The Intuitive Guide to

Doing Link Budgets

Charan Langton

Mountcastle Academic

Contents

The Intuitive Guide to Doing Link Budgets

The authors gratefully acknowledge the many websites and their authors, whose work we used to understand and explain the concepts in this book. If we inadvertently left out any credits, please let us know.

The figures in this book were done using Matlab Simulink. Book site for MATLAB code and supplemental material (TBD): www.complextoreal.com/linkbudgetbook
Contact email: charldsp@gmail.com or use comment page on complextoreal.com

The Library of Congress cataloging data has been applied for.

Print ISBN-13: 9780913063262
1. Link budgets 2. Link capacity 3. link analysis analysis 4. Communications system design

Draft review copy, October 2019
Not for general posting or transmission. 1 2 3 4 5 6
Printed in the United States of America

Book cover design by Josep Blanco
Book interior design and Latex typesetting by Patricio Prada
Editing by Rena Tishman, Grima Sharma, Paul McGhee

Chapter 1

Key concepts of communication links

Communication science as a subsystem of the OSI model

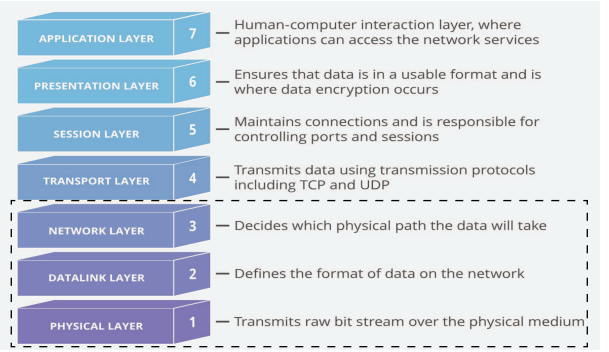


Figure 1.1: Conceptualizing a information network in seven functional layers, called the Open System Interconnection (OSI).

The study of communications science can be organized in a framework provided by the Open System Interconnection (OSI) model of information transfer. The OSI model separates major functions of an information system in *layers*. These layers are organized according to fields of expertise. This facilitates the study and design of the system in manageable units by groups of people with specific knowledge in their particular area. Each layer of the OSI is essentially an independent field and requires a different way of thinking from the other layers. Even the science required for each level is different. The math, the terminology, and the metrics of performance are all specific to each level.

The OSI model is not an official standard, but is instead an abstraction of a functional structure that might be required to build large-scale information networks. Not all of

these functional layers are present as separate organizations in real networks. They are often combined to create a shallower management structure.

For example, the top layer of the information system in the OSI model, the **application** layer, is the final presentation of the information to the end user. You may see this final product as a *movie*, an *email* or a *text message* you receive. This requires an entirely different set of expertise (application programming) than the bottom layer of the OSI, the **physical** layer, which requires expertise in digital/analog hardware design. These two groups of people do not have much in common in their areas of expertise, yet both are needed for you to download a file from the Internet or to watch a program via Dish network. This layering of functions allows people of different expertise to work independently to build an inter-connected system, and to do so in independent groups.

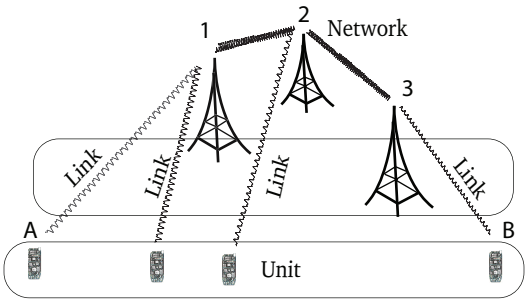


Figure 1.2: The study of network at layers 1 to 3, includes three main functional areas. These are, the unit, the link and the network. A unit is physically limited and its work occurs at baseband, low frequencies and low powers. The connection of units is under the link layer. A link design consists of conversion of baseband information to symbol, modulation and coding. The network then aggregates the many individual links and manages the resources, so that many conversations can take place over the network in an efficient manner.

The study of **communications systems** may be thought of as a sub-system of the OSI model. Its study generally encompasses just the bottom three layers of the OSI model, which are called, **physical, datalink and network**. In keeping with the principle of functional separation between the OSI layers, here too, the three layers of relevance to the study of communication systems, also have little crossover in terms of knowledge and expertise. We will examine the design and analysis of a communication system along these three layers but we will differ slightly from the OSI abstraction. The three layers are:

1. Physical (Unit)
2. Datalink (Link)
3. Network (Network)

In general the field of communications can be conceptualized in three fields of study as shown in Fig. 1.3. (The names in parentheses are what we choose to call them in this treatise.) There is the **physical** unit. We can think of it as a hardware device,

all the way from logic and chip design to an amplifier or a filter. For this layer, the design and analysis puts us in the realm of electromagnetic and circuit theory. If you are working at this level, you will need a good understanding of circuit theory, material properties, analog design and electromagnetism. In this layer, we are designing and assessing circuit boards, and other hardware components such as a filters, amplifiers, antenna etc.

The **link layer** moves up the abstraction level. Now we step out of one hardware unit and make connection with other hardware units. The link layer covers the connection of one hardware unit with an another, usually just a single connection between a sender and a receiver. If we call the units, **transmitter** and a **receiver**, then the link is the system that connects them. The link can be thought of as the highway between these units. This highway can be implemented with a wire or be un-wired. The un-wired connections are called, wi-fi, wireless, rf, IR, Bluetooth, Zigbee etc. In general the signal transmission over this highway takes the form of *waves*, hence the people working on this layer are often called waveform experts. These waves can be either continuous/analog or discrete/digital. Most signals contain both analog and digital processing even though they may be referred to as solely digital systems such as cell phone or satellite communications. In general, we can say that native signals are **digital** but what travels through the highway is **analog**. Layer two is the most mathematically oriented of the communication sciences, while the physical layer also involves a great deal of empirical science.

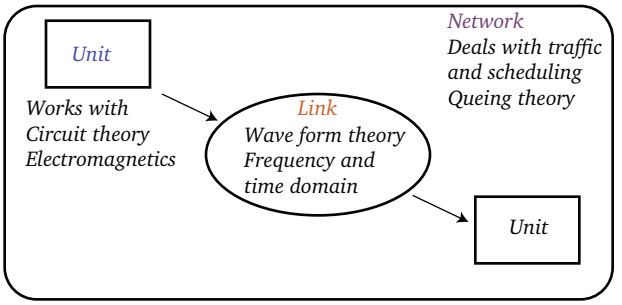


Figure 1.3: The study of communication science includes three main functional areas. These are unit, link and network. A unit is usually a physical component. The connection of units is done under the study of the link layer. A link design consists of conversion of baseband information to symbol, modulation and coding. The network then aggregates the many individual links and manages the resources, so that many conversations can take place over the network in an efficient manner.

The study and design of the link layer, requires understanding waveform theory and the mathematical tools that are used to understand and design it. Fourier analysis is one of these tools. The concept of frequency and time domain is another such tool. Electromagnetic noise theory and non-linear behavior also come into play. How is a signal distorted as it is going from one unit to the next? How do signals interact with each other in various types of channels? As part of a link design, we study subjects

like symbol generation, modulation, channel effects, power required etc. These are in a category called the **channel effects**.

We define a link as a *single connection*. As we move up to multiple links, we come to the concept of a network. This third layer is the integration of various individual communications *links* into a network. The links are dynamic, they can come and go, whereas some applications such as satellite TV, is alive and static all the time. A cell network has many base stations, many users, the issues of how we design a network so that the total capacity is used efficiently, each user is uniquely identified, and does not cause interference to other users, are all issues of network design. These include study and management of **congestion**, **multiple access** and **control**. Here congestion management, resource allocation and queuing theory are the primary drivers as method of examining the performance of the system. Performance metrics such as **throughput** and **delay** are often invoked at the network layer, whereas at link layer, we are concerned with power and distortion, with performance metrics of E_bN_0 and BER.

In this book, we will be looking mostly at the *link* and *network* layers. At the link level, we are concerned with how to convert real human or machine communication to a form suitable for electronic media. This is done through a topic we will term **waveform analysis**. Where at the network level, we might talk about throughput and memory, in link layer, the primary thing of interest is the **signal**. We look at properties of signals, baseband signals, carriers and modulated carriers, we use for communicating information.

For the purposes of identifying the direction of data flow, links are characterized in two ways: **uplink** and **downlink** as shown in Fig. 1.4. Often these are also called **forward or reverse link**. If user A wants to talk to user B, then his transmission to the tower is called the *uplink* and this transmission from the tower to the receiver A is called the *downlink*. If the receiver is located in a far off region, requiring the tower to route the signal to the tower where the receiver is located, this is also part of the uplink but treated differently, since it is a different frequency band, and uses some form of multiple access and multiplexing. Each of these paths are an independent design process.

An uplink from a cell phone will have different modulation, frequency etc. from the downlink. When modulation does not change, the links are usually considered **bent-pipe** such as in most satellite systems. However, if the link is demodulated at an intermediate point, and re-transmitted, it is part of an **on-board processing system** and is functionally an independent link and analyzed separately. New generation of satellite links are of this type, as are all cell phone links.

A complete link analysis is done as part of **Link Budget Analysis**. This is usually done in conjunction with design of the units and is often an interactive process with the waveform analysis at the link level. The issues of congestion and interference at the

network layer also come into play but are done separately and then added into the link budget. We will look at this functionality in Chapter 4.

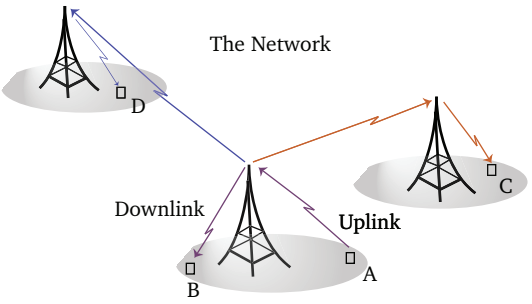


Figure 1.4: A network manages many transient communications links within it. A link is one specific communication characterized as an uplink or a downlink.

What is a link budget

You are planning a vacation. You estimate that you will need $1000 dollars to pay for the hotels, restaurants, food etc.. You start your vacation and watch the money get spent at each stop. When you get home, you pat yourself on the back for a job well done because you still have $50 left in your wallet.

We do something similar with communication links, called creating a link budget. The traveler is the signal and instead of dollars it starts out with "power". It spends its power (or attenuates) as it travels, be it wired or wireless.

Just as you can use a credit card along the way for extra infusion of money, the signal attenuates at each hop and then gets extra power infusion along the way from antennas and intermediate amplifiers. The designer hopes that the signal will complete its trip with just enough power to be decoded at the receiver with the desired signal quality.

In our example, we started our trip with $1000 because we wanted a budget vacation. But what if our goal was a first-class vacation with stays at five-star hotels, best shows and travel by QE2? A $1000 budget would not be enough and possibly we will need instead $5000. The quality of the trip *desired* determines how much money we need to take along and how much intermediate infusions we will need on the way.

For modern systems, the quality is measured by bit error rate (BER) of the final, received data, because the native data is itself discrete. Lower BER, the better, but not all signals need really low error rates. If we want our signal to have a low BER, we would start it out with higher power and then make sure that along the way it has enough power available at every stop to maintain this BER.

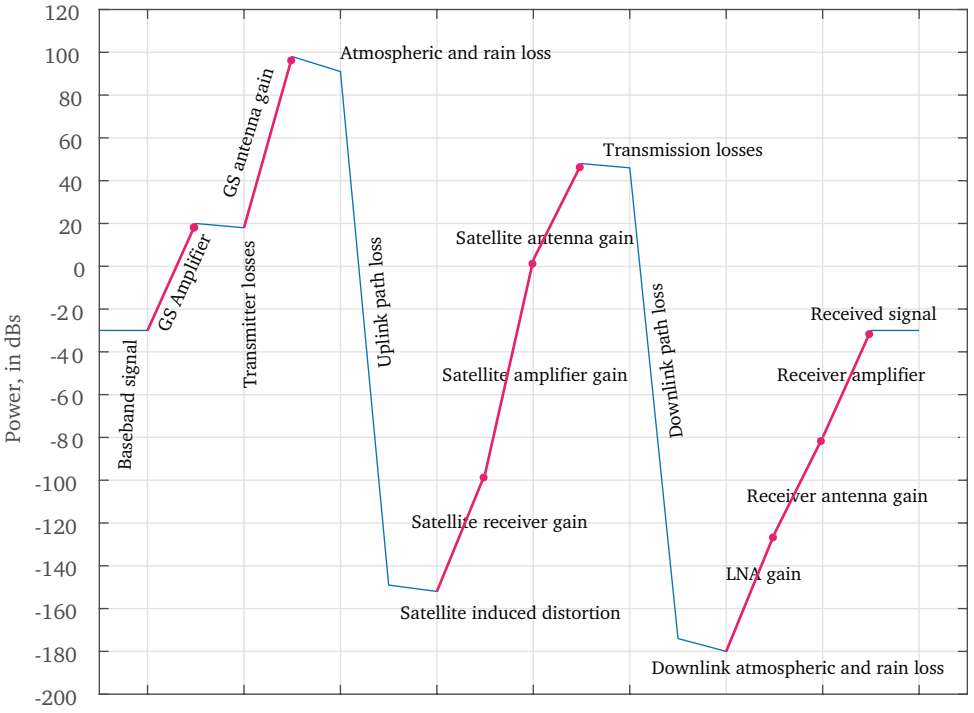


Figure 1.5: The red lines show the injection of power by the amplifiers and the antennas, and the blue lines show the attenuation at various points in a line of sight satellite uplink and downlink. We need to have the signal arrive at a given threshold of power. Any less and, the receiver will not be able to decode it. Any more would be wasted. This process of arriving at the threshold required power, while considering all intermediate losses and gains, is called doing the link budget.

In Fig. 1.5 we see how a typical satellite signal is gaining and losing power, finally arriving at the destination with just enough power to be decoded at the desired BER. This is essentially what we are doing in a link budget, determining each of these additions and subtractions of power and attenuation so that at the final receiver, the signal has enough EbN0 to be decoded correctly for a given quality specification.

What we need to know about signals

When we talk about communications, we are talking about transfer of some finite amount of information from one source (usually a transmitter) to a sink (receiver). Whether right up close or to far-off destinations, the electronic communications use *signals* of some sort. In a broad sense, a signal is a packet of electromagnetic energy, the transfer of which is conceptualized as a wave or a sequence of discrete information transmitted as pulses. This signal describes a relationship of one parameter such as an amplitude, to an independent parameter, such as time. This relationship can be discrete or continuous.

Which type of signal is used for communication, depends on how far the signal has to travel and how the information is *coded* into waves, and the **medium** or its generic name, the **channel**, through which it must travel. On the receiving side, the signal is decoded and the *information* is recovered. There are various distinct steps a signal goes through from the source to the sink.

We distinguish the word **channel** from **link**. The link is transnational connection from the sender to another. The signal will travel in a link through various types of channels. A channel is physical path, offering a particular pattern of signal degradation. Some channels are linear, and hence the loss is linear, say based on path length. For example, when the signal is going from the sky to the ground, that is a **line of sight channel** with its own specific power-in vs. power out behavior. When the same signal is going through an amplifier, that is also a channel, often called a **non-linear channel**. A link, then is said to be impacted by all the channels a signal must travel through.

We will now discuss some important aspects of signals as they relate to doing link budgets.

The modulated signal

In general sense, the process of communications consists of sending some information from place *One* to place *Two*. Assume that we want to send a voice message. Of course, we need something other than yelling to deliver our message. But, even yelling requires us to amplify the message, albeit using our lungs and then using air to push the sound to the receiver. And if there are walls between us and the receiver, well, you can see how quickly that idea of transmitting information in its **native** or **baseband form** falls apart. Not enough lung power power to get over the obstructions!

From this basic concept of message transmission, we can see that we need some sort of way to transfer information (our own voice) to someone not within earshot. Our un-amplified voice is just not enough to overcome the barriers. We need some other way to *carry* our information. For this we use a signal called, the **carrier**. You may think of it as a sound wave, a cart, a car, a truck, or any medium that carries the payload of information you want to transfer.

In communications, the carrier is a very simple concept. We all learned it in ninth grade. It is merely a pure sine (or a cosine) wave. Thats it. It is an analog signal generated usually by a fairly accurate oscillator. The carrier frequency is usually dictated to us based on our application. A hand-held phone, for example uses a carrier of frequency 900 MHz The carrier power is a function of the carrier amplitude, generally related by its square. It can be huge such as the UK electric power signal level of 220V or smaller such as ours in the U.S. at 110V. Those are power signals, whereas communications signal levels are usually much much smaller than these numbers, often in microvolts.

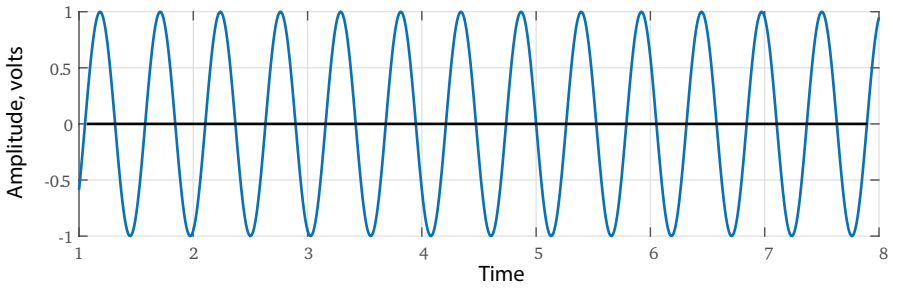


Figure 1.6: This is a carrier. A carrier is just a sinusoid of a selected frequency with an arbitrary starting phase. Its power is related to its amplitude and the time scale is related to its frequency.

A carrier can be made to convey information by changing or **modulating** any of its three parameters, the frequency, amplitude or its phase. When the phase is manipulated in response to information, we call that **phase modulation**. By the same idea, we have frequency and amplitude modulation. The carrier, by itself, since it is a single frequency signal, has *zero* bandwidth. But when we modify the carrier, with information, it becomes "loaded" with information. The "loaded carrier" is still analog and is called the **modulated signal** and is no longer a pure sinusoid. It is a sinusoid that is abruptly changing some aspect of itself in response to the data.

In Fig. 1.8 (a), we see a sinusoid carrier and in (b), the same carrier after it has been phase-modulated. In this case, the phase modulation has changed the phase of the sinusoid abruptly at period boundaries in response to the data, shown in the red plot line in (a). Other modulations may change the sinusoid, or the carrier wave in different ways such as changing its amplitude or its frequency, but it remains an analog signal whereas the modulating data itself may have been discrete. The creation of the modulated carrier hence forms the boundary where digital data becomes analog, or we go from digital to analog communications.

In satellite communications and in fact even in the cellular networks, QPSK or phase modulation is the main choice. It is theoretically capable to carrying 2 bits per symbol. Higher order modulation can increase on this capacity. More details on this topic are covered in chapter 2.

QPSK and many of the phase shift keying modulations are called digital modulation. This is because they take digital data and *modulate* or *change* an analog signal. Hence one side is digital and the other analog. The digital part resides only inside the hardware. The radio frequency, the carrier carrying the information is always analog. When we talk about digital wi-fi communications, it is only *partly* digital, the rf part being always analog. There are some exceptions to this but they are for close-range communications only.

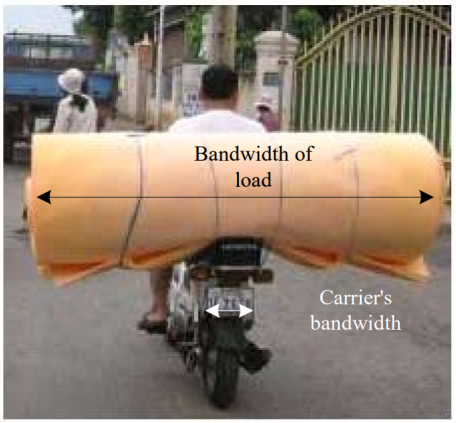


Figure 1.7: Our motor cycle guy is the carrier and what he is asked to carry is the information load.

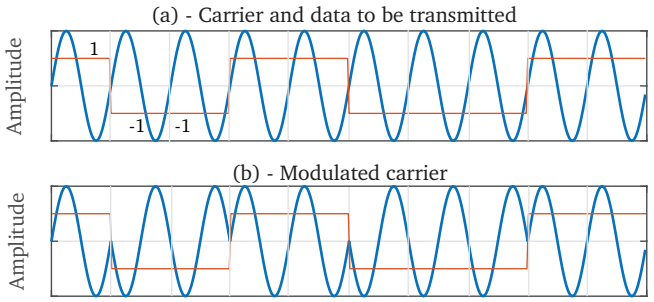


Figure 1.8: (a) A carrier (in blue) and the data sequence, (in red: 1 -1 -1 1 1 -1 -1 -1 1 1) to be transmitted. (b) The carrier is modulated with data. In that it changes its phase by 180° *at the end of each cycle, in response to the data. The signal in (b) is considered an analog signal although containing a lot of abrupt transitions.*

Higher order modulation

In a QPSK signal, both the I and the Q signals are represented by just two levels. By increasing the signal levels, we can represent more symbols. 8-PSK uses a three level signal to create eight symbol constellation. Each symbol represents three bits instead of two. All constellations where the symbols are all located on a circle are called M-PSK modulation. Both 8-PSK and 16-PSK modulation, when un-shaped, are true constant-envelope signals. However, nearly all wideband signals require some shaping so no signal, save a sinusoid is truly constant-envelope.

The category of signals called M-APSK allow varying both the amplitude and the phase, so they are a hybrid. These are not constant-envelope signal and hence, an issue with amplifiers, such as traveling wave tube amplifiers, TWTA. However, they offer better throughput and are called higher-order modulation. With better amplifiers, with

pre-distortion and other techniques, non-linear effects can be controlled making these modulation useful. We discuss these in more detail in chapter 2.

Pulse shaping

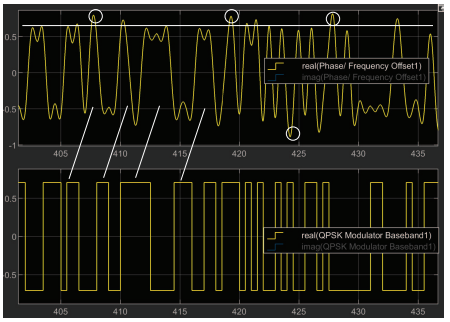
In section above, we see that the data is in bits and the phase modulation leads to large phase changes at the bounderies. The Fourier transform of a square pulse tells us that the required bandwidth for such a shape is very large. The way to get around this is to pulse shape the signal, using, most commonly, the ***raised cosine shaping***.

The raised cosine shape essentially replaces the square pulse, as we see in the figure above. The main parameter of raised cosine shaping is a parameter called the roll off factor, α. Most satellite signals are shaped using a alpha value of 0.15 to 0.25. This shaping is actually split in two parts, hence is instead called ***root raised cosine***. The transmit signal is multiplied by a RRC shape and then the signal is again multiplied with the same response at the receive side, which acts as a matched filter. Hence the signal is shaped to reduce the sharp transitions, however this leads to reduction in the symbol rate carrying capacity of the link. If a bandwidth B can carry two symbols per Hz for an ideal QPSK signal, then a shaped signal can only carry $2/(1+\alpha)$ symbols. However, this shaping is essential because a pure QPSK signal of square pulses will require far too much bandwidth. We discuss this topic in Chapter 2 in more detail.

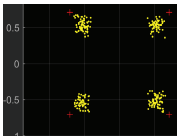
A small α is preferred as it is more more bandwidth efficient, however this increases the signal excursion, it makes a constant-envelope QPSK signal into one that has some envelope variation. This results in some powerloss and greater degradation due to the amplifier amplitude non-linearity. Fig. 1.9 shows how a root-raised cosine signal looks like in time domain.

Need for amplification

All electromagnetic signals attenuate as they travel. Receivers, to do their job properly, must receive a signal at some minimum level of power. Hence, the most important goal in the design of a link is to transmit it with enough power so that a receiver will be able to sense it. We can we do this in two ways, one, by an active amplifier, and the other, by an antenna. Most links use both to provide the total amplification. A signal is first amplified with a physical amplifier and then again via an antenna. Both of these methods are of course *amplifiers*, however, the term **amplifier** is usually reserved for the physical devise providing **active** amplification and the antenna, although also an amplifier, is just the *antenna*. It is said to be a **passive** amplifier. The use of an antenna is a significant part of the total amplification, often providing a much larger portion of it.



(a) RRC shaped QPSK signal above with actual data below.
Note excursion above average for certain highlighted symbols.



(b) Constellation diagram for both a shaped signal and the reference QPSK signal. in red dots.

Figure 1.9: in (a) top figure, we see a shaped RRC signal ($\alpha = 0.24$)*, whereas in lower trace, we see the QPSK raw data. The delay is due to the shaping filter. In (b) we see the side view of this signal, called the constellation diagram, showing the excursion of the symbols, as compared to the red dots of the data which is static. Low roll-off factor results in higher excursion for this shape. Note that some power loss occurs when the signal goes through the RRC filter, although it is quite small.*

Besides power amplification, antennas also provide the ability to control the *directionality* of the power transmission and *polarization*. Polarization can be utilized to improve throughput.

1.3.4.1 How antennas work

All physical objects, not at absolute zero, emit electromagnetic energy. Most of this is unintentional and results in what is called the **universal cosmic radiation**. Physicist characterize this cosmic or thermal noise in a mathematical sense as waves. This noise is characterized as a composite signal consisting of many frequencies, all the way from 0 Hz to Gamma rays of very very large frequencies. So fundamental is this phenomena that we can think of all physical devices, including our own bodies as passive noise generators, and indeed all of us, a form of an antenna. The energy being broadcasted by these devices/antennas is a function of its **temperature**, which can be thought of as a measure of the energy of the device.

Unless some measure is taken, an electromagnetic source, passive or active, will spit out energy in *all* directions, hence the term *omni-directional antenna* is used to describe unmanaged transmittal of energy. Such source/antenna is transmitting energy in all directions equally, at once, all its power is hence expended in a sphere radiating out from it. The power reaching a point, a certain distance away from the source, will be quite small because the total power is spread over the surface of a sphere. The signal from such a source, hence disperses and falls off in power with the square of the distance, being also a function of the frequency chosen.

The ratio of the received power to transmitted power (called the **gain**, or relative increase in power) for a clear line of sight omni-directional antenna, called an **omni antenna**, is given by:

$$\text{Gain of an omni antenna} = \frac{(4\pi r)^2}{\lambda^2} \tag{1.1}$$

Here λ is the carrier wavelength and r is the distance from the source to the sink. Clearly since λ is inversely proportional to the frequency, ($\lambda = c/f$) a higher frequency will result in a larger received power. All this energy is spreading out in form of a sphere and only a very small amount hits the sink as represented by the little patch in Fig. 1.10. If all the energy were to be concentrated in a particular direction, say in a narrow beam of α radians, the power received at the same little patch will be a function of $2r^2/\alpha$ instead of being a function of πr^2. This is called the **gain** of a **directional antenna**. These gains can be huge, hence we nearly always use **directional antennas**. All communications links are power-limited, i.e. we want to use small physical amplifiers, and using directional antennas is one way to reduce power requirements. However if the location of the receiver is not known or changes rapidly, then omni antennas become necessary. In addition to parabolic reflectors and omni antennas, communication system often use phased array antennas which can create a variety of coverage patterns.

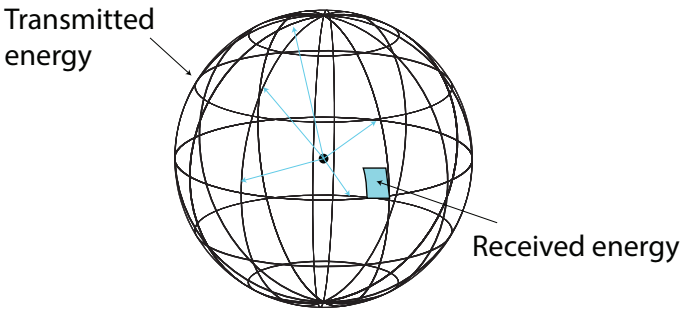


Figure 1.10: An omni antenna transmits in all directions, but a receiver only captures a very small part of that energy depending on its surface area or size of the receiving antenna. Hence omni transmitting antennas waste a lot of energy if there is only one immobile receiver located in a particular spot.

The directional antenna gain is a function of the square of the frequency; the higher the frequency, the higher the gain. It is also a function of the diameter of the reflector. So to transmit a modulated carrier signal of frequency 1 kHz vs. 1 GHz, would require an antenna that would have to be 10^6 times larger in area. So even doubling, tripling the antenna size won't make up for the advantage offered by the use of a higher frequency. Hence high frequencies are advantageous in that they provide higher antenna gains but at the same time also increase path-loss. The use of a high frequency carrier signals

allows us to use much smaller antennas. In satellites, a Ku-band signal requires only a 0.3 meter dish vs. C-band which requires a dish of app. 2 meters. This is also one reason why we keep pushing for the use of higher frequencies. We can reduce the antenna size!

Directional antenna gains have a pattern in that they have a main beam of some particular solid angle as well as smaller side beams which are called **side lobes**. The side lobes are, of course, undesired, and waste transmit energy. We see an example of the gain pattern of a reflector antenna in Fig. 1.11. Here we see four smaller side lobes, albeit 20 dB below the main lobe towards directions that end up becoming interference for others.

Not all of the *gain* from antenna, as we see in Fig. 1.11(a) reaches the intended source, presumed to be located at the boresight. The ratio of the power in the mainbeam vs. the total gain is called the **directivity** of the antenna. An antenna type, called the horn antenna provides the best directivity and a version of it called, the horn-fed reflector is used in satellite communications, where high directivity is paramount. The total gain and the directivity of such antennas are usually quite similar, with only a small loss. The directivity, hence is the useful power, and is the quantity we use in doing link budgets.

Another parameter, called the **beamwidth** is defined in both the horizontal and vertical planes. Beamwidth is the angular separation between the half power points (3 dB points) in the radiation pattern of the antenna in any plane. Another important parameter is the **front-to-back power ratio**. This specifies how much power is leaked in the unwanted back direction. Hence, a good antenna will have a ratio of 20 dB or less. The level of the side lobes is extremely important as well. The side lobe levels cause adjacent and co-channel interference and usually the goal is to keep them around 30 dB below the main lobe.

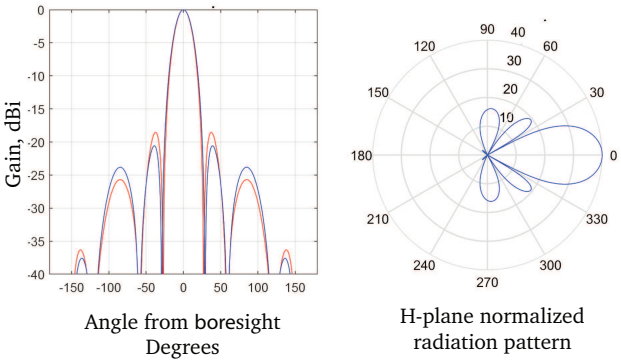


Figure 1.11: (a) An antenna radiation pattern as a function of the angle from the bore sight. Maximum power is at the bore sight. There are side lobes of significant levels at app. 40° *on each side. We see this same behavior in looking at the H-pattern in a polar plot. Some energy is leaking in these side lobes.*

Signal polarization

It turns out that we can arrange two real (modulated) signals at 90◦ to each other as shown in Fig. 1.12. Here the two modulated carriers are orthogonal to each other and we can pretty quickly see that if we orient our antenna along one or the other of these directions, we can extract these two signals independently. If perfectly aligned, these two signals don't interfere with each other. We can think of them as independent signals although both are on exact same frequency. This idea, called **polarization** allows us to isolate the signals from each other, as well as, allowing us to **reuse the assigned frequency**. Satellite nearly always use this concept to double the data rate capacity.

The vertical direction is defined as being vertical to the surface of the earth and horizontal, similarly as being parallel to the surface of the earth. When a signal is going at a angle to the earth's surface such as a satellite signal, these definitions do not make a lot of sense. But for many terrestrial signals, they work well enough and are used conventionally. Specific types of antennas are used for each type of polarization. Its obvious that the plane of the antenna, or its bore sight, must be aligned with the transmitted plane for the largest gain. The gain deteriorates rapidly away from the bore sight.

Horizontal polarization
Most TV antennas in the US are designed to receive horizontally polarized signals. Here the signal is oriented in the vertical direction such that its E field is located in the horizontal direction.

Vertical polarization
A short dipole antenna (which is basically a center-fed open wire) is an example of an antenna with a vertical field and which best receives signals with vertical polarization. Most of the mobile antennas (such as on cars), as well as wi-fi links, tend to use this polarization. This keeps the interference caused by the TV signals in same frequency bands, to a minimum.

Slant polarization Here the E field is at a angle (often at 45◦). The receiving antenna must be aligned to the slant angle for proper reception.

Circular and elliptical polarization A circular polarized signal means that the energy is being radiated in both vertical and horizontal planes at once. Imagine the energy going out in a cork-screw fashion. When the power in both the horizontal and vertical directions is equal, we get a circular polarization. When there is an imbalance, the transmitted wave appears to be elliptical and this is called **elliptical polarization**. The ratio between the minor and major axis of the polarization ellipse is described as the **axial ratio**, often given in dBs. A circularly polarized antenna will have an axial ratio of 0 dB and a elliptically polarized antenna will have an axial ratio around 1.7 dB, or a 3 to 2 power distribution. Satellite antennas are nearly always elliptical. The reason

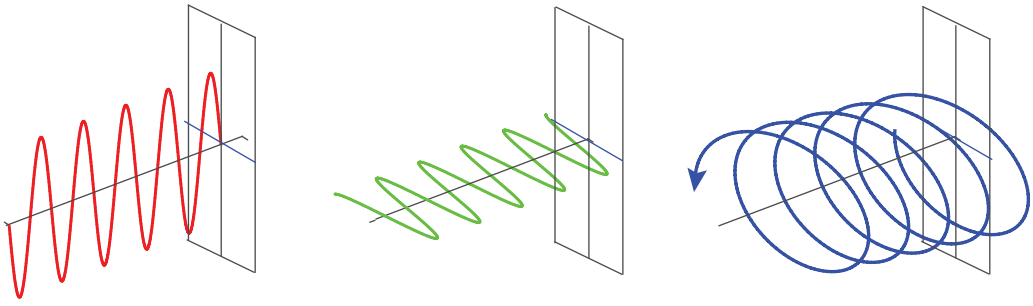


Figure 1.12: (a) The signal is aligned with the vertical electrical field of the antenna. (b The signal is aligned with the horizontal plane. (c) Energy is transmitted and received in both planes simultaneously resulting in a rotating vector. If moving in clockwise, it is RHS sense and hence rotational sense must align with the receiving antenna.)

is that, an antenna can be pointed across large spans of the sky such that its main lobe is always pointed directly at the satellite. Note that a circularly polarized antenna can indeed receive linearly polarized signals but the power will be 3 dB or more smaller.

Multibeam frequency reuse

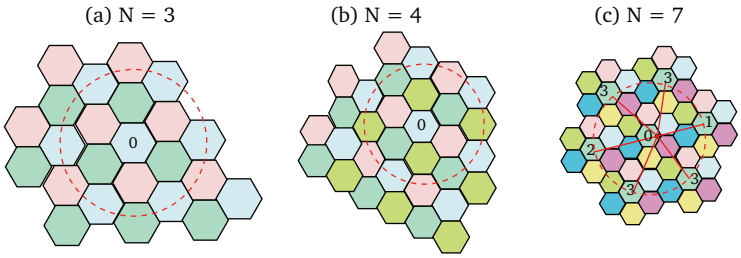


Figure 1.13: (a) A 3-cell configuration, (b) a 4-cell configuration and a typical one with 7-cell reuse. Since same total bandwidth is divided by N, the cell sizes usually get smaller with increasing N. In each case, a target spot (marked 0) is surround by six nearest interferes, of same color cells, on the circle in this figure.

For regions with differing data needs, **multibeam systems** are better suited, particularly for communication satellites. The coverage area of a satellite is divided into a cluster of spot beams that share the bandwidth of the system. The total frequency band is spatially partitioned in a given cluser typically of 3, or 7 spot beams as shown in Fig. 1.13. This pattern is then replicated to cover the non-circular geographic regions, such as Europe, US, etc. with a very large number of spot beams. The cells using the same frequencies, are some distance away from each other, but due to the antenna spill-over, or the side lobe performance, will get some part of the signal transmitted from the satellite, intended for the adjacent same-frequency cells. These types of architectures are also called **cellular systems**. In principal, it involves the re-use of the same frequency, with interference protection provided by the physical distances between the same-frequency

cells.

$$N = i^2 + j^2 + ij \tag{1.2}$$

Due to geometrical limitations, it turns out that the tessellation of cells, so that the cells of the same frequency do not touch each other, come in only a few unique configurations, as given by Eq. 1.2. The symbols, i and j are integers (including zero). For $i = 1, j = 1$, we get $N = 3$, for $i = 2$ and $j = 0$, we get $N = 4$ and for $i = 2$ and $j = 1$, we get, $N = 7$. Larger configurations are possible but not used commonly. The most common of these are based on reuse factor, N of 3, 4 and 7. N represents the number of cells in each cluster with the same center frequency, within the total allocated bandwidth. Since no cells of the same frequency touch each other, the received signal interference among them, due to the spatial separation is minimized, but not eliminated and must be accounted for in a link budget.

For any N, we find that the interferers are aligned in a circle around the target beam. There are various *rings* or *tiers* located around the center of the target beam. Each tier interferes with the target beam, with the closest tier adding the most significant amount of interference. In Fig.1.13 (c), we see that the target beam at station 0 has six foes in the first tier. Assume the user is located at the beam edge, the most disadvantageous spot. The beam to cell 1 is the closest, at app. a distance of $(D - R)$, where D is the distance to the tier, and R is the radius of the cell, another spot beam at 2 that is the farthest at $(D + R)$ and four other cells, numbered 3s, all about the same distance from it. These distances can be calculated fairly easily from the hexagon geometry.

We can calculate the **carrier to noise ratio**, C/I based on the contribution of just the first tier beams of the same frequency, by EQ. 1.3. Here n represents the attenuation factor for the side lobes of the spot beam antenna at the satellite, because that is the level of signals that enters into the target beam. The C/I ratio, also called the **co-channel interference, CIR** for this simple case is given by

$$CIR = \frac{1}{[(\frac{D}{R} - 1)^{-n} + (\frac{D}{R} + 1)^{-n} + 4(\frac{D}{R})^{-n}]} \tag{1.3}$$

An approximation for the value of $\frac{D}{R}$ is given in Ref.1, [Schwartz] is $\sqrt{3N}$, with N the reuse factor. The value of n depends on how fast the antenna side lobes are falling off in the direction of the interferer or at approximately 60∘ from the main lobe. Take for example the antenna pattern shown in Fig. 1.11. The first lobe is approximate 20 dB below the main lobe. However at app 60∘, the side lobe level is quite low, on the order of about -35 dB. Hence the exponent for this case will be 3.5. This is the target exponent for most satellite designs with typical value between 3 and 4. Higher side lobes lead to smaller C/I, which will swamp the link C/N. This parameter tends to make a large contribution to the link in terms of degradation.

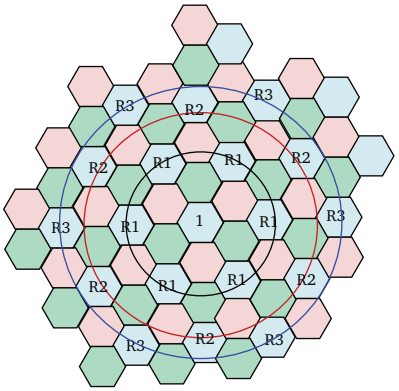


Figure 1.14: A 3-cell system. The interferers are aligned in rings around the target cell. Not all rings are the same distance from the target cell, but vary. Note that ring 3 is closer to ring 2 than than is ring 1 to the target.

Tiers farther out will add to this number, but depending on the geometry, since not all spots may be the same size, nor may have uniformly decreasing side lobes, we get an uneven contribution from farther out tiers, where some tiers may even add a higher interference than a closer one. Generally though the overall number is not likely to change much more than about a dB or so as result of the far-tiers. In the table below we give the values calculated using the Eq. 1.3.

Table 1.1: App. C/I calculations

N	n	D/R	**CIR, dB**
3	3	3.00	7.0
3	4	3.00	11.8
4	3	3.46	8.8
4	4	3.46	14.3
7	3	4.58	12.4
7	4	4.58	19.1

This is what happens to a spot beam receiver since the radiation is falling down on it from the satellite, but what about its converse at the satellite receiver which is also receiving signals from all these spots, and many at the same frequency. This is managed by the phased array antennas which can be tuned, not just to a particular frequency, but also to a certain direction. The received signal from an interferer, although it may be at the same frequency, is by the same principal as for the ground beams, at a side lobe of the transmitter and hence is already quite low in amplitude. These numbers, in form of C/I and the **directivity** figures are made available by the antenna design specialists to the person doing the link budgets. The calculation of these numbers, is usually not within the purview of the system designer. For high-level link budgets, general estimates as shown in Table 1.1 can be used as a starting point.

The question of bandwidth

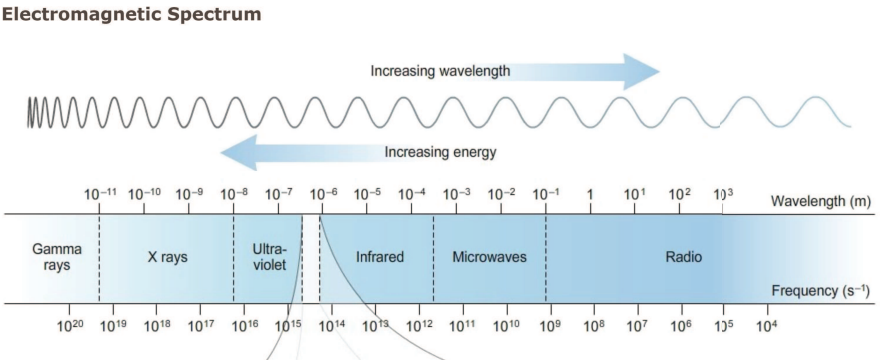


Figure 1.15: The electromagnetic spectrum covers all frequencies, from zero to ∞. But only the range up to 10^{14} is usable at the present time. A majority of our wi-fi communications takes place just in the microwave region.

Think of the electromagnetic spectrum as a highway of many lanes. Although a very wide highway, for practical purposes, its width is finite. The part we can use at the present time is from baseband (near the zero frequency) to the optical range at the other end. Radio communications such as wi-fi type of communications, including satellite communications take place mostly in the microwave region shown in the middle area of Fig. 1.15.

Just as the car-pool lanes in a highway, electromagnetic lanes are often set aside for specific uses. International bodies have been setup to manage the spectrum and assign these lanes. These bodies assign certain widths of this space to each user. For example, a satellite may be assigned a 2 GHz wide lane, with a center frequency in the Ku-band (12 to 18 GHz) The width of the lane, called the bandwidth, is a range of frequencies one is allocated on the electromagnetic highway, specified in Hz.

Bandwidth can be imagined as the span of a signal's frequency content, sort of as the *fatness* of a signal. It is the main currency by which links are valued. Bandwidths determine how much data can be transmitted in one second. Note that the bandwidth of a carrier signal is zero, because a carrier is of course composed only of a single frequency. A carrier signal has a center frequency but it has *zero bandwidth*, whereas an information signal to convey information, needs to contain many different frequencies and it is this span of the frequency content that is called the **data bandwidth**.

If a data signal is modulated on to a carrier, what is the bandwidth of the modulated signal? The modulated signal takes on the bandwidth of the information signal it is carrying, like the guy on the motorcycle in Fig. 1.7. He is the modulated signal and his **bandwidth** went from near zero, without the load, to the size of the mattress which is his "information" signal.

The bandwidth of a modulated signal, however is also a function of the modulation scheme. Nyquist tells us that one needs a minimum of 1 Hz (at baseband) to transmit one symbol of information. A **symbol** in its simplest sense can be a single bit. In this case, we need at least one Hz of bandwidth for every bit. But higher order modulations can use a symbol to mean more than 1 bit to create bandwidth-efficient modulation techniques. (More about these in Chapter 3.)

Bandwidth, in common usage is a ambiguous term and takes on many different interpretations. Here we briefly look at some of the important ways the parameter *bandwidth* of a signal is used.

Bandwidth as specified by regulatory bodies - The role of Government bodies is to assign a span of frequencies to each type of system. Take for example, a LTE system allocation as shown in Fig. 1.16. It is assigned two separate bands, one for out-bound and the other for return. The LTE channel 1, the downlink, has a center frequency of 1950 MHz and the uplink is located some ways away (a common practice) at a center frequency of 2140 MHz. Each is assigned a 60 MHz of bandwidth. Within this bandwidth, many users of a given system have to fit and this job of further allocation within this allowed band falls on the system licensee. In addition to specifying exactly the frequency band, the regulatory bodies also specify how much energy (specified in terms of flux density) you can transmit in your assigned band and also how much you can safely leak into the adjacent bands. Typical names of these official bandwidth are **assigned bandwidth** and **occupied bandwidth**.

Depending on the application, the assigned bandwidths for FCC, ITU etc. must follow certain rules. Certain absolute limits are assigned and must meet the specified flux density inside the assigned band and the emissions outside of the band. Signal design and link analysis requires meeting both limitations of how much power you can transmit in-band and, the spectral spreading to control out-of-band (OOB) leakage.

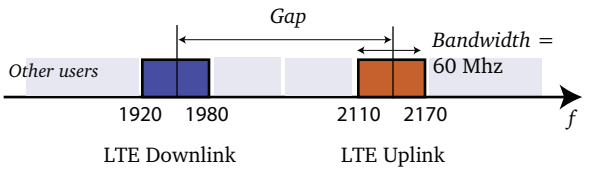


Figure 1.16: The assignment of LTE channels, their center frequencies for both uplink and downlink and the bandwidths are controlled by regulatory bodies. LTE channel 1 is assigned the frequencies and bandwidths as shown here.

3-dB Bandwidth - This is also called the **Half Power Bandwidth**. Between some specified frequencies, f_1 and f_2, the power of a signal drops to a -3 dB level from the maximum in-band power. Filters are often specified using this definition of bandwidth. In figure Fig. 1.17 we see this form of bandwidth definition. The passband, the quantity $(f_1 - f_2)$ is between these two frequencies. In addition to the 3 dB reference points

as offsets from the center frequency, there are often 10 dB and 30 dB points to define similar cut-off frequencies.

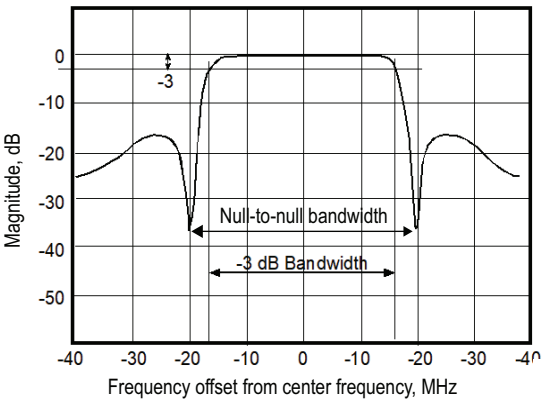


Figure 1.17: The bandwidth of a filter is often defined as the 3 dB points of its passband.

Absolute bandwidth - Here all of the power of the signal is confined in a finite frequency range from f_1 and f_2. This is a rarely specified form of bandwidth, but we do often talk about signals requiring absolute bandwidth. For example, a signal consisting of square waves, its absolute bandwidth is said to be infinite. In practice, we can rarely create a signal that has a fully contained bandwidth.

Null-to-null bandwidth - We measure this bandwidth from first null on the high side to the first null on the low side. For baseband systems, the low side would be zero. For a square wave signal, this bandwidth is 2 times the symbol rate. The spectrum of a square wave consists of a sinc-squared form and the first null occurs at twice the symbol rate on each side. For real signals, nulls are, of course, never fully null (or achieve zero), so some low point maybe thought of as a null point as shown in Fig. 1.17.

99% Power bandwidth - Another way bandwidth is defined is by stating the amount of power that must be contained inside a given band. The common points specified ar 99% or 99.5% power. For a square wave signal, this bandwidth is quite large. For 20 dB level, the bandwidth is app. 20 times the symbol rate. For a 50-dB bandwidth, the bandwidth extends to nearly 200 times the symbol rate. How wide this bandwidth is a function of the signal shape, or its modulation scheme.

Equivalent Noise Bandwidth (EqNb)) - This is a definition used in analysis and tends to be confusing to most. Here we imagine a brickwall filter that contains the same amount of energy as a real filter. EqNb is defined as the bandwidth of the brickwall filter which contains the same integrated noise power as that of the real filter. The height of this filter is same as the maximum level of the real filter. We write the equivalent noise

bandwidth (EqNb) as

$$E_E qNb = \int_0^\infty |H(j\omega)^2| d\omega = H_{max}^2 2\pi\Delta f \tag{1.4}$$

$$\Delta f = \frac{1}{2\pi} \int_0^\infty \frac{|H(j\omega)^2|}{H_{max}^2} d\omega \tag{1.5}$$

EqNb is calculated by integrating the PSD over each band (which means multiplying the PSD with the bin size), summing all the products and then dividing by the maximum PSD. (See sample Excel worksheet at website used to calculate the EqNb in Fig. 1.18.)

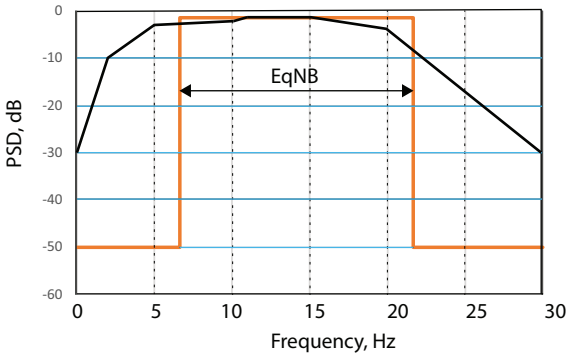


Figure 1.18: The equivalent noise bandwidth (EqNb) of a signal is the bandwidth required for a brickwall filter to capture all the energy of the real filter. Its peak is same as the signal PSD.

Analog or digital

Most signals in nature are analog. Examples of analog signals are sound, noise, light, heat, and electronic communication signals going through air (or space). A very important type of analog signal is the one we use to transmit information, the *carrier*. Once modulated, the signal takes on digital or discrete data but, retains its analog nature. The data or information up to the modulator can either be digital or analog. But once that data is modulated, the signal is considered analog. The so called *digital communications* then is referring to the processing of the data prior to transmission and after the reception at the receiver.

1.3.8.1 Discrete vs. digital signals

In general terms, a **discrete signal** is *continuous in amplitude* but is *discrete in time*. This means that it can have any value whatsoever for its amplitude but is defined or measured only at *uniform* time intervals. Hence, the term *discrete* applies to the time dimension and not to the *amplitude*.

A discrete signal is often confused with the term **digital signal**. Although in common language they are thought of as the same thing, a digital signal is a special type of discrete signal. Like any discrete signal, it is defined only at specific time intervals, but its amplitude is constrained to specific values. There are binary digital signals where the amplitude is limited to only two values, $\{+1, -1\}$ or $\{0, 1\}$. A M-level signal can take on just one of 2^M preset amplitudes only. Hence, a *digital* signal is a specific type of discrete signal with *constrained amplitudes*.

The noise, once added to an analog signal is indistinguishable from the signal itself. But in digital processing, we demodulate the signal to a specific threshold value and hence noise is mostly eliminated. The presence of noise in an analog signal distorts the signal amplitude, whereas in digital communications, the problem shows up in the form of bit errors. The qualitative effect of bit errors is the same as noise in an analog sense. Coding and decoding, a concept applicable only to discrete signals, can reduce the bit error level to near nothing, whereas integrated noise in an analog signal is impossible to eradicate.

A satellite signal, hence is digital up to the point of modulation, is analog after modulation, and maintains this nature through amplification, and its travel through the sky to the point of reception where it is demodulated. Once demodulated, we can talk about digital signal metrics such as Eb/N0 and bit error rates. The metric for the analog signal is the term signal to noise ratio, given either as SNR or CNR. In satellite communications, this is the term C/N. Once the signal is decoded, then depending on the codes used, the analog signal metric C/N is converted to the digital metric, EbN0 to be compared to the required specification of the receiver.

Constant-envelope signal

What makes a signal constant-envelope? A constant envelope signal is one where the signal peak to average power ratio (PAPR) is equal to 1, or 0 dB.

Let's take a sine wave of amplitude of 1.0, with a RMS value of 0.707. It varies in amplitude from 1.0 to -1.0. The mean power of a such a signal is the square of its RMS value, or 1.0. The maximum amplitude is also 1.0, so the ratio of the maximum power (square of the maximum amplitude) to the mean power is 1, and in dBs, that is zero dB. This ratio is called the **Peak to average power ratio** or PAPR. The sine wave then is the gold standard of a constant-envelope signal.

As long as this parameter for a chosen signal type is near 1.0 (or zero dB), signal distortion by a high power amplifier (HPA) or a traveling wave tyube amplifier (TWTA) is negligible. However, as this ratio increases, which it does when we have a shaped signal or we use modulation, different parts of the signal come out amplified by different scale factors, hence changing not just the scaling but also the shape of the signal.

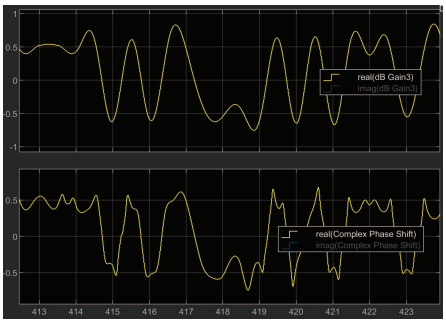


Figure 1.19: We see in this figure the amount of distortion even a constant envelope signal can suffer when through a higly non-linear device such as a TWT. Top trace is the signal prior to the amplifier and the lower trace is the output signal.

A pure QPSK signal seen in Fig. 1.9 (a), lower trace, does not vary in amplitude, and hence, is a constant-envelope signal. However, we don't transmit such signals. The signal as shape with a root-raised cosine baseband filter (shown here for roll-off of 0.24), is no longer purely constant-envelope. However, PAPR is still fairly low, so we consider it a quasi constant-envelope signal.

This property is desired because of the non-linearity of the high power amplifiers used in satellites, ground towers and gateways. The non-linearity distorts the signals. One way to manage this effect is to use **constant-envelope** signals. Examples are BPSK/QPSK, and most all M-PSK signals, MSK, PM and FM are all considered mostly constant-envelope, whereas 16QAM, and 16APSK signals are not. Shaped signal such as root-raised cosine QPSK signal, are not 100% constant-envelope but nearly so and cause the least amount of trouble, and hence their wide usage.

When more than one signal (even when it is a constant-envelope signal) is injected into a HPA simultaneously, and even though each is individually constant-envelope, the sum, as shown in Fig. 1.20 is not a constant envelope signal. A 2-signal (that is the sum of two signals) often has smaller PAPR than a 3-signal case which is then somewhat smaller than a 4-signal case, on up to about 7 signals, where thereafter, there might as well be 100s of signals, the PAPR approaches 6-8 dB and the amplifier distortion reaches a peak. Such signals suffer the most distortion, particularly when the HPA is operated near the maximum power. Hence, as a general rule, multi-carrier signals are not operated at maximum input power, whereas single carrier signals can be.

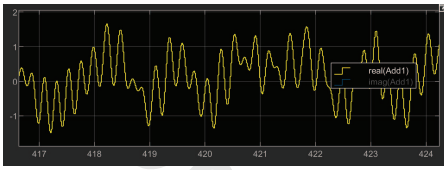


Figure 1.20: The sum of two RRC QPSK signals is clearly no longer constant-evelope.

1.3.9.1 Random signals

A carrier is a perfect periodic and unchanging wave that is modified by random information. This randomness of the information signal brings a forever changing nature to the carrier. The spectrum of such a signal is changing from moment to moment. We can talk about a particular spectrum of the signal, but it is just one picture of reality in time. The tools used to assess the spectrum of random signals is the DFT of the auto-correlation function of a selected part of the total signal, with length in power of 2. The formal name for computing the spectrum of such signals is called the **Autopower**. The alternate method is to compute the DFT of the signal and squaring it, called the **Periodogram**. The first method is considered the mathematically valid form for random signals. These considerations are important because much of the regulatory analysis such as out-of-band emissions analyses are specified based on the simulated and measured spectrum. In chapter 4, we discuss more about this issue.

Power amplifier parameters

An amplifier is specified by its maximum power output. This is often given in units of Watts, or as dBW or for smaller units, in dBms. Also given is the ***gain*** of the amplifier. The gain, a non-dimensional scale factor, tells us the amount of amplification that is provided at various input power levels. Hence an amplifier rated with maximum power output of 200 Watts and a gain of 30 dB would provide a 1000 times amplification. The amplifiers are able to operate over a large range of input power, and this range is called its **dynamic range**.

The relationship of the amplifier power-out vs. power-in is not linear. A typical power-in vs. power-out plot is shown in Fig. 1.21(a). Generally at lower power levels, the amplifier amplification as evidenced by the gain plot, is pretty nearly constant, as we see in Fig. 1.21. This is called the **small signal range** of operation. We note that the power-out peaks at about 0.2 Watts of input power. Thereafter it begins to decline. The term AM/AM is often used to describe this relationship. The figure in Fig. 1.21(b) shows the same graph plotted in dB form, by converting each input and output power to a $10log10$ or a dB scale. The x-axis is now power in dBms and the y-axis is in dBW. These values are normalized to their respective maximum levels. The amplification or, the gain begins to decrease, however, the total output power is still increasing. At a point called the ***saturation point***, the amplification becomes negative, i.e. the delta increase in output power is less than the delta increase in the input power. It seems reasonable that the amplifier should not be operated beyond this point. Not only does it not result in more power, it is also severely distorts most types of modulated signals.

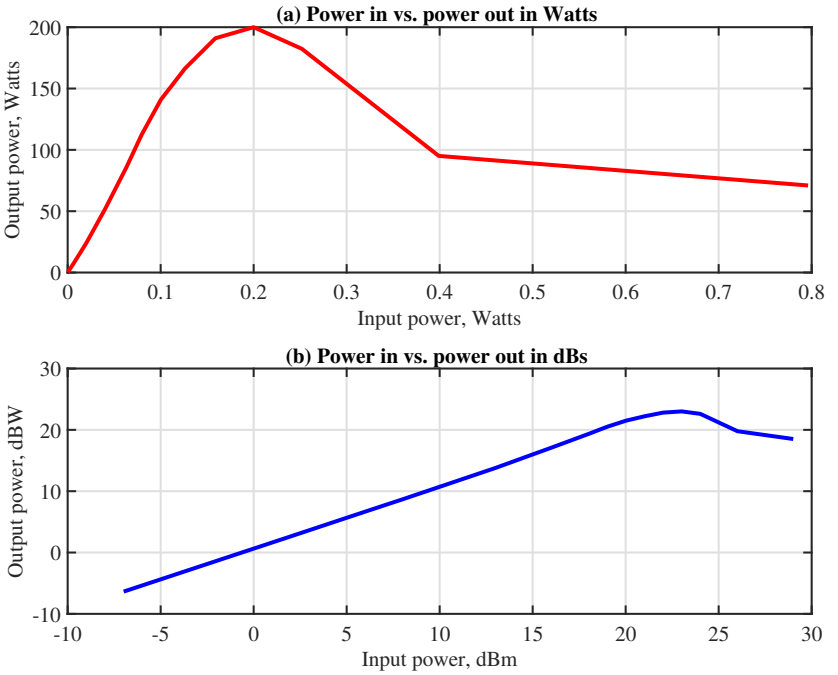


Figure 1.21: (a) Input power vs. output power of a typical HPA. (b) The same graph plotted in dBs takes on the familiar shape.

1.3.10.1 Input and Output backoff

Each HPA is designed to take in only a certain range of power after which it is said to saturate. i.e. the output in response to the input is no longer increasing. The input power at which the maximum power is obtained, a reference point, is called the input saturation power. If a transmitter transmits, instead at a lower power level, then this reduction in power from the maximum or the reference point is called the ***input backoff*** (IBO). Hence IBO at any point x in time is given as

$$\text{IBO}_x = 10\log_{10}\frac{\text{Input power}_x}{\text{Input power}_{sat}} \tag{1.6}$$

For example, a HPA saturates at 0.1 watts input power. Then at a point, where the input power is 0.03 W, the IBO would be -5.2 dB per relationship above. The IBO is always a negative number below the saturation point and positive above. So a 3 dB input power backoff means that the input power is now half of the saturation-inducing power. A 0 dB IBO means that a HPA is being fed the maximum power for which it is rated to produce the maximum output power. Note, that although the IBO is computed as a negative number, in discussion, people usually leave off the sign. Hence a 3 dB backoff implies a computed number of -3 dB. A small IBO colloquially, is numerically a

larger number, not smaller. The IBO values, above the saturation point are often called overdrive. Hence, a +1 dB IBO would be referred to 1 dB overdrive and not as plus 1 dB IBO.

We show the power transfer curve of Fig. 1.21(b) in the form called the ***IBO vs. OBO form***. If the input power is backed off, then obviously the output power would also be *backed off* from its maximum, and the maximum rated power would not be delivered. The reduction in the output power due to the reduction in the input power or input-backoff is called the ***Output Backoff***, (OBO). The output backoff is computed the same way by normalizing the y-axis values by the maximum output power. At the maximum output power, we would then see a value of 0 dB OBO. In this type of plot, both IBO and OBO are independently normalized and are 0 dB at the saturation point.

$$\text{OBO}_x = 10\log_{10}\frac{\text{Output power}_x}{\text{Max. output power}_{sat}} \tag{1.7}$$

The question one might ask, why use this nomenclature when real powers in watts and dBm would suffice. The main reason is that this method allows us to compare all sorts of amplifiers of various saturation powers, frequencies. We can compare the normalized curves to see which has better gain and better non-linear behavior. Another advantage is that we can refer to the operating setting as a 3 dB backoff point, instead of remembering what the maximum is for each amplifier and then giving an absolute number for the power level. This methodology is exclusively used for high power traveling wave tube amplifier used for satellite systems as well as towers and other HPAs. Small class-C amplifier use a slightly different methodology, particularly when specifying non-itineraries.The 1 dB compression point, as a specification is not commonly used in the satellite industry. It is more commonly used in the cell phone business since the operating range is usually quite small and fixed. The optimization of power also is not nearly as critical in industrial applications as it is for the satellites.

These power-in, power-out plots are developed by the manufacturers using a single continuous wave (CW) carrier. Hence sometimes a particular point is called the CW OBO, or the CW IBO, to refer to the operating point, which is to recognize that the actual power developed by a modulated carrier is different than a CW carrier, but the only calibrated point we have is that of a CW carrier. Hence that is what is used to set the operating point of a HPA.

However, real signals that use these amplifiers are rarely just a sinusoid. They are modulated signals of a finite bandwidth. The manufacturer of these amplifiers have no idea what type of signal a user might use, so in order to keep the specifications simple, the use of a sinusoid to establish the power transfer curve is standard. If we were to use a modulated signal, we would not get the same result as the idealized case of the power specs provided by the manufacturer. The maximum power generated by

a modulated signal will almost certainly be something less than the maximum stated by the manufacturer data. A 200W TWTA will not deliver 200W for a QPSK signal. This reduction from the stated maximum is due to the use of a modulated signal and is called the **modulation loss**, and we shall later discuss how it is computed for each type of modulation. This additional loss must be accounted for in a link budget, since we use the maximum CW EIRP as a starting point.

The type of non-linearity presented by the amplifiers is primarily an amplitude-based non-linearity. Hence it is often said that if a constant-envelope signal is used, it suffers no distortion. However that is in theory only. Real signals are never totally constant-envelope, but shaped pulses of finite bandwidth. All such signals are distorted by the HPAs, and this distortion is a factor how much the signal differs from a constant-envelope signal.

The HPA impacts the link in two ways, and both are a function of the operating point, i.e. where we set the input power. One is, the delivered power or the gain and the other is, the distortion at that point. We write the net amplifier effect on the link as function of the IBO as

$$\text{Net gain}_{IBO} = OBO_{IBO} - \text{Distortion}_{IBO} \tag{1.8}$$

In addition to amplitude, amplifiers can also cause the phase to shift. A non-linearized amplifier or a TWT has a very pronounced phase non-linearity, varying from 0 degrees all the way to about 50 degrees at saturation. The general rule of thumb is that every ten degrees phase increase will increase the degradation by about 0.1 dB. Linearized tubes improve the phase shift quite a bit, with a maximum phase shift from the linear range to saturation to about 10 degrees or less. They usually have this characteristic response as shown in Fig. 1.22(b), red curve. The phase shift is near zero in the small signal region, then begins to dip near saturation (usually at the point at which the tube is normalized/linearized) and then increases again.

1.3.10.2 Linearization

Linearization improves the power-out to power-in relationship by making it linear up to a certain point of operation. This is done by pre-distorting the signal by an LCAMP ahead of the TWT. The pattern of linearization is matched to a particular tube and improves not just the amplitude but also the phase shift out of it. Linearized tubes offer better performance for multi-carrier signals but it is not always clear if they are needed for single carriers. Another method of improving the performance is to add a limiter after the TWT. Fig. 1.22 shows the difference in the performance of a linearized vs. a non-linearized tube.

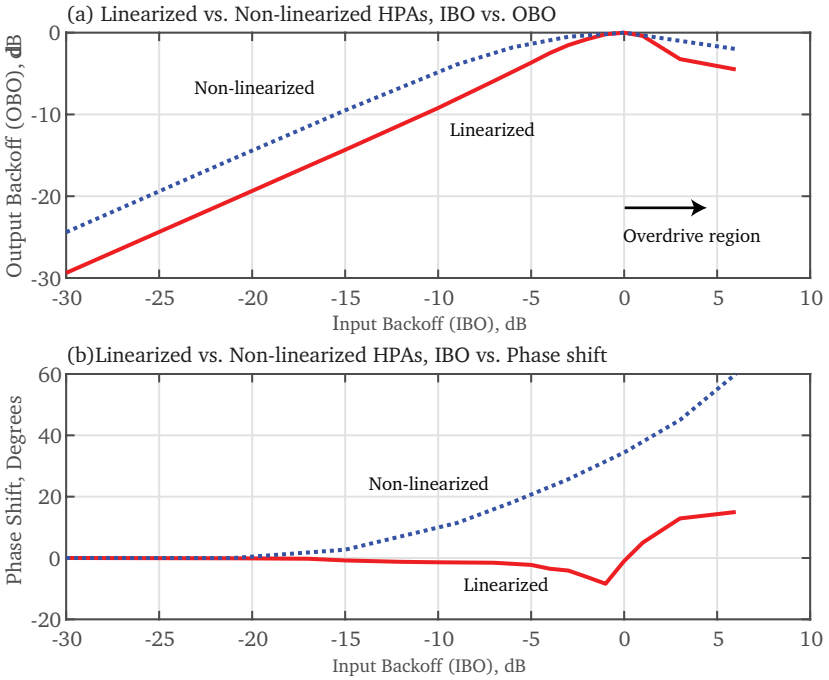


Figure 1.22: The power transfer and phase shift behavior of a traveling wave tube amplifier as a function of input power backoff.

1.3.10.3 Measures of amplifier non-linearity

The transmitter non-linearities produce two types of effects. The first type, called the **harmonic distortion**, produces replicas of the signal at integer multiple of the center frequency. A 2 GHz signal would also appear as copies at 4, 6 8 GHz etc. Bandpass filters are used to eliminate these copies from causing issues for other users at these frequencies. The second and more important effects are called the **inter-modulation distortion**. Unlike the harmonic distortion effects, these cause in-band distortion and can not be filtered out. Its as if the signal has melted a bit. This often called **spectral spreading**, **spectral regrowth** etc. because of the distinctive spectrum changes that occur. So whats the problem with this spreading? Well, first of all, there are adjacent signals and this spreading acts as noise for them. Secondly, the in-band signal also suffers from shape distortion which requires higher EbN0 to overcome. This is precious and limited quantity and hence we do not like what the TWTs and other such amplifiers including the SSPA and Gan amplifiers do to certain types of signals. We see in Fig. 1.24, what happens to the signal in time-domain when it goes through a TWTA, resulting in major spectral changes.

The effect of inter-modulation non-linearity on the signal, i.e. both the spreading of the noise into adjacent spaces, as well as time domain shape distortions, get larger as we operate closer to saturation. So we have a problem. On one hand we want the maximum output power and on the other hand we want the attendant distortion to be as

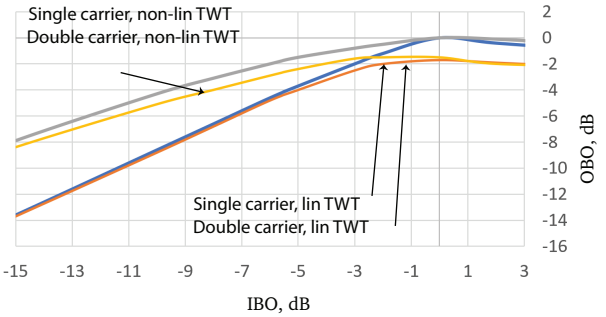


Figure 1.23: The total output power in presence of more than one carrier. Notice that both the linearized and a non-linearized TWTA output app. 1.7 dB less total power for a dual carrier case . This is for a CW case. Modulated signals, depending on the modulation, suffer even bigger losses. Hence all multi-carrier inputs will lead to a power loss.

(a) Single RRC shaped QPSK signal Input

(b) TWT output at 0.5 dB OBO

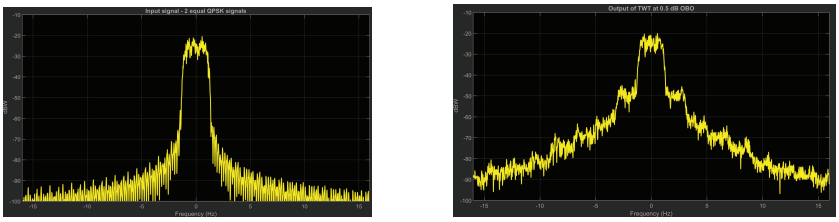


Figure 1.24: (a) A single RRC-shaped QPSK signal into a linearized TWT, (b) output signal at 0.5 dB OBO. The spectrum shape change is obvious. Its impact is a degradation in the required EbNo.

low as possible. There is a special non-linear relationship between these two issues and there exists an optimum operating point (OOP) for the HPA power where the net C/N0 after accounting for the increase in output power and the associated reduction in net power due to the degradation, is at a maximum for a particular link. At this **optimum operating point**, the link requires the least power to deliver a particular BER. There is a special analysis that needs to be done to find this optimum operating point for each type of link with the particular modulation being used and the number of carriers in the link. We will describe this methodology in Chapter 3. This analysis is a function of the modulation type, the number of carriers, the filtering being used as well as the particular specifications of the amplifier. Generally, this analysis requires time-domain simulation and is done by the satellite manufacturers, during the design phase.

In Fig. 1.27 we see the result of the effect of the non-linearity on the required EbN0. In theory we would not expect the required EbN0 to change as a function of power, but it does so for signals through a HPA, as shown in this figure. The required EbN0 increases, the closer we get to the operating point, OBO. When we combine the effect of the increasing EbN0 with the output power, we can plot the net effect, the CN0 as in (b). The inflection point forms the optimum operating point for this signal configuration. This analysis was done for a linearized TWT, and it shows the OPP for a

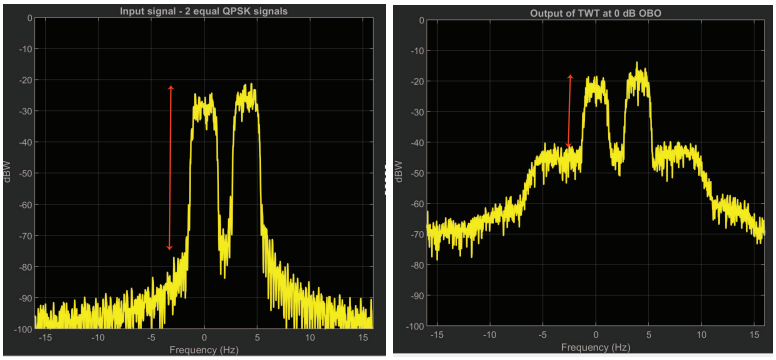


Figure 1.25: (a) Two equal QPSK signals into a non-linearized TWT, (b) output signal at 2 dB OBO. The output signal shows spreading (seen as shoulders on the side) as well as reduction in noise rejection. The noise floor for the two signals has moved up a great deal after going through the TWT, from -90 dB to app. -60 dB.

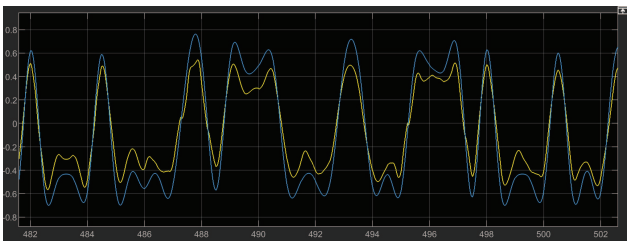


Figure 1.26: Time domain distortion of a QPSK signal (RRC shaped) through a TWT at 2 dB OBO.

single 8PSK carrier as at, 0.3 dB, 0.75 dB and 1.6 dB respectively for the one, two and three carrier case. Similar analysis are done for actual signal configuration to determine this operating point, this value is then used in the link budget.

Often, system operators use rules of thumb to specify the operating point. But simulation tools are available to compute these precisely and there is no need to operate at large OBOs of the past, such as 3 dB. A majority of signal configurations require a good deal less OBO than what is often assumed and mentioned in literature and books.

To provide a metric to assess the non-linearity of a HPA, the manufacturers provide two main metrics. These are called **C/3IM** and **noise power ratio, NPR**. The C/3IM forms the least or the minimum distortion bound when two carriers are used, and NPR which is the other extreme of being equivalent to a large number of carriers, is the other side of the behavior envelope. These two measures envelope the whole of the non-linear behavior of the TWTA, SSPAs and other high power amplifiers in the satellite industry.

1.3.10.4 Phase non-linearity

The phase non-linearity of a non-linearized tube is huge. It does cause measurable degradation, as a signal may suffer as much as 30 degrees of phase shift over a very narrow power range. Linearized tubes however have narrowed this to a just a few

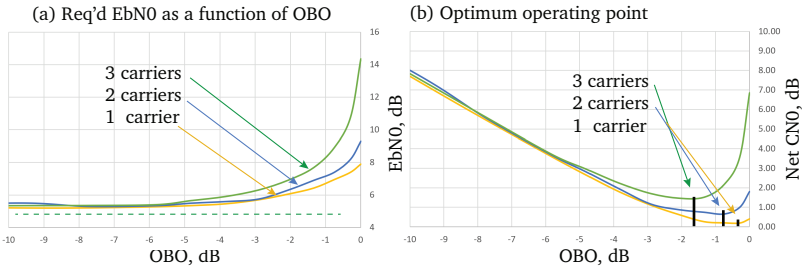


Figure 1.27: (a) Three signal configurations of 1, 2 and 3 equal 8PSK carriers through a TWT. The EbN0 is flat up to point and then as non-linearity becomes significant, it shoots up. In (b) we plot CN0 degradation which includes this effect and see that the C/N0 degradation increases as OBO is increased. The saddle point of each, is the OOP for that configuration.

degrees and offer a far better performance, hence are now used in a majority of the cases. However, using linearization should be done with consideration of the signal being used. A single carrier QPSK signal can often be better off with a non linearized TWT as linearization does lead to some power loss. Limiters can also improve performance in many cases, which may seem counter-intuitive.

1.3.10.5 Bandwidth flatness

Often a TWTA will be used to amplify signals located very far apart in the spectrum, such as in Ka band, where certain signals can be as much as 2 GHz apart. Hence a TWTA must have a fairly flat frequency response over a large region. In general they do. The response over a typical 72 MHz is pretty flat, but when larger bandwidths are used, the analysis needs to consider the rolloff in gain response in the band of interest. And example of a typical frequency response is show here in Fig.**??**.

1.3.10.6 C/3IM test

Here two equal amplitude CW signals are injected into a TWT, as shown in Fig. 1.28 (a). In (b), (c) and(d), we see in addition to the two input tones, new frequency components at integer multiples of the frequency separation between the tones. The first set on each side is called the third order modulation product or 3IM. The one after that, the fifth order and so on. The third-order product is nearly always the highest. The delta difference between the carrier signal and the 3IM product is called the ratio, C/3IM. It is indicative of the non-linearity present in the amplifier. It is an indication of what would happen to a signal that contains a wideband of frequencies. This metric is of course a function of the operating point of the HPA, hence a function of the IBO(or the OBO).

A question often asked is why is it called the third-order and not the second order, etc. The reason is that in this type of non-linearity, represented often by a power series, the second order effects fall as integer multiple of the individual frequencies and hence

are far out of the range of interest and are easily filtered out by a harmonic filter, only the odd-order terms remain in the bandwidth of interest and must be dealt with. But note that power does leak out into these even order terms and is lost forever. Hence a TWTA will never produce its rated maximum power for a non-constant envelope signal.

This behavior depends on the amount of non-linearity present. A non-linearized TWT has a more adverse behavior as we see in the table. Its C/3IM values are worse than one for a linearized TWT. Hence, this metric is used to compare the non-linearity of TWTs. However, this test only demonstrates what happens when just two frequencies are present. Another test, which measures the behavior in a multiple signal scenarios is called the noise power ratio as described below.

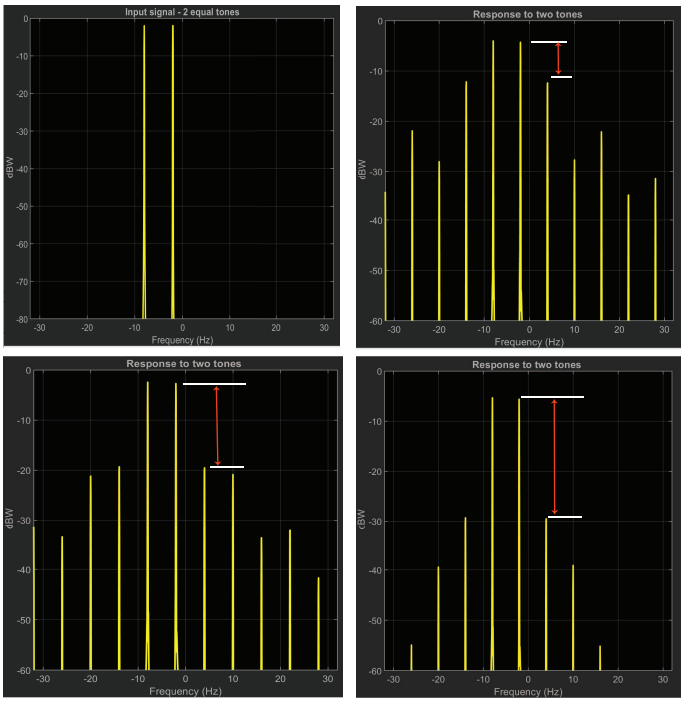


Figure 1.28: (a) Two equal CW tones into a linearized TWT, (b) output signal at 0 dB IBO, (c) output at 2 dB IBO. (d) output at 6 dB IBO. Note that as the input backoff is increased, the delta between the tones and adjacent third order inter-modulation product increases. This delta, a ratio of powers, is called the metric C/3IM.

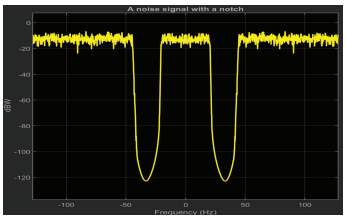
Table 1.2: C/3IM of a linearized vs. a non-linearized TWT.

	C/3IM, dB	
OBO, dB	**Lin TWT**	**Nlin TWT**
0	8.02	7.08
-0.5	14.61	9.355
-1	22.34	12.105
-2	23.78	15.64

1.3.10.7 Noise Power Ratio, NPR

The other extreme end is the case of multi-carriers. This behavior is encompassed by a test setup called the ***noise power ratio*** or NPR. In this case a noise signal is injected into the amplifier. A noise signal has a wide bandwidth, hence this is equivalent to a multi-carrier signal. The signal is put through a notch filter prior to the amplifier. Had the amplifier been perfectly linear, the spectrum of the signal after the amplification, would look exactly as the input with a notch in the spectrum. However, when such a signal goes through a HPA with non-linear behavior, the notch fills up with noise. The amount of filling (from the top) is a measure of the non-linearity introduced by the amplifier. The amount of this fill is a function of the IBO/OBO level.

(a) A noise signal with a notch



(b) TWT output at 2 dB OBO, NPR of app. 13 dB.

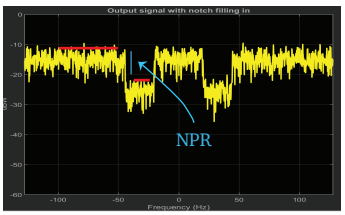


Figure 1.29: (a) A notch filter is added after a pure noise signal, hence we see large notches in the spectrum. When this signal goes through the TWTA, the notch fills up with noise in (b). The delta difference in the signal level in the un-notched region and the notch is termed the noise power ratio or the NPR. It is given in dBs and varies with the power level or the back-off. It tells us how a TWTA will behave when a multi-carrier signal is put through it.

The C3IM and the NPR form the two ends of non-linear effect. In most cases, HPAs have no more than 10 carriers and hence, the true behavior of a multi-carrier case falls somewhere between these two cases.

Note, that often, these are given as a function of the OBO and not IBO, the reason being that the system operators have a dial that sets the desired OBO and not the IBO. This is done for convenience and for maintaining a certain output power level. The designer needs to know the non-linearity to expect at a given OBO where the system will be operated, hence these graphs are given usually as a function of the OBO and not the IBO.

How does NPR compare with C/3IM? This is not a really valid question because they are not equivalent. The two parameters look at different aspects of the non-linearity, but it is a question that is often asked. We can say in general that NPR is always a smaller number than the C/3IM, because it shows the response to many carriers vs. the two tones for the C/3IM. A behavior for one particular TWTA is shown below.

There are two other infrequently used parameters sometimes specified in the satellite industry. In author's opinion, these are not useful metrics of performance of a HPA, that

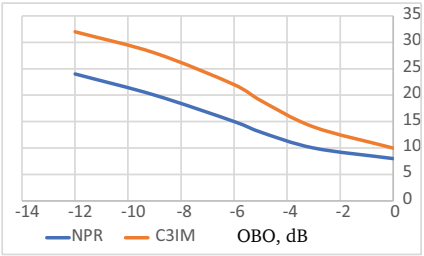


Figure 1.30: A comparison of C3IM and NPR tells us that the two-carrier non-linearity metric, C3IM is usually results in a smaller degradation. Both of these are just indication of what a user of the HPA might expect and are not in themselves a measure of what an actual signal will experience. That depends on many other things, primarily on the type of signal shaping and modulation, as well as the number of carriers.

are not already covered by the NPR and the C/3IM They are rarely used other than to to verify the performance of the TWTAs. However, the first one (AM-PM conversion) is essentially a duplicate of the phase curve and the second (AM-PM transfer) is encompassed by the C/3IM and the NPR specifications and hence both of these are not useful and should be dropped by the satellite industry to save cost and time.

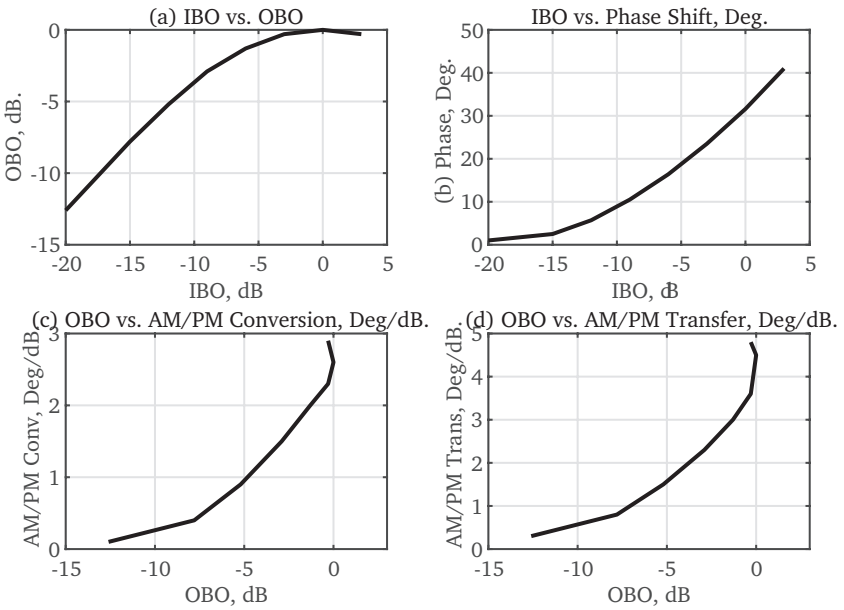


Figure 1.31: The AM/PM conversion and transfer coefficients for a non-linearized TWTA.

1.3.10.8 AM-PM Conversion coefficient

This specification specifies the phase shift change of a single carrier over a drive level. It is given in degrees per dB of change in the OBO. To some extent this test is duplicate of the phase shift information, but is given in a rate instead of as a degrees phase shift at a given drive level. In Fig. /reffig:twtaconv(c) we see a representative graph of this

parameter for a non-linearized TWTA. It is not a useful parameter and is redundant. In Fig. 1.31, we give the AM/AM conversion and AM/PM transfer graphs for a non-linearized TWTA.

1.3.10.9 AM-PM Transfer coefficient

This parameter is a measure of a second order non-linearity using two carriers. It measures the effect of increasing drive level of a smaller carrier on the phase of a larger adjacent carrier through the same HPA. Two carriers are used, a larger carrier with 1 dB amplitude modulation and a second carrier that is unmodulated. The drive level of the smaller carrier is increased from -20 dB to saturation and the change in the phase of the larger modulated signal is monitored. The result of this test is to some extent an attempt to quantify the phase distortion that might result in a multi-carrier signal scenario as adjacent signals change in power levels.

In this chapter we covered some items that are important in understanding and doing link budgets. Another significant topic is the signal itself, its bandwidth, modulation and the ultimate capacity a signal can achieve. We discuss some of these issues in the next chapter.

References: To be added.

Chapter 2

Symbols, bits, quality and capacity

Symbol and Bit rate

The prime motivator for doing link budgets is to determine if our hoped-for data throughput can indeed be achieved by the design we have created. To maximize the total network throughput, our goal for each link is to be able to transmit through it, as much data as possible. In addition to the throughput as our desired goal, we also want a certain signal quality, a parameter that is most commonly specified by the **bit error rate** (BER). In addition to the BER, we also want the network to be reliable and available in most commonly encountered environments. The **system availability** parameter is designated as a percentage, such as 99%, 99.5% etc., which specifies the percent of time, the system must be *available*, barring such events as solar eclipse, outages due to major malfunction etc. The **availability** indicates the sensitivity of the design to external events such as severe atmospheric disturbances that can cause the data rate to fall below a minimum specified throughput. The metric of data throughput is the information transfer rate via the link. The term *throughput*, also given in terms bit rate, may also be specified for a group of links, a beam, or the whole network. This number is based on the gross data rate and often includes code bits and all protocol overheads. For example, it may be said that a particular satellite system can provide a throughput of 4 Gbps or that a wi-fi link can provide 500 Mbps, whereas the actual information transfer may only be 75% or smaller, due to coding and other overheads.

Despite our emphasis on *bit rates*, a communications system, in fact, operates on the basis of ***symbols***, and not bits. A bit, if it were to be represented as an electromagnetic signal, i.e. a pulse, would require far too high a bandwidth and, hence is not a physically useful form of signaling for wide-band networks. In a communication system, what we transmit are ***symbols***, used as a proxy for an assigned number of bits. These symbols are small segments of a sinusoid, differentiated by their amplitude or starting phase. Shannon's theorem tells us that a link has a capacity to transmit a certain number of

symbols per second, which is deemed to be the ***capacity*** of the link. From here, the choice of the ***modulation*** determines how many bits each symbol represents.

What is a symbol

Think of the Morse code. The length of a beep is a ***symbol***, or an entity that represents a certain information. Morse code has just two symbols, a long beep and a short beep. By using a combination of these beeps, i.e. symbols, letters and words can be formed by a preassigned set of mapping. In electromagnetic signaling, the symbol is a small piece of a sinusoid wave.

As we know, a sinusoid has three parameters, frequency, phase and amplitude. The frequency and the bandwidth at which we transmit for a given link (or even a network), is usually a regulatory stipulation. The other two aspects, can then be used to *code* a symbol. In digital form of signaling, it is the phase that is manipulated to code symbols, and as such this type of coding is called **Phase modulation** or **Phase shift keying** (PSK). The number, M in M-PSK, tells us how many symbols are being used. A modulation referred to as 8PSK, then will have eight fundamental symbols, which can be manipulated to create words, letters, numbers etc.

In Fig. 2.1, we see, two symbols used by the 2-PSK, or **BPSK modulation** as defined per Eq. 2.1. Note that these are just one period of a sinusoid at a certain frequency, $f_0 = 1/T_s$. (For graphing purposes, we have assumed that f_0 and T_s are unity. The x-axis in this figure is the argument of the cosine or the quantity ωt. Since f_0 is 1.0, then the phase is obtained by multiplying the x-axis by π.) The first symbol in (a), called s_0, is a single period of a cosine which at the time of its origin has a phase of zero π, or 0°. the second symbol, s_1, is a single period of the same cosine with a starting phase of π or 180°.

Imagine, you are a receiver, and you are catching from the air, little cards, each with one of these pictures. You can easily decode the sequence by just mapping what you see with their bit representation. We assume that no one modified the drawings on the way to you, nor changed their order, which in fact does happen to electromagnetic symbols as they fly through the air!

$$\begin{aligned} s_0 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s) \\ s_1 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s + \pi) \end{aligned} \tag{2.1}$$

The BPSK transmitter is using an electromagnetic blip in form of a little sinusoidal wave, lasting T_s seconds, to indicate a bit of information. Note that only one sole parameter distinguishes symbol 0 from symbol 1, and, that is the *phase*. Hence this scheme, has a *single* degree of freedom. The amplitude of the symbol expressed as

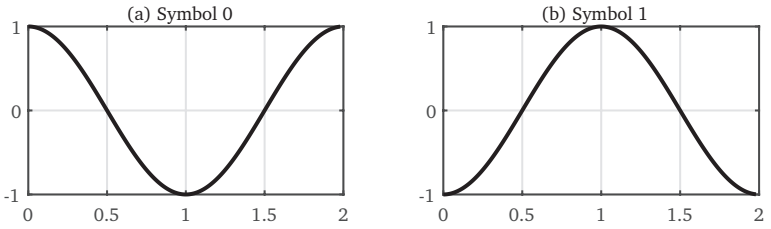


Figure 2.1: (a) Symbol 0, (b) Symbol 1. Both are based on a cosine basis function. Symbol 0 has a starting phae of 0°and the second sysmbol has a stating phase of 180°.

$\sqrt{\frac{2E_s}{T_s}}$, is function of the power we wish to impart to the signal, (See chapter 3). For didactic purposes, we will assume that this term is equal 1.0 in this section. In real life, both T_s and A_s are much smaller than 1.

The only degree of freedom, this scheme has is the phase. Hence the number of symbols possible is $M = 2^n$, where n is the number of degrees of freedom. The term n can also be thought of as the number of bits a symbol represents. The term M is 2, hence this M-PSK modulation is called 2PSK, or BPSK. This modulation represents 1 bit per symbol.

To transmit bits, we take the incoming bits and since we can not actually transmit the *bits*, we transmit what we can, and that is these *symbols*, s_0 and s_1. Let's assume that the transmit rate is one bit per second and hence, the symbol transmit rate is also one symbol per second. We use the term, R_s for the symbol rate. Since each symbol represents just one bit, the bit rate, R_b is equal to the symbol rate. Now we map a bit

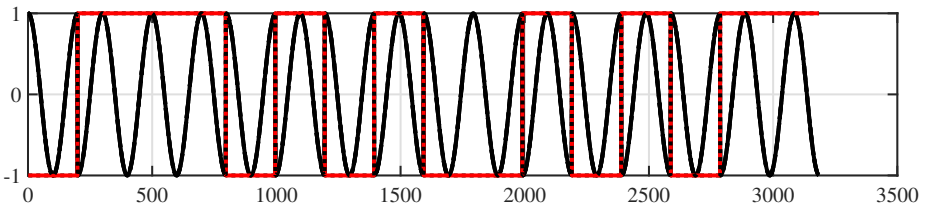


Figure 2.2: We see a symbol stream that is created from bit sequence, 0111010100101011, (shown in red). Note that at each symbol change, from one symbol to another, the phase of the carrier changes by 180°. This is a BPSK modulated signal.

stream of $0,1,1,1,0,1,0,1,0,0,1,0,1,0,1,1$ to a symbol stream of $s0,s1,s1,s1,s0,s1,s0,s1,s0,s0,s1,s0,s1,s0,s1,s1$. In Fig. 2.2, we see what these symbols look like when arrayed as a continuous stream. Note that symbols change after one period, in response to the incoming bit. This is an *analog* signal that will be transmitted in lieu of the bits. The identity of the bits has disappeared, hence, we call this a process of mapping, **phase modulation**. It consists of mapping the incoming bits into phase differentiated symbols, as pre-defined by Eq. 2.1.

From symbols to modulation

A signal looks incomprehensible in time domain. Hence, in signal processing, we have devised a system of transforms to to understand signals. Time to frequency transforms via FFT, Z-transfroms etc. are one way. The other way is by suppressing the time dimension. Modulation is best examined with a **constellation diagram**, as shown in Fig. 2.3. The constellation diagram is also called the IQ diagram. This is a view of the modulated signal without the carrier, hence, this is called a baseband view. The radius of the circle represents the amplitude of the signal. The expression of the amplitude is given by $\sqrt{\frac{2E_s}{T_s}}$, and we will discuss why it is so in chapter 3. The location of the symbol on this circle, tells us the phase which is assigned to it, from 0°to 360°, going anti-clockwise. The x-axis in this diagram is called the **In-phase** or the **I axis** and the y-axis is called the **quadrature** or the **Q axis**. The circle in this diagram represents the amplitude of the carrier, hence its power. We do not see time, but only the phase as the points (symbols) move around this circle in a radial fashion.

For a BPSK signal, symbol 0 and symbol 1 both have the same amplitude but are polar opposites in phase. The distance between these two symbols on the constellation diagram is twice the amplitude or the diameter of the circle. This distance, called the ***inter-symbol distance***, is an important parameter in assessing the error sensitivity of a modulation scheme. The smaller this distance between the neighboring symbols, the greater is the error sensitivity to noise and other sources of distortion. The closer the symbols are, the *harder* it is for the receiver to tell them apart. This in intuitive, as although we can easily distinguish visually between a phase of 0 and 180 degrees, we would be hard pressed to distinguish two signals that are just 20 degrees apart.

Due to the large inter-symbol distance, BPSK is the best of the digital phase modulations, when it comes to error sensitivity. But this immunity comes at a price of lower throughput. The BPSK bit rate can definitely be improved upon. Other modulation methods, where we allow the symbol greater phase choices and/or amplitude, can provide better bit rates, but of course, at increased bit error sensitivity because the symbols would have to be placed closer together, assuming the power is to stay the same.

By taking advantage of the magic of orthogonal sinusoids, we can transmit two BPSK signals at the same time. The symbols are defined now with *two* orthogonal basis functions, one a cosine as in Eq. 2.1 and another one, a sine. The data rate doubles, as we now send two symbols instead of just one, in the same time period as before. This is called the quadrature representation of a signal, since a sine and a cosine are in quadrature (or 90°)to each other and are called the I and the Q channels. But we don't actually transmit two signals, this is just a mathematical concept. The real signal, the one that is transmitted, is a single signal that is the sum of these two sine and cosine or the I and the Q channels.

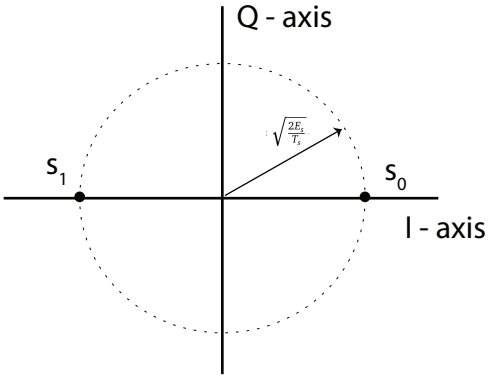


Figure 2.3: Imagine this signal in a 2D space, with amplitude of the circle, the amplitude of the signal and angular position on the circle as the phase of that symbol. Then symbol 0 is at 0 degree point and the symbol 1 sits on the opposite end at the 180° point.

QPSK is, hence, essentially two orthogonal BPSK signals. Q stands for 4, and it also called a 4-PSK modulation. Here we have, four symbols instead of two. To do that, we keep the amplitude the same, but we allow four phases, 0°, 90°, 180°and 270°, or alternately the preferred set of 45°, 135°, 225°, 315°. Plotted on an constellation (or an IQ) diagram, they would appear as in Fig. 2.5. The following expressions define these four symbols.

$$\begin{aligned} s_0 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s + \pi/4) & s_1 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s + 3\pi/4) \\ s_2 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s - \pi/4) & s_3 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s - 3\pi/4) \end{aligned} \tag{2.2}$$

Note that there is nothing particular about the starting phase of $\pi/4$ or the 45° for symbol s_0. We could have started at any angle, just so as long as each signal is shifted by $\pi/2$ or 90° from the previous one. 45° is just conventional,probably due to hardware limitations. Similarly, BPSK can also be oriented in any direction, as long as the two symbols fall diametrically opposed to each other. We see the time-domain shape of these four symbols in Fig. 2.4.

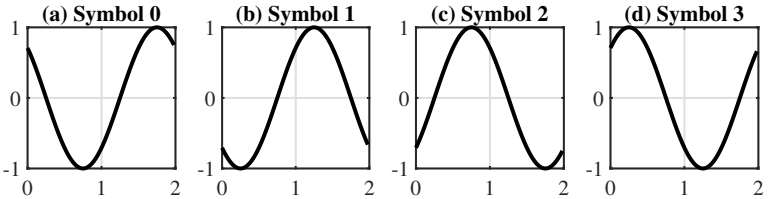


Figure 2.4: The four symbols of QPSK in time-domain

How do we send the same bit sequence as in Fig. 2.2 using QPSK? First we group these in bits in pairs, as underlined here. 01 11 01 01 00 10 10 11. Then we use the bit

Lorem ipsum

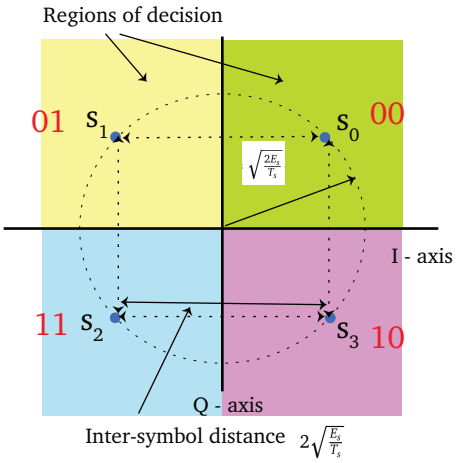


Figure 2.5: The four symbols of QPSK shown on a constellation diagram.

to symbol mapping plan per Fig. 2.5 to create the modulated analog signal as follows.

$$\begin{aligned} &\text{Bit stream: } 01\ 11\ 01\ 01\ 00\ 10\ 10 \\ &\text{Mapped symbols: } s1\ s2\ s1\ s1\ s0\ s3\ s3\ s2 \end{aligned} \tag{2.3}$$

In Fig. 2.6 we see the resulting QPSK **modulated carrier** that would be transmitted for this bit sequence. The number of symbols transmitted is now one-half the number of bits. Hence we can think of QPSK as the first of the ***bit-efficient modulations***. Note

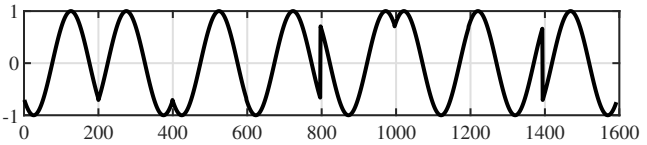


Figure 2.6: QPSK modulated signal created from bit sequence 0111010100101011.

in Fig. 2.5 that the inter-symbol distance has decreased from $2\sqrt{\frac{2E_s}{T_s}}$ for BPSK to $2\sqrt{\frac{E_s}{T_s}}$ for QPSK. The inter-symbol distance determines the BER a signal will experience by the following general equation for a signal in an AWGN channel, where d is the minimum inter-symbol distance for a given modulation. The error probability, P_e is proportional to the square of this distance.

$$P_e = \frac{1}{2}\, erfc\, \sqrt{\frac{d^2}{4N_0}} \tag{2.4}$$

In Fig. 2.5, we see how a receiver makes a decision about what was sent. It has a very simple decision to make. It examines the location of the received symbol on this

diagram by measuring the phase shift. As long as the symbol is still in its original ***region of decision***, which for a QPSK is one of the quadrants, it will make a correct decision and map the symbol to the correct set of bits. However, if the disturbance is large enough to have displaced the transmitted symbol, in phase and amplitude, to a neighboring region (the most likely scenario), the receiver will make an incorrect decision, resulting in an error.

As more symbols are added to a constellation in a bid to increase the bit-efficiency, the region of decision gets smaller. This is the dual sword of bit efficient modulations. By increasing the number of bits mapped to each symbol, we can indeed increase the throughput, but at the same time, in noisy channels, even smaller distortion in phase and amplitudes are much more likely to result in error. Of course, we can increase the power of the signal, which will make the circle bigger and increase the distance between the symbols, but here we want to compare how increasing bit-efficiency results in greater errors for a given power.

The course of an error is the result of having moved a transmitted symbol out of its original region of decision, hence the receiver mistaking one symbol for an another. This gives us the following relationship between the symbol error rate (SER) and the bit error rate(BER) depending on the modulation order being use.

$$BER = \frac{SER}{M} \tag{2.5}$$

Here M is number of bits assigned per symbol. For BPSK, M = 1, for QPSK, M = 2, etc. This linear equation assumes that only one bit error will happen per symbol, in other words, the symbol s_0 is far more likely to be confused with adjacent symbol s_1, than symbol s_2 which is farther in a distance and region sense. Confusing symbol s_0 with s_1, will cause only one bit error (00 vs. 01) whereas confusing symbol s_0 with s_2 with cause a 2 bit error (00 vs. 11). This is assumed to be less likely in a AWGN channel.

Packing more bits into a symbol

To increase the throughput, we can use a symbol to represent more than two bits and in fact as many as we want. But, we find that the BER of the resulting signal increases rapidly. Imagine a situation, you are looking from a high window into a playground, trying to spot your child. Its far easier when there are fewer children in the area than when the place in packed. In the domain of signal processing, decoding a signal correctly in such a case requires significantly more power. (Note that larger power causes the radius of the IQ diagram to increase, hence increasing the inter-symbol distance, they key parameter.) If the system is bandwidth-limited and but has plenty of power, then these higher order modulations may make sense, but, for satellites which are power limited (due to their being weight-limited), we usually do not use these modulation

scheme presently. But this is changing as high-power amplifier non-linearities improve. We can, by this analogy of assigning more bits to each symbol, create the next signal, 8PSK, by assigning three bits to each symbol. This means, we need eight different symbols, created by the following expressions.

$$\begin{aligned} s_0 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s + \pi/8) & s_1 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s + 3\pi/8) \\ s_2 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s + 5\pi/4) & s_3 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s + 7\pi/4) \\ s_4 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s - \pi/8) & s_5 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s - 3\pi/8) \\ s_6 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s - 5\pi/8) & s_7 &= \sqrt{\tfrac{2E_s}{T_s}}\cos(2\pi/T_s - 7\pi/4) \end{aligned} \tag{2.6}$$

In Fig. 2.7, we see these 8 symbols. Each symbol differs from its neighbor by 45°, as opposed to 90° for QPSK.

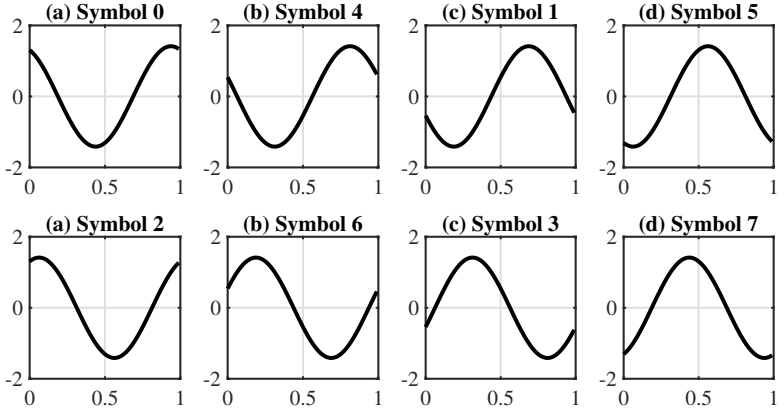


Figure 2.7: 8PSK symbols. Each is 45 degrees shifted from the previous.

How do we send a bit sequence 0111010100101011 in 8-PSK? The sequence consists of 5 symbols (3 bits per symbol) plus one extra bit which we will ignore. The order in which these symbols would be transmitted is

$$\begin{aligned} \text{Bit stream: } & 011\ 101\ 010\ 010\ 101 \\ \text{Mapped symbols: } & s2 \quad s4 \quad s5 \quad s5 \quad s4 \end{aligned} \tag{2.7}$$

In Fig. 2.8 we see the modulated carrier that would be transmitted for this sequence. Note that the abrupt phase transitions at the symbol boundary range from $\pi/8$ to $7\pi/8$. If we draw its constellation diagram, we get the figure with eight points shown as Fig. 2.9(a).

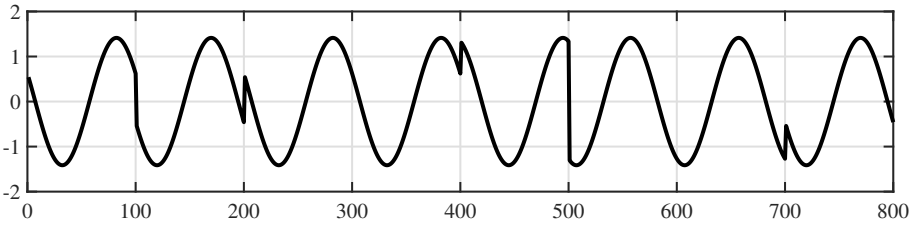


Figure 2.8: 8PSK signal in response to a bit sequence of 0111010100101011

2.1.3.1 Higher-order modulations

Of course, we can go further, by having symbols represent four bits instead of three or even more. A four-bit representation requires that the signal (the symbol) have sixteen different unique shapes. In Fig. 2.9, we see in (a) the 8PSK constellation diagram, along with three other schemes that are possible for a 16 symbol modulation. The first idea is to lay them all out in a circle, same as 8PSK, as in (b). The inter-symbol distance then gets smaller. In (c), we lay out the symbols in rows and columns. Because, now not all symbols have the same amplitude (the inner rows vs, rows the outer rows), this type of scheme is generally called a **quadrature amplitude modulation** or QAM. Then we have in (d), something called **amplitude phase shift keying**, APSK. The 16APSK scheme is the most promising of the three ideas. The last two, (c) and (d) are considered non-constant envelope, while the first two are considered constant envelope. (By constant envelope we mean that the average symbol amplitude, hence the power of the signal is constant.)

Let's examine the inter-symbol distance of the four constellations in Fig. 2.9. As you can see, as the constellation density, i.e. the number of symbols is increasing, this distance between neighboring symbols is decreasing. 16APSK inter-symbol distance is smaller than 16QAM. The region of decision is getting smaller. We would ordinarily prefer 16QAM over 16APSK, but 16QAM also has issues of its own. Most amplifiers cause amplitude non-linearities which greatly distort the higher power (i.e. larger amplitude) symbols in the cornerls. The alternate, scheme, 16APSK in (c), is also not constant-envelope, but, it has no high-power corner symbols, as such, is superior to 16QAM, when high amplifier efficiency is desired.

These ideas can be extended as far as we like. The only limitation in obtaining bit-efficiency, i.e. the number of symbols we use, is the receiver's ability to distinguish each of them and to correctly map them to the bits.

2.1.3.2 Gray coding

There is an another issue about mapping of bits to the symbols that we should discuss. In AWGN channels, the most common errors occur when a symbol is confused for one of its neighbors. So if we map the bits to symbol in such a way that adjacent mapping

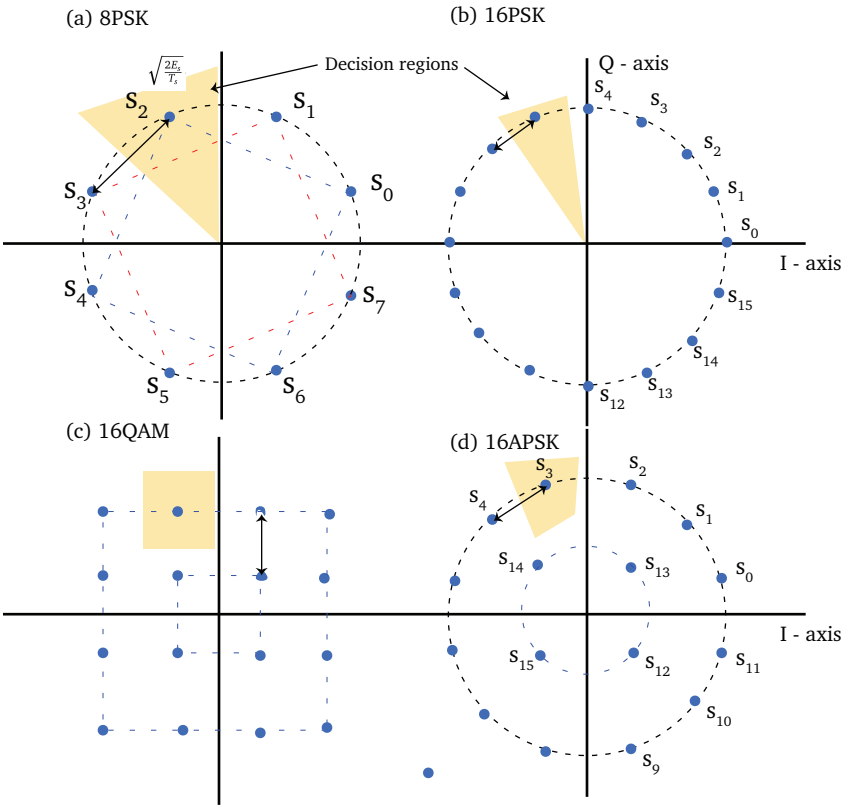


Figure 2.9: (a) 8PSK symbols, all of equal amplitude. (b) 16APSK with sixteen symbols, and a typical region of decision in yellow (which extends further out than shown), (c) 16QAM where 16 symbols are arranged in an array and (d) 16APSK, where same sixteen symbols are arranged in two concentric circles.

differs in as few bits as possible, less number of errors will be reported per mistaken symbol.

In Fig. 2.10(a), we see QPSK mapping that proceeds along the binary numbering scheme. We see that symbol 3 and symbol 0 differ in two bits, 00 vs. 11. So if either of these symbols is decoded incorrectly, it will lead to a 2-bit error. In (b), a coding method, called **Gray coding** is used to assign the bits to the symbols. Here we note, that all adjacent symbol assignments to bits are different from each other in only one bit. We see the same thing in (c) and (d) for 8PSK, In MPSK modulations, it is always possible to apply Gray coding, but not so for APSK modulations. We see in (e), for 16APSK, the bit mapping of the inner symbols are two-bit different from neighboring symbols. Hence, we are not able to use Eq. 2.5 to estimate the 16APSK BER from the SER.

How many symbols per second can we transmit on a given channel? By the laws of physics, this is a physical constant called the channel capacity. However, the important

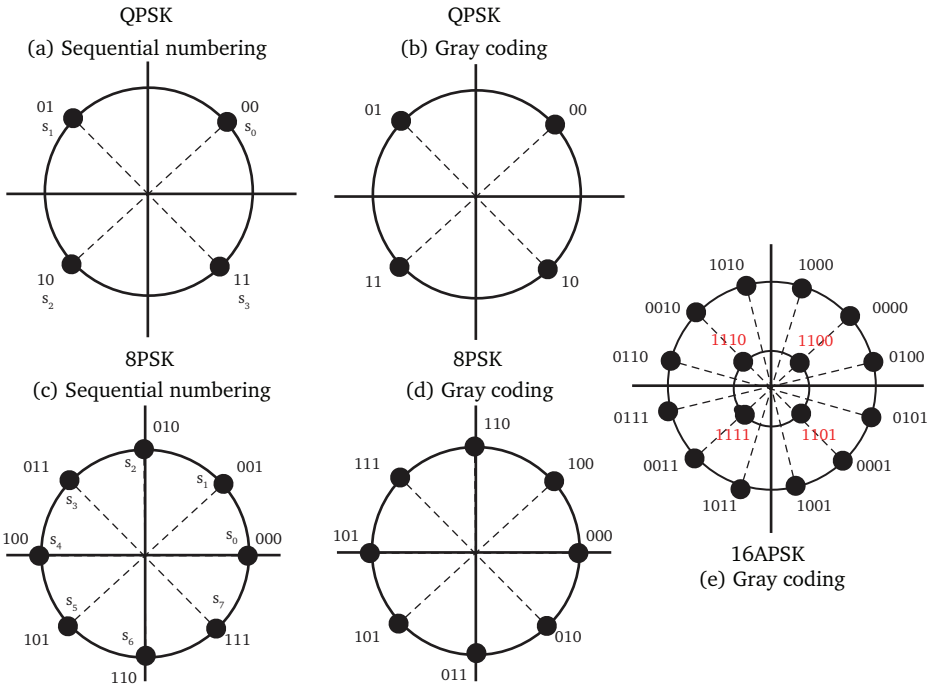


Figure 2.10: (a) QPSK symbols named in binary sequential manner, (b) QPSK symbols mapped with Gray coding, (c) 8PSK mapped with binary sequential numbering, and (d) same with Gray coding. Gray coding keeps adjacent bit mapping delta to a minimum.

thing to keep in mind is that the same number of symbols are going through the communications system in each of these modulations, and that the communication system is not "working any harder" in a higher modulation system. Its just those more numerously shaped symbols present decoding complexity.

How to compare modulations

So now a question can be asked, if higher order modulations are so great, then why don't we always use them? This question can be answered by looking at the **required** for these modulations. This metric, the required , tells us what it will *cost* to use a higher order modulation. Nothing is free, as we know. If we want a bigger, faster car, we have to pay more. Same here. We pay in terms of higher which is mostly all about larger transmit power. Transmit power tends to be a limited quantity for many communication systems. Yes, a cell phone could transmit with 32APSK, which would be awesome for watching hi-def movies, but then it will also have to be the size of your old wireless phone and not the tiny thing you can put in your pocket. Perhaps as newer technologies are developed for amplifiers, this may indeed be possible.

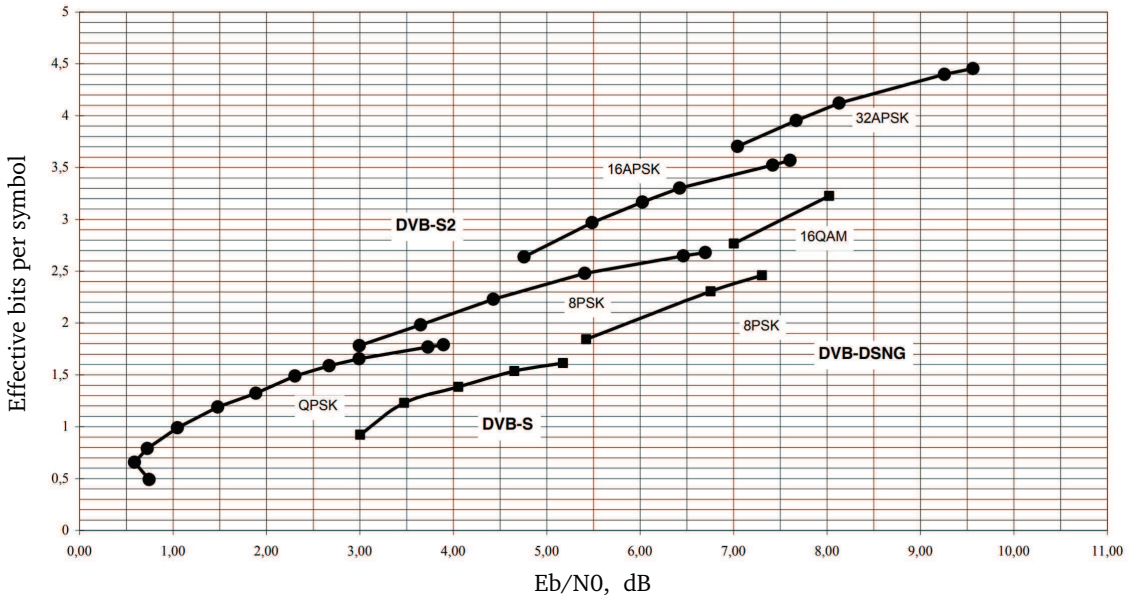


Figure 2.11: Bit efficiency that can be achieved by varying the code rates for MPSK modulations in an AWGN channel at a BER of 10^{-7}. *Source: DVB-S2 standard, document A171-1*

In Fig. 2.11, we see a set of plots for several modulations part of the ITU DVB-S2 standard. The x-axis in this figure shows the required (the power required) to achieve a certain BER. The y-axis is the **bit efficiency** achieved, indicating how many bits can actually be received per symbol. The points along a particular modulation plot line indicate the code rate used. As the code rate increases, i.e. less coding overhead, the required increases for a given BER. QPSK is shown to have bit efficiency, from less than 1 bit per symbol to a maximum of 1.75 bits per symbol, while as we know it should be able to achieve 2 bits per symbol. Same for 8PSK, which is seen to provide no where near the theoretical value of 3 bits per symbol. Simulation of these modulations with noise tells us that in real channels, we are not able to achieve the theoretical efficiency, This plot, however, is still an ambitious plot. It includes *only* the effect of Gaussian noise, and coding. This graph does not include many other effects such as, system non-linearities, pulse shaping, filtering, interference effects, all of these in fact reduce these numbers even further.

The system designer of course wants the highest bit efficiency possible, but limitations of hardware and the **available** , which is the main topic of this book, limit the choices. The selection of modulation depends a great deal on the *available* of the link to meet the *required* of the chosen option.

How we measure signal quality

There is no one quality measure for a communication system. The quality measure depends on the OSI functionality layer. At the *network* level, we consider **throughput**, **delay**, **memory** requirements, and **availability** as the prime metrics of quality. While at the *link* level, we judge the signal quality by the **bit error rate** of the signal. This, of course, implies that we can count the *bit errors* and hence the signal is presumed to be a discrete or a digital signal at the time of measurement. In a communication system, not all parts of a link are in digital form. Portions, such as the modulated signal, are analog. The figure of merit for this portion is the signal is **carrier to noise ratio, CNR**. The BER is the measure of the final signal quality as the baseband data itself is nearly always a discrete signal now.

When a bank transmits money over the line to an another bank, they want the numbers to go through perfectly accurately, hence an "error-free" transmission. Typically a data transmission at about 10^{-9} BER is considered nearly error free. A BER of 10^{-9} is equivalent to about one error per ten million pages of text transmitted. Rightfully so, we consider this a "quasi error-free" rate using this term.

A typical satellite link can provide, for a regulatory-compliant power level, a BER of only about 10^{-2} or more. This may come as a shock to you. Are satellite signals of such poor quality? Yes indeed, the raw channel error rate is often this high, but not just for satellite links, same is true for all wi-fi links. For cell phones, the uncoded links are operated at even worse BER levels. However, coding comes to the rescue and is used to clean up the signal before delivery. So the customer receives a fairly clean signal and knows nothing of the channel BER. The delivered error rate can be as good as one desires, by changing the code or the code rate being utilized. This effects the throughput of the link, but not the power required, which is presumed to be a fixed quantity. Links often have adaptive coding systems, and, these can be changed as channel environment dictates, to obtain or to improve a given signal quality, if the available declines due to environmental conditions.

Error correction coding (ECC) is nearly always used on all types of links now. There are many different varieties of ERC, suitable for all different types of channels. In all cases, the use of coding is a trade-off between power needed and the bit rate achieved. Coding is accomplished by adding redundant bits to a packet of information bits. The bits used to convey coding can not be used for information, effectively reducing the usable bit rate for a given bandwidth and power level. Hence we now have to consider two types of data rate, a gross data rate that includes the coding bits and the true information bit rate, which is smaller than the gross bit rate. All this because, the total symbol rate (hence the bit rate) is limited by the bandwidth assigned to the link by the Shannon limit. It can not be increased without breaking natural or regulatory constraints.

The BER, a link experiences is a function of the ratio of the bit energy, E_b, and the noise-density, N_0 of the signal. The E_bN_0, the ratio, is a non-dimensional parameter. We can think of E_bN_0 as equivalent to the term *net power*. It is independent of the bit rate, bandwidth and actual power of the signal, so it is very useful as a performance benchmark. Generally, the larger the *available* E_bN_0, the better.

You are going on a trip. You think about all the contingencies and you want to carry just exactly enough money for all your trip needs. Same with link power planning, our main goal in doing a link budget is to allocate just enough power to the link and not much extra. The money you start a trip with, can be thought of as the available resource which depletes over the trip. Similarly, we want to figure out what is the minimum E_bN_0 (the spendable resource) we can provide to the signal so that the desired quality measure, the BER, is achieved at the end.

In doing link budgets, we need to deal with two types of E_bN_0. One called the **required** E_bN_0 and an another called the **available** E_bN_0. The available E_bN_0, hence is like the money in your pocket for the trip and the required E_bN_0 is what you actually need to complete your trip. The two should match, with some left over, called **margin** in link budget parlance.

2.2.0.1 Required E_b/N_0

The main goal in doing link budgets is to determine the **available** E_bN_0 of a link and to compare it to the **required** E_bN_0. The link budget analysis for a given link tells us the available E_bN_0 and not the *required* E_bN_0. The required E_bN_0 is determined via an entirely different process, not usually considered a part of doing link budgets. The required E_bN_0 can be determined either by analysis or by using manufacturer provided receiver data. We take a look at both of these methods.

For a QPSK signal in an additive white-Gaussian-noise (AWGN) channel, this equation that can be used to compute the **required** BER is given by

$$BER = \frac{1}{2}\text{erfc}(\sqrt{E_b/N_0}) \tag{2.8}$$

This formula says that the BER of a signal in an Additive White Gaussian Noise (AWGN) environment is related to its received E_bN_0 by the function, *erfc*. The function erfc, called the *complimentary error function* describes the cumulative probability distribution of a Gaussian distribution. This formula applies only for estimating the BER in Gaussian channels. A majority of channels are not purely Gaussian and the BER is actually a lot worse than what is estimated by this expression. In Fig. 2.12, we see how much larger is the required power in fading channels. When coding is used, the The E_bN_0 is something different again. Unfortunately, there are no closed equations such as Eq. 2.8 for coded links and we have to use empirically derived data. The E_bN_0 needed to achieve a desired

BER (based on the appropriate analytical expression or test data for that type of link) is what we call the ***required*** E_bN_0 for the link. It is a hard requirements and it is this number we try to satisfy in our link design.

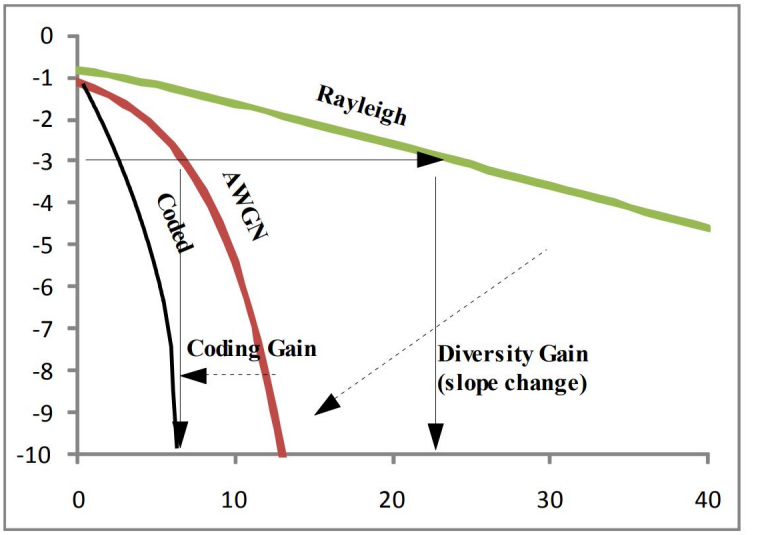


Figure 2.12: The BER under AWGN vs. fading channels.

The Eq. 2.8 plotted as a function of E_bN_0 has what is often called the *waterfall* shape when plotted on a log-log scale as shown in Fig. **??** and Fig. 2.13 (c). The AWGN channel forms the best possible estimate for the required The required E_bN_0 and is sort of an ideal limit, often used as a benchmark. However, it can be improved on by using error correction on the baseband data. The BER is inversely related to E_bN_0 as we see in all three of these link types. Notice that the BER curves are always given with E_bN_0 as the independent parameter on the x-axis. This allows us to compare all sorts of links, with high or low power. However, each curve applies uniquely to one type of modulation only in the given channel type.

Alternately, the receiver manufacturers also provide data on required E_bN_0 for various code sets, usually for AWGN channels with random and burst error patterns. These numbers are used to set the required E_bN_0 benchmark, to be met by the design. The required E_bN_0 in manufacturer data is given as a function of the modulation and coding. We see in table below, a list of required E_bN_0 values for a QPSK signal using code sets of various rates to achieve a BER of 10^{-7}. This data comes from the ITU standard and can be used if such data for the specific receiver being used is not available. We note in this table that as the code rate increases, the required E_bN_0 increases with it. For example, code rate of 0.5 would have an overhead rate of 50% and hence it requires smaller E_bN_0 of 4 dB vs. the code rate of 0.88, where the overhead rate has decreased to 12% and now the signal requires a 3 dB higher E_bN_0 to achieve the same BER.

Table 2.1: Required E_bN_0 for an AWGN channel at BER $= 10^{-7}$

Modulation	Code rate	Bit Eff.	Req. E_sN_0	Req. E_bN_0
QPSK	1/4	0.49	-2.3	0.80
QPSK	1/3	0.66	-1.24	0.59
QPSK	2/5	0.79	-0.3	0.73
QPSK	1/2	0.99	1	1.05
QPSK	3/5	1.19	2.23	1.48
QPSK	2/3	1.32	3.1	1.89
QPSK	3/4	1.49	4.03	2.31
QPSK	4/5	1.59	4.68	2.67
QPSK	5/6	1.65	5.18	2.99
QPSK	8/9	1.77	6.2	3.73
QPSK	9/10	1.79	6.42	3.89
8PSK	3/5	1.78	5.5	3.00
8PSK	2/3	1.98	6.62	3.65
8PSK	3/4	2.23	7.91	4.43
8PSK	5/6	2.48	9.35	5.41
8PSK	8/9	2.65	10.69	6.46
8PSK	9/10	2.68	10.98	6.70
16APSK	2/3	2.64	8.97	4.76
16APSK	3/4	2.97	10.21	5.49
16APSK	4/5	3.17	11.03	6.03
16APSK	5/6	3.30	11.61	6.42
16APSK	8/9	3.52	12.89	7.42
16APSK	9/10	3.57	13.13	7.61
32APSK	3/4	3.70	12.73	7.04
32APSK	4/5	3.95	23.64	17.67
32APSK	5/6	4.12	14.28	8.13
32APSK	8/9	4.40	15.69	9.26
32APSK	9/10	4.45	16.05	9.56

$E_bN_0 = E_sN_0$ - 10 log10 (Bit eff.)

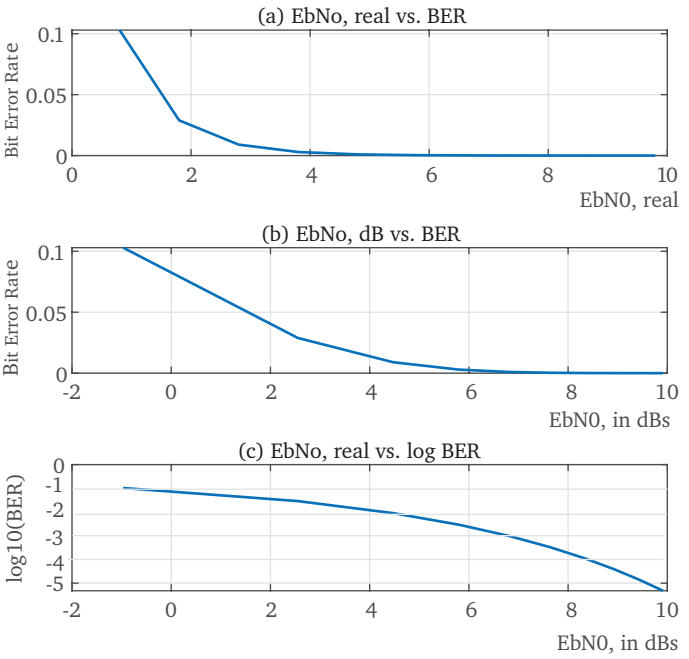


Figure 2.13: The actual Bit Error Rate of a QPSK uncoded link is a function of its available E_bN_0. (a) BER vs. the E_bN_0 in real terms. (b) BER vs. E_bN_0 in dBs (c) log BER and E_bN_0 in dBs.

2.2.0.2 Available E_b/N_0

When we are doing link budgets, we are computing the **available** E_bN_0. This consists of examining the injected power into the link at various points in the link. It also requires that we estimate the noise and distortion generated by the various elements along the way. Distortions reduce the net power. The net power is then converted to the term **available** E_bN_0.

The power is injected into the link by the amplifiers and antennas. That's pretty much it, as far as the power input is concerned. This is our spending currency. A signal spends *power*. The rest are the expenditure of power, also called **attenuation** and **degradation**. We can think of *attenuation* and *degradation* like this: if the signal is subject to some sort of *scale-changing* phenomena, then this is considered a linear change and we call it ***power attenuation***. If the phenomena causes a non-linear change in the *shape* of the signal, i.e. phase or power-related amplitude changes, then this is called ***power degradation***.

In time domain, a signal *shape change* is called **distortion**. Its power domain effect is called **degradation**. A linear effect, causing only *scale change*, is often called **attenuation** in both the time domain and the power domain. An example of an attenuation is path loss in a line of sight link. This particular effect just attenuates the signal amplitude.

Degradation on the other hand, are often a function of the power level. An example is high power amplifiers, which can cause both the amplitude *and* the phase *distortions*. For HPAs the extent of the distortion is usually a function of the amplifier back-off. Two main types of degradation coming from the HPAs are **modulation loss**, **linearizer effects** and the **inter-modulation (Im) loss**.

In Fig. 1.5, we see how for a typical satellite line of site, the power is added by the amplifiers at the ground and then again at the satellite. In between the power is being spent due to linear and non-linear effects.

So the process of doing link budgets is to first obtain the required E_bN_0 for a given specified BER, modulation and code rate, and then figure out a way to provide this to the link. The main goal at the end is to have enough margin, i.e the available E_bN_0 should be larger than the required E_bN_0 by a certain quantity, say a 1 or 2 dB margin.

We give here the general equation for computing the delivered power of a link. It contains plus terms for power and minus terms for losses and degradations. This process is called budgeting. We will discuss how to compute each of these terms in detail in chapters 3 and 4 and how to convert the net power to the available E_bN_0. This equation tells us that the power is added by the transmitter via the quantity EIRP and the receiver gain, via the term G_r. Then we subtract the various degradations or losses as appropriate for the link. The result is the net power received.

$$\text{Net power received} = EIRP + G_r - \underline{\text{Losses}} \tag{2.9}$$

Where the losses in above expression include these terms.

$$\begin{aligned}\text{Losses} = \text{Fee space loss} + \text{PathLoss} + \text{Modloss} + \text{HPALoss} + \text{AtmosLoss} \\ + \text{PolarizationLoss} + \text{FilterLosses} + \text{Backoff} + \text{OtherLosses}\end{aligned} \tag{2.10}$$

Coded links

For most systems, the available power for transmission tends to be at a premium. This impetus has led to nearly all the advancements in communications and signal processing. However, if power is pulled back, the bit error rate increases. **Error correction coding** (ECC) allows us to improve the BER for the same power. However, ECC comes at a cost of reducing the throughput. Coding involves adding overhead bits to the information signal. Since the symbol rate (and bit rate) capacity for a link is pretty much a hard limit, we displace information bits with overhead bits and end up reducing the net information rate. Hence, the improvement in BER comes at a cost of information rate reduction. Not only that, we also increase complexity of the design by adding a coder/buffer at transmit and a decoder/buffer at the receive side. Hardware complexity is of course an important part of the link development process.

There are two main categories of coding, 1. forward error correction (FEC) and 2. bi-directional of ACK/NCK form of coding. In the FEC form, coding adds extra bits to the data packet and the receiver does all the work in figuring out which bits are received incorrectly. There is no communication or information transfer between the sender and the receiver about what is sent. In ACK/NCK, the second form, the receiver checks the received data based on some rule, such as parity bits and, if the packet is found to be in error, it sends a message back to the transmitter, called a NCK message, in which case the transmitter backs up to the corrupted block and resends it. This reduces overhead due to simpler coding (smaller overhead) but slows down the data transfer due to the round trip delays for the acknowledgment messages and the retransmission of packets. Such FEC is used only when delays are small (such as hard disk data transfer) and usually not when data rate is large, such as in wi-fi towers and satellite communications, or when delays introduced would be unacceptable to a customer.

The most common error pattern in communications links are the random bit errors which are uniformly distributed through a long message. A one-bit error here, a two-bit error there etc. The common FEC codes are best for managing these random bit errors. A typical error threshold is $p = .05$ or less and hence, the probability of getting more than one error at a time gets progressively smaller. Parity check codes and convolutional codes are examples of codes that can correct a small number of consecutive errors quite well. They are said to be best with *random errors*.

The other problematic error pattern is a large number of consecutive errors, called **burst errors**. In some cases, these are also called **erasure errors** when distortion results in complete loss of information. The **block codes** can process a large number of bits at once, and are good at detecting and correcting burst errors. An another class of codes are used for carrier acquisition and synchronization. **Barker codes** are an example of this type of coding.

Code sets

For satellite links, the present recommendation is to use concatenation of Reed-Solman, and LDPC codes with interleaving in between. LDPC codes codes rplace the convolutional codes used prior to the current set. Both the LDPC nad the RS codes are considered block codes, which operate on large number of bits at a time. A typical block size is 64,000 bits. This entails delay but the coding gains nearly approach the theoretical possible gains.

The inner code in a set is a LDPC code, which is able to correct both random and burst errors. The RS code on the outside takes care of residual errors which under some conditions cause an error floor.

In Reed Solomon codes, the bits are transmitted in blocks. The blocks are encoded in “symbols” (not at all the same thing as the modulation symbols!). A code symbol

may consist of 8-bits or anything else. A commonly used code encodes 223 eight-bit data symbols plus 32 eight-bit parity symbols into a 255-symbol block. This is written as a RS(255, 233) code. It can correct up to 16 symbol errors per block, which translates to app. 16 errors out 256 bits. Because RS codes can correct a whole symbol consisting of 8 bits at a time, it makes them well-suited for channels where errors occur in bursts. BCH is one algorithm for decoding the RS codes.

The conceptual idea is to encode each block with a given polynomial. As the polynomial is known to the decoder, the decoder can then correct the whole group of bits (a symbol) if they deviate from the expected polynomial.

Whereas in RS coding, we have an polynomial (which is really an equation) for each block, what if we just send, not 255 equations, but a smaller number on random bits in the message, say only on $x_2, x_{72}, x_{213}, x_{244}$ bits. The message length of this polynomial can also be variable, sometimes 10 bits and sometimes only 3, unlike for RS code where it is always equal to the length of the block. This makes LDPC codes "low density". The decoding is faster and surprisingly works just as well, and in fact better than RS codes. Together this code set, approaches near-theoretical bounds.

Coding Gain

Let's take a look at the three curves in Fig. 2.14. The plot line 1 is for an uncoded QPSK signal, while the plot line 3 is for the same signal using a code. The coding moves the BER curve to the left. This is always so or otherwise there would be no point in using coding. What this says is that we can get the same BER, along the horizontal line, but for a smaller . Of course what it does not say that this has also reduced the effective data rate. But we won't worry about the reduced data rate, we can live with it.

We see in Fig. 2.14 that for a desired BER of 10^{-6}, an uncoded QPSK signal requires an of app. 10.6 dB. But, with coding, the requirement decreases to 3.8 dB. This difference between uncoded and coded requirements is called the **coding gain**. We define the term ***coding gain*** at the desired BER as

$$\text{Coding Gain} = \text{Uncoded } E_bN_0 - \text{Coded } E_bN_0 \tag{2.11}$$

The coding gain is a function of the BER level chosen. So before we calculate the coding gain we need to know what BER level is desired for the data. Most commonly a BER of 10^{-3} is chosen for voice and 10^{-9} for data links. However, we need to be clear about whether the BER under discussion is for the uncoded channel or the coded channel.

The coded BER curves (plot lines 3 and 4) shown below are usually developed by testing or simulation. Determining these by analysis is often an intractable problem. There are no closed form equations we can use to come up with these curves for a given

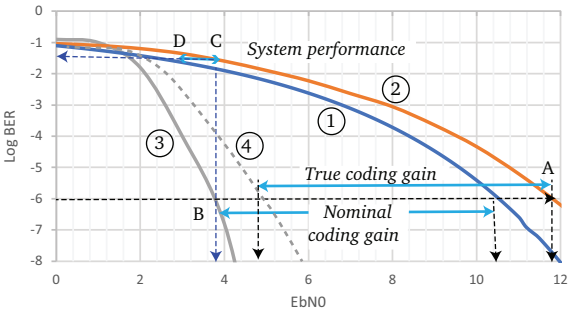


Figure 2.14: vs. BER. Plot line 1 is the required E_bN_0 in an AWGN channel with ideal hardware. The plot line 2 shows the E_bN_0 required an actual system containing degradation in addition to the AWGN channel. Plot line 3, gives the relationship in an ideal AWGN channel using coding. This plot line is a complement to plot line 1. Plot line 4 is the true performance of the real hardware with coding and is the complement of plot line 2.

code set. There are many published sources for vs. BER for various types of codes that can be used when estimating the coding gain. Those produced by the companies that build code chips are of greater validity than those produced or included in academic papers and books, including this one.

The effect of coding on E_bN_0 The main effect of coding is, of course, on the quality of the signal. With coding, we can achieve nearly error-free transmission for a lot less Eb/N0. Let's compare a signal with and without coding requiring a information bit error rate of 10_5. A QPSK signal in AWGN channel will require a E_bN_o of 9.6 dB to deliver this BER. We find that we can achieve the same BER using coding (depending on the code rate) with a E_bN_0 any where from 2 dB to 8 dB, resulting in a coding gain as large as 7.6 dB to as small as 1.6 dB, depending on the throughput loss we are willing to accept. The coding gain, hence, is defined as the difference between what E_bN_0 is needed for an uncoded signal vs. what is needed with a coded signal for the same BER.

In doing link budgets which is covered in the next chapter, the main goal is to make available to the link a E_bN_0 that will result in the desired BER for a given application. Note that this is the final BER, and not the input channel bit error rate, which we will denote as **cBER** (with *c* for the channel) to differentiate it from the information bit error rate, written as plain BER.

Setting a target BER

Often the desired BER for a TV signal is somewhere around 10^{-6}, whereas data signals need at least 10^{-8}, and voice signals around a minimum of 10^{-2}. What is the significance of these BER levels? First of all, not all types of communications require a super clean signal. Phone calls can get by with a fair amount noise and distortion, as our brain, the best error correction device ever made, is able to take a signal that has a BER of

app. 10^{-1} and still make out what the other person is saying. Hence, cell phones can get by with low E_bN_0 when in a phone call mode. A threshold of satisfaction is established for each application by the system owner, depending on the fussiness about the quality of the signal desired and even more importantly, the desire to keep the device size small. We see this relationship in Fig. 2.14. Plot line 1 is the ideal performance of a link in a AWGN channel computed from Eq. 2.8. In plot line 2, we see the BER vs. E_bN_0 of a real system which has many hardware components such as non-linear amplifiers, filters etc. We see degradation from the ideal as the horizontal distance between point C and point D for a system designed to provide a channel BER of app. $10^{-1.5}$. This degradation is a function of the parameters of the link such as amplifier operating point, pulse shaping, noise figure of hardware used, phase error specifications, single vs. multi-carrier frequency usage and many others. This is computed by the system designer usually via time-domain simulation using Monte Carlo methods.

In Fig. 2.14, plot line 3 shows what the required E_bN_0 would be if coding is used. This plot line is based on a code set of a convolutional code of rate 1/2 with RS. All code sets have their own individual behavior and would be different than what is shown here. Plots for each code set are developed by the code manufacturing companies and are available from them. One can also generate them by simulation but this is not an easy task. Note that all such data available from the manufacturer assumes (in similar fashion as the plot line 1) that the system itself is ideal. The only sources of error simulated are Gaussian noise, and some form of simulated environment representing burst errors. If we did a simulation for the actual system with real hardware using this same code set, we might get the plot line shown by 4. We see that it no longer provides the same performance as the one advertised by the manufacturer. The coding gain according to most text books and manufacturers is the distance between the ideal plot lines and not between those for your particular system. This they don't know and leave it up to the system designer to figure out. However, the two numbers, the true coding gain as shown here and the nominal or ideal coding gain are pretty close, with true coding gain always some what smaller.

If we need a BER of 10^{-6}, then from Fig. 2.14, we need a received E_bN_0 of 11.9 dB (vs. 10.5 dB for an ideal system). (See end of nominal coding gain line.) Assume that we are using a hypothetical code set, which tells us that we can obtain the same BER for a E_bN_0 of only 3.9 dB, a saving of app. 8 dB, point B. This saving is called the **nominal coding gain**. However it comes at a cost that the effective bit rate is only half what it might have been.

In doing link budget we need to be clear about the difference between the channel BER (cBER) and the information BER (iBER), as these are a function of the code set selected. The channel BER is not a number any one actually measures. Since the signal, up until the receiver, is an analog signal, this BER does not actually exist, it is just a

conceptual idea that is used to decide where to perform the system simulation in order to determine the actual system degradation and coding gains.

Relationship of channel BER and information BER

We take a certain data stream and add overhead bits to it, as created by the code being used. The rule is that the lower the code rate, the larger the overhead. A code rate of 1/3 will produce 3 bits for every one info bit. Two bits out of three are hence overhead. A code rate of 7/8 will produce just one overhead bit for 7 info bits. The performance of these codes is related to these rates. A lower code rate will produce an info bit stream after decoding withe overhead, the better the final BER. But we do have here *two* different BERs. Let us take a data stream, which contains both information and overhead bits. When this data stream goes through the channel at a given E_bN_0, it suffers from errors, for a AWGN channel, by the relationship given by Eq. For a signal transmitted with a 3 dB E_bN_0, will result in a BER of app. $10^{-1.5}$. Assume we used a particular code set and hence the receiver is able to clean up this signal even further to 10^{-10}. Hence the first BER is the ***channel BER*** and the second the **information BER**. But, we think, that the information BER is too clean, we don't need that much. So we allow the channel BER to be a bit higher by reducing the SNR. But, how do we know the final relationship between the channel BER and the information BER? This is a very difficult question to answer as each code set behaves differently and usually non-linearly. Empirical experience tell us that for satellite likes this number is around 0.01 and 0.05 for most wifi terrestrial links.

Nyquist, Shannon and the information carrying capacity of signals

As far as a communications engineer is concerned, information is defined as a quantity called a *bit*. This is a pretty easy concept to intuit. A yes or a no, in or out, up or down, a 0 or a 1, these are all a form of information bits. In communications, a bit is a unit of information that can be conveyed by some change in the signal, be it amplitude, phase or frequency. We might issue a pulse of a fixed amplitude. The pulse of a positive amplitude can represent a 0 bit and one of a negative amplitude, a 1 bit.

Let's think about the transmission of information along the concepts of a highway as shown in Fig. 2.15. The width of this highway which is really the full electromagnetic spectrum, is infinite. The useful part of this highway is divided in lanes by regulatory bodies. We have very many lanes of different widths, which separate the multitude of users, such as satellites, microwave towers, wifi etc. In communications, we need to confine these signals in lanes. The specific lanes (i.e. of a certain center frequency) and their widths are assigned by ITU internationally and represent the allocated bandwidth.

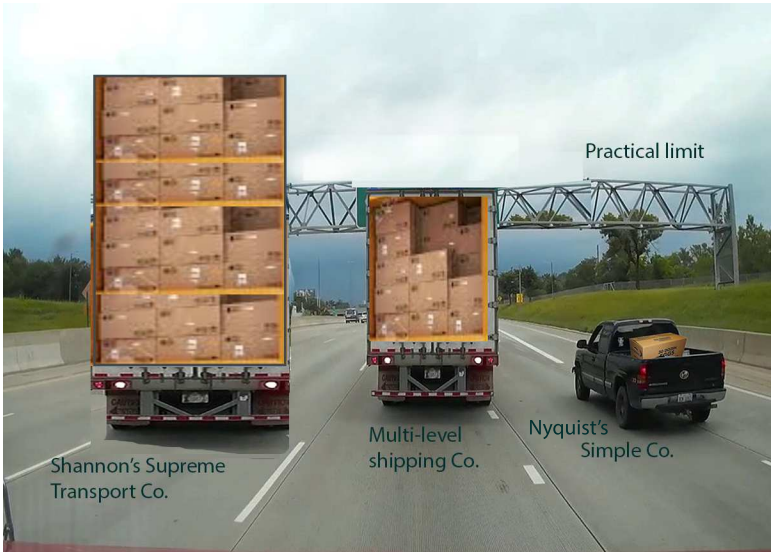


Figure 2.15: The information highway

Within this width, the users can do what they want. In our highway analogy, we want our transporting truck to be the width of the lane, or in other words, we want the signal to fill the whole allocated bandwidth. But just as in a real highway, guard bands or space between the lanes is often required.

The other parameter in this conceptual framework is the quality of the road, which we equate with the amount of noise encountered. This noise is anything that can disturb a signal. If the road is rough, it can toss-out our cargo, or cause bit errors, in signal-speak. Or it can be so smooth that nothing gets damaged, and we are home free.

Then we have the issues of packing our cargo in an efficient manner on the truck. The Nyquist theorem gives us our first cut at this number and tells us how much we can carry. But from Hartley's theorem we learn that we can actually carry a lot of stuff, if we just pack it smartly. We call this *smart packing* a form of **multi-level signaling**. The simplest case is just one layer and that is what is the baseline assumption in the **Nyquist limit**. We should be able to pack our cargo as high as we want as long the noise does not knock the whole pile down. This stacking idea increases the total number of bits we can transmit, or the throughput, but of course, makes the cargo more susceptible to road surface problems. The higher the stack, the easier it is for it to fall down. Hence multi-level signaling is more sensitive to noise.

Then there is the issue of the size of the *truck engine*. We can equate that with the carrier power, S. We can tell intuitively that a multilevel shipping scenario will require a higher power engine. The ratio of the power and noise is the term, SNR.

Using four parameters of a channel, its *bandwidth*, the *efficiency of stacking*, the *noise* likely to be encountered, and the the engine *power*, we can now discuss **channel capacity**, i.e. the ability to transport a certain number of bits over this highway in a given channel without upsetting too many of the cargo bits. Shannon's theorem gives us a absolute limit for any SNR. It is considered akin to the speed of light. But just as knowing the speed of light and building a rocket that can actually do it are two different problems, this theorem also tells us nothing about how we might achieve such a capacity. Many practical obstacles stand in our way and we can rarely achieve Shannon's capacity.

Nyquist Rate

We first state the simplest case as theorized by Nyquist. In 1927, Nyquist developed a thesis that in order to reconstruct a signal from its samples, the analog signal must be sampled at least two times its highest frequency. From this comes the ideas that the maximum digital data rate we can transmit is no more *symbols* than two times per Hz of bandwidth. Nyquist relationship assumes that you have all the power you need to transmit these symbols and the channel experiences no noise and hence suffers no errors. This concept also does not tell us what type of signal we should use. It is obvious we can not use a pulse. The reason is that a narrow pulse has a very large bandwidth. This case implies that whatever method we choose to transmit the symbol, it must fit inside the allocated bandwidth of B Hz.

What kind of signal can we use? Can we use a square pulse such as the one shown in Fig. 2.16(a)? Perhaps, but the problem is that if we take the Fourier transform (FT) of a square pulse, what we get is a very wide sinc function that does not decay quickly. FT of a sinc pulse alternately is a square pulse, then why not use the sinc pulse as our time-domain signal which confines it to the bandwidth B. Problem solved! However, sinc pulses are hard to build in hardware and a variant called root-raised cosine pulses are often used. But assume that we do use the sinc pulse, then the ideal data rate possible for a signal of low-pass bandwidth, B is as given by the Nyquist theorem here.

$$R_s = 2B_l \text{ Low-pass} \tag{2.12}$$

$$R_s = B_b \text{ Band-pass} \tag{2.13}$$

It may appear from the equation above that a lowpass signal has higher capacity than a bandpass signal given the same bandwidth. But this confusion comes from the fact that frequency is defined *always to be a positive quantity*. When a signal spectrum is around the zero frequency, as it is for a lowpass signal, only the positive half of the spectrum, B_l, is used in the above calculation. When a signal is centered at a higher frequency, the whole spectral range, B_p, is used. The bandpass bandwidth is twice the lowpass bandwidth for most digital signaling. The spectral content is exactly the same for both definitions, and hence, so is the symbol capability. Nyquist states that we can send one

symbol for every one Hz of (bandpass) bandwidth. The key word here is *symbol*. By allowing the symbol to represent more than one bit such as QPSK, 8PSK etc., we can do better. This observation and its incorporation into the Nyquist rate is called the Hartley theorem. About a year after Nyquist formulated the limit, Hartley, using a previously available idea that a generic symbol could in fact represent more than one bit, modified the Nyquist limit by the addition of a multi-level factor in the capacity equation.

In describing a multi-level signaling, we use two terms, M and n. The term M, in Eq. (2.14) is the number of alternative symbols that a receiver will have to read and discriminate. We can also think of this number as the *levels of a signal* for description purposes. Each symbol represents $N = log_2 M$ bits. n is also called the **modulation efficiency** or **bit efficiency** and is measured in bits/symbol.

$$\begin{aligned} R_b &= B_b \ log_2(M) \\ R_b &= B_b N n \\ \eta_B &= \frac{R_b}{B_b} = n \end{aligned} \tag{2.14}$$

In Fig. 2.16 we see a two level signal. A two-level signal with$M = 2$, and $n = 1$ is one we know as a BPSK signal. A four-level signal with $M = 4$ and $n = 2$, is a QPSK signal.

At a minimum M is 2 and we can increase it as powers of 2. However, n, only changes as a function of log_2 of M. A 8-level ($M = 8$) vs. a 4-level ($M = 4$) signal only increases the data rate capacity from 2 times the bandwidth to 3 times and a 16-level signal, raises it from 3 times to 4 times the bandwidth.

The term M, is called the **modulation order**. Modulations that use M $>$ 4 are often called **higher-order modulation**. We went from a generic idea of capacity in *symbols* in Eq. (2.13) to capacity in bits/sec in Eq. (2.14). The term M in this equation brings a finer resolution to information and allows us to form more complex symbols that can represent any number of n bits to increase the throughput. When we take the bandwidth to the other side, then the term R_b/B_p is called the **spectral efficiency** with units of bps/Hz, denoted by η. It is same as the term, n.

$$\eta_B = \frac{R_b}{B_b} = n \tag{2.15}$$

The Nyquist capacity is for a *single* noiseless or a single-in, single-out (SISO) channel. Multi-in, multi-out (MIMO) is a higher dimension case which can raise capacity in fading channels. Its most common form shows it as for the low-pass bandwidth case. This is

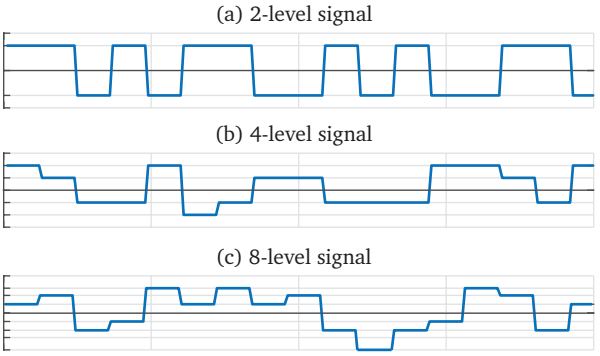


Figure 2.16: A multi-level signal can take on M levels of both amplitudes and phase. (a) This signal takes on only two amplitude levels and is known as a BPSK signal. (b) This signal takes on four different amplitudes and is known as a QPSK signal. (c) This signal takes on 8 levels and can be thought of as a 8PSK signal, although a true 8PSK signal is built using a complex signal.

often a source of confusion. With the use of a larger number of symbols, noise will add uncertainty to the true level of the signal, making it harder for a receiver to discriminate between the large number of symbols. With current technology, presently it is very difficult to go beyond a factor of 1024.

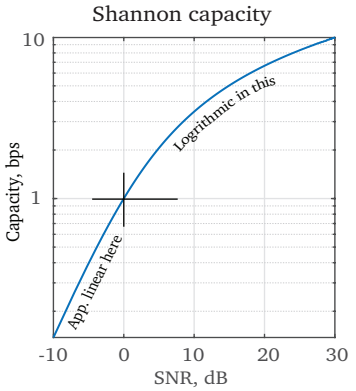


Figure 2.17: Shannon capacity in bits/s as a function of SNR. It has two ranges, the one below 0 dB SNR and one above. For SNR > 0, the limit increases slowly.

What Shannon said

Now we come to Shannon's theorem, developed around the same time as Hartley and is often jointly called the **Shannon-Hartley theorem**. The theorem sets out the maximum data rate that may be achieved in any channel when noise is present. Shannon gives this as the maximum rate at which data can be sent *without errors*. This rate, called C in Eq. (2.16), is given in bits per second and is called the **channel capacity**, or the **Shannon**

capacity. The two main independent parameters are the bandwidth (equivalent to the bandpass bandwidth) in Hz, and SNR as a non-dimensional ratio of signal power and the total noise power in that bandwidth. As far as we know, this rate represents a physical limit, similar to the speed of light as being a limit for how fast things can move. This is capacity in a noisy and bandwidth-limited channel, and is always a larger number than the Nyquist, which is a surprising result.

The noise behavior assumed in this expression, is additive white Gaussian noise (AWGN). Although in a majority of channels, such as wi-fi, the noise type is much more destructive than AWGN, the equation gives a way to estimate what is ideally possible. Under non-AWGN cases, the physical limit on data rate is likely much smaller than the Shannon limit.

$$C = B\ log_2(1 + \text{SNR}) \tag{2.16}$$

Symbol and bit rate capacity of a channel

The Shannon limit is a comprehensive relationship in that it takes into account three of the four most important parameters, the bandwidth, the carrier power and the noise level. It does not account for signal levels because it is already in terms of bits of information. The maximum level M can in fact be calculated by equating the Nyquist rate with the Shannon capacity for a given SNR and bandwidth.

In Fig. 2.17, we see the Shannon limit plotted as a function of the channel signal to noise ratio. This figure is plotted assuming a channel bandwidth of 1 Hz. The x-axis is in terms of signal power to noise ratio in dBs or as log_{10}SNR. The y-axis gives the maximum bit rate possible for a signal of bandwidth 1 Hz. By normalizing the bandwidth, we write an alternate form of the capacity limit called the spectral efficiency, η_B as

$$\eta_B < log_2(1 + \text{SNR}) \tag{2.17}$$

In Fig. 2.17, we note that as SNR increases, there are two ranges of behavior. At very low SNR, the spectral efficiency increases linearly with increasing power, but slows down to a logarithmic behavior for high SNR. Hence increasing SNR brings diminishing returns. However, this graph plotted from Eq. (2.17) is misleading. The capacity does not decrease linearly in a SNR range below 0 dB. An alternate behavior actually takes place. As SNR decreases, the effect of noise increases rapidly. We can write the SNR in this form, where N_0 is noise density, total noise power is noise density times the bandwidth and P_s is the received signal power. We convert SNR to $E_B N_0$, an alternate measure of signal that is independent of bandwidth and measures the distribution of total power to individual bits in the signal.

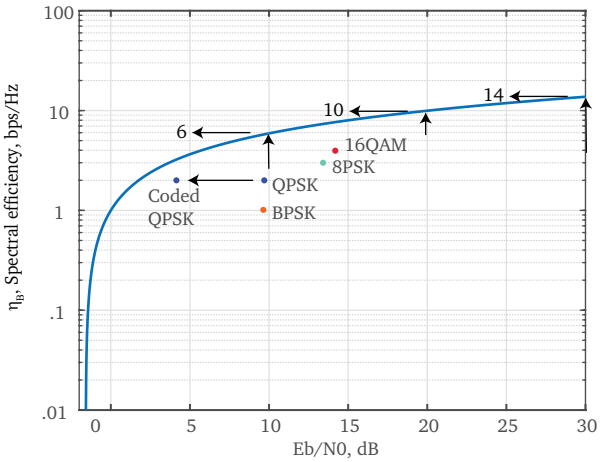


Figure 2.18: Shannon capacity limit

$$\begin{aligned} P_n &= N_0 B \\ SNR &= \frac{P_s}{P_n} = \frac{P_s}{N_0 B} \end{aligned} \tag{2.18}$$

We rewrite Eq. (2.17) by setting $P_s = E_b R_b$, where R_b is the data rate. The data rate R_b when divided by the bandwidth is the bit efficiency, we defined in Eq. (2.17).

$$\eta_B < log_2(1 + \frac{E_b}{N_0}\frac{R_b}{B}) = log_2(1 + \eta_B \frac{E_b}{N_0}) \tag{2.19}$$

Now we write this equation by taking log of both sides.

$$\frac{E_b}{N_0} \geq \frac{2_B^{\eta} - 1}{\eta_B} \tag{2.20}$$

Letting the efficiency go to zero, we get the following limit.

$$\frac{E_b}{N_0} \geq \lim_{\eta_B \to 0} \frac{2_B^{\eta} - 1}{\eta_B} = ln(2) = -1.59 \text{ dB} \tag{2.21}$$

Hence we plot the full Shannon capacity curve in Fig. 2.18 for a digital signal using $E_b N_0$ as the signal parameter using Eq. (2.21). There is hard limit on the left at -1.6 dB beyond which no communication is possible. We see that it takes ten times the power (from 10 dB to 20 dB) to increase the capacity by only two-thirds, from 6 to 10. It takes a 100 times increase in power (from 10 dB to 30 dB) to approximately double the rate from 6 to 14.

Under the Shannon capacity curve, in Fig. 2.18, there are pointers at the operational points of various PSK modulations. QPSK, for example requires a E_bN_0 of 9.6 dB to provide a BER of 10^{-5}. This is approximately, 7 to 8 db away from the Shannon limit (horizontally). However, if a code, such a RS/LDPC set is used, the same signal can be transmitted at the same BER with a much lower E_bN_0 of app. 3-4 dB, taking us much closer to the Shannon capacity. Other codes can do even better. This is true for all higher order modulations. They are all, in uncoded form, typically about 7-10 dB away from the Shannon capacity but with ever improving coding technology, can be brought within a dB or less of the Shannon capacity number.

Hence we see that the capacity of a digital signal is first constrained by the Shannon-Hartley theorem and then by the Nyquist theorem. The Nyquist theorem sets the first limit which we often use as a starting point and then we try to reduce the distance from it to the Shannon capacity using either by coding or increasing the levels of signaling.