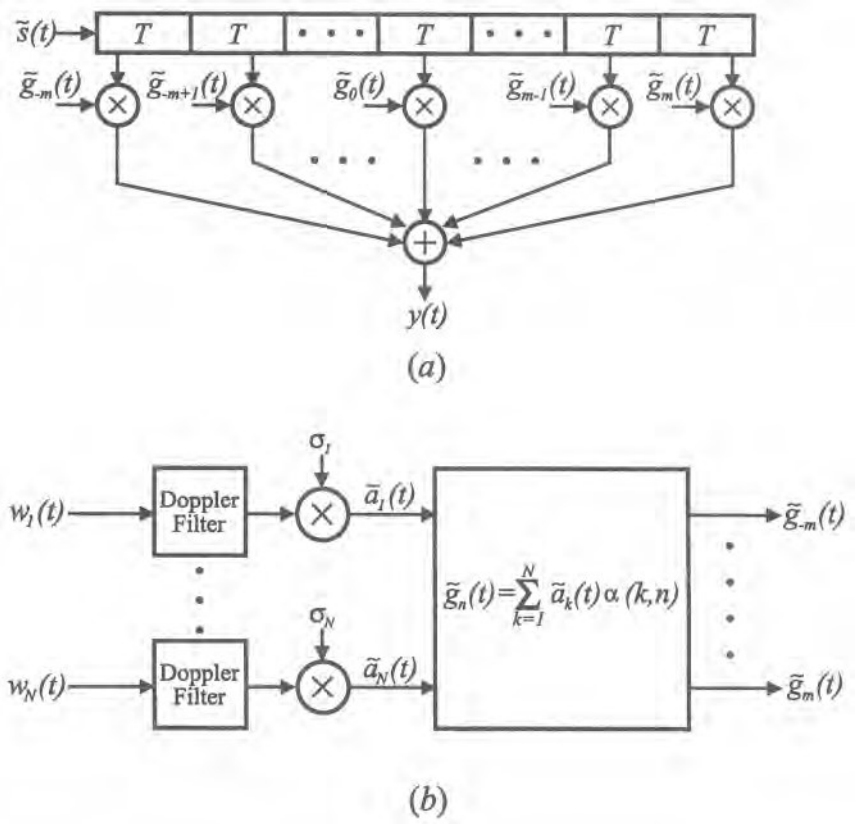


Figure 9.16. Uniformly spaced TDL model for a discrete multipath channel. (a) TDL structure; (b) generation of tap-gain processes: input processes are independent, zero-mean Gaussian processes.

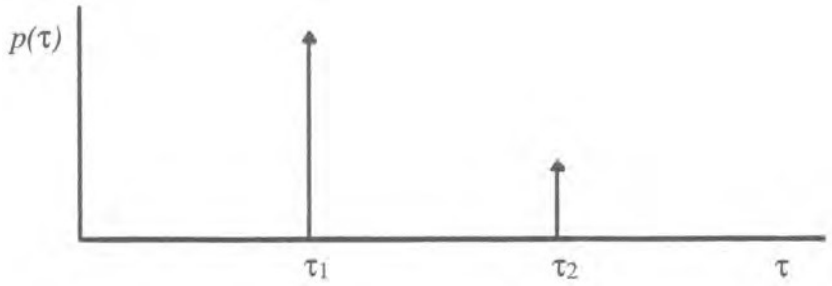


Figure 9.17. A simple two-ray model for Example 9.1.2.

$\Delta\tau \ll 1$, then the two paths can be combined and the model can be treated as frequency-nonselective; if $\Delta\tau > 0.1$ or so, there will be considerable intersymbol interference in the channel and it is treated as frequency-selective.

To illustrate the calculation of the tap-gain functions, let us assume that $\Delta\tau = 0.75$. The tap-gain functions in this case are obtained by filtering two uncorrelated white Gaussian

noise processes and then transforming them to tap-gain processes according to Equation (9.1.61) as

$$\begin{bmatrix} \tilde{g}_{-3}(t) \\ \tilde{g}_{-2}(t) \\ \tilde{g}_{-1}(t) \\ \tilde{g}_{0}(t) \\ \tilde{g}_{1}(t) \\ \tilde{g}_{2}(t) \\ \tilde{g}_{3}(t) \end{bmatrix} = \begin{bmatrix} \text{sinc}(0.0+3) & \text{sinc}(0.75+3) \\ \text{sinc}(0.0+2) & \text{sinc}(0.75+2) \\ \text{sinc}(0.0+1) & \text{sinc}(0.75+1) \\ \text{sinc}(0.0+0) & \text{sinc}(0.75+0) \\ \text{sinc}(0.0-1) & \text{sinc}(0.75-1) \\ \text{sinc}(0.0-2) & \text{sinc}(0.75-2) \\ \text{sinc}(0.0-3) & \text{sinc}(0.75-3) \end{bmatrix} \begin{bmatrix} \tilde{a}_1(t) \\ \tilde{a}_2(t) \end{bmatrix} = \begin{bmatrix} 0.0 & -0.060 \\ 0.0 & 0.082 \\ 0.0 & -0.128 \\ 1.0 & 0.300 \\ 0.0 & -0.900 \\ 0.0 & -0.180 \\ 0.0 & 0.100 \end{bmatrix} \begin{bmatrix} \tilde{a}_1(t) \\ \tilde{a}_2(t) \end{bmatrix}$$

The preceding equation shows the coefficients of the transformation for only seven taps. The reader can see that these coefficients will be negligible for higher order tap gains and hence they can be ignored and the TDL model can be truncated to seven taps. ■

9.1.3.5.3. Generation of Tap-Gain Processes The structural models considered thus far require two main steps for their implementation. First, we generate a set of white (discrete-time) Gaussian processes. The methods for doing so were discussed in detail in Chapter 7. Second, we need to shape the power spectral density of these processes to assume the shape of the Doppler spectrum at each tap location. In Chapter 7 we discussed means for generating Gaussian processes with arbitrary spectral density. Given a more or less arbitrary scattering function, we could in principle produce the shaping necessary to replicate $S(\tau, kB^{-1})$, $k = 0, 1, \ldots, N$. This could be a burdensome task, both in engineering manpower and computationally. For this reason, the simplified model introduced earlier is often used [as well as the fact that $S(\tau, \nu)$ may not be available]. The common Doppler spectral shape adopted in the simplified approach is also typically an even function of frequency, which means that only a real filter is necessary to do the shaping. This implies also that the real and imaginary parts of the complex white processes are independent of one another.

Thus, in the simplified model, the shaping filter (impulse response) is real and has a pure amplitude transfer function $H(f) = \sqrt{S(f)}$, where $S(f)$ is the Doppler (power) spectrum. For the reference channel models discussed latter, three spectra are specified, depending on the physical environment: a flat spectrum, a Gaussian spectrum, and Jakes spectrum. We now show how to obtain the corresponding shaping filters. The first two are rather evident, but the third is only a bit more complicated. The generic form of the model is shown in Figure 9.18.

For a flat spectrum, clearly nothing needs to be done, except for proper scaling and bandlimiting to the desired process bandwidth. If the flat spectrum is, say, $S_f(f) = A, |f| \leq B$, the shaping filter is simply $H_f(f) = \sqrt{A}, |f| \leq B$, an ideal lowpass filter.

For the Gaussian spectrum $S_G(f)$, let us say $S_G(f) = A^{-kf^2}$, with k chosen to set the desired bandwidth. The shaping filter is then $H_G(f) = \sqrt{S_G(f)} = \sqrt{A}\exp(-\frac{1}{2}kf^2)$, which is

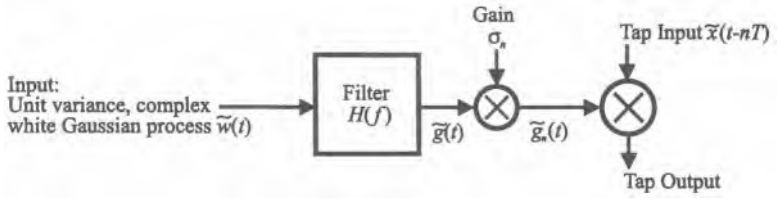


Figure 9.18. Generic block diagram for shaping the power spectrum of the nth tap-gain process.

still Gaussian, and the impulse response of which, $h_G(t)$, [inverse Fourier transform of $H_G(f)$] is also Gaussian, i.e., $h_G(t) = \sqrt{2\pi A/k}\exp[-(2\pi^2/k)t^2]$.

Finally, let us consider the Jakes spectrum [previously given in (9.1.23) and displayed in Figure 9.8a]:

$$S_J(f) = \frac{A}{[1-(f/f_d)^2]^{1/2}} \tag{9.1.62}$$

where f_d is the largest Doppler frequency.

To obtain the shaping filter response, we have

$$S_J(f) = S_f(f)|H_J(f)|^2 \tag{9.1.63}$$

where $S_J(f)$ is the Jakes spectrum, $S_f(f)$ is the flat spectrum, and $H_J(f)$ is the Jakes frequency response.

The frequency response is then, assuming $S_f(f) = 1$,

$$H_J(f) = [S_J(f)]^{1/2} = \frac{A^{1/2}}{[1-(f/f_d)^2]^{1/4}} \tag{9.1.64}$$

Since $H_J(f)$ is a real and symmetric function, the impulse response of the filter $h_J(t)$ can be derived from the *cosine transform*, which can be found, e.g., in Ref. 24.

The derived impulse response is then

$$\begin{aligned} h_J(t) &= F^{-1}(H_J(f)) = A^{1/2}2^{1/4}\pi^{1/2}\Gamma(3/4)f_d x^{-1/4}J_{1/4}(x) \\ &= A^{1/2}\cdot 2.583 f_d x^{-1/4}J_{1/4}(x) \end{aligned} \tag{9.1.65}$$

where $J_{1/4}(\cdot)$ is the fractional Bessel function and $x = 2\pi f_d|t|$. Equation (9.1.65) can be well approximated by an FIR filter (see Case Study IV in Chapter 12).

The filter gain $A^{1/2}$ is chosen such that $h_J(t)$ has the normalized power of 1. The individual tap gains $\tilde{g}_n$ then have to be properly scaled to account for the different powers of the taps (see, e.g., equ. 9.1.38b). It should be noted that the bandwidth of the tap-gain processes of slowly time-varying channels will be considerably smaller than the bandwidth of the signals that flow through them. Simulation of the filter at the higher rate would likely lead to computational inefficiencies as well as stability problems. These potential difficulties can be avoided by applying multirate techniques. For example, the tap-gain filter could be modeled at a lower rate, compatible with the filter bandwidth, and interpolation would then be used at the output of the filter to produce denser samples at the rate of the signal coming into the tap. An example application is developed in Case Study IV in Chapter 12.

9.1.3.6. Indoor Wireless Channels

Fading properties of indoor wireless channels are quite different from those of mobile (vehicular) channels due to differences in the physical environment and the consequent differences in propagation mechanisms. In mobile wireless channels, 5–10 (and perhaps fewer) most significant paths are usually responsible for more than 80% of power delivered

between the transmitter and receiver antennas. Rays along those paths undergo diffraction around vertical or horizontal edges of buildings, reflection off building surfaces, ground reflection, vegetation scattering, and so forth. If the dominant path to the receiver is shadowed by an obstacle, the total power may attenuate significantly, a phenomenon that earlier we called shadow fading.

In the case of indoor wireless channels, on the other hand, the number of rays contributing to the total receiver power is much larger than in mobile wireless channels due to the fact that rays in the horizontal plane usually undergo multiple reflections and transmissions at interior or exterior walls of buildings as well as reflection off furniture. Rays in the vertical plane may also experience multiple reflections at ceilings and floors and may also bounce off furniture. Lateral diffraction is involved in propagation when the transmitter and receiver antennas are located in building hallways, but it is rare that diffracted paths overwhelm reflected or transmitted paths.

There are a number of statistical models for the indoor wireless channel, based on extensive measurements in buildings[25–31] and also coming into use is the deterministic channel modeling technique, which combines ray tracing with a building-specific database.[32] By and large, proposed indoor models fall under the "discrete" multipath category, as defined earlier, and can be considered special cases of (9.1.5). However, the statistical characterization and parameter values are dependent on the specific physical environment, as might be expected.

The distinction between shadow fading and multipath (or "fast") fading is not as appropriate in the indoor channel due to the confined environment and the much lower speed of the mobile. The latter consideration, in fact, leads to a Doppler spectrum which is modeled as flat (and generally so specified in the reference channel models for PCS; see the Appendix). In certain environments, such as an office building, propagation will almost always be "shadowed," while in buildings with more open plans there may be either shadowed or line-of-sight (LOS) paths. However, because of the relatively low speed of the mobiles (which may also be stationary for relatively long periods of time), it is appropriate to develop models that are location-specific, depending only on whether a location is LOS to the transmitter or shadowed. In the following, a brief description of some models will be introduced as examples of modeling constructs.

9.1.3.6.1. Factory and Open-Plan-Building Model Extensive wideband measurement campaigns for indoor multipath channel modeling around 1 GHz have been made in various types of factory buildings by different workers (see, e.g., Refs. 25–28). Factories may be considered particular instances of environments that can be called "open plan," others being warehouses, stores, supermarkets, etc. In Refs. 25 and 26, for example, the complex lowpass-equivalent channel impulse response is represented by the

$$\tilde{c}(\tau, t) = \sum_{k=1}^{L} a_k e^{-j\theta_k} \delta(t - \tau_k) \tag{9.1.66}$$

In (9.1.66), the path phases θ_k are *a priori* assumed to be identically and independently distributed (i.i.d.) uniformly on $[0, 2\pi]$. The other model parameters, L (the effective number of paths), the $\{a_k\}$ (the path amplitudes), and the τ_k (the path delays), are to be empirically characterized. It turns out to be convenient to develop empirical distributions in two receiver location categories: LOS and OBS (obstructed), the latter synonymous with "shadowed." The number of multipath components L was found to have approximately a normal distribution.

The measurements also supported an empirical model for the pdf of component arrival times τ_k, which turns out to be a piecewise linear function of excess delay, assuming the multipath components in a given profile arrive independently, but with different probabilities[26] (Excess delay is simply the delay measured with respect to the first-arriving multipath component.) Amplitude fading statistics of multipath components were also obtained both in large and small scale. In a large-scale model, the cumulative distribution of the local averaged powers of individual multipath components turns out to be log-normally distributed about the mean value of local averaged powers with a standard deviation of 5.4 dB. The statistics of small-scale fading which occurs with small changes in receiver location is described by a log-normal distribution about the local mean power of individual multipath components where the standard deviation is a random variable having a distribution function[26] of exponential form.

The local averaged impulse response was used to compute path loss and multipath parameters such as mean excess delay and rms delay spread with an assumption of wide-sense stationary properties of the channel over space in the range of about 10 wavelengths. With the uniform i.i.d. assumption on the phases θ_k, the total received power will be simply the superposition of averaged power delivered along each path. This hypothesis was shown to account for the good agreement between wideband and CW path loss measurements. The total received power having a path distance of d varies randomly depending on the propagation environment between the transmitter and receiver antennas. One can express a general path loss model[29] as

$$\mathrm{PL}(d) = \mathrm{PL}\,(d_0) + 10n\,\log(d/d_0) + X_\sigma \tag{9.1.67}$$

where d is the distance, PL the path loss in decibels, and n the path loss index; X_σ in decibels is a random variable following approximately a Gaussian distribution with standard deviations of 5–9 dB.[29] depending on the building environment. Measured data showed that the path loss index is 2.2 on the average. Based on the measured multipath spreads ranging from 40 to 800 ns, rms delay spread was found to have values from 30 to 300 ns.[25] Measured results showed that a uniform distribution having a mean of 100 ns is a reasonable approximation.

Some sample measurements of a delay power profile are given in Figure 9.19.[31] Figures 9.19a and 9.19b show single delay power profiles in a retail store. Such measurements are typically taken multiple times in the vicinity of a point, to understand the small-scale variations. Figures 9.19c and 9.19d show two such types of measurements taken in a retail store at 16 different points separated by one-quarter wavelength along a straight line in localized spots characterized as LOS and "cluttered" reception, respectively. Figure 9.20 shows some additional measurements in the same environment. A model such as (9.1.66) with appropriate statistical descriptions of the parameters will imitate measured data to the extent that the model is accurate.

9.1.3.6.2. Office Building Model In order to characterize the multipath propagation properties, wideband measurements using 10-ns, 1.5-GHz, radarlike pulses were made within a medium-size two-story office building.[30] The authors proposed a statistical model of the indoor radio channel which appears to fit the measurements very well. In this model the received signal rays arrive in clusters, where the cluster arrival times are modeled as a Poisson arrival process with some fixed rate λ. Within each cluster, subsequent rays also arrive according to a Poisson process with another fixed rate Γ with the constraint that $\lambda \gg \Gamma$.

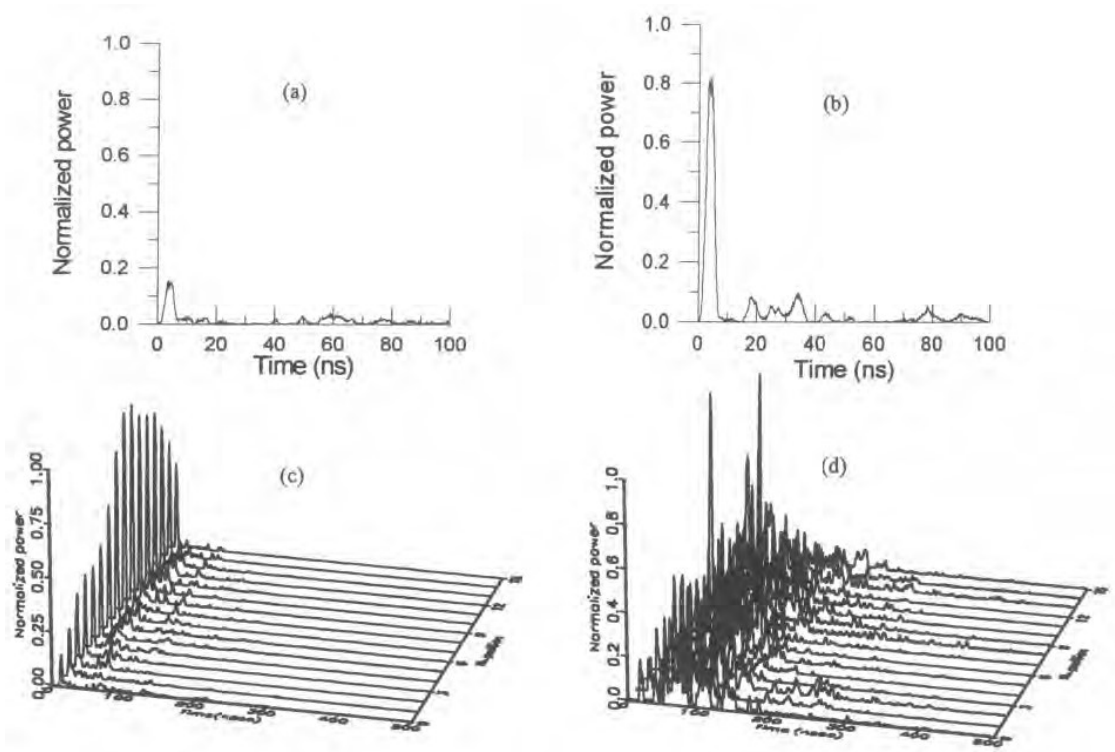


Figure 9.19. Illustration of the measurement of single delay power profiles and scattering functions in a retail store.

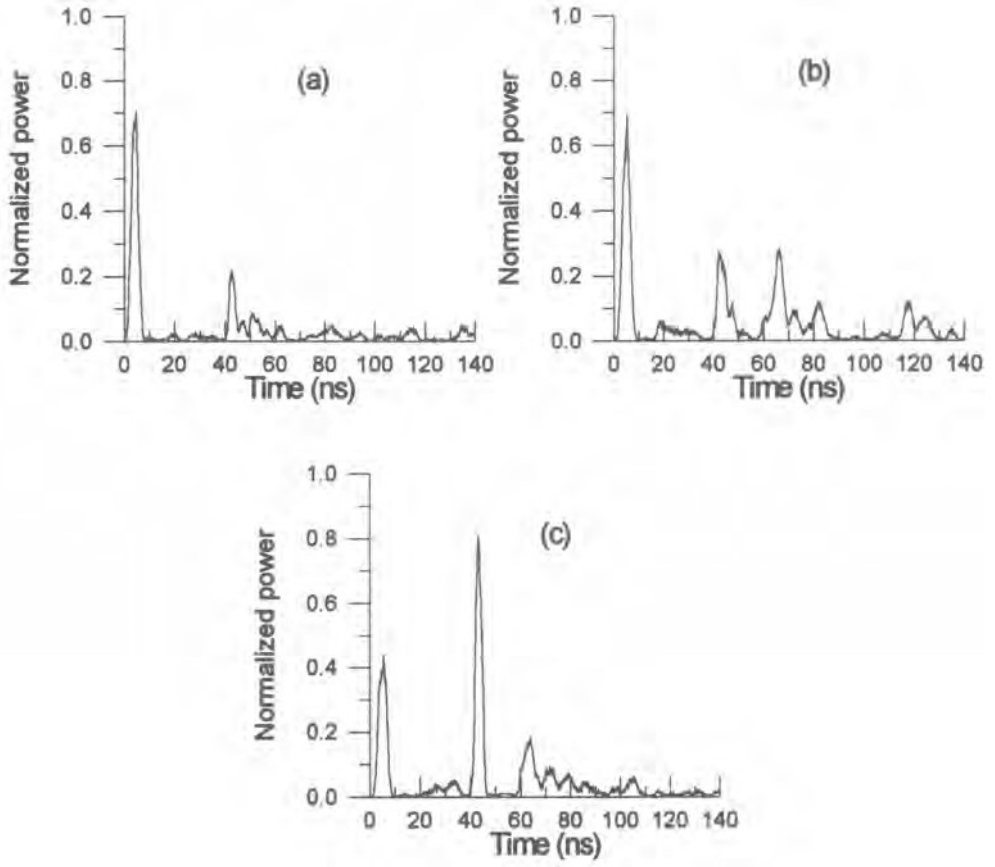


Figure 9.20. Additional measurements of delay power profiles with lightly obstructed paths in a retail store.

Hence the complex lowpass channel impulse response is given as

$$\tilde{c}(\tau, t) = \sum_{l=1}^{\infty} \sum_{k=1}^{\infty} a_{kl} e^{-j\theta_{kl}} \delta(t - T_l - \tau_{kl}) \tag{9.1.68}$$

In addition to an i.i.d. assumption on the phases θ_{kl}, uniform over $(0, 2\pi)$, the mean square values are monotonically decreasing functions of T_l and τ_{kl} as expressed in the following equation:

$$\overline{a_{kl}^2} \equiv \overline{a^2(T_l, \tau_{kl})} = \overline{a^2(0, 0)} \exp^{(-T_l/\Gamma)} \exp^{(-T\tau_{kl}/\gamma)} \tag{9.1.69}$$

where $\overline{a^2(0, 0)}$ is the average power of the first ray of the first cluster, and Γ and γ are power-delay time constants for the clusters and the rays, respectively. The probability distribution of the normalized amplitudes of rays is independent of the associated delays and is stated to follow the Rayleigh probability density function.[30] However, Kim *et al.*[31] reported that the distribution of amplitudes is close to a Ricean distribution with K factor of 5 for LOS paths between receivers and transmitters as shown in Figure 9.21. On the other hand, the distri-

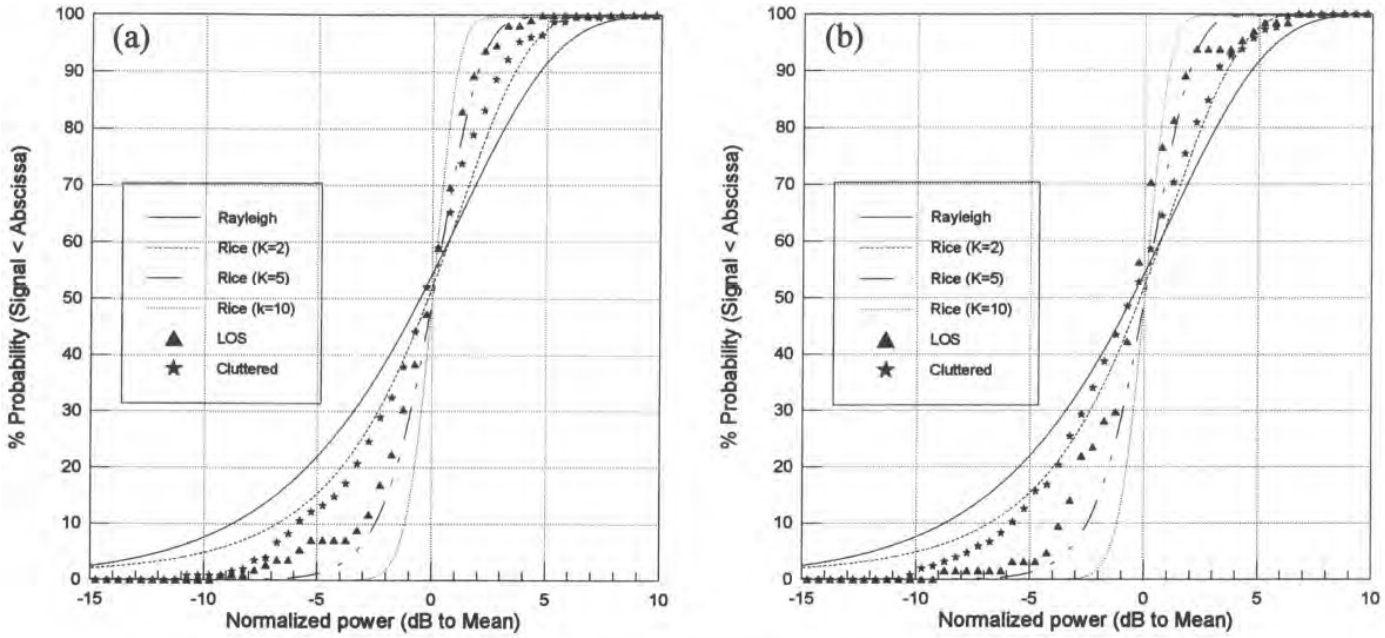


Figure 9.21. Comparison of cumulative distributions of measured amplitudes with analytical models. (a) Engineering building; (b) retail store.

bution for non-LOS paths and heavily cluttered environments around receivers and transmitters does seem to follow the Ricean distribution with a K factor of 2, which is very close to the Rayleigh distribution as suggested by Saleh and Valenzuela.[30] In principle, rays and clusters extend over an infinite time as shown in equation (9.1.67). However, the sum over l stops at L when $\exp(-T_l/\Gamma) \ll 1$, and that over k stops at K when $\exp(-\tau_{kl}/\gamma) \ll 1$, so that in practice, one uses a truncated version:

$$\tilde{c}(\tau, t) = \sum_{l=1}^{L} \sum_{k=1}^{K} a_{kl} e^{-j\theta_{kl}} \delta(t - T_l - \tau_{kl}) \tag{9.1.70}$$

The numbers L and K were not given a statistical characterization, although they do vary depending on the environment. Rather, the authors[30] propose to let these numbers "float", that is, to be set by an optimization procedure that produces the closest agreement between the measurements and the model.

The measured rms delay spread within rooms of the building had a median value of 25 ns and a maximum value of 50 ns, which implies that rms delay spread is dependent on the building size as well as the extent of obstruction of line-of-sight paths. The path loss index seems to be between 3 and 4, which is higher than that of the factory model. This may be due to the existence of interior walls in the office building, which is less common in factory buildings.

Some typical examples of measured delay power profiles in an office building are shown in Figure 9.22.

9.1.3.6.3. Ray-Tracing Prediction Model An understanding of the characteristics of indoor wireless communication channels is essential for the design and deployment of wireless communication systems. Measurements of CW and pulsed signals are the most common measurement methods to model empirically the radio channel, as described in the previous sections. However, statistical analysis based on measurements may not provide for optimum deployment of the communication systems in a specific building. In order to avoid expensive measurements in individual buildings prior to installation, or adjustments afterward, theoretical prediction models have been developed to predict the path loss and delay spread from the building floor plan.

A computer-based tool called Wireless System Engineering (WiSE)[32] has been developed for designing an indoor and campus-sized wireless system. It uses algorithms from computational geometry and optimization to decide where to place base-station transceivers. The propagation model of WiSE is a 3D ray-tracing process that in principle follows all paths from a base station to a specific point. A brute-force approach here requires too much computation time to be feasible for real buildings. Accordingly, a modified algorithm was developed to reduce the number of path components that have to be evaluated by focusing on useful rays only. This propagation model predicts the local mean of signal power received at any given point with an i.i.d. assumption on the phase angle of each ray component. The building database, that is, the locations and radio characteristics of all walls, floors, and ceilings, must be acquired and digitized. Given building data and system parameters (for example, antenna type, antenna location and elevation, radiation power, and so forth), WiSE determines system performance (for example, received power or delay spread) throughout the building or at specific points.

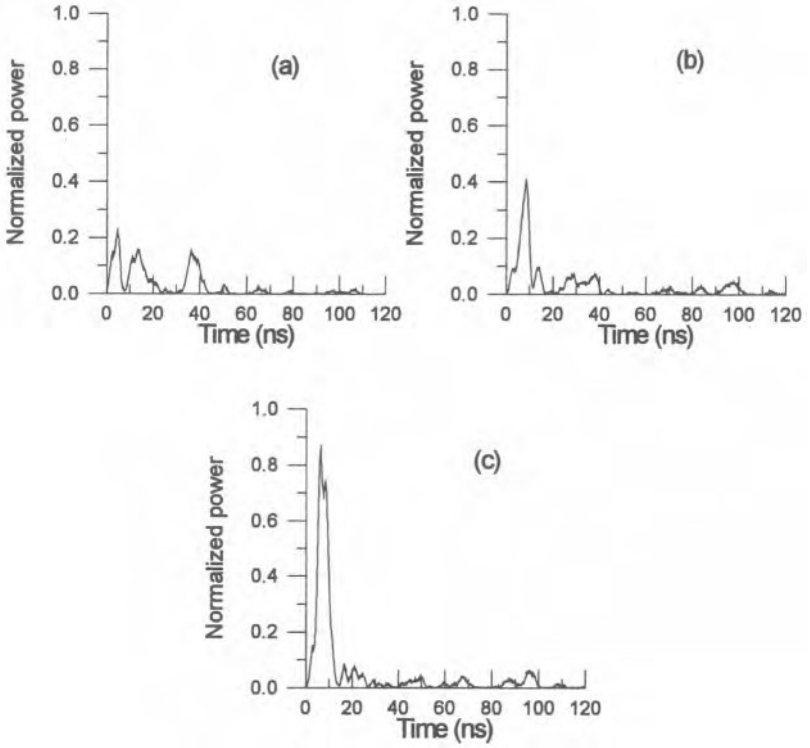


Figure 9.22. Measurements of power delay profiles for a site with lightly obstructed paths in an office building.

9.1.3.7. Radio-Relay Line-of-Sight (LOS) Discrete Multipath Fading Channel Model

There are many possible models for a multipath channel, depending on the application, the frequency band, and the physical environment. The same assumptions may not apply, or apply equally well, in all circumstances. In this section we present another type of multipath model, developed explicitly for terrestrial communication links between fixed antenna towers. There are, again, conceptually different possible models for this environment, but here we limit ourselves to the presentation of the most accepted model, the so-called Rummler model.[33]

The channel is a line-of-sight (LOS) radio channel where the multipath fading is caused by stratification of the atmosphere under certain climatic conditions. Because of the use of larger antennas, the field of view of the antenna is limited to very small angles of arrival, and hence the smaller number of multipath components. Also, since the antennas are fixed, the only time variations in the channel characteristics are due to changes in the atmospheric conditions. These variations can be considered very slow compared to the channel bandwidths, which will be on the order of tens of MHz. Hence, the Rummler model is a multipath model with very slow fading. This model is widely used for terrestrial microwave links operating in the frequency range of 2–18 GHz between fixed towers. The model derivation is based on a set of assumptions discussed shortly. Measured data were used to obtain the values and statistical characterization of the model parameters.

Given the geometry of the link and antenna parameters, Rummler hypothesized a three-ray model of the form

$$y(t) = s(t) + \alpha s(t - \tau_1) + \beta s(t - \tau_2) \tag{9.1.71}$$

where $s(t)$ and $y(t)$ are the bandpass input and output, respectively. In terms of the complex envelopes, the model takes the form

$$\tilde{y}(t) = \tilde{s}(t) + \alpha\tilde{s}(t - \tau_1)e^{-j2\pi f_c\tau_1} + \beta\tilde{s}(t - \tau_2)e^{-j2\pi f_c\tau_2}$$

where f_c is the carrier frequency.

The lowpass-equivalent transfer function of the channel is given by

$$H(f) = 1 + \alpha e^{-j2\pi(f_c - f)\tau_1} + \beta e^{-j2\pi(f_c - f)\tau_2} \tag{9.1.72}$$

Since we are dealing with lowpass-equivalent channels, it is convenient to substitute $f \rightarrow (f - f_c)$, where f is now the lowpass frequency. Then $|f| \leq B/2$, where B is the bandwidth of the channel.

To achieve a unique fit of the model parameters to the empirical data, Rummler introduced a few modifications. The first modification of the model is based on the assumption that over the bandwidth of interest one has $B\tau_1 \ll 1$, which causes frequency nonselective (flat) fading. Hence $\exp^{(-j2\pi f\tau_1)} \approx 1$, and (9.1.72) can be written as

$$H(f) = 1 + \alpha + \beta e^{-j2\pi f\tau_2} \tag{9.1.73}$$

The above equation has three random parameters, α, β, and τ_2. Notice that, because of the assumptions made, the three-path model is effectively a two-path model.

Considerations related to fitting the model parameters to experimental data lead us to identify in the model a "notch" frequency parameter f_0 for which $|H(f)|$ is minimum. Thus, substituting $\tau_2 = [(f - f_0)/f]\tau$, where τ is a constant delay, we get the final form of the model

$$H(f) = a[1 - be^{-j2\pi(f - f_0)\tau}] \tag{9.1.74}$$

where we can regard $a \simeq 1 + \alpha$ as the overall attenuation parameter, τ the relative path delay of the second ray, and $b = -\beta/(1 + \alpha)$ and $2\pi f_0\tau$ as the relative amplitude and phase of that ray, respectively. Since (9.1.74) is a model, not the physical reality, the choice of the parameter τ is somewhat arbitrary, so long as the ultimate model is good. It turns out that a fixed value of τ on the order of $(6B)^{-1}$ allows empirical fitting of the remaining free parameters in a way that properly accounts for the observed channel behavior. The value chosen in this model is $\tau = 6.3\text{ns}$.

The squared amplitude response of this channel is

$$|H(f)|^2 = a^2\{1 + b^2 - 2b\cos[2\pi(f - f_0)\tau]\} \tag{9.1.75a}$$

and the group delay is given by the derivative of the phase characteristic,

$$D(f) = \frac{-1}{2\pi}\frac{d\phi(f)}{df} = \frac{\tau}{2\pi}\frac{b\{b - \cos[2\pi(f - f_0)\tau]\}}{1 + b^2 - 2b\cos[2\pi(f - f_0)\tau]} \tag{9.1.75b}$$

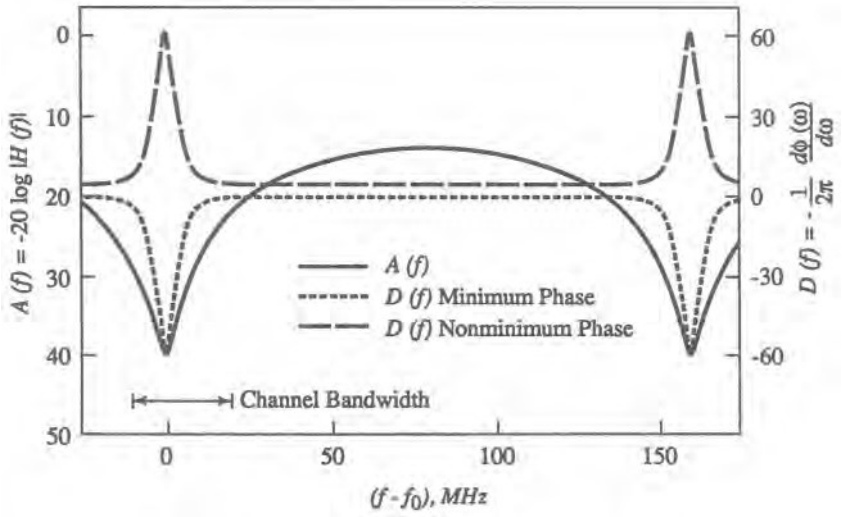


Figure 9.23. Typical amplitude and delay responses for Rummler's LOS fading channel model.

If $b < 1$, the transfer function is *minimum-phase* since it has zeros only in the left half of the s plane. For $b \geq 1$, the transfer function is *nonminimum-phase* because its zeros are in the right half of the s plane. Typical amplitude and delay responses based on (9.1.75) are shown in Figure 9.23.

The lowpass-equivalent channel impulse response, the Fourier transform of (9.1.74), is

$$\tilde{h}(t) = a[\delta(t) - be^{-j2\pi f_0 \tau}\delta(t - \tau)] \tag{9.1.76}$$

Equation (9.1.76) corresponds to a two-tap TDL model shown in Figure 12.10 in Case Study I of Chapter 12.

The parameters a and b are random variables whose distributions have been empirically (approximately) determined. It is more convenient to describe the distributions of logarithmic forms. Thus, we define, for minimum-phase fades,

$$A_1 = -20 \log a \quad \text{[dB]} \tag{9.1.77a}$$

$$B_1 = -20 \log(1 - b) \quad \text{[dB]} \tag{9.1.77b}$$

For nonminimum-phase fades, the range of b is unbounded and it is convenient to rephrase (9.1.74) as[34]

$$H'(f) = ab[e^{-j2\pi(f - f_0)\tau} - 1/b] \tag{9.1.78}$$

and we can now define

$$A_2 = -20 \log(ab) \quad \text{[dB]} \tag{9.1.79a}$$

$$B_2 = -10 \log(1 - 1/b) \quad \text{[dB]} \tag{9.1.79b}$$

The minimum- and nonminimum-phase fades are equally likely to occur and can be assumed to have identical distributions[34] as follows:

1. Parameters B_1 and B_2 are both exponentially distributed with mean 3.8 dB.
2. Parameters A_1 and A_2 are Gaussian random variables with mean

$$\mu = 24.6\frac{B^4 + 500}{B^4 + 800} \quad \text{[dB]} \tag{9.1.80}$$

where $B = B_1$ for A_1 and $B = B_2$ for A_2. The standard deviation of A_1 and A_2 is 5 dB. Note that, because of (9.1.80), parameters a and b are correlated, i.e., the mean of A depends on the value of B.

The distribution of the phase $\theta = 2\pi f_0 \tau$ has a range of $[-\pi, \pi]$ and a constant density on each section $|\theta| > \pi/2$ and $|\theta| < \pi/2$ with

$$P\{|\theta| < \pi/2\} = 5P\{|\theta| > \pi/2\} \tag{9.1.81}$$

The notch frequency is given by

$$f_0 = \frac{\theta}{2\pi\tau} \qquad \text{where} \quad \tau = 6.3 \text{ ns} \tag{9.1.82}$$

The application of the Rummler model is described in Case Study I in Chapter 12. We note that although this model falls under the general description of (9.1.1), the statistical characterization is quite different from that of the mobile channel, and these again are different from the characterization of indoor channels that we discussed in the previous section.

9.2. The Almost Free-Space Channel

Virtually any channel will be benign if the signal has sufficiently small bandwidth, and conversely almost any benign channel will exhibit nontrivial distortion when the signal has sufficiently large bandwidth. Propagation in or through the atmosphere is an extremely complex phenomenon,[35–37] which can take on a wide range of behavior depending on circumstances. At one end of this range, the atmosphere has frequently been regarded as well approximated by free space, i.e., an ideal channel. This approximation is fairly good for satellite systems using large ground antennas operating in the 4- to 6-GHz range and with elevation angles that are not too small. However, as we increase the carrier frequency and increase the bandwidth correspondingly, a nonnegligible filtering effect begins to manifest itself. We will refer to such a case as the almost free-space channel. In this channel we consider only this effective filtering to exist; that is, we assume there is no multipath or scintillation. Although this filtering characteristic is in reality time-varying, it is reasonable to treat it on a quasistatic basis because at the very wide bandwidths for which the filtering effect becomes significant the channel does vary very slowly with respect to the signal. The time variations can be separately modeled as phase noise whose spectrum is extremely narrow.[38] We consider briefly three contributors to this filtering effect: (a) the clear-air atmosphere, in which significant filtering can occur around specific absorption "lines"; (b) the rainy

atmosphere, in which absorption is a function of frequency; and (c) phase distortion due to the ionosphere. In the rainy atmosphere we also briefly consider the depolarizing effect of rain.

9.2.1. Clear-Air Atmospheric (Tropospheric) Channel

In a clear-air atmosphere, an electromagnetic wave interacts with the oxygen and water vapor that are present in a way that depends on the frequency of the wave. At certain frequencies there are resonances, resulting in peaks of absorption.[39] An example of atmospheric absorption curves is shown in Figure 9.24. Clearly, over a large enough bandwidth, the atmosphere acts as a filter. The possible effects of such a filter have been reported in Ref. 40. In order to simulate this effect, one needs to know the transfer function $H(f)$. If one has a set of measurements, it is straightforward to include them as a filter in a simulation, as discussed in Chapter 4. An analytical form is preferable, however, and for this purpose, one can use Liebe's model,[39]

$$H(f) = H_0 \exp[j0.02096f(10^6 + N)L] \tag{9.2.1}$$

where N is the complex refractivity in parts per million (a function of f), $N = N_0 + D(f) + jN''(f)$; H_0 is a constant determined from table lookup; N_0 is the frequency-independent refractivity; $D(f)$ is the refractive absorption; $N''(f)$ is the absorption; and L is the distance in kilometers.

The above parameters are dependent on frequency and on atmospheric conditions, namely, temperature, barometric pressure, and relative humidity. Values of these parameters are tabulated in Ref. 39. Liquid water in clouds or fog will also affect the absorption. Of course, the major liquid water effect occurs in rain, which we discuss next.

9.2.2. The Rainy-Atmospheric Channel

At microwave frequencies, say above 10 GHz, rain can become a dominant effect on atmospheric propagation. Certainly, at high enough frequencies and rain rate, rain attenuation is much more significant than atmospheric absorption, except possibly at the resonance lines. The effect of rain is well known[41,42] and is usually accounted for as attenuation which is

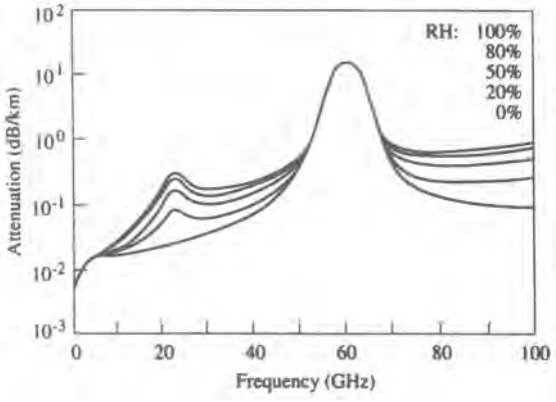


Figure 9.24. Atmospheric absorption at sea level as a function of frequency and relative humidity (from Ref. 40, © IEEE, 1984).

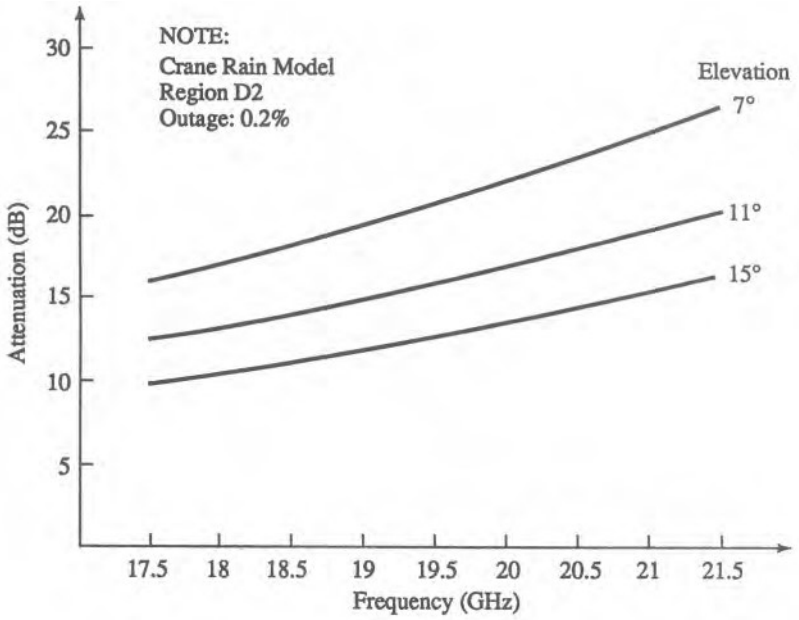


Figure 9.25. Rain attenuation as a function of frequency.

independent of frequency across the bandwidth of the signal. However, as mentioned, for sufficiently wide bandwidth this attenuation will be frequency dependent, and hence effectively act as a filter. Figure 9.25 illustrates this point. As an example, we have computed rain attenuation as a function of frequency over the band from 17.7 to 21.2 GHz allocated to the space-to-earth path for satellite communications. At these frequencies rain attenuation is not only a major attenuating factor, but, as can be seen, it exhibits a significant effective gain slope across the band. The curves in Figure 9.25 have been computed using the Crane model[41] for three different elevation angles and rainfall corresponding to an outage probability of 0.2%. The receiving station is assumed at Washington, D.C., which has a relatively temperate climate. The effects would be more severe in rainier climates. One could interpret Figure 9.25 as a filter for use in simulation and simply input points from the curves.

Another effect of rain which can be significant in dual-polarized systems is the depolarization of radio waves.[36,37,42] That is, some of the energy in each polarization is transformed into energy of the opposite (orthogonal) polarization. Let

$$S_1(t) = \rho_1(t)\cos[\omega_1 t + \phi_1(t)] \tag{9.2.2a}$$

$$S_2(t) = \rho_2(t)\cos[\omega_2 t + \phi_2(t)] \tag{9.2.2b}$$

represent signals of polarization 1 and 2, respectively; these polarizations could be linear, say horizontal and vertical, or left hand and right-hand circular; the carrier frequencies f_1 and f_2 may not be the same, but the spectra overlap. The simplest model for depolarization takes the form

$$R_1(t) = \alpha_{11}S_1(t) + \alpha_{21}S_2(t) \tag{9.2.3a}$$

$$R_2(t) = \alpha_{22}S_2(t) + \alpha_{12}S_1(t) \tag{9.2.3b}$$

where $R_1(t)$ and $R_2(t)$ are the signals received of polarizations 1 and 2, respectively, and the coefficients represent the relative magnitude of each term. Basically, the cross-polarized leakage introduces interference into each signal. This interference is referred to as XPI. In signal 1, for example, the XPI is $20 \log(\alpha_{11}/\alpha_{21})$. Situations of this type are easily simulated, and there is also a good deal of literature on the effect of interference on the BER of digital signals, as well as on methods to compensate for XPI (see, e.g., Ref. 43 for a brief overview of the subject). However, simulation becomes increasingly simpler than analysis as the system departs further from the ideal. Equations (9.2.3) can be generalized by replacing each of the coefficients α_{ij} by a linear frequency-dependent transfer characteristic. The result is

$$R_1(t) = h_{11}(t) * S_1(t) + h_{21}(t) * S_2(t) \tag{9.2.4a}$$

$$R_2(t) = h_{22}(t) * S_2(t) + h_{12}(t) * S_1(t) \tag{9.2.4b}$$

which, again, is straightforward to simulate if we know the impulse responses $h_{ij}(t)$.

9.2.3. The Ionospheric Phase Channel

In the lower frequency bands, say on the order of a few hundred megahertz and below, the effect of the ionosphere on propagation is extremely complex and best characterized as a time-varying multipath channel. As such, it is perhaps more appropriately modeled by the methods discussed in Section 9.1. At frequencies above a few hundred megahertz, for example, in the bands allocated to satellite communications, and assuming the absence of anomalous condition (solar flares, nuclear events, extremely low evaluation angles), the ionosphere can be approximately modeled by an all-pass filter with a nonideal phase characteristic.

It can be shown[44] that the phase shift experienced by a wave of frequency f due to free electrons in the ionosphere, over and above the free-space propagation lag, is given by

$$\phi(f) = \frac{2\pi 40 \times 10^6}{cf} \int_{s_1}^{s_2} N_e(s)\, ds \quad \text{(rad)} \tag{9.2.5}$$

where c is the speed of light (cm/s), N_e is the areal electron concentration (electrons/cm^2) at any point along the path s, and the integral represents the integrated ("columnar") electron density along the signal path. The differential phase shift between any two frequencies f_0 and $f_0 + \Delta f$ is therefore given by

$$\psi(\Delta f) = \phi(f_0 + \Delta f) - \phi(f_0) \tag{9.2.6a}$$

$$= \frac{2\pi 40 \times 10^6 \Delta f}{cf_0(f_0 + \Delta f)} \int_{s_1}^{s_2} N_e(s)\, ds \tag{9.2.6b}$$

$$= K(f_0, N_e, s) \frac{\Delta f}{f_0 + \Delta f} \tag{9.2.6c}$$

in which we have lumped into the constant K all the factors not depending on Δf. In fact, since we can consider Δf to be the departure from any given center frequency f_0, it can be seen that (with $\Delta f \rightarrow f$)

$$\psi(f) = K\frac{f}{f_0 + f} \tag{9.2.7}$$

is the complex lowpass-equivalent filtering characteristic. We can simplify (9.2.7) still further by defining the normalized frequency $\nu = f/f_0$, so that

$$\psi(\nu) = K\frac{\nu}{1 + \nu} \tag{9.2.8}$$

Obviously, the actual phase characteristic depends in scale on f_0^{-1} and on the integrated electron density, which can vary by two or three orders of magnitude depending upon path length (elevation angle), time of day, or solar activity. The frequency-dependent characteristic $v/(1 + v)$ is very simple and can be entered directly as-is (in sampled form) for use in simulation; this function is plotted in Figure 9.26. However, it is even simpler to implement if we realize, as the figure suggests, that this function is dominated by parabolic phase. That is, if we expand $\psi(\nu)$ in a power series, say, or a least-squares polynomial decomposition, then the ν^2 term is the major one (other than v, which does not lead to distortion) for typical

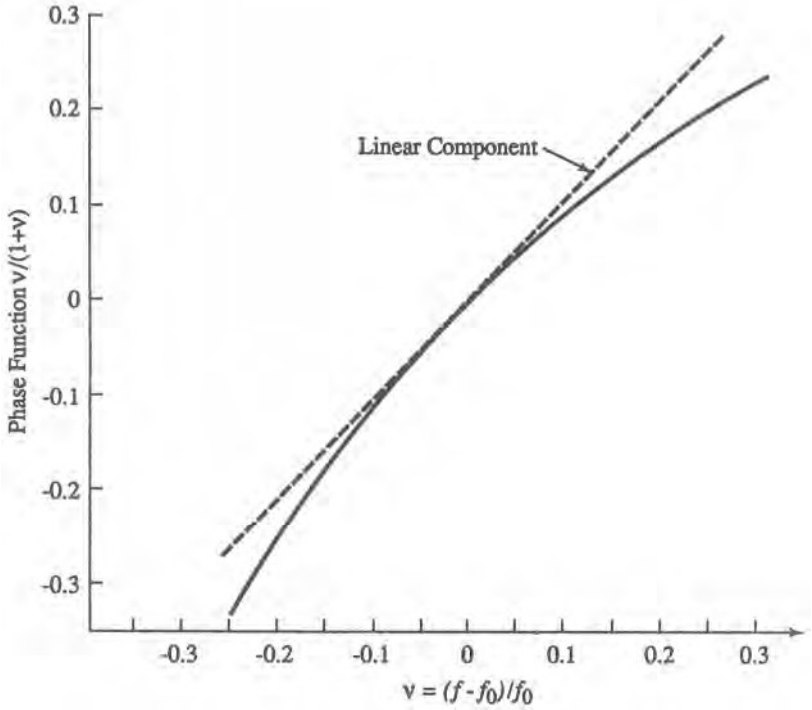


Figure 9.26. Ionospheric lowpass-equivalent phase characteristic.

fractional bandwidth values, say $|\nu| \leq B/f_0 < 0.1$. Because of the simple form of $\psi(\nu)$ it is easy to demonstrate the point using a Taylor series about $\nu = 0$. Thus, one obtains

$$\psi(\nu)/K = \nu/(1+\nu) = \nu - \nu^2 + R_3 \tag{9.2.9}$$

where the remainder is

$$R_3 \leq M\nu^3/3!$$

and $M \geq |\psi'''(u)/K|$, where $|u| \leq B$. Since $\psi'''(u)/K = 6(1+u)^4$, its maximum value is reached when $u = -B$, which we can take as -0.1. Thus, $R_3 = \nu^3/(0.9)^4 \approx 1.5\nu^3$, which means that at $\nu = -0.1$, the magnitude of the remainder relative to the squared term is $1.5\nu^3/\nu^2 = 1.5\nu = 0.15$. This means, basically, that the parabolic phase term ν^2 dominates over any reasonable fractional bandwidth. Hence, a simple expedient to model the effect of the ionosphere is to replace it by a parabolic phase functional filter, as discussed in Section 8.9. Notice in this particular case that the parabola is concave down.

9.3. Conducting and Guided Wave Media

In many situations of interest the medium (channel) between transmitter and receiver physically confines the signal within or through solid boundaries. Such media include wires, cables, waveguides, or optical fibers, whose transfer functions are generally frequency dependent. Of course, these elements are also used to interconnect devices within a transmit or receiving facility. Often the effect of these media is ignored, and indeed, depending upon the bandwidth of the signal, the frequency dependence may or may not be significant. When it is significant, that frequency dependence should be taken into account when simulating a system. These media possess a wide range of physical characteristics, depending upon material, shape, and generally the specific design. In this section, for illustrative purposes, we set down the transfer function for two important cases, rectangular waveguides and optical fibers.

9.3.1. Rectangular Waveguide Medium

At microwave frequencies, waveguides often serve as the connections between various elements of a system, e.g., between a receiving antenna and a low-noise amplifier. Waveguide runs can sometimes have appreciable length, and since the transfer function is directly proportional to length, the distorting effect may not be negligible.

For rectangular waveguides, it can be shown[45] that the transfer function is as follows. The amplitude characteristic $A(f) = 20\log|H(f)|$ is given by

$$A(f) = \frac{\alpha[(A/2B)(f/f_c)^{1.5} + (f/f_c)^{-0.5}]}{[(f/f_c)^2 - 1]^{1/2}} \quad \frac{\text{dB}}{\text{m}} \tag{9.3.1}$$

where α is a constant depending upon the material and the large dimension; A is the waveguide's large dimension; B is the waveguide's small dimension; and f_c is the cutoff frequency, $f_c = c/2A$, where c is the speed of light.

The phase characteristic is given by

$$\beta(f) = (360/c)f[1 - (f_c/f)^2]^{1/2} \quad \text{deg/m} \tag{9.3.2}$$

The functions $A(f)$ and $\beta(f)$, multiplied by the waveguide length, can be used directly in simulation, for example, by sampling them in frequency and storing them as tabular filters. However, as with the ionospheric phase filter, it is more illuminating to extract the complex lowpass-equivalent filter. We do this now for the phase characteristic (9.3.2).

Assume that the bandpass signal is centered on a carrier frequency f_0 so that we can write $f = f_0 + \Delta f$. Equation (9.3.2) then transforms to

$$\psi(\Delta f) = \left(\frac{360}{c}\right) f_0 \left(1 + \frac{\Delta f}{f_0}\right) \left[1 - \frac{(f_c/f_0)^2}{(1 + \Delta f/f_0)^2}\right]^{1/2} \tag{9.3.3}$$

which is the complex lowpass equivalent, and is plotted in Figure 9.27 as a function of the normalized lowpass frequency $\nu = \Delta f/f_0$. For plotting purposes we have assumed $f_0/f_c = 1.6$. Again, the figure suggests that this function is essentially parabolic, except for the linear phase term.

This can be demonstrated readily through the following steps, using $\nu = \Delta f/f_0$. First we use the fact that

$$(1 + \nu)^{-1} = 1 - \nu + \nu^2 - \nu^3 \cdots \approx 1 - \nu \qquad (\nu \leq 0.1)$$

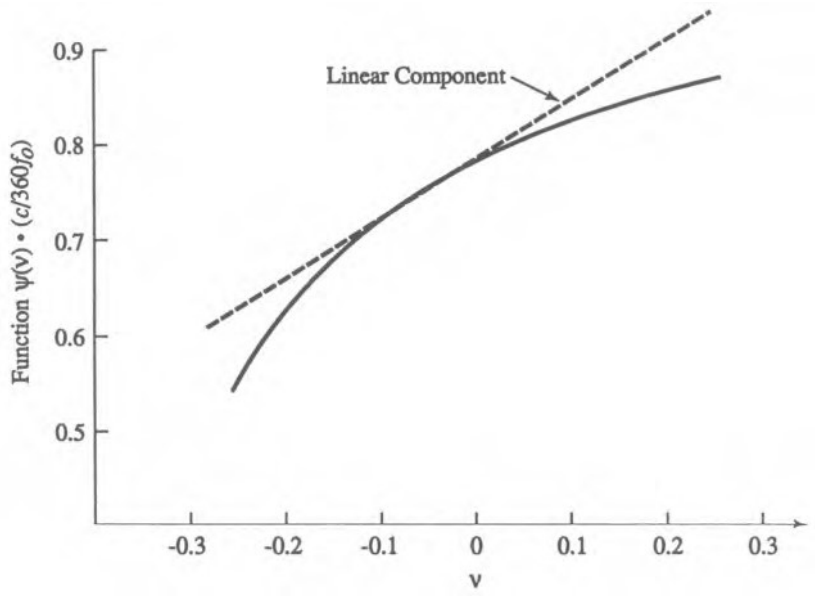


Figure 9.27. Rectangular waveguide lowpass-equivalent phase characteristic.

so that $(1+\nu)^{-2} \approx (1-\nu)^2$. We also use the approximation $(1-x)^{1/2} \approx 1-(x/2)$ for x small. Actually, the approximation is fairly good even for x not so small. Using the two approximations indicated yields

$$\psi(\nu) \approx \frac{360}{c} f_0 \left\{ 1 - \frac{1}{2}\nu_c^2 + \nu + \frac{1}{2}\nu\nu_c^2 + \frac{1}{2}\nu^2\nu_c^2 - \frac{1}{2}\nu^3\nu_c^2 + \cdots \right\}$$

where $\nu_c = f_c/f_0$. The constant and linear terms produce no distortion, and it can be seen that for $\nu \leq 0.1$ the parabolic term dominates.

9.3.2. The Fiber Optic Channel

Optical pulses propagating along an optical fiber are attenuated and distorted to an extent that depends on the chemical composition of the materials used to construct the fiber, as well as on the structure of the fiber.[46–48] Both manifestations are important because the nature of the detected optical signal depends generally both on the detector input signal level as well as on the transfer function of the channel. The main distortion is dispersion, of which there are two types, chromatic dispersion and intermodal dispersion. The former describes the effect due to the index of refraction in the fiber, which is a nonlinear function of wavelength.[47] In more familiar terms, the fiber has a nonlinear phase characteristic. Intermodal dispersion is an effect that is seen in multimode fibers and results from the fact that a large number of ray path lengths is supported by the fiber, generally with different delays.[47] Thus, intermodal dispersion can be seen as a form of distortion due to internal multipath. Splices between fibers introduce additional mode mixing and modeling difficulties.

A mathematical model for the baseband transfer function of an optical fiber can be approximated by (see, e.g., Ref. 49)

$$H(f) = \int_{-\infty}^{\infty} S(\lambda)L(\lambda)H_{\text{im}}(f)H_c(\lambda, f)\, d\lambda \tag{9.3.4}$$

where $S(\lambda)$ is the source spectrum as a function of wavelength; $L(\lambda)$ is the reciprocal of loss as a function of wavelength; $H_{\text{im}}(f) = \exp(-\sigma_{\text{im}}^2\omega^2/2 - j\omega t_d)$ (intermodal dispersion transfer function, normalized to unity gain at the center frequency); σ_{im} is the rms pulse spreading; $\omega = 2\pi f$; t_d is the fiber time delay; and $H_c(\lambda, f) = e^{-j\omega l T(\lambda)}$; l is the fiber length and $T(\lambda)$ is the group delay as a function of wavelength λ. Figure 9.28 shows for reference a typical loss curve $L^{-1}(\lambda)$ for a fiber.

The model (9.3.4) is valid when the source spectral width is large in comparison to the data rate. The model is linear in optical power (hence the name "power domain model" is sometimes used). For single-mode fibers there are no intermodal dispersion effects, so the transfer function for single-mode fibers can be written as

$$H_{\text{sm}}(f) = \int_{-\infty}^{\infty} S(\lambda)L(\lambda)\, \exp[-j\omega l T(\lambda)]\, d\lambda \tag{9.3.5}$$

The simulation model of the single-mode fiber would typically be constructed from the source spectrum, the loss, and the dispersion curves in tabular form. The source spectrum function

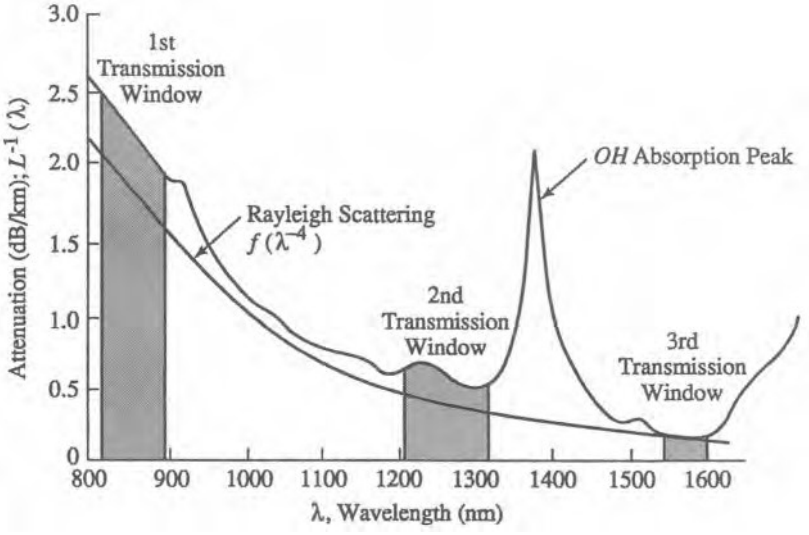


Figure 9.28. Typical loss curve as a function of wavelength for an optical fiber (from J. C. Daly, ed., *Fiber Optics*, © CRC Press, Boca Raton, Florida, 1984, reprinted with permission).

must be normalized such that $\int S(\lambda)\, d\lambda = 1$. For light emitting diodes LEDs, it is often appropriate to use a Gaussian shape to approximate the source spectrum,[46,47,49,50] namely,

$$S(\lambda) = \frac{1}{\sigma_s (2\pi)^{1/2}} \exp\left(\frac{-(\lambda - \lambda_s)^2}{2\sigma_s^2}\right) \tag{9.3.6}$$

where σ_s is the rms spectral width and λ_s is the source center wavelength.

Fiber loss as a function of wavelength would typically be input from a tabular data file created by the user. The function $L(\lambda)$ in (9.3.9) can be expressed as

$$L(\lambda) = 10^{-a(\lambda)l/10}$$

where $a(\lambda)$ is the fiber loss in dB/km and l and is the fiber length in km.

Chromatic dispersion $d(\lambda)$ is defined as the derivative of the group delay function $T(\lambda)$,

$$T(\lambda) = \int_{-\infty}^{\infty} d(u)\, du = D(\lambda) + c_0$$

where $d(\lambda)$ is the dispersion as a function of wavelength, $D'(\lambda) = d(\lambda)$, and c_0 is the constant of integration, $c_0 = T(\lambda_u) - D(\lambda_u)$. Here $T(\lambda_u)$ is a specified value of group delay at a specified wavelength λ_u. The chromatic dispersion as a function of wavelength can be modeled either as tabular data or a closed-form approximation.

Chromatic dispersion for silica fibers is sometimes approximated by (see e.g., Ref. 49)

$$\frac{dT(\lambda)}{d\lambda} = \frac{\delta}{c}\left[\frac{\lambda - \lambda_0}{\lambda^2}\right] \tag{9.3.7}$$

where δ is a dimensionless constant (typically around 0.047), c is the speed of light, and λ_0 is the zero-dispersion wavelength. The group delay for this approximation is

$$T(\lambda) = \frac{\delta}{c}\left[\ln\left(\frac{\lambda}{\lambda_0}\right) + \frac{\lambda_0' - \lambda}{\lambda}\right] \tag{9.3.8}$$

Two closed-form expressions for dispersion that are also used as fiber models (comité consultatif international de telephoné et telegraph CCITT standards) are as follows. For a standard (unshifted) fiber the dispersion is given by

$$d(\lambda) = \frac{1}{4}\delta_0(\lambda - \lambda_0^4/\lambda^3) \tag{9.3.9}$$

which results in the following group delay expression:

$$T(\lambda) = \frac{1}{8}\delta_0(\lambda - \lambda_0^2/\lambda)^2 \tag{9.3.10}$$

where the group delay evaluated at λ_0 is zero.

For a dispersion-shifted fiber the dispersion approximation is

$$d(\lambda) = \delta_0(\lambda - \lambda_0) \tag{9.3.11}$$

which corresponds to the following group delay function:

$$T(\lambda) = \frac{1}{2}\delta_0(\lambda - \lambda_0)^2 \tag{9.3.12}$$

In these CCITT standard expressions, the zero-dispersion slope is given by δ_0. A typical value for δ_0 is 0.087 ps/nm^2/km.

The transfer function model (9.3.4) or (9.3.5) can be implemented using the dispersion approximations (9.3.7), (9.3.9), or (9.3.11), or any other appropriate expression, or using tabular input. Some details on the computation of the integral can be found in Ref. 51. Of course, once the integral has been evaluated for a particular fiber, $H(f)$ can be stored as a tabular filter for simulation purposes.

When the width of the source spectrum becomes comparable to the modulation bandwidth, the power domain model becomes inadequate,[52] and a model for the single-mode fiber that assumes linearity in the electric field must be used. Coherent lightwave systems, which are under development, utilize lasers with very narrow bandwidths, making this model useful for analysis of the effects of chromatic dispersion in these systems.[53] The complex lowpass-equivalent transfer function of the fiber is then given by

$$H(f) = \exp\left(-j\pi d(\lambda_s)\frac{\lambda_s^2}{c}f^2\right) = \exp(-j\alpha f^2) \tag{9.3.13}$$

where c is the speed of light, λ_s is the source operating wavelength, and α is a constant which lumps all of the factors in the exponent. Thus, we recognize that a model of this medium is again a parabolic phase filter.

9.4. Finite-State Channel Models

The terminology "finite-state channel" is used to denote all the elements of a communication system that lie between any two points *a* and *b* in the system where the input entering at point *a* is a symbol sequence $\{X_k\}$ and the output of the system at *b* is another symbol sequence related to the input sequence $\{Y_k\}$ (see Figure 9.29). Usually, a will be the output of the channel encoder at the transmitter and *b* will be the input to the decoder in the receiver. The relationship between the input and output sequences will be affected by the distortion and noise introduced by filters and other elements that are present between *a* and *b*, and the physical channel. In a binary communication system with hard decision decoding, both $\{X_k\}$ and $\{Y_k\}$ will be binary sequences and transmission errors caused by all the elements in between *a* and *b* including the physical channel will cause $\{Y_k\}$ to be different than $\{X_k\}$ at least occassionally.

The nomenclature "finite state" implies a visualization of the channel (defined between some two points *a* and *b*) as being in one of several (but typically few) identifiable conditions, or "states". The transition from one state to another is governed by some probabilistic rule, which is part of the model. The model description is completed by specifying, for each state, an error generation mechanism. (Sometimes the terminology "discrete channel model" is also used in this context, the word "discrete" referring to the input and output alphabets at *a* and *b*. We avoid this usage here to prevent possible confusion with the use of "discrete" in the description "discrete *multipath* channels," where, of course, it implies a countable number of reflections.) Finite state channel models fall into two categories. The first category, called *memoryless* models, can be considered a degenerate case of the general class because such models have only a single state. Memoryless models are used to model the transmission errors or the transitions from input to the output under the assumption that there is no temporal correlation in the transition mechanism. That is, the probability of transition (or error) for the nth input symbol is not affected by what happened to any other input symbol. Such models are applicable to channels in which there is no ISI or fading and the noise is AWGN. In a hard decision binary system, this assumption implies that the bit errors are uncorrelated.

The second class of finite-state models are applicable to situations where the transitions from the input symbols to the output symbols are temporally correlated, i.e., the probability of transition for the *n*th symbol is correlated with the transitions of the preceding and or the following symbols. This will be the case in systems that have fading and impulse noise. Temporal correlation of signal amplitude variations due to fading will cause the transmission errors to be correlated. Errors will tend to occur in bursts and these channels are referred to as *burst error channels* or *channels with memory.*

Finite-state channel models are probabilistic models which are computationally more efficient than waveform-level models. The increased efficiency comes from two factors. The finite-state models are simulated at the symbol rate, whereas a waveform-level model will be

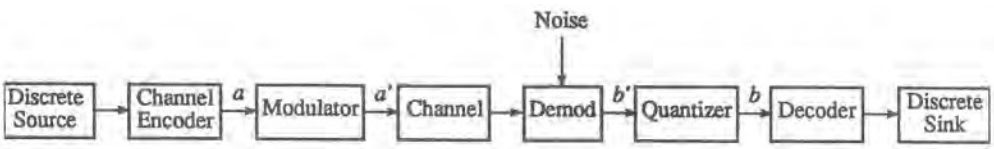


Figure 9.29. Definition of a finite-state or discrete channel: *a-b*; the waveform channel lies between a'-b'.

simulated at 8–16 times the symbol rate. While each individual block is simulated in detail in a waveform-level model, there is a high level of abstraction in the finite-state model. These two factors might contribute to several orders of magnitude savings in computational burden.

In simple cases, finite-state channel models can be derived analytically from the models of the underlying components between the input *a* and the output *b*. In most cases, however, finite-state channel models are derived from simulated or measured error patterns between the input *a* and the output *b*. Finite-state channel models are used to design and analyze error control coders, interleavers, etc., and also for network simulation.

Modeling of *memoryless* channels is rather straightforward. For example, in the case of a memoryless channel with binary input and binary output, all we need to know to characterize the channel model is the bit error probability. Simulation of this channel involves drawing a single random number to decide if a given bit will suffer a transmission error or not. Thus, if we have to simulate the transmission of 1 million bits through this channel, all we have to do is generate 1 million uniform random numbers. The entire system represented by this channel is simulated efficiently. (Compare this with a waveform-level simulation, which will involve generating the sampled values of a 1 million-bit-long waveform and processing it though all the functional blocks in the channel. Of course, we might have to do this a few times to derive the BER to be used in the discrete channel model, but this is usually done at a higher BER than is typical in a coded system.)

Finite-state channels with *memory* are harder to model. The temporally correlated error generation mechanism is usually modeled by a discrete-time Markov sequence[54–56] in which a state model is used to characterize the various states of the channel and a set of transition probabilities is used to capture the progression of the channel between various states. Each state will also have associated with it a set of input-to-output symbol transition probabilities. Thus the model now is more complex, with more parameters. The model structure and parameter values are estimated from either simulated or measured error patterns. Simulation of the finite-state model consists in generating a random number prior to the transmission of each symbol to determine the channel state and then drawing another random number to determine the input-to-output transition. While the model and parameter estimation procedures are complex, simulation of the Markov model is straightforward and very efficient.

We now look at the details of the finite-state channel models, beginning with the memoryless models. In this chapter, we only look at the construction of the model itself. The utilization of the model to derive various statistical measures of performance related to error occurrences will be taken up in Chapter 11.

9.4.1. Finite-State Memoryless Models

In a finite-state memoryless channel (recalling that it has only one state) the mapping of input to output is instantaneous and is described by a set of transition probabilities. A standard example of a simple binary memoryless channel model called the binary symmetric channel (BSC), is shown in Figure 9.30.

The input to the channel is an independent sequence of binary digits $\{X_k\}$ and the output is a binary sequence $\{Y_k\}$. A binary input of '0' is received correctly as '0' with a probability $1 - P_e$ and incorrectly as '1' with an error probability of P_e. The channel is symmetric in that zeros and ones are affected the same way. For the BSC, we can also express the input/output relationship as

$$Y_k = X_k \oplus \varepsilon_k$$

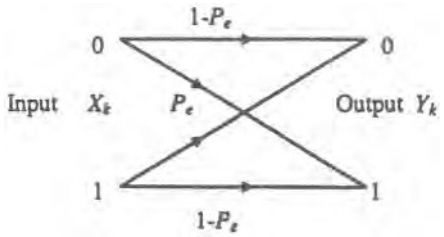


Figure 9.30. A simple discrete channel model: the binary symmetric channel (BSC).

where ε_k is an *independent* sequence of errors, $\varepsilon_k = 1$ indicating a transmission error. The only parameter of this model is the probability of error P_e, which can be easily estimated via simulations or from measurements.

A slightly more complicated model for the binary channel is the nonsymmetric model (Figure 9.3la) for systems in which the probability of error may be different for zeros and ones, as in some optical communication systems. This model is characterized by a set of input-to-output transition probabilities α_{ij}, $i, j = 0, 1$, and the channel is still memoryless. These transition probabilities are estimated using a detailed waveform-level simulation using a normalized count of the crossovers. When the finite-state channel model is simulated, a random number is drawn for each bit passing through the channel to determine whether the bit will suffer a transmission error or not.

Extension of this model to the case where the input and output belong to an M-ary alphabet is shown in Figure 9.31b. Another example, shown in Figure 9.31c, is used for memoryless channels in which the output of the channel is quantized to a different number of levels than the number of input levels as in soft decision decoding. In these models the transition probabilities are estimated from simulated or measured data as

$$\hat{\alpha}_{ij} = \frac{n_{ij}}{N_i}$$

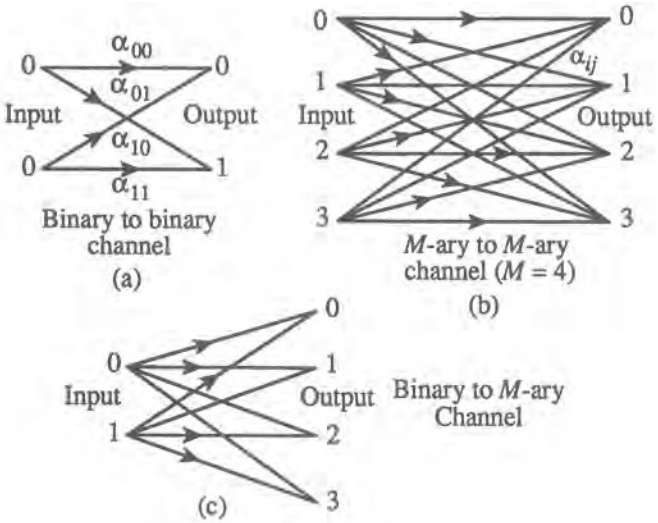


Figure 9.31. Other examples of discrete memoryless channel models.

where N_i is the number of occurrences of the ith input symbol and n_{ij} is the number of times the ith input symbol appears as the jth output.

Simulation of the channels shown in Figure 9.31 will involve once again the generation of a single random number for each input symbol to determine the input-to-output transition.

9.4.2. Finite-State Models with Memory: Hidden Markov Models (HMM)

For channels with memory the most commonly used model is the discrete-time, finite-state Markov model. There are several reasons for the popularity of Markov models. These models are analytically tractable, their theory is well established in the statistical literature, and they have been applied to communication problems, for example, to model the output of discrete information sources. Also, most importantly, computationally efficient techniques are available for estimating the parameters of Markov sequences from simulated or measured error patterns.

To set the stage for Markov models, consider a fading channel in which the received signal strength is above an acceptable threshold part of the time and below the threshold during a deep fade. Ignoring the in-between stages, we can assume the channel to be in either one of the following two "states": (1) a good state in which the received signal level is strong enough that the probability of transmission error for a binary communication system operating over the channel is almost zero, and (2) a bad state in which the received signal level is so low that the probability of error approaches 0.5. As time progresses, the channel goes from good state to bad state and vice versa. The rate of transition and the length of stay in each of the two states will depend on the temporal correlation of the fading process. If tune is measured in increments of a symbol (bit) tune, then we can construct a discrete channel model as follows. This model is referred to as the Gilbert model, discussed again later.

At the beginning of each symbol (bit) interval, the channel is one of the two states. If the channel is in a good state, then the transmitted bit suffers no transmission error. On the other hand, if the channel is in a bad state, the probability of transmission error is 0.5. Prior to the transmission of each new bit, the channel may change state or remain in the current state. This transition between states takes place with a set of transition probabilities a_{ij}. A graphical representation of the state transition diagram of the model is shown in Figure 9.32

Note that when the channel is in good state, there are no transmission errors, and when the channel is in bad state, it will produce a high concentration of errors. Also, the error sequence will be statistically dependent since the state of the channel during the $(n + 1)$th bit interval (and hence the probability of occurrence of an error) will depend on the state the channel was in during the nth bit interval. Also note that each particular state corresponds to a

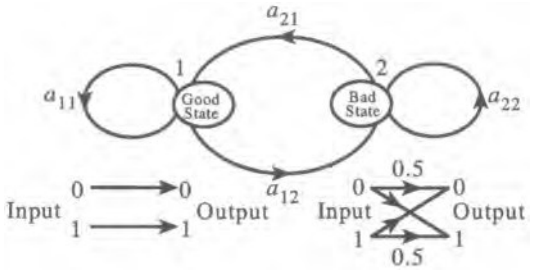


Figure 9.32. Example of the state transition diagram for a two-state Markov model (the Gilbert model).

different BSC model. By comparison, a memoryless channel has only one state and one BSC model.

A more general version of the above model can be constructed with a larger number of states, with many good and bad states with varying degrees of "goodness" or "badness" corresponding to increasingly more severity of fading and/or noise. Such an approach leads to the more general Markov model.

9.4.2.1. N-State Markov Model

An N-state Markov (process) model for a discrete communication channel is defined by the following parameters:

- Set of states: $\{1,2,3,\ldots,N\}$.
- State at time $t = S_t$.
- Set of state probabilities:

$$\pi_i(t) = \text{Probability of being in state } i \text{ at time } t$$
$$= P[S_t = i], \qquad i = 1, 2, \ldots, N$$

- Set of *state transition* probabilities:

$$\text{Probability of going from state } i \text{ at time } t \text{ to state } j \text{ at time } t + 1$$
$$= a_{ij}(t) = P[S_{t+1} = j | S_t = i], \qquad i, j = 1, 2, \ldots, N$$

(time increment is usually equal to the symbol or bit duration).

- Set of *input-to-output transition* (symbol error) probabilities for each state:

$$\text{Error symbols: } E = \{e_1, e_2, \ldots, e_M\}; \quad E = \{1, 0\} \text{ in the binary case}$$
$$b_i(e_k) = P[\text{error symbol } e_k \text{ occurring} | S_t = i]$$

These parameters define a discrete-time Markov process operating at a transition rate

of two sequences: the sequence of states $\{S_t\}$ and a sequence of error symbols $\{E_t\}$, where t is the time index, which can be indexed over the integer set $\{0,1,2,\ldots\}$. Normally, only the input and the output of the channel and hence the error sequence can be observed, and the state sequence itself cannot be easily observed in a physical channel. Hence, the state sequence is "hidden" or not visible from external observations, and such a Markov model is called a *hidden Markov model* or HMM.

The theory of HMMs is well developed[57,58] and HMM models have been used in a number of applications including the modeling of symbol sequences emitted by discrete information sources and speech coding. For example, the occurrence of a sequence of letters of the English alphabet in a typical passage can be modeled as a hidden Markov sequence. Similarly, the sampled values of an audio or video waveform can also be modeled by an HMM. Such models have been developed and used extensively to design source encoders in communication systems. Procedures have also been developed for estimating the parameters of such models from data. These techniques are directly applicable to modeling, analysis, and simulation of discrete (finite-state) communication channels. Also, the HMM models can be

used to evaluate the capacity of a discrete channel and for the design of optimal error control coding techniques.

9.4.2.2. First-Order Markov Process

The Markov property is defined as

$$P[S_{t+1}|S_t, S_{t-1}, \ldots] = P[S_{t+1}|S_t, S_{t-1}, \ldots, S_{t-m-1}] \tag{9.4.1}$$

for a Markov process with memory m. For a first-order Markov process

$$P[S_{t+1}|S_t, S_{t-1}, \ldots] = P[S_{t+1}|S_t]$$

The Markov property implies that the future behavior of the system depends on the current state and the m previous states, but not on the past states beyond $t - m - 1$ or how the system got there. For channel models we will use a first-order Markov model. (For applications such as modeling the symbol sequences in an English language text, a second- or third-order model may be adequate since the influence of the preceding symbols on the probability of occurrence of a later symbol does not span more than two or three symbols.)

9.4.2.3. Stationarity

Since Stationarity is usually assumed in modeling the analog portion of the communication channel, it is common to assume Stationarity of the Markov model for discrete channels also. Stationarity implies that the parameters of the model, i.e., the probabilities $\pi_i(t)$, $a_{ij}(t)$, and $b_i(e_k)$, do not depend on t. In this case, we have

$$P[S_{t+1} = i] = \pi_i = \sum_{k=1}^{N} P[S_{t+1} = i|S_t = k]P[S_t = k]$$

$$= \sum_{k=1}^{N} a_{ki}\pi_k \tag{9.4.2}$$

In terms of vector notation,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{N1} & \cdot & \cdots & a_{NN} \end{bmatrix}, \qquad \Pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{bmatrix}$$

we can write Equation (9.4.2) in the alternate form

$$\Pi = A^T\Pi \tag{9.4.3}$$

Using Equation (9.4.3), we can show that for a for a first-order Markov process, the n-stage transition matrix is given by A^n and also that $A = A^n$.

Equation (9.4.3) along with the constraint that

$$\sum_{i=1}^{N} \pi_i = 1$$

implies that the π_i are uniquely determined from the a_{ij}, and hence the Markov model is completely defined by the matrix A of the state transition probabilities and B, where B is the matrix of input-to-output symbol transition (i.e., error) probabilities,

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1M} \\ \vdots & \vdots & \vdots \\ b_{N1} & \cdots & b_{NM} \end{bmatrix}, \qquad b_{ik} = b_i(e_k)$$

that is, $b_i(e_k)$ is the probability of observing the error symbol e_k from state i.

■ *Example 9.4.1.* A Markov Channel Model. Lets assume a quaternary ($M = 4$) communication system with a channel represented by a two state ($N = 2$) Markov model as in Figure 9.32 with a *state transition matrix*

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \end{bmatrix};$$

The *state probabilities* Π are calculated from

$$\Pi = \mathbf{A}^T \Pi$$

or

$$\begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} 0.9 & 0.8 \\ 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix};$$

and

$$\pi_1 + \pi_2 = 1$$

The two equations above result in $N = 2$ linearly independent equations, from which the state probabilities are evaluated:

$$\pi_1 = 8/9 \quad \text{and} \quad \pi_2 = 1/9$$

The Input-to-Output Transition Probabilities. For a quaternary system the error symbols are defined as:

$$\mathbf{E} = \{e_1, e_2, e_3, e_4\}$$

where $e_k = S_r - S_t$, and S_r = received symbol, S_t = transmitted symbol.

The error symbols are then

$$e_1 = 0\ (00) \qquad \text{(no error)}$$
$$e_2 = 1\ (01)$$
$$e_3 = 2\ (10)$$
$$e_4 = 3\ (11)$$

or

$$\mathbf{E} = \{0,\ 1,\ 2,\ 3\}$$

Let's assume that for state $\mathbf{S}_1$ (good state) the probabilities $b_{ik} = b_i(e_k)$ are

$$b_{11} = 0.8$$
$$b_{12} = 0.1$$
$$b_{13} = 0.1$$
$$b_{14} = 0.0$$

and for $\mathbf{S}_2$ (bad state) the *input-to-output transition probabilities* are:

$$b_{21} = 0.0$$
$$b_{22} = 0.4$$
$$b_{23} = 0.4$$
$$b_{24} = 0.2$$

The *input-to-output transition probabilities* $N \times M$ matrix is then:

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.1 & 0.1 & 0.0 \\ 0.0 & 0.4 & 0.4 & 0.2 \end{bmatrix}$$

The Average Probability of Error. The average probability of error e_k is

$$P_{\text{ave},k} = \mathbf{\Pi}^T \mathbf{b}_k$$

where $\mathbf{b}_k$ is the is the kth column of the matrix $\mathbf{B}$. Thus the probabilities of the channel producing the specific error e_k are:

$$P_{\text{ave},1} = 8/9 \times 0.8 + 1/9 \times 0.0 = 0.711,$$
$$P_{\text{ave},2} = 8/9 \times 0.1 + 1/9 \times 0.4 = 0.133,$$
$$P_{\text{ave},3} = 8/9 \times 0.1 + 1/9 \times 0.4 = 0.133,$$
$$P_{\text{ave},4} = 8/9 \times 0.0 + 1/9 \times 0.2 = 0.022.$$

9.4.3. Types of Hidden Markov Models: Gilbert and Fritchman Model

The characterization of channels using discrete-time, finite-state Markov sequences has been proposed by Gilbert,[59] Fritchman,[60] and others. The Gilbert model is a two-state model with a good error-tree state and a bad state with an error probability of p. This model was used by Gilbert to calculate the capacity of a channel with burst errors. Parameters of the model can be estimated from data using the procedures we will present later in this section.

The Fritchman model, first proposed in 1967, is now receiving a considerable amount of attention since it seems to be very suitable for modeling burst errors in mobile radio channels and also because it is relatively easy to estimate the parameters of the Fritchman model from burst error distributions. For binary channels, Fritchman's framework divides the state space into k good states and $N - k$ bad states. The good states represent error-free transmissions and the bad states always produce a transmission error. Hence the entries in the B matrix are zeros and ones and they need not be estimated.

The state transition matrix A in this model can be partitioned as

$$A = \begin{bmatrix} A_{GG} & A_{GB} \\ A_{BG} & A_{BB} \end{bmatrix}$$

where the submatrices represent the transition probabilities between various good and bad states. For this model it is possible to derive analytically the expressions for burst error distributions in terms of the model parameters and use these expressions to estimate the parameters of the model from empirical (measured or simulated) burst error distributions.

Let $O = \{O_1, O_2, \ldots, O_T\}$ be an error sequence, $O_k = 1$ indicating that the kth transmitted bit suffered a transmission error and $O_k = 0$ indicating error-free transmission. Also, let the notation $(0^m|1)$ denote the event of observing m or more consecutive (i.e., a burst of) error-free transmissions following an error, and $(1^m|0)$ represent the event of observing m or more consecutive errors following a good period. Fritchman showed that the probabilities of occurrence of these two events are given by the sum of weighted exponentials

$$P(0^m|1) = \sum_{i=1}^{k} f_i \lambda_i^{m-1} \tag{9.4.4a}$$

and

$$P(1^m|0) = \sum_{i=k+1}^{N} f_i \lambda_i^{m-1} \tag{9.4.4b}$$

where $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ and $(\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_N)$ are the eigenvalues of A_{GG} and A_{BB}, respectively, and the f_i are functions of the a_{ij}. From Equation (9.4.4) we can obtain the probability of obtaining exactly m zeros (or ones) as

$$P(0^{m-1}|1) - P(0^m|1) = \sum_{i=1}^{k} f_i \lambda_i^{m-1}(1 - \lambda_i) \tag{9.4.5a}$$

and

$$P(1^{m-1}|0) - P(1^m|0) = \sum_{i=k+1}^{N} f_i \lambda_i^{m-1}(1 - \lambda_i) \tag{9.4.5b}$$

The Fritchman model can be interpreted as equivalent to a Markov process with a state transition probability matrix

$$\tilde{A} = \begin{bmatrix} \Lambda_{GG} & A_{GB} \\ A_{BG} & \Lambda_{BB} \end{bmatrix}, \qquad \Lambda_{GG} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix} \quad \text{and} \quad \Lambda_{BB} = \begin{bmatrix} \lambda_{k+1} & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}$$

Note that in this equivalent model there are no transitions within the set of $N - k$ good states and no transitions within the set of bad states. Such transitions are indistinguishable from the observed error sequence since a transition from one good state to another good state still produces no errors and the transition is not observable from the output.

The Fritchman model is not unique except when there is only one bad state. This is so because, in the general case, the error-free run distribution $P(0^m|1)$ and the error burst distribution $P(1^m|0)$ do not specify the statistical dependence of the error-free runs and the error bursts. In the case of a single-error-state model (Figure 9.33), Fritchman showed that

$$P(0^m|1) = \sum_{k=1}^{N-1} \frac{a_{Nk}(a_{kk})^m}{a_{kk}}, \qquad m \geq 1 \tag{9.4.6}$$

From Equation (9.46) it can be seen that in the case of a single-error-state model, the error-free run uniquely specifies the $2(N - 1)$ model parameters. An empirical procedure is used to fit an $N - 1$ mixture of exponentials as in Equation (9.4.6) to the (simulated or measured) error-free run distribution. Figure 9.34 shows an example from Fritchman's paper.

While Fritchman's model is applicable to discrete channels with simple burst error distributions, it may not be adequate to characterize very complex burst error patterns, which will require more than one error state in the model. In such cases it will be very difficult to estimate the model parameters from the burst error distributions alone. Fortunately, it is possible to find a maximum likelihood estimate of the parameters using iterative techniques.

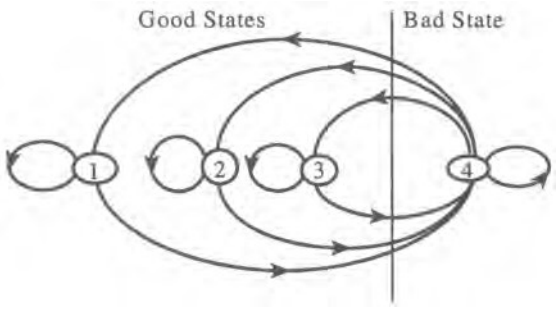


Figure 9.33. A simplified Fritchman model with a single bad state.

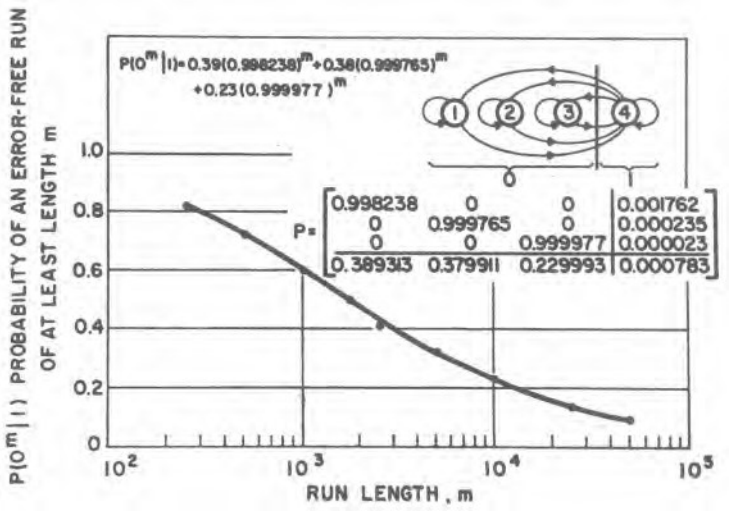


Figure 9.34. Example of model fitting to error-free run distribution (from Ref. 60, © IEEE, 1967).

9.4.4. Estimation of the Parameters of a Markov Model

The Markov model for a discrete channel is described by the $N \times N$ state transition matrix A and the $N \times M$ input-to-output transition (error probability) matrix B. An iterative procedure for estimating these parameters $\Gamma = \{A, B\}$ from a given set of a simulated or measured error sequence $\bar{O} = \{O_1, \ldots, O_t, \ldots, O_T\}$ was developed by Baum and Welch.[61] This iterative algorithm, described e.g., in Ref. 62, is designed to converge to the maximum likelihood estimator of $\Gamma = \{A, B\}$, i.e., the Γ that maximizes $P[\bar{O}|\Gamma)$. The algorithm consists of the following steps.

Step 0. Start with an initial model $\Gamma = \{A, B\}$.

Step 1. With $\Gamma = \{A, B\}$ as the model, compute the "forward variables" $\alpha_t(i) = P[O_1, O_2, \ldots, O_t, s_t = i|\Gamma]$ and the "backward variables" $\beta_t = P[O_{t+1}, O_{t+2}, \ldots, O_T|s_t = i, \Gamma)$ for $t = 1, 2, \ldots, T$ and $i = 1, 1, 2, \ldots, N$ as follows:

Forward variables (see Figure 9.35a for details)

Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad i = 1, 2, \ldots, N$$

Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad 1 \le t \le T-1;\ 1 \le j \le N$$

Termination:

$$P[\bar{O}|\Gamma) = \sum_{i=1}^{N} \alpha_T(i)$$

Note:

$$\sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} P[O_1, \ldots, O_T, s_T = i|\Gamma] = P[\bar{O}|\Gamma]$$

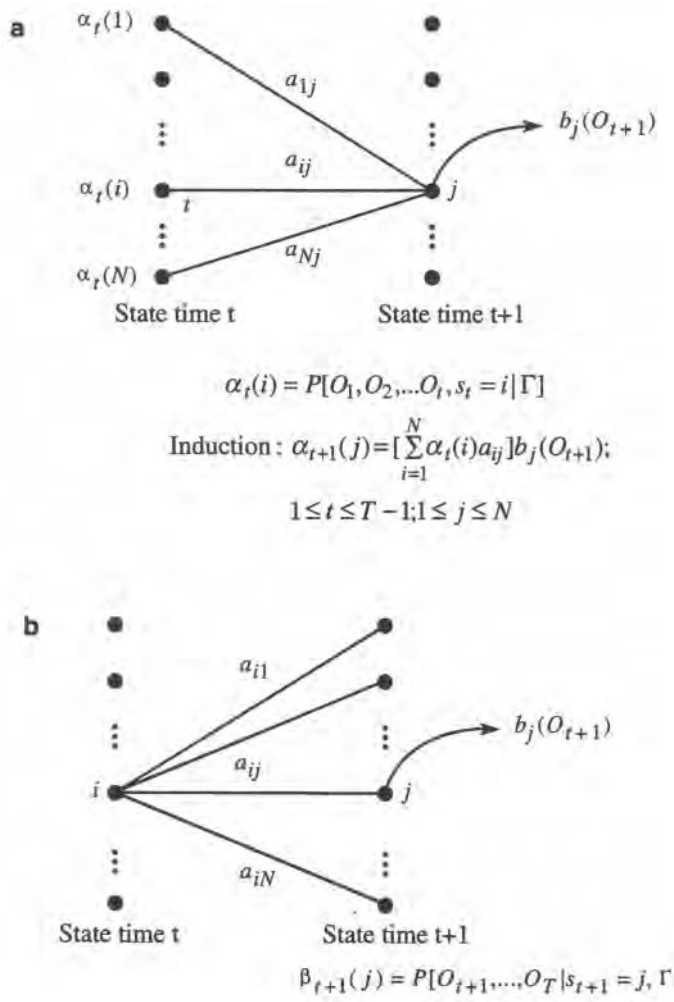


Figure 9.35. The process of estimating the parameters of a hidden Markov model, (a) Computation of the forward variables; (b) computation of the backward variables.

Backward variables (see Figure 9.35b for details)

Initialization:

$$\beta_T(i) = 1, \qquad i = 1, 2, \ldots, N$$

Induction:

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) b_j(O_{t+1}) a_{ij}, \qquad 1 \le t \le T-1; \ 1 \le j \le N$$

Step 2. Compute the state transition probabilities given $\bar{O}$ and Γ and update the parameters A, B according to

$$\gamma_t(i) = P[s_t = i|\bar{O}, \Gamma] = \frac{\alpha_t(i)\beta_t(i)}{P[\bar{O}|\Gamma]}, \qquad i = 1, 2, \ldots, N$$

$$\xi_t(i,j) = P[s_t = i, s_{t+1} = j|\bar{O}, \Gamma]$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P[\bar{O}, \Gamma]}$$

$$\hat{a}_{ij} = \frac{\text{expected \# of transitions from } i \text{ to } j}{\text{expected \# of transitions from } i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\hat{b}_j(e_k) = \frac{\text{expected \# of times } e_k \text{ is emitted from state } j}{\text{expected number of transitions from state } j}$$

$$= \frac{\sum_{t=1, O_t=e_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

Step 3. Go back to step 1 with the new values of $\hat{\Gamma} = \{\hat{A}, \hat{B}\}$ (i.e., $\hat{\Gamma} = \Gamma$ obtained in step 2, and repeat until convergence, i.e., until successive values of $P[\bar{O}|\hat{\Gamma}]$ differ very little. (The Baum–Welch algorithm is guaranteed to converge to the maximum likelihood solution; proof is also given in Ref. 62.)

The Baum–Welch algorithm is one of many reestimation algorithms available for iterative computation of the likelihood ratio estimate of the model parameters. While the algorithm is computationally efficient and converges rapidly, the overall computational requirements are still high since the forward and backward variables are computed for each error symbol in the given error sequence, which could be several thousand symbols long, and potentially in the millions. Several ways of improving the computational efficiency have been proposed recently. One of the methods, proposed by Turin and Sondhi,[63] involves the computation of the forward and backward variables using the following matrix version (for the binary channel with good and bad states):

$$A = \begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix}$$

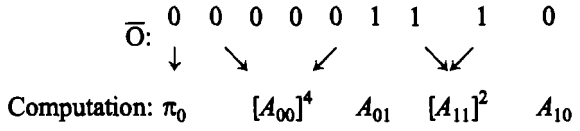
$$\bar{O}: \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0$$

$$\text{Computation: } \pi_0 \quad [A_{00}]^4 \quad A_{01} \quad [A_{11}]^2 \quad A_{10}$$

in which the powers of submatrices involved in the computations can be precomputed and reused. Another modification[64] is based on the fact that for a general Markov model there is an equivalent Fritchman-like model with good states and bad sates and an A matrix that has the form

$$A = \begin{bmatrix} \Lambda_{00} & A_{01} \\ A_{10} & \Lambda_{11} \end{bmatrix}$$

where Λ_{00} and Λ_{11} are diagonal matrices. With this model, the channel remains in the same state during a burst and changes state only at the end of a burst. Hence, all the variables are computed only at time steps involving the change of error symbol, i.e., at the beginning of each burst rather than once every symbol. The computations now take the form



Note that the computations during long bursts of 1s and 0s now involve raising the power of a diagonal matrix rather than raising the power of arbitrary matrices, and matrix multiplications occur only at the transition from one burst to another. Thus the computations are very efficient and long error bursts can be processed without excessive storage and computational requirements.

Irrespective of the algorithms used, the Markov model for a discrete channel has to be computed for different values of the parameters of the underlying physical channel. If the discrete channel model is for a Rummler channel, then the Markov model has to be developed from measurements or waveform-level simulations for different values of the parameters of the Rummler channel and SNR. Thus a parametrized set of Markov models for an underlying Rummler channel can be developed and used for the design and analysis of error control coders, interleavers, etc.

The development and use of Markov models for burst error channels (channels with memory) is currently an active area of research. The interested reader may find recent articles on this topic in the *IEEE Transactions on Communications*, *the IEEE Transactions on Vehicular technology*, and other journals and conference proceedings. Markov models are also being developed for burst errors at the higher layers of communication networks, for example, packet errors at the transport layer in a network.

Simulation of the Markov Model. Once the Markov model is derived, simulation of the model is relatively easy. The model is simulated either at the symbol rate in the system or slower. If the model is simulated at the symbol rate, then a uniform random number in the interval [0, 1] is drawn at the beginning of each symbol interval to determine the state transition according to the entries in A, and then a second number is drawn to determine the input-to-output symbol transition according to the entries in the B matrix corresponding to the new state.

9.5. Methodology for Simulating Communication Systems Operating over Fading Channels

In Chapter 2, we discussed the methodology of simulation from a broader perspective. Some of the ideas discussed there will become clearer when placed in a specific context. We can illustrate various aspects of the methodology, and in particular the notion of partitioning a problem into more manageable subproblems, by considering the simulation of a communication system operating over a fading channel. Consider a typical cellular radio system for voice communication. The performance of the cellular radio interface ("air interface") is usually evaluated at three different levels.

At the physical level the performance of the *uncoded* system is evaluated using a detailed waveform-level simulation. The main objective of this level of simulation is the comparative evaluation of various modulation, equalization, and filtering schemes and optimization of system parameters. The output of this level of simulation is the uncoded bit error rate performance and a description of the burst error characteristics of the radio channel including all the effects of waveform-level processing blocks such as modulators, filters, and equalizers. The burst error characteristic of the channel is captured in a Markov model which is parameterized in terms of fading parameters and SNR.

At this point, it is useful to elaborate a bit on our definition of "level". Because we are dealing with a sampled analog waveform, the waveform level corresponds to a continuous-amplitude, discrete-time model. Of course, the amplitude is not truly continuous, but can be so considered for present purposes. The next level of simulation will be of the *coded* system using the Markov model. Here, the signal can be more precisely described as a discrete-amplitude, discrete-time sequence. Because the signal amplitudes belong to a discrete alphabet, we sometimes refer to this level as "discrete". However, for the reason mentioned earlier, we avoid this usage here. Instead we will refer to this level of simulation as the *symbol level*, which in the most general case will imply not only legitimate symbol values, but quantized amplitudes as well (e.g., soft decisions). At the symbol level, comparative performance of coding, interleaving, and scrambling schemes are evaluated and the parameters of this portion of the system are optimized. The output of this level of simulation is again the BER characteristics of the coded system as a function of fading parameters and SNR.

The third level of performance evaluation deals with the speech coder and the voice quality. This will involve simulating the voice coder with a finite-state channel model for the coded system, and evaluating the voice quality subjectively. This level can also be described as symbol-level simulation. The distinction between this and the previous level will be in the nature of the finite-state model. For example, a memoryless model can often be used here under the assumption that the interleaver in conjunction with the error control coder will randomize burst errors. Otherwise, we will still need a Markov model at this level. Of course, for signals other than speech, the third level will be tailored to the signal in question, if indeed there is a third level at all.

The ultimate measure of performance for voice communication in a cellular radio system is the voice quality as measured by subjective listening tests, which produce a mean opinion survey (MOS). The MOS ranges from 1 to 5, with 5 being an excellent rating. Typically, the quality of service of a cellular radio system will be specified in terms of the requirement that 99% of the time the MOS is better than 2.0 (i.e., the probability of system outage or unacceptable quality of service below an MOS of 2 is 1%). The overall methodology used to carry

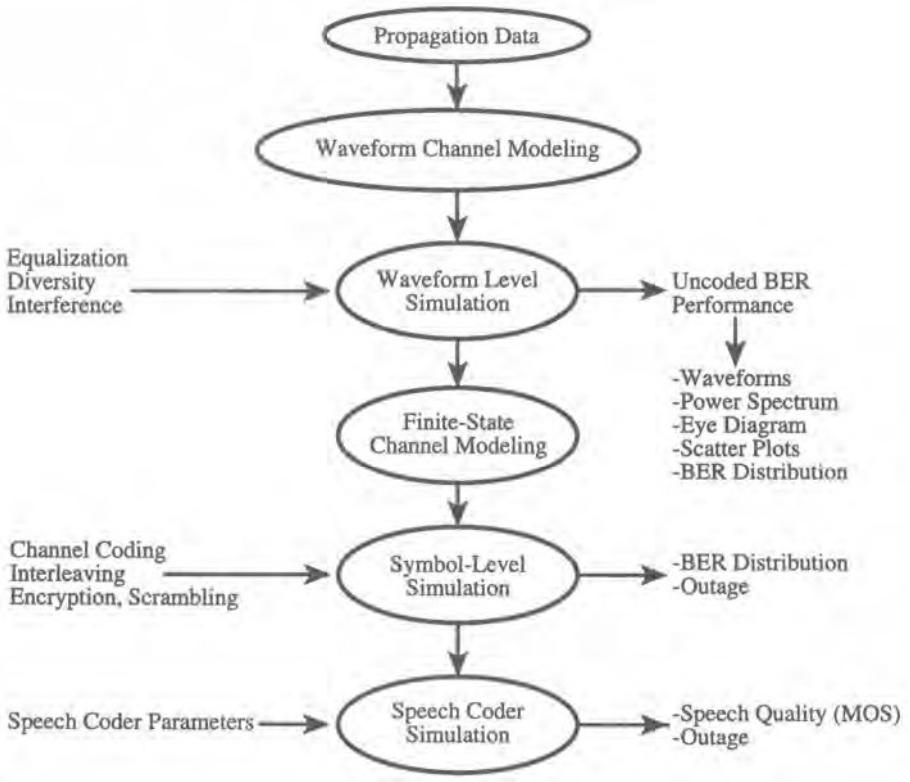


Figure 9.36. Organization of the methodology for simulating the performance of a voice communication system operating over a fading channel.

out the simulations leading up to the calculation of outage probability performance is shown in Figure 9.36; the details are as follows.

9.5.1. Waveform-Level Simulation

Waveform-level simulation of the system provides a detailed simulation of all the signal processing functions in the transmitter and receiver and it includes blocks such as modulator, demodulators, channel, and equalizer. Interference and noise are included in the simulations either through Monte Carlo or semianalytic techniques as described elsewhere in the book. The waveform-level simulation is used to compare different methods of modulation, equalization, etc., and it is also used to evaluate subsystem performance such as equalizer convergence rate. Also, optimization of parameters such as number of taps in the equalizer, tap spacing, and filter bandwidths are done using waveform-level simulations.

The output of waveform-level simulations includes spectral plots, eye diagrams, and BER curves. The performance analysis includes the following conditions: selected snapshots of the channel and optimized system components such as equalizers and filters; dynamic fading channels and optimized system components; cochannel interference, adjacent channel interference, and noise.

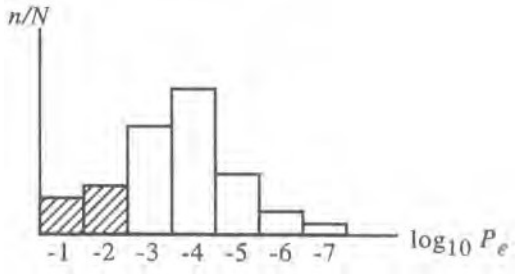


Figure 9.37. Distribution of BER in an uncoded system: the shaded area represents the outage probability at a BER threshold of 10^{-3} (i.e., n/N, where n is the number of channels out of N simulated channels that produced a BER> 10^{-3}).

In order to compute the overall quality of service in the system as measured by MOS and outage probability at a particular MOS level, the simulations at this level have to be carried out over thousands of static channel conditions and a distribution of the uncoded link BER has to be obtained. This distribution, usually shown as a histogram of error probabilities (see Figure 9.37 for an example), is used later in symbol-level simulations to deduce the MOS and outage probability for the entire system.

Unfortunately, this poses a significant challenge since the number of channels simulated may be large. For, if the fading statistics of the channel are described by a higher order multidimensional distribution (distributions for the number of multipath components, delay power profiles, urban versus suburban conditions, and a range of SNR values), it is easy to see that a large number of samples have to be drawn from this distribution to produce a statistically significant sample set of channel conditions. Since we have to estimate the BER for each channel condition via simulations, considerable simplification has to be made in the simulation model. Quasianalytical and other fast methods of BER estimation have to used in order to reduce the computational burden associated with computing the BER distribution to a reasonable level.

Waveform-level simulation is also used to extract a hidden Markov type finite-state channel model for use at the next level. These finite-state channel models have to be parametrized by channel rms delay spread, mobility (Doppler bandwidth), signal power, noise, and interference as well as by what other system parameters are included in the tradeoff space. Assuming that each of the m parameters is quantized to n levels, a library of m^n finite-state channel models has to be created from the waveform-level simulations. Fortunately, the BER for the uncoded portion of the system may be as high as 10^{-2} (compared to a typical BER of 10^{-6} for the coded system) and hence each waveform-level simulation will consist of only a few thousand symbols.

9.5.2. Symbol-Level Simulation

The finite-state channel models developed through waveform-level simulations emphasize the bursty nature of errors in the physical channel due to fading, noise, and interference. These models are used to analyze the comparative performance of different types of error control coding, interleaving, scrambling, and encryption schemes. Note that all of these operations are algorithmic operations on symbol sequences and hence the finite-state channel model is most appropriate for evaluating and optimizing these operations and for

evaluating the performance of the coded portion of the communication system. While a representative set of finite-state Markov models will be useful for producing the dynamic evolution of error bursts and their impact on the dynamic performance of the coded system, burst error distributions rather than the actual error sequence may be adequate to obtain the BER distribution of the coded system. There are many analytical bounds available for computing the coded system BER as a function of the burst error distribution in the uncoded channel (see also Chapter 11).

The output of this level of simulation is the distribution of average BER in each simulated coded link, i.e., $P[P_e > P_0]$, for various levels of error probability thresholds, where P_e is the (average) decoded error probability. With an interleaver, it can usually be assumed that errors produced in the coded link are independent and hence we need only the average BER for each coded channel that is simulated. Note, however, that the notion of a BER distribution is meaningful or applicable only in a quasistatic environment, i.e., one where the channel state is relatively unchanging for a large number of symbols.

9.5.3. Speech Coder Simulation

Finally, the voice quality performance of the end-to-end link is evaluated by relating the average BER in the coded link to the MOS. This is done by sending digitized voice over the voice encoder and decoder and inserting errors in the coded voice bit stream at various BERs. The output of the decoder is recorded and played back to a set of listeners and their subjective assessment of voice quality is noted as a function of injected BER. A typical example is shown in Table 9.2. Using the MOS measure shown in this table, we can map the error probability distribution in the coded link to the distribution of MOS over the entire link. From the distribution of MOS we can calculate the outage probability at various MOS thresholds. If the underlying channel models have been parametrized in terms of rms delay spread, SNR, etc., then we can relate overall outage probability of the link to these parameters.

9.6. Summary

In this chapter we developed models, simulation techniques, and simulation methodologies for communication channels. Simulation models for communication channels take one of three forms: (1) transfer functions (2) TDL with tap-gain functions which are stationary random processes, or (3) discrete-time Markov sequences. The first two are waveform-level

Table 9.2. Mean Opinion Survey (MOS) Values as a Function of BER

Average BER	MOS
10^{-5}	5.0
10^{-4}	4.5
5×10^{-3}	3.5
10^{-3}	3.0
5×10^{-2}	2.0
10^{-2}	1.0

simulation models and the third one is used for simulating the discrete portion of the communication channel.

The most difficult part of modeling and simulation of communication systems is the modeling part. Several examples were presented with an emphasis on deriving the TDL and the Markov models. Once the models and their parameters are specified, simulation is straightforward.

In the case of fading channels, the simulation methodology used will depend on the end objective. For detailed design of subsystems and for design optimization, waveform-level simulation with dynamic channel models might be appropriate. For performance estimation (BER and outage probability), static channel models and hierarchical techniques are most appropriate, where applicable.

9.7. Appendix Reference Models for Mobile Channels

Discrete channel models are widely used for indoor and outdoor wireless system simulations. Given the large number of both mathematical and empirical models that have been proposed recently, the designer of communication systems is faced with the difficult problem of choosing a set of channel models that will represent the channels over which the communication system will operate satisfactorily. Fortunately, some guidance on the choice of which models to use has been provided by international standards bodies, which specify a set of *representative* channels for analyzing and simulating the performance of different communication systems. Such *standard* or *reference* models are also very important for comparison of performance of the different systems provided by industry. Below we present examples of reference channel models for three different applications.

9.A.1. Reference Channel Models for GSM Applications

Groupe Special Mobile, GSM, is one of the worldwide standardized systems for mobile communications, originally developed for the 1- to 2-GHz frequency range[65]. The multiple access technique is time division multiplexing (TDMA), and the individual channels are 200 kHz wide.

Reference channel models have been developed for GSM applications, as shown in Tables 9.A.1–9.A.6. The recommended GSM models are discrete multipath models, specified for three different environments: rural, hilly, and urban. For the first, a 6-ray (path) model is

Table 9.A.1. Typical Profile for Rural Areas

Tap number	Relative time (μs)		Average relative power (dB)		Doppler spectrum
	(1)	(2)	(1)	(2)	
1	0.0	0.0	0.0	0.0	Ricean
2	0.1	0.2	−4.0	−2.0	Jakes
3	0.2	0.4	−8.0	−10.0	Jakes
4	0.3	0.6	−12.0	−20.0	Jakes
5	0.4	—	−16.0	—	Jakes
6	0.5	—	−20.0	—	Jakes

Table 9.A.2. Typical Profile for Hilly Terrain

Tap number	Relative time (μs)		Average relative power (dB)		Doppler spectrum
	(1)	(2)	(1)	(2)	
1	0.0	0.0	−10.0	−10.0	Jakes
2	0.1	0.2	−8.0	−8.0	Jakes
3	0.3	0.4	−6.0	−6.0	Jakes
4	0.5	0.6	−4.0	−4.0	Jakes
5	0.7	0.8	0.0	0.0	Jakes
6	1.0	2.0	0.0	0.0	Jakes
7	1.3	2.4	−4.0	−4.0	Jakes
8	15.0	15.0	−8.0	−8.0	Jakes
9	15.2	15.2	−9.0	−9.0	Jakes
10	15.7	15.8	−10.0	−10.0	Jakes
11	17.2	17.2	−12.0	−12.0	Jakes
12	20.0	20.0	−14.0	−14.0	Jakes

Table 9.A.3. The Reduced Profile for Hilly Terrain (Six Taps)

Tap number	Relative time (μs)		Average relative power (dB)		Doppler spectrum
	(1)	(2)	(1)	(2)	
1	0.0	0.0	0.0	0.0	Jakes
2	0.1	0.2	−1.5	−2.0	Jakes
3	0.3	0.4	−4.5	−4.0	Jakes
4	0.5	0.6	−7.5	−7.0	Jakes
5	15.0	15.0	−8.0	−6.0	Jakes
6	17.2	17.2	−17.7	−12.0	Jakes

Table 9.A.4. Typical Profile for Urban Area

Tap number	Relative time (μs)		Average relative power (dB)		Doppler spectrum
	(1)	(2)	(1)	(2)	
1	0.0	0.0	−4.0	−4.0	Jakes
2	0.1	0.2	−3.0	−3.0	Jakes
3	0.3	0.4	0.0	0.0	Jakes
4	0.5	0.6	−2.6	−2.0	Jakes
5	0.8	0.8	−3.0	−3.0	Jakes
6	1.1	1.2	−5.0	−5.0	Jakes
7	1.3	1.4	−7.0	−7.0	Jakes
8	1.7	1.8	−5.0	−5.0	Jakes
9	2.3	2.4	−6.5	−6.0	Jakes
10	3.1	3.0	−8.6	−9.0	Jakes
11	3.2	3.2	−11.0	−11.0	Jakes
12	5.0	5.0	−10.0	−10.0	Jakes

Table 9.A.5. The Reduced Profile for Urban Area (Six Taps)

Tap number	Relative time (μs)		Average relative power (dB)		Doppler spectrum
	(1)	(2)	(1)	(2)	
1	0.0	0.0	−3.0	−3.0	Jakes
2	0.2	0.2	0.5	0.0	Jakes
3	0.5	0.6	−2.0	−2.0	Jakes
4	1.6	1.6	−6.0	−6.0	Jakes
5	2.3	2.4	−8.0	−8.0	Jakes
6	5.0	5.0	−10.0	−10.0	Jakes

Table 9.A.6. Profile for Equalization Test (Six Tap Setting)

Tap number	Relative time (μs)	Average relative power (dB)	Doppler spectrum
1	0.0	0.0	Jakes
2	3.2	0.0	Jakes
3	6.4	0.0	Jakes
4	9.6	0.0	Jakes
5	12.8	0.0	Jakes
6	16.0	0.0	Jakes

given, while for the latter two environments, two sets of models are offered, a 12-ray model and a 6-ray model. These correspond to a tapped-delay-line structural model with either 12 or 6 taps. The 6-tap model is offered as a computationally less demanding alternative, which may be used in particular in simulations that include interfering signals. Whenever possible, however, the full configuration (12 taps) should be used. For each model in the tables, two equivalent alternative tap settings (weights) and relative delays are indicated by (1) and (2) atop the appropriate columns.

It should be noted that for the standard signaling rate of 271 kb/s, the symbol duration is about $T_{\text{sym}} = 3.7\,\mu\text{s}$, but some of the differential delays in the models are as small as $0.1\,\mu\text{s}$, i.e., 74 times smaller then the symbol duration. To simulate properly such a delay would imply using a simulation sampling rate of 74 samples/symbol, which is obviously burdensome, or using an alternative approach such as outlined in the preceding section, where the model is "filtered" to produce a constant-spacing TDL.

Doppler Spectrum Types. As can be seen from Tables 9.A. 1–9.A.6, the GSM channel models specify two Doppler spectra:

1. The classical (Jakes) spectrum

$$S(f) = \frac{A}{[1 - (f/f_{\text{d}})^2]^{-1/2}}, \quad f \in [-f_{\text{d}}, f_{\text{d}}]$$

where $f_{\text{d}} = v/\lambda$, v/λ represents the vehicle speed (in m/s), and λ (in m) is the wavelength.

2. The Ricean spectrum, which is the sum of a classical Doppler spectrum and one direct line-of-sight path

$$S(f) = \frac{0.41}{2\pi f_d[1-(f/f_d)^2]^{-1/2}} + 0.91\delta(f-0.7f_d), \qquad f \in [-f_d, f_d]$$

9.A.2. Reference Models for PCS Applications

For PCS communication systems operating in the 2-GHz band the standards bodies[66] have agreed on a set of discrete models for both indoor and outdoor environments. These models are summarized in Tables 9.A.7–9.A.9. In Table 9.A.7 the ray strength for each tap is defined as

$$\rho_k = E[|a_k|^2] \tag{9A.2.1}$$

and is given in Table 9.A.8. The proposed indoor models are given in Table 9.A.9. The weights of the individual rays are not specified.

It should noted again that the differential delay of 50 ns in the indoor environment is very small compared to the symbol times proposed for PCS systems. Hence, either the fading should be treated as frequency-nonselective, or the bandlimited TDL model with symbol time spacing should be used for simulations.

Table 9.A.7. Parameters of Outdoor Models

Environment	$\tau_1(\mu s)$	$\tau_2(\mu s)$	$\tau_3(\mu s)$	Spectrum	f_d(Hz)
Pedestrian	0	1.5	14.5	Flat	12
Wireless loop	0	1.5	14.5	Gaussian	12
Vehicular	0	1.5	14.5	Jakes	180

Table 9.A.8. Relative Tap Strengths of Outdoor Models

Tap number k	Tap strength (dB) 10 log ρ_k
1	0
2	−3
3	−6

Table 9.A.9. Parameters of Indoor Models

Environment	Tap spacing (ns)	Number of taps	Tap spectrum Tap spectrum	Doppler f_d (Hz)
Residential	50	2	Gaussian	3
Office	50	4	Gaussian	3
Commercial	50	12	Flat	30

9.A.3. Reference Channel Models for UMTS-IMT-2000 Applications

The next generation of wireless technology, designed to support broadband application, is in the planning stages. The International Telecommunication Union (ITU) is promoting a standard called IMT-2000, designed to support "universal mobile communications systems" (UMTS). The standard involves every aspect of system design and test and is voluminous[67] Here we provide an extract only of the propagation models for the terrestrial component of a system. These models include both path loss models (along the lines given in Section 9.1.2) and channel impulse response models (along the lines discussed in Section 9.1.3).

For terrestrial environments, the propagation effects are divided into three types of model: mean path loss, slow variation about the mean due to shadowing and scattering, and the rapid variation of the signal due to multipath effects. The channel impulse response is modeled using a tapped delay line as discussed in Section 9.1.3.

9.A.3.1. Path Loss Models

Equations are given for *mean* path loss as a function of distance for different terrestrial environments. The slow variation is taken to be log-normally distributed with a particular standard deviation and a specified form for the decorrelation length for the vehicular test environment.

9.A.3.1.1. Path Loss Model for Indoor Office Test Environment The indoor path loss model is given by

$$L = 37 + 30 \log_{10} R + 18.3 n^{(n+2)/(n+1)-0.46} \quad \text{[dB]}$$

where R is the transmitter-receiver separation in meters and n is the number of floors in the path.

Note: L shall in no circumstances be less than the free-space loss. A log-normal shadow fading standard deviation of $\sigma = 12$ dB can be expected.

9.A.3.1.2. Path Loss Model for Outdoor-to-Indoor and Pedestrian Test Environments The following should be used for outdoor-to-indoor and pedestrian test environments:

$$L = 40 \log_{10} R + 30 \log_{10} f + 49 \quad \text{[dB]}$$

where R is the base station–mobile station separation in kilometers and f is the carrier frequency of 2000 MHz for IMT-2000/FPLMTS application.

Note: L shall in no circumstances be less than the free-space loss. This model is valid for non-line-of-sight case only and describes worst-case propagation. Log-normal shadow fading with a standard deviation of 10 dB for outdoor users and 12 dB for indoor users is assumed. The average building penetration loss is 12 dB with a standard deviation of 8 dB.

9.A.3.1.3. Path Loss Model for Vehicular Test Environments This model is applicable for the test scenarios in urban and suburban areas outside the high-rise core where structures are of nearly uniform height:

$$L = 40(1 - 4 \times 10^{-3} \Delta h_b) \log_{10} R - 18 \log_{10} \Delta h_b + 21 \log_{10} f + 80 \quad \text{[dB]}$$

where R is the base station–mobile station separation in kilometers, f is the carrier frequency of 2000 MHz for UMTS, and Δh_b is the base station antenna height, in meters, measured from the average rooftop level.

Note 1: The path loss model is valid for a range of Δh_b from 0 to 50 m.

Note 2: L shall in no circumstances be less than the free-space loss. This model is valid for the non-line-of-sight case only and describes worst-case propagation. Log-normal shadow fading with a standard deviation of 10 dB is assumed in both urban and suburban areas.

The base station antenna height is fixed at $\Delta h_b = 15$ m above the average rooftop, but each system may specify an alternate base station antenna height to optimize coverage and spectrum efficiency in their proposal. Considering a carrier of $f = 2000$ MHz and a fixed base station antenna height of $\Delta h_b = 15$ m, the above path loss formula becomes

$$L = 128.1 + 37.6 \log_{10} R \quad \text{[dB]}$$

9.A.3.1.4. Decorrelation Length of the Long-Term Fading The long-term (log-normal) fading in the logarithmic scale around the mean path loss L dB is characterized by a Gaussian distribution with zero mean and some standard deviation. Due to the slow fading process versus distance Δx, adjacent fading values are correlated. The normalized autocorrelation function $R(\Delta x)$ can be described with sufficient accuracy by an exponential function

$$R(\Delta x) = \exp\left[-\frac{|\Delta x|}{d_{\text{cor}}} \ln 2\right]$$

The decorrelation length d_{cor} depends on the environment. This concept can be applied in the vehicular test environment with $d_{\text{cor}} = 20$ m. Although the evaluation of decorrelation length may not be fully valid in the outdoor-to-indoor and pedestrian environments, this concept is still to be applied with a $d_{\text{cor}} = 5$ m.

9.A.3.2. Channel Impulse Response Model

For each terrestrial test environment, a channel impulse response model based on a tapped delay line is given. The model is characterized by the number of taps, the time delay relative to the first tap, the average power relative to the strongest tap, and the Doppler spectrum of each tap. A majority of the time, rms delay spreads are relatively small, but occasionally, there are worst-case multipath characteristics that lead to much larger delay spreads. Measurements in outdoor environments show that rms delay spread can vary by over an order of magnitude within the same environment. Although large spreads occur relatively infrequently, they can have a major impact on system performance. To evaluate accurately the relative performance of candidate systems, it is desirable to model the variability of delay spread as well as the locations where delay spread is relatively large.

As this delay spread variability cannot be captured using a single tapped-delay-line model, up to two multipath channels are defined for each test environment. Within one test environment, channel A is the low-delay-spread case that occurs frequently, and channel B is a median-delay-spread case that also occurs frequently. Each of these two channels is expected to be encountered for some percentage of the time in a given test environment. Table 9.A.10 gives the percentage of time that the particular channel may be encountered with the associated rms average delay spread for channel A and channel B for each terrestrial test

Table 9.A.10. Parameters for Channel Impulse Response Model

Test environment	rms A (ns)	p(A) (%)	rms B (ns)	p(B) (%)
Indoor office	35	50	100	45
Outdoor-to-indoor and pedestrian	45	40	750	55
Vehicular, high antenna	370	40	4000	55

Table 9.A.11. Indoor Office Test Environment Tapped-Delay-Line Parameters

Tap number	Channel A		Channel B		Doppler spectrum
	Rel. delay (ns)	Avg. power (dB)	Rel. delay (ns)	Avg. power (dB)	
1	0	0	0	0	Flat
2	50	−3.0	100	−3.6	Flat
3	110	−10.0	200	−7.2	Flat
4	170	−18.0	300	−10.8	Flat
5	290	−26.0	500	−18.0	Flat
6	310	−32.0	700	−25.2	Flat

Table 9.A.12. Outdoor-to-Indoor Test Environment Tapped-Delay-Line Parameters

Tap number	Channel A		Channel B		Doppler spectrum
	Rel. delay (ns)	Avg. power (dB)	Rel. delay (ns)	Avg. power (dB)	
1	0	0	0	0	Jakes
2	110	−9.7	200	−0.9	Jakes
3	190	−19.2	800	−4.9	Jakes
4	410	−22.8	1200	−8.0	Jakes
5	—	—	2300	−7.8	Jakes
6	—	—	3700	−23.9	Jakes

Table 9.A.13. Vehicular Test Environment Tapped-Delay-Line Parameters

Tap number	Channel A		Channel B		Doppler spectrum
	Rel. delay (ns)	Avg. power (dB)	Rel. delay (ns)	Avg. power (dB)	
1	0	0.0	0	−2.5	Jakes
2	310	−1.0	300	0.0	Jakes
3	710	−9.0	8,900	−12.8	Jakes
4	1090	−10.0	12,900	−10.0	Jakes
5	1730	−15.0	17,100	−25.2	Jakes
6	2510	−20.0	20,000	−16.0	Jakes

environment. The 5% of the time not accounted for in Table corresponds to occurrences of large delay spread.

Tables 9.A.11–9.A.13 describe the tapped-delay-line parameters for each of the terrestrial test environments. For each tap of the channels three parameters are given: the time delay relative to the first tap, the average power relative to the strongest tap, and the Doppler spectrum of each tap. A small variation ±3% in the relative time delay is allowed so that the channel sampling rate can be made to match some multiple of the link simulation rate.

References

1. W. C. Jakes (ed.), *Microwave Mobile Communications*, Wiley, New York (1974).
2. J. D. Parsons, *The Mobile Radio Propagation Channel*, Halsted Press, New York, (1992).
3. R. Steele (ed.), *Mobile Radio Communications*, Pentech Press, London (1994).
4. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York (1993).
5. B. Sklar, Rayleigh fading channels in mobile radio communications, Parts I and II, *IEEE Commun. Mag.* **35**(9), 136–155 (1997).
6. W. Y. C. Lee, *Mobile Cellular Communications*, McGraw-Hill, New York (1989).
7. T. S. Rappaport, *Wireless Communications*, Prentice-Hall, Upper Saddle River, New Jersey (1996).
8. V Erceg *et al*., Comparisons of a computer-based propagation prediction tool with experimental data collected in urban microcellular environments, *IEEE J. Select. Areas Commun.* **15**(4), 677–684 (1997).
9. S. C. Kim *et al*., Radio propagation measurement and prediction using three-dimensional ray tracing in urban environments at 908 MHz and 1.9 GHz, *IEEE Trans. Vehic. Technol.* **VT-43**(3), 931–946 (1999).
10. M. Hata Empirical formulae for propagation loss in land mobile radio services, *IEEE Trans. Vehic. Technol.* **VT-29**(3), 317–325 (1980).
11. Y. Okumura, E. Ohmori, and K. Fukuda, Field strength and its variability in VHF and UHF land mobile radio service, *Rev. Elec. Commun. Lab.* **16**(9, 10), 825–873 (1968).
12. COST-231, Urban Transmission Loss Models for Mobile Radio in the 900 MHz and 1800 MHz Bands (rev. 2), COST 231 TD(90), 119 Rev. 2, Den Haag (1991).
13. S. Y. Seidel *et al*., Path loss, scattering and multipath delay statistics in four European cities for digital cellular and microcellular radiotelephone, *IEEE Trans. Vehic. Technol.* **VT-40**(4), 721–730 (1991).
14. P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* **CS-11**(4), 360–393 1963.
15. B. Glance and L. J. Greenstein, Frequency-selective fading effects in digital mobile radio with diversity combining, *IEEE Trans. Commun.* **COM-31**(9), 625–635 (1983).
16. R. H. Clarice, A statistical theory of mobile radio reception, *Bell Syst. Tech. J.* **47**(6), 957–1000 (1968).
17. F. Amoroso, Use of DS/SS signaling to mitigate Rayleigh fading in a dense scatter environment, *IEEE Pers. Commun.* **3**(2), 52–61 (1996).
18. D. C. Cox, Delay-Doppler characteristics of multipath propagation at 910 MHz in a suburban mobile radio environment, *IEEE Trans. Ant. Prop.* **AP-20**(9), 625–635 (1972).
19. H. L. Van Trees, *Detection, Estimation and Modulation Theory*, Part 3, Wiley, New York (1968).
20. S. A. Fechtel, A novel approach to modeling and efficient simulation of frequency-selective fading radio channels, *IEEE J. Select. Areas Commun.* **11**(3), 422–431 (1993).
21. M. B. Priestly, *Spectral Analysis and Time Series*, Vol. 2, Academic Press, London (1981).
22. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York (1991).
23. B. N. Datta, *Numerical Linear Algebra and Applications*, Brooks Cole, Pacific Grove, California (1995).
24. H. Bateman, *Table of Integral Transforms*, Vol. I, McGraw-Hill, New York (1954).
25. T. S. Rappaport, Characterization of UHF multipath radio channels in factory buildings, *IEEE Trans. Ant. Prop.* **37**(8), 1058–1069 (1989).
26. T. S. Rappaport, S. Y. Seidel, and K. Takamizawa, Statistical channel impulse response models for factory and open plan building radio communication system design, *IEEE Trans. Commun.* **39**(5), 794–807 (1991).
27. R. Ganesh and K. Pahlavan, On the modeling of fading multipath indoor radio channels, in *Proceedings of Global Telecommunications Conference* pp. 1346–1350 (1989).
28. R. Ganesh and K. Pahlavan, Statistical modelling and computer simulation of Indoor radio channel, *Proc. IEE* **138**(3), 153–161 (1991).

29. J. B. Anderson, T. S. Rappaport, and S. Yoshida, Propagation measurements and models for wireless communications channels, *IEEE Commun. Mag.* 42–49 (1995).
30. A. A. M. Saleh and R. A. Valenzuela, A statistical model for indoor multipath propagation, *IEEEJ. Select. Areas Commun.* **SAC-54**(2), 128–137 (1987).
31. S. C. Kim, H. L. Bertoni, and M. Stem, Pulse propagation characteristics at 2.4 GHz inside buildings, *IEEE Trans. Vehic. Technol* **45**(3), 579–592 (1996).
32. S. F. Fortune *et al*., Wise design of indoor wireless systems: Practical computation and optimization, *IEEE Comput. Sci. Eng.* **1995**(Spring), 58–68.
33. W. D. Rummler, A new selective fading model: Application to propagation data, *Bell Syst. Tech. J.* **58**(5), 1037–1071 (1979).
34. P. Balaban, Statistical models for amplitude and delay of selective fading, *ATT Tech. J.* **64**(10), 2525–2250 (1985).
35. G. H. Millman, Atmospheric and Extraterrestrial Effects on Radio Wave Propagation, General Electric Co., Technical Information Series, No. R61EMH29 (June 1961).
36. W. L. Flock, Propagation Effects on Satellite Systems at Frequencies below 10 GHz, NASA Reference Publication 1108 (December 1983).
37. L. J. Ippolito, R. D. Kaul, and R. G. Wallace, Propagation Effects Handbook for Satellite Systems Design, NASA Reference Publication 1082(03), (June, 1983).
38. P. Lo, J. Haddon, H. O'Neill, and E. Vilar, Computation of rain induced scintillations on satellite down-links at microwave frequencies, *IEE Conference on Communications,* Publication 219.
39. H. J. Liebe, Modeling attenuation and phase of radio waves in air at frequencies below 1000 GHz, *Radio Sci.* **16**(6), 1183–1199 (1981).
40. K. S. Shanmugan, M. S. McKinley, V S. Frost, E. M. Friedman, and J. C. Holtzman, Wideband digital transmission through the atmosphere at EHF frequencies: Effects of refractive dispersion, *Proc. Globecom Conference*, Atlanta, Georgia (1984).
41. R. K. Crane, Prediction of attenuation by rain, *IEEE Trans. Commun.* **COM-28,** 1717–1733 (1980).
42. L. J. Ippolito, Jr., *Radiowave Propagation in Satellite Communications*, Van Nostrand Reinhold, New York (1986).
43. D. M. Jansky and M. C. Jeruchim, *Communication Satellites in the Geostationary Orbit*, 2nd ed., Artech House, Norwood, Massachusetts (1987).
44. G. H. Millman and M. C. Arabadjis, Tropospheric and Ionospheric Phase Perturbations and Doppler Frequency Shift Effects, General Electric Co., Technical Information Series No. R84EMH003 (August 1984).
45. S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, Wiley, New York (1965).
46. J. C. Daly, (ed.), *Fiber Optics*, CRC Press, Boca Raton, Florida (1984).
47. G. Keiser, *Optical Fiber Communications*, McGraw-Hill, New York (1983).
48. S. Karp, R. M. Gagliardi, S. E. Moran, and L. B. Stotts, *Optical Channels: Fibers. Clouds, Water, and the Atmosphere*, Plenum Press, New York (1988).
49. D. G. Duff, Computer-aided design of digital lightwave systems, *IEEE J. Select. Areas Commun.* **SAC-2**(1), 171–185 (1984).
50. P. K. Cheo, *Fiber Optics Devices and Systems*, Prentice-Hall, Englewood Cliffs, New Jersey (1985).
51. J. K. Townsend, Computer simulation of digital lightwave communication links, Ph. D. thesis, University of Kansas, Lawrence, Kansas (June 1988).
52. J. Gimlett and N. Cheung, Dispersion penalty analysis for LED/single-mode fiber transmission systems, *IEEE J. Lightwave Technol.* **LT-4**, 1381–1392 (1986).
53. A. Elrefaie, R. E. Wagner, D. A. Atlas, and D. G. Daut, Chromatic dispersion limitations in coherent optical fiber transmission systems, *IEE Electron. Lett.* **23**(14), 756–758 (1987).
54. L. N. Kanal and A. R. K. Sastry, Markov models for channels with memory and their application to error control, *Proc. IEEE*, **66**, 724–744, (1978).
55. F. Swarts and H. C. Ferreira, Markov Characterization of digital fading mobile channels, *IEEE Trans. on Vehicular Tech.*, **43**, 977–985, (1994).
56. C. Chao and Y. Yao, HMM models for the bust error characteristics of Viterbi decoding, *Proceedings of ICC'93*, 751–756, (1993).
57. L. R. Rabiner and B. H. Huang, An introduction to hidden Markov models, *IEEE ASSP Mag.* **1986** (January), 4–16.
58. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, An introduction to the theory and application of Markov processes to automatic speech recognition, *Bell Syst. Tech. J.* **62**, 1035–1074 (1983).
59. E. N. Gilbert, Capacity of a burst noise channel, *Bell Syst. Tech. J.* **39**, 1253–1265 (1960).

60. B. D. Fritchman, A binary channel characterization using partitioned Markov chains, *IEE Trans. Inform. Theory* **IT-13,** 221–227 (1967).
61. L. E. Baum. *et. al*., A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Math. Statistics*, **11,** 164–171, (1970).
62. W. Turin, *Digital Transmission Systems: Performance Analysis and Modeling*, McGraw-Hill, New York (1998).
63. W. Turin and M. M. Sondhi, Modeling of error sources in digital transmission systems, *IEEE J. Sect. Areas Commun.* **11**, 340–347 (1993).
64. S. Sivaprakasam and K. Sam Shanmugan, A Markov model for burst error channels, *IEEE Trans. Commun.* **43**, 1347–1355 (1995).
65. ETSI, *GSM Recommendation* 05.05, *Radio Transmission and Reception, Annex 3*, 13–16 (November 1988).
66. ANSI J-STD-008, *Personal Station–Base Station Compatibility Requirements for 1.8 to 2.0 GHz Code Division Multiple Access (CDMA) Personal Communications Systems* (March 1995).
67. Recommendation ITU-R M.1225, *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000, Annex 2* (1997).

This page intentionally left blank

10

Estimation of Parameters in Simulation

As mentioned in Chapter 2, a Monte Carlo simulation run can be viewed as a statistical experiment which is the software counterpart of an experiment on an actual system. The objective of the experiment is to allow us to make inferences about one or more of the properties of the signal as it passes through the system. For purposes of this book, we shall place such properties into one of two categories that we shall refer to as *parameters* and *performance measures*, respectively. By the latter, we mean measures related in some way to the fidelity of transmission. Here, too, we can subdivide this topic into two broad subtopics, one of which deals with the signal-to-noise ratio and the other with error performance in digital systems. This subject will be treated in the next chapter. In the present chapter we will deal with extracting various signal parameters that may be of interest. We are defining here "parameters" to be properties of the signal not directly phrased in term of performance measures. Such parameters may be of interest for a number of reasons. Specifically, the parameters in which we shall be interested are the average level, average power, probability distribution and density, power spectral density, delay, and phase shift.

The observations (measurements) consist of discretely spaced values of a finite-duration segment of a waveform at some point in the system. In a Monte Carlo simulation, the waveform is a sample path of a random process. Hence the measurements are inherently random. This means the inferences made can only be statistical. In Chapter 2 we alluded to alternative simulation methodologies that may not be interpretable as random experiments. Such methodologies will be particularly in evidence in the next chapter. Of course, in such cases, the standard language of statistical estimation may not apply strictly, but we may still use such terms as "estimation" in a connotative sense. More will be said about this in the next chapter.

We have emphasized the point that the main commodity consumed in simulation is run time, and, particularly in Monte Carlo simulation, this will often be the simulator's ultimate limitation. That is, the number of waveform samples to be observed will be restricted by run-time considerations (this, too, will be examined in the next chapter), whereas in the great majority of cases, a corresponding experiment on the real system in real time will have no such limitations. Our focus in this and the next chapter, therefore, will be to highlight these run-time constraints in estimating quantities of interest.

10.1. Preliminaries

In this section we set the stage for what is to come in this and the next chapter by first reviewing some basic ideas and definitions and establishing notation.

10.1.1. Random Process Model: Stationarity and Ergodicity

Several aspects of random process theory necessary for this book have been reviewed in Chapter 6. The notions of this theory that are of particular relevance in this chapter are those of stationarity and ergodicity. These ideas allow us to attribute more general significance to a single measurement than merely the measurement itself. The key to such an interpretation is the visualization of a random process as an ensemble of functions of time and embedding a simulation run within such a model. As we know from Chapter 6, the ergodic property basically means that we can equate a time average over a particular sample function (over $-\infty < t < \infty$) with an ensemble average. Consequently, this means we can equate a time average over a particular sample function with that over any other sample function in the ensemble. (Ergodicity can therefore be thought of as establishing a transitivity property among members of the ensemble.) This is the basis for attributing "global" significance to a measurement performed on a single sample function (i.e., a single simulation run). Ergodicity requires that the process be stationary. This means that the statistical properties of estimators are time-invariant.

There is an important class of processes in communications for which ergodicity and stationarity may not be unconditionally true. This is the class for which every sample function has associated with it an explicit or implicit (hidden) periodicity. Every modulated carrier, for example, exhibits an explicit periodicity in the carrier wave itself, although, in the simulation context, this periodicity will not manifest itself whenever we use the complex envelope representation. Samples of waveforms have implicit periodicity associated with them because of the clock inherent in the sampling process. The general term given to these processes is that they are *cyclostationary*. If, for such processes, sample path properties are equal to the corresponding ensemble properties, then the processes are said to be *cycloergodic*. As before, the main issue here relates to the generalizability of a measurement. For such processes, ergodicity may or may not hold, depending upon the model assumed for certain parameters of a process. Suppose, for example, that $X(t) = A\cos(\omega_c t + \theta)$, where θ is a random variable with uniform distribution over $(-\pi, \pi)$; then $X(t)$ is ergodic in the mean and autocorrelation function (or "weakly" ergodic). Similarly, $X(t) = A(t)\cos(\omega_c t + \theta)$ is weakly ergodic if $A(t)$ is ergodic. The key is the inclusion (or interpretation) of θ, which is often overlooked. The specific issues involved have been discussed in detail by Gardner.[1–3]

It is generally not simple to ascertain ergodicity, except for Gaussian processes (see Section 6.7.3). We take it as granted that any estimate computed in a single simulation run can have a meaningful general interpretation if the necessary properties, such as stationarity or ergodicity, do apply. It is important to make clear that such ensemble properties are desirable only for the purpose of placing a simulation run in a larger context. They are not necessary for the mechanization of a simulation run, nor are they required to make a simulation run useful.

10.1.2. Basic Notation and Definitions

It is useful to review the notational conventions established in Chapter 6 (Section 6.7.1) and adapt them to the present context. As we have seen, a random process can be looked upon

as a collection of time functions, and we shall denote the process $X(t)$. Any given specific function in this ensemble is labeled $x(t)$ or $\boldsymbol{x_i(t)}$ if the set of functions is denumerable. When we perform a simulation run, we are dealing with one such $x(t)$.

In general, we use $X(t)$ to represent the waveform of interest (voltage or current) at some point in the system. We are also often interested in some function of $X(t)$, say $Y(t) = g[X(t)]$; forexample, $\boldsymbol{Y(t) = X^2(t)}$ or $\boldsymbol{Y(t) = |X(t)|}$. Our usual objective in simulation is to estimate a property or parameter Q of $Y(t)$. This we do by a suitable operation on an observation of $Y(t)$.

We define an *observation* either as a finite-time "slice" of the process

$$\mathbf{Y}_T(t) = \begin{cases} Y(t), & 0 \le t \le T \\ 0, & \text{elsewhere} \end{cases} \tag{10.1.1a}$$

or as the sequence of samples

$$\mathbf{Y}_N = (Y_1, Y_2, \ldots, Y_N) \tag{10.1.1b}$$

where

$$Y_i = Y((i-1)T_s)$$

$\boldsymbol{T_s}$ is the simulation sampling interval, and $\boldsymbol{(N-1)T_s \approx T}$. The specific interval $[0, T]$ is so denoted for convenience, but by the stationarity assumption the interval could be located anywhere in time. Unless there is reason to distinguish between the continuous-time observation (10.1.1a) and the discrete-time observation (10.1.1b), and unless there is a need to state the length (T or N) of the interval, we shall denote an observation simply as $\mathbf{Y}$ and refer to it as a vector. Note that $\mathbf{Y}$ is a random vector. Once a sample function $y(t)$ has been chosen, the corresponding vector $\mathbf{y}$ is a nonrandom quantity and can be referred to as a measurement. Of course, in the simulation context we are always obliged to deal with discrete-time observations or measurements.

Any function of the random vector $\mathbf{Y}$, $G(\mathbf{Y})$, is a random variable and is called a *statistic*. If by the operation G we attempt to estimate the characteristic Q of the process, then we say that

$$\hat{Q} = G(\mathbf{Y}) \tag{10.1.2}$$

is an estimator of Q (or, sometimes, a point estimator) and a specific value $G(\mathbf{y})$ computed from the measurement $\mathbf{y}$ is called an estimate. The caret is standardly used to denote an estimator. In principle, there can be many estimators for the same characteristic Q. There are established procedures for determining "good" estimators, and we refer the reader to the many excellent texts in statistics (e.g., Refs. 4–10). We shall therefore take it for granted that the estimators we choose have a reasonable basis.

For almost all cases of interest, the estimator can be expressed as a weighted time average:

$$\hat{Q} = \frac{1}{N}\sum_{i=1}^{N} w_i Y_i \tag{10.1.3}$$

We shall also consistently denote a time average with angular brackets

$$\hat{Q} = \langle \mathbf{wY} \rangle_N \tag{10.1.4}$$

where the subscript N will be made explicit only if required for clarity. The product of vectors, as in $\mathbf{wY}$, is interpreted as the ordinary scalar, or "dot" product, namely, the summation in (10.1.3). Since $\hat{Q}$ is a random variable (because it depends on the chosen sample function) it has an associated distribution and pdf, say $f_{\hat{Q}}(q; N)$, which are properties of the ensemble. Hence, all average properties (moments) of $\hat{Q}$ are determined by $f_{\hat{Q}}$, and such averages will be indicated by the expectation operator E. Since we normally expect the estimation process to improve as N increases, an essential attribute of an estimator is that it converges to the true value as $N \to \infty$. Such an estimator is termed *consistent*. By the ergodic property, $\hat{Q}$ will tend to Q as $N \to \infty$ for almost every sample function.

Of course we are always limited to a finite, and perhaps small, number of samples, and it is under these conditions that we wish to determine the quality of an estimator, where quality means closeness to the true value in some sense. We consider this question next.

10.1.3. Quality of an Estimator: Bias, Variance, Confidence Interval, and Time–Reliability Product

The quality of an estimator, as just stated, can be judged by how close it is to the true value of the parameter. Since $\hat{Q}$ is a random variable, "closeness" must be measured in a probabilistic sense. In this book, three measures of quality of an estimator are of interest: the bias, the variance, and the confidence interval. The latter two are generally used as alternatives to describe the "spread" of an estimator. The confidence interval is more descriptive, but generally difficult to obtain, especially for small samples. One then has to be content with a cruder, but still useful measure, namely, the variance. A measure that we shall find useful, the time–reliability product, combines the variance with the run-time measure.

One general point that is worthwhile remembering is that the degree to which we are able to make progress in formulating expressions and arriving at useful results is very much dependent upon what we know or are willing to assume about the properties of the processes under examination.

10.1.3.1. Bias of an Estimator

An estimator is said to be unbiased if

$$\begin{aligned} E(\hat{Q}) &= \int_{-\infty}^{\infty} q f_{\hat{Q}}(q; N)\, dq \\ &= \int_{-\infty}^{\infty} G(\mathbf{y}) f_Y(\mathbf{y})\, d\mathbf{y} = \int_{-\infty}^{\infty} G[g(\mathbf{x})] f_X(\mathbf{x})\, d\mathbf{x} \end{aligned} \tag{10.1.5}$$

is equal to Q. The equality between the integrals is a consequence of the fundamental theorem of expectation.[10] This allows us to compute averages without explictly knowing the distribution $f_{\hat{Q}}$, which is generally hard to find, and rely only on the underlying distribution, which can be assumed known. If $E(\hat{Q}) \neq Q$, the estimator is said to be biased. The value of an unbiased estimator is equal to the true value, "on the average," but the value of any specific

estimate obtained from a simulation run may not be equal to the true value. Unbiasedness is an important property, particularly for a BER estimator. Heuristically, that property guarantees that estimates will generally be clustered around the true value.

10.1.3.2. Variance of an Estimator

The variance of an estimator,

$$\sigma^2(\hat{Q}) = E(\hat{Q}^2) - E^2(\hat{Q}) \tag{10.1.6a}$$

$$= \int_{-\infty}^{\infty} q^2 f_{\hat{Q}}(q; N)\, dq - E^2(\hat{Q}) \tag{10.1.6b}$$

is a measure of the dispersion about its expected value. The smaller the variance, the better the estimator, for a given N. In fact, if $\sigma^2(\hat{Q}) \to 0$, and $\hat{Q}$ is unbiased, then every observation is arbitrarily close to the true value of the parameter. Generally, $\sigma^2(\hat{Q}) \to 0$ only as $N \to \infty$, and we are then faced with a tradeoff between estimator variance and sample size N, the latter being directly proportional to run time in the simulation context.

10.1.3.3. Confidence Interval

The confidence interval is the most descriptive measure of the quality of an estimator because it quantifies the measure of spread with an associated probability. Let $h_1(\hat{Q}), h_2(\hat{Q})$ be two functions of the estimator such that, with high probability, the interval (h_1, h_2) brackets the true value Q. The difference $h_1 - h_2$ is the width of the confidence interval. The probability associated with the condition $h_2 \le Q \le h_1$ is called the confidence level and is usually denoted $1 - \alpha$. Thus, a confidence interval and confidence level are defined through

$$P[h_2(\hat{Q}) \le Q \le h_1(\hat{Q})] = 1 - \alpha \tag{10.1.7}$$

Typical values for α are 0.05 and 0.01, corresponding to "95% confidence level" and "99% confidence level," respectively. The left side of (10.1.7) is often labeled a "fiduciary" probability because Q is not a random variable, but an unknown parameter. The proper interpretation of (10.1.7) is that, with probability $1 - \alpha$, the random interval defined by (h_1, h_2) will bracket the true value.

To illustrate the procedure for obtaining a confidence interval, assume that the pdf of the estimator $f_{\hat{Q}}(q)$ is known. We assume that $f_{\hat{Q}}$ is unimodal and that the mode is at or near Q, as illustrated in Figure 10.1. Let $\sigma_{\hat{Q}}$ be the SD of $\hat{Q}$ and let d_1 and d_2 be positive constants such that

$$P[\hat{Q} - Q \le -d_1\sigma_{\hat{Q}}] = \alpha_1 \tag{10.1.8a}$$

$$P[\hat{Q} - Q \ge d_2\sigma_{\hat{Q}}] = \alpha_2 \tag{10.1.8b}$$

as shown in the figure. Combining (10.1.8a) and (10.1.8b), we get

$$P[\hat{Q} - d_1\sigma_{\hat{Q}} \le Q \le \hat{Q} + d_2\sigma_{\hat{Q}}] = 1 - \alpha \tag{10.1.9}$$

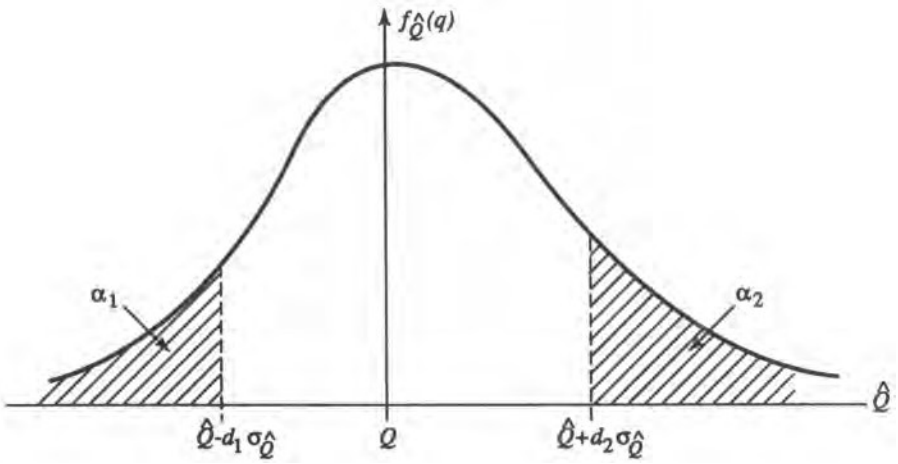


Figure 10.1. Definition of confidence interval and confidence level.

which is in the form (10.1.7), where $\alpha = \alpha_1 + \alpha_2$. The confidence interval $[\hat{q} - d_1\sigma_{\hat{Q}}, \hat{q} + d_2\sigma_{\hat{Q}}]$ is constructed around the observed value $\hat{q}$ and has length $(d_1 + d_2)\sigma_{\hat{Q}}$. For given $1 - \alpha$ it is desirable to find d_1, d_2 so as to minimize this length, although it is usually adequate to take $d_1 = d_2$. ($f_{\hat{Q}}$ may not be symmetric, but if it is, the interval length is minimized if $d_1 = d_2$.) If $\hat{Q}$ is the sample mean of N variates, say

$$\hat{Q} = \frac{1}{N}\sum_{k=1}^{N} Z_k \tag{10.1.10}$$

where the Z_k are i.i.d. with SD σ, then $\sigma_{\hat{Q}} = \sigma/\sqrt{N}$.

The construction of (10.1.9) requires knowledge of $f_{\hat{Q}}$, which is difficult to obtain for many estimates. This is why one often settles for the variance (usually much easier to find) as an estimator of spread. However, there are some important cases when the estimator distribution can be obtained exactly or approximately. One case is when the estimator is a linear function of samples of a Gaussian process because the estimator then has a normal distribution. For example, if in (10.1.10), $Z_k \in N(\mu, \sigma^2)$, then $\hat{Q} \in N(\mu, \sigma^2/N)$ and a confidence interval for the mean can be established if σ is known. In the same situation, we cannot construct a confidence interval if σ is not known. In that case the well-known alternative is to use Student's t-distribution. The variable

$$t = \frac{\hat{Q} - \mu}{S}(N - 1)^{1/2} \tag{10.1.11}$$

where $\hat{Q}$ is given by (10.1.10) and

$$S^2 = \frac{1}{N}\sum_{k=1}^{N}(Z_k - \hat{Q})^2 \tag{10.1.12}$$

is independent of σ, and its distribution is tabulated in many texts (see, e.g., Refs. 5, 10, and 11). We can thus find a confidence interval for μ by setting up the expression

$$P[t_1 \le t \le t_2) = P[\hat{Q} - t_2 S/(N - 1)^{1/2} \le \mu \le \hat{Q} - t_1 S/(N - 1)^{1/2}]$$

In probably most cases of interest in simulation it is not necessary to use the t-distribution because it converges rapidly to the normal $N(0, 1)$ for N greater than about 30.

In the situation when $\hat{Q}$ is the sum of a large number of variables, for example, (10.1.10) with N large, under the conditions of the central limit theorem (see Chapter 6) $f_{\hat{Q}}$ tends to the normal distribution and this distribution can be used as an approximation to obtain a confidence interval.

10.1.3.4. Time–Reliability Product

A measure of the goodness of an estimator that is also useful in the simulation context is the time–reliability product,

$$\zeta \triangleq N\sigma^2(\hat{Q})$$

which is the product of the number of observations and the variance of the estimator. This one figure embodies the tradeoff between the "reliability" of the estimator and the "time" of observation N, which is proportional to run time. It is useful for comparing one estimator against another and will be made use of in Chapter 11.

10.1.3.5. Normalized Measures

Measures of goodness such as bias and variance are most useful when referenced to the quantities being measured. For example, a bias of magnitude 10^{-6} seems very good on its face, but renders useless the estimation of a quantity (such as an error probability) suspected to be on the order of 10^{-7}. Therefore, measures of estimator goodness normalized to the true value of a parameter (or possibly its estimated value) are generally more meaningful. For example, we might define a normalized bias as $b(\hat{Q}) = (Q - \hat{Q})/Q$ or simply $\hat{Q}/Q$. Ideally, the first of these is zero, while the second is unity. Similarly, we can define a relative precision or normalized standard deviation as $\sigma(\hat{Q})/Q$. These relative measures are especially useful when the quantities to be estimated are small, as in error probability estimation, where the absolute measures may give us a false sense of comfort, as in Figure 10.1.

10.2. Estimating the Average Level of a Waveform

We may sometimes wish to estimate the average level of a waveform within or prior to a simulation run. The main utility of such a measurement will typically be for calibration purposes. For example, if two system elements are intended to be ac-coupled, the output of the coupling network should have a long-term zero average. An observation at that node during simulation will confirm this, or possibly uncover a "bug."

10.2.1. Form of the Estimator

As before, $X(t)$ is the waveform (process) of interest. An illustrative waveform is shown in Figure 10.2, along with samples $X(kT_s) \equiv X_k$ separated by the simulation sampling interval T_s. The natural estimator of the average level is the sample mean, namely,

$$\langle \mathbf{X} \rangle = \frac{1}{N} \sum_{k=1}^{N} X_k \tag{10.2.1}$$

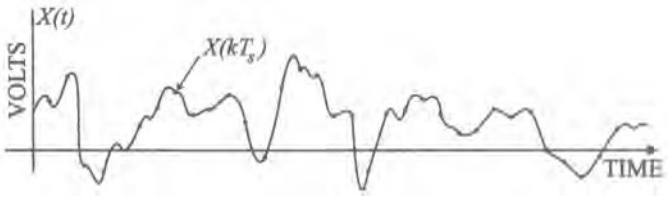


Figure 10.2. A segment of a sampled waveform $\{X(kT_s)\}$: its average is $\overline{X} = (1/N)\sum_{k=1}^{N} X(kT_s)$.

10.2.2. Expected (Mean) Value of the Estimator

As we discussed, an estimator is a random variable. Its expected (or mean) value is thus of interest and is given by

$$E(\langle \mathbf{X} \rangle) = E\left(\frac{1}{N}\sum_{k=1}^{N} X_k\right) = \frac{1}{N}\sum_{k=1}^{N} E(X_k) = \frac{1}{N}\sum_{k=1}^{N} E(X) = E(X) \tag{10.2.2}$$

where we have used successively the distributive property of expectation and stationarity. So, as is well known, the average value of the sample is the same as that of the population, irrespective of N.

Although we cannot do a continuous-time measurement in the simulation context, the objective of the discrete sum may be to estimate the average of a continuous-time process, which ideally would be estimated by

$$\langle \mathbf{X}_T \rangle = \frac{1}{T}\int_0^T X(t)\, dt$$

The average over all possible sample functions is simply

$$E(\langle \mathbf{X}_T \rangle) = \frac{1}{T}\int_0^T E(X(t))\, dt = E(X)$$

Because of stationarity, $E(X(t))$ is a constant, so that the expectation is independent of t. Hence, the (ensemble) average of the measurement in this particular instance is the same whether a continuous-time average were possible or whether a discrete average is taken.

10.2.3. Variance of the Estimator

The variance of the estimator is the ensemble average of the square of the difference between the value of the measurement and its expected value. Symbolically we have

$$\mathrm{Var}(\langle \mathbf{X} \rangle) = E(\langle \mathbf{X} \rangle - E(\langle \mathbf{X} \rangle))^2 \tag{10.2.3a}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C_{XX}(i,j) \tag{10.2.3b}$$

where C_{XX} is the autocovariance of the process,

$$C_{XX}(i,j) = E([X_i - E(X_i)][X_j - E(X_j)]) \tag{10.2.4}$$

By virtue of stationarity, $E(X_i) = E(X_j) = E(X)$ and

$$C_{XX}(i,j) = C_{XX}(i-j) \tag{10.2.5}$$

Thus, substituting (10.2.5) into (10.2.3b) and relabeling $i - j = k$, we have

$$\mathbf{Var}(\langle \mathbf{X} \rangle) = \frac{\sigma_X^2}{N} + \frac{2}{N^2} \sum_{k=1}^{N-1} (N-k) C_{XX}(kT_s) \tag{10.2.6}$$

where $C_{XX}(0) = \sigma_X^2$ is the variance of the process, and for clarity we have made time explicit by including the sampling interval T_s.

Not much more can be said without making specific assumptions about C_{XX}. Let us assume that $C_{XX}(\tau)$ essentially vanishes for $\tau > \tau_0$, with $\tau_0 = |k_0|T_s$. For physical processes, this will always be at least approximately true. Suppose T_s is sufficiently larger than τ_0 so that we can safely assume $C_{XX}(kT_s) = 0$ for all $k \geq 1$. Then the sum in (10.2.6) is zero, so that

$$\mathbf{Var}(\langle \mathbf{X} \rangle) = \sigma_X^2 / N \tag{10.2.7}$$

Hence $\mathbf{Var}(\langle \mathbf{X} \rangle) \to 0$ as $N \to \infty$, which implies, from (10.2.3a), that

$$\lim_{N \to \infty} E(\langle \mathbf{X} \rangle - E(\langle \mathbf{X} \rangle))^2 = 0$$

Heuristically, we can interpret this as saying that the average for almost every sample path tends to the ensemble average as the number of observations becomes large.

Now suppose $T_s \to 0$. Two situations present themselves here. Since $NT_s = T$, we can assume that the observation interval T remains constant, in which case $N \to \infty$, or we can assume that N remains finite, in which case $T \to 0$. Actually, neither of these cases is very meaningful in the simulation context because, in the first instance, the number of samples is infinite, and in the second, all the samples are essentially at the same instant, namely, completely redundant. So this latter case is not of much interest, but there is something to be learned from the first case ($NT_s = T$, with T fixed).

In particular, it can be shown (see Chapter 6 and Ref. 1) that, as $T_s \to 0$, Equation (10.5.3) takes the form

$$\lim_{T_s \to 0} \mathbf{Var}(\langle \mathbf{X} \rangle) = \frac{2}{T} \int_0^T \left(1 - \frac{\tau}{T}\right) C_{XX}(\tau)\, d\tau \tag{10.2.8}$$

Since $C_{XX}(\tau) \leq C_{XX}(0) = \sigma_X^2$, we can bound (10.2.8) as

$$\mathbf{Var}(\langle \mathbf{X} \rangle) \leq \frac{2\tau_0}{T} \sigma_X^2, \quad \tau_0 < T \tag{10.2.9a}$$

$$\leq 2\sigma_X^2, \quad \tau_0 \geq T \tag{10.2.9b}$$

Condition (10.2.9b) confirms the fact that the condition $\tau_0 \geq T$ is to be avoided, as might be expected. Condition (10.2.9a) may be rephrased as

$$\mathbf{Var}(\langle \mathbf{X} \rangle) \leq \frac{\sigma_X^2}{N_e} \tag{10.2.9c}$$

where

$$N_e \triangleq T/2\tau_0$$

can be viewed as an "effective" number of independent samples, by analogy with (10.2.6). Therefore, even though there is an infinite number of samples, they do not contribute independently to the reliability of the estimate. In fact, since we are limited to a finite number of simulation observations, it seems from the foregoing that the best strategy is to space the samples far enough apart that there is little or no correlation between adjacent samples. We note that the reliability of the estimates does not explicitly depend on T_s, which would appear to give us the freedom to choose T_s large enough so that samples are independent. But, in fact, the sampling theorem (Chapter 3) constrains us from using too large a value for T_s. Thus, it is not generally possible to choose T_s so that it both satisfies the sampling theorem and produces uncorrelated samples.

However, there is one important special case where both requirements are satisfied. This case is that for which the waveform $X(t)$ has a power spectral density (PSD) given in Figure 10.3. The autocorrelation function corresponding to this PSD is given by

$$R_{XX}(\tau) = 2\,AB[\sin(2\pi B\tau)/2\pi B\tau] \tag{10.2.10}$$

The sampling theorem for random processes (Section 6.10) stipulates that samples separated by $1/2B$ are sufficient to represent the process. But, if $\tau = 1/2B$, $R_{XX}(1/2B) = 0$; hence the samples are uncorrelated.

What is a suitable value of $\mathbf{Var}(\langle \mathbf{X} \rangle)$? Generally, we would like $\sigma_X/[\sqrt{N}E(\mathbf{X})]$ to be sufficiently small, which means choosing N sufficiently large. The main problem here is when $E(X) = 0$. In that case, we have to accept an absolute value of $\mathbf{Var}(\langle \mathbf{X} \rangle)$ as being a good

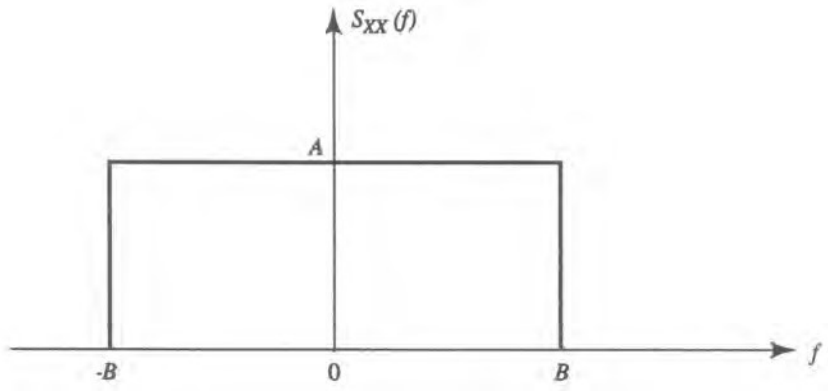


Figure 10.3. Power spectral density of a bandlimited lowpass process.

criterion, or perhaps measure $\sigma_X/\sqrt{N}$ with respect to a nonnegative parameter such as σ_X itself.

10.2.4. Mixture (Signal Plus Noise) Processes

In simulation we are almost always interested in the mixture of two or more processes, typically signal plus noise, signal plus interference, etc. Thus, if

$$X(t) = \sum_{i=1}^{M} V_i(t) \tag{10.2.11}$$

and if the reasonable assumption is made that the $V_i(t)$ are mutually independent, then (10.2.6) becomes

$$\mathrm{Var}(\langle \mathbf{X} \rangle) = \sum_{i=1}^{M} \frac{\sigma_{V_i}^2}{N} + \frac{2}{N^2} \sum_{i=1}^{M} \sum_{k=1}^{N-1} (N-k) C_{VV}^{(i)}(kT_s) \tag{10.2.12}$$

where $C_{VV}^{(i)}$ is the autocovariance of $V_i(t)$. Thus, rather similar results as before can be expected, but now we have to pay attention to the properties of the component processes. Of course, we can still think in terms of a single covariance function which is the sum of the C_{VV}, and in that sense the formulation (10.2.6) is unchanged.

10.2.5. Confidence Interval Conditioned on the Signal

Let us assume that

$$\mathbf{X} = \mathbf{S} + \mathbf{N} \tag{10.2.13}$$

where $N(t)$ is a zero-mean Gaussian random process and $S(t)$ the signal process. In this case we can obtain a confidence interval for $\langle \mathbf{X} \rangle$ by conditioning upon the signal. Conditioning upon the signal means, in the conceptualization of the ensemble, that the same $s(t)$ is always generated, while the natural range of realizations of $N(t)$ is permitted.

Let $\mathbf{s} = (s_1, s_2, \ldots, s_M)$ represent the vector of measured signal values. Then the conditional sample mean is

$$Z \triangleq \langle \mathbf{X} | \mathbf{s} \rangle = \frac{1}{M} \sum_{k=1}^{M} s_k + \frac{1}{M} \sum_{k=1}^{M} N_k \tag{10.2.14a}$$

$$= \langle \mathbf{s} \rangle + \frac{1}{M} \sum_{k=1}^{M} N_k \tag{10.2.14b}$$

Since the N_k are assumed normal, the summation in (10.2.14b) is also normal with zero mean, and variance given by

$$\sigma_Z^2 = \frac{\sigma_N^2}{M} + \frac{2}{M^2} \sum_{k=1}^{M-1} (M-k) C_{NN}(kT_s) \tag{10.2.15}$$

Thus, Z is normal with pdf

$$f_Z(z) = \frac{1}{(2\pi)^{1/2}\sigma_Z} \exp\left[\frac{(z - \langle \mathbf{s} \rangle)^2}{2\sigma_Z^2}\right] \tag{10.2.16}$$

Hence, a confidence interval for $\langle \mathbf{s} \rangle$ can be constructed from

$$P[Z - d_1\sigma_Z \leq \langle \mathbf{s} \rangle \leq Z + d_1\sigma_Z] = 1 - \alpha \tag{10.2.17}$$

by using for Z the observed value z, and d_1 is given by the implicit relation

$$\frac{1}{(2\pi)^{1/2}} \int_{-d_1}^{d_1} e^{-u^2/2}\, du = 1 - \alpha \tag{10.2.18}$$

Note that the interval has been chosen to be symmetrical because in this case it produces the shortest interval. The *length* of the interval is $2d_1\sigma_Z$. It is desired that this length tend to zero as N becomes large, and generally it is realistic to assume that the requisite properties hold.

■ *Remark.* In practice, one often uses *test signals* to probe a system and it is useful to think of such a signal as the one that the experiment is conditioned upon. For a digital system, the test signal is typically a PN sequence, while for an analog system it is often a sine wave, or several sine waves, or perhaps a square wave. All of these signals are, of course, periodic, so that if the observation interval is a multiple of a period and if T_s is chosen suitably, the resulting $\langle \mathbf{s} \rangle$ will be (almost) exactly the true average for that signal. Whether the test signal is representative of the ensemble, or even a member of the ensemble, is another question. Indeed, actual systems are often tested and specified in terms of one or a few test signal(s), and as far as the system engineer is concerned, the test-signal "ensemble" may be the one of real interest; it is presumed that the connection between a certain level of performance of the test signal and that of the ensemble has been established. ■

■ *Summary of the Simulation Procedure.* The procedure for obtaining the average of a waveform is quite straightforward (see Figure 10.2).

1. Collect samples $\{X(kT_s)\} = \{X_k\}, k = 1, 2, \ldots, N$
2. Sum the samples and divide by N, as in (10.2.1). ■

10.3. Estimating the Average Power (Mean-Square Value) of a Waveform

The average power of a waveform is an important attribute, whether in analog or digital transmission. The waveform of interest will normally consist of signal and noise. From the link performance standpoint the signal-to-noise ratio (SNR) is the quantity of interest. We shall look into estimating SNR in the next section. Here, we focus on the total average power. The average power per se is generally not a performance barometer. Instead, like the average level, the average power is a parameter of interest at a more detailed level of system design. For example, it is typical to specify the input drive level to an amplifier, hence one must be able to verify that the requisite operating point is indeed achieved.

A problem that is distinct from, but closely related to, the estimation of the average power is the estimation of the energy over a specified interval T_0. This problem is of interest

in digital communications. The average power P_{av} is the average of the finite-time average power $P_k(T_0)$, which is the energy over the kth interval divided by the duration T_0. If the energy in the kth interval $\xi_k(T_0)$ is intended to be independent of k, as in a constant-envelope signal, there is an obvious connection between average power and energy. In actuality, the energy will fluctuate from one interval to the next because of envelope variation due to filtering and because of the presence of noise. (Of course, the energy may vary intentionally in different intervals, as in PAM or QAM constellations not on a circle.) Nevertheless, the average energy may still be of interest in order, say, to calibrate the E_b/N_0 axis of a BER curve.

In some systems the energy $\xi_k(T_0)$ is *inherently* random. For example, in a direct detection optical receiver, the output of an avalanche photodetector in a given interval is a random variable even with fixed input intensity. In such a system, while the average output may still be useful information, what is of real interest is the distribution of the output given a type of symbol. In such cases, the problem is one of estimating a distribution, not a parameter. This is addressed in Section 10.4.

10.3.1. Form of the Estimator for Average Power

As before, the waveform $X(t)$ is sampled at intervals T_s to yield $\mathbf{X}$. We choose the average power estimator to be of the following form:

$$P_N(\mathbf{X}) = \frac{1}{N}\sum_{k=1}^{N} X_k^2 \tag{10.3.1}$$

We can use the results of the preceding section if we define the auxiliary variable $Y(t) = X^2(t)$, in which case the problem becomes formally one of estimating the mean of $Y(t)$, namely,

$$\langle \mathbf{Y} \rangle = \frac{1}{N}\sum_{k=1}^{N} Y_k \tag{10.3.2}$$

Refer to Figure 10.4, which shows an illustrative segment of a waveform and its samples.

10.3.2. Expected Value of the Estimator

It is clear that

$$E\langle \mathbf{Y} \rangle = E(Y) \tag{10.3.3a}$$

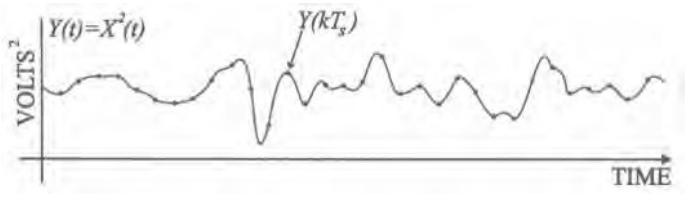


Figure 10.4. A segment of the square of a waveform $\{Y(kT_s)\}$, where $Y(t) = X^2(t)$: its average is $\overline{Y} = (1/N)\sum_{k=1}^{N} Y(kT_s)$.

for a stationary process. Hence

$$E(P_N(\mathbf{X})) = E(X^2) \tag{10.3.3b}$$

so that the expected value of the finite-sample estimator of average power is equal to the mean-square value of the process, independent of N.

In the continuous-time case,

$$E(\langle \mathbf{Y} \rangle) = \frac{1}{T}\int_0^T E(Y(t))\,dt$$

$$= E(Y) = E(X^2)$$

so that (if the underlying process is continuous) the expected value is properly given by the sampled version of the process. Hence the temporal average is an unbiased estimator of the related ensemble parameter.

10.3.3. Variance of the Estimator

We can write, directly from (10.2.6),

$$\mathrm{Var}(\langle \mathbf{Y} \rangle) = \frac{\sigma_Y^2}{N} + \frac{2}{N^2}\sum_{k=1}^{N-1}(N-k)C_{YY}(kT_s) \tag{10.3.4}$$

so that all the properties of $\langle \mathbf{X} \rangle$ also apply to $\langle \mathbf{Y} \rangle$ if C_{YY} has the same characteristics as C_{XX}. Similarly, if we let $T_s \to 0, NT_s \to T$, then from (10.2.8)

$$\mathrm{Var}(\langle \mathbf{Y} \rangle) = \frac{2}{T}\int_0^T \left(1 - \frac{\tau}{T}\right) C_{YY}(\tau)\,d\tau \tag{10.3.5}$$

The estimator $P_N(X)$ will converge in mean square to P_{av} for almost every sample path $x(t)$ if $\mathrm{Var}(\langle \mathbf{Y} \rangle) \to 0$ as $N \to \infty$. This will happen in the case of (10.3.5) if[1]

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T C_{YY}(\tau)\,d\tau = 0 \tag{10.3.6}$$

or, in the case of (10.3.4), if the discrete equivalent of (10.3.6) holds, namely,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} C_{YY}(kT_s) = 0 \tag{10.3.7}$$

Of course, the constraints on C_{YY} also imply constraints on the related parameters of $X(t)$. Specifically, since $Y = X^2$, we have

$$C_{YY}(\tau) = E(X^2(t)X^2(t+\tau)) - R_{XX}^2(0) \tag{10.3.8}$$

where $R_{XX}(\tau)$ is the autocorrelation function of $X(t)$. Note that $R_{XX}(0) = E(X^2)$, the very quantity we are trying to estimate. Nevertheless, one could, in principle, evaluate C_{YY} in terms of the known properties of $X(t)$, and hence compute variance as a function of N. In practice, (10.3.8) is not simple to evaluate, and requires knowledge of the four-dimensional joint density of the process $X(t)$. If $X(t)$ is a Gaussian random process, one can show that

$$C_{YY}(\tau) = 2C_{XX}^2(\tau) + 4E^2(X)C_{XX}(\tau) \tag{10.3.9}$$

Notice that $X(t)$ and $X(t+\tau)$ become independent as $\tau \to \infty$ for most processes of interest, hence both (10.3.8) and (10.3.9) tend to zero as $\tau \to \infty$, which for most processes of interest is a necessary condition for (10.3.6) to hold. [Even for a singular process like $X(t) = a, -\infty < t < \infty$, for which $X(t)$ and $X(t+\tau)$ do not become independent as $\tau \to \infty$, $C_{YY}(\infty) \to 0$. In fact, for this example, $C_{YY}(\tau) = 0$.] Of course, if $C_{XX}(\tau)$ is identically zero for some $\tau > \tau_0$, then (10.3.6) is guaranteed.

■ *Example 10.3.1.* Let us suppose that NT_s and C_{YY} are such that the first term in (10.3.4) dominates. Then, the relative error is given by

$$\varepsilon = \frac{[\mathrm{Var}((Y))]^{1/2}}{E(Y)} \approx \frac{\sigma_Y}{\sqrt{N}E(Y)}$$

$$= \left\{\frac{2}{N}\left[1 - \frac{E^4(X)}{R_{XX}^2(0)}\right]\right\}^{1/2} \tag{10.3.10}$$

To get a numerical sense of (10.3.10), suppose $X(t) = A + N(t)$ with $N(t)$ a zero-mean process with variance σ^2. Then, $E(X) = A$, and $E(X^2) = R_{XX}(0) = A^2 + \sigma^2$. Substituting in (10.3.10) yields

$$\varepsilon = \left\{\frac{2}{N}\left[1 - \frac{1}{1 + 2(\sigma/A)^2 + (\sigma/A)^4}\right]\right\}^{1/2}$$

which is plotted in Figure 10.5. If A is only moderately larger than the fluctuating component, e.g., $A/\sigma = 10$, it can be seen that comparatively few observations will yield a reliable estimate. In the reverse situation, e.g., $A/\sigma = 1$, a fairly large number of observations may have to be made if the fractional error is required to be small, less than 1%, say. ■

■ *Remark on Confidence Interval.* We previously mentioned the difficulty of obtaining confidence intervals. This is even more so in the present instance. Even if we condition P_N on the signal, we can show that the conditional variable is a sum of independent but not identically distributed noncentral gamma random variables. The labor involved with evaluating this distribution is clearly disproportionate with the (relatively low) accuracy that we generally need for an average power estimate. Hence, the variance is quite sufficient for our purposes. ■

■ *Summary of the Simulation Procedure.* The simulation procedure for estimating the average power should be quite clear from the preceding, namely:

1. Collect samples $\{X_k\}, k = 1, 2, \ldots, N$
2. Square each sample $\{Y_k = X_k^2\}, k = 1, 2, \ldots, N$ (see Figure 10.4).
3. Sum the samples Y_k and divide by N, as in (10.3.1). ■

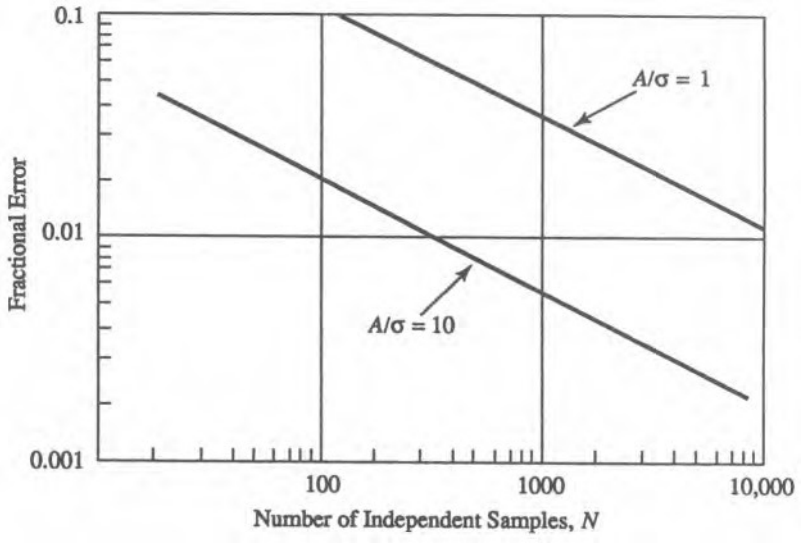


Figure 10.5. Relationship between the fractional error of the average power estimate and the number of observations for Example 10.3.1.

10.4. Estimating the Probability Density or Distribution Function of the Amplitude of a Waveform

In communications applications the probability density function (pdf) or the cumulative distribution function (CDF) of the values of a waveform is usually not, in itself, a measure of performance. Nevertheless, these functions tell us about the statistical nature of the process under observation and thus help in understanding the system. In digital communications, partial knowledge of these functions is necessary for the determination of error probability. Also of interest in characterizing some communications links is the temporal distribution of errors, for example, the probability of a given number of errors in a block of specified size. Empirically obtained distributions can impart quantitative or qualitative information. One can obtain insight by observing that a distribution "appears" to be of such and such a form. To obtain this information, only a procedure is needed. However, to make quantitative inferences we must be able to make quantitative statements about the accuracy of the estimator resulting from that procedure.

In this section we give only a short treatment of this subject, focusing on the histogram method, which is particularly simple to implement in software. More extensive treatments can be found, for example, in Refs. 5 and 11–14.

10.4.1. The Empirical Distribution

The *empirical distribution* $\hat{F}_X(x)$ is the classical estimator for an underlying distribution $F_X(x)$. (To simplify the notation in this section we shall drop the subscript X on F.) Let

$\mathbf{X} = (X_1, X_2, \ldots, X_N)$ be an observation of the random vector $\mathbf{X}$ written in *increasing order* $(X_k \leq X_{k+1})$. Then, we define

$$\hat{F}(x) = \begin{cases} 0, & x < X_1 \\ k/N, & X_k \leq x < X_{k+1}, \quad k = 1, 2, \ldots, N-1 \\ 1, & x \geq X_N \end{cases}$$

$\hat{F}(x)$ is a binomially distributed random variable with the probability of "success" $p = F(x)$. The distribution of $\hat{F}(x)$ is therefore

$$P[\hat{F}(x) = m/N] = \frac{N!}{m!(N-m)!}[F(x)]^m[1 - F(x)]^{N-m} \tag{10.4.1}$$

This emphasizes that $\hat{F}(x)$ is a random variable, and it is of interest to quantify how close $\hat{F}(x)$ is to $F(x)$. We do not discuss this in detail here, but simply mention the following important theorems.[5]

We expect that as $N \to \infty$, $\hat{F}(x)$ approaches $F(x)$ in some sense. In fact, Glivenko's theorem[5] (also known as the Glivenko–Cantelli theorem) states that

$$P\left(\lim_{N\to\infty} D_N = 0\right) = 1$$

where

$$D_N = \sup_{-\infty<x<\infty} |\hat{F}_N(x) - F(x)|$$

and $\hat{F}_N(x)$ is exactly the same as $\hat{F}(x)$ with an explicit subscript N indicating the number of samples. The theorems of Kolmogorov and Smirnov[5] give explicit formulas for evaluating the limit of $\sqrt{D_N}$ as $N \to \infty$, and a very interesting formulation for the same thing has been given in terms of a "Brownian bridge".[14] The theorem of Renyi[5] also gives explicit formulas for evaluating the more useful statistic $[\hat{F}_N(x) - F(x)]/F(x)$.

The empirical distribution when plotted has a familiar staircase-type appearance, and is easy to produce within a simulation (see Figure 10.6). Note that in general the location of the risers of the staircase will not be equally spaced. However, we can enforce equispaced transitions by grouping the data in "cells" defined accordingly. (Of course this will come about naturally if we define the cells sufficiently finely.) This grouping results in a more immediately interpretable visual display and is typically implemented as discussed below.

10.4.2. The Empirical Probability Density Function—Histogram

One is often interested in obtaining the density function directly, rather than the CDF. The "natural" empirical pdf may be thought to be the derivative of $\hat{F}(x)$. This would result in a set of delta functions located at the risers of $\hat{F}(x)$. However, an essentially equivalent but visually more pleasing alternative is the *histogram*. The histogram is basically a bar chart, as illustrated in Figure 10.7, which shows the actual or relative frequency of events observed within equispaced intervals. The histogram is an estimator of the first-order pdf. Similar ideas

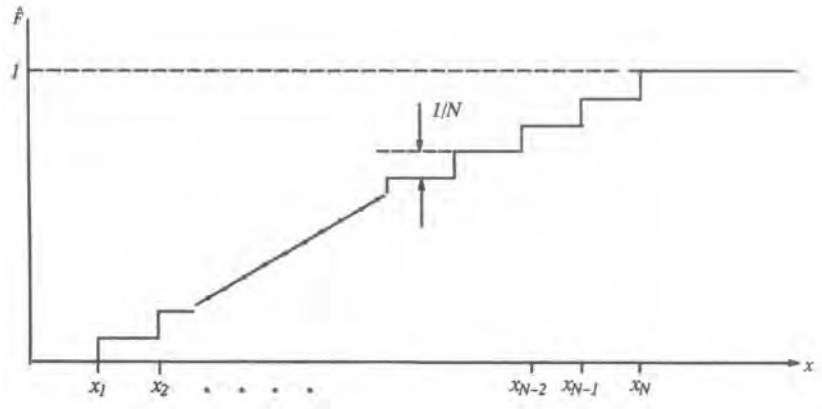


Figure 10.6. Illustration of the empirical distribution.

are used to obtain estimates of the multidimensional pdf. Sometimes we may be interested in determining if it is reasonable to suppose that the underlying population has a pdf of a given form, like a normal distribution. This is a test-of-hypothesis problem which has been briefly addressed in Chapter 6 as a "goodness-of-fit" test.

10.4.2.1. Form of the Estimator

Let the ith interval be centered on x_{c_i}, and let W be the width of the interval. Hence the ith interval will be denoted

$$\Delta_i = \left\{ x : x_{c_i} - \frac{W}{2} < x \leq x_{c_i} + \frac{W}{2} \right\}$$

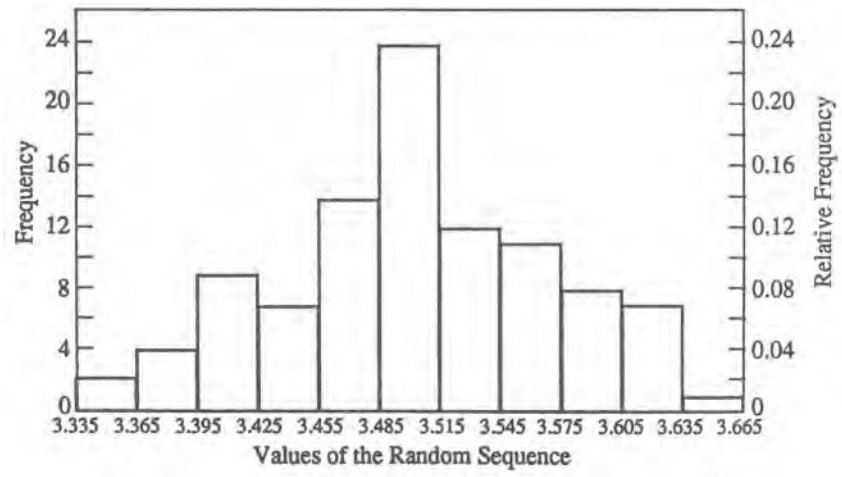


Figure 10.7. Construction of a histogram (from Ref. 11, © John Wiley & Sons, 1998, reproduced by permission).

so that the (normalized) histogram $\hat{f}(x)$, namely, the estimator of $f(x)$, is defined by

$$\hat{f}_i(x) = \frac{N_i}{NW}, \qquad x \in \Delta_i \tag{10.4.2}$$

and N_i is the number of observations (out of N) for which $x \in \Delta_i$. It is clear that $W \sum_i \hat{f}_i(x) = 1$, as it should.

10.4.2.2. Expectation of the Estimator

The expected value of $\hat{f}_i(x)$ is given by

$$E(\hat{f}_i(x)) = \frac{1}{NW} E(N_i), \qquad x \in \Delta_i$$

From the discussion in Section 10.4.1, N_i is a binomial random variable with probability of "success" p_i given by

$$p_i = \int_{x \in \Delta_i} f(x)\, dx$$

so that

$$E(N_i) = Np_i$$

whence

$$E(\hat{f}_i(x)) = p_i/W \tag{10.4.3}$$

In general, (10.4.3) will not equal $f(x)$ for all $x \in \Delta_i$.This is obvious because $\hat{f}_i(x)$ is constant within the cell, whereas for most phenomena of interest this is not so for $f(x)$. Thus, $\hat{f}(x)$ is a biased estimator. A quantitative estimate of the bias can be obtained by approximating $f(x)$ by the first three terms of a Taylor series expansion about the center of the cell, namely,

$$f(x) \approx f(x_{c_i}) + f'(x_{c_i})(x - x_{c_i}) + \tfrac{1}{2} f''(x_{c_i})(x - x_{c_i})^2, \qquad x \in \Delta_i$$

Hence, we have

$$E(\hat{f}_i(x)) \approx \frac{1}{W} \int_{x \in \Delta_i} [f(x_{c_i}) + f'(x_{c_i})(x - x_{c_i}) + \tfrac{1}{2} f''(x_{c_i})(x - x_{c_i})^2]\, dx$$

or

$$E(\hat{f}_i(x)) \approx f(x_{c_i}) + f''(x_{c_i}) \frac{W^2}{24}, \qquad x \in \Delta_i$$

Thus, the bias of $\hat{f}_i(x)$ as an estimator for $f(x_{c_i})$ is approximately

$$b(\hat{f}_i(x)) = E(\hat{f}_i(x) - f(x_{c_i})) \approx f''(x_{c_i})\frac{W^2}{24} \tag{10.4.4}$$

which evidently increases with cell size width W.

■ *Example 10.4.1.* A brief example is helpful in visualizing the magnitude of the bias. Suppose $f(x)$ is Gaussian,

$$f(x) = \frac{1}{(2\pi)^{1/2}\sigma} e^{-x^2/2\sigma^2}$$

Then, we have

$$f''(x) = f(x)\left[\left(\frac{x}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right]$$

A more meaningful indicator of bias than (10.4.4) itself is the bias relative to the true value, or the normalized bias

$$\varepsilon_b = b(\hat{f}_i(x))/f(x_{c_i})$$

Hence, we have

$$\varepsilon_b = \left\{\left(\frac{x_{c_i}}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right\}\frac{W^2}{24}$$

Let $x_{c_i} = iW$ and suppose we set $W = \alpha\sigma$. Then, we have

$$\varepsilon_b = \frac{\alpha^2}{24}[(i\alpha)^2 - 1]$$

This allows one to size $W(= \alpha\sigma)$ for a desired ε_b and a desired location (i) in the tail of the density. ■

10.4.2.3. Variance of the Estimator

Since N_i is a binomial random variable, its variance is $Np_i(1 - p_i)$, so that the variance of $\hat{f}_i(x)$ is given by

$$\text{Var}[\hat{f}_i(x)] = \left(\frac{1}{NW}\right)^2 Np_i(1 - p_i)$$

Approximating p_i by

$$p_i = \int_{x\in\Delta_i} f(x)\,dx \approx Wf(x_{c_i})$$

results in

$$\text{Var}[\hat{f}_i(x)] = \frac{1}{NW} f(x_{c_i})[1 - Wf(x_{c_i})]$$

The normalized standard error ε_r for estimating $f(x_{c_i})$ is

$$\varepsilon_r = \frac{\{\text{Var}[\hat{f}_i(x)]\}^{1/2}}{f(x_{c_i})} = \left[\frac{1 - Wf(x_{c_i})}{NWf(x_{c_i})}\right]^{1/2} \tag{10.4.5}$$

Notice that the effect of cell width is opposite to that on the bias. Here, for given x_{c_i}, as $W \to 0$ we see that $\varepsilon_r \to [NWf(x_{c_i})]^{-1/2}$, hence the variance and standard error increase as W decreases. If W is fixed at a value that yields acceptable bias, then it is clear that in order to maintain ε_r at some given desirable value, it is necessary that N increase inversely with $f(x_{c_i})$. In other words, in order to maintain a fixed (relative) accuracy for $\hat{f}(x_{c_i})$ when $f(x_{c_i})$ is in the tails we must increase N so that $Nf(x_{c_i})$ remains constant. It may be seen that if we choose $W \propto 1/\sqrt{N}$, both the bias and standard error will tend to zero as $N \to \infty$.

■ *Remark.* In digital communications the performance criterion is error probability, which is given by the area under $f(x)$ for x in the "tail" region, say $x > x_T$, the "threshold." This is equivalent to the value of the CDF $F(x_T)$ or $1 - F(x_T)$ depending upon the symbol detected. (This is discussed in detail in the next chapter.) It seems sometimes to be assumed that it is easier to estimate $f(x)$ and integrate the estimator $\hat{f}(x)$ than directly to estimate $F(x_T)$. This misimpression is probably due to the fact that for a relatively large N the mode of the density can be estimated quite accurately. This follows from the expression for ε_r in (10.4.5). This allows a good analytical fit of pdf around the mode, which perhaps leads to a false sense of security about extrapolating the form near the mode to the tails. The fact is, as we have seen, that the accuracy in the tails decreases for a given N, and is proportional to $N^{-1/2}$. This is precisely the dependence on N of the accuracy for estimating a probability directly. This is hardly surprising since a probability is the integral of a density. In fact, a probability estimated by integrating the empirical density function is nothing more than the empirical distribution $\hat{F}(x)$ evaluated at the appropriate value of x. (Actually, this last statement is approximate, but sufficiently good for moderately small values of the cell width W.) Now, the distribution of $\hat{F}(x)$ is given by (10.4), which has expected value $F(x)$ and variance $F(x)[1 - F(x)]/N$. Thus, for given x and N, we should expect the same accuracy whether estimating $\hat{F}$ directly or by integrating $\hat{f}$. ■

■ *Summary of the Simulation Procedure (Histogram).* The construction of a histogram can be summarized as follows:

1. Collect N sampled values of a waveform.
2. Choose an interval width W and a reference interval center x_{c_0}.
3. The ith interval is centered at $x_{c_0} + iW = x_{c_i}$.
4. Count the number of samples in $x_{c_i} \pm W/2$; this is N_i.
5. Calculate the estimate $\hat{f}_i = N_i/NW$, all i. ■

10.5. Estimating the Power Spectral Density (PSD) of a Process

The power spectral density (PSD) embodies the frequency-domain properties of a process. Probably the single most important frequency domain attribute is "bandwidth,"

which is derivable from the PSD. (We note that bandwidth is not uniquely defined[15,16]. Traditionally, the PSD of a particular modulation method has been obtained by analytical means, and a great deal of attention has been paid to this problem; see, e.g., Refs. 17–26. However, derivation of the PSD in an analytical or computable form can pose significant difficulties when practical equipment impairments or nonlinearities are present. Thus, it is often more straightforward to determine the PSD "experimentally," using the simulated system. Such a measurement implies a statistical approach to determine properties of the PSD estimator.

There is a rich body of literature on spectral estimation (see, e.g., Refs. 11 and 27–31), much of which deals with "model-based" techniques that are designed to cope with "system identification" issues. One such issue is concerned with the resolution of close "spikes." Another issue is that the signal may be observable for an interval so short (corresponding to less than a few hundred samples) that "conventional" techniques could not produce a meaningful estimate of PSD. Neither of the above issues presents a problem in simulation, where we typically have available thousands of samples. For simulation application, therefore, it is sufficient to take recourse only to the straightforwardly implemented "classical" techniques.

10.5.1. Form of the Estimator

There are two "classical" techniques of spectral estimation, based on alternative definitions of power spectral density. These two definitions are equivalent for a wide-sense stationary process.

10.5.1.1. The Correlogram or Indirect Method

The first definition of the PSD $S_{XX}(f)$ of the process $X(t)$ is as the Fourier transform of the autocorrelation function $R_{XX}(\tau)$. For a continuous function, the classical definition is[17,32]

$$S_{XX}(f) = \int_{-\infty}^{\infty} R_{XX}(\tau)\ \exp(-j2\pi f\tau)\, d\tau, \qquad -\infty < f < \infty \tag{10.5.1}$$

In simulation we must deal with samples of $X(t)$. The discrete-time version of (10.5.1) is given by[30]

$$P_{XX}(f) = T_s \sum_{k=-\infty}^{\infty} R_{XX}(k)\ \exp(-j2\pi fkT_s), \qquad |f| \le \tfrac{1}{2}f_s \tag{10.5.2}$$

where $R_{XX}(k)$ is the discrete-time autocorrelation function (or sequence)

$$\begin{aligned} R_{XX}(k) &\triangleq R_{XX}(kT_s) = E[X(nT_s + kT_s)X(nT_s)] \\ &\triangleq E[X(n+k)X(n)] \end{aligned} \tag{10.5.3}$$

(In much of the literature dealing with discrete-time systems and digital signal processing, the time element is not explicit. But since we are dealing here with approximation of continuous systems, the time element is taken into account with the factor T_s, which is necessary in order

to keep the units correct.) Notice that we have changed notation from $S_{XX}(f)$ for the continuous PSD to $P_{XX}(f)$ for the discrete-time PSD. It is well known that sampling replicates the spectrum at frequency intervals to $1/T_s$ (see Chapters 3 and 6). We assume that T_s has been chosen to control the *aliasing error* to an acceptable level, and take it as our objective to determine $P_{XX}(f)$ as best we can.

In practice we have only a finite number of observations $\{X(n)\}$. A natural estimator of (10.5.2) is to replace the true autocorrelation sequence by an estimator, or *correlogram*, $\hat{R}_{XX}(k)$. Hence, the PSD estimator can be written as

$$\hat{P}_{XX}(f) = T_s \sum_{k=-L}^{L} \hat{R}_{XX}(k) \exp(-j2\pi fkT_s) \tag{10.5.4}$$

where the index k is referred to as the *lag*, and the maximum lag used or available is L. It should be noted that $\hat{P}_{XX}$ is periodic in f and that we are interested in the first zone, $|f| \leq 1/2T_s$. To evaluate (10.5.4), we first need the autocorrelation estimate. There are two basic approaches to this, the one most commonly used being

$$\hat{R}_{XX}(k) = \frac{1}{N} \sum_{n=0}^{N-k-1} X(n+k)X^*(n) \tag{10.5.5}$$

where $k = 0, 1, \ldots, N-1$ is the number of observed samples, and X^* is the complex conjugate if X is complex. [The form (10.5.5) is customary in this context. However, if one wanted to estimate the autocorrelation function itself, (10.5.5) yields a biased estimate. To produce an unbiased estimate, replace $1/N$ in front of (10.5.5) by $1/(N-k-1)$.] Because $\hat{R}_{XX}(k) = \hat{R}^*_{XX}(-k)$, one can obtain $\hat{R}_{XX}$ for negative lags. As k increases, fewer samples are included in the summation. Hence, the estimator reliability deteriorates as k increases. For this reason, it has been suggested[31] that the maximum lag L in (10.5.4) be limited to about $N/10$. The $\hat{P}_{XX}(f)$ is given for any f in the interval $|f| \leq 1/2T_s$. In practice, we calculate $\hat{P}_{XX}(f)$ only for a discrete set $f_i = i/KT_s$, $0 \leq i \leq K$, in order to take advantage of the FFT algorithm (see Section 10.5.5).

10.5.1.2. The Periodogram or Direct Method

An alternative definition for the PSD is the following[32]:

$$S_{XX}(f) = \lim_{T \to \infty} E[\hat{S}_{XX}(f, T)] \tag{10.5.6}$$

where

$$\hat{S}_{XX}(f, T) = \frac{|X_T(f)|^2}{T} \tag{10.5.7}$$

and

$$X_T(f) = \int_0^T X(t) \exp(-j2\pi ft)\, dt \tag{10.5.8}$$

is the finite-time Fourier transform of a sample function of the process. When the process is wide-sense stationary, (10.5.6) is equivalent to (10.5.1). A "natural" estimate of $S_{XX}(f)$ is suggested by (10.5.6) if we simply omit the limiting and expectation operations. That is, $\hat{S}_{XX}(f, T)$ is an estimate for $S_{XX}(f)$. The $\hat{S}_{XX}(f, T)$ is referred to as the *periodogram* or the *sample spectrum.* The discrete-time version of $\hat{S}_{XX}(f, T)$ is denoted $\tilde{P}_{XX}(f)$, and is given by

$$\tilde{P}_{XX}(f) = \frac{1}{NT_s} |X_N(f)^2| \tag{10.5.9}$$

where

$$X_N(f) = T_s \sum_{n=0}^{N-1} X(n) \exp(-j2\pi fnT_s) \tag{10.5.10}$$

is the discrete Fourier transform of the observed data sequence. (The tilde here stands for estimator, as the caret usually does. We use a different symbol here to distinguish the two estimators.) It can be seen that X_N, hence $\tilde{P}_{XX}$, is periodic in f, but we are interested only in the fundamental period. Again, to speedily compute $\tilde{P}_{XX}$ one normally relies on the FFT algorithm to determine discretely spaced values $\tilde{P}_{XX}(f_k), f_k = k/KT_s, k = 0, 1, \ldots, K-1$. If $K > N - 1$, zero padding is implied.

■ *Remark on the Relationship between Direct and Indirect Methods.* If, in the indirect method the maximum lag L is set to $N - 1$ the largest possible value, the estimators $\hat{P}_{XX}(f)$ and $\tilde{P}_{XX}(f)$ are identical.[30] While the same observed sequence forms the basis of both procedures, there is the possibility of further manipulation of either form of the estimator which may or may not have a ready interpretation in the other form. Although $\hat{P}_{XX}(f)$ and $\tilde{P}_{XX}(f)$ contain the essential ideas of PSD estimation, as estimators they suffer from certain drawbacks, which can be remedied to some extent by auxiliary techniques referred to as windowing and averaging (smoothing). These techniques are customarily applied in practice and they are discussed in the next section. ■

10.5.2. Modified Form of the Estimator: Windowing and Averaging

The term *window* implies a function $w(t)$ that "lets in" or colors certain portions of another function $X(t)$. Any finite observation is implicitly a windowed sample function of a process. In that case the (time) window is inescapable and induces distortion which can be mitigated to some extent by intentionally superimposing another window. There are two basic types of windows: *data windows*, denoted $d(n)$, that are used to modify directly (multiply) the observed data sequence $X(n)$; and *lag windows*, denoted $\lambda(k)$, that are used to modify directly the autocorrelation sequence $\hat{R}_{XX}(k)$. Data windows are also referred to as tapering functions. The transform of a data window is called a *frequency window* and denoted $W(f)$, and the transform of a lag window is called a *spectral window* denoted $\Omega(f)$.

If we apply a data window $d(n)$ to $X(n)$ in (10.5.10), we have

$$X_N(f;\, d) = T_s \sum_{n=0}^{N-1} X(n)d(n) \exp(-j2\pi fnT_s) \tag{10.5.11}$$

$$= X_N(f) \circledast W(f) \tag{10.5.12}$$

where $\circledast$ represents circular convolution, and the presence of the window is made explicit in the argument of X_N. Therefore the PSD estimator is

$$\tilde{P}_{XX}(f;\ d) = \frac{1}{(NT_s)\xi}|X_N(f;\ d)|^2 \tag{10.5.13}$$

where

$$\xi = T_s \sum_{n=0}^{N-1} d^2(n) \tag{10.5.14}$$

is the discrete-time window energy. Division by ξ maintains the expected value of the sample average power unchanged; obviously, this operation can be avoided if $d(n)$ is prenormalized so that $\xi = 1$.

The effect of the window is to convolve the discrete-time Fourier transform of the data with the frequency window. This makes it somewhat difficult to visualize the net effect on $\tilde{P}_{XX}(f)$ from sample to sample. However, a more easily interpretable result is obtainable in an average sense. One can show that

$$E[\tilde{P}_{XX}(f;\ d)] = \sum_{k=-\infty}^{\infty} T_s\lambda(k)R_{XX}(k)\ \exp(-j2\pi fkT_s) \tag{10.5.15}$$

where $R_{XX}(k)$ is the true autocorrelation sequence, and

$$\lambda(k) = \frac{1}{N}\sum_{n=-\infty}^{\infty} d(n)\ d(n+k) \tag{10.5.16}$$

It can be seen that $\lambda(k)$ acts as a lag window. Hence, in an average sense, the data window is equivalent to a lag window which is the data window's autocorrelation function.

Equation (10.5.15) can be expressed as

$$E[\tilde{P}_{XX}(f;\ d)] = P_{XX}(f) \circledast \Omega(f) \tag{10.5.17}$$

where $\Omega(f)$ is the discrete Fourier transform of $\lambda(k)$. Note that $\lambda(k)R_{XX}(k)$ is the effective autocorrelation $R_{XX}(k;\ d)$ of the windowed sequence, and that $R_{XX}(0,\ d)$ is the average power of the process. Hence, this average power will be unaffected if we set $\lambda(0) = 1$, which we can do simply by dividing (10.5.17) by

$$\lambda(0) = (1/N)\sum d^2(n) = (1/NT_s)\xi$$

It can also be shown that

$$\Omega(f) = (1/NT_s)|W(f)|^2 \tag{10.5.18}$$

Therefore, the properly normalized version of (10.5.17) is

$$E[\tilde{P}_{XX}(f;\ d)] = P_{XX}(f) \circledast \Omega(f)/\lambda(0) = P_{XX}(f) \circledast |W(f)|^2/\xi \tag{10.5.19}$$

Table 10.1. Definitions of Typical N-Point Discrete-Time Windows[a]

Window name	Discrete-time function $d[n]$ or $\lambda[k]$	Frequency response $W(f)$ or $\Omega(f)$
Rectangular (uniform)	1	$D_N(f)$
Triangle (Bartlett)	$1 - \lvert t[n]\rvert$	$2[(N-1)T]^{-1}D^2_{(N-1)/2}(f)$
Squared cosine (Hann)	$0.50 + 0.50\cos(2\pi t[n])$	$0.50D_N(f) + 0.25[D_N(f-(NT)^{-1}) + D_N(f+(NT)^{-1})]$
Raised cosine (Hamming)	$0.54 + 0.46\cos(2\pi t[n])$	$0.54D_N(f) + 0.23[D_N(f-(NT)^{-1}) + D_N(f+(NT)^{-1})]$
Weighted cosines (Nuttall, $R=3$)	$\sum_{r=0}^{R} \alpha_r \cos(2\pi r t[n])$	$\sum_{r=0}^{R} 0.5\alpha_r[D_N(f-r(NT)^{-1}) \times D_N(f+r(NT)^{-1})]$
Truncated Gaussian ($\alpha = 2.5$)	$p\{-2(\alpha t[n])^2\}$	$(\sqrt{2\pi}/2\alpha)\exp[-2(\pi fT/\alpha)^2] * D_N(f)$

[a]From S. L. Marple, Jr., Frequency-Time Signal Processing, course notes. © 1991, S. L. Marple, Jr., reproduced with permission.

An alternative approach to windowing is to define a window in the lag domain explicitly, rather than in the time domain. That is, obtain the autocorrelation sequence from unwindowed data, then apply a lag window $\lambda(k)$.

Much attention has been paid in the literature to windows.[11,27–31] For illustration we show here a few of the more common ones. Table 10.1 catalogs the formulas for some data or lag windows and their transforms, while Figure 10.8 illustrates these forms in the "sample" and frequency domains. Maple[31] also provides listings and a disk for programs that compute the PSD through both approaches. In Table 10.1, T is the sampling interval. The data window $d[n]$ is defined over the index range $0 \le n \le N-1$ and N may be even or odd. The lag window $\lambda[k]$ is defined over the index range $-(N-1)/2 \le k \le (N-1)/2$ and N is always odd. The function $t[n] = 2n/(N-1) - 1$ for $d[n]$ and $t[k] = 2k/(N-1)$ for $\lambda[k]$. The function $D_N(f) = T\exp[-j2\pi fT(N-1)]\sin(\pi fTN)/\sin(\pi fT)$ for $d[n]$ and $D_N(f) = T\sin(\pi fTN)/\sin(\pi fT)$ for $\lambda[k]$.

A window distorts the true spectrum. The customary measures of that distortion are bias and variability (in the usual statistical sense) and resolution. The latter is the ability to distinguish two relatively closely spaced spectral lines one from the other. The choice of window rests on balancing bias and resolution, as improving one will generally worsen the other. In simulation we can usually obtain a long enough record to satisfy good performance on both counts. We recall that windows are also beneficial in the filtering context, as discussed in Chapter 3.

We need to introduce one more operation that is necessary for the periodogram approach, and this is the smoothing or averaging of periodograms. The operation means simply to average $\tilde{P}_{xx}(f)$ or $\tilde{P}_{xx}(f; d)$ calculated for a number of data segments. In general, the individual segments do not have to be contiguous. In fact, there appear to be advantages in using overlapped segments. The periodogram obtained by averaging unwindowed nonoverlapping segments is called the Bartlett periodogram, $\tilde{P}_B(f)$, and the periodogram obtained by averaging windowed overlapped segments is called the Welch periodogram, $\tilde{P}_w(f)$.[33] Let $\tilde{P}_{XX,m}(f; d)$ be the periodogram obtained from the mth data segment; then,

$$\tilde{P}_w(f) = \frac{1}{M}\sum_{m=0}^{M-1} \tilde{P}_{XX,m}(f;\ d) \tag{10.5.20}$$

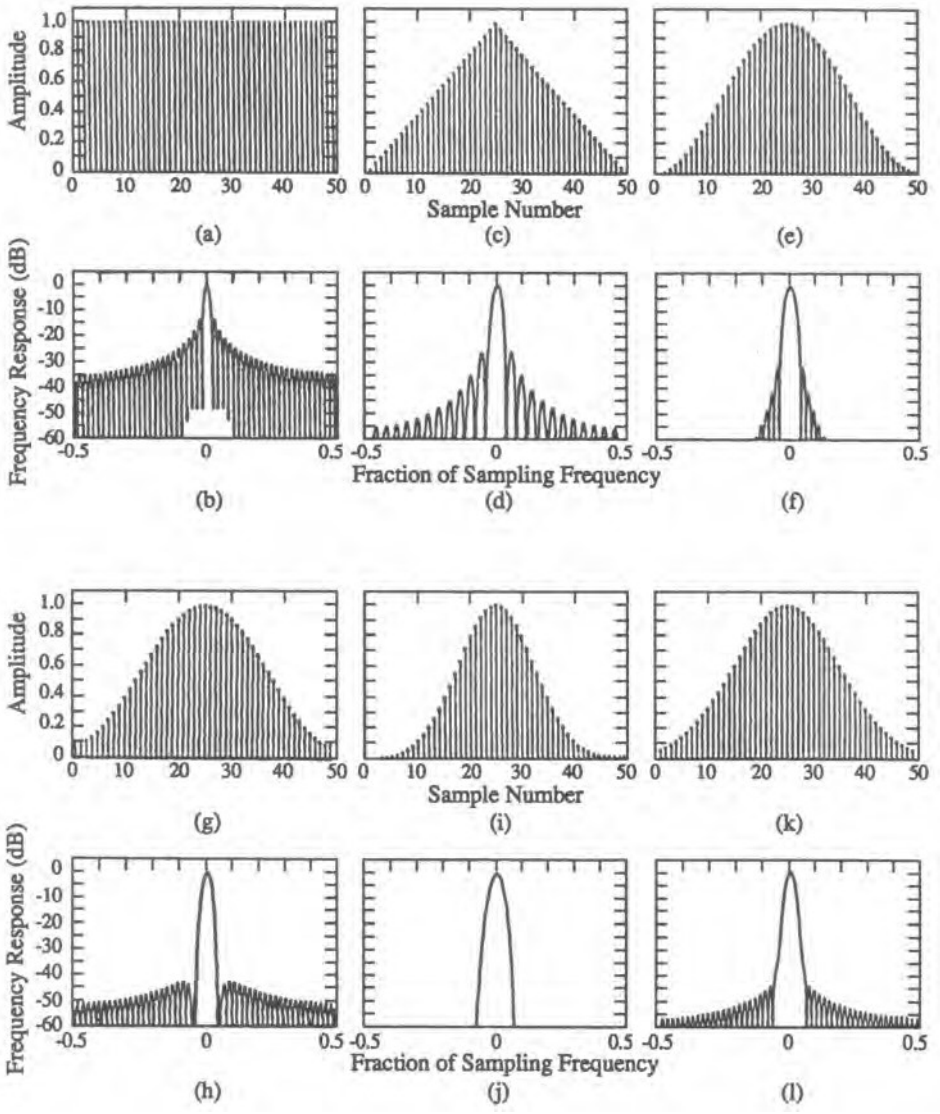


Figure 10.8. Typical discrete-time windows (upper) and log of their DTFT magnitude (lower), (a, b) Rectangular; (c, d), triangular; (e, f), Hann; (g, h), Hamming; (i, j), Nuttal; (k, 1), truncated Gaussian. (From S. L. Marple, Jr., Frequency-Time Signal Processing, course notes, © 1991, S. L. Marple Jr., reproduced with permission.)

where M is the total number of segments, $M \geq N/J$, where J is the number of samples per segment. Another popular smoothing approach uses a running average in the frequency domain[11] (see also problem P10.9.)

10.5.3. Expected Value of the Estimator

According to (10.5.19), the expected value of $\tilde{P}_{XX}(f; d)$ is not $P_{XX}(f)$, hence the periodogram is a biased estimate. However, $\tilde{P}_{XX}(f; d)$ is asymptotically unbiased because for

any reasonable window, including the implicit rectangular window, $|W(f)|^2 \to \delta(f)$ as $N \to \infty$. Hence, it will generally be true that (with $\xi = 1$),

$$\lim_{N\to\infty} E[\tilde{P}_{XX}(f;\ d)] = P_{XX}(f) \tag{10.5.21}$$

The same statements are also true for the Bartlett or Welch periodograms, keeping in mind that the window is J samples rather than N samples long. Hence (10.5.21) is true for $\tilde{P}_B(f)$ or $\tilde{P}_w(f)$ if both $J, N \to \infty$.

It can also be shown, for the periodogram estimate with applied lag window $\lambda(k)$, that there is an effective lag window $(1 - |k|/N)\lambda(k) = \lambda_e(k)$ with transform $\Omega_e(f)$ such that

$$E[\hat{P}_{XX}(f;\ \lambda)] = P_{XX}(f) \circledast \Omega e\ (f) \tag{10.5.22}$$

Even if no lag window is explicitly applied, there is an implied window, as before. Notice that if $L = N - 1$ and no applied window is used, then the triangular window $(1 - |k|/N)$ is precisely the convolution of the rectangular data window, as expected.

10.5.4. Variance of the Estimator

For an arbitrary process, the variance of the estimator is difficult to evaluate, but for a Gaussian process, one can show[32] that

$$E[\hat{S}^2_{XX}(f, T)] \geq 2\{E[S_{XX}(f, T)]\}^2 \tag{10.5.23}$$

By definition of variance we then have

$$\text{Var}\ [\hat{S}_{XX}(f, T)] \geq \{E[\hat{S}_{XX}(f, T)]\}^2 \tag{10.5.24}$$

which states that, *independently of* T, the standard deviation of the estimated spectrum is at least as large as the expected value of the estimated spectrum itself. Thus, the unmodified periodogram is not only a poor estimator, it cannot be improved by increasing the length of observation. Essentially, the reason is that the conventional periodogram ignores the expectation operation in (10.5.6) which makes it a consistent estimator. This is the motivation for the Bartlett and Welch periodograms, which form "pseudoensembles" over which to average. For discrete-time functions it can be shown (see, e.g., Ref. 31) that the variance of the periodogram of a (discrete) zero-mean white Gaussian process is

$$\begin{aligned}\text{Var}[\tilde{P}_{XX}(f)] &= P^2_{XX}(f)\left[1 + \left(\frac{\sin(2\pi f T_s N)}{N \sin(2\pi f T_s)}\right)^2\right] \\ &\geq P^2_{XX}(f)\end{aligned} \tag{10.5.25}$$

which reconfirms the more general result (10.5.24).

For a Gaussian process, it is possible to find the distribution of the periodogram, from which we can calculate confidence intervals if desired. This is because, in either the continuous or discrete form, $X_T(f)$ or $X_N(f)$ is a linear combination of normal variables, hence is itself normal. The periodogram, which is proportional to $|X_T|^2$ or $|X_N|^2$ is then, by definition, chi-square distributed. (See, e.g., Refs. 11 and 30 for a derivation.)

If the Bartlett or Welch periodograms can be considered to be obtained from independent segments, then, from (10.5.20) we have

$$\mathrm{Var}[\tilde{P}_w(f)] = \frac{1}{M}\mathrm{Var}[\tilde{P}_{XX,m}(f;\ d)] \tag{10.5.26a}$$

and since, from the previous discussion, the variance of the individual periodograms $\mathrm{Var}[\tilde{P}_{XX,m}(f; d)]$ is approximately $P_{XX}^2(f)$, we have that

$$\mathrm{Var}[\tilde{P}_w(f)] \approx \frac{1}{M}P_{XX}^2(f) \tag{10.5.26b}$$

Thus, segmenting the data improves the variance by a factor equal to the number of segments. If there is overlap in the Welch periodogram, the segments will not be independent, but there will be more segments, hence (10.5.2b) will still be approximately correct. It should be noted that in order to drive the variance (10.5.2b) to zero as N becomes indefinitely large, M should increase also. An appropriate scheme is to choose $M \propto \sqrt{N}$.

In the correlogram approach, the above results still apply if the maximum possible lag is used, since, in that case, the estimator is identical to the periodogram. If a maximum lag $L < N - 1$ is used, the ratio N/L can be thought of as having the same effect as segmenting does in the periodogram approach, namely, reducing the variance by that number. This is the basis of the statement made earlier that L should not be larger than about $N/10$.

10.5.5. Some Considerations on Implementing PSD Estimators: Summary of the Simulation Procedure

Use of the FFT is standard because of processing speed. If we observe N samples *per segment*, or time interval NT_s, the FFT yields N frequency samples $f_i = i/NT_s$, $i = 0, 1, \ldots, N-1$. The bandwidth of a (truly bandlimited) baseband process that could be observed without aliasing is $1/2T_s$, and the resolution is $1/NT_s$. The number N in almost every FFT routine is a power of 2. If the number of observations cannot be made a power of 2, one can contrive to do so by adding zeros to the end of the observed data sequence. This is referred to as *zero-padding*. The discrete Fourier transform is unaffected at any f if we simply add zeros to $X(n)$. However, for discretely spaced f, this may not be the case and the frequency locations may not be the same, depending upon the relationship between N and N_p, the padded sequence length. If N_p/N is an integer, however, the original frequency components will appear at their original values. There will be additional components as well, which can be considered interpolated values.

For the actual simulation procedure, either the direct or indirect method can be used, although the Welch periodogram appears to be the most convenient to implement. The following summarizes the procedure to be used for either approach.

10.5.5.1. Welch Periodogram Procedure (Direct Method)

1. Collect N total data points $X(n)$, $n = 0, 1, \ldots, N-1$.
2. Select a segmentation and overlap scheme, namely, choose J samples per segment and shift S between segments so that the number of segments $M = [(N-J)/(S+1)]$