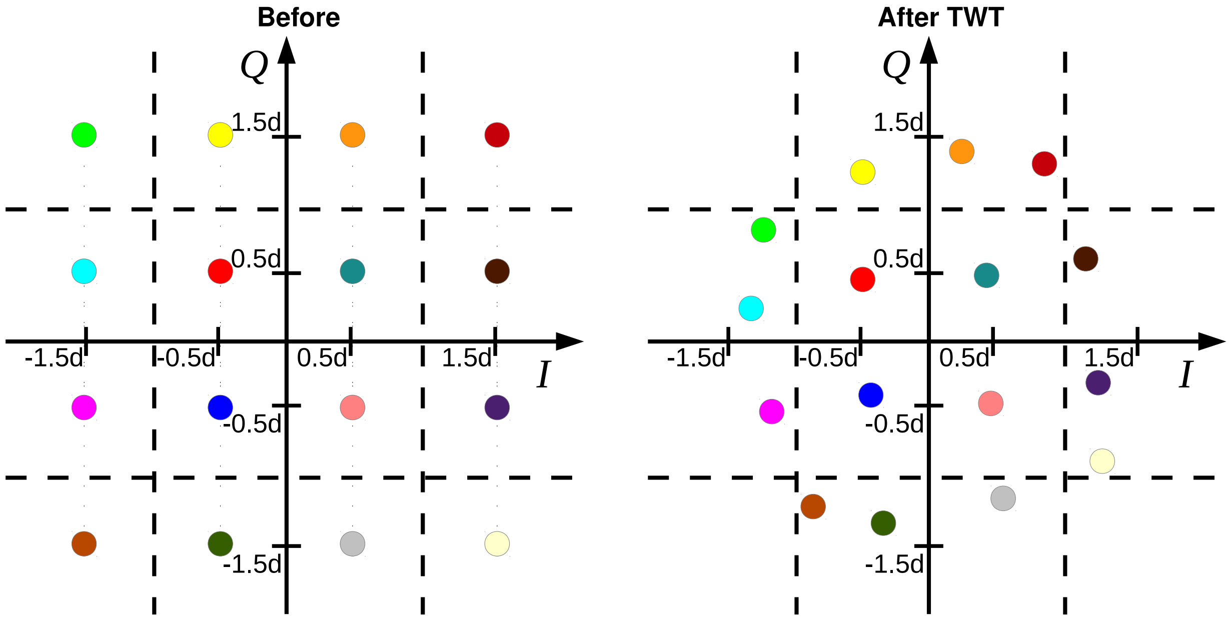
The effect of these non-linearities on the various M-ary digital modulation schemes can be illustrated with constellation diagrams:



Here, on the left, the outer states are drawn in towards the inner states but the inner states are not scaled in the same way. This can pose a difficult problem for the receiver as received outer symbols may be confused with the inner state symbols. On the right the phase distortion has the effect of twisting the received symbols one way or the other, again particularly for the outer states. This may not pose such a problem for PSK modulation schemes as all states lie at the same distance from the origin. However QAM may have higher errors if either of the distortions are present. Therefore QAM modulation schemes may require the amplifier to be run at less than the maximum amplification to help prevent the non-linearities occurring.

5.2 Combined Affect on QAM

Characterized by nonlinearities in the phase. This might be seen after simulating a Travelling Wave Tube (TWT).



Similar effects can be seen with SSPAs although less pronounced especially for phase.
Outer states pulled in towards inner states. This might be a problem for QAM, however QAM also requires lower transmission power for the same resistance to noise (see dB relative to BPSK).

6 SNR for Digital Communication

Analogue communications use Signal to Noise Ratio (SNR)

$$\text{SNR} = \frac{\text{signal power}}{\text{noise power}}$$

Digital systems use:

$$\begin{aligned}\frac{\text{bit energy}}{\text{noise power spectral density}} &= \frac{\text{amount of power for each bit}}{\text{amount of noise across bandwidth}}\\ &= \frac{\text{signal power} \times \text{bit time}}{\frac{\text{noise power}}{\text{bandwidth}}}\end{aligned}$$

Analogue communications use Signal to Noise Ratio (SNR)

$$\text{SNR} = \frac{\text{signal power}}{\text{noise power}}$$

Digital systems use:

$$\begin{aligned}\frac{\text{bit energy}}{\text{noise power spectral density}} &= \frac{\text{signal power} \times \text{bit time}}{\frac{\text{noise power}}{\text{bandwidth}}}\\ &= \frac{\text{signal power} \times \frac{1}{\text{bit rate}}}{\frac{\text{noise power}}{\text{bandwidth}}}\\ &= \frac{\text{signal power}}{\text{noise power}} \times \frac{\text{bandwidth}}{\text{bit rate}}\end{aligned}$$

Analogue communications use Signal to Noise Ratio (SNR)

$$\text{SNR} = \frac{S}{N}$$

Digital systems use:

$$\begin{aligned}\frac{E_b}{N_0} &= \frac{S \times T_b}{\frac{N}{W}}\\ &= \frac{S \times \frac{1}{R}}{\frac{N}{W}}\\ &= \frac{S}{N} \times \left(\frac{W}{R}\right).\end{aligned}$$

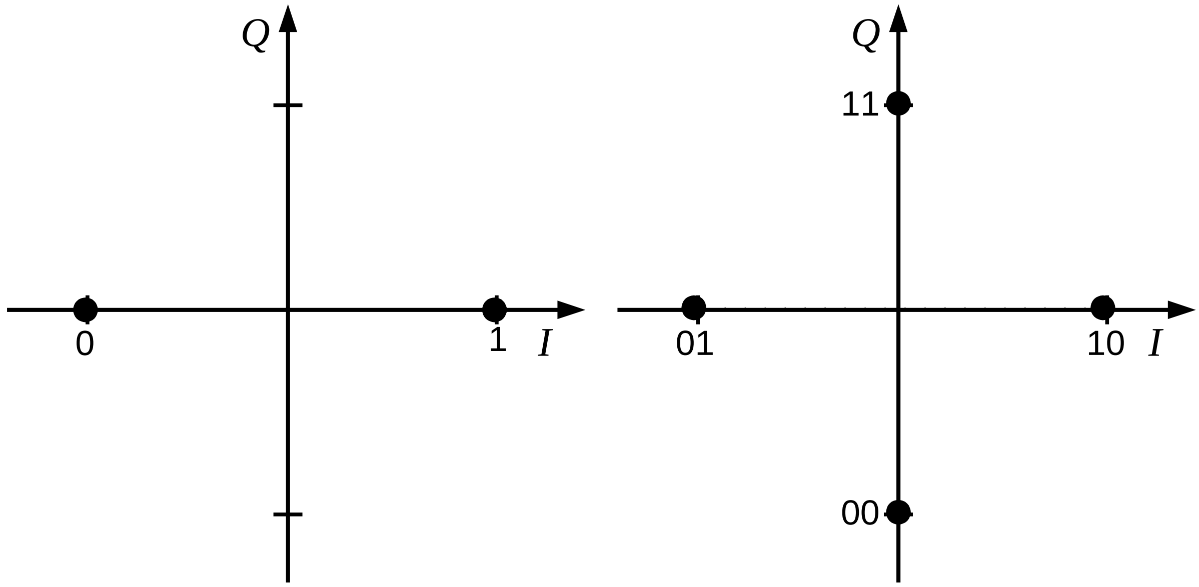
$$\begin{array}{ccccc}\frac{E_b}{N_0} & = & \frac{S}{N} & \times & \left(\frac{W}{R}\right).\\ \downarrow & & \downarrow & & \downarrow\\ \frac{\text{bit energy}}{\text{noise power spectral density}} & = & \text{SNR} & \times & \frac{\text{bandwidth}}{\text{bit rate}}\end{array}$$

SNR is normalized by: bandwidth to bit rate ratio.

Also dimensionless.

Useful for comparing performance of different digital communication processes.

6.1 BPSK and QPSK



BPSK QPSK

Same Bit Error Probability.

BPSK	QPSK
$\times 1$ BPSK signal	$\times 2$ orthogonal BPSK signals
$\times 1$ A amplitude signal	$\times 2$ $A/\sqrt{2}$ signal amplitudes
S Average Power	$S/2$ Average Power
R Bit Rate	$R/2$ Bit Rate
$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right)$	$\frac{E_b}{N_0} = \frac{S/2}{N}\left(\frac{W}{R/2}\right)$ $= \frac{S}{N}\left(\frac{W}{R}\right)$

Same Bit Error Probability.

7 Summary

- Modulation is a way of preparing a signal for transmission
- It needs to make efficient use of available bandwidth
- Overcome problems with noise and other sources of distortion