Chapter 14

MODULATION

INTRODUCTION

As we have seen in previous three chapters, different types of media need different types of electromagnetic signals to carry information from the source to the destination. In chapters 9 and 10 we discussed analog and digital baseband signals. Now that we discussed the characteristics of various media we can now discuss how to construct signals that take maximum advantage of each media's capabilities. In almost all cases, the source information is impressed upon a carrier-wave (essentially a sinusoid of a certain frequency) by *changing or modifying some characteristic of the sinusoidal wave*. This process is called modulation. The original source signal (e.g., audio, voltage pulse train carrying digital information) is called the *baseband signal*. Modulation has the effect of moving the baseband signal spectrum to be centered frequencies around the frequency of the carrier. The resulting modulated signal is considered a *bandpass* signal. Other processes that modify the original information bearing signal are sometimes called modulation—for example, the representation of sampled signals by the amplitude, position or width of a pulse as described in Chapter 10.

Consider a general sinusoid of frequency f_c which we will refer to as the *carrier frequency*. Recall from previous chapters (2, 6, 9) that we can write this sinusoidal carrier signal as:

$$c(t) = A\cos(2\pi f_c t + \varphi) \tag{14.1}$$

Here, A is called the *amplitude* and φ the *phase* of the carrier. Before this carrier is transmitted, data are used to modulate or change its amplitude, frequency, phase or some combination of these as we will see later. We discuss the need for this in Section 14.1. We consider the continuous change of the characteristics of a carrier or analog modulation in Section 14.2. In particular, we discuss amplitude and frequency modulation. We discuss discrete changes in the characteristics of the carrier (digital modulation) in 14.3 along with the methods of representing and analyzing the performance of these modulation schemes. We consider binary modulation schemes and multi-level modulation schemes here. In this section, we also describe a special case of digital modulation that is very important for transmission of information using modems—quadrature amplitude modulation (QAM) and tradeoffs between data rate and signal bandwidth. Digital subscriber lines and emerging wireless local area networks use multiple carriers to carry information—a technique called

orthogonal frequency division multiplexing (OFDM). Current and emerging generations of cellular wireless communications use multiple layers of modulation resulting in spreading of the spectrum of a signal. Both of these complex modulation schemes are briefly discussed in 14.4.

14.1 WHY MODULATION?

Why do we need modulation? As the reader saw in the last three chapters, all media used for transmission act as filters that attenuate different frequencies by different values. So it is beneficial to move the spectrum of a signal to a frequency that is less susceptible to attenuation over a given medium. We considered an example of *amplitude modulation* in Chapter 6 that does this frequency translation. Some media have characteristics that severely distort digital waveforms in such cases, it is necessary to send digital information in analog form using sinusoidal carriers. We saw this restriction in Chapter 13 when we discussed the electromagnetic spectrum which is regulated and different applications are allowed transmissions only in specific parts of the electromagnetic spectrum. We also saw in Chapter 13 that the size of the antenna depends on the wavelength of the signal that is transmitted or received. Higher frequencies (smaller wavelengths) can reduce the size of the antenna and thus the transceiver. Again, this makes it necessary to move the spectrum to a higher frequency range. In processing signals, circuits are sometimes designed to best operate in only a certain range of frequencies. The same circuit may have to be used used to process signals that occupy different frequency bands. In such a case we once again translate the spectrum, this time to an *intermediate frequency* (IF) that is in the range where the circuit operates best. We will consider *separating different transmissions* from different sources in Chapter 15 (multiplexing). A common way of separating such different transmissions is to use separated frequency bands for these transmissions. Once again, this implies that the spectrum of a signal must be shifted to the range of frequencies that it is allowed to occupy. Modulation is necessary in all of the above scenarios.

14.2 ANALOG MODULATION

In analog modulation, the characteristics of the modulated sinusoid (such as amplitude, frequency or phase) can take a continuum of values depending on the source of the information. The two common forms of analog modulation are *amplitude modulation (AM)* and *frequency modulation (FM)* which is specific form of more general *angle modulation*. Most of us are familiar with AM and FM commercial radio stations. These radio transmissions make use of amplitude and frequency modulation respectively. In North America, the 525 kHz to 1715 kHz band is used for AM transmissions and the 87.8 MHz–108 MHz band is used for FM transmissions. An AM channel is 10 kHz wide and an FM channel is 0.2 MHz wide. Note that the commercial radio systems adopted the names of the modulation schemes that they emply, but there are other systems that also use amplitude and frequency modulation. For example, analog video transmission for television makes use of a combination of both AM and FM. The analog cell phone systems use FM as the modulation scheme. We discuss AM in Section 14.2.1 and FM in Section 14.2.2.

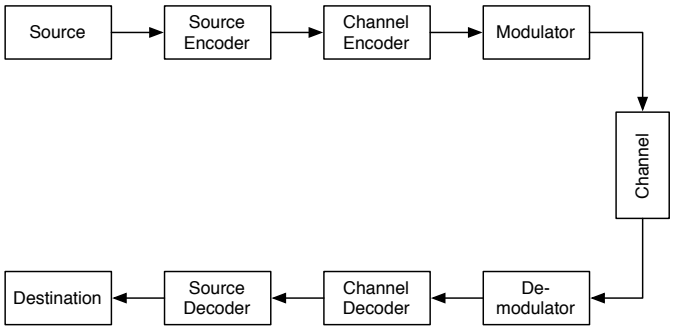

Figure 14.7 Block diagram of a communication system.

MHz = 200 kHz. This fits reasonably well with the computation of transmission bandwidth using Carson's rule. FM receivers use an IF of 10.7 MHz to recover the message signal.

Noise and FM signals. Noise analysis for general FM signals is fairly complex and described in detail in [1]. FM signals exhibit an improvement in SNR at the receiver output over the receiver input SNR by a factor that is around $\frac{3k_f^2 P}{f_{max}^2}$ where P is the average power in the message signal. Observe that the improvement in SNR is a function of the frequency sensitivity k_f which in turn affects the signal bandwidth through the frequency deviation Δf. The improvement factor can be shown to be proportional to D^2 where $D = \frac{\Delta f}{f_{max}}$. The transmission bandwidth B_T is approximately proportional to D. hence, we can say that the output signal-to-noise ratio is improved quadratically whenever the transmission bandwidth is increased.

The above reasoning is true only when the carrier power is large compared to the noise power. FM receivers also exhibit the threshold effect. Improvements are not seen when the signal-to-noise ratio is below a threshold value. Below the threshold value, an FM receiver cannot function. Initially, there may be clicks in the received audio and these degrade to a crackle or sputter.

FM however has an inherent ability to minimize the effects of interference. If two FM signals at the same carrier frequency are received, an FM receiver *captures* the stronger signal and rejects the weaker signal. This capture effect is useful in packet radio applications. FM was also the modulation scheme of choice in the first generation (1G) or analog cellular systems in the USA, Europe and Japan.

14.3 DIGITAL MODULATION

In Chapter 8, we briefly considered a model of a digital communication system where we have a source, a source encoder, and a channel encoder on the transmitter side and the corresponding channel decoder, source decoder and destination on the receiver side. We reproduce this communication system in Figure 14.7 with the addition of a *modulator*

block on the transmitter side and a *demodulator* block on the receiver side. Our goal in this section is to consider the details of the modulator and demodulator blocks. In Figure 14.7, we see that information is sent to the modulator *after* both source and channel encoding. The source and channel encoders typically assume that the source produces a discrete alphabet of information.

We considered digital signals in Chapter 10. In the case of digital signals, the information is in the form of a finite set of *discrete* symbols called an alphabet. For example, if the alphabet is binary, the two possible symbols are 0 and 1 and information is simply a long sequence of 0's and 1's. A "binary" digital signal represents the "zero symbol" using a specific signal that lasts for a duration of T_s seconds and the "one symbol" using another specific signal that also typically lasts for the same duration of T_s seconds. Since one bit is transmitted every T_s seconds, the bit rate is $\frac{1}{T_s}$ bps.

If the number of possible symbols is M, we call it an M-ary alphabet and the corresponding signal is an M-ary digital signal. While it is possible to have any arbitrary value for M, most systems are constructed such that $M = 2^k$. In such a case, we can think of each one of the M symbols as *containing* k bits. For instance, if m_1, m_2, m_3, and m_4 are the symbols of a 4-ary system, we can associate the "dibit" 00 to m_1, 01 to m_2, 10 to m_3 and 11 to m_4. Each one of the M symbols is usually represented by a unique signal that lasts for T_s seconds. The message signal is thus one of M discrete possibilities. We call T_s the *symbol duration* and $\frac{1}{T_s}$ as the symbol rate (expressed in units of baud). The bit rate (if $M = 2^k$) will be $\frac{k}{T_s}$ bits per second.

Example. Each symbol occupies 1 μs, then the symbol rate is 1 M symbol/s or 1 Mbaud. If each symbol carries 4 bits ($k = 4$ or it is a $M = 16$-ary alphabet) and the bit rate is 4 Mbps.

In a manner similar to analog modulation, the message is mapped to the amplitude, frequency, phase (or a combination of these) of the carrier. Note however that there are a finite and discrete number of messages and each message has a corresponding amplitude, frequency or phase value. Thus there are a discrete number of carriers with specific values of amplitude, frequency and phase values corresponding to a given alphabet. If the message is mapped only to the amplitude of the carrier, the modulation is called *amplitude shift keying* or ASK. If the message is mapped only to the frequency of the carrier, the modulation is called *frequency shift keying* or FSK. If the message is mapped only to the phase of the carrier, the modulation is called *phase shift keying* or PSK. A hybrid of amplitude and phase mapping is called *quadrature amplitude modulation* (QAM).

In analog modulation we were interested in the SNR, but recall from Chapter 10 that in digital modulation we are intersted in the bit error rate (BER) as a function of the ratio of the energy per bit (E_b) to the value of the noise PSD (N_0) given by $\frac{E_b}{N_0}$. Also, like analog modulation we would like to maximize the efficiency with which we use the available bandwidth in digital modulation we quantify the spectral efficiency for the amount of bandwidth W required to transmit at a given data rate R quantified as $\eta = \frac{R}{W}$ bps/Hz. Ideally, we would like to get the lowest bit error rate while spending the smallest amount of energy for transmitting a bit and have the ability to simultaneously transmit at the highest possible data rate in the given bandwidth. As we will see later, there are tradeoffs between the BER for a given $\frac{E_b}{N_0}$ and η.

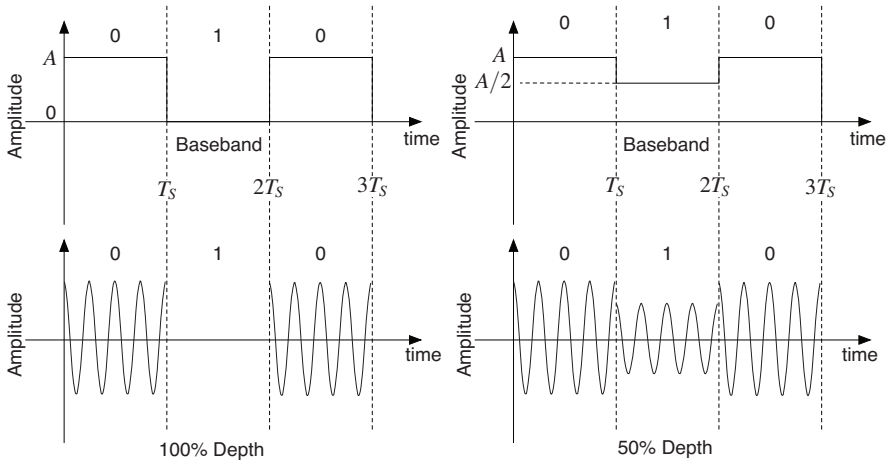

Figure 14.8 Binary Amplitude Shift Keying.

In what follows, we first describe the signaling schemes for binary and M-ary alphabets. We then briefly revisit the idea of a matched filter and consider the impact of noise in digital modulation schemes. We then consider a geometric representation of signals and noise leading to the idea of a "signal constellation." With a signal constellation, it is possible to easily understand the performance of many digital modulation schemes under a unified framework.

14.3.1 Binary Modulation Schemes

In the case of binary modulation schemes, the alphabet has two values "0" and "1." In ASK, a "0" is mapped to one amplitude value and a "1" is mapped to another amplitude value. In FSK, a "0" is mapped to one frequency value and a "1" is mapped to another frequency value. In PSK, a "0" is mapped to one phase value and a "1" is mapped to another phase value. We call these modulation schemes BASK, BFSK and BPSK respectively to denote that the alphabet is binary. We discuss these schemes in more detail below.

Binary Amplitude Shift Keying (BASK). In BASK, the binary symbols last for T_s seconds each and are characterized by the amplitude of the carrier. In the general case,

$$s_i(t) = A_i \cos(2\pi f_c t + \phi),\ 0 \leq t \leq T_s \text{ for } i = 1, 2 \tag{14.25}$$

The transmitter will transmit $s_1(t)$ when the bit is zero and $s_2(t)$ when the bit is one. When $A_1 = A$ and $A_2 = 0$, we refer to the modulation scheme as having 100% depth. This scheme (where the two amplitude values are A and 0) is also called *unipolar modulation* or *on-off keying*. When $A_1 = A$ and $A_2 = \frac{A}{2}$, we say that the modulation depth is 50%. Both of these schemes are shown in Figure 14.8.

Recall that we are interested in the BER as a function of E_b/N_0 as one of the performance measures. Towards this goal, let us perform some simple calculations. The

energy in the "zero" bit is given by:

$$\begin{aligned} E_{zero} &= \int_0^{T_s} s_1^2(t)\,dt = A_1^2 \int_0^{T_s} \cos^2(2\pi f_c t + \phi)\,dt \\ &= \frac{A_1^2}{2} \int_0^{T_s} [1 + \cos(2\pi(2 \times f_c)t + 2\phi)]\,dt \\ &\approx \frac{A_1^2}{2} T_s \end{aligned} \tag{14.26}$$

The approximation is an equality if $f_c = \frac{k}{T_s}$ and is a close approximation if $f_c \gg \frac{1}{T_s}$ even if f_c is not a multiple of $\frac{1}{T_s}$. The energy in the "one" bit is similarly equal to:

$$E_{one} = \frac{A_2^2}{2} T_s \tag{14.27}$$

The *average* energy per bit is given by:

$$E_{b,av} = \frac{T_s}{4}\left[A_1^2 + A_2^2\right] \tag{14.28}$$

Here we assume that the number of "0"s and the number of "1"s in a transmission are equal. For the two special cases in Figure 14.8, the average energy per bit can be calculated to be $E_b = \frac{A^2 T_s}{4}$ and $E_b = \frac{5A^2 T_s}{16}$ respectively. Note that the average energy per bit with 50% modulation is higher. One may expect that the bit error rate with 50% modulation is lower since more energy is being expended. But the error rates are worse as we will see later because it is easier to make a mistake as to which bit was transmitted. Thus wherever BASK is employed, it is common to use on-off keying.

Generation of the on-off keyed signal is fairly simple. The carrier is simply multiplied by a baseband unipolar signal (see Figure 14.1). The baseband signal can be recovered at the receiver using the same techniques as AM (envelope detection or coherent detection).

On-off keying is not a very popular modulation scheme. As we will see later, it is fairly inefficient in terms of the BER performance as a function of the energy consumed. Historically, on-off keying was used for transmitting Morse codes on RF carriers. It is now used in devices that need to be extremely simple—some examples are television remotes, RF-ID tags and infra-red links.

Binary Frequency Shift Keying (BFSK). In BFSK, the binary symbols last for T_s seconds each and are characterized by the frequency of the carrier. In the general case,

$$s_i(t) = A\cos(2\pi f_i t),\ 0 \le t \le T_s \text{ for } i = 1, 2 \tag{14.29}$$

where we assume that the phase of the carrier is zero for simplicity. The transmitter will transmit $s_1(t)$ when the bit is zero and $s_2(t)$ when the bit is one. Figure 14.9 shows an example of the transmission of three bits using BFSK.

It is important to note that any two arbitrary frequencies f_1 and f_2 cannot be used to represent the binary digits. The two frequencies must be separated by at least $\frac{1}{T_s}$ to

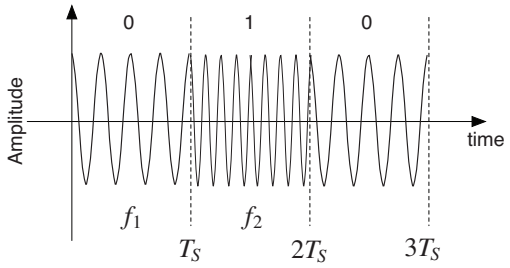

Figure 14.9 Binary Frequency Shift Keying.

ensure that the receiver is able to differentiate between the frequencies. The receiver typically determines which frequency was transmitted by correlating the received signal with locally generated carriers at the two frequencies f_1 and f_2 and picking the larger of the two correlations. Consider the correlation between $s_1(t)$ and $s_2(t)$ which involves multiplication of the two signals and integration over one symbol period given by:

$$\begin{aligned}\int_0^{T_s} s_1(t)s_2(t)\,dt &= A^2 \int_0^{T_s} \cos(2\pi f_1 t)\cos(2\pi f_2 t)dt \\ &= \frac{A^2}{2}\int_0^{T_s} \left[\cos\left(2\pi(f_1 - f_2)t\right) + \cos\left(2\pi(f_1 + f_2)t\right)\right] dt \end{aligned} \quad (14.30)$$

The term $\int \cos(2\pi(f_1 + f_2)t)$ has a frequency close to two times f_i and it can be filtered out using a low-pass filter. The term $\int \cos(2\pi(f_1 - f_2)t)$ needs to be made as small as possible. Specifically, orthogonal FSK ensures that there is no correlation between $s_1(t)$ and $s_2(t)$, that is:

$$\int_0^{T_s} s_1(t)s_2(t)\,dt = 0 \quad (14.31)$$

In the case of orthogonal FSK, $f_1 - f_2 = \frac{1}{T_s}$ so that the integration in $\int \cos(2\pi(f_1 - f_2)t)$ is over exactly one period resulting in a zero value. This idea also becomes important in the discussion of orthogonal frequency division multiplexing (OFDM) later on in this chapter (see Section 14.4.1).

As we saw in the case of BASK, the energy per bit in BFSK (for both a "0" and a "1") can be calculated to be $\frac{A^2 T_s}{2}$ which is also the average energy per bit. As we will see later, BFSK has a better BER performance than BASK for the same average $\frac{E_b}{N_0}$. Note also that an ASK signal, like AM has an envelope that is varying with time. However, an FSK signal, like FM, has a constant envelope providing robustness against amplitude fluctuations.

Example. A classic example of FSK is the old 300 baud modems that used Manchester signaling (Chapter 10). Recall that a Manchester pulse consists of two "half" pulses. The duration of one bit in these modems was $T_s = 1.67$ ms for a data rate of 600 bps. The frequency used to represent one of the half pulses was $f_1 = 1.5$ kHz and the frequency used to represent the other half pulse was $f_2 = 1.8$ kHz.

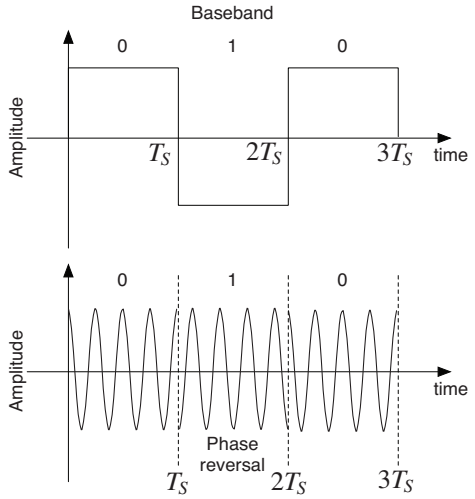

Figure 14.10 Binary Phase Shift Keying.

FSK has been used in early modems and FAX machines. FSK was used for signaling purposes in the early analog cellular systems (like the Advanced Mobile Phone System—AMPS). Recently, FSK has found applications in low power wireless networks like Bluetooth, Zigbee and sensor networks.

Binary Phase Shift Keying (BPSK). In BPSK, the binary symbols last for T_s seconds each and are characterized by the phase of the carrier. In the general case,

$$s_i(t) = A\cos(2\pi f_c t + \phi_i),\ 0 \le t \le T_s \text{ for } i = 1, 2 \tag{14.32}$$

The transmitter will transmit $s_1(t)$ when the bit is zero and $s_2(t)$ when the bit is one. Figure 14.10 shows an example of the transmission of three bits using BPSK.

Since the phase can only be between 0 and 2π radians, the maximum possible phase difference between the two bits is π. It is common to assume that $\phi_1 = 0$ and $\phi_2 = \pi$ in which case, the two signals will be:

$$\begin{aligned} s_1(t) &= A\cos(2\pi f_c t),\ 0 \le t \le T_s \\ s_2(t) &= \cos(2\pi f_c t + \pi) = -A\cos(2\pi f_c t),\ 0 \le t \le T_s \end{aligned} \tag{14.33}$$

From the equations and Figure 14.10, we can see that there is a reversal of phase when the bit changes and so, this scheme is also called *phase reversal keying*. This interesting result shows that $s_1(t) = -s_2(t)$ and we can view BPSK as BASK where $A_1 = A$ and $A_2 = -A$. However, it is more appropriate to consider this as phase modulation as we will see later. In the case of baseband signals without modulation, BPSK is equivalent to *antipodal* or bipolar signaling with non-return-to-zero (NRZ) pulses. The transmitter can be simply implemented as a multiplication of the baseband antipodal signal and the carrier at frequency f_c.

In a manner similar to BASK, we can compute the average energy per bit in the case of BPSK. We can show that the average energy per bit is $\frac{A^2 T_s}{2}$. Also, BPSK signals, like phase modulation, have a constant envelope. Consequently, they are robust to amplitude fluctuations compared to BASK signals. As we will see later, they also have the best BER performance of all binary modulation schemes for a given energy per bit.

BPSK is used as a robust modulation scheme in many applications. In 802.11 wireless local area networks, although different modulation schemes are used depending on the transmission rate, the header of each frame is always transmitted using BPSK to ensure its successful reception. Second generation cellular CDMA systems use what is called "dual BPSK" for transmissions from the cell phone tower to mobile devices (downlink). Here, there are two BPSK signals, one using a cosine and the other a sine that are transmitted simultaneously. Each of these signals carries the same data. The reason why this is possible is because the sine and cosine are orthogonal to one another. This fact is also exploited in M-ary modulation schemes.

14.3.2 M-ary Modulation Schemes

M-ary alphabets are used to improve the spectral efficiency of a telecommunications system by sending multiple data bits using one symbol. In M-ary modulation schemes, the source produces one of M symbols m_i for $i = 1, 2, 3, \cdots, M$. The alphabet m_i is mapped to a signal $s_i(t)$ that lasts for T_s seconds.

In the case of M-ASK, there are M different amplitude values of the carrier. The signal will be represented by:

$$s_i(t) = A_i \cos(2\pi f_c t + \phi),\ 0 \le t \le T_s \text{ for } i = 1, 2, 3, \cdots, M \tag{14.34}$$

M-ASK is also called *pulse amplitude modulation* like its baseband counterpart in Chapter 9. It is common to assume that $A_i = (2i - 1 - M)d$ where $2d$ is the difference between two consecutive signal amplitudes. For now, let us just note that d is some integer value. In Section 14.3.4, we will define the *distance* between two signals when we represent signals as vectors. This value $2d$ will be the distance between adjacent signals. We will also see that the BER depends on half the distance (in this case d) between adjacent signals.

Example. Let $M = 4$ and $d = 1$. The four signal amplitudes will be -3, -1, 1 and 3 V. The M-ASK signals will be:

$$\begin{aligned}
s_1(t) &= \cos(2\pi f_c t + \phi),\ 0 \le t \le T_s \\
s_2(t) &= -\cos(2\pi f_c t + \phi),\ 0 \le t \le T_s \\
s_3(t) &= 3\cos(2\pi f_c t + \phi),\ 0 \le t \le T_s \\
s_4(t) &= -3\cos(2\pi f_c t + \phi),\ 0 \le t \le T_s
\end{aligned}$$

Note that once again, a composite M-ASK signal over many symbol durations does not have a constant amplitude.

In the case of M-FSK, M different carrier frequencies are used to represent the M alphabet symbols. Care must be taken to choose the frequencies such that there is no interference

between adjacent frequency carriers. The M-FSK signal will be given by:

$$s_i(t) = A\cos\left[2\pi f_c t + 2\pi\left(i - \frac{M}{2}\right)\Delta f t\right],\ 0 \le t \le T_s$$
$$\text{for } i = 1, 2, 3, \cdots, M \tag{14.35}$$

The M frequencies will be $f_c + \left(i - \frac{M}{2}\right)\Delta f$ for $i = 1, 2, 3, \cdots, M$. This ensures that the frequencies are equally distributed on either side of f_c. The parameter Δf denotes the separation between two adjacent frequencies and is typically a multiple of $\frac{1}{T_s}$. 4-FSK modulation is employed in systems such as Bluetooth.

In the case of M-PSK, there are M different carrier phases that represent the M alphabet symbols. The M-PSK signal is given by:

$$s_i(t) = A\cos(2\pi f_c t + \phi_i),\ 0 \le t \le T_s \text{ for } i = 1, 2, 3, \cdots, M \tag{14.36}$$

The phase ϕ_i is typically given by $\phi_i = \frac{2\pi}{M}(i-1)+$ constant. Variations of QPSK are used in almost all wireless communication systems like 802.11 wireless LANs and cellular telephony (CDMA and digital TDMA in North America).

Example. Suppose $M = 4$ and the constant is zero. The four phases are $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$. This scheme is commonly called *quadriphase shift keying* or QPSK. If the constant is $\frac{\pi}{4}$, the four phases will be $\frac{\pi}{4}$, $\frac{3\pi}{4}$, $\frac{5\pi}{4}$ and $\frac{7\pi}{4}$. Both schemes are equivalent. In the case of $\frac{\pi}{4}$-QPSK, a variation of QPSK, the symbols are picked alternatively from these two schemes (constant $= 0$ and constant $= \frac{\pi}{4}$) to reduce the amount of discontinuity between adjacent symbols. This helps in keeping the sidelobes of the spectrum of the signal confined to low levels.

Quadrature Amplitude Modulation. An important observation that impacts the bandwidth of modulation schemes is that a sine and a cosine at the same frequency are orthogonal. So it is possible to transmit a carrier at a frequency f_c and another carrier at the same frequency f_c with a phase shift of 90° and be able to differentiate between the two of them easily. This approach enables us to double the symbol rate without doubling the bandwidth required for the transmission. This concept where both a sine and a cosine are simultaneously used for transmitting information is called *quadrature modulation*. The cosine is called the *in-phase* component and the sine is called the *quadrature-phase* component.

If different (multiple positive and negative) amplitudes are used with the two phase-shifted carriers, the modulation scheme is called *quadrature amplitude modulation* (QAM). QAM is a popular bandwidth efficient modulation scheme used in many practical systems. The general M-QAM signal for an M-ary alphabet can be written as:

$$s_i(t) = A_{i,I}\cos(2\pi f_c t) + A_{i,Q}\sin(2\pi f_c t),\ 0 \le t \le T_s$$
$$\text{for } i = 1, 2, 3, \cdots, M \tag{14.37}$$

where the subscripts I and Q refer to the in-phase and quadrature-phase components. Note that we can also write the QAM signal as:

$$s_i(t) = A_i\cos(2\pi f_c t + \phi_i),\ 0 \le t \le T_s$$
$$\text{for } i = 1, 2, 3, \cdots, M \tag{14.38}$$

where $A_i = \sqrt{A_{i,I}^2 + A_{i,Q}^2}$ and $\phi_i = -\tan^{-1}\left(\frac{A_{i,Q}}{A_{i,I}}\right)$. So it is possible for us to think of QAM as a mix of both amplitude and phase shift keying since the message m_i is mapped to a carrier with amplitude A_i and phase ϕ_i. Like M-ASK, it is common in M-QAM to pick the in-phase and quadrature-phase amplitudes such that they are of the form $(2i - 1 - M)d$ where $2d$ is the difference between two consecutive amplitude values and is a measure of the distance between adjacent signals.

QAM is employed in all voice-band modems. QAM is also used in digital subscriber lines. Traditionally, QAM has not been used in wireless systems because of its dependence on the amplitude which will be affected by fading. However, recently, QAM is being considered in wireless communications as well in OFDM based systems.

14.3.3 Demodulation

So far we have described signals associated with modulation schemes analytically. We have also qualitatively described how ASK, FSK and PSK signals can be generated at the transmitter but we have not delved into the details of the transmitter. We will consider a similar approach for understanding the reception and demodulation of signals, and noise analysis of the various digital modulation schemes. We will consider these aspects at a fairly high level without considering what happens at the circuit or electronics level. The subject of transceiver design is fairly involved. A description of the transceivers is available in [1], [2].

In most cases, we assume that the received signal is only corrupted by additive white Gaussian noise (AWGN) with a flat two-sided power spectral density of value $\frac{N_0}{2}$. If the transmitted signal is $s_i(t)$ for some $i \in 1, 2, 3, \cdots, M$, the received signal will be:

$$r(t) = s_i(t) + n(t),\ 0 \leq t \leq T_s \tag{14.39}$$

The goal of the receiver is to determine what $s_i(t)$ was transmitted given that $r(t)$ was received. If the receiver can correctly determine what $s_i(t)$ was, it can determine m_i and thus recover the transmitted information. However, $r(t)$ is corrupted by noise and it is possible that the receiver will sometimes determine that the transmitted signal was $s_j(t)$ where $j \neq i$ when $s_i(t)$ was actually transmitted. We refer to this outcome as an error in reception. The goal of the receiver is to reduce the probability of error to as small a value as possible.

The common metric that is used for performance comparisons is the ratio of the energy per bit (E_b) to the noise power spectral density value N_0. Consider the example of an on-off BASK signal (100% modulation) with two bits as shown in Figure 14.11. The symbol duration is 1s. The transmitted signal consists of a bit "0" and a bit "1." The figure also shows the received signal for different values of $\frac{E_b}{N_0}$. As the $\frac{E_b}{N_0}$ reduces, even the visual difference between the "0" bit and the "1" bit reduces. Remember that a receiver must automatically and electronically detect which bit was transmitted. Visual clues are useful for people, who after all do not sense voltages that well :-). As the received signal gets noisier, detecting which signal was transmitted becomes harder.

So how does the receiver decide which symbol was transmitted in a given time unit of T_s seconds? The first step in the receiver will be *demodulation* of the received signal where the baseband signal is extracted from the carrier. Consider a received BASK signal

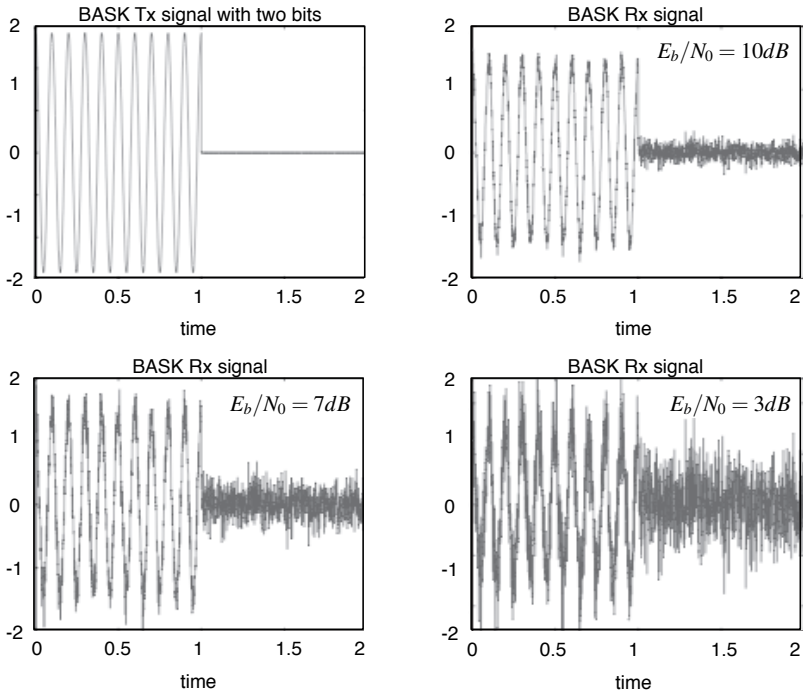

Figure 14.11 Noisy BASK signals with different $\frac{E_b}{N_0}$ values.

where there are no amplitude fluctuations except for AWGN. The received signal over one symbol duration will be of the form:

$$r(t) = A_i \cos(2\pi f_c t) + n(t),\ 0 \leq t \leq T_s \tag{14.40}$$

To recover the number A_i, the receiver will coherently demodulate the signal in a manner similar to AM as shown in Figure 14.12 (a). That is, the receiver computes:

$$\begin{aligned} Z &= \int_0^{T_s} r(t)\cos(2\pi f_c t)\,dt \\ &= \int_0^{T_s} A_i \cos^2(2\pi f_c t)\,dt + \int_0^{T_s} n(t)\cos(2\pi f_c t)\,dt \\ &= \frac{A_i}{2} + \int_0^{T_s} n(t)\cos(2\pi f_c t)\,dt = \frac{A_i}{2} + \mathbf{n} \end{aligned} \tag{14.41}$$

where $\mathbf{n} = \int_0^{T_s} n(t)\cos(2\pi f_c t)\,dt$ is a Gaussian random variable (with zero mean and a variance that is a function of $\frac{N_0}{2}$) that *adds* to the desired quantity $\frac{A_i}{2}$. We will discuss the

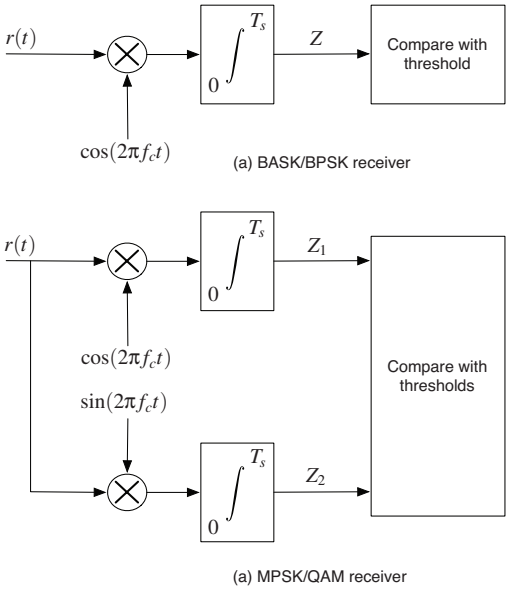

Figure 14.12 Receivers for BASK/BPSK and MPSK.

characteristics of this noise later. To decide which symbol was transmitted, the receiver may simply test whether the value of Z is above or below a threshold.

Example. In the case of BASK, under noise-free conditions, Z would be either $\frac{A_i}{2}$ or 0. Since **n** can be positive or negative, there is a finite probability that an error is made in the decision. But this error can be in picking a "0" when a "1" was transmitted or vice versa because **n** is symmetric. Thus, from a common sense perspective, if Z is above $\frac{A_i}{4}$, the receiver decides that a "0" was transmitted and a "1" otherwise. In the case of MASK, it is possible to define similar thresholds that will enable the receiver to decide which amplitude was actually transmitted.

Example. In the case of BPSK, the receiver needs to detect whether the phase is 0° or 180°. A simple way of determining this would be to perform the following computation:

$$Z = \int_0^{T_s} r(t)\cos(2\pi f_c t)\,dt \tag{14.42}$$

The above computation is similar to the coherent demodulation of BASK shown in Figure 14.12. Depending on the phase, the computed number Z will be as follows:

$$Z = \begin{cases} A\int_0^{T_s}\cos^2(2\pi f_c t)\,dt + \int_0^{T_s} n(t)\cos(2\pi f_c t)\,dt & \text{if the phase is } 0^\circ \\ -A\int_0^{T_s}\cos^2(2\pi f_c t)\,dt + \int_0^{T_s} n(t)\cos(2\pi f_c t)\,dt & \text{if the phase is } 180^\circ \end{cases} \tag{14.43}$$

Note that in the noiseless case ($n(t) = 0$), Z will have a large positive value ($\frac{A}{2}$) when the phase is 0° and a large negative value ($-\frac{A}{2}$) when the phase is 180°. The threshold for comparison will be zero. That is, the receiver decides that the "0" bit was transmitted if Z is positive and the "1" bit was transmitted if Z is negative. The noise term is similar to the noise term $\mathbf{n}$ in (14.41). The effect of the noise term is to change the value of Z. If the noise is positive and the phase was 180°, the value of Z may be moved towards zero and in some cases, Z may become positive resulting in an error in the detected bit.

For MPSK, the situation becomes complex because the receiver has to decide between a large number of possible carrier phases. The received signal will be:

$$r(t) = A\cos(2\pi f_c t + \phi_i) + n(t) \tag{14.44}$$

Simply determining the polarity of Z will not be sufficient. Instead, the receiver will multiply the received signal by both a sine and a cosine carrier (that are locally generated) as shown in Figure 14.12 (b). Let us consider the computation of the receiver outputs below. Multiplication by a local cosine followed by integration over T_s seconds yields:

$$\begin{aligned} Z_1 &= A\int_0^{T_s} \cos(2\pi f_c t + \phi_i)\cos(2\pi f_c t)\,dt + \int_0^{T_s} n(t)\cos(2\pi f_c t)\,dt \\ &= \frac{A}{2}\cos(\phi_i) + n_1 \end{aligned} \tag{14.45}$$

Here n_1 is the noise component at the output that adds to the desired signal component. Multiplication by a local sine followed by integration over T_s seconds yields:

$$\begin{aligned} Z_2 &= A\int_0^{T_s} \cos(2\pi f_c t + \phi_i)\sin(2\pi f_c t)\,dt + \int_0^{T_s} n(t)\sin(2\pi f_c t)\,dt \\ &= \frac{A}{2}\sin(\phi_i) + n_2 \end{aligned} \tag{14.46}$$

In the noiseless case, the receiver makes use of the ordered pair $(Z_1, Z_2) = \left(\frac{A}{2}\cos\phi_i, \frac{A}{2}\sin\phi_i\right)$ to decide which symbol was transmitted.

Example. In the case of QPSK described previously with constant = 0, let us suppose that $\phi_i = 90^\circ = \frac{\pi}{2}$. Then the receiver computes (Z_1, Z_2) as $\left(\frac{A}{2}\cos\left(\frac{\pi}{2}\right), \frac{A}{2}\sin\left(\frac{\pi}{2}\right)\right) = \left(0, \frac{A}{2}\right)$. Thus the receiver decides that the carrier phase is $\frac{\pi}{2}$ if Z_1 is close to zero and Z_2 is positive. The other possibilities are $\left(\frac{A}{2}, 0\right)$, $\left(-\frac{A}{2}, 0\right)$ and $\left(0, -\frac{A}{2}\right)$. These four possibilities correspond to the four phases $\frac{\pi}{2}, 0, \pi$ and $\frac{3\pi}{2}$ respectively. Again, the impact of n_1 and n_2 will be to possibly shift the values of Z_1 and Z_2 such that the decision is erroneous.

In the case of QAM, the quantities Z_1 and Z_2 can take on multiple values and appropriate thresholds are necessary for deciding which amplitude and which phase was transmitted. In the case of FSK, it is typical to choose the frequencies f_i such that the carriers are orthogonal. As described previously, the receiver will multiply the received signal with locally generated carriers of all M frequencies, integrate the product over T_s seconds and pick the one with the largest output as the transmitted frequency.

Note that in all of the above cases, the receiver *multiplies* the received signal $r(t)$ by a locally generated cosine, sine or both and *integrates* the product over a duration of T_s seconds. This process is called *correlation* (see Section 7.3.4 in Chapter 7) and the result of the correlation is a single number Z or two numbers Z_1 and Z_2. We can think of the output of the receiver as a *vector* $\mathbf{Z}$ with two components Z_1 and Z_2. Such a vector has two dimensions because of the two components. Note that we can also represent the output noise $\mathbf{n}$ as a vector with two components n_1 and n_2.

Common receiver architectures make use of a *matched filter* or a *correlator*. In either case, the continuous time signal $r(t)$ is mapped into a single *received vector* $\mathbf{Z}$ with N dimensions in general. It is possible to view the M possible transmit signals $s_i(t)$ for $i = 1, 2, 3, \cdots, M$ also as M vectors $\mathbf{s_i}$ each of dimension N. The representation of the M signals as vectors $\mathbf{s_i}$ in N-dimensional space is called the *signal constellation* corresponding to the modulation scheme (we will discuss some examples in Section 14.3.5). We will see that in the case of ASK, $N = 1$, in the case of PSK and QAM, $N = 2$, and in the case of FSK, $N = M$. For ASK, PSK and QAM, the signal constellation is the same as the phasor diagram of the signals (ignoring the $2\pi f_c t$). Noise being a random quantity cannot be completely represented by a finite dimension vector. However, it can be shown that the receiver performance depends only on the noise components in those dimensions that the signal exists. Hence it is sufficient to consider only an N-dimensional noise vector. For the transmitted signal $s_i(t)$, we can then write in vector notation:

$$\mathbf{Z} = \mathbf{s_i} + \mathbf{n} \tag{14.47}$$

The receiver computes which of the M possible transmit vectors $\mathbf{s_i}$ is *closest* to the received vector $\mathbf{Z}$. Under the AWGN assumption, we can show that the closest transmit vector also corresponds to the most likely transmitted signal. Let us suppose that the receiver computes the received vector to be $\mathbf{Z}$ and determines that the vector $\mathbf{s_k}$ is closest to it. Then it decides that $s_k(t)$ was transmitted, which in turn implies that the message m_k was transmitted. Thus the estimated message is $\hat{m} = m_k$.

The interested reader can look at the mathematical details of representing signals and noise as vectors this in the following section. Otherwise, you can skip this section and go to the next section.

14.3.4 Signal and Noise Representation

An important method of characterizing and analyzing digital communication systems is to use signal space analysis. Given the finite set of waveforms $s_i(t), i = 1, 2, 3, \cdots, M$ that last for a duration T_s seconds each, it is possible to establish that they can be represented as elements of a *finite vector space* (see Chapter 6) spanned by a set of orthonormal basis functions $\varphi_1(t), \varphi_2(t), \cdots, \varphi_N(t)$ where $N \leq M$. It is also possible to represent the received signal that has additive white Gaussian noise as an element of the same vector space. In this section, we provide a preliminary treatment of mapping signals into vectors. This provides a general technique for analyzing the bit-error rate performance of several modulation schemes.

Communication Signals as Vectors in a Vector Space. Consider a communication system as shown in Figure 14.13. In this simplified communication system, we assume that the

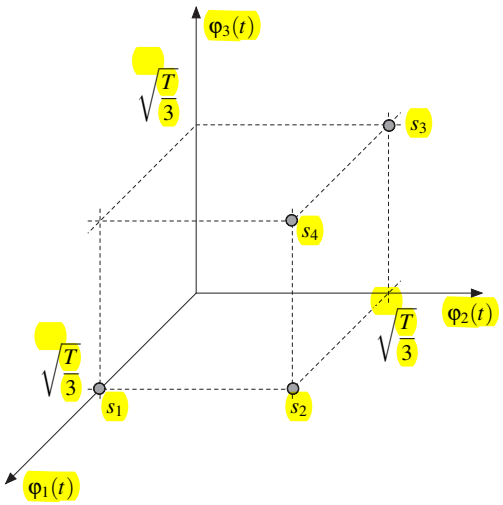

Figure 14.16 Signals computed to determine the orthonormal basis.

Now that we know all the coefficients s_{ij}, we can easily express the signals as vectors as follows:

$$\mathbf{s_1} = \left(\sqrt{\frac{T}{3}}, 0, 0\right) \quad \mathbf{s_2} = \left(\sqrt{\frac{T}{3}}, \sqrt{\frac{T}{3}}, 0\right) \tag{14.69}$$

$$\mathbf{s_3} = \left(0, \sqrt{\frac{T}{3}}, \sqrt{\frac{T}{3}}\right) \quad \mathbf{s_4} = \left(\sqrt{\frac{T}{3}}, \sqrt{\frac{T}{3}}, \sqrt{\frac{T}{3}}\right) \tag{14.70}$$

The four signals are plotted in the vector subspace spanned by $\varphi_1(t), \varphi_2(t)$ and $\varphi_3(t)$ as shown in Figure 14.16.

14.3.5 Signal Constellations and Performance Analysis

We summarize the significant results from the above subsection that are relevant from this point onwards. If a modulation scheme is transmitting one of M signals $s_i(t)$ each of duration T_s seconds, it is possible to decompose these signals into an $N(\leq M)$ dimensional vector space spanned by N basis functions $\varphi_j(t)$ also lasting for T_s seconds. We can write:

$$s_i(t) = \sum_{j=1}^{N} s_{ij}\varphi_j(t) \quad \text{OR} \quad \mathbf{s_i} = (s_{i1}, s_{i2}, \cdots, s_{iN}) \tag{14.71}$$

We can compute the components of $s_i(t)$, namely s_{ij} using the following expression:

$$s_{ij} = \int_0^{T_s} s_i(t)\varphi_j(t)\,dt \tag{14.72}$$

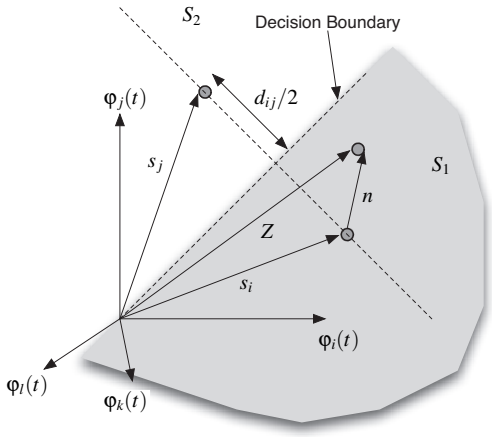

Figure 14.17 Decision at the receiver.

The above operation can be recognized as a *correlation operation* where we are correlating the signal $s_i(t)$ with $\varphi_j(t)$ which is what the receiver does with the received signal as we saw earlier in the discussion of demodulation. The representation of all of the M possible transmit signals in an N-dimensional space is called the *signal constellation* of the modulation scheme.

We can represent the part of noise that impacts performance as an N dimensional vector $\mathbf{n} = (n_1, n_2, \cdots, n_N)$. The components of $\mathbf{n}$ are independent and identically distributed (a normal distribution with mean zero and variance $\frac{N_0}{2}$. If the signal $s_i(t)$ was transmitted, the received signal $r(t) = s_i(t) + n(t)$. We can write $r(t)$ as a vector $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_N)$ where:

$$\begin{aligned} Z_j &= \int_0^{T_s} r(t)\varphi_j(t)\,dt \\ &= \int_0^{T_s} s_i(t)\varphi_j(t)\,dt + \int_0^{T_s} n(t)\varphi_j(t)\,dt \\ &= s_{ij} + n_j \end{aligned} \tag{14.73}$$

In vector notation,

$$\mathbf{Z} = \mathbf{s_i} + \mathbf{n} \tag{14.74}$$

The receiver computes the *distance* between $\mathbf{Z}$ and every possible transmit signal $\mathbf{s_i}$. It then picks the signal $\mathbf{s_k}$ that is closest to $\mathbf{Z}$ as the transmitted signal. Figure 14.17 shows the process for two possible transmit signals $\mathbf{s_i}$ and $\mathbf{s_j}$. The received signal vector is $\mathbf{Z}$ which is the sum of the transmitted signal vector $\mathbf{s_i}$ and the noise vector $\mathbf{n}$. The dashed line denoted "decision boundary" separates the signal space into two regions $\mathcal{S}_1$ and $\mathcal{S}_2$. If $\mathbf{Z}$ falls in $\mathcal{S}_1$, it is closer to $\mathbf{s_i}$ which is what is desired. If however $\mathbf{Z}$ falls in $\mathcal{S}_2$, it is closer to $\mathbf{s_j}$ which is not the transmitted signal. In such a case, an error is made at the receiver. You can see from

Figure 14.17, that the error depends upon the characteristics of the noise vector **n** and *the distance d_{ij} between the signals* $\mathbf{s_i}$ *and* $\mathbf{s_j}$ *in signal space*. In particular, if the noise vector has a component that is larger than $\frac{d_{ij}}{2}$ in the correct direction along the line joining $\mathbf{s_i}$ and $\mathbf{s_j}$, an error is made at the receiver.

For binary modulation schemes, we have two signals $\mathbf{s_1}$ and $\mathbf{s_2}$. Let the distance between the two signals be d and let the noise be additive, white and Gaussian with a two-sided PSD of $\frac{N_0}{2}$. Then, it is possible to show that the probability that the receiver picks $\mathbf{s_1}$ given that $\mathbf{s_2}$ was transmitted or vice versa is given by:

$$P_e = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{d^2}{4N_0}}\right) \tag{14.75}$$

Notice that this is also the probability of bit error since each signal in a binary modulation scheme carries one bit.

Next let us see how we can represent signals from various example modulation schemes as vectors and determine the performance for simple cases.

Binary Modulation Schemes. BPSK and ASK signals are one dimensional. From (14.32) and (14.34), we can see that the signals depend only on $\cos(2\pi f_c t)$. It can be shown that:

$$\varphi_1(t) = \sqrt{\frac{2}{T_s}}\cos(2\pi f_c t), \tag{14.76}$$

is the single basis function for BPSK and all ASK signals. You can easily compute the energy in $\varphi_1(t)$ and see that it equals one.

Let us consider BPSK. The energy in any of the $s_i(t)$ in BPSK was $\frac{A^2 T_s}{2}$ which is the energy in one bit. That is,

$$E_b = \frac{A^2}{2}T_s \quad \Rightarrow \quad A = \sqrt{\frac{2E_b}{T_s}} \tag{14.77}$$

We can rewrite the signals for BPSK as:

$$s_i(t) = \pm A\cos((2\pi f_c t),\ 0 \le t \le T_s \quad \text{or}$$

$$s_i(t) = \pm\sqrt{\frac{2E_b}{T_s}}\cos(2\pi f_c t),\ 0 \le t \le T_s \tag{14.78}$$

Clearly, $s_1(t) = \sqrt{E_b}\varphi_1(t)$ and $s_2(t) = -\sqrt{E_b}\varphi_1(t)$. In vector notation, $\mathbf{s_1} = \sqrt{E_b}$ and $\mathbf{s_2} = -\sqrt{E_b}$. The signal constellation corresponding to BPSK is shown in Figure 14.18 (a). Similarly, it is possible to determine the signal constellations of BASK with 100% and 50% depths. For BASK (100%), the the average energy per bit assuming that zeros and ones occur with equal probability is:

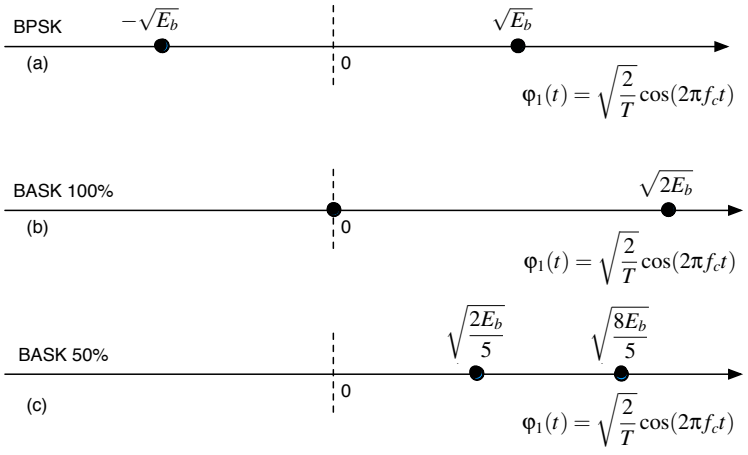

Figure 14.18 Signal constellations of BPSK and BASK.

$$E_b = \frac{1}{2} \times \left[\frac{A^2}{2} T_s + 0 \right] \quad \Rightarrow \quad A = \sqrt{\frac{4E_b}{T_s}} \tag{14.79}$$

Clearly, $s_1(t) = \sqrt{2E_b}\varphi_1(t)$ and $s_2(t) = 0 \times \varphi_1(t)$. In vector notation, $\mathbf{s_1} = \sqrt{2E_b}$ and $\mathbf{s_2} = 0$. The signal constellation corresponding to BASK (100%) is shown in Figure 14.18 (b). For BASK (50%), the average energy per bit assuming that zeros and ones occur with equal probability is:

$$E_b = \frac{1}{2} \times \left[\frac{A^2}{2} T_s + \frac{A^2}{8} T_s \right] \quad \Rightarrow \quad A = \sqrt{\frac{16E_b}{5T_s}} \tag{14.80}$$

Clearly, $s_1(t) = \sqrt{\frac{8E_b}{5}}\varphi_1(t)$ and $s_2(t) = \sqrt{\frac{2E_b}{5}} \times \varphi_1(t)$. In vector notation, $\mathbf{s_1} = \sqrt{\frac{8E_b}{5}}$ and $\mathbf{s_2} = \sqrt{\frac{2E_b}{5}}$. The signal constellation corresponding to BASK (100%) is shown in Figure 14.18 (c). Notice that the *signal distance* between points in the constellation successively reduces as we move from BPSK to the two BASK schemes. The noise vector can be smaller and still create errors as the distance reduces. Thus, the BER for the same average E_b/N_0 per bit will be larger for BASK.

Figure 14.19 shows the signal constellation of orthogonal BFSK. Even though this is a binary modulation scheme, there are two basis functions corresponding to the two frequencies. They are orthogonal to one another. The amplitude of each carrier is $A = \sqrt{\frac{2E_b}{T_s}}$ in a manner similar to BPSK. The energy in each symbol and the average energy per bit is E_b.

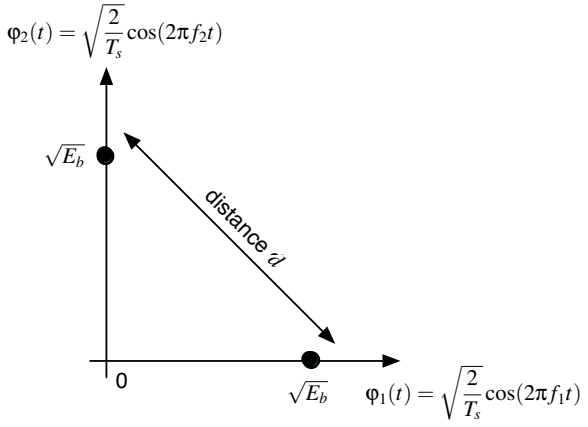

Figure 14.19 Signal constellation of BFSK.

Using (14.75), we can now compute the probability of error for the binary modulation schemes. The distances between the signal points are:

$$\begin{aligned} \text{Distance for BPSK} &= 2\sqrt{E_b} \\ \text{Distance for BASK–100\%} &= \sqrt{2E_b} \\ \text{Distance for BASK–50\%} &= \sqrt{\frac{2E_b}{5}} \\ \text{Distance for BFSK} &= \sqrt{2E_b} \end{aligned} \tag{14.81}$$

The probabilities of bit error are:

$$P_e(\text{BPSK}) = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{14.82}$$

$$P_e(\text{BASK (100\%)}) = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{2N_0}}\right) \tag{14.83}$$

$$P_e(\text{BASK (50\%)}) = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{10N_0}}\right) \tag{14.84}$$

$$P_e(\text{BFSK}) = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{2N_0}}\right) \tag{14.85}$$

Figure 14.20 shows the bit error rate curves for the different binary modulation schemes. The x-axis has $\frac{E_b}{N_0}$ in dB and the y-axis has the probability of error on a logarithmic scale. You can see that BPSK has the best performance of the four schemes and BASK (50%) has

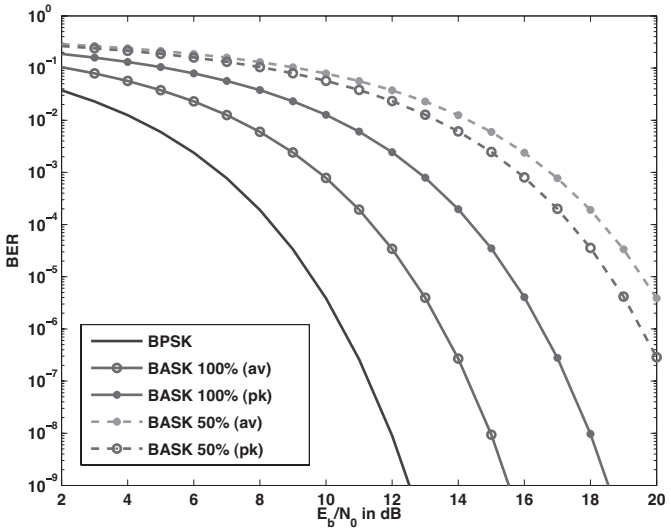

Figure 14.20 Bit error rates for binary modulation schemes.

the worst performance. Recall that the objective of the modulation scheme is to achieve as small a bit error rate as possible while expending as little energy as possible. For a bit error rate of 10^{-5}, BPSK needs an $\frac{E_b}{N_0}$ of approximately 10 dB, BFSK and BASK (100%) need an $\frac{E_b}{N_0}$ of 13 dB (this is 3 dB larger) and BASK (50%) is around 19 dB. The reason for this is BASK (50%) expends energy in both symbols used for transmission, but the symbols are alike (distance between them in signal space is small). For two dimensional (binary) modulation, BPSK is the optimal modulation scheme. No other modulation scheme can perform as well in an AWGN channel.

Peak Vs Average $\frac{E_b}{N_0}$**.** In the previous discussion, we have considered BPSK and BASK where the modulation schemes have the same average energy per bit. As we discussed in Chapter 10, photonic systems are limited by the peak power output of LEDs and lasers and it is therefor more practical and useful to compare the performance as a function of the peak power or peak signal-to-noise ratio for photonics. If we are to consider BPSK and BASK with the *same peak power*, the signal constellations will be quite different. Consider the following three signals for $0 \leq t \leq T$:

$$\text{BPSK: } s(t) = \pm\sqrt{\frac{2E_b}{T}}\cos(2\pi f_c t) \tag{14.86}$$

$$\text{BASK}-100\%\text{: } s(t) = 0 \text{ OR } \sqrt{\frac{2E_b}{T}}\cos(2\pi f_c t) \tag{14.87}$$

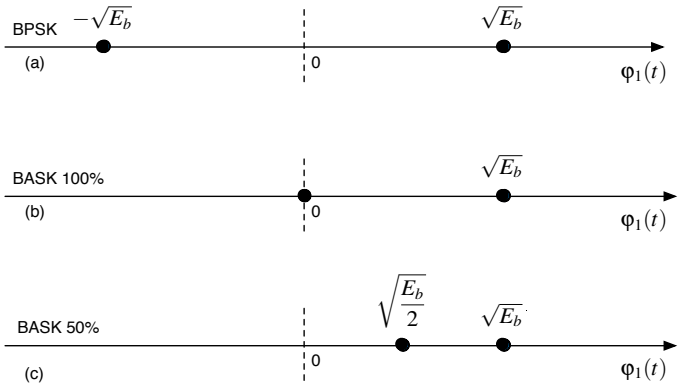

Figure 14.21 Signal constellation of BFSK.

$$\text{BASK}-50\%\text{: } s(t) = \sqrt{\frac{E_b}{T}}\cos(2\pi f_c t) \text{ OR } \sqrt{\frac{2E_b}{T}}\cos(2\pi f_c t) \tag{14.88}$$

These three signals all have the same peak power. The constellations are shown in Figure 14.21. The distances between the signal points in these constellations are $2\sqrt{E_b}$, $\sqrt{E_b}$ and $\sqrt{\frac{E_b}{2}}$ respectively. The probabilities of error in the three cases will be:

$$P_e(\text{BPSK}) = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{14.89}$$

$$P_e(\text{BASK (100\%)}) = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{4N_0}}\right) \tag{14.90}$$

$$P_e(\text{BASK (50\%)}) = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{8N_0}}\right) \tag{14.91}$$

The probabilities of error are also shown in Figure 14.20. Observe that if the peak powers are the same, but the average power is lower, the BER performance of BASK (100%) is worse. We caution the reader that this is only an artifact of the presentation of the BER curves as a function of E_b/N_0. If you are making the right comparison using the right metric, the results from the different curves will be the same as demonstrated by this example.

Example. Consider BASK with 100% modulation depth. The amplitudes of the signals associated with the "0" and "1" symbols are 1 V and 0 V respectively. The symbol lasts for 1 s. The energy in the signal corresponding to "0" is $\frac{1}{2}$ J. The energy in the signal corresponding to "1" is 0 J. The peak power in the BASK signal is $\frac{1}{2}$ W. The average power

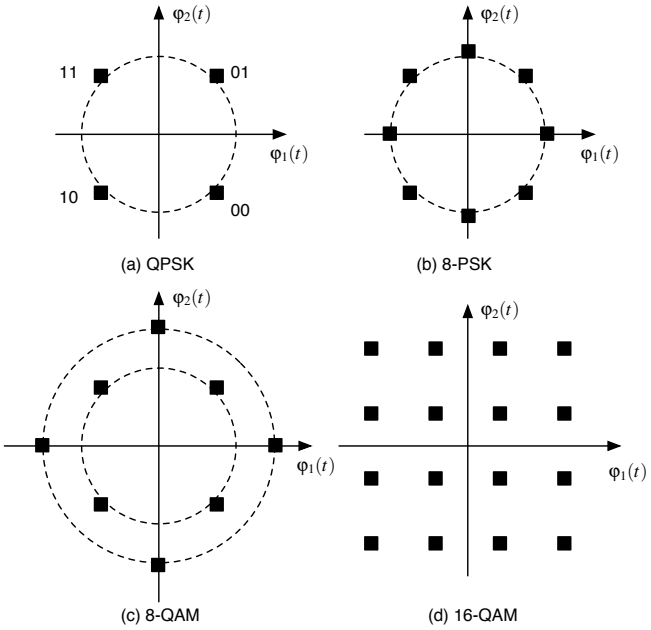

Figure 14.22 Signal constellations of QPSK, 8-PSK, 8-QAM and 16-QAM.

is $\frac{1}{4}$ W. Although this is not a realistic value of N_0, let us assume that it is 0.025 W/Hz. Then the average $\frac{E_b}{N_0} = \frac{0.25}{0.025} = 10$ (10 dB) and the peak $\frac{E_b}{N_0} = \frac{0.5}{0.025} = 20$ (13 dB). From Figure 14.20, the BER using the average BASK 100% curve is 3×10^{-4} at $\frac{E_b}{N_0} = 10$ dB. Note that the BER using the peak BASK 100% curve is also 3×10^{-4} at but at $\frac{E_b}{N_0} = 13$ dB.

M-ary Modulation Schemes. For PSK modulation with $M > 2$ and QAM, there are two basis functions given by:

$$\varphi_1(t) = \sqrt{\frac{2}{T}} \cos(2\pi f_c t)$$
$$\varphi_2(t) = \sqrt{\frac{2}{T}} \sin(2\pi f_c t) \tag{14.92}$$

Figure 14.22 shows signal constellations for QPSK, 8-PSK, 8-QAM and 16-QAM. The figure does not show the exact coordinates of the signal points. We leave this as an exercise to the reader. However, some remarks are necessary.

- The norm of the vector (essentially the length of the line joining the origin to the signal point) is a measure of the energy E_s in that signal point or symbol. Note that each symbol carries $k = \log_2 M$ bits. The energy per bit will be $\frac{E_s}{k}$ for *that particular*

symbol. While calculating the average energy per bit for a given modulation scheme is straightforward (assuming that symbols are equally likely), it can be cumbersome.

- As the number M of signal points increases, to keep E_s approximately uniform across symbols, it will become necessary to reduce the distance between symbols. In the case of M-PSK, all signal points are at the same distance from the origin. This implies that the signal points get closer as M increases. It is easy to show that the distance between any two adjacent signal points in M-PSK is proportional to $\sin(\frac{\pi}{M})$.
- Determining the probability of error is not trivial because of the following. Suppose the signal $\mathbf{s_1}$ was transmitted. The receiver could decide that any of the remaining $M-1$ signals were transmitted depending on what the noise does to the received signal (the direction and length of the noise vector). Although the noise vector has a Gaussian distribution, all signal points in the constellations are not at the same distance from one another. Thus the probabilities of making an error between one signal point and each of the other signal points will be different. Moreover, each signal carries k bits. Depending on which signal was incorrectly decided upon as the transmitted signal, some of the k bits could be in error and others may not.
- The way in which symbols are mapped into bits plays a role in the probability of error. It can be shown that in an AWGN channel, it is more likely that errors are made between adjacent signal points in the constellation. It is thus common to make use of what is called a *gray code* where adjacent symbols differ in the smallest number of bits (usually by one bit). If a gray code is used (see Figure 14.22 (a) for gray code with QPSK), the probability of bit error can be reduced.
- There is a tradeoff between increasing M and the probability of error for a given $\frac{E_b}{N_0}$. In the cases of both MPSK and QAM, as M increases (except for $M=4$), the probability of error also increases for a given $\frac{E_b}{N_0}$ compared to BPSK. However, we can send more bits per second in the same bandwidth as the BPSK signal (discussed in Section 14.3.6). The approximate equation for the probability of bit error of M-PSK as a function of $\frac{E_b}{N_0}$ are as follows:

$$P_e(\text{M-PSK}) \approx \frac{1}{k}\text{erfc}\left(\sin\left(\frac{\pi}{M}\right)\sqrt{\frac{kE_b}{N_0}}\right) \tag{14.93}$$

  For M-QAM, it is common to use bounds on the symbol error probability rather than the bit error probability.
- In QAM, all signal points do not have the same E_s. You can see that some signal points are closer to the origin than others. The peak power in a QAM signal is thus quite different from the average power requiring the use of highly linear amplifiers to prevent distortion.

A second class of M-ary signals correspond to M-FSK. They are M dimensional signals if the M carriers with different frequencies are orthogonal to one another. The advantage of this scheme is that the probability of error reduces as M increases unlike M-PSK or