Nyquist, Shannon and the information carrying capacity of signals

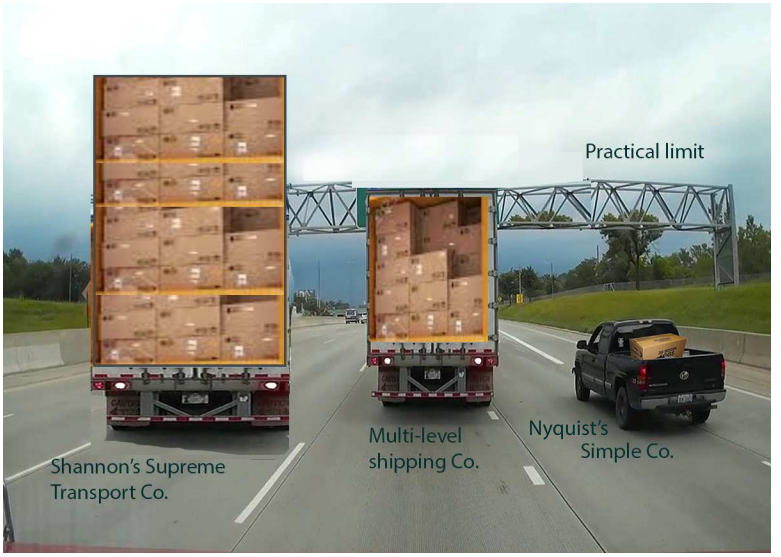

Figure 1: The information highway

There is whole science called the information theory. 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. 1. 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. 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, gaurd 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 assumed 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.

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 it the term, SNR.

Using these 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.

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(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 well. FT of a sinc pulse alternately is a square pulse, then why not use the sinc pulse as our time-domain signal and 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 data rate possible in a signal of low-pass bandwidth, B is as given by the Nyquist theorem here.

$$\begin{aligned} R_s &= 2B_l \ \text{ Low-pass} \\ R_s &= B_b \ \text{ Band-pass} \end{aligned} \tag{1}$$

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. An example is the bit rates used in satellite communications. A transponder of bandwidth 36 Mhz is assumed to be able to transmit at most 36 M symbols/s. When QPSK is used, this number doubles to 72. However, the roll-off of a root-raised cosine pulse knocks this number down by another 25% to app. 54 Mbps.

Nyquist rate tells us that we can send 1 symbol for every Hz of (bandpass) bandwidth. The key word here is *symbol*. By allowing the symbol to represent more than one bit, 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 and modified the Nyquist limit by the addition of a multi-level factor.

In describing a multi-level signaling, we use two terms, M and N. The term M, in Eq. (2) 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.

$$\begin{aligned} R_b &= B_b \ log_2(M) \\ R_b &= B_b N \\ \eta_B &= \frac{R_b}{B_b} = N \end{aligned} \tag{2}$$

In Fig. 2 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. (1) to capacity in bits/sec in Eq. (2). 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.

$$\eta_B = \frac{R_b}{B_b} = N \tag{3}$$

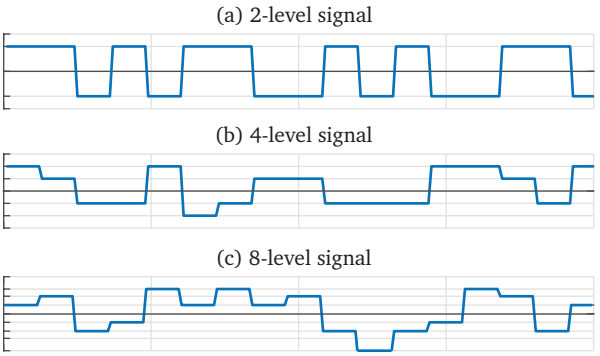

Figure 2: A multi-level signal can take on M levels of both amplitudes and phase. (a) This signal takes on only two amplitude levels and is known as a BPSK signal. (b) This signal takes on four different amplitudes and is known as a QPSK signal. (c) This signal takes on 8 levels and can be thought of as a 8PSK signal, although a true 8PSK signal is built using a complex signal.

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.

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, it is very difficult to go beyond a factor of 1024.

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

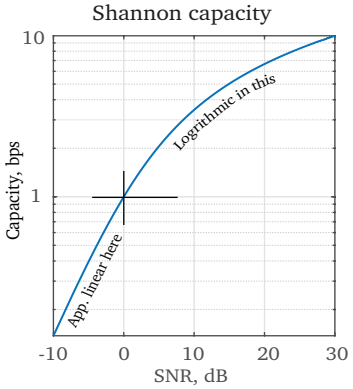

Figure 3: Shannon capacity in bits/s as a function of SNR. It has two ranges, the one below 0 dB SNR and one above. For SNR > 0, the limit increases slowly.

C in Eq. (4), 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 large 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.

$$C = B\ log_2(1 + \text{SNR}) \tag{4}$$

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 infact be calculated by equating the Nyquist rate with the Shannon capacity for a given SNR and bandwidth.

In Fig. 3, 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 from of capacity limit called the spectral efficiency, η_B as

$$\eta_B < log_2(1 + \text{SNR}) \tag{5}$$

In Fig. 3, 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. (5) 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 contribution of noise increases. 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.

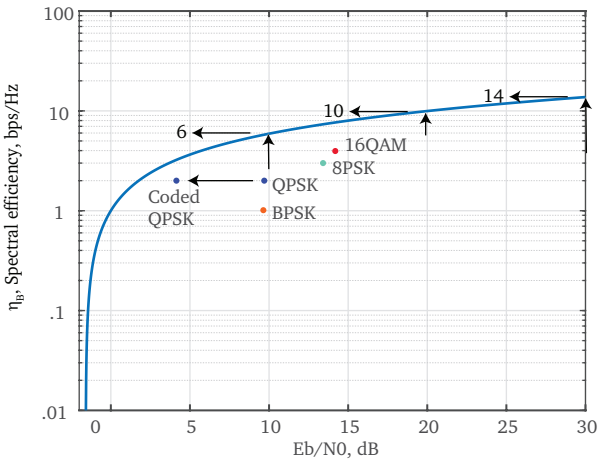

Figure 4: Shannon capacity limit

$$\begin{aligned} P_n &= N_0 B \\ SNR &= \frac{P_s}{P_n} = \frac{P_s}{N_0 B} \end{aligned} \tag{6}$$

We rewrite Eq. (5) 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. (5).

$$\eta_B < log_2(1 + \frac{E_b}{N_0}\frac{R_b}{B}) = log_2(1 + \eta_B \frac{E_b}{N_0}) \tag{7}$$

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

$$\frac{E_b}{N_0} \geq \frac{2_B^{\eta} - 1}{\eta_B} \tag{8}$$

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

$$\frac{E_b}{N_0} \geq \lim_{\eta_B \to 0} \frac{2_B^{\eta} - 1}{\eta_B} = ln(2) = -1.59dB \tag{9}$$

Hence we plot the full Shannon capacity curve in Fig. 4 for a digital signal using E_bN_0 as the signal parameter using Eq. (9). 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. 4, we have put marks 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/LDP 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 coding or increasing the levels of signaling.

 www.complextoreal.com Comments: charldsp@gmail.com