INFORMATION THEORY

The primary goal of a communication system is to convey information-bearing messages from an information source to a destination over a communication channel. All real channels are subject to noise and other channel impairments that limit communication system performance. The receiver attempts to reproduce transmitted messages from the received distorted signals as accurately as possible.

In 1948, Shannon proposed a mathematical theory for the communication process. This theory, known as information theory, deals with the fundamental limits on the representation and transmission of information. Information theory was a remarkable breakthrough in that it provided a quantitative measure for the rather vague and qualitative notion of the amount of information contained in a message. Shannon suggested that the amount of information conveyed by the occurrence of an event is related to the uncertainty associated with it and was defined to be inversely related to the probability of occurrence of that event. Information theory also provides fundamental limits on the transmission of information and on the representation of information. These fundamental limits are employed as benchmarks and are used to evaluate the performance of practical systems by determining how closely these systems approach the fundamental limits.

In his celebrated work, Shannon laid the foundation for the design and analysis of modern communication systems. He proved that nearly error-free information transmission over a noisy communication link is possible by encoding signals prior to transmission over the link and by decoding the received signals. He only provided an existence proof stating that such procedures exist but did not specify an approach to design the best encoders and decoders. Also, he did not discuss the implementation complexity. These results have provided the impetus for researchers to try to design encoding and decoding procedures that approach the fundamental limits given by information theory.

While information theory was primarily developed as a mathematical model for communications, it has had an impact on a wide variety of fields that include physics, chemistry, biology, psychology, linguistics, statistics, economics, and computer science. For example, languages provide a means for communication between human beings, and application of information theory to linguistics arises naturally. Examples of application of information theory to computer science include the design of efficient decision trees and introduction of redundancy in computer systems to attain fault-tolerant computing.

COMMUNICATION SYSTEM MODEL

The main components of a digital communication system are shown in Fig. 1. The source is assumed to be a digital source in that a symbol from a finite alphabet is generated in discrete time. An analog source can be converted to a digital source by sampling and quantization. Data from the source are processed by the source encoder, which represents the source data in an efficient manner. The objective of the source encoding operation is to represent the source output in a compact form with as high fidelity as possible (i.e., with as little information loss as possible). The sequence of source codewords generated by the source encoder is fed to the channel encoder, which yields the sequence of channel codewords. The channel encoder adds redundancy to provide error control capabilities. The goal is to exploit the redundancy in the most effective manner by achieving a high degree of error control capability for a specified amount of redundancy. In some encoding schemes, the input data stream is divided into blocks of fixed length, and then some additional symbols are added to each block to yield channel codewords. These codes are known as block codes. In the class of codes known as tree codes, the encoding process exhibits memory in that a block of input data stream is encoded based on the past blocks also. In either case, the output of the channel encoder is a string of symbols to be transmitted. The modulator converts source codeword symbols to analog waveforms suitable for transmission over the channel. The received waveforms are distorted due to noise and other interference processes present over the channel. The demodulator converts the received waveform into symbols and then furnishes received words to the channel decoder. Due to channel noise, the received word may be in error. The channel decoder exploits the redundancy introduced at the channel encoder to detect and/ or correct errors in the received word. This corrected word is the best estimate of the source codeword, which is delivered to the destination after performing the inverse of the source encoding operation. Information theory is based on a probabilistic model of this communication system.

ENTROPY

Let the discrete random variable S represent the output of a source generating a symbol every signaling interval in a

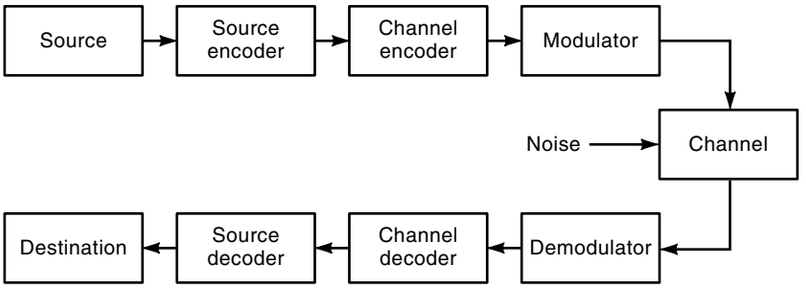

Figure 1. Block diagram of a communication system.

statistically independent manner. This discrete memoryless source (DMS) is assumed to generate symbols from a fixed finite alphabet $\{s_1, \ldots, s_K\}$ with probabilities $P(S = s_k) = p_k$, $k = 1, \ldots, K$. The amount of information gained after observing the symbol s_k is defined by the logarithmic function

$$I(s_k) = \log(1/p_k)$$

It is inversely related to the probability of a symbol occurrence. The base of the logarithm is usually taken to be 2 and the unit is called a bit. In this article, the base of all logarithms is assumed to be 2. Some properties of $I(s_k)$ are as follows:

1. If the outcome of an event is certain, no information gain occurs; that is,

$$I(s_k) = 0 \qquad \text{if} \qquad p_k = 1$$

2. Information gain from the occurrence of an event is nonnegative; that is,

$$I(s_k) \geq 0 \qquad \text{for} \qquad 0 \leq p_k \leq 1$$

3. Occurrence of less probable events results in more information gain; that is,

$$I(s_k) > I(s_\ell) \qquad \text{if} \qquad p_k < p_\ell$$

The average information per source symbol for a DMS is obtained by determining the average of $I(s_1), \ldots, I(s_K)$.

$$H(S) = \sum_{k=1}^{K} p_k \log(1/p_k)$$

This quantity is known as the entropy of the DMS. It characterizes the uncertainty associated with the source and is a function of source symbol probabilities. The entropy is bounded as $0 \leq H(S) \leq \log_2 K$. The lower bound is attained when one of the symbols occurs with probability one and the rest with probability zero. The upper bound is realized when all the symbols are equally likely.

Example: Consider a binary DMS whose output symbols are zero and one with associated probabilities of occurrence given by p_0 and p_1, respectively. The entropy is given by

$$H(S) = -p_0 \log p_0 - p_1 \log p_1$$

It is plotted in Fig. 2 as a function of p_0. Note that $H(S)$ is zero when $p_0 = 0$ or 1. This corresponds to no uncertainty. When $p_0 = \frac{1}{2}$, $H(S) = 1$. This corresponds to maximum uncertainty since symbols 0 and 1 are equally likely.

SOURCE CODING

One of the important problems in communications is an efficient representation of symbols generated by a DMS. Each symbol s_k is assigned a binary codeword of length ℓ_k. For an efficient representation, it is desirable to minimize the average codeword length $\overline{L}$, where

$$\overline{L} = \sum_{k=1}^{K} p_k \ell_k$$

Shannon's first theorem, also known as the source coding theorem, provides a fundamental limit on $\overline{L}$ in terms of the entropy of the source.

Source Coding Theorem: Given a DMS with entropy $H(S)$, the average codeword length $\overline{L}$ for any source encoding scheme is bounded as

$$\overline{L} \geq H(S)$$

Thus, entropy of a DMS provides a fundamental limit on the average number of bits per source symbol necessary to repre-

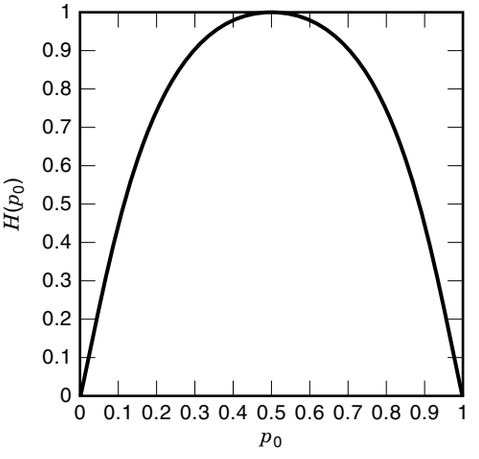

Figure 2. Binary entropy function.

Table 1. Illustration of Huffman Coding Algorithm

	Probabilities at Different Stages					
Source Symbols	1	2	3	4	5	Codewords
s_0	0.3	0.3	0.45	0.55	1.0	11
s_1	0.25	0.25	0.3	0.45		10
s_2	0.25	0.25	0.25			01
s_3	0.1	0.2				001
s_4	0.1					000

sent the DMS. Based on this lower bound on $\overline{L}$, we can express the coding efficiency of a source encoder as

$$\eta = \frac{H(S)}{\overline{L}}$$

A source encoder that is able to attain the lower bound has an efficiency of one.

An important requirement for source codes is that they be uniquely decodable so that perfect reconstruction is possible from the encoded binary sequence. One class of uniquely decodable codes is the class of prefix-free codes. In these codes, no codeword is a prefix of any other codeword. Huffman code is an example of such a source code in which $\overline{L}$ approaches $H(S)$. This code is optimum in that no other uniquely decodable code has a smaller $\overline{L}$ for a given DMS. The basic procedure for Huffman coding can be summarized as follows:

1. Arrange the source symbols in decreasing order of probabilities.
2. Assign a 0 and a 1 to the two source symbols with lowest probability.
3. Combine the two source symbols into a new symbol with probability equal to the sum of two original probabilities. Place this new symbol in the list according to its probability.
4. Repeat this procedure until there are only two source symbols in the list. Assign a 0 and a 1 to these two symbols.
5. Find the codeword for each source symbol by working backwards to obtain the binary string assigned to each source symbol.

Example: Consider a DMS with an alphabet consisting of five symbols with source probabilities, as shown in Table 1. Different steps of the Huffman encoding procedure and the resulting codewords are also shown. Codewords have been obtained by working backward on the paths leading to individual source symbol.

In this case,

$$\begin{aligned} H(S) = &-0.3\log 0.3 - 0.25\log 0.25 \\ &- 0.25\log 0.25 - 0.1\log 0.1 - 0.1\log 0.1 \\ = &\ 2.1855 \text{ bits/symbol} \end{aligned}$$

and $\overline{L} = 2.2$ bits/symbol. Thus, $\overline{L} > H(S)$ and $\eta = 0.9934$.

MUTUAL INFORMATION

Let X and Y be two discrete random variables that take values from $\{x_1, \ldots, x_J\}$ and $\{y_1, \ldots, y_K\}$, respectively. The conditional entropy $H(X|Y)$ is defined as

$$H(X|Y) = \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_j, y_k)\log[1/p(x_j|y_k)]$$

This quantity represents the amount of uncertainty remaining about X after observing Y. Since $H(X)$ represents the original uncertainty regarding X, information gained regarding X by observing Y is obtained by the difference of $H(X)$ and $H(X|Y)$. This quantity is defined as the mutual information $I(X; Y)$.

$$I(X;Y) = H(X) - H(X|Y)$$

Some important properties of $I(X; Y)$ are as follows:

1. The mutual information is symmetric with respect to X and Y; that is,

$$I(X;Y) = I(Y;X)$$

2. The mutual information is nonnegative; that is,

$$I(X;Y) \geq 0$$

3. $I(X; Y)$ is also given as

$$I(X;Y) = H(Y) - H(Y|X)$$

RELATIVE ENTROPY

The relative entropy or discrimination is a measure of the distance between two probability distributions. Let $p(\cdot)$ and $q(\cdot)$ be two probability mass functions. Then relative entropy or Kullback Leibler distance between the two is defined as

$$D(p\|q) = \sum_{k=1}^{K} p(x_k) \log \frac{p(x_k)}{q(x_k)}$$

The relative entropy is always nonnegative and is zero only if p and q are identical.

The mutual information $I(X; Y)$ can be interpreted as the relative entropy between the joint distribution $p(x_j, y_k)$ and the product distribution $p(x_j)\, p(y_k)$. That is,

$$I(X;Y) = D(p(x_j, y_k)\|p(x_j)p(y_k))$$

CHANNEL CAPACITY

Consider a discrete channel with input X and output Y, where X and Y are discrete random variables taking values from $(x_1, \ldots, x_J)$ and $(y_1, \ldots, y_K)$, respectively. This channel is known as a discrete memoryless channel (DMC) if the output symbol at any time depends only on the corresponding input symbol and not on any prior ones. This channel can be completely characterized in terms of channel transition probabilities, $p(y_k|x_j)$; $j = 1, \ldots, J$; $k = 1, \ldots, K$.

Example: An important example of a DMC is the binary symmetric channel (BSC) shown in Fig. 3. In this case, both the input and the output take values from {0, 1} and the two types of errors (receiving a zero when a one is sent, and receiving a one when a zero is sent) are equal.

For a DMC, mutual information $I(X; Y)$ is the amount of input source uncertainty reduced after observing the output. The channel capacity of a DMC is defined as the maximum mutual information for any signaling interval, where the maximization is performed over all possible input probability distributions. That is,

$$C = \max_{\{p(x_j)\}} I(X;Y)$$

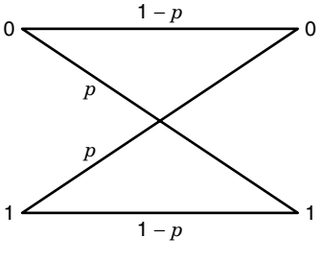

Figure 3. Binary symmetric channel.

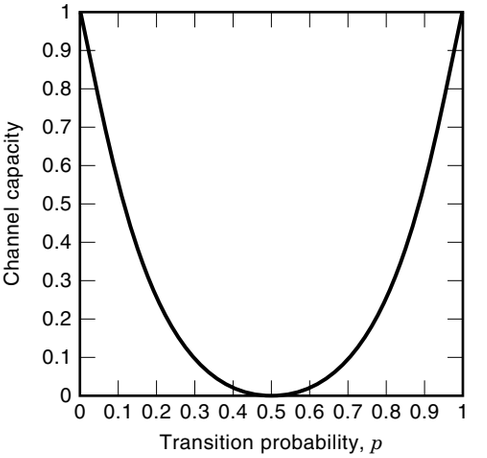

Figure 4. Capacity of a binary symmetric channel.

Channel capacity is a function only of the channel transition probabilities and its units are bits per channel use.

Example: The capacity of a BSC as a function of the error probability p is given by

$$C = 1 - H(p)$$

and is shown in Fig. 4. When $p = 0$ or $p = 1$, the channel capacity is maximum and is equal to 1 bit. Note that $p = 1$ also corresponds to a deterministic channel in that a zero is always received as a one and vice versa. When $p = \frac{1}{2}$, the channel is very noisy and the capacity is zero.

CHANNEL CODING THEOREM

To combat the effects of noise during transmission, the incoming data sequence from the source is encoded into a channel input sequence by introducing redundancy. At the receiver, the received sequence is decoded to reconstruct the data sequence. Shannon's second theorem, also known as the channel coding theorem or the noisy coding theorem, provides the fundamental limits on the rate at which reliable information transmission can take place over a DMC.

Channel Coding Theorem

(i) Let a DMS with entropy $H(S)$ produce a symbol every T_s seconds. Let a DMC have capacity C and be used once every T_c seconds. Then, if

$$\frac{H(S)}{T_s} \leq \frac{C}{T_c}$$

there exists a coding scheme with which source output can be transmitted over the channel and be reconstructed at the receiver with an arbitrarily small probability of error. Here, error refers to the event that a transmitted symbol is reconstructed incorrectly.

(ii) Conversely, if

$$\frac{H(S)}{T_s} > \frac{C}{T_c}$$

it is not possible to transmit data with an arbitrarily small probability of error.

It must be emphasized that the foregoing result only states the existence of "good" codes but does not provide methods to construct such codes. Development of efficient codes has remained an active area of research and is discussed elsewhere in this volume. In error-control coding, redundant symbols are added to the transmitted information at the transmitter to provide error detection and error correction capabilities at the receiver. Addition of redundancy implies increased data rate and thus an increased transmission bandwidth.

DIFFERENTIAL ENTROPY

Thus far, only discrete random variables were considered. Now we define information theoretic quantities for continuous random variables. Consider a continuous random variable X with probability density function $f(x)$. Analogous to the entropy of a discrete random variable, the differential entropy of a continuous random variable X is defined as

$$h(x) = \int_{-\infty}^{\infty} f(x) \log[1/f(x)]\, dx$$

Example: For a Gaussian random variable with probability density function,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}$$

the differential entropy can be computed to be

$$h(x) = \frac{1}{2} \log 2\pi e\sigma^2 \text{ bits}$$

In an analogous manner, mutual information for two continuous random variables X and Y can be defined as

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy$$

CAPACITY OF GAUSSIAN CHANNELS

Earlier, the fundamental limit on error-free transmission over a DMC was presented. Here we present the channel capacity theorem for band-limited and power-limited Gaussian channels. This theorem is known as Shannon's third theorem or as the Shannon–Hartley theorem. It is an extremely important result with great practical relevance because it expresses the channel capacity in terms of system parameters channel bandwidth, average signal power, and noise power spectral density.

Channel Capacity Theorem: The capacity of a band-limited additive white Gaussian noise (AWGN) channel is given by

$$C = B \log\left(1 + \frac{P}{N_0 B}\right) \text{ bits/s}$$

where B is the bandwidth of the channel, P is the average transmitted signal power, and the noise power spectral density is equal to $N_0/2$.

The capacity provides a fundamental limit on the rate at which information can be transmitted with arbitrarily small probability of error. Conversely, information cannot be transmitted at a rate higher than C bits/s with arbitrarily small probability of error irrespective of the coding scheme employed.

RATE DISTORTION THEORY

Previously, the problem of source coding that required perfect reconstruction of a DMS was considered. It was seen that the entropy provided the minimum rate at which perfect reconstruction is possible. A question arises as to what happens when the allowed rate is less than the lower bound. Also, what if the source is continuous, because a finite representation of such a source can never be perfect? These questions give rise to rate distortion theory. A distortion measure needs to be defined to quantify the distance between the random variable and its representation. For a given source distribution and distortion measure, the fundamental problem in rate distortion theory is to determine the minimum achievable expected distortion at a given rate. An equivalent problem is to find the minimum rate required to attain a given distortion. This theory is applicable to both continuous and discrete random variables.

Consider a source with alphabet $\mathscr{X}$ that produces a sequence of independent identically distributed random variables $X_1, X_2, \ldots$. Let $\hat{X}_1, \hat{X}_2, \ldots$ be the corresponding reproductions with reproduction alphabet denoted as $\hat{\mathscr{X}}$. The single-letter distortion measure $d(x, \hat{x})$ is a mapping d: $\mathscr{X} \times \hat{\mathscr{X}} \rightarrow R^+$ from the source alphabet-reproduction alphabet pair into the set of nonnegative real numbers. It quantifies the distortion when x is represented by $\hat{x}$. Two commonly used distortion measures are as follows:

Hamming distortion measure:

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

Squared error distortion measure:

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

The single-letter distortion measure can be extended to define the distortion measure for n-tuples as follows:

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

This is the average of the per symbol distortion over the elements of the n-tuple.

Now we consider the encoding of the source output sequence of length n, X^n, and then its decoding to yield $\hat{X}^n$. To accomplish this we define a $(2^{nR}, n)$ rate distortion code that consists of an encoding function and a decoding function, as

given by

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$$
$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$$

where R is the number of bits available to represent each source symbol. The expected distortion for this rate distortion code is given by

$$D_n = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n)))$$

where $p(\cdot)$ is the probability density function associated with the source.

A rate distortion pair (R, D) is said to be achievable if there exists a rate distortion code with rate R such that

$$\lim_{n \to \infty} D_n \leq D$$

The rate distortion function $R(D)$ is the infimum of rates R such that (R, D) is achievable for a given D. Next, we present the fundamental theorem of rate distortion theory.

Rate Distortion Theorem: The rate distortion function for an independent identically distributed source X with distribution $p(x)$ and bounded distortion function $d(x, \hat{x})$ is given by

$$R(D) = \min_{p(\hat{x}|x):\ \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X})$$

Thus, $R(D)$ is the minimum achievable rate at distortion D. Conversely, if R is less than $R(D)$, we cannot achieve a distortion less than or equal to D.

Example: Consider a binary source that produces an output of 1 with probability p. For the Hamming distortion measure, its $R(D)$ is given by

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

It is illustrated in Fig. 5 for $p = 0.5$.

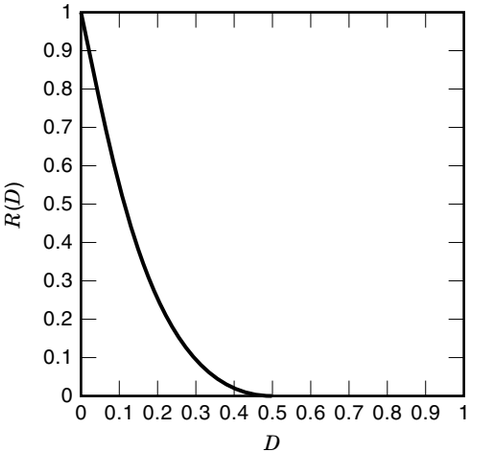

Figure 5. Rate distortion function for the binary source.

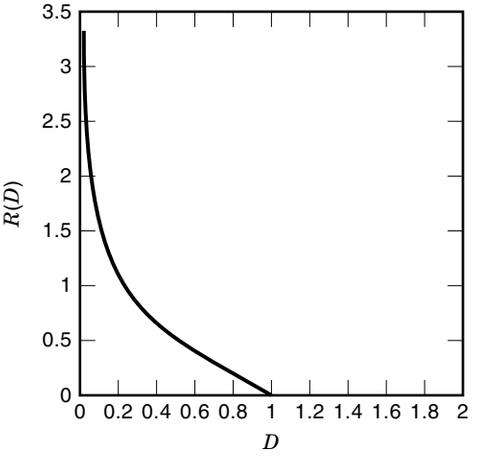

Figure 6. Rate distortion function for the Gaussian source.

Example: Consider a zero-mean Gaussian source with variance σ^2. For the squared error distortion measure, the rate distortion function is given by

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

It is plotted in Fig. 6.

The rate distortion function $R(D)$ is a nonincreasing convex function of D. For the binary source, when $D = 0$, the minimum rate required for perfect reconstruction is given by $H(p)$. As D increases, minimum required rate R decreases. Similar observations can also be made for the Gaussian source.

ACKNOWLEDGMENT

I would like to thank Qian Zhang for his help in the preparation of this article. This article was written while the author was a Visiting Scientist at the Air Force Research Laboratory at Rome Research Site, AFRL/IFG, 525 Brooks Road, Rome, NY 13441-4505.

BIBLIOGRAPHY

For a detailed discussion of information theory and its applications, the reader is referred to the sources listed below. Recent results on this topic are published in the *IEEE Transactions on Information Theory.*

T. Berger, *Rate Distortion Theory,* Englewood Cliffs, NJ: Prentice-Hall, 1971.

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

R. E. Blahut, *Theory and Practice of Error Control Codes,* Reading, MA: Addison-Wesley, 1983.

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

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

R. W. Hamming, *Coding and Information Theory,* Englewood Cliffs, NJ: Prentice-Hall, 1980.

S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications,* Englewood Cliffs, NJ: Prentice-Hall, 1983.

M. Mansuripur, *Introduction to Information Theory,* Englewood Cliffs, NJ: Prentice-Hall, 1987.

R. J. McEliece, *The Theory of Information Theory and Coding,* Reading, MA: Addison-Wesley, 1977.

C. E. Shannon, A Mathematical Theory of Communication, *Bell Syst. Techn. J.,* vol. 27, pp. 379–423 (part I), and pp. 623–656 (Part II), 1949.

PRAMOD K. VARSHNEY
Syracuse University

INFORMATION THEORY. See MAXIMUM LIKELIHOOD DETECTION.