J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*

INFORMATION THEORY OF STOCHASTIC PROCESSES

This article starts by acquainting the reader with the basic features in the design of a data communication system and discusses, in general terms, how the information theory of stochastic processes can aid in this design process. At the start of the data communication system design process, the communication engineer is given a source, which generates information, and a noisy channel through which this information must be transmitted to the end user. The communication engineer must then design a data communication system so that the information generated by the given source can be reliably transmitted to the user via the given channel. System design consists in finding an encoder and decoder through which the source, channel, and end user can be linked as illustrated in Fig. 1.

To achieve the goal of reliable transmission, the communication engineer can use discrete-time stochastic processes to model the sequence of source outputs, the sequence of channel inputs, and the sequence of channel outputs in response to the channel inputs. The probabilistic behavior of these processes can then be studied over time. These behaviors will indicate what level of system performance can be achieved by proper encoder/decoder design. Denoting the source in Fig. 1 by (S and denoting the channel in Fig. 1 by C, one would like to know the rate $R(S)$ at which the source generates information, and one would like to know the maximum rate $R(C)$ at which the channel can reliably transmit information. If $R(S) \leq R(C)$, the design goal of reliable transmission of the source information through the given channel can be achieved.

Information theory enables one to determine the rates $R(S)$ and R (C). Information theory consists of two subareas—*source coding theory* and *channel coding theory*. Source coding theory concerns itself with the computation of R (S) for a given source model S, and channel coding theory concerns itself with the computation of $R(C)$ for a given channel model C.

Suppose that the source generates an output U_i at each discrete instant of time $i = 1, 2, 3, \ldots$. The discrete-time stochastic process $\{U_\mathrm{i}: i \geq 1\}$ formed by these outputs may obey an information-theoretic property called the *asymptotic equipartition property*, which will be discussed in the section entitled "Asymptotic Equipartition Property." The asymptotic equipartition property will be applied to source coding theory in the section entitled "Application to Source Coding Theory." If the asymptotic equipartition property is satisfied, there is a nice way to characterize the rate $R(S)$ at which the source S generates information over time.

Suppose that the channel generates a random output Y_i at time i in response to a random input X_i at time i, where $i = 1, 2, 3, \ldots$. The discrete-time stochastic process $\{(X_\mathrm{i}, Y_\mathrm{i}): i \geq 1\}$ consisting of the channel input–output pairs (called a *channel pair process*) may obey an information-theoretic property called the *information stability property*, which shall be discussed in the section entitled "Information Stability Property." The information stability property will be applied to channel coding theory in the section entitled "Application to Channel Coding Theory." If sufficiently many channel pair processes obey the information stability property, there will be a nice way to characterize the rate R (C) at which the channel C can reliably transmit information.

In conclusion, the information theory of stochastic processes consists of the development of the asymptotic equipartition property and the information stability property. In this article we discuss these properties, along with their applications to source coding theory and channel coding theory.

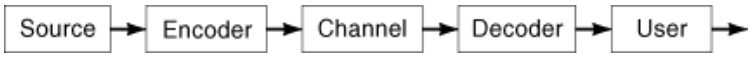

Fig. 1. Block diagram of data communication system.

Asymptotic Equipartition Property

If the asymptotic equipartition property holds for a random sequence $\{Ui: i \geq 1\}$, then, for large n, the random vector $(U_1, U_2, \ldots, U_n)$ will be approximately uniformly distributed. In order to make this idea precise, we must first discuss the concept of entropy.

Entropy. Let U be a discrete random variable. We define a nonnegative random variable $h(U)$, which is a function of U, so that

$$h(U) = -\log \Pr[U = u]$$

whenever $U = u$. The logarithm is taken to base two (as are all logarithms in this article). Also, we adopt the convention that $h(U)$ is defined to be zero, whenever $\Pr[\ U = u] = 0$. The random variable $h(U)$ is called the *self-information* of U.

The expected value of $h(U)$ is called the *entropy* of U and is denoted $H(U)$. In other words,

$$H(U) = E[h(U)] = \sum_u -\Pr[U = u] \log \Pr[U = u]$$

where E (here and elsewhere) denotes the expected value operator. Certainly, $H(U)$ satisfies

$$0 \leq H(U) \leq \infty$$

We shall only be interested in the finite entropy case in which $H(U) < \infty$. One can deduce that U has finite entropy if U takes only finitely many values. Moreover, the bound

$$H(U) \leq \log N \qquad (1)$$

holds in this case, where N is the number of values of U. To see why Eq. (1) is true, we exploit *Shannon's inequality*, which says

$$-\sum_u p(u) \log\ p(u) \leq -\sum_u p(u) \log q(u) \qquad (2)$$

whenever $\{p(u)\}$ and $\{q(u)\}$ are probability distributions on the space in which U takes its values. In Shannon's inequality, take

$$\begin{aligned} p(u) &= \Pr[U = u] \\ q(u) &= 1/N \end{aligned}$$

for each value u of U, thereby obtaining Eq. (1). If the discrete random variable U takes on a countably infinite number of values, then $H(U)$ may or may not be finite, as the following examples show.

Example 1. Let the set of values of U be 2, 3, 4, ...}, and let

$$\Pr[U = u] = \frac{C}{u(\log u)^2}$$

for every value u of U, where C is the normalization constant that makes these probabilities sum to one. It can be verified that $H(U) = \infty$.

Example 2. Let U follow a geometric distribution

$$\Pr[U = u] = p^{u-1}(1-p), \quad u = 1, 2, 3, \ldots$$

where p is a parameter satisfying $0 < p < 1$. It can be verified that

$$H(U) = \frac{-p \log p - (1-p)\log(1-p)}{1-p} < \infty$$

We are now ready to discuss the asymptotic equipartition property. Let $\{U_i\colon i \geq 1\}$ be a discrete-time stochastic process, in which each random variable U_i is discrete. For each positive integer n, let U^n denote the random vector $(U_1, U_2, \ldots, U_n)$. (This notational convention shall be in effect throughout this article.) We assume that the process $\{U_i\colon i \geq 1\}$ obeys the following two properties:

(1) $H(U^n) < \infty$, $n \geq 1$.
(2) The sequence $\{H(U^n)/n\colon n \geq 1\}$ has a finite limit.

Under this assumption, we can define a nonnegative real number $\bar{H}$ by

$$\bar{H} \triangleq \lim_{n\to\infty} n^{-1} H(U^n)$$

The number $\bar{H}$ is called the *entropy rate* of the process $\{U_i\colon i \geq 1\}$. Going further, we say that the process $\{U_i\colon i \geq 1\}$ obeys the *asymptotic equipartition property* (AEP) if

$$\lim_{n\to\infty} Pr[|n^{-1}h(U^n) - \bar{H}| > \epsilon] = 0, \quad \forall \epsilon > 0 \qquad (3)$$

What does the AEP tell us? Let ε be a fixed, but arbitrary, positive real number. The AEP implies that we may find, for each positive integer n, a set E_n consisting of certain n-tuples in the range of the random vector U^n, such that the sets $\{E_n\}$ obey the following properties:

$$2^{-n(\bar{H}+\epsilon)} \leq Pr[U^n = u^n] \leq 2^{-n(\bar{H}-\epsilon)}$$

(2.3) $\lim n\to\infty \Pr[U^N \in E_N] = 1$. For each n, if u^n is an n-tuple in E_n, then (2.5) For sufficiently large n, if $|E_n|$ is the number of n-tuples in E_n, then

$$2^{n(\bar{H}-\epsilon)} \leq |E_n| \leq 2^{n(\bar{H}+\epsilon)}$$

In loose terms, the AEP says that for large n, U^n can be modeled approximately as a random vector taking roughly 2^{nH} equally probable values. We will apply the AEP to source coding theory in the section entitled "Application to Source Coding Theory."

Example 3. Let $\{U_i : i \geq 1\}$ consist of independent and identically distributed (IID) discrete random variables. Letting $H(U_1) < \infty$, assumptions (2.1) and (2.2) hold, and the entropy rate is $\bar{H} = H(U_1)$. By the law of large numbers, the AEP holds.

Example 4. Let $\{U_i: i \geq 1\}$ be a stationary, ergodic homogeneous Markov chain with finite state space. Assumptions (2.1) and (2.2) hold, and the entropy rate is given by $\bar{H} = H(U^2) - H(U_1)$. Shannon (1) proved that the AEP holds in this case.

Extensions. McMillan (2) established the AEP for a stationary ergodic process $\{U_i: i \geq 1\}$ with finite alphabet. He established L^1 convergence, namely, he proved that

$$\lim_{n\to\infty} E[|n^{-1}h(U^n) - \bar{H}|] = 0$$

which is a stronger notion of convergence than the notion of convergence in Eq. (3). In the literature, McMillan's result is often referred to as the Shannon–McMillan Theorem. Breiman (3) proved almost sure convergence of the sequence $\{n^{-1}h(U^n): n \geq 1\}$ to the entropy rate $\bar{H}$:, for a stationary ergodic finite alphabet process $\{U_i : i \geq 1\}$. This is also a notion of convergence that is stronger than Eq. (3). Breiman's result is often referred to as the Shannon–McMillan–Breiman Theorem. Gray and Kieffer (4) proved that a type of nonstationary process called an asymptotically mean stationary process obeys the AEP. Verdú and Han (5) extended the AEP to a class of information sources called flat-top sources. Many other extensions of the AEP are known. Most of these results fall into one of the three categories described below.

(1) *AEP for Random Fields.* A random field $\{U_g: g \in G\}$ is given in which G is a countable group, and there is a finite set A such that each random variable U_g takes its values in A. A sequence $\{F_n: n \geq 1\}$ of growing finite subsets of G is given in which, for each n, the number of elements of F_n is denoted by $|F_n|$. For each n, let U_n denote the random vector

$$U^{F_n} \triangleq (U_g : g \in F_n)$$

One tries to determine conditions on $\{U_g\}$ and $\{F_n\}$ under which the sequence of random variables $\{|F_n|^{-1}h(U^{\mathrm{Fn}}): n \geq 1\}$ converges to a constant. Results of this type are contained in Refs. (6) (L^1 convergence) and (7) (almost sure convergence).

(2) *Entropy Stability for Stochastic Processes.* Let $\{U_i: i \geq 1\}$ be a stochastic process in which each random variable U_i is real-valued. For each $n = 1, 2, \ldots$, suppose that the distribution of the random vector U_{n} is absolutely continuous, and let F_n be its probability density function. For each n, let g_{n} be an n-dimensional probability density function different from F_n. One tries to determine conditions on $\{U_i\}$ and $\{g_{\mathrm{n}}\}$ under which the sequence of random variables

$$\left\{n^{-1}\log\frac{f_n(U^n)}{g_n(U^n)} : n \geq 1\right\}$$

converges to a constant. A process $\{U_i: i \geq 1\}$ for which such convergence holds is said to exhibit the *entropy stability property* (with respect to the sequence of densities $\{g_{\mathrm{n}}\}$). Perez (8) and Pinsker [(9), Sections 7.6, 8.4, 9.7, 10.5, 11.3] were the first to prove theorems showing that certain types of processes $\{U_i: i \geq 1\}$

exhibit the entropy stability property. Entropy stability has been studied further (10 11 12 13 14,15. In the textbook (16), Chapters 7 and 8 are chiefly devoted to entropy stability.

(3) *Entropy Stability for Random Fields.* Here, we describe a type of result that combines types (i) and (ii). As in (i), a random field $\{U_g: g \in G\}$ and subsets $\{F_n: n \geq 1\}$ are given, except that it is now assumed that each random variable U_g is real-valued. It is desired to find conditions under which the sequence of random variables

$$\left\{ |F_n|^{-1} \log \frac{f_n(U^{F_n})}{g_n(U^{F_n})} : n \geq 1 \right\}$$

converges to a constant, where, for each n, F_n is the probability density function of the $|F_n|$-dimensional random vector U^Fn and g_n is some other $|F_n|$-dimensional probability density function. Tempelman (17) gave a result of this type.

Further Reading. In this article, we have focused on the application of the AEP to communication engineering. It should be mentioned that the AEP and its extensions have been exploited in many other areas as well. Some of these areas are ergodic theory (18,19), differentiable dynamics (20), quantum systems (21), statistical thermodynamics (22), statistics (23), and investment theory (24).

Information Stability Property

The information stability property is concerned with the asymptotic information-theoretic behavior of a pair process, that is, a stochastic process $\{(X_i, Y_i): i \geq 1\}$ consisting of pairs of random variables. In order to discuss the information stability property, we must first define the concepts of mutual information and information density.

Mutual Information. Let X, Y be discrete random variables. The mutual information between X and Y, written $I(X; Y)$, is defined by

$$I(X;Y) \triangleq \sum_{x,y} \Pr[X = x, Y = y] \log \frac{\Pr[X = x, Y = y]}{\Pr[X = x] \Pr[Y = y]}$$

where we adopt the convention that all terms of the summation in which $\Pr[X = x, Y = y] = 0$ are taken to be zero. Suppose that X, Y are random variables that are not necessarily discrete. In this case, the mutual information $I(X; Y)$ is defined as

$$I(X;Y) = \sup_{(X_d, Y_d)} I(X_d; Y_d)$$

where the supremum is taken over all pairs of random variables (X_d, Y_d) in which X_d, Y_d are discrete functions of X, Y, respectively. From Shannon's inequality, Eq. (2), $I(X; Y)$ is either a nonnegative real number or is $+\infty$. We shall only be interested in mutual information when it is finite.

Example 5. Suppose X and Y are independent random variables. Then $I(X;Y) = 0$. The converse is also true.

Example 6. Suppose X is a discrete random variable. The inequality

$$I(X;Y) \leq \min[H(X), H(Y)]$$

always holds. From this inequality, we see that if $H(X)$ or $H(Y)$ is finite, then $I(X;Y)$ is finite. In particular, we see that $I(X;Y)$ is finite if either X or Y take finitely many values.

Example 7. Suppose X, Y are real-valued random variables, with variances $\sigma^2\mathrm{x} > 0$, $\sigma^2{}_\mathrm{y} > 0$, respectively. Let (X, Y) have a bivariate Gaussian distribution, and let ρ_{xy} be the correlation coefficient, defined by

$$\rho_{xy} \triangleq \frac{E[XY] - E[X]E[Y]}{\sigma_x \sigma_y}$$

It is known (9, p. 123) that

$$I(X;Y) = -(1/2)\log(1 - \rho_{xy}^2)$$

In this case, we conclude that $I(X;Y) < \infty$ if and only if $-1 < \rho\mathrm{xy} < 1$.

Example 8. Suppose X and Y are real-valued random variables, and that (X, Y) has an absolutely continuous distribution. Let $f(X, Y)$ be the density function of (X, Y), and let $f(X)$ and $g(Y)$ be the marginal densities of X, Y, respectively. It is known (9, p. 10) that

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \log \frac{f(x,y)}{f(x)g(y)}\, dx\, dy \qquad (4)$$

Information Density. We assume in this discussion that X, Y are random variables for which $I(X;Y) < \infty$. The *information density* $i(X;Y)$ of the pair (X, Y) shall be defined to be a random variable, which is a function of (X, Y) and for which

$$I(X;Y) = E[i(X;Y)] \qquad (5)$$

In other words, the expected value of the information density is the mutual information. Let us first define the information density for the case in which X and Y are both discrete random variables. If $X = X$ and $Y = Y$, we define

$$i(X;Y) \triangleq \begin{cases} \log \dfrac{\Pr[X = x, Y = y]}{\Pr[X = x]\Pr[Y = y]}, & \Pr[X = x, Y = y] > 0 \\ 0, & \text{otherwise} \end{cases}$$

Now suppose that X, Y are not necessarily discrete random variables. The information density of the pair (X, Y) can be defined (16, Chap. 5) as the unique random variable $I(X;Y)$ such that, for any $\varepsilon > 0$, there exist discrete random variables X^ε, Y^ε, functions of X, Y, respectively, such that

$$E[|i(X', Y') - i(X;Y)|] < \epsilon$$

whenever X', Y' are discrete random variables such that

- X^ε is a function of X' and X' is a function of X.
- Y^ε is a function of Y' and Y' is a function of Y.

Example 9. In Example 8, if $I(X; Y) < \infty$, then

$$i(X;Y) = \log \frac{f(X,Y)}{f(X)g(Y)}$$

Example 10. If X is a discrete random variable with finite entropy, then

$$I(X;X) = H(X)$$

$$i(X;X) = h(X)$$

We are now ready to discuss the information stability property. Let $\{(X_i, Y_i)\colon i \geq 1\}$ be a pair process satisfying the following two properties:

(1) (10.1)$I(X^n; Y^n) < \infty$, $n \geq 1$.
(2) (10.2)The sequence $\{n^{-1} I(X^n; Y^n)\colon n \geq 1\}$ has a finite limit.

We define the *information rate* of the pair process $[(X_i, Y_i)\colon I \geq 1\ ]$ to be the nonnegative real number

$$\bar{I} \triangleq \lim_{n\to\infty} n^{-1} I(X^n;Y^n)$$

A pair process $[(X_i, Y_i)\colon I \geq 1\ ]$ satisfying (10.1) and (10.2) is said to obey the *information stability property* (ISP) if

$$\lim_{n\to\infty} \Pr[|n^{-1} i(X^n;Y^n) - \bar{I}| > \epsilon] = 0, \quad \forall \epsilon > 0$$

We give some examples of pair processes obeying the ISP.

Example 11. Let the stochastic process $[X_i : I \geq 1\ ]$ and the stochastic process $[Y_i : I \geq 1\ ]$ be statistically independent. For every positive integer n, we have $I(X^n; Y^n) = 0$. It follows that the pair process $[(X_i, Y_i)\colon I \geq 1\ ]$ obeys the ISP and that the information rate is zero.

Example 12. Let us be given a semicontinuous stationary ergodic channel through which we must transmit information. "Semicontinuous channel" refers to the fact that the channel generates an infinite sequence of random outputs $[Y_i\}$ from a continuous alphabet in response to an infinite sequence of random inputs $\{X_i\}$ from a discrete alphabet. "Stationary ergodic channel" refers to the fact that the channel pair process $\{(X_i, Y_i)\}$ will be stationary and ergodic whenever the sequence of channel inputs $\{X_i\}$ is stationary and ergodic. Suppose that $\{X_i\}$ is a stationary ergodic discrete-alphabet process, which we apply as input to our given channel. Let $[Y_i\ ]$ be the resulting channel output process. In proving a channel coding theorem (see the section entitled "Application to Channel Coding Theory"), it could be useful to know whether the stationary and ergodic pair process $\{(X_i, Y_i)\colon I \geq 1\}$ obeys the information stability property. We quote a result that allows us to conclude that the ISP holds in this type of situation. Appealing to Theorems 7.4.2 and 8.2.1 of (9), it is known that a stationary and ergodic pair process $[(X_i, Y_i)\colon I \geq 1\ ]$ will obey the ISP provided that X_1 is discrete with $H(X_1) < \infty$. The proof of this fact in (9) is too complicated to discuss here. Instead, let us deal with the special case in which we assume that Y_1 is also discrete with $H(Y_1) < \infty$. We easily deduce that $[(X_i,Y_i)\colon I \geq 1\ ]$ obeys the ISP. For we can write

$$n^{-1} i(X^n;Y^n) = n^{-1} h(X^n) + n^{-1} h(Y^n) - n^{-1} h(X^n, Y^n) \quad (6)$$

for each positive integer n. Due to the fact that each of the processes $\{X_i\}$, $\{Y_i\}$, $\{(X_i,Y_i)\}$ obeys the AEP, we conclude that each of the three terms on the right hand side of Eq. (6) converges to a constant as $n \to \infty$. The left side of Eq. (6) therefore must also converge to a constant as $n \to \infty$.

Example 13. An IID pair process $[(X_i,Y_i): I \geq 1]$ obeys the ISP provided that $I(X_1; Y_1) < \infty$. In this case, the information rate is given by $\tilde{I} = I(X_1; Y_1)$. This result is evident from an application of the law of large numbers to the equation

$$n^{-1} i(X^n; Y^n) = n^{-1} \sum_{i=1}^{n} i(X_i; Y_i)$$

This result is important because this is the type of channel pair process that results when an IID process is applied as input to a memoryless channel. (The memoryless channel model is the simplest type of channel model—it is discussed in Example 21.)

Example 14. Let $[(X_i,Y_i): I \geq 1]$ be a Gaussian process satisfying (10.1) and (10.2). Suppose that the information rate of this pair process satisfies $\bar{I} > 0$. It is known that the pair process obeys the ISP (9, Theorem 9.6.1).

Example 15. We assume that $[(X_i,Y_i): I \geq 1]$ is a stationary Gaussian process in which, for each I, the random variables X_i and Y_i are real-valued and have expected value equal to zero. For each integer $k \geq 0$, define the trix

$$R(k) = \begin{bmatrix} R_{1,1}(k) & R_{1,2}(k) \\ R_{2,1}(k) & R_{2,2}(k) \end{bmatrix} = \begin{bmatrix} E[X_1 X_{k+1}] & E[X_1 Y_{k+1}] \\ E[Y_1 X_{k+1}] & E[Y_1 Y_{k+1}] \end{bmatrix}$$

Assume that

$$\sum_{k=0}^{\infty} |R_{i,j}(k)| < \infty, \quad i,j = 1,2$$

Following (25, p. 85), we define the spectral densities

$$\begin{bmatrix} S_{1,1}(\omega) & S_{1,2}(\omega) \\ S_{2,1}(\omega) & S_{2,2}(\omega) \end{bmatrix} = \sum_{k=-\infty}^{\infty} R(k) \exp(-j\omega k), \quad -\pi \leq \omega \leq \pi \tag{7}$$

where in Eq. (7), for $k < 0$, we take $R(k) = R(-k)^{\mathrm{T}}$. Suppose that

$$\int_{-\pi}^{\pi} \log\left(1 - \frac{|S_{1,2}(\omega)|^2}{S_{1,1}(\omega) S_{2,2}(\omega)}\right) d\omega < \infty$$

where the ratio $|S_{1,2}(\omega)|^2 / S_{1,1}(\omega) S_{2,2}(\omega)$ is taken to be zero whenever $S_{1,2}(\omega) = 0$. It is known (9, Theorem 10.2.1) that the pair process $[(X_i,Y_i): I \geq 1]$ satisfies (10.1) and (10.2), and that the information rate $\tilde{I}$ is expressible as

$$\bar{I} = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log\left(1 - \frac{|S_{1,2}(\omega)|^2}{S_{1,1}(\omega) S_{2,2}(\omega)}\right) d\omega \tag{8}$$

Furthermore, we can deduce that $[(X_i, Y_i)\colon I \geq 1\,]$ obeys the ISP. For, if $\tilde{I} > 0$, we can appeal to Example 14. On the other hand, if $\tilde{I} = 0$, Eq. (8) tells us that the processes $\{X_i\}$ and $\{Y_i\}$ are statistically independent, upon which we can appeal to Example 11.

Example 16. Let $\{(X_i,Y_i)\}\colon I \geq 1\,]$ be a stationary ergodic process such that, for each positive integer n,

$$\Pr[Y_1 \in A_1, Y_2 \in A_2, \ldots, Y_n \in A_n | X_1, X_2, \ldots, X_n] = \prod_{i=1}^{n} \Pr[Y_i \in A_i | X_i] \tag{9}$$

holds almost surely for every choice of measurable events A_1, A_2, ..., A_n. [The reader not familiar with the types of conditional probability functions on the two sides of Eq. (9) can consult (26, Chap. 6).] In the context of communication engineering, the stochastic process $[\,Y_i : i \geq 1\,]$ may be interpreted to be the process that is obtained by passing the process $[\,X_i : i \geq 1\,]$ through a memoryless channel (see Example 21). Suppose that $I(X_1; Y_1) < \infty$. Then, properties (10.1) and (10.2) hold and the information stability property holds for the pair process $[(X_i,Y_i)\colon i \geq 1\,]$ (14, 27).

Example 17. Let $[(X_i,Y_i)\colon i \geq 1\,]$ be a stationary ergodic process in which each random variable X_i is real-valued and each random variable Y_i is real-valued. We suppose that (10.1) and (10.2) hold and we let $\tilde{I}$ denote the information rate of the process $[(X_i,Y_i)\colon i \geq 1\,]$. A *quantizer* is a mapping Q from the real line into a finite subset of the real line, such that for each value q of Q, the set $[\,r\colon Q(r) = q\,]$ is a subinterval of the real line. Suppose that Q is any quantizer. By Example 12, the pair process $[(Q(X_i),Q(Y_i))\colon i \geq 1\,]$ obeys the ISP; we will denote the information rate of this process by $\tilde{I}_Q$. It is known that $[(X_i,Y_i)\colon I \geq 1\,]$ satisfies the information stability property if

$$\bar{I} = \sup_Q \bar{I}_Q \tag{10}$$

where the supremum is taken over all quantizers Q. This result was first proved in (9, Theorem 8.2.1). Another proof of the result may be found in (28), where the result is used to prove a source coding theorem. Theorem 7.4.2 of (9) gives numerous conditions under which Eq. (10) will hold.

Example 18. This example points out a way in which the AEP and the ISP are related. Let $[\,X_i : I \geq 1\,]$ be any process satisfying (2.1) and (2.2). Then the pair process $\{(X_i,X_i)\colon i \geq 1\}$ satisfies (10.1) and (10.2). The entropy rate of the process $[\,X_i : i \geq 1\,]$ coincides with the information rate of the process $(X_i,X_i)\colon i \geq 1\,]$. The AEP holds for the process $[\,X_i : i \geq 1\,]$ if and only if the ISP holds for the pair process $[(X_i,X_i)\colon i \geq 1\,]$. To see that these statements are true, the reader is referred to Example 10.

Further Reading. The exhaustive text by Pinsker (9) contains many more results on information stability than were discussed in this article. The text by Gray (16) makes the information stability results for stationary pair processes in (9) more accessible and also extends these results to the bigger class of asymptotically mean stationary pair processes. The text (9) still remains unparalleled for its coverage of the information stability of Gaussian pair processes. The paper by Barron (14) contains some interesting results on information stability, presented in a self-contained manner.

Application to Source Coding Theory

As explained at the start of this article, source coding theory is one of two principal subareas of information theory (channel coding theory being the other). In this section, explanations are given of the operational significance of the AEP and the ISP to source coding theory.

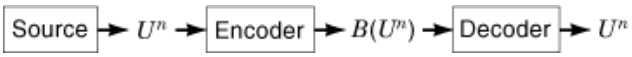

Fig. 2. Lossless source coding system.

An information source generates data samples sequentially in time. A fixed *abstract information source* is considered, in which the sequence of data samples generated by the source over time is modeled abstractly as a stochastic process $[\, U_i : i \geq 1\,]$. Two coding problems regarding the given abstract information source shall be considered. In the problem of *lossless source coding*, one wishes to assign a binary codeword to each block of source data, so that the source block can be perfectly reconstructed from its codeword. In the problem of *lossy source coding*, one wishes to assign a binary codeword to each block of source data, so that the source block can be approximately reconstructed from its codeword.

Lossless Source Coding. The problem of lossless source coding for the given abstract information source is considered first. In lossless source coding, it is assumed that there is a finite set A (called the *source alphabet*) such that each random data sample U_i generated by the given abstract information source takes its values in A. The diagram in Fig. 2 depicts a *lossless source coding system* for the block $U^n = (U_1, U_2, \ldots, U_n)$, consisting of the first n data samples generated by the given abstract information source.

As depicted in Fig. 2, the lossless source coding system consists of *encoder* and *decoder*. The encoder accepts as input the random source block U^n and generates as output a random binary codeword $B(U^n)$. The decoder perfectly reconstructs the source block U^n from the codeword $B(U^n)$. A nonnegative real number R is called an *admissible lossless compression rate* for the given information source if, for each $\delta > 0$, a Fig. 2 te system can be designed for fficiently large n so that

$$\lim_{n\to\infty} \Pr[n^{-1}|B(U^n)| \leq R + \delta] = 1 \qquad (11)$$

where $|\, B(U^n)|$ denotes the length of the codeword $B(U^n)$.

Let us now refer back to the start of this article, where we talked about the rate $R(S)$ at which the information source S in a data communication system generates information over time (assuming that the information must be losslessly transmitted). We were not precise in the beginning concerning how $R(S)$ should be defined. We now define $R(S)$ to be the minimum of all admissible lossless compression rates for the given information source S.

As discussed earlier, if the communication engineer must incorporate a given information source S into the design of a data communication system, it would be advantageous for the engineer to be able to determine the rate $R(S)$. Let us assume that the process $\{U_i : i \geq 1\}$ modeling our source S obeys the AEP. In this case, it can be shown that

$$R(S) = \bar{H} \qquad (12)$$

where $\bar{H}$ is the entropy rate of the process $\{U_i\}$. We give here a simple argument that $\bar{H}$ is an admissible lossless compression rate for the given source, using the AEP. [This will prove that $R(S) \leq \bar{H}$. Using the AEP, a proof can also be given that $R(S) \geq \bar{H}$, thereby completing the demonstration of Eq. (12), but we omit this proof.] Let A^n be the set of all n-tuples from the source alphabet A. For each $n \geq 1$, we may pick a subset E_n of A^n so that properties (2.3) to (2.5) hold. [The ε in (2.4) and (2.5) is a fixed, but arbitrary, positive real number.] Let F_n be the set of all n-tuples in A^n, which are not contained in E_n. Because of property (2.5), for sufficiently large n, we may assign each n-tuple in E_n a unique binary codeword of length $1 + \lceil n(\bar{H} + \varepsilon)\rceil$, so that each codeword begins with 0. Letting $|A|$ denote the number of symbols in A, we may assign each n-tuple in F_n a unique binary codeword of length $1 + \lceil \mathrm{CR}n \log |A|\rceil$, so that each codeword begins with 1. In this way, we have a lossless

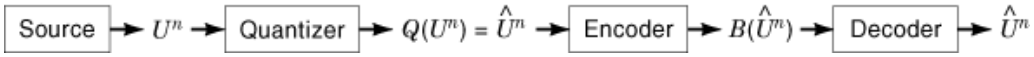

Fig. 3. Lossy source coding system.

codeword assignment for all of A^n, which gives us an encoder and decoder for a Fig. 2 lossless source coding system. Because of property (2.3), Eq. (11) holds with $R = \bar{H}$ and $\delta = 2\varepsilon$. Since ε (and therefore δ) is arbitrary, we can conclude that $\bar{H}$ is an admissible lossless compression rate for our given information source.

In view of Eq. (12), we see that for an abstract information source modeled by a process $\{U_i : i \geq 1\}$ satisfying the AEP, the entropy rate $\bar{H}$ has the following operational significance:

- No $R < \bar{H}$ is an admissible lossless compression rate for the given source.
- Every $R \geq \bar{H}$ is an admissible lossless compression rate for the given source.

If the process $\{U_i : i \geq 1\}$ does not obey the AEP, then Eq. (12) can fail, even when properties (2.1) and (2.2) are true and thereby ensure the existence of the entropy rate $\bar{H}$. Here is an example illustrating this phenomenon.

Example 19. Let the process $\{U_i : i \geq 1\}$ modeling the source S have alphabet $A = \{0, 1\}$ and satisfy, for each positive integer n, the following properties:

$$\Pr[U^n = u^n] = \begin{cases} (1/2)(1 + 2^{-n}), & u^n \text{ all zeros} \\ (1/2)2^{-n}, & \text{otherwise} \end{cases}$$

Properties (2.1) and (2.2) are satisfied and the entropy rate is $\bar{H} = \frac{1}{2}$. Reference 29 shows that $R(S) = 1$.

Extensions. The determination of the minimum admissible lossless compression rate $R(S)$, when the AEP does not hold for the process [$U_i : I \geq 1$] modeling the abstract source S, is a problem that is beyond the scope of this article. This problem was solved by Parthasarathy (29) for the case in which [$U_i : I \geq 1$] is a stationary process. For the case in which [$U_i : I \geq 1$] is nonstationary, the problem has been solved by Han and Verdú (30, Theorem 3).

Lossy Source Coding. The problem of lossy coding of a given abstract information source is now considered. The stochastic process $[U_i: I \geq 1]$ is again used to model the sequence of data samples generated by the given information source, except that the source alphabet A is now allowed to be infinite. Figure 3 depicts a *lossy source coding system* for the source block $U^n = (U_1, U_2, \ldots, U_n)$.

Comparing Fig. 3 to Fig. 2, we see that what distinguishes the lossy system from the lossless system is the presence of the quantizer in the lossy system. The quantizer in Fig. 3 is a mapping Q from the set of n-tuples A^n into a finite subset $Q(A^n)$ of A_n. The quantizer Q assigns to the random source block U^n a block

$$Q(U^n) = \hat{U}^n = (\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n) \in Q(A^n) \subset A^n$$

The encoder in Fig. 3 assigns to the quantized source block $\hat{U}^n$ a binary codeword $B^{(\hat{U}^n)}$ from which the decoder can perfectly reconstruct $\hat{U}^n$. Thus the system in Fig. 3 reconstructs not the original source block U_n, but $\hat{U}^n$, a quantized version of U^n.

In order to evaluate how well lossy source coding can be done, one must specify for each positive integer n a nonnegative real-valued function ρ_n on the product space $A^n \times A^n$ (called a *distortion measure*). The quantity $\rho_n (U_n, \hat{U}^n)$ measures how closely the reconstructed block $\hat{U}^n$ in Fig. 3 resembles the source block U^n. Assuming that ρ_n is a jointly continuous function of its two arguments, which vanishes whenever the arguments are equal, one goal in the design of the lossy source coding system in Fig. 3 would be:

- *Goal 1.* Ensure that ρ_n (U_n,$\hat{U}^n$) is sufficiently close to zero.

However, another goal would be:

- *Goal 2.* Ensure that the length | $B^{(\hat{U}^n)}$| of the codeword $B^{(\hat{U}^n)}$ is sufficiently small.

These are conflicting goals. The more closely one wishes $\hat{U}^n$ to resemble U_n [corresponding to a sufficiently small value of ρ_n (U_n,$\hat{U}^n$)], the more finely one must quantize U^n, meaning an increase in the size of the set $Q(A^n)$, and therefore an increase in the length of the codewords used to encode the blocks in $Q(A^n)$. There must be a trade-off in the accomplishment of Goals 1 and 2. To reflect this trade-off, two figures of merit are used in lossy source coding. Accordingly, we define a pair (R, D) of nonnegative real numbers to be an *admissible rate-distortion pair* for lossy coding of the given abstract information source, if, for any $\varepsilon > 0$, the Fig. 3 system can be designed for sufficiently large n so that

$$\lim_{n\to\infty} \Pr[\rho_n(U^n, \hat{U}^n) \leq D + \epsilon] = 1 \tag{13}$$

$$\lim_{n\to\infty} \Pr[n^{-1}|B(\hat{U}^n)| \leq R + \epsilon] = 1 \tag{14}$$

We now describe how the information stability property can allow one to determine admissible rate-distortion pairs for lossy coding of the given source. For simplicity, we assume that the process [$U_i : I \geq 1$] modeling the source outputs is stationary and ergodic. Suppose we can find another process {$V_i : I \geq 1$] such that

- The pair process [(U_i,V_i) : $I \geq 1$] is stationary and ergodic.
- There is a finite set $\hat{A} \subset A$ such that each V_i takes its values in $\hat{A}$.

Appealing to Example 12, the pair process [(U_i,V_i): $I \geq 1$] satisfies the information stability property. Let $\tilde{I}$ be the information rate of this process. Assume that the distortion measures [ρ_n] satisfy

$$\rho_n((u_1, u_2, \ldots, u_n), (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)) = n^{-1} \sum_{i=1}^{n} \rho_1(u_i, \hat{u}_i)$$

for any pair of n-tuples (u_1, …, U_n), ($\hat{u}_1$, …, $\hat{u}_n$) from A_n. (In this case, the sequence of distortion measures [ρ_n] is called a *single letter fidelity criterion.*) Let $D = E[\rho_1(U_1, V_1)]$. Via a standard argument (omitted here) called a random coding argument [see proof of Theorem 7.2.2 of (31)], information stability can be exploited to show that the pair ($\tilde{I}$, D) is an admissible rate-distortion pair for our given abstract information source. [It should be pointed out that the random coding argument not only exploits the information stability property but also exploits the property that

$$\lim_{n\to\infty} \Pr[\rho_n(U^n, V^n) \leq D + \epsilon] = 1, \quad \forall \epsilon > 0$$

which is a consequence of the *ergodic theorem* [(32), Chap. 3]].

Example 20. Consider an abstract information source whose outputs are modeled as an IID sequence of real-valued random variables [U_i: $I \geq 1$]. This is called the *memoryless* source model. The squared-error single letter fidelity criterion [ρ_n] is employed, in which

$$\rho_n((u_1, u_2, \ldots, u_n), (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)) = n^{-1} \sum_{i=1}^{n} (u_i - \hat{u}_i)^2$$

It is assumed that $E[U^2_1] < \infty$. For each $D > 0$, let $R(D)$ be the class of all pairs of random variables (U, V) in which

- U has the same distribution as U_1.
- V is real-valued.
- $E[(U - V)^2] \leq D$.

The *rate distortion function* of the given memoryless source is defined by

$$r(D) \triangleq \min\{I(U;V): (U,V) \in \mathsf{P}(D)\}, \quad D \geq 0$$

Shannon (33) showed that any (R, D) satisfying $R \geq r(D)$ is an admissible rate-distortion pair for lossy coding of our memoryless source model. A proof of this can go in the following way. Given the pair (R, D) satisfying $R \geq r(D)$, one argues that there is a process [$V_i : I \geq 1$] for which the pair process [U_i,V_i): $I \geq 1$] is independent and identically distributed, with information rate no bigger than R and with $E[(U_1 - V_1)^2] \leq D$. A random coding argument exploiting the fact that [(U_i, V_i): $I \geq 1$] obeys the ISP (see Example 13) can then be given to conclude that (R, D) is indeed an admissible rate-distortion pair. Shannon (33) also proved the converse statement, namely, that *any* admissible rate-distortion pair (R, D) for the given memoryless source model must satisfy $R \geq r(D)$. Therefore the set of admissible rate-distortion pairs for the memoryless source model is the set

$$\{(R, D): R \geq r(D)\} \tag{15}$$

Extensions. The argument in Example 20 exploiting the ISP can be extended [(31), Theorem 7.2.2] to show that for any abstract source whose outputs are modeled by a stationary ergodic process, the set in Eq. (15) coincides with the set of all admissible rate-distortion pairs, provided that a single letter fidelity criterion is used, and provided that the rate-distortion function $r(D)$ satisfies $r(D) < \infty$ for each $D > 0$. [The rate-distortion function for this type of source must be defined a little differently than for the memoryless source in Example 20; see (31) for the details.] Source coding theory for an abstract source whose outputs are modeled by a stationary nonergodic process has also been developed. For this type of source model, it is customary to replace the condition in Eq. (13) in the definition of an admissible rate-distortion pair with the condition

$$\limsup_{n \to \infty} E[\rho_n(U^n, \hat{U}^n)] \leq D + \epsilon$$

A source coding theorem for the stationary nonergodic source model can be proved by exploiting the information stability property, provided that the definition of the ISP is weakened to include pair processes [(U_i, V_i): $I \geq 1$] for which the sequence [$n^{-1} I(U^n; V^n)$: $n \geq 1$] converges to a nonconstant random variable. However, for this source model, it is difficult to characterize the set of admissible rate-distortion pairs by use of the ISP. Instead, Gray and Davisson (34) used the ergodic decomposition theorem (35) to characterize this

set. Subsequently, source coding theorems were obtained for abstract sources whose outputs are modeled by asymptotically mean stationary processes; an account of this work can be found in Gray (16).

Further Reading. The theory of lossy source coding is called *rate-distortion theory*. Reference (31) provides excellent coverage of rate-distortion theory up to 1970. For an account of developments in rate-distortion theory since 1970, the reader can consult (36,37).

Application to Channel Coding Theory

In this section, explanations are given of the operational significance of the ISP to channel coding theory. To accomplish this goal, the notion of an abstract channel needs to be defined. The description of a completely general abstract channel model would be unnecessarily complicated for the purposes of this article. Instead, an abstract channel model is chosen that will be simple to understand, while of sufficient generality to give the reader an appreciation for the concepts that shall be discussed.

We shall deal with a semicontinuous channel model (see Example 12) in which the channel input phabet is finite and the channel output alphabet is the real line. We proceed to give a precise formulation of this channel model. We fix a finite set A, from which inputs to our abstract channel are to be drawn. For each positive integer n, let A^n denote the set of all n-tuples $X^n = (X_1, X_2, \ldots, X_n)$ in which each $X_i \in A$, and let R^n denote the set of all n-tuples $Y^n = (Y_1, Y_2, \ldots, Y_n)$ in which each $Y_i \in R$, the set of real numbers. For each $n \geq 1$, a function F_n is given that maps each 2 n-tuple $(X^n, Y^n) \in A^n \times R^n$ into a nonnegative real number $F_n(Y^n|X^n)$ so that the following rules are satisfied:

- For each $X^n \in A^n$, the mapping $Y^n \to F_n(Y^n|X^n)$ is a jointly measurable function of n variables.
- For each $X^n \in A^n$,

$$\int \int \cdots \int_{R^n} f_n(y^n|x^n) dy^n = 1$$

For each $n \geq 2$, each $(x_1, x_2, \ldots, x_n) \in A^n$, and each $(y_1, \ldots, y_n - 1) \in R^n - 1$,

$$\int_{-\infty}^{\infty} f_n(y_1, y_2, \ldots, y_n|x_1, x_2, \ldots, x_n)\, dy_n = f_{n-1}(y_1, \ldots, y_{n-1}|x_1, \ldots, x_{n-1}) \quad (16)$$

We are now able to describe how our abstract channel operates. Fix a positive integer n. Let $X^n \in A^n$ be any n-tuple of channel inputs. In response to X^n, our abstract channel will generate a random n-tuple of outputs from R^n. For each measurable subset En of R^n, let Pr[En|x^n] denote the conditional probability that the channel output n-tuple will lie in En, given that the channel input is X^n. This conditional probability is computable via the formula

$$\Pr[E_n|x^n] = \int \int \cdots \int_{E_n} f_n(y^n|x^n)\, dy^n$$

We now need to define the notion of a channel code for our abstract channel model. A *channel code* for our given channel is a collection of pairs $[(x(i), E(i)): i = 1, 2, \ldots, 2^k]$ in which

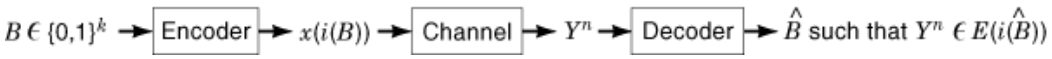

Fig. 4. Implementation of a (k, n) channel code.

(1) k is a positive integer.
(2) For some positive integer n,

- $x(1), x(2), \ldots, X(2^k)$ are n-tuples from A^n.
- $E(1), E(2), \ldots, E(2^k)$ are subsets of R^n, which form a partition of R^n.

The positive integer n given by (ii) is called the *number of channel uses* of the channel code, and the positive integer k given by (i) is called the *number of information bits* of the channel code. We shall use the notation $\boldsymbol{c}_n$ as a generic notation to denote a channel code with n channel uses. Also, a channel code shall be referred to as a (k,n) channel code if the number of channel uses is n and the number of information bits is k. In a channel code $\{(x(i), E(i))\}$, the sequences $\{x(i)\}$ are called the *channel codewords*, and the sets $\{E(i)\}$ are called the *decoding sets*.

A (k,n) channel code $\{(x(i), E(i)): i = 1, 2, \ldots, 2^k\}$ is used in the following way to transmit data over our given channel. Let $\{0, 1\}^k$ denote the set of all binary k-tuples. Suppose that the data that one wants to transmit over the channel consists of the k-tuples in $\{0, 1\}^k$. One can assign each k-tuple $B \in [0, 1]^k$ an integer index $I = I(B)$ satisfying $1 \leq I \leq 2^k$, which uniquely identifies that k-tuple. If the k-tuple B is to be transmitted over the channel, then the *channel encoder* encodes B into the channel codeword $X(I)$ in which $I = I(B)$, and then $x(i)$ is applied as input to the channel. At the receiving end of the channel, the *channel decoder* examines the resulting random channel output n-tuple Y^n that was received in response to the channel codeword $x(i)$. The decoder determines the unique random integer J such that $Y^n \in E(J)$ and decodes Y^n into the random k-tuple $\hat{B} \in \{0,1\}^k$ whose index is J. The transmission process is depicted in Fig. 4.

There are two figures of merit that tell us the performance of the (k, n) channel code $\boldsymbol{c}_n$ depicted in Fig. 4, namely, the *transmission rate* $R(\boldsymbol{c}_n)$ and the *error probability* $e(\boldsymbol{c}_n)$. The transmission rate measures how many information bits are transmitted per channel use and is defined by

$$R(\mathbf{c}_n) \stackrel{\Delta}{=} \frac{k}{n}$$

The error probability gives the worst case probability that $\hat{B}$ in Fig. 4 will not be equal to B, over all possible $B \in \{0, 1\}^k$. It is defined by

$$e(\mathbf{c}_n) \stackrel{\Delta}{=} \max_{B \in \{0,1\}^k} \{1 - \Pr[E(i(B))|x(i(B))]\}$$

It is desirable to find channel codes that simultaneously achieve a large transmission rate and a small error probability. Unfortunately, these are conflicting goals. It is customary to see how large a transmission rate can be achieved for sequences of channel codes whose error probabilities $\to 0$. Accordingly, an *admissible transmission rate* for the given channel model is defined to be a nonnegative number R for which there exists

a sequence of channel codes [$\boldsymbol{c}_{\mathrm{n}} : n = 1, 2, \ldots$] satisfying both of the following:

$$\liminf_{n\to\infty} R(\mathbf{c}_n) \geq R$$

$$\lim_{n\to\infty} e(\mathbf{c}_n) = 0$$

We now describe how the notion of information stability can tell us about admissible transmission rates for our channel model. Let [$X_i : i \geq 1$] be a sequence of random variables taking their values in the set A, which we apply as inputs to our abstract channel. Because of the consistency crite- rion, Eq. (16), the abstract channel generates, in response to [X_i: $i \geq 1$], a sequence of real-valued random outputs [Y_i: $i \geq 1$] for which the distribution of the pair process [(X_i,Y_i): $i \geq 1$] is uniquely specified by

$$\begin{aligned}\Pr[(X_1, X_2, \ldots, X_n) = x^n,\ (Y_1, Y_2, \ldots, Y_n) \in E_n] \\ = \Pr[(X_1, X_2, \ldots, X_n) = x^n]\Pr[E_n|x^n]\end{aligned}$$

for every positive integer n, every n-tuple $X^n \in A^n$, and every measurable set En $\subset R^n$. Suppose the pair process [(X_i,Y_i): $i \geq 1$] obeys the ISP with information rate $\tilde{I}$. Then a standard argument [see (38), proof of Lemma 3.5.2] can be given to show that $\tilde{I}$ is an admissible transmission rate for the given channel model.

Using the notation introduced earlier, the *capacity*$R(C)$ of an abstract channel C is defined to be the maximum of all admissible transmission rates. For a given channel C, it is useful to determine the capacity $R(C)$. (For example, as discussed at the start of this article, if a data communication system is to be designed using a given channel, then the channel capacity must be at least as large as the rate at which the information source in the system generates information.) Suppose that an abstract channel C possesses at least one input process [$X_i : i \geq 1$] for which the corresponding channel pair process [(X_i,Y_i): $i \geq 1$] obeys the ISP. Define R_{ISP} (C) to be the supremum of all information rates of such processes [(X_i,Y_i): $I \geq 1$]. By our discussion in the preceding paragraph, we have

$$R(C) \geq R_{\mathrm{ISP}}(C)$$

For some channels C, one has $\mathscr{R}C$ ($\mathscr{R}$) = $\mathscr{R}$ $_{\mathrm{ISP}}(C)$. For such a channel, an examination of channel pair processes satisfying the ISP will allow one to determine the capacity.

Examples of channels for which this is true are the memoryless channel (see Example 21 below), the finite-memory channel (39), and the finite-state indecomposable channel (40). On the other hand, if $\mathscr{R}$ (C) > $\mathscr{R}_{\mathrm{ISP}}(C)$ for a channel C, the concept of information stability cannot be helpful in determining the channel capacity—some other concept must be used. Examples of channels for which $\mathscr{R}(C) > \mathscr{R}_{\mathrm{ISP}}(C)$ holds, and for which the capacity $\mathscr{R}(C)$ has been determined, are the $\tilde{I}$ continuous channels (41), the weakly continuous channels (42), and the historyless channels (43). The authors of these papers could not use information stability to determine capacity. They used instead the concept of "information quantiles," a concept beyond the scope of this article. The reader is referred to Refs. 41–43 to see what the information quantile concept is and how it is used.

Example 21. Suppose that the conditional density functions [$F_n : n = 1, 2, \ldots$] describing our channel satisfy

$$f_n(y^n|x^n) = \prod_{i=1}^{n} f_1(y_i|x_i)$$

for every positive integer n, every n-tuple $x^n = (x_1, \ldots, x_n)$ from A^n, and every n-tuple $Y^n = (Y_1, \ldots, Y_n)$ from R^n. The channel is then said to be *memoryless*. Let R^* be the nonnegative real number defined by

$$R^* \triangleq \sup_{(X,Y)} I(X;Y) \tag{17}$$

where the supremum is over all pairs (X, Y) in which X is a random variable taking values in A, and Y is a real-valued random variable whose conditional distribution given X is governed by the function f_1. (In other words, we may think of Y as the channel output in response to the single channel input X.) We can argue that R^* is an admissible transmission rate for the memoryless channel as follows. Pick a sequence of IID channel inputs $[\, X_i : I \geq 1\,]$ such that if $[\, Y_i : I \geq 1\,]$ is the corresponding sequence of random channel outputs, then $I(X_1; Y_1) = R^*$. The pairs $[(X_i,Y_i): i \geq 1\,]$ are IID, and the process $[(X_i,Y_i): i \geq 1\,]$ obeys the ISP with information rate $\tilde{I} = R^*$ (see Example 13). Therefore R^* is an admissible transmission rate. By a separate argument, it is well known that the converse is also true; namely, every admissible transmission rate for the memoryless channel is less than or equal to R^* (1). Thus the number R^* given by Eq. (17) is the capacity of the memoryless channel.

Final Remarks

It is appropriate to conclude this article with some remarks concerning the manner in which the separate theories of source coding and channel coding tie together in the design of data communication systems. In the section entitled "Lossless Source Coding," it was explained how the AEP can sometimes be helpful in determining the minimum rate $R(S)$ at which an information source S can be losslessly compressed. In the section entitled "Application to Channel Coding Theory," it was indicated how the ISP can sometimes be used in determining the capacity $R(C)$ of a channel C, with the capacity giving the maximum rate at which data can reliably be transmitted over the channel. If the inequality $R(S) \leq R(C)$ holds, it is clear from this article that reliable transmission of data generated by the given source S is possible over the given channel C. Indeed, the reader can see that reliable txsransmission will take place for the data communication system in Fig. 1 by taking the encoder to be a two-stage encoder, in which a good source encoder achieving a compression rate close to $R(S)$ is followed by a good channel encoder achieving a transmission rate close to $R(C)$. On the other hand, if $R(S) > R(C)$, there is no encoder that can be found in Fig. 1 via which data from the source S can reliably be transmitted over the channel C [see any basic text on information theory, such as (44), for a proof of this result]. One concludes from these statements that in designing a reliable encoder for the data communication system in Fig. 1, one need onlyonsider the two-stage encoders consisting of a good source encoder followed by a good channel encoder. This principle, which allows one to break down the problem of encoder design in communication systems into the two separate simpler problems of source encoder design and channel encoder design, has come to be called "Shannon's separation principle," after its originator, Claude Shannon.

Shannon's separation principle also extends to lossy transmission of source data over a channel in a data communication system. In Fig. 1, suppose that the data communication system is to be designed so that the data delivered to the user through the channel C must be within a certain distance D of the original data generated by the source S. The system can be designed if and only if there is a positive real number R such that (1) (R, D) is an admissible rate-distortion pair for lossy coding of the source S in the sense of the "Lossy Source Coding" section, and (2) $R \leq R(C)$. If R is a positive real number satisfying (1) and (2), Shannon's separation principle tells us that the encoder in Fig. 1 can be designed as a two-stage encoder consisting of source encoder followed by channel encoder in which:

- The source encoder is designed to achieve the compression rate R and to generate blocks of encoded data that are within distance D of the original source blocks.
- The channel encoder is designed to achieve a transmission rate close to $R(C)$.

It should be pointed out that Shannon's separation principle holds only if one is willing to consider arbitrarily complex encoders in communication systems. [In defining the quantities $R(S)$ and $R(C)$ in this article, recall that no constraints were placed on how complex the source encoder and channel encoder could be.] It would be more realistic to impose a complexity constraint specifying how complex an encoder one is willing to use in the design of a communication system. With a complexity constraint, there could be an advantage in designing a "combined source–channel encoder" which combines data compression and channel error correction capability in its operation. Such an encoder for the communication system could have the same complexity as two-stage encoders designed according to the separation principle but could afford one a better data transmission capability than the two-stage encoders. There has been much work in recent years on "combined source–channel coding," but a general theory of combined source–channel coding has not yet been put forth.

BIBLIOGRAPHY

1. C. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, **27**: 379–423, 623–656, 1948.
2. B. McMillan, The basic theorems of information theory, *Ann. Math. Stat.*, **24**: 196–219, 1953.
3. L. Breiman, The individual ergodic theorem of information theory, *Ann. Math. Stat.*, **28**: 809–811, 1957.
4. R. Gray J. Kieffer, Asymptotically mean stationary measures, *Ann. Probability*, **8**: 962–973, 1980.
5. S. Verdú T. Han, The role of the asymptotic equipartition property in noiseless source coding, *IEEE Trans. Inf. Theory*, **43**: 847–857, 1997.
6. J. Kieffer, A generalized Shannon-McMillan theorem for the action of an amenable group on a probability space, *Ann. Probability*, **3**: 1031–1037, 1975.
7. D. Ornstein B. Weiss, The Shannon-McMillan-Breiman theorem for a class of amenable groups, *Isr. J., Math.*, **44**: 53–60, 1983.
8. A. Perez, Notions généralisées d′incertitude, d′entropie et d′information du point de vue de la théorie de martingales, *Trans. 1st Prague Conf. Inf. Theory, Stat. Decision Funct., Random Process.*, pp. 183–208, 1957.
9. M. Pinsker, *Information and Information Stability of Random Variables and Processes*, San Francisco: Holden-Day, 1964.
10. A. Ionescu Tulcea Contributions to information theory for abstract alphabets, *Ark. Math.*, **4**: 235–247, 1960.
11. A. Perez, Extensions of Shannon-McMillan's limit theorem to more general stochastic processes, *Trans. 3rd Prague Conf. Inf. Theory*, pp. 545–574, 1964.
12. S. Moy, Generalizations of Shannon-McMillan theorem, *Pac. J. Math.*, **11**: 705–714, 1961.
13. S. Orey, On the Shannon-Perez-Moy theorem, *Contemp. Math.*, **41**: 319–327, 1985.
14. A. Barron, The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem, *Ann. Probability*, **13**: 1292–1303, 1985.
15. P. Algoet T. Cover, A sandwich proof of the Shannon- McMillan-Breiman theorem, *Ann. Probability*, **16**: 899–909, 1988.
16. R. Gray, *Entropy and Information Theory*, New York: Springer-Verlag, 1990.
17. A. Tempelman, Specific characteristics and variational principle for homogeneous random fields, *Z. Wahrschein. Verw. Geb.*, **65**: 341–365, 1984.
18. D. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems*, Yale Math. Monogr. 5, New Haven, CT: Yale University Press, 1974.
19. D. Ornstein B. Weiss, Entropy and isomorphism theorems for actions of amenable groups, *J. Anal. Math.*, **48**: 1–141, 1987.
20. R. Mañé *Ergodic Theory and Differentiable Dynamics*, Berlin and New York: Springer-Verlag, 1987.

21. M. Ohya, Entropy operators and McMillan type convergence theorems in a noncommutative dynamical system, *Lect. Notes Math.*, **1299**, 384–390, 1988.
22. J. Fritz, Generalization of McMillan's theorem to random set functions, *Stud. Sci. Math. Hung.*, **5**: 369–394, 1970.
23. A. Perez, Generalization of Chernoff's result on the asymptotic discernability of two random processes, *Colloq. Math. Soc. J. Bolyai*, No. 9, pp. 619–632, 1974.
24. P. Algoet T. Cover, Asymptotic optimality and asymptotic equipartition properties of log-optimum investme, *Ann. Probability*, **16**: 876–898, 1988.
25. A. Balakrishnan, *Introduction to Random Processes in Engineering*, New York: Wiley, 1995.
26. R. Ash, *Real Analysis and Probability*, New York: Academic Press, 1972.
27. M. Pinsker, Sources of messages, *Probl. Peredachi Inf.*, **14**, 5–20, 1963.
28. R. Gray J. Kieffer, Mutual information rate, distortion, and quantization in metric spaces, *IEEE Trans. Inf. Theory*, **26**: 412–422, 1980.
29. K. Parthasarathy, Effective entropy rate and transmission of information through channels with additive random noise, *Sankhyā, Ser. A*, **25**: 75–84, 1963.
30. T. Han S. Verdú, Approximation theory of output statistics, *IEEE Trans. Inf. Theory*, **39**: 752–772, 1993.
31. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice–Hall, 1971.
32. W. Stout, *Almost Sure Convergence*, New York: Academic Press, 1974.
33. C. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE Natl. Conv. Rec.*, Part 4, pp. 142–163, 1959.
34. R. Gray L. Davisson, Source coding theorems without the ergodic assumption, *IEEE Trans. Inf. Theory*, **20**: 502–516, 1974.
35. R. Gray L. Davisson, The ergodic decomposition of stationary discrete random processes, *IEEE Trans. Inf. Theory*, **20**: 625–636, 1974.
36. J. Kieffer, A survey of the theory of source coding with a fidelity criterion, *IEEE Trans. Inf. Theory*, **39**: 1473–1490, 1993.
37. T. Berger J. Gibson, Lossy source coding, *IEEE Trans. Inf. Theory*, **44**: 2693–2723, 1998.
38. R. Ash, *Information Theory*, New York: Interscience, 1965.
39. A. Feinstein, On the coding theorem and its converse for finite-memory channels, *Inf. Control*, **2**: 25–44, 1959.
40. D. Blackwell, L. Breiman, A. Thomasian, Proof of Shannon's transmission theorem for finite-state indecomposable channels, *Ann. Math. Stat.*, **29**: 1209–1220, 1958.
41. R. Gray D. Ornstein, Block coding for discrete stationary d? continuous noisy channels, *IEEE Trans. Inf. Theory*, **25**: 292-306, 1979.
42. J. Kieffer, Block coding for weakly continuous channels, *IEEE Trans. Inf. Theory*, **27**, 721–727, 1981.
43. S. Verdú T. Han, A general formula for channel capacity, *IEEE Trans. Inf. Theory*, **40**: 1147–1157, 1994.
44. T. Cover J. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

READING LIST

R. Gray L. Davisson, *Ergodic and Information Theory*, Benchmark Pap. Elect. Eng. Comput. Sci. Vol. 19, Stroudsburg, PA: Dowden, Hutchinson, & Ross, 1977.

IEEE Transactions of Information Theory, Vol. 44, No. 6, October, 1998. (Special issue commemorating fifty years of information theory.)

JOHN C. KIEFFER
University of Minnesota