Binary with N Zero Substitution (BNZS) Signaling

A class of line codes similar to HDBN is the **binary with N zero substitution, or BNZS** code, where if N zeros occur in succession, they are replaced, by one of the two special sequences containing some **1**s to increase timing content. There are deliberate bipolar violations just as in HDBN. Binary with eight-zero substitution (B8ZS) is used in DS1 signals of the digital telephone hierarchy in Chapter 6. It replaces any string of eight zeros in length with a sequence of ones and zeros containing two bipolar violations. Such a sequence is unlikely to be counterfeited by errors, and any such sequence received by a digital channel bank is replaced by a string of eight logic zeros prior to decoding. The sequence used as a replacement consists of the pattern **000VB0VB**. Similarly, in **B6ZS** code used in DS2 signals, a string of six zeros is replaced with **0VB0VB**, and DS3 signal features a three-zero B3ZS code. The B3ZS code is slightly more complex than the others in that either **B0V** or **00V** is used, the choice being made so that the number of **B** pulses between consecutive **V** pulses is odd. These BNZS codes with $N = 3, 6$, or 8 involve bipolar violations and must therefore be carefully replaced by their equivalent zero strings at the receiver.

There are many other transmission (line) codes, too numerous to list here. A list of codes and appropriate references can be found in Bylanski and Ingram.[3]

7.3 PULSE SHAPING

The PSD $S_y(f)$ of a digital signal $y(t)$ can be controlled by a choice of line code or by $P(f)$, the pulse shape. In the last section we discussed how the PSD is controlled by a line code. In this section we examine how $S_y(f)$ is influenced by the pulse shape $p(t)$, and we learn how to shape a pulse $p(t)$ to achieve a desired $S_y(f)$. The PSD $S_y(f)$ is strongly and directly influenced by the pulse shape $p(t)$ because $S_y(f)$ contains the term $|P(f)|^2$. Thus, in comparison to the nature of the line code, the pulse shape is a more direct and potent factor in terms of shaping the PSD $S_y(f)$.

7.3.1 Intersymbol Interferences (ISI) and Effect

In the last section, we used a simple half-width rectangular pulse $p(t)$ for the sake of illustration. Strictly speaking, in this case the bandwidth of $S_y(f)$ is infinite, since $P(f)$ has infinite bandwidth. But we found that the essential bandwidth of $S_y(f)$ was finite. For example, most of the power of a bipolar signal is contained within the essential band 0 to R_b Hz. Note, however, that the PSD is small but is still nonzero in the range $f > R_b$ Hz. Therefore, when such a signal is transmitted over a channel of bandwidth R_b Hz, a significant portion of its spectrum is transmitted, but a small portion of the spectrum is suppressed. In Sec. 3.5 and Sec. 3.6, we saw how such a spectral distortion tends to spread the pulse (dispersion). Spreading of a pulse beyond its allotted time interval T_b will cause it to interfere with neighboring pulses. This is known as **intersymbol interference** or **ISI**.

ISI is *not* noise. ISI is caused by nonideal channels that are not distortionless over the entire signal bandwidth. In the case of half-width rectangular pulse, the signal bandwidth is, strictly speaking, infinity. ISI, as a manifestation of channel distortion, can cause errors in pulse detection if it is large enough.

To resolve the difficulty of ISI, let us review briefly our problem. We need to transmit a pulse every T_b interval, the kth pulse being $a_k p(t-kT_b)$. The channel has a finite bandwidth, and we are required to detect the pulse amplitude a_k correctly (i.e., without ISI). In our discussion so far, we have considered time-limited pulses. Since such pulses cannot be band-limited, part of their spectra is suppressed by a band-limited channel. This causes pulse distortion (spreading out) and, consequently, ISI. We can try to resolve this difficulty by using pulses that are band-limited to begin with so that they can be transmitted intact over a band-limited channel. But band-limited pulses cannot be time-limited. Obviously, various pulses will overlap and cause ISI. Thus, whether we begin with time-limited pulses or band-limited pulses, it appears that ISI cannot be avoided. It is inherent in the finite transmission bandwidth. Fortunately, there is an escape from this blind alley. Pulse amplitudes can be detected correctly despite pulse spreading (or overlapping), if there is no ISI at the decision-making instants. This can be accomplished by a properly shaped band-limited pulse. To eliminate ISI, Nyquist proposed three different criteria for pulse shaping,[4] where the pulses are allowed to overlap. Yet, they are shaped to cause zero (or controlled) interference with all the other pulses at the decision-making instants. Thus, by limiting the noninterference requirement only at the decision-making instants, we eliminate the need for the pulse to be totally nonoverlapping. We shall consider only the first two criteria. The third is much less useful than the first two criteria,[5] and hence, will not be considered here.

7.3.2 Nyquist's First Criterion for Zero ISI

In the first method, Nyquist achieves zero ISI by choosing a pulse shape that has a nonzero amplitude at its center (say $t = 0$) and zero amplitudes at $t = \pm nT_b$ ($n = 1, 2, 3, \ldots$), where T_b is the separation between successive transmitted pulses (Fig. 7.11a). Thus,

$$p(t) = \begin{cases} 1 & t = 0 \\ 0 & t = \pm nT_b \end{cases} \qquad \left(T_b = \frac{1}{R_b}\right) \tag{7.23}$$

A pulse satisfying this criterion causes zero ISI at all the remaining pulse centers, or signaling instants as shown in Fig. 7.11a, where we show several successive pulses (dashed) centered at $t = 0, T_b, 2T_b, 3T_b, \ldots$ ($T_b = 1/R_b$). For the sake of convenience, we have shown all pulses

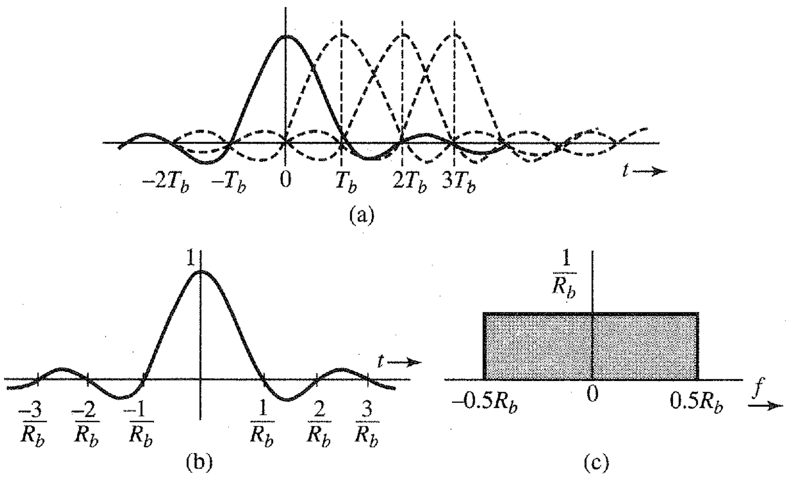

Figure 7.11 The minimum bandwidth pulse that satisfies Nyquist's first criterion and its spectrum.

to be positive.* It is clear from this figure that the samples at $t = 0, T_b, 2T_b, 3T_b, \ldots$ consist of the amplitude of only one pulse (centered at the sampling instant) with no interference from the remaining pulses.

Now transmission of R_b bit/s requires a theoretical minimum bandwidth $R_b/2$ Hz. It would be nice if a pulse satisfying Nyquist's criterion had this minimum bandwidth $R_b/2$ Hz. Can we find such a pulse $p(t)$? We have already solved this problem (Example 6.1 with $B = R_b/2$), where we showed that there exists one (and only one) pulse which meets Nyquist's criterion (7.23) and has a bandwidth $R_b/2$ Hz. This pulse, $p(t) = \text{sinc}\,(\pi R_b t)$, (Fig. 7.11b) has the property

$$\text{sinc}\,(\pi R_b t) = \begin{cases} 1 & t = 0 \\ 0 & t = \pm n T_b \end{cases} \qquad \left(T_b = \frac{1}{R_b}\right) \tag{7.24a}$$

Moreover, the Fourier transform of this pulse is

$$P(f) = \frac{1}{R_b}\Pi\left(\frac{f}{R_b}\right) \tag{7.24b}$$

which has a bandwidth $R_b/2$ Hz as seen from Fig. 7.11c. We can use this pulse to transmit at a rate of R_b pulses per second without ISI, over a bandwidth only $R_b/2$.

This scheme shows that we can attain the theoretical limit of performance by using a sinc pulse. Unfortunately, this pulse is impractical because it starts at $-\infty$. We will have to wait an infinite time to generate it. Any attempt to truncate it would increase its bandwidth beyond $R_b/2$ Hz. But even if this pulse were realizable, it would have an undesirable feature: namely, it decays too slowly at a rate $1/t$. This causes some serious practical problems. For instance, if the nominal data rate of R_b bit/s required for this scheme deviates a little, the pulse amplitudes will not vanish at the other pulse centers. Because the pulses decay only as $1/t$, the cumulative interference at any pulse center from all the remaining pulses is of the form $\sum(1/n)$. It is well known that the infinite series of this form does not converge and can add up to a very large value. A similar result occurs if everything is perfect at the transmitter but the sampling rate at the receiver deviates from the rate of R_b Hz. Again, the same thing happens if the sampling instants deviate a little because of pulse time jitter, which is inevitable even in the most sophisticated systems. This scheme therefore fails unless everything is perfect, which is a practical impossibility. And all this is because sinc $(\pi R_b t)$ decays too slowly (as $1/t$). The solution is to find a pulse $p(t)$ that satisfies Eq. (7.23) but decays faster than $1/t$. Nyquist has shown that such a pulse requires a bandwidth $kR_b/2$, with $1 \leq k \leq 2$.

This can be proved as follows. Let $p(t) \Longleftrightarrow P(f)$, where the bandwidth of $P(f)$ is in the range $(R_b/2,\ R_b)$ (Fig. 7.12a). The desired pulse $p(t)$ satisfies Eq. (7.23). If we sample $p(t)$ every T_b seconds by multiplying $p(t)$ by $\delta_{T_b}(t)$, (an impulse train), then because of the property (7.23), all the samples, except the one at the origin, are zero. Thus, the sampled signal $\overline{p}(t)$ is

$$\overline{p}(t) = p(t)\delta_{T_b}(t) = \delta(t) \tag{7.25}$$

Following the analysis of Eq. (6.4) in Chapter 6, we know that the spectrum of a sampled signal $\overline{p}(t)$ is ($1/T_b$ times) the spectrum of $p(t)$ repeating periodically at intervals of the sampling

* Actually, a pulse corresponding to **0** would be negative. But considering all positive pulses does not affect our reasoning. Showing negative pulses would make the figure needlessly confusing.

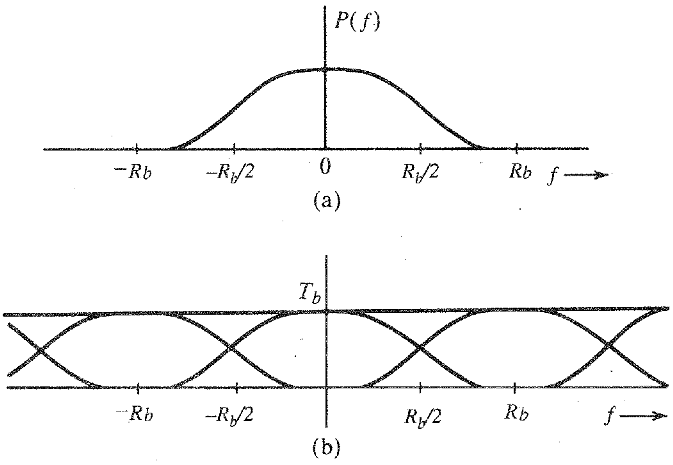

Figure 7.12 Derivation of the zero ISI Nyquist criterion pulse.

frequency R_b. Therefore, the Fourier transform of both sides of Eq. (7.25) yields

$$\frac{1}{T_b}\sum_{n=-\infty}^{\infty} P(f - nR_b) = 1 \qquad \text{where} \qquad R_b = \frac{1}{T_b} \tag{7.26}$$

or

$$\sum_{n=-\infty}^{\infty} P(f - nR_b) = T_b \tag{7.27}$$

Thus, the sum of the spectra formed by repeating $P(f)$ spaced R_b apart is a constant T_b, as shown in Fig. 7.12b.*

Consider the spectrum in Fig. 7.12b over the range $0 < f < R_b$. Over this range only two terms $P(f)$ and $P(f - R_b)$ in the summation in Eq. (7.27) are involved. Hence

$$P(f) + P(f - R_b) = T_b \qquad 0 < f < R_b$$

Letting $x = f - R_b/2$, we have

$$P\,(x + 0.5R_b) + P\,(x - 0.5R_b) = T_b \qquad |x| < 0.5R_b \tag{7.28a}$$

or, alternatively,

$$P\left(x + \frac{R_b}{2}\right) + P\left(x - \frac{R_b}{2}\right) = T_b \qquad |x| < 0.5R_b \tag{7.28b}$$

Use of the conjugate symmetry property [Eq. (3.11)] on Eq. (7.28) yields

$$P\left(\frac{R_b}{2} + x\right) + P^*\left(\frac{R_b}{2} - x\right) = T_b \qquad |x| < 0.5R_b \tag{7.29}$$

* Observe that if $R_b > 2B$, where B is the bandwidth (in hertz) of $P(f)$, the repetitions of $P(f)$ are nonoverlapping, and condition (7.27) cannot be satisfied. For $R_b = 2B$, the condition is satisfied only for the ideal low-pass $P(f)[p(t) = \text{sinc}\,(\pi R_b t)]$, which is not realizable. Hence, we must have $B > R_b/2$.

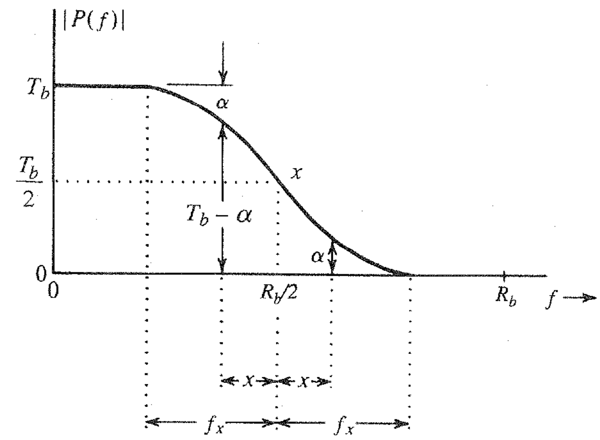

Figure 7.13 Vestigial (raised-cosine) spectrum.

If we choose $P(f)$ to be real-valued and positive then only $|P(f)|$ needs to satisfy Eq. (7.29). Because $|P(f)|$ is real, Eq. (7.29) implies

$$\left|P\left(\frac{R_b}{2}+x\right)\right| + \left|P\left(\frac{R_b}{2}-x\right)\right| = T_b \qquad |x| < 0.5R_b \tag{7.30}$$

Hence, $|P(f)|$ should be of the form shown in Fig. 7.13. This curve has an odd symmetry about the set of axes intersecting at point α [the point on $|P(f)|$ curve at $f = R_b/2$]. Note that this requires that

$$|P(0.5R_b)| = 0.5|P(0)|$$

The bandwidth, in hertz, of $P(f)$ is $0.5R_b + f_x$, where f_x is the bandwidth in excess of the minimum bandwidth $R_b/2$. Let r be the ratio of the excess bandwidth f_x to the theoretical minimum bandwidth $R_b/2$:

$$\begin{aligned} r &= \frac{\text{excess bandwidth}}{\text{theoretical minimum bandwidth}} \\ &= \frac{f_x}{0.5R_b} \\ &= 2f_xT_b \end{aligned} \tag{7.31}$$

Observe that because f_x cannot be larger than $R_b/2$,

$$0 \leq r \leq 1 \tag{7.32}$$

In terms of frequency f, the theoretical minimum bandwidth is $R_b/2$ Hz, and the excess bandwidth is $f_x = rR_b/2$ Hz. Therefore, the bandwidth of $P(f)$ is

$$B_T = \frac{R_b}{2} + \frac{rR_b}{2} = \frac{(1+r)R_b}{2} \tag{7.33}$$

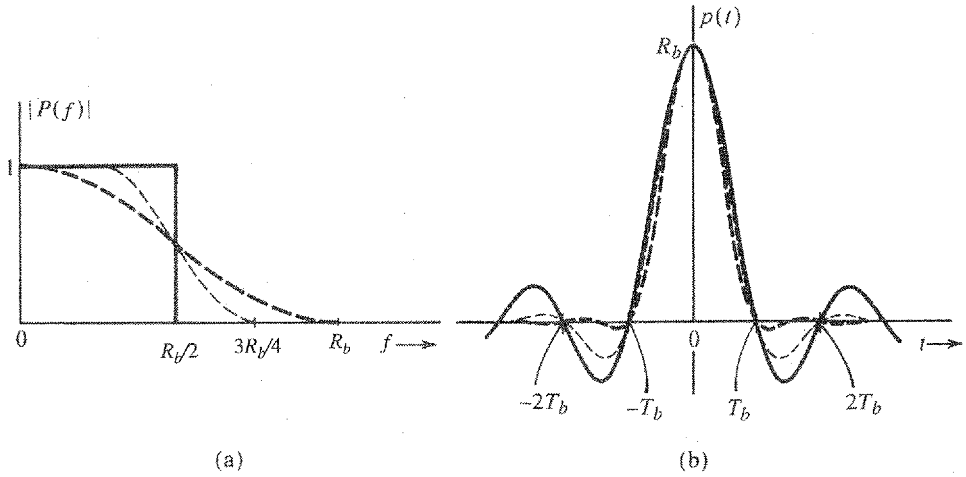

Figure 7.14 Pulses satisfying Nyquist's first criterion: solid curve, ideal $f_x = 0$ ($r = 0$); light dashed curve, $f_x = R_b/4$ ($r = 0.5$); heavy dashed curve, $f_x = R_b/2$ ($r = 1$).

The constant r is called the **roll-off factor** and is also expressed in terms of percent. For example, if $P(f)$ is a Nyquist first criterion spectrum with a bandwidth that is 50% higher than the theoretical minimum, its roll-off factor $r = 0.5$ or 50%.

A filter having an amplitude response with the same characteristics is required in the vestigial sideband modulation discussed in Sec. 4.5 [Eq. (4.26)]. For this reason, we shall refer to the spectrum $P(f)$ in Eqs. (7.29) and (7.30) as a **vestigial spectrum**. The pulse $p(t)$ in Eq. (7.23) has zero ISI at the centers of all other pulses transmitted at a rate of R_b pulses per second. A pulse $p(t)$ that causes zero ISI at the centers of all the remaining pulses (or signaling instants) is the Nyquist first criterion pulse. We have shown that a pulse with a vestigial spectrum [Eq. (7.29) or Eq. (7.30)] satisfies the Nyquist's first criterion for zero ISI.

Because $0 \leq r < 1$, the bandwidth of $P(f)$ is restricted to the range $R_b/2$ to R_b Hz. The pulse $p(t)$ can be generated as a unit impulse response of a filter with transfer function $P(f)$. But because $P(f) = 0$ over a frequency band, it violates the Paley-Wiener criterion and is therefore unrealizable. However, the vestigial roll-off characteristic is gradual, and it can be more closely approximated by a practical filter. One family of spectra that satisfies Nyquist's first criterion is

$$P(f) = \begin{cases} 1, & |f| < \dfrac{R_b}{2} - f_x \\ \dfrac{1}{2}\left[1 - \sin \pi \left(\dfrac{f - R_b/2}{2f_x}\right)\right] & \left|f - \dfrac{R_b}{2}\right| < f_x \\ 0, & |f| > \dfrac{R_b}{2} + f_x \end{cases} \tag{7.34}$$

Figure 7.14a shows three curves from this family, corresponding to $f_x = 0$ ($r = 0$), $f_x = R_b/4$ ($r = 0.5$) and $f_x = R_b/2$ ($r = 1$). The respective impulse responses are shown in Fig. 7.14b. It can be seen that increasing f_x (or r) improves $p(t)$; that is, more gradual cutoff reduces the oscillatory nature of $p(t)$ and causes it to decay more rapidly in time domain. For the case of the maximum value of $f_x = R_b/2$ ($r = 1$), Eq. (7.34) reduces to

$$P(f) = \frac{1}{2}(1 + \cos \pi f T_b)\, \Pi\left(\frac{f}{2R_b}\right) \tag{7.35a}$$

$$= \cos^2\left(\frac{\pi f T_b}{2}\right) \Pi\left(\frac{f T_b}{2}\right) \tag{7.35b}$$

This characteristic of Eq. (7.34) is known in the literature as the **raised-cosine** characteristic, because it represents a cosine raised by its peak amplitude. Eq. (7.35) is also known as the **full-cosine roll-off** characteristic. The inverse Fourier transform of this spectrum is readily found as (see Prob 7.3-8)

$$p(t) = R_b \frac{\cos \pi R_b t}{1 - 4R_b^2 t^2} \operatorname{sinc}(\pi R_b t) \tag{7.36}$$

This pulse is shown in Fig. 7.14b ($r = 1$). We can make several important observations about the raised-cosine pulse. First, the bandwidth of this pulse is R_b Hz and has a value R_b at $t = 0$ and is zero not only at all the remaining signaling instants but also at points midway between all the signaling instants. Second, it decays rapidly, as $1/t^3$. As a result, the raised-cosine pulse is relatively insensitive to deviations of R_b, sampling rate, timing jitter, and so on. Furthermore, the pulse-generating filter with transfer function $P(f)$ [Eq. (7.35b)] is closely realizable. The phase characteristic that goes along with this filter is very nearly linear, so that no additional phase equalization is needed.

It should be remembered that it is the pulses received at the detector input that should have the form for zero ISI. In practice, because the channel is not ideal (distortionless), the transmitted pulses should be shaped so that after passing through the channel with transfer function $H_c(f)$, they will be received with the proper shape (such as raised-cosine pulses) at the receiver. Hence, the transmitted pulse $p_i(t)$ should satisfy

$$P_i(f)H_c(f) = P(f)$$

where $P(f)$ has the vestigial spectrum in Eq. (7.30). For convenience, the transfer function $H_c(f)$ as a channel may also include a receiver filter designed to reject interference and other out-of-band noises.

Example 7.1 Determine the pulse transmission rate in terms of the transmission bandwidth B_T and the roll-off factor r. Assume a scheme using Nyquist's first criterion.

From Eq. (7.33)

$$R_b = \frac{2}{1+r} B_T$$

Because $0 \leq r \leq 1$, the pulse transmission rate varies from $2B_T$ to B_T, depending on the choice of r. A smaller r gives a higher signaling rate. But the pulse $p(t)$ decays slowly, creating the same problems as those discussed for the sinc pulse. For the raised-cosine pulse $r = 1$ and $R_b = B_T$, we achieve half the theoretical maximum rate. But the pulse decays faster as $1/t^3$ and is less vulnerable to ISI.

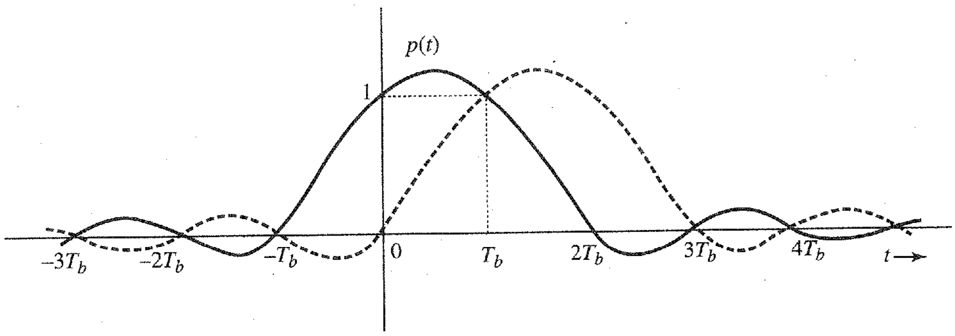

Figure 7.15 Communication using controlled ISI or Nyquist second criterion pulses.

7.3.3 Controlled ISI or Partial Response Signaling

The Nyquist criterion pulse requires in a bandwidth somewhat larger than the theoretical minimum. If we wish to further reduce the pulse bandwidth, we must find a way to widen the pulse $p(t)$ (the wider the pulse, the narrower the bandwidth). Widening the pulse may result in interference (ISI) with the neighboring pulses. However, in the binary transmission with just two possible symbols, a known and controlled amount of ISI may be possible to remove or compensate because there are only a few possible interference patterns.

Consider a pulse specified by (see Fig. 7.15):

$$p(nT_b) = \begin{cases} 1 & n = 0, 1 \\ 0 & \text{for all other } n \end{cases} \tag{7.37}$$

This leads to a known and controlled ISI from the kth pulse to the very next transmitted pulse. We use polar signaling by means of this pulse. Thus, **1** is transmitted by $p(t)$ and **0** is transmitted by using the pulse $-p(t)$. The received signal is sampled at $t = nT_b$, and the pulse $p(t)$ has zero value at all n except for $n = 0$ and 1, where its value is 1 (Fig. 7.15). Clearly, such a pulse causes zero ISI with all the pulses except the succeeding pulse. Therefore, we need to worry about the ISI with the succeeding pulse only. Consider two such successive pulses located at 0 and T_b, respectively. If both pulses were positive, the sample value of the resulting signal at $t = T_b$ would be 2. If the both pulses were negative, the sample value would be -2. But if the two pulses were of opposite polarity, the sample value would be 0. With only these three possible values, the signal sample clearly allows us to make correct decision at the sampling instants. The decision rule is as follows. If the sample value is positive, the present bit is **1** and the previous bit is also **1**. If the sample value is negative, the present bit is **0** and the previous bit is also **0**. If the sample value is zero, the present bit is the opposite of the previous bit. Knowledge of the previous bit then allows the determination of the present bit.

Table 7.1 shows a transmitted bit sequence, the sample values of the received signal $x(t)$ (assuming no errors causes by channel noise), and the detector decision. This example also indicates the error detecting property of this scheme. Examination of samples of the waveform $y(t)$ in Table 7.1 shows that there are always an even number of zero-valued samples between two full-valued samples of the same polarity and an odd number of zero-valued samples between two full-valued samples of opposite polarity. Thus, the first sample value of $x(t)$ is 2, and the next full-valued sample (the fourth sample) is 2. Between these full-valued samples

TABLE 7.1
Transmitted Bits and the Received Samples in Controlled ISI Signaling

Information sequence	1	1	0	1	1	0	0	0	1	0	1	1	1
Samples $y(kT_b)$	1	2	0	0	2	0	−2	−2	0	0	0	2	2
Detected sequence	1	1	0	1	1	0	0	0	1	0	1	1	1

of the same polarity, there are an even number (i.e., 2) of zero-valued samples. If one of the sample values is detected wrong, this rule is violated, and the error is detected.

The pulse $p(t)$ goes to zero at $t = -T_b$ and $2T_b$, resulting in the pulse width (of the primary lobe) 50% higher than that of the first criterion pulse. This pulse broadening in the time domain leads to reduction of its bandwidth. This is the second criterion proposed by Nyquist. This scheme of controlled ISI is also known as **correlative** or **partial-response** scheme. A pulse satisfying the second criterion in Eq. (7.37) is also known as the **duobinary pulse**.

7.3.4 Example of a Duobinary Pulse

If we restrict the pulse bandwidth to $R_b/2$, then following the procedure of Example 7.1, we can show that (see Prob 7.3-9) only the following pulse $p(t)$ meets the requirement in Eq. (7.37) for the duobinary pulse:

$$p(t) = \frac{\sin(\pi R_b t)}{\pi R_b t(1 - R_b t)} \tag{7.38}$$

The Fourier transform $P(f)$ of the pulse $p(t)$ is given by (see Prob 7.3-9)

$$P(f) = \frac{2}{R_b} \cos\left(\frac{\pi f}{R_b}\right) \Pi\left(\frac{f}{R_b}\right) e^{-j\pi f/R_b} \tag{7.39}$$

The pulse $p(t)$ and its amplitude spectrum $|P(f)|$ are shown in Fig. 7.16.* This pulse transmits binary data at a rate of R_b bit/s and has the theoretical minimum bandwidth $R_b/2$ Hz. Equation (7.38) shows that this pulse decays rapidly with time as $1/t^2$. This pulse is not ideally realizable because $p(t)$ is noncausal and has infinite duration [because $P(f)$ is band-limited]. However, it decays rapidly (as $1/t^2$), and therefore can be closely approximated.

It may come as a surprise that we are able to achieve the theoretical rate using the duobinary pulse. In fact, it is an illusion. The theoretical rate of transmission is 2 pieces of independent information per second per hertz bandwidth. We have achieved this rate for binary information. Here is the catch! A piece of binary information does not qualify as an independent piece of information because it cannot take on an arbitrary value. It must be selected from a finite set. The duobinary pulse would fail if the pulses were truly independent pieces of information, that is, if the pulses were to have arbitrary amplitudes. The scheme works only because the binary pulses take on finite known values, and hence, there are only a finite (known) number of interference patterns between pulses, which permits correct determination of pulse amplitudes despite interference.

* The phase spectrum is linear with $\theta_p(f) = -\pi f T_b$.

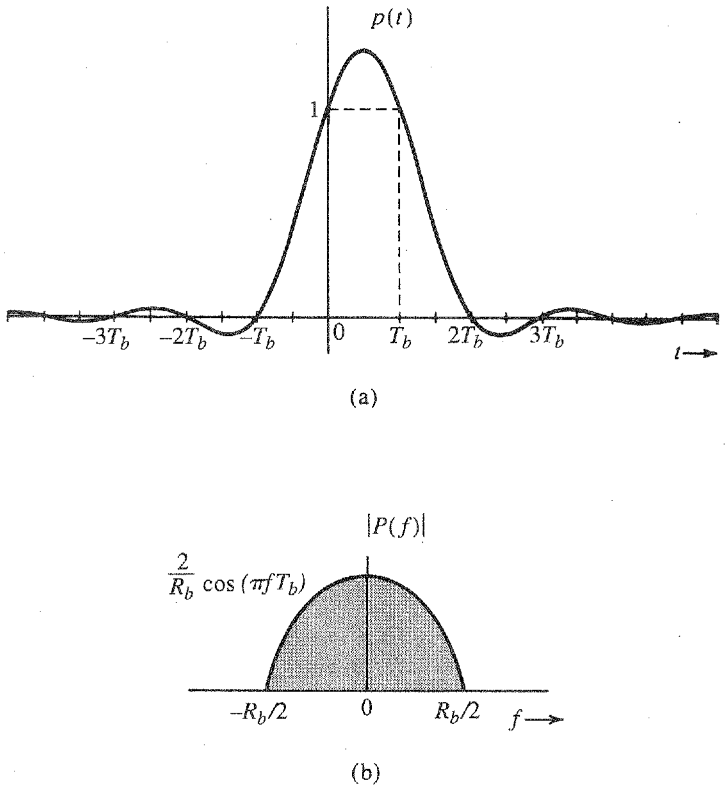

Figure 7.16 (a) The minimum bandwidth pulse that satisfies the duobinary pulse criterion and (b) its spectrum.

7.3.5 Pulse Relationship between Zero-ISI, Duobinary, and Modified Duobinary

Now we can establish the simple relationship between a pulse $p_a(t)$ satisfying the first Nyquist criterion (zero ISI) and a duobinary pulse $p_b(t)$ (with controlled ISI). From Eqs. (7.23) and (7.37), it is clear that $p_a(kT_b)$ and $p_b(kT_b)$ only differ for $k = 1$. They have identical sample values for all other integer k. Therefore, one can easily construct a pulse $p_b(t)$ from $p_a(t)$ by

$$p_b(t) = p_a(t) + p_a(t - T_b)$$

This addition is the "controlled" ISI or partial-response signaling that we deliberately introduced to reduce the bandwidth requirement. To see what effect "duobinary" signaling has on the spectral bandwidth, consider the relationship of the two pulses in the frequency domain:

$$P_b(f) = P_a(f)[1 + e^{-j2\pi f T_b}] \tag{7.40a}$$

$$|P_b(f)| = |P_a(f)|\sqrt{2(1 + \cos(2\pi f T_b)}2\,|\cos(\pi f T_b)| \tag{7.40b}$$

We can see that partial-response signaling is actually forcing a frequency null at $2\pi f T_b = \pi$ or, equivalently $f = 0.5/T_b$. Therefore, conceptually we can see how partial-response signaling provides an additional opportunity to reshape the PSD or the transmission bandwidth. Indeed, duobinary signaling, by forcing a frequency null at $0.5/T_b$, forces its essential bandwidth to be at the minimum transmission bandwidth needed for a data rate of $1/T_b$ (as discussed in Sec. 6.1.3).

In fact, many physical channels such as magnetic recording have a zero gain at dc. Therefore, it makes no sense for the baseband signal to have any dc component in its PSD. Modified partial-response signaling is often adopted to force a null at dc. One notable example is the so-called **modified duobinary** signaling that requires

$$p_c(nT_b) = \begin{cases} 1 & n = -1 \\ -1 & n = 1 \\ 0 & \text{for all other integers } n \end{cases} \tag{7.41}$$

A similar argument indicates that $p_c(t)$ can be generated from any pulse $p_a(t)$ satisfying the first Nyquist criterion via

$$p_c(t) = p_a(t + T_b) - p_a(t - T_b)$$

Equivalently, in the frequency domain, the duobinary pulse is

$$P_c(f) = 2jP_a(f) \sin(2\pi f T_b)$$

which uses $\sin(2\pi f T_b)$ to force a null at dc to comply with the physical channel constraint.

7.3.6 Detection of Duobinary Signaling and Differential Encoding

For the controlled ISI method of duobinary signaling, Fig. 7.17 shows the basic transmitter diagram. We now take a closer look at the relationship of all the data symbols at the baseband and the detection procedure. For binary message bit $I_k = 0$, or 1, the polar symbols are simply

$$a_k = 2I_k - 1$$

Under the controlled ISI, the samples of the transmission signal $y(t)$ are

$$y(kT_b) = b_k = a_k + a_{k-1} \tag{7.42}$$

The question for the receiver is how to *detect* I_k from $y(kT_b)$ or b_k. This question can be answered by first considering all the possible values of b_k or $y(kT_b)$. Because $a_k = \pm 1$, then $b_k = 0, \pm 2$. From Eq. (7.42), it is evident that

$$\begin{aligned} b_k &= 2 && \Rightarrow a_k = 1 && \text{or } I_k = 1 \\ b_k &= -2 && \Rightarrow a_k = -1 && \text{or } I_k = 0 \\ b_k &= 0 && \Rightarrow a_k = -a_{k-1} && \text{or } I_k = 1 - I_{k-1} \end{aligned} \tag{7.43}$$

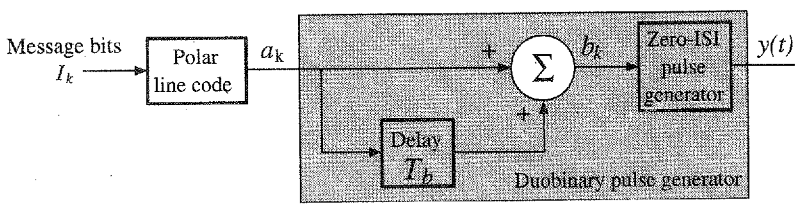

Figure 7.17 Equivalent duobinary signaling.

Therefore, a simple detector of duobinary signaling is to first detect all the bits I_k corresponding to $b_k = \pm 2$. The remaining $\{b_k\}$ are zero-valued samples that imply transition: that is, the current digit is **1** and the previous digit is **0**, or vice versa. This means the digit detection must be is based on the previous digit. An example of this digit-by-digit detection was shown in Table 7.1. The disadvantage of the detection method in Eq. (7.43) is that when $y(kT_b) = 0$, the current bit decision depends on the previous bit decision. If the previous digit were detected incorrectly, then the error would tend to propagate, until a sample value of ± 2 appears. To mitigate this error propagation problem, we apply a effective mechanism known as **differential coding**.

Figure 7.18 illustrates a duobinary signal generator by introducing an additional differential encoder prior to partial-response pulse generation. As shown in Fig. 7.18, differential encoding is a very simple step that changes the relationship between line code and the message bits. Differential encoding generates a new binary sequence

$$p_k = I_k \oplus p_{k-1} \qquad \text{modulo 2}$$

with the assumption that the precoder initial state is either $p_0 = 0$ or $p_0 = 1$. Now, the precoder output enters a polar line coder and generates

$$a_k = 2p_k - 1$$

Because of the duobinary signaling $b_k = a_k + a_{k-1}$ and the zero-ISI pulse generator, the samples of the received signal $y(t)$ without noise become

$$\begin{aligned} y(kT_b) = b_k &= a_k + a_{k-1} \\ &= 2(p_k + p_{k-1}) - 2 \\ &= 2\,(p_{k-1} \oplus I_k + p_{k-1} - 1) \\ &= \begin{cases} 2(1 - I_k) & p_{k-1} = 1 \\ 2(I_k - 1) & p_{k-1} = 0 \end{cases} \end{aligned} \tag{7.44}$$

Based on Eq. (7.44), we can summarize the direct relationship between the message bits and the sample values as

$$y(kT_b) = \begin{cases} 0 & I_k = 1 \\ \pm 2 & I_k = 0 \end{cases} \tag{7.45}$$

This relationship serves as our basis for a symbol-by-symbol detection algorithm. In short, the decision algorithm is based on the current sample $y(kT_b)$. When there is no noise, $y(kT_b) = b_k$ and the receiver decision is

$$I_k = \frac{2 - |y(kT_b)|}{2} \tag{7.46}$$

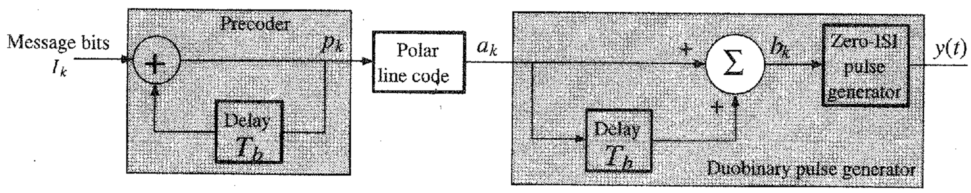

Figure 7.18 Differential encoded duobinary signaling.

TABLE 7.2
Binary Duobinary Signaling with Differential Encoding

Time k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
I_k		1	1	0	1	1	0	0	0	1	0	1	1	1
p_k	0	1	0	0	1	0	0	0	0	1	1	0	1	0
a_k	−1	1	−1	−1	1	−1	−1	−1	−1	1	1	−1	1	−1
b_k		0	0	−2	0	0	−2	−2	−2	0	2	0	0	0
Detected bits		1	1	0	1	1	0	0	0	1	0	1	1	1

Therefore, the incorporation of differential encoding with duobinary signaling not only simplifies the decision rule but also makes the decision independent of the previous digit and eliminates error propagation. In Table 7.2, the example of Table 7.1 is recalculated with differential encoding. The decoding relationship of Eq. (7.45) is clearly shown in this example.

The differential encoding defined for binary information symbols can be conveniently generalized to nonbinary symbols. When the information symbols I_k are M-ary, the only change to the differential encoding block is to replace "modulo 2" with "modulo M." Similarly, other generalized partial-response signaling such as the modified duobinary must also face the error propagation problem at its detection. A suitable type of differential encoding can be similarly adopted to prevent error propagation.

7.3.7 Pulse Generation

A pulse $p(t)$ satisfying a Nyquist criterion can be generated as the unit impulse response of a filter with transfer function $P(f)$. This will not always be easy. A better method is to generate the waveform directly, using a transversal filter (tapped delay line) discussed here. The pulse $p(t)$ to be generated is sampled with a sufficiently small sampling interval T_s (Fig. 7.19a), and the filter tap gains are set in proportion to these sample values in sequence, as shown in Fig. 7.19b. When a narrow rectangular pulse with the width T_s, the sampling interval, is applied at the input of the transversal filter, the output will be a staircase approximation of $p(t)$. This output, when passed through a low-pass filter, is smoothed out. The approximation can be improved by reducing the pulse sampling interval T_s.

It should be stressed once again that the pulses arriving at the detector input of the receiver need to meet the desired Nyquist criterion. Hence, the transmitted pulses should be so shaped that after passing through the channel, they are received in the desired (Nyquist) form. In practice, however, pulses need not be shaped rigidly at the transmitter. The final shaping can be carried out by an equalizer at the receiver, as discussed later (Sec. 7.5).

7.4 SCRAMBLING

In general, a scrambler tends to make the data more random by removing long strings of **1**s or **0**s. Scrambling can be helpful in timing extraction by removing long strings of **0**s in binary data. Scramblers, however, are primarily used for preventing unauthorized access to the data, and they are optimized for that purpose. Such optimization may actually result in generation of a long string of zeros in the data. The digital network must be able to cope with these long zero strings by using the zero replacement techniques discussed in Sec. 7.2.