Chapter 1

Digital Communications

1.1 Introduction

For the past 25-30 years digital communications system having been replacing their corresponding analog counterparts. While it has long been known that digital systems offer tremendous advantages over analog systems, only in the last 20 years has technology advanced to the point to allow digital systems to be a cost-effective alternative. Cell-phones, cordless telephones, military systems, and land-line communications have all seen digital technology replace analog. Among the advantages that digital systems offer are

- Increased fidelity through error correction coding
- Mixed information types on a single system
- The opportunity for bandwidth reduction via source coding
- More flexible trade-offs between bandwidth and performance
- Opportunities to improve performance via advanced signal processing

A block diagram of a typical digital communications system is shown in Figure 1.1. Digital Communications systems can transmit either information that is inherently digital (e.g., computer files) or information that is analog but has been converted to a digital format through sampling and quantization. Digital information can also be compressed through source coding and/or encryption. Source coding removes redundancy in the information to reduce the requirements for transmission while encryption provides security. This new information stream can then be further encoded to provide protection against errors during transmission. This operation expands the data by adding redundancy that can be exploited to detect and possibly correct bit errors.

Modulation accepts the coded bit stream and converts the bits to symbols for transmission using electromagnetic waves. This operation typically involves modulating either the frequency, phase or amplitude of a sinusoidal carrier. The resulting signal is launched into the channel which will attenuate the signal. The most benign channel will attenuate the signal based on the distance traveled in a manner which is constant in time and frequency. In many systems, however, the channel may also have a limited bandwidth, which can cause frequency distortion. In wireless systems, mobility combined with multipath can induce a time-varying channel which is typically referred to as a *fading* channel since the time-varying attenuation can be quite severe resulting in a "fade" of the received signal. When multipath components have relative arrival times that are separated by more than the symbol duration, the attenuation experienced can also vary across frequency. Typically this "frequency-selective" fading must be combated through equalization techniques, although spread spectrum systems offer a specific advantage over this type of distortion as we will see.

At the receiver, the reverse operations must be performed. After filtering, down-conversion and demodulation the receiver will typically perform equalization, if necessary, followed by decoding. This is in turn followed by decryption, if employed, source decoding and, if the information is analog, digital-to-analog conversion.

We will discuss several of these steps in detail shortly. However, before we do so, we will review two concepts that will prove to be quite useful in the analysis of communication systems: bandpass signals and random processes. Note that the coverage here is intended to merely reacquaint the reader with these topics. For a more detailed description of general digital communications, the reader is encouraged to refer to one or more of the many fine texts in the area [3, 4, 5, 7, 8].

1.2 Bandpass Signals

Communication signals typically involve carrier modulation and are thus bandpass signals. However, the use of bandpass signals for analysis is somewhat awkward. Thus, typically we will use the complex baseband representation of bandpass signals. We will briefly describe different representations of bandpass signals. Consider a signal $z(t)$ which has non-zero frequency components only in a relatively small region $|f - f_o| \leq W$ where $W << f_o$. We can define f_o as the center frequency and W as the signal bandwidth. This is considered a bandpass signal. A bandpass signal can be represented as

$$z(t) = z_I(t) \cos 2\pi f_o t - z_Q(t) \sin 2\pi f_o t \tag{1.1}$$

where $z_I(t)$ and $z_Q(t)$ are real baseband signals and are called the in-phase and quadrature components of $z(t)$ respectively. This representation is termed the in-phase and quadrature representation of bandpass signals and is useful for developing transmitter and receiver structures. However, it isn't particularly intuitive when it comes to understanding modulation. Bandpass signals can also be represented as

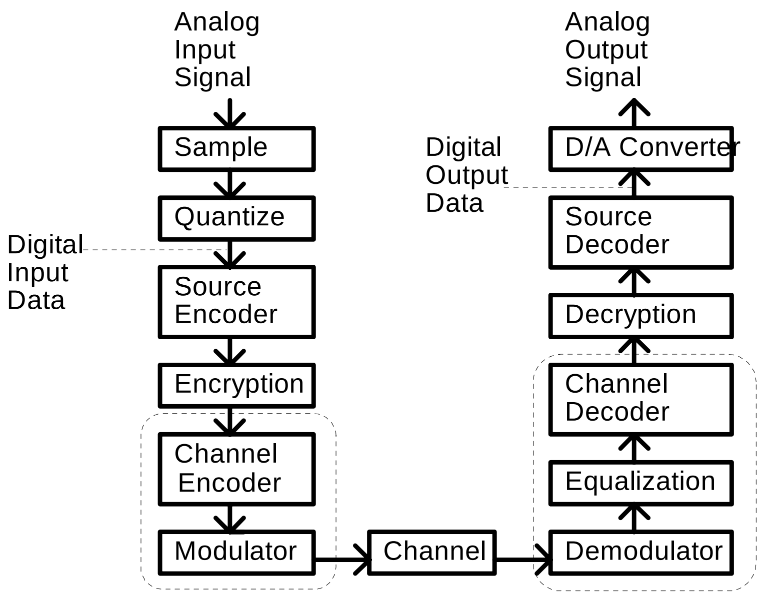

Figure 1.1: Block Diagram of a Typical Digital Communications System

$$z(t) = A(t)\cos\left(2\pi f_o t + \theta(t)\right) \tag{1.2}$$

where $A(t)$ and $\theta(t)$ are real baseband signals that represent the amplitude and phase modulation respectively. It is relatively straightforward to show that

$$A(t) = \sqrt{z_I^2(t) + z_Q^2(t)} \tag{1.3}$$

$$\theta(t) = \tan^{-1}\left(\frac{z_Q(t)}{z_I(t)}\right) \tag{1.4}$$

Finally, a bandpass signal can be written in complex baseband representation as

$$z(t) = \mathrm{Re}\left\{\widetilde{z}(t)e^{j2\pi f_o t}\right\} \tag{1.5}$$

where $\widetilde{z}(t)$ is a complex baseband signal and can be written as

$$\widetilde{z}(t) = z_I(t) + jz_Q(t) \tag{1.6}$$

We will find that this last representation is the most useful in terms of analysis, since it provides us a means for eliminating the awkward carrier component.

1.3 Random Processes

The use of random processes is fundamental to the study of communication systems. In this section we briefly review random processes. The reader is referred to [1, 2, 6] for a detailed discussion. A random process is a collection of time functions where each time function in the set (termed a sample function) has at least one parameter that is the instantiation of a random variable. Each sample function is one realization of the random process and is one member of the ensemble of sample functions. A random process is described by an N-fold joint distribution function:

$$f_{X(t_1),\ldots X(t_n)}(x_1,\ldots x_n;t_1,\ldots t_n) = P\left\{X(t_1)\leq x_1;\ldots X(t_n)\leq x_n\right\} \tag{1.7}$$

where $x_i = X\left(t_i\right)$ is a random variable. If the first order density function $f_{X(t_1)}(x_1;t_1)$ is independent of time t_1, that is

$$f_{X(t_1)}(x_1;t_1) = f_{X(t_1+\tau)}(x_1;t_1+\tau) \tag{1.8}$$

for any τ then the random process is called *first order stationary.* A result of first order stationarity is that

$$\mathrm{E}\left\{X(t)\right\} = \mu_X = \text{constant} \tag{1.9}$$

where $E\{\cdot\}$ is the expectation operator and μ_X is termed the mean. In other words, we may measure the mean of the random process at any time t. A strictly stationary process is a random process for which all n, all $(t_1,t_2,\ldots t_n)$ and all τ,

$$f_{X(t_1),\ldots X(t_n)}(x_1,\ldots .x_n) = f_{X(t_1+\tau),\ldots X(t_n+\tau)}(x_1,\ldots x_n) \tag{1.10}$$

Strict stationarity is not typically required for proper analysis. A more relaxed requirement is Wide-Sense Stationarity. A wide-sense stationary process is one for which

- $\mathrm{E}\left\{X(t)\right\} = \mathrm{E}\left\{X(t+\tau)\right\} = \mu_X$
- $\mathrm{E}\left\{X(t_1)X(t_2)\right\} = \mathrm{E}\left\{X(t_1)X(t_1+\tau)\right\} = R_X(\tau)$

  In this book we will typically assume that random processes are wide-sense stationary.

Another important class of random processes is the class of ergodic random processes. An ergodic random process is one for which the time and ensemble averages are interchangeable. For example

$$\mathrm{E}\left\{(X(t))\right\} = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} x(t)dt \tag{1.11}$$

and

$$\mathrm{E}\left\{(X(t)X(t+\tau))\right\} = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} x(t)x(t+\tau)dt \tag{1.12}$$

Note that for strict ergodicity, all ensemble averages must equal their corresponding time averages, not just the first and second order averages. Further, it should be noted that ergodicity requires stationarity and is thus a more strict requirement than strict stationarity. We will commonly assume ergodicity although it is not strictly necessary in many cases.

Another important class of random processes is the class of Gaussian random processes. A Gaussian random process is a process $X(t)$ for which the random variables $\{X(t_i)\}_{i=1}^{n}$ (i.e., n time samples of the random process) have a joint Gaussian density function. That is

$$f_X\left(x_1,\ldots x_n;t_1,\ldots t_n\right)=\frac{1}{\sqrt{(2\pi)^n\left|\mathbf{C}_X\right|}}exp\left\{-(1/2)\left[\mathbf{x}-\overline{\mathbf{x}}\right]^T\mathbf{C}_X^{-1}\left[\mathbf{x}-\overline{\mathbf{x}}\right]\right\} \tag{1.13}$$

where

$$\left[\mathbf{x}-\overline{\mathbf{x}}\right]=\left[\begin{array}{c} x_1-\overline{x_1}\\ x_2-\overline{x_2}\\ \cdots\\ x_n-\overline{x_n}\end{array}\right], \tag{1.14}$$

$\overline{x_i}$ is the mean of x_i,

$$\mathbf{C}_X=\left[\begin{array}{cccc} C_{11} & C_{12} & \ldots & C_{1n}\\ C_{21} & C_{22} & \ldots & C_{2n}\\ \vdots & \vdots & & \vdots\\ C_{n1} & C_{n2} & \ldots & C_{nn}\end{array}\right], \tag{1.15}$$

is the covariance matrix and

$$C_{ij}=\mathrm{E}\left\{\left(x_i-\overline{x_i}\right)\left(x_j-\overline{x_j}\right)\right\} \tag{1.16}$$

The most relevant examples of Gaussian random processes are Additive White Gaussian Noise and the Rayleigh fading channel. Note that a *White* process is a process that has a spectral density $S(f)$ which is constant. In other words based on the Wiener-Khintchine Theorem, the autocorrelation function $R_{xx}(\tau)=\mathcal{F}\left\{R_{xx}(\tau)\right\}=\delta(\tau)$.

1.4 Sampling

The analog-to-digital process consists primarily of sampling and quantization. These two procedures convert an analog signal to a signal which is discrete in time and amplitude. These values can then be converted directly to data bits or further processed to compress the information (e.g., in vocoders). Sampling is the process of converting an analog signal to one which is discrete in time. If samples are taken at a uniform rate, it was shown by Nyquist that the original signal can be reconstructed from samples that are taken at a rate f_s provided that

$$f_s\geq 2B \tag{1.17}$$

where B is the absolute bandwidth of the signal of interest. This is true of both baseband and bandpass signals. However, if B is the bandwidth of the baseband equivalent signal and B_T is the bandpass bandwidth, the sampling rate can be written as

$$\begin{aligned} f_s &\geq 2B_T \\ &\geq 4B \end{aligned} \tag{1.18}$$

In terms of the complex baseband we can say that we must sample the real component which has bandwidth B at $f_s \geq 2B$ and we must sample the imaginary component which also has bandwidth B at $f_s \geq 2B$ for a total sampling rate of $f_s \geq 4B$.

We can also view sampling through the interpolation formula which states that a signal can be reconstructed from its samples as

$$x(t) = \sum_n x\left(\frac{n}{2B}\right) \frac{\sin 2\pi B\,(t - n/2B)}{2\pi B\,(t - n/2B)} \tag{1.19}$$

To see this, recall that the sampling theorem tells us that we can completely represent a band-limited signal $x(t)$ by samples of the signal $x(nT_s)$ taken at an appropriate frequency $f_s = \frac{1}{T_s}$. Formally, if we assume an impulse-sampled signal we can write the sampled signal as

$$\begin{aligned} x_s(t) &= \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s) \\ &= x(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \end{aligned} \tag{1.20}$$

where $\delta(t) = \begin{cases} 1 & t = 0 \\ 0 & else \end{cases}$. Now taking the Fourier transform of (1.20) results in

$$\begin{aligned} X_s(f) &= X(f) * \mathcal{F}\left\{ \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \right\} \\ &= X(f) * \frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_s}\right) \\ &= \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X\left(f - \frac{n}{T_s}\right) \end{aligned} \tag{1.21}$$

where $*$ is the convolution operation. Now we see that sampling in the time domain results in spectral replication in the frequency domain. If $f_s > 2B$ where B is the bandwidth of the signal, there will be no overlapping in frequency of the sampled signal and thus no distortion. Now if we low-pass filter the signal *i.e.,* we multiply the spectrum by a brick wall filter with response $H(f) = \begin{cases} T_s & |f| < B \\ 0 & else \end{cases}$, we obtain the original spectrum:

$$\begin{aligned} Y(f) &= H(f)X_s(f) \\ &= X(f) \end{aligned} \tag{1.22}$$

Looking at this in the time domain we have

$$\begin{aligned} y(t) &= h(t) * x_s(t) \\ &= 2BT_s \text{sinc}\,(2Bt) * \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s) \\ &= \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \text{sinc}\,(2Bt - n) \\ &= \sum_{n=-\infty}^{\infty} x\left(\frac{n}{T_s}\right) \text{sinc}\left(\frac{t}{T_s} - n\right) \end{aligned} \tag{1.23}$$

Thus, if we use the sinc function for interpolating the signal, we can reconstruct the signal perfectly. Note that $\text{sinc}\,(x) = 0$ for $x = 1, 2, 3, \ldots n$. If the signal is sampled at the Nyquist rate ($f_s = 2B$) an ideal brick wall filter is needed. However, such a filter is impossible to build in practice since it requires an infinite impulse response. Realistic systems have finite roll-off and will thus introduce some finite amount of aliasing, although it may be negligible.

1.5 Quantization

Quantization is the second step in the analog-to-digital process. Whereas sampling converts a signal which is continuous in time to one which is discrete in time, quantization converts a signal which is continuous in amplitude to one which is discrete in amplitude. This is required simply because we cannot represent an infinite number of values with a finite number of digital words. Unlike the sampling process, quantization necessarily introduces distortion into the resulting digital signal. Specifically, at each sampling instance there is some error introduced

$$e_i = s(t_i) - f_q\left\{s(t_i)\right\} \tag{1.24}$$

where $f_q\left\{x\right\}$ is the quantization function. If the input signal $s(t)$ is uniformly distributed (note that in general $s(t)$ is not a deterministic function, but rather a random process) and a uniform quantizer is used, the resulting signal-to-noise ratio can be written as

$$\frac{S}{N} = 4n \tag{1.25}$$

where n is the number of bits per sample used in the quantization process, $S = \mathrm{E}\left\{s^2(t)\right\}$ is the average signal power and $N = \mathrm{E}\left\{e_i^2\right\}$ is the resulting average error power or noise power.

1.6 Modulation

Modulation is the process of converting data bits into symbols for transmission. Typically symbols are sinusoids of radio frequency (RF) with a specific set of phases, amplitudes and/or frequencies. For example, one of the simplest modulation schemes is what is referred to as on-off keying (OOK). This can also be called binary amplitude shift keying (BASK). In this scheme we use the mapping

$$\begin{array}{lcll} b=0 & \rightarrow & s(t)=0 & 0 \leq t \leq T_s \\ b=1 & \rightarrow & s(t)=A\cos(\omega_c t) & 0 \leq t \leq T_s \end{array} \tag{1.26}$$

where b is the information bit to be transmitted, $s(t)$ is the transmitter symbol, and ω_c is the carrier frequency in radians per second. In other words the amplitude of the sinusoid carries the information. An amplitude of 0 represents a binary '0' and an amplitude of A represents a binary '1' and thus this scheme is called Amplitude Shift Keying. Information can also be encoded into the phase or frequency of a sinusoid. When the former is done, it is called Phase Shift Keying or PSK. In Binary PSK or BPSK, the mapping is

$$\begin{array}{lcll} b=0 & \rightarrow & s(t)=A\cos(\omega_c t) & 0 \leq t \leq T_s \\ b=1 & \rightarrow & s(t)=A\cos(\omega_c t+\pi) & 0 \leq t \leq T_s \end{array} \tag{1.27}$$

is used. In Frequency Shift Keying or FSK, the information is encoded in the frequency of the sinusoid. For Binary FSK or BFSK we use the mapping

$$\begin{array}{lcl} b=0 & \rightarrow & s(t)=A\cos(\omega_1 t) \\ b=1 & \rightarrow & s(t)=A\cos(\omega_2 t) \end{array} \quad 0 \leq t \leq T_s \tag{1.28}$$

where ω_1 and ω_2 are typically chosen such that the two symbols are orthogonal over a symbol period and are centered around the nominal carrier frequency ω_c.

1.6.1 M-ary Modulation

The previous modulation schemes are known as *binary* modulation schemes since there are only two symbols, one for each bit value. It is also possible to map n bits per symbol. This is called M-ary modulation where there are $M = 2^n$ different symbols. For example, with M-PSK the modulation is still contained in the phase of the sinusoid, but there are M different possible phase values corresponding to $M = 2^n$ different groups of binary '1's and '0's. The M symbols can be defined as

$$s_i(t) = A\cos\left(\omega_c t + (i-1)\tfrac{2\pi}{M}\right) \quad 0 \leq t \leq T_s \tag{1.29}$$

where $i = 1, 2, ...M$. M-ary ASK and M-ary FSK are also possible where each group of $n = \log_2 M$ bits is mapped to one of M symbols each with a different amplitude or frequency respectively.

1.6.2 Pulse Shaping

Each of the above described modulation schemes can be written in the form

$$x(t) = \sum_i s_i\left(t - iT_s\right) \tag{1.30}$$

where i is the symbol index and T_s is the symbol period. Since the symbols each last over a single symbol period, we can think of each symbol as being a sinusoid with a particular amplitude, phase and frequency multiplied by a square pulse of width T_s and centered at $\left(i - \frac{1}{2}\right)T_s$. A square pulse has a spectrum that is related to the sinc function. Because of the slow decay of the sidelobes of the sinc function, other pulse shapes are often used rather than a simple rectangular pulse. The optimal pulse shape is a sinc function, since its spectral properties are related to a square wave (i.e., it has minimum bandwidth). However, since the sinc function has infinite duration in both the positive and negative time directions, it is impossible to implement in practice. A function which has good bandwidth properties and reasonable implementation complexity is the raised cosine pulse.

1.6.3 Performance

The performance of a modulation scheme is specified in terms of the probability of making a bit error for a specified signal-to-noise ratio (SNR) in an Additive White Gaussian Noise (AWGN) channel. Since modulation schemes can have different numbers of bits per symbol, and bits are the fundamental unit of information, the SNR per bit or $\frac{E_b}{N_o}$ is typically used as the basis for comparison.

Let us first consider a binary modulation scheme with two symbols $s_1(t)$ and $s_2(t)$. Since there are two symbols, we can completely represent the two symbols with at most two basis functions $\phi_1(t)$ and $\phi_2(t)$ where by definition

$$\int_0^{T_s} \phi_1^2(t)dt = \int_0^{T_s} \phi_2^2(t)dt = 1 \tag{1.31}$$

and

$$\int_0^{T_s} \phi_1\phi_2(t)dt = 0 \tag{1.32}$$

The two signals can be represented as

$$\begin{aligned} s_1(t) &= s_{11}\phi_1(t) + s_{12}\phi_2(t) \\ s_2(t) &= s_{21}\phi_1(t) + s_{22}\phi_2(t) \end{aligned} \tag{1.33}$$

where

$$s_{ij} = \int_0^T s_i(t)\phi_j(t)dt \tag{1.34}$$

Now since we can choose the first basis function arbitrarily, let's choose

$$\phi_1(t) = \frac{1}{\sqrt{E_1}} s_1(t) \tag{1.35}$$

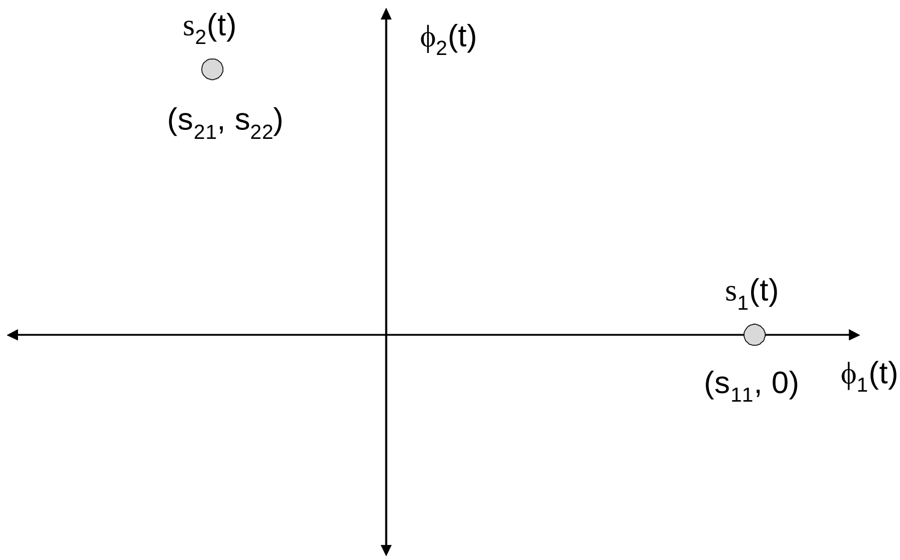

Figure 1.2: Distances for Constellation Diagram for Arbitrary Binary Modulation Scheme

where E_1 is the energy in symbol 1. Once the basis functions are defined, a signal space plot (sometimes termed a constellation diagram) can be drawn which represents the symbols in k-dimensional space (k is the number of basis function which is two in this example). An example constellation diagram is plotted in Figure 1.2 for arbitrary $s_1(t)$ and $s_2(t)$.

Assume that the received signal is $r(t) = s(t) + n(t)$ where $n(t)$ is additive white Gaussian noise with power spectral density $P_N(f) = \frac{N_o}{2}$. We wish to examine the receiver which minimizes the probability of symbol error.

The optimum (maximum SNR) receiver is one which correlates the received signal with the two basis functions (or a matched filter receiver). The outputs of the two correlators can be represented in vector form as

$$\mathbf{z} = \mathbf{s}_i + \mathbf{n} \tag{1.36}$$

where the vectors are of length 2 corresponding to the outputs of the two correlators. Now the maximum likelihood receiver is the one which maximizes the *a posteriori* probability. That is

$$\hat{\mathbf{s}} = \max_{\mathbf{s}} P\left(\mathbf{s}\,|\mathbf{z}\right) \tag{1.37}$$

In general, the *a posteriori* probability is difficult to calculate. However, using Bayes Theorem

$$P\left(\mathbf{s}\,|\mathbf{z}\right) = \frac{P\left(\mathbf{z}\,|\mathbf{s}\right) P\left(\mathbf{s}\right)}{P\left(\mathbf{z}\right)} \tag{1.38}$$

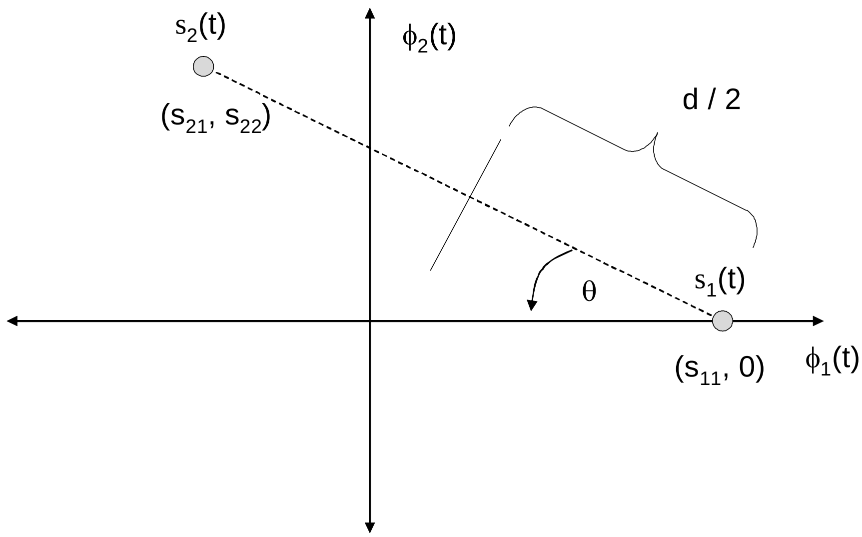

Figure 1.3: Distance Properties for Arbitrary Binary Modulation Scheme

Now, since $P(\mathbf{z})$ is constant regardless of the choice of $\mathbf{s}$, we can write

$$P(\mathbf{s}|\mathbf{z}) = P(\mathbf{z}|\mathbf{s})\,P(\mathbf{s}) \tag{1.39}$$

Further, if the symbols are all equally likely:

$$\hat{\mathbf{s}} = \max_{\mathbf{s}} P(\mathbf{z}|\mathbf{s}) \tag{1.40}$$

Now, since the noise $\mathbf{n}$ is a vector of uncorrelated Gaussian noise samples we can write

$$P(\mathbf{z}|\mathbf{s}_i) = \frac{1}{2\pi\sigma^2} e^{-\frac{(z_1 - s_{i1})^2}{2\sigma^2}} e^{-\frac{(z_2 - s_{i2})^2}{2\sigma^2}} \tag{1.41}$$

where z_1 and z_2 are the two correlator outputs and σ is the standard deviation of the Gaussian noise. Further, from the above equation we can see that maximizing $P(\mathbf{z}|\mathbf{s}_i)$ is equivalent to minimizing $(z_1 - s_{i1})^2 + (z_2 - s_{i2})^2$ which is equivalent to choosing the symbol which is closest to the received vector $\mathbf{z}$.

Now an error will occur if the noise vector projected onto the line connecting the two symbols is greater than one half the distance between the two symbols (see Figure 1.3). The noise projected onto the line connecting the two symbols is

$$\nu = n_1 \cos\theta + n_2 \sin\theta \tag{1.42}$$

where θ is the angle between the line connecting the two symbols and the x-axis as is shown in Figure 1.3. Further, the distance between the two symbols is

$$d = \sqrt{(s_{11} - s_{21})^2 + (s_{22})^2} \tag{1.43}$$

Now, n_1 and n_2 are zero mean Gaussian random variables (since they are the outputs of linear filters with Gaussian inputs). The variance of the noise terms is

$$\begin{aligned}
\mathrm{E}\left\{n_1^2\right\} &= \mathrm{E}\left\{n_2^2\right\} = \mathrm{E}\left\{\int_0^T \int_0^T n(t)\phi_1(t)n(\tau)\phi_1(\tau)dtd\tau\right\} \\
&= \int_0^T \mathrm{E}\left\{n^2(t)\right\}\phi_1^2(t)\,dt \\
&= \frac{N_o}{2} \qquad (1.44)
\end{aligned}$$

where we have used the fact that the basis functions have unit energy. Thus, we can easily show that

$$\mathrm{E}\left\{\nu^2\right\} = \frac{N_o}{2}\frac{1}{2} + \frac{N_o}{2}\frac{1}{2} = \frac{N_o}{2} \qquad (1.45)$$

Thus, the probability of error can be found as

$$\begin{aligned}
P_e &= P\left\{\nu > \frac{d}{2}\right\} \\
&= \int_{d/2}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{N_o/2}} e^{-x^2/N_o} dx \qquad (1.46)
\end{aligned}$$

Making the substitution $y = \sqrt{2/N_o}x$

$$\begin{aligned}
P_e &= \int_{\sqrt{2/N_o}d/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&= Q\left(\sqrt{\frac{d^2}{2N_o}}\right) \qquad (1.47)
\end{aligned}$$

where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$ is the standard Q-function. Now, from Figure 1.3 we have

$$\begin{aligned}
d^2 &= (s_{11} - s_{21})^2 + (s_{22})^2 \\
&= s_{11}^2 - 2s_{11}s_{21} + s_{21}^2 + s_{22}^2 \\
&= E_1 + E_2 - 2s_{11}s_{21} \\
&= E_1 + E_2 - 2\sqrt{E_1}\int_0^T s_2(t)\phi_1(t)dt \\
&= E_1 + E_2 - 2\int_0^T s_2(t)s_1(t)dt \\
&= E_1 + E_2 - 2\rho_{12} \qquad (1.48)
\end{aligned}$$

where ρ_{12} is the correlation between $s_1(t)$ and $s_2(t)$. Substituting we have

$$\begin{aligned} P_e &= Q\left(\sqrt{\frac{E_1 + E_2 - 2\rho_{12}}{2N_o}}\right) \\ &= Q\left(\sqrt{\frac{E_b - \rho_{12}}{N_o}}\right) \end{aligned} \tag{1.49}$$

where E_b is the average energy per bit (which is the same as the average energy per symbol). Now, let's examine a few specific examples of binary modulation. If the modulation scheme is BPSK, there is a single basis function and $s_1(t) = -s_2(t)$. Thus, $E_1 = E_2 = E_b$ and $\rho_{12} = -E_b$. Thus,

$$P_e^{BPSK} = Q\left(\sqrt{\frac{2E_b}{N_o}}\right) \tag{1.50}$$

If the modulation scheme is BFSK, the two symbols are orthogonal, $E_1 = E_2 = E_b$ and thus $\rho_{12} = 0$ and

$$P_e^{BFSK} = Q\left(\sqrt{\frac{E_b}{N_o}}\right) \tag{1.51}$$

If the modulation scheme is BASK, $\rho_{12} = 0$ and

$$P_e^{BASK} = Q\left(\sqrt{\frac{E_b}{N_o}}\right) \tag{1.52}$$

From equation (1.49) we can see that the minimum probability of error occurs when the signals are antipodal ($\rho_{12} = -E_b$). Thus, BPSK provides the minimum probability of error for binary modulation.

1.6.4 Non-coherent Demodulation

The performance of the modulation schemes considered so far assume that a coherent reference is available at the receiver. While all of the modulation schemes considered can be demodulated coherently, it is very often more practical to build non-coherent receivers. FSK and ASK easily allow envelope detection which doesn't require a coherent reference. Further, PSK allows for differential encoding and detection. In Differential PSK (DPSK) the change in phase represents the data. A receiver does not need a coherent reference since it must only compare the current signal phase with the previous signal phase. The benefit of such schemes is that they allow for much simpler (and thus inexpensive) receivers. The downside is that they provide inferior performance. It can be shown that the performance of binary DPSK can be approximated by [3]

$$P_b \approx \frac{1}{2} e^{-\frac{E_b}{N_o}} \tag{1.53}$$

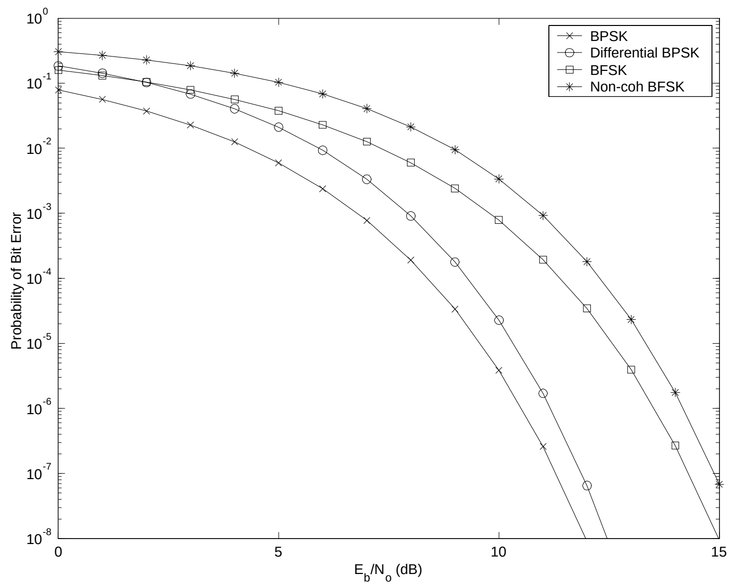

Figure 1.4: Comparison of the Performance of Coherent and Non-coherent (or differentially coherent) Reception

Similarly, the performance of non-coherent BASK and BFSK can be approximated by [3]

$$P_b \approx \frac{1}{2} e^{-\frac{1}{2}\frac{E_b}{N_o}} \tag{1.54}$$

A comparison of coherent BPSK and BFSK with non-coherent BFSK and binary DPSK is plotted in Figure 1.4. As we can see, non-coherent demodulation costs the system approximately 1dB in performance at high SNR with larger losses experienced at lower E_b/N_o values.

1.6.5 M-ary Modulation

The probability of error for M-ary modulation schemes can be found in a similar manner by using the Union Bound [3]. The Union Bound says that the probability of the union of several events is less than or equal to the sum of the individual event probabilities. We can use this in determining the probability of symbol error for M-ary modulation symbols by bounding the error probability by the sum of the individual pair-wise error probabilities:

$$\Pr\left(\hat{s} \neq s_i \middle| s = s_i\right) \leq \sum_{j \neq i} \Pr\left(\hat{s} = s_j \middle| s = s_i\right) \tag{1.55}$$

Using this approximation we can bound the probability of symbol error for MPSK using the error probability given above for binary modulation. Specifically, by realizing that each symbol has two adjacent neighbors and noting that the distance between symbols is $d = 2\sqrt{E_s}\sin(\pi/M)$, we can write the probability of symbol error using the Union Bound as

$$P_s \leq 2\mathrm{Q}\left(\sqrt{\frac{2E_b}{N_o}\log_2 M}\sin\frac{\pi}{M}\right) \tag{1.56}$$

Similarly, the probability of error for MFSK can be found as

$$P_s \leq (M-1)\mathrm{Q}\left(\sqrt{\frac{E_b}{N_o}\log_2 M}\right) \tag{1.57}$$

1.7 Bandwidth Efficiency and Energy Efficiency

The two main characteristics of modulation schemes that need to be considered by the communications engineer are bandwidth efficiency and energy efficiency. Bandwidth efficiency can be defined as the bit rate per bandwidth or bits/sec/Hz. First let us consider two-dimensional modulation schemes such as PSK, ASK, and QAM. In these modulation schemes, there are only two dimensions (i.e., two basis functions). If square pulses are used, the null-to-null bandwidth is

$$B = 2R_s \tag{1.58}$$

where R_s is the symbol rate and is related to the bit rate by $R_s = \frac{R_b}{\log_2 M}$. Thus, the bandwidth efficiency in terms of null-to-null bandwidth can be written as

$$\begin{aligned}
\eta_{bw} &= \frac{R_b}{B} \\
&= \frac{R_b}{2R_s} \\
&= \frac{\log_2(M)\,R_b}{2R_b} \\
&= \frac{\log_2(M)}{2}
\end{aligned} \tag{1.59}$$

Thus, we have

$$\eta_{bw} = \tfrac{\log_2(M)}{2} \quad M\mathrm{PSK}, M\mathrm{QAM}, M\mathrm{ASK} \tag{1.60}$$

If optimal pulse shaping is used, i.e., sinc pulses, then bandwidth efficiency is (in terms of absolute bandwidth) $\eta_{bw} = \log_2(M)$. Thus, for schemes with a fixed number of dimensions, bandwidth efficiency increases with M. Conversely, for orthogonal modulation schemes (e.g., MFSK) bandwidth efficiency decreases with M. This is due to the fact that orthogonal modulation schemes require an additional dimension for each additional symbol. The null-to-null bandwidth in this case depends on whether or not coherent demodulation is used. In the best case (coherent demodulation) the minimum frequency separation is $\Delta f = \frac{1}{2T_s}$ where T_s is the symbol duration. Since there are $M-1$ intervals and R_s Hz on either end (assuming square pulses), the null-to-null bandwidth is approximately

$$B = \frac{(M-1)\,R_s}{2} + R_s \tag{1.61}$$

Since $R_s = \frac{R_b}{\log_2 M}$,

$$\eta_{bw} = \tfrac{2\log_2(M)}{M+1} \quad M\text{FSK} \tag{1.62}$$

Thus, while for PSK, QAM and ASK bandwidth efficiency increases with M, the opposite is true for orthogonal (or bi-orthognal) signals.

Energy efficiency is typically defined as the value of $\frac{E_b}{N_o}$ required to obtain a specified probability of bit error. The probability of symbol error for MPSK and MFSK are given in equations (1.56) and (1.57) respectively. For MPSK the probability of bit error can be found by assuming the use of Gray coding. Gray coding maps bits to symbols such that adjacent symbols differ by one bit only. Since the symbol error probability is dominated by the nearest neighbors, the probability of bit error, P_b, can be approximated as

$$P_b \approx \frac{P_s}{M} \tag{1.63}$$

since one of M bits will be in error for each symbol error. Since in FSK all symbols are nearest neighbors, all bit error patterns are equally likely and thus $P_b = P_s \frac{M}{2(M-1)} \approx \frac{P_s}{2}$. Using these relationships and the probability of symbol error expressions from equations (1.56) and (1.57), we can see that for PSK (and for ASK and QAM) the probability of symbol error increases with M. This is due to the fact that the symbols move closer together as we increase M since the number of dimensions in the symbol space is fixed. FSK on the other hand behaves in an opposite manner. Energy efficiency improves with M which may seem counter-intuitive at first glance. As M increases, the distances between symbols remains constant in terms of the energy per symbol, although the number of nearest neighbors increases. This increases the probability of symbol error directly with M. However, since the number bits per symbol increases, the distance between symbols in terms of E_b increases with $\log_2 M$ and thus the argument inside the Q-function increases with $\log_2 M$. The net result is that probability of bit error improves with M, thus improving energy efficiency. A comparison of MPSK and MFSK is plotted in Figure 1.5. An overall comparison of MPSK, MFSK and QAM is provided in Table 1.1 for various values of M. Note that (a) QAM provides better energy efficiency than PSK for the same bandwidth efficiency, and (b) the performance of FSK assumes coherent reception and is also applicable to any M-ary orthogonal modulation scheme.

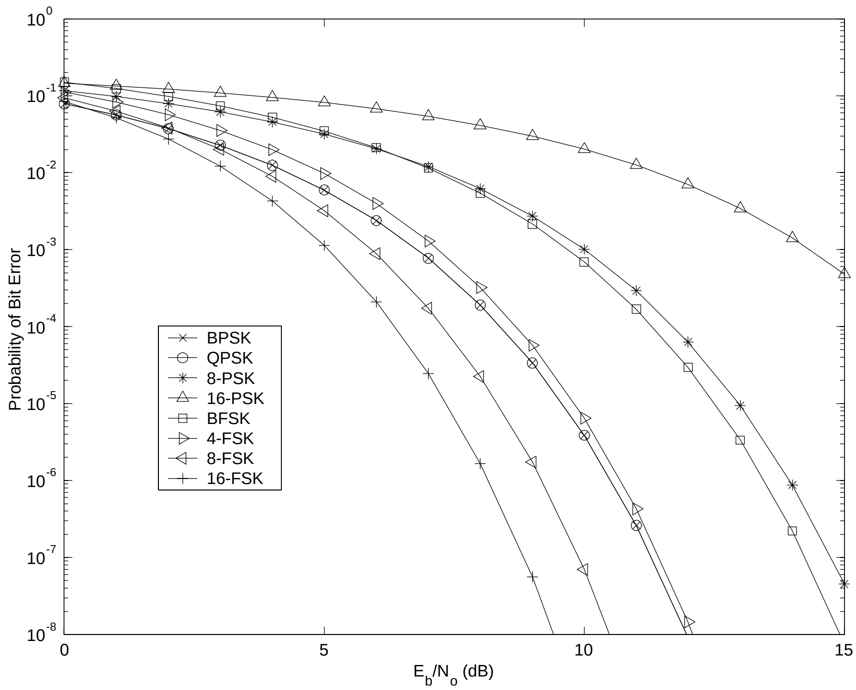

Figure 1.5: Performance of MPSK and MFSK in AWGN with Coherent Reception

Table 1.1: Comparison of Modulation Schemes

Modulation Scheme	Energy Efficiency (E_b/N_o for $P_b = 10^{-5}$)	BW Efficiency (b/s/Hz)
BPSK	9.5dB	$\frac{1}{2}$
QPSK	13.5dB	1
8-PSK	18dB	$\frac{3}{2}$
16-PSK	23dB	2
BFSK	12.5dB	$\frac{2}{5}$
4-FSK	10dB	$\frac{4}{7}$
8-FSK	8dB	$\frac{6}{11}$
16-FSK	7dB	$\frac{8}{19}$
4-QAM	9.5dB	2
16-QAM	13dB	$\frac{5}{2}$
64-QAM	17.5dB	3

1.8 Conclusion

In this chapter we have provided a brief overview of digital communications. Specifically, we have discussed sampling, quantization, modulation, bandwidth efficiency and energy efficiency. Throughout this book we will assume that the reader has a basic understanding of digital communications. For those readers who feel that they need additional review, please refer to the following excellent text books on Digital Communications [3, 4, 5, 7, 8].

Bibliography

[1] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, NY, third edition, 1991.

[2] P.Z. Peebles. *Probability, Random Variables, and Random Signal Principles.* McGraw-Hill, New York, NY, second edition, 1987.

[3] J.G. Proakis. *Digital Communications.* McGraw-Hill, New York, NY, fourth edition, 2001.

[4] M.B. Pursley. *Introduction to Digital Communications.* Prentice Hall, Upper Saddle River, NJ, 2005.

[5] T. S. Rappaport. *Wireless Communications: Principles and Practice.* Prentice Hall, Upper Saddle River, NJ, second edition, 2002.

[6] K.S. Shanmugan and A.M. Breipohl. *Random Signals: Detection, Estimation, and Data Analysis.* John Wiley and Sons, New York, NY, second edition, 1988.

[7] B. Sklar. *Digital Communications.* Prentice Hall, Upper Saddle River, NJ, 2000.

[8] R.E. Ziemer and R.L. Peterson. *Introduction to Digital Communications.* Prentice Hall, Upper Saddle River, NJ, second edition, 2001.