634–648, Aug. 1971.
[3] B. S. Atal and S. F. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Amer. Statist. Ass.*, vol. 50, pp. 637–655, Aug. 1971.
[4] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, pp. 1973–1986, Oct. 1970.
[5] G. M. Jenkins and G. S. Watts, *Spectral Analysis and Its Applications.* San Francisco, Calif.: Holden–Day, 1969, ch. 7.
[6] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968, ch. 9.
[7] T. Berger, *Rate Distortion Theory, A Mathematical Basis for Data Compression.* Englewood Cliffs, N.J.: Prentice–Hall, 1971.
[8] J. E. Gunn, "The application of modern estimation techniques to the speech data reduction problem," Ph.D. dissertation, Information and Control Sciences Center, Southern Methodist Univ., Dallas, Tex., 1972.
[9] R. A. McDonald and P. M. Schultheiss, "Information rates of Gaussian signals under criteria constraining the error spectrum," *IEEE Proc.*, vol. 52, pp. 415–416, Apr. 1964.
[10] T. J. Goblick, "Theoretical limitations on the transmission of data from analog sources," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 558–567, Oct. 1965.
[11] H. L. Van Trees, *Detection, Estimation and Modulation, Part II: Nonlinear Modulation Theory.* New York: Wiley, 1971, ch. 5.
[12] H. L. Royden, *Real Analysis.* New York: Macmillan, 1968.
[13] B. J. Bunin and J. K. Wolf, "Convergence to the rate-distortion function for Gaussian sources," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 65–71, Jan. 1971.
[14] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications.* Berkeley, Calif.: Univ. California Press, 1958.
[15] J. B. O'Neal, Jr., "Predictive quantizing systems for the transmission of television signals," *Bell Syst. Tech. J.*, vol. 45, pp. 689–720, May 1966.
[16] C. E. Shannon and W. Weaver, *A Mathematical Theory of Communication.* Urbana, Ill.: Univ. Illinois Press, 1949.
[17] L. Davisson, "Theory of data compression," Ph.D. dissertation, Dep. Eng., Univ. Calif., Los Angeles, Calif., Sept. 1964.
[18] J. B. O'Neal, Jr., "Entropy coding in speech and television DPCM signals," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 758–760, Nov. 1971.
[19] J. B. O'Neal and R. W. Stroh, "A speech encoder-multiplexer feasibility study," Air Force Office of Scientific Research, prepared under Contract F44620-70-C-0122.

Noiseless Coding of Correlated Information Sources

DAVID SLEPIAN AND JACK K. WOLF

***Abstract*—Correlated information sequences $\cdots,X_{-1},X_0,X_1,\cdots$ and $\cdots,Y_{-1},Y_0,Y_1,\cdots$ are generated by repeated independent drawings of a pair of discrete random variables X,Y from a given bivariate distribution $p_{XY}(x,y)$. We determine the minimum number of bits per character R_X and R_Y needed to encode these sequences so that they can be faithfully reproduced under a variety of assumptions regarding the encoders and decoders. The results, some of which are not at all obvious, are presented as an admissible rate region $\mathscr{R}$ in the R_X–R_Y plane. They generalize a similar and well-known result for a single information sequence, namely $R_X \geq H(X)$ for faithful reproduction.**

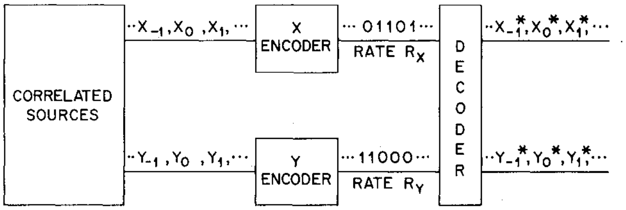
CORRELATED SOURCES
$\cdots X_{-1}, X_0, X_1,\cdots$
X ENCODER
$\cdots 01101 \cdots$
RATE R_X
$\cdots Y_{-1}, Y_0, Y_1,\cdots$
Y ENCODER
$\cdots 11000 \cdots$
RATE R_Y
DECODER
$\cdots X_{-1}^*, X_0^*, X_1^*\cdots$
$\cdots Y_{-1}^*, Y_0^*, Y_1^*\cdots$

Fig. 1. Correlated source coding configuration.

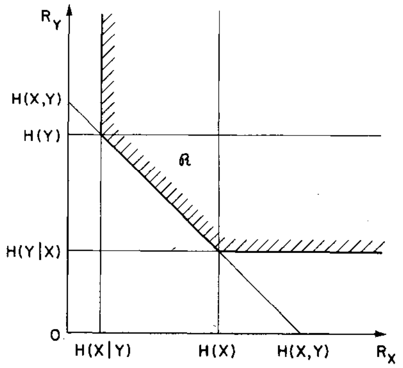

Fig. 2. Admissible rate region $\mathscr{R}$ corresponding to Fig. 1.

I. Introduction

Notation and Problem Statement

IN THIS PAPER, we generalize, to the case of two correlated sources, certain well-known results on the noiseless coding of a single discrete information source. Typical of the situations considered is that depicted in Fig. 1. Here the two correlated information sequences $\cdots,X_{-1},X_0,X_1,\cdots$ and $\cdots,Y_{-1},Y_0,Y_1,\cdots$ are obtained by repeated independent drawings from a discrete bivariate distribution $p(x,y)$. The encoder of each source is constrained to operate without knowledge of the other source, while the decoder has available both encoded binary message streams. We determine the minimum number of bits per source character required for the two encoded message streams in order to ensure accurate reconstruction by the decoder of the outputs of both information sources. The results are presented as an allowed two-dimensional rate region $\mathscr{R}$ for the two encoded message streams as shown in Fig. 2. Note that in $\mathscr{R}$ for this case we can have both $R_X < H(X)$ and $R_Y < H(Y)$ al-

Manuscript received August 25, 1972; revised December 28, 1972.

D. Slepian is with the University of Hawaii, Honolulu, Hawaii, and the Bell Laboratories, Murray Hill, N.J. 07974.
J. K. Wolf was with the University of Hawaii, Honolulu, Hawaii, on leave from the Polytechnic Institute of Brooklyn, Brooklyn, N.Y.

though each encoder sees only its own source. All these notions will be made precise in the following discussion.

First, however, let us review briefly some results for a single source that have long been known. Let X be a discrete random variable taking values in the set $\mathscr{A} = \{1,2,\cdots,A\}$. Denote the probability distribution of X by $p_X(x) = \Pr[X = x]$, $x \in \mathscr{A}$. Now let $\boldsymbol{X} = (X_1,X_2,\cdots,X_n)$ be a sequence of n independent realizations of X so that the probability distribution for the random n-vector $\boldsymbol{X}$ is given by

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr[\boldsymbol{X} = \boldsymbol{x}] = \prod_{i=1}^{n} p_X(x_i) \qquad (1)$$

$$\boldsymbol{x} = (x_1,x_2,\cdots,x_n) \in \mathscr{A}^n, \qquad x_i \in \mathscr{A}, \quad i = 1,2,\cdots,n.$$

Here, we introduce the symbol $\mathscr{A}^n$ to stand for the collection of the A^n different n-vectors $\boldsymbol{x}$, each of whose components is a member of $\mathscr{A}$. We regard $\boldsymbol{X}$ as a block of n successive characters from the output of an information source producing characters independently with letter distribution $p_X(x)$.

While little can be said about individual letters produced by this information source, for large n the composition of blocks of n letters tends to be fixed. In a typical long block, one can expect about $np_X(1)$ occurrences of letter 1, about $np_X(2)$ occurrences of letter 2, etc. The probability of such a typical long sequence is, therefore,

$$\begin{aligned} p_T &= p_X(1)^{np_X(1)} p_X(2)^{np_X(2)} \cdots p_X(A)^{np_X(A)} \\ &= \exp[np_X(1) \log p_X(1)] \cdots \exp[np_X(A) \log p_X(A)] \\ &= \exp[-nH(X)] \end{aligned}$$

where

$$H(X) \equiv -\sum_{1}^{A} p_X(i) \log p_X(i) \qquad (2)$$

is called the *entropy* of the source or of the random variable X. These simple observations lead to the useful, though imprecise, statement characterizing the long blocks of such a source: there are only $N_T = \exp[nH(X)]$ likely blocks of length n; each has probability $\exp[-nH(X)]$. This in turn suggests that we can accurately transmit the output of the information source using only $R = (1/n) \log N_T = H(X)$ natural units (nats) of information per character and that at least this rate is required of any transmission scheme that allows accurate recovery of the source output.

These intuitive coding notions can be made precise as follows. An encoder $\mathscr{C}(n,M)$ is any single-valued function $i = f(\boldsymbol{x})$ from the n-vectors $\boldsymbol{x}$ of $\mathscr{A}^n$ to the integers of the set $\mathscr{M} \equiv (1,2,\cdots,M)$. A decoder $\mathscr{D}(n,M)$ is a single-valued function $\boldsymbol{x} = \boldsymbol{g}(i)$ from the integers $i \in \mathscr{M}$ to the vectors $\boldsymbol{x} \in \mathscr{A}^n$. Associated with a source and a particular encoder and decoder pair are the rate R of the encoded messages defined by $R = (1/n) \log M$, and the two random variables $I \equiv f(\boldsymbol{X})$ and $\boldsymbol{X}^* \equiv \boldsymbol{g}(I)$ called, respectively, the encoded message number and the decoded block. We think of the encoder as producing the integer I after observing the n source characters $\boldsymbol{X}$. Then R units of information per source character suffice to communicate the value of I to the decoder. The decoder then produces the message block $\boldsymbol{X}^*$ as its estimate of $\boldsymbol{X}$.

A rate R is said to be *admissible* if for every $\varepsilon > 0$ there exists for some $n = n(\varepsilon)$ an encoder $\mathscr{C}(n,\lfloor\exp(nR)\rfloor)$ and a decoder $\mathscr{D}(n,\lfloor\exp(nR)\rfloor)$ such that $\Pr[\boldsymbol{X}^* \neq \boldsymbol{X}] < \varepsilon$. Otherwise R is called *inadmissible*. Here the symbol $\lfloor x \rfloor$ denotes the largest integer not greater than x. We shall make frequent use of the following well-known theorem. (See, for example, [1, p. 43] or [3, p. 45] for equivalent results.)

Theorem 1: If $R > H(X)$, R is admissible. If $R < H(X)$, R is inadmissible. In this latter case there exists a $\delta > 0$ independent of n such that for every encoder–decoder pair $\mathscr{C}(n,\lfloor\exp(nR)\rfloor)$, $\mathscr{D}(n,\lfloor\exp(nR)\rfloor)$, $\Pr[\boldsymbol{X}^* \neq \boldsymbol{X}] > \delta > 0$. Stated in less formal terms the theorem asserts that for $\eta > 0$ one can achieve arbitrarily small decoding error probability with block codes transmitting at a rate $R = H(X) + \eta$; block codes using a rate $R = H(X) - \eta$ cannot have arbitrarily small probability of error.

We now seek to generalize these notions to correlated sources. Let X and Y be discrete random variables taking values in the sets $\mathscr{A}_X = \{1,2,\cdots,A_X\}$ and $\mathscr{A}_Y = \{1,2,\cdots, A_Y\}$, respectively. Denote their joint probability distribution by

$$p_{XY}(x,y) = \Pr[X = x \text{ and } Y = y], \qquad x \in \mathscr{A}_X, \quad y \in \mathscr{A}_Y. \qquad (1)$$

Next let $(X_1,Y_1), (X_2,Y_2),\cdots,(X_n,Y_n)$ be a sequence of n independent realizations of the pair of random variables (X,Y). Denote by $\boldsymbol{X}$ the sequence $X_1,X_2,\cdots,X_n$ and by $\boldsymbol{Y}$ the sequence $Y_1,Y_2,\cdots,Y_n$. The probability distribution for this correlated pair of vectors is

$$P_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y}) = \Pr[\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}] = \prod_{i=1}^{n} p_{XY}(x_i,y_i) \qquad (2)$$

$$\boldsymbol{x} = (x_1,x_2,\cdots,x_n) \in \mathscr{A}_X{}^n$$

$$\boldsymbol{y} = (y_1,y_2,\cdots,y_n) \in \mathscr{A}_Y{}^n$$

where $\mathscr{A}_X{}^n$ is the set of $A_X{}^n$ distinct n-vectors whose components are in $\mathscr{A}_X$ and $\mathscr{A}_Y{}^n$ is defined analogously. We regard $\boldsymbol{X}$ as a block of n-characters produced by one of two correlated information sources. $\boldsymbol{Y}$ is the corresponding block produced by the other source.

When it comes to encoding the outputs of these correlated sources a number of possibilities of interest present themselves depending upon the information available to the encoders and decoders. Sixteen cases that we shall consider are shown in Fig. 3. Each setting of the switches S_1,S_2,S_3,S_4 yields a new case. It is convenient to associate with switch S_i a state variable s_i taking the value 0 if the switch is open and the value 1 if the switch is closed, $i = 1,2,3,4$. The quadruple $s_1s_2s_3s_4$, always listed in that order, will be used to specify the setting of the switches. Thus 0101 means that switches S_1 and S_3 are open while S_2 and S_4 are closed. The setting 0011 corresponds to Fig. 1.

An X-encoder $\mathscr{C}_X(n,s_2,M_X)$ is a single-valued function from $\mathscr{A}_X{}^n \times \mathscr{A}_Y{}^n$ to the set of integers $\mathscr{M}_X = \{1,2,\cdots,M_X\}$ of the form $i_X = f_X(\boldsymbol{x},s_2\boldsymbol{y})$. Similarly a Y-encoder $\mathscr{C}_Y(n,s_1,$

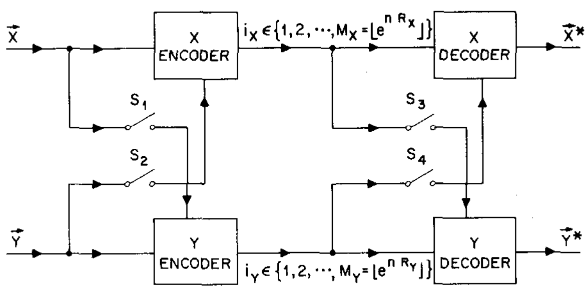
$\vec{X}$
X ENCODER
$i_X \in \{1,2,\cdots,M_X=\lfloor e^{nR_X}\rfloor\}$
X DECODER
$\vec{X}^*$
S_1
S_2
S_3
S_4
$\vec{Y}$
Y ENCODER
$i_Y \in \{1,2,\cdots,M_Y=\lfloor e^{nR_Y}\rfloor\}$
Y DECODER
$\vec{Y}^*$

Fig. 3. Sixteen cases of correlated source coding.

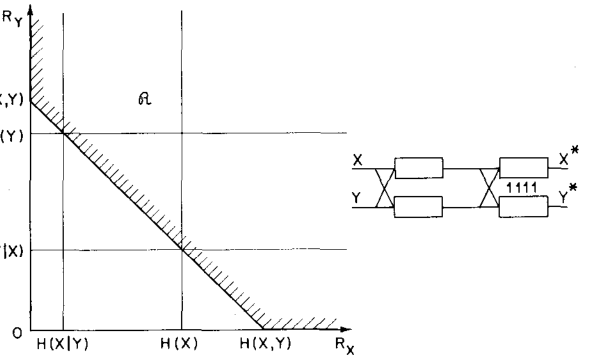

Fig. 4. Admissible rate region.

M_Y) is a single-valued function of the form $i_Y = f_Y(s_1\boldsymbol{x},\boldsymbol{y})$ from $\mathscr{A}_X{}^n \times \mathscr{A}_Y{}^n$ to the set $\mathscr{M}_Y = \{1,2,\cdots,M_Y\}$. Decoders are defined analogously: $\mathscr{D}_X(n,s_4,M_X,M_Y)$ is any single-valued function of the form $\boldsymbol{x}^* = g_X(i_X,s_4i_Y)$ from $\mathscr{M}_X \times \mathscr{M}_Y$ to $\mathscr{A}_X{}^n$ while a Y-decoder $\mathscr{D}_Y(n,s_3,M_X,M_Y)$ is a single-valued function of the form $\boldsymbol{y}^* = g_Y(s_3i_X,i_Y)$ from $\mathscr{M}_X \times \mathscr{M}_Y$ to $\mathscr{A}_Y{}^n$.

Associated with these coders and decoders are rates $R_X = (1/n)\log M_X$, $R_Y = (1/n)\log M_Y$, and random variables $I_X \equiv f_X(\boldsymbol{X},s_2\boldsymbol{Y})$, $I_Y \equiv f_Y(s_1\boldsymbol{X},\boldsymbol{Y})$, $\boldsymbol{X}^* = g_X(I_X,s_4I_Y)$ and $\boldsymbol{Y}^* = g_Y(s_3I_X,I_Y)$. I_X and I_Y are called the *encoded X-message number* and the *encoded Y-message number*, respectively, while $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ are the corresponding decoded blocks. We think of the two encoders as producing the integers I_X and I_Y after n correlated source pairs X,Y have been generated. R_X units of information per source character suffice to transmit I_X to the X-decoder: R_Y units suffice to transmit I_Y to the Y-decoder. The decoders then produce the estimates $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ of the input sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$. The role played by the switches S_1,S_2,S_3,S_4 has been incorporated here into the arguments of various coding functions f_X,f_Y,g_X, and g_Y. Thus, for example, if S_2 is open, $s_2 = 0$ and then $I_X = f_X(\boldsymbol{X},0)$ depends only on the X-sequence.

We are at last in a position to define our problem. A pair of nonnegative numbers R_X,R_Y is said to be an *admissible rate point* for the case $s_1s_2s_3s_4$ if for every $\varepsilon > 0$ there exists for some $n = n(\varepsilon)$ encoders $\mathscr{C}_X(n,s_2,M_X)$, $\mathscr{C}_Y(n,s_1,M_Y)$, and decoders $\mathscr{D}_X(n,s_4,M_X,M_Y)$, $\mathscr{D}_Y(n,s_3,M_X,M_Y)$ with $M_X = \lfloor\exp(nR_X)\rfloor$, $M_Y = \lfloor\exp(nR_Y)\rfloor$, such that, $\Pr[\{\boldsymbol{X}^* \neq \boldsymbol{X}\} \cup \{\boldsymbol{Y}^* \neq \boldsymbol{Y}\}] < \varepsilon$. Otherwise, the pair (R_X,R_Y) is called an *inadmissible rate point*. We denote by $\mathscr{R}$ the closure of the set of admissible rate points. In this paper, we determine the admissible rate region $\mathscr{R}$ for the 16 cases of Fig. 3 for the correlated source described.

II. Discussion of Results

To describe the admissible rate region $\mathscr{R}$ for the various cases of Fig. 3, we must first introduce the marginal and conditional distributions of X and Y, namely,

$$
\begin{aligned}
p_X(x) &= \sum_y p_{XY}(x,y)\\
p_Y(y) &= \sum_x p_{XY}(x,y)\\
p_{X\mid Y}(x \mid y) &= p_{XY}(x,y)/p_Y(y), \qquad p_Y(y) \neq 0\\
p_{Y\mid X}(y \mid x) &= p_{XY}(x,y)/p_X(x), \qquad p_X(x) \neq 0 \qquad (3)
\end{aligned}
$$

and the usual associated information-theoretic numbers

$$
\begin{aligned}
H(X,Y) &= -\sum_x\sum_y p_{XY}(x,y)\log p_{XY}(x,y)\\
H(X) &= -\sum_x p_X(x)\log p_X(x)\\
H(Y) &= -\sum_y p_Y(y)\log p_Y(y)\\
H(Y \mid X) &= -\sum_x p_X(x)\sum_y p_{Y\mid X}(y \mid x)\log p_{Y\mid X}(y \mid x)\\
H(X \mid Y) &= -\sum_y p_Y(y)\sum_x p_{X\mid Y}(x \mid y)\log p_{X\mid Y}(x \mid y). \qquad (4)
\end{aligned}
$$

The regions are described in terms of these quantities.

The 16 cases are covered by Figs. 4–9. Each figure shows a region $\mathscr{R}$ and certain switching configurations that have $\mathscr{R}$ as region of admissible rates. Figs. 5–7 each serve as well for the switch settings shown at the right of the drawing when every X in the figure is replaced by Y and every Y is replaced by X, including those on the small block diagrams.

The cases vary in novelty and interest. For instance, the case 1111 shown in Fig. 4 contains nothing new. To obtain the results shown there, we have only to regard the pair (X,Y) as a new discrete random variable taking on A_XA_Y values and apply Theorem 1. This will be explained in full below.

The case 0011 shown in Fig. 8 is by far the most interesting and novel of our results. Consider for a moment a point near the corner of $\mathscr{R}$ given by $R_Y = H(Y) + \varepsilon$, $R_X = H(X \mid Y) + \varepsilon$, where $\varepsilon > 0$ is thought of as very small. By Theorem 1, a Y-encoder transmitting at this rate R_Y and a Y-decoder exist that allow the Y-source outputs to be recovered with arbitrarily small error probability. We can suppose then that the joint X–Y decoder shown has available the Y outputs. In view of the normal interpretation of $H(X \mid Y)$ as the "uncertainty of X given Y," it is most satisfying then that the X-encoder need only produce a message stream with information rate $R_X = H(X \mid Y) + \varepsilon$. But how can this be done? What properties of $\boldsymbol{X}$ alone must the X-encoder examine and transmit (for it cannot observe the Y source) at the rate $H(X \mid Y) < H(X)$ to allow reconstruction of the $\boldsymbol{X}$ sequence when $\boldsymbol{Y}$ is at last seen at the decoder? The answer is not clear. We obtain our results by a random coding argument which somewhat generalizes that used in the usual noisy channel coding theorem. Since

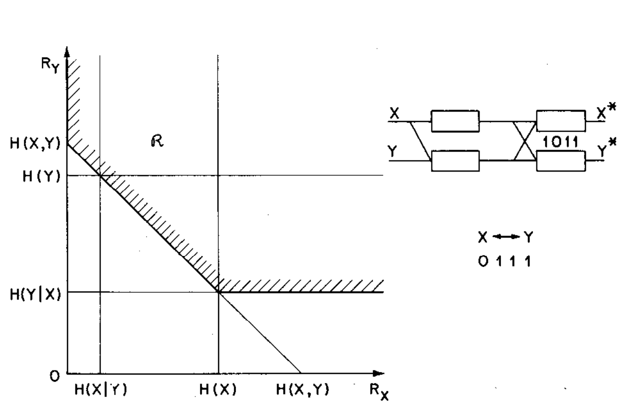

Fig. 5. Admissible rate region.

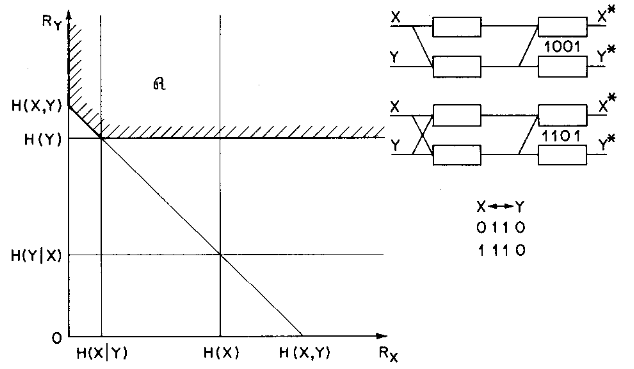

Fig. 6. Admissible rate region.

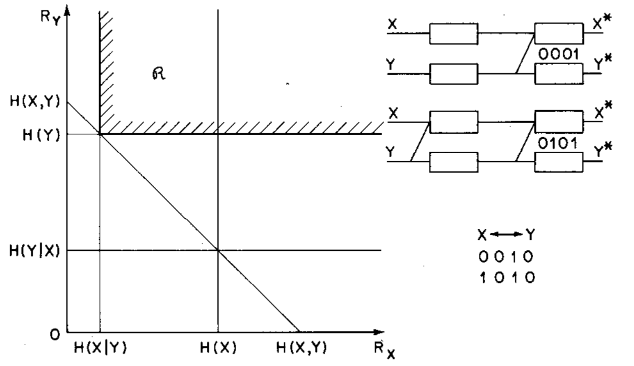

Fig. 7. Admissible rate region.

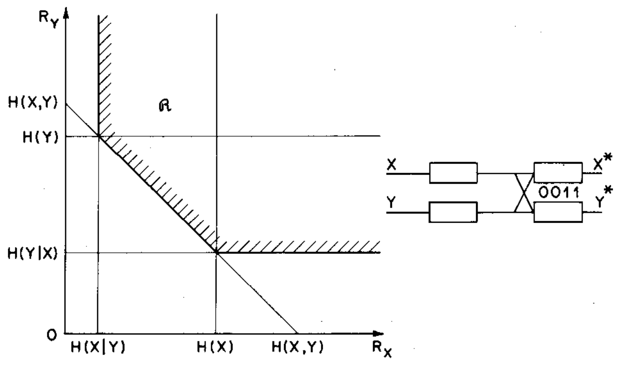

Fig. 8. Admissible rate region.

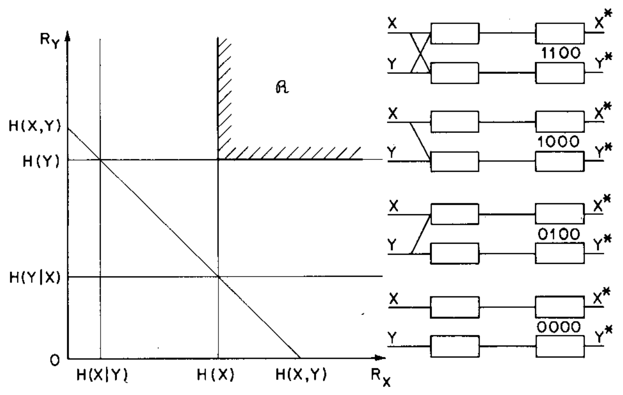

Fig. 9. Admissible rate region.

this is one of the principal contributions of the paper, we turn now to treat this case in some detail, then later proceed to establish more general machinery that allows treatment of the remaining 15 cases with more dispatch.

III. The Case 0011

In this section we prove the following.

Theorem 2:

$$\begin{aligned} R_X &= H(X \mid Y) + \varepsilon_X, \qquad \varepsilon_X > 0 \\ R_Y &= H(Y) + \varepsilon_Y, \qquad \varepsilon_Y > 0 \end{aligned} \tag{5}$$

is an admissible rate point for the case 0011.

In the course of the proof, we shall need the quantities

$$P_{Y \mid X}(\mathbf{y} \mid \mathbf{x}) = \Pr[\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}] = \prod_{i=1}^{n} p_{Y \mid X}(y_i \mid x_i) \tag{6}$$

with notation as in (1)–(3). We shall also use

$$\begin{aligned} I(X,Y) &\equiv H(X) + H(Y) - H(X,Y) = H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \end{aligned} \tag{7}$$

with the H given by (4).

The concepts behind the formal proof that follows are these. By Theorem 1, we know that R_Y is large enough to allow nearly error-free transmissions of the Y-sequences. We shall accordingly think of the n-vector $\mathbf{Y}$ as known as the decoder.

Now, from the fact that $p_{XY}(x,y) = p_{Y \mid X}(y \mid x)p_X(x)$, we can think of the Y-sequences of the correlated source as being generated by applying successive characters of the X-sequence as inputs to a discrete memoryless channel with transition probabilities $p_{Y \mid X}(y \mid x)$. The coding theorem for such a channel tells us that for large n and any $\varepsilon > 0$, there exists a decoder and a code $\mathscr{X}_1$ composed of $M = \lfloor \exp[n(I(X,Y) - \varepsilon)] \rfloor$ n-vectors $\mathbf{x}_{11},\mathbf{x}_{12},\cdots,\mathbf{x}_{1M}$ that can be used as inputs to this channel and decoded with little error probability when the output $\mathbf{Y}$ is seen. Now it turns out that we can find many other codes for this channel, say $\mathscr{X}_2,\mathscr{X}_3,\cdots,\mathscr{X}_{M'}$, each with its own decoder, each of the same size as $\mathscr{X}_1$, and each enjoying the same small probability of error. Our scheme then for encoding the X-sequences of our correlated sources X and Y is to transmit to the decoder the index of the first code in the series $\mathscr{X}_1$, $\mathscr{X}_2,\cdots,\mathscr{X}_M$, that contains X. The X-Y decoder can then use the decoder appropriate for that code $\mathscr{X}_i$ of the $p_{Y\mid X}(y \mid x)$ channel to determine X. There are $\exp\{n[H(X) + \eta]\}$ highly likely X sequences, so that if the codes $\mathscr{X}_1,\mathscr{X}_2,\cdots$, $\mathscr{X}_M$, were disjoint, it would require

$$M' = \exp\{n[H(X) + \eta]\}/\exp\{n[I(X,Y) - \varepsilon]\}$$
$$= \exp\{n[H(X \mid Y) + \varepsilon_X]\}$$

code books to be certain that X was contained in one of them. Although the codes are not disjoint, we shall use just this many.

Let us turn now to the formal proof. Let n be a positive integer, later to be chosen very large, and set

$$M_X = \lfloor \exp(nR_X) \rfloor \qquad M_Y = \lfloor \exp(nR_Y) \rfloor \tag{8}$$

with the R', given as in (5). X- and Y-encoders for the case at hand are functions $i_X = f_X(x)$ from $\mathscr{A}_X{}^n$ to $\mathscr{M}_X = \{1,2,\cdots,M_X\}$ and $i_Y = f_Y(y)$ from $\mathscr{A}_Y{}^n$ to $\mathscr{M}_Y = \{1,2,\cdots,M_Y\}$. A decoder is a pair of functions $x^* = g_X(i_X,i_Y)$ from $\mathscr{M}_X \times \mathscr{M}_Y$ to $\mathscr{A}_X{}^n$ and $y^* = g_Y(i_X,i_Y)$ from $\mathscr{M}_X \times \mathscr{M}_Y$ to $\mathscr{A}_Y{}^n$. In the proof of Theorem 2, it suffices to restrict our attention to coding and decoding functions of a very special form.

To describe the Y-encoder, we must first define the list $\mathscr{T}(\varepsilon,n)$ of typical Y-sequences of length n. Here $\varepsilon > 0$ is a parameter that will stay fixed throughout the rest of this section. Let k be the smallest integer greater than $\sqrt{A_Y/\varepsilon}$ where as before the Y alphabet is $\mathscr{A}_Y = \{1,2,\cdots,A_Y\}$. Let $f_i(y)$ be the number of occurrences of the integer i among the list of components $y_1,y_2,\cdots,y_n$ of y. A Y-sequence y is contained in $\mathscr{T}(\varepsilon,n)$ if

$$|f_i(y) - np_Y(i)| < k\sqrt{np_Y(i)[1 - p_Y(i)]},$$
$$i = 1,2,\cdots,A_Y. \tag{9}$$

If (9) is violated for any i, y is called *atypical* and is not a member of $\mathscr{T}(\varepsilon,n)$. The following facts about typical sequences are well-known. (See, for example, [2, pp. 14–16] for a very readable account.)

Theorem 3: 1) $\Pr[Y \in \mathscr{T}(\varepsilon,n)] > 1 - \varepsilon$. 2) There exists an $A > 0$ such that for every $y \in \mathscr{T}(\varepsilon,n)$

$$\exp[-nH(Y) - A\sqrt{n/\varepsilon}]$$
$$< P_Y(y) < \exp[-nH(Y) + A\sqrt{n/\varepsilon}].$$

Here A is independent of n and ε. 3) The number N of members of $\mathscr{T}(\varepsilon,n)$ is $e^{n[H(Y)+\delta_n(\varepsilon)]}$, where

$$\lim_{n\to\infty} \delta_n(\varepsilon) = 0.$$

We now assume that n is chosen sufficiently large so that $\delta_n(\varepsilon) < \varepsilon_Y$ of (5). Then the number of sequences y in $\mathscr{T}(\varepsilon,n)$ satisfies

$$N \le M_Y \tag{10}$$

with M_Y given by (5) and (8).

The Y-encoder for the correlated sources X and Y is constructed as follows. Number the vectors of $\mathscr{T}(\varepsilon,n)$ to obtain the list $y_1,y_2,\cdots,y_N$. Adjoin to this sequence any $M_Y - N$ other vectors of $\mathscr{A}_Y{}^n$ (not necessarily distinct), labeling them $y_{N+1},y_{N+2},\cdots,y_{M_Y}$. We denote the list $y_1,y_2,\cdots,y_{M_Y}$ by $\mathscr{L}$. Now define the Y-encoder by

$$f_Y(y) = \begin{cases} \text{smallest index } i \text{ such that } y = y_i, & \text{if } y \in \mathscr{L} \\ 1, & \text{if } y \notin \mathscr{L}. \end{cases} \tag{11}$$

The mapping is from $\mathscr{A}_Y{}^n$ to $\mathscr{M}_Y$ as required of a Y-encoder.

The X-encoders are of a very special form. Let $\mathscr{X}_i = \{x_{i1},x_{i2},\cdots,x_{iM}\}$ be a list of M vectors of $\mathscr{A}_X{}^n$, $i = 1,2,\cdots$, M_X. The vectors in these lists need not be distinct. We call each list $\mathscr{X}_i$ an *X-code* and we call the collection $\mathscr{X}$ of M_X X-codes an *X-supercode*. We shall specify how M is to be chosen later. The X-encoders we consider are of the form

$$f_X(x) = \begin{cases} 1, & \text{if } x \notin \mathscr{X} \\ \text{smallest index } i \text{ such that } x \in \mathscr{X}_i, \\ \quad \text{if } x \in \mathscr{X}. \end{cases} \tag{12}$$

To define the decoding functions we set

$$g_Y(i_X,i_Y) = y_{i_Y} \tag{13}$$

for all $(i_X,i_Y) \in \mathscr{M}_X \times \mathscr{M}_Y$. The X-decoder is somewhat more complicated. Denote by $j(i_X,i_Y)$ the smallest index j such that

$$P_{Y\mid X}(y_{i_Y} \mid x_{i_Xj}) \ge P_{Y\mid X}(y_{i_Y} \mid x_{i_Xk}),$$
$$k = 1,2,\cdots,M. \tag{14}$$

Then the X-decoder is given by

$$g_X(i_X,i_Y) = x_{i_X,j(i_X,i_Y)}, \tag{15}$$

for all $(i_X,i_Y) \in \mathscr{M}_X \times \mathscr{M}_Y$.

As in the introduction, we introduce the random variables

$$I_X = f_X(X)$$
$$I_Y = f_Y(Y)$$
$$X^* = g_X(I_X,I_Y)$$
$$Y^* = g_Y(I_X,I_Y).$$

We wish to show that for every $\varepsilon' > 0$ there exists an X supercode $\mathscr{X}$ such that

$$P_e(\mathscr{X}) \equiv \Pr[\{X^* \ne X\} \cup \{Y^* \ne Y\}] < \varepsilon'. \tag{16}$$

We cannot exhibit such an X-supercode explicitly, but we will establish the existence of one by the now familiar random coding argument. We average $P_e(\mathscr{X})$ over an ensembles $\mathscr{E}$ of X-supercodes and show that this average, $\overline{P_e(\mathscr{X})}$, is less than ε'. At least one member of the ensemble must then satisfy (16).

A supercode of $\mathscr{E}$ is specified by particular values of the M_XM random vectors X_{ij}, $i = 1,\cdots,M_X$, $j = 1,\cdots,M$. The values lie in $\mathscr{A}_X{}^n$. The probability structure of $\mathscr{E}$ is specified by

$$\Pr\left[\mathbf{X}_{ij} = \mathbf{x}_{ij},\ i = 1,\cdots,M_X,\ j = 1,\cdots,M\right] = \prod_{i=1}^{M_X}\prod_{j=1}^{M} P_{\mathbf{X}}(\mathbf{x}_{ij}) \tag{17}$$

where

$$P_{\mathbf{X}}(\mathbf{x}) = \prod_{1}^{n} p_X(x_i) \tag{18}$$

in the notation of (2) and (3). Stated otherwise, the vectors of the supercode are drawn component by component independently from the marginal $p_X(x)$ of the given joint distribution $p_{XY}(x,y)$.

Suppose now all supercodes are enumerated and listed $-\mathscr{X}^{(1)}, \mathscr{X}^{(2)},\cdots$. The average error probability we seek is

$$\begin{aligned}\overline{P_e(\mathscr{X})} &= \sum_j \Pr\left(\mathscr{X} = \mathscr{X}^{(j)}\right)\\ &\quad\cdot \Pr\left[\{\mathbf{X}^* \neq \mathbf{X}\} \cup \{\mathbf{Y}^* \neq \mathbf{Y}\} \mid \mathscr{X} = \mathscr{X}^{(j)}\right]\\ &= \sum_j \Pr\left[\{\mathbf{X}^* \neq \mathbf{X}\} \cup \{\mathbf{Y}^* \neq \mathbf{Y}\} \text{ and } \mathscr{X} = \mathscr{X}^{(j)}\right].\end{aligned}$$

This last sum can be interpreted as the probability P_e that $(\mathbf{X}^*,\mathbf{Y}^*) \neq (\mathbf{X},\mathbf{Y})$ in the joint experiment of choosing an X-supercode $\mathscr{X}$ from $\mathscr{E}$ and independently choosing an $\mathbf{X}$ and $\mathbf{Y}$ from $p_{\mathbf{XY}}(\mathbf{x},\mathbf{y})$ to use with that supercode. We proceed to upperbound this quantity.

Let

$$\begin{aligned}P_1 &= \Pr\left[\mathbf{Y} \notin \mathscr{L}\right]\\ P_2 &= \Pr\left[\mathbf{X} \notin \mathscr{X}\right]\\ P_3 &= \Pr\left[\mathbf{Y} \in \mathscr{L}, \mathbf{X} \in \mathscr{X}, \mathbf{X}^* \neq \mathbf{X}\right].\end{aligned} \tag{19}$$

Then clearly

$$\overline{P_e(\mathscr{X})} \leq P_1 + P_2 + P_3. \tag{20}$$

That

$$P_1 < \varepsilon \tag{21}$$

follows directly from statement 1) of Theorem 3 about typical sequences and from the fact that $\mathscr{L}$ includes $\mathscr{T}(\varepsilon,n)$.

We now show that if n is large enough and

$$M = \lfloor \exp\{n[I(X,Y) - \tfrac{1}{2}\varepsilon_X]\}\rfloor, \tag{22}$$

where we assume $I - \frac{1}{2}\varepsilon_X > 0$, then

$$P_2 < 2\varepsilon. \tag{23}$$

We first note that

$$\begin{aligned}P_2 = \Pr\left[\mathbf{X} \notin \mathscr{X}\right] &= \sum_{\mathbf{x}} \Pr\left[\mathbf{X} = \mathbf{x}\right] \Pr\left[\mathbf{X} \notin \mathscr{X} \mid \mathbf{X} = \mathbf{x}\right]\\ &= \sum_{\mathbf{x}} P_{\mathbf{X}}(\mathbf{x})[1 - P_{\mathbf{X}}(\mathbf{x})]^{MM_X}\end{aligned} \tag{24}$$

with $P_{\mathbf{X}}(\mathbf{x})$ given by (18). From Theorem 3, the set of X-sequences of length n can be divided into two parts, one $\mathscr{T}_X{}^c(\varepsilon,n)$ of probability $< \varepsilon$ and a disjoint part $\mathscr{T}_X(\varepsilon,n)$ such that

$$\exp\left\{-n\left[H(X) + \frac{A'}{\sqrt{n\varepsilon}}\right]\right\} \leq P_{\mathbf{X}}(\mathbf{x}), \tag{25}$$

for all $\mathbf{x} \in \mathscr{T}_X(\varepsilon,n)$. Here $A' > 0$ is independent of n and ε. We write (24) as

$$\begin{aligned}P_2 &= \sum_{\mathbf{x} \in \mathscr{T}_X} P_{\mathbf{X}}(\mathbf{x})[1 - P_{\mathbf{X}}(\mathbf{x})]^{MM_X}\\ &\quad + \sum_{\mathbf{x} \in \mathscr{T}_X^c} P_{\mathbf{X}}(\mathbf{x})[1 - P_{\mathbf{X}}(\mathbf{x})]^{MM_X}\\ &\leq \left[1 - \exp\left\{-n\left[H(X) + \frac{A'}{\sqrt{n\varepsilon}}\right]\right\}\right]^{MM_X}\\ &\quad \cdot \sum_{\mathbf{x} \in \mathscr{T}_X} P_{\mathbf{X}}(\mathbf{x}) + \sum_{\mathbf{x} \in \mathscr{T}_X^c} P_{\mathbf{X}}(\mathbf{x})\\ &\leq \left[1 - \exp\left\{-n\left[H(X) + \frac{A'}{\sqrt{n\varepsilon}}\right]\right\}\right]^{MM_X} + \varepsilon\\ &= Z + \varepsilon.\end{aligned} \tag{26}$$

Now

$$\begin{aligned}\log Z &= MM_X \log\left[1 - \exp\left\{-n\left[H(X) + \frac{A'}{\sqrt{n\varepsilon}}\right]\right\}\right]\\ &\leq -\exp\left(n[I(X,Y) - \tfrac{1}{2}\varepsilon_X]\right)\exp\left\{n[H(X \mid Y) + \varepsilon_X]\right\}\\ &\quad \cdot \exp\left\{-n\left[H(X) + \frac{A'}{\sqrt{n\varepsilon}}\right]\right\}\\ &= -\exp\left\{n\left[\frac{1}{2}\varepsilon_X - \frac{A'}{\sqrt{n\varepsilon}}\right]\right\}\end{aligned}$$

on using (5), (7), (8), and (22). Thus, for any fixed ε, Z approaches zero as n becomes large. We can therefore choose n sufficiently large to make $Z < \varepsilon$ and then, from (26), (23) is true.

For the remaining term in (20) we have

$$\begin{aligned}P_3 &= \Pr\left[\mathbf{Y} \in \mathscr{L}, \mathbf{X} \in \mathscr{X}, \mathbf{X}^* \neq \mathbf{X}\right]\\ &\leq \Pr\left[\mathbf{X} \in \mathscr{X}, \mathbf{X}^* \neq \mathbf{X}\right]\\ &= \sum_{\mathbf{xy}} P_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \Pr\left[\mathbf{X} \in \mathscr{X}, \mathbf{X}^* \neq \mathbf{X} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\right]\\ &= \sum_{\mathbf{xy}} P_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) A(\mathbf{x},\mathbf{y}).\end{aligned} \tag{27}$$

Here

$$\begin{aligned}A(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M_X} \Pr\big[&\mathbf{X} \notin \mathscr{X}_1, \mathbf{X} \notin \mathscr{X}_2,\cdots,\mathbf{X} \notin \mathscr{X}_{i-1},\\ &\mathbf{X} \in \mathscr{X}_i, \mathbf{X}^* \neq \mathbf{X} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\big]\end{aligned}$$

and the term $i = 1$ is to be interpreted as $\Pr[\mathbf{X} \in \mathscr{X}_1, \mathbf{X}^* \neq \mathbf{X} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}]$. Now

$$A(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M_X} [1 - P_{\mathbf{X}}(\mathbf{x})]^{(i-1)M} \sum_{j=1}^{M} B_{ij}(\mathbf{x},\mathbf{y}) \tag{28}$$

with

$$\begin{aligned}B_{ij}(\mathbf{x},\mathbf{y}) = \Pr\big[&\mathbf{X}_{i1} \neq \mathbf{x}, \mathbf{X}_{i2} \neq \mathbf{x},\cdots,\mathbf{X}_{ij-1} \neq \mathbf{x}, \mathbf{X}_{ij} = \mathbf{x},\\ &\mathbf{X}^* \neq \mathbf{X} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\big].\end{aligned}$$

The terms B_{i1} are to be interpreted as $\Pr[\mathbf{X}_{i1} = \mathbf{x}, \mathbf{X}^* \neq \mathbf{X} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}]$. But

$$B_{ij}(\mathbf{x},\mathbf{y}) = [1 - P_{\mathbf{X}}(\mathbf{x})]^{j-1} P_{\mathbf{X}}(\mathbf{x}) C_{ij}(\mathbf{x},\mathbf{y}) \tag{29}$$

with $C_{ij}(x,y) \leq \Pr\,[P_{Y\mid X}(y \mid X_{i\alpha}) \geq P_{Y\mid X}(y \mid x)$, for some $\alpha \neq j \mid \mathscr{D}]$ where $\mathscr{D} \equiv \{X_{i1} \neq x,\cdots,X_{ij-1} \neq x, X_{ij} = x, X = x, Y = y\}$.

$$C_{ij}(x,y) \leq \Pr\left[\bigcup_{\alpha\neq j} \{P_{Y\mid X}(y \mid X_{i\alpha}) \geq P_{Y\mid X}(y \mid x)\} \mid \mathscr{D}\right]$$

$$\leq \left[\sum_{\alpha\neq j} \Pr\,[P_{Y\mid X}(y\mid X_{i\alpha}) \geq P_{Y\mid X}(y\mid x) \mid \mathscr{D}]\right]^{\rho} \quad (30)$$

where $0 \leq \rho \leq 1$. Here we have used the now common union bound, (see [1, p. 136]),

$$\Pr\left[\bigcup_{i} \{\mathscr{A}_i\}\right] \leq \left[\sum_{i} \Pr\,[\mathscr{A}_i]\right]^{\rho}.$$

But if $\alpha > j$

$$\Pr\,[P_{Y\mid X}(y \mid X_{i\alpha}) \geq P_{Y\mid X}(y \mid x) \mid \mathscr{D}]$$
$$= \Pr\,[P_{Y\mid X}(y \mid X_{i\alpha}) \geq P_{Y\mid X}(y \mid x) \mid X = x, Y = y] \quad (31)$$

while if $\alpha < j$

$$\Pr\,[P_{Y\mid X}(y\mid X_{i\alpha}) \geq P_{Y\mid X}(y\mid x) \mid \mathscr{D}]$$
$$= \Pr\,[P_{Y\mid X}(y\mid X_{i\alpha}) \geq P_{Y\mid X}(y\mid x) \mid X_{i\alpha} \neq x, X = x, Y = y]$$
$$= \frac{\Pr\,[P_{Y\mid X}(y\mid X_{i\alpha}) \geq P_{Y\mid X}(y\mid x), X_{i\alpha} \neq x \mid X = x, Y = y]}{\Pr\,[X_{i\alpha} \neq x \mid X = x, Y = y]}$$
$$\leq a^{-1} \Pr\,[P_{Y\mid X}(y\mid X_{i\alpha}) \geq P_{Y\mid X}(y\mid x) \mid X = x, Y = y] \quad (32)$$

where

$$a \equiv \Pr\,[X_{i\alpha} \neq x \mid X = x, Y = y] = [1 - P_X(x)] \leq 1. \quad (33)$$

Thus, (30)–(32) give

$$C_{ij}(x,y) \leq [\{(j-1)a^{-1} + M - j\} \Pr\,[P_{Y\mid X}(y \mid X_{i\alpha}) \geq P_{Y\mid X}(y \mid x) \mid X = x, Y = y]]^{\rho}$$
$$= [(j-1)a^{-1} + M - j]^{\rho} \left[\sum_{x_{i\alpha}}{}' P_X(x_{i\alpha})\right]^{\rho}$$

where the sum is over the set of values of $x_{i\alpha}$ for which $P_{Y\mid X}(y \mid x_{i\alpha}) \geq P_{Y\mid X}(y \mid x)$. Since on this set $P_{Y\mid X}(y \mid x_{i\alpha})/P_{Y\mid X}(y \mid x) \geq 1$, we have

$$C_{ij}(x,y) \leq [(j-1)a^{-1} + M - j]^{\rho} \cdot \left[\sum_{x'} P_X(x') \left[\frac{P_{Y\mid X}(y \mid x')}{P_{Y\mid X}(y \mid x)}\right]^{s}\right]^{\rho}$$

for any $s \geq 0$.

Let us now assemble these results. Equations (27)–(29) and (34) give

$$P_3 < \sum_{xy} P_{XY}(x,y) \sum_{i=1}^{M_X} a^{(i-1)M} \cdot \sum_{j=1}^{M} a^{j-1}P_X(x)[(j-1)a^{-1} + M - j]^{\rho} \cdot \left[\sum_{x'} P_X(x') \left[\frac{P_{Y\mid X}(y \mid x')}{P_{Y\mid X}(y \mid x)}\right]^{s}\right]^{\rho}. \quad (34)$$

In the Appendix it is shown that

$$\sum_{i=1}^{M_X} a^{(i-1)M} \sum_{j=1}^{M} a^{j-1}[(j-1)a^{-1} + M - j]^{\rho} \leq \frac{M^{\rho}}{P_X(x)}. \quad (35)$$

Insert this bound for the sums into (34) and write $P_{XY}(x,y) = P_X(x)P_{Y\mid X}(y \mid x)$. There results

$$P_3 \leq \sum_{xy} P_X(x)P_{Y\mid X}(y \mid x)^{1-\rho s}M^{\rho} \cdot \left[\sum_{x'} P_X(x')P_{Y\mid X}(y \mid x')^{s}\right]^{\rho}.$$

Choose $s = 1/(1 + \rho)$ to obtain

$$P_3 \leq T \equiv M^{\rho} \sum_{y} \left[\sum_{x} P_X(x)P_{Y\mid X}(y \mid x)^{1/(1+\rho)}\right]^{1+p}. \quad (36)$$

The quantity T is well known to information theorists. It plays a central role in the Gallager bound for error probability of a noisy memoryless channel. Recalling (6), (18), and (22), we have

$$T = \left[\exp\left\{n\rho\left[I(X,Y) - \frac{\varepsilon_X}{2}\right]\right\}\right] \cdot \left[\sum_{y}\left[\sum_{x} p_X(x)p_{Y\mid X}(y \mid x)^{1/(1+\rho)}\right]^{1+\rho}\right]^{n}$$
$$\leq \exp\left\{-np\left[\frac{1}{\rho}V(\rho) - I(X,Y) + \frac{\varepsilon_X}{2}\right]\right\}.$$

Here

$$V(\rho) = -\log \sum_{y} \left[\sum_{x} p_X(x)p_{Y\mid X}(y \mid x)^{1/(1+\rho)}\right]^{1+\rho}$$

is seen to be analytic in the neighborhood of $\rho = 0$, and indeed $V(0) = 0$. The function $E(\rho) = V(\rho)/\rho$ is also analytic in this neighborhood and by L'Hôpital's rule and a straightforward calculation one finds

$$E(0) = \left.\frac{dV}{d\rho}\right|_{\rho=0} = -\sum_{y} p_Y(y) \log p_Y(y) + \sum_{xy} p_{XY}(x,y) \log p_{Y\mid X}(y \mid x) = I(X,Y)$$

by (4) and (7).

Since $E(\rho)$ is analytic at $\rho = 0$, there exists a $\rho_0 > 0$ such that $|E(\rho_0) - I(X,Y)| < \varepsilon_X/4$ whence (36) becomes

$$P_3 \leq \exp\,(-n\rho_0\varepsilon_X/4).$$

It is now seen that by choosing n sufficiently large, $P_3 < \varepsilon$. Combined with (23), (21), and (20), this shows that $P_e \leq 4\varepsilon$. Now choose $\varepsilon = \varepsilon'/4$. We have then shown that $\overline{P_e(\mathscr{X})} < \varepsilon'$. There must therefore exist a code in the ensemble for which (16) holds. Thus Theorem 2 is proved.

IV. DETERMINATION OF THE REGIONS $\mathscr{R}$

Table I lists twelve theorems whose applications in connection with Fig. 10 give immediately the admissible rate region $\mathscr{R}$ for the 16 cases. In Table I, the symbol x for the state of a switch means that the theorem holds both when

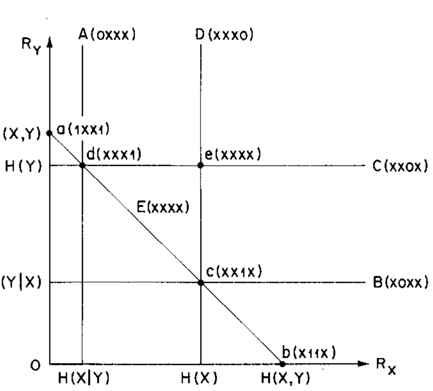

Fig. 10. Lines and points of Table I.

TABLE I
THEOREMS TO DETERMINE $\mathscr{R}$

$s_1s_2s_3s_4$	Theorem Name	Theorem	
		It is necessary that:	
0xxx	A	$R_X \geq H(X \mid Y)$	
x0xx	B	$R_Y \geq H(Y \mid X)$	
xx0x	C	$R_Y \geq H(Y)$	
xxx0	D	$R_X \geq H(X)$	
xxxx	E	$R_X + R_Y \geq H(X,Y)$	
		It is sufficient that:	
1xx1	a	$R_X = 0$	$R_Y = H(X,Y) + \varepsilon_{XY}$
x11x	b	$R_X = H(X,Y) + \varepsilon_{XY}$	$R_Y = 0$
xx1x	c	$R_X = H(X) + \varepsilon_X$	$R_Y = H(Y \mid X) + \varepsilon_Y$
xxx1	d	$R_X = H(X \mid Y) + \varepsilon_X$	$R_Y = H(Y) + \varepsilon_Y$
xxxx	e	$R_X = H(X) + \varepsilon_X$	$R_Y = H(Y) + \varepsilon_Y$
		$\varepsilon_X, \varepsilon_Y, \varepsilon_{XY} > 0$	
		Bit stuffing:	
xxxx	f	$(R_X,R_Y) \in \mathscr{R} \Rightarrow (R_X + \delta_X, R_Y + \delta_Y) \in \mathscr{R}$	
		$\delta_X, \delta_Y \geq 0$	
		Limited time sharing:	
xxxx	g	If $(R_X,R_Y) \in \mathscr{R}$, $(R_X',R_Y') \in \mathscr{R}$ and $R_X + R_Y = H(X,Y)$ and $R_X' + R_Y' = H(X,Y)$, then $(R_X'',R_Y'') \in \mathscr{R}$, $R_X'' = \lambda R_X + (1 - \lambda)R_X'$ $R_Y'' = \lambda R_Y + (1 - \lambda)R_Y'$, $0 \leq \lambda \leq 1$	

the switch is open and when the switch is closed. For example, Theorem A asserts that if switch S_1 is open the $\mathscr{R}_X$ coordinate of an admissible rate point must be at least as large as $H(X \mid Y)$. Theorem E asserts that the coordinates of all admissible rate points, independent of the switch settings, must satisfy $R_X + R_Y \geq H(X,Y)$. Theorem a says that if switches S_1 and S_4 are closed, then $R_X = 0$, $R_Y = H(X,Y) + \varepsilon_{XY}$ is an admissible point for every $\varepsilon_{XY} > 0$. Since $\mathscr{R}$ is defined as the closure of the set of admissible rate points, the theorem also asserts that $R_X = 0$, $R_Y = H(X,Y)$ is in $\mathscr{R}$.

On Fig. 10 certain lines and points are labeled with the names of theorems of Table I. The corresponding switching states are affixed. These points and lines can be used along with Theorems f and g to determine immediately the boundary of $\mathscr{R}$ for any of the 16 cases. The bit stuffing Theorem f then shows that the interior of the boundary can be filled in to obtain $\mathscr{R}$. We give several illustrations.

1) For the switch setting 1011, we see at once from Table I, scanning the columns of switching states, that Theorems B E a c d e apply. The first two show that $\mathscr{R}$ cannot extend below the line labeled B on Fig. 10 nor below the line labeled E there. The next four applicable theorems show that the points a, c, d, and e all lie in $\mathscr{R}$. Theorem f then shows that points above a on the R_Y-axis are in $\mathscr{R}$ as well as all points on B to the right of c. Theorem g shows that the line segment $\overline{ac}$ is in $\mathscr{R}$. The region as given in Fig. 5 is thus established.

2) For the setting 0001, Table I shows that A B C E d e all apply. Locating the lines ABCE on Fig. 10, we see that $\mathscr{R}$ can neither extend to the left of line A nor below line C. The point d is in $\mathscr{R}$, and by Theorem f so is every point to the right of it on line C and every point above it on line A. The region $\mathscr{R}$ of Fig. 7 is thus established.

Many of the theorems of Table I are trivial and we do not belabor them.

Theorem E: The pair of random variables X,Y can be regarded as a single random variable Z taking A_XA_Y values. The entropy of this variable is $H(Z) = H(X,Y)$. Any encoding of the pair (X,Y) as described by Fig. 3 can also be regarded as an encoding of Z by indexing the M_XM_Y possible pairs of values (i_X,i_Y) for I_X and I_Y and taking this index of (I_X,I_Y) as the value of I_Z. If (R_X,R_Y) were admissible and $R_X + R_Y < H(X,Y)$, say $R_X + R_Y = H(X,Y) - \delta$, the construction just mentioned would show the existence of Z codes with $M_Z = M_XM_Y = \lfloor \exp(nR_X) \rfloor \lfloor \exp(nR_Y) \rfloor \leq \lfloor \exp[n(R_X + R_Y)] \rfloor = \lfloor \exp[n(H(X,Y) - \delta)] \rfloor$ values for the channel symbols that had error $< \varepsilon$. But this contradicts Theorem 1 as applied to the variable Z. Q.E.D.

Theorem A: Let switch S_1 be open and suppose that $R_X = H(X \mid Y) - \delta$ and $R_Y = R_2$ is an admissible rate point. We first show that this implies that $R_X = H(X \mid Y) - \delta$, $R_Y = H(Y) + \delta/2$ is also an admissible rate point.

Let $\mathscr{C}_X, \mathscr{C}_Y, \mathscr{D}_X, \mathscr{D}_Y$ be encoders and decoders that employ coded message rates $R_X = H(X \mid Y) - \delta$ and $R_Y = R_2$ and that achieve error probability ε. We replace $\mathscr{C}_Y$ by an encoder $\mathscr{C}_Y'$ that produces coded messages at a rate $R_2 = H(Y) + \delta/2$ by using a list of typical Y sequences. We know by Theorem 1 that for large enough n such an encoder and a corresponding decoder $\mathscr{D}_Y'$ exist, ones that reproduce the Y sequence with arbitrary accuracy. We now consider a new decoder $\mathscr{D}_Y''$ consisting of $\mathscr{D}_Y'$ followed by $\mathscr{C}_Y$ and $\mathscr{D}_Y$. The scheme $\mathscr{C}_X, \mathscr{C}_Y', \mathscr{D}_X, \mathscr{D}_Y''$ signals at rates $R_X = H(X \mid Y) - \delta$ and $R_Y = H(Y) + \delta/2$ with error probability $< \varepsilon$. But $R_X + R_Y = H(X \mid Y) + H(Y) - \delta/2 = H(X,Y) - \delta/2$, contrary to Theorem E. Therefore, we must have $R_X \geq H(X \mid Y)$ for an admissible point.

Theorem B: Theorem A with X and Y interchanged.

Theorems C and D: Follow directly from Theorem 1.

Theorem a: Theorem 1 applied to $Z = (X,Y)$.

Theorem b: Theorem a with X and Y interchanged.

Theorem c: Theorem 2.

Theorem d: Theorem 2 with X and Y interchanged.

Theorem e: Follows from Theorem 1.

Theorem f: This theorem follows from the simple observation that for any encoder, say $\mathscr{C}_X(n,s_2,M_X)$, mapping elements of $\mathscr{A}_X{}^n \times \mathscr{A}_Y{}^n$ onto integers of the set $\mathscr{M}_X = \{1,2,\cdots,M_X\}$ we can always trivially increase the range of the mapping by increasing M_X (and hence R_X). The new values of the enlarged set $\mathscr{M}_X{}'$ never occur as values of I_X in this new mapping and so the decoder can be defined arbitrarily for these values. The error probability remains unchanged.

Theorem g: A result somewhat stronger than Theorem 1 is the following. If $R > H(X)$, then for every $\varepsilon > 0$ there exists an $n_0(\varepsilon)$ and an encoder $\mathscr{C}(n_0,\lfloor\exp(n_0R)\rfloor)$ and a decoder $\mathscr{D}(n_0,\lfloor\exp(n_0R)\rfloor)$ such that $\Pr[X^* \neq X] < \varepsilon$. Furthermore, for each integer $n > n_0$ there exists an encoder $\mathscr{C}(n,\lfloor\exp(nR)\rfloor)$ and a decoder $\mathscr{D}(n,\lfloor\exp(nR)\rfloor)$ such $\Pr[X^* \neq X] < \varepsilon$. This result is implicit in proofs of Theorem 1 that compute explicit bounds for error probability such as the one given by Jelinek [3, sec. 5.2, p. 86]. Examination of Section III then shows that a correspondingly strengthened statement of Theorem 2 is also possible: if $R_X = H(X \mid Y) + \varepsilon_X$, $R_Y = H(Y) + \varepsilon_Y$ then for every $\varepsilon > 0$ there exist coders and decoders for every n greater than some $n_0(\varepsilon)$ for which $\Pr[X^* \neq X \text{ and } Y^* \neq Y] < \varepsilon$. We call a rate point (R_X,R_Y) *strongly admissible* if for every $\varepsilon > 0$ and all sufficiently large n there exist encoders and decoders using block length n for which $\Pr[X^* \neq X$ and $Y^* \neq Y] < \varepsilon$. When points a, b, c, d, or e are in $\mathscr{R}$, i.e., when the switch settings are appropriate, they are strongly admissible rate points.

For strongly admissible rate points (R_X,R_Y) and $(R_X{}',R_Y{}')$, Theorem g is shown as follows. There are encoders and decoders for all block lengths n greater than $n_0(\varepsilon/2)$ that use encoded message rates R_X and R_Y with error $<\varepsilon/2$. Similarly there are encoders and decoders for all block lengths n' greater than $n_0{}'(\varepsilon/2)$ that use encoded message rates $R_X{}'$ and $R_Y{}'$ with error $<\varepsilon/2$. Let λ, $0 \leq \lambda \leq 1$, be rational and choose n'' so large that $n = \lambda n''$ and $n' = (1-\lambda)n''$ are both integers and $n \geq n_0$ and $n' \geq n_0{}'$. Now encode $X - Y$ sequences by alternately using first block length n and the code with rate (R_X,R_Y) and then block length n' with the code of rate $(R_X{}',R_Y{}')$. This operation can be regarded as the use of a single block code of length $n'' = n + n'$. For this new code, $M_X{}'' = M_XM_X{}' = \lfloor\exp(\lambda n''R_X)\rfloor\lfloor\exp[(1-\lambda)n''R_X{}']\rfloor \leq \lfloor\exp(n''R_X{}'')\rfloor$. As in the proof of Theorem f, we can artificially increase $M_X{}''$ so that $M_X{}'' = \lfloor\exp(n''R_X{}'')\rfloor$. A similar calculation holds for $M_Y{}''$. The error probability for this n'' code is less than $1 - (1-\varepsilon/2)^2 = \varepsilon - (\varepsilon/4)^2 \leq \varepsilon$. This establishes Theorem g for rational λ. But $\mathscr{R}$ was defined as the closure of all admissible points and since the rationals are dense in the reals, Theorem g is established.

A stronger form of Theorem g is indeed true, for examination of Figs. 4–9 shows that $\mathscr{R}$ is convex for all 16 cases. Thus we have the following theorem.

Theorem h: If $(R_X,R_Y) \in \mathscr{R}$ and $(R_X{}',R_Y{}') \in \mathscr{R}$, then for every $\lambda, 0 \leq \lambda \leq 1, (R_X{}'',R_Y{}'') \in \mathscr{R}$ where $R_X{}'' = \lambda R_X + (1-\lambda)R_X{}', R_Y{}'' = \lambda R_Y + (1-\lambda)R_Y{}'$.

V. Commentary

Many topics for further research on joint coding of correlated sources suggest themselves. We mention a few.

How does the foregoing extend to N correlated sources instead of two? The number of switch settings grows rapidly with N. Many cases are easy extensions of our results for $N = 2$, but basically new situations arise too. For example, when $N = 3$ consider the case where the X decoder sees I_X and I_Y, the Y decoder sees I_Y and the Z decoder sees I_Y and I_Z. What is the admissible rate region then if the encoder sees all three message sources?

How does the foregoing extend to a rate-distortion theory of correlated sources? The probability-of-error criterion is then replaced by more general measures of decoder output fidelity. A rate-distortion theory would permit a generalization from discrete-valued to continuous-valued random variables.

The design of block codes of given length n to have small error probability is a more difficult and more interesting problem for correlated sources than for a single source, where the problem is solved by providing a list of typical sequences. Here, in cases such as 0011 one wants to take advantage of the known correlation of the sources. Are there better methods than timesharing to achieve rates along the line E between c and d?

What is the theory of variable-length encodings for correlated sources? How does one generalize the Huffman code, say, in the case 0011 to encode X sequences of length n with fewest bits on the average when $R_Y = H(Y)$?

How does the theory extend for correlated sources that are not independent drawings of pairs of correlated variables?

These are but a few of the many interesting problems to be solved in this area.

Acknowledgment

The authors gratefully acknowledge many valuable discussions with Prof. N. T. Gaarder on this research.

Appendix

Here we establish the inequality (35). Jensen's theorem (see, for example, [1, (4.4.4) and (4.4.5), p. 85]) asserts that if $g(x)$ is convex up for $a \leq x \leq b$, i.e., if $g''(x) \leq 0$, for $a \leq x \leq b$, then

$$\sum_1^M p_j g(x_j) \leq g\left(\sum_1^M p_j x_j\right)$$

where $a \leq x_1 \leq x_2 \cdots \leq x_M \leq b$ and

$$\sum_1^M p_j = 1$$

with $p_j \geq 0, j = 1,\cdots,M$. We apply this theorem to the following function

$$g(x) = x^\rho, \qquad 0 \leq \rho \leq 1$$

which is convex up for $x \geq 0$, taking

$$p_j = \frac{1-a}{1-a^M} a^{j-1}$$

$$x_j = [(j-1)a^{-1} + M - j]$$

$$j = 1,2,\cdots,M$$

with a given by (33), so that $0 < a < 1$. We find

$$\frac{1-a}{1-a^M} \sum_{j=1}^{M} a^{j-1}[(j-1)a^{-1} + M - j]^\rho$$

$$\leq \left[\frac{1-a}{1-a^M} \sum_{j=1}^{M} a^{j-1}[(j-1)a^{-1} + M - j]\right]^\rho$$

$$= \left[\frac{1-a}{1-a^M}\left\{(M - a^{-1}) \sum_{j=1}^{M} a^{j-1} + (a^{-1} - 1) \sum_{j=1}^{M} ja^{j-1}\right\}\right]^\rho$$

$$= \left[\frac{1-a}{1-a^M}\left\{M \frac{1-a^{M-1}}{1-a}\right\}\right]^\rho.$$

Here we have evaluated the sums using the two formulas

$$\sum_{j=1}^{M} a^{j-1} = \frac{1-a^M}{1-a}$$

$$\sum_{j=1}^{M} ja^{j-1} = \frac{d}{da}\left(\frac{1-a^{M+1}}{1-a}\right) = \frac{1-(M+1)a^M + Ma^{M+1}}{(1-a)^2} \tag{A-1}$$

and performed some algebraic simplification. It follows then that

$$\sum_{j=1}^{M} a^{j-1}[(j-1)a^{-1} + M - j]^\rho \leq \frac{1-a^M}{1-a} M^\rho \left(\frac{1-a^{M-1}}{1-a^M}\right)^\rho$$

$$\leq \frac{1-a^M}{1-a} M^\rho \tag{A-2}$$

since if $0 < a < 1$, $[1 - a^{M-1}]/[1 - a^M] < 1$.

Returning to (35), from (A-1) we have, on replacing M by M_X and a by a^M,

$$\sum_{i=1}^{M_X} a^{M(i-1)} = \frac{1-a^{MM_X}}{1-a^M}.$$

Combining this with (A-2) gives

$$\sum_{i=1}^{M_X} a^{M(i-1)} \sum_{j=1}^{M} a^{j-1}[(j-1)a^{-1} + M - j]^\rho$$

$$\leq \frac{1-a^{MM_X}}{1-a^M} \frac{1-a^M}{1-a} M^\rho$$

$$\leq \frac{1}{1-a} M^\rho = \frac{M^\rho}{P_X(x)} \tag{35}$$

on using the definition (33) and the fact that $1 - a^{MM_X} \leq 1$.

Q.E.D.

References

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

[2] R. Ash, *Information Theory.* New York: Interscience, 1965.

[3] F. Jelinek, *Probabilistic Information Theory.* New York: McGraw-Hill, 1968.

A New Class of Lower Bounds to Information Rates of Stationary Sources via Conditional Rate-Distortion Functions

ROBERT M. GRAY

***Abstract*—A new class of lower bounds to rate-distortion functions of stationary processes with memory and single-letter vector-valued distortion measures is derived. This class is shown to include or imply all other well-known lower bounds to rates of such sources and distortion measures. The derivation is based on the definition and properties of the conditional rate-distortion function. In addition to providing a unified and intuitive approach to lower bounds, this approach yields several interesting relations among conditional, joint, and marginal rates that are similar to and sometimes identical with the analogous relations among the corresponding entropies.**

I. Introduction

The evaluation of rate-distortion functions is most difficult for precisely that class of sources for which the theory is potentially most useful—sources with memory. This has led during the last few years to the development and study of several lower bounds to rate-distortion functions of various classes of processes with memory. The best

Manuscript received May 2, 1972; revised September 6, 1972. This work was supported in part by the National Science Foundation under Grants GK-5452 and GK-31630 and by the Joint Service Program at Stanford Electronics Laboratory under U.S. Navy Contract N00014-67-A-0112-004.

The author is with the Department of Electrical Engineering, Stanford University, Stanford, Calif. 94305.