Digital Communication Systems

S Le Goff

October 2023

School of Engineering

Newcastle University

Contents

6. Transmission of Signals over Wireless Channels

In communication engineering, the signal to be transmitted is generally a low-frequency signal, called the *message* or the *baseband signal*, which contains frequency components ranging from 0 to B, where B is the *bandwidth* of this message.

The wireless transmission of such signal would require the use of a very long antenna. In fact, a basic principle of wireless communications is that the minimum length, L_{min}, of the antenna is inversely proportional to the signal frequency:

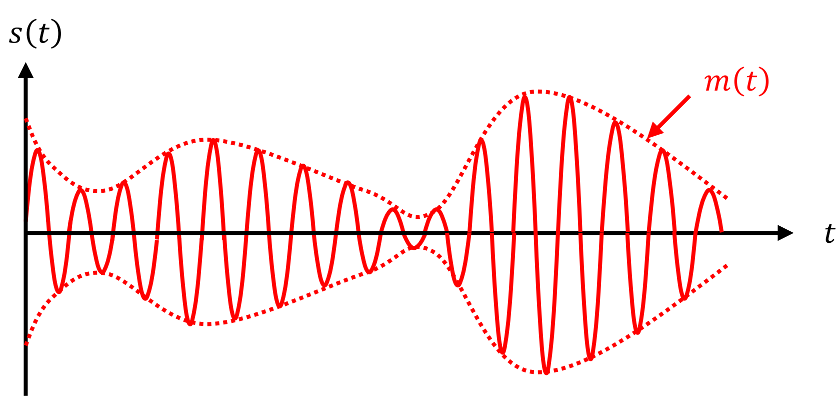
$$L_{min} \sim \frac{c}{10f},$$

where $c \sim 3 \times 10^8$ m/s denotes the speed of light and f is the electromagnetic wave frequency.

To be able to use an antenna of reasonable length, it is therefore necessary to increase the frequency of the message in an artificial way: we must employ a *modulation* technique. Modulation simply consists of mixing the message, $m(t)$, with a high-frequency sinusoidal signal called the *carrier*. The resulting signal, $s(t)$, is called *modulated signal*.

• **Amplitude modulation**

The amplitude modulation operation consists of multiplying the message $m(t)$ with the carrier signal $\cos(2\pi f_0 t)$. The expression of the modulated signal is then $s(t) = m(t) \cdot \cos(2\pi f_0 t)$.



In the frequency domain, the effect of amplitude modulation is to shift the power spectral density, i.e., the spectrum, $M(f)$ of the message from around the zero frequency to around the carrier frequency f_0.

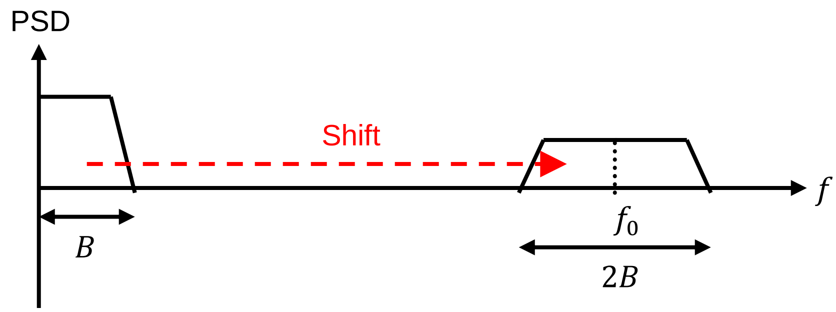
To demonstrate this, consider a particular frequency component f_1 in $M(f)$, i.e., a particular signal $\cos(2\pi f_1 t)$ among all those constituting the message $m(t)$. When mixed with the carrier $\cos(2\pi f_0 t)$, the component $\cos(2\pi f_1 t)$ becomes

$$\cos(2\pi f_1 t) \cdot \cos(2\pi f_0 t) = \frac{1}{2}[\cos[2\pi(f_0 - f_1)t] + \cos[2\pi(f_0 + f_1)t]].$$

This result indicates that the frequency component f_1 is shifted to frequencies $(f_0 - f_1)$ and $(f_0 + f_1)$, but its amplitude has been halved.

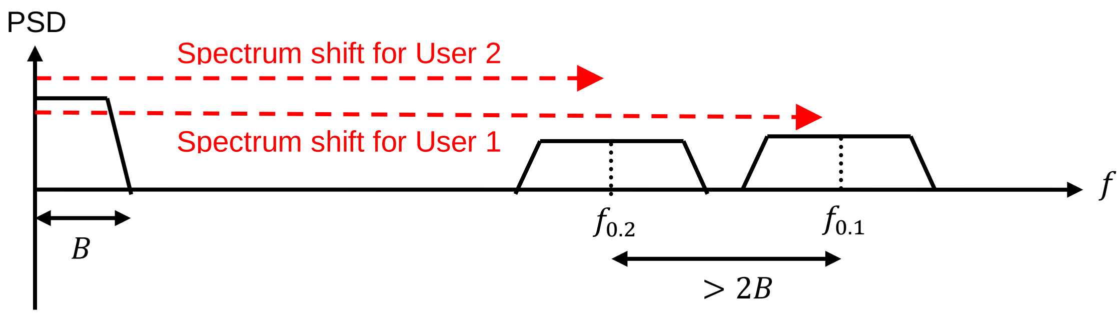
If we apply this simple calculation to all frequency components of $M(f)$, we obtain the spectrum $S(f)$ of the modulated signal. After modulation, the energy of the signal is thus located around the carrier frequency f_0.

The spectrum is symmetrical around f_0. We remark that the bandwidth occupied by the modulated signal is twice that occupied by the message.



In addition to making the transmission possible using an antenna of reasonable size, the modulation operation allows for an equitable sharing of the spectrum by all users according to the international regulations regarding bandwidth allocations.

Consider, without loss of generality, the case of a communication systems with two users. We have seen that User 1 can transmit its message in a frequency range $[f_{0,1} - B, f_{0,1} + B]$ and another User 2 can do the same in another frequency range $[f_{0,2} - B, f_{0,2} + B]$.

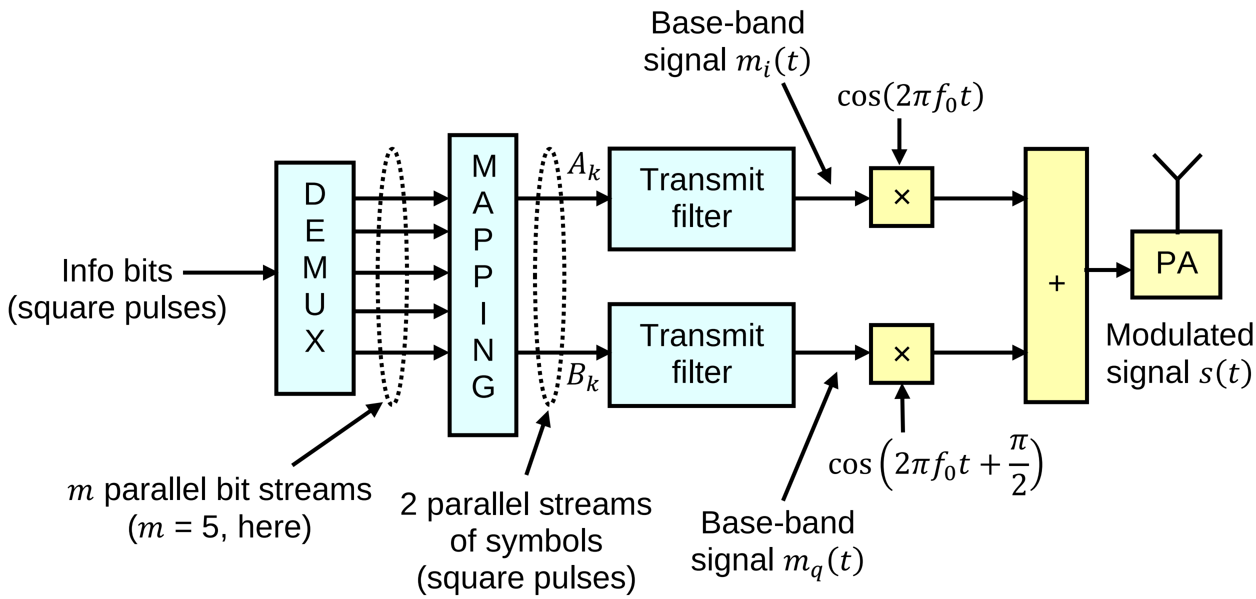


If both frequency ranges do not overlap, i.e., the condition $f_{0,1} - B > f_{0,2} + B$ is satisfied (assuming $f_{0,1} > f_{0,2}$), the modulated signals of both users can be fully separated at the receiver side by using a mixer followed by a loss-pass filter, meaning that there is going to be no interference between these signals. In other words, two (or more) users can share the same communication channel without interfering with each other.

Here, the two modulated signals, that are "mixed" during transmission over the same radio channel, can be separated at the receiver side provided that these signals use carrier frequencies $f_{0,1}$ and $f_{0,2}$ that satisfy the inequality $f_{0,1} - f_{0,2} > 2B$. This channel sharing technique is known as frequency-division multiple access (FDMA).

7. Digital Communication Transmitters

We have represented below the generic structure of a digital communication transmitter. We are going to explain its operation throughout this chapter.

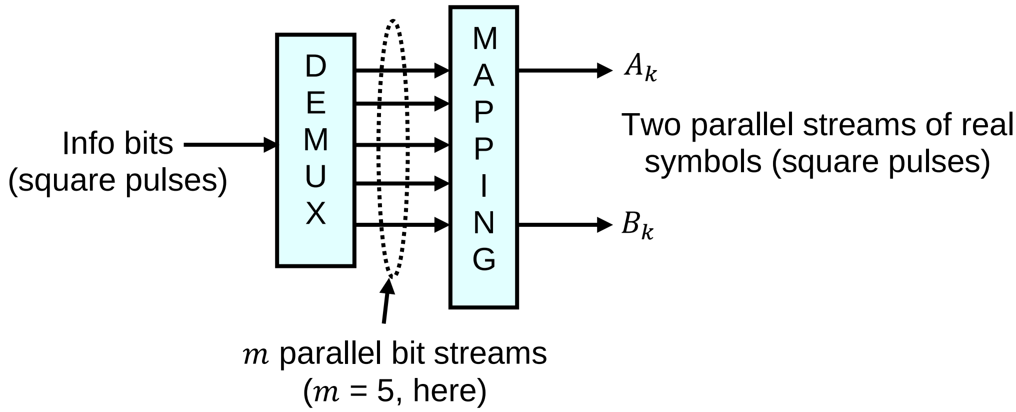


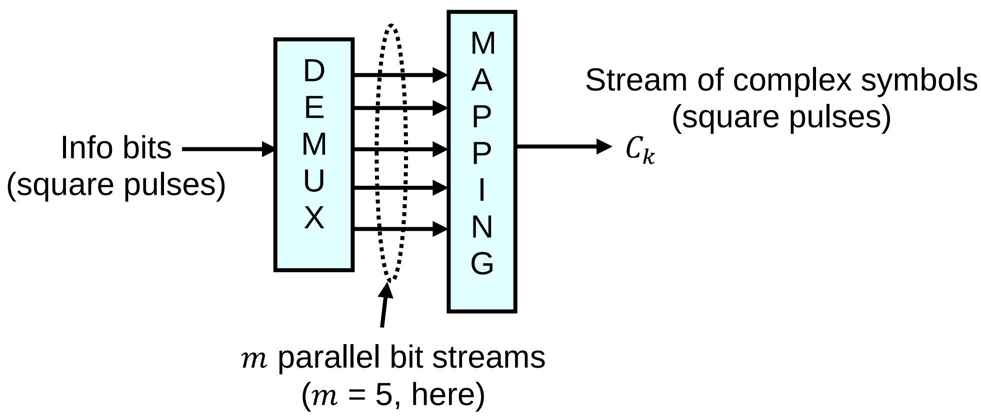
The sequence of information bits to be transmitted is de-multiplexed into m parallel bit streams. Let $(x_{1,k}, x_{2,k}, \ldots, x_{m,k})$ denote the vector of m parallel bits at time kT, where k is an integer and T is the duration of a symbol.

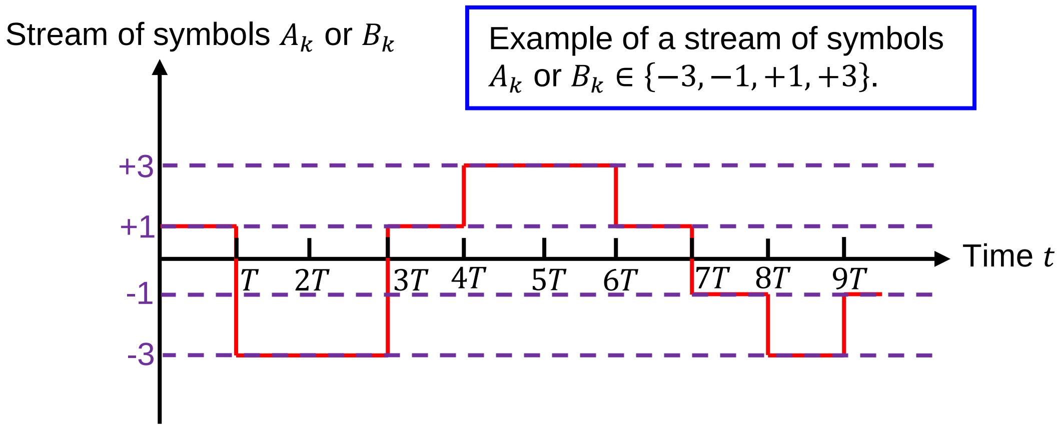
• **Mapping**

The mapping operation converts this m -bit vector $(x_{1,k}, x_{2,k}, \ldots, x_{m,k})$ into a pair of real symbols denoted as A_k and B_k, or equivalently a complex symbol $C_k = A_k + jB_k$. The complex symbol C_k can take $M = 2^m$ different values.

At the mapping block output, we thus have two streams composed at time kT of real symbols A_k and B_k, which is equivalent to a single stream of complex symbols C_k.

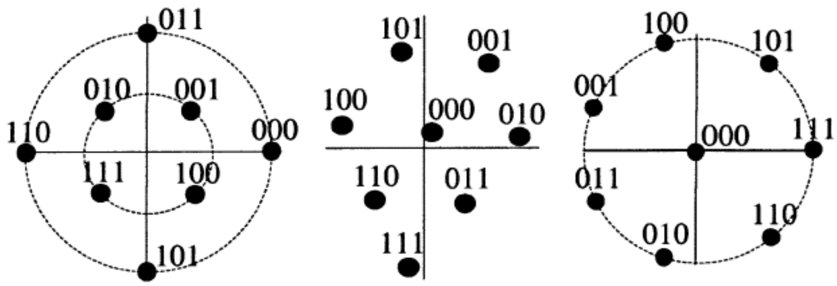

Info bits
(square pulses)
DEMUX
MAPPING
A_k
B_k
Two parallel streams of real symbols (square pulses)
m parallel bit streams
(m = 5, here)


Info bits
(square pulses)
DEMUX
MAPPING
Stream of complex symbols
(square pulses)
C_k
m parallel bit streams
(m = 5, here)

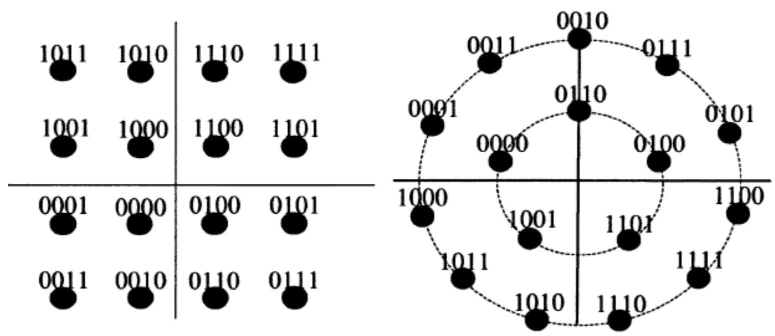

Stream of symbols A_k or B_k
Example of a stream of symbols A_k or $B_k \in \{-3, -1, +1, +3\}$.
+3
+1
-1
-3
T $2T$ $3T$ $4T$ $5T$ $6T$ $7T$ $8T$ $9T$
Time t

A *signal constellation* provides a complete graphical description of the mapping function by specifying the possible values of C_k as well as the one-to-one correspondence between a m-bit vector $(x_{1,k}, x_{2,k}, \ldots, x_{m,k})$ and an M-ary complex symbol C_k.

Example of constellations with $M = 8$ signal points ($m = 3$) --------------------------------------

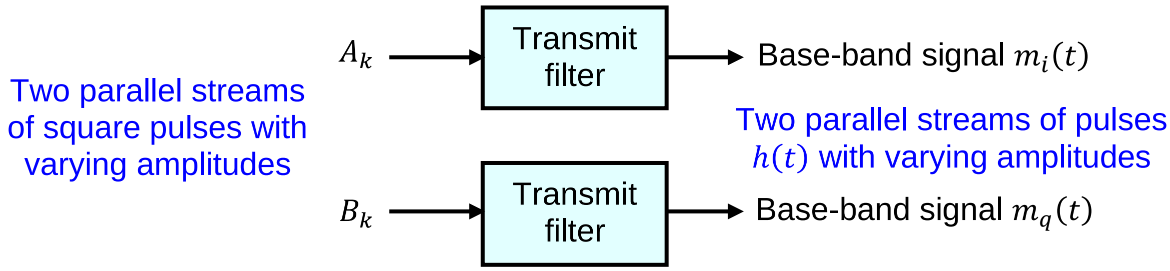


Example of constellations with $M = 16$ signal points ($m = 4$) ------------------------------------



• **Transmit filter**

Filtering is mandatory within the transmitter to constrain the bandwidth of the signal to that dictated by regulation or by the practical necessity to co-habit with other users on adjacent channel frequencies. This implies that we cannot use simple square pulses to transmit symbols A_k and B_k because they require excessive bandwidth to be transmitted. In fact, strictly speaking, square pulses have an infinite bandwidth.



Fourier transform of a square pulse $h_s(t)$ of unit amplitude and duration T:
$H_s(f) = T \cdot \mathrm{sinc}(\pi f T) \cdot e^{-j\pi f T}$.

To minimize the bandwidth of the transmitted signal, we must use a pulse shape that require less bandwidth than square pulses. Let us denote as $h(t)$ the pulse that we wish to use in order to carry symbols A_k and B_k over the channel. The task of the transmit filter thus consists of converting each square pulse into a pulse $h(t)$.

The transfer function of the transmit filter must thus be $\frac{H(f)}{H_s(f)}$, where $H(f)$ is the Fourier transform of the desired pulse $h(t)$, and $H_s(f) = T \cdot \mathrm{sinc}(\pi f T) \cdot e^{-j\pi f T}$ denotes the Fourier transform of a square pulse $h_s(t)$ of unit amplitude and duration T.

At the output of both transmit filters, the expressions of the base-band signals are

$$m_i(t) = \sum_{k=0}^{+\infty} A_k \cdot h(t - kT),$$

and

$$m_q(t) = \sum_{k=0}^{+\infty} B_k \cdot h(t - kT).$$

Both expressions can also be written in a more compact way as a single complex signal $m(t)$ given by

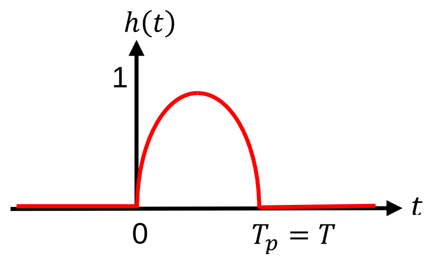
$$m(t) = m_i(t) + j m_q(t) = \sum_{k=0}^{+\infty} C_k \cdot h(t - kT),$$

where $C_k = A_k + jB_k$ is the M-ary symbol transmitted at time kT.

The three signals $m_i(t)$, $m_q(t)$, and $m(t)$ are said to be *base-band* since their spectra are centered on the zero frequency.

Example ---

Assume the transmission of five successive symbols A_k between the times $t = 0$ and $4T$: $A_0 = +1, A_1 = -1, A_2 = +1, A_3 = +1, A_4 = -1$. To this end, we use the pulse shown below.

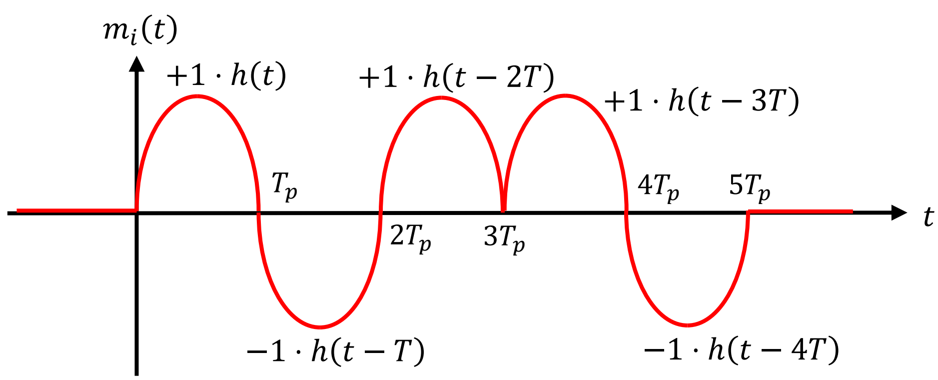


The duration T_p of this pulse is equal to the duration T of symbols A_k and B_k.

The expression of the corresponding base-band signal $m_i(t)$ is therefore given by

$$m_i(t) = A_0 \cdot h(t) + A_1 \cdot h(t-T) + A_2 \cdot h(t-2T) + A_3 \cdot h(t-3T) + A_4 \cdot h(t-4T)$$

$$\rightarrow m_I(t) = +1 \cdot h(t) - 1 \cdot h(t-T) + 1 \cdot h(t-2T) + 1 \cdot h(t-3T) - 1 \cdot h(t-4T).$$

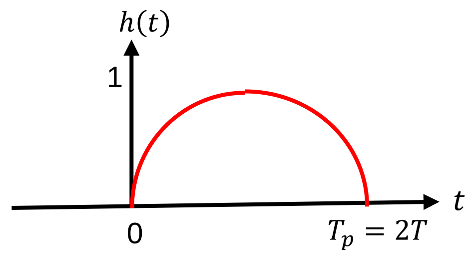


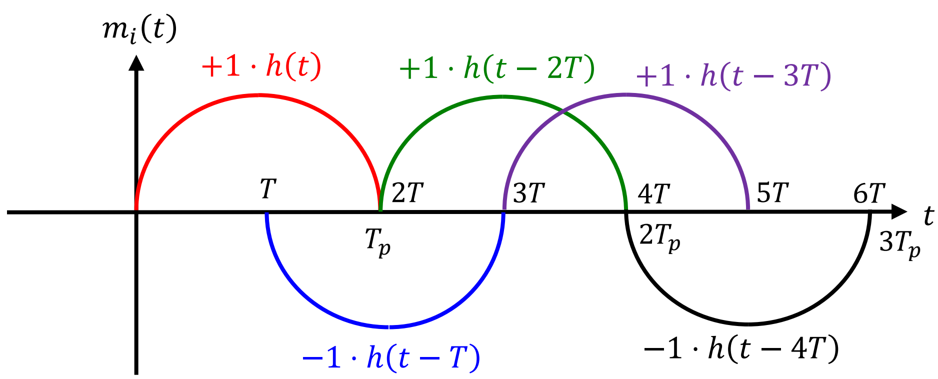
Example ---

The way to minimise the bandwidth of the base-band signal consists of employing slow-varying pulses that have a long duration $T_p > T$. Let us reconsider the previous example by now using pulses with a duration T_p which is twice that of symbols A_k and B_k.

As in the previous example, the expression of the base-band signal $m_i(t)$ is given by

$$m_i(t) = A_0 \cdot h(t) + A_1 \cdot h(t-T) + A_2 \cdot h(t-2T) + A_3 \cdot h(t-3T) + A_4 \cdot h(t-4T)$$
$$\rightarrow m_I(t) = +1 \cdot h(t) - 1 \cdot h(t-T) + 1 \cdot h(t-2T) + 1 \cdot h(t-3T) - 1 \cdot h(t-4T).$$





We see that the pulses overlap because each of them starts before the previous one had sufficient time to go back to zero. The signal $m_i(t)$ depends on two adjacent pulses. For example, from time $t = 2T$ to $3T$, we have $m_i(t) = -1 \cdot h(t-T) + 1 \cdot h(t-2T)$. This could constitute a major problem when trying to recover the individual pulses and the symbol associated with them at the receiver side. This issue, known as inter-symbol interference (ISI), is often encountered in communication engineering.

A closer look at this example indicates that the ISI is, in fact, not really harmful here despite the overlap between two adjacent pulses. We must understand that, in digital communications, the task of the receiver consists of recovering the sequence of transmitted symbols C_k rather than the base-band signal $m(t)$. In this example, sampling $m_i(t)$ at times $t = pT$, where p is an integer, allows us to recover the corresponding symbols A_p perfectly because the contribution of the adjacent pulses at these particular times is zero. Hence, there is no ISI.

We conclude that lengthening the duration of each pulse beyond the symbol duration T will reduce the bandwidth required for transmitting symbols C_k. However, doing so may also result in some interference between adjacent symbols. We must always be aware of that issue as ISI could potentially cause a great deal of confusion when trying to recover the transmitted symbols, Note that, in our example above, it is possible to completely overcome ISI by sampling the base-band signal at carefully chosen times.

• **Symbol and bit rates**

The speed at which information is transmitted is measured by the symbol and, more importantly, the bit rate. The *symbol rate*, D_s, is given by

$$D_s = \frac{1}{T} \text{ symbols/sec,}$$

whereas the *bit rate*, D_b, is expressed as

$$D_b = \frac{m}{T} \text{ bits/sec.}$$

Note that what really matters in practice is the bit rate as our goal is to transmit information bits

• **Power spectral density of the base-band signal**

It is interesting to determine the power spectral density (PSD), $M(f)$, of the base-band signal $m(t)$. In the context of these lecture notes, we can use the following expression:

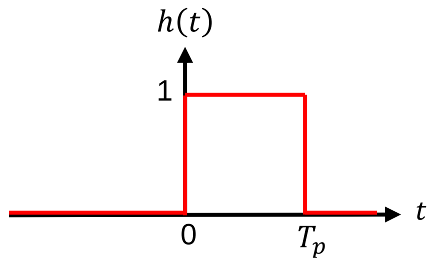
$$M(f) = K \cdot |H(f)|^2,$$

where K is a constant and $H(f)$ is the Fourier transform of the pulse $h(t)$.

This equation indicates that the bandwidth B_{bb} of the base-band signal is equal to that of $H(f)$.

Example ---

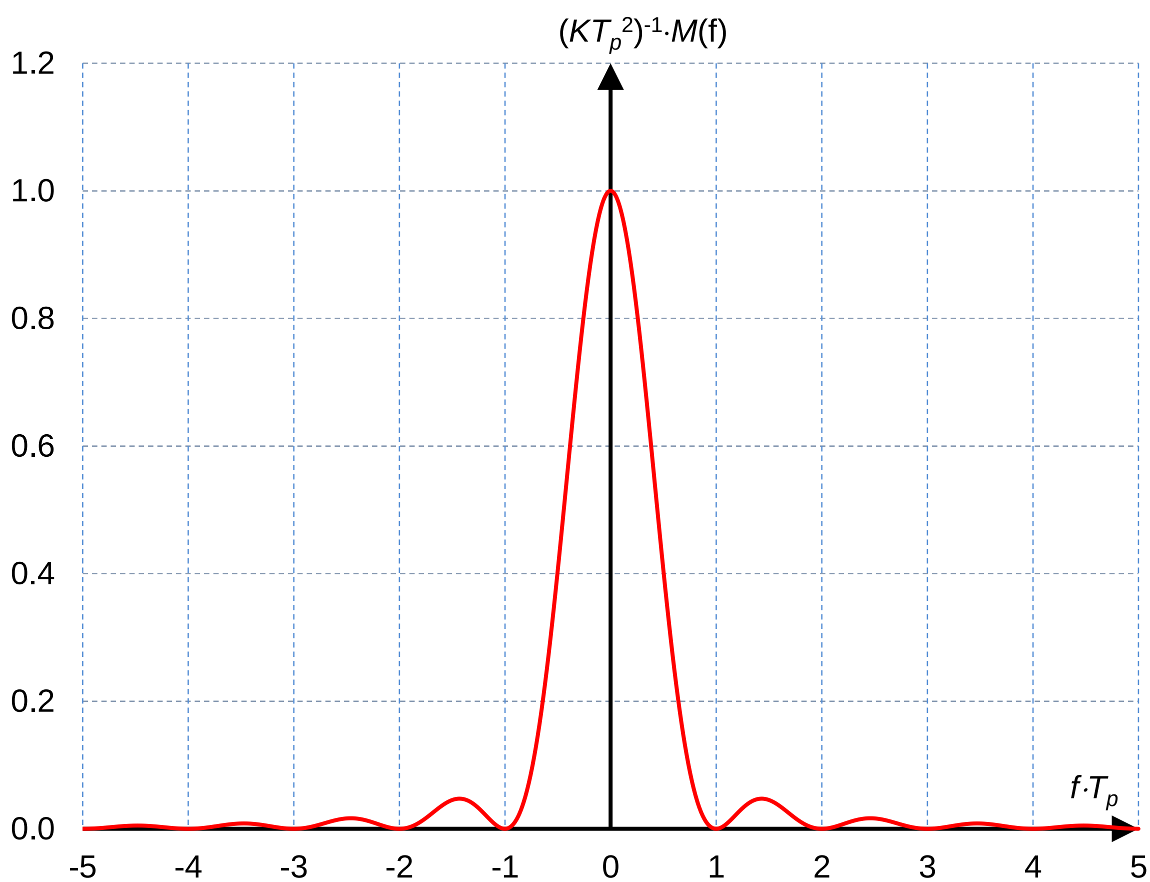
Assume that $h(t)$ is a square pulse of unit amplitude and duration T_p.



The Fourier transform of $h(t)$ is given by $H(f) = \int_{-\infty}^{+\infty} h(t) \cdot e^{-j2\pi ft} \; dt = \int_0^{T_p} e^{-j2\pi ft} \; dt = \left[\frac{e^{-j2\pi ft}}{-j2\pi f}\right]_0^{T_p} = -\frac{1}{j2\pi f} \cdot \left(e^{-j2\pi fT_p} - e^{-j2\pi f0}\right) = -\frac{1}{j2\pi f} \cdot \left(e^{-j\pi fT_p} - e^{+j\pi fT_p}\right) \cdot e^{-j\pi fT_p} = \frac{\sin(\pi fT_p)}{\pi f} \cdot e^{-j\pi fT_p} = T_p \cdot \mathrm{sinc}(\pi fT_p) \cdot e^{-j\pi fT_p}$.

Finally, the PSD of the base-band signal $m(t)$ is thus given by

$$M(f) = K \cdot |H(f)|^2 = K \cdot T_p^2 \cdot \mathrm{sinc}^2(\pi fT_p).$$

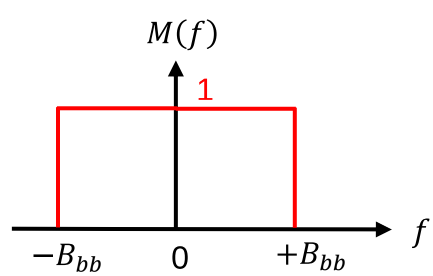


We see that $M(f)$ is centered on the zero frequency, which is indeed the definition of a base-band signal. It can also be seen that most of the signal energy is located inside the main lobe of the spectrum ranging from $f = 0$ to $\frac{1}{T_p}$, and almost all of it is included in the main and second lobes, i.e., in the spectral range from $f = 0$ to $\frac{2}{T_p}$. Therefore, although the bandwidth of this signal is, strictly speaking, infinite, it is probably reasonable, in practice, to consider that the bandwidth B_{bb} is given by $B_{bb} \sim \frac{2}{T_p}$.

To minimise B_{bb}, we need to maximise the pulse duration T_p. However, if the latter becomes greater than the symbol duration T, there may be inter-symbol interference. So, the best choice consists of ensuring that the square pulse duration is equal to the symbol duration: $T_p = T$. In this case, we can write $B_{bb} \sim \frac{2}{T}$.

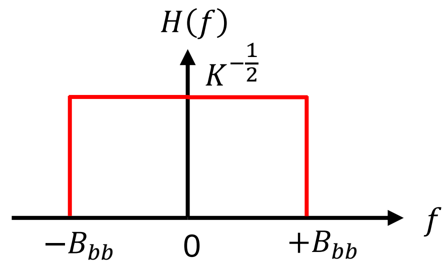
Example --

We would like the PSD of the base-band signal $m(t)$ to be as shown below.



What pulse do we need to use to have such an *ideal* PSD?

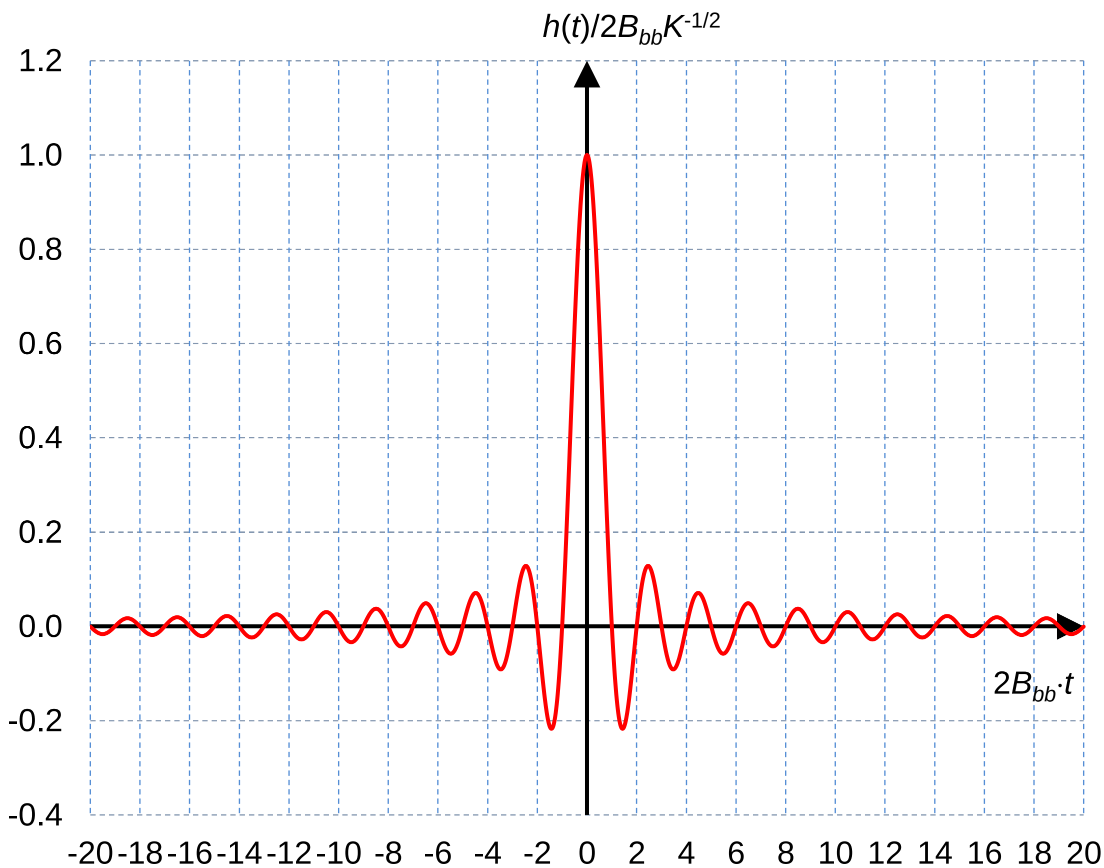
We should have $|H(f)|^2 = \frac{1}{K} \cdot M(f)$. Assuming that $H(f)$ is real, we can write $H(f) = \frac{1}{\sqrt{K}} \cdot \sqrt{M(f)}$.

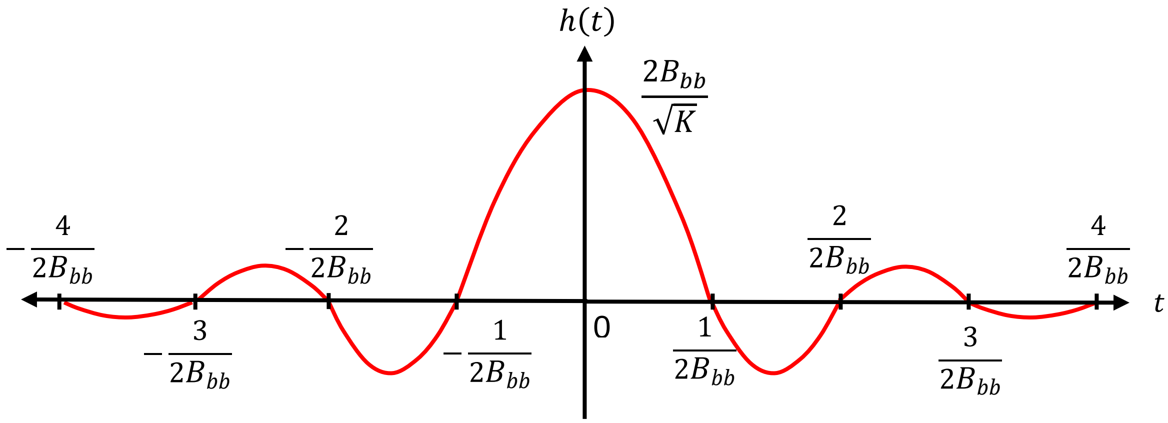


The inverse Fourier transform of $H(f)$ is given by

$h(t) = \int_{-\infty}^{+\infty} H(f) \cdot e^{+j2\pi ft}\, df = \frac{1}{\sqrt{K}} \cdot \int_{-B_{bb}}^{+B_{bb}} e^{+j2\pi ft}\, df = \frac{1}{\sqrt{K}} \cdot \left[\frac{e^{+j2\pi ft}}{j2\pi t}\right]_{-B_{bb}}^{+B_{bb}} = \frac{1}{\sqrt{K}} \cdot \frac{1}{j2\pi t} \cdot \left(e^{+j2\pi B_{bb} t} - e^{-j2\pi B_{bb} t}\right) = \frac{1}{\sqrt{K}} \cdot \frac{\sin(2\pi B_{bb} t)}{\pi t} = \frac{2B_{bb}}{\sqrt{K}} \cdot \mathrm{sinc}(2\pi B_{bb} t).$

The pulse $h(t)$ must thus be a sine cardinal function.





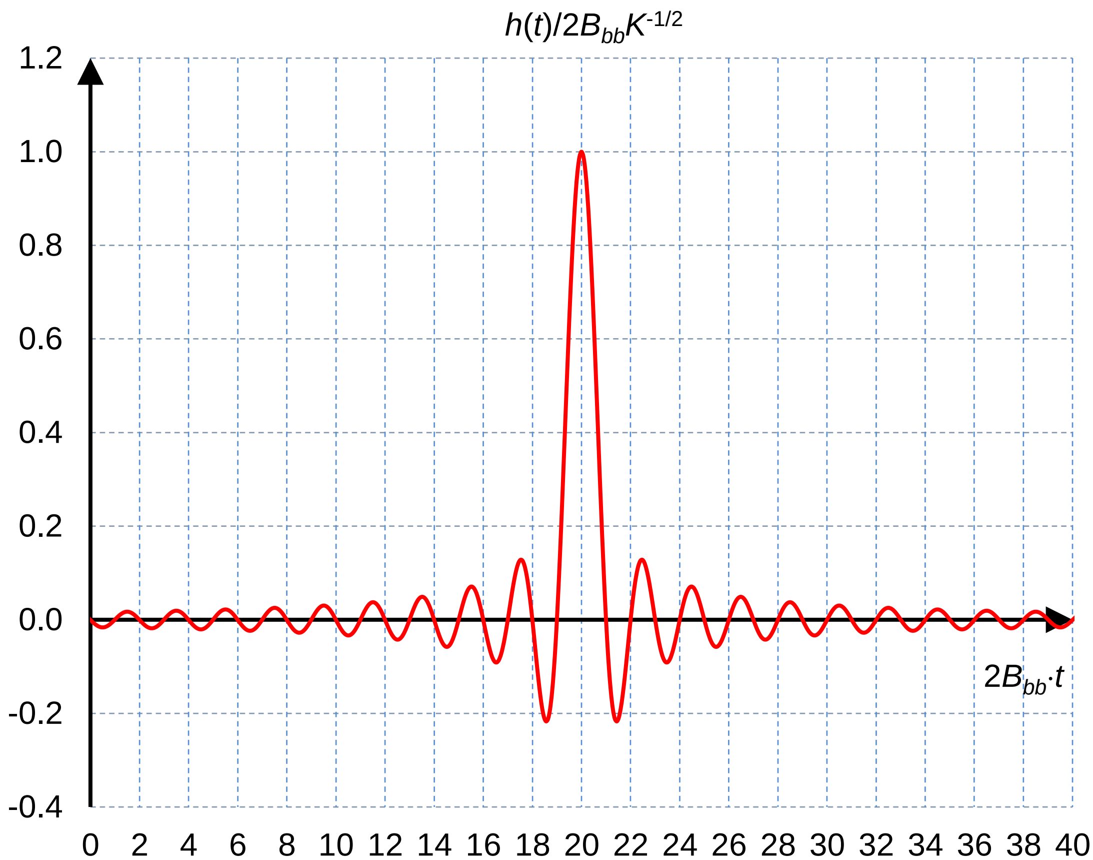
Due to its very good spectral behaviour, a sine cardinal pulse seems to be an excellent choice for carrying symbols C_k. However, the problem with this pulse is that its duration is, strictly speaking, infinite. It has neither a start time nor an end time. In practice, we can solve this issue by truncating the pulse duration. For instance, we can replace it with the following (causal) pulse signal:

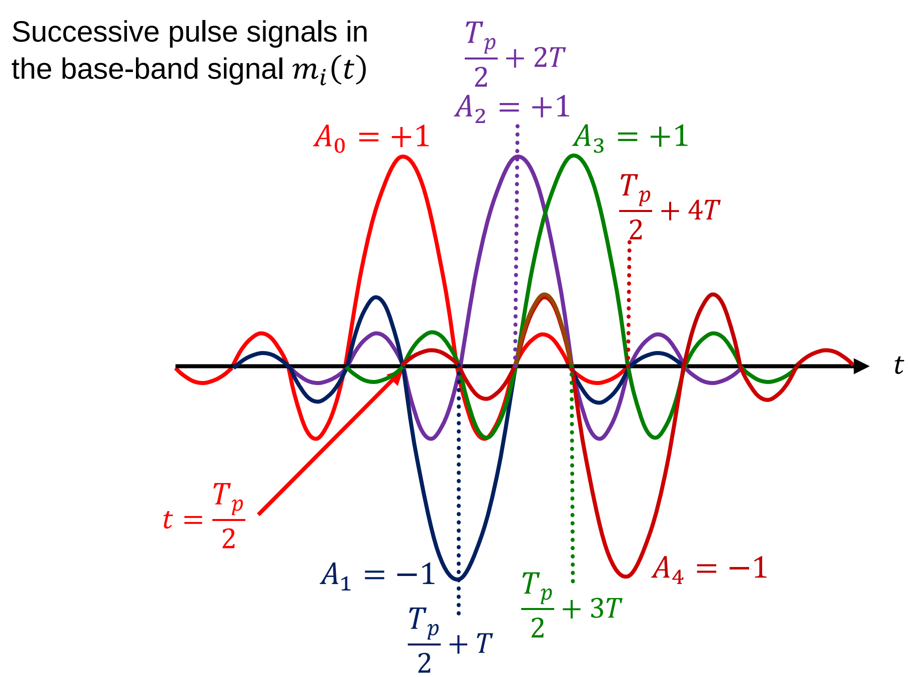
- $h(t) = 0$ for $t < 0$,
- $h(t) = \frac{2B_{bb}}{\sqrt{K}} \cdot \text{sinc}\left(2\pi B_{bb}\left(t - \frac{T_p}{2}\right)\right)$ for $0 < t < T_p = \frac{40}{2B_{bb}}$,
- $h(t) = 0$ for $t > T_p$.

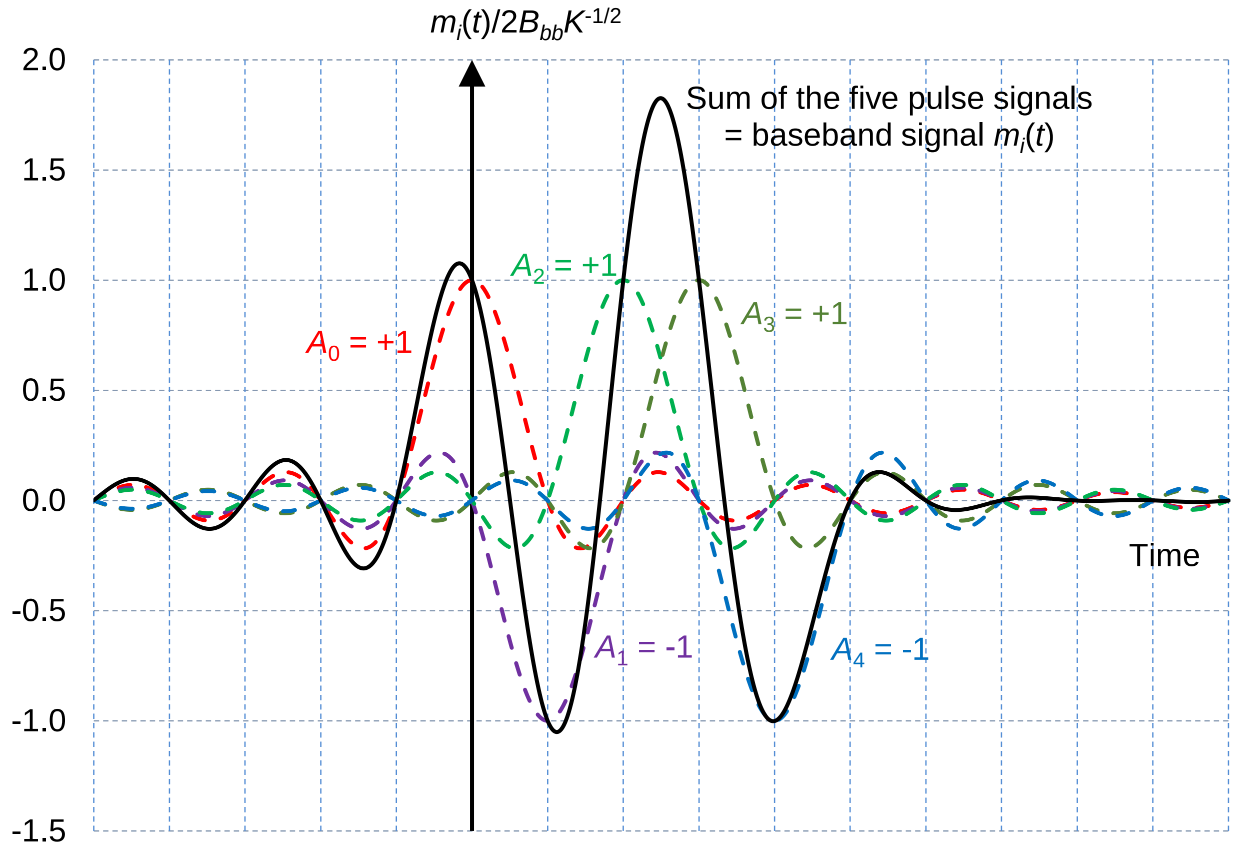
This seems to be a very long pulse signal. It has the potential to cause inter-symbol interference if the symbol duration T is less than $T_p = \frac{40}{2B_{bb}}$.

If we choose the symbol duration T to be equal to $\frac{1}{2B_{bb}}$, which is forty times smaller than T_p, we however notice something very interesting: pulses do overlap as expected, but, in spite of that, there is no ISI.

This is illustrated in the example below where we have sketched the signal $m_i(t)$, assuming the transmission of five successive symbols A_k for $k = 0$ to 4 as follows: $A_0 = +1$, $A_1 = -1$, $A_2 = +1$, $A_3 = +1$, and $A_4 = -1$.


$h(t)/2B_{bb}K^{-1/2}$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
$2B_{bb}\cdot t$


Successive pulse signals in the base-band signal $m_i(t)$
$\frac{T_p}{2}+2T$
$A_2=+1$
$A_0=+1$
$A_3=+1$
$\frac{T_p}{2}+4T$
t
$t=\frac{T_p}{2}$
$A_1=-1$
$\frac{T_p}{2}+3T$
$A_4=-1$
$\frac{T_p}{2}+T$



In digital communications, the task of the receiver consists of recovering the sequence of transmitted symbols rather than the full base-band signal.

If the symbol duration T is chosen so that $T = \frac{1}{2B_{bb}}$, sampling $m_i(t)$ at times $t = \frac{T_p}{2} + pT$ will allow for the perfect recovery of the corresponding symbol A_p, $p \in \{0, 1, 2, 3, 4\}$, because the contribution of the adjacent sine cardinal pulses at that time is zero.

We can demonstrate this mathematically as shown below.

The expression of the base-band signal $m_i(t)$ is $m_i(t) = \sum_{k=0}^{4} A_k \cdot h(t - kT) = \frac{2B_{bb}}{\sqrt{K}} \cdot \sum_{k=0}^{4} A_k \cdot \mathrm{sinc}\left(2\pi B_{bb}\left(t - \frac{T_p}{2} - kT\right)\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot \sum_{k=0}^{4} A_k \cdot \mathrm{sinc}\left(2\pi B_{bb}\left(t - \frac{T_p}{2} - \frac{k}{2B_{bb}}\right)\right)$.

By sampling $m_i(t)$ at time $t = \frac{T_p}{2} + pT$, we obtain a sample of the base-band signal which is expressed as $m_i\left(\frac{T_p}{2} + pT\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot \sum_{k=0}^{4} A_k \cdot \mathrm{sinc}\left(2\pi B_{bb}\left(\frac{T_p}{2} + pT - \frac{T_p}{2} - \frac{k}{2B_{bb}}\right)\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot \sum_{k=0}^{4} A_k \cdot \mathrm{sinc}\left(2\pi B_{bb}\left(\frac{p-k}{2B_{bb}}\right)\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot \sum_{k=0}^{4} A_k \cdot \mathrm{sinc}\big(\pi(p-k)\big)$.

Let us develop this equation to better understand it: $m_i\left(\frac{T_p}{2} + pT\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot \Big(A_0 \cdot \mathrm{sinc}(\pi p) + A_1 \cdot \mathrm{sinc}\big(\pi(p-1)\big) + A_2 \cdot \mathrm{sinc}\big(\pi(p-2)\big) + A_3 \cdot \mathrm{sinc}\big(\pi(p-3)\big) + A_4 \cdot \mathrm{sinc}\big(\pi(p-4)\big)\Big)$.

Therefore, we can finally write $m_i\left(\frac{T_p}{2}\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot A_0$, $m_i\left(\frac{T_p}{2} + T\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot A_1$, $m_i\left(\frac{T_p}{2} + 2T\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot A_2$, $m_i\left(\frac{T_p}{2} + 3T\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot A_3$, and $m_i\left(\frac{T_p}{2} + 4T\right) = \frac{2B_{bb}}{\sqrt{K}} \cdot A_4$.

It has been possible to recover the individual symbols without any ISI despite the overlapping between adjacent sine cardinal pulses.

The only condition for the recovery to be possible is that we have the following relationship between symbol duration and bandwidth: $T = \frac{1}{2B_{bb}}$ or, equivalently, $B_{bb} = \frac{1}{2T}$. In practice, there is, in fact, no restriction on the symbol duration. We can freely choose the value of T and the bandwidth required for transmission of the base-band signal will then be given by

$$B_{bb} = \frac{1}{2T}.$$

• **Bandwidth and symbol/bit rates**

Both previous examples have shown that the bandwidth B_{bb} of the base-band signal depends on the shape of the pulse: the smoother the pulse, the smaller the bandwidth.

These examples also indicate that B_{bb} is inversely proportional to the symbol duration T.

Consequently, the expressions $D_s = \frac{1}{T}$ symbols/sec and $D_b = \frac{m}{T}$ bits/sec show that the bandwidth B_{bb} is proportional to the symbol rate D_s and the bit rate D_b. We now understand where the term *broadband* comes from. *Broadband* is another name for *high symbol/bit rates*.

With the kind of pulses that are used in practical communication systems, the bandwidth B_{bb} required for the transmission of $m(t)$ is given by

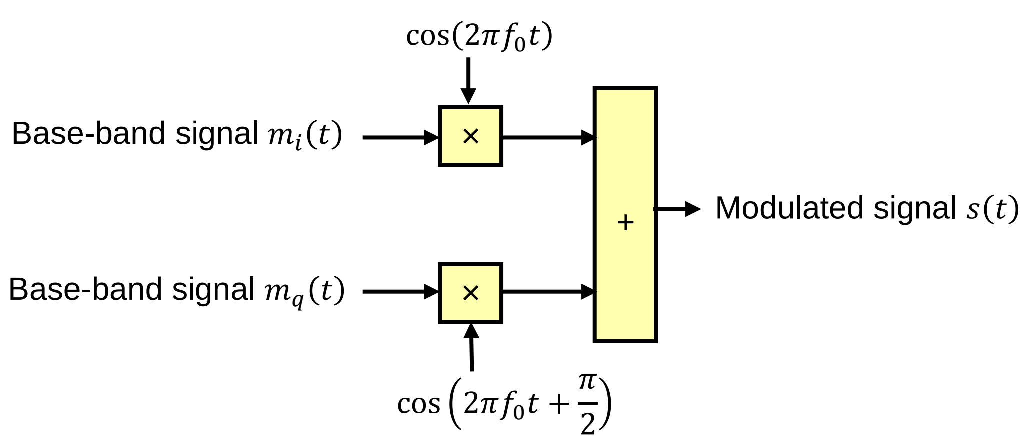
$$B_{bb} = \frac{1+\alpha}{2T},$$

where α is a fixed system parameter, called roll-off factor, ranging from 0 to 1. This crucial result will be demonstrated later.

Let us for now focus on the bit rate as our ultimate goal in digital communications is to transmit bits. The previous expression can be re-written as $B_{bb} = \frac{(1+\alpha)\cdot D_b}{2m}$.

The amount of bandwidth required for transmitting the base-band signal $m(t)$ is clearly proportional to the bit rate D_b.

- **Modulation**



The equation of the modulated signal $s(t)$ is given by

$$s(t) = \sum_{k=0}^{+\infty} A_k \cdot h(t-kT) \cdot \cos(2\pi f_0 t) + B_k \cdot h(t-kT) \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right),$$

where f_0 designates the carrier frequency, $\cos(2\pi f_0 t)$ is the in-phase carrier component, and $\cos\left(2\pi f_0 t + \frac{\pi}{2}\right)$ is the quadrature carrier component.

It is worth mentioning that this equation can be re-written in several different ways which are as follows:

$$s(t) = m_i(t) \cdot \cos(2\pi f_0 t) + m_q(t) \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right),$$

$$s(t) = m_i(t) \cdot \cos(2\pi f_0 t) - m_q(t) \cdot \sin(2\pi f_0 t),$$

$$s(t) = \Re\left\{m(t) \cdot e^{j2\pi f_0 t}\right\},$$

and

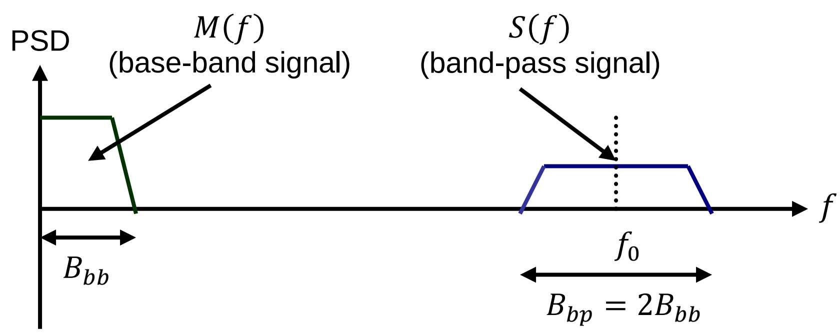
$$s(t) = \Re\left\{\sum_{k=0}^{+\infty} C_k \cdot h(t-kT) \cdot e^{j2\pi f_0 t}\right\},$$

where $e^{j2\pi f_0 t}$ designates the complex carrier component. The two last equations are particularly elegant and many communication engineers often intentionally omit the real part when using them and simply write that the modulated signal is expressed as

$$s(t) = m(t) \cdot e^{j2\pi f_0 t} = \sum_{k=0}^{+\infty} C_k \cdot h(t-kT) \cdot e^{j2\pi f_0 t}.$$

When this happens, readers are expected to remember that only the real part is actually transmitted.

As previously seen, the mixing operation shifts the spectrum of $m(t)$ from the frequency 0 to the frequency f_0. Thus, the modulated signal $s(t)$ is also referred to as the *band-pass signal*.



The bandwidth B_{bp} occupied by the band-pass signal is twice that occupied by the base-band signal. With the pulses $h(t)$ used in practice, the bandwidth needed for the transmission of the modulated signal is therefore equal to

$$B_{bp} = 2B_{bb} = \frac{(1+\alpha)\cdot D_b}{m}.$$

- **Spectral efficiency**

At this stage, we need to introduce an important parameter called *spectral efficiency* and defined as

$$\eta = \frac{D_b}{B_{bp}}.$$

The value of η, expressed in bits/sec/Hz, measures the bit rate that can be transmitted per Hz of bandwidth. A higher value of η indicates that the transmission system can transmit a higher bit rate while keeping the required bandwidth unchanged, which is obviously a good thing. It is easy to show that the spectral efficiency is simply expressed as

$$\eta = \frac{m}{1+\alpha}.$$

This equation indicates that η depends on both parameters m and α. Generally, communication engineers assume that the system is designed with $\alpha = 0$, which is obviously the most optimistic scenario in terms of spectral efficiency but also the most unrealistic one as we will later see. They thus consider that $\eta = m$ bits/sec/Hz.

We are now going to introduce three different types of digital modulation techniques that are commonly used in practice.

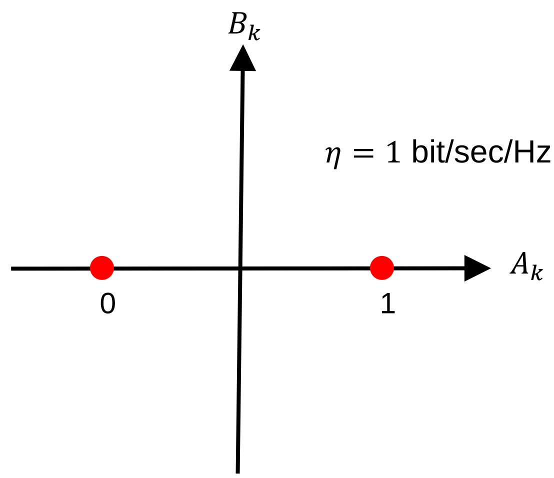
- **M-state amplitude shift keying (M-ASK) modulation**

For these modulation schemes, the symbol C_k is a real symbol that can take M possible values: $C_k = A_k \in \{\pm 1, \pm 3, \pm 5, \dots\}$. Therefore, we can write

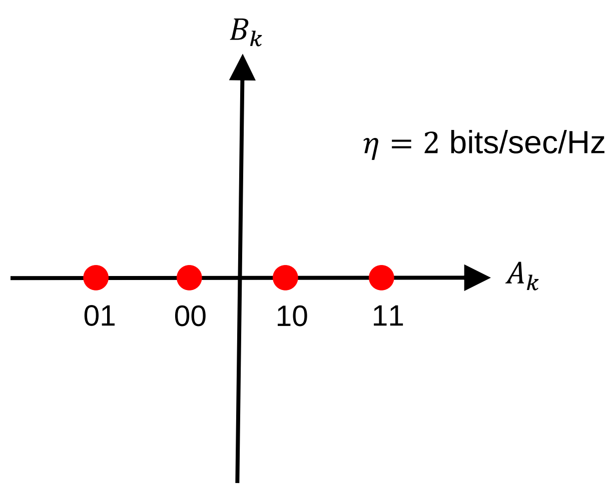
$$s(t) = \sum_{k=0}^{+\infty} A_k \cdot h(t - kT) \cdot \cos(2\pi f_0 t),$$

which clearly corresponds to a classical analogue amplitude modulation for which the message is the in-phase base-band signal $m_i(t)$.

Constellation of 2-ASK



Constellation of 4-ASK with *Gray mapping*

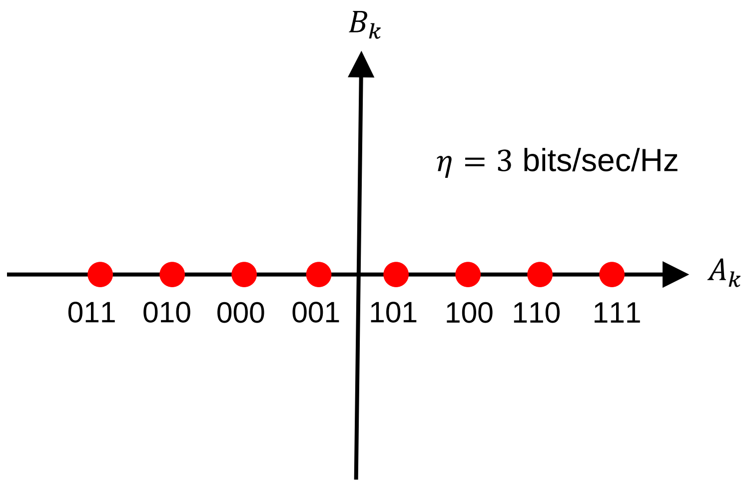


Gray mapping is a type of one-to-one correspondence between m-bit vectors and symbols C_k in which the m-bit vectors associated with two nearest neighbour signal points in the constellation always differ by only one bit.

Gray mapping is commonly used by communication engineers as it minimizes the bit error probability, i.e., the probability of erroneous detection of a transmitted bit at the receiver output.

In fact, the erroneous detection of a transmitted symbol C_k at the receiver side is almost always due to some confusion between two nearest neighbour signal points in the constellation. When this happens, Gray mapping ensures that only one bit in the m-bit vector associated with the transmitted symbol C_k is erroneously detected despite the erroneous symbol detection.

Constellation of 8-ASK with Gray mapping



We can also have 16-ASK, 32-ASK, 64-ASK, and so on.

- **M-state phase shift keying (M-PSK) modulation**

For these modulation schemes, the complex symbol C_k has a constant modulus, i.e., C_k can be written in the form $e^{j\Phi_k}$ with $\Phi_k \in \left\{0, \frac{2\pi}{M}, \frac{4\pi}{M}, \dots, \frac{2\pi(M-1)}{M}\right\}$.

Therefore, we can write

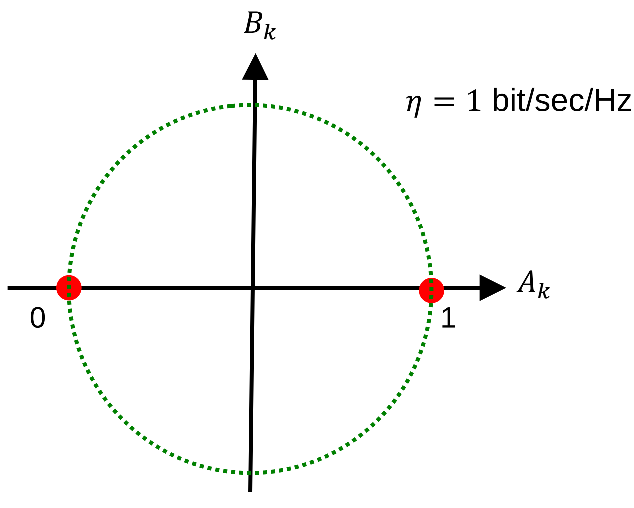
$$s(t) = \Re\left\{\sum_{k=0}^{+\infty} h(t-kT) \cdot e^{j(2\pi f_0 t + \Phi_k)}\right\},$$

or, equivalently,

$$s(t) = \sum_{k=0}^{+\infty} h(t-kT) \cdot \cos(2\pi f_0 t + \Phi_k),$$

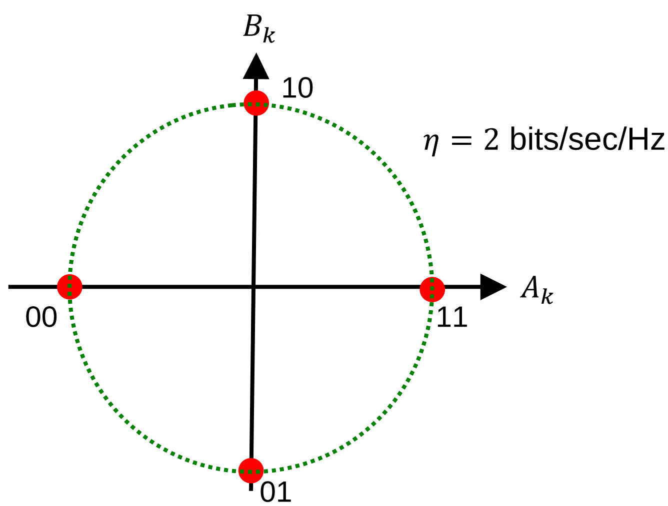
which clearly corresponds to a classical phase modulation for which the phase of the carrier can only take discrete values.

Constellation of 2-PSK



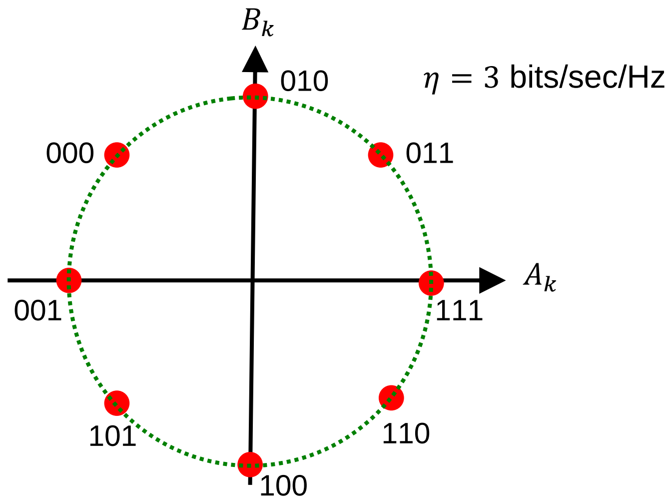
2-PSK is also known as binary PSK (BPSK). It is identical to 2-ASK.

Constellation of 4-PSK with Gray mapping



4-PSK is also known as quaternary PSK (QPSK).

Constellation of 8-PSK with Gray mapping

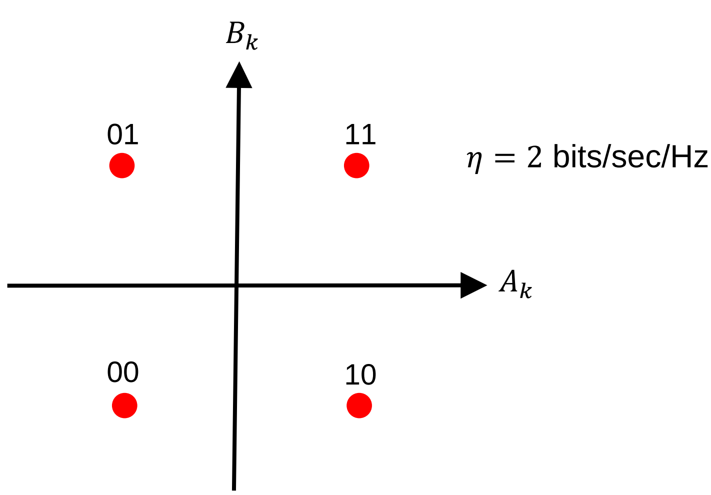


We can also have 16-PSK, 32-PSK, 64-PSK, and so on.

• **M-state quadrature amplitude modulation (M-QAM)**

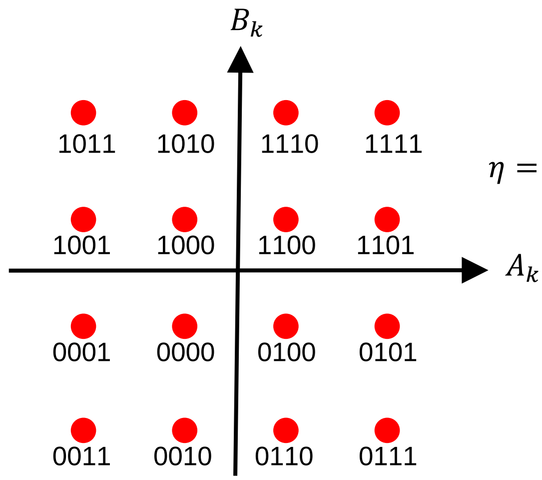
For such modulation schemes, the symbol $C_k = A_k + jB_k$ is a complex symbol that can take M possible values, with $A_k, B_k \in \{\pm 1, \pm 3, \pm 5, \ldots\}$. The equation of the modulated signal $s(t)$ is given by $s(t) = \sum_{k=0}^{+\infty} A_k \cdot h(t-kT) \cdot \cos(2\pi f_0 t) + B_k \cdot h(t-kT) \cdot \cos\left(2\pi f_0 t + \frac{\pi}{2}\right)$.

Constellation of 4-QAM with Gray mapping

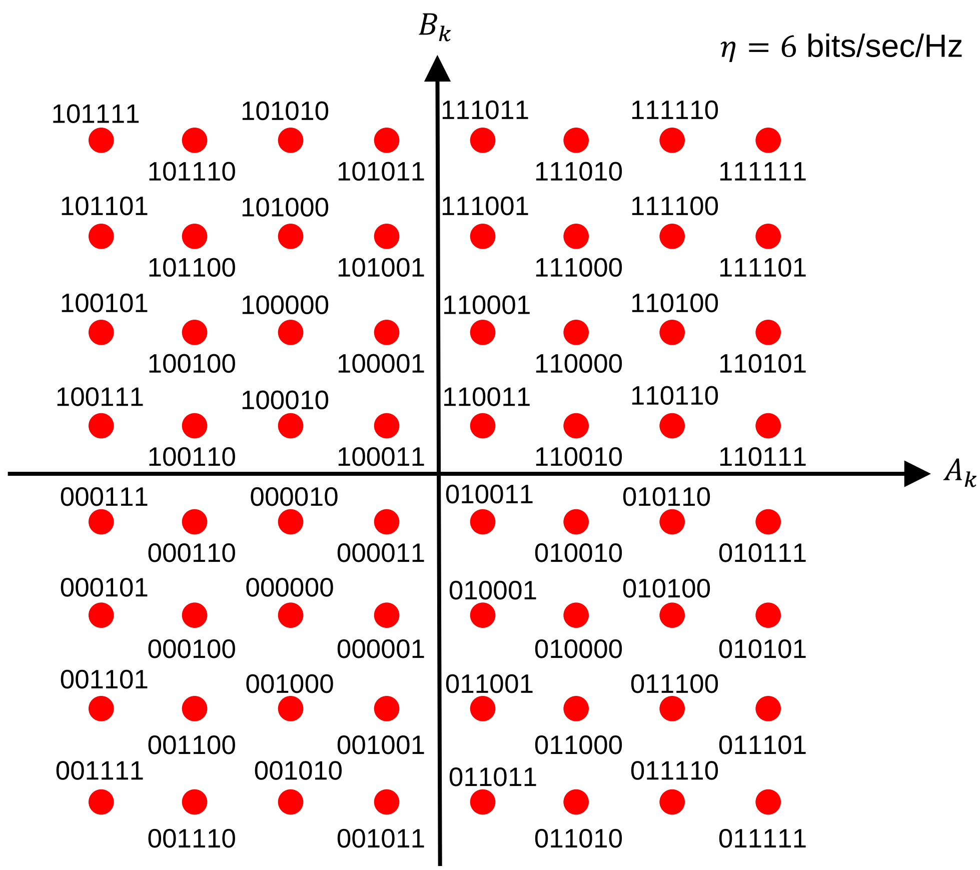


4-QAM is identical to QPSK.

Constellation of 16-QAM with Gray mapping



Constellation of 64-QAM with Gray mapping



We can also have 256-QAM, 1024-QAM, 4096-QAM, and so on.

Example ---

Assume that the bandwidth of our system is equal to $B_{bp} = 100$ kHz. What are the maximum bit rates that can be achieved when using QPSK, 8-PSK, 16-QAM, and 64-QAM?

- QPSK: $D_{b,max} = m \cdot B_{bp}$ = 2 bits/s/Hz × 100 kHz = 200 kbits/s;
- 8-PSK: $D_{b,max} = m \cdot B_{bp}$ = 3 bits/s/Hz × 100 kHz = 300 kbits/s;
- 16-QAM: $D_{b,max} = m \cdot B_{bp}$ = 4 bits/s/Hz × 100 kHz = 400 kbits/s;
- 64-QAM: $D_{b,max} = m \cdot B_{bp}$ = 6 bits/s/Hz × 100 kHz = 600 kbits/s.

What are the actual bit rates that are achieved when using QPSK, 8-PSK, 16-QAM, and 64-QAM and a system with a roll-off factor $\alpha = 1$?

- QPSK: $D_b = \frac{m}{1+\alpha} \cdot B_{bp}$ = 1 bits/s/Hz × 100 kHz = 100 kbits/s;
- 8-PSK: $D_b = \frac{m}{1+\alpha} \cdot B_{bp}$ = 1.5 bits/s/Hz × 100 kHz = 150 kbits/s;
- 16-QAM: $D_b = \frac{m}{1+\alpha} \cdot B_{bp}$ = 2 bits/s/Hz × 100 kHz = 200 kbits/s;
- 64-QAM: $D_b = \frac{m}{1+\alpha} \cdot B_{bp}$ = 3 bits/s/Hz × 100 kHz = 300 kbits/s.

Example ---

Assume that we want to design a communication system that achieves a bit rate $D_b = 1.2$ Mbits/s. What are the minimum bandwidths that are required when using QPSK, 8-PSK, 16-QAM, and 64-QAM?

- QPSK: $B_{bp,min} = \frac{D_b}{m}$ = 1.2 Mbits/s ÷ 2 bits/s/Hz = 600 kHz;
- 8-PSK: $B_{bp,min} = \frac{D_b}{m}$ = 1.2 Mbits/s ÷ 3 bits/s/Hz = 400 kHz;
- 16-QAM: $B_{bp,min} = \frac{D_b}{m}$ = 1.2 Mbits/s ÷ 4 bits/s/Hz = 300 kHz;

• 64-QAM: $B_{bp,min} = \frac{D_b}{m}$ = 1.2 Mbits/s ÷ 6 bits/s/Hz = 200 kHz.

What are the actual bandwidths that are required when using QPSK, 8-PSK, 16-QAM, and 64-QAM and a system with a roll-off factor $\alpha = 0.5$?

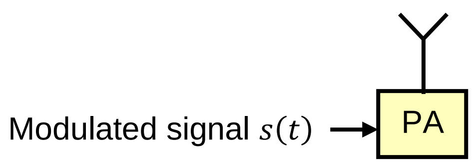
• QPSK: $B_{bp} = (1+\alpha) \cdot \frac{D_b}{m}$ = (1.5 × 1.2 Mbits/s) ÷ 2 bits/s/Hz = 900 kHz;

• 8-PSK: $B_{bp} = (1+\alpha) \cdot \frac{D_b}{m}$ = (1.5 × 1.2 Mbits/s) ÷ 3 bits/s/Hz = 600 kHz;

• 16-QAM: $B_{bp} = (1+\alpha) \cdot \frac{D_b}{m}$ = (1.5 × 1.2 Mbits/s) ÷ 4 bits/s/Hz = 450 kHz;

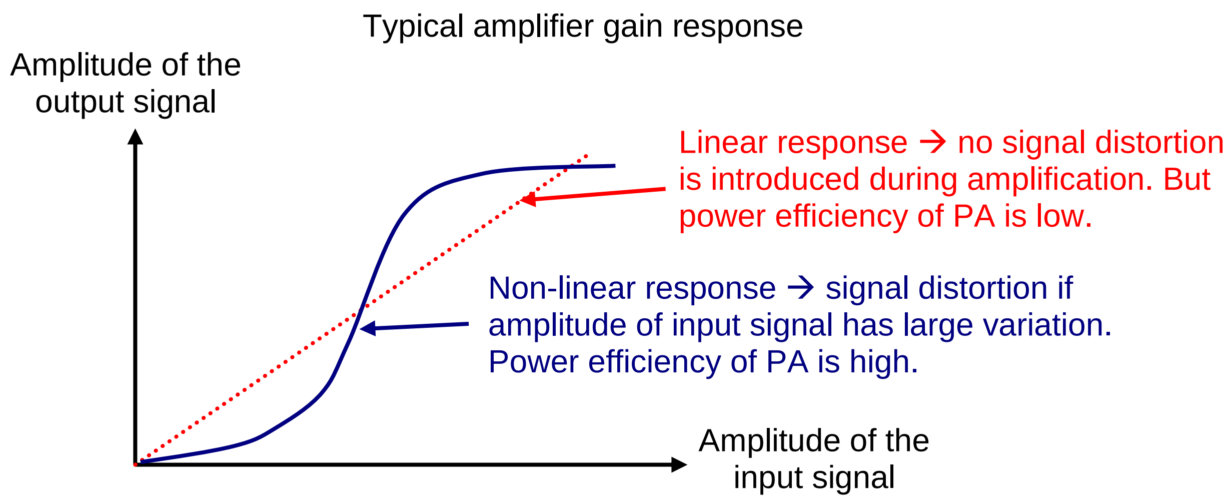
• 64-QAM: $B_{bp} = (1+\alpha) \cdot \frac{D_b}{m}$ = (1.5 × 1.2 Mbits/s) ÷ 6 bits/s/Hz = 300 kHz.

• **Power amplification at the transmitter output**



The power of the modulated signal s(t) needs to be amplified so as to generate an electromagnetic wave that has sufficient power to reach the receiver. The power amplifier (PA) is a high-frequency electronic circuit, typically operating in the GHz frequency range, that consumes a significant proportion of the power supplied to the transmitter. Ideally, this radio-frequency (RF) power amplifier should, as all other types of amplifiers, be as linear as possible in order to introduce no significant distortion on the amplified signal.

On the other hand, in portable devices operating from batteries (e.g., smart phones), the amount of power wasted as heat in the RF amplifier should be minimized. In other words, the RF amplifier should have a power-efficiency close to 100%, meaning that most of the power drawn from the battery should be delivered to the load. Unfortunately, the most power-efficient amplifiers tend to be highly non-linear.



When using non-linear amplifiers, it is preferable to employ modulated signals that have a constant envelope, such as PSK signals, as they will experience minimal distortion when passing through the power amplifier. Remember that, in PSK, the expression of the modulated signal is given by $s(t) = \sum_{k=0}^{+\infty} h(t - kT) \cdot cos(2\pi f_0 t + \Phi_k)$, which shows that the information to be transmitted is contained in the phase of the carrier wave. This information is thus not sensitive to an amplitude distortion of the carrier wave.

On the other hand, ASK and QAM are not ideal modulation schemes for use with non-linear power amplifiers because, in this case, the information is contained in the amplitude of the carrier wave and can thus be affected by any amplitude distortion of this wave. The distortion effect is worsened when the number of possible amplitude levels of the modulated signal is increased.

Students who are interested in knowing more about this topic should take a module on the design of RF/microwave analogue electronic circuits.

Finally, the power amplifier output is connected to the antenna whose task is to radiate the RF signal as an electromagnetic wave into free space.