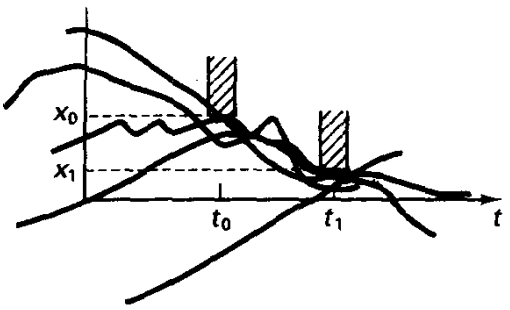


Figure 2.5.1b $F_{X_0,X_1}(x_0, x_1; t_0, t_1)$ is measure of sample paths for which $x(t_0) \le x_0$ and $x(t_1) \le x_1$.

This pair of random variables also has associated moments, notably a correlation, that depends not only on the time separation, but in general on absolute time as well. To carry this further, we can imagine nth-order descriptions, for any n, by thinking of the collection of random variables defined by inspecting the process at n time instants, $t_0, t_1, \ldots, t_{n-1}$.

We have alluded to the possibility that the description of the process may depend on the choice of reference times, as well as the time differences. If the nth-order distribution function is independent of time reference, that is, the functional form is identical if all time instants t_i are shifted by a common amount Δ, for any Δ, then we say the process is ***n*th-order stationary**. If the process is stationary for all n, it is ***strictly stationary***. In the latter cases, any probabilistic question we could pose about the process would produce the same answer whether computed now or an arbitrary time earlier or later.

Two such probabilistic averages we could consider are the ***mean function***,

$$E[X(t)] = m_X(t) \tag{2.5.1a}$$

and the ***autocorrelation function***

$$R_X(t, t+\tau) = E[X(t)X(t+\tau)], \tag{2.5.1b}$$

which is the correlation between the random variables $X(t)$ and $X(t+\tau)$. Notice that letting $\tau = 0$ in (2.5.1b) gives $R_X(t, t+\tau) = E[X^2(t)]$, which is the ***mean-square value*** of the process at time t. Electrical engineers often refer to this quantity as the *(instantaneous) power* of the process, since the mean-square value is the electrical power if $X(t)$ is a voltage signal appearing across a 1-ohm resistance.

In terms of probability density functions, these process statistics would be computed as

$$m_X(t) = E[X(t)] = \int x f_X(x; t)\, dx \tag{2.5.2a}$$

and

$$\begin{aligned} R_X(t, t+\tau) &= E[X(t)X(t+\tau)] \\ &= \iint x_0 x_1 f_{X_0,X_1}(x_0, x_1; t, t+\tau)\, dx_0\, dx_1. \end{aligned} \tag{2.5.2b}$$

(We emphasize that these integrals are not integrals over time, but over values of the random variables.) For a strictly stationary process, (2.5.2a) would produce a constant m_X, while (2.5.2b) would produce a result depending only on τ, since the joint p.d.f. is only a function of time difference τ and not absolute time t.

2.5.1 Wide-sense Stationarity, Autocorrelation Function, and Power Spectral Density

Strict stationarity is a stronger property than we generally require for systems analysis and difficult to establish in any application of the theory. A weaker condition, adequate for most applications in communications and signal processing, is that of wide-sense stationarity. A process $X(t)$ is ***wide-sense stationary***[13] if its mean and autocorrelation function are independent of absolute time; that is, for any t and τ

$$E[X(t)] = m_X \tag{2.5.3}$$

and

$$E[X(t)X(t+\tau)] = R_X(\tau). \tag{2.5.4}$$

The utility of this condition is twofold. In stable, time-invariant systems, input processes that are wide-sense stationary produce output processes which are also wide-sense stationary, and the required first and second moment functions of the output process are often easily found. Second, if the process in question is Gaussian, then wide-sense stationarity implies strict-sense stationarity, since we have seen in Section 2.3 that the nth-order distribution for a Gaussian process depends solely on the mean vector and covariance matrix, both of which are time invariant given wide-sense stationarity.

Now that we have introduced a way of abstractly visualizing stochastic processes, the question next arises, "How are processes actually specified?" There are three principal ways.

1. We can describe the ensemble in functional form, with each sample function having a dependence on some set of random variables.
2. We can decree the process to have certain statistical behavior; for example, we may assume that receiver noise in a communication link is Gaussian and wide-sense stationary with a given mean and autocorrelation function.
3. We can construct the process phenomenologically, for example, by specifying a random process $X(t)$ to be the ensemble of binary valued waveforms that switch values with probability $\frac{1}{2}$ every microsecond, with the transition time closest to $t = 0$ uniformly distributed over a 1-microsecond interval. (The reason for the randomization of the transition instants will become clear shortly.)

We proceed now to study an example of each type of specification, encountering some subtleties and seeing more clearly the description of random processes.

Example 2.16 Sinusoidal Processes

Let $X(t)$ be the process defined by $X(t) = A\sin(\omega_0 t)$, where $A \sim U[-1, 1]$, and ω_0 is a fixed value. Thus the ensemble is a set of sinusoids all crossing zero at common times, but having random amplitude A, viewed as the outcome of some experiment. Two sample functions are shown in Figure 2.5.2a. Consider the first-order probability density function: at $t = 0$, or multiples of a period later, all the sample functions have zero value, hence

[13] Wide-sense stationarity is also referred to as weak-sense stationarity.

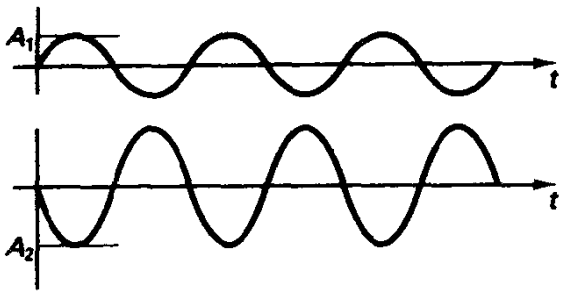


Figure 2.5.2a Two sample functions of process for Example 2.16.

the p.d.f. at these times is the Dirac impulse; that is, $f_X(x, t = n2\pi/\omega_0) = \delta(x - 0)$. At times corresponding to a quarter-period later than these instants, the process values are $X(\omega, t) = A(\omega)$, so $X(t)$ has a probability density uniform on $[-1, 1]$. Further study shows the density is always uniform, but with support that is expanding and collapsing, as in Figure 2.5.2b. Thus, the process described is not even first-order stationary. Note, however, that the mean function $m_X(t)$ is zero for all time, a result of the symmetry of the time-varying density about zero.

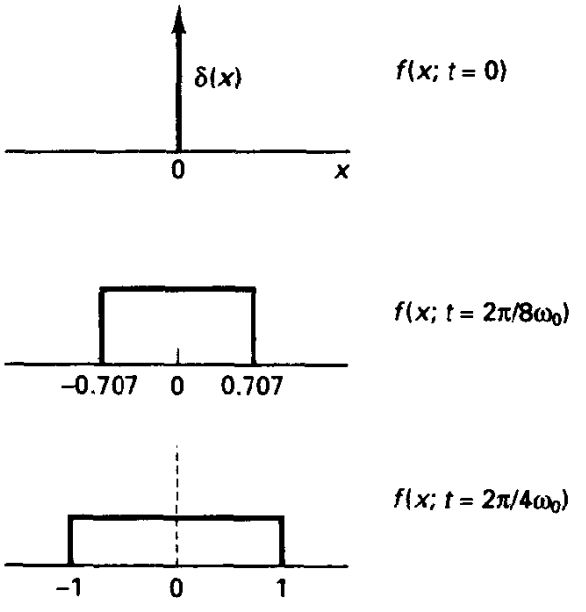


Figure 2.5.2b First-order (time-varying) density functions for Example 2.16.

A seemingly innocuous change of the process description gives the stationarity issue quite a different outcome. Let $X(t) = A\sin(\omega_0 t + \Theta)$, with A and ω_0 as before, but with Θ independent of A and uniformly distributed on $[0, 2\pi)$. This merely applies a uniformly distributed zero-crossing time to each preceding sample function. The first moment function is

$$\begin{aligned} E[X(t)] &= E[A\sin(\omega_0 t + \Theta)] \\ &= E[A\sin(\omega_0 t)\cos(\Theta) + A\cos(\omega_0 t)\sin(\Theta)] \\ &= \sin(\omega_0 t)E[A\cos(\Theta)] + \cos(\omega_0 t)E[A\sin(\Theta)], \end{aligned} \tag{2.5.5}$$

which follows from pulling nonrandom quantities outside of expectations. Because A and Θ are independent, we find $E[A\cos(\Theta)] = E[A]E[\cos(\Theta)] = 0$. (Both expectations are

zero in the final step, but either is sufficient.) Similar argument for $E[A\sin(\Theta)]$ yields a time-invariant mean, $E[X(t)] = 0$.

Now consider the autocorrelation function:

$$\begin{aligned} R_X(t, t+\tau) &= E[X(t)X(t+\tau)] \\ &= E[A\sin(\omega_0 t + \Theta)A\sin(\omega_0(t+\tau) + \Theta)] \\ &= E\left[\frac{A^2}{2}\cos(\omega_0\tau)\right] - E\left[\frac{A^2}{2}\cos(2\omega_0 t + \omega_0\tau + 2\Theta)\right] \\ &= E\left[\frac{A^2}{2}\right]\cos(\omega_0\tau) - E\left[\frac{A^2}{2}\right]E\left[\cos(2\omega_0 t + \omega_0\tau + 2\Theta)\right]. \end{aligned} \tag{2.5.6}$$

The second expectation is zero, as before, because of the uniformly distributed phase angle, Θ, while the first expectation involves the second moment of the uniform random variable A. Since $E[A^2] = \frac{1}{3}$,

$$R_X(t, t+\tau) = R_X(\tau) = \frac{1}{6}\cos(\omega_0\tau), \tag{2.5.7}$$

showing that the process is now at least wide-sense stationary. Further analysis would show that in fact the process is strict-sense stationary as well.

Example 2.17 White Gaussian Noise

As an example of the second method of specifying processes, we *define* $X(t)$ to be Gaussian, with zero mean and with autocorrelation function given by $R_X(\tau) = (N_0/2)\delta(\tau)$, where $N_0/2$ is an arbitrary constant.[14] Saying the process is Gaussian means any nth-order density function is of Gaussian form, (2.3.11). Furthermore, each random variable (sample) has zero mean, and variables (samples) at distinct time instants are uncorrelated by definition of the autocorrelation function. Because these variables are jointly Gaussian, they are independent. There is a technical problem with this process in that $X(t)$ has an infinite mean-square value (recall from the definition of autocorrelation function that $R_X(0) = E[X^2(t)]$), or equivalently infinite power, and thus this process cannot exist in the physical sense. However, this noise process serves as the archetypal model for noise processes in communication theory, as discussed in Chapter 3.

Example 2.18 Random Binary Waveform

Imagine an infinite set of binary random number generators, which produce outputs A or $-A$ every T seconds, called the bit duration. We model the values associated with each generator as a equiprobable binary random variable X_n, with successive values defined to be *independent*. This provides a construction of a random sequence, X_n. Further to each sample function, or generator, we assign a random time offset α, which over the ensemble is uniformly distributed on the interval $[0, T)$ and independent of X_n for all n. The random binary wave is then defined as

$$X(t) = \sum_{n=-\infty}^{\infty} X_n \text{rect}\left(\frac{t - nT - \alpha}{T}\right) \tag{2.5.8}$$

[14]The reason for the inclusion of the factor of 2 in the constant will be clear shortly; $N_0/2$ is conventional communication theory notation for the intensity of white noise processes.

where rect$[t/T]$ denotes the unit-height rectangle pulse on $[0, T)$ seconds. Figure 2.5.3a shows several sample functions of this process. A physical realization of this process is a collection of binary random waveform generators, each clocked at the same rate, but with different time offsets.

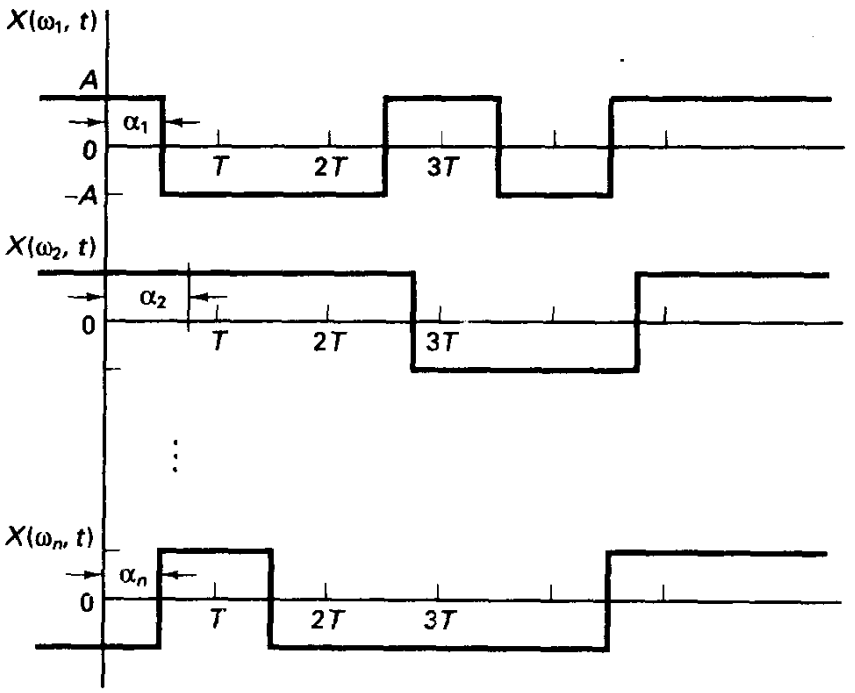


Figure 2.5.3a Sample functions of random binary wave. Each sample function assigned a random switching epoch α.

We now proceed to derive the mean and autocorrelation functions for this random process. The mean value function, since the random variable $X(t)$ is binary valued, is

$$\begin{aligned} E[X(t)] &= A \cdot P(X(t) = A) + (-A) \cdot P(X(t) = -A) \\ &= \frac{A}{2} - \frac{A}{2} = 0. \end{aligned} \tag{2.5.9}$$

Also, the autocorrelation function is

$$\begin{aligned} R_X(t, t+\tau) &= E[X(t)X(t+\tau)] \\ &= A^2 \cdot P[X(t) = X(t+\tau)] - A^2 \cdot P[X(t) \neq X(t+\tau)] \\ &= A^2 2 \cdot P[X(t) = X(t+\tau)] - 1. \end{aligned} \tag{2.5.10}$$

The latter probability requires some careful interpretation—we are seeking the probability of the event that the sample functions are identical in sign when examined τ seconds apart.

First, consider the case $\tau \geq T$, which implies that for each sample function there has been at least one switching instant in $[t, t+\tau)$. This really means that the random variables attached to the two observation times are independent r.v.'s. Thus, $P(X(t) = X(t+\tau)) = \frac{1}{2}$; that is, just as many sample functions have identical signs as opposite signs at the two time instants, on average. Then for $\tau \geq T$, $R_X(t, t+\tau) = 0$.

For $\tau < T$, the probability that the samples have identical sign is

$$P(X(t) = X(t+\tau)) = 1 \cdot P[\text{no switch instant in } (t, t+\tau)] + \frac{1}{2} \cdot P(\text{switch instant in } (t, t+\tau))$$

$$= 1 \cdot \left(1 - \frac{|\tau|}{T}\right) + \frac{1}{2} \cdot \frac{|\tau|}{T}$$

$$= 1 - \frac{|\tau|}{2T}, \qquad |\tau| < T. \tag{2.5.11}$$

We have used the fact that the probability of a switch instant occurring in an interval of length τ is τ/T for $\tau \leq T$.

Substituting results for the two cases into (2.5.10) produces

$$R_X(\tau) = \begin{cases} A^2\left(1 - \dfrac{|\tau|}{T}\right), & |\tau| \leq T, \\ 0, & \text{otherwise}, \end{cases} \tag{2.5.12}$$

which is the triangular autocorrelation function shown in Figure 2.5.3b. In other words, the random binary waveform process exhibits zero correlation for time separations longer than one bit interval and linearly decreasing correlation for time separations less than one bit duration. Note also that $R_X(0) = A^2$, the mean-square value of the process.

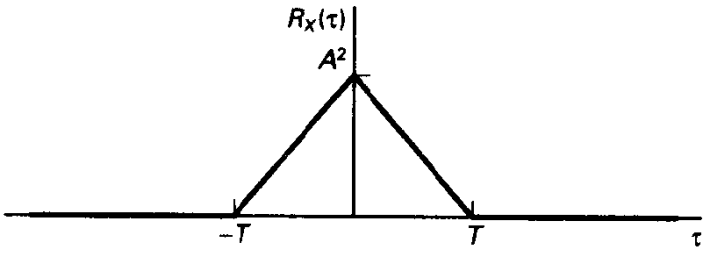


Figure 2.5.3b Autocorrelation function of random binary wave.

Communication engineers have long found it useful to describe signals in the frequency domain through the use of Fourier series for periodic signals and the Fourier transform for aperiodic signals. A frequency-domain statistical description for stochastic processes is provided by the ***power spectral density***, or simply power spectrum, $G_X(f)$, defined to be the Fourier transform of the autocorrelation function $R_X(\tau)$:

$$G_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau}\, d\tau. \tag{2.5.13a}$$

$G_X(f)$ is an even, nonnegative function of the frequency variable f. (See Exercise 2.5.1.) The inverse transform relation is

$$R_X(\tau) = \int_{-\infty}^{\infty} G_X(f) e^{j2\pi f\tau}\, df. \tag{2.5.13b}$$

Equations (2.5.13a) and (2.5.13b) are called the Wiener–Khintchine relations.

Since we have given $R_X(0)$ the significance of power in the electrical sense, (2.5.13b) shows that power is equivalently the integral of the function $G_X(f)$; hence the appropriateness of the name power spectral density, for it conveys the distribution in

frequency of the power in a random process. Specifically, $2G_X(f)\,df$ is the power of the signal located in the infinitesimal frequency range $(f, f + df)$. Power spectral densities may include impulse components (called spectral lines), which must correspond to periodicities in autocorrelation function. For example, the autocorrelation function derived in Example 2.16 was $R_X(\tau) = (A^2/2)\cos(\omega_0\tau)$. The Fourier transform of this function gives

$$G_X(f) = \frac{A^2}{4}[\delta(f - f_0) + \delta(f + f_0)], \tag{2.5.14}$$

where $f_0 = \omega_0/2\pi$. This reveals, as expected, that the entire power in the random signal is localized at one frequency. Furthermore, integration of (2.5.14) gives that this total power is $A^2/2$, as expected.

Example 2.19 White Noise (continued)

Having defined white Gaussian noise, we can now see how the name "white" derives. Recall the autocorrelation function was specified as $R_X(\tau) = (N_0/2)\delta(\tau)$. The Fourier transform of the Dirac impulse is a constant for all frequencies: $G_X(f) = N_0/2$ watts/hertz, meaning that the process has equal power in every incremental band of frequencies from audio frequencies through the x-ray region and beyond! In analogy with white light, said to contain an equal mix of all visible colors, or frequencies, we refer to the spectrum as white. This also reveals the total power difficulty cited previously, since the integral of the power spectrum is infinite. Given this situation, we may ask, "Why even consider such a process?" The answer is that white noise serves as an appropriate model when the noise process has a power spectrum that is wide compared to that of the signal of interest and constant over this region. The actual noise process, however, has finite power by virtue of its spectrum decaying to zero well outside the region of interest.

To compute the noise power contained in any finite band of frequency, $[f_1, f_1 + B]$, we integrate the (two-sided) power spectral density over both the positive and negative frequency regions, obtaining $(N_0/2)(B + B) = N_0B$ watts.

It is frequently misunderstood that "white" and "Gaussian" are synonymous. It is quite possible, however, to find Gaussian stochastic processes with nonwhite power spectrum; similarly, a process may have a constant power spectral density, but not have Gaussian density functions.

2.5.2 Stochastic Processes in Linear Systems

A linear, time-invariant, continuous-time system can be specified by its response to a unit impulse $\delta(t)$, which is called the ***impulse response***, $h(t)$. Equivalently, we may specify the system's ***frequency response function***, $H(f)$, which is the Fourier transform of the impulse response. The use of these functions in system analysis for deterministic signals should be quite familiar; in particular, if $x(t)$ is an input to a linear system and $y(t)$ is its output, then the convolution integral relates these as

$$y(t) = \int_{-\infty}^{\infty} x(t - \tau)h(\tau)\,d\tau. \tag{2.5.15a}$$

Alternatively, we may express the input/output relation in the frequency domain by

$$Y(f) = X(f)H(f), \tag{2.5.15b}$$

where $Y(f)$ and $X(f)$ are the (generally complex) Fourier transforms of the output and input signals, respectively.

The same expressions pertain for stochastic inputs, but they are not useful in themselves, for the resultant functions in either time or frequency domains must be interpreted as stochastic processes themselves. What we can do, however, is give a *statistical description* of the output process in terms of the input description and the system response.

If we take expectations of both sides of (2.5.15a) and then interchange the order of time integration and expectation, we find that

$$\begin{aligned} m_Y &= m_X \int_{-\infty}^{\infty} h(\tau)\, d\tau \\ &= m_X H(0). \end{aligned} \tag{2.5.16}$$

This ratifies what we would probably anticipate: the mean or d.c. value of the output process is the mean of the input, scaled by the zero-frequency gain of the system.

Continuing with the autocorrelation, we write the output autocorrelation as

$$\begin{aligned} R_Y(t, t+\tau) &= E[Y(t)Y(t+\tau)] \\ &= E\left[\int X(t-\alpha)h(\alpha)\, d\alpha \int X(t+\tau-\beta)h(\beta)\, d\beta\right] \\ &= \iint E[X(t-\alpha)X(t+\tau-\beta)]h(\alpha)h(\beta)\, d\alpha\, d\beta. \end{aligned} \tag{2.5.17}$$

The expectation in the integrand is just $R_X(\tau+\alpha-\beta)$, since the process $X(t)$ is wide sense stationary. We substitute this into (2.5.17) and then recognize after change of variables that the integral is an iterated convolution operation, obtained by first convolving the autocorrelation function with the impulse response and then convolving this with the time-reversed impulse response. Specifically,

$$R_Y(t, t+\tau) = R_X(\tau) * h(\tau) * h(-\tau) = R_Y(\tau), \tag{2.5.18}$$

where again $*$ denotes the convolution operation. Note that the output process is wide sense stationary if the input process is (and the system is stable and time invariant).

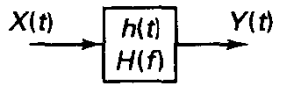
This input/output behavior is more easily comprehended in the frequency domain, obtained by taking the Fourier transform of both sides of (2.5.18):

$$G_Y(f) = G_X(f)H(f)H^*(f) = G_X(f)|\, H(f)\, |^2. \tag{2.5.19}$$

(Here the superscript $*$ denotes conjugation of a complex variable.) This reveals that the output spectrum is shaped according to the input spectrum, weighted in frequency by the power response of the linear network. Such effects underlie the ability of a filter in a receiver to pass certain portions of the frequency band and perhaps reject large amounts of unwanted noise. Figure 2.5.4 summarizes the important relations for wide-sense stationary signals acted on by linear systems.

At this point, we can reconsider stochastic sequences and merely state analogous input/output relationships for linear discrete-time systems. Detailed treatments are found in [1], [3], and [4]. Distribution functions and p.d.f.'s are defined at discrete points in time. The autocorrelation sequence of a wide-sense stationary random sequence $\{X_n\}$ is defined as

$$R_X(k) = E[X_n X_{n+k}], \tag{2.5.20a}$$



$$Y(t) = \int X(t-\tau)h(\tau)\,d\tau$$
$$m_Y = m_X H(0)$$
$$R_Y(\tau) = R_X(\tau) * h(\tau) * h(-\tau)$$
$$G_Y(f) = G_X(f)|H(f)|^2$$

Figure 2.5.4 Input/output relations for linear time-invariant system excited by stationary random process.

and the power spectrum of the discrete-time process is defined as the ***discrete-time Fourier transform*** of the autocorrelation sequence:

$$G_X(f) = \sum_{k=-\infty}^{\infty} R_X(k)e^{-j2\pi fk} \tag{2.5.20b}$$

The input/output relation for power spectra in linear, time-invariant, discrete-time systems can be shown to be

$$G_Y(f) = G_X(f)|H(f)|^2, \tag{2.5.21}$$

where $H(f)$ is the discrete-time Fourier transform of the system's impulse response sequence, $\{h_n\}$:

$$H(f) = \sum_{n=-\infty}^{\infty} h_n e^{-j2\pi fn}. \tag{2.5.22}$$

The power spectra for discrete-time sequences are seen to be periodic with period equaling the inverse of the sampling interval, a manifestation of the aliasing phenomenon. The exercises provide some simple applications of these properties of stochastic signals in linear systems.

2.5.3 Time Averages versus Ensemble Averages

We have been characterizing random processes by their ***ensemble averages***, that is, by imagining the collection of random variables defined by examining the entire ensemble at various time instants. Thus, a statement about the autocorrelation function for a random process is a statement about the behavior of the ensemble samples taken at two fixed time instants. This would be computed by

$$E[X(t)X(t+\tau)] = \int\int x_1 x_2 f(x_1, x_2; t, t+\tau)\,dx_1 dx_2. \tag{2.5.23}$$

In the practical situation of, say, transmitting a message through a communications link, we presume that we are provided a sample function from the source (the message), and the channel provides some random corruption in the form of noise, time-varying channel characteristics, or the like, but we deal with *one* sample function from this large process. The logical question is whether ensemble averages tell us anything about similar averages obtained from a single sample function of the process. For example, we will devote considerable attention to predicting the probability of error associated with various kinds of signaling alternatives studied in the rest of the book; this mathematics

is implicitly an ensemble viewpoint. Can we assume that the average error probability measured over time from one sample function will be equivalent? Or can we count errors associated with any single sample function we are given and perform probability calculations based on this empirical data?

We shall be a bit circular here and say that, if a process possesses certain ***ergodic*** properties, ensemble averages of traditional probability theory may be equated with time averages, in the limit of long averaging time. To illustrate, recall the usual way to estimate the mean of a process, through time averaging, is by means of a T-second sliding average of the sample function:

$$<X(t)>_T = \frac{1}{T}\int_{t-T/2}^{t+T/2} X(\omega, s)\, ds. \tag{2.5.24}$$

(We employ the brackets $<>$ to denote time averaging.) Notice that the time average is itself a random process; the results are dependent on the sample function for which averaging is performed. However, for a stationary mean-ergodic process we can claim that

$$\lim_{T\to\infty} <X(t)>_T = m_x, \tag{2.5.25}$$

where the equality will be interpreted in the mean-square sense; that is, the second moment of the difference between the time averaged estimate and the ensemble mean m_X goes to zero as T increases. Similarly, other time-averaged statistics, such as variance, correlation, and histograms/distribution functions, when averaged a sufficiently long time, will approach the corresponding ensemble quantity if an appropriate ergodic property holds. This equivalence is what makes probability theory a useful tool for engineering applications, but it is important to understand the conceptual leap we make when equating ensemble averages with time averages.

What makes a process have ergodic properties then? Basically, the requirement is that any single sample function of the process, over time, should reflect the nature of the ensemble. Probabilists express this as a *mixing* property, and there are various technical requirements for processes to possess ergodic properties in various forms. For example, for $X(t)$ to be "ergodic in the mean," meaning that (2.5.25) holds, we require that the autocorrelation function be absolutely integrable, that is, $\int | R_X(\tau) | \, d\tau < \infty$ [1]. We shall not dwell further on this issue, but assume that the requisite conditions hold for the processes of interest to possess whatever ergodic properties are needed. To indicate the subtlety involved here, we discuss a strictly stationary process that is not ergodic in the mean.

Example 2.20 A Stationary, But Not Ergodic, Process

Consider the random process formed by observing the waveform output of nominally 5-volt power supplies from a stockroom. Some indeed produce 5-volt terminal voltage, some produce 4.8 volts, others produce 5.1 volts, and so on. Some may be defective, in which case the output waveform is always zero. We assume that each produces a constant voltage for all time, so the ensemble is a set of fixed-voltage waveforms, as indicated in Figure 2.5.5. Notice that the ensemble average mean of the process may be 4.9 volts (reflecting some defective supplies), whereas a time-averaged estimate (2.5.24) converges immediately to the voltage associated with the particular power supply under test. Thus, even the simplest time average cannot be equated with the corresponding ensemble average here, and the process is not ergodic in the mean. However, it is clear that the process is

stationary, since the statistical description is certainly time invariant. In fact, the p.d.f. for n samples taken at *any* distinct set of time points is

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_X(x_1)\delta(x_2 - x_1)\delta(x_3 - x_1)\cdots\delta(x_n - x_1), \tag{2.5.26}$$

signifying that the n samples are identical (for each sample function), with a marginal p.d.f. $f_X(x)$.

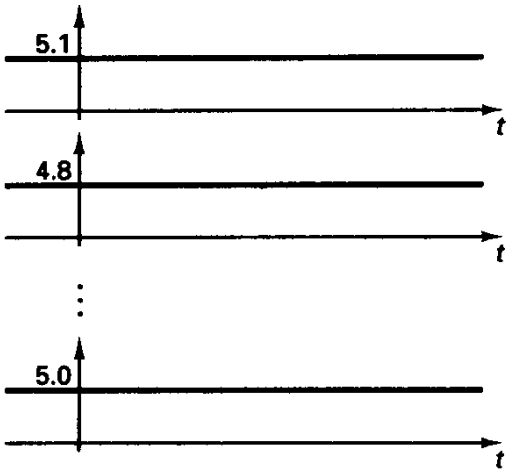


Figure 2.5.5 Sample functions of power supply process, Example 2.20.

2.5.4 Karhunen–Loeve Series Representation for Random Processes

In the study of deterministic signals, especially periodic signals, orthogonal series expansions have played a prominent role in analysis. The preeminent case is the representation of signals by weighted sums of sinusoids or complex exponentials, the familiar Fourier series. This choice has special appeal in linear systems analysis, since sinusoids are eigenfunctions of linear systems in the steady-state case; that is, sinusoidal inputs produce sinusoidal outputs at the same frequency, but with different amplitude and phase. The latter quantities are specified by the transfer function of the linear system.

An ***orthonormal series expansion*** for a deterministic signal $x(t)$ over some interval (T_i, T_f) is of the form

$$x(t) = \sum_n x_n \phi_n(t), \qquad T_i < t < T_f, \tag{2.5.27}$$

where x_n are ***expansion coefficients*** and $\{\phi_n(t)\}$ denotes a set of ***orthonormal basis functions*** over the time interval (T_i, T_f) specified for the expansion in (2.5.27). In general, we understand the sum to involve an infinite number of terms, although in some cases we shall encounter in Chapter 3, a finite sum provides an *exact* representation. We shall say more about the convergence in (2.5.27) shortly.

Orthonormality of a set of functions requires

$$\int_{T_i}^{T_f} \phi_i(t)\phi_k(t)\,dt = \delta_{ik} = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \end{cases} \tag{2.5.28}$$

where δ_{ik} denotes the Kronecker delta function. Examples of orthonormal sets include nonoverlapping rectangular pulses and the Fourier set of sinusoids and cosinusoids having frequencies $2\pi m/T$, where $T = T_f - T_i$ is the length of the expansion interval.

Because of orthogonality of the basis functions, it follows that the expansion coefficients can be computed separately, in any order, as

$$x_n = \int_{T_i}^{T_f} x(t)\phi_n(t)\,dt. \tag{2.5.29}$$

The representation in (2.5.27) thus provides an association between a waveform $x(t)$ and its expansion coefficients, $\{x_n\}$, once a basis set has been adopted.

Literal equality in (2.5.27) cannot be expected for an arbitrary class of signals, since two signals differing in only a finite number of points would have identical expansion coefficients by (2.5.29) and hence identical right-hand sides in (2.5.27), yet the two functions being represented differ. For our purposes, however, it is adequate that, as more terms are added to the expansion, the integral-square error diminishes to zero. Specifically, let $x_N(t)$ represent a finite series N-term expansion as in (2.5.27). If we find that

$$\lim_{N\to\infty} \int_{T_i}^{T_f} [x(t) - x_N(t)]^2 dt = 0 \tag{2.5.30}$$

for all signals $x(t)$ in some class, say the class of finite-energy signals, then we say the set $\{\phi_n(t)\}$ is ***complete*** for the prescribed class. For example, the set of complex exponentials form a complete set with respect to the class of bounded functions $x(t)$ having a finite number of discontinuities and extrema on $[0, T]$.

The previous discussion has pertained to deterministic signals, but the same concept is applicable, with care, to stochastic processes. For example, consider the N-term expansion

$$X_N(t) = \sum_{n=1}^{N} X_n \phi_n(t) \tag{2.5.31}$$

as an N-term approximation of the random process $X(t)$ over some interval. As before, we envision computing the expansion coefficients as in (2.5.29). Here, however, we must interpret the coefficients as random variables. Furthermore, the issue of convergence is more subtle. We say the set of basis functions is complete here if

$$\lim_{N\to\infty} E[X(t) - X_N(t)]^2 = 0, \tag{2.5.32}$$

which is to say that the mean-square value of the approximation error approaches zero for all points in time. Sometimes this is referred to as mean-square stochastic convergence, or "limit in the mean" convergence.

With deterministic signal expansions, there is some latitude in the choice of basis set. Usually, the choice is driven by convenience or by special behavior, such as that of sinusoids in linear networks. In the case of stochastic processes, a convenient choice is one that makes the expansion coefficient r.v.'s ***uncorrelated.*** Thus, we have in mind a set of orthogonal basis functions that induces the statistical result

$$E[X_m X_n] = \lambda_m \delta_{mn}, \tag{2.5.33}$$

where δ_{mn} is the Kronecker delta function and λ_m is the mean-square value of the mth expansion coefficient. To see what this requires of the basis set, we write

$$\begin{aligned} E[X_m X_n] &= E\left[\int_{T_i}^{T_f} X(t)\phi_m(t)\,dt \int_{T_i}^{T_f} X(s)\phi_n(s)\,ds\right] \\ &= \int\int E[X(t)X(s)]\phi_m(t)\phi_n(s)\,dt\,ds. \end{aligned} \tag{2.5.34}$$

Assuming that $X(t)$ is wide sense stationary, we can write the expectation as $R_X(t-s)$, necessitating from (2.5.33)

$$\lambda_m \delta_{mn} = \int \phi_m(t)\left[\int R_X(t-s)\phi_n(s)\,ds\right] dt. \tag{2.5.35}$$

For this to hold for a given m and all n, the basis functions must satisfy the integral equation

$$\lambda_m \phi_m(t) = \int_{T_i}^{T_f} R_X(t-s)\phi_m(s)\,ds, \qquad m = 0, 1, \ldots. \tag{2.5.36}$$

The possible solutions $\phi_m(t)$ are known as ***eigenfunctions*** of the integral equation, and the λ_m are the corresponding ***eigenvalues.*** Obviously, the solutions depend on the correlation structure of the process.

Oddly, our main interest is not in solving this integral equation, although we will consider two important cases shortly. More important is the fact that orthonormal solution sets do exist[15] (in general, a countable infinity of solutions), and when these orthonormal bases are employed to expand the random process, uncorrelated coefficients are indeed obtained. The corresponding expansion of the form (2.5.27) is known as the ***Karhunen–Loeve (K–L) expansion*** of a random process [8].

It may also be seen from (2.5.29) that the X_n coefficients have zero mean if $X(t)$ has zero mean and that the variance of X_n is λ_n, the eigenvalue attached to the nth solution. Furthermore, the sum of all the eigenvalues equals the power of the process. (These results are developed in Exercise 2.5.5). In the important case when $X(t)$ is a Gaussian process, the uncorrelatedness of the coefficients renders them independent as well, providing additional analytical simplicity.

Example 2.21 Karhunen–Loeve Expansion for Band-limited White Noise Process

Let $X(t)$ be a stationary, zero-mean process with the spectrum shown in Figure 2.5.6. We refer to such processes as ***ideal band-limited processes***, or band-limited white noise processes. We note that the bandwidth of the process is B hertz and that the total power of the process is $N_0 B$. The autocorrelation function for this process is obtained by computing the inverse Fourier transform of the power spectrum and is found to be

$$R_X(\tau) = N_0 B \frac{\sin(\pi B \tau)}{\pi B \tau}. \tag{2.5.37}$$

The Karhunen–Loeve basis functions for this case are solutions to the integral equation (2.5.36), with (2.5.37) substituted, and are known as ***prolate spheroidal wave functions*** [9].

---

[15] See, for example, Courant and Hilbert, *Methods of Mathematical Physics*, Interscience, New York, 1953.

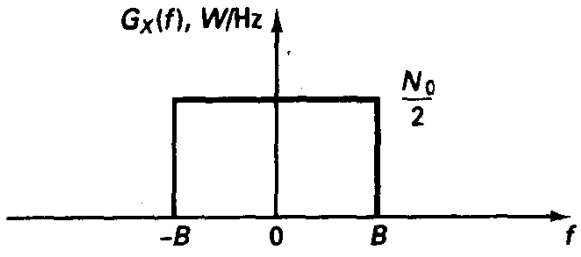


Figure 2.5.6 Power spectral density for ideal band-limited process.

These functions are infinite-duration waveforms, orthogonal over the entire real line, and over $[-T/2, T/2)$, as we require, are strictly band limited and a complete orthonormal set for the set of finite-energy band-limited signals. Although not simple to express, these functions are illustrated in [9] for differing BT products. Figure 2.5.7a presents the first four basis functions for a $BT = 1.27$, and we can visualize both forms of orthogonality. The eigenvalue λ_n represents the energy in the waveform $\phi_n(t)$ over the interval of length T. For small indexes the eigenvalues tend to have value near $N_0/2$, but it is noted that the eigenvalues drop sharply for indexes greater than $\lceil 2BT+1 \rceil$. Two sets of eigenvalue profiles are shown in Figure 2.5.7b. In other words, when the K–L decomposition of a process is limited to B hertz and T seconds, only about $2BT+1$ expansion coefficients have significant energy. We might say that the signal lies in a space with $2BT+1$ dimensions. It is also known that the argument "hardens" as the time–bandwidth product, BT, increases. That is, the transition from large eigenvalues to insignificant values is sharper as BT increases; this is seen in Figure 2.5.7b.

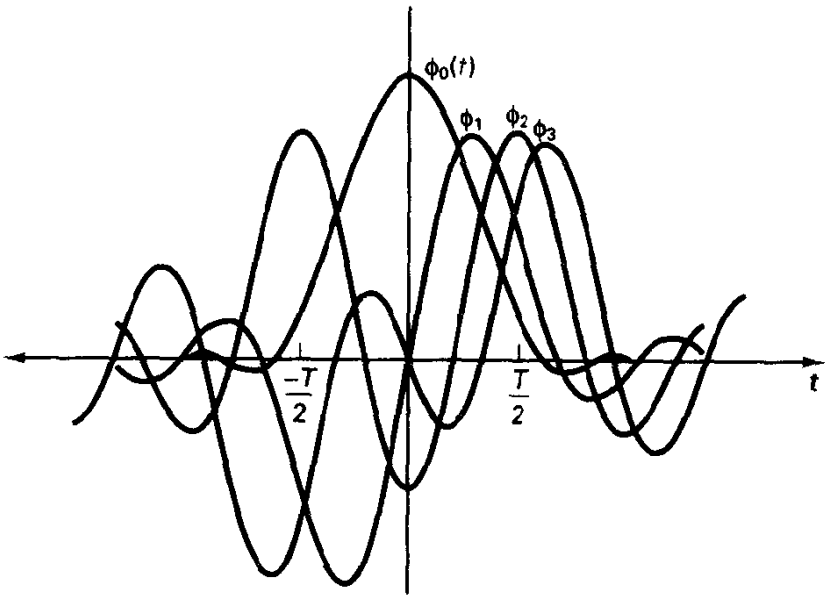


Figure 2.5.7a Orthogonal basis functions $\phi_i(t)$ for K–L expansion of ideal band-limited process; $BT = 1.27$ (taken from Slepian and Landau [9]).

We might notice a certain similarity of the basis functions shown in Figure 2.5.7a to the traditional sine/cosine basis set. For random processes whose power spectra are rational functions in f^2, it is known that for large BT the eigenfunctions indeed approach sinusoids, and the frequencies of these sinusoids approach multiples of $2\pi/T$; that is, the Fourier basis emerges. Other suitable basis sets for large BT are the sine functions, of the form

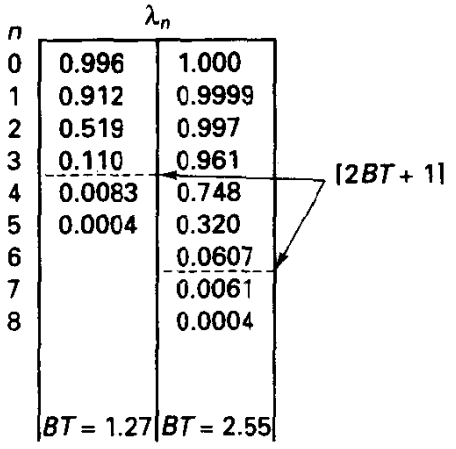


Figure 2.5.7b Eigenvalue profiles for K–L expansion of ideal band-limited process [9].

$\sin(\pi Bt)/(\pi Bt)$ and time translates by $\frac{1}{2}B$ seconds. These are band limited and orthogonal, and roughly $2BT$ may be squeezed into an interval of length T. Here the expansion coefficients become just the samples of the random process, taken at rate $2B$ per second, and the synthesis function is provided by injecting these samples into an ideal low-pass filter.

Example 2.22 K–L Series Representation of White Noise

Recall that the autocorrelation function of white noise is the Dirac impulse, $R_X(\tau) = (N_0/2)\delta(\tau)$. Substitution into (2.5.36) and invocation of the sifting property reveal the degenerate result that *any* orthonormal set of functions provides uncorrelated coefficients, making the formal selection of the basis set not important. Furthermore, the variance of each coefficient is $\lambda_n = N_0/2$, the noise spectral density. The existence of an unbounded number of coefficients with equal variance again exposes the infinite-power dilemma. Although we will often invoke white noise as a useful model, we should really have in mind a large, but finite, bandwidth signal as in the previous example. [Technically, the sum in (2.5.27) does not converge in mean square in this case, because the original process does not have finite power; this does not nullify the claims just made about what occurs when white noise is projected onto orthonormal functions.]

2.5.5 Markov Models

Markov processes play a key role in modeling the statistical dependencies of many random process situations. Our interest will be in their use in descriptions of discrete information sources, for describing channels having memory, and as descriptions of certain channel encoding operations.

A random sequence $\{X_k\}$ is called ***first-order Markov*** if

$$f(x_k|x_{k-1}, x_{k-2}, \ldots x_0) = f(x_k|x_{k-1}); \tag{2.5.38}$$

that is, the density function of the random variable X_k, conditioned on the entire past, can be expressed exactly through conditioning only on the most recent symbol.[16] The beauty of Markov models is that p.d.f.'s for any collection of random variables can be obtained

[16]The definition can be extended to jth-order Markov behavior if the conditioning can be reduced to the j most recent symbols.

by knowing the marginal p.d.f. for the first variable and applying the conditional density function iteratively in chain-rule fashion to build joint density functions. For example, a third-order p.d.f. can be constructed as $f(x_3, x_2, x_1) = f(x_1)f(x_2|x_1)f(x_3|x_2)$.

An important special case of Markov processes is ***finite-state Markov sequences***, or Markov chains as these are known in the probability and operations research field. We define a finite-state Markov system to have a finite number of internal states, designated $0, 1, \ldots, S-1$, among which the system evolves in time. We let the state at time k be denoted σ_k, and at regular time instants the state transitions to another state (or perhaps itself) according to a set of conditional probabilities:

$$a_{ij} = P(\sigma_{k+1} = j \mid \sigma_k = i). \tag{2.5.39}$$

We assume these transition probabilities are not time dependent. The conditional probabilities can be conveniently summarized by a state-transition-probability matrix **A**, having dimension $S \times S$:

$$\mathbf{A} = [a_{ij}]. \tag{2.5.40}$$

Note that since the entries of this matrix are (conditional) probabilities, rows must sum to 1. An equivalent description is provided by a state-transition diagram, as in Figure 2.5.8, with arcs labeled according to the probability of making the indicated transition.

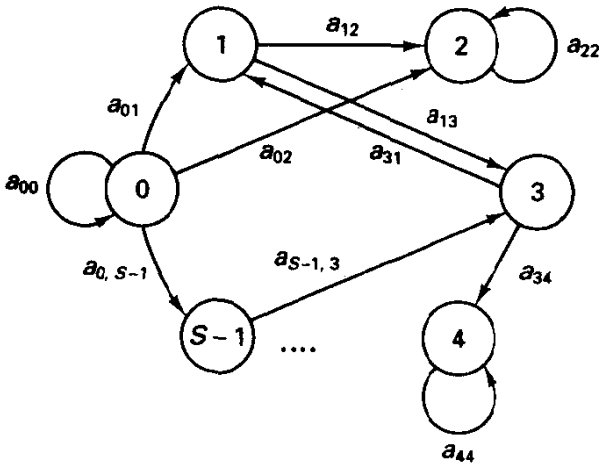


Figure 2.5.8 State transition diagram for discrete-time Markov system. Arc labels are transition probabilities.

Given a probability distribution on states at time k, $\mathbf{p}_k = [P(\sigma_k = 0), P(\sigma_k = 1), \ldots, P(\sigma_k = S-1)]$, we consider the probabilistic evolution of system state in the future. We visualize probability as a commodity that must be conserved in a state graph, and we realize that the probability of being in state j at time $k+1$ is given by

$$P(\sigma_{k+1} = j) = \sum_{i=0}^{S-1} P(\sigma_k = i)a_{ij}. \tag{2.5.41}$$

This relation holds for other states as well, and we may represent the evolution of the state probabilities in the matrix equation

$$\mathbf{p}_{k+1} = \mathbf{p}_k\mathbf{A}. \tag{2.5.42}$$

If the Markov system is well connected and regular (essentially meaning that there are no dead-end or absorbing states) and does not exhibit periodic behavior,[17] then, regardless of initial probability distribution on states, $\mathbf{p}_0$, the system reaches, asymptotically in time, a steady-state distribution, and the state sequence is asymptotically (as time evolves) stationary. The same conditions ensure that any sample function of the random state sequence, observed over sufficient time, exhibits the steady-state ensemble average statistics.

The steady-state solution is obtained by requiring the state probabilities at time k and $k+1$ to be equal (the definition of steady state). This implies we must solve the linear system

$$\mathbf{p} = \mathbf{pA} \tag{2.5.43}$$

subject to the constraint that the elements of $\mathbf{p}$ sum to 1.

Example 2.23 Gilbert–Elliot Model for a Bursty Channel

Some binary channels have a tendency to exhibit bursts of transmission errors, wherein the channel error probability is 1/2 during burst error conditions and very small (10^{-5} say) during nominally good periods. (Such effects occur due to sporadic strong noise, loss of synchronization in a receiver, signal fading, and the like.) A classical model for such channels is the Gilbert–Elliot model [10] in which we assign the channel to be in one of two states: 0 for "good" and 1 for "bad." Through measurements, we might find that the state transition probabilities are

$$\begin{aligned} a_{00} &= 0.99, \qquad a_{01} = 0.01, \\ a_{10} &= 0.10, \qquad a_{11} = 0.90. \end{aligned} \tag{2.5.44}$$

Thus, the system tends to persist in either state, but more so in the good state. The state transition diagram is shown in Figure 2.5.9, from which it is clear that the state process is recurrent. The steady-state probabilities of being in the good or bad state are given, respectively, by

$$\begin{aligned} P_0 &= 0.99P_0 + 0.10P_1, \\ P_1 &= 0.90P_0 + 0.01P_1, \end{aligned} \tag{2.5.45a}$$

together with

$$P_0 + P_1 = 1. \tag{2.5.45b}$$

The two equations in (2.5.45a) are dependent; either combined with (2.5.45b) yields $P_0 = \frac{10}{11}, P_1 = \frac{1}{11}$.

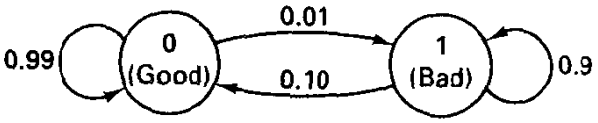


Figure 2.5.9 Channel state diagram for Gilbert–Elliot bursty binary channel.

Notice that we have determined only the state probabilities; the average probability of channel error is something different and yet to be discussed.

Having modeled the state of a Markov system, we now wish to specify an ***action***, or output, of the system at each time by a production rule which is state dependent. The actions may correspond to the production of a source character in a digital message,

[17] Such Markov chains are said to be recurrent, or ergodic [4].

the production of channel errors in the preceding example, or perhaps the generation of channel code symbols in a certain form of finite-state Markov encoder.

More specifically, we let the system produce one of B actions at time k and designate this action as $b_k \in \{0, 1, \ldots, B-1\}$. The probability that a given action occurs is conditional on the state σ_k, and to completely define the model, we simply specify $P(b_k = j \mid \sigma_k = i)$ for all $j = 0, 1, \ldots, B-1$ and $i = 0, 1, \ldots, S-1$.

The combination of the Markov dynamics for the system state and the production rule dependent on state imbues the action process with Markovian nature, as defined by (2.5.38). It is possible for the output to correspond with the current (or next) state with certainty, in which case we may as well label the states as the actions. However, our present model is more general, allowing for example, a digital source to have two states (perhaps alphanumeric data and English prose), but a much larger set of actions or outputs.

We can compute any joint probability of interest from this formulation simply by finding the steady-state probabilities for system states and then using the conditional probabilities for the various actions. As a special case, the marginal probability of the system output $b_k = j$ is

$$P(b_k = j) = \sum_{i=0}^{S-1} P(\sigma_k = i) P(b_k = j \mid \sigma_k = i)). \tag{2.5.46}$$

Example 2.23 (continued)

Given the earlier specification, we have that the two actions of the channel are error ($b_k = 1$) and no error ($b_k = 0$). The conditional probabilities of these, given the two states of the channel, are

$$\begin{aligned} P(\text{error} \mid \text{good}) &= 10^{-5}, & P(\text{no error} \mid \text{good}) &= 1 - 10^{-5} \\ P(\text{error} \mid \text{bad}) &= 0.5, & P(\text{no error} \mid \text{bad}) &= 0.5 \end{aligned} \tag{2.5.47}$$

Substitution into (2.5.45) yields the average error probability for the channel as $P(\text{error}) = P(\text{good})10^{-5} + P(\text{bad})0.5 = 0.04546$. This is the long-term error probability, which would be measured by counting errors, assuming that our modeling is accurate; it is important to notice the fact, however, that the errors tend to cluster when designing the digital communication system. Exercise 2.5.10 involves calculating the probability of two consecutive errors; this is certainly not the square of the marginal error probability calculated previously, which would be correct if channel errors were independent. Effective error control techniques would need to anticipate this error clustering phenomenon.

2.6 STATISTICAL DECISION THEORY

Demodulation and decoding of noisy signals are a direct application of statistical decision theory. In the more general setting, we are given a finite set of possible hypotheses about an experiment, along with observations related statistically to the various hypotheses, and the theory provides rules for making best decisions (according to some performance criterion) about which hypothesis is likely to be true. The general theory has applications in many fields of social and physical sciences, including economic policy and the assessment of drug efficacy on illness.

In digital communications, the hypotheses are the possible messages, and the observables are the outputs of a probabilistic channel. The schematic situation is depicted

in Figure 2.6.1. Usually we will assume that the observables are continuous random variables, or random vectors, and thus we express the influence of the channel through probability density functions. Conversion to the discrete random variable case is done in an obvious manner.

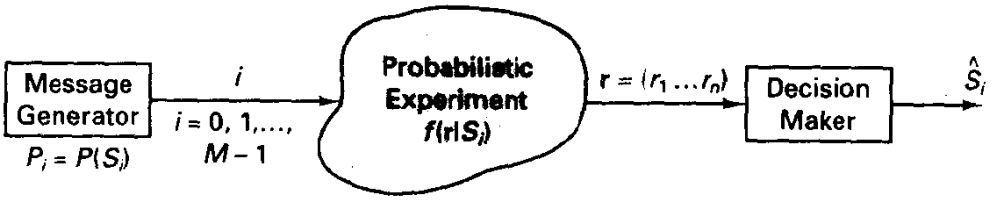


Figure 2.6.1 Statistical decision theory setup.

2.6.1 Minimum Probability of Error Policies

Suppose we have M possible ***hypotheses*** (signals), labeled by $S_i, i = 0, 1, \ldots, M-1$, associated with a probabilistic experiment. We also adopt ***prior probabilities*** on the hypotheses, denoted P_i. We assume the ***observable*** of the experiment is some collection of n real values, denoted by the vector $\mathbf{r} = (r_1, r_2, \ldots, r_n)$, and we presume we are given, or can compute, ***conditional probability densities*** $f(\mathbf{r} \mid S_i)$ or $P(\mathbf{r} \mid S_i)$, depending on whether the observation is a continuous or discrete random vector. Based on $\mathbf{r}$, the decision maker produces a decision $\hat{S}_i$. We are interested in the best decision-making algorithm in the sense of minimizing $P(\hat{S}_i \neq S_i)$, the probability of decision error.

As a side note, this problem may be generalized by weighing differently the costs of various kinds of errors and then finding the policy that minimizes expected weighted cost. In a radar detection setting, for example, we may wish to penalize errors of the missed-target variety more heavily than false-alarm errors. However, in digital communications it is customary to assign unit cost to all error conditions and zero cost to correct decisions, whence the expected cost is the probability of decision error.

The observation vector $\mathbf{r}$ may be regarded as a point in some observation space, perhaps R^n or the space of binary n-tuples. Conceptually, it is helpful to view the decision maker as partitioning the observation space into decision zones, as shown in Figure 2.6.2 for a case with three hypotheses. We label the decision zones $D_i, i = 0, 1, \ldots, M-1$, and agree that the decision is in favor of hypothesis S_i if $\mathbf{r} \in D_i$. Note that

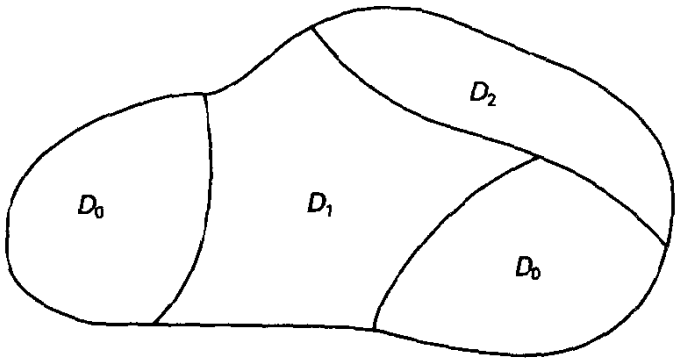


Figure 2.6.2 Abstract partition of observation space for $M = 3$.

in general the individual decision regions are not required to be "connected" regions in observation space (observe D_0 in Figure 2.6.2), and it is even possible that no points in the observation space will be assigned to a given hypothesis, tantamount to never accepting that hypothesis. The task now is to define the partition boundaries optimally, which in effect gives a rule for processing $\mathbf{r}$ to obtain the best decision $\hat{S}_i$. Following this development, we will see how to implement the decision maker in certain common cases.

We are interested in minimizing $P(\hat{S}_i \neq S_i) = P(\epsilon)$, where ϵ designates the error event. First, let us consider the probability of error, conditioned on S_i being the true hypothesis. Then

$$\begin{aligned} P(\epsilon \mid S_i) &= P(\mathbf{r} \in D_i^c \mid S_i) \\ &= \int_{D_i^c} f(\mathbf{r} \mid S_i)\, d\mathbf{r}, \end{aligned} \tag{2.6.1}$$

where D_i^c denotes the complement of the ith decision region, and the integral is interpreted as an n-dimensional integral, or n-fold summation in the case of discrete r.v. observations.

The average probability of error is then

$$\begin{aligned} P(\epsilon) &= \sum_{i=0}^{M-1} P_i P(\epsilon \mid S_i) \\ &= \sum_{i=0}^{M-1} P_i \int_{D_i^c} f(\mathbf{r} \mid S_i)\, d\mathbf{r}. \end{aligned} \tag{2.6.2}$$

We now state the optimal way to partition observation space.

> Assign $\mathbf{r}$ to that D_i for which $P_i f(\mathbf{r} \mid S_i)$ is maximum.

If ties occur in this assignment, an arbitrary choice among those decision regions that are tied may be made.

Obviously, the decision maker need not formally compute decision boundaries and then determine which cell D_j the vector $\mathbf{r}$ falls into, but instead need compute only $P_i f(\mathbf{r} \mid S_i)$, $i = 0, 1, \ldots, M-1$, and choose that index i with the largest result. Again, in the case of ties, an arbitrary tie-breaking rule is permissible. Thus we claim that the optimal decision rule is

$$\hat{S}_i = \arg\max_{S_i} P_i f(\mathbf{r} \mid S_i). \tag{2.6.3}$$

(We read "arg max" as the operator producing the argument that maximizes the function indicated.) For discrete observations, we merely replace the p.d.f.'s in (2.6.3) with the appropriate conditional probabilities.

The proof of this rule's optimality, which we now provide for the two-hypothesis case, is by contradiction. Suppose we adopt (2.6.3) as a decision procedure, which implies some associated $P(\epsilon)$ by (2.6.2). Now make an arbitrary change of the boundary, moving a piece Δ of observation space formerly in D_0, say, to D_1, as indicated in Figure 2.6.3.

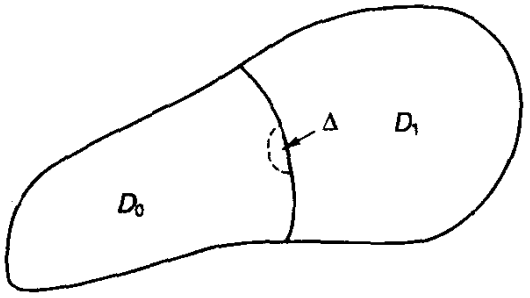


Figure 2.6.3 Perturbation of decision boundary: $\delta P = \int_\Delta P_0 f(\mathbf{r} \mid S_0) - P_1 f(\mathbf{r} \mid S_1)\, d\mathbf{r}$.

The new error probability is

$$\tilde{P}(\epsilon) = P(\epsilon) + \delta P = P(\epsilon) + \int_\Delta [P_0 f(\mathbf{r} \mid S_0) - P_1 f(\mathbf{r} \mid S_1)]\, d\mathbf{r}. \tag{2.6.4}$$

Over the region Δ, the integrand is nonnegative; otherwise, Δ would not have formerly been assigned to D_0, so the error probability associated with the new partition must be at least as large as the original. Exercise 2.6.1 takes up the extension to the M-ary case.

Using Bayes's rule in mixed form, we may write the posterior probability for the hypothesis S_i, given the observation $\mathbf{r}$, as

$$P(S_i \mid \mathbf{r}) = \frac{P_i f(\mathbf{r} \mid S_i)}{f(\mathbf{r})} \tag{2.6.5}$$

However, the denominator on the right-hand side in (2.6.5) does not involve i, and maximizing (2.6.5) is equivalent to maximizing $P_i f(\mathbf{r} \mid S_i)$. In fact, maximization over i of any *monotonic* function of the product $P_i f(\mathbf{r} \mid S_i)$ is optimal. [Keep in mind we are not ultimately interested in the value of $P_i f(\mathbf{r} \mid S_i)$, but only the index i that maximizes the expression.] Often the proper choice of the monotonic function can simplify the calculation considerably, as we will see shortly.

Because of its equivalence with maximizing (2.6.5), the rule stated in (2.6.3) is known as a ***maximum a posteriori,*** or ***MAP, detector.***

If the prior probabilities are equal, as is normally assumed to be the case in digital transmission (otherwise the message should be further coded), then the optimal policy is to maximize $f(\mathbf{r} \mid S_i)$ over choices of message index i. This conditional density function is called the ***likelihood*** of $\mathbf{r}$, given S_i, and in this case the detector is referred to as ***maximum likelihood***, or ***ML.***

In summary, the rules are as follows:

$$\boxed{\begin{aligned} &\text{MAP}: \hat{S}_i = \arg\max_{S_i} P_i f(\mathbf{r} \mid S_i) \\ &\text{ML}: \hat{S}_i = \arg\max_{S_i} f(\mathbf{r} \mid S_i) \end{aligned}} \tag{2.6.6}$$

Clearly, if the priors P_i are equal, both procedures produce the same decision $\hat{S}_i$ for any specific $\mathbf{r}$. For unequal P_i, the two decision rules may produce different results, but we are assured that if the prior probabilities are correctly known the MAP rule will

have lower error probability than the ML procedure. If the priors are unknown, the usual choice is to use the ML policy.

Before proceeding, it is appropriate to emphasize the universality of these decision rules. Any digital communication decision task, whether involving a simple nonencoded binary signaling problem or an elaborate error control coding technique, ultimately reverts back to these procedures. The only steps in question are (1) how to formulate the required conditional probability density functions and (2) how to efficiently locate the maximizing S_i.

2.6.2 Irrelevant Data and Sufficient Statistics

In many everyday decisions we are presented with data, or observations, that have no bearing on our alternative choices. We could say such observations are irrelevant. Also, there are situations where the raw data itself are not essential for optimal decisions, but some reduced statistic, or function of the data, is adequate. A financial officer, in assessing whether to risk a loan, is perhaps only interested in our bottom-line assets, not how these are distributed among real estate, automobiles, savings account, and the like. Similar cases occur in digital communications as well, and it is important to recognize them, for much simpler processing can result.

Let $\mathbf{r}$ be an observation vector, related probabilistically to a choice of signals through $f(\mathbf{r} \mid S_i)$. We are interested in maximizing $P_i f(\mathbf{r} \mid S_i)$ over i. Suppose we partition the observation into two vectors $\mathbf{a}$ and $\mathbf{b}$ so that $\mathbf{r}$ is equivalent, within a permutation, to $(\mathbf{a}, \mathbf{b})$. Then we wish to maximize

$$P_i f(\mathbf{a}, \mathbf{b} \mid S_i) = P_i f(\mathbf{b} \mid \mathbf{a}, S_i) f(\mathbf{a} \mid S_i). \tag{2.6.7}$$

It may happen, through judicious choice of the partition, that $f(\mathbf{b} \mid \mathbf{a}, S_i)$ is invariant to S_i, in which case the middle term on the right-hand side in (2.6.7) is only a multiplying factor that scales equally for all hypotheses, and thus $\mathbf{b}$ may be safely disregarded as ***irrelevant.*** Only $\mathbf{a}$ is essential to the decision process. It is sometimes easy to recognize irrelevant data in a decision problem. If certain data are not irrelevant by inspection, simplification of the optimal decision rules will often expose irrelevant data. Exercises 2.6.7 and 2.6.8 will help to clarify the concept.

A closely related and more profound idea is that of a ***sufficient statistic.*** Instead of merely partitioning the observations as before, we can introduce a vector-valued function (or statistic) $\mathbf{g}(\mathbf{r})$ of the data $\mathbf{r}$ and think of the observation as $\tilde{\mathbf{r}} = (\mathbf{g}(\tilde{\mathbf{r}}), r_1, r_2, \ldots, r_n)$. (Certainly, we have not improved or diminished our decision-making ability by adding a completely determined relation on the original data to our observation vector.) We again wish to maximize

$$P_i f(\tilde{\mathbf{r}} \mid S_i) = P_i f(r_1, \ldots, r_n \mid \mathbf{g}(\mathbf{r}), S_i) f(\mathbf{g}(\mathbf{r}) \mid S_i). \tag{2.6.8}$$

Again, if, and only if, the middle term is invariant to S_i, then $(r_1, \ldots, r_n)$ as it stands may be ignored, with only $\mathbf{g}(\mathbf{r})$ preserved for the decision. $\mathbf{g}(\mathbf{r})$ is then termed a sufficient statistic for the problem. Ideally, $\mathbf{g}(\mathbf{r})$ will be a simple, perhaps scalar, function of the data. Finding sufficient statistics is often done by intuition, or by direct formulation of

the MAP/ML decision equations, simplifying where possible. If a sufficient statistic is proposed, it may be tested by asking whether the second term on the right in (2.6.8) is indeed constant with respect to i.

Example 2.24 Two Binary Signals in Independent Gaussian Noise

Let there be two equiprobable signals (hypotheses), formed as follows: to send S_0 we transmit $(-1,-1,-1)$ and to send S_1 we transmit $(1,1,1)$. The units could be volts or any other physical unit. We also assume both signals are equally likely to have been transmitted. The observation $\mathbf{r} = (r_1, r_2, r_3)$ is the result of adding independent Gaussian noise in each coordinate position of the signal, with the noise mean and variance taken as 0 and 1, respectively. In block diagram form we have the channel shown in Figure 2.6.4.

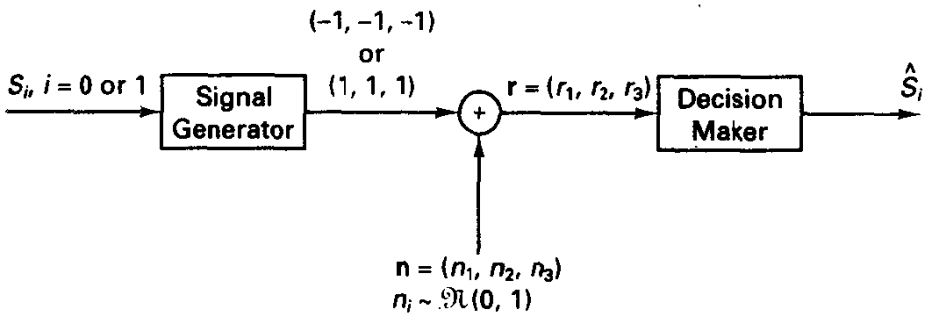


Figure 2.6.4 Binary signaling and detection in independent Gaussian noise.

By independence of the noise, the variables r_i, conditioned on either signal, are independent, and the required conditional probability densities may be written as

$$f(\mathbf{r} \mid S_0) = \prod_{j=1}^{3} f(r_j \mid S_0) = \prod_{j=1}^{3} \frac{1}{(2\pi)^{1/2}} e^{-(r_j+1)^2/2}$$
$$f(\mathbf{r} \mid S_1) = \prod_{j=1}^{3} f(r_j \mid S_1) = \prod_{j=1}^{3} \frac{1}{(2\pi)^{1/2}} e^{-(r_j-1)^2/2} \tag{2.6.9}$$

The assumption of equal prior probabilities implies that the ML test is optimal, and in the binary case we may compare the two likelihoods and decide in favor of the larger. We express this as

$$f(\mathbf{r} \mid S_1) \underset{S_0}{\overset{S_1}{\gtrless}} f(\mathbf{r} \mid S_0), \tag{2.6.10}$$

where the symbols attached to the inequalities denote the decision produced by the given inequality sense.

In this case we can further simplify (2.6.10) by taking logarithms of both sides. (Again, we may apply any monotonic function to the decision statistic without altering the decision, and the logarithm is such a function.) Doing so, we obtain the equivalent rule

$$\sum_{j=1}^{3} (r_j - 1)^2 \underset{S_1}{\overset{S_0}{\gtrless}} \sum_{j=1}^{3} (r_j + 1)^2. \tag{2.6.11}$$

[Note the reversal of the inequality sense in (2.6.11).] By recognizing the two sides of the comparison as squares of Euclidean distances, we can interpret the decision rule geometrically as ***choose that signal closest in Euclidean distance to the observed r.***

The decision is even simpler than (2.6.11) indicates, for upon expansion of the sums and cancellation of common terms from both sides, we can reduce the test to

$$T = \sum_{j=1}^{3} r_j \underset{S_0}{\overset{S_1}{\gtrless}} 0. \tag{2.6.12}$$

Equation (2.6.12) in effect tests to see on which side of the dividing plane $r_1 + r_2 + r_3 = 0$ the observation $\mathbf{r}$ lies. This is consistent with nearest-signal decoding obtained previously. The geometry of the problem and the dividing plane are indicated in Figure 2.6.5. This plane clearly is the partition boundary abstractly indicated in Figure 2.6.2.

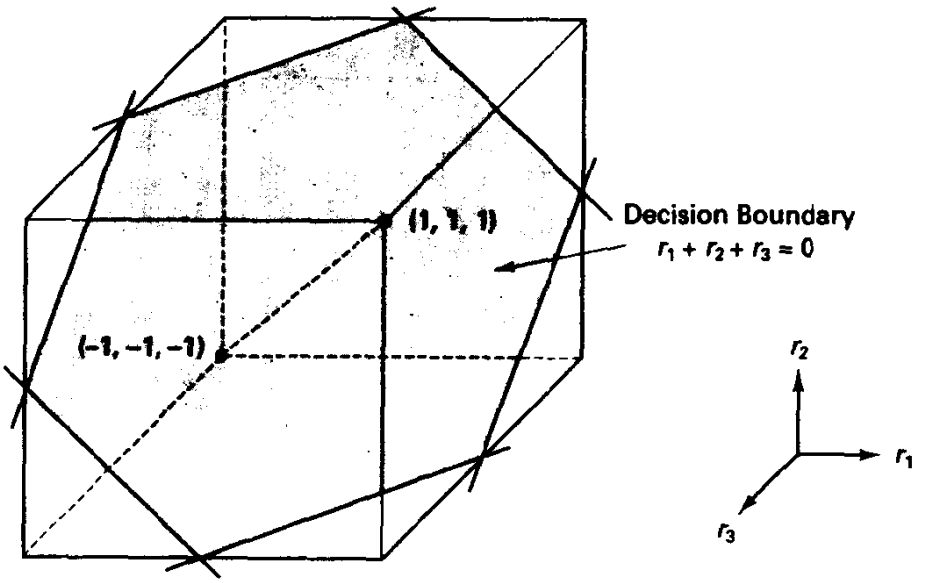


Figure 2.6.5 Optimal decision boundary is plane bisecting line connecting signals.

This example reveals that the data influence the decision only through the arithmetic sum and is an example of a sufficient statistic, discussed previously. $T = \sum r_j$ contains the essential ingredients for the optimal decision; it is not important to maintain r_1 by itself, for example. To formally confirm that the T is indeed a sufficient statistic, we could verify that $f(r_1, r_2, r_3 \mid r_1 + r_2 + r_3, S_i)$ is invariant to i by manipulating the p.d.f. using Bayes's rule.

Next, we compute the performance of this detector. The probability of error may be expressed using the law of total probability as

$$P(\epsilon) = \frac{1}{2} P(\epsilon \mid S_0) + \frac{1}{2} P(\epsilon \mid S_1). \tag{2.6.13}$$

Because of the symmetry evident in Figure 2.6.4, $P(\epsilon \mid S_0) = P(\epsilon \mid S_1)$, so $P(\epsilon) = P(\epsilon \mid S_0)$.

Conditioned on message S_0, $T = \sum r_j$ is Gaussian with mean -3 and variance 3. (Recall that in a sum of random variables, means are always additive, and the variances are

additive here by independence.) Thus,

$$\begin{aligned} P(\epsilon \mid S_0) &= P(T > 0 \mid S_0) \\ &= \int_0^\infty \frac{1}{(2\pi 3)^{1/2}} e^{-(t+3)^2/6}\, dt \\ &= \int_{3^{1/2}}^\infty \frac{1}{(2\pi)^{1/2}} e^{-y^2/2}\, dy, \end{aligned} \tag{2.6.14}$$

where the last step follows by change of variables, $y = (t+3)/3^{1/2}$. We previously defined the latter integral in terms of the $Q(x)$ function, (2.2.12). Thus, by (2.6.13) and (2.6.14), $P(\epsilon) = Q((3)^{1/2}) = 0.042$ from a table of $Q(x)$.

It is helpful to think of the signals corresponding to S_0 and S_1 as triplication of a basic signal, -1 and $+1$, respectively. Using three such transmissions is superior to one or two, and four, five, or more repetitions would lessen the error probability still further. This is a demonstration of the ability to "average out" the additive noise, at the expense of increased transmission time, fundamentally the law of large numbers at work. Prior to Shannon's work, this kind of ***repetition coding*** was thought to be the only way to improve the reliability. We now know much better ways to achieve high reliability without drastically sacrificing system throughput.

Example 2.25 Suboptimum Detection Applied to Example 2.24

Given the problem formulation of Example 2.24, it is tempting to think that the following procedure is best: make a binary decision on $r_i, i = 1, 2, 3$, based on the sign of r_i. This produces a vector of ± 1 values, and the decision can be based on a majority vote.

The decision boundary in three-dimensional space corresponding to this procedure is shown in Figure 2.6.6, which we note is vaguely similar to the optimal separating surface

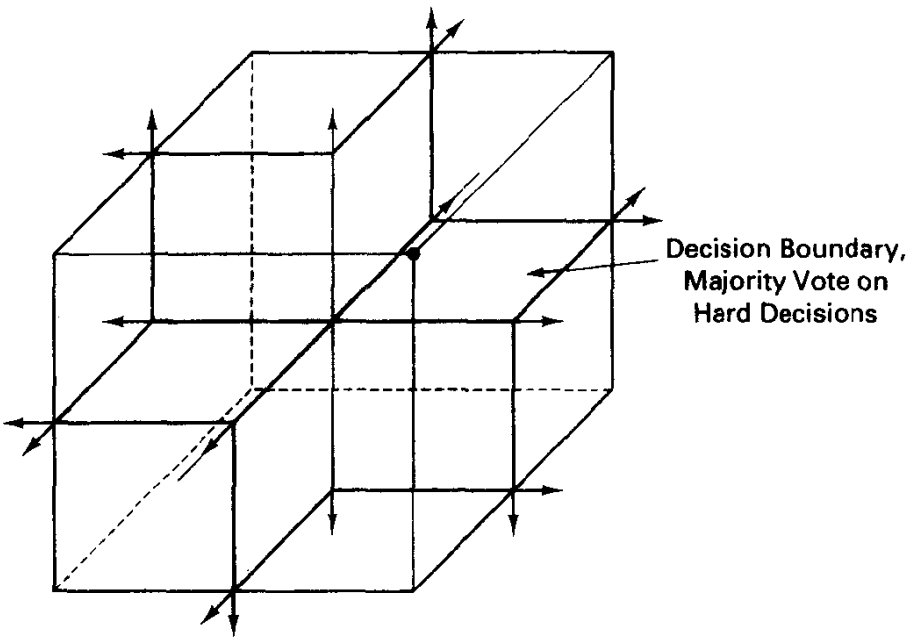


Figure 2.6.6 Separating surface defined by binary decision on each variable with majority voting.

of Figure 2.6.5. The error probability is, by virtue of symmetry,

$$\begin{aligned} P(\epsilon) &= P(\epsilon \mid S_0) \\ &= P[\text{two or more } r_i \geq 0 \mid (-1,-1,-1)\text{sent}] \\ &= C_2^3 p^2(1-p) + C_3^3 p^3, \end{aligned} \tag{2.6.15}$$

where p is the probability a single coordinate is decided incorrectly. Since this is a Gaussian noise setting, the coordinate error probability is

$$\begin{aligned} p &= \int_0^\infty \frac{1}{(2\pi)^{1/2}} e^{-(r+1)^2/2}\, dr \\ &= Q(1) = 0.1587 \end{aligned} \tag{2.6.16}$$

after a change of variables. Thus, upon substitution in (2.6.15), $P(\epsilon) = 0.0675$, which we note is larger than obtained with the optimal decision policy. In essence, in performing what we shall eventually call ***hard-decision decoding***, the decoder has discarded important likelihood information contained in the size of the observations.

It is apparent that some noise vectors cause hard-decision decoding to err, while ML decoding succeeds. For example, selection of message $(-1,-1,-1)$ combined with noise vector $\mathbf{n} = (1.1, 1.2, -0.1)$ produces $\mathbf{r} = (0.1, 0.2, -1.1)$, which is decoded correctly by ML decoding but not by majority voting. Equivalently, the point $\mathbf{r}$ is on the proper side of the ML decision surface, but on the wrong side of the majority-voting surface. We should not conclude, however, that the ML decoder will never err when the suboptimum decoder is correct; consider transmission of the same message with $\mathbf{n} = (0.1, 0.2, 2.0)$, which might be said to include one especially bad noise sample. Here majority voting is correct, while the ML test errs. We simply conclude that the probability of the former situation is greater than that of the latter noise types, under the adopted model, and thus the superior performance of the ML decoder on average.

Example 2.26 Photon Counting

Let's revisit the optical PPM modulation technique introduced in Figure 1.1.2. There are $M = 8$ message hypotheses in each signaling interval T_s, and the message is communicated by transmitting optical energy in one of the 8 slots. Suppose the receiver is a direct detection system, essentially counting optical-frequency photons in each slot. Thus, the observation is the 8-vector of photon counts, $(k_1, k_2, \ldots, k_8)$.

We model the signal's photon arrival process as a Poisson point process with mean arrival rate λ_s photons per unit time. The average energy per slot attached to such a signal would be $hf\lambda_s T$, since hf is the energy of a photon with frequency f, and $\lambda_s T$ is the mean number of signal photons per slot.[18]

Because of background radiation, slots without signal can register photon counts. Furthermore, due to the quantum effect, slots designated as signal bearing may produce zero photon counts! We let the Poisson rate parameter be λ_n for such slots. In a signal-bearing slot, the Poisson parameter will be $\lambda_s + \lambda_n$. In either case, the number of counts is a ***Poisson random variable*** with probability mass function

$$P_K(k \mid \text{signal}) = \frac{[(\lambda_s + \lambda_n)T]^k e^{-(\lambda_s+\lambda_n)T}}{k!}, \qquad k = 0, 1, \ldots, \tag{2.6.17a}$$

[18] h is Planck's constant, $6.6 \cdot 10^{-34}$

and

$$P_K(k \mid \text{no signal}) = \frac{(\lambda_n T)^k e^{-\lambda_n T}}{k!}, \qquad k = 0, 1, \ldots. \tag{2.6.17b}$$

Figure 2.6.7 illustrates these two p.m.f.'s for a case with $\lambda_n T = 1, \lambda_s T = 3$.

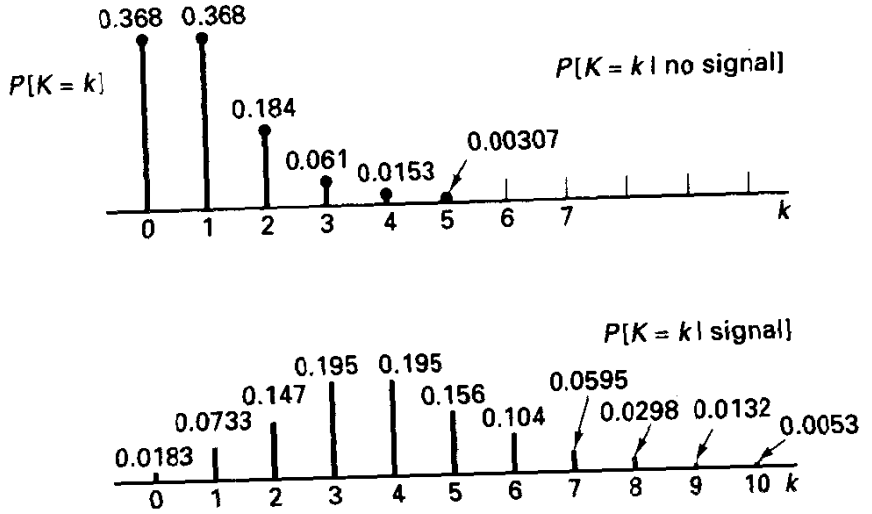


Figure 2.6.7 Probability mass functions for Poisson random variables of Example 2.26.

Finally, we claim that counts in the various slots will be independent r.v.'s (this is a basic property of the Poisson point process). This allows us to construct the conditional p.d.f. for the observation, under each hypothesis, as

$$P_{K_1,K_2,\ldots,K_8}(k_1, k_2, \ldots, k_8 \mid \text{signal in slot } i) = \frac{[(\lambda_s + \lambda_n)T)]^{k_i} e^{-(\lambda_s+\lambda_n)T}}{k_i!} \cdot \prod_{p \neq i} \frac{(\lambda_n T)^{k_p} e^{-\lambda_n T}}{k_p!}. \tag{2.6.18}$$

The ML rule then reduces to, after simple algebraic manipulation,

$$\underset{i}{\text{maximize}} \left[\frac{\lambda_s + \lambda_n}{\lambda_n}\right]^{k_i} \tag{2.6.19}$$

which in turn implies that the decision should be in favor of the slot with the largest photon count. Thus, a sufficient statistic is the *index* of the slot with largest count; no other data are necessary. If ties exist, we can break the tie in any reasonable way.

Visualizing decision regions in observation space is difficult here due to the eight-dimensional space involved. However, the decision region for D_0 is the set of all 8-tuples for which k_1 is the largest, and so on. Observation vectors for which two or more slot counts tie can be arbitrarily assigned to one of the competing choices.

To evaluate the error probability of this decision process, we can assume that energy was transmitted in the first slot. Then we would need to calculate the probability that *any* count $k_2, k_3, \ldots, k_8$ exceeds k_1 (or equals k_1 to be pessimistic toward tie breaking).

It is easier instead to calculate the probability of correct decision. We can do this by first conditioning on a specific value for k_1 and calculating the probability that all the other

counts are less than k_1; we then weight these conditional error probabilities by $P(K_1 = k_1)$ and sum. Exercise 2.6.5 pursues this further.

A slightly different interpretation of the optimal decision rule is provided by the calculation of ***likelihood ratios.*** We recall $f(\mathbf{r} \mid S_i)$ is the likelihood for signal S_i, and we define

$$L_i(\mathbf{r}) = \frac{f(\mathbf{r} \mid S_i)}{f(\mathbf{r} \mid S_0)}, \qquad i = 1, 2, \ldots, M-1 \tag{2.6.20}$$

to be the likelihood ratio, with respect to signal S_0, for the ith signal. Note $L_i(\mathbf{r})$ is a scalar function of a vector random variable $\mathbf{r}$, mapping observation space to the interval $[0, \infty)$.

One way of implementing the optimal decision policy is shown in Figure 2.6.8. We first compute the $M-1$ likelihood ratios, forming the likelihood ratio vector $\mathbf{L} = (L_1, L_2, \ldots, L_{M-1})$, which can be viewed as a point in $(M-1)$-dimensional space, called *likelihood ratio space.* The likelihood ratio vector can be viewed as a random $(M-1)$-dimensional vector obtained by a nonlinear transformation on the observation vector $\mathbf{r}$.

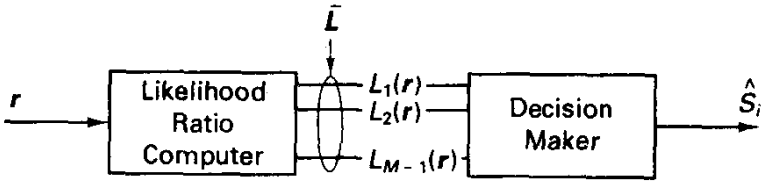


Figure 2.6.8 Likelihood ratio version of optimal processor.

The decision is based on $\mathbf{L}$ as follows:

1. Choose S_0 if all components of $\mathbf{L}$ are less than 1.
2. Otherwise, choose the index of the largest entry in L.

It should be clear that this procedure is equivalent to computing the likelihoods and then choosing the index of the largest. Thus, equivalent decisions are made by properly partitioning either observation space or likelihood ratio space. Stated another way, the likelihood ratio vector $\mathbf{L}$ is always a sufficient statistic for the decision problem.

An appealing geometric aspect of the likelihood ratio perspective is that, whereas decision boundaries in observation space are usually oddly shaped regions, the decision regions in likelihood-ratio space are always defined by *fixed hyperplanes* in $(M-1)$-dimensional space, regardless of the probability densities $f(\mathbf{r} \mid S_i)$ of the problem. Figures 2.6.9a and 2.6.9b illustrate likelihood ratio spaces for $M = 3$ and $M = 4$ cases, along with the separating planes.

We could just as well work with the log-likelihoods, since the logarithm is a monotone increasing function of its argument, and define

$$\mathbf{Z} = [\log_e L_1(\mathbf{r}), \log_e L_2(\mathbf{r}), \ldots, \log_e L_{M-1}(\mathbf{r})] \tag{2.6.21}$$

and decide as follows:

1. Choose S_0 if all components of $\mathbf{Z}$ are negative.

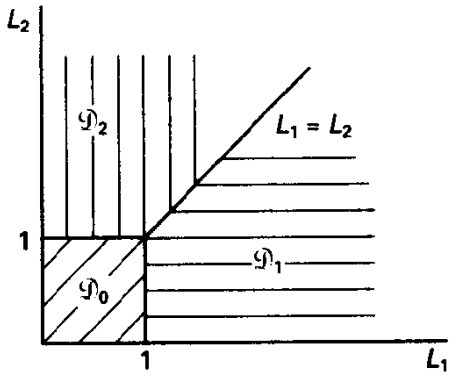


Figure 2.6.9a Decision regions in likelihood ratio space, $M = 3$.

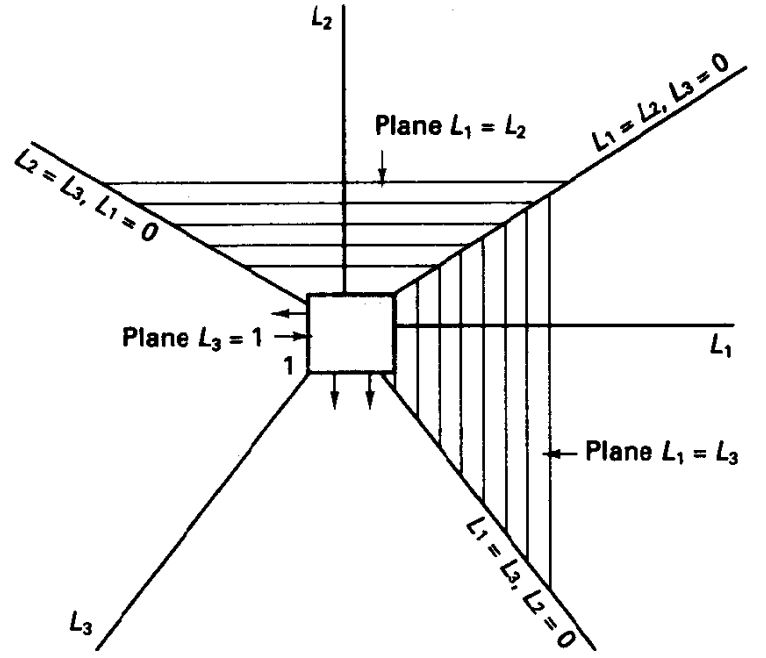


Figure 2.6.9b View into D_3 in likelihood ratio space for $M = 4$; region is in positive orthant and bounded by three planes.

2. Or choose the index of the most positive entry in $\mathbf{Z}$.

Obviously, $\mathbf{Z}$ is a sufficient statistic as well.

We conclude this section by developing a general bound on error probability for a binary decision problem, linking the concepts of likelihood ratios and the Chernoff bound developed in Section 2.4. For a two-hypothesis problem, we note that a sufficient statistic is

$$Z_1 = \log_e \left[\frac{f(\mathbf{r} \mid S_1)}{f(\mathbf{r} \mid S_0)} \right] \tag{2.6.22}$$

and the test is simply to compare Z_1 with 0. Thus, the error probability, given that S_0 was selected, is

$$P(\epsilon \mid S_0) = P(Z_1 > 0 \mid S_0) \leq \min_{s>0} E[e^{sZ_1} \mid S_0] \tag{2.6.23}$$

where the last step follows from a Chernoff bound. Substituting the definition of the random variable Z_1, we obtain

$$P(\epsilon \mid S_0) \leq \min_s E\left[\frac{f(\mathbf{r} \mid S_1)}{f(\mathbf{r} \mid S_0)} \mid S_0\right]^s \tag{2.6.24}$$

(We should interpret the expectation as being with respect to the random vector $\mathbf{r}$, conditioned upon selection of S_0.) Thus, the conditional error probability becomes, from the definition of expectation,

$$\begin{aligned} P(\epsilon \mid S_0) &\leq \min_{s>0} \int \left[\frac{f(\mathbf{r} \mid S_1)}{f(\mathbf{r} \mid S_0)}\right]^s f(\mathbf{r} \mid S_0)\, d\mathbf{r} \\ &= \min_s \int f(\mathbf{r} \mid S_0)^{1-s} f(\mathbf{r} \mid S_1)^s\, d\mathbf{r}. \end{aligned} \tag{2.6.25}$$

A similar expression follows for the conditional error probability given S_1 is selected for transmission, except the conditional density functions are interchanged. Assuming the two messages have equal prior probabilities, we then have that the unconditional error probability is bounded by

$$P(\epsilon) \leq \frac{1}{2} \min_{s>0} \left[\int f(\mathbf{r} \mid S_0)^{1-s} f(\mathbf{r} \mid S_1)^s\, d\mathbf{r} + \int f(\mathbf{r} \mid S_1)^{1-s} f(\mathbf{r} \mid S_0)^s\, d\mathbf{r}\right]. \tag{2.6.26}$$

The minimization with respect to s can be performed in principle, once the two p.d.f.'s are specified, but at least the bound is valid when we set $s = \frac{1}{2}$, leaving the compact expression

$$P(\epsilon) \leq \int f(\mathbf{r} \mid S_0)^{1/2} f(\mathbf{r} \mid S_1)^{1/2}\, d\mathbf{r}. \tag{2.6.27}$$

This bound depends only on the form of the two density functions and does not require formal description of the decision regions and the ability to perform integrals over complicated decision zones. This type of performance bound will be encountered again in Chapter 4 in our introduction to coded communications. It is known that this bound gives the tightest exponential form for error probability, if one exists, provided the optimization with respect to s is accomplished. Exercise 2.6.10 treats this approach for the two-signal problem introduced in Example 2.24.

2.7 CONCEPTS OF INFORMATION THEORY FOR DISCRETE ALPHABETS

In popular usage the term *information* is broadly understood but elusive to define. However, information has a precise meaning to a communication theorist, expressed solely in terms of probabilities of source messages and actions of the channel. In this section we develop the Shannon notion of information by introducing various entropy (or uncertainty) measures associated with the communication process and then define information exchange as a reduction in entropy. Following this, we demonstrate through source and channel coding theorems that these measures are, in fact, important quantities for communications purposes. Our initial treatment is confined to discrete-alphabet situations; extension to the case of continuous random variables and processes is made in Section 2.9.

2.7.1 Entropy for Discrete Random Variables

Consider a discrete scalar random variable X, which we might regard as an output of a discrete message source. Suppose the variable X can assume one of K outcomes, labeled $x_i, i = 0, 1, \ldots, K-1$, with probabilities specified by $P_X(x_i)$. As shorthand notation, these will also be designated P_i. We define the ***entropy*** of the random variable X to be

$$\begin{aligned} H(X) &= \sum_{i=0}^{K-1} P_X(x_i) \log \frac{1}{P_X(x_i)} \\ &= \sum_i P_i \log \frac{1}{P_i} = -\sum_i P_i \log P_i. \end{aligned} \tag{2.7.1}$$

The selection of H to denote entropy is now conventional and dates to Boltzmann's work in the field of statistical thermodynamics. Before developing the properties of this entropy function and in fact justifying its usefulness, we observe that $H(X)$ is the expected value of the random variable $\log(1/P_i)$, which some authors denote as the self-information of the outcome x_i. However, we shall reserve the meaning of information to be distinctly tied to a reduction in entropy, rather than an intrinsic property of messages.

Although by no means justifying entropy as important for the communications process, we can argue that it is a proper measure of ***prior uncertainty*** of an experiment. We begin by denoting $H(X) = H(P_0, P_1, \ldots, P_{K-1})$ to explicitly indicate the functional dependence on probabilities $P_0, \ldots, P_{K-1}$. Now we stipulate some properties that an uncertainty measure should possess.

Property 1 Continuity

$H(P_0, \ldots, P_{K-1})$ should be continuous in all its variables; that is, small changes in the probability model should have small effect on the uncertainty.

Property 2 Monotonicity

$H(1/K, 1/K, \ldots, 1/K)$ should be monotone increasing in K, meaning that, if the outcomes are equiprobable, increasing their number should increase the uncertainty.

Property 3 Additivity for independent experiments

If the random variable X is bivariate, $X = (Y, Z)$, with Y and Z independent, then we desire that $H(X) = H(Y) + H(Z)$. In words, the uncertainty of a joint experiment that involves independent random variables should be the sum of the respective uncertainties of the component experiments.

Property 4 Grouping

Suppose the outcomes x_i are assigned to two disjoint events A and B as shown in Figure 2.7.1, with probabilities $P_A = P_0 + P_1 + \cdots + P_l$ and $P_B = P_{l+1} + \cdots + P_{K-1}$, respectively. We desire that the overall uncertainty be representable in hierarchical manner: there is a component of overall uncertainty due to the uncertainty of group membership and one due to uncertainty remaining after group membership is identified.

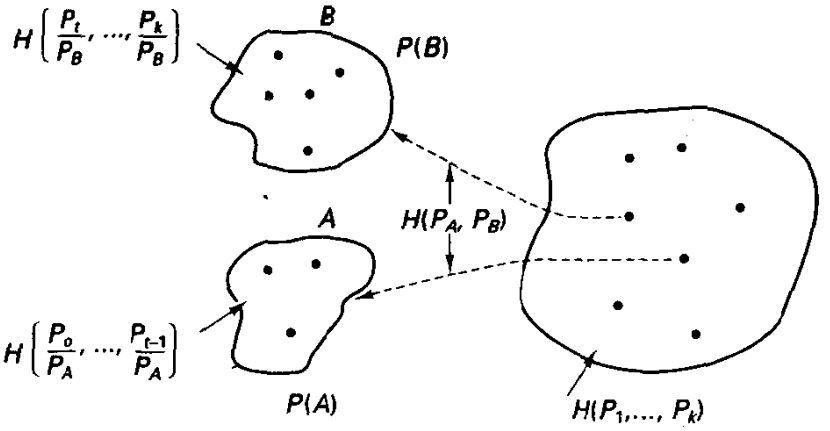


Figure 2.7.1 Interpretation of grouping axiom.

The first term is the uncertainty (entropy) of a *binary* experiment with probabilities P_A and P_B, while the second is just the total of probability-weighted uncertainties attached to the groups, or subexperiments. Functionally, this requires

$$H(P_0, \ldots, P_{K-1}) = H(P_A, P_B) + P_A H\left(\frac{P_0}{P_A}, \ldots, \frac{P_t}{P_A}\right) + P_B H\left(\frac{P_{t+1}}{P_B}, \ldots, \frac{P_{K-1}}{P_B}\right) \tag{2.7.2}$$

The fact that the function in (2.7.1) satisfies these four properties is readily demonstrated. The more interesting fact is that (2.7.1) is the *only* such functional of probabilities (to within a scale factor associated with choice of logarithm base) that satisfies these four desired properties. Proof of this uniqueness is provided in Shannon's original work [11].

Regarding the logarithm base in (2.7.1), two choices are prevalent in information theory: base 2, in which case the units of entropy are ***bits***, and base e where the units are ***nats***, for natural units. Because the practice of information theory occurs in a world of binary machines, *we shall assume base 2 logarithms throughout, unless otherwise stated.* Also regarding logarithms, we define $0 \cdot \log_2 0$ to be zero. Equivalently, we could omit zero probability events from the definition of entropy.

Next, we develop a further property of entropy:

$$0 \leq H(X) \leq \log K. \tag{2.7.3}$$

Equality on the left-hand side occurs when (and only when) one of the messages or outcomes has probability 1 (note this is not equivalent to saying the sample space contains only one outcome), in which case there is zero uncertainty according to the definition. Equality on the right-hand side of (2.7.3) is obtained if and only if[19] the outcomes are equiprobable. The calculus of variations furnishes a direct way to demonstrate that the equiprobable assignment attains an extremum; then we can verify the solution is in fact a maximum. A proof that is less direct, but more useful in the entire development, is

[19]This will often be abbreviated as iff.

based on the ***information theory inequality***, so called because of its frequent appearance in proofs of information-theoretic results.

Lemma (Information Theory Inequality). For $z > 0$, $\log z \leq (z-1)/\log_e 2$, with equality iff $z = 1$.

Proof. It is simple to verify the equality condition and then that $\log z$ is a convex $\cap$ function of z for positive z by differentiation.[20] Graphical interpretation of the lemma is shown in Figure 2.7.2.

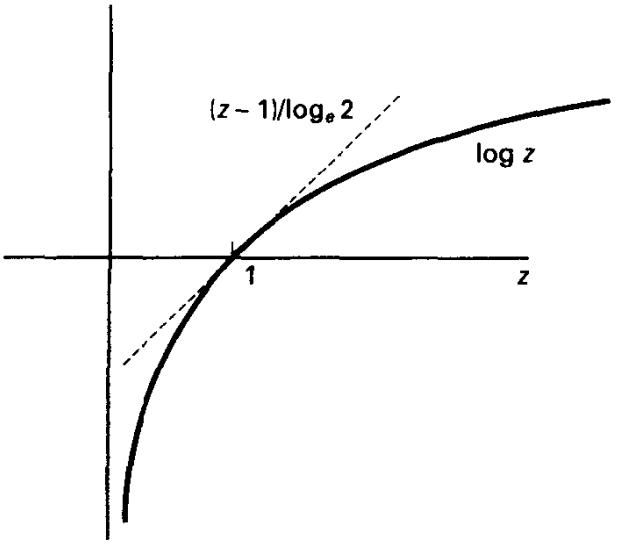


Figure 2.7.2 Illustration of information theory lemma.

To prove $H(X) \leq \log K$, with equality iff $P_i = 1/K$, we show $H(X) - \log K \leq 0$.

$$H(X) - \log K = \sum_i P_i \log \frac{1}{P_i} - \left(\sum_i P_i\right) \log K. \tag{2.7.4a}$$

We consider the sum to only involve terms for which $P_i > 0$, so that the previous lemma may be applied, and obtain

$$\begin{aligned} H(X) - \log K &= \sum_i P_i \log \frac{1}{KP_i} \\ &\leq \frac{\sum_i P_i\left(\frac{1}{KP_i} - 1\right)}{\log_e 2} \\ &= \frac{\left(\sum_{i=0}^{K-1} \frac{1}{K} - \sum_{i=1}^{K} P_i\right)}{\log_e 2} = 0. \end{aligned} \tag{2.7.4b}$$

The proof also demonstrates that $H(X) = \log K$ iff $P_i = 1/K$ for all i.

[20] A function $f(y)$ of a scalar variable y is convex $\cap$ if $\alpha f(y_1) + (1-\alpha) f(y_2) \leq f[\alpha y_1 + (1-\alpha) y_2]$ for any y_1, y_2 in the domain of the function and any $0 \leq \alpha \leq 1$. Equivalently, provided it exists, the second derivative of the function is less than or equal to zero.

Example 2.27 Entropy Function for Binary and Ternary Sources

In the case of a binary random variable taking on two values with probability p and $1 - p$, substitution into (2.7.1) yields

$$H(X) = -p \log p - (1 - p) \log(1 - p). \tag{2.7.5}$$

This function will be subsequently referred to as the ***binary entropy function*** and denoted by $h_2(p)$. The binary entropy function is sketched in Figure 2.7.3a as a function of the single parameter p.

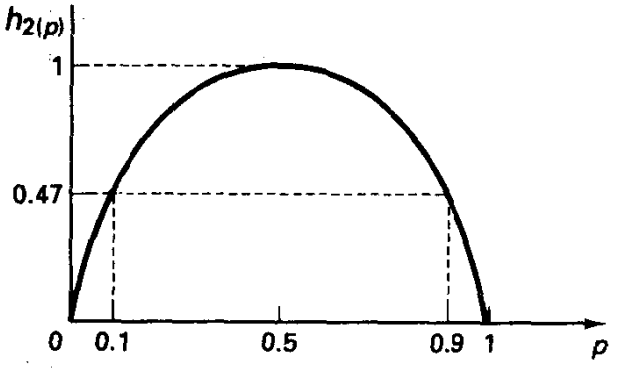


Figure 2.7.3a Binary entropy function, $h_2(p)$.

For ternary ($K = 3$) sources, the entropy is a function of two variables, since the third probability is constrained by the first two. This is also illustrated in Figure 2.7.3b. Note in each case the location of the maximizing probability assignment (that is, equiprobable), as well as the convexity of the surface over the region of probability assignments. We shall not prove the convexity property; the interested reader is referred to Gallager's text [12].

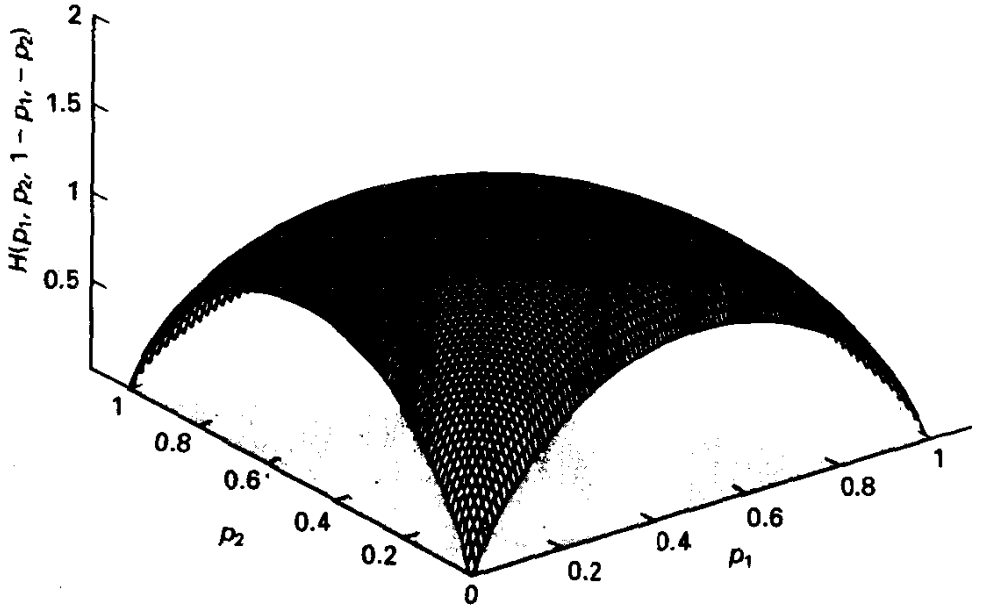


Figure 2.7.3b Ternary entropy function versus p_1, p_2. Maximum is at $p_1 = p_2 = \frac{1}{3}$.

2.7.2 Joint and Conditional Entropy

We now consider a bivariate discrete random variable (X, Y) having a specified joint distribution $P(x_i, y_j), i = 0, 1, \ldots, K-1, j = 0, 1, \ldots, J-1$, from which marginal and conditional probabilities may be derived. The joint entropy, $H(X, Y)$ is, in keeping with (2.7.1),

$$H(X, Y) = \sum_i \sum_j P(x_i, y_j) \log \frac{1}{P(x_i, y_j)}. \tag{2.7.6}$$

Property 3 held that if X and Y are independent then the joint entropy is the sum of the individual entropies. In general, however,

$$H(X, Y) \leq H(X) + H(Y) \tag{2.7.7}$$

with equality only when X and Y are independent (see Exercise 2.7.1).

Next we consider the uncertainty (or entropy) associated with X, given that $Y = y_j$ is specified. Again, using the earlier definition, we express this as a weighted sum of $\log[1/P(x_i \mid y_j)] = -\log[P(x_i \mid y_j)]$:

$$H(X \mid Y = y_j) = -\sum_{i=0}^{K-1} P(x_i \mid y_j) \log P(x_i \mid y_j), \tag{2.7.8}$$

which remains a function of the conditioning outcome y_j. Then, to obtain the ***conditional entropy*** $H(X \mid Y)$, we average with respect to Y:

$$\begin{aligned} H(X \mid Y) &= \sum_j P(y_j) H(X \mid Y = y_j) \\ &= -\sum_i \sum_j P(x_i, y_j) \log P(x_i \mid y_j). \end{aligned} \tag{2.7.9}$$

This may be interpreted in the communication context by letting X be the input to a noisy transmission channel and Y the output. Then $H(X \mid Y)$ will be the uncertainty about the input message after the channel output is observed, averaged over input selections and channel actions. $H(X \mid Y)$ is sometimes called the ***equivocation.***

2.7.3 Mutual Information

Shannon defined information as follows: the ***(average) mutual information*** shared between random variables X and Y is

$$I(X; Y) = H(X) - H(X \mid Y); \tag{2.7.10}$$

that is, the information Y reveals about X is the prior uncertainty in X, less the posterior uncertainty about X after Y is specified. From this definition we have

$$\begin{aligned} I(X; Y) &= -\sum_i P(x_i) \log P(x_i) + \sum_i \sum_j P(x_i, y_j) \log P(x_i \mid y_j) \\ &= -\sum_i \sum_j P(x_i, y_j) \log P(x_i) + \sum_i \sum_j P(x_i, y_j) \log P(x_i \mid y_j) \\ &= \sum_i \sum_j P(x_i, y_j) \log \left[\frac{P(x_i \mid y_j)}{P(x_i)} \right] \end{aligned} \tag{2.7.11}$$

Using the definition of conditional probability, we can write (2.7.11) as

$$I(X;Y) = \sum_i \sum_j P(x_i, y_j) \log \left[\frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right], \tag{2.7.12}$$

which explicitly indicates that $I(X;Y) = I(Y;X)$; the information function is symmetric, and X gives the same information about Y as Y does about X! The symmetry is also stated by

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X). \tag{2.7.13}$$

Mutual information is nonnegative, being zero only when X and Y are independent. (This is consistent with a heuristic notion of information: if one random experiment is independent of another, knowing the outcome of either provides no information, or no reduction in uncertainty, about the outcome of the other.) To demonstrate this formally, we again appeal to the information inequality:

$$\begin{aligned} -I(X;Y) &= \sum_i \sum_j P(x_i, y_j) \log \left[\frac{P(x_i)}{P(x_i \mid y_j)} \right] \\ &\leq \frac{\sum_i \sum_j P(x_i \mid y_j) P(y_j) \left[\frac{P(x_i)}{P(x_i \mid y_j)} - 1 \right]}{\log_e 2} \\ &= \frac{\sum_i \sum_j P(x_i)P(y_j) - \sum_i \sum_j P(x_i \mid y_j)P(y_j)}{\log_e 2} = 0. \end{aligned} \tag{2.7.14}$$

[We have again considered the original double summation in (2.7.14) to be over those (i, j) indexes for which $P(x_i, y_j) > 0$ so that the previous lemma may be applied.] Equality occurs if, and only if, $P(x_i) = P(x_i \mid y_j)$ for all i, j such that $P(y_j) \neq 0$, which is to say if, and only if, X and Y are independent.

As a further consistency check, since information is nonnegative, (2.7.10) and the result of (2.7.14) give that entropy is always at least as large as a conditional entropy, so conditioning typically reduces uncertainty; that is,

$$H(X \mid Y) \leq H(X). \tag{2.7.15}$$

Example 2.28 Binary Symmetric Channel Revisited

Consider again the system of Example 2.3, shown in Figure 2.7.4 in schematic form. The joint probabilities of the four outcomes are tabulated. We proceed to calculate $H(X)$ and $I(X;Y)$. First,

$$\begin{aligned} H(X) &= -0.4 \log 0.4 - 0.6 \log 0.6 \\ &= 0.97 \text{ bit/symbol}, \end{aligned}$$

which is virtually as large as the 1 bit/symbol entropy for the equiprobable binary random variable. To find $I(X;Y)$, we could determine $H(X \mid Y)$ and subtract this from $H(X)$, or we can use the alternate form $I(X;Y) = H(Y) - H(Y \mid X)$. Using the latter method, we find that $H(Y) = 0.982$ bit, and from (2.7.8), $H(Y \mid X) = 0.469$ bit (note this is the *uncertainty about the output* Y, *conditioned on either input value* x_i, *by symmetry*). Thus, $I(X;Y) = 0.982 - 0.469 = 0.513$ bit/channel usage. We also remark that if the input

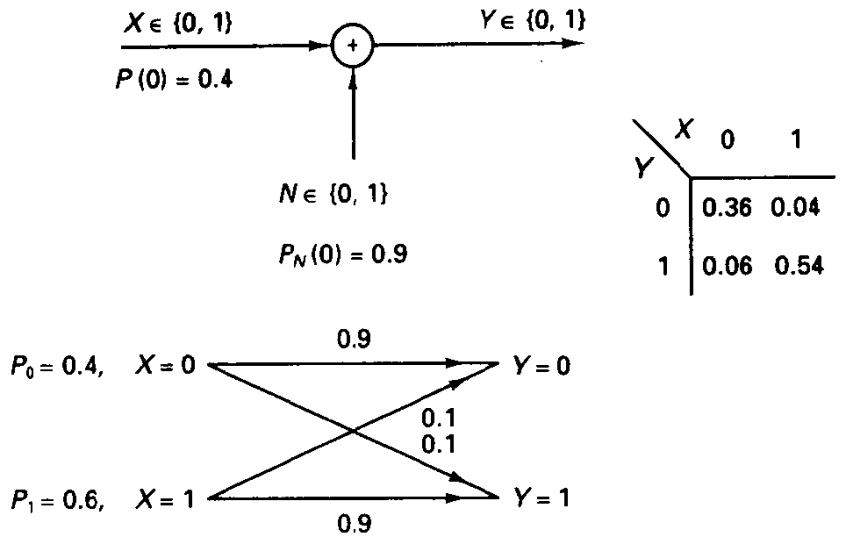


Figure 2.7.4 Channel models for Example 2.23.

selection was equiprobable the mutual information increases slightly to 0.531 bit/channel usage. Assuming capacity relates to information-passing ability, it appears the channel here is not capable of throughput anything near 1 bit per use, but roughly half that!

Notice from (2.7.12) that average mutual information is an expectation of a random variable $\log(P(x_i, y_j)/P(x_i)P(y_j))$, which is a function of the original random variable outcomes. Thus, we could define this latter quantity as the random information $I(x_i; y_j)$ associated with specific outcomes, and indeed this is a possible starting definition for the development we have just presented. In this context, if the output y_j specifies the input x_i with probability 1; that is, $P(x_i \mid y_j) = 1$, then $I(x_i; y_j) = I(x_i)$ and

$$I(x_i) = -\log P(x_i), \tag{2.7.16}$$

which is called the ***self-information*** of the outcome x_i. The entropy then can be defined as the expected value of this random variable, again giving (2.7.1). This notion, however, fosters confusion of information with uncertainty—a proper interpretation is that information is a reduction in uncertainty.

Whereas $I(X; Y)$ is nonnegative, the *event information* $I(x_i; y_j)$ can be negative. In the previous example, $I(x = 0; y = 1)$ has a negative value, since the joint probability of this event is smaller than the product of marginal probabilities. We should simply interpret such situations to be negatively informing, or misleading. On the whole, however, one *experiment* or random variable is not misleading about another, although it may provide, at worst, zero information.

2.7.4 Discrete Channels and Channel Capacity

The previous information relations have been developed in a general probabilistic setting. To pursue communications applications, we introduce the notion of a discrete channel having an input alphabet of size M and an output alphabet of size Q, both finite. Such

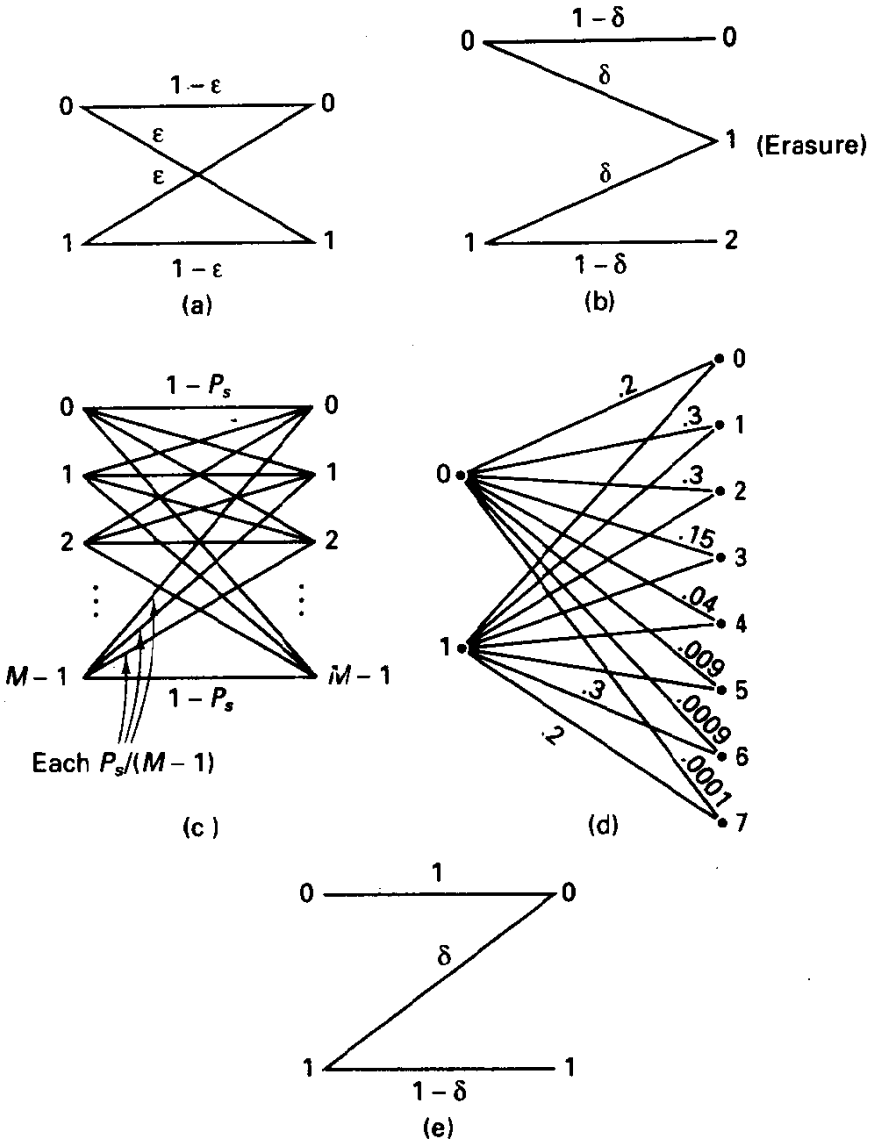


Figure 2.7.5 Some discrete channels: (a) Binary symmetric channel (BSC); (b) binary erasure channel (BEC); (c) M-ary uniform channel (MUC); (d) $M = 2$, $Q = 8$ channel; (e) Z channel.

a channel is specified by its transition probabilities, $P(y_j \mid x_i)$, typically depicted with a line diagram of the form shown in Figure 2.7.5. There we show several simple but prominent discrete channel models:

1. The binary symmetric channel (BSC) previously introduced in Example 2.3
2. The binary erasure channel (BEC)
3. The M-ary uniform channel (MUC)
4. A finely quantized channel with $M = 2$, $Q = 8$
5. The Z-channel

The BEC model arises in situations where "errors" are not made per se; instead, channel dropouts or side information that a decision on a given symbol would be very unreliable produces an ***erasure output.*** This is perfectly acceptable when the message is a redundant sequence of symbols, since other transmissions can perhaps resolve the message ambiguity. The MUC is simply the M-ary extension of the BSC, with the symbol error probability P_s distributed uniformly among $M-1$ error possibilities. The $M=2$, $Q=8$ example illustrates that the channel output alphabet may be much larger than the input alphabet, as in the case of a binary input, Gaussian noise channel that is finely quantized in the demodulator. Finally, the Z-channel is asymmetric in its action on inputs, in distinction to the other models. While channels such as this are of largely academic interest, the Z-channel does surface as a model for optical communications (photon counting with no background radiation) and for semiconductor memory error processes, in which memory cell errors of type $0 \to 1$, say, predominate over the alternative error type.

Given a discrete channel model, imagine the transmission of *one* symbol. (We call this a *channel use*.) The input is randomly selected according to a distribution designated $P(x_i)$, and every such choice of distribution induces an average mutual information given by an alternative form of (2.7.11):

$$I(X;Y) = \sum_i \sum_j P(x_i)P(y_j \mid x_i) \log \left[\frac{P(y_j \mid x_i)}{\sum_k P(x_k)P(y_j \mid x_k)} \right]. \tag{2.7.17}$$

Here we have written mutual information in a form so that the dependence on the channel, $P(y_j \mid x_i)$, and on the input distribution, $P(x_i)$, is explicitly shown.

The most celebrated quantity of information theory is the ***channel capacity***, C, defined as the maximum mutual information over all input distributions:

$$\boxed{C = \max_{P(x)} I(X;Y) \text{ bits/channel use.}} \tag{2.7.18}$$

We emphasize that the number C is only a function of the channel description.

Determination of C requires a constrained maximization of a function $I(X;Y)$ that is convex $\cap$ over the space of input probability vectors, [12]. Thus, standard numerical optimization methods can determine C for arbitrary channels (see [13] for an iterative algorithm that finds C and the maximizing distribution). In practice, however, we are most frequently dealing with channels that are *symmetric* in the following sense: if we write the transition probabilities in a M by Q matrix, we can partition the matrix, perhaps after rearranging columns, into submatrices such that within each submatrix all rows are permutations of each other, and likewise for columns. This construction for the BEC is shown in Figure 2.7.6. By this definition all channels of Figure 2.7.5 except the Z-channel are symmetric. (Some outwardly symmetric channels are not, as shown in Exercise 2.7.7). For such symmetric channels, we have the following results (see for example Gallager [12]):

1. The equiprobable input assignment, $P(x_i) = 1/M$, achieves C, that is, produces the largest mutual information.
2. The resulting capacity C is the mutual information between any specific input x_i which has $p(x_i) \to 0$ and the output r.v. Y.

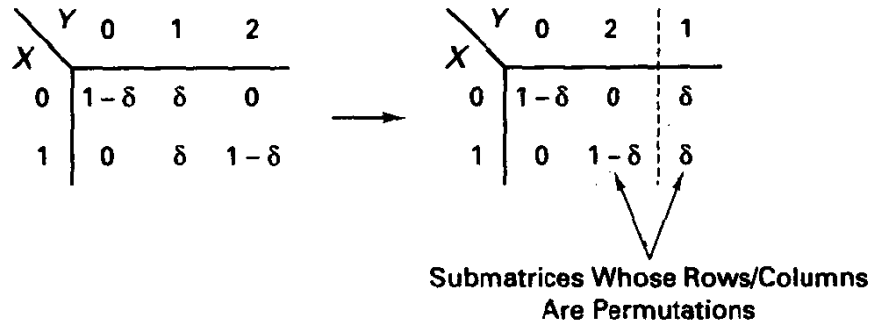


Figure 2.7.6 Reordering outputs to show BEC is symmetric.

Computation of C for the symmetric channels of Figure 2.7.5 is straightforward using $C = H(X) - H(X \mid Y)$ under the equiprobable input distribution. The results are

$$\text{BSC:} \quad C_{\text{BSC}} = 1 + \epsilon \log(\epsilon) + (1 - \epsilon)\log(1 - \epsilon) \tag{2.7.19a}$$

$$= 1 - h_2(\epsilon) \text{ bits/channel use}$$

$$\text{BEC:} \quad C_{\text{BEC}} = 1 - \delta \text{ bits/channel use} \tag{2.7.19b}$$

$$\text{MUC:} \quad C_{\text{MUC}} = \log M + P_s \log \frac{P_s}{(M-1)} \tag{2.7.19c}$$

$$+ (1 - P_s)\log(1 - P_s) \text{ bits/channel use}$$

The expression for the finely quantized channel involves all the transition probabilities and cannot be put into compact form. For the asymmetric Z-channel, optimization of the input probabilities must be performed to establish C. These issues are addressed in the exercises.

2.7.5 Sequence Transmission

To this point we have assumed ***one-shot*** transmission, where a single input symbol X produces a single output Y. To model real digital communication systems that send sequences, and to look ahead toward coded transmission, we consider now the sending of an N-tuple $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, which produces a channel response $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$. We assume feedback from output to input is not allowed to influence the selection of future inputs. As before, inputs $X_t, t = 1, \ldots, N$, are chosen from an alphabet of size M and outputs Y_t are in a Q-ary set. We make the further strong assumption that the channel acts on each input in *independent* fashion and refer to such a channel as ***memoryless.*** While certainly not always valid (intersymbol interference and fading are two causes of channel memory), the memoryless assumption serves our immediate need of developing the necessary concepts related to channel capacity. The memoryless assumption means that

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{N} P(y_t \mid x_t). \tag{2.7.20}$$

We have from the definition of mutual information that

$$I(\mathbf{X};\mathbf{Y}) = \sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) \log\left[\frac{P(\mathbf{y} \mid \mathbf{x})}{P(\mathbf{y})}\right]$$

$$= \sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) \left[\log \prod_{t=1}^{N} P(y_t \mid x_t) - \log P(\mathbf{y})\right] \tag{2.7.21}$$

$$= \sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) \left[\sum_{t=1}^{N} \log P(y_t \mid x_t)\right] - \sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) \log P(\mathbf{y})\Bigg]$$

(The outer sums indexed by $\mathbf{x}$ or $\mathbf{y}$ are understood as N-fold sums.) The first term in the final form of (2.7.21) can be interpreted as an expected value of a sum of random variables, which is the sum of expected values, that is, $\sum H(Y_t \mid X_t)$. The second term on the other hand is, upon summing out the $\mathbf{x}$ variable, the unconditional entropy $H(\mathbf{Y})$. We have shown, however, that this entropy is less than or equal to the sum of entropies for each symbol, with equality if symbols are independent. Thus, we have that

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &\le \sum_{t=1}^{N} H(Y_t) - H(Y_t \mid X_t) \\ &= \sum_{t=1}^{N} I(X_t; Y_t), \end{aligned} \tag{2.7.22}$$

with overall equality iff the Y_t variables are independent. The latter is true if and only if the inputs X_t are independent, given the memoryless channel model. Furthermore, each information term in (2.7.22) is upper bounded by channel capacity C, by definition. Thus, we have the inequality string

$$I(X;Y) \le \sum_{t} I(X_t; Y_t) \le NC. \tag{2.7.23}$$

In other words, the average mutual information shared by N-tuples over a discrete memoryless channel is no larger than the sum of the average (scalar) mutual informations, with equality when the input sequence is independent. This sum, in turn, is at most NC, where C is the maximum mutual information achievable in a single use of the channel. To have end-to-end equality in (2.7.23), we must choose the components of $\mathbf{X}$ independently according to a distribution that achieves capacity for the one-shot problem; that is,

$$P(\mathbf{x}) = \prod_{t=1}^{N} P(x_t), \tag{2.7.24}$$

where $P(x)$ achieves capacity in the sense of (2.7.18).

If maximizing mutual information between sequences is our goal, we find that for a memoryless channel there is no benefit in using dependent channel inputs, and in general this implies a penalty because the input entropy is diminished by such dependencies.

Another important relation for mutual information is associated with a cascade of channels (Figure 2.7.7). The boxes or channels may be visualized as processors that are part of the communication equipment, as well as the physical channel. We wish to show that the end-to-end average mutual information $I(X; Z)$ can be no larger than either of