Poor, H.V., Looney, C.G., Marks II, R.J., Verdú, S., Thomas, J.A., Cover, T.M.
"Information Theory"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

73

Information Theory

H. Vincent Poor
Princeton University

Carl G. Looney
University of Nevada

R. J. Marks II
University of Washington

Sergio Verdú
Princeton University

Joy A. Thomas
IBM

Thomas M. Cover
Stanford University

73.1 Signal Detection

H. Vincent Poor

The field of signal detection and estimation is concerned with the processing of information-bearing signals for the purpose of extracting the information they contain. The applications of this methodology are quite broad, ranging from areas of electrical engineering such as automatic control, digital communications, image processing, and remote sensing, into other engineering disciplines and the physical, biological, and social sciences.

There are two basic types of problems of interest in this context. *Signal detection* problems are concerned primarily with situations in which the information to be extracted from a signal is discrete in nature. That is, signal detection procedures are techniques for deciding among a discrete (usually finite) number of possible alternatives. An example of such a problem is the demodulation of a digital communication signal, in which the task of interest is to decide which of several possible transmitted symbols has elicited a given received signal. *Estimation* problems, on the other hand, deal with the determination of some numerical quantity taking values in a continuum. An example of an estimation problem is that of determining the phase or frequency of the carrier underlying a communication signal.

Although signal detection and estimation is an area of considerable current research activity, the fundamental principles are quite well developed. These principles, which are based on the theory of statistical inference, explain and motivate most of the basic signal detection and estimation procedures used in practice. In this section, we will give a brief overview of the basic principles underlying the field of signal detection. Estimation is treated elsewhere this volume, notably in Section 16.2. A more complete introduction to these subjects is found in Poor [1994].

General Considerations

The basic principles of signal detection can be conveniently discussed in the context of decision-making between two possible statistical models for a set of real-valued measurements, $Y_1, Y_2, \ldots, Y_n$. In particular, on observing $Y_1, Y_2, \ldots, Y_n$, we wish to decide whether these measurements are most consistent with the model

$$Y_k = N_k, \qquad k = 1, 2, \ldots, n \tag{73.1}$$

or with the model

$$Y_k = N_k + S_k, \quad k = 1, 2, \ldots, n \tag{73.2}$$

where $N_1, N_2, \ldots, N_n$ is a random sequence representing noise, and where $S_1, S_2, \ldots, S_n$ is a sequence representing a (possibly random) signal.

In deciding between Eqs. (73.1) and (73.2), there are two types of errors possible: a *false alarm,* in which (73.2) is falsely chosen, and a *miss,* in which (73.1) is falsely chosen. The probabilities of these two types of errors can be used as performance indices in the optimization of rules for deciding between (73.1) and (73.2). Obviously, it is desirable to minimize both of these probabilities to the extent possible. However, the minimization of the **false-alarm probability** and the minimization of the **miss probability** are opposing criteria. So, it is necessary to effect a trade-off between them in order to design a signal detection procedure. There are several ways of trading off the probabilities of miss and false alarm: the **Bayesian detector** minimizes an average of the two probabilities taken with respect to prior probabilities of the two conditions (73.1) and (73.2), the *minimax* detector minimizes the maximum of the two error probabilities, and the **Neyman-Pearson detector** minimizes the miss probability under an upper-bound constraint on the false-alarm probability.

If the statistics of noise and signal are known, the Bayesian, minimax, and Neyman-Pearson detectors are all of the same form. Namely, they reduce the measurements to a single number by computing the **likelihood ratio**

$$L(Y_1, Y_2, \ldots, Y_n) \triangleq \frac{p_{S+N}(Y_1, Y_2, \ldots, Y_n)}{p_N(Y_1, Y_2, \ldots, Y_n)} \tag{73.3}$$

where p_{S+N} and p_N denote the probability density functions of the measurements under signal-plus-noise (73.2) and noise-only (73.1) conditions, respectively. The likelihood ratio is then compared to a *decision threshold,* with the signal-present model (73.2) being chosen if the threshold is exceeded, and the signal-absent model (73.1) being chosen otherwise. Choice of the decision threshold determines a trade-off of the two error probabilities, and the optimum procedures for the three criteria mentioned above differ only in this choice.

There are several basic signal detection structures that can be derived from Eqs. (73.1) to (73.3) under the assumption that the noise sequence consists of a set of independent and identically distributed (i.i.d.) Gaussian random variables with zero means. Such a sequence is known as **discrete-time white Gaussian noise.** Thus, until further notice, we will make this assumption about the noise. It should be noted that this assumption is physically justifiable in many applications.

Detection of Known Signals

If the signal sequence $S_1, S_2, \ldots, S_n$ is known to be given by a specific sequence, say $s_1, s_2, \ldots, s_n$ (a situation known as *coherent detection*), then the likelihood ratio (73.3) is given in the white Gaussian noise case by

$$\exp\left\{\left(\sum_{k=1}^{n} s_k Y_k - \frac{1}{2}\sum_{k=1}^{n} s_k^2\right) / \sigma^2\right\} \tag{73.4}$$

where σ^2 is the variance of the noise samples. The only part of (73.4) that depends on the measurements is the term $\Sigma_{k=1}^{n} s_k Y_k$ and the likelihood ratio is a monotonically increasing function of this quantity. Thus, optimum detection of a coherent signal can be accomplished via a correlation detector, which operates by comparing the quantity

$$\sum_{k=1}^{n} s_k Y_k \tag{73.5}$$

to a threshold, announcing signal presence when the threshold is exceeded.

Note that this detector works on the principle that the signal will correlate well with itself, yielding a large value of (73.5) when present, whereas the random noise will tend to average out in the sum (73.5), yielding a relatively small value when the signal is absent. This detector is illustrated in Fig. 73.1.

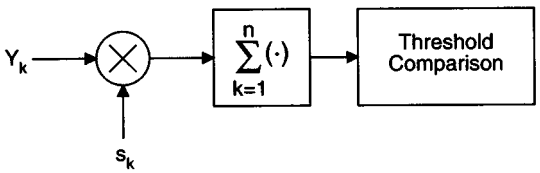

FIGURE 73.1 Correlation detector for a coherent signal in additive white Gaussian noise.

Detection of Parametrized Signals

The correlation detector cannot usually be used directly unless the signal is known exactly. If, alternatively, the signal is known up to a short vector $\boldsymbol{\theta}$ of random parameters (such as frequencies or phases) that are independent of the noise, then an optimum test can be implemented by threshold comparison of the quantity

$$\int_{\Lambda} \exp\left\{\left(\sum_{k=1}^{n} s_k(\boldsymbol{\theta}) Y_k - \frac{1}{2}\sum_{k=1}^{n} \left[s_k(\boldsymbol{\theta})\right]^2\right) / \sigma^2\right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{73.6}$$

where we have written $S_k = s_k(\boldsymbol{\theta})$ to indicate the functional dependence of the signal on the parameters, and where Λ and $p(\boldsymbol{\theta})$ denote the range and probability density function, respectively, of the parameters.

The most important example of such a parametrized signal is that in which the signal is a modulated sinusoid with random phase; i.e.,

$$S_k = a_k \cos(\omega_c k + \theta), \qquad k = 1, 2, \ldots, n \tag{73.7}$$

where $a_1, a_2, \ldots, a_n$ is a known amplitude modulation sequence, ω_c is a known (discrete-time) carrier frequency, and the random phase θ is uniformly distributed in the interval $[-\pi,\pi]$. In this case, the likelihood ratio is a monotonically increasing function of the quantity

$$\left[\sum_{k=1}^{n} a_k \cos(\omega_c k) Y_k\right]^2 + \left[\sum_{k=1}^{n} a_k \sin(\omega_c k) Y_k\right]^2 \tag{73.8}$$

Thus, optimum detection can be implemented via comparison of (73.8) with a threshold, a structure known as an **envelope detector.** Note that this detector correlates the measurements with two orthogonal components of the signal, $a_k \cos(\omega_c k)$ and $a_k \sin(\omega_c k)$. These two correlations, known as the in-phase and quadrature components of the measurements, respectively, capture all of the energy in the signal, regardless of the value of θ. Since θ is unknown, however, these two correlations cannot be combined coherently, and thus they are combined noncoherently via (73.8) before the result is compared with a threshold. This detector is illustrated in Fig. 73.2.

Parametrized signals also arise in situations in which it is not appropriate to model the unknown parameters as random variables with a known distribution. In such cases, it is not possible to compute the likelihood ratio (73.6) so an alternative to the likelihood ratio detector must then be used. (An exception is that in which the likelihood ratio detector is invariant to the unknown parameters—a case known as *uniformly most powerful detection.*) Several alternatives to the likelihood ratio detector exist for these cases.

One useful such procedure is to test for the signal's presence by threshold comparison of the *generalized likelihood ratio,* given by

$$\max_{\boldsymbol{\theta} \in \Lambda} L_{\boldsymbol{\theta}}(Y_1, Y_2, \ldots, Y_n) \tag{73.9}$$

where $L_{\boldsymbol{\theta}}$ denotes the likelihood ratio for Eqs. (73.1) and (73.2) for the known-signal problem with the parameter vector fixed at $\boldsymbol{\theta}$. In the case of white Gaussian noise, we have

$$L_{\boldsymbol{\theta}}(Y_1, Y_2, \ldots, Y_n) = \exp\left\{\left(\sum_{k=1}^{n} s_k(\boldsymbol{\theta}) Y_k - \frac{1}{2}\sum_{k=1}^{n}\left[s_k(\boldsymbol{\theta})\right]^2\right) / \sigma^2\right\} \tag{73.10}$$

It should be noted that this formulation is also valid if the statistics of the noise have unknown parameters, e.g., the noise variance in the white Gaussian case.

One common application in which the generalized likelihood ratio detector is useful is that of detecting a signal that is known except for its time of arrival. That is, we are often interested in signals parametrized as

$$s_k(\theta) = a_{k-\theta} \tag{73.11}$$

where $\{a_k\}$ is a known finite-duration signal sequence and where θ ranges over the integers. Assuming white Gaussian noise and an observation interval much longer than the duration of $\{a_k\}$, the generalized likelihood ratio detector in this case announces the presence of the signal if the quantity

$$\max_{\theta} \sum_{k} a_{k-\theta} Y_k \tag{73.12}$$

exceeds a fixed threshold. This type of detector is known as a *matched filter,* since it can be implemented by filtering the measurements with a digital filter whose pulse response is a time-reversed version of the known

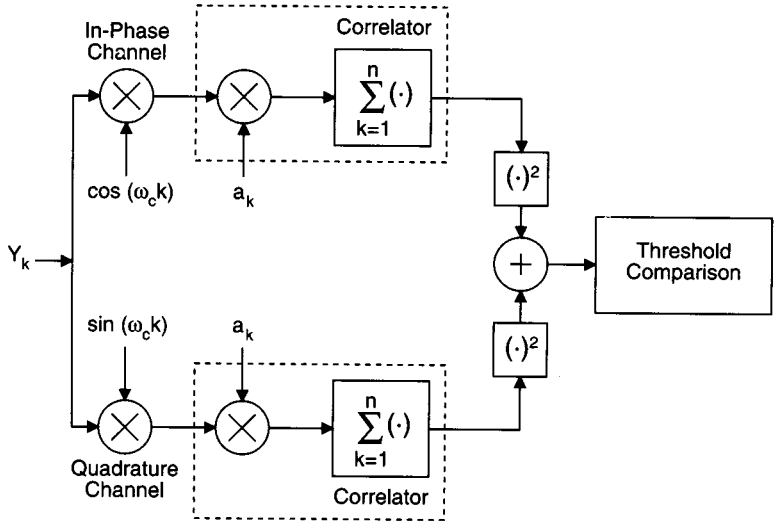

FIGURE 73.2 Envelope detector for a noncoherent signal in additive white Gaussian noise.

signal $\{a_k\}$ (hence it is "matched" to the signal), and announcing the signal's presence if the filter output exceeds the decision threshold at any time.

Detection of Random Signals

In some applications, particularly in remote sensing applications such as sonar and radio astronomy, it is appropriate to consider the signal sequence $S_1, S_2, \ldots, S_n$ itself to be a random sequence, statistically independent of the noise. In such cases, the likelihood ratio formula of (73.6) is still valid with the parameter vector $\boldsymbol{\theta}$ simply taken to be the signal itself. However, for long measurement records (i.e., large n), (73.6) is not a very practical formula except in some specific cases, the most important of which is the case in which the signal is Gaussian.

In particular, if the signal is Gaussian with zero-mean and autocorrelation sequence $r_{k,l} \triangleq E\{S_k S_l\}$, then the likelihood ratio is a monotonically increasing function of the quantity

$$\sum_{k=1}^{n} \sum_{l=1}^{n} q_{k,l} Y_k Y_l \tag{73.13}$$

with $q_{k,l}$ the element in the kth row and lth column of the positive-definite matrix

$$\mathbf{Q} \triangleq \mathbf{I} - (\mathbf{I} + \mathbf{R}/\sigma^2)^{-1} \tag{73.14}$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix, and R is the covariance matrix of the signal, i.e., it is the $n \times n$ matrix with elements $r_{k,l}$.

Note that (73.13) is a quadratic function of the measurements; thus, a detector based on the comparison of this quantity to a threshold is known as a **quadratic detector.** The simplest form of this detector results from the situation in which the signal samples are, like the noise samples, i.i.d. In this case, the quadratic function (73.13) reduces to a positive constant multiple of the quantity

$$\sum_{k=1}^{n} Y_k^2 \tag{73.15}$$

A detector based on (73.15) simply measures the energy in the measurements and then announces the presence of the signal if this energy is large enough. This type of detector is known as a *radiometer.*

Thus, radiometry is optimum in the case in which both signal and noise are i.i.d. Gaussian sequences with zero means. Since in this case the presence of the signal is manifested only by an increase in energy level, it is intuitively obvious that radiometry is the only way of detecting the signal's presence. More generally, when the signal is correlated, the quadratic function (73.13) exploits both the increased energy level and the correlation structure introduced by the presence of the signal. For example, if the signal is a narrowband Gaussian process, then the quadratic function (73.13) acts as a narrowband radiometer with bandpass characteristic that approximately matches that of the signal. In general, the quadratic detector will make use of whatever spectral properties the signal exhibits.

If the signal is random but not Gaussian, then its optimum detection [described by (73.6)] typically requires more complicated nonlinear processing than the quadratic processing of (73.13) in order to exploit the distributional differences between signal and noise. This type of processing is often not practical for implementation, and thus approximations to the optimum detector are typically used. An interesting family of such detectors uses cubic or quartic functions of the measurements, which exploit the higher-order spectral properties of the signal [Mendel, 1991]. As with deterministic signals, random signals can be parametrized. In this case, however, it is the distribution of the signal that is parametrized. For example, the power spectrum of the signal of interest may be known only up to a set of unknown parameters. Generalized likelihood ratio detectors (73.9) are often used to detect such signals.

Deciding Among Multiple Signals

The preceding results have been developed under the model (73.1)–(73.2) that there is a single signal that is either present or absent. In digital communications applications, it is more common to have the situation in which we wish to decide between the presence of two (or more) possible signals in a given set of measurements. The foregoing results can be adapted straightforwardly to such problems. This can be seen most easily in the case of deciding among known signals. In particular, consider the problem of deciding between two alternatives:

$$Y_k = N_k + s_k^{(0)}, \qquad k = 1, 2, \ldots, n \tag{73.16}$$

and

$$Y_k = N_k + s_k^{(1)}, \qquad k = 1, 2, \ldots, n \tag{73.17}$$

where $s_1^{(0)}, s_2^{(0)}, \ldots, s_n^{(0)}$ and $s_1^{(1)}, s_2^{(1)}, \ldots, s_n^{(1)}$ are two known signals. Such problems arise in data transmission problems, in which the two signals $s^{(0)}$ and $s^{(1)}$ correspond to the waveforms received after transmission of a logical "zero" and "one," respectively. In such problems, we are generally interested in minimizing the *average probability of error,* which is the average of the two error probabilities weighted by the prior probabilities of occurrence of the two signals. This is a Bayesian performance criterion, and the optimum decision rule is a straightforward extension of the correlation detector based on (73.5). In particular, under the assumptions that the two signals are equally likely to occur prior to measurement, and that the noise is white and Gaussian, the optimum decision between (73.16) and (73.17) is to choose the model (73.16) if $\Sigma_{k=1}^{n} s_k^{(0)} Y_k$ is larger than $\Sigma_{k=1}^{n} s_k^{(1)} Y_k$, and to choose the model (73.17) otherwise.

More generally, many problems in digital communications involve deciding among M equally likely signals with $M > 2$. In this case, again assuming white Gaussian noise, the decision rule that minimizes the error probability is to choose the signal $s_1^{(j)}, s_2^{(j)}, \ldots, s_n^{(j)}$, where j is a solution of the maximization problem

$$\sum_{k=1}^{n} s_k^{(j)} Y_k = \max_{0 \le m \le M-1} \sum_{k=1}^{n} s_k^{(m)} Y_k \tag{73.18}$$

There are two basic types of digital communications applications in which the problem (73.18) arises. One is in *M-ary data transmission,* in which a symbol alphabet with M elements is used to transmit data, and a decision among these M symbols must be made in each symbol interval [Proakis, 1983]. The other type of application in which (73.18) arises is that in which data symbols are correlated in some way because of intersymbol interference, coding, or multiuser transmission. In such cases, each of the M possible signals represents a frame of data symbols, and a joint decision must be made about the entire frame since individual symbol decisions cannot be decoupled. Within this latter framework, the problem (73.18) is known as *sequence detection.* The basic distinction between M-ary transmission and sequence detection is one of degree. In typical M-ary transmission, the number of elements in the signaling alphabet is typically a small power of 2 (say 8 or 32), whereas the number of symbols in a frame of data could be on the order of thousands. Thus, solution of (73.18) by exhaustive search is prohibitive for sequence detection, and less complex algorithms must be used. Typical digital communications applications in which sequence detection is necessary admit dynamic programming solutions to (73.18) (see, e.g., Verdú [1993]).

Detection of Signals in More General Noise Processes

In the foregoing paragraphs, we have described three basic detection procedures: correlation detection of signals that are completely known, envelope detection of signals that are known except for a random phase, and quadratic detection for Gaussian random signals. These detectors were all derived under an assumption of white Gaussian noise. This assumption provides an accurate model for the dominant noise arising in many communication channels. For example, the thermal noise generated in signal processing electronics is adequately described as being white and Gaussian. However, there are also many channels in which the statistical behavior of the noise is not well described in this way, particularly when the dominant noise is produced in the physical channel rather than in the receiver electronics.

One type of noise that often arises is noise that is Gaussian but not white. In this case, the detection problem (73.1)–(73.2) can be converted to an equivalent problem with white noise by applying a linear filtering process known as *prewhitening* to the measurements. In particular, on denoting the noise covariance matrix by Σ, we can write

$$\Sigma = \mathbf{C}\mathbf{C}^T \tag{73.19}$$

where $\mathbf{C}$ is an $n \times n$ invertible, lower-triangular matrix and where the superscript T denotes matrix transposition. The representation (73.19) is known as the *Cholesky decomposition.* On multiplying the measurement vector $\mathbf{Y} \triangleq (Y_1, Y_2, \ldots, Y_n)^T$ satisfying (73.1)–(73.2) with noise covariance Σ, by $\mathbf{C}^{-1}$, we produce an equivalent (in terms of information content) measurement vector that satistifies the model (73.1)–(73.2) with white Gaussian noise and with the signal conformally transformed. This model can then be treated using the methods described previously.

In other channels, the noise can be modeled as being i.i.d. but with an amplitude distribution that is not Gaussian. This type of model arises, for example, in channels dominated by impulsive phenomena, such as urban radio channels. In the non-Gaussian case the procedures discussed previously lose their optimality as defined in terms of the error probabilities. These procedures can still be used, and they will work well under many conditions; however, there will be a resulting performance penalty with respect to optimum procedures based on the likelihood ratio. Generally speaking, likelihood-ratio-based procedures for non-Gaussian noise channels involve more complex nonlinear processing of the measurements than is required in the standard detectors, although the retention of the i.i.d. assumption greatly simplifies this problem. A treatment of methods for such channels can be found in Kassam [1988].

When the noise is both non-Gaussian and dependent, the methodology is less well developed, although some techniques are available in these cases. An overview can be found in Poor and Thomas [1993].

Robust and Nonparametric Detection

All of the procedures outlined in the preceding subsection are based on the assumption of a known (possibly up to a set of unknown parameters) statistical model for signals and noise. In many practical situations it is

not possible to specify accurate statistical models for signals or noise, and so it is of interest to design detection procedures that do not rely heavily on such models. Of course, the parametrized models described in the foregoing paragraphs allow for uncertainty in the statistics of the observations. Such models are known as *parametric* models, because the set of possible distributions can be parametrized by a finite set of real parameters.

While parametric models can be used to describe many types of modeling uncertainty, composite models in which the set of possible distributions is much broader than a parametric model would allow are sometimed more realistic in practice. Such models are termed *nonparametric models.* For example, one might be able to assume only some very coarse model for the noise, such as that it is symmetrically distributed. A wide variety of useful and powerful detectors have been developed for signal-detection problems that cannot be parametrized. These are basically of two types: *robust* and *nonparametric.* Robust detectors are those designed to perform well despite small, but potentially damaging, nonparametric deviations from a nominal parametric model, whereas nonparametric detectors are designed to achieve constant false-alarm probability over very wide classes of noise statistics.

Robustness problems are usually treated analytically via minimax formulations that seek best worst-case performance as the design objective. This formulation has proven to be very useful in the design and characterization of robust detectors for a wide variety of detection problems. Solutions typically call for the introduction of light limiting to prevent extremes of gain dictated by an (unrealistic) nominal model. For example, the correlation detector of Fig. 73.1 can be made robust against deviations from the Gaussian noise model by introducing a soft-limiter between the multiplier and the accumulator.

Nonparametric detection is usually based on relatively coarse information about the observations, such as the algebraic signs or the ranks of the observations. One such test is the *sign test,* which bases its decisions on the number of positive observations obtained. This test is nonparametric for the model in which the noise samples are i.i.d. with zero median and is reasonably powerful against alternatives such as the presence of a positive constant signal in such noise. More powerful tests for such problems can be achieved at the expense of complexity by incorporating rank information into the test statistic.

Distributed and Sequential Detection

The detection procedures discussed in the preceding paragraphs are based on the assumption that all measurements can and should be used in the detection of the signal, and moreover that no constraints exist on how measurements can be combined. There are a number of applications, however, in which constraints apply to the information pattern of the measurements.

One type of constrained information pattern that is of interest in a number of applications is a network consisting of a number of distributed or local decision makers, each of which processes a subset of the measurements, and a *fusion center,* which combines the outputs of the distributed decision makers to produce a global detection decision. The communication between the distributed decision makers and the fusion center is constrained, so that each local decision maker must reduce its subset of measurements to a summarizing local decision to be transmitted to the fusion center. Such structures arise in applications such as the testing of large-scale integrated circuits, in which data collection is decentralized, or in detection problems involving very large data sets, in which it is desirable to distribute the computational work of the detection algorithm. Such problems lie in the field of *distributed detection.* Except in some trivial special cases, the constraints imposed by distributing the detection algorithm introduce a further level of difficulty into the design of optimum detection systems. Nevertheless, considerable progress has been made on this problem, a survey of which can be found in Tsitsiklis [1993].

Another type of nonstandard information pattern that arises is that in which the number of measurements is potentially infinite, but in which there is a cost associated with taking each measurement. This type of model arises in applications such as the synchronization of wideband communication signals. In such situations, the error probabilities alone do not completely characterize the performance of a detection system, since consideration must also be given to the cost of sampling. The field of *sequential detection* deals with the optimization of detection systems within such constraints. In sequential detectors, the number of measurements taken becomes a random variable depending on the measurements themselves. A typical performance criterion for optimizing such a system is to seek a detector that minimizes the expected number of measurements for given levels of miss and false-alarm probabilities.

The most commonly used sequential detection procedure is the *sequential probability ratio test,* which operates by recursive comparison of the likelihood ratio (73.3) to two thresholds. In this detector, if the likelihood ratio for a given number of samples exceeds the larger of the two thresholds, then the signal's presence is announced and the test terminates. Alternatively, if the likelihood ratio falls below the smaller of the two thresholds, the signal's absence is announced and the test terminates. However, if neither of the two thresholds is crossed, then another measurement is taken and the test is repeated.

Detection with Continuous-Time Measurements

Note that all of the preceding formulations have involved the assumption of discrete-time (i.e., sampled-data) measurements. From a practical point of view, this is the most natural framework within which to consider these problems, since implementations most often involve digital hardware. However, the procedures discussed in this section all have continuous-time counterparts, which are of both theoretical and practical interest. Mathematically, continuous-time detection problems are more difficult than discrete-time ones, because they involve probabilistic analysis on function spaces. The theory of such problems is quite elegant, and the interested reader is referred to Poor [1994] or Grenander [1981] for more detailed exposition.

Continuous-time models are of primary interest in the front-end stages of radio frequency or optical communication receivers. At radio frequencies, continuous-time versions of the models described in the preceding paragraphs can be used. For example, one may consider the detection of signals in continuous-time Gaussian white noise. At optical wavelengths, one may consider either continuous models (such as Gaussian processes) or point-process models (such as Poisson counting processes), depending on the type of detection used (see, e.g., Snyder and Miller [1991]). In the most fundamental analyses of optical detection problems, it is sometimes desirable to consider the quantum mechanical nature of the measurements [Helstrom, 1976].

Defining Terms

Bayesian detector: A detector that minimizes the average of the false-alarm and miss probabilities, weighted with respect to prior probabilities of signal-absent and signal-present conditions.

Correlation detector: The optimum structure for detecting coherent signals in the presence of additive white Gaussian noise.

Discrete-time white Gaussian noise: Noise samples modeled as independent and identically distributed Gaussian random variables.

Envelope detector: The optimum structure for detecting a modulated sinusoid with random phase in the presence of additive white Gaussian noise.

False-alarm probability: The probability of falsely announcing the presence of a signal.

Likelihood ratio: The optimum processor for reducing a set of signal-detection measurements to a single number for subsequent threshold comparison.

Miss probability: The probability of falsely announcing the absence of a signal.

Neyman-Pearson detector: A detector that minimizes the miss probability within an upper-bound constraint on the false-alarm probability.

Quadratic detector: A detector that makes use of the second-order statistical structure (e.g., the spectral characteristics) of the measurements. The optimum structure for detecting a zero-mean Gaussian signal in the presence of additive Gaussian noise is of this form.

Related Topics

16.2 Parameter Estimation • 70.3 Spread Spectrum Communications

References

U. Grenander, *Abstract Inference*, New York: Wiley, 1981.

C.W. Helstrom, *Quantum Detection and Estimation Theory*, New York: Academic Press, 1976.

S.A. Kassam, *Signal Detection in Non-Gaussian Noise*, New York: Springer-Verlag, 1988.

J.M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and systems theory: Theoretical results and some applications," *Proc. IEEE,* vol. 79, pp. 278–305, 1991.
H.V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed., New York: Springer-Verlag, 1994.
H.V. Poor and J. B. Thomas, "Signal detection in dependent non-Gaussian noise," in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.
J.G. Proakis, *Digital Communications*, New York: McGraw-Hill, 1983.
D.L. Snyder and M.I. Miller, *Random Point Processes in Time and Space*, New York: Springer-Verlag, 1991.
J. Tsitsiklis, "Distributed detection," in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.
S. Verdú, "Multiuser detection," in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.

Further Information

Except as otherwise noted in the accompanying text, further details on the topics introduced in this section can be found in the textbook:
Poor, H.V. *An Introduction to Signal Detection and Estimation,* 2nd ed., New York: Springer-Verlag, 1994.

The bimonthly journal, *IEEE Transactions on Information Theory,* publishes recent advances in the theory of signal detection. It is available from the Institute of Electrical and Electronics Engineers, Inc., 345 East 47th Street, New York, NY 10017.

Papers describing applications of signal detection are published in a number of journals, including the monthly journals *IEEE Transactions on Communications, IEEE Transactions on Signal Processing,* and the *Journal of the Acoustical Society of America.* The IEEE journals are available from the IEEE, as above. The *Journal of the Acoustical Society of America* is available from the American Institute of Physics, 335 East 45th Street, New York, NY 10017.

73.2 Noise

Carl G. Looney

Every information signal $s(t)$ is corrupted to some extent by the superimposition of extra-signal fluctuations that assume unpredictable values at each time instant t. Such undesirable signals were called **noise** due to early measurements with sensitive audio amplifiers.

Noise sources are (1) *intrinsic*, (2) *external*, or (3) *process induced*. Intrinsic noise in conductors comes from thermal agitation of molecularly bound ions and electrons, from microboundaries of impurities and grains with varying potential, and from transistor junction areas that become temporarily depleted of electrons/holes. External electromagnetic interference sources include airport radar, x-rays, power and telephone lines, communications transmissions, gasoline engines and electric motors, computers and other electronic devices; and also include lightning, cosmic rays, plasmas (charged particles) in space, and solar/stellar radiation (conductors act as antennas). Reflective objects and other macroboundaries cause multiple paths of signals. Process-induced errors include measurement, quantization, truncation, and signal generation errors. These also corrupt the signal with noise power and loss of resolution.

Statistics of Noise

Statistics allow us to analyze the spectra of noise. We model a noise signal by a **random** (or *stochastic*) **process** $N(t)$, a function whose realized value $N(t) = x_t$ at any time instant t is chosen by the outcome of the random variable $N_t = N(t)$. $N(t)$ has a probability distribution for the values x it can assume. Any particular trajectory $\{(t, x_t)\}$ of outcomes is called a **realization** of the noise process. The *first-order statistic* of $N(t)$ is the *expected value* $\mu_t = E[N(t)]$. The *second-order statistic* is the *autocorrelation function* $R_{NN}(t, t + \tau) = E[N(t)N(t + \tau)]$, where $E[-]$ is the expected value operator. **Autocorrelation** measures the extent to which noise random variables $N_1 = N(t_1)$ and $N_2 = N(t_2)$ at times t_1 and t_2 depend on each other in an average sense.

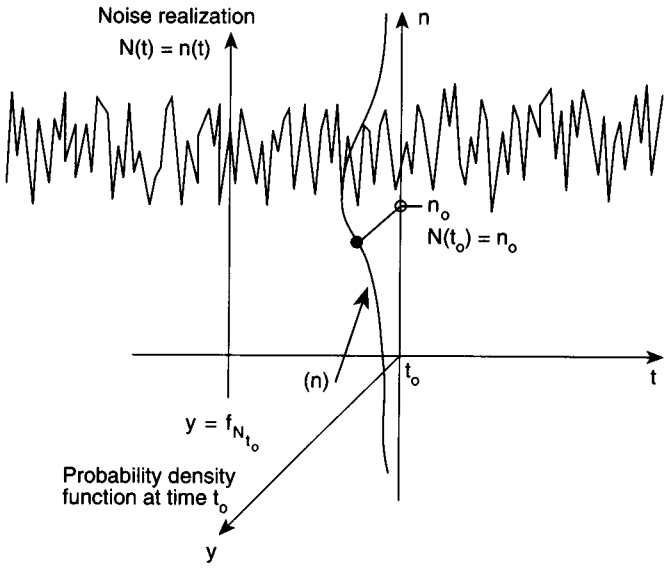

FIGURE 73.3 A noise process.

When the first- and second-order statistics do not change over time, we call the noise a **weakly** (or *wide-sense*) **stationary process**. This means that: (1) $E[N(t)] = \mu_t = \mu$ is constant for all t, and (2) $R_{NN}(t, t+\tau) = E[N(t)N(t+\tau)] = E[N(0)N(\tau)] = R_{NN}(\tau)$ for all t [see Brown, 1983, p. 82; Gardner, 1990, p. 108; or Peebles, 1987, p. 153 for properties of $R_{NN}(\tau)$]. In this case the autocorrelation function depends only on the *offset* τ. We assume hereafter that $\mu = 0$ (we can subtract μ, which does not change the autocorrelation). When $\tau = 0$, $R_{NN}(0) = E[N(t)N(t+0)] = E[(N(t))^2] = \sigma_N^2$, which is the fixed variance of each random variable N_t for all t. Weakly stationary (ws) processes are the most commonly encountered cases and are the ones considered here. *Evolutionary* processes have statistics that change over time and are difficult to analyze.

Figure 73.3 shows a realization of a noise process $N(t)$, where at any particular time t, the probability density function is shown coming out of the page in a third dimension. For a ws noise, the distributions are the same for each t. The most mathematically tractable noises are *Gaussian* ws processes, where at each time t the probability distribution for the random variable $N_t = N(t)$ is Gaussian (also called *normal*). The first- and second-order statistics completely determine Gaussian distributions, and so ws makes their statistics of all orders stationary over time also. It is well known [see Brown, 1983, p. 39] that linear transformations of Gaussian random variables are also Gaussian random variables. The probability density function for a Gaussian random variable N_t is $f_N(x) = \{1/[2\pi\sigma_N{}^2]^{1/2}\exp[-(x-\mu_N)^2/2\sigma_N^2]$, which is the familiar bell-shaped curve centered on $x = \mu_N$. The standard Gaussian probability table [Peebles, 1987, p. 314] is useful, e.g., $\Pr[-\sigma_N < N_t < \sigma_N) = 2\Pr[0 < N_t < \sigma_N) = 0.8413$ from the table.

Noise Power

The noise signal $N(t)$ represents voltage, so the autocorrelation function at offset 0, $R_{NN}(0) = E[N(t)N(t)]$ represents expected power in volts squared, or watts per ohm. When $R = 1\ \Omega$, then $N(t)N(t) = N(t)[N(t)/R] = N(t)I(t)$ volt-amperes = watts (where $I(t)$ is the current in a 1-Ω resistor). The Fourier transform $\mathbf{F}[R_{NN}(\tau)]$ of the autocorrelation function $R_{NN}(\tau)$ is the power spectrum, called the **power spectral density function** (psdf), $S_{NN}(w)$ in W/(rad/s). Then

$$S_{NN}(w) = \int_{-\infty}^{\infty} R_{NN}(\tau)e^{-jws}d\tau = \mathbf{F}[R_{NN}(\tau)]$$
$$R_{NN}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{NN}(w)e^{jws}dw = \mathbf{F}^{-1}[S_{NN}(w)] \tag{73.20}$$

The psdf at frequency f is defined to be the expected power that the voltage $N(t)$, bandlimited to an incremental band df centered at f, would dissipate in a 1-Ω resistance, divided by df.

Equations (73.20) are known as the *Wiener-Khinchin* relations that establish that $S_{NN}(w)$ and $R_{NN}(\tau)$ are a Fourier transform pair for ws random processes [Brown, 1983; Gardner, 1990, p. 230; Peebles, 1987]. The psdf $S_{NN}(w)$ has units of W/(rad/s), whereas the autocorrelation function $R_{NN}(\tau)$ has units of watts. When $\tau = 0$ in the second integral of Eq. (73.20), the exponential becomes $e^0 = 1$, so that $R_{NN}(0)$ $(= E[N(t)^2] = \sigma_N^2)$ is the integral of the psdf $S_{NN}(w)$ over all radian frequencies, $-\infty < w < \infty$. The rms (root-mean-square) voltage is $N_{\text{rms}} = \sigma_N$ (the *standard deviation*). The power spectrum in W/(rad/s) is a density that is summed up via an integral over the radian frequency band w_1 to w_2 to obtain the total power over that band.

$$P_{NN}(w_1, w_2) = \frac{1}{2\pi}\int_{w1}^{w2} S_{NN}(w)\cdot dw \quad \text{watts}$$
$$P_{NN} = \sigma_N^2 = E[N(t)^2] = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{NN}(w)\cdot dw \quad \text{watts} \tag{73.21}$$

The variance $\sigma_N^2 = R_{NN}(0)$ is the mean instantaneous power P_{NN} over all frequencies at any time t.

Effect of Linear Transformations on Autocorrelation and Power Spectral Density

Let $h(t)$ be the impulse response function of a time-invariant linear system L and $H(w) = \mathbf{F}[h(t)]$ be its transfer function. Let an input noise signal $N(t)$ have autocorrelation function $R_{NN}(\tau)$ and psdf $S_{NN}(w)$. We denote the output noise signal by $Y(t) = L[N(t)]$. The Fourier transforms $Y(w) \equiv \mathbf{F}[Y(t)]$ and $N(w) \equiv \mathbf{F}[N(t)]$ do not exist, but they are not needed. The output $Y(t)$ of a linear system is ws whenever the input $N(t)$ is ws [see Gardner, 1990, p. 195; or Peebles, 1987, p. 215]. The output psdf $S_{YY}(w)$ and autocorrelation function $R_{YY}(\tau)$ are given by, respectively,

$$S_{YY}(w) = |H(w)|^2 S_{NN}(w),\ R_{YY}(\tau) = \mathrm{F}^{-1}[S_{YY}(w)] \tag{73.22}$$

[see Gardner, 1990, p. 223]. The output noise power is

$$\sigma_Y^2 = P_{YY} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{YY}(w)dw = \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(w)|^2 S_{NN}(w)dw \tag{73.23}$$

White, Gaussian, and Pink Noise Models

White noise [see Brown, 1983; Gardner, 1990, p. 234; or Peebles, 1987] is a theoretical model $W(t)$ of noise that is ws with zero mean. It has a constant power level n_o over all frequencies (analogous to white light), so its psdf is $S_{WW}(w) = n_o$ W/(rad/s), $-\infty < w < \infty$. The inverse Fourier transform of this is the impulse function $R_{WW}(\tau) = (n_o)\delta(\tau)$, which is zero for all offsets except $\tau = 0$. Therefore, white noise $W(t)$ is a process that is uncorrelated over time, i.e., $E[W(t_1)W(t_2)] = 0$ for t_1 not equal to t_2. Figure 73.4(a) shows the autocorrelation and psdf for white noise where the offset is $s = \tau$. A *Gaussian white noise* is white noise such that the probability distribution of each random variable $W_t = W(t)$ is Gaussian. When two Gaussian random variables W_1 and W_2 are *uncorrelated*, i.e., $E[W_1W_2] = 0$, they are independent [see Gardner, 1990, p. 37]. We use Gaussian models because of the *central limit theorem* that states that the sum of a number of random variables is approximately Gaussian.

Actual circuits attenuate signals above cut-off frequencies, and also the power must be finite. However, for white noise, $P_{WW} = R_{NN}(0) = \infty$, so we often truncate the white noise spectral density (psdf) at cut-offs $-w_c$ to w_c. The result is known as *pink noise*, $P(t)$, and is usually taken to be Gaussian because linear filtering of any white noise (through the effect of the central limit theorem) tends to make the noise Gaussian [see Gardner,

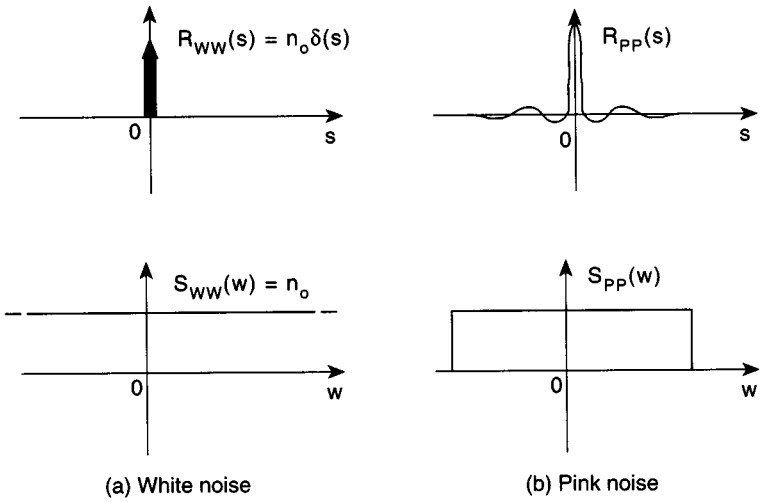

FIGURE 73.4 Power transform pairs for white and pink noise.

Figure 73.5 not available

FIGURE 73.5 Thermal noise in a resistor.

1990, p. 241]. Figure 73.4(b) shows the sinc function $R_{PP}(s) = \mathrm{F}^{-1}[S_{PP}(w)]$ for pink noise. Random variables P_1 and P_2 at times t_1 and t_2 are correlated only for t_1 and t_2 close.

Thermal Noise as Gaussian White Noise

Brown observed in 1828 that pollen and dust particles moved randomly when suspended in liquid. In 1906, Einstein analyzed such motion based on the random walk model. Perrin confirmed in 1908 that the thermal activity of molecules in a liquid caused irregular bombardment of the much larger particles. It was predicted that charges bound to thermally vibrating molecules would generate electromotive force (emf) at the open terminals of a conductor, and that this placed a limit on the sensitivity of galvanometers. Thermal noise (also called *Johnson noise*) was first observed by J. B. Johnson at Bell Laboratories in 1927. Figure 73.5 displays white noise as seen in the laboratory on an oscilloscope.

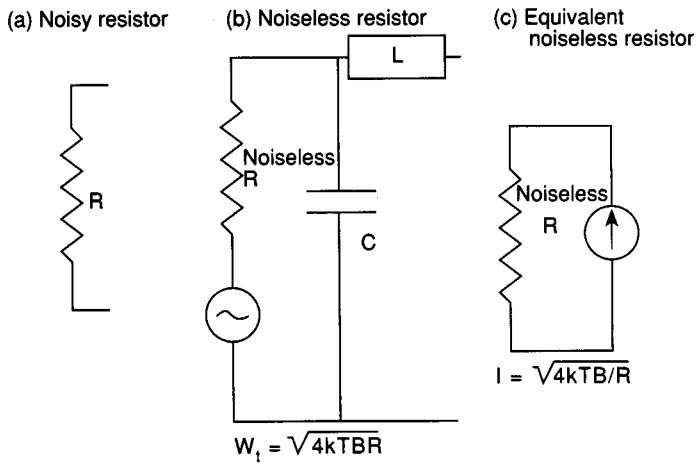

FIGURE 73.6 Thermal noise in a resistor.

The voltage $N(t)$ generated thermally between two points in an open circuit conductor is the sum of an extremely large number of superimposed, independent electronically and ionically induced microvoltages at all frequencies up to f_c = 6,000 GHz at room temperature [see Gardner 1990, p. 235], near infrared. The mean relaxation time of free electrons is $1/f_c = 0.5 \times 10^{-10}/T$ s, so at room temperature of T = 290K, it is 0.17 ps (1 picosecond = 10^{-12} s). The values of $N(t)$ at different times are uncorrelated for time differences (offsets) greater than $\tau_c = 1/f_c$. The expected value of $N(t)$ is zero. The power is fairly constant across a broad spectrum, and we cannot sample signals at picosecond periods, so we model Johnson noise $N(t)$ with Gaussian white noise $W(t)$. Although $\mu = E[W(t)] = 0$, the average power is positive at temperatures above 0K, and is $\sigma_W^2 = R_{WW}(0)$ [see the right side of Eq. (73.21)]. A disadvantage of the white noise model is its infinite power, i.e., $R_{WW}(0) = \sigma_W^2 = \infty$, but it is valid over a limited bandwidth of B Hz, in which case its power is finite.

In 1927, Nyquist [1928] theoretically derived thermal noise power in a resistor to be

$$P_{WW}(B) = 4kTRB \text{ (watts)} \tag{73.24}$$

where R is resistance (ohms), B is the frequency bandwidth of measurement in Hz (all emf fluctuations outside of B are ignored), $P_{WW}(B)$ is the mean power over B (see Eq. 73.21), and Boltzmann's constant is $k = 1.38 \times 10^{-23}$ J/K [see Ott, 1988; Gardner, 1990, p. 288; or Peebles, 1987, p. 227]. Under external emf, the thermally induced collisions are the main source of resistance in conductors (electrons pulled into motion by an external emf at 0K meet no resistance). The rms voltage is $W_{\text{rms}} = \sigma_W = [(4kTRB)]^{1/2}$ V over a bandwidth of B Hz.

Planck's radiation law is $S_{NN}(w) = (2h|f|)/[\exp(h|f|/kT) - 1]$, where $h = 6.63 \times 10^{-34}$ J/s is Planck's constant, and f is the frequency [see Gardner, 1990, p. 234]. For $|f|$ much smaller than $kT/h = 6.04 \times 10^{12}$ Hz $\approx$ 6,000 GHz, the exponential above can be approximated by $\exp(\text{h}|f|/\text{kT}) = 1 + \text{h}|f|/\text{kT}$. The denominator of $S_{NN}(w)$ becomes $h|f|/kT$, so $S_{NN}(w) = (2h|f|)/(h|f|/kT) = 2kT$ W/Hz in a 1-Ω resistor. Over a resistance of R Ω and a bandwidth of B Hz (positive frequencies), this yields the total power $P_{WW}(B) = 2BRS_{NN}(w) = 4kTRB$ W over the two-sided frequency spectrum. This is Nyquist's result.

Thermal noise is the same in a 1000-Ω carbon resistor as it is in a 1000-Ω tantalum thin-film resistor [see Ott, 1988]. While the intrinsic noise may never be less, it may be higher because of other superimposed noise (described in later sections). We model the thermal noise in a resistor by an internal source (generator), as shown in Fig. 73.6. Capacitance cannot be ignored at high f, but pure reactance (C or L) cannot dissipate energy, and so cannot generate thermal noise. The white noise model $W(t)$ for thermal noise $N(t)$ has a constant psdf $S_{WW}(w) = n_o$ W/(rad/s) for $-\infty < w < \infty$. By Eq. 73.21, the white noise mean power over the frequency bandwidth B is

$$P_{WW}(B) = \frac{1}{2\pi}\int_{-2\pi B}^{2\pi B} S_{WW}(w)dw = n_o(4\pi B/2\pi) = 2n_oB \tag{73.25}$$

Solving for the constant n_o, we obtain $n_o = P_{WW}(B)/2B$, which we put into Eq. (73.20) to get the spectral density as a function of temperature and resistance using Nyquist's result above.

$$S_{WW}(w) = n_o = P_{WW}(B)/4\pi B = 4kTR2\pi B/4\pi B = 2kTR \text{ watts/(rad/s)} \qquad (73.26)$$

Some Examples

The parasitic capacitance in the terminals of a resistor may cause a roll-off of about 20 dB/octave in actual resistors [Brown, 1983, p. 139]. At 290K (room temperature), we have $2kT = 2 \times 1.38 \times 10^{-23} \times 290 = 0.8 \times 10^{-20}$ W/Hz due to each ohm [see Ott, 1988]. For $R = 1$ MΩ (10^6 Ω), $S_{WW}(w) = 0.8 \times 10^{-14}$. Over a band of 10^8 Hz, we have $P_{WW}(B) = S_{WW}(w)B = 0.8 \times 10^{-14} \times 10^8 = 0.8 \times 10^{-6}$ W = 0.8 μW by Eqs. (73.24) and (73.26). In practice, parasitic capacitance causes thermal noise to be bandlimited (pink noise). Now consider Fig. 73.6(b) and let the temperature be 300K, $R = 10^6$ Ω, $C = 1$ pf (1 picofarad = 10^{-12} farads), and assume L is 0H. By Eq. (73.26), the thermal noise power is

$$S_{WW}(w) = 2kTR = 2 \times 1.38 \times 10^{-23} \times 300 \times 10^6 = 828 \times 10^{-17} \text{ W/Hz}$$

The power across a bandwidth $B = 10^6$ is $P_{WW}(B) = S_{WW}(w)B = 8280 \times 10^{-12}$ W, so the rms voltage is $W_{\text{rms}} = [P_{WW}(B)]^{1/2} = 91$ μV.

Now let $Y(t)$ be the output voltage across the capacitor. The transfer function can be seen to be $H(w) = \{I(w)(1/jwC)\}/\{I(w)[R + (1/jwC)]\} = (1/jwC)/[R + 1/jwC] = 1/[1 + jwRC]$ (where $I(w)$ is the Fourier transform of the current). The output psdf [see Eq. (73.22)] is

$$S_{YY}(w) = |H(w)|^2 S_{WW}(w) = (1/[1 + w^2R^2C^2])S_{WW}(w)$$

Integrating $S_{YY}(w) = (1/[1 + w^2R^2C^2])S_{WW}(w)$ over all radian frequencies $w = 2\pi f$ [see Eq. (73.21)], we obtain the antiderivative $(828 \times 10^{-17})(1/RC)\text{atan}(RCw)/2\pi$. Upon substituting the limits $w = \pm\infty$, this becomes $828 \times 10^{-17}[\pi/2 + \pi/2]/2\pi RC = 414 \times 10^{-17}(1/2\text{RC}) = 207 \times 10^{-17} \times 10^6 = 2070 \times 10^{-12}$ W/Hz. Then $\sigma_Y^2 = E[Y(t)^2] = P_{YY}(-\infty,\infty) = 2070 \times 10^{-12}$ W, so $Y_{\text{rms}}(\text{t}) = \sigma_Y = [P_{YY}(-\infty,\infty)]^{1/2} = 45.5$ μV. The half-power (cut-off) radian frequency is $w_c = 1/RC = 10^6$ rad/s, or $f_c = w_c/2\pi = 159.2$ kHz. Approximating $S_{YY}(w)$ by the rectangular spectrum $S_{YY}(w) = n_o, -10^6 < w < 10^6$ rad/s (0 elsewhere), we have that $R_{YY}(\tau) = (w_c/\pi)\text{sinc}(w_c\tau)$, which has the first zeros at $|w_c\tau| = \pi$, that is $|\tau| = 1/(2f_c)$ [see Fig. 73.4(b)]. We approximate the autocorrelation by $R_{YY}(\tau) = 0$ for $|s| \geq 1/2f_c$.

Measuring Thermal Noise

In Fig. 73.7, the thermal noise from a noisy resistor R is to be measured, where R_L is the measurement load. The incremental noise power in R over an incremental frequency band of width df is $P_{WW}(df) = 4kTRdf$ W, by Eq. (73.24). $P_{YY}(df)$ is the integral of $S_{YY}(w)$ over df by Eqs. (73.21), where $S_{YY}(w) = |H(w)|^2 S_{WW}(w)$, by Eq. (73.22). In this case, the transfer function $H(w)$ is nonreactive and does not depend upon the radian frequency (we can factor it out of the integral). Thus,

$$P_{YY}(df) = \int_{-df}^{df} |H(f)|^2 \, (2kTR)df = \{R_L/(R + R_L)^2\}(4kTRdf)$$

To maximize the power measured, let $R_L = R$. The *incremental available power* measured is then $P_{YY}(df) = 4kTR^2df/(4R^2) = kTdf$ [see Ott, 1988, p. 201; Gardner, 1990, p. 288; or Peebles, 1987, p. 227]. Thus, we have the result that incremental available power over bandwidth df depends only on the temperature T.

$$P_{YY}(df) = kTdf \qquad (\text{output power over } df) \qquad (73.27)$$

Albert Einstein used statistical mechanics in 1906 to postulate that the mean kinetic energy per degree of freedom of a particle, $(1/2)mE[v^2(t)]$, is equal to $(1/2)kT$, where m is the mass of the particle, $v(t)$ is its

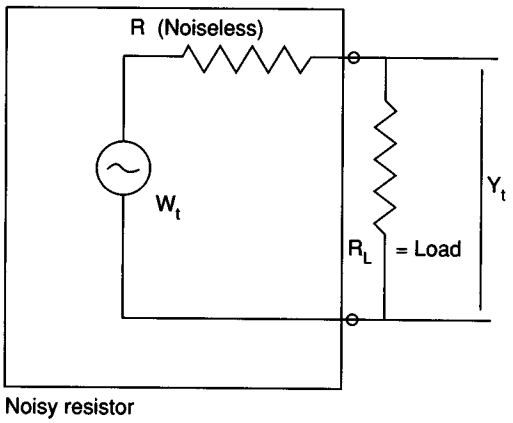

FIGURE 73.7 Measuring thermal noise voltage.

instantaneous velocity in a single dimension, k is Boltzmann's constant, and T is the temperature in kelvin. A shunt capacitor C is charged by the thermal noise in the resistor [see Fig. 73.6(b), where L is taken to be zero]. The average potential energy stored is $(1/2)CE[W(t)^2]$. Equating this to $1/2kT$ and solving, we obtain the mean square power

$$E[W(t)^2] = kT/C \tag{73.28}$$

For example, let $T = 300$K and $C = 50$ pf, and recall that $k = 1.38 \times 10^{-23}$ J/K. Then $E[W(t)^2] = kT/C = 82.8 \times 10^{-12}$, so that the input rms voltage is $\{E[W(t)^2]\}^{1/2} = 9.09$ μV.

Effective Noise and Antenna Noise

Let two series resistors R_1 and R_2 have respective temperatures of T_1 and T_2. The total noise power over an incremental frequency band df is $P_{\text{Total}}(df) = P_{11}(df) + P_{22}(df) = 4kT_1R_1\,df + 4kT_2R_2\,df = 4k(T_1R_1 + T_2R_2)df$. By putting

$$T_E = (T_1R_1 + T_2R_2)/(R_1 + R_2) \tag{73.29}$$

we can write $P_{\text{Total}}(df) = 4kT_E(R_1 + R_2)df$. T_E is called the *effective noise temperature* [see Gardner, 1990, p. 289; or Peebles, 1987, p. 228]. An antenna receives noise from various sources of electromagnetic radiation, such as radio transmissions and harmonics, switching equipment (such as computers, electrical motor controllers), thermal (blackbody) radiation of the atmosphere and other matter, solar radiation, stellar radiation, and galaxial radiation (the ambient noise of the universe). To account for noise at the antenna output, we model the noise with an equivalent thermal noise using an effective noise temperature T_E. The incremental available power (output) over an incremental frequency band df is $P_{YY}(df) = kT_E\,df$, from Eq. (73.27). T_E is often called *antenna temperature*, denoted by T_A. Although it varies with the frequency band, it is usually virtually constant over a small bandwidth.

Noise Factor and Noise Ratio

In reference to Fig. 73.8(a), we define the *noise factor* F = (*noise power output of actual device*)/(*noise power output of ideal device*), where (noise power output of ideal device) = (power output due to thermal noise source). The noise source is taken to be a noisy resistor R at a temperature T, and all output noise measurements must be taken over a resistive load R_L (reactance is ignored). Letting $P_{WW}(B) = 4kTRB$ be the open circuit thermal noise power of the source resistor over a frequency bandwidth B, and noting that the gain of the device is G, the output power due to the resistive noise source becomes $G^2P_{WW}(B) = 4kTRBG^2/R_L$. Now let $Y(t)$ be the output voltage measured at the output across R_L. Then the noise factor is

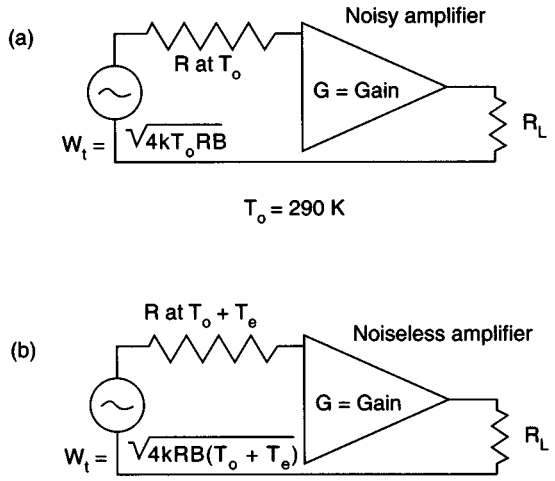

FIGURE 73.8 Equivalent input noise and noise factor.

$$F = (P_{YY}(B)/R_L)/(G^2P_{WW}(B)/R_L) = (P_{YY}(B))/(4kTRBG^2) \tag{73.30}$$

F is seen to be independent of R_L, but not R. To compare two noise factors, the same source must be used. In the ideal noiseless case, $F = 1$, but as the noise level in the device increases, F increases. Because this is a power ratio, we may take the logarithm, called the *noise ratio*, which is

$$N_F = 10\log_{10}(F) = 10\log_{10}(P_{YY}(B)) - 10\log_{10}(4kTRBG^2) \tag{73.31}$$

The noise power output $P_{YY}(B)$ of an actual device is a superposition of the amplified source thermal noise $G^2P_{WW}(B)$ and the device noise, i.e., $P_{YY}(B) = G^2P_{WW}(B)$ + (device noise). The output noise across R_L can be measured by putting a single frequency (in the passband) source generator $S(t)$ as input. First, $S(t)$ is turned off, and the output rms voltage $Y(t)$ is measured and the output power $P_{Y(W)}(B)$ is recorded. This is the sum of the thermal available power and the device noise. Next, $S(t)$ is turned on and adjusted until the output power doubles, i.e., until the output power $P_{Y(W)}(B) + P_{Y(S)}(B) = 2P_{Y(W)}(B)$. This $P_{SS}(B)$ is recorded. Solving for $P_{Y(S)}(B) = P_{Y(W)}(B)$, we substitute this in $F = P_{Y(W)}(B)/(G^2P_{WW}(B))$ to obtain

$$F = P_{Y(S)}(B)/(G^2 \cdot P_{WW}(B)) = (G^2P_{SS}(B))/(G^24kTRB) = P_{SS}(B)/4kTRB \tag{73.32}$$

A better way is to input white noise $W(t)$ in place of $S(t)$ (a noise diode may be used). The disadvantages of noise factors are (1) when the device has low noise relative to thermal noise, the noise factor has value close to 1; (2) a low resistance causes high values; and (3) increasing the source resistance decreases the noise factor while increasing the total noise in the circuit [Ott, 1988, p. 216]. Thus, accuracy is not good. For cascaded devices, the noise factors can be conveniently computed [see Buckingham, 1985, p. 67; or Ott, 1988, p. 228].

Equivalent Input Noise

Shot noise (see below) and other noise can be modeled by equivalent thermal noise that would be generated in an input resistor by increased temperature. Recall that the (maximum) incremental available power (output) in a frequency bandwidth df is $P_{WW}(df) = kTdf$ from Eq. (73.27). Figure 73.8(b) presents the situation. Let the resistor be the noise source at temperature T_o with thermal noise $W(t)$. Then $E[W(t)^2] = 4kT_oRdf$, by Eq. (73.24) (Nyquist's result). Let the open circuit output noise power at R_L be $E[Y(t)^2]$. The incremental available noise power $P_{YY}(df)$ at the output ($R_L = R$) can be considered to be due to the resistor R having a higher temperature and an ideal (noiseless) device, usually an amplifier. We must find a temperature T_e at which a pseudothermal

noise power $E[W_e(t)^2] = 4kT_eRdf$ yields the extra "input" noise power. Let $V(t) = W(t) + W_e(t)$. Then $P_{VV}(df) = 4kT_oRdf + 4kT_eRdf = 4k(T_o + T_e)Rdf$ W, from Eq. (73.24). T_e is called the *equivalent input noise temperature*. It is related to the noise factor F by $T_e = 290(F - 1)$. In cascaded amplifiers with gains $G_1, G_2, \ldots$ and equivalent input noise temperatures $T_{e1}, T_{e2}, \ldots$, the total equivalent input noise temperature is

$$T_{e(\text{Total})} = T_{e1} + T_{e2}/G_1 + T_{e3}/G_1G_2 + \ldots \tag{73.33}$$

[see Gardner, 1990, p. 289].

Other Electrical Noise

Thermal noise and shot noise (which can be modeled by thermal noise with equivalent input noise) are the main noise sources. Other noises are discussed in the following paragraphs.

Shot Noise

In a conductor under an external emf, there is an average flow of electrons, holes, photons, etc. In addition to this induced net flow and thermal noise, there is another effect. The potential differs across the boundaries of metallic grains and particles of impurities, and when the kinetic energy of electrons exceeds this potential, electrons jump across the barrier. This summed random flow is known as *shot noise* [see Gardner, 1990, p. 239; Ott, 1988, p. 208]. The shot effect was analyzed by Schottky in 1918 as $I_{\text{sh}} = (2qI_{\text{dc}}B)^{1/2}$, where $q = 1.6 \times 10^{-19}$ coulombs per electron, I_{dc} = average dc current in amperes, and B = noise bandwidth (Hz).

Partition Noise

Partition noise is caused by a parting of the flow of electrons to different electrodes into streams of randomly varying density. Suppose that electrons from some source S flow to destination electrodes A and B. Let $n(A)$ and $n(B)$ be the average numbers of electrons per second that go to nodes A and B respectively, so that $n(S) = n(A) + n(B)$ is the average total number of electrons emitted per second. It is a success when an electron goes to A, and the probability of success on a single trial is p, where

$$p = n(A)/n(S),\ 1 - p = n(B)/n(S) \tag{73.34}$$

The current to the respective destinations is $I(A) = n(A)q$, $I(B) = n(B)q$, where q is the charge of an electron, so that $I(A)/I(S) = p$ and $I(B)/I(S) = 1 - p$. Using the binomial model, the average numbers of successes are $E[n(A)] = n(S)p$ and $E[n(B)] = n(S)(1 - p)$. The variance is $\text{Var}(n(A)) = n(S)p(1 - p) = \text{Var}(n(B))$ (from the binomial formula for variance). Therefore, substitution yields

$$\text{Var}(I(A)) = q^2[n(S)p(1 - p)] = q^2n(S)\{I(A)I(B)/[I(A) + I(B)]\} \tag{73.35}$$

Partition noise applies to pentodes, where the source is the cathode, A is the anode (success), and B is the grid. For transistors, the source is the emitter, A is the collector, and B represents recombination in the base. In photo devices, a photoelectron is absorbed, and either an electron is emitted (a success) or not. Even a partially silvered mirror can be considered to be a partitioner: the passing of a photon is a success and reflection is a failure. While the binomial model applies to partitions with destinations A and B, multinomial models are analogous for more than two destinations.

Flicker, Contact, and Burst Noise

J.B. Johnson first noticed in 1925 that noise across thermionic gates exceeded the expected shot noise at lower frequencies. It is most noticeable up to about 2 kHz. The psdf of the extra noise, called *flicker noise*, is

$$S(f) = I^2/\alpha f,\ f > 0 \tag{73.36}$$

where I is the dc current flowing through the device and f is the positive frequency. Empirical values of α are about 1 to 1.6 for different sources. These sources vary but include the irregularity of the size of macro regions

of the cathode surface, impurities in the conducting channel, and generation and recombination noise in transistors. In the early days of transistors, this generation-recombination was of great concern because the materials were not of high purity. Flicker noise occurs in thin layers of metallic or semiconducting material, solid state devices, carbon resistors, and vacuum tubes [see Buckingham, 1985, p. 143]. It includes *contact noise* because it is caused by fluctuating conductivity due to imperfect contact between two surfaces, especially in switches and relays. Flicker noise may be high at low frequencies.

Burst noise is also called *popcorn noise:* audio amplifiers sound like popcorn popping in a frying pan background (thermal noise). Its characteristic is $1/f^n$ (usually $n = 2$), so its power density falls off rapidly, where f is frequency. It may be problematic at low frequencies. The cause is manufacturing defects in the junction of transistors (usually a metallic impurity).

Barkhousen and Other Noise

Barkhousen noise is due to the variations in size and orientation of small regions of ferromagnetic material and is especially noticeable in the steeply rising region of the hysteresis loop. There is also secondary emission, photo and collision ionization, etc.

Measurement and Quantization Noise

Measurement Error

The measurement X_t of a signal $X(t)$ at any t results in a measured value $X_t = x$ that contains error, and so is not equal to the true value $X_t = x_T$. The probability is higher that the magnitude of $e = (x - x_T)$ is closer to zero. The bell-shaped Gaussian probability density $f(e) = [1/(2\pi\sigma^2]^{1/2}\exp(-e^2/2\pi\sigma)$ fits the error well. This noise process is stationary over time. The expected value is $\mu_e = 0$, the mean-square error is σ_e^2, and the rms error is σ_e. Its instantaneous power at time t is σ_e^2. To see this, the error signal $e(t) = (x - x_T)$ has instantaneous power per Ω of

$$P_i = e(t)\,i(t) = e(t)[\,e(t)/R] = e^2(t) \tag{73.37}$$

where $R = 1\ \Omega$ and $i(t)$ is the current. The average power is the summed instantaneous power over a period of time T, divided by the time, taken in the limit as $T \to \infty$, i.e.,

$$P_{\text{ave}} = \lim_{T\to\infty}(1/T)\int_0^T e^2(t)dt$$

This average power can be determined by sampling on known signal values and then computing the sample variance (assuming ergodicity: see Gardner [1990, p. 163]). The error and signal are probabilistically independent (unless the error depends on the values of X). The signal-to-noise power ratio is computed by $S/N = P_{\text{signal}}/P_{\text{ave}}$.

Quantization Noise

Quantization noise is due to the digitization of an exact signal value $v_t = v(t)$ captured at sampling time t by an A/D converter. The binary representation is $b_{n-1}b_{n-2}\ldots b_1b_0$ (an n-bit word). The n-bit digitization has 2^n different values possible, from 0 to 2^n–1. Let the voltage range be R. The *resolution* is $dv = R/2^n$. Any voltage v_t is coded into the nearest lower binary value x_b, where the error $e = x_t - x_b$ satisfies $0 \le e \le dv$. Thus, the errors e are distributed over the interval $[0, dv]$ in an equally likely fashion that implies the uniform distribution on $[0, dv]$. The expected value of $e = e_t = e(t)$ at any time is $\mu_e = dv/2$, and the variance is $\mu_e^2 = dv^2/12$ (the variance of a uniform distribution on an interval $[a,b]$ is $\sigma = (b - a)^2/12$). Thus the noise is ws and the power of quantization noise is

$$\begin{aligned}\sigma_e^2 &= \int_0^{dv}(e - dv/2)^2(1/dv)de \\ &= (e - dv/2)^3/3dv\,\Big|_0^{dv} = [(dv)^3 + (dv)^3]/24dv = dv^2/12\end{aligned} \tag{73.38}$$

We can find the signal-to-noise voltage ratio for the total range R via $R/(dv/(12)^{1/2}) = 2^n dv/(dv/(12)^{1/2}) = 2^n(12)^{1/2}$. The power ratio is the square of this, which is $(2^{2n})(12)$. In decibels this becomes $(S/N)_{\text{dB}} = 10\log_{10}(2^{2n}\cdot 12) = 10\log_{10}(12) + 20n\log_{10}(2) = 10.8 + 6.02n$. Thus, quantization S/N power ratio depends directly upon the number of bits n in that the higher S/N power ratio is better, just as we would have expected.

Coping with Noise

External interference is ubiquitous. Intrinsic noise is present up to the incremental available power at temperatures above absolute zero, and other intrinsic noises depend on material purity and connection integrity. Processing error is always introduced in some form.

External Sources

Standard defenses are (1) shielding of lines and circuits, (2) twisted wire pairs or coaxial cables, (3) short lines and leads, (4) digital regeneration at waypoints of digital signals, (5) narrowband signals, (6) correlation of received signals with multipaths, and (7) adaptive notch filtering to eliminate interference at known frequencies; e.g., the second harmonic of 60-Hz ac power lines may interfere with biological microvoltage measurements but could be eliminated via adaptive notch filtering. Ferrite beads can dampen interference [Barnes, 1987]. Digital signal processing, spectral shaping filters [see Brown, 1983], and frequency-shift filters [see Gardner, 1990, p. 400] can be used to lower noise power. Kalman filtering is a powerful estimation method, and frequency-shift filtering is a newer technique for discriminating against both measurement error (e.g., in system identification applications) and extrinsic sources of both noise and interference [Gardner, 1990, p. 400].

Intrinsic Sources

Strategies for minimizing intrinsic noise are (a) small bandwidth B, (b) small resistances R, (c) low temperature T (higher temperatures can be devastating), (d) low voltage and currents (CMOS transistors), (e) modern materials of high purity, (f) wrapped wire resistors (thermal noise is the same, but other noise will be less), (g) fewer and better connections (of gold), (h) smaller circuits of lower power, and (i) shunt capacitors to reduce noise bandwidth. Greater purity of integrated circuit materials nowadays essentially reduces intrinsic noise to thermal noise. Better design and materials are the keys to lower noise.

Processing Sources

Processing errors can be reduced by using higher resolution of analog-to-digital converters, i.e., more bits to represent each value. This lowers the quantization error power. Measurement error can be reduced while using the same instruments by taking multiple measurements and averaging. Other estimation/correlation can yield better values (e.g., the Global Positioning System location determination can be reduced from meters to a few centimeters by multiple measurement estimation).

Defining Terms

Autocorrelation: A function associated with a random signal $X(t)$ that is defined on pairs of time instants t_1 and t_2 and whose value is the expected value of the product of the random variables $X(t_1)$ and $X(t_2)$, i.e., $R_{XX}(t_1,t_2) = E[X(t_1)X(t_2)]$. For weakly stationary random signals, it depends only on the offset $\tau = t_2 - t_1$, so we write $R_{XX}(\tau) = E[X(t)X(t+\tau)]$.

Noise: A signal $N(t)$ whose value at any time t is randomly selected by events beyond our control. At any time instant t, $N(t)$ is a random variable N_t with a probability distribution that determines the relative frequencies at which N_t assumes values. The statistics of the family of random variables $\{N_t\}$ may be constant (stationary) over time (the usual case) or may vary.

Power spectral density: The Fourier transform of the power $X^2(t)$ does not necessarily exist, but it does for $X_T^2(t)/2T$ ($X_T(t) = 0$ for $|t| > T$, $= X(t)$ elsewhere), for any $T > 0$. Letting $T \to \infty$, the expected value of the Fourier transforms $E[\mathrm{F}[X_T^2(t)/2T] = \mathrm{F}[E[X_T^2(t)]/2T$ goes to the limit of the average power in $X(t)$ over $-T$ to T, known as the power spectral density function $S_{xx\,(w)}$. Summed up over all frequencies, it gives the total power in the signal $X(t)$.

Random process: (signal): A signal that is either a noise, an interfering signal $s(t)$, or a sum of these such as $X(t) = s_1(t) + \ldots + s_m(t) + N_1(t) + \ldots + N_n(t)$.

Realization: A trajectory $\{(t, x_t)\colon X(t) = x_t\}$ determined by the actual outcomes $\{x_t\}$ of values from a random signal $X(t)$, where $X(t) = x_t$ at each instant t. A trajectory is also called a *sample function* of $X(t)$.

Weakly stationary (ws) random process (signal): A random signal whose first- and second-order statistics remain stationary (fixed) over time.

Related Topic

15.2 Speech Enhancement and Noise Reduction

References

J. R. Barnes, *Electronic System Design: Interference and Noise Control,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering,* New York: Wiley, 1983.

M. J. Buckingham, *Noise in Electronic Devices and Systems,* New York: Halstead Press, 1985.

W. A. Gardner, *Introduction to Random Processes,* 2nd ed., New York: McGraw-Hill, 1990.

J. B. Johnson, "Thermal agitation of electricity in conductors," *Phys. Rev.,* vol. 29, pp. 367–368, 1927.

J. B. Johnson, "Thermal agitation of electricity in conductors," *Phys. Rev.,* vol. 32, pp. 97–109, 1928.

H. Nyquist, "Thermal agitation of electric charge in conductors," *Phys. Rev.,* vol. 32, pp. 110–113, 1928.

H. W. Ott, *Noise Reduction Techniques in Electronic Systems,* 2nd ed., New York: Wiley-Interscience, 1988.

P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles,* 2nd ed., New York: McGraw-Hill, 1987.

Further Information

The IEEE Individual Learning Program, *Random Signal Analysis with Random Processes and Kalman Filtering,* prepared by Carl G. Looney (IEEE Educational Activities Board, PO Box 1331, Piscataway, NJ 08855-1331, 1989) contains a gentle introduction to estimation and Kalman filtering.

Also see H. M. Denny, *Getting Rid of Interference,* IEEE Video Conference, Educational Activities Dept., Piscataway, NJ, 08855-1331, 1992.

73.3 Stochastic Processes

Carl G. Looney

Introduction to Random Variables

A *random variable* (rv) A is specified by its *probability density function* (pdf)

$$f_A(a) = \lim_{\varepsilon\to 0}\ (1/\varepsilon)P[a - (\varepsilon/2) < A \le a + (\varepsilon/2)]$$

In other words, the rectangular area $\varepsilon \cdot f_A(a)$ approximates the probability $P[(A \le a + (\varepsilon/2)] - P[a - (\varepsilon/2) < A]$. The joint pdf of two rv's A and B is specified by

$$f_{AB}(a,b) = \lim_{\varepsilon\to 0}\ (1/\varepsilon^2)P[\,a - \varepsilon < A \le a + (\varepsilon/2) \text{ and } b - \varepsilon < B \le b + (\varepsilon/2)]$$

A similar definition holds for any finite number of rv's.

The *expected value* $E[A]$, or *mean* μ_A, of a rv A is the first moment of the pdf, and the *variance* of A is the second centralized moment, defined respectively by

$$\mu_A = E[A] \equiv \int_{-\infty}^{\infty} a f_A(a)da \tag{73.39a}$$

$$\sigma_A^2 = E[(A-\mu_A)^2] \equiv \int_{-\infty}^{\infty} (a-\mu_A)^2 f_A(a)da \tag{73.39b}$$

The square root of the variance is the *standard deviation*, which is also called the *root mean square* (rms) error. The *covariance* of two rv's A and B is the second-order centralized joint moment

$$\sigma_{AB} = E[(A-\mu_A)(B-\mu_B)] \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a-\mu_A)(b-\mu_B) f_{AB}(a,b)dadb \tag{73.40}$$

The noncentralized second moments are the *mean-square value* and the *correlation*, respectively,

$$E[A^2] = \int_{-\infty}^{\infty} a^2 f_A(a)da = \sigma_A^2 + \mu_A^2, \quad E[AB] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} abf_{AB}(a,b)dadb = \sigma_{AB} + \mu_A\mu_B$$

A set of rv's A, B, and C is defined to be *independent* whenever their joint pdf factors as

$$f_{ABC}(a,b,c) = f_A(a)f_B(b)f_C(c) \tag{73.41}$$

for all a, b, and c, and similarly for any finite set of rv's. A *weak independence* holds when the second moment of the joint pdf, the correlation, factors as $E[AB] = E[A]E[B]$, so that $\sigma_{AB} = 0$, in which case the rv's are said to be *uncorrelated*. The covariance of A and B is a measure of how often A and B vary together (have the same sign), how often they vary oppositely (different signs), and by how much, on the average over trials of outcomes. To standardize so that units do not influence the measure of dependence, we use the *correlation coefficient*

$$\rho_{AB} \equiv \sigma_{AB}/\sigma_A\sigma_B$$

The accuracy of approximating a rv A as a linear function of another rv B, $A \approx cB + d$, for real coefficients c and d, is found by minimizing the mean-square error $\varepsilon = E\{[A - (cB + d)]^2\}$. Upon squaring and taking the expected values, we can obtain $\varepsilon_{\min} = \sigma_A^2(1 - |\rho_{AB}|^2)$, which shows $|\rho_{AB}|$ to be a measure of the degree of linear relationship between A and B. Because $\varepsilon_{\min} \geq 0$, this shows that $|\rho_{AB}| \leq 1$, which demonstrates the Cauchy-Schwarz inequality

$$|E[AB]| \leq \{E[A^2]E[B^2]\}^{1/2} \tag{73.42}$$

When $|\rho_{AB}| = 1$, then knowledge of one of A or B completely determines the other ($c \neq 0$), and so A and B are completely dependent, while $|\rho_{AB}| = 0$ indicates there is no linear relationship, i.e., that A and B are uncorrelated.

An important result is the *fundamental theorem of expectation*: if $g(\cdot)$ is any real function, then the expected value of the rv $B = g(A)$ is given by

$$E[B] = E[g(A)] = \int_{-\infty}^{\infty} g(a)f_A(a)da \tag{73.43}$$

Stochastic Processes

A **stochastic** (or *random*) **process** is a collection of random variables $\{X_t : t \in T\}$, indexed on an ordered set T that is usually a subset of the real numbers or integers. Examples are the Dow-Jones averages $D(t)$ at each time t, the pressure $R(x)$ in a pipe at distance x, or a noise voltage $N(t)$ at time t. A process is thus a *random function* $X(t)$ of t whose value at each t is drawn randomly from a range of outcomes for the rv $X_t = X(t)$ according to a probability distribution for X_t. A trajectory $\{x_t : t \in T\}$ of outcomes over all $t \in T$, where $X_t = x_t$ is the realized value at each t, is called a **sample function** (or *realization*) of the process. A stochastic process $X(t)$ has mean

TABLE 73.1 Continuous/Discrete Classification of Stochastic Processes

	X Values	
T Values	Continuous	Discrete
Continuous	Continuous stochastic processes	Discrete valued stochastic processes
Discrete	Continuous random sequences	Discrete valued random sequences

value $E[X(t)] = \mu(t)$ at time t, and **autocorrelation function** $R_{XX}(t, t+\tau) = E[X(t)X(t+\tau)]$ at times t and $t+\tau$, the correlation of two rv's at two times offset by τ. When $\mu(t) = 0$ for all t, the autocorrelation function equals the *autocovariance function* $C_{XX}(t, t+\tau) = E[(X(t) - \mu(t))(X(t+\tau) - \mu(t+\tau))]$.

A process $X(t)$ is completely determined by its joint pdf's $f_{X(t(1)) \ldots X(t(n))}(x(t_1), \ldots, x(t_n))$ for all time combinations $t_1, \ldots, t_n$ and all positive integers n (where $t(j) = t_j$). When the rv's $X(t)$ are *iid* (independent, identically distributed), then knowledge of one pdf yields the knowledge of all joint pdf's. This is because we can construct the joint pdf by factorization, per Eq. (73.41).

Classifications of Stochastic Processes

The ordered set T can be continuous or discrete, and the values that $X(t)$ assumes at each t may also be continuous or discrete, as shown in Table 73.1.

In another classification, a stochastic process $X(t)$ is *deterministic* whenever an entire sample function can be determined from an initial segment $\{x;\ t \le t_1\}$ of $X(t)$. Otherwise, it is *nondeterministic* [see Brown, 1983, p. 79; or Gardner, 1990, p. 304].

Stationarity of Processes

A stochastic process is *nth order* (*strongly*) *stationary* whenever all joint pdf's of n and fewer rv's are independent of all translations of times $t_1, \ldots, t_n$ to times $\tau + t_1, \ldots, \tau + t_n$. The case of $n = 2$ is very useful. Another type of process is called **weakly stationary** (ws), or *wide-sense stationary*, and is defined to have first- and second-order moments that are independent of time (see Section 73.2 on noise). These satisfy (1) $\mu(t) = \mu$ (constant) for all t, and (2) $R_{XX}(t, t+\tau) = R_{XX}(t+s, t+s+\tau)$ for all values of s. For $s = -t$, this yields $R_{XX}(t, t+\tau) = R_{XX}(0, 0+\tau)$, which is abbreviated to $R_{XX}(\tau)$. $X(t)$ is *uncorrelated* whenever $C_{XX}(\tau) = 0$ for τ not zero [we say $X(t)$ has *no memory*]. If $X(t)$ is correlated, then $X(t_1)$ depends on values $X(t)$ for $t \ne t_1$ [$X(t)$ has *memory*].

Some properties of autocorrelation functions for ws processes follow. First, $|R_{XX}(\tau)| \le R_{XX}(0)$, $-\infty < \tau < \infty$, as can be seen from Eq. (73.42) with $|R_{XX}(\tau)|^2 = E[X(0)X(\tau)] \le E[X(0)^2]E[X(\tau)^2] = R_{XX}(0)R_{XX}(\tau)$. Next, $R_{XX}(\tau)$ is real and even, i.e., $R_{XX}(-\tau) = R_{XX}(\tau)$, which is evident from substituting $s = t - \tau$ in $E[X(s)X(s+\tau)]$and using time independence. If $X(t)$ has a periodic component, then $R_{XX}(\tau)$ will have that same periodic component, which follows from the definition. Finally, if $X(t)$ has a nonzero mean μ and no periodic components, then the variance goes to zero (the memory fades) and so $\lim_{\tau\to\infty} R_{XX}(\tau) \to 0 + \mu^2 = \mu^2$.

Gaussian and Markov Processes

A process $X(t)$ is defined to be *Gaussian* if for every possible finite set $\{t_1, \ldots, t_n\}$ of times, the rv's $X(t_1), \ldots, X(t_n)$ are *jointly Gaussian*, which means that every linear combination $Z = a_1X(t_1) + \ldots + a_nX(t_n)$ is a Gaussian rv, defined by the Gaussian pdf

$$f_Z(z) = \left[1/(\sigma_Z\sqrt{2\pi})\right] \exp\{-(z - \mu_Z)^2/2\sigma_Z^2\} \tag{73.44}$$

In case the n rv's are *linearly independent*, i.e., $Z = 0$ only if $a_1 = \cdots = a_n = 0$, the joint pdf has the Gaussian form [see Gardner, 1990, pp. 39–40]

$$f_{X(t(1))\ldots X(t(n))}(x_1, \ldots, x_n) = [1/(2\pi)^{n/2}|C|^{1/2}] \cdot \exp\{-(\mathrm{x} - \boldsymbol{\mu})^t C^{-1}(\mathrm{x} - \boldsymbol{\mu})\} \quad (73.45)$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is a column vector, $\mathbf{x}^t$ is its transpose, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is the mean vector, C is the *covariance matrix*

$$C = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \vdots & \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{pmatrix} \quad (73.46)$$

and $|C|$ is the determinant of C. If $X(t_1), \ldots, X(t_n)$ are linearly dependent, then the joint pdf takes on a form similar to Eq. (73.45), but contains impulses [see Gardner, 1990, p. 40].

A weakly stationary Gaussian process is strongly stationary to all orders *n:* all Gaussian joint pdf's are completely determined by their first and second moments by Eq. (73.45), and those moments are time independent by weak stationarity, and so all joint pdf's are also. Every second-order strongly stationary stochastic process $X(t)$ is also weakly stationary because the time translation independence of the joint pdf's determines the first and second moments to have the same property. However, non-Gaussian weakly stationary processes need not be strongly second-order stationary.

Rather than with pdf's, a process $X(t)$ may be specified in terms of conditional pdf's

$$f_{X(t(1))\ldots X(t(n))}(x_1, \ldots, x_n) = f_{X(t(n))|X(t(n-1))}(x_n|x_{n-1}) \cdot \ldots \cdot f_{X(t(2))|X(t(1))}(x_2|x_1) f_{X(t(1))}(x_1)$$

by successive applications of Bayes' law, for $t_1 < t_2 < \cdots < t_n$. The conditional pdf's satisfy

$$f_{A|B}(a|b) = f_{AB}(a,b)/f_B(b) \quad (73.47)$$

The conditional factorization property may satisfy

$$f_{X(t(n))|X(t(n-1))\ldots X(t(1))}(x_n|x_{n-1}, \ldots, x_1) = f_{X(t(n))|X(t(n-1))}(x_n|x_{n-1}) \quad (73.48)$$

which indicates that the pdf of the process at any time t_n, given values of the process at any number of previous times $t_{n-1}, \ldots, t_1$, is the same as the pdf at t_n given the value of the process at the most recent time t_{n-1}. Such an $X(t)$ is called a *first-order Markov process*, in which case we say the process remembers only the previous value (the previous value has influence). In general, an *nth-order Markov* process remembers only the n most recent previous values. A first-order Markov process can be fully specified in terms of its first-order conditional pdf's $f_{X(t)|X(s)}(x_t, x_s)$ and its unconditional first-order pdf at some initial time t_0, $f_{X(t(0))}(x_0)$.

Examples of Stochastic Processes

Figure 73.9 shows two sample functions of nonstationary processes. Now consider the discrete time process $X(k) = A$, for all $k \geq 0$, where A is a rv (a *random initial condition*) that assumes a value 1 or –1 with respective probabilities p and $1 - p$ at $k = 0$. This value does not change, once the initial random draw is done at $k = 0$. This stochastic sequence has two sample functions only, the constant sequences $\{-1\}$ and $\{1\}$. The expected value of $X(k)$ at any time k is $E[X(k)] = E[A] = p \cdot 1 + (1 - p) \cdot (-1) = 2p - 1$, which is independent of k. The autocorrelation function is, by definition, $E[X(k)X(k + m)] = E[A \cdot A] = E[A^2] = p \cdot 1^2 + (1 - p) \cdot (-1)^2 = 1$ which is also independent of time k. Thus $X(k)$ is perfectly correlated for all time (the process has *infinite memory*). This process is deterministic.

For another example, put $X(t) = (c) \cdot \cos(wt + \Phi)$, where Φ is the uniform rv on $(-\pi, \pi)$. Then $X(t)$ is a function of the rv Φ (as well as t), so by use of Eq. (73.39a), we obtain

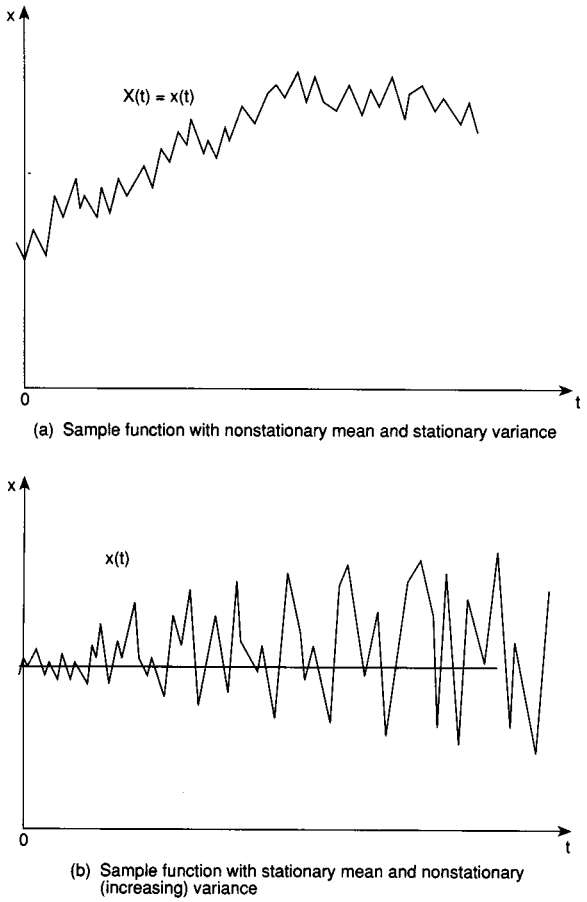

FIGURE 73.9 Examples of nonstationary processes.

$$E[X(t) = c \cdot \int_{-\pi}^{\pi} \cos(wt + \phi) f_{\Phi}(\phi) d\phi = (c/2\pi) \sin(wt + \phi)\Big|_{-\pi}^{\pi} = 0$$

Therefore, the mean does not vary with time t. The autocorrelation is

$$\begin{aligned} R_{XX}(t, t + \tau) &= E[(c) \cdot \cos(wt + \Phi)(c) \cdot \cos(wt + w\tau + \Phi)] \\ &= c^2 E[\cos(wt + \Phi) \cos(wt + w\tau + \Phi)] \\ &= c^2 \int_{-\pi}^{\pi} \cos(wt + \phi) \cos(wt + w\tau + \Phi) f_{\Phi}(\phi) d\phi \\ &= (c^2/2) \int_{-\pi}^{\pi} \{\cos(2wt + 2\phi + w\tau) + \cos(w\tau)\}(1/2\pi) d\phi \\ &= (c^2/4\pi) \cdot \{\sin(\Theta + 2\pi) - \sin(\Theta - 2\pi) + \cos(w\tau) \cdot 2\pi\} \\ &= (c^2/4\pi) \cdot \{\cos(w\tau) \cdot 2\pi\} = (c^2/2) \cos(w\tau) \end{aligned}$$

[using $\cos(x)\cos(y) = \frac{1}{2}\{\cos(x + y) + \cos(x - y)\}$ and letting $\Theta = 2wt + 2\Phi + w\tau$]. Therefore, $X(t)$ is ws. The autocorrelation is periodic in the offset variable τ.

Now consider the example $X(t) = A\cos(2\pi f_0 t)$ for each t, where f_0 is a constant frequency, and the amplitude A is a random initial condition as given above. There are only two sample functions here: (1) x(t) = $\cos(2\pi f_0 t)$ and (2) $x(t) = -\cos(2\pi f_0 t)$. A related example is $X(t) = A\cos(2\pi f_0 t + \Phi)$, where A is given above, the phase Φ is the uniform random variable on $[0,\pi]$, and A and Φ are independent. Again, Φ and A do not depend on time (initial random conditions). Thus, the sample functions for $X(t)$ are $x(t) = \pm\cos(2\pi ft + \phi)$, where $\Phi = \phi$ is the value assumed initially. There are infinitely many sample functions because of the phase. Equation (73.39b) and the independence of A and Φ yield

$$\begin{aligned} E[X(t)] &= E[A\cos(2\pi ft + \Phi)] = E[A]E[g(\Phi)] = \mu_A \int_0^{\pi} \cos(2\pi ft + \phi)(1/\pi)d\phi \\ &= (\mu_A/\pi)\sin(2\pi t + \phi)\Big|_{\phi=0}^{\pi} = (\mu_A/\pi)[\sin(2\pi t + \pi) - \sin(2\pi t)] \\ &= (\mu_A/\pi)[\sin(-2\pi t) - \sin(2\pi t)] = (-2\mu_A/\pi)\sin(2\pi t) \end{aligned}$$

which is dependent upon time. Thus, $X(t)$ is not ws.

Next, let $X(t) = [a + S(t)]\cos[2\pi ft + \Phi]$, where the signal $S(t)$ is a nondeterministic stochastic process. This is an amplitude-modulated sine wave carrier. The carrier $\cos[2\pi ft + \Phi]$ has random initial condition Φ and is deterministic. Because $S(t)$ is nondeterministic, $X(t)$ is also. The expected value $E[X(t)] = E[a + S(t)]E[\cos(2\pi ft + \Phi)]$ can be found as above by independence of $S(t)$ and Φ.

Finally, let $X(t)$ be uncorrelated ($E[X(t)X(t + \tau)] = 0$ for τ not zero) such that each rv $X(t) = X_t$ is Gaussian with zero mean and variance $\sigma^2(t) = t$, for all $t > 0$. Any realized sample function $x(t)$ of $X(t)$ cannot be predicted in any average sense based on past values (uncorrelated Gaussian random variables are independent). The variance grows in an unbounded manner over time, so $X(t)$ is neither stationary nor deterministic. This is called the *Wiener* process.

A useful model of a ws process is that for which $\mu = 0$ and $R_{XX}(\tau) = \sigma_X^2\exp(-\alpha|\tau|)$. If this process is also Gaussian, then it is strongly stationary and all of its joint pdf's are fully specified by $R_{XX}(\tau)$. In this case it is also a first-order Markov process and is called the *Ornstein-Uhlenbeck* process [see Gardner, 1990, p. 102]. Unlike white noise, many real-world ws stochastic processes are correlated ($|R_{XX}(t, t + \tau)| > 0$) for $|\tau| > 0$. The autocorrelation either goes to zero as τ goes to infinity, or else it has periodic or other nondecaying memory. We consider ws processes henceforth [for nonstationary processes, see Gardner, 1990]. We will also assume without loss of generality that $\mu = 0$.

Linear Filtering of Weakly Stationary Processes

Let the ws stochastic process $X(t)$ be the input to a linear time-invariant stable filter with impulse response function $h(t)$. The output of the filter is also a ws stochastic process and is given by the convolution

$$Y(t) = h(t) * X(t) = \int_{-\infty}^{\infty} h(s)X(t-s)ds \tag{73.49}$$

The mean of the output process is obtained by using the linearity of the expectation operator [see Gardner, 1990, p. 32]

$$\begin{aligned} \mu_Y &= E[Y(t)] = E\left[\int_{-\infty}^{\infty} h(s)X(t - s)ds\right] = \int_{-\infty}^{\infty} h(s)E[X(t - s)]ds = \int_{-\infty}^{\infty} h(s)\mu_X ds \\ &= \mu_X \int_{-\infty}^{\infty} h(s)ds = \mu_X \cdot H(0) \end{aligned} \tag{73.50}$$

where $H(f) = \int_{-\infty}^{\infty} h(t)\,e^{-j2\pi ft}dt$ is the filter transfer function and $H(0)$ is the dc gain of the filter.

The autocorrelation of the output process, obtained by using the linearity of $E[\cdot]$, is

$$\begin{aligned} R_{YY}(\tau) &= E[Y(t)Y(t+\tau)] = E\left[\int_{-\infty}^{\infty} h(v)X(t-v)dv \int_{-\infty}^{\infty} h(u)X(t+\tau-u)du\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[X(t-v)X(t+\tau-u)]h(v)h(u)dvdu \\ &= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} R_{XX}(\tau-u+v)h(u)du\right\}h(v)dv \\ &= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} R_{XX}([\tau-(-v)]-u)h(u)du\right\}h(-v)(dv) \\ &= \int_{-\infty}^{\infty}\left\{R_{XX}(\tau+v) * h(\tau+v)\right\}h(-v)dv \\ &= [R_{XX}(\tau) * h(\tau)] * h(-\tau) = R_{XX}(\tau) * [h(\tau) * h(-\tau)] = R_{XX}(\tau) * r_h(\tau) \end{aligned} \tag{73.51}$$

where $r_h(\tau) = \int_{-\infty}^{\infty} h(\tau + u)h(u)du$. However, $r_h(\tau)$ has Fourier transform $H(f)H^*(f) = |H(f)|^2$, because the Fourier transform of the convolution of two functions is the product of their Fourier transforms, and the Fourier transform of $h(-\tau)$ is the complex conjugate $H^*(f)$ of the Fourier transform $H(f)$ of $h(\tau)$. Thus, the Fourier transform of $R_{YY}(\tau)$, denoted by $\mathrm{F}\{R_{YY}(\tau)\}$, is

$$\mathbf{F}\{R_{YY}(\tau)\} = \mathbf{F}\{R_{XX}(\tau) * h(\tau) * h(-\tau)\} = \mathbf{F}\{R_{XX}(\tau)\} \cdot H(f)H^*(f) = \mathbf{F}\{R_{XX}(\tau)\} \cdot |H(f)|^2$$

Upon defining the functions

$$S_{XX}(f) \equiv \mathbf{F}\{R_{XX}(\tau)\},\ S_{YY}(f) \equiv \mathbf{F}\{R_{YY}(\tau)\} \tag{73.52}$$

we can also determine $R_{YY}(\tau)$ via the two steps

$$S_{YY}(f) = S_{XX}(f) \cdot |H(f)|^2 \tag{73.53}$$

$$R_{YY}(\tau) = \mathbf{F}^{-1}\{S_{YY}(f)\} = \int_{-\infty}^{\infty} S_{YY}(f)e^{j2\pi f\tau}df \tag{73.54}$$

Equations (73.52) define the *power spectral density functions* (psdf's) $S_{XX}(f)$ for $X(t)$ and $S_{YY}(f)$ for $Y(t)$. Thus, $R_{XX}(\tau)$ and $S_{XX}(f)$ are Fourier transform pairs, as are $R_{YY}(\tau)$ and $S_{YY}(f)$ (see Eq. 73.20). Further, the psdf $S_{XX}(f)$ of $X(t)$ is a power spectrum (in an average sense). If $X(t)$ is the voltage dropped across a 1-Ω resistor, then $X^2(t)$ is the instantaneous power dissipation in the resistance. Consequently, $R_{XX}(0) = E[X^2(t)]$ is the expected power dissipation over all frequencies, i.e., by Eq. (73.54) with $\tau = 0$, we have

$$R_{XX}(0) = \int_{-\infty} S_{XX}(f)df$$

We want to show that when we pass $X(t)$ through a narrow bandpass filter with a bandwidth δ centered at the frequency $\pm f_0$, the expected power at the output terminals, divided by the bandwidth δ, is $S_{XX}(f_0)$ in the limit as $\delta \to 0$. This shows that $S_{XX}(f)$ is a density function (whose area is the total expected power over all frequencies, just as the area under a pdf is the total probability). This result that $R_{XX}(\tau)$ and $S_{XX}(f)$ are a Fourier transform pair is known as the *Wiener-Khinchin* relation [see Gardner, 1990, p. 230].

To verify this relation, let $H(f)$ be the transfer function of an ideal bandpass filter, where

$$H(f) = 1, |f - f_0| < \delta/2; \qquad H(f) = 0, \text{ otherwise}$$

Let Y(t) be the output of the filter. Then Eqs. (73.54) and (73.53) provide

$$E[Y^2(t)] = R_{YY}(0) = \int_{-\infty}^{\infty} S_{YY}(f)df = \int_{-\infty}^{\infty} S_{XX}(f) \mid H(f) \mid^2 df$$

$$= \int_{f_0-\delta/2}^{f_0+\delta/2} S_{XX}(f)df + \int_{-f_0-\delta/2}^{-f_0+\delta/2} S_{XX}(f)df$$

Dividing by 2δ and taking the limit as $\delta \to 0$ yields $(1/2)S_{XX}(f_0) + (1/2)S_{XX}(-f_0)$, which becomes $S_{XX}(f_0)$ when we use the fact that psdf's are even and real functions (because they are the Fourier transforms of autocorrelation functions, which are even and real).

For example, let $X(t)$ be white noise, with $S_{XX}(f) = N_0$, being put through a first-order linear time-invariant system with respective impulse response and transfer functions

$$h(t) = \exp\{-\alpha t\},\ t \geq 0;\ h(t) = 0,\ t < 0 \qquad H(f) = 1/[\alpha + j2\pi f],\ \text{all } f$$

The temporal correlation of $h(t)$ with itself is $r_h(\tau) = (1/2\alpha)\exp\{-\alpha|\tau|\}$, so the power transfer function is $|H(f)|^2 = 1/[\alpha^2 + (2\pi f)^2]$. The autocorrelation for the input $X(t)$ is

$$R_{XX}(\tau) = \int_{-\infty}^{\infty} N_0 e^{j2\pi f\tau} df = N_0\delta(\tau)$$

which is an impulse. It follows (see Eq. 73.22) that the output $Y(t)$ has respective autocorrelation and psdf

$$R_{YY}(\tau) = [N_0\delta(\tau)] * [(1/2\alpha)\, e^{-\alpha|\tau|}] = (N_0/2\alpha)\, e^{-\alpha|\tau|},\ S_{YY}(f) = N_0/[\alpha^2 + (2\pi f)^2]$$

The output expected power $E[Y^2(t)]$ can be found from either one of

$$E[Y^2(t)] = R_{YY}(0) = N_0/2\alpha \qquad \text{or} \qquad E[Y^2(t)] = \int_{-\infty}^{\infty} S_{YY}(f)df = N_0/2\alpha$$

If the input $X(t)$ to a linear system is Gaussian, then the output will also be Gaussian [see Brown, 1983; or Gardner, 1990]. Thus, the output of a first-order linear time-invariant system driven by Gaussian white noise is the Ornstein–Uhlenbeck process, which is also a first-order Markov process.

For another example, let $X(t) = A\cos(w_o t + \Theta)$, where the random amplitude A has zero mean, the random phase Θ is uniform on $[-\pi, \pi]$, and A and Θ are independent. As before, we obtain $R_{XX}(\tau) = \sigma_A{}^2\cos(w_o\tau)$, from which it follows that $S_{XX}(f) = (\sigma_A{}^2/2)[\delta(f - w_o/2\pi) + \delta(f + w_o/2\pi)]$. These impulses in the psdf, called *spectral lines*, represent positive amounts of power at discrete frequencies.

Cross-Correlation of Processes

The *cross-correlation function* for two random processes $X(t)$ and $Y(t)$ is defined via

$$R_{XY}(t,\ t+\tau) \equiv E[X(t)Y(t+\tau)] \tag{73.55}$$

Let both processes be ws with zero means, so the covariance coincides with the correlation function. We say that two ws processes $X(t)$ and $Y(t)$ are *jointly ws* whenever $R_{XY}(t,t+\tau) = R_{XY}(\tau)$. In case $Y(t)$ is the output of a filter with impulse response $h(t)$,we can find the cross-correlation $R_{XY}(\tau)$ between the input and output via

$$R_{XY}(\tau) = E[X(t)Y(t+\tau)] = E[X(t)\int_{-\infty}^{\infty} h(u)X(t+\tau-u)du]$$
$$= \int_{-\infty}^{\infty} h(u)E[X(t)X(t+\tau-u)]\,du \tag{73.56}$$
$$= \int_{-\infty}^{\infty} h(u)R_{XX}(\tau-u)du = R_{XX}(\tau) * h(\tau)$$

Cross-correlation functions of ws processes satisfy (1) $R_{XY}(-\tau) = R_{YX}(\tau)$, (2) $|R_{XY}(\tau)|^2 \le R_{XX}(0)R_{YY}(0)$, and (3) $|R_{XY}(\tau)| \le (1/2)[R_{XX}(0) + R_{YY}(0)]$. The first follows from the definition, while the second comes from expanding $E[\{Y(t+\tau) - \alpha X(t)\}^2] \ge 0$. The third comes from the fact that the geometric mean cannot exceed the arithmetic mean [see Peebles, 1987, p. 154].

Taking the Fourier transform of the leftmost and rightmost sides of Eqs. (73.56) yields

$$S_{XY}(f) = S_{XX}(f)H(f) \tag{73.57}$$

The Fourier transform of the cross-correlation function is the *cross-spectral density function*

$$S_{XY}(f) = \int_{-\infty}^{\infty} R_{XY}(\tau)e^{-j2\pi f\tau}d\tau \tag{73.58}$$

According to Gardner [1990, p. 228], this is a *spectral correlation density function* that does not represent power in any sense.

Equation (73.57) suggests a method for identifying a linear time-invariant system. If the system is subjected to a ws input $X(t)$ and the power spectral density of $X(t)$ and the cross-spectral density of $X(t)$ and the output $Y(t)$ are measured, then the ratio yields the system transfer function

$$H(f) = S_{XY}(f)/S_{XX}(f) \tag{73.59}$$

In fact, it can be shown that this method gives the best linear time-invariant model of the (possibly time varying and nonlinear) system in the sense that the time-averaged mean-square error between the outputs of the actual system and of the model, when both are subjected to the same input, is minimized [see Gardner, 1990, pp. 282–286].

As an application, suppose that an undersea sonar-based device is to find the range to a target, as shown in Fig. 73.10, by transmitting a sonar signal $X(t)$ and receiving the reflected signal $Y(t)$. If v is the velocity of the sonar signal, and τ_o is the offset that maximizes the cross-correlation $R_{XY}(\tau)$, then the range (distance) d can be determined from $d = v\tau_o/2$ (note that the signal travels twice the range d).

Coherence

When $X(t)$ and $Y(t)$ have no spectral lines at f, the finite spectral correlation $S_{XY}(f)$ is actually a spectral covariance and the two associated variances are $S_{XX}(f)$ and $S_{YY}(f)$. We can normalize $S_{XY}(f)$ to obtain a *spectral correlation coefficient* $\Upsilon_{XY}(f)$ defined by

$$\Upsilon_{XY}(f)^2 = |S_{XY}(f)|^2/S_{XX}(f)S_{YY}(f) \tag{73.60}$$

We call $\Upsilon_{XY}(f)$ the *coherence function*. It is a measure of the power correlation of $X(t)$ and $Y(t)$ at each frequency f. When $Y(t) = X(t)*h(t)$, it has a maximum: by Eqs. (73.53), (73.59), and (73.60), $|\Upsilon_{XY}(f)|^2 = |S_{XX}(f) \cdot H(f)|^2/[S_{XX}(f) \cdot S_{XX}(f) \cdot |H(f)|^2] = 1$. In the general case we have

$$|S_{XY}(f)| \le [S_{XX}(f)S_{YY}(f)]^{1/2} \tag{73.61}$$

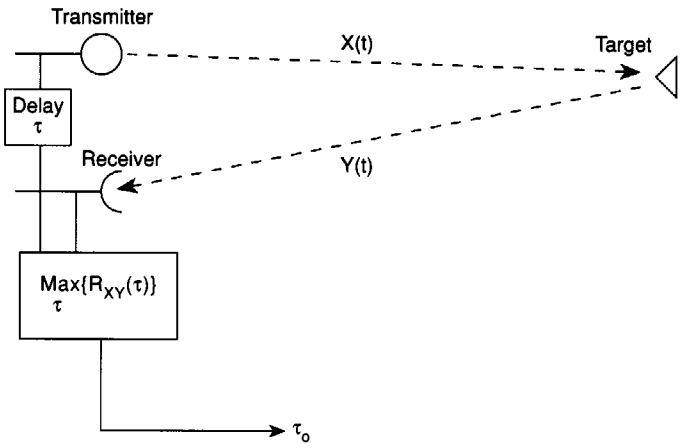

FIGURE 73.10 A sonar range finder.

Upon minimizing the mean-square error $\varepsilon = E[(Y(t) - X(t)*h(t))^2]$ over all possible impulse response functions $h(t)$, the optimal one, $h_o(t)$, has transfer function

$$H_o(f) = S_{XY}(f)/S_{XX}(f) \tag{73.62}$$

Further, the resultant minimum value is given by

$$\varepsilon_{\min} = \int_{-\infty}^{\infty} S_{YY}(f)[1 - |\Upsilon_{XY}(f)|^2]\,df$$

[see Gardner, 1990, pp. 434–436; or Bendat and Piersol, 1986]. At frequencies f where $|\Upsilon_{XY}(f)| \approx 1$, $\varepsilon_{\min} \approx 0$. Thus $1 - |\Upsilon_{XY}(f)|^2$ is the mean-square proportion of $Y(t)$ not accounted for by $X(t)$, while $|\Upsilon_{XY}(f)|^2$ is the proportion due to $X(t)$. When $Y(t) = X(t)*h(t)$, $\varepsilon_{\min} = 0$.

The optimum system $H_o(f)$ of Eq. (73.62) is known as the *Wiener filter* for minimum mean-square error estimation of one process $Y(t)$ using a filtered version of another process $X(t)$ [see Gardner, 1990; or Peebles, 1987, p. 262].

Ergodicity

When the **time average**

$$\lim_{T\to\infty}(1/T)\int_{-T/2}^{T/2} X(t)dt$$

exists and equals the corresponding expected value $E[X(t)]$, then the process $X(t)$ is said to possess an *ergodic property associated with the mean*. There are ergodic properties associated with the mean, autocorrelation (and power spectral density), and all finite-order joint moments, as well as finite-order joint pdf's. If a process has all possible ergodic properties, it is said to be an *ergodic process*.

Let $Y(t) = g[X(t + t_1), \ldots, X(t + t_n)]$, where $g[\cdot]$ is any nonrandom real function, so that $Y(t)$ is a function of a finite number of time samples of a strongly stationary process. For example, let (1) $Y(t) = X(t + t_1)X(t + t_2)$, $E[Y(t)] = R_{XX}(t_1 - t_2)$ and (2) $Y(t) = 1$ if $X(t) < x$, $Y(t) = 0$, otherwise, so that

$$E[Y(t)] = 1\cdot P(X(t) < x) + 0\cdot P(X(t) \geq x) = P(X(t) < x) = \int_{-\infty}^{1} f_{X(t)}(z)dz$$

We want to know under what conditions the mean-square error between the time average

$$\langle Y(t) \rangle_T \equiv (1/T)\int_{-T/2}^{T/2} Y(t)dt$$

and the expected value $E[Y(t)]$ will converge to zero. It can be shown that a necessary and sufficient condition for the mean-square ergodic property

$$\lim_{T\to\infty} E[\{\langle Y(t) \rangle_T - E[Y(t)]\}^2] = 0 \tag{73.63}$$

to hold is that

$$\lim_{T\to\infty} (1/T)\int_0^T C_{YY}(\tau)d\tau = 0 \tag{73.64}$$

For example, if $C_{YY}(\tau) \to 0$ as $\tau \to \infty$, then Eq. (73.64) will hold, and thus Eq. (73.63) will also, where $C_{YY}(\tau)$ is the covariance function of $Y(t)$. As long as the two sets of rv's $\{X(t + t_1), \ldots, X(t + t_n)\}$ and $\{X(t + t_1 + \tau), \ldots, X(t + t_n + \tau)\}$ become independent of each other as $\tau \to \infty$, the above condition holds, so Eq. (73.63) holds [see Gardner, 1990, pp. 163–174].

In practice, if $X(t)$ exhibits ergodicity associated with the autocorrelation, then we can estimate $R_{XX}(\tau)$ using the time average

$$\langle X(t)X(t+\tau) \rangle_T \equiv (1/T)\int_{-T/2}^{T/2} X(t)X(t+\tau)dt \tag{73.65}$$

In this case the mean-square estimation error $E[\{\langle X(t)X(t+\tau)\rangle_T - R_{XX}(\tau)\}^2]$ will converge to zero as T increases to infinity, and the power spectral density $S_{XX}(f)$ can also be estimated via time averaging [see Gardner, 1990, pp. 230–231].

Defining Terms

Autocorrelation function: A function $R_{XX}(t, t + \tau) = E[X(t)X(t + \tau)]$ that measures the degree to which any two rv's $X(t)$ and $X(t + \tau)$, at times t and $t + \tau$, are correlated.

Coherence function: A function of frequency f that provides the degree of correlation of two stochastic processes at each f by the ratio of their cross-spectral density function to the product of their power spectral density functions.

Power spectral density function: The Fourier transform of the autocorrelation function of a stochastic process $X(t)$, denoted by $S_{XX}(f)$. The area under its curve between f_1 and f_2 represents the total power over all t in $X(t)$ in the band of frequencies f_1 to f_2. Its dimension is watts per Hz.

Sample function: A real-valued function $x(t)$ of t where at each time t the value $x(t)$ at the argument t was determined by the outcome of a rv $X_t = x(t)$.

Stochastic process: A collection of rv's $\{X_t; t \in T\}$, where T is an ordered set such as the real numbers or integers [$X(t)$ is also called a random function, on the domain T].

Time average: Any real function $g(t)$ of time has average value g_{ave} on the interval $[a,b]$ such that the rectangular area $g_{ave}(b - a)$ is equal to the area under the curve between a and b, i.e., $g_{ave} = [1/(b - a)] \int_a^b g(t)dt$. The time average of a sample function $x(t)$ is the limit of its average value over $[0, T]$ as T goes to infinity.

Weakly stationary: The property of a stochastic process $X(t)$ whose mean $E[X(t)] = \mu(t)$ is a fixed constant μ over all time t, and whose autocorrelation is also independent of time in that $R_{XX}(t, t + \tau) = R_{XX}(s + t, s + t + \tau)$ for any s. Thus, $R_{XX}(t, t + \tau) = R_{XX}(0, \tau) = R_{XX}(\tau)$.

Related Topic

16.1 Spectral Analysis

References

The author is grateful to William Gardner of the University of California, Davis for making substantial suggestions.

J.S. Bendat and A.G. Piersol, *Random Data: Analysis and Measurement,* 2nd ed., New York: Wiley-Interscience, 1986.

R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering,* New York: Wiley, 1983.

W. A. Gardner, *Introduction to Random Processes,* 2nd ed., New York: McGraw-Hill, 1990.

P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles,* 2nd ed., New York: McGraw-Hill, 1987.

Further Information

The IEEE Individual Learning Package, *Random Signal Analysis with Random Processes and Kalman Filtering,* prepared for the IEEE in 1989 by Carl G. Looney, IEEE Educational Activities Board, PO Box 1331, Piscataway, NJ 08855-1331.

R. Iranpour and P. Chacon, *Basic Stochastic Processes: The Mark Kac Lectures,* New York: Macmillan, 1988.

A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* 3rd ed., New York: Macmillan, 1991.

73.4 The Sampling Theorem

R. J. Marks II

Most signals originating from physical phenomena are analog. Most computational engines, on the other hand, are digital. Transforming from analog to digital is straightforward: we simply sample. Regaining the original signal from these samples and assessing the information lost in the sampling process are the fundamental questions addressed by the **sampling theorem**.

The fundamental result of the sampling theorem is, remarkably, that a bandlimited signal is uniquely specified by its sufficiently close equally spaced samples. Indeed, the sampling theorem illustrates how the original signal can be regained from knowledge of the samples and the sampling rate at which they were taken.

Popularization of the sampling theorem is credited to Shannon [1948] who, in 1948, used it to show the equivalence of the information content of a bandlimited signal and a sequence of discrete numbers. Shannon was aware of the pioneering work of Whittaker [1915] and Whittaker's son [1929] in formulating the sampling theorem. Kotel'nikov's [1933] independent discovery in the then Soviet Union deserves mention. Higgins [1985] credits Borel [1897] with first recognizing that a signal could be recovered from its samples.

Surveys of sampling theory are in the widely cited paper of Jerri [1977] and in two books by the author [1991, 1993]. Marvasti [1987] has written a book devoted to nonuniform sampling.

The Cardinal Series

If a signal has finite energy, the minimum **sampling rate** is equal to two samples per period of the highest frequency component of the signal. Specifically, if the highest frequency component of the signal is B Hz, then the signal, $x(t)$, can be recovered from the samples by

$$x(t) = \frac{1}{\pi} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \frac{\sin[\pi(2Bt - n)]}{2Bt - n} \tag{73.66}$$

The frequency B is also referred to as the signal's bandwidth and, if B is finite, $x(t)$ is said to be bandlimited. The signal, $x(t)$, is here being sampled at a rate of 2 B samples per second. If sampling were done at a lower

rate, the replications would overlap and the information about $X(\omega)$ [and thus $x(t)$] is irretrievably lost. Undersampling results in *aliased* data. The minimum sampling rate at which **aliasing** does not occur is referred to as the **Nyquist rate** which, in our example, is $2B$. Eq. (73.66) was dubbed the **cardinal series** by the junior Whittaker [1929].

A signal is bandlimited in the low-pass sense if there is a $B > 0$ such that

$$X(\omega) = X(\omega)\Pi\left(\frac{\omega}{4\pi B}\right) \tag{73.67}$$

where the gate function $\Pi(\xi)$ is one for $\xi \le 1/2$ and is otherwise zero, and

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{73.68}$$

is the **Fourier transform** of $x(t)$. That is, the spectrum is identically zero for $|\omega| > 2\pi B$. The B parameter is referred to as the signal's bandwidth. The inverse Fourier transform is

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\, e^{j\omega t}d\omega \tag{73.69}$$

The sampling theorem reduces the normally continuum infinity of ordered pairs required to specify a function to a countable—although still infinite—set. Remarkably, these elements are obtained directly by sampling.

How can the cardinal series interpolate uniquely the bandlimited signal from which the samples were taken? Could not the same samples be generated from another bandlimited signal? The answer is no. Bandlimited functions are smooth. Any behavior deviating from smooth would result in high-frequency components which in turn invalidates the required property of being bandlimited. The smoothness of the signal between samples precludes arbitrary variation of the signal there.

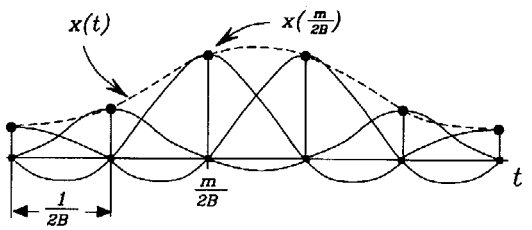

FIGURE 73.11 Illustration of the interpolation that results from the cardinal series. A sinc function, weighted by the sample, is placed at each sample bottom. The sum of the sincs exactly generates the original bandlimited function from which the samples were taken.

Let's examine the cardinal series more closely. Evaluate Eq. (73.74) at $t = m/2B$. Since $\text{sinc}(n)$ is one for $n = 0$ and is otherwise zero, only the sample at $t = m/2B$ contributes to the interpolation at that point. This is illustrated in Fig. 73.11, where the reconstruction of a signal from its samples using the cardinal series is shown. The value of $x(t)$ at a point other than a sample location [e.g., $t = (m + 1/2)/2B$] is determined by all of the sample values.

Proof of the Sampling Theorem

Borel [1897] and Shannon [1948] both discussed the sampling theorem as the Fourier transform dual of the Fourier series. Let $x(t)$ have a bandwidth of B. Consider the periodic signal

$$Y(\omega) = \sum_{n=-\infty}^{\infty} X(\omega - 4\pi nB) \tag{73.70}$$

The function $Y(\omega)$ is a periodic function with period $4\pi B$. From Eq. (73.67) $X(\omega)$ is zero for $\omega > 2\pi B$ and is thus finite in extent. The terms in Eq. (73.70) therefore do not overlap. Periodic functions can be expressed as a Fourier series.

$$Y(\omega) = \sum_{n=-\infty}^{\infty} \alpha_n \exp\left(\frac{-jn\omega}{2B}\right) \tag{73.71}$$

where the Fourier series coefficients are

$$\alpha_n = \frac{1}{4\pi B} \int_{-2\pi B}^{2\pi B} Y(\omega) \exp\left(\frac{jn\omega}{2B}\right) d\omega$$

or

$$\alpha_n = \frac{1}{2B} x\left(\frac{n}{2B}\right) \tag{73.72}$$

where we have used the inverse Fourier transform in Eq. (73.69). Substituting into the Fourier series in Eq. (73.71) gives

$$Y(\omega) = \frac{1}{2B} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \exp\left(\frac{-jn\omega}{2B}\right) \tag{73.73}$$

Since a period of $Y(\omega)$ is $X(\omega)$, we can get back the original spectrum by

$$X(\omega) = Y(\omega) \prod\left(\frac{\omega}{4\pi B}\right)$$

Substitute Eq. (73.73) and inverse transforming gives, using Eq. (73.69),

$$x(t) = \frac{1}{4\pi B} \int_{-2\pi B}^{2\pi B} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \exp\left(\frac{-jn\omega}{2B}\right) e^{j\omega t} d\omega$$

or

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \text{sinc}\,(2Bt - n) \tag{73.74}$$

where

$$\text{sinc}\,(t) = \frac{\sin \pi t}{\pi t}$$

is the inverse Fourier transform of $\Pi\,(\omega/2\pi)$. Eq. (73.74) is, of course, the cardinal series.

The sampling theorem generally converges uniformly, in the sense that

$$\lim_{N\to\infty} \left| x(t) - x_N(t) \right|^2 = 0$$

where the truncated cardinal series is

$$x_N(t) = \sum_{n=-N}^{N} x\left(\frac{n}{2B}\right) \operatorname{sinc}(2Bt - n) \tag{73.75}$$

Sufficient conditions for uniform convergence are [Marks, 1991]

1. the signal, $x(t)$, has finite energy, E,

$$E = \int_{-\infty}^{\infty} \left| x(t) \right|^2 dt < \infty$$

2. or $X(\omega)$ has finite area,

$$A = \int_{-\infty}^{\infty} \left| X(\omega) \right| d\omega < \infty$$

Care must be taken in the second case, though, when singularities exist at $\omega = \pm 2\pi B$. Here, sampling may be required to be strictly greater than $2B$. Such is the case, for example, for the signal, $x(t) = \sin(2\pi Bt)$. Although the signal is bandlimited, and although its Fourier transform has finite area, all of the samples of $x(t)$ taken at $t = n/2B$ are zero. The cardinal series in Eq. (73.74) will thus interpolate to zero everywhere. If the sampling rate is a bit greater than $2B$, however, the samples are not zero and the cardinal series will uniformly converge to the proper answer.

The Time-Bandwidth Product

The cardinal series requires knowledge of an infinite number of samples. In practice, only a finite number of samples can be used. If most of the energy of a signal exists in the interval $0 \le t \le T$, and we sample at the Nyquist rate of $2B$ samples per second, then a total of $S = \langle 2BT \rangle$ samples are taken. ($\langle\theta\rangle$ denotes the largest number not exceeding θ.) The number S is a measure of the degrees of freedom of the signal and is referred to as its **time-bandwidth product.** A 5-min single-track audio recording requiring fidelity up to 20,000 Hz, for example, requires a minimum of $S = 2 \times 20{,}000 \times 5 \times 60 = 12$ million samples. In practice, audio sampling is performed well above the Nyquist rate.

Sources of Error

Exact interpolation using the cardinal series assumes that (1) the values of the samples are known exactly, (2) the sample locations are known exactly, and (3) an infinite number of terms are used in the series. Deviation from these requirements results in interpolation error due to (1) data noise, (2) jitter, and (3) truncation, respectively. The effect of data error on the restoration can be significant. Some innocently appearing sampling theorem generalizations, when subjected to performance analysis in the presence of data error, are revealed as ill-posed. In other words, a bounded error on the data can result in unbounded error on the restoration [Marks, 1991].

Data Noise

The source of data noise can be the signal from which samples are taken, or from round-off error due to finite sampling precision. If the noise is additive and random, instead of the samples

$$x\left(\frac{n}{2B}\right)$$

we must deal with the samples

$$x\left(\frac{n}{2B}\right)+\xi\left(\frac{n}{2B}\right)$$

where $\xi(t)$ is a stochastic process. If these noisy samples are used in the cardinal series, the interpolation, instead of simple $x(t)$, is

$$x(t)\ +\ \eta(t)$$

where the interpolation noise is

$$\eta(t)=\sum_{n=-\infty}^{\infty}\xi\left(\frac{n}{2B}\right)\mathrm{sinc}\,(2Bt-n)$$

If $\xi(t)$ is a zero mean process, then so is the interpolation noise. Thus, the noisy interpolation is an unbiased version of $x(t)$. More remarkably, if $\xi(t)$ is a zero-mean (wide-sense) stationary process with uncertainty (variance) σ^2, then so is $\eta(t)$. In other words, *the uncertainty at the sample point locations is the same as at all points of interpolation* [Marks, 1991].

Truncation

The truncated cardinal series is in Eq. (73.75). A signal cannot be both bandlimited and of finite duration. Indeed, a bandlimited function cannot be identically zero over any finite interval. Thus, other than the rare case where an infinite number of the signal's zero crossings coincide with the sample locations, truncation will result in an error.

The magnitude of this **truncation error** can be estimated through the use of Parseval's theorem for the cardinal series that states

$$\begin{aligned}E\ &=\ \int_{-\infty}^{\infty}\left|\,x(t)\,\right|^2dt\\ &=\ \frac{1}{2B}\sum_{-\infty}^{\infty}\left|x\left(\frac{n}{2B}\right)\right|^2\end{aligned}$$

The energy of a signal can thus be determined directly from either the signals or the samples. The energy associated with the truncated signal is

$$E_N\ =\ \frac{1}{2B}\sum_{-N}^{N}\left|x\left(\frac{n}{2B}\right)\right|^2$$

If $E - E_N << E$, then the truncation error is small.

Jitter

Jitter occurs when samples are taken near to but not exactly at the desired sample locations. Instead of the samples $x\,(n/2W)$, we have the samples

$$x\left(\frac{n}{2W} - \sigma_n\right)$$

where σ_n is the jitter offset of the nth sample. For jitter, the σ_n's are not known. If they were, an appropriate nonuniform sampling theorem [Marks, 1993; Marvasti, 1987] could be used to interpolate the signal.

Using the jittered samples in the cardinal series results in an interpolation that is not an unbiased estimate of $x(t)$. Indeed, if the probability density function of the jitter is the same at all sample locations, the expected value of the jittered interpolation is the convolution of $x(t)$ with the probability density function of the jitter. This bias can be removed by inverse filtering at a cost of decreasing the signal-to-noise ratio of the interpolation [Marks, 1993].

Generalizations of the Sampling Theorem

There exist numerous generalizations of the sampling theorem [Marks, 1991; Marks, 1993].

1. **Stochastic processes.** A wide-sense stationary stochastic process, $\chi(t)$, is said to be bandlimited if its autocorrelation, $R_\chi(t)$, is a bandlimited function. The cardinal series

$$\hat{\chi}(t) = \sum_{n=-\infty}^{\infty} \chi\left(\frac{n}{2B}\right) \operatorname{sinc}(2Bt - n)$$

converges to $\chi(t)$ in the sense that

$$E\left[\left|\hat{\chi}(t) - \chi(t)\right|^2\right] = 0$$

where E denotes expectation.
2. **Nonuniform sampling.** There exist numerous scenarios wherein interpolation can be performed from samples that are not spaced uniformly. Marvasti [1987] devotes a book to the topic.
3. **Kramer's generalization.** Kramer generalized the sampling theorem to integral transforms other than Fourier, for example, to Legendre and Laguerre transforms.
4. **Papoulis' generalization.** Shannon noted that a bandlimited signal could be restored when sampling was performed at half the Nyquist rate if, at every sample location, a sample of the signal's derivative were also taken. Recurrent nonuniform sampling is where P samples are spaced the same in every P Nyquist intervals. Another sampling scenario is when a signal and its Hilbert transform are both sampled at half their respective Nyquist rates. Restoration of the signal from these and numerous other sampling scenarios are subsumed in an eloquent generalization of the sampling theorem by Papoulis.
5. **Lagrangian interpolation.** Lagrangian interpolation is a topic familiar in numerical analysis. An Nth order polynomial is fit to $N + 1$ arbitrarily spaced sample points. If an infinite number of samples are equally spaced, then Lagrangian interpolation is equivalent to the cardinal series.
6. **Trigonometric polynomials.** All periodic bandlimited signals can be expressed as trigonometric polynomials (i.e., a Fourier series with a finite number of terms). If the series has M terms, then the signal has M degrees of freedom which can be determined from M samples taken within a single period.
7. **Multidimensional sampling theorems.** Multidimensional signals, such as images, require dimensional extensions of the sampling theorem. The sampling of the signal now requires geometrical interpretation. Uniform sampling of an image, for example, can either be done on a rectangular or hexagonal grid. The minimum sampling density for one geometry may differ from that of another. The smallest sampling density that does not result in aliasing can be achieved, in many cases, with a number of different uniform sampling geometries and is referred to as the Nyquist density. Interestingly, sampling can sometimes be performed below the Nyquist density with nonuniform sampling geometries such that the multidimensional signal can be restored. Such is not the case for one dimension.

8. **Continuous sampling.** When a signal is known on one or more disjoint intervals, it is said to have been continuously sampled. Divide the time line into intervals of T. Periodic continuous sampling assumes that the signal is known on each interval over an interval of αT where α is the duty cycle. Continuously sampled signals can be accurately interpolated even in the presence of aliasing. Other continuously sampled cases, each of which can be considered as a limiting case of continuously periodically sampled restoration, include
   (a) **Interpolation.** The tails of a signal are known and we wish to restore the middle.
   (b) **Extrapolation.** We wish to generate the tails of a function with knowledge of the middle.
   (c) **Prediction.** A signal for $t > 0$ is to be estimated from knowledge of the signal for $t < 0$.

Final Remarks

Since its popularization in the late 1940s, the sampling theorem has been studied in depth. More than 1000 papers have been generated on the topic [Marks, 1993]. Its understanding is fundamental in matching the largely continuous world to digital computation engines.

Defining Terms

Aliasing: A phenomenon that occurs when a signal is undersampled. High-frequency information about the signal is lost.

Cardinal series: The formula by which samples of a bandlimited signal are interpolated to form a continuous time signal.

Fourier transform: The mathematical operation that converts a time-domain signal into the frequency domain.

Jitter: A sample is temporally displaced by an unknown, usually small, interval.

Kramer's generalization: A sampling theory based on other than Fourier transforms and frequency.

Lagrangian interpolation: A classic interpolation procedure used in numerical analysis. The sampling theorem is a special case.

Nyquist rate: The minimum sampling rate that does not result in aliasing.

Papoulis' generalization: A sampling theory applicable to many cases wherein signal samples are obtained either nonuniformly and/or indirectly.

Sampling rate: The number of samples per second.

Sampling theorem: Samples of a bandlimited signal, if taken close enough together, exactly specify the continuous time signal from which the samples were taken.

Signal bandwidth: The maximum frequency component of a signal.

Time bandwidth product: The product of a signal's duration and bandwidth approximates the number of samples required to characterize the signal.

Truncation error: The error that occurs when a finite number of samples are used to interpolate a continuous time signal.

Related Topic

8.5 Sampled Data

References

E. Borel, "Sur l'interpolation," *C.R. Acad. Sci. Paris,* vol. 124, pp. 673–676, 1897.

J. R. Higgins, "Five short stories about the cardinal series," *Bull. Am. Math. Soc.,* vol. 12, pp. 45–89, 1985.

A. J. Jerri, "The Shannon sampling theorem—its various extension and applications: a tutorial review," *Proc. IEEE,* vol. 65, pp. 1565–1596, 1977.

V. A. Kotel'nikov, "On the transmission capacity of 'ether' and wire in electrocommunications," *Izd. Red. Upr. Svyazi RKKA* (Moscow), 1933.

R. J. Marks II, *Introduction to Shannon Sampling and Interpolation Theory,* New York: Springer-Verlag, 1991.

R. J. Marks II, Ed., *Advanced Topics in Shannon Sampling and Interpolation Theory,* New York: Springer-Verlag, 1993.
F. A. Marvasti, *A Unified Approach to Zero-Crossing and Nonuniform Sampling,* Oak Park, Ill.: Nonuniform, 1987.
C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal,* vol. 27, pp. 379, 623, 1948.
E. T. Whittaker, "On the functions which are represented by the expansions of the interpolation theory," *Proc. Royal Society of Edinburgh,* vol. 35, pp. 181–194, 1915.
J. M. Whittaker, "The Fourier theory of the cardinal functions," *Proc. Math. Soc. Edinburgh,* vol. 1, pp. 169–176, 1929.
A. I. Zayed, Advances in Shannon's Sampling Theory, Boca Raton, Fla.: CRC Press, 1993.

Further Information

An in-depth study of the sample theorem and its numerous variations is provided in R. J. Marks II, Ed., *Introduction to Shannon Sampling and Interpolation Theory,* New York:Springer-Verlag, 1991.

In-depth studies of modern sampling theory with over 1000 references are available in R. J. Marks II, Ed., *Advanced Topics in Shannon Sampling and Interpolation Theory,* New York: Springer-Verlag, 1993.

The specific case of nonuniform sampling is treated in the monograph by F. A. Marvasti, *A Unified Approach to Zero-Crossing and Nonuniform Sampling*, Oak Park, Ill:Nonuniform, 1987.

The sampling theorem is treated generically in the *IEEE Transactions on Signal Processing*. For applications, topical journals are the best source of current literature.

73.5 Channel Capacity

Sergio Verdú

Information Rates

Tens of millions of users access the Internet daily via standard telephone lines. **Modems** operating at data rates of up to 28,800 bits per second enable the transmission of text, audio, color images, and even low-resolution video. The progression in modern technology for the standard telephone channel shown in Fig. 73.12 exhibits, if not the exponential increases ubiquitous in computer engineering, then a steady slope of abut 825 bits per second per year.

Few technological advances can result in as many time-savings for worldwide daily life as advances in modem information rates. However, modem designers are faced with a fundamental limitation in the maximum transmissible information rate. Every communication channel has a number associated with it called **channel capacity**, which determines the maximum information rate that can flow through the channel regardless of the complexity of the transmitting and receiving devices. Thus, the progression of modem rates shown in Fig. 73.12 is sure to come to a halt. But, at what rate? Answering this question for any communication channel model is one of the major goals of information theory—a discipline founded in 1948 by Claude E. Shannon [Shannon, 1948].

Communication Channels

The communication channel is the set of devices and systems that connects the transmitter to the receiver. The transmitter and receiver consist of an **encoder** and **decoder**, respectively, which translate the information stream produced by the source into a signal suitable for channel transmission and vice versa (Fig. 73.13). For example, in the case of the telephone line, two communication channels (one in each direction) share the same physical channel that connects the two modems. That physical channel usually consists of twisted copper wires at both ends and a variety of switching and signal processing operations that occur at the telephone exchanges. The modems themselves are not included in the communication channel. A microwave radio link is another example of a communication channel that consists of an amplifier and an antenna (at both ends) and a certain portion of the radio spectrum. In this case, the communication channel model does not fully correspond with the physical channel. Why not, for example, view the antenna as part of the transmitter rather than the channel

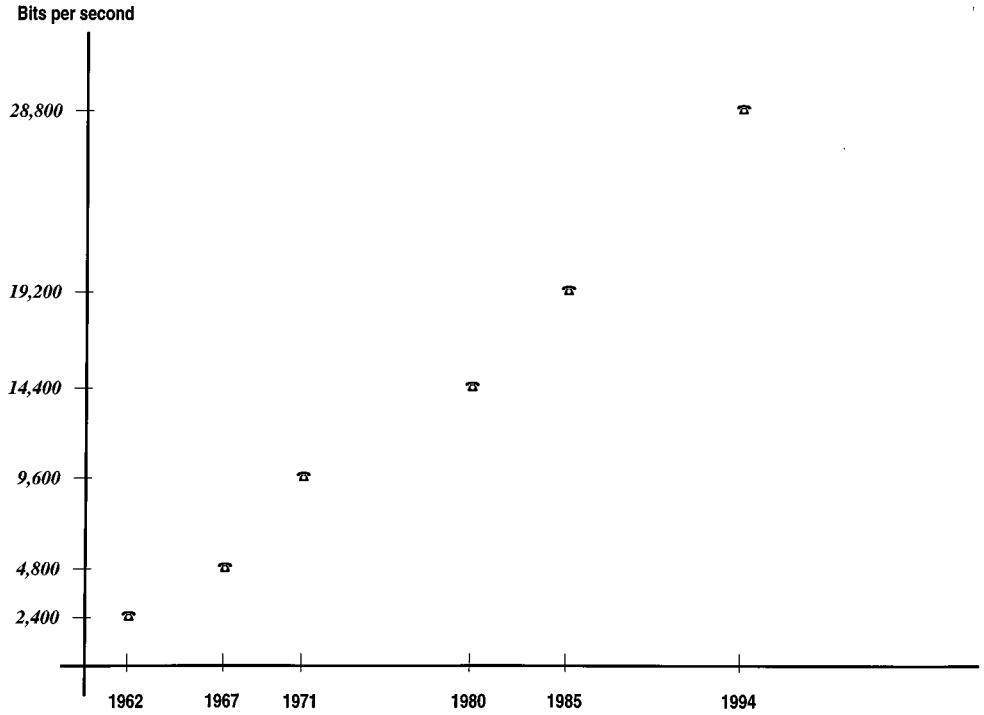

FIGURE 73.12 Information rates of modems for telephone channels.

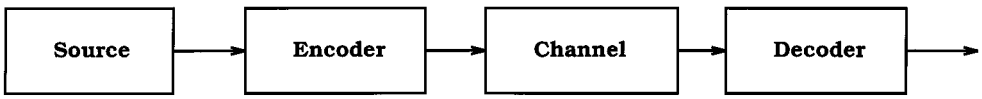

FIGURE 73.13 Elements of a communication system.

(Fig. 73.13)? Because considerations other than the optimization of the efficiency of the link are likely to dictate the choice of antenna size. This illustrates that the boundaries encoder–channel and channel–decoder in Fig. 73.13 are not always uniquely defined. This suggests an alternative definition of a channel as that part of the communication system that the designer is unable or unwilling to change.

A channel is characterized by the probability distributions of the output signals given every possible input signal. Channels are divided into (1) **discrete-time channels** and (2) **continuous-time channels** depending on whether the input/output signals are sequences or functions of a real variable. Some examples are as follows.

Example 1: Binary Symmetric Channel

A discrete-time **memoryless channel** with binary inputs and outputs (Fig. 73.14) where the probabilities that 0 and 1 are received erroneously are equal.

Example 2: Z-Channel

A discrete-time memoryless channel with binary inputs and outputs (Fig. 73.15) where 0 is received error-free.

Example 3: Erasure Channel

A discrete-time memoryless channel with binary inputs and ternary outputs (Fig. 73.16). The symbols 0 and 1 cannot be mistaken for each other but they can be "erased".

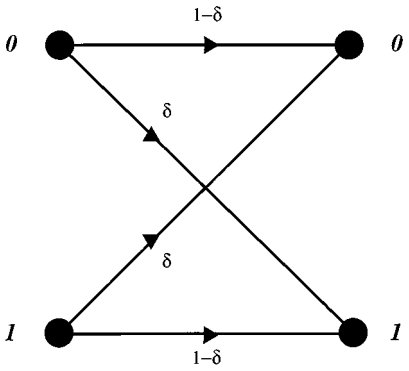

FIGURE 73.14 Binary symmetric channel.

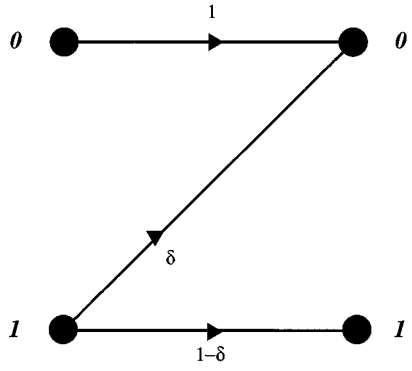

FIGURE 73.15 Z-channel.

Example 4: White Gaussian Discrete-Time Channel

A discrete-time channel whose output sequence is given by

$$y_i = x_i + n_i \tag{73.76}$$

where x_i is the input sequence and n_i is a sequence of independent Gaussian random variables with equal variance.

Example 5: Linear Continuous-Time Gaussian Channel

A continuous-time channel whose output signal is given by (Fig. 73.17)

$$y(t) = h(t) * x(t) + n(t) \tag{73.77}$$

where $x(t)$ is the input signal, $n(t)$ is a stationary Gaussian process, and $h(t)$ is the impulse response of a linear time-invariant system. The telephone channel is typically modeled by Eq. (73.77).

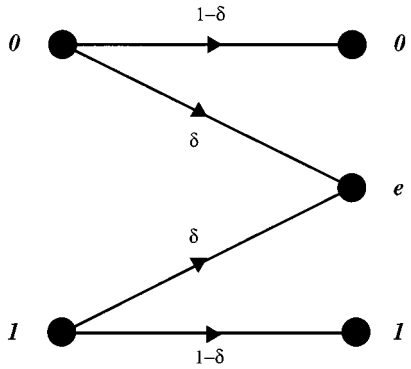

FIGURE 73.16 Erasure channel.

The goal of the encoder (Fig. 73.13) is to convert strings of binary data (messages) into channel-input signals. Source strings of m bits are translated into channel input strings of n symbols (with $m \leq n$) for discrete channels, and into continuous-time signals of duration T for continuous-time channels. The channel code (or more precisely the codebook) is the list of 2^m **codewords** (channel input signals) that may be sent by the encoder. The **rate** of the code is equal to the logarithm of its size divided by the duration of the codewords. Thus, the rate is equal to

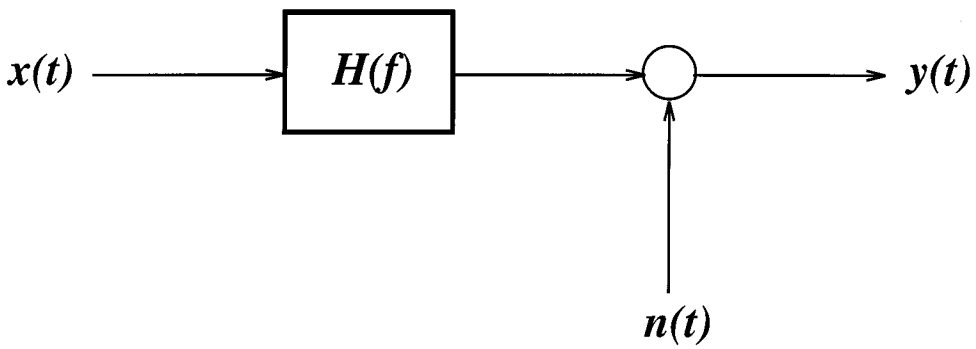

FIGURE 73.17 Linear continuous-time gaussian channel.

$$\frac{m}{n} \quad \text{bits per channel use}$$

for a discrete-time channel, whereas it is equal to

$$\frac{m}{T} \quad \text{bits per second}$$

for a continuous-time channel.

Once a codeword has been chosen by the encoder, the channel probabilistic mechanisms govern the distortion suffered by the transmitted signal. The role of the decoder is to recover the transmitted binary string (message) upon reception of the channel-distorted version of the transmitted codeword. To that end, the decoder knows the codebook used by the encoder. For most channels (including those above) there is a nonzero probability that the best decoder (maximum likelihood decoder) selects the wrong message. Thus, for a given channel the two figures of merit and of interest are the rate and the probability of error. The higher the tolerated probability of error, the higher the allowed rate; however, computing the exact tradeoff is a formidable task unless the code size either is very small or tends to infinity. The latter case was the one considered by Shannon and treated in the following section.

Reliable Information Transmission: Shannon's Theorem

Shannon [1948] considered the situation in which the codeword duration grows without bound. Channel capacity is the maximum rate for which encoders and decoders exist whose probability of error vanishes as the codewords get longer and longer.

Shannon's Theorem [Shannon, 1948] The capacity of a discrete memoryless channel is equal to

$$C = \max_X I(X; Y), \tag{73.78}$$

where $I(X; Y)$ stands for the input-output mutual information, which is a measure of the dependence of the input and the output defined as the divergence between the joint input/output distribution and the product of its marginals, $D(P_{XY}||P_X P_Y)$. For any pair of probability mass functions P and Q defined on the same space, divergence is an asymmetric measure of their similarity:

$$D(P||Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}. \tag{73.79}$$

Divergence is zero if both distributions are equal; otherwise it is strictly positive. The maximization in Eq. (73.78) is over the set of input distributions. Although, in general, there is no closed-form solution for that optimization problem, an efficient algorithm was obtained by Blahut and Arimoto in 1972 [Blahut, 1987]. The distribution that attains the maximum in Eq. (73.78) determines the statistical behavior of optimal codes and, thus, is of interest to the designer of the encoder. For the discrete memoryless channels mentioned above, the capacity is given by the following examples.

Example 1: Binary Symmetric Channel

$$C = 1 - \delta \log \frac{1}{\delta} - (1 - \delta) \log \frac{1}{1 - \delta}$$

attained by an equiprobable distribution and shown in Fig. 73.18.

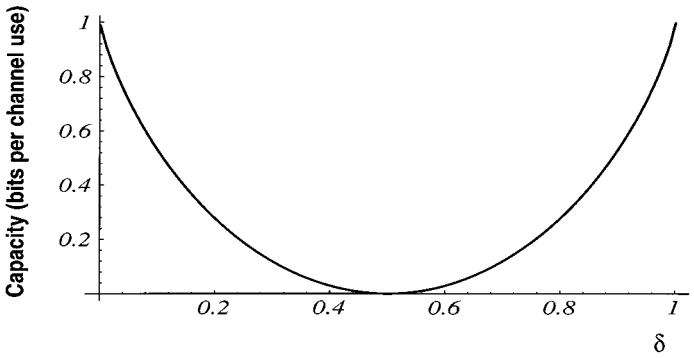

FIGURE 73.18 Capacity of the binary symmetric channel as a function of crossover probability.

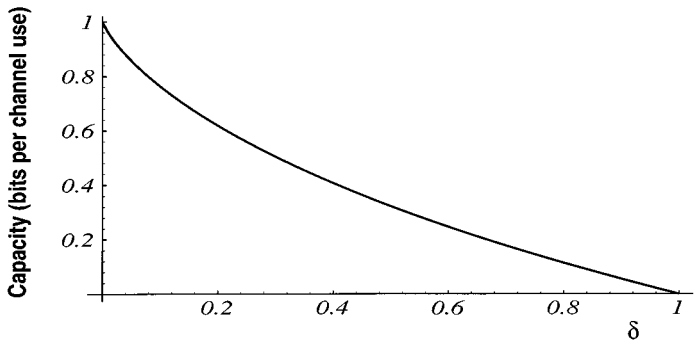

FIGURE 73.19 Capacity of the Z-channel.

Example 2: Z-Channel

$$C = \log\left(1 - \delta^{\frac{1}{1-\delta}} + \delta^{\frac{\delta}{1-\delta}}\right)$$

attained for a distribution whose probability mass at 0 ranges from 1/2 ($\delta = 0$) to $1/e$ ($\delta \to 1$) (Fig. 73.19).

Example 3: Erasure Channel

$$C = 1 - \delta$$

attained for equiprobable inputs.

Oftentimes the designer is satisfied with not exceeding a certain fixed level of bit error rate, ϵ, rather than the more stringent criterion of vanishing probability of selecting the wrong block of data. In such a case, it is possible to transmit information at a rate equal to capacity times

$$\left(1 - \epsilon \log \frac{1}{\epsilon} - (1 - \epsilon) \log \frac{1}{1 - \epsilon}\right)^{-1}$$

which is shown in Fig. 73.20.

If, contrary to what we have assumed thus far, the message source in Fig. 73.13 is not a source of pure bits, the significance of capacity can be extended to show that as long as the source entropy (see Chapter 73.6 on

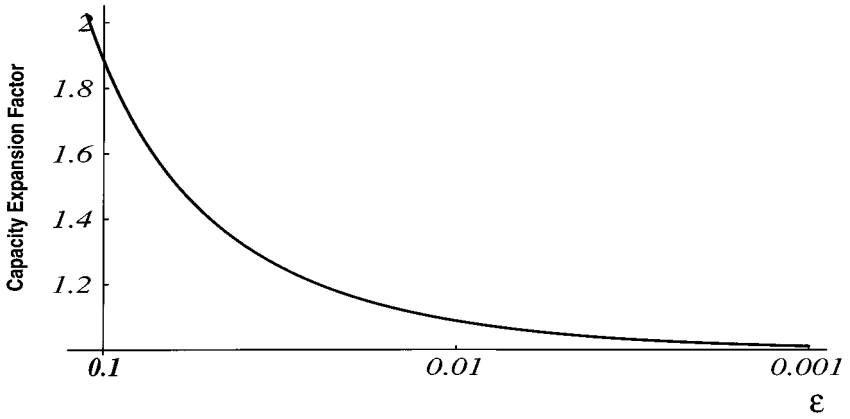

FIGURE 73.20 Capacity expansion factor as a function of bit-error-rate.

Data Compression) is below the channel capacity, an encoder/decoder pair exists that enables arbitrarily reliable communication. Conversely, if the source entropy is above capacity, then no such encoder/decoder pair exists.

Bandwidth and Capacity

The foregoing formulas for discrete channels do not lead to the capacity of continuous-time channels such as Example 5. We have seen that in the case of the telephone channel whose bandwidth is approximately equal to 3 kHz, capacity is lower bounded by 28,800 bits per second. How does bandwidth translate into capacity? The answer depends on the noise level and distribution. For example, if in the channel of Example 5, the noise is absent, capacity is infinite regardless of bandwidth. We can encode any amount of information as the binary expansion of a single scalar, which can be sent over the channel as the amplitude or phase of a single sinusoid; knowing the channel transfer function, the decoder can recover the transmitted scalar error-free. Clearly, such a transmission method is not recommended in practice because it hinges on the non-physical scenario of noiseless transmission.

In the simplest special case of Example 5, the noise is white, the channel has an ideal flat transfer function with bandwidth B (in Hz), and the input power is limited. Then, the channel capacity is equal to

$$C = B\ SNR^{dB} \log_2 10^{0.1} \quad \text{bits per second} \tag{73.80}$$

where $\log_2 10^{0.1} = 0.33$ and SNR^{dB} is equal to the optimum signal-to-noise ratio (in dB) of a linear estimate of a flat input signal given the channel output signal. Such an optimum signal-to-noise ratio is equal to one plus the power alloted to the input divided by the noise power in the channel band, i.e.,

$$SNR^{dB} = 10 \log_{10} \left(1 + \frac{P}{BN_0}\right).$$

It is interesting to notice that as the bandwidth grows, the channel capacity does not grow without bound. It tends to

$$\frac{P}{N_0} \log_2 e \qquad \text{bits per second}$$

where $\log_2 e = 1.44$. This means that the energy per bit necessary for reliable communication is equal to 0.69 times the noise power spectral density level. When the channel bandwidth is finite, the energy necessary to send one bit of information is strictly larger. The energy required to send one bit of information reliably can

be computed for other (non-Gaussian) channels even in cases where expressions for channel capacity are not known [Verdú, 1990].

When the channel transfer function $H(f)$ and/or noise spectral density $N(f)$ are not flat, the constant in Eq. (73.80) no longer applies. The so-called water-filling formula [Shannon, 1949] gives the channel capacity as

$$C = \frac{1}{2}\int \log\left(1 + \frac{\max\{0,\ w - M(f)\}}{M(f)}\right) df$$

where w is chosen so that

$$\int \max\{0,\ w - M(f)\}\, df = P,$$

and

$$M(f) = \frac{N(f)}{|H(f)|^2}.$$

The linear Gaussian-noise channel is a widely used model for space communication (in the power limited region) and for the telephone channel (in the bandwidth limited region). Thanks to the prevalence of digital switching and digital transmission in modern telephone systems, not only signal-to-noise ratios have improved over time but the Gaussian-noise model in Example 5 becomes increasingly questionable because quantization is responsible for a major component of the channel distortion. Therefore, future improvements in modem speeds are expected to arise mainly from finer modeling of the channel.

Due to the effect of time-varying received power (fading), several important channels fall outside the scope of Example 5 such as high-frequency radio links, tropospheric scatter links, and mobile radio channels.

Channel Coding Theorems

In information theory, the results that give a formula for channel capacity in terms of the probabilistic description of the channel are known as channel coding theorems. They typically involve two parts: an achievability part, which shows that codes with vanishing error probability exist for any rate below capacity; and a converse part, which shows that if the code rate exceeds capacity, then the probability of error is necessarily bounded away from zero. Shannon gave the first achievability results in [Shannon, 1948] for discrete memory channels. His method of proof, later formalized as the method of "typical sequences" (e.g., [Cover and Thomas, 1991]), is based on showing that the average probability of error of a code chosen at random vanishes with blocklength. Other known achievability proofs such as Feinstein's [1954], Gallager's [1968], and the method of types [Csiszar and Korner, 1981] are similarly non-constructive. The discipline of coding theory deals with constructive methods to design codes that approach the Shannon limit (see Chapter 71.1). The first converse channel coding theorem was not given by Shannon, but by Fano in 1952. A decade after Shannon's pioneering paper, several authors obtained the first channel coding theorems for channels with memory [Dobrushin, 1963]. The most general formula for channel capacity known to date can be found in [Verdú and Han, 1994]. The capacity of channels with feedback was first considered by Shannon in 1961 [Shannon, 1961], with later developments for Gaussian channels summarized in [Cover and Thomas, 1991]. In his 1961 paper [Shannon, 1961], Shannon founded the discipline of multiuser information theory by posing several challenging channels with more than one transmitter and/or receiver. In contrast to the **multiaccess channel** (one receiver) which has been solved in considerable generality, the capacities of channels involving more than one receiver, such as **broadcast channels** [Cover, 1972] and **interference channels** remain unsolved except in special cases.

Channel capacity has been shown to have a meaning outside the domain of information transmission [Han and Verdú, 1993]: it is the minimum rate of random bits required to generate any input random process so that the output process is simulated with arbitrary accuracy.

Defining Terms

Blocklength: The duration of a codeword, usually in the context of discrete-time channels.

Channel capacity: The maximum rate for which encoders and decoders exist whose probability of error vanishes as the codewords get longer and longer.

Codeword: Channel-input signal chosen by the encoder to represent the message.

Communication channel: Set of devices and systems that connect the transmitter to the receiver, not subject to optimization.

Broadcast channel: A communication channel with one input and several outputs each connected to a different receiver such that possibly different messages are to be conveyed to each receiver.

Discrete memoryless channel: A discrete-time memoryless channel where each channel input and output takes a finite number of values.

Discrete-time channel: A communication channel whose input/output signals are sequences of values. Its capacity is given in terms of bits per "channel use".

Continuous-time channel: A communication channel whose input/output signals are functions of a real variable (time). Its capacity is given in terms of bits per second.

Interference channel: A channel with several inputs/outputs such that autonomous transmitters are connected to each input and such that each receiver is interested in decoding the message sent by one and only one transmitter.

Memoryless channel: A channel where the conditional probability of the output given the current input is independent of all other inputs or outputs.

Multiaccess channel: A channel with several inputs and one output such that autonomous transmitters are connected to each input and such that the receiver is interested in decoding the messages sent by all the transmitters.

Decoder: Mapping from the set of channel-output signals to the set of messages.

Maximum-likelihood decoder: A decoder which selects the message that best explains the received signal, assuming all messages are equally likely.

Encoder: Mapping from the set of messages to the set of input codewords.

Modem: Device that converts binary information streams into electrical signals (and vice-versa) for transmission through the voiceband telephone channel.

Rate: The rate of a code is the number of bits transmitted (logarithm of code size) per second for a continuous-time channel or per channel use for a discrete-time channel.

Related Topics

70.1 Coding • 73.4 The Sampling Theorem

References

C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.,* 27, 379–423, 623–656, July–Oct. 1948.

R. E. Blahut, *Principles of Information Theory.* Reading, Mass.: Addison-Wesley, 1987.

S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Information Theory,* IT-36(5), 1019–1030, Sept. 1990.

C. E. Shannon, "Communication in the presence of noise," *Proc. Institute of Radio Engineers,* 37, 10–21, 1949.

T. M. Cover and J. A. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

A. Feinstein, "A new basic theorem of information theory," *IRE Trans. PGIT,* pp. 2–22, 1954.

R. G. Gallager, *Information Theory and Reliable Communication,* New York: Wiley, 1968.

I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* New York: Academic Press, 1981.
R. L. Dobrushin, *General Formulation of Shannon's Main Theorem in Information Theory,* American Mathematical Society Translations, pp. 323–438, 1963.
S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. on Information Theory,* 40(4), 1147–1157, July 1994.
C. E. Shannon, "Two-way communication channels," *Proc. 4th. Berkeley Symp. Math. Statistics and Prob.,* pp. 611–644, 1961.
T. M. Cover, "Broadcast channels," *IEEE Trans. on Information Theory,* pp. 2–14, Jan. 1972.
T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. on Information Theory,* IT-39, 752–772, May 1993.

Further Information

The premier journal and conference in the field of information theory are the *IEEE Trans. on Information Theory* and the *IEEE International Symposium on Information Theory,* respectively. *Problems of Information Transmission* is a translation of a Russian-language journal in information theory. The newsletter of the IEEE Information Theory Society regularly publishes expository articles.

73.6 Data Compression

Joy A. Thomas and Thomas M. Cover

Data compression is a process of finding the most efficient representation of an information source in order to minimize communication or storage. It often consists of two stages—the first is the choice of a (probabilistic) model for the source and the second is the design of an efficient coding system for the model. In this section, we will concentrate on the second aspect of the compression process, though we will touch on some common sources and models in the last subsection.

Thus, a data compressor (sometimes called a source coder) maps an information source into a sequence of bits, with a corresponding decompressor, that given these bits provides a reconstruction of the source. Data compression systems can be classified into two types: *lossless,* where the reconstruction is exactly equal to the original source, and *lossy,* where the reconstruction is a distorted version of the original source. For lossless data compression, the fundamental lower bound on the rate of the data compression system is given by the entropy rate of the source. For lossy data compression, we have a tradeoff between the rate of the compressor and the distortion we incur, and the fundamental limit is given by the rate distortion function, which is discussed later in this section.

Shannon [1948] was the first to distinguish the probabilistic model that underlies an information source from the semantics of the information. An information source produces one of many possible messages; the goal of communication is to transmit an unambiguous specification of the message so that the receiver can reconstruct the original message. For example, the information to be sent may be the result of a horse race. If the recipient is assumed to know the names and numbers of the horses, then all that must be transmitted is the number of the horse that won. In a different context, the same number might mean something quite different, e.g., the price of a barrel of oil. The significant fact is that the difficulty in communication depends only on the length of the representation. Thus, finding the best (shortest) representation of an information source is critical to efficient communication.

When the possible messages are all equally likely, then it makes sense to represent them by strings of equal length. For example, if there are 32 possible equally likely messages, then each message can be represented by a binary string of 5 bits. However, if the messages are not equally likely, then it is more efficient on the average to allot short strings to the frequently occurring messages and longer strings to the rare messages. Thus, the Morse code allots the shortest string (a dot) to the most frequent letter (E) and allots long strings to the infrequent letters (e.g., dash, dash, dot, dash for Q). The minimum average length of the representation is a fundamental quantity called the entropy of the source, which is defined in the next subsection.

Entropy

An information source will be represented by a random variable X, which takes on one of a finite number of possibilities $i \in \mathcal{X}$ with probability $p_i = \Pr(X = i)$. The entropy of the random variable X is defined as

$$H(X) = -\sum_{i \in \mathcal{X}} p_i \log p_i \tag{73.81}$$

where the log is to base 2 and the entropy is measured in *bits*. We will use logarithms to base 2 throughout this chapter.

Example 73.1 Let X be a random variable that takes on a value 1 with probability θ and takes on the value 0 with probability $1 - \theta$. Then $H(X) = -\theta \log \theta - (1 - \theta) \log (1 - \theta)$. In particular, the entropy of a fair coin toss with $\theta = \frac{1}{2}$ is 1 bit.

This definition of entropy is related to the definition of entropy in thermodynamics. It is the fundamental lower bound on the average length of a code for the random variable.

A **code** for a random variable X is a mapping from $\mathcal{X}$, the range of X, to the set of finite-length binary strings. We will denote the code word corresponding to i by $C(i)$, and the length of the code word by l_i. The average length of the code is then $L(C) = \Sigma_i p_i l_i$.

A code is said to be *instantaneous* or *prefix-free* if no code word is a prefix of any other code word. This condition is sufficient (but not necessary) to allow a sequence of received bits to be parsed unambiguously into a sequence of code words.

Example 73.2 Consider a random variable X taking on the values $\{1, 2, 3\}$ with probabilities (0.5, 0.25, 0.25). An instantaneous code for this random variable might be (0, 10, 11). Thus, a string 01001110 can be uniquely parsed into 0, 10, 0, 11, 10, which decodes to the string $x = (1, 2, 1, 3, 2)$. Note that the average length of the code is 1.5 bits, which is the same as the entropy of the source.

For any instantaneous code, the following property of binary trees called the *Kraft inequality* [Cover and Thomas, 1991].

$$\sum_i 2^{-l_i} \leq 1 \tag{73.82}$$

must hold. Conversely, it can be shown that given a set of lengths that satisfies the Kraft inequality, we can find a set of prefix-free code words of those lengths.

The problem of finding the best source code then reduces to finding the optimal set of lengths that satisfies the Kraft inequality and minimizes the average length of the code. Simple calculus can then be used to show [Cover and Thomas, 1991] that the average length of any instantaneous code is larger than the entropy of the random variable, i.e., the minimum of $\Sigma p_i l_i$ over all l_i satisfying $\Sigma\, 2^{-l_i} \leq 1$ is $-\Sigma p_i \log p_i$. Also, if we take $l_i = \lceil \log 1/p_i \rceil$ (where $\lceil t \rceil$ denotes the smallest integer greater than or equal to t), we can verify that this choice of lengths satisfies the Kraft inequality and that

$$L(C) = \sum_i p_i \left\lceil \log \frac{1}{p_i} \right\rceil < \sum_i p_i \left(\log \frac{1}{p_i} + 1 \right) = H(X) + 1 \tag{73.83}$$

The optimal code can only have a shorter length, and therefore we have the following theorem:

Theorem 73.1 Let L^* be the average length of the optimal instantaneous code for a random variable X. Then

$$H(X) \leq L^* < H(X) + 1 \tag{73.84}$$

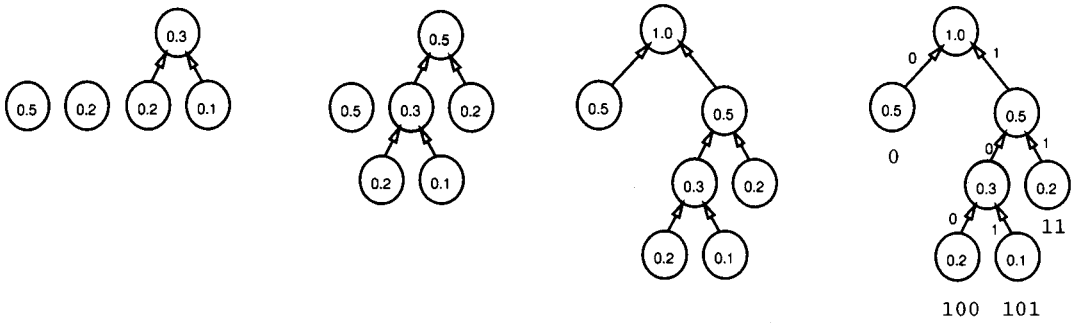

FIGURE 73.21 Example of the Huffman algorithm.

This theorem is one of the fundamental theorems of information theory. It identifies the entropy as the fundamental limit for the average length of the representation of a discrete information source and shows that we can find representations with average length within one bit of the entropy.

The Huffman Algorithm

The choice of code word lengths $l_i = \lceil \log 1/p_i \rceil$ (called the *Shannon code lengths*) is close to optimal, but not necessarily optimal, in terms of average code word length. We will now describe an algorithm (the **Huffman algorithm**) that produces an instantaneous code of minimal average length for a random variable with distribution $p_1, p_2, \ldots, p_m$. The algorithm is a greedy algorithm for building a tree from the bottom up.

Step 1. Arrange the probabilities in decreasing order so that $p_1 \geq p_2 \geq \ldots \geq p_m$.
Step 2. Form a subtree by combining the last two probabilities p_{m-1} and p_m to a single node of weight $p'_{m-1} = p_{m-1} + p_m$.
Step 3. Recursively execute Steps 1 and 2, decreasing the number of nodes each time, until a single node is obtained.
Step 4. Use the tree constructed above to allot code words.

The algorithm for tree construction is illustrated for a source with distribution (0.5, 0.2, 0.2, 0.1) in Fig. 73.21. After constructing the tree, the leaves of the tree (which correspond to the symbols of $\mathcal{X}$) can be assigned code words that correspond to the paths from the root to the leaf. We will not give a proof of the optimality of the Huffman algorithm; the reader is referred to Gallager [1968] or Cover and Thomas [1991] for details.

Entropy Rate

The entropy of a sequence of random variables $X_1, X_2, \ldots, X_n$ with joint distribution $p(x_1, x_2, \ldots, x_n)$ is defined analogously to the entropy of a single random variable as

$$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} p(x_1, x_2, \ldots, x_n)\log p(x_1, x_2, \ldots, x_n) \qquad (73.85)$$

For a stationary process $X_1, X_2, \ldots$, we define the *entropy rate* $\mathcal{H}(\mathcal{X})$ of the process as

$$\mathcal{H}(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \qquad (73.86)$$

It can be shown [Cover and Thomas, 1991] that the entropy rate is well defined for all stationary processes. In particular, if $X_1, X_2, \ldots, X_n$ is a sequence of independent and identically distributed (i.i.d.) random variables, then $H(X_1, X_2, \ldots, X_n) = nH(X_1)$, and $\mathcal{H}(\mathcal{X}) = H(X_1)$.

In the previous subsection, we showed the existence of a prefix-free code having an average length within one bit of the entropy. Now instead of trying to represent one occurrence of the random variable, we can form a code to represent a block of n random variables. In this case, the average code length is within one bit of $H(X_1, X_2, \ldots, X_n)$. Thus, the average length of the code per input symbol satisfies

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{L_n^*}{n} < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n} \tag{73.87}$$

Since $[H(X_1, X_2, \ldots, X_n)]/n \to \mathcal{H}(\mathcal{X})$, we can get arbitrarily close to the entropy rate by using longer and longer block lengths. Thus, the entropy rate is the fundamental limit for data compression for stationary sources, and we can achieve rates arbitrarily close to this limit by using long blocks.

All the above assumes that we know the probability distribution that underlies the information source. In many practical examples, however, the distribution is unknown or too complex to be used for coding. There are various ways to handle this situation:

- Assume a simple distribution and design an appropriate code for it. Use this code on the real source. If an estimated distribution $\hat{p}$ is used when in fact the true distribution is p, then the average length of the code is lower bounded by $H(X) + \Sigma_x p(x) \log [p(x)/\hat{p}(x)]$. The second term, which is denoted $D(p \| \hat{p})$ is called the *relative entropy* or the *Kullback Leibler distance* between the two distributions.
- Estimate the distribution empirically from the source and adapt the code to the distribution. For example, with *adaptive Huffman coding*, the empirical distribution of the source symbols is used to design the Huffman code used for the source.
- Use a *universal coding algorithm* like the **Lempel–Ziv algorithm** (see the subsection "Lempel–Ziv Coding").

Arithmetic Coding

In the previous subsections, it was shown how we could construct a code for a source that achieves an average length within one bit of the entropy. For small source alphabets, however, we have efficient coding only if we use long blocks of source symbols. For example, if the source is binary, and we code each symbol separately, we must use 1 bit per symbol, irrespective of the entropy of the source. If we use long blocks, we can achieve an expected length per symbol close to the entropy rate of the source.

It is therefore desirable to have an efficient coding procedure that works for long blocks of source symbols. Huffman coding is not ideal for this situation, since it is a bottom-up procedure with a complexity that grows rapidly with the block length. Arithmetic coding is an incremental coding algorithm that works efficiently for long block lengths and achieves an average length within one bit of the entropy for the block.

The essential idea of arithmetic coding is to represent a sequence $x^n = x_1, x_2, \ldots, x_n$ by the cumulative distribution function $F(x^n)$ (the sum of the probability of all sequences less than x^n) expressed to an appropriate accuracy. The cumulative distribution function for x^n is illustrated in Fig. 73.22. We can use any real number in the interval $[F(x^n) - p(x^n), F(x^n)]$ as the code for x^n. Expressing $F(x^n)$ to an accuracy of $\lceil \log 1/p(x^n) \rceil$ will give us a code for the source. The receiver can draw the cumulative distribution function, draw a horizontal line corresponding to the truncated value $\lfloor F(x^n) \rfloor$ that was sent, and read off the corresponding x^n. (This code is not prefix-free but can be easily modified to construct a prefix-free code [Cover and Thomas, 1991]). To implement arithmetic coding, however, we need efficient algorithms to calculate $p(x^n)$ and $F(x^n)$ to the appropriate accuracy based on a probabilistic model for the source. Details can be found in Langdon [1984] and Bell et al. [1990].

Lempel–Ziv Coding

The Lempel–Ziv algorithm [Ziv and Lempel, 1978] is a universal coding procedure that does not require knowledge of the source statistics and yet is asymptotically optimal. The basic idea of the algorithm is to construct a table or dictionary of frequently occurring strings and to represent new strings by pointing to their prefixes in the table. We first parse the string into sequences that have not appeared so far. For example, the

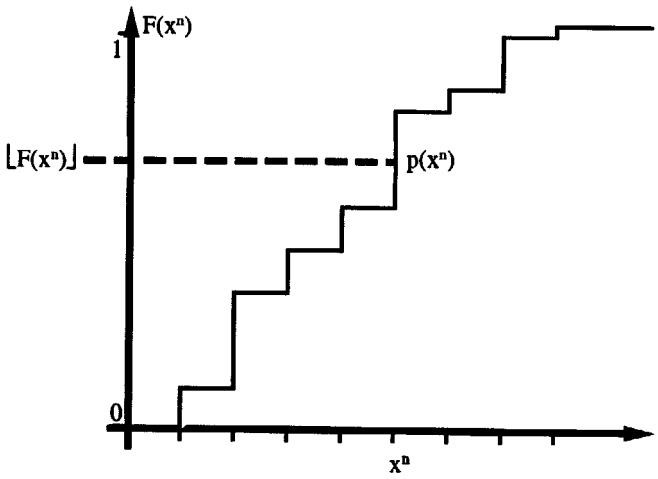

FIGURE 73.22 Cumulative distribution function for sequences x^n.

binary string 11010011011100 is parsed into 1,10,100,11,0,111,00. Then instead of sending the bits of each phrase, we send a pointer to its prefix and the value of the last bit. Thus, if we use three bits for the pointer, we will represent this string by (000,1), (001,0), (010,0), (001,1), (000,0), (100,1), (101,0), etc. For this short example, the algorithm has not compressed the string—it has in fact expanded it.

The very surprising fact is that, as Lempel and Ziv have shown, the algorithm is asymptotically optimal for any stationary ergodic source. This is expressed in the following theorem [Ziv and Lempel, 1978; Cover and Thomas, 1991]:

Theorem 73.2 Let L_n be the length of the Lempel–Ziv code for n symbols drawn from a stationary ergodic process $X_1, X_2, \ldots, X_n$ with entropy rate $\mathcal{H}(\mathcal{X})$. Then

$$\frac{L_n}{n} \to \mathcal{H}(\mathcal{X}) \quad \text{with probability 1} \tag{73.88}$$

Thus, for long enough block lengths, the Lempel–Ziv algorithm (which does not make any assumptions about the distribution of the source) does as well as if we knew the distribution in advance and designed the optimal code for this distribution.

The algorithm described above is only one of a large class of similar adaptive dictionary-based algorithms, which are all rather loosely called Lempel–Ziv. These algorithms are simple and fast and have been implemented in both software and hardware, e.g., in the *compress* command in UNIX and the *PKZIP* command on PCs. On ASCII text files, the Lempel–Ziv algorithm achieves compressions on the order of 50%. It has also been implemented in hardware and has been used to "double" the capacity of data storage media or to "double" the effective transmission rate of a modem. Many variations on the basic algorithm can be found in Bell et al. [1990].

Rate Distortion Theory

An infinite number of bits are required to describe an arbitrary real number, and therefore it is not possible to perfectly represent a continuous random variable with a finite number of bits. How "good" can the representation be? We first define a distortion measure, which is a measure of the distance between the random variable and its representation. We can then consider the trade-off between the number of bits used to represent a random variable and the distortion incurred. This trade-off is represented by the **rate distortion function** $R(D)$, which represents the minimum rate required to represent a random variable with a distortion D.

We will consider a discrete information source that produces random variables $X_1, X_2, \ldots, X_n$ that are drawn i.i.d. according to $p(x)$. (The results are also valid for continuous sources.) The encoder of the rate distortion

system of rate R will encode a block of n outputs X^n as an index $f(X^n) \in \{1, 2, \ldots, \lfloor 2^{nR} \rfloor\}$. (Thus, the index will require R bits/input symbol.) The decoder will calculate a representation $\hat{X}^n(f(X^n))$ of X^n. Normally, the representation alphabet $\hat{X}$ of the representation is the same as the source alphabet $\mathcal{X}$, but that need not be the case.

Definition: A *distortion function* or *distortion measure* is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow R^+ \tag{73.89}$$

from the set of source alphabet–reproduction alphabet pairs into the set of nonnegative real numbers. The distortion $d(x, \hat{x})$ is a measure of the cost of representing the symbol x by the symbol $\hat{x}$.

Examples of common distortion functions are

- *Hamming (probability of error) distortion.* The Hamming distortion is given by

$$d(x, \hat{x}) = \begin{cases} 0 \text{ if } x = \hat{x} \\ 1 \text{ if } x \neq \hat{x} \end{cases} \tag{73.90}$$

  and thus $Ed(X, \hat{X}) = \Pr(X \neq \hat{X})$.
- *Squared error distortion.* The squared error distortion

$$d(x, \hat{x}) = (x - \hat{x})^2 \tag{73.91}$$

  is the most popular distortion measure used for continuous alphabets. Its advantages are its simplicity and its relationship to least squares prediction. However, for information sources such as images and speech, the squared error is not an appropriate measure for distortion as perceived by a human observer.

The distortion between sequences x^n and $\hat{x}^n$ of length n is defined by

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{73.92}$$

For a rate distortion system, the expected distortion D is defined as

$$D = Ed(X^n, \hat{X}^n(f(X^n))) = \sum_{x^n} p(x^n) d(x^n, \hat{X}^n(f(x^n))) \tag{73.93}$$

Definition: The rate distortion pair (R,D) is said to be achievable if there exists a rate distortion code of rate R with expected distortion D. The *rate distortion function* $R(D)$ is the infimum of rates R such that (R,D) is achievable for a given D.

Definition: The *mutual information* $I(X, \hat{X})$ between random variables X and $\hat{X}$, with joint probability mass function $p(x, \hat{x})$ and marginal probability mass functions $p(x)$ and $p(\hat{x})$ is defined as

$$I(X; \hat{X}) = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \hat{\mathcal{X}}} p(x, \hat{x}) \log \frac{p(x, \hat{x})}{p(x)p(\hat{x})} \tag{73.94}$$

The mutual information is a measure of the amount of information that one random variable carries about another.

The main result of rate distortion theory is contained in the following theorem, which provides a characterization of the rate distortion function in terms of the mutual information of joint distributions that satisfy the expected distortion constraint:

Theorem 73.3 The rate distortion function for an i.i.d. source X with distribution $p(x)$ and distortion function $d(x, \hat{x})$ is

$$R(D) = \min_{p(\hat{x}|x):\Sigma_{(x,\hat{x})}p(x)p(\hat{x}|x)d(x,\hat{x})\le D} I(X; \hat{X}) \tag{73.95}$$

We can construct rate distortion codes that can achieve distortion D at any rate greater than $R(D)$, and we cannot construct such codes at any rate below $R(D)$.

The proof of this theorem uses ideas of random coding and long block lengths as in the proof of the channel capacity theorem. The basic idea is to generate a code book of 2^{nR} reproduction code words $\hat{X}^n$ at random and show that for long block lengths, for any source sequence, it is very likely that there is at least one code word in this code book that is within distortion D of that source sequence. See Gallager [1968] or Cover and Thomas [1991] for details of the proof.

Example 73.3 (*Binary source*) The rate distortion function for a Bernoulli (p) source (a random variable that takes on values $\{0, 1\}$ with probabilities p, $1 - p$) with Hamming distortion is given by

$$R(D) = \begin{cases} H(p) - H(D), & 0 \le D \le \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\} \end{cases} \tag{73.96}$$

where $H(p) = -p \log p - (1 - p)\log (1 - p)$ is the binary entropy function.

Example 73.4 (*Gaussian source*) The rate distortion function for a Gaussian random variable with variance σ^2 and squared error distortion is

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \le D \le \sigma^2 \\ 0, & D > \sigma^2 \end{cases} \tag{73.97}$$

Thus, with nR bits, we can describe n i.i.d. Gaussian random variables $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ with a distortion of $\sigma^2 2^{-2R}$ per symbol.

Quantization and Vector Quantization

The rate distortion function represents the lower bound on the rate that is needed to represent a source with a particular distortion. We now consider simple algorithms that represent a continuous random variable with a few bits. Suppose we want to represent a single sample from a continuous source. Let the random variable to be represented be X and let the representation of X be denoted as $\hat{X}(X)$. If we are given R bits to represent X, then the function $\hat{X}$ can take on 2^R values. The problem of optimum **quantization** is to find the optimum set of values for $\hat{X}$ (called the reproduction points or code points) and the regions that are associated with each value $\hat{X}$ in order to minimize the expected distortion.

For example, let X be a Gaussian random variable with mean 0 and variance σ^2, and assume a squared error distortion measure. In this case, we wish to find the function $\hat{X}(X)$ such that $\hat{X}$ takes on at most 2^R values and minimizes $E(X - \hat{X}(X))^2$. If we are given 1 bit to represent X, it is clear that the bit should distinguish whether $X > 0$ or not. To minimize squared error, each reproduced symbol should be at the conditional mean of its region. If we are given 2 bits to represent the sample, the situation is not as simple. Clearly, we want to divide the real line into four regions and use a point within each region to represent the samples within that region.

We can state two simple properties of optimal regions and reconstruction points for the quantization of a single random variable:

- Given a set of reconstruction points, the distortion is minimized by mapping a source random variable X to the representation $\hat{X}(w)$ that is closest to it (in distortion). The set of regions defined by this mapping is called a Voronoi or Dirichlet partition defined by the reconstruction points.
- The reconstruction points should minimize the conditional expected distortion over their respective assignment regions.

These two properties enable us to construct a simple algorithm to find a "good" quantizer: we start with a set of reconstruction points, find the optimal set of reconstruction regions (which are the nearest neighbor regions with respect to the distortion measure), then find the optimal reconstruction points for these regions (the centroids of these regions if the distortion measure is squared error), and then repeat the iteration for this new set of reconstruction points. The expected distortion is decreased at each stage in the algorithm, so the algorithm will converge to a local minimum of the distortion. This algorithm is called the *Lloyd algorithm* [Gersho and Gray, 1992].

It follows from the arguments of rate distortion theory that we will do better if we encode long blocks of source symbols rather than encoding each symbol individually. In this case, we will consider a block of n symbols from the source as a vector-valued random variable, and we will represent these n-dimensional vectors by a set of 2^{nR} code words. This process is called **vector quantization** (VQ). We can apply the Lloyd algorithm to design a set of representation vectors (the code book) and the corresponding nearest neighbor regions. Instead of using the probability distribution for the source to calculate the centroids of the regions, we can use the empirical distribution from a training sequence. Many variations of the basic vector quantization algorithm are described in Gersho and Gray [1992].

Common information sources like speech produce continuous waveforms, not discrete sequences of random variables as in the models we have been considering so far. By sampling the signal at twice the maximum frequency present (the Nyquist rate), however, we convert the continuous time signal into a set of discrete samples from which the original signal can be recovered (the sampling theorem). We can then apply the theory of rate distortion and vector quantization to such waveform sources as well.

Kolmogorov Complexity

In the 1960s, the Russian mathematican Kolmogorov considered the question "What is the intrinsic descriptive complexity of a binary string?" From the preceding discussion, it follows that if the binary string were a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$, then on the average it would take $nH(X)$ bits to represent the sequence. But what if the bits were the first million bits of the binary expansion of π? In that case, the string appears random but can be generated by a simple computer program. So if we wanted to send these million bits to another location which has a computer, we could instead send the program and ask the computer to generate these million bits. Thus, the descriptive complexity of π is quite small.

Motivated by such considerations, Kolmogorov defined the complexity of a binary string to be the length of the shortest program for a universal computer that generates that string. (This concept was also proposed independently and at about the same time by Chaitin and Solomonoff.)

Definition: The ***Kolmogorov complexity*** $K_{\mathcal{U}}(x)$ of a string x with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p) \tag{73.98}$$

the minimum length over all programs that print x and halt. Thus $K_{\mathcal{U}}(x)$ is the shortest description length of x over all descriptions interpreted by computer $\mathcal{U}$.

A universal computer can be thought of as a Turing machine that can simulate any other universal computer. At first sight, the definition of Kolmogorov complexity seems to be useless, since it depends on the particular

computer that we are talking about. But using the fact that any universal computer can simulate any other universal computer, any program for one computer can be converted to a program for another computer by adding a constant length "simulation program" as a prefix. Thus, we can show [Cover and Thomas, 1991] that for any two universal computers, $\mathcal{U}$ and $\mathcal{A}$,

$$\left|K_{\mathcal{U}}(x) - K_{\mathcal{A}}(x)\right| < c \tag{73.99}$$

where the constant c, though large, does not depend on the string x under consideration. Thus, Kolmogorov complexity is universal in that it does not depend on the computer (up to a constant additive factor).

Kolmogorov complexity provides a unified way to think about problems of data compression. It is also the basis of principles of inference (Occam's razor: "The simplest explanation is the best") and is closely tied with the theory of computability.

Data Compression in Practice

The previous subsections discussed the fundamental limits to compression for a stochastic source. We will now consider the application of these algorithms to some practical sources, namely, text, speech, images, and video. In real applications, the sources may not be stationary or ergodic, and the distributions underlying the source are often unknown. Also, in addition to the efficiency of the algorithm, important considerations in practical applications include the computational speed and memory requirements of the algorithm, the perceptual quality of the reproductions to a human observer, etc. A considerable amount of research and engineering has gone into the development of these algorithms, and many issues are only now being explored. We will not go into the details but simply list some popular algorithms for the different sources.

Text

English text is normally represented in ASCII, which uses 8 bits/character. There is considerable redundancy in this representation (the entropy rate of English is about 1.3 bits/character). Popular compression algorithms include variants of the Lempel–Ziv algorithm, which compress text files by about 50% (to about 4 bits/character).

Speech

Telephone quality speech is normally sampled at 8 kHz and quantized at 8 bits/sample (a rate of 64 kbits/s) for uncompressed speech. Simple compression algorithms like adaptive differential pulse code modulation (ADPCM) [Jayant and Noll, 1984] use the correlation between adjacent samples to reduce the number of bits used by a factor of two to four or more with almost imperceptible distortion. Much higher compression ratios can be obtained with algorithms like linear predictive coding (LPC), which model speech as an autoregressive process, and send the parameters of the process as opposed to sending the speech itself. With LPC-based methods, it is possible to code speech at less than 4 kbits/s. At very low bit rates, however, the reproduced speech sounds synthetic.

Images

A single high-quality color image of 1024 by 1024 pixels with 24 bits per pixel represents about 3 MB of storage in an uncompressed form, which will take more than 14 minutes to transmit over a 28800-baud modem. It is therefore very important to use compression to save storage and communication capacity for images. Many different algorithms have been proposed for image compression, and standards are still being developed for compression of images. For example, the popular GIF standard uses a patented version of Lempel–Ziv coding, and the JPEG standard being developed by the Joint Photographic Experts Group uses an 8 by 8 discrete cosine transform (DCT) followed by quantization (the quality of which can be chosen by the user) and Huffman coding. Newer compression algorithms using wavelets or fractals offer higher compression than JPEG. The compression ratios achieved by these algorithms are very dependent on the image being coded. The lossless compression methods achieve compression ratios of up to about 3:1, whereas lossy compression methods achieve ratios up to 50:1 with very little perceptible loss of quality.

Video

Video compression methods exploit the correlation in both space and time of the sequence of images to improve compression. There is a very high correlation between successive frames of a video signal, and this can be exploited along with methods similar to those used for coding images to achieve compression ratios up to 200:1 for high-quality lossy compression. Standards for full-motion video and audio compression are being developed by the Moving Pictures Experts Group (MPEG). Applications of video compression techniques include video-conferencing, multimedia CD-ROMs, and high-definition TV.

A fascinating and very readable introduction to different sources of information, their entropy rates, and different compression algorithms can be found in the book by Lucky [1989]. Implementations of popular data compression algorithms including adaptive Huffman coding, arithmetic coding, Lempel–Ziv and the JPEG algorithm can be found in Nelson and Gailly [1995].

Defining Terms

Code: A mapping from a set of messages into binary strings.

Entropy: A measure of the average uncertainty of a random variable. For a random variable with probability distribution $p(x)$, the entropy $H(X)$ is defined as $\Sigma_x - p(x) \log p(x)$.

Huffman coding: A procedure that constructs the code of minimum average length for a random variable.

Kolmogorov complexity: The minimum length description of a binary string that would enable a universal computer to reconstruct the string.

Lempel-Ziv coding: A dictionary-based procedure for coding that does not use the probability distribution of the source and is nonetheless asymptotically optimal.

Quantization: A process by which the output of a continuous source is represented by one of a set of discrete points.

Rate distortion function: The minimum rate at which a source can be described to within a given average distortion.

Vector quantization: Quantization applied to vectors or blocks of outputs of a continuous source.

Related Topics

17.1 Digital Image Processing • 69.5 Digital Audio Broadcasting

References

T. Bell, J. Cleary, and I. Witten, *Text Compression*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

T. M. Cover and J. A. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

R. Gallager, *Information Theory and Reliable Communication,* New York: Wiley, 1968.

A. Gersho and R. Gray, *Vector Quantization and Source Coding,* Boston: Kluwer Academic, 1992.

N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video,* Englewood Cliffs, N.J.: Prentice-Hall, 1984.

G. Langdon, "An introduction to arithmetic coding," *IBM Journal of Research and Development,* vol. 28, pp. 135–149, 1984.

R. Lucky, *Silicon Dreams: Information, Man and Machine,* New York: St. Martin's Press, 1989.

M. Nelson and J. Gailly, *The Data Compression Book,* 2nd ed., San Mateo, Calif.: M & T Books, 1995.

C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal,* vol. 27, pp. 379–423, 623–656, 1948.

J. Ziv and A. Lempel, "Compression of individual sequences by variable rate coding," *IEEE Trans. Inform. Theory,* vol. IT–24, pp. 530–536, 1978.

Further Information

Discussion of various data compression algorithms for sources like speech and images can be found in the *IEEE Transactions on Communications* and the *IEEE Transactions on Signal Processing*, while the theoretical underpinnings of compression algorithms are discussed in the *IEEE Transactions on Information Theory*.

Some of the latest developments in the areas of speech and image coding are described in a special issue of the *IEEE Journal on Selected Areas in Communications*, June 1992. It includes an excellent survey by N.S. Jayant of current work on signal compression, including various data compression standards.

Special issues of the *IEEE Proceedings* in June 1994 and February 1995 also cover some of the recent developments in data compression and image and video coding.

A good starting point for current information on compression on the World Wide Web is the FAQ for the newsgroup comp.compression, which can be found at

http://www.cis.ohio-state.edu/hypertext/faq/usenet/compression-faq/top.html.