Chapter 6

Digital Modulation

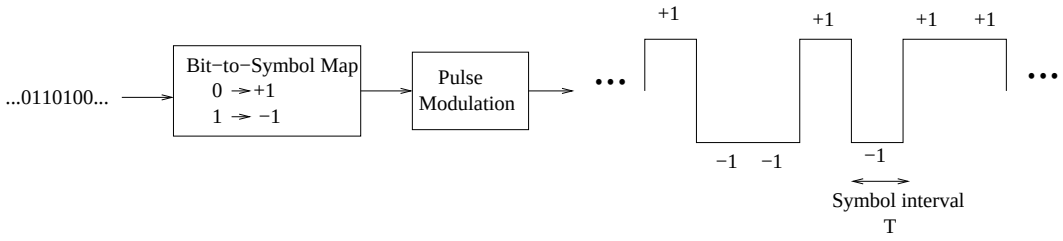


Figure 6.1: Running example: Binary antipodal signaling using a timelimited pulse.

Digital modulation is the process of translating bits to analog waveforms that can be sent over a physical channel. Figure 6.1 shows an example of a baseband digitally modulated waveform, where bits that take values in $\{0, 1\}$ are mapped to symbols in $\{+1, -1\}$, which are then used to modulate translates of a rectangular pulse, where the translation corresponding to successive symbols is the symbol interval T. The modulated waveform can be represented as

$$u(t) = \sum_n b[n]p(t - nT) \tag{6.1}$$

where $\{b[n]\}$ is a sequence of *symbols* in $\{-1, +1\}$, and $p(t)$ is the *modulating pulse.* This is an example of a widely used form of digital modulation termed *linear modulation,* where the transmitted signal depends linearly on the symbols to be sent. Our treatment of linear modulation in this chapter generalizes this example in several ways. The modulated signal in Figure 6.1 is a baseband signal, but what if we are constrained to use a passband channel (e.g., a wireless cellular system operating at 900 MHz)? One way to handle this to simply translate this baseband waveform to passband by upconversion; that is, send $u_p(t) = u(t) \cos 2\pi f_c t$, where the carrier frequency f_c lies in the desired frequency band. However, what if the frequency occupancy of the passband signal is strictly constrained? (Such constraints are often the result of guidelines from standards or regulatory bodies, and serve to limit interference between users operating in adjacent channels.) Clearly, the timelimited modulation pulse used in Figure 6.1 spreads out significantly in frequency. We must therefore learn to work with modulation pulses which are better constrained in frequency. We may also wish to send information on both the I and Q

components. Finally, we may wish to pack in more bits per symbol; for example, we could send 2 bits per symbol by using 4 levels, say $\{\pm 1, \pm 3\}$.

Plan: We first develop an understanding of the structure of linearly modulated signals, using the binary modulation in Figure 6.1 to lead into variants of this example corresponding to different signaling constellations which can be used for baseband and passband channels. We discuss how to quantify the bandwidth of linearly modulated signals by computing the power spectral density. Since the receiver does not know the bits being sent, they can be modeled as random, which implies that the modulated signal is a random process. We compute the power spectral density for this random process in order to determine how bandwidth depends on the choice of modulation pulse and the statistics of the transmitted symbols. With these basic insights in place, we turn to a discussion of modulation for bandlimited channels, treating signaling over baseband and passband channels in a unified framework using the complex baseband representation. We note, invoking Nyquist's sampling theorem to determine the degrees of freedom offered by bandlimited channels, that linear modulation with a bandlimited modulation pulse can be used to fill all of these degrees of freedom. We discuss how to design bandlimited modulation pulses based on the Nyquist criterion for intersymbol interference (ISI) avoidance. Finally, we discuss some other forms of modulation, including orthogonal and biorthogonal modulation.

6.1 Signal Constellations

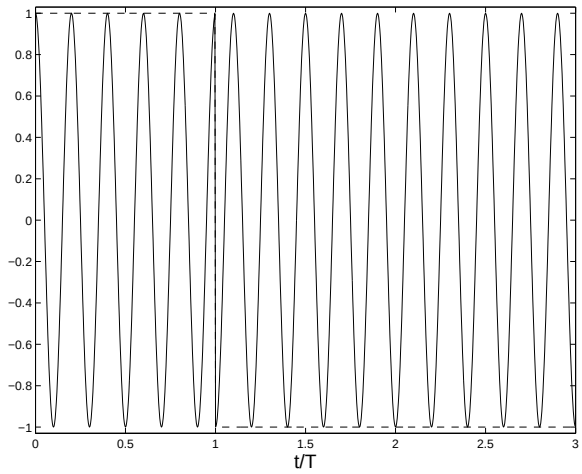


Figure 6.2: BPSK illustrated for $f_c = \frac{4}{T}$ and symbol sequence $+1, -1, -1$. The solid line corresponds to the passband signal $u_p(t)$, and the dashed line to the baseband signal $u(t)$. Note that, due to the change in sign between the first and second symbols, there is a phase discontinuity of π at $t = T$.

The signal in Figure 6.1 is a baseband waveform: while it is timelimited and hence cannot be strictly bandlimited, it is approximately bandlimited to a band around DC. Now, if we are given a passband channel over which to send the information encoded in this waveform, one easy approach is to send the passband signal

$$u_p(t) = u(t) \cos 2\pi f_c t \tag{6.2}$$

where f_c is the carrier frequency. That is, the modulated baseband signal is sent as the I component of the passband signal. To see what happens to the passband signal as a consequence of the modulation, we plot it in Figure 6.2. For the nth symbol interval $nT \leq t < (n+1)T$, we have $u_p(t) = \cos 2\pi f_c t$ if $b[n] = +1$, and $u_p(t) = -\cos 2\pi f_c t = \cos(2\pi f_c t + \pi)$ if $b[n] = -1$. Thus, binary antipodal modulation switches the phase of the carrier between two values 0 and π, which is why it is termed Binary Phase Shift Keying (BPSK) when applied to a passband channel:

We know from Chapter 2 that any passband signal can be represented in terms of two real-valued baseband waveforms, the I and Q components.

$$u_p(t) = u_c(t)\cos 2\pi f_c t - u_s(t)\sin 2\pi f_c t$$

The complex envelope of $u_p(t)$ is given by $u(t) = u_c(t) + ju_s(t)$. For BPSK, the I component is modulated using binary antipodal signaling, while the Q component is not used, so that $u(t) = u_c(t)$. However, noting that the two signals, $u_c(t)\cos 2\pi f_c t$ and $u_s(t)\sin 2\pi f_c t$ are orthogonal regardless of the choice of u_c and u_s, we realize that we can modulate both I and Q components independently, without affecting their orthogonality. In this case, we have

$$u_c(t) = \sum_n b_c[n]p(t-nT), \quad u_s(t) = \sum_n b_s[n]p(t-nT)$$

The complex envelope is given by

$$u(t) = u_c(t) + ju_s(t) = \sum_n \left(b_c[n] + jb_s[n]\right)p(t-nT) = \sum_n b_[n]p(t-nT) \tag{6.3}$$

where $\{b[n] = b_c[n] + jb_s[n]\}$ are complex-valued symbols.

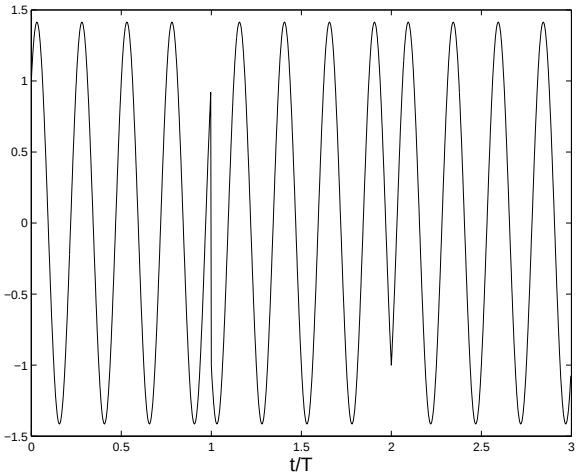


Figure 6.3: QPSK illustrated for $f_c = \frac{4}{T}$, with symbol sequences $\{b_c[n]\} = \{+1, -1, -1\}$ and $\{b_s[n]\} = \{-1, +1, -1\}$. The phase of the passband signal is $-\pi/4$ in the first symbol interval, switches to $3\pi/4$ in the second, and to $-3\pi/4$ in the third.

Let us see what happens to the passband signal when $b_c[n], b_s[n]$ each take values in $\{\pm 1 \pm j\}$. For the nth symbol interval $nT \leq t < (n+1)T$:
$u_p(t) = \cos 2\pi f_c t - \sin 2\pi f_c t = \sqrt{2}\cos\left(2\pi f_c t + \pi/4\right)$ if $b_c[n] = +1, b_s[n] = +1$;
$u_p(t) = \cos 2\pi f_c t + \sin 2\pi f_c t = \sqrt{2}\cos\left(2\pi f_c t - \pi/4\right)$ if $b_c[n] = +1, b_s[n] = -1$;

$u_p(t) = -\cos 2\pi f_c t - \sin 2\pi f_c t = \sqrt{2}\cos(2\pi f_c t + 3\pi/4)$ if $b_c[n] = -1, b_s[n] = +1$;
$u_p(t) = -\cos 2\pi f_c t + \sin 2\pi f_c t = \sqrt{2}\cos(2\pi f_c t - 3\pi/4)$ if $b_c[n] = -1, b_s[n] = -1$.
Thus, the modulation causes the passband signal to switch its phase among four possibilities, $\{\pm\pi/4, \pm 3\pi/4\}$, as illustrated in Figure 6.3, which is why we call it Quadrature Phase Shift Keying (QPSK).

Equivalently, we could have seen this from the complex envelope. Note that the QPSK symbols can be written as $b[n] = \sqrt{2}e^{j\theta[n]}$, where $\theta[n] \in \{\pm\pi/4, \pm 3\pi/4\}$. Thus, over the nth symbol, we have

$$u_p(t) = \mathrm{Re}\left(b[n]e^{j2\pi f_c t}\right) = \mathrm{Re}\left(\sqrt{2}e^{j\theta[n]}e^{j2\pi f_c t}\right) = \sqrt{2}\cos(2\pi f_c t + \theta[n]), \quad nT \le t < (n+1)T$$

This indicates that it is actually easier to figure out what is happening to the passband signal by working with the complex envelope. We therefore work in the complex baseband domain for the remainder of this chapter.

In general, the complex envelope for a linearly modulated signal is given by

$$u(t) = \sum_n b[n]p(t - nT)$$

where $b[n] = b_c[n] + jb_s[n] = r[n]e^{j\theta[n]}$ can be complex-valued. We can view this as $b_c[n]$ modulating the I component and $b_s[n]$ modulating the Q component, or as scaling the envelope by $r[n]$ and switching the phase by $\theta[n]$. The set of values that each symbol can take is called the signaling *alphabet,* or *constellation.* We can plot the constellation in a two-dimensional plot, with the x-axis denoting the real part $b_c[n]$ (corresponding to the I component) and the y-axis denoting the imaginary part $b_s[n]$ (corresponding to the Q component). Indeed, this is why linear modulation over passband channels is also termed *two-dimensional* modulation. Note that this provides a unified description of constellations that can be used over both baseband and passband channels: for physical baseband channels, we simply constrain $b[n] = b_c[n]$ to be real-valued, setting $b_s[n] = 0$.

Figure 6.4 shows some common constellations. Pulse Amplitude Modulation (PAM) corresponds to using multiple amplitude levels along the I component (setting the Q component to zero). This is often used for signaling over physical baseband channels. Using PAM along both I and Q axes corresponds to Quadrature Amplitude Modulation (QAM). If the constellation points lie on a circle, they only affect the phase of the carrier: such signaling schemes are termed Phase Shift Keying (PSK). When naming a modulation scheme, we usually indicate the number of points in the constellations. BPSK and QPSK are special: BPSK (or 2PSK) can also be classified as 2PAM, while QPSK (or 4PSK) can also be classified as 4PAM.

Each symbol in a constellation of size M can be uniquely mapped to $\log_2 M$ bits. For a symbol rate of $1/T$ symbols per unit time, the *bit rate* is therefore $\frac{\log_2 M}{T}$ bits per unit time. Since the transmitted bits often contain redundancy due to a channel code employed for error correction or detection, the *information rate* is typically smaller than the bit rate. The choice of constellation for a particular application depends on considerations such as power-bandwidth tradeoffs and implementation complexity. We shall discuss these issues once we develop more background.

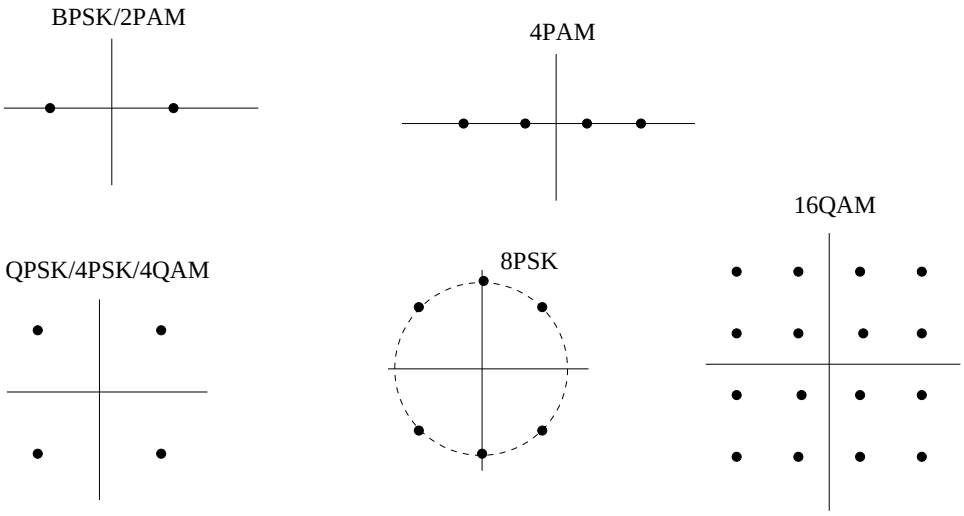


Figure 6.4: Some commonly used constellations. Note that 2PAM and 4PAM can be used over both baseband and passband channels, while the two-dimensional constellations QPSK, 8PSK and 16QAM are for use over passband channels.

6.2 Bandwidth Occupancy

Bandwidth is a precious commodity, hence it is important to quantify the frequency occupancy of communication signals. To this end, consider the complex envelope of a linearly modulated signal (the two-sided bandwidth of this complex envelope equals the physical bandwidth of the corresponding passband signal), which has the form:

$$u(t) = \sum_n b[n]p(t - nT) \tag{6.4}$$

The complex-valued symbol sequence $\{b[n]\}$ is modeled as a discrete-time random process. Modeling the sequence as random at the transmitter makes sense because the latter does not control the information being sent (e.g., it depends on the specific computer file or digital audio signal being sent). Since this information is mapped to the symbols in some fashion, it follows that the symbols themselves are also random rather than deterministic. Modeling the symbols as random at the receiver makes even more sense, since the receiver by definition does not know the symbol sequence (otherwise there would be no need to transmit). Thus, in the running example in Figure 6.1, we might choose to model the symbols as i.i.d., taking values ± 1 with equal probability. Of course, depending on the application, such an i.i.d. model may not be strictly applicable. For example, we may wish to enforce constraints such as not having too long a run of symbols of the same sign (which is unlikely but possible), in order to avoid loss of synchronism or DC bias. We may also wish to insert dependencies in the symbol sequence in order to shape the spectrum of the modulated signal (this is explored in the problems). Despite such caveats, it is usually reasonable to model $\{b[n]\}$ as being i.i.d., taking values with equal probability from the signaling constellation. The constellation itself is typically chosen such that the center of mass of the points is at the origin (this is the case for all the constellations in Figure 6.4, for example), so that choosing all points with equal probability does yield zero DC.

Let us now compute the time-averaged PSD for a sample path of the form (6.4). Recall from Chapter 4 that the steps for computing the PSD for a finite-power signal $u(t)$ are as follows:

(a) timelimit to a finite observation interval of length T_o to get a finite energy signal $u_{T_o}(t)$;
(b) compute the Fourier transform $U_{T_o}(f)$, and hence obtain the energy spectral density $|U_{T_o}(f)|^2$;
(c) estimate the PSD as $\hat{S}_u(f) = \frac{|U_{T_o}(f)|^2}{T_0}$, and take the limit $T_0 \to \infty$ to obtain $S_u(f)$.

Consider the observation interval $[-NT, NT]$, which fits roughly $2N$ symbols. Unlike our running example, in general, the modulation pulse $p(t)$ need not be timelimited to the symbol duration T. However, we can neglect the edge effects caused by this, since we eventually take the limit as the observation interval gets large. Thus, we can write

$$u_{T_o}(t) \approx \sum_{n=-N}^{N} b[n]p(t-nT)$$

Taking the Fourier transform, we obtain

$$U_{T_o}(f) = \sum_{n=-N}^{N} b[n]P(f)e^{-j2\pi fnT}$$

The energy spectral density is therefore given by

$$|U_{T_o}(f)|^2 = U_{T_o}(f)U^*_{T_o}(f) = \sum_{n=-N}^{N} b[n]P(f)e^{-j2\pi fnT} \sum_{m=-N}^{N} b^*[m]P^*(f)e^{j2\pi fmT}$$

where we need to use two different dummy variables, n and m, for the summations corresponding to $U_{T_o}(f)$ and $U^*_{T_o}(f)$, respectively. Thus,

$$|U_{T_o}(f)|^2 = |P(f)|^2 \sum_{m=-N}^{N} \sum_{n=-N}^{N} b[n]b^*[m]e^{-j2\pi(m-n)fT}$$

and the PSD is estimated as

$$\hat{S}_u(f) = \frac{|U_{T_o}(f)|^2}{2NT} = \frac{|P(f)|^2}{T}\{\frac{1}{2N}\sum_{m=-N}^{N} \sum_{n=-N}^{N} b[n]b^*[m]e^{-j2\pi f(n-m)T}\} \tag{6.5}$$

Thus, the PSD factors into two components: the first is a term $\frac{|P(f)|^2}{T}$ that depends on the spectrum of the modulation pulse $p(t)$, while the second term (in curly brackets) Even without simplifying further, we see that the PSD factors into two components depends on the symbol sequence $\{b[n]\}$. Let us now work on simplifying the latter. Grouping terms of the form $m = n-k$ for each fixed k, we can rewrite this term as

$$\frac{1}{2N}\sum_{m=-N}^{N} \sum_{n=-N}^{N} b[n]b^*[m]e^{-j2\pi f(n-m)T} = \sum_{k} \frac{1}{2N}\sum_{n=-N}^{N} b[n]b^*[n-k]e^{-j2\pi fkT} \tag{6.6}$$

Note that we have been deliberately sloppy about the limits of summation in n on the right-hand side to avoid messy notation. Actually, since $-N \le m = n-k \le N$, we have the constraint $-N+k \le n \le N+k$ in addition to the constraint $-N \le n \le N$. Thus, the summation in n should go from $n = -N$ to $n = N+k$ for $k < 0$, and $n = -N+k$ to $n = N$ for $k \ge 0$. However,

these edge effects become negligible when we let N get large while keeping k fixed. When we do this, we get

$$\lim_{N\to\infty} \frac{1}{2N} \sum_{n=-N}^{N} b[n]b^*[n-k] = R_b[k]$$

where $R_b[k] = \overline{b[n]b^*[n-k]}$ is the empirical autocorrelation function of the symbol sequence. Thus, as we take the limit $N \to \infty$ in (6.5), we obtain

$$S_u(f) = \frac{|P(f)|^2}{T} \sum_k R_b[k] e^{-j2\pi fkT} \tag{6.7}$$

Thus, we see that the PSD depends both on the modulating pulse $p(t)$ and on the properties of the symbol sequence $\{b[n]\}$. We explore how the dependence on the symbol sequence can be exploited for shaping the spectrum in the problems. However, for most settings of interest, the symbol sequence can be modeled as uncorrelated and zero mean, In this case, $R_b[k] = 0$ for $k \neq 0$. In order to emphasize the importance of the expression for PSD in this special case, we state the result, which follows by specializing (6.7), as a theorem.

Theorem 6.2.1 (PSD of a linearly modulated signal) *Consider a linearly modulated signal*

$$u(t) = \sum_n b[n]p(t-nT)$$

where the symbol sequence $\{b[n]\}$ is zero mean and uncorrelated with $R_b[0] = \overline{|b[n]|^2} = \sigma_b^2$. Then the PSD is given by

$$S_u(f) = \frac{|P(f)|^2}{T}\sigma_b^2 \tag{6.8}$$

and the power of the modulated signal is

$$P_u = \frac{\sigma_b^2 ||p||^2}{T} \tag{6.9}$$

where $||p||^2$ denotes the energy of the modulating pulse.

The PSD expression follows from specializing (6.7). The expression for power follows from integrating the PSD:

$$P_u = \int_{-\infty}^{\infty} S_u(f)df = \frac{\sigma_b^2}{T}\int_{-\infty}^{\infty} |P(f)|^2 df = \frac{\sigma_b^2}{T}\int_{-\infty}^{\infty} |p(t)|^2 dt = \frac{\sigma_b^2||p||^2}{T}$$

where we have used Parseval's identity.

An intuitive interpretation of this theorem is as follows. Every T time units, we send a pulse of the form $b[n]p(t-nT)$ with average energy spectral density $\sigma_b^2|P(f)|^2$, so that the PSD is obtained by dividing this by T. The same reasoning applies to the expression for power: every T time units, we send a pulse $b[n]p(t-nT)$ with average energy $\sigma_b^2||p||^2$, so that the power is obtained by dividing by T. The preceding intuition does not apply when successive symbols are correlated, in which case we get the more complicated expression (6.7) for the PSD.

Bandwidth definitions: Once we know the PSD, we can define the bandwidth of u in a number of ways:

3 dB bandwidth: For symmetric $S_u(f)$ with a maximum at $f = 0$, the *3 dB bandwidth* B_{3dB} is defined by $S_u(B_{3dB}/2) = S_u(-B_{3dB}/2) = \frac{1}{2}S_u(0)$. That is, the 3 dB bandwidth is the size of the interval between the points at which the PSD is 3 dB, or a factor of $\frac{1}{2}$, smaller than its maximum value.

Fractional power containment bandwidth. This is the size of the smallest interval that contains a given fraction of the power. For example, for symmetric $S_u(f)$, the 99% fractional power containment bandwidth B is defined by

$$\int_{-B/2}^{B/2} S_u(f)df = 0.99P_u = 0.99\int_{-\infty}^{\infty} S_u(f)df$$

(replace 0.99 in the preceding equation by any desired fraction γ to get the corresponding γ power containment bandwidth).

Time/frequency normalization: Before we discuss examples in detail, let us simplify our life by making a simple observation on time and frequency scaling. Suppose we have a linearly modulated system operating at a symbol rate of $1/T$, as in (6.4). We can think of it as a normalized system operating at a symbol rate of one, where the unit of time is T. This implies that the unit of frequency is $1/T$. In terms of these new units, we can write the linearly modulated signal as

$$u_1(t) = \sum_n b[n]p_1(t-n)$$

where $p_1(t)$ is the modulation pulse for the normalized system. For example, for a rectangular pulse timelimited to the symbol interval, we have $p_1(t) = I_{[0,1]}(t)$. Suppose now that the bandwidth of the normalized system (computed using any definition that we please) is B_1. Since the unit of frequency is $1/T$, the bandwidth in the original system is B_1/T. Thus, in terms of determining frequency occupancy, we can work, without loss of generality, with the normalized system. What we are really doing is working with t/T and fT in the original system.

It is instructive to work through the algebra to make the preceding argument concrete. We know that we can go from the normalized system (operating at symbol rate one) to the original system (operating at symbol rate $1/T$) by time-scaling u_1 by T; that is, by setting $u(t) = u_1(t/T)$. This means that $u(\alpha T) = u_1(\alpha)$, so that when αT units of time elapse in the original system, α units of time elapse in the normalized system. The time scaling yields

$$u(t) = u_1(t/T) = \sum_n b[n]p_1(t/T - n) = \sum_n b[n]p_1\left(\frac{t-nT}{T}\right) = \sum_n b[n]p(t-nT)$$

where $p(t) = p_1(t/T)$ is the time-scaled modulation pulse. For example, for $p_1(t) = I_{[0,1]}(t)$, we obtain $p(t) = I_{[0,T]}(t)$ after time scaling. What does this say about the PSD? Suppose that the PSD of the normalized system is $S_{u_1}(f)$. Then the PSD of the original system is proportional to $S_{u_1}(fT)$, where the proportionality constant is chosen based on our normalization assumptions. If we define the normalized system simply as a time-scaled version of the original system, the signal power in both systems is the same, hence we must choose the proportionality constant such that $\int_{-\infty}^{\infty} S_u(f)df = \int_{-\infty}^{\infty} S_{u_1}(f)df$. This implies (check for yourself!) that

$$S_u(f) = TS_{u_1}(fT)$$

Since bandwidth depends on the shape of the PSD rather than its scaling, and since the proportionality constant would change if we decided to scale amplitude as well as time, we feel free to be sloppy about the preceding proportionality constant and often set $S_u(f) = S_{u_1}(fT)$.

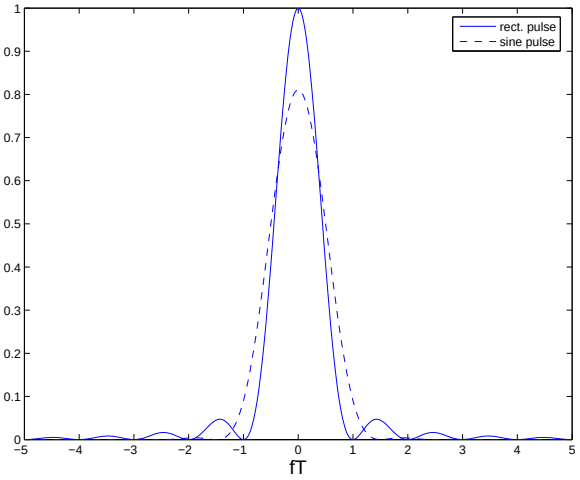


Figure 6.5: PSD corresponding to rectangular and sine timelimited pulses. The main lobe of the PSD is broader for the sine pulse, but its 99% power containment bandwidth is much smaller.

Rectangular pulse: Without loss of generality, consider a normalized system with $p_1(t) = I_{[0,1]}(t)$, for which $P_1(f) = \text{sinc}(f)e^{-j\pi f}$. For $\{b[n]\}$ i.i.d., taking values ± 1 with equal probability, we have $\sigma_b^2 = 1$. Applying (6.8), we obtain

$$S_{u_1}(f) = \sigma_b^2 \text{sinc}^2(f) \tag{6.10}$$

Integrating, or applying (6.9), we obtain $P_u = \sigma_b^2$. The scale factor of σ_b^2 is not important, since it drops out for any definition of bandwidth. We therefore set it to $\sigma_b^2 = 1$. The PSD for the rectangular pulse, along with that for a sine pulse introduced shortly, is plotted in Figure 6.5. Note that the PSD for the rectangular pulse has much fatter tails, which does not bode well for its bandwidth efficiency. For the fractional power containment bandwidth with fraction γ, we have the equation

$$\int_{-B_1/2}^{B_1/2} \text{sinc}^2 f df = \gamma \int_{-\infty}^{\infty} \text{sinc}^2 f df = \gamma \int_0^1 1^2 dt = \gamma$$

using Parseval's identity. We therefore obtain, using the symmetry of the PSD, that the bandwidth is the numerical solution to the equation

$$\int_0^{B_1/2} \text{sinc}^2 f df = \gamma/2 \tag{6.11}$$

For example, for $\gamma = 0.99$, we obtain $B_1 = 10.2$, while for $\gamma = 0.9$, we obtain $B_1 = 0.85$. Thus, if we wish to be strict about power containment (e.g., in order to limit adjacent channel interference in wireless systems), the rectangular timelimited pulse is a very poor choice. On the other hand, in systems where interference or regulation are not significant issues (e.g., low-cost wired systems), this pulse may be a good choice because of its ease of implementation using digital logic.

Example 6.2.1 (Bandwidth computation): A passband system operating at a carrier frequency of 2.4 GHz at a bit rate of 20 Mbps. A rectangular modulation pulse timelimited to the symbol interval is employed.

(a) Find the 99% and 90% power containment bandwidths if the constellation used is 16-QAM.
(b) Find the 99% and 90% power containment bandwidths if the constellation used is QPSK.
Solution:
(a) The 16-QAM system sends 4 bits/symbol, so that the symbol rate $1/T$ equals $\frac{20 \text{ Mbits/sec}}{4 \text{ bits/symbol}} = 5$ Msymbols/sec. Since the 99% power containment bandwidth for the normalized system is $B_1 = 10.2$, the required bandwidth is $B_1/T = 51$ MHz. Since the 90% power containment for the normalized system is $B_1 = 0.85$, the required bandwidth B_1/T equals 4.25 MHz.
(b) The QPSK system sends 2 bits/symbol, so that the symbol rate is 10 Msymbols/sec. The bandwidths required are therefore double those in (a): the 99% power containment bandwidth is 102 MHz, while the 90% power containment bandwidth is 8.5 MHz.
Clearly, when the criterion for defining bandwidth is the same, then 16-QAM consumes half the bandwidth compared to QPSK for a fixed bit rate. However, it is interesting to note that, for the rectangular timelimited pulse, a QPSK system where we are sloppy about power leakage (90% power containment bandwidth of 8.5 MHz) can require far less bandwidth than a system using a more bandwidth-efficient 16-QAM constellation where we are strict about power leakage (99% power containment bandwidth of 51 MHz). This extreme variation of bandwidth when we tweak definitions slightly is because of the poor frequency domain containment of the rectangular timelimited pulse. Thus, if we are serious about limiting frequency occupancy, we need to think about more sophisticated designs for the modulation pulse.

Smoothing out the rectangular pulse: A useful alternative to using the rectangular pulse, while still keeping the modulating pulse timelimited to a symbol interval, is the sine pulse, which for the normalized system equals

$$p_1(t) = \sqrt{2}\sin(\pi t)\ I_{[0,1]}(t)$$

Since the sine pulse does not have the sharp edges of the rectangular pulse in the time domain, we expect it to be more compact in the frequency domain. Note that we have normalized the pulse to have unit energy, as we did for the normalized rectangular pulse. This implies that the power of the modulated signal is the same in the two cases, so that we can compare PSDs under the constraint that the area under the PSDs remains constant. Setting σ_b^2 and using (6.8), we obtain (see Problem 6.5):

$$S_{u_1}(f) = |P_1(f)|^2 = \frac{8}{\pi^2}\frac{\cos^2 \pi f}{\left(1-4f^2\right)^2} \tag{6.12}$$

Proceeding as we did for obtaining (6.11), the fractional power containment bandwidth for fraction γ is given by the formula:

$$\int_0^{B_1/2} \frac{8}{\pi^2}\frac{\cos^2 \pi f}{\left(1-4f^2\right)^2} df = \gamma/2 \tag{6.13}$$

For $\gamma = 0.99$, we obtain $B_1 = 1.2$, which is an order of magnitude improvement over the corresponding value of $B_1 = 10.2$ for the rectangular pulse.

While the sine pulse has better frequency domain containment than the rectangular pulse, it is still not well-suited for use over *strictly* bandlimited channels. We discuss pulse design for such channels next.

6.3 Design for Bandlimited Channels

Suppose that you are told to design your digital communication system so that the transmitted signal fits between 2.39 and 2.41 GHz; that is, you are given a passband channel of bandwidth 20 MHz at a carrier frequency of 2.4 GHz. Any signal that you transmit over this band has a complex envelope with respect to 2.4 GHz that occupies a band from -10 MHz to 10 MHz. Similarly, the passband channel (modeled as an LTI system) has an impulse response whose complex envelope is bandlimited from -10 MHz to 10 MHz. In general, for a passband channel or signal of bandwidth W, with an appropriate choice of reference frequency, we have a corresponding complex baseband signal spanning the band $[-W/2, W/2]$. Thus, we restrict our design to the complex baseband domain, with the understanding that the designs can be translated to passband channels by upconversion of the I and Q components at the transmitter, and downconversion at the receiver. Also, note that the designs specialize to physical baseband channels if we restrict the baseband signals to be real-valued.

6.3.1 Nyquist's Sampling Theorem and the Sinc Pulse

Our first step in understanding communication system design for such a bandlimited channel is to understand the structure of bandlimited signals. To this end, suppose that the signal $s(t)$ is bandlimited to $[-W/2, W/2]$. We can now invoke Nyquist's sampling theorem (proof postponed to later in this section) to express the signal in terms of its samples at rate W.

Theorem 6.3.1 (Nyquist's sampling theorem) Any signal $s(t)$ bandlimited to $[-\frac{W}{2}, \frac{W}{2}]$ can be described completely by its samples $\{s(\frac{n}{W})\}$ at rate W. The signal $s(t)$ can be recovered from its samples using the following interpolation formula:

$$s(t) = \sum_{n=-\infty}^{\infty} s\left(\frac{n}{W}\right) p\left(t - \frac{n}{W}\right) \tag{6.14}$$

where $p(t) = \text{sinc}(Wt)$.

Degrees of freedom: What does the sampling theorem tell us about digital modulation? The interpolation formula (6.14) tells us that we can interpret $s(t)$ as a linearly modulated signal with symbol sequence equal to the samples $\{s(n/W)\}$, symbol rate $1/T$ equal to the bandwidth W, and modulation pulse given by $p(t) = \text{sinc}(Wt) \leftrightarrow P(f) = \frac{1}{W} I_{[-W/2,W/2]}(f)$. Thus, linear modulation with the sinc pulse is able to exploit all the "degrees of freedom" available in a bandlimited channel.

Signal space: If we signal over an observation interval of length T_o using linear modulation according to the interpolation formula (6.14), then we have approximately WT_o complex-valued samples. Thus, while the signals we send are continuous-time signals, which in general, lie in an infinite-dimensional space, the set of possible signals we can send in a finite observation interval of length T_o live in a complex-valued vector space of *finite* dimension WT_o, or equivalently, a real-valued vector space of dimension $2WT_o$. Such geometric views of communication signals as vectors, often termed *signal space concepts,* are particularly useful in design and analysis, as we explore in more detail in Chapter 7.

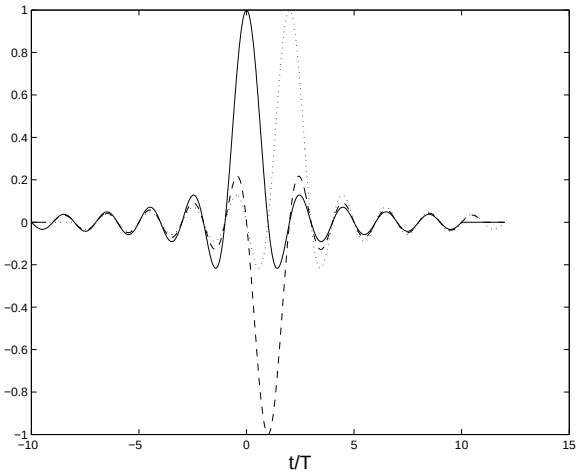


Figure 6.6: Three successive sinc pulses (each pulse is truncated to a length of 10 symbol intervals on each side) modulated by +1,-1,+1. The actual transmitted signal is the sum of these pulses (not shown). Note that, while the pulses overlap, the samples at $t = 0, T, 2T$ are equal to the transmitted bits because only one pulse is nonzero at these times.

The concept of Nyquist signaling: Since the sinc pulse is not timelimited to a symbol interval, in principle, the symbols could interfere with each other. The time domain signal corresponding to a bandlimited modulation pulse such as the sinc spans an interval significantly larger than the symbol interval (in theory, the interval is infinitely large, but we always truncate the waveform in implementations). This means that successive pulses corresponding to successive symbols which are spaced by the symbol interval (i.e., $b[n]p(t - nT)$ as we increment n) overlap with, and therefore can interfere with, each other. Figure 6.6 shows the sinc pulse modulated by three bits, +1,-1,+1. While the pulses corresponding to the three symbols do overlap, notice that, by sampling at $t = 0$, $t = T$ and $t = 2T$, we can recover the three symbols because exactly one of the pulses is nonzero at each of these times. That is, at sampling times spaced by integer multiples of the symbol time T, there is no *intersymbol interference.* We call such a pulse *Nyquist* for signaling at rate $\frac{1}{T}$, and we discuss other examples of such pulses soon. Designing pulses based on the Nyquist criterion allows us the freedom to expand the modulation pulses in time beyond the symbol interval, while ensuring that there is no ISI at appropriately chosen sampling times despite significant overlap between successive pulses.

The problem with sinc: Are we done then? Should we just use linear modulation with a sinc pulse when confronted with a bandlimited channel? Unfortunately, the answer is no: just as the rectangular timelimited pulse decayed too slowly in frequency, the rectangular bandlimited pulse, corresponding to the sinc pulse in the time domain, decays too slowly in time. Let us see what that does. Figure 6.7 shows a plot of the modulated waveform for a bit sequence of alternating sign. At the correct sampling times, there is no ISI. However, if we consider a small timing error of $0.25T$, the ISI causes the sample value to drop drastically, making the system more vulnerable to noise. What is happening is that, when there is a small sampling offset, we can make the ISI add up to a large value by choosing the interfering symbols so that their contributions all have signs opposite to that of the desired symbol at the sampling time. Since the sinc pulse decays as $1/t$, the ISI created for a given symbol by an interfering symbol which is n symbol intervals away decays as $1/n$, so that, in the worst-case, the contributions from the interfering symbols

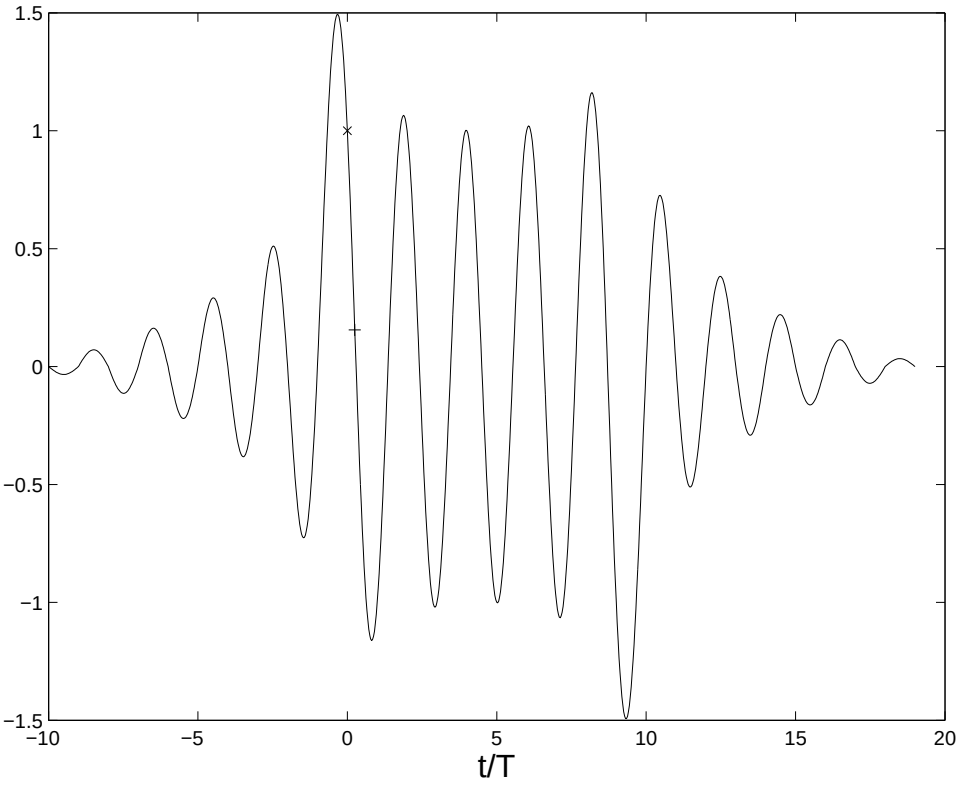


Figure 6.7: The baseband signal for 10 BPSK symbols of alternating signs, modulated using the sinc pulse. The first symbol is +1, and the sample at time $t = 0$, marked with 'x', equals +1, as desired (no ISI). However, if the sampling time is off by $0.25T$, the sample value, marked by '+', becomes much smaller because of ISI. While it still has the right sign, the ISI causes it to have significantly smaller noise immunity.

roughly have the form $\sum_n \frac{1}{n}$, a series that is known to diverge. Thus, in theory, if we do not truncate the sinc pulse, we can make the ISI arbitrarily large when there is a small timing offset. In practice, we do truncate the modulation pulse, so that we only see ISI from a finite number of symbols. However, even when we do truncate, as we see from Figure 6.7, the slow decay of the sinc pulse means that the ISI adds up quickly, and significantly reduces the margin of error when noise is introduced into the system.

While the sinc pulse may not be a good idea in practice, the idea of using bandwidth-efficient Nyquist pulses is a good one, and we now develop it further.

6.3.2 Nyquist Criterion for ISI Avoidance

Nyquist signaling: Consider a linearly modulated signal

$$u(t) = \sum_n b[n]p(t - nT)$$

We say that the pulse $p(t)$ is Nyquist (or satisfies the Nyquist criterion) for signaling at rate $\frac{1}{T}$ if the symbol-spaced samples of the modulated signal are equal to the symbols (or a fixed scalar multiple of the symbols); that is, $u(kT) = b[k]$ for all k, so that there is no ISI at appropriately chosen sampling times.

In the time domain, it is quite easy to see what is required to satisfy the Nyquist criterion. The samples $u(kT) = \sum_n b[n]p(kT - nT) = b[k]$ (or a scalar multiple of $b[k]$) for all k if and only if $p(0) = 1$ (or some nonzero constant) and $p(mT) = 0$ for all integers $m \neq 0$. However, for design of bandwidth efficient pulses, it is important to characterize the Nyquist criterion in the frequency domain. This is given by the following theorem.

Theorem 6.3.2 (Nyquist criterion for ISI avoidance): *The pulse $p(t) \leftrightarrow P(f)$ is Nyquist for signaling at rate $\frac{1}{T}$ if*

$$p(mT) = \delta_{m0} = \begin{cases} 1 & m = 0 \\ 0 & m \neq 0 \end{cases} \tag{6.15}$$

or equivalently,

$$\frac{1}{T} \sum_{k=-\infty}^{\infty} P(f + \frac{k}{T}) = 1 \quad \text{for all } f \tag{6.16}$$

The proof of this theorem is postponed to Section 6.3.5, where we show that both the Nyquist sampling theorem, Theorem 6.3.1, and the preceding theorem are based on the same mathematical result, that the samples of a time domain signal have a one-to-one mapping with the sum of translated (or *aliased*) versions of its Fourier transform.

In this section, we explore the design implications of Theorem 6.3.2. In the frequency domain, the translates of $P(f)$ by integer multiples of $1/T$ must add up to a constant. As illustrated by Figure 6.8, the minimum bandwidth pulse for which this happens is one that is ideal bandlimited over an interval of length $1/T$.

Minimum bandwidth Nyquist pulse: The minimum bandwidth Nyquist pulse is

$$P(f) = \begin{cases} T & |f| \leq \frac{1}{2T} \\ 0 & else \end{cases}$$

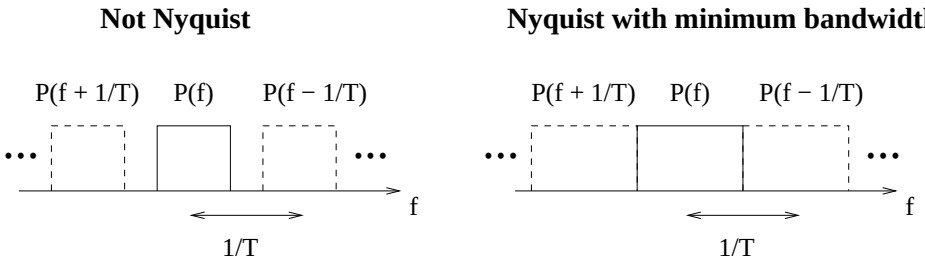


Figure 6.8: The minimum bandwidth Nyquist pulse is a sinc.

corresponding to the time domain pulse

$$p(t) = \text{sinc}(t/T)$$

As we have already discussed, the sinc pulse is not a good choice in practice because of its slow decay in time. To speed up the decay in time, we must expand in the frequency domain, while conforming to the Nyquist criterion. The trapezoidal pulse depicted in Figure 6.8 is an example of such a pulse.

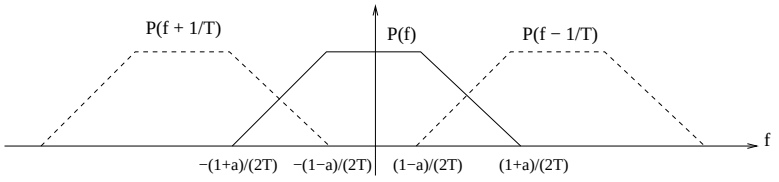


Figure 6.9: A trapezoidal pulse which is Nyquist at rate $1/T$. The (fractional) excess bandwidth is a.

The role of excess bandwidth: We have noted earlier that the problem with the sinc pulse arises because of its $1/t$ decay and the divergence of the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$, which implies that the worst-case contribution from "distant" interfering symbols at a given sampling instant can blow up. Using the same reasoning, however, a pulse $p(t)$ decaying as $1/t^b$ for $b > 1$ should work, since the series $\sum_{n=1}^{\infty} \frac{1}{n^b}$ does converge for $b > 1$. A faster time decay requires a slower decay in frequency. Thus, we need *excess bandwidth,* beyond the minimum bandwidth dictated by the Nyquist criterion, to fix the problems associated with the sinc pulse. The (fractional) excess bandwidth for a linear modulation scheme is defined to be the fraction of bandwidth over the minimum required for ISI avoidance at a given symbol rate. In particular, Figure 6.9 shows that a trapezoidal pulse (in the frequency domain) can be Nyquist for suitably chosen parameters, since the translates $\{P(f+k/T)\}$ as shown in the figure add up to a constant. Since trapezoidal $P(f)$ is the convolution of two boxes in the frequency domain, the time domain pulse $p(t)$ is the product of two sinc functions (see Problem 6.1 for details). Since each sinc decays as $1/t$, the product decays as $1/t^2$, which implies that the worst-case ISI with timing mismatch is indeed bounded.

Raised cosine pulse: Replacing the straight line of the trapezoid with a smoother cosine-shaped curve in the frequency domain gives us the raised cosine pulse shown in Figure 6.11,

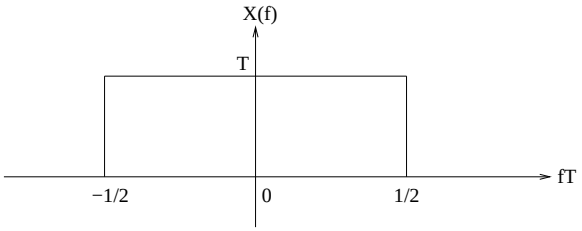
which has a faster, $1/t^3$, decay in the time domain.

$$P(f) = \begin{cases} T & |f| \leq \frac{1-a}{2T} \\ \frac{T}{2}[1 - \sin((|f| - \frac{1}{2T})\frac{\pi T}{a})] & \frac{1-a}{2T} \leq |f| \leq \frac{1+\alpha}{2T} \\ 0 & |f| > \frac{1+a}{2T} \end{cases}$$

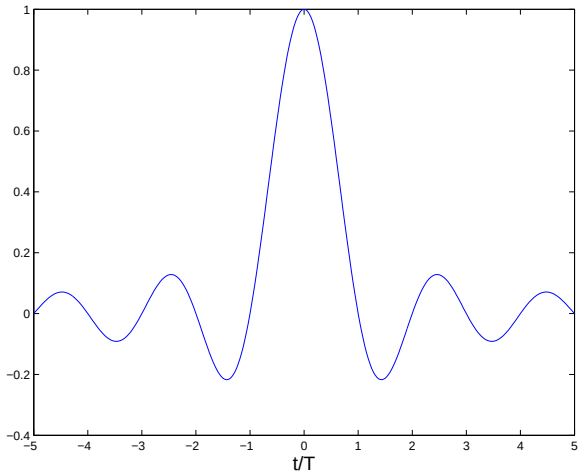
where a is the fractional excess bandwidth, typically chosen in the range where $0 \leq a < 1$. As shown in Problem 6.9, the time domain pulse $s(t)$ is given by

$$p(t) = \text{sinc}(\frac{t}{T}) \frac{\cos \pi a \frac{t}{T}}{1 - \left(\frac{2at}{T}\right)^2}$$

This pulse inherits the Nyquist property of the sinc pulse, while having an additional multiplicative factor that gives an overall $(\frac{1}{t^3})$ decay with time. The faster time decay compared to the sinc pulse is evident from a comparison of Figures 6.11(b) and 6.10(b).



(a) Frequency domain boxcar



(b) Time domain sinc pulse

Figure 6.10: Sinc pulse for minimum bandwidth ISI-free signaling at rate $1/T$. Both time and frequency axes are normalized to be dimensionless.

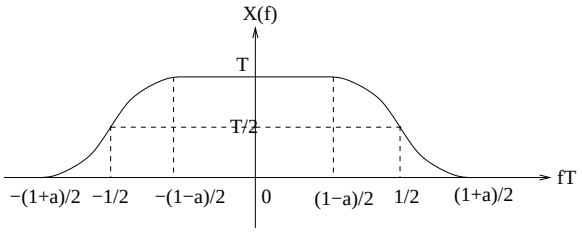
6.3.3 Bandwidth efficiency

We define the *bandwidth efficiency* of linear modulation with an M-ary alphabet as

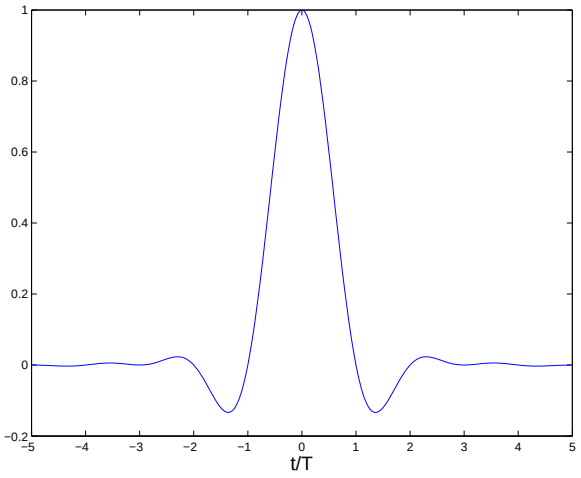
$$\eta_B = \log_2 M \text{ bits/symbol}$$

The Nyquist criterion for ISI avoidance says that the minimum bandwidth required for ISI-free transmission using linear modulation equals the symbol rate, using the sinc as the modulation pulse. For such an idealized system, we can think of η_B as bits/second per Hertz, since the symbol rate equals the bandwidth. Thus, knowing the bit rate R_b and the bandwidth efficiency η_B of the modulation scheme, we can determine the symbol rate, and hence the minimum required bandwidth B_{min}. as follows:

$$B_{min} = \frac{R_b}{\eta_B}$$



(a) Frequency domain raised cosine



(b) Time domain pulse (excess bandwidth $a = 0.5$)

Figure 6.11: Raised cosine pulse for minimum bandwidth ISI-free signaling at rate $1/T$, with excess bandwidth a. Both time and frequency axes are normalized to be dimensionless.

This bandwidth would then be expanded by the excess bandwidth used in the modulating pulse. However, this is not included in our definition of bandwidth efficiency, because excess bandwidth is a highly variable quantity dictated by a variety of implementation considerations. Once we decide on the fractional excess bandwidth a, the actual bandwidth required is

$$B = (1+a)B_{min} = (1+a)\frac{R_b}{\eta_B}$$

6.3.4 The Nyquist criterion at the link level

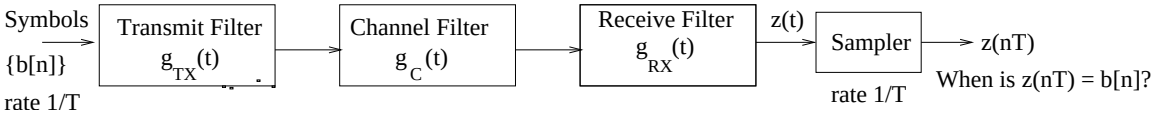


Figure 6.12: Nyquist criterion at the link level.

Figure 6.12 shows a block diagram for a link using linear modulation, with the entire model expressed in complex baseband. The symbols $\{b[n]\}$ are passed through the transmit filter to obtain the waveform $\sum_n b[n]g_{TX}(t-nT)$. This then goes through the channel filter $g_C(t)$, and then the receive filter $g_{RX}(t)$. Thus, at the output of the receive filter, we have the linearly modulated signal $\sum_n b[n]p(t-nT)$, where $p(t) = (g_{TX} * g_C * g_{RX})(t)$ is the cascade of the transmit, channel and receive filters. We would like the pulse $p(t)$ to be Nyquist at rate $1/T$, so that, in the absence of noise, the symbol rate samples at the output of the receive filter equal the transmitted symbols. Of course, in practice, we do not have control over the channel, hence we often assume an ideal channel, and design such that the cascade of the transmit and receive filter, given by $(g_{TX} * g_{RX})(t) G_{TX}(f) G_{RX}(f)$ is Nyquist. One possible choice is to set G_{TX} to be a Nyquist pulse, and G_{RX} to be a wideband filter whose response is flat over the band of interest. Another choice that is even more popular is to set $G_{TX}(f)$ and $G_{RX}(f)$ to be square

roots of a Nyquist pulse. In particular, the square root raised cosine (SRRC) pulse is often used in practice.

Square root Nyquist pulses and their time domain interpretation: A pulse $g(t) \leftrightarrow G(f)$ is defined to be square root Nyquist at rate $1/T$ if $|G(f)|^2$ is Nyquist at rate $1/T$. Note that $P(f) = |G(f)|^2 \leftrightarrow p(t) = (g * g_{MF})(t)$, where $g_{MF}(t) = g^*(-t)$. The time domain Nyquist condition is given by

$$p(mT) = (g * g_{MF})(mT) = \int g(t)g^*(t - mT)dt = \delta_{m0} \tag{6.17}$$

That is, a square root Nyquist pulse has an autocorrelation function that vanishes at nonzero integer multiples of T. In other words, the waveforms $\{g(t - kT, k = 0, \pm 1, \pm 2, ...\}$ are orthonormal, and can be used to provide a basis for constructing more complex waveforms, as we see later.

6.3.5 Proofs of the Nyquist theorems

We have used Nyquist's sampling theorem, Theorem 6.3.1, to argue that linear modulation using the sinc pulse is able to use all the degrees of freedom in a bandlimited channel. On the other hand, Nyquist's criterion for ISI avoidance, Theorem 6.3.2, tells us, roughly speaking, that we must have enough degrees of freedom in order to avoid ISI (and that the sinc pulse provides the minimum such degrees of freedom). As it turns out, both theorems are based on the same mathematical relationship between samples in the time domain and aliased spectra in the frequency domain, stated in the following theorem.

Theorem 6.3.3 (Sampling and Aliasing): *Consider a signal $s(t)$, sampled at rate $\frac{1}{T_s}$. Let $S(f)$ denote the spectrum of $s(t)$, and let*

$$B(f) = \frac{1}{T_s} \sum_{k=-\infty}^{\infty} S(f + \frac{k}{T_s}) \tag{6.18}$$

denote the sum of translates of the spectrum. Then the following observations hold:
(a) $B(f)$ is periodic with period $\frac{1}{T_s}$.
(b) The samples $\{s(nT_s)\}$ are the Fourier series for $B(f)$, satisfying

$$s(nT_s) = T_s \int_{-\frac{1}{2T_s}}^{\frac{1}{2T_s}} B(f)e^{j2\pi fnTs}\, df \tag{6.19}$$

$$B(f) = \sum_{n=-\infty}^{\infty} s(nT_s)e^{-j2\pi fnT_s} \tag{6.20}$$

Remark: Note that the signs of the exponents for the frequency domain Fourier series in the theorem are reversed from the convention in the usual time domain Fourier series (analogous to the reversal of the sign of the exponent for the inverse Fourier transform compared to the Fourier transform).

Proof of Theorem 6.3.3: The periodicity of $B(f)$ follows by its very construction. To prove (b), apply the the inverse Fourier transform to obtain

$$s(nT_s) = \int_{-\infty}^{\infty} S(f) e^{j2\pi f n T_s} df$$

We now write the integral as an infinite sum of integrals over segments of length $1/T$

$$s(nT_s) = \sum_{k=-\infty}^{\infty} \int_{\frac{k-\frac{1}{2}}{T_s}}^{\frac{k+\frac{1}{2}}{T_s}} S(f) e^{j2\pi f n T_s} df$$

In the integral over the kth segment, make the substitution $\nu = f - \frac{k}{T_s}$ and rewrite it as

$$\int_{-\frac{1}{2T_s}}^{\frac{1}{2T_s}} S(\nu + \frac{k}{T_s}) e^{j2\pi(\nu + \frac{k}{T_s}) n T_s} d\nu = \int_{-\frac{1}{2T_s}}^{\frac{1}{2T_s}} S(\nu + \frac{k}{T_s}) e^{j2\pi \nu n T_s} d\nu$$

Now that the limits of all segments and the complex exponential in the integrand are the same (i.e., independent of k), we can move the summation inside to obtain

$$\begin{aligned} s(nT_s) &= \int_{-\frac{1}{2T_s}}^{\frac{1}{2T_s}} \left(\sum_{k=-\infty}^{\infty} S(\nu + \frac{k}{T_s}) \right) e^{j2\pi \nu n T_s} d\nu \\ &= T_s \int_{-\frac{1}{2T_s}}^{\frac{1}{2T_s}} B(\nu) e^{j2\pi \nu n T_s} d\nu \end{aligned}$$

proving (6.19). We can now recognize that this is just the formula for the Fourier series coefficients of $B(f)$, from which (6.20) follows. □

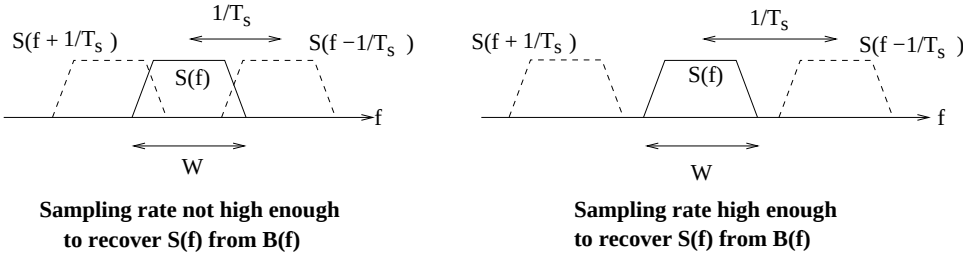


Figure 6.13: Recovering a signal from its samples requires a high enough sampling rate for translates of the spectrum not to overlap.

Inferring Nyquist's sampling theorem from Theorem 6.3.3: Suppose that $s(t)$ is bandlimited to $[-\frac{W}{2}, \frac{W}{2}]$. The samples of $s(t)$ at rate $\frac{1}{T_s}$ can be used to reconstruct $B(f)$, since they are the Fourier series for $B(f)$. But $S(f)$ can be recovered from $B(f)$ if and only if the translates $S(f - \frac{k}{T_s})$ do not overlap, as shown in Figure 6.13. This happens if and only if $\frac{1}{T_s} \geq W$. Once this condition is satisfied, $\frac{1}{T_s} S(f)$ can be recovered from $B(f)$ by passing it through an ideal bandlimited filter $H(f) = I_{[-W/2.W/2]}(f)$. We therefore obtain that

$$\frac{1}{T_s} S(f) = B(f) H(f) = \sum_{n=-\infty}^{\infty} s(nT_s) e^{-j2\pi f n T_s} I_{[-W/2.W/2]}(f) \tag{6.21}$$

Noting that $I_{[-W/2.W/2]}(f) \leftrightarrow W\text{sinc}(Wt)$, we have

$$e^{-j2\pi fnT_s} I_{[-W/2.W/2]}(f) \leftrightarrow W\text{sinc}\left(W(t-nT_s)\right)$$

Taking inverse Fourier transforms, we get the interpolation formula

$$\frac{1}{T_s}s(t) = \sum_{n=-\infty}^{\infty} s(nT_s)W\text{sinc}\left(W(t-nT_s)\right)$$

which reduces to (6.14) for $\frac{1}{T_s} = W$. This completes the proof of the sampling theorem, Theorem 6.3.1. □

Inferring Nyquist's criterion for ISI avoidance from Theorem 6.3.3: A Nyquist pulse $p(t)$ at rate $1/T$ must satisfy $p(nT) = \delta_{n0}$. Applying Theorem 6.3.3 with $s(t) = p(t)$ and $T_s = T$, it follows immediately from (6.20) that $p(nT) = \delta_{n0}$ (i.e., the time domain Nyquist criterion holds) if and only if

$$B(f) = \frac{1}{T}\sum_{k=-\infty}^{\infty} P(f + \frac{k}{T_s}) = 1$$

In other words, if the Fourier series only has a DC term, then the periodic waveform it corresponds to must be constant. □

6.3.6 Linear modulation as a building block

Linear modulation can be used as a building block for constructing more sophisticated waveforms, using discrete-time sequences modulated by square root Nyquist pulses. Thus, one symbol would be made up of multiple "chips," linearly modulated by a square root Nyquist "chip waveform." Specifically, suppose that $\psi(t)$ is square root Nyquist at a chip rate $\frac{1}{T_c}$. N chips make up one symbol, so that the symbol rate is $\frac{1}{T_s} = \frac{1}{NT_c}$, and a symbol waveform is given by linearly modulating a code vector $\mathbf{s} = (s[0], ..., s[N-1])$ consisting of N chips, as follows:

$$s(t) = \sum_{n=0}^{N} s[k]\psi(t - kT_c)$$

Since $\{\psi(t - kT_c)\}$ are orthonormal (see (6.17)), we have simply expressed the code vector in a continuous time basis. Thus, the continuous time inner product between two symbol waveforms (which determines their geometric relationships and their performance in noise, as we see in the next chapter) is equal to the discrete time inner product between the corresponding code vectors. Specifically, suppose that $s_1(t)$ and $s_2(t)$ are two symbol waveforms corresponding to code vectors $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively. Then their inner product satisfies

$$\langle s_1, s_2\rangle = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} s_1[k]s_2^*[l]\int \psi(t-kT_c)\psi^*(t-lT_c)dt = \sum_{k=0}^{N-1} s_1[k]s_2^*[k] = \langle \mathbf{s}_1, \mathbf{s}_2\rangle$$

where we have use the orthonormality of the translates $\{\psi(t - kT_c)\}$. This means that we can design discrete time code vectors to have certain desired properties, and then linearly modulate square root Nyquist chip waveforms to get symbol waveforms that have the same desired properties. For example, if $\mathbf{s}_1$ and $\mathbf{s}_2$ are orthogonal, then so are $s_1(t)$ and $s_2(t)$; we use this in the next section when we discuss orthogonal modulation.

Examples of square root Nyquist chip waveforms include a rectangular pulse timelimited to an interval of length T_c , as well as bandlimited pulses such as the square root raised cosine. From Theorem 6.2.1, we see that the PSD of the modulated waveform is proportional to $|\Psi(f)|^2$ (it is typically a good approximation to assume that the chips $\{s[k]\}$ are uncorrelated). That is, the bandwidth occupancy is determined by that of the chip waveform ψ.

6.4 Orthogonal and Biorthogonal Modulation

While linear modulation with larger and larger constellations is a means of increasing bandwidth efficiency, we shall see that orthogonal modulation with larger and larger constellations is a means of increasing power efficiency (while making bandwidth efficiency smaller). Consider first M-ary frequency shift keying (FSK), a classical form of orthogonal modulation in which one of M sinusoidal tones, successively spaced by Δf, are transmitted every T units of time, where $\frac{1}{T}$ is the symbol rate. Thus, the bit rate is $\frac{\log_2 M}{T}$, and for a typical symbol interval, the transmitted passband signal is chosen from one of M possibilities:

$$u_{p,k}(t) = \cos\left(2\pi(f_0 + k\Delta f)t\right) \ , \ 0 \le t \le T, \quad k = 0, 1, ..., M-1$$

where we typically have $f_0 \gg \frac{1}{T}$. Taking f_0 as reference, the corresponding complex baseband waveforms are

$$u_k(t) = \exp\left(j2\pi k\Delta f t\right) \ , \ 0 \le t \le T, \quad k = 0, 1, ..., M-1$$

Let us now understand how the tones should be chosen in order to ensure orthogonality. Recall that the passband and complex baseband inner products are related as follows:

$$\langle u_{p,k}, u_{p,l}\rangle = \frac{1}{2}\mathrm{Re}\langle u_k, u_l\rangle$$

so we can develop criteria for orthogonality working in complex baseband. Setting $k = l$, we see that

$$||u_k||^2 = T$$

For two adjacent tones, $l = k + 1$, we leave it as an exercise to show that

$$\mathrm{Re}\langle u_k, u_{k+1}\rangle = \frac{\sin 2\pi\Delta f T}{2\pi\Delta f}$$

We see that the minimum value of Δf for which the preceding quantity is zero is given by $2\pi\Delta f T = \pi$, or $\Delta f = \frac{1}{2T}$.

From the point of view of the receiver, it means that when there is an incoming wave at the kth tone, then correlating against the kth tone will give a large output, but correlating against the $(k+1)$th tone will give zero output (in the absence of noise) if the tone spacing is $\frac{1}{2T}$. However, this assumes a *coherent* system in which the tones we are correlating against are synchronized in phase with the incoming wave. What happens if they are 90° out of phase? Then correlation of the kth tone with itself yields

$$\int_0^T \cos\left(2\pi(f_0 + k\Delta f)t\right)\cos\left(2\pi(f_0 + k\Delta f)t + \frac{\pi}{2}\right)dt = 0$$

(by orthogonality of the cosine and sine), so that the output we desire to be large is actually zero! In order to be robust to such variations, we must use *noncoherent* reception, which we describe next.

Noncoherent reception: Let us develop the concept of noncoherent reception in generality, because it is a concept that is useful in many settings, not just for orthogonal modulation. Suppose that we transmit a passband waveform, and wish to detect it at the receiver by correlating it against the receiver's copy of the waveform. However, the receiver's local oscillator may not be synchronized in phase with the phase of the incoming wave. Let us denote the receiver's copy of the signal as

$$u_p(t) = u_c(t)\cos 2\pi f_c t - u_s(t)\sin 2\pi f_c t$$

and the incoming passband signal as

$$y_p(t) = y_c(t)\cos 2\pi f_c t - y_s(t)\sin 2\pi f_c t = u_c(t)\cos\left(2\pi f_c t + \theta\right) - u_s(t)\sin\left(2\pi f_c t + \theta\right)$$

Using the receiver's local oscillator as reference, the complex envelope of the receiver's copy is $u(t) = u_c + ju_s(t)$, while that of the incoming wave is $y(t) = u(t)e^{j\theta}$. Thus, the inner product

$$\langle y_p, u_p\rangle = \frac{1}{2}\mathrm{Re}\langle y, u\rangle = \frac{1}{2}\mathrm{Re}\langle ue^{j\theta}, y\rangle = \frac{1}{2}\mathrm{Re}\left(||u||^2 e^{j\theta}\right) = \frac{||u||^2}{2}\cos\theta$$

Thus, the output of the correlator is degraded by the factor $\cos\theta$, and can actually become zero, as we have already observed, if the phase offset $\theta = \pi/2$. In order to get around this problem, let us look at the complex baseband inner product again:

$$\langle y, u\rangle = \langle ue^{j\theta}, y\rangle = e^{j\theta}||u||^2$$

We could ensure that this output remains large regardless of the value of θ if we took its *magnitude,* rather than the real part. Thus, noncoherent reception corresponds to computing $|\langle y, u\rangle|$ or $|\langle y, u\rangle|^2$. Let us unwrap the complex inner product to see what this entails:

$$\langle y, u\rangle = \int y(t)u^*(t)dt = \int (y_c(t)+jy_s(t))(u_c(t)-ju_s(t))dt = (\langle y_c, u_c\rangle + \langle y_s, u_s\rangle)+j\left(\langle y_s, u_c\rangle - \langle y_c, u_s\rangle\right)$$

Thus, the noncoherent receiver computes the quantity

$$|\langle y, u\rangle|^2 = \left(\langle y_c, u_c\rangle + \langle y_s, u_s\rangle\right)^2 + \left(\langle y_s, u_c\rangle - \langle y_c, u_s\rangle\right)^2$$

In contrast, the coherent receiver computes

$$\mathrm{Re}\langle y, u\rangle = \langle y_c, u_c\rangle + \langle y_s, u_s\rangle$$

That is, when the receiver LO is synchronized to the phase of the incoming wave, we can correlate the I component of the received waveform with the I component of the receiver's copy, and similarly correlate the Q components, and sum them up. However, in the presence of phase asynchrony, the I and Q components get mixed up, and we must compute the magnitude of the complex inner product to recover all the energy of the incoming wave. Figure 6.14 shows the receiver operations corresponding to coherent and noncoherent reception.

Back to FSK: Going back to FSK, if we now use noncoherent reception, then in order to ensure that we get a zero output (in the absence of noise) when receiving the kth tone with a noncoherent receiver for the $(k+1)$th tone, we must ensure that

$$|\langle u_k, u_{k+1}\rangle| = 0$$

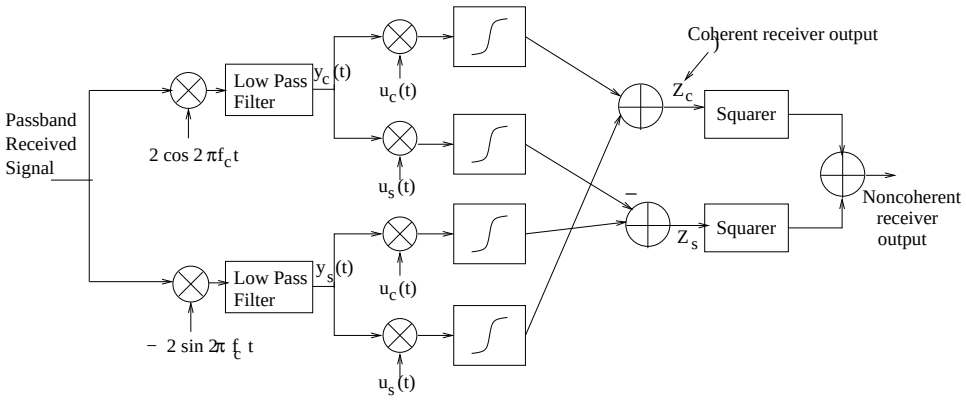


Figure 6.14: Structure of coherent and noncoherent receivers.

We leave it as an exercise to show that the minimum tone spacing for noncoherent FSK is $\frac{1}{T}$, which is double that required for orthogonality in coherent FSK. The bandwidth for coherent M-ary FSK is approximately $\frac{M}{2T}$, which corresponds to a time-bandwidth product of approximately $\frac{M}{2}$. This corresponds to a complex vector space of dimension $\frac{M}{2}$, or a real vector space of dimension M, in which we can fit M orthogonal signals. On the other hand, M-ary noncoherent signaling requires M complex dimensions, since the complex baseband signals must remain orthogonal even under multiplication by complex-valued scalars.

Summarizing the concept of orthogonality: To summarize, when we say "orthogonal" modulation, we must specify whether we mean coherent or noncoherent reception, because the concept of orthogonality is different in the two cases. For a signal set $\{s_k(t)\}$, orthogonality requires that, for $k \neq l$, we have

$$\begin{array}{ll} \mathrm{Re}(\langle s_k, s_l \rangle) = 0 & \textbf{coherent orthogonality criterion} \\ \langle s_k, s_l \rangle = 0 & \textbf{noncoherent orthogonality criterion} \end{array} \tag{6.22}$$

Bandwidth efficiency: We conclude from the example of orthogonal FSK that the bandwidth efficiency of orthogonal signaling is $\eta_B = \frac{\log_2(2M)}{M}$ bits/complex dimension for coherent systems, and $\eta_B = \frac{\log_2(M)}{M}$ bits/complex dimension for noncoherent systems. This is a general observation that holds for any realization of orthogonal signaling. In a signal space of complex dimension D (and hence real dimension $2D$), we can fit $2D$ signals satisfying the coherent orthogonality criterion, but only D signals satisfying the noncoherent orthogonality criterion. As M gets large, the bandwidth efficiency tends to zero. In compensation, as we see in Chapter 7, the power efficiency of orthogonal signaling for large M is the "best possible."

Orthogonal Walsh-Hadamard codes

Section 6.3.6 shows how to map vectors to waveforms while preserving inner products, by using linear modulation with a square root Nyquist chip waveform. Applying this construction, the problem of designing orthogonal waveforms $\{s_i\}$ now reduces to designing orthogonal code vectors $\{\mathbf{s}_i\}$. Walsh-Hadamard codes are a standard construction employed for this purpose, and can be constructed recursively as follows: at the nth stage, we generate 2^n orthogonal vectors, using the 2^{n-1} vectors constructed in the $n-1$ stage. Let $\mathbf{H}_n$ denote a matrix whose rows are 2^n orthogonal codes obtained after the nth stage, with $H_0 = (1)$. Then

$$\mathbf{H}_n = \begin{pmatrix} \mathbf{H}_{n-1} & \mathbf{H}_{n-1} \\ \mathbf{H}_{n-1} & -\mathbf{H}_{n-1} \end{pmatrix}$$

We therefore get

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad \text{etc.}$$

The signals $\{s_i\}$ obtained above can be used for noncoherent orthogonal signaling, since they satisfy the orthogonality criterion $\langle s_i, s_j \rangle = 0$ for $i \neq j$. However, just as for FSK, we can fit twice as many signals into the same number of degrees of freedom if we used the weaker notion of orthogonality required for coherent signaling, namely $\text{Re}(\langle s_i, s_j \rangle = 0$ for $i \neq j$. It is easy to check that for M-ary Walsh-Hadamard signals $\{s_i, i = 1, ..., M\}$, we can get $2M$ orthogonal signals for coherent signaling: $\{s_i, js_i, i = 1, ..., M\}$. This construction corresponds to independently modulating the I and Q components with a Walsh-Hadamard code; that is, using passband waveforms $s_i(t) \cos 2\pi f_c t$ and $-s_i(t) \sin 2\pi f_c t$, $i = 1, ..., M$.

Biorthogonal modulation

Given an orthogonal signal set, a biorthogonal signal set of twice the size can be obtained by including a negated copy of each signal. Since signals s and $-s$ cannot be distinguished in a noncoherent system, biorthogonal signaling is applicable to coherent systems. Thus, for an M-ary Walsh-Hadamard signal set $\{s_i\}$ with M signals obeying the noncoherent orthogonality criterion, we can construct a coherent orthogonal signal set $\{s_i, js_i\}$ of size $2M$, and hence a biorthogonal signal set of size $4M$, e.g., $\{s_i, js_i, -s_i, -js_i\}$. These correspond to the passband waveforms $\pm s_i(t) \cos 2\pi f_c t$ and $\pm s_i(t) \sin 2\pi f_c t$, $i = 1, ..., M$.

Problems

Problem 6.1 Consider the trapezoidal pulse of excess bandwidth a shown in Figure 6.9.
(a) Find an explicit expression for the time domain pulse $p(t)$.
(b) What is the bandwidth required for a passband system using this pulse operating at 100 Mbps using 64QAM, with an excess bandwidth of 25%?

Problem 6.2 Consider a pulse $s(t) = \text{sinc}(at)\text{sinc}(bt)$, where $a \geq b$.
(a) Sketch the frequency domain response $S(f)$ of the pulse.
(b) Suppose that the pulse is to be used over an ideal real baseband channel with one-sided bandwidth 400 Hz. Choose a and b so that the pulse is Nyquist for 4-PAM signaling at 1200 bits/sec and exactly fills the channel bandwidth.
(c) Now, suppose that the pulse is to be used over a passband channel spanning the frequencies 2.4-2.42 GHz. Assuming that we use 64-QAM signaling at 60 Mbits/sec, choose a and b so that the pulse is Nyquist and exactly fills the channel bandwidth.
(d) Sketch an argument showing that the magnitude of the transmitted waveform in the preceding settings is always finite.

Problem 6.3 Consider the pulse $p(t)$ whose Fourier transform satisfies:

$$P(f) = \begin{cases} 1, & 0 \leq |f| \leq A \\ \frac{B-|f|}{B-A}, & A \leq |f| \leq B \\ 0, & \text{else} \end{cases}$$

where $A = 250KHz$ and $B = 1.25MHz$.
(a) **True or False** The pulse $p(t)$ can be used for Nyquist signaling at rate 3 Mbps using an 8-PSK constellation.
(b) **True or False** The pulse $p(t)$ can be used for Nyquist signaling at rate 4.5 Mbps using an 8-PSK constellation.

Problem 6.4 Consider the pulse

$$p(t) = \begin{cases} 1 - \frac{|t|}{T}, & 0 \leq |t| \leq T \\ 0, & \text{else} \end{cases}$$

Let $P(f)$ denote the Fourier transform of $p(t)$.
(a) **True or False** The pulse $p(t)$ is Nyquist at rate $\frac{1}{T}$.
(b) **True or False** The pulse $p(t)$ is square root Nyquist at rate $\frac{1}{T}$. (i.e., $|P(f)|^2$ is Nyquist at rate $\frac{1}{T}$).

Problem 6.5 Show that the Fourier transform of the pulse $p(t) = \sin \pi t I_{[0,1]}(t)$ is given by

$$P(f) = \frac{2\cos(\pi f)\ e^{-j\pi f}}{\pi(1-4f^2)}$$

Problem 6.6 Consider the pulse

$$p(t) = \begin{cases} 1 - \frac{|t|}{T}, & 0 \leq |t| \leq T \\ 0, & \text{else} \end{cases}$$

Let $P(f)$ denote the Fourier transform of $p(t)$.
(a) **True or False** The pulse $p(t)$ is Nyquist at rate $1/T$.
(b) **True or False** The pulse $p(t)$ is square root Nyquist at rate $1/T$. (i.e., $|P(f)|^2$ is Nyquist at rate $1/T$).

Problem 6.7 Consider the pulse $p(t)$ whose Fourier transform satisfies:

$$P(f) = \begin{cases} 1, & 0 \leq |f| \leq A \\ \frac{B-|f|}{B-A}, & A \leq |f| \leq B \\ 0, & \text{else} \end{cases}$$

where $A = 250KHz$ and $B = 1.25MHz$.
(a) **True or False** The pulse $p(t)$ can be used for Nyquist signaling at rate 3 Mbps using an 8-PSK constellation.
(b) **True or False** The pulse $p(t)$ can be used for Nyquist signaling at rate 4.5 Mbps using an 8-PSK constellation.

Problem 6.8 (True or False) Any pulse timelimited to duration T is square root Nyquist (up to scaling) at rate $1/T$.

Problem 6.9 In this problem, we derive the time domain response of the frequency domain raised cosine pulse. Let $R(f) = I_{[-\frac{1}{2},\frac{1}{2}]}(f)$ denote an ideal boxcar transfer function, and let $C(f) = \frac{\pi}{2a}\cos(\frac{\pi}{a}f)I_{[-\frac{a}{2},\frac{a}{2}]}$ denote a cosine transfer function.
(a) Sketch $R(f)$ and $C(f)$, assuming that $0 < a < 1$.
(b) Show that the frequency domain raised cosine pulse can be written as

$$S(f) = (R * C)(f)$$

(c) Find the time domain pulse $s(t) = r(t)c(t)$. Where are the zeros of $s(t)$? Conclude that $s(t/T)$ is Nyquist at rate $1/T$.
(d) Sketch an argument that shows that, if the pulse $s(t/T)$ is used for BPSK signaling at rate $1/T$, then the magnitude of the transmitted waveform is always finite.

Problem 6.10 (Effect of timing errors) Consider digital modulation at rate $1/T$ using the sinc pulse $s(t) = \text{sinc}(2Wt)$, with transmitted waveform

$$y(t) = \sum_{n=1}^{100} b_n s(t - (n-1)T)$$

where $1/T$ is the symbol rate and $\{b_n\}$ is the bit stream being sent (assume that each b_n takes one of the values ± 1 with equal probability). The receiver makes bit decisions based on the samples $r_n = y((n-1)T)$, $n = 1, ..., 100$.
(a) For what value of T (as a function of W) is $r_n = b_n$, $n = 1, ..., 100$?
Remark: In this case, we simply use the sign of the nth sample r_n as an estimate of b_n.
(b) For the choice of T as in (a), suppose that the receiver sampling times are off by .25 T. That is, the nth sample is given by $r_n = y((n-1)T + .25T)$, $n = 1, ..., 100$. In this case, we do have ISI of different degrees of severity, depending on the bit pattern. Consider the following bit pattern:

$$b_n = \begin{cases} (-1)^{n-1} & 1 \le n \le 49 \\ (-1)^n & 50 \le n \le 100 \end{cases}$$

Numerically evaluate the 50th sample r_{50}. Does it have the same sign as the 50th bit b_{50}?
Remark: The preceding bit pattern creates the worst possible ISI for the 50th bit. Since the sinc pulse dies off slowly with time, the ISI contributions due to the 99 other bits to the 50th sample sum up to a number larger in magnitude, and opposite in sign, relative to the contribution due to b_{50}. A decision on b_{50} based on the sign of r_{50} would therefore be wrong. This sensitivity to timing error is why the sinc pulse is seldom used in practice.
(c) Now, consider the digitally modulated signal in (a) with the pulse $s(t) = \text{sinc}(2Wt)\text{sinc}(Wt)$.

For ideal sampling as in (a), what are the two values of T such that $r_n = b_n$?
(d) For the smaller of the two values of T found in (c) (which corresponds to faster signaling, since the symbol rate is $1/T$), repeat the computation in (b). That is, find r_{50} and compare its sign with b_{50} for the bit pattern in (b).
(e) Find and sketch the frequency response of the pulse in (c). What is the excess bandwidth relative to the pulse in (a), assuming Nyquist signaling at the same symbol rate?
(f) Discuss the impact of the excess bandwidth on the severity of the ISI due to timing mismatch.

Problem 6.11 (OQPSK and MSK) Linear modulation with a bandlimited pulse can perform poorly over nonlinear passband channels. For example, the output of a passband hardlimiter (which is a good model for power amplifiers operating in a saturated regime) has constant envelope, but a PSK signal employing a bandlimited pulse has an envelope that passes through zero during a 180 degree phase transition, as shown in Figure 6.15. One way to alleviate this

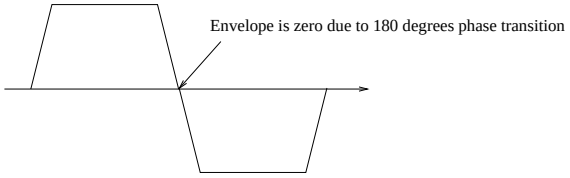


Figure 6.15: The envelope of a PSK signal passes through zero during a 180 degree phase transition, and gets distorted over a nonlinear channel.

problem is to not allow 180 degree phase transitions. Offset QPSK (OQPSK) is one example of such a scheme, where the transmitted signal is given by

$$s(t) = \sum_{n=-\infty}^{\infty} b_c[n]p(t - nT) + jb_s[n]p(t - nT - \frac{T}{2}) \tag{6.23}$$

where $\{b_c[n]\}$, $b_s[n]$ are ± 1 BPSK symbols modulating the I and Q channels, with the I and Q signals being staggered by half a symbol interval. This leads to phase transitions of at most 90 degrees at integer multiples of the *bit time* $T_b = \frac{T}{2}$. Minimum Shift Keying (MSK) is a special case of OQPSK with timelimited modulating pulse

$$p(t) = \sqrt{2}\sin(\frac{\pi t}{T})I_{[0,T]}(t) \tag{6.24}$$

(a) Sketch the I and Q waveforms for a typical MSK signal, clearly showing the timing relationship between the waveforms.
(b) Show that the MSK waveform has constant envelope (an extremely desirable property for nonlinear channels).
(c) Find an analytical expression for the PSD of an MSK signal, assuming that all bits sent are i.i.d., taking values ± 1 with equal probability. Plot the PSD versus normalized frequency fT.
(d) Find the 99% power containment normalized bandwidth of MSK. Compare with the minimum Nyquist bandwidth, and the 99% power containment bandwidth of OQPSK using a rectangular pulse.
(e) Recognize that Figure 6.5 gives the PSD for OQPSK and MSK, and reproduce this figure, normalizing the area under the PSD curve to be the same for both modulation formats.

Problem 6.12 (FSK tone spacing) Consider two real-valued passband pulses of the form

$$s_0(t) = cos(2\pi f_0 t + \phi_0) \quad 0 \le t \le T$$
$$s_1(t) = cos(2\pi f_1 t + \phi_1) \quad 0 \le t \le T$$

where $f_1 > f_0 \gg 1/T$. The pulses are said to be *orthogonal* if $\langle s_0, s_1 \rangle = \int_0^T s_0(t)s_1(t)dt = 0$.
(a) If $\phi_0 = \phi_1 = 0$, show that the minimum frequency separation such that the pulses are orthogonal is $f_1 - f_0 = \frac{1}{2T}$.
(b) If ϕ_0 and ϕ_1 are arbitrary phases, show that the minimum separation for the pulses to be orthogonal regardless of ϕ_0, ϕ_1 is $f_1 - f_0 = 1/T$.
Remark: The results of this problem can be used to determine the bandwidth requirements for coherent and noncoherent FSK, respectively.

Problem 6.13 (Walsh-Hadamard codes)
(a) Specify the Walsh-Hadamard codes for 8-ary orthogonal signaling with noncoherent reception.
(b) Plot the baseband waveforms corresponding to sending these codes using a square root raised cosine pulse with excess bandwidth of 50%.
(c) What is the fractional increase in bandwidth efficiency if we use these 8 waveforms as building blocks for biorthogonal signaling with coherent reception?

Problem 6.14 (Bandwidth occupancy as a function of modulation format) We wish to send at a rate of 10 Mbits/sec over a passband channel. Assuming that an excess bandwidth of 50% is used, how much bandwidth is needed for each of the following schemes: QPSK, 64-QAM, and 64-ary noncoherent orthogonal modulation using a Walsh-Hadamard code.

Problem 6.15 Consider 64-ary orthogonal signaling using Walsh-Hadamard codes. Assuming that the chip pulse is square root raised cosine with excess bandwidth 25%, what is the bandwidth required for sending data at 20 Kbps over a passband channel assuming (a) coherent reception, (b) noncoherent reception.