18.8 Channel Capacity

The main challenge in designing the physical layer of a digital communications system is approaching the *channel capacity*. By channel capacity we mean the maximum number of information bits that can be reliably transferred through that channel in a second. For example, the capacity of a modern telephone channel is about 35,000 bits per second (35 Kb/s); it is possible to transfer information at rates of up to 35 kilobits per second without error, but any attempt at perfectly transferring more data than that will surely fail.

Why is there a maximal channel capacity? Why can't we push data as fast as we wish through a digital link? One might perhaps believe that the faster data is transmitted, the more errors will be made by the receiver; instead we will show that data can be received essentially without error up to a certain rate, but thereafter errors invariably ensue. The maximal rate derives from two factors, noise and finite bandwidth. Were there to be no noise, or were the channel to have unlimited bandwidth, there would be unlimited capacity as well. Only when there are both noise *and* bandwidth constraints is the capacity finite. Let us see why this is the case.

Assume there is absolutely no noise and that the channel can support some range of signal amplitudes. Were we to transmit a constant signal of some allowable amplitude into a nonattenuating noiseless channel, it would emerge at the receiver with precisely the same amplitude. An ideal receiver would be able measure this amplitude with arbitrary accuracy. Even if the channel does introduce attenuation, we can precisely compensate for it by a constant gain. There is also no fundamental physical reason that this measurement cannot be performed essentially instantaneously. Accordingly we can achieve errorless recovery of an infinite amount of information per second. For example, let's assume that the allowable signal amplitudes are those between 0 and 1 and that we wish to transmit the four bits 0101. We simply define sixteen values in the permissable range of amplitudes, and map the sixteen possible combinations of four bits onto them. The simplest mapping method considers this string of bits as a value between 0 and 1, namely the binary fraction 0.0101_2. Since this amplitude may be precisely measured by the receiver in one second, we can transfer at least four bits per second through the channel. Now let's try to transmit eight bits (e.g., 01101001). We now consider this as the binary fraction 0.01101001_2 and transmit a constant signal of this amplitude. Once again this can be exactly retrieved in a second and thus the channel capacity is above eight bits per second. In similar fashion we could take the complete works of Shakespeare,

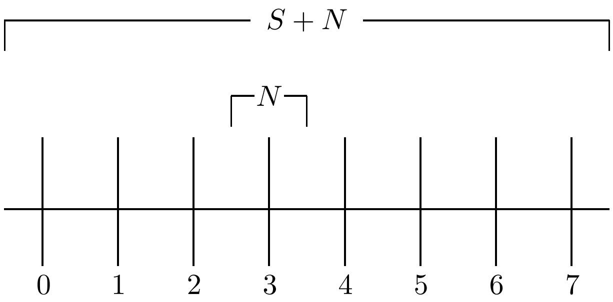

Figure 18.10: The effect of noise on amplitude resolution. The minimum possible spacing between quantization levels is the noise amplitude N, and the total spread of possible signal values is the peak-to-peak signal amplitude S plus the noise N. The number of levels is thus the ratio between the signal-plus-noise and the noise, and the number of bits is the base-two logarithm of this ratio.

encode the characters as bytes, and represent the entire text as a single (rather lengthy) number. Normalizing this number to the interval between 0 and 1 we could, in principle, send the entire text as a single voltage in one second through a noiseless channel. This demonstrates that the information-carrying capacity of a noiseless channel is infinite.

What happens when there *is* noise? The precision to which the amplitude can be reliably measured at the receiver is now limited by the noise. We can't place quantization levels closer than the noise amplitude, since the observed signals would not be reliably distinguishable. As is clarified by Figure 18.10 the noise limits the number of bits to the base-two logarithm of the signal-plus-noise-to-noise ratio, $\text{SNNR} = \text{SNR} + 1$.

Of course, even if the noise limits us to sending b bits at a time, we can always transmit more bits by using a time varying signal. We first send b bits, and afterwards another b bits, then yet another b, and so on. Were the channel to be of unlimited bandwidth we could abruptly change the signal amplitude as rapidly as we wish. The transmitted waveform would be piecewise constant with sharp jumps at the transitions. The spectral content of such jump discontinuities extends to infinite frequency, but since our channel has infinite bandwidth the waveform is received unaltered at the receiver, and once again there is no fundamental limitation that hinders our receiver from recovering all the information. So even in the presence of noise, with no bandwidth limitation the channel capacity is effectively infinite.

Signals that fluctuate rapidly cannot traverse a channel with finite bandwidth without suffering the consequences. The amount of time a signal must remain relatively constant is inversely proportional to the channel bandwidth, and so when the bandwidth is BW our piecewise constant signal cannot vary faster than BW times per second. Were we to transfer an NRZ signal through a noisy finite-bandwidth channel we would transfer BW bits per second. By using the maximum number of levels the noise allows, we find that we can send $\mathrm{BW}\log_2 \mathrm{SNNR}$ bits per second. Slightly tightening up our arguments (see the exercises at the end of this section) leads us to Shannon's celebrated channel capacity theorem.

Theorem: The Channel Capacity Theorem
Given a transmission channel bandlimited to BW by an ideal band-pass filter, and with signal-to-noise ratio SNR due to additive white noise:

• there is a way of transmitting digital information through this channel at a rate up to

$$C = \mathrm{BW}\log_2(\mathrm{SNR}+1) \tag{18.16}$$

bits per second, which allows the receiver to recover the information with negligible error;

• at any transmission rate above C bits per second rate no transmission method can be devised that will eliminate all errors;

• the signal that attains the maximum information transfer rate is indistinguishable from white noise filtered by the channel band-pass filter. ■

As an example of the use of the capacity theorem, consider a telephone line. The SNR is about 30 dB and the bandwidth approximately 3.5 KHz. Since SNR $\gg 1$ we can approximate

$$C = \mathrm{BW}\ \log_2(\mathrm{SNR}+1) \approx \mathrm{BW}\log_2 \mathrm{SNR} = \mathrm{BW}\,\frac{\mathrm{SNR}_{dB}}{10\log_{10}2} \approx \mathrm{BW}\,\frac{\mathrm{SNR}_{dB}}{3}$$

and so C is about 35 Kb/s.

What the channel theorem tells us is that under about 35 Kb/s there *is* some combination of modulation and error correcting techniques that can transfer information essentially error-free over telephone lines. We will see later that V.34 modems presently attain 33.6 Kb/s, quite close to the theoretical limit. There will occasionally be errors even with the best modem, but these are caused by deviations of the channel from the conditions of the theorem, for example, by short non-white noise spikes. The reader who presently uses 56 Kb/s modems or perhaps DSL modems that transmit over telephone lines at rates of over 1 Mb/s can rest assured these modems exploit more bandwidth than 3.5 KHz.

The last part of the capacity theorem tells us that a signal that optimally fills the channel has no structure other than that imposed by the channel. This condition derives from the inverse relation between predictability and information. Recall from Section 5.2 that white noise is completely unpredictable. Any deviation of the signal from whiteness would imply some predictability, and thus a reduction in information capacity. Were the signal to be of slightly narrower bandwidth, this would mean that it obeys the difference equation of a band-pass filter that filters it to this shape, an algebraic connection between sample values that needlessly constrains its freedom to carry information.

The channel capacity theorem as expressed above is limited by two conditions, namely that the bandwidth is filtered by an *ideal* band-pass filter, and that the noise is completely *white*. However, the extension to arbitrary channels with arbitrary stationary noise is (at least in principle) quite simple. Zoom in on some very small region of the channel's spectrum; for a small enough region the attenuation as a function of frequency will be approximately constant and likewise the noise spectrum will be approximately flat. Hence for this small spectral interval the channel capacity theorem holds and we can compute the number of bits per second that could be transferred using only this part of the total spectrum. Identical considerations lead us to conclude that we can find the capacities of all other small spectral intervals. In principle we could operate independent modems at each of these spectral regions, dividing the original stream of bits to be transmitted between the different modems. Hence we can add the information rates predicted by the capacity theorem for all the regions to reach an approximate prediction for the entire spectrum. Let's call the bandwidth of each spectral interval δf, and the signal-to-noise ratio in the vicinity of frequency f we shall denote $\mathrm{SNR}(f)$. Then

$$C = \sum_f \log_2(\mathrm{SNR}(f) + 1)\ \delta f$$

and for this approximation to become exact we need only make the regions infinitesimally small and integrate instead of adding.

$$C = \int \log_2(\mathrm{SNR}(f) + 1)\ df \tag{18.17}$$

We see that for the general case the channel capacity depends solely on the frequency-dependent signal-to-noise ratio.

From the arguments that lead up to the capacity theorem it is obvious that the SNR mentioned in the theorem is to be measured at the receiver, where the decisions must be made. It is not enough to specify the transmitted power at the frequency of interest $P(f)$ (measured in watts per Hz), since for each small spectral region it is this transmitted power times the line attenuation $A(t)$ that must be compared to the noise power $N(t)$ (also in watts per Hz) at that frequency. In other words, the SNR is $P(f)A(t)/N(f)$, and the total information rate to be given by the following integral.

$$C = \int \log_2 \left(\frac{P(f)A(t)}{N(f)} + 1 \right) \, df \tag{18.18}$$

Unfortunately, equation (18.18) is not directly useful for finding the maximal information capacity for the common case where we are given the line attenuation $A(t)$, the noise power distribution $N(f)$ and the total transmitted power P.

$$P = \int P(f) \, df \tag{18.19}$$

In order to find the maximal capacity we have to know the optimal transmitter power distribution $P(f)$. Should we simply take the entire power at the transmitter's disposal and spread it equally across the entire spectrum? Or can we maximize the information rate of an arbitrary channel by transmitting more power where the attenuation and noise are greater? A little thought leads us to the conclusion that the relevant quantity is the noise-to-attenuation ratio $N(f)/A(f)$. In regions where this ratio is too high we shouldn't bother wasting transmitted power since the receiver SNR will end up being low anyway and the contribution to the capacity minimal. We should start spending power where the N/A ratio is lower, and expend the greatest amount of power where the ratio is lowest and thus the received SNR highest.

In other words, we should distribute the power according to

$$P(f) = \begin{cases} \Theta - \frac{N(f)}{A(f)} & \frac{N(f)}{A(f)} < \Theta \\ 0 & \frac{N(f)}{A(f)} > \Theta \end{cases} \tag{18.20}$$

where the value of Θ is determined by the requirement (18.19) that the total Power should equal P. Gallager called this the 'water pouring criterion'. To understand this name, picture the attenuation to noise distribution ratio as an irregularly shaped bowl, and the total amount of power to be transmitted as the amount of water in a pitcher (Figure 18.11). Maximizing signal capacity is analogous to pouring the water from the pitcher into the bowl. Where the bowl's bottom is too high no water remains, where the bowl is low the height of water is maximal.

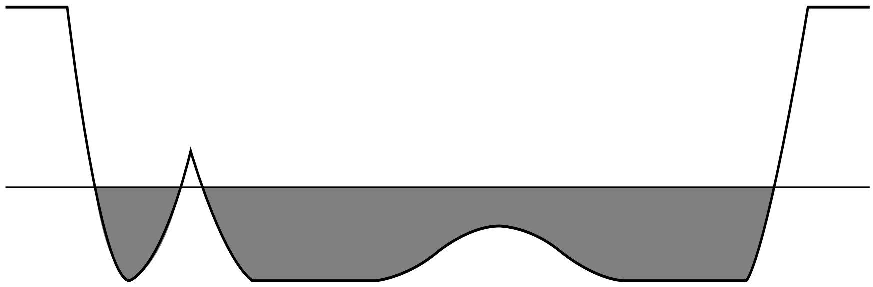

Figure 18.11: The water pouring criterion states that the information rate is maximized when the amount of power available to be transmitted is distributed in a channel in the same way as water fills an irregularly shaped bowl.

With the water pouring criterion the generalized capacity theorem is complete. Given the total power and the attenuation-to-noise ratio, we 'pour water' using equation (18.20) to find the power distribution of the signal with the highest information transfer rate. We can then find the capacity using the capacity integral (18.18). Modern modems exploit this generalized capacity theorem in the following way. During an initialization phase they probe the channel, measuring the attenuation-to-noise ratio as a function of frequency. One way of doing this is to transmit a set of equal amplitude, equally spaced carriers and measuring the received SNR for each. This information can then be used to tailor the signal parameters so that the power distribution approximates water pouring.

EXERCISES

18.8.1 SNR always refers to the power ratio, not the signal value ratio. Show that assuming the noise is uncorrelated with the signal, the capacity should be proportional to $\frac{1}{2}\log_2 SNR$.

18.8.2 Using the sampling theorem, show that if the bandwidth is W we can transmit $2W$ pulses of information per second. Jump discontinuities will not be passed by a finite bandwidth channel. Why does this not affect the result?

18.8.3 Put the results of the previous examples together and prove Shannon's theorem.

18.8.4 When the channel noise is white its power can be expressed as a *noise power density* N_0 in watts per Hz. Write the information capacity in terms of BW and N_0.