ECT 305
ANALOG AND DIGITAL COMMUNICATION

Module-4

G-S Procedure and Effects in the Channel

Prepared By
Sithara Jeyaraj
Assistant professor
Dept of ECE,MACE

Syllabus

- Gram-Schmitt procedure.
  - Signal space.
- Baseband transmission through AWGN channel.
  - Mathematical model of ISI. Nyquit criterion for zero ISI.
  - Signal modeling for ISI, Raised cosine and Square-root raised cosine spectrum,
  - Partial response signalling and duobinary coding.
- Equalization.
  - Design of zero forcing equalizer.
- Vector model of AWGN channel.
  - Matched filter and correlation receivers. MAP receiver,
  - Maximum likelihood receiver and probability of error.
  - Capacity of an AWGN channel
  - (Expression only) -- significance in the design of communication schemes..

Baseband transmission through AWGN channel.

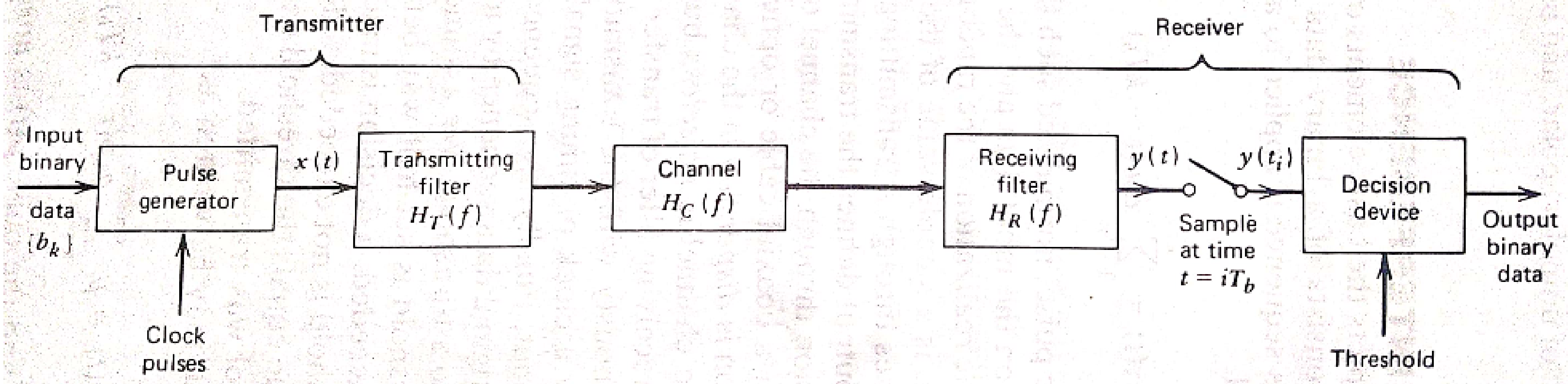


Figure 6.5 Baseband binary data transmission system.

Baseband transmission through AWGN channel.

- Consider the figure which depicts the basic elements of a baseband PAM system.
- The input signal consists of a binary data sequence $\{bk\}$ with a bit duration of T_b seconds.
- This sequence is applied to a pulse generator , producing the discrete PAM signal.

$$x(t) = \sum_{k=-\infty}^{\infty} a_k \, v(t - kT_b)$$

- Where $v(t)$ denotes the basic pulse
- The coefficients a_k depends on the input data and the type of the format used.
- The PAM signal $x(t)$ passes through a transmitting filter of transfer function $H_T(f)$
- The resulting filter output defines the transmitted signal which is modified as a result of transmission through the channel of transfer function $H_c(f)$
- The channel may represent a coaxial cable or optical fiber where a major source of system degradation is dispersion in the channel.
- We assume that channel is noiseless but dispersive.

Baseband transmission through AWGN channel.

- The channel output is passed through a receiving filter of the transfer function $H_R(f)$
- This filter output is sampled synchronously with the transmitter with sampling instants being determined by a clock or timing signal that is usually extracted from the receiving filter output.
- Finally the sequence of samples thus obtained is used to reconstruct the original data sequence by a means of decision device.
- Each sample is compared to a threshold.
- We assume that symbols 1 and 0 are equally likely and the threshold is set half way between their representation levels
- If threshold is exceeded , a decision is made favour of symbol 1
- If threshold is not exceeded , a decision is made favour of symbol 0
- If the sample value equals the threshold exactly the flip of a fair coin will determine which symbol was transmitted.

Intersymbol Interference

- The receiving filter output may be written as

$$y(t) = \mu \sum_{k=-\infty}^{\infty} a_k\, p(t - kT_b)$$

- Where μ is a scaling factor and pulse p(t) is normalized such that p(0)=1
- The output y(t) is produced in response to the binary data waveform applied to the input of the transmitting filter
- Specially the Pulse $\mu p(t)$ is the response of the cascade connection of the transmitting filter , the channel and receiving filter which is produced by the pulse $v(t)$ applied to the input of this cascade connection
- Therefore we may relate $p(t)$ to $v(t)$ in the frequency domain by writing

$$\mu P(f) = V(f)\, H_T(f) H_c(f) H_R(f)$$

- Where $P(f)$ and $V(f)$ are fourier transform of $p(t)$ and $v(t)$

Intersymbol Interference

- The receiving filter output y(t) is sampled at time $t_i = iT_b$ yielding

$$y(t_i) = \mu \sum_{k=-\infty}^{\infty} a_k \, p(iT_b - kT_b)$$

$$y(t_i) = \mu a_i + \mu \sum_{\substack{k=-\infty \\ k \neq i}}^{\infty} a_k \, p(iT_b - kT_b)$$

- The first tem is produced by the ith transmitted bit
- The second term represents the residual effect of all other transmitted bits on the decoding of the ith bit;
- this residual effect is called intersymbol interference (IS)
- In physical terms ISI arises because of imperfections in the overall frequency response of the system
- When a short pulse of duration Tb seconds is transmitted through a bandlimited system the frequency components constituting the input pulses are differentially attenuated and more significantly differentially delayed by the system.

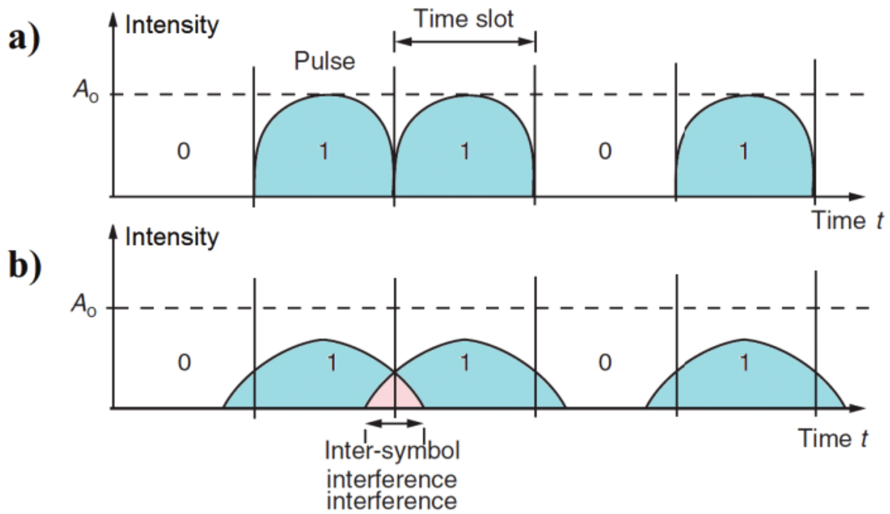
Intersymbol Interference

- Consequently the pulse appearing at the output of the system dispersed over an interval longer than Tb seconds
- This when a sequence of short pulses are transmitted through the system , one pulse every Tb seconds the dispersed Reponses originating from different symbols intervals will interfere with each other there by resulting in intersymbol interference
- In the absence of ISI we observe

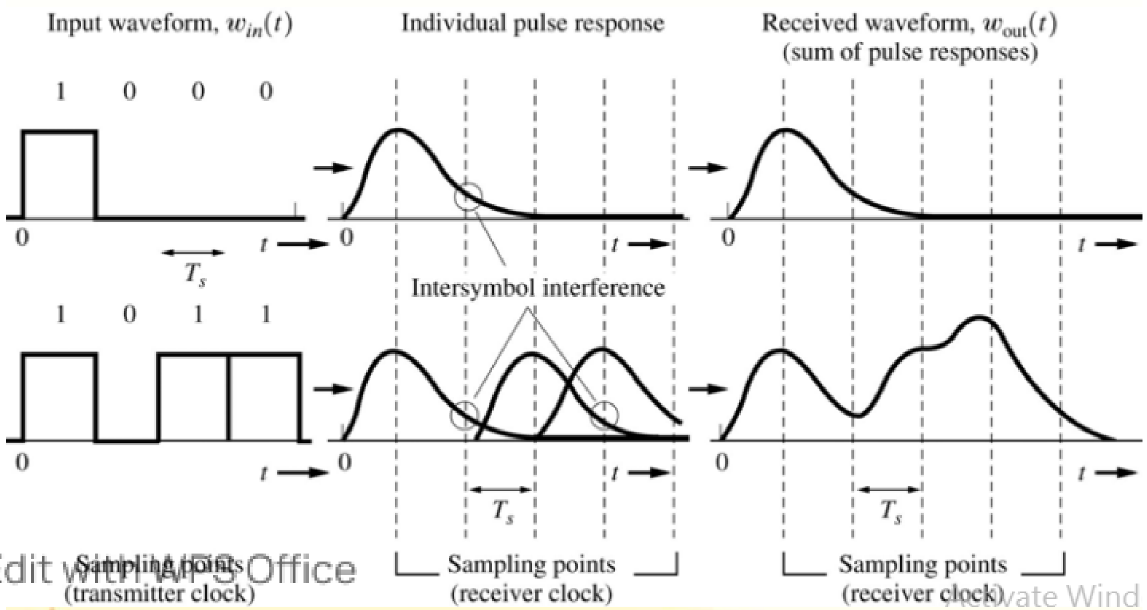
$$y(t_i) = \mu a_i$$

- Which shows that under these conditions the ith transmitted bit can be decoded correctly.
- The presence of ISI in the system however introduces errors in the decision device at the receiver output
- Therefore in the design of the transmitting and receiving filters, the objective is to minimize the effects of ISI and thereby deliver the digital data to its destination with the smallest error rate possible.

Intersymbol Interference



Examples of ISI on received pulses in a binary communication system.



Nyquist's Criterion For Distoritonless Baseband Binary Transmission

- Typically the transfer function of the channel and the transmitted pulse shape are specified and the problem is to determine the transfer functions of the transmitting and receiving filters so as to reconstruct the transmitted data sequence $\{b_k\}$
- The receiving does this by extracting and then decoding the corresponding sequence of weights $\{a_k\}$ from the output y(t)
- Except for a scaling factor y(t)is determined by the a_k and the received pulse p(t)
- The extraction involves sampling the output y(t) at some time $t_i = iT_b$
- The decoding requires that the weighted pulse contribution $a_k\ p(iT_b - kT_b)$ for $k = i$, be free from ISI due to the overlapping tails of all other weighted pulse contributions represented by $k \neq i$
- This in turn requires that we control the received pulse p(t) as shown by

Equation A

$$p(iT_\mathrm{b} - kT_\mathrm{b}) = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases}$$

- This condition the assures perfect reception in the absence of noise .

Nyquist's Criterion For Distoritonless Baseband Binary Transmission

- We consider the condition into frequency domain.
- Consider the sequence of samples {p(nTb)} where n=0,$\pm$1,$\pm$2,….
- Recall that sampling in time domain produces periodicity in the frequency domain.

$$P_\delta(f) = R_b \sum_{n=-\infty}^{\infty} P(f - R_b)$$ **Equation B**

- Where R_b=1/$T_b$is the bit rate
- $P_\delta(f)$ is the fourier transform of an infinite periodic sequence of delta functions of period T_b and whose strengths are weighted the respective sample values of p(t)
- That is $P_\delta(f)$ is given by

$$P_\delta(f) = \int_{-\infty}^{\infty} \sum_{m=-\infty}^{\infty} [p(mT_b)\,\delta(t - mT_b)\exp(-j2\pi f t)dt$$ **Equation C**

Nyquist's Criterion For Distoritonless Baseband Binary Transmission

- Let the interger m=i-k
- Then i=k corresponds to m=0 and likewise $i \neq k$corresponds to m $\neq$0
- Accordingly imposing the condition in equation A on the sample values of p(t) in the integral of the above equation we get

$$P_\delta(f) = \int_{-\infty}^{\infty} p(0)\,\delta(t)\exp(-\mathrm{j}2\pi f t)\mathrm{dt} = \mathrm{p}(0)$$ **Equation D**

- Where we made use of sifting property delta function
- Since p(0) =1 by normalization we thus see that from the above equation B and D that the condition for zero intersymbols interference is satisfied if

$$\boxed{\sum_{n=-\infty}^{\infty} P(f - R_b) = T_b}$$ **Equation E**

- Thus equation A and E constitutes the Nyquist criterion for distortion less baseband transmission in the absence of noise.
- It provides a method for constructing bandlimited functions to overcome the effects of ISI

Ideal Solution

- A frequency function occupying the narrowest that satisfies equation E is obtained by permitting only one non zero component in the series for each f in the range extending from $-W$ to W where W denotes half the bit rate

$$W = \frac{R_b}{2}$$

- That is we specify $P(f)$ as

$$P(f) = \begin{cases} \frac{1}{2W}, & -W < f < W \\ 0, & |f| > W \end{cases} = \frac{1}{2W}\mathrm{rect}\left(\frac{f}{2W}\right)$$

- In this solution no frequencies of absolute value exceeding half the bitrate are needed
- Hence, from Fourier-transform pair, we find that a signal waveform that produces zero ISI is defined by the sinc function:

$$p(t) = \frac{\sin(2\pi W t)}{2\pi W t}$$

$$= \mathrm{sinc}(2Wt)$$

Ideal Solution

- The special value of the bit rate Rb = 2W is called the Nyquist rate and W is itself called the Nyquist bandwidth.
- Correspondingly, the baseband pulse p(t) for distortionless transmission is called the ideal Nyquist pulse, ideal in the sense that the bandwidth requirement is one half the bit rate
- Figure 8.4 shows plots of P(f) and p(t).
- In part a of the figure, the normalized form of the frequency function P(f) is plotted for positive and negative frequencies.
- In part b of the figure, we have also included the signaling intervals and the corresponding centered sampling instants.
- The function p(t) can be regarded as the impulse response of an ideal low-pass filter with passband magnitude response 1/2W and bandwidth W.
- The function p(t) has its peak value at the origin and goes through zero at integer multiples of the bit duration Tb.

Ideal Solution

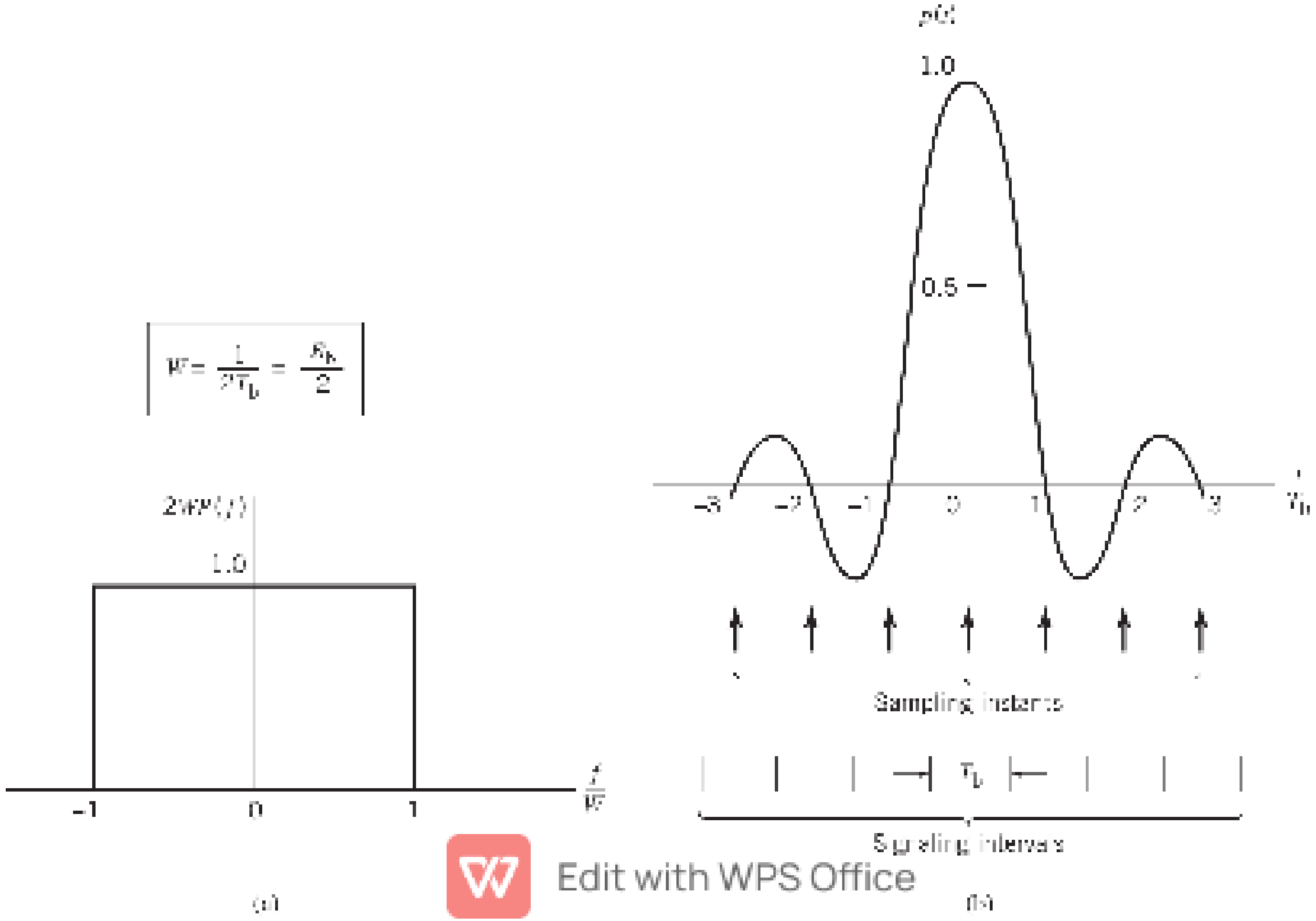


Figure 8.4 (a) Ideal magnitude response. (b) Ideal basic pulse shape.

Ideal Solution

- It is apparent, therefore, that if the received waveform y(t) is sampled at the instants of time t = 0, $\pm$Tb, $\pm$2Tb, …., then the pulses defined by ai p(t – iTb) with amplitude ai and index i = 0, $\pm$1, $\pm$2, … will not interfere with each other.
- This condition is illustrated in Figure 8.5 for the binary sequence 1011010.

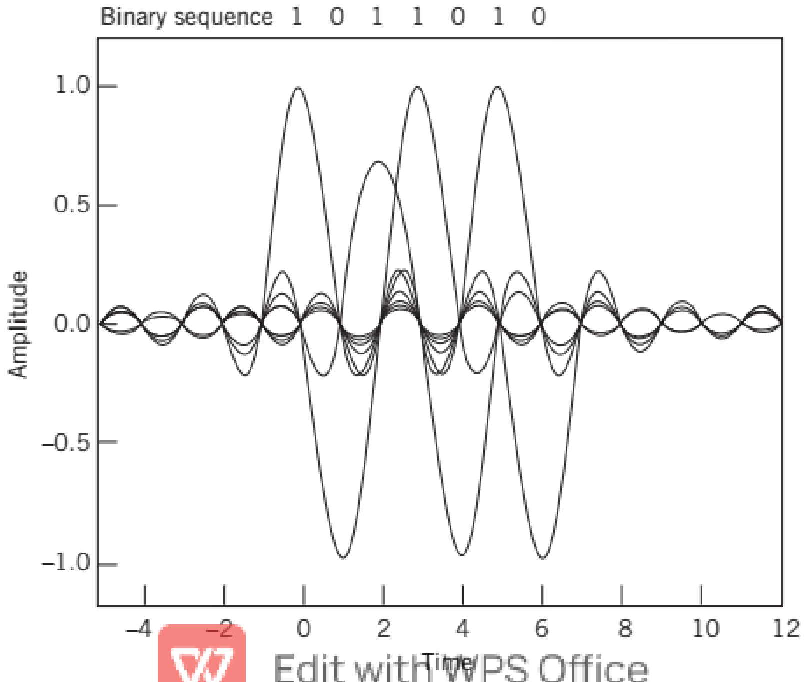


Figure 8.5 A series of sinc pulses corresponding to the sequence 1011010.

Practical Difficulties

- Although the use of the ideal Nyquist pulse does indeed achieve economy in bandwidth, in that it solves the problem of zero ISI with the minimum bandwidth possible, there are two practical difficulties that make it an undesirable objective for signal design:
  - It requires that the magnitude characteristic of P(f) be flat from –W to +W, and zero elsewhere. This is physically unrealizable because of the abrupt transitions at the band edges $\pm$W,
  - The pulse function p(t) decreases as 1/|t| for large | t|, resulting in a slow rate of decay. This is also caused by the discontinuity of P(f) at $\pm$W. Accordingly, there is practically no margin of error in sampling times in the receiver.

Ideal Solution

To evaluate the effect of the timing error alluded to under point 2, consider the sample of $y(t)$ at $t = \Delta t$, where Δt is the timing error. To simplify the exposition, we may put the correct sampling time t_i equal to zero. In the absence of noise, we thus have from the first line of (8.10):

$$\begin{aligned} y(\Delta t) &= \sum_{k=-\infty}^{\infty} a_k p(\Delta t - kT_b) \\ &= \sum_{k=-\infty}^{\infty} a_k \left\{ \frac{\sin[2\pi W(\Delta t - kT_b)]}{2\pi W(\Delta t - kT_b)} \right\} \end{aligned} \tag{8.19}$$

Since $2WT_b = 1$, by definition, we may reduce (8.19) to

$$y(\Delta t) = a_0 \operatorname{sinc}(2W\Delta t) + \frac{\sin(2\pi W\Delta t)}{\pi} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{(-1)^k a_k}{2W\Delta t - k} \tag{8.20}$$

The first term on the right-hand side of (8.20) defines the desired symbol, whereas the remaining series represents the ISI caused by the timing error Δt in sampling the receiver output $y(t)$. Unfortunately, it is possible for this series to diverge, thereby causing the receiver to make erroneous decisions that are undesirable.

Raised-Cosine Spectrum

- We may overcome the practical difficulties encountered with the ideal Nyquist pulse by extending the bandwidth from the minimum value W = Rb/2 to an adjustable value between W and 2W.
- In effect, we are trading off increased channel bandwidth for a more robust signal design that is tolerant of timing errors.
- Specifically, the overall frequency response P(f) is designed to satisfy a condition more stringent than that for the ideal Nyquist pulse, in that we retain three terms of the summation on the left-hand side of equation E and restrict the frequency band of interest to [–W, W], as shown by

$$P(f) + P(f-2W) + P(f+2W) = \frac{1}{2W}, \qquad -W \le f \le W$$

- where, on the right-hand side, we have set Rb = 1/2W
- We may now devise several band-limited functions that satisfy above equation.

Raised-Cosine Spectrum

- A particular form of P(f)that embodies many desirable features is provided by a raised-cosine (RC) spectrum.
- This frequency response consists of a flat portion and a roll-off portion that has a sinusoidal form, as shown by:
- **Equation A**

$$P(f) = \begin{cases} \dfrac{1}{2W}, & 0 \le |f| < f_1 \\ \dfrac{1}{4W}\left\{1 + \cos\left[\dfrac{\pi}{2W\alpha}(|f| - f_1)\right]\right\}, & f_1 \le |f| < 2W - f_1 \\ 0, & |f| \ge 2W - f_1 \end{cases}$$

- we have introduced a new frequency f1 and a dimensionless parameter α, which are related by

$$\alpha = 1 - \frac{f_1}{W}$$

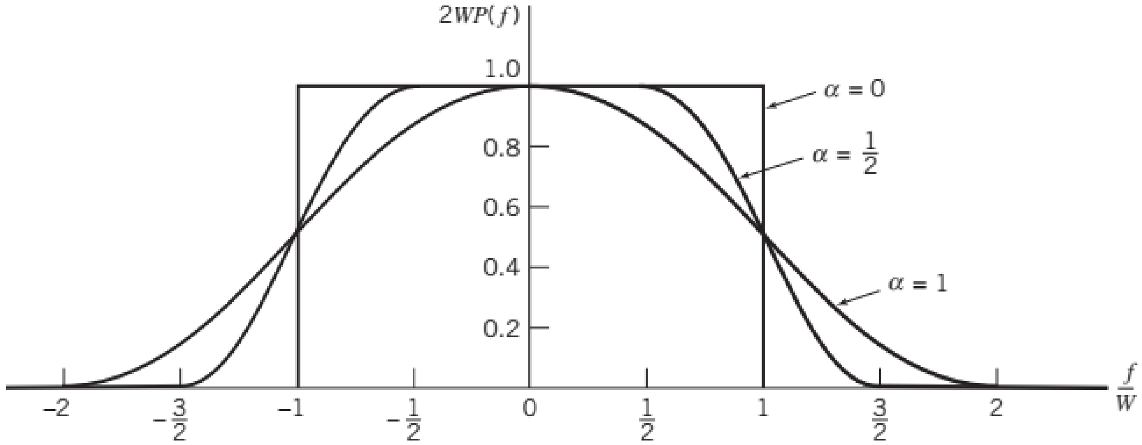
Raised-Cosine Spectrum

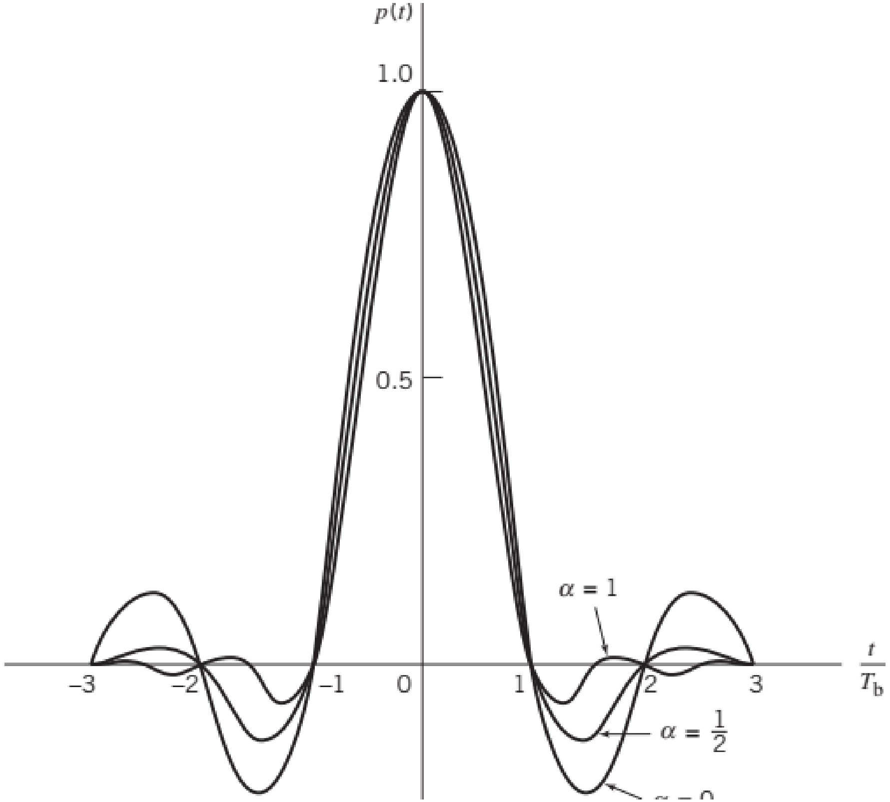
- The parameter α is commonly called the roll-off factor; it indicates the excess bandwidth over the ideal solution, W.
- Specifically, the new transmission bandwidth is defined by

$$\begin{aligned} B_T &= 2W - f_1 \\ &= W(1+\alpha) \end{aligned}$$

- The frequency response P(f), normalized by multiplying it by the factor 2W, is plotted in Figure 8.6a for α = 0, 0.5, and 1.
- We see that for α = 0.5 or 1, the frequency response P(f) rolls off gradually compared with the ideal Nyquist pulse (i.e., α = 0) and it is therefore easier to implement in practice.
- This roll-off is cosine-like in shape, hence the terminology "RC spectrum."
- Just as importantly, the P(f) exhibits odd symmetry with respect to the Nyquist bandwidth W, which makes it possible to satisfy the frequency-domain condition of (8.15).

Raised-Cosine Spectrum

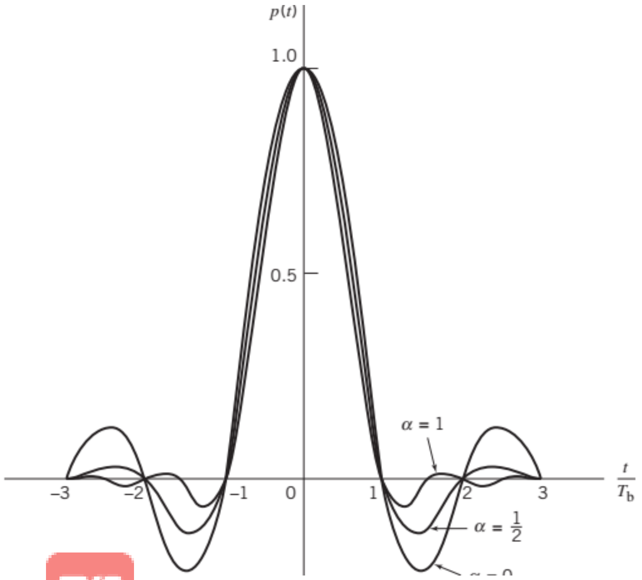




Raised-Cosine Spectrum

- The time response p(t) is naturally the inverse Fourier transform of the frequency response P(f). Hence, transforming the P(f) defined in (8.22) into the time domain, we obtain

$$p(t) = \text{sinc}(2Wt)\frac{\cos(2\pi\alpha Wt)}{1-16\alpha^2W^2t^2}$$



Raised-Cosine Spectrum

The time response $p(t)$ consists of the product of two factors: the factor sinc$(2Wt)$ characterizing the ideal Nyquist pulse and a second factor that decreases as $1/|t|^2$ for large $|t|$. The first factor ensures zero crossings of $p(t)$ at the desired sampling instants of time $t = iT_b$, with i equal to an integer (positive and negative). The second factor reduces the tails of the pulse considerably below those obtained from the ideal Nyquist pulse, so that the transmission of binary data using such pulses is relatively insensitive to sampling time errors. In fact, for $\alpha = 1$ we have the most gradual roll-off, in that the amplitudes of the oscillatory tails of $p(t)$ are smallest. Thus, the amount of ISI resulting from timing error decreases as the roll-off factor α is increased from zero to unity.

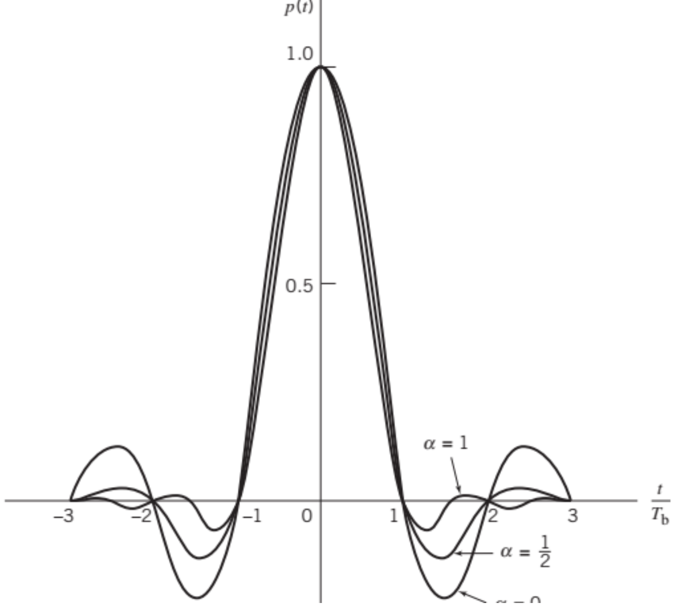
The special case with $\alpha = 1$ (i.e., $f_1 = 0$) is known as the *full-cosine roll-off* characteristic, for which the frequency response of (8.22) simplifies to

$$P(f) = \begin{cases} \frac{1}{4W}\left[1 + \cos\left(\frac{\pi f}{2W}\right)\right], & 0 < |f| < 2W \\ 0, & |f| \geq 2W \end{cases} \tag{8.26}$$

Correspondingly, the time response $p(t)$ simplifies to

$$p(t) = \frac{\text{sinc}(4Wt)}{1 - 16W^2t^2} \tag{8.27}$$

Raised-Cosine Spectrum



The time response of (8.27) exhibits two interesting properties:

1. At $t = +T_b/2 = \pm 1/4W$, we have $p(t) = 0.5$; that is, the pulse width measured at half amplitude is exactly equal to the bit duration T_b.
2. There are zero crossings at $t = \pm 3T_b/2$, $\pm 5T_b/2$, … in addition to the usual zero crossings at the sampling times $t = \pm T_b$, $\pm 2T_b$, … .

These two properties are extremely useful in extracting *timing information* from the received signal for the purpose of synchronization. However, the price paid for this desirable property is the use of a channel bandwidth double that required for the ideal Nyquist channel for which $\alpha = 0$: simply put, there is "no free lunch."

Square-Root Raised-Cosine Spectrum

- A more sophisticated form of pulse shaping uses the square-root raised-cosine (SRRC)spectrum rather than the conventional RC spectrum of Equation A.
- Specifically, the spectrum of the basic pulse is now defined by the square root of the right-hand side of this equation.
- Thus, using the trigonometric identity

$$\cos^2\theta = \frac{1}{2}(1 + \cos 2\theta)$$

$$\theta = \frac{\pi}{2W\alpha}(|f| - f_1) = \frac{\pi}{2W\alpha}[|f| - W(1-\alpha)]$$

- To avoid confusion, we use $G(f)$ as the symbol for the SRRC spectrum, and so we may write

$$G(f) = \begin{cases} \frac{1}{\sqrt{2W}}, & 0 \le |f| \le f_1 \\ \frac{1}{\sqrt{2W}}\cos\left\{\frac{\pi}{4W\alpha}[|f| - W(1-\alpha)]\right\}, & f_1 \le |f| < 2W - f_1 \\ 0, & |f| \ge 2W - f_1 \end{cases} \quad (8.33)$$

Square-Root Raised-Cosine Spectrum

If, now, the transmitter includes a *pre-modulation filter* with the transfer function defined in (8.33) and the receiver includes an identical *post-modulation* filter, then under ideal conditions the overall pulse waveform will experience the squared spectrum $G^2(f)$, which is the regular RC spectrum. In effect, by adopting the SRRC spectrum $G(f)$ of (8.33) for pulse shaping, we would be working with $G^2(f) = P(f)$ in an overall transmitter–receiver sense. On this basis, we find that in wireless communications, for example, if the channel is affected by both fading and AWGN and the pulse-shape filtering is partitioned equally between the transmitter and the receiver in the manner described herein, then effectively the receiver would maximize the output SNR at the sampling instants.

Square-Root Raised-Cosine Spectrum

The inverse Fourier transform of (8.33) defines the *SRRC shaping pulse*:

$$g(t) = \frac{\sqrt{2W}}{1-(8\alpha Wt)^2}\left\{\frac{\sin[2\pi W(1-\alpha)t]}{2\pi Wt} + \frac{4\alpha}{\pi}\cos[2\pi W(1+\alpha)t]\right\} \tag{8.34}$$

The important point to note here is the fact that the SRRC shaping pulse $g(t)$ of (8.34) is radically different from the conventional RC shaping pulse of (8.25). In particular, the new shaping pulse has the distinct property of satisfying the *orthogonality constraint under T-shifts*, described by

$$\int_{-\infty}^{\infty} g(t)g(t-nT)\,\mathrm{d}t = 0 \qquad \text{for } n = \pm 1, \pm 2, \ldots \tag{8.35}$$

where T is the symbol duration. Yet, the new pulse $g(t)$ has exactly the same excess bandwidth as the conventional RC pulse.

Square-Root Raised-Cosine Spectrum

It is also important to note, however, that despite the added property of orthogonality, the SRRC shaping pulse of (8.34) lacks the zero-crossing property of the conventional RC shaping pulse defined in (8.25).

Figure 8.9a plots the SRRC spectrum $G(f)$ for the roll-off factor $\alpha = 0, 0.5, 1$; the corresponding time-domain plots are shown in Figure 8.9b. These plots are naturally different from those of Figure 8.6 for nonzero α. The following example contrasts the waveform of a specific binary sequence using the SRRC shaping pulse with the corresponding waveform using the regular RC shaping pulse.

Square-Root Raised-Cosine Spectrum

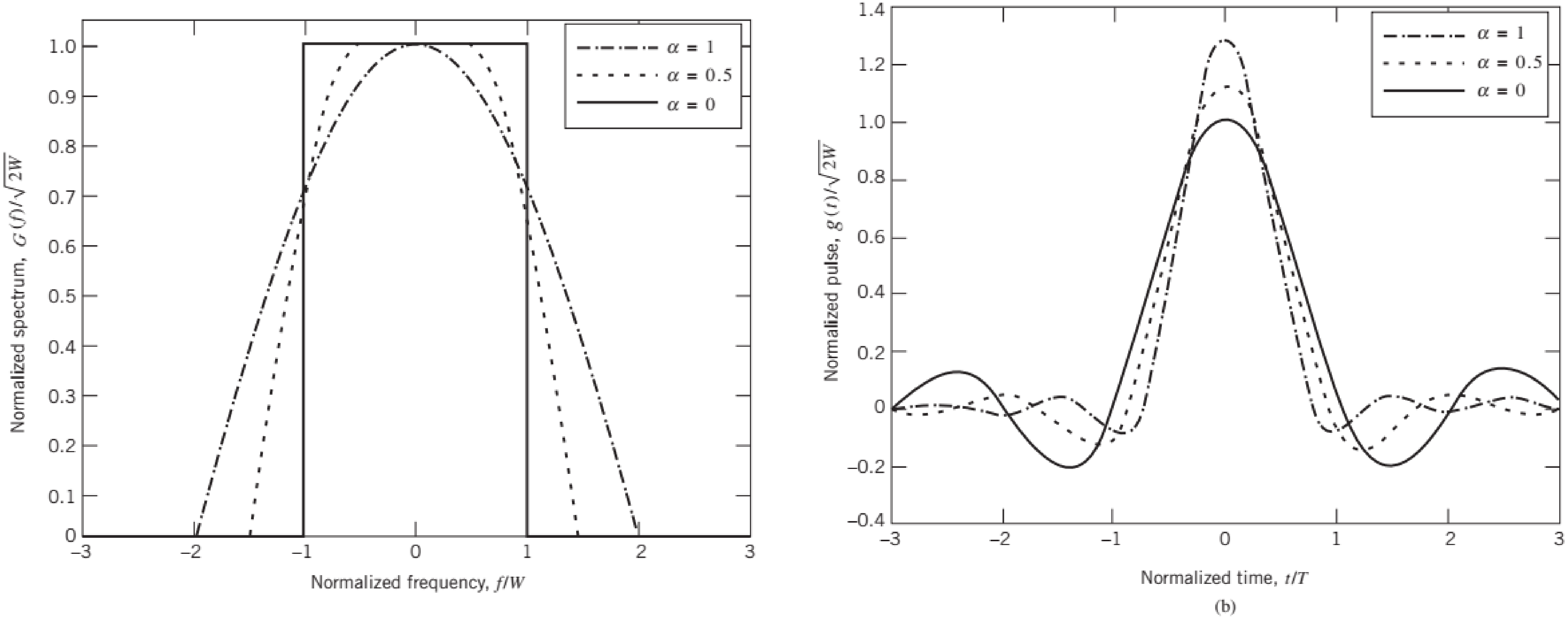


Figure 8.9 (a) $G(f)$ for SRRC spectrum. (b) $g(t)$ for SRRC pulse.

Partial response signalling

- Thus far we have treated ISI as undesirable phenomenon that produces a degradation in system performance.
- By adding ISI to the transmitted signal in a controlled manner it is possible to achieve a bit rate of 2W bits per second in a channel bandwidth W Hertz
- Such schemes are called **correlative coding or partial-response signaling schemes**
- Since ISI introduced into the transmitted signal is known, its effect can be interpreted at the receiver in a deterministic way.
- Thus correlative level coding may be regarded as a practical method of achieving the theoretical maximum signalling rate of 2W symbols per second in a bandwidth of W Hertz as postulated by Nyquist using realizable filters

Duobinary Signalling

- The basic idea of correlative level coding will be illustrated by considering the specific example of duobinary signalling where duo implies doubling of the transmission capacity of a straight binary systems
- This particular form of correlative level coding is also called class I partial response
- Consider a binary input sequence $\{b_k\}$ consisting of uncorrelated binary symbols 1 and 0 each having duration Tb
- As before this sequence is applied to a pulse amplitude modulator producing a two level sequence of short pulses whose amplitude a_k is defined by

$$a_k = \begin{cases} +1 & \text{if symbol } b_k \text{ is } 1 \\ -1 & \text{if symbol } b_k \text{ is } 0 \end{cases}$$

- When this sequence is applied to a duobinary encoder it is converted into a three level output namely -2,0and +2.
- To produce this transformation we use the scheme shown below

Duobinary Signalling

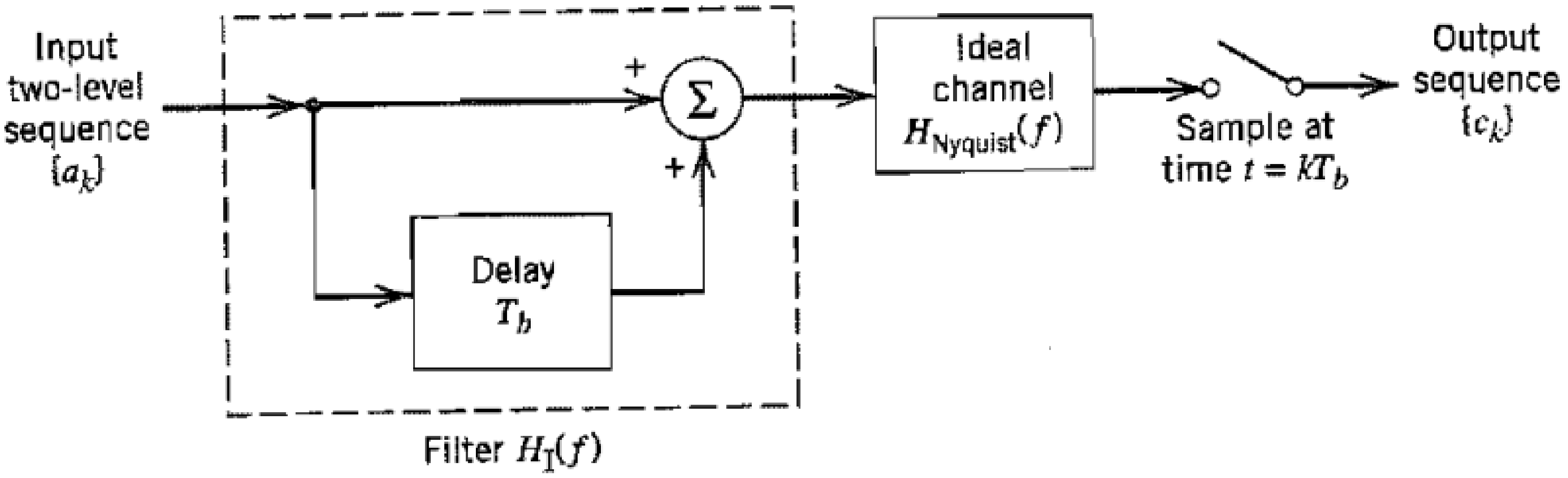


FIGURE 4.11 Duobinary signaling scheme.

Duobinary Signalling

- The two level sequence $\{a_k\}$ is first passed through a simple filter involving a single delay element and summer
- For every unit impulse applied to the input of this filter , we get two unit impulses spaced Tb seconds apart at the filter output
- We may therefore express the duobinary coder output c_k as the sum of the present input pulse a_k and its previous value a_{k-1} as shown by

$$c_k = a_k + a_{k-1}$$

- One of the effects of the transformation described by the above equation is to change the input sequence $\{a_k\}$ of uncorrelated two level pulses into a sequence $\{c_k\}$ of correlated three level pulses
- This correlation between the adjacent pulses may be viewed as introducing ISI into the transmitted signal in an artificial manner
- However the ISI so introduced is under the designers control which is the basis of correlative coding

Duobinary Signalling

An ideal delay element, producing a delay of T_b seconds, has the frequency response $\exp(-j2\pi f T_b)$, so that the frequency response of the simple delay-line filter in Figure 4.11 is $1 + \exp(-j2\pi f T_b)$. Hence, the overall frequency response of this filter connected in cascade with an ideal Nyquist channel is

$$\begin{aligned} H_I(f) &= H_{\text{Nyquist}}(f)[1 + \exp(-j2\pi f T_b)] \\ &= H_{\text{Nyquist}}(f)[\exp(j\pi f T_b) + \exp(-j\pi f T_b)] \exp(-j\pi f T_b) \\ &= 2H_{\text{Nyquist}}(f) \cos(\pi f T_b) \exp(-j\pi f T_b) \end{aligned} \tag{4.67}$$

where the subscript I in $H_I(f)$ indicates the pertinent class of partial response. For an ideal Nyquist channel of bandwidth $W = 1/2T_b$, we have (ignoring the scaling factor T_b)

$$H_{\text{Nyquist}}(f) = \begin{cases} 1, & |f| \leq 1/2T_b \\ 0, & \text{otherwise} \end{cases} \tag{4.68}$$

Duobinary Signalling

Thus the overall frequency response of the duobinary signaling scheme has the form of a half-cycle cosine function, as shown by

$$H_I(f) = \begin{cases} 2\cos(\pi f T_b)\exp(-j\pi f T_b), & |f| \leq 1/2T_b \\ 0, & \text{otherwise} \end{cases} \tag{4.69}$$

for which the magnitude response and phase response are as shown in Figures 4.12*a* and 4.12*b*, respectively. An advantage of this frequency response is that it can be easily approximated, in practice, by virtue of the fact that there is continuity at the band edges.

From the first line in Equation (4.67) and the definition of $H_{\text{Nyquist}}(f)$ in Equation (4.68), we find that the impulse response corresponding to the frequency response $H_I(f)$

Duobinary Signalling

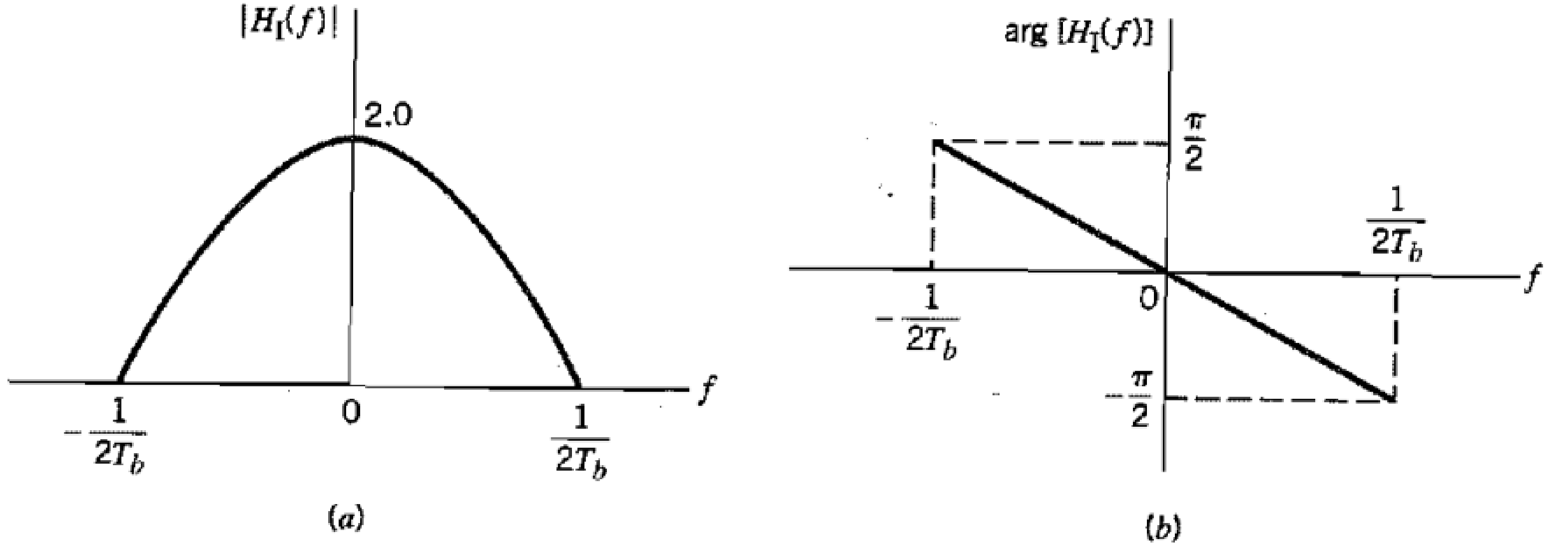


FIGURE 4.12 Frequency response of the duobinary conversion filter. (*a*) Magnitude response. (*b*) Phase response.

Duobinary Signalling

consists of two sinc (Nyquist) pulses that are time-displaced by T_b seconds with respect to each other, as shown by (except for a scaling factor)

$$
\begin{aligned}
h_I(t) &= \frac{\sin(\pi t/T_b)}{\pi t/T_b} + \frac{\sin[\pi(t - T_b)/T_b]}{\pi(t - T_b)/T_b} \\
&= \frac{\sin(\pi t/T_b)}{\pi t/T_b} - \frac{\sin(\pi t/T_b)}{\pi(t - T_b)/T_b} \\
&= \frac{T_b^2 \sin(\pi t/T_b)}{\pi t(T_b - t)}
\end{aligned}
\tag{4.70}
$$

Duobinary Signalling

The impulse response $h_I(t)$ is plotted in Figure 4.13, where we see that it has only *two* distinguishable values at the sampling instants. The form of $h_I(t)$ shown here explains why we also refer to this type of correlative coding as partial-response signaling. The response to an input pulse is spread over more than one signaling interval; stated in another way, the response in any signaling interval is "partial." Note also that the tails of $h_I(t)$ decay as $1/|t|^2$, which is a faster rate of decay than the $1/|t|$ encountered in the ideal Nyquist channel.

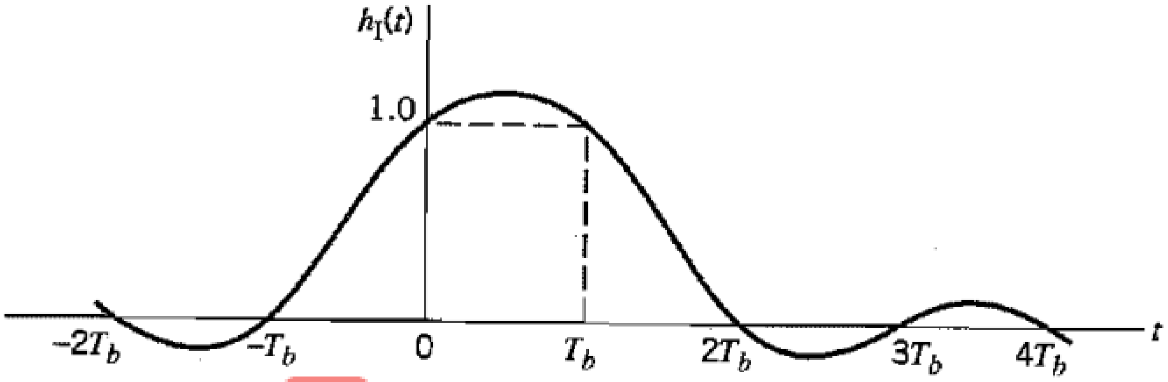


FIGURE 4.13 Impulse response of the duobinary conversion filter.

Duobinary Signalling

The original two-level sequence $\{a_k\}$ may be detected from the duobinary-coded sequence $\{c_k\}$ by invoking the use of Equation (4.66). Specifically, let $\hat{a}_k$ represent the *estimate* of the original pulse a_k as conceived by the receiver at time $t = kt_b$. Then, subtracting the previous estimate $\hat{a}_{k-1}$ from c_k, we get

$$\hat{a}_k = c_k - \hat{a}_{k-1} \tag{4.71}$$

It is apparent that if c_k is received without error and if also the previous estimate $\hat{a}_{k-1}$ at time $t = (k-1)T_b$ corresponds to a correct decision, then the current estimate $\hat{a}_k$ will be

correct too. The technique of using a stored estimate of the previous symbol is called *decision feedback.*

We observe that the detection procedure just described is essentially an inverse of the operation of the simple delay-line filter at the transmitter. However, a major drawback of this detection procedure is that once errors are made, they tend to *propagate* through the output because a decision on the current input a_k depends on the correctness of the decision made on the previous input a_{k-1}.

Duobinary Signalling

A practical means of avoiding the error-propagation phenomenon is to use *precoding* before the duobinary coding, as shown in Figure 4.14. The precoding operation performed on the binary data sequence $\{b_k\}$ converts it into another binary sequence $\{d_k\}$ defined by

$$d_k = b_k \oplus d_{k-1} \tag{4.72}$$

where the symbol $\oplus$ denotes *modulo-two addition* of the binary digits b_k and d_{k-1}. This addition is equivalent to a two-input EXCLUSIVE OR operation, which is performed as follows:

$$d_k = \begin{cases} \text{symbol 1} & \text{if either symbol } b_k \text{ or symbol } d_{k-1} \text{ (but not both) is 1} \\ \text{symbol 0} & \text{otherwise} \end{cases} \tag{4.73}$$

The precoded binary sequence $\{d_k\}$ is applied to a pulse-amplitude modulator, producing a corresponding two-level sequence of short pulses $\{a_k\}$, where $a_k = \pm 1$ as before. This sequence of short pulses is next applied to the duobinary coder, thereby producing the sequence $\{c_k\}$ that is related to $\{a_k\}$ as follows:

$$c_k = a_k + a_{k-1} \tag{4.74}$$

Duobinary Signalling

Note that unlike the linear operation of duobinary coding, the precoding described by Equation (4.72) is a *nonlinear* operation.

The combined use of Equations (4.72) and (4.74) yields

$$c_k = \begin{cases} 0 & \text{if data symbol } b_k \text{ is } 1 \\ \pm 2 & \text{if data symbol } b_k \text{ is } 0 \end{cases} \tag{4.75}$$

which is illustrated in Example 4.3. From Equation (4.75) we deduce the following decision rule for detecting the original binary sequence $\{b_k\}$ from $\{c_k\}$:

$$\begin{aligned} &\text{If } |c_k| < 1, \quad \text{say symbol } b_k \text{ is } 1 \\ &\text{If } |c_k| > 1, \quad \text{say symbol } b_k \text{ is } 0 \end{aligned} \tag{4.76}$$

Duobinary Signalling

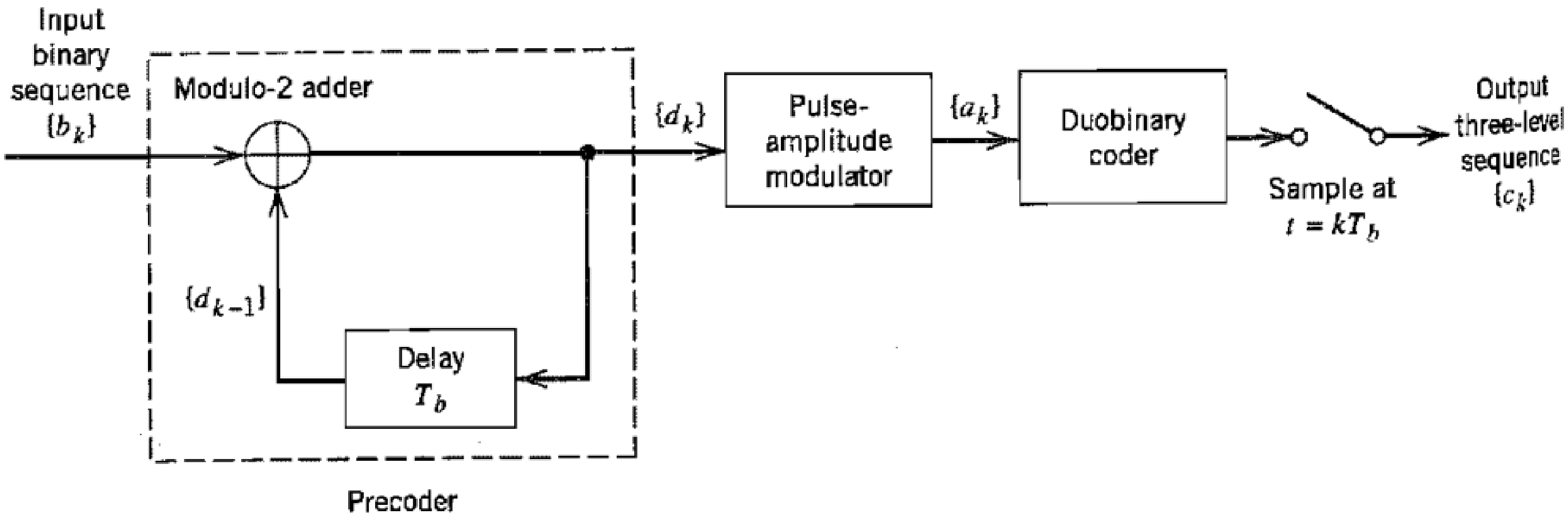


FIGURE 4.14 A precoded duobinary scheme; details of the duobinary coder are given in Figure 4.11.

Duobinary Signalling

When $|c_k| = 1$, the receiver simply makes a random guess in favor of symbol 1 or 0. According to this decision rule, the detector consists of a rectifier, the output of which is compared in a decision device to a threshold of 1. A block diagram of the detector is shown in Figure 4.15. A useful feature of this detector is that no knowledge of any input sample other than the present one is required. Hence, error propagation cannot occur in the detector of Figure 4.15.

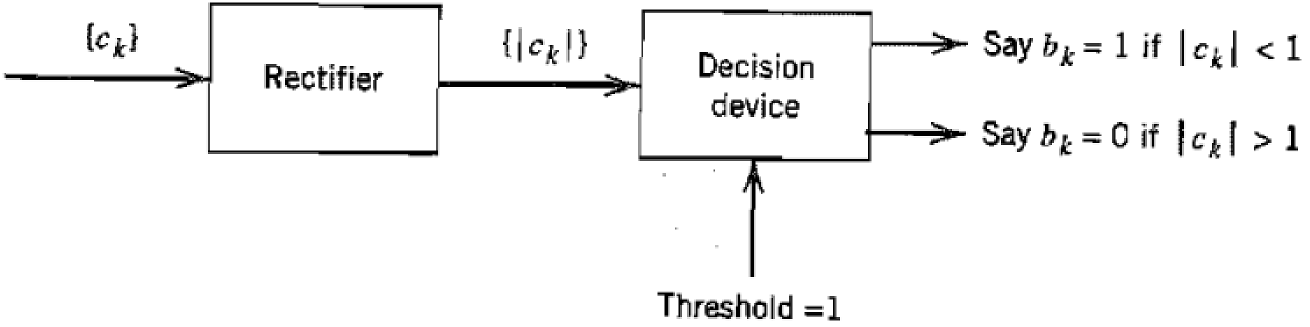


FIGURE 4.15 Detector for recovering original binary sequence from the precoded duobinary coder output.

Duobinary Signalling

Example 4.3 Duobinary Coding with Precoding

Consider the binary data sequence 0010110. To proceed with the precoding of this sequence, which involves feeding the precoder output back to the input, we add an extra bit to the precoder output. This extra bit is chosen arbitrarily to be 1. Hence, using Equation (4.73), we find that the sequence $\{d_k\}$ at the precoder output is as shown in row 2 of Table 4.1. The polar representation of the precoded sequence $\{d_k\}$ is shown in row 3 of Table 4.1. Finally, using Equation (4.74), we find that the duobinary coder output has the amplitude levels given in row 4 of Table 4.1.

To detect the original binary sequence, we apply the decision rule of Equation (4.76), and so obtain the binary sequence given in row 5 of Table 4.1. This latter result shows that, in the absence of noise, the original binary sequence is detected correctly.

Duobinary Signalling

TABLE 4.1 *Illustrating Example 4.3 on duobinary coding*

Binary sequence $\{b_k\}$		0	0	1	0	1	1	0
Precoded sequence $\{d_k\}$	1	1	1	0	0	1	0	0
Two-level sequence $\{a_k\}$	+1	+1	+1	−1	−1	+1	−1	−1
Duobinary coder output $\{c_k\}$		+2	+2	0	−2	0	0	−2
Binary sequence obtained by applying decision rule of Eq. (4.76)		0	0	1	0	1	1	0

Conversion of the Continuous AWGN Channel into a Vector Channel

- Suppose that the $s_i(t)$ is not any signal, but specifically the signal at the receiver side, defined in accordance with an AWGN channel:
- So the output of the correlator (Fig. 5.3b) can be defined as:

$$x(t) = s_i(t) + w(t),$$

$$\begin{Bmatrix} 0 \le t \le T \\ i = 1, 2, \ldots, M \end{Bmatrix} \quad (5.28)$$

$$\begin{aligned} x_i &= \int_0^T x(t)\phi_j(t)\,dt \\ &= s_{ij} + w_i, \end{aligned}$$

$$j = 1, 2, \ldots, N \quad (5.29)$$

$$s_{ij} = \int_0^T s_i(t)\phi_i(t)\,dt \quad (5.30)$$

$$w_i = \int_0^T w(t)\phi_i(t)\,dt \quad (5.31)$$

deterministic quantity

contributed by the transmitted signal $s_i(t)$

random quantity

sample value of the variable W_i due to noise

Now,

- Consider a random process $X^1(t)$, with $x^1(t)$, a sample function which is related to the received signal $x(t)$ as follows:
- Using 5.28, 5.29 and 5.30 and the expansion 5.5 we get:

$$x'(t) = x(t) - \sum_{j=1}^{N} x_j \phi_i(t) \qquad (5.32)$$

$$x'(t) = x(t) - \sum_{j=1}^{N} (s_{ij} + w_j)\phi_j(t)$$

$$= w(t) - \sum_{j=1}^{N} w_j \phi_j(t)$$

$$= w'(t) \qquad (5.33)$$

which means that the sample function $x^1(t)$ depends only on the channel noise!

- The received signal can be expressed as:

$$x(t) = \sum_{j=1}^{N} x_j \phi_i(t) + x'(t)$$

$$= \sum_{j=1}^{N} x_j \phi_i(t) + w'(t) \qquad (5.34)$$

NOTE: *This is an expansion similar to the one in 5.5 but it is random, due to the additive noise.*

Statistical Characterization

- The received signal (output of the correlator of Fig.5.3b) is a random signal.
- To describe it we need to use statistical methods – mean and variance.
- The assumptions are:
  - X(t) denotes a random process, a sample function of which is represented by the received signal x(t).
  - X_j denotes a random variable whose sample value is represented by the correlator output x_j, j = 1, 2, …N.
  - We have assumed AWGN, so the noise is Gaussian, so X(t) is a Gaussian process and being a Gaussian RV, X_j is described fully by its mean value and variance.

Mean Value

- Let W_j, denote a random variable, represented by its sample value w_j, produced by the j^{th} correlator in response to the Gaussian noise component w(t).
- So it has zero mean (by definition of the AWGN model)
- ...then the *mean of X_j* depends only on s_{ij}:

$$\mu_{X_j} = E\left[X_j\right]$$
$$= E\left[s_{ij} + W_j\right]$$
$$= s_{ij} + E[W_j]$$
$$\mu_{X_j} = s_{ij} \qquad (5.35)$$

Variance

- Starting from the definition, we substitute using 5.29 and 5.31

$$\sigma_{x_i}^2 = \mathrm{var}[X_j]$$
$$= E\left[(X_j - s_{ij})^2\right]$$
$$= E\left[W_j^2\right] \qquad (5.36)$$

$$w_i = \int_0^T w(t)\,\phi_i(t)\,dt \qquad (5.31)$$

$$\sigma_{x_i}^2 = E\left[\int_0^T W(t)\,\phi_j(t)\,dt \int_0^T W(u)\,\phi_j(u)\,du\right]$$
$$= E\left[\int_o^T \int_0^T \phi_j(t)\,\phi_i(u)\,W(t)\,W(u)\,dtdu\right] \qquad (5.37)$$

$$\sigma_{x_i}^2 = \int_o^T \int_0^T \phi_i(t)\,\phi_j(u)\,E[W(t)\,W(u)]\,dtdu$$
$$= E\left[\int_o^T \int_0^T \phi_j(t)\,\phi_i(u)\,R_w(t,u)\,dtdu\right] \qquad (5.38)$$

Autocorrelation function of the noise process

- It can be expressed as: (because the noise is stationary and with a constant power spectral density)

$$R_w(t,u) = \frac{N_0}{2}\delta(t-u) \qquad (5.39)$$

- After substitution for the variance we get:

$$\sigma_{x_i}^2 = \frac{N_0}{2}\int_o^T \int_0^T \phi_i(t)\phi_j(u)\delta(t-u)\,dt\,du$$

$$= \frac{N_0}{2}\int_0^T \phi_j^2(t)\,dt \qquad (5.40)$$

- And since $\varphi_j(t)$ has unit energy for the *variance* we finally have:

$$\sigma_{x_i}^2 = \frac{N_0}{2} \quad \text{for all } j \qquad (5.41)$$

- *Correlator outputs, denoted by X_j have variance equal to the power spectral density $N_0/2$ of the noise process $W(t)$.*

Properties (without proof)

- X_j are mutually uncorrelated
- X_j are statistically independent (follows from above because X_j are Gaussian)
- and for a memoryless channel the following equation is true:

$$f_x(x/m_i) = \prod_{j=1}^{N} f_{x_j}(x_j/m_i), \qquad i = 1,2,\ldots,M \qquad (5.44)$$

- **<u>Define (construct)</u>** a vector **X** of N random variables, X_1, X_2, …X_N,
- whose elements are independent Gaussian RV with mean values s_{ij}, (output of the correlator, deterministic part of the signal defined by the signal transmitted)
- and variance equal to $N_0/2$ (output of the correlator, random part, calculated noise added by the channel).
- then the X_1, X_2, …X_N , elements of **X** are statistically independent.
- So, we can express the conditional probability of X, given $s_i(t)$ (correspondingly symbol m_i) as a product of the conditional density functions (f_X) of its individual elements f_{xj}.

NOTE: This is equal to finding an expression of the probability of a received symbol given a specific symbol was sent, assuming a memoryless channel

- …that is:

$$f_{x}(x/m_i) = \prod_{j=1}^{N} f_{x_j}(x_j/m_i), \qquad i=1,2,....,M \qquad (5.44)$$

- where, the ***vector x*** and the *scalar x_j*, are sample values of the random vector ***X*** and the random variable ***X_j.***

$$f_x(x/m_i) = \prod_{j=1}^{N} f_{x_j}(x_j/m_i), \qquad i=1,2,\ldots,M \qquad (5.44)$$

Vector x and *scalar* x_j are sample values of the random vector **X** and the random variable X_j

Vector x is called *observation vector*
Scalar x_j is called *observable element*

- Since, each X_j is Gaussian with mean s_j and variance $N_0/2$

$$f_{X_j}(x_j|m_i) = \frac{1}{\sqrt{\pi N_0}} \exp\left[-\frac{1}{N_0}(x_j - s_{ij})^2\right], \qquad \begin{array}{l} j = 1, 2, \ldots, N \\ i = 1, 2, \ldots, M \end{array} \qquad (5.45)$$

- we can substitute in 5.44 to get 5.46:

$$f_x(x/m_i) = (\pi N_0)^{-N/2} \exp\left[-\frac{1}{N_0}\sum_{j=1}^{N}(x_j - s_{ij})^2\right], \qquad i = 1, 2, \ldots, M \qquad (5.46)$$

- If we go back to the formulation of the received signal through a AWGN channel 5.34

$$x(t) = \sum_{j=1}^{N} x_j \phi_i(t) + x'(t)$$
$$= \sum_{j=1}^{N} x_j \phi_i(t) + w'(t) \qquad (5.34)$$

The vector that we have constructed fully defines this part

Only projections of the noise onto the basis functions of the signal set $\{s_i(t)\}^M_{i=1}$ affect the significant statistics of the detection problem

Finally,

- The AWGN channel, is equivalent to an N-dimensional vector channel, described by the observation vector

$$x = s_i + w, \qquad i = 1, 2, \ldots\ldots, M \qquad (5.48)$$

Likelihood Functions

- The conditional probability density functions $f_X(x|m_i)$, $I = 1, 2, 3, \ldots M$ are the very characterization of the AWGN channel.
- They express the *functional dependence* of the ***observation vector x*** on the ***transmitted message symbol m_i***. (known as the transmitted message symbol)

62/26

However,

- If we have the observation vector given, and we want to define the transmitted message signal, then we have the reverse situation
- We introduce the "likelihood function" $L(m_i)$ as:

$$L(m_i) = f_X(x/m_i), \qquad i = 1, 2, \ldots.., M \qquad (5.49)$$

- Or log likelihood function ..$l(m_i)$ as:

$$l(m_i) = \log L(m_i), \qquad i = 1, 2, \ldots.., M \qquad (5.50)$$

63/26

Log-Likelihood Function of AWGN Channel

- Substitute 5.46 into 5.50:

Vector presentation of the AWGN channel

$$f_x(x/m_i) = (\pi N_0)^{-N/2} \exp\left[-\frac{1}{N_0}\sum_{j=1}^{N}(x_j - s_{ij})^2\right], \qquad i = 1,2,\ldots,M \quad (5.46)$$

$$l(m_i) = \log L(m_i), \qquad i = 1, 2, \ldots, M \qquad (5.50)$$

- where s_{ij}, j = 1, 2, 3, ..N are the elements of the ***signal vector s_i***, representing the ***message symbol m_i***.

64/26

So,

$$l(m_i) = -\frac{1}{N_0}\sum_{j=1}^{N}(x_j - s_{ij})^2, \qquad i = 1, 2, \ldots\ldots, M \qquad (5.51)$$

which is the log likelihood function of the AWGN channel..

65/26

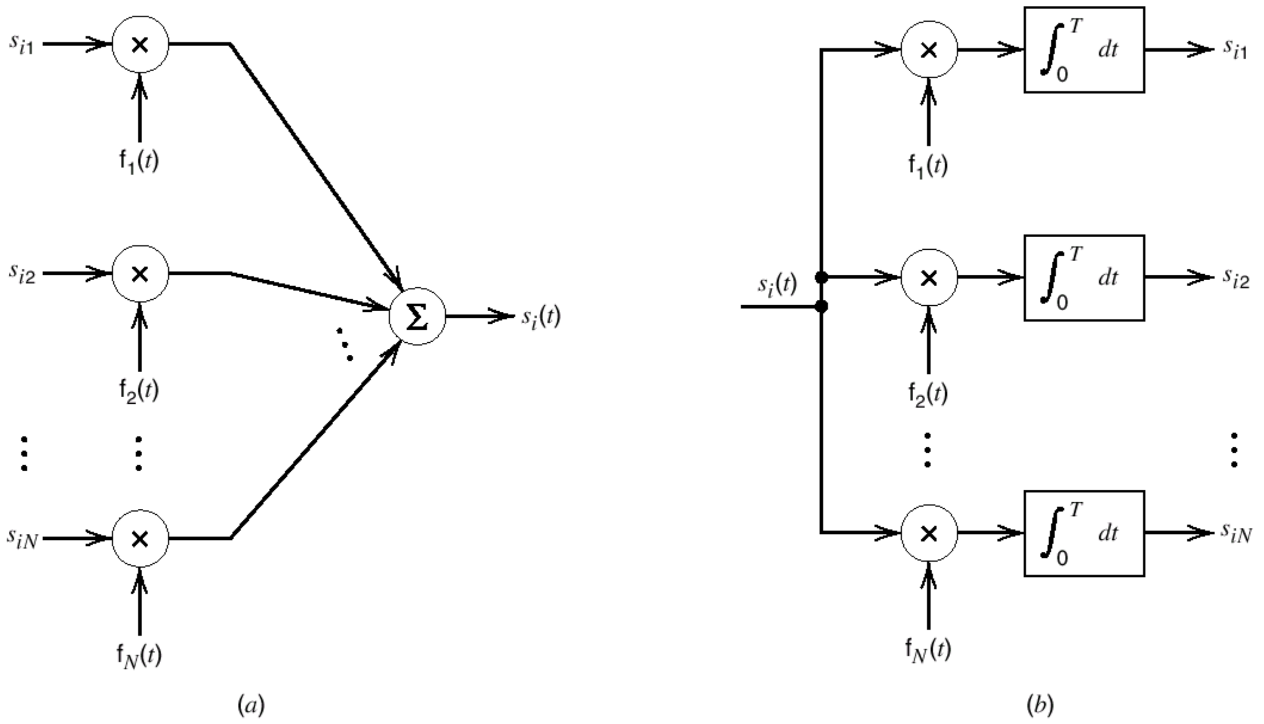
5.5 Maximum Likelihood Decoding

- Defining the problem
  - Suppose that in each time slot duration of T seconds, one of M possible signals, $s_1(t)$, $s_2(t)$, …$s_M(t)$ is transmitted with equal probability, 1/M.
  - the vector representation, the signal $s_i(t)$, i=1, 2, …M is applied to a bank of correlators, with a common input and supplied with a suitable set of N orthogonal basis functions, N. The resulting output defines the signal vector $\boldsymbol{s_i}$.
  - We represent each signal $s_i(t)$ as a point in the Euclidian space, N ≤ M (referred to as *transmitted signal point* or *message point*).The set of message points corresponding to the set of transmitted signals $s_i(t)$ {i = 1 to M} is called *signal constellation*.

66/26

Figure 5.3

(a) Synthesizer for generating the signal $s_i(t)$. (b) Analyzer for generating the set of signal vectors $\{s_i\}$.

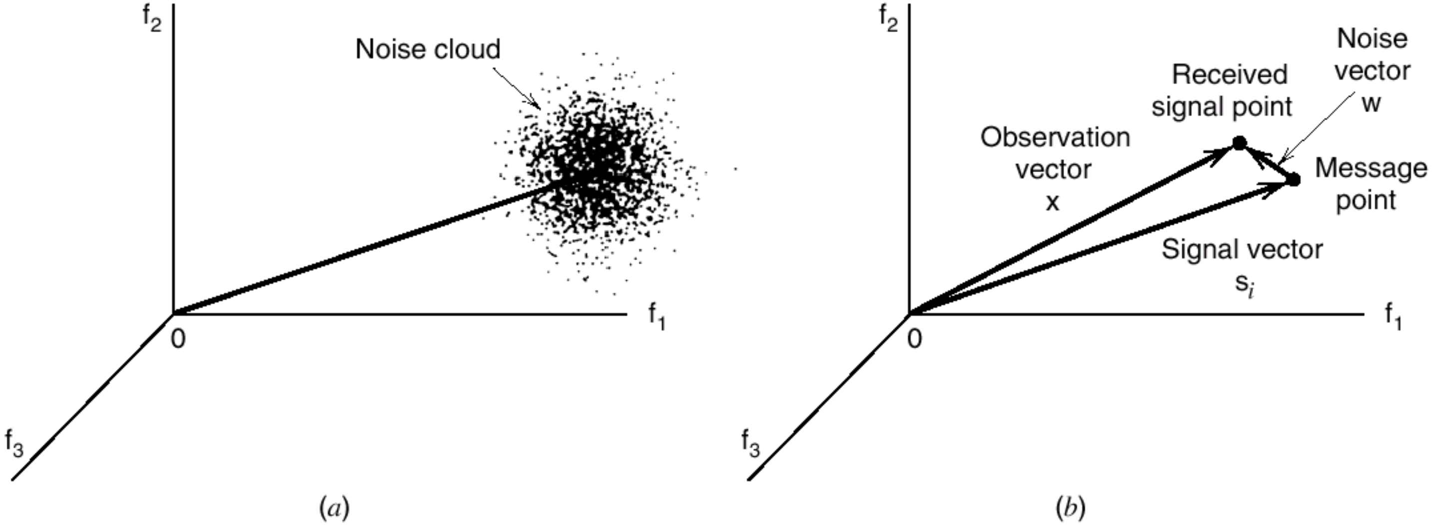


67/26

- The received signal ***x(t)*** is applied to a bank of N correlators (Fig. 5.3b) and the correlator outputs define the *observation vector* ***x***.
- On the receiving side the representation of the received signal ***x(t)*** is complicated by the additive noise ***w(t)***.
- As we discussed the previous class, the *vector* ***x*** *differs from the vector* $\mathbf{s}_i$ by the noise vector ***w***.
- However only the portion of it which interferes with the detection process is of importance to us, and this is fully described by ***w(t)***.

68/26

- Based on the *observation vector* ***x*** we may represent the received signal signal ***x(t)*** by a point in the same Euclidian space used to represent the transmitted signal.



69/26

- **Signal Detection Problem**
- For a given observation vector x we have to make a estimate m' = m_i
- The decision is based on the criterion to minimize the probability of error in mapping each observation vector into a decision.
- Suppose that given the observation vector x we make decision $\hat{m} = m_i$
- The probability of error in this decision which denoted by $P_e(m_i|\mathrm{x})$ is simply

$$\begin{aligned} P_e(m_i|\,\mathbf{x}) &= P(m_i \text{ not sent}\,|\,\mathbf{x}) \\ &= 1 - P(m_i \text{ sent}\,|\,\mathbf{x}) \end{aligned}$$

Maximum a posteriori probability (MAP Rule)

So the optimum decision rule is:

Set $\hat{m} = m_i$ if

$$P(m_i \text{ sent} / x) \geq P(m_k \text{ sent}/x) \quad \text{for all } k \neq i \qquad (5.53)$$

Where k=1,2,…M

- The same rule can be more explicitly expressed using the *a priori* probabilities of the transmitted signals as:

$$\text{Set } \hat{m} = m_i \text{ if } \frac{p_k \, f_X(x/m_k)}{f_X(x)} \text{ is maximum for } k = i \qquad (5.54)$$

Conditional pdf of observation vector X given m_k was transmitted

a priori probability of transmitting m_k

Unconditional pdf of observation vector X

72/26

- Thus we can conclude, according to the definition of likelihood functions, the likelihood function $l(m_k)$ will be maximum for k = i.
- So the decision rule using the likelihood function will be formulated as:

Maximum Likelihood rule

$$\text{Set } \hat{m} = m_i \quad \text{if}$$

$$l(m_k) \text{ is maximum for } k = i \qquad (5.55)$$

- For a graphical representation of the maximum likelihood rule we introduce the following:
  - Observation space – Z, N-dimensional, consisting of all possible observation vectors x
  - Z is partitioned into M decision regions, Z_1, Z_2, .. Z_M

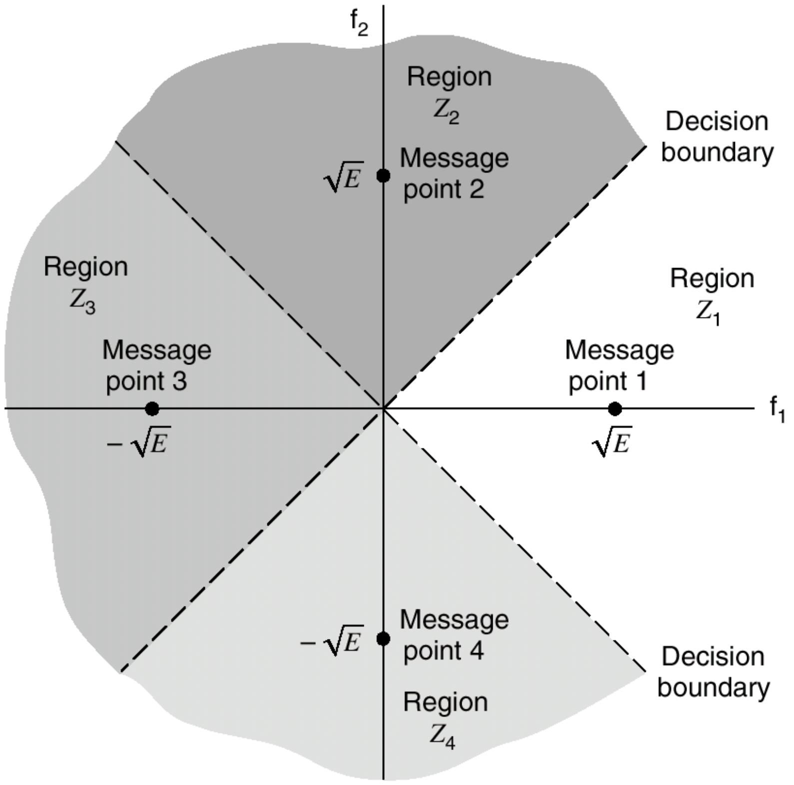
$$\text{Observation vector x lies region } Z_i \text{ if}$$

$$l(m_k) \text{ is maximum for } k = i \qquad (5.56)$$

73/26

Figure 5.8

Illustrating the partitioning of the observation space into decision regions for the case when N 2 and M 4; it is assumed that the M transmitted symbols are equally likely.



74/26

For the AWGN channel..

- Based on the log-likelihood function, of the AWGN channel, $l(m_k)$ will be max when the term $\sum_{j=1}^{N} (x_j - s_{kj})^2$ is minimized by k = i.
- Decision rule for AWGN:

  Observation vector x lies region Z_i if

  $$\sum_{j=1}^{N} (x_j - s_{kj})^2, \text{ is minimum for } k = i \qquad (5.57)$$

- Or using Euclidian space notation

  Observation vector x lies region Z_i if

  the Euclidean distance $\|x - s_k\|$

  is minimum for $k = i$ (5.59)

Finally,

- (5.59) states that the maximum likelihood decision rule is simply to choose the message point closest to the received signal point.
- After few re-organizations we get: (*left as homework brain gymnastic exercise for you*)

Observation vector x lies in region Z_i if

$$\sum_{j=1}^{N} x_j s_{kj} - \frac{1}{2} E_k \text{ is maximum for } k = i \qquad (5.61)$$

$$E_k = \sum_{j=1}^{N} s_{kj}^2 \qquad (5.62)$$

Energy of the transmitted signal $s_k(t)$

76/26

Correlation Receiver

From the material presented in the previous sections, we find that for an AWGN channel and for the case when the transmitted signals $s_1(t), s_2, \ldots, s_M(t)$ are equally likely, the optimum receiver consists of two subsystems, which are detailed in Figure 5.9 and described here:

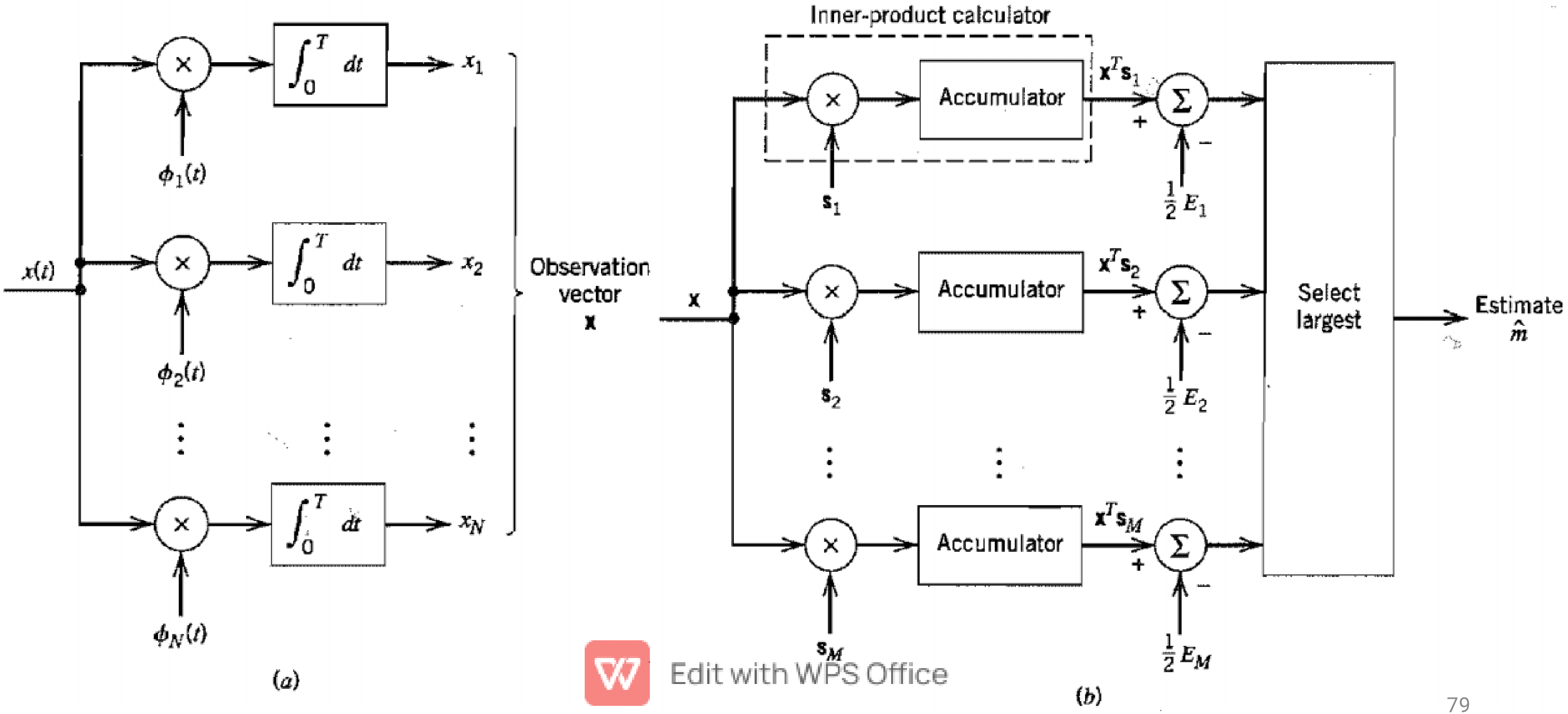
1. The *detector* part of the receiver is shown in Figure 5.9*a*. It consists of a bank of M *product-integrators* or *correlators*, supplied with a corresponding set of coherent reference signals or orthonormal basis functions $\phi_1(t), \phi_2(t), \ldots, \phi_N(t)$ that are generated locally. This bank of correlators operates on the received signal $x(t)$, $0 \le t \le T$, to produce the observation vector $\mathbf{x}$.

Correlation Receiver

2. The second part of the receiver, namely, the *signal transmission decoder* is shown in Figure 5.9*b*. It is implemented in the form of a maximum-likelihood decoder that operates on the observation vector $\mathbf{x}$ to produce an estimate, $\hat{m}$, of the transmitted symbol m_i, $i = 1, 2, \ldots, M$, in a way that would minimize the average probability of symbol error. In accordance with Equation (5.61), the N elements of the observation vector $\mathbf{x}$ are first multiplied by the corresponding N elements of each of the M signal vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M$, and the resulting products are successively summed in accumulators to form the corresponding set of inner products $\{\mathbf{x}^T\mathbf{s}_k \mid k = 1, 2, \ldots, M\}$. Next, the inner products are corrected for the fact that the transmitted signal energies may be unequal. Finally, the largest in the resulting set of numbers is selected, and an appropriate decision on the transmitted message is made.

The optimum receiver of Figure 5.9 is commonly referred to as a *correlation receiver*.

Correlation Receiver



5.7 Probability of Error

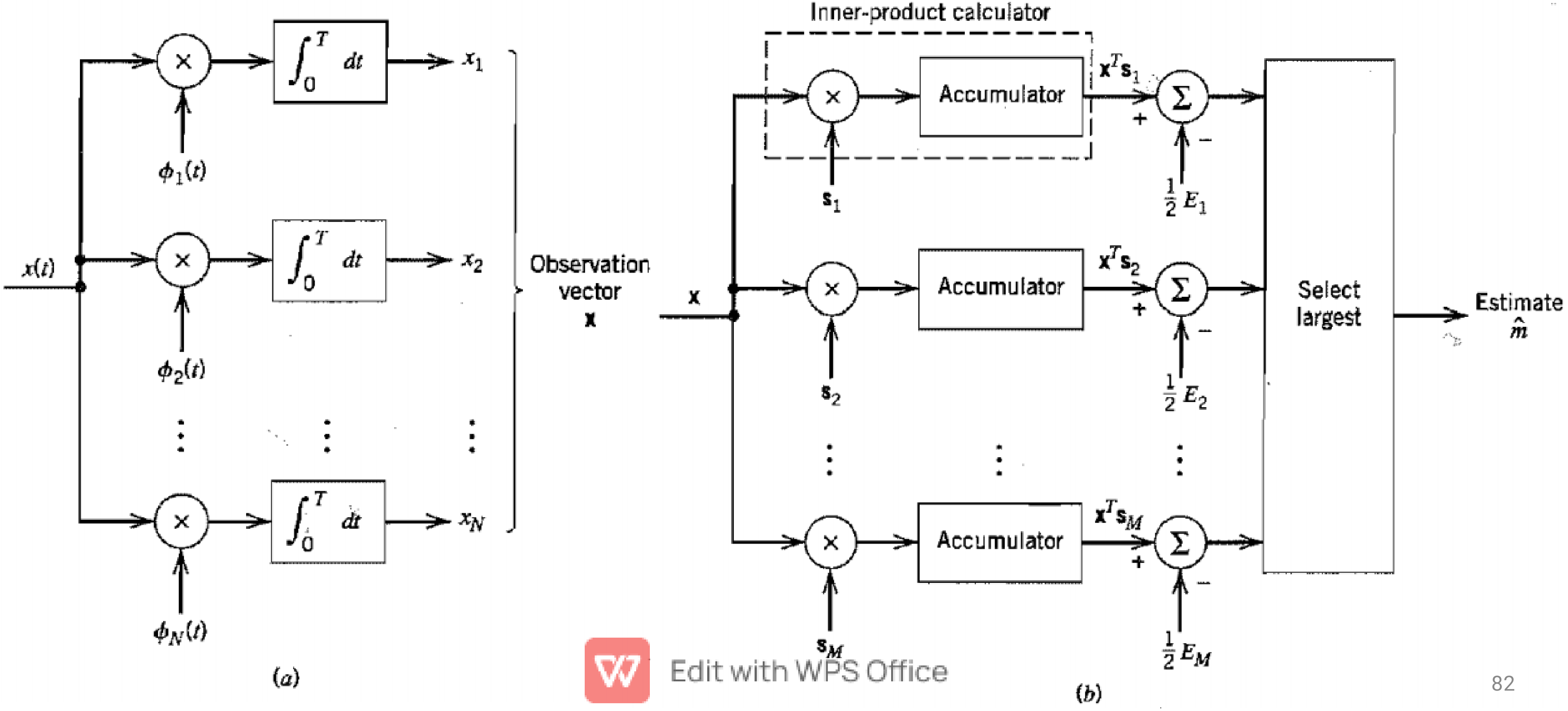
- To complete the statistical characterization of the correlation receiver (Fig. 5.9) we need to discuss its *noise performance*.
- Using the assumptions made before, we can define the average probability of error P_e as:

$$
\begin{aligned}
P_e &= \sum_{i=1}^{M} p_i \, P(\mathbf{x} \text{ does not lie in } Z_i \,|\, m_i \text{ sent}) \\
&= \frac{1}{M} \sum_{i=1}^{M} P(\mathbf{x} \text{ does not lie in } Z_i \,|\, m_i \text{ sent}) \\
&= 1 - \frac{1}{M} \sum_{i=1}^{M} P(\mathbf{x} \text{ lies in } Z_i \,|\, m_i \text{ sent})
\end{aligned}
$$

- Using the likelihood function this can be re-written as:

$$P_e = 1 - \frac{1}{M} \sum_{i=1}^{M} \int_{Z_i} f_x(x \mid m_i)\, dx \qquad (5.68)$$

Correlation Receiver



3.8 MATCHED FILTER RECEIVER

Since each of the orthonormal basis functions $\phi_1(t), \phi_2(t), \ldots, \phi_N(t)$ is assumed to be zero outside the interval $0 \leqslant t \leqslant T$, the use of multipliers shown in Fig. 3.10*a* may be avoided. This is desirable because analog multipliers are usually hard to build. Consider, for example, a linear filter with impulse response $h_j(t)$. With the received signal $x(t)$ used as the filter input, the resulting filter output, $y_j(t)$, is defined by the convolution integral:

$$y_j(t) = \int_{-\infty}^{\infty} x(\tau)h_j(t - \tau)d\tau \tag{3.76}$$

Suppose we now set the impulse response

$$h_j(t) = \phi_j(T - t) \tag{3.77}$$

Then the resulting filter output is

$$y_j(t) = \int_{-\infty}^{\infty} x(\tau)\phi_j(T - t + \tau)d\tau \tag{3.78}$$

M

Sampling this output at time $t = T$, we get

$$y_j(T) = \int_{-\infty}^{\infty} x(\tau)\phi_j(\tau)d\tau \tag{3.79}$$

and since $\phi_j(T)$ is zero outside the interval $0 \leqslant t \leqslant T$, we finally get

$$y_j(T) = \int_0^T x(\tau)\phi_j(\tau)d\tau \tag{3.80}$$

We note that $y_j(T) = x_j$, where x_j is the jth correlator output produced by the received signal $x(t)$ in Fig. 3.10*a*. Thus the detector part of the optimum receiver may also be implemented as in Fig. 3.11.

A filter whose impulse response is a time-reversed and delayed version of some signal $\phi_j(t)$, as in Eq. 3.77, is said to be *matched* to $\phi_j(t)$. Correspondingly, the optimum receiver based on the detector of Fig. 3.11 is referred to as the *matched filter receiver*.*

For a matched filter operating in real time to be physically realizable, it must be causal. That is to say, its impulse response must be zero for negative time, as shown by

$$h_j(t) = 0 \qquad t < 0$$

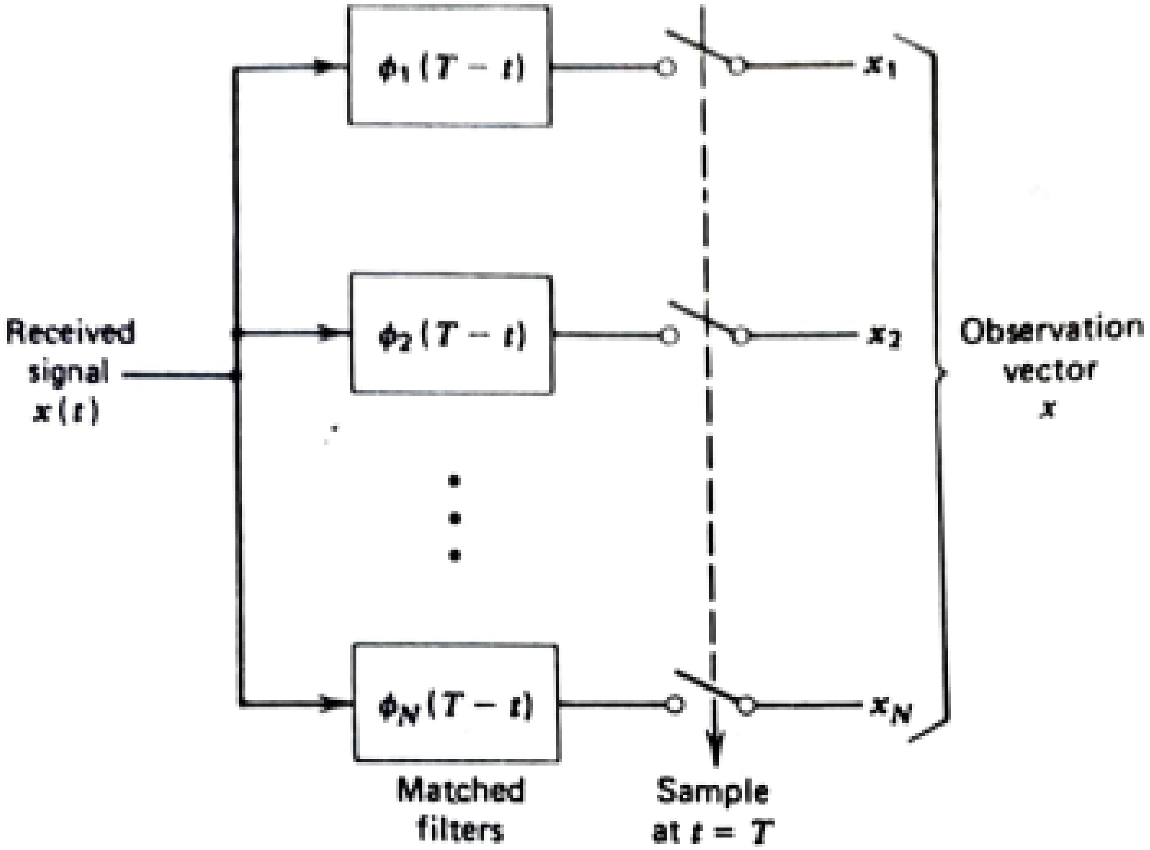


Figure 3.11 Detector part of matched filter receiver; the vector receiver part is as shown in Fig. 3.10*b*.

With $h_j(t)$ defined in terms of $\phi_j(t)$ as in Eq. 3.77, we see that the causality condition is satisfied provided that the signal $\phi_j(t)$ is zero outside the interval $0 \leq t \leq T$.

(1) Maximization of Output Signal-to-noise Ratio

We may gain further insight into the operation of a matched filter by using output signal-to-noise ratio as the optimality criterion for deriving the matched filter from first principles.

Consider then a linear filter of impulse response $h(t)$, with an input that consists of a known signal, $\phi(t)$, and an additive noise component, $w(t)$, as shown in Fig. 3.12. We may thus write

$$x(t) = \phi(t) + w(t) \qquad 0 \leqslant t \leqslant T \tag{3.81}$$

where T is the observation instant. In particular, we may choose $\phi(t)$ to be one of the orthonormal basis functions. The $w(t)$ is the sample function of a white Gaussian noise process of zero mean and power spectral density $N_0/2$. Since the filter is linear, the resulting output, $y(t)$, may be expressed as

$$y(t) = \phi_o(t) + n(t) \tag{3.82}$$

where $\phi_o(t)$ and $n(t)$ are produced by the signal and noise components of the input $x(t)$, respectively. A simple way of describing the requirement that the

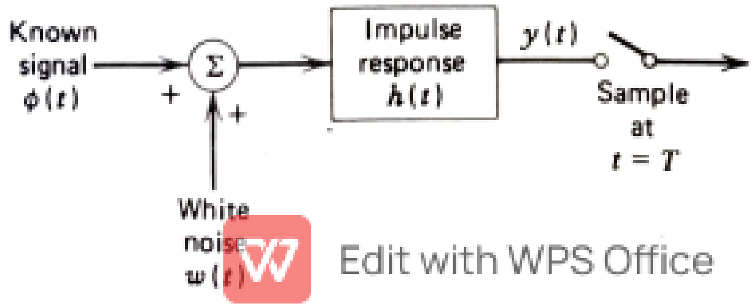


Figure 3.12 Illustrating the condition for derivation of the matched filter.

Matched Filter

output signal component $\phi_o(t)$ be considerably greater than the output noise component $n(t)$ is to have the filter make the instantaneous power in the output signal $\phi_o(t)$, measured at time $t = T$, as large as possible compared with the average power of the output noise $n(t)$. This is equivalent to maximizing the *output signal-to-noise ratio* defined as

$$(\mathrm{SNR})_O = \frac{|\phi_o(T)|^2}{E[n^2(t)]} \tag{3.83}$$

We now show that this maximization occurs when the filter is matched to the known signal $\phi(t)$ at the input.

Let $\Phi(f)$ denote the Fourier transform of the known signal $\phi(t)$, and $H(f)$ denote the transfer function of the filter. Then the Fourier transform of the output signal $\phi_o(t)$ is equal to $H(f)\Phi(f)$, and $\phi_o(t)$ is itself given by the inverse Fourier transform

$$\phi_o(t) = \int_{-\infty}^{\infty} H(f)\Phi(f)\exp(j2\pi ft)\,df \tag{3.84}$$

Matched Filter

Hence, when the filter output is sampled at time $t = T$, we may write

$$|\phi_o(T)|^2 = \left| \int_{-\infty}^{\infty} H(f)\Phi(f)\exp(j2\pi fT)df \right|^2 \tag{3.85}$$

Consider next the effect of the noise $w(t)$ alone on the filter output. The power spectral density $S_N(f)$ of the output noise $n(t)$ is equal to the power spectral density of the input noise $w(t)$ times the squared magnitude of the transfer function $H(f)$. Since $w(t)$ is drawn from a process that is white with constant power spectral density $N_0/2$, it follows that

$$S_N(f) = \frac{N_0}{2}|H(f)|^2 \tag{3.86}$$

The average power of the output noise $n(t)$ is therefore

$$E[n^2(t)] = \int_{-\infty}^{\infty} S_N(f)df$$

$$= \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2 df \tag{3.87}$$

Matched Filter

Thus, substituting Eqs. 3.85 and 3.87 into Eq. 3.83, we may rewrite the expression for the output signal-to-noise ratio as

$$(\mathrm{SNR})_O = \frac{\left| \int_{-\infty}^{\infty} H(f)\Phi(f)\exp(j2\pi f T)\,df \right|^2}{\frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2\,df} \tag{3.88}$$

Our problem is to find, while holding the Fourier transform $\Phi(f)$ of the input signal fixed, the form of the transfer function $H(f)$ of the filter that makes $(\mathrm{SNR})_O$ a maximum. To find the solution to this optimization problem, we

Matched Filter

apply a mathematical result known as Schwarz's* inequality to the numerator of Eq. 3.88. Accordingly, we may write

$$\left| \int_{-\infty}^{\infty} H(f)\Phi(f)\exp(j2\pi fT)df \right|^2 \leq \int_{-\infty}^{\infty} |H(f)|^2 df \int_{-\infty}^{\infty} |\Phi(f)|^2 df \qquad (3.89)$$

Using this relation in Eq. 3.88, we may simplify the output signal-to-noise ratio as

$$(\mathrm{SNR})_O \leq \frac{2}{N_0} \int_{-\infty}^{\infty} |\Phi(f)|^2 df \qquad (3.90)$$

The right side of Eq. 3.90 is uniquely defined by two quantities:

1. The signal energy given by (in accordance with Rayleigh's energy theorem)

$$\int_{-\infty}^{\infty} |\phi(t)|^2 \, dt = \int_{-\infty}^{\infty} |\Phi(f)|^2 \, dt$$

2. The noise power spectral density $N_0/2$.

Matched Filter

As such, the right side of Eq. 3.90 does not depend on the transfer function $H(f)$. Consequently, the output signal-to-noise ratio will be a maximum when $H(f)$ is chosen so that the equality holds; that is

$$(\mathrm{SNR})_{O,\max} = \frac{2}{N_0}\int_{-\infty}^{\infty} |\Phi(f)|^2 df \qquad (3.91)$$

For this condition, $H(f)$ assumes its optimum value denoted as $H_{opt}(f)$. From Schwarz's inequality, we also find that, except for a scaling factor, the optimum value of this transfer function is defined by

$$H_{opt}(f) = \Phi^*(f)\exp(-j2\pi fT) \qquad (3.92)$$

where $\Phi^*(f)$ is the complex conjugate of the Fourier transform of the input signal $\phi(t)$. This relation states that, except for the necessary time delay factor $\exp(-j2\pi fT)$, the transfer function of the optimum filter is the same as the complex conjugate of the spectrum of the input signal.

Matched Filter

Equation 3.92 specifies the matched filter in the frequency domain. To characterize it in the time domain, we take the inverse Fourier transform of $H_{opt}(f)$ in Eq. 3.92 to obtain the impulse response of the matched filter as

$$h_{opt}(t) = \int_{-\infty}^{\infty} \Phi^*(f)\exp[-j2\pi f(T - t)]df$$

Since for a real-valued signal $\phi(t)$ we have $\Phi^*(f) = \Phi(-f)$, we may also write

$$\begin{aligned} h_{opt}(t) &= \int_{-\infty}^{\infty} \Phi(-f)\exp[-j2\pi f(T - t)]df \\ &= \phi(T - t) \end{aligned} \qquad (3.93)$$

Equation 3.93 shows that the impulse response of the optimum filter is a time-reversed and delayed version of the input signal $\phi(t)$; that is, it is matched to the input signal. Note that the only assumptions we have made about the input noise $w(t)$ are that it is additive, stationary, and white with zero mean and power spectral density $N_0/2$.

The result given in Eq. 3.93 was derived by maximizing the output signal-to-noise ratio of the receiver. This result is identical, except for a minor change in notation, to that given in Eq. 3.77, which was derived earlier by minimizing the average probability of symbol error at the receiver output. The two derivations, however, were performed under somewhat different sets of conditions. Collecting these two sets of conditions together, we may now make the following important statement concerning these two criteria for optimum receiver design. Specifically, maximization of the output signal-to-noise ratio is equivalent to minimization of the average probability of symbol error under two assumptions:

1. The additive white noise at the receiver input is stationary with Gaussian statistics.
2. The a priori probabilities of the transmitted signals are known.

EXAMPLE 3 MATCHED FILTER FOR RF PULSE

Consider a rectangular RF pulse of duration T seconds and unit energy, as shown by (see Fig. 3.13*a*)

$$\phi(t) = \begin{cases} \sqrt{\dfrac{2}{T}} \cos(2\pi f_c t) & 0 \leqslant t \leqslant T \\ 0 & \text{elsewhere} \end{cases} \tag{3.94}$$

where f_c is a large integral multiple of $1/T$. The impulse response of a filter matched to $\phi(t)$ is therefore

$$\begin{aligned} h_{opt}(t) &= \phi(T - t) \\ &= \begin{cases} \sqrt{\dfrac{2}{T}} \cos(2\pi f_c t) & 0 \leqslant t \leqslant T \\ 0 & \text{elsewhere} \end{cases} \end{aligned} \tag{3.95}$$

which, in this example, works out to be the same as the signal $\phi(t)$ itself. The corresponding filter output, $\phi_o(t)$, is determined by convolving $h_{opt}(t)$ with $\phi(t)$; we thus obtain

$$\phi_o(t) = \begin{cases} (t/T) \cos(2\pi f_c t) & 0 \leqslant t \leqslant T \\ (2 - t/T) \cos(2\pi f_c t) & T \leqslant t \leqslant 2T \\ 0 & \text{elsewhere} \end{cases} \tag{3.96}$$

which is shown sketched in Fig. 3.13*b*. As expected, the filter output attains its maximum value at time $t = T$.

When a unit impulse of current is applied to the idealized parallel tuned circuit shown in Fig. 3.14*a*, the resulting voltage response is equal to the impulse response of the circui*, as shown by

$$h(t) = \frac{1}{C} \cos\left(\frac{t}{\sqrt{LC}}\right) \qquad 0 \leqslant t \leqslant \infty \tag{3.97}$$

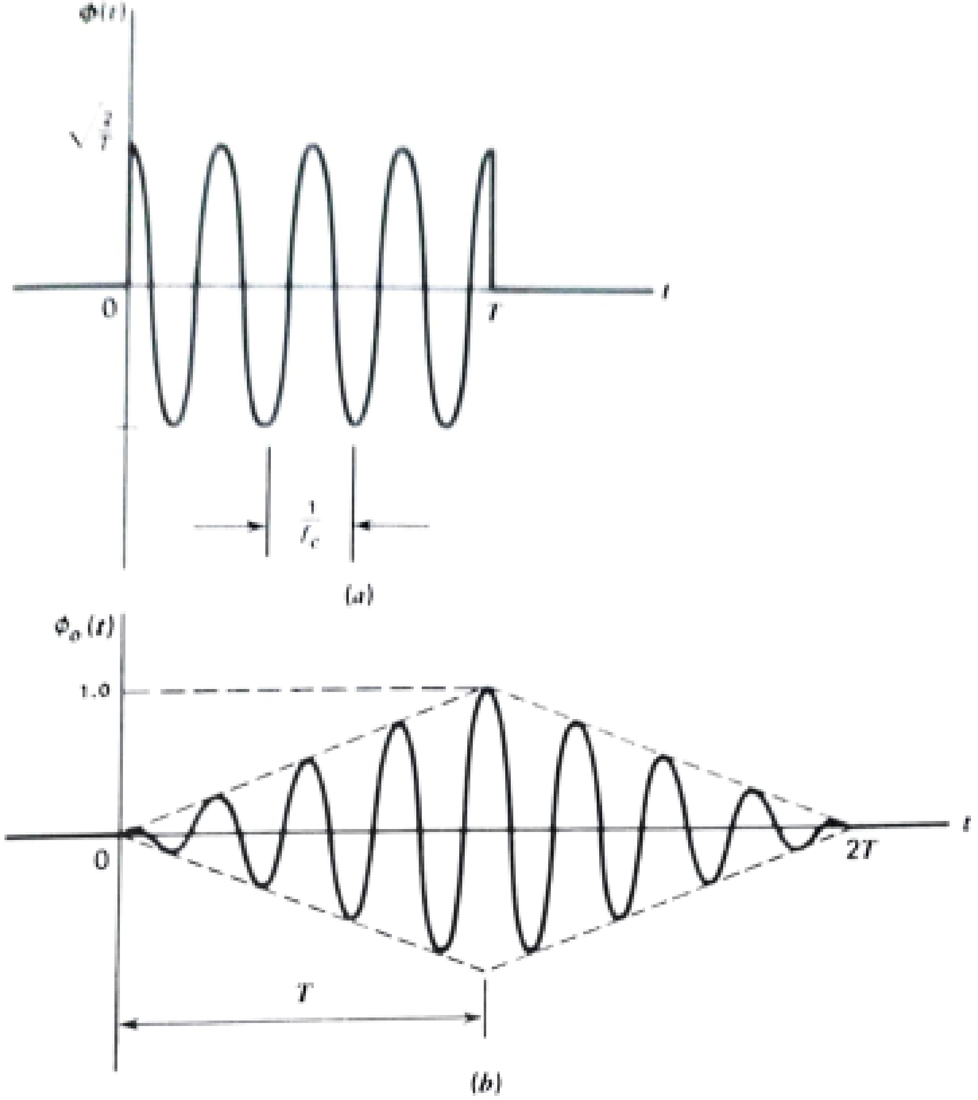
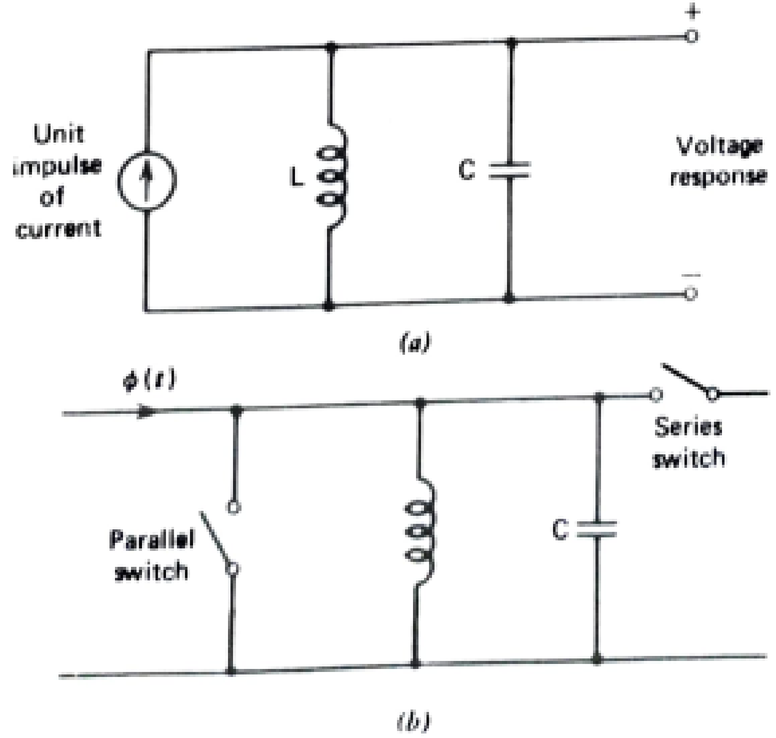


Figure 3.13 (*a*) RF pulse input. (*b*) Matched filter output.



(*a*) Parallel tuned LC circuit. (*b*) Integrate-and-dump circuit.

where it is assumed that the initial energy in both L and C at time $t = 0$ is zero. Suppose we choose the circuit parameters L and C such that

$$\frac{1}{\sqrt{LC}} = 2\pi f_c$$

and

$$\frac{1}{C} = \sqrt{\frac{2}{T}}$$

We then find that the response $h(t)$ of the tuned circuit coincides with $h_{opt}(t)$ over the interval $0 \leq t \leq T$. However, this is not so for $t > T$. Thus, the matched filtering operation for the RF pulse $\phi(t)$ of Fig. 3.13*a* may be implemented by using the modification shown in Fig. 3.14*b*. The parallel switch closes briefly at time $t = 0$, dumping any residual energy in the filter. This ensures that signal energy received before $t = 0$ does not contribute to the output at time $t = T$. The series switch then closes briefly at time $t = T$, thereby sampling the filter output at the right time. This form of a matched filter is called an *integrate-and-dump filter*. Note that such a filter is not time-invariant; however, it does exhibit the desired impulse response as long as the timing of the two switches is properly synchronized with respect to the input RF pulse $\phi(t)$.

Properties of Matched Filter

PROPERTY 1

The spectrum of the output signal of a matched filter with the matched signal as input is, except for a time delay factor, proportional to the energy spectral density of the input signal.

Let $\Phi_o(f)$ denote the Fourier transform of the filter output $\phi_o(t)$. Then

$$\begin{aligned}\Phi_o(f) &= H_{opt}(f)\Phi(f) \\ &= \Phi^*(f)\Phi(f)\exp(-j2\pi fT) \\ &= |\Phi(f)|^2 \exp(-j2\pi fT)\end{aligned} \qquad (3.100)$$

which is the desired result.

PROPERTY 2

The output signal of a matched filter is proportional to a shifted version of the autocorrelation function of the input signal to which the filter is matched.

This property follows directly from Property 1, recognizing that the autocorrelation function and energy spectral density of a signal form a Fourier transform pair. Thus, taking the inverse Fourier transform of Eq. 3.100, we may express the matched-filter output as

$$\phi_o(t) = R_\phi(t - T) \tag{3.101}$$

where $R_\phi(\tau)$ is the autocorrelation function of the input $\phi(t)$ for lag τ. Equation 3.101 is the desired result. Note that at time $t = T$, we have

$$\phi_o(T) = R_\phi(0) = E \tag{3.102}$$

where E is the signal energy. That is, in the absence of noise, the maximum value of the matched-filter output, attained at time $t = T$, is proportional to the signal energy.

PROPERTY 3

The output signal-to-noise ratio of a matched filter depends only on the ratio of the signal energy to the power spectral density of the white noise at the filter input.

To demonstrate this property, consider a filter matched to an input signal $\phi(t)$. From Property 2, the maximum value of the filter output, at time $t = T$, is proportional to the signal energy E. Substituting Eq. 3.92 in Eq. 3.87 gives the average output noise power as

$$E[n^2(t)] = \frac{N_0}{2} \int_{-\infty}^{\infty} |\Phi(f)|^2 df = \frac{N_0}{2} E \tag{3.103}$$

where we have made use of Rayleigh's energy theorem. Therefore, the output signal-to-noise ratio has the maximum value

$$(\text{SNR})_{O,\max} = \frac{E^2}{N_0 E/2} = \frac{2E}{N_0} \tag{3.104}$$

Properties of Matched Filter

This result is perhaps the most important parameter in the calculation of the performance of signal processing systems using matched filters. From Eq.

3.104 we see that dependence on the waveform of the input $\phi(t)$ has been completely removed by the matched filter. Accordingly, in evaluating the ability of a matched-filter receiver to combat additive white Gaussian noise, we find that all signals that have the same energy are equally effective. Note that the signal energy E is in joules and the noise spectral density $N_0/2$ is in watts per hertz, so that the ratio $2E/N_0$ is dimensionless; however, the two quantities have different physical meaning. We refer to E/N_0 as the *signal energy-to-noise density ratio*.

PROPERTY 4

The matched-filtering operation may be separated into two matching conditions; namely, spectral phase matching that produces the desired output peak at time T, and spectral amplitude matching that gives this peak value its optimum signal-to-noise density ratio.

In polar form, the spectrum of the signal $\phi(t)$ being matched may be expressed as

$$\Phi(f) = |\Phi(f)|\exp[j\theta(f)]$$

where $|\Phi(f)|$ is the amplitude spectrum and $\theta(f)$ is the phase spectrum of the signal. The filter is said to be *spectral phase matched* to the signal $\phi(t)$ if the transfer function of the filter is defined by*

$$H(f) = |H(f)|\exp[-j\theta(f) - j2\pi f T]$$

where $|H(f)|$ is real and nonnegative and T is a positive constant. The output of such a filter is

$$\begin{aligned}\phi_o'(t) &= \int_{-\infty}^{\infty} H(f)\Phi(f)\exp(j2\pi f t)df \\ &= \int_{-\infty}^{\infty} |H(f)||\Phi(f)|\exp[j2\pi f(t - T)]df\end{aligned}$$

Properties of Matched Filter

where the product $|H(f)||\Phi(f)|$ is real and nonnegative. The spectral phase matching ensures that all the spectral components of the output $\phi_o'(t)$ add constructively at time $t = T$, thereby causing the output to attain its maximum value, as shown by

$$\phi_o'(t) \leqslant \phi_o'(T) = \int_{-\infty}^{\infty} |\Phi(f)||H(f)|\,df$$

For *spectral amplitude matching*, we choose the amplitude response $|H(f)|$ of the filter to shape the output for best signal-to-noise ratio at $t = T$ by using

$$|H(f)| = |\Phi(f)|$$

and the standard matched filter is the result.

Equalization

- Equalizer is a system which is designed to have a frequency response that is the inverse of the frequency response of the channel, so that when it is kept in cascade with the channel, the overall frequency response will be flat and the signal distortion caused by the channel is eliminated
- Incase of digital communications, what is important is that during each time slot, the receiver should be able to decide correctly whether what was transmitted during that time slot was a binary 1 or a binary 0.

The Zero-forcing Equalizer

- In this approach to the optimization of the receiver , the channel is totally ignored and the receiver is viewed as zero forcing equalizer followed by the decision making device.
- The equalizer is used to force the inter symbol interference to be zero at all sampling instants t=kT at which the channel output taken through the equalizer is sampled except for k=0 for which the distorted but desired pulse occurs.
- A transversal (tapped delay) equalizer with (2N+1) taps is used and the samples of the equalizer output $p_{eq}(t)$are taken at regular intervals of T

The Zero-forcing Equalizer

- The tap gains or weights of the equalizer $c_i, i = -N \ to \ +N$ are so chosen that

$$p_{eq}(t_k) = \begin{cases} 1 & \text{for } k = 0 \\ 0 & \text{for } k = \pm 1, \pm 2, \pm 3, \ldots, \pm N \end{cases} \tag{10.50}$$

This means that we are forcing N-zero values to exist on each side of the peak of $p_{eq}(t)$, where $p_{eq}(t)$ is the output pulse from the equalizer and $p_{eq}(t_k)$ is the kth sample of $p_{eq}(t)$, taken at $t = kT$. It should be noted that it is required that $p_{eq}(t)$, the output pulse satisfies the Nyquist criterion or, the controlled ISI criterion as the case may be.

If, as shown in Fig. 10.17, $p_r(t)$ is the output pulse from the channel which is given as input to the zero-forcing equalizer, then, since the equalizer output pulse is the sum of all the delayed pulses from the outputs of the various delay elements of the equalizer, we have

$$p_{eq}(t) = \sum_{n=-N}^{N} c_n p_r(t - nT - NT) \tag{10.51}$$

The Zero-forcing Equalizer

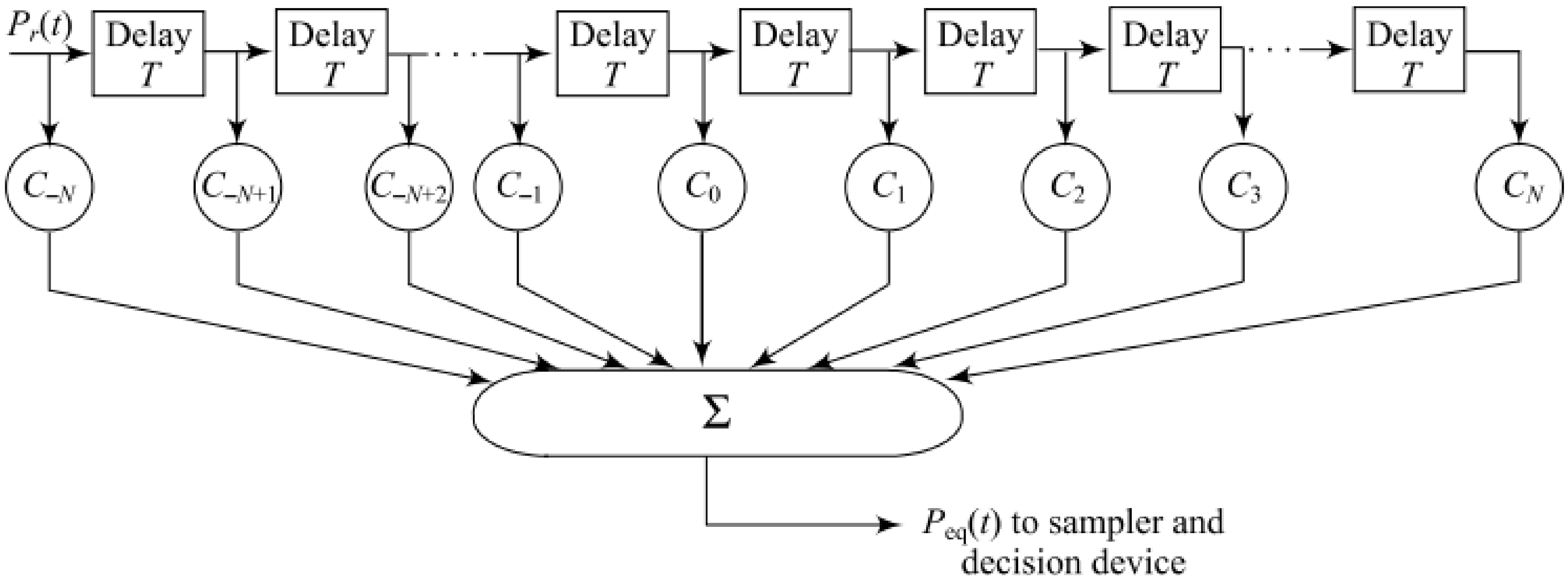


Fig. 10.17 *Zero-forcing equalizer*

The Zero-forcing Equalizer

The $-NT$ term in the argument of $p_r(\cdot)$ in Eq. (10.51), represents a constant time delay which we may ignore during the analysis and re-introduce at the end.

$$\therefore \quad p_{eq}(t) = \sum_{n=-N}^{N} c_n p_r(t - nT) \tag{10.52}$$

Now, this $p_{eq}(t)$ is sampled at $t = kT$, $k = 0, \pm1, \pm2, \ldots$. Hence, at $t = kT$, we have

$$p_{eq}(kT) = \sum_{n=-N}^{N} c_n p_r(kT - nT); \quad k = 0, \pm1, \pm2, \ldots \tag{10.53}$$

For convenience in notation, let us drop T for the time being, so that

$$p_{eq}(k) = \sum_{n=-N}^{N} c_n p_r(k - n); \quad k = 0, \pm1, \pm2, \ldots \tag{10.54}$$

Equation (10.54) represents a set of an infinite number of simultaneous equations with only $(2N + 1)$ variables, viz., c_ns, $n = -N$ to N, which we have to determine with the constraint that

$$p_{eq}(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k = \pm1, \pm2, \pm3, \ldots, \pm N \end{cases} \tag{10.55}$$

The Zero-forcing Equalizer

It is not possible to solve this set of an infinite number of equations. However, since $p_{eq}(t)$ satisfies the Nyquist criterion, or the controlled ISI criterion, the pre-cursor and post-cursor amplitudes of $p_{eq}(t)$ decay rather rapidly. So we can as well modify the constraint (Eq. (10.55)) and limit the number of samples on either side of $k = 0$, to N values only. Thus, we are implicitly assuming that the ISI is limited only to N preceding and N succeeding values – an assumption that is quite justifiable if $N > 2$. Thus, we rewrite the constraint equation in Eq. (10.55) as

$$p_{eq}(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k = \pm 1, \pm 2, \ldots, \pm N \end{cases} \tag{10.56}$$

Once this modified constraint condition is imposed on the samples of $p_{eq}(t)$, as shown in Eq. (10.54), we get only $(2N + 1)$ simultaneous linear equations involving $(2N + 1)$ variables, viz., the c_ns. We can write these $(2N + 1)$ equations in matrix form as follows.

The Zero-forcing Equalizer

$$\begin{bmatrix} p_r(0) & p_r(-1) & \cdot & \cdot & \cdot & p_r(-2N) \\ p_r(1) & p_r(0) & \cdot & \cdot & \cdot & p_r(-2N+1) \\ \cdots\cdots & \cdots\cdots\cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ p_r(N-1) & p_r(N-2) & \cdot & \cdot & \cdot & p_r(-N-1) \\ p_r(N) & p_r(N-1) & \cdot & \cdot & \cdot & p_r(-N) \\ p_r(N+1) & p_r(N) & \cdot & \cdot & \cdot & p_r(-N+1) \\ \cdots\cdots & \cdots\cdots\cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ p_r(2N-1) & p_r(2N-2) & \cdot & \cdot & \cdot & p_r(1) \\ p_r(2N) & p_r(2N-1) & \cdot & \cdot & \cdot & p_r(0) \end{bmatrix} \begin{bmatrix} c_{-N} \\ c_{-N+1} \\ \cdots \\ c_{-1} \\ c_0 \\ c_1 \\ \cdots \\ c_{N-1} \\ c_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 1 \\ 0 \\ \cdots \\ 0 \\ 0 \end{bmatrix} \tag{10.57}$$

The zero-forcing equalizer's tap gains, or weights are obtained by solving this matrix equation for c_ns.

The Zero-forcing Equalizer

Remarks

(i) *The constant time delay, NT, in Eq. (10.51) which was ignored in the rest of the analysis has been re-introduced for drawing Fig.10.18(b).*

(ii) *As stated earlier, the zero-forcing approach totally ignores the channel noise. This, in fact, leads to an overall performance degradation owing to 'noise enhancement'.*

(iii) *Determination of zero-forcing tap-gains, i.e., c_ns using Eq. (10.57) requires that we know the $p_r(\cdot)$ values in the $(2N + 1) \times (2N + 1)$ square matrix. In the case of changes in channel characteristics, as happens in switched telephone links and slowly changing radio links, an a priori knowledge of $p_r(\cdot)$ values is not possible. In such cases, the tap gains are directly adjusted on-line by using a training sequence that is transmitted before the actual message sequence is transmitted.*

The Zero-forcing Equalizer

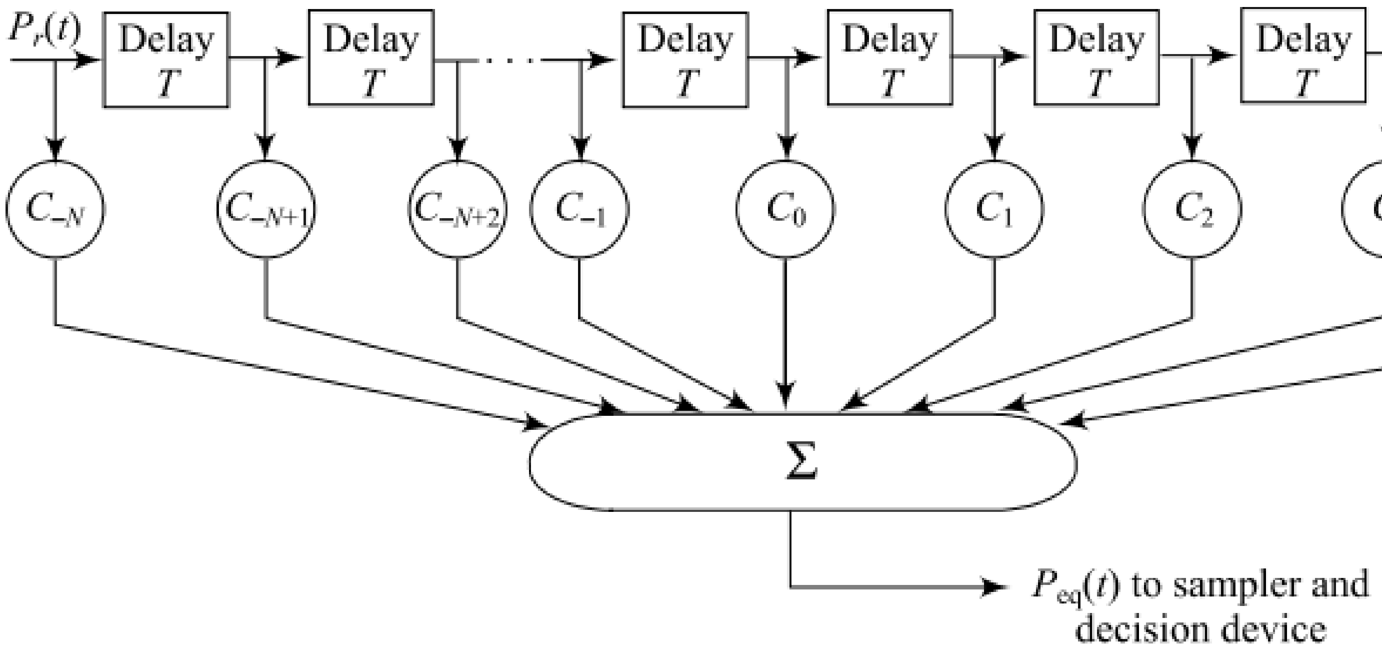


Fig. 10.17 *Zero-forcing equalizer*

The Zero-forcing Equalizer

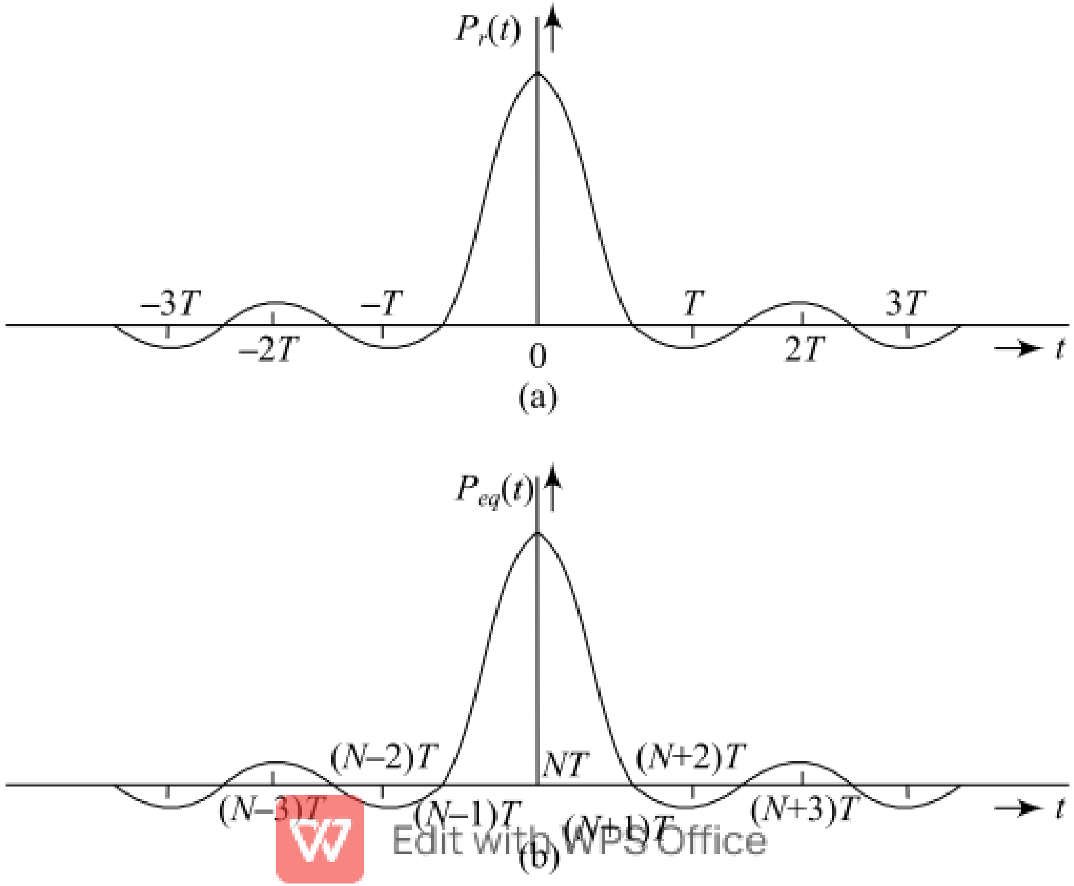


Fig. 10.18 *The input pulse $p_r(t)$ and the output pulse $p_{eq}(t)$*

The Zero-forcing Equalizer

Example 10.6 A zero-forcing equalizer is to be designed using 3 taps. Assume that the input pulse $p_r(t)$ to the equalizer is as shown in Fig. 10.18(a) in which $p_r(-2T) = 0.08$, $p_r(-T) = -0.25$, $p_r(0) = 1$, $p_r(T) = -0.20$ and $p_r(2T) = 0.10$

Solution Here, $2N + 1 = 3 \therefore N = 1$. Substituting the given values of $p_r(kT)$ in Eq. (10.57), we get

$$\begin{bmatrix} p_r(0) & p_r(-1) & p_r(-2) \\ p_r(1) & p_r(0) & p_r(-1) \\ p_r(2) & p_r(1) & p_r(0) \end{bmatrix} = \begin{bmatrix} 1 & -0.25 & 0.08 \\ -0.20 & 1 & -0.25 \\ 0.10 & -0.20 & 1 \end{bmatrix}$$

Hence, for 'zero forcing', the tap-gains c_ns must satisfy the following equation (Eq. (10.57)):

$$\begin{bmatrix} 1 & -0.25 & 0.08 \\ -0.20 & 1 & -0.25 \\ 0.10 & -0.20 & 1 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Solving the above set of three linear equations for c_{-1}, c_0 and c_1, we get

$$c_{-1} = 0.90825; \quad c_0 = 3.4386; \quad c_1 = -0.6075$$

Error rate due to noise

In Section 3.8 we presented a qualitative discussion of the effect of channel noise on the performance of a binary PCM system. Now that we are equipped with the matched filter as the optimum detector of a known pulse in additive white noise, we are ready to derive a formula for the error rate in such a system due to noise.

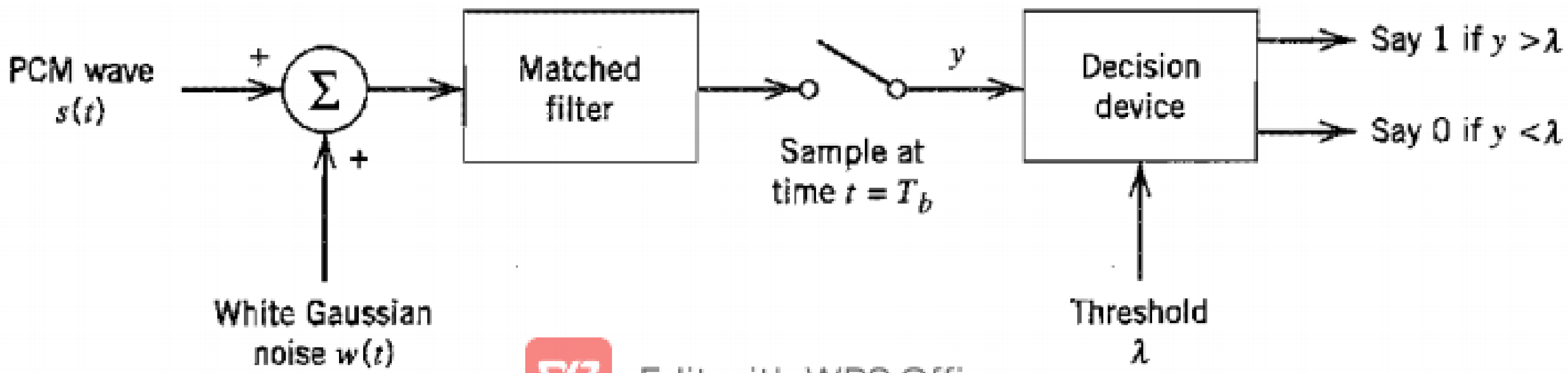
To proceed with the analysis, consider a binary PCM system based on *polar non-return-to-zero (NRZ) signaling*. In this form of signaling, symbols 1 and 0 are represented by positive and negative rectangular pulses of equal amplitude and equal duration. The channel noise is modeled as *additive white Gaussian noise* $w(t)$ of zero mean and power spectral density $N_0/2$; the Gaussian assumption is needed for later calculations. In the signaling interval $0 \leq t \leq T_b$, the received signal is thus written as follows:

$$x(t) = \begin{cases} +A + w(t), & \text{symbol 1 was sent} \\ -A + w(t), & \text{symbol 0 was sent} \end{cases} \tag{4.21}$$

Error rate due to noise

where T_b is the *bit duration*, and A is the *transmitted pulse amplitude*. It is assumed that the receiver has acquired knowledge of the starting and ending times of each transmitted pulse; in other words, the receiver has prior knowledge of the pulse shape, but not its polarity. Given the noisy signal $x(t)$, the receiver is required to make a decision in each signaling interval as to whether the transmitted symbol is a 1 or a 0.

The structure of the receiver used to perform this decision-making process is shown in Figure 4.4. It consists of a matched filter followed by a sampler, and then finally a



Error rate due to noise

decision device. The filter is matched to a rectangular pulse of amplitude A and duration T_b, exploiting the bit-timing information available to the receiver. The resulting matched filter output is sampled at the end of each signaling interval. The presence of channel noise $w(t)$ adds randomness to the matched filter output.

Let y denote the sample value obtained at the end of a signaling interval. The sample value y is compared to a preset *threshold* λ in the decision device. If the threshold is exceeded, the receiver makes a decision in favor of symbol 1; if not, a decision is made in favor of symbol 0. We adopt the convention that when the sample value y is exactly equal to the threshold λ, the receiver just makes a guess as to which symbol was transmitted; such a decision is the same as that obtained by flipping a fair coin, the outcome of which will not alter the average probability of error.

There are two possible kinds of error to be considered:

Error rate due to noise

There are two possible kinds of error to be considered:

1. Symbol 1 is chosen when a 0 was actually transmitted; we refer to this error as an *error of the first kind.*
2. Symbol 0 is chosen when a 1 was actually transmitted; we refer to this error as an *error of the second kind.*

To determine the average probability of error, we consider these two situations separately.

Suppose that symbol 0 was sent. Then, according to Equation (4.21), the received signal is

$$x(t) = -A + w(t), \qquad 0 \leq t \leq T_b \tag{4.22}$$

Correspondingly, the matched filter output, sampled at time $t = T_b$, is given by (in light of Example 4.1 with kAT_b set equal to unity for convenience of presentation)

$$\begin{aligned} y &= \int_0^{T_b} x(t)\, dt \\ &= -A + \frac{1}{T_b} \int_0^{T_b} w(t)\, dt \end{aligned} \tag{4.23}$$

which represents the sample value of a random variable Y. By virtue of the fact that the noise $w(t)$ is white and Gaussian, we may characterize the random variable Y as follows:

Error rate due to noise

- The random variable Y is Gaussian distributed with a mean of $-A$.
- The variance of the random variable Y is

$$
\begin{aligned}
\sigma_Y^2 &= E[(Y+A)^2] \\
&= \frac{1}{T_b^2} E\left[\int_0^{T_b}\int_0^{T_b} w(t)w(u)\,dt\,du\right] \\
&= \frac{1}{T_b^2}\int_0^{T_b}\int_0^{T_b} E[w(t)w(u)]\,dt\,du \\
&= \frac{1}{T_b^2}\int_0^{T_b}\int_0^{T_b} R_W(t,u)\,dt\,du
\end{aligned}
\qquad (4.24)
$$

where $R_W(t, u)$ is the autocorrelation function of the white noise $w(t)$. Since $w(t)$ is white with a power spectral density $N_0/2$, we have

$$R_W(t,u) = \frac{N_0}{2}\delta(t-u) \qquad (4.25)$$

Error rate due to noise

where $\delta(t-u)$ is a time-shifted delta function. Hence, substituting Equation (4.25) into (4.24) yields

$$\begin{aligned}\sigma_Y^2 &= \frac{1}{T_b^2}\int_0^{T_b}\int_0^{T_b}\frac{N_0}{2}\,\delta(t-u)\,dt\,du \\ &= \frac{N_0}{2T_b}\end{aligned} \tag{4.26}$$

where we have used the sifting property of the delta function and the fact that its area is unity. The conditional probability density function of the random variable Y, given that symbol 0 was sent, is therefore

$$f_Y(y|0) = \frac{1}{\sqrt{\pi N_0/T_b}}\exp\left(-\frac{(y+A)^2}{N_0/T_b}\right) \tag{4.27}$$

Error rate due to noise

This function is plotted in Figure 4.5(*a*). Let p_{10} denote the *conditional probability of error, given that symbol 0 was sent*. This probability is defined by the shaded area under the curve of $f_Y(y|0)$ from the threshold λ to infinity, which corresponds to the range of values assumed by y for a decision in favor of symbol 1. In the absence of noise, the matched filter output y sampled at time $t = T_b$ is equal to $-A$. When noise is present, y occasionally assumes a value greater than λ, in which case an error is made. The probability of this error, conditional on sending symbol 0, is defined by

$$\begin{aligned} p_{10} &= P(y > \lambda \,|\, \text{symbol 0 was sent}) \\ &= \int_{\lambda}^{\infty} f_Y(y|0)\, dy \qquad (4.28) \\ &= \frac{1}{\sqrt{\pi N_0/T_b}} \int_{\lambda}^{\infty} \exp\left(-\frac{(y+A)^2}{N_0/T_b}\right) dy \end{aligned}$$

Error rate due to noise

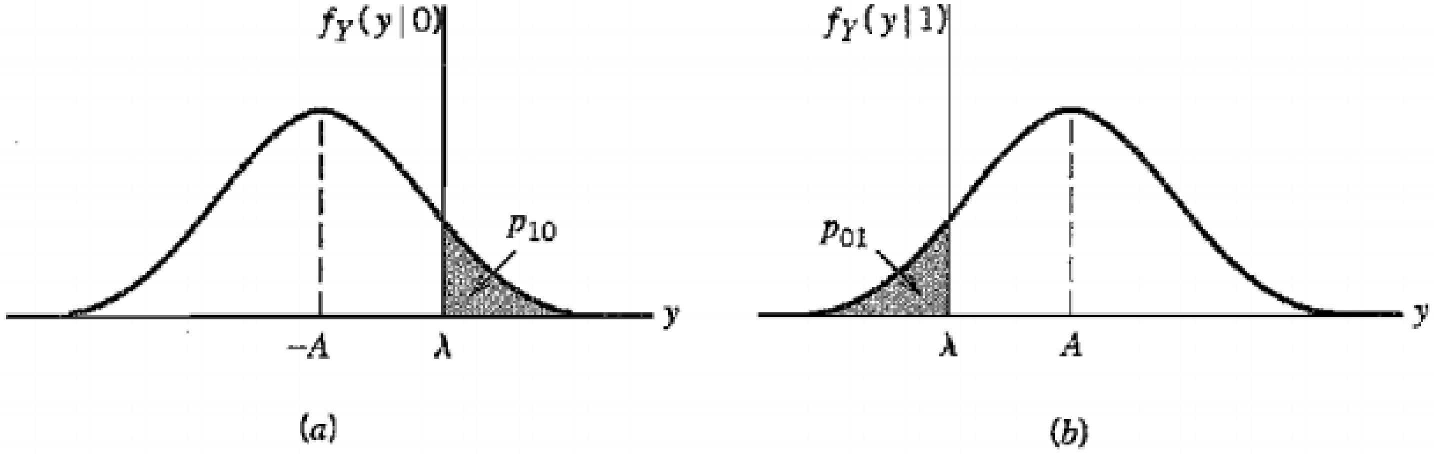


FIGURE 4.5 Noise analysis of PCM system. (*a*) Probability density function of random variable Y at matched filter output when 0 is transmitted. (*b*) Probability density function of Y when 1 is transmitted.

Error rate due to noise

To reformulate the conditional probability of error p_{10} in terms of the complementary error function, we first define a new variable

$$z = \frac{y + A}{\sqrt{N_0/T_b}}$$

Accordingly, we may rewrite Equation (4.28) in the compact form

$$\begin{aligned} p_{10} &= \frac{1}{\sqrt{\pi}} \int_{(A+\lambda)/\sqrt{N_0/T_b}}^{\infty} \exp(-z^2)\, dz \\ &= \frac{1}{2} \operatorname{erfc}\left(\frac{A + \lambda}{\sqrt{N_0/T_b}}\right) \end{aligned} \tag{4.31}$$

Assume next that symbol 1 was transmitted. This time the Gaussian random variable Y represented by the sample value y of the matched filter output has a mean $+A$ and variance $N_0/2T_b$. Note that, compared to the situation when symbol 0 was sent, the mean of the random variable Y has changed, but its variance is exactly the same as before. The conditional probability density function of Y, given that symbol 1 was sent, is therefore

$$f_Y(y|1) = \frac{1}{\sqrt{\pi N_0/T_b}} \exp\left(-\frac{(y - A)^2}{N_0/T_b}\right) \tag{4.32}$$

which is plotted in Figure 4.5*b*. Let p_{01} denote the *conditional probability of error, given that symbol 1 was sent*. This probability is defined by the shaded area under the curve of $f_Y(y|1)$ extending from $-\infty$ to the threshold λ, which corresponds to the range of values assumed by y for a decision in favor of symbol 0. In the absence of noise, the matched filter output y sampled at time $t = T_b$ is equal to $+A$. When noise is present, y occasionally assumes a value less than λ, and an error is then made. The probability of this error, conditional on sending symbol 1, is defined by

$$\begin{aligned} p_{01} &= P(y < \lambda | \text{symbol 1 was sent}) \\ &= \int_{-\infty}^{\lambda} f_Y(y|1)\, dy \\ &= \frac{1}{\sqrt{\pi N_0/T_b}} \int_{-\infty}^{\lambda} \exp\left(-\frac{(y-A)^2}{N_0/T_b}\right) dy \end{aligned} \tag{4.33}$$

To express p_{01} in terms of the complementary error function, this time we define a new variable

$$z = \frac{A - y}{\sqrt{N_0/T_b}}$$

Accordingly, we may reformulate Equation (4.33) in the compact form

$$\begin{aligned} p_{01} &= \frac{1}{\sqrt{\pi}} \int_{(A-\lambda)/\sqrt{N_0/T_b}}^{\infty} \exp(-z^2)\, dz \\ &= \frac{1}{2} \operatorname{erfc}\left(\frac{A - \lambda}{\sqrt{N_0/T_b}}\right) \end{aligned} \tag{4.34}$$

Error rate due to noise

Having determined the conditional probabilities of error, p_{10} and p_{01}, our next task is to derive the formula for the *average probability of symbol error*, denoted by P_e. Here we note that these two possible kinds of error are mutually exclusive events in that if the receiver, at a particular sampling instant, chooses symbol 1, then symbol 0 is excluded from appearing, and vice versa. Let p_0 and p_1 denote the *a priori* probabilities of transmitting symbols 0 and 1, respectively. Hence, the *average probability of symbol error* P_e in the receiver is given by

$$\begin{aligned} P_e &= p_0 p_{10} + p_1 p_{01} \\ &= \frac{p_0}{2}\operatorname{erfc}\left(\frac{A+\lambda}{\sqrt{N_0/T_b}}\right) + \frac{p_1}{2}\operatorname{erfc}\left(\frac{A-\lambda}{\sqrt{N_0/T_b}}\right) \end{aligned} \tag{4.35}$$

From Equation (4.35) we see that P_e is in fact a function of the threshold λ, which immediately suggests the need for formulating an *optimum threshold* that minimizes P_e. For this optimization we use *Leibniz's rule*.

Error rate due to noise

Consider the integral

$$\int_{a(u)}^{b(u)} f(z, u)\, dz$$

Leibniz's rule states that the derivative of this integral with respect to u is

$$\frac{d}{du}\int_{a(u)}^{b(u)} f(z, u)\, dz = f(b(u), u)\,\frac{db(u)}{du} - f(a(u), u)\,\frac{d(a(u)}{du} + \int_{a(u)}^{b(u)} \frac{\partial f(z, u)}{\partial u}\, dz$$

For the problem at hand, we note from the definition of the complementary error function in Equation (4.29) that

$$f(z, u) = \frac{2}{\sqrt{\pi}}\exp(-z^2)$$

$$a(u) = u$$

$$b(u) = \infty$$

The application of Leibniz's rule to the complementary error function thus yields

$$\frac{d}{du}\operatorname{erfc}(u) = -\frac{1}{\sqrt{\pi}}\exp(-u^2) \qquad (4.36)$$

Error rate due to noise

Hence, differentiating Equation (4.35) with respect to λ by making use of the formula in Equation (4.36), then setting the result equal to zero and simplifying terms, we obtain the optimum threshold as

$$\lambda_{\text{opt}} = \frac{N_0}{4AT_b} \log\left(\frac{p_0}{p_1}\right) \tag{4.37}$$

For the special case when symbols 1 and 0 are equiprobable, we have

$$p_1 = p_0 = \frac{1}{2}$$

in which case Equation (4.37) reduces to

$$\lambda_{\text{opt}} = 0$$

This result is intuitively satisfying as it states that, for the transmission of equiprobable binary symbols, we should choose the threshold at the midpoint between the pulse heights $-A$ and $+A$ representing the two symbols 0 and 1. Note that for this special case we also have

$$p_{01} = p_{10}$$

Error rate due to noise

A channel for which the conditional probabilities of error p_{01} and p_{10} are equal is said to be *binary symmetric*. Correspondingly, the average probability of symbol error in Equation (4.35) reduces to

$$P_e = \frac{1}{2}\,\mathrm{erfc}\left(\frac{A}{\sqrt{N_0/T_b}}\right) \tag{4.38}$$

Now the *transmitted signal energy per bit* is defined by

$$E_b = A^2 T_b \tag{4.39}$$

Accordingly, we may finally formulate the average probability of symbol error for the receiver in Figure 4.4 as

$$P_e = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{4.40}$$

which shows that *the average probability of symbol error in a binary symmetric channel depends solely on E_b/N_0, the ratio of the transmitted signal energy per bit to the noise spectral density.*

Error rate due to noise

Using the upper bound of Equation (4.30) on the complementary error function, we may correspondingly bound the average probability of symbol error for the PCM receiver as

$$P_e < \frac{\exp(-E_b/N_0)}{2\sqrt{\pi E_b/N_0}} \tag{4.41}$$

The PCM receiver of Figure 4.4 therefore exhibits an *exponential* improvement in the average probability of symbol error with increase in E_b/N_0.

This important result is further illustrated in Figure 4.6 where the average probability of symbol error P_e is plotted versus the dimensionless ratio E_b/N_0. In particular, we see that P_e decreases very rapidly as the ratio E_b/N_0 is increased, so that eventually a very "small increase" in transmitted signal energy will make the reception of binary pulses almost error free, as discussed previously in Section 3.8. Note, however, that in practical terms the increase in signal energy has to be viewed in the context of the bias; for example, a 3-dB increase in E_b/N_0 is much easier to implement when E_b has a small value than when its value is orders of magnitude larger.

Error rate due to noise

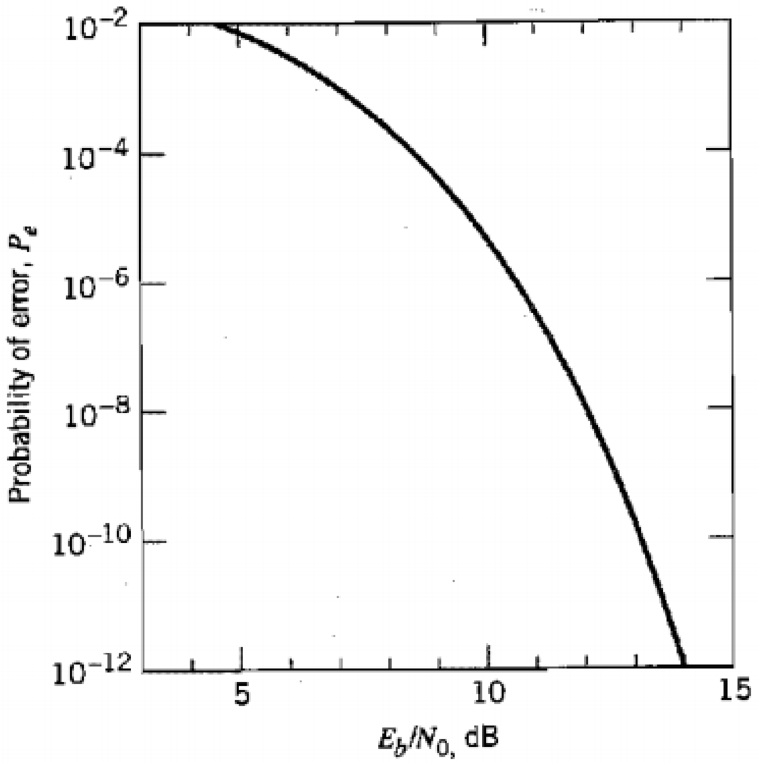


FIGURE 4.6 Probability of error in a PCM receiver.

Channel Capacity Theorem

Based on the formula of Equation (9.95), we may now state Shannon's third (and most famous) theorem, the *information capacity theorem*,[10] as follows:

The information capacity of a continuous channel of bandwidth B hertz, perturbed by additive white Gaussian noise of power spectral density $N_0/2$ and limited in bandwidth to B, is given by

$$C = B \log_2\left(1 + \frac{P}{N_0 B}\right) \text{ bits per second}$$

where P is the average transmitted power.

The information capacity theorem is one of the most remarkable results of information theory for, in a single formula, it highlights most vividly the interplay among three key system parameters: channel bandwidth, average transmitted power (or, equivalently, average received signal power), and noise power spectral density at the channel output. The dependence of information capacity C on channel bandwidth B is *linear*, whereas its dependence on signal-to-noise ratio P/N_0B is *logarithmic*. Accordingly, *it is easier to increase the information capacity of a communication channel by expanding its bandwidth than increasing the transmitted power for a prescribed noise variance.*

The theorem implies that, for given average transmitted power P and channel bandwidth B, we can transmit information at the rate of C bits per second, as defined in Equation (9.95), with arbitrarily small probability of error by employing sufficiently complex encoding systems. It is not possible to transmit at a rate higher than C bits per second by any encoding system without a definite probability of error. Hence, the channel capacity theorem defines the *fundamental limit* on the rate of error-free transmission for a power-limited, band-limited Gaussian channel. To approach this limit, however, the transmitted signal must have statistical properties approximating those of white Gaussian noise.