Source Coding with Side Information and a Converse for Degraded Broadcast Channels

RUDOLF F. AHLSWEDE AND JÁNOS KÖRNER

Abstract—Let $\{(X_i,Y_i)\}_{i=1}^{\infty}$ be a memoryless correlated source with finite alphabets, and let us imagine that one person, encoder 1, observes only $X^n = X_1,\cdots,X_n$ and another person, encoder 2, observes only $Y^n = Y_1,\cdots,Y_n$. The encoders can produce encoding functions $f_n(X^n)$ and $g_n(Y^n)$, respectively, which are made available to the decoder. We determine the rate region in case the decoder is interested only in knowing $Y^n = Y_1,\cdots,Y_n$ (with small error probability). In Section II of the paper we give a characterization of the capacity region for degraded broadcast channels (DBC's), which was conjectured by Bergmans [11] and is somewhat sharper than the one obtained by Gallager [12].

I. Introduction

THIS PAPER consists of two parts. The first part deals with two problems concerning correlated information sources, and the second part deals with the degraded broadcast channel (DBC).

The first problem we shall consider is the following: two correlated discrete memoryless information sources (DMS's) emit the random variables (RV's) (X_i,Y_i) at a time instant i. A decoder has the task to provide us with a λ-code of the first n outputs of the source $\{Y_i\}$. This decoder is allowed to observe a suitable code of $X^n = X_1,X_2,\cdots,X_n$, however, the rate of this code is limited by some constant $c > 0$. We ask for the minimum rate of that additional code of $Y^n = Y_1,Y_2,\cdots,Y_n$ he has to know in order to provide us with a code of Y^n having prescribed error probability λ. The encoders of X^n and Y^n can only observe the sources they have to encode. By proving a coding theorem and a weak converse result, we shall determine the region of achievable rate pairs. Some related questions are also treated.

As far as we know, the first problem about encoding correlated sources was that of finding the "common information" contained in them. Common information was first meant as some common part of the total amount of information contained separately in each of two correlated sources $\{X_i\}$ and $\{Y_i\}$ and which can, therefore, be encoded by any of them independently from the knowledge of the actual outcomes of the other source. The problem of finding such a code was stated independently by I. Csiszár and D. Slepian. P. Gács and J. Körner [2] showed that no "common code" of two correlated sources exploiting the correlation can be constructed, even if the probability of coincidence of the two codes is an arbitrary $\varepsilon > 0$ rather than one. In other words, common codes can only use deterministic interdependence of the sources. This result was later sharpened under somewhat stronger assumptions by H. S. Witsenhausen [3]. The present authors came to the problem stated above in discussions with G. Tusnády and P. Gács about source coding and, in particular, the results in [2] during the 1972 European Meeting of Statisticians in Budapest. Meanwhile, an important contribution to coding of correlated sources had been made by D. Slepian and J. K. Wolf [4]. In [4] the outputs of two correlated sources are encoded independently, and a decoder that has to reproduce the outputs of both sources with small error probability has both coded messages available. Slepian and Wolf determined the region of achievable rates. Our present problem has been studied by A. D. Wyner and J. Ziv [9] and by A. D. Wyner [10] who proved a (weak) converse result in the special case of two symmetrically correlated binary RV's. No direct result establishing, for an arbitrary $c > 0$, the region of achievable rates has been proved.

In the second part of the present paper we shall deal with the DBC with two components. Broadcast channels were first considered by T. M. Cover [8]. Cover's first paper created immediate interest, and P. Bergmans in [11] described a coding scheme for the DBC, which he conjectured to be optimal. The corresponding (weak) converse result was obtained for the special case of binary symmetric broadcast channels by Wyner and Ziv [9] and by Wyner [10]. Gallager [12] proved a coding theorem and weak converse for arbitrary DBC's. The characterization obtained by him is somewhat weaker than the one proposed by Bergmans [11], and we prove the latter to be true also. This answers a question left open in [12] and, in the terminology used there, amounts to proving the concavity $\{\cap\}$ of the function C_2 (C_1) (see [12, formula (7), p. 5]).[1]

II. Source Coding Problem with Side Information

A. Statement of Problem and Auxiliary Results

A discrete memoryless correlated source is a sequence $\{(X_i,Y_i)\}_{i=1}^{\infty}$ of independent and identically distributed pairs of discrete random variables (DRV's). At time instant i, the source emits (X_i,Y_i).

Let us imagine that one person, encoder 1, observes only X^n and another person, encoder 2, observes only Y^n. The

Manuscript received April 6, 1974; revised March 25, 1975. This work was supported by the National Science Foundation under Grant 40492.

R. F. Ahlswede was with the Department of Mathematics, Ohio State University, Columbus, Ohio. He is now with the University of Göttingen, Göttingen, Germany.

J. Körner is with the Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary.

[1] Meanwhile, the strong converse has been proved. The result is contained in the paper "Bounds on conditional probabilities with applications in multi-user communication," by R. Ahlswede, P. Gács, and J. Körner (submitted to *Z. Wahrscheinlichkeitstheorie u. verw. Geb.*).

encoders can produce encoding functions $f_n(X^n)$ and $g_n(Y^n)$, respectively. A third person, the decoder, has $f_n(X^n)$ and $g_n(Y^n)$ available, however, he is only interested in knowing $Y^n = Y_n, \cdots, Y_n$ with small error probability λ. Using his knowledge of $f_n(X^n)$ and $g_n(Y^n)$, he shall construct a decoding function $V_n(f_n(X^n), g_n(Y^n))$ such that

$$\Pr\{V_n(f_n(X^n), g_n(Y^n)) = Y^n\} \geq 1 - \lambda. \tag{1.1}$$

A pair of nonnegative real numbers (R_1, R_2) is called an *achievable* pair of rates if, for any $\delta > 0$, $0 < \lambda \leq 1$, and sufficiently large n, there exist encoding functions f_n of X^n and g_n of Y^n and a decoding function $V_n = V_n(f_n(X^n), g_n(Y^n))$ such that

$$\Pr(V_n(f_n(X^n), g_n(Y^n)) = Y^n) \geq 1 - \lambda \tag{1.2}$$

and

$$\begin{aligned} \|f_n(X^n)\| &\leq \exp\{(R_1 + \delta)n\} \\ \|g_n(Y^n)\| &\leq \exp\{(R_2 + \delta)n\}. \end{aligned} \tag{1.3}$$

(Here we have used the notation $\|Z\|$ for the cardinality of the range of a function Z). We denote the region of all achievable pairs of rates by $\mathscr{R}$. Our main goal was to characterize $\mathscr{R}$, and the answer is given in Theorem 2 in Section II-D.

In what follows we prove some convexity properties of the functions we shall deal with in the sequel. All the RV's in this paper are supposed to take finitely many different values. The terminology is that of [13]. $H(Z)$ is the entropy of the RV Z, $H(X \mid Y)$ stands for the average conditional entropy of X given Y, etc.

Lemma 1: a) Let X,Y be a pair of DRV's with joint distribution $\Pr(X = x, Y = y) = Q(x,y)$. If U is any DRV such that U,X,Y form a Markov chain, then the function

$$T(c) = \inf_{H(X|U) \geq c} H(Y \mid U)$$

is convex $(\cup)$ in c.

b) Let U,X,Y be arbitrary DRV's such that U,X,Y form a Markov chain in this order and such that $\Pr(Y = y \mid X = x) = Q(y \mid x)$, for a given stochastic matrix $\{Q(y \mid x) \mid x \in \mathscr{X}, y \in \mathscr{Y}\}$, then the function

$$G(c) = \inf_{H(X|U) \geq c} H(Y \mid U)$$

is convex $(\cup)$ in c.

Proof: The function $G(c)$ as introduced here is equal to the gerbator of a noisy channel with transmission matrix Q (see [5]). Part b) is, therefore, equivalent to [5, theorem 1] but is not needed in the present paper and is only stated for comparison. In order to prove part a), we first observe that "conditioning" means to take convex linear combinations. Indeed let (U_i,X_i,Y_i), $i = 1,2$, be two triples of RV's satisfying our assumptions. For any α, $0 \leq \alpha \leq 1$, we introduce a new triple of RV's $(\tilde{U},\tilde{X},\tilde{Y})$ such that, with probability α, $(\tilde{U},\tilde{X},\tilde{Y})$ equals (U_1,X_1,Y_1) and, with probability $(1 - \alpha)$, it equals (U_2,X_2,Y_2). Furthermore, let I be another RV ranging over the set $\{1,2\}$ and yielding $\Pr\{I = 1\} = \alpha$. In this notation, we have the equations

$$H(\tilde{X} \mid \tilde{U},I) = \alpha H(X_1 \mid U_1) + (1 - \alpha)H(X_2 \mid U_2) \tag{1.4}$$

and

$$H(\tilde{Y} \mid \tilde{U},I) = \alpha H(Y_1 \mid U_1) + (1 - \alpha)H(Y_2 \mid U_2). \tag{1.5}$$

Since $\{(I,\tilde{U}),\tilde{X},\tilde{Y}\}$ is again a Markov chain, the result follows from (1.4), (1.5), and the definition of $T(\cdot)$.

Lemma 2: Let $\{(X_i,Y_i)\}_{i=1}^{\infty}$ be a discrete memoryless correlated source and let U be any DRV such that U,X^n,Y^n form a Markov chain. Then, for every $n \in N$ and every $c \geq 0$,

a) $$T_n(c) \triangleq \inf_{(1/n)H(X^n|U) \geq c} \frac{1}{n} H(Y^n \mid U) = T(c)$$

b) $$J_n(c) \triangleq \sup_{(1/n)H(Y^n|U) \leq c} \frac{1}{n} H(X^n \mid U) = J_1(c).$$

Furthermore, if X and Y are not independent, one has

$$J(c) = T^{-1}(c).$$

Proof: We first show part a) by arguments similar to the ones used in [5] and then we deduce b) from a). We can write (see [5])

$$\begin{aligned} H(Y^n \mid U) &= \sum_{i=1}^{n} H(Y_i \mid UY_1,Y_2,\cdots,Y_{i-1}) \\ &\geq \sum_{i=1}^{n} H(Y_i \mid UY_1,\cdots,Y_{i-1},X_1,\cdots,X_{i-1}) \\ &= \sum_{i=1}^{n} H(Y_i \mid UX_1,\cdots,X_{i-1}). \end{aligned} \tag{i.6}$$

By the definition of $T(c)$, we have

$$\begin{aligned} &H(Y_i \mid U, X_1 = x_1,\cdots,X_{i-1} = x_{i-1}) \\ &\qquad \geq T(H(X_i \mid U, X_1 = x_1,\cdots,X_{i-1} = x_{i-1})) \end{aligned}$$

and hence by the convexity of $T(c)$ we get for the expected values $H(Y_i \mid UX_1,\cdots,X_{i-1}) \geq T(H(X_i \mid UX_1,\cdots,X_{i-1}))$. This and (1.6) yield

$$H(Y^n \mid U) \geq \sum_{i=1}^{n} T(H(X_i \mid UX_1,\cdots,X_{i-1})). \tag{1.7}$$

It follows from the convexity of T and from (1.7) that

$$\begin{aligned} H(Y^n \mid U) &\geq nT\left(\frac{1}{n}\sum_{i=1}^{n} H(X_i \mid UX_1,\cdots,X_{i-1})\right) \\ &= nT\left(\frac{1}{n} H(X^n \mid U)\right). \end{aligned}$$

Statement a) follows now, because T is monotonically increasing. Statement b) is now easily derived from a). If X and Y are not independent, T is not constant and as a convex monotonically increasing function is certainly strictly increasing. Since $T_n(c) = T(c)$, $T_n(c)$ is also strictly increasing and in this case we have that J_n is the inverse function of T_n for $n \in N$. Hence $J_n \equiv J_1 \equiv T^{-1}$. Part b)

is not needed in the present paper and is only included for the sake of completeness.

Lemma 3: Let $\mathcal{P}_n$ be the set of all probability n-vectors $\boldsymbol{p} = (p_1, \cdots, p_n)$ and let $f_j(\boldsymbol{p})$, $j = 1, \cdots, k$, be continuous functions on $\mathcal{P}_n$. Then, to any probability measure μ on (the Borel subsets of) $\mathcal{P}_n$ there exist $(k+1)$ elements $\boldsymbol{p}_i$ of $\mathcal{P}_n$ and constants $\alpha_i \geq 0, i = 1, \cdots, k+1$ with $\sum_{i=1}^{k+1} \alpha_i = 1$ such that

$$\int f_j(\boldsymbol{p})\, d\mu = \sum_{i=1}^{k+1} \alpha_i f_j(\boldsymbol{p}_i), \qquad j = 1, \cdots, k.$$

Proof: $F = \{(f_1(\boldsymbol{p}), \cdots, f_k(\boldsymbol{p})) \mid \boldsymbol{p} \in \mathcal{P}_n\}$ is an image of a compact set under a continuous function and, therefore, a compact subset of E^k. Since $(\int f_1(\boldsymbol{p})\, d\mu, \cdots, \int f_k(\boldsymbol{p})\, d\mu)$ belongs to the convex closure of F, by Carathéodory's theorem it can be represented as the convex combination of at most $k+1$ extremal points of that convex closure. The latter clearly belong to F and the proof is complete.

Finally, we shall need the following important result of D. Slepian and J. K. Wolf [4].

Theorem (Slepian-Wolf): Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a discrete memoryless correlated source. For every sufficiently large n and any λ, $0 < \lambda \leq 1$, $\delta > 0$, there exist encoding functions $f_n(X^n)$ and $g_n(Y^n)$ such that one can construct a decoding function $V_n(f_n(X^n), g_n(Y^n))$ yielding

$$\Pr\{V_n(f_n(X^n), g_n(Y^n)) = (X^n, Y^n)\} \geq 1 - \lambda \quad (1.8)$$

and

$$\|f_n(X^n)\| \leq \exp\{n(H(X) + \delta)\}$$

$$\|g_n(Y^n)\| \leq \exp\{n(H(Y|X) + \delta)\}.$$

In the next section we shall include the fairly simple proof of this theorem which was presented in [6]. Knowledge of this proof is not necessary for an understanding of the later sections.

B. Simple Proof of Coding Theorem of Slepian-Wolf

We shall make use of a well-known theorem originally due to A. Feinstein [7]. We state it in the special form that will serve us in the sequel.

Feinstein's Fundamental Theorem: Let (X^n, Y^n) be the first n outcomes of a correlated DMS with joint distribution $\{Q(x,y)\}$. A sequence of pairs $\{\mathcal{B}_i, y_i\}$, $1 \leq i \leq M$, is an ε-code, if $\mathcal{B}_i \subseteq \mathcal{X}^n$, $\mathcal{B}_i \cap \mathcal{B}_j = \phi$, for $i \neq j$, $y_i \in \mathcal{Y}^n$, and

$$\Pr(X^n \in \mathcal{B}_i \mid Y^n = y_i) \geq 1 - \varepsilon.$$

If we fix a set $\mathcal{Y}_0^n \subseteq \mathcal{Y}^n$ and select all our codewords y_i from $\mathcal{Y}_0^n$ and, furthermore, if $\{\mathcal{B}_i, y_i\}$, $1 \leq i \leq M$, is a maximal ε-code one can select from $\mathcal{Y}_0^n$, then

$$M \geq \exp\{n[I(X \wedge Y) - \delta]\} \cdot [\varepsilon \cdot Q(\mathcal{Y}_0^n) - Q(A_n(\delta))]$$

where $Q(\mathcal{Y}_0^n) = Q(\mathcal{X}^n, \mathcal{Y}_0^n)$,

$$A_n(\delta) = \left\{(x^n, y^n) \mid \left|\frac{1}{n} \log \frac{Q(x^n, y^n)}{Q(x^n) \cdot Q(y^n)} - I(X \wedge Y)\right| > \delta\right\}$$

and $I(X \wedge Y)$ is the average mutual information of the RV's X and Y.

Since, for any fixed $\delta > 0$, $Q(A_n(\delta)) \to 0$ as $n \to \infty$, we have the following

$$M \geq \exp\{n[I(X \wedge Y) - \delta]\} \cdot \frac{\varepsilon}{2} \cdot Q(\mathcal{Y}_0^n).$$

We shall prove that, for every $\varepsilon > 0$ and sufficiently large n, one can keep constructing ε-codes with disjoint sets of codewords containing more than $\varepsilon^2/4 \cdot \exp\{n[I(X \wedge Y) - \delta]\}$ elements each. The proof consists of an iterative application of Feinstein's theorem. Let us choose

$$\mathcal{Y}_\delta^n(1) \triangleq \left\{y^n \mid y^n \in \mathcal{Y}^n, \left|\frac{1}{n} \log Q(y^n) + H(Y)\right| < \delta\right\}.$$

By the weak law of large numbers, $Q^n(\mathcal{Y}_\delta^n(1)) \to 1$ and, therefore, $Q^n(\mathcal{Y}_\delta^n(1)) > 1 - \varepsilon$, for sufficiently large n. Hence, in a first step, we can construct an ε-code with

$$M(1) = \frac{\varepsilon}{2}(1 - \varepsilon) \exp(n(I(X \wedge Y) - \delta))$$

codewords. Since the set $C^n(1)$ consisting of all these codewords has probability less than

$$\frac{\varepsilon}{2}(1 - \varepsilon) \exp[n(I(X \wedge Y) - \delta)] \cdot \exp - [n(H(Y) - \delta)]$$

$$= \frac{\varepsilon}{2}(1 - \varepsilon) \exp - [nH(Y \mid X)]$$

and at least

$$\frac{\varepsilon}{2}(1 - \varepsilon) \exp n[I(X \wedge Y) - \delta] \cdot \exp - [n(H(Y) + \delta)]$$

$$= \frac{\varepsilon}{2}(1 - \varepsilon) \cdot \exp - [n(H(Y \mid X) + 2\delta)],$$

putting $\mathcal{Y}_\delta^n(2) \triangleq \mathcal{Y}_\delta^n(1) - C^n(1)$ we get

$$(1 - \varepsilon)\left(1 - \frac{\varepsilon}{2} \exp[-nH(Y \mid X)]\right)$$

$$\leq Q^n(\mathcal{Y}_\delta^n(2))$$

$$\leq (1 - \varepsilon)\left(1 - \frac{\varepsilon}{2} \exp[-n(H(Y \mid X) + 2\delta)]\right). \quad (2.1)$$

Now we apply Feinstein's theorem to the set $\mathcal{Y}_\delta^n(2)$, and we get a code $C^n(2)$ with

$$M(2) = \exp\{n[I(X \wedge Y) - \delta]\}$$

$$\cdot \frac{\varepsilon}{2}(1 - \varepsilon)\left[\left(1 - \frac{\varepsilon}{2} \exp(-nH(Y \mid X))\right)\right]$$

codewords.

After having constructed the disjoint sets of codewords $C^n(1), C^n(2), \cdots, C^n(L-1)$, we obtain a set $\mathcal{Y}_\delta^n(L) \triangleq \mathcal{Y}_\delta^n(L-1) - C^n(L-1)$. For the probability of this set

one easily obtains the bounds

$$(1-\varepsilon)\left(1-\frac{\varepsilon}{2}\exp\left(-nH(Y\mid X)\right)\right)^{L-1} \le Q^n(\mathscr{Y}_\delta^n(L)) \le (1-\varepsilon)\left(1-\frac{\varepsilon}{2}\exp\left(-n[H(Y\mid X)+2\delta]\right)\right)^{L-1}. \quad (2.2)$$

Using the elementary inequality $(1-\mu x)^L \le \exp(-\mu L x)$, for $0 \le x \le 1$, one can write the right side of (2.2) as the following

$$Q^n(\mathscr{Y}_\delta^n(L)) \le (1-\varepsilon)\exp\left[-\frac{\varepsilon}{2}2^{-n(H(Y\mid X)+2\delta)}\right]^{L-1}$$

and hence choosing $L = [\exp(n(H(Y\mid X)+\delta_1))]$ with any $\delta_1 > 2\delta$ we get $Q^n(\mathscr{Y}_\delta^n(2^{n(H(Y\mid X)+\delta_1)})) \to 0$.

On the other hand, recognizing that, for $0 \le x$, $\lim_{x\to 0}(1-\mu x)^x = \exp(-(1/\mu))$, it becomes clear from the left side of (2.2) that, for $L = [2^{nH(Y\mid X)}]$, the probability of $\mathscr{Y}_\delta^n(\exp(nH(Y\mid X)))$ is bounded away from zero independently of n. Summarizing what we have obtained so far, we know that, for any $\varepsilon > 0$, $\delta > 0$, and every sufficiently large n, one can construct $[\exp[n(H(Y\mid X)+\delta_n)]]$ disjoint ε-codes with $0 < \delta_n < 3\delta$. Each of these codes contains at least

$$\frac{\varepsilon^2}{2}\cdot(1-\varepsilon)\exp(n(I(X\wedge Y)-\delta)) > \frac{\varepsilon^2}{4}\exp(n(I(X\wedge Y)-\delta))$$

codewords, and the union of all the codeword sets has probability greater than $1-\varepsilon$.

Now we are ready to prove the Slepian–Wolf theorem. For the given λ and δ we choose

$$\varepsilon < \frac{\lambda-\delta}{2} \quad (2.3)$$

and construct the $[\exp[n(H(Y\mid X)+\delta_n)]]$ disjoint sets of codewords described above. Let f_n be a function defined on $\mathscr{X}^n$ which is invertible on

$$\mathscr{X}_\delta^n(1) \triangleq \left\{x^n \mid x^n \in \mathscr{X}^n, \frac{1}{n}\left|\log Q^n(x^n) + H(X)\right| < \delta\right\}$$

and takes any constant value on $\mathscr{X}^n - \mathscr{X}_\delta^n(1)$. Thus $f_n(X^n)$ is a code of error probability less than δ of X^n, for sufficiently large n. We define a code g_n of Y^n as follows:

$$g_n(Y^n) \triangleq \begin{cases} J, & \text{if } Y^n = y^n \text{ and } y^n \in C^n(J), \quad \text{for } 1 \le J \le L \\ 0, & \text{if } Y^n = y^n \text{ and } y^n \text{ does not belong to any of the codes we have constructed.} \end{cases}$$

Since we have constructed $[\exp(nH(Y\mid X)+\delta_n)]$ codes, this means that

$$\frac{1}{n}\log\|g_n(Y^n)\| \le H(Y\mid X)+3\delta = H(Y\mid X)+\delta_1$$

where we write $\delta_1 = 3\delta$. On the other hand, it is obvious that

$$\frac{1}{n}\log\|f_n(X^n)\| \le H(X)+\delta < H(X)+\delta_1.$$

The function V_n, by means of which we shall obtain a DRV, $\tilde{Y}^n = V_n(f_n(X^n), g_n(Y^n))$ with $\Pr(\tilde{Y}^n = Y^n) \ge 1-\lambda$, is now the following: suppose that the random output of the correlated source was (x^n, y^n). The decoder has available the corresponding values of the functions f_n and g_n, i.e., $f_n(x^n)$ and $g_n(y^n)$. With a probability greater than $1-\delta$, he can decode x^n correctly from $f_n(x^n)$. With a probability at least $1-\varepsilon$, y^n is a codeword in one of the $[\exp n(H(Y\mid X)+\delta_n)]$ codes we have constructed. Thus, with a probability greater than $1-\varepsilon-\delta$, the decoder is in the position to do the following: he looks for that particular decoding set of the $g^n(y^n)$th code which $x^n \in \mathscr{X}^n$ belongs to, and doing so, with probability greater than $1-\varepsilon$, he will decide on the right y^n since we have constructed ε-codes. Thus, all together with a probability greater than $1-2\varepsilon-\delta$, our decoder will know correctly which y^n was the actual value of the DRV Y^n. Proceeding as described, the decoder has constructed an RV $\tilde{Y}^n$ such that

$$\Pr(\tilde{Y}^n = Y^n) \ge 1-2\varepsilon-\delta > 1-\lambda.$$

The last inequality follows from (2.3) and the theorem is proved.

C. Characterization of Rate Region

Theorem 1: Let $\{(X_i,Y_i)\}_{i=1}^\infty$ be a discrete memoryless correlated source. The *rate region* $\mathscr{R}$ of the source coding problem with side information (as described in Section II-A) *equals*

$$\mathscr{R}^* = \left\{\left(\frac{1}{n}H(f_n(X^n)), \frac{1}{n}H(Y^n\mid f_n(X^n)) \mid n \in N, f_n: \mathscr{X}^n \to N\right\}.$$

Proof: The relation $\mathscr{R} \supset \mathscr{R}^*$ is immediately obtained by applying the Slepian–Wolf theorem of Section II-A to the supersource $\{(U_t,\tilde{Y}_t)\}_{t\in N}$, where the $(U_t,\tilde{Y}_t)$, $t \in N$, are independent identically distributed RV's with the same distribution as the pair $((f_n(X^n),Y^n))$.

Suppose now that $(f_n(X^n), g_n(Y^n))$ is a pair of encoding functions and that $V_n = V_n(f_n(X^n),g_n(Y^n))$ is a decoding function such that

$$\Pr(V_n(f_n(X^n),g_n(Y^n)) = Y^n) = 1-\lambda \quad (3.1)$$

and

$$\|f_n(X^n)\| \le \exp\{R_1 n\}, \quad \|g_n(Y^n)\| \le \exp\{R_2 n\}. \quad (3.2)$$

We shall prove below, with the use of Fano's lemma, that then

$$R_2 \ge \frac{1}{n}[H(Y^n\mid f_n(X^n)) - h(\lambda) - n\lambda\log\|Y\|] \quad (3.3)$$

and hence $\mathscr{R} \subset \mathscr{R}^*$ follows by choosing λ arbitrary small and n sufficiently large. Let us denote $f_n(X^n)$ by U. Clearly, for *any value* u of U, we have

$$R_2 \geq \frac{1}{n} \log \|g_n(Y^n)\| \geq \frac{1}{n} H(g_n(Y^n) \mid U = u)$$

$$\geq \frac{1}{n} H(V_n(f_n(X^n), g_n(Y^n)) \mid U = u)$$

$$\geq \frac{1}{n} I(Y^n \wedge V_n \mid U = u)$$

$$= \frac{1}{n}[H(Y^n \mid U = u) - H(Y^n \mid U = u, V_n)]. \quad (3.4)$$

Using Fano's Lemma we conclude that

$$H(Y^n \mid U = u, V_n) \leq h(\lambda(u)) + \lambda(u) \cdot \log \|Y\| \cdot n \quad (3.5)$$

where $\lambda(u)$ is the conditional probability

$$\Pr\{V_n \neq Y_n \mid U = u\}.$$

The last two inequalities yield

$$R_2 \geq \frac{1}{n}[H(Y^n \mid U = u) - h(\lambda(u)) - \lambda(u) \cdot n \cdot \log \|Y\|]. \quad (3.6)$$

Taking the expected value on the right side of our inequality and observing that the concavity of the entropy function implies $E(h(\lambda(U))) \leq h(\lambda)$, we finally obtain inequality (3.3) which we wanted to prove.

D. Single Letter Characterization of Rate Region

Now we deduce from Theorem 1 a simple characterization of the rate region $\mathscr{R}$, which actually could be used for its numerical determination.

Theorem 2 (Coding Theorem and Converse): The rate region $\mathscr{R}$ for the correlated source coding problem as described in Section II-A equals

$$\mathscr{R}^{**} = \{(R_1, R_2) \mid R_1 \geq I(X \wedge U), R_2 \geq H(Y \mid U), U \text{ DRV and } U, X, Y \text{ Markov chain}\}.$$

Furthermore, the region is already obtained if we limit the cardinality of the range of U by the constraint

$$\|U\| \leq \|X\| + 2.$$

Proof: We use Theorem 1 and show first that $\mathscr{R}^* \subset \mathscr{R}^{**}$. Suppose that $R_1 = (1/n)H(Z)$ and that $R_2 = (1/n)H(Y^n \mid Z)$, where $Z = f_n(X^n)$. We have to show that there exists a Markov chain (U, X, Y) such that $R_1 \geq I(X \wedge U)$ and $R_2 \geq H(Y \mid U)$. Clearly,

$$n \cdot R_1 = H(Z) \geq I(Z \wedge X^n) = H(X^n) - H(X^n \mid Z). \quad (4.1)$$

Using the independence of the X_i, one can write

$$\frac{1}{n} H(X^n \mid Z) \geq H(X) - R_1. \quad (4.2)$$

Since T is increasing, we also have

$$T\left(\frac{1}{n} H(X^n \mid Z)\right) \geq T(H(X) - R_1). \quad (4.3)$$

On the other hand, Lemma 2 yields $R_2 = (1/n)H(Y^n \mid Z) \geq T[(1/n)H(X^n \mid Z)]$, and the two last inequalities give

$$R_2 \geq T(H(X) - R_1) = \inf_{H(X \mid U) \geq H(X) - R_1} H(Y \mid U) \quad (4.4)$$

where the equality holds by definition of T. Rewriting the constraint equality in the form $I(X \wedge U) \geq R_1$, we finally get

$$R_2 \geq \inf_{I(X \wedge U) \geq R_1} H(Y \mid U) \quad (4.5)$$

as was to be shown.

Now we prove that $\mathscr{R}^* \supset \mathscr{R}^{**}$. We have to show that, for every Markov chain $U \to X \to Y$ and for every $\varepsilon > 0$, there exists an n and a function f_n on $\mathscr{X}^n$ such that

$$-\frac{1}{n} H(f_n(X^n)) \leq I(X \wedge U) + \varepsilon \quad (4.6)$$

$$\frac{1}{n} H(Y^n \mid f_n(X^n)) \leq H(Y \mid U) + \varepsilon. \quad (4.7)$$

Clearly we can assume that, for some $\delta > 0$,

$$H(U) - I(X \wedge U) > \delta \quad (4.8)$$

because otherwise $I(X \wedge U) = H(U)$ and thus, U being a deterministic function of X, the choice $f_n(X^n) = U^n$ would do. Suppose further that $\varepsilon < \delta$.

We now make use of the notions and simple properties of typical sequences and of generated sequences, both defined with const. $\cdot \sqrt{n}$ deviation (see [4, ch. 3]). Denote by $\mathscr{T}(\mathscr{U}^n)$ the set of those typical sequences in $\mathscr{U}^n$, by $\mathscr{G}(\mathscr{X}^n, u^n)$ [or $\mathscr{G}(\mathscr{Y}^n, x^n, u^n)$] the elements of $\mathscr{X}^n$ [or $\mathscr{Y}^n$] generated by $u^n \in \mathscr{U}^n$ [or $(x^n, u^n) \in \mathscr{X}^n \times \mathscr{U}^n$]. Furthermore, define

$$\mathscr{G}^*(\mathscr{Y}^n, u^n) = \cup \mathscr{G}(\mathscr{Y}^n, x^n, u^n), \qquad x^n \in \mathscr{G}(\mathscr{X}^n, u^n)$$

and choose the constant before the square root such that for $u^n \in \mathscr{T}(\mathscr{U}^n)$, $x^n \in \mathscr{G}(\mathscr{X}^n, u^n)$,

$$\lim_{n \to \infty} \Pr(Y^n \in \mathscr{G}^*(\mathscr{Y}^n, u^n) \mid X^n = x^n) = 1 \quad (4.9)$$

and

$$\left| \frac{1}{n} \log \|\mathscr{G}^*(\mathscr{Y}^n, u^n)\| = H(Y/U) \right| \leq \frac{c_1}{\sqrt{n}} \quad (4.10)$$

for a suitable constant c_1.

The *maximal code construction* (see [14, ch. 3]) leads, for any λ, $0 < \lambda < 1$, to a code $\{(v_j, D_j)\}_{j=1}^N$ with the following properties:

a) $v_j \in \mathscr{T}(\mathscr{U}^n)$, $j = 1, 2, \cdots, N$;
b) $D_j \subset \mathscr{G}(\mathscr{X}^n, v_j)$, $D_j \cap D_j, = \phi$, $j \neq j$, $j = 1, \cdots, N$;
c) $\Pr(X^n \in D_j \mid U^n = v_j) \geq 1 - \lambda$, $j = 1, \cdots, N$;
d) $\Pr(X^n \in D_0) \leq (1 - \lambda) \Pr(U^n \notin \{v_1, \cdots, v_N\}) + \lambda \cdot \Pr(U^n \in \{v_1, \cdots, v_N\})$;

where $D_0 = \mathscr{X}^n - \bigcup_{j=1}^N D_j$ (maximality of the code).

Wolfowitz's strong converse of the coding theorem (see [14, ch. 3]) implies, for large n, that

e) $1/n \log (N + 1) \leq I(X \wedge U) + \varepsilon$.

Define

$$f_n(X^n) = j, \qquad \text{for } X^n \in D_j, j = 0,1,\cdots,N. \quad (4.11)$$

This function satisfies (4.6), because $(1/n)H(f_n(X^n)) \leq (1/n) \log (N + 1)$. Now we verify (4.7). For $i = 1,2,\cdots,N$, we use the estimate

$$H(Y^n \mid f_n(X^n) = i) \leq \log \|\mathcal{G}^*(\mathcal{Y}_1^n v_i)\| + \delta_i^n \log \frac{1}{\delta^n} + n \cdot \delta_i^n \cdot \log \|\mathcal{Y}\| \quad (4.12)$$

where

$$\delta_i^n = \Pr (Y^n \notin \mathcal{G}^*(\mathcal{Y}^n, v_i) \mid f_n(X^n) = i)$$

and $\lim_{n\to\infty} \delta_i^n = 0$ by (4.9). Using this and (4.10), we get for $i = 1,2,\cdots,N$ and n large

$$\frac{1}{n} H(Y^n \mid f_n(X^n) = i) \leq H(Y \mid U) + \frac{\varepsilon}{2} \quad (4.13)$$

and, therefore, also

$$\frac{1}{n} H(Y^n \mid f_n(X^n)) \leq H(Y \mid U) + \frac{\varepsilon}{2} + \log \|\mathcal{Y}\| \cdot \Pr (f_n(X^n) = 0). \quad (4.14)$$

It follows from a), e), the choice $\varepsilon < \delta$, and (4.8) that $\Pr (U^n \in \{v_1,\cdots,v_N\}) \to 0$ when $n \to \infty$. Part d) and the fact that we can choose λ arbitrarily close to one, imply that $\Pr (X^n \in D_0)$ can be made arbitrary small for n sufficiently large. This combined with c) and (4.14) gives the result.

Finally, we show that the cardinality of the range of U can be bounded by $\|X\| + 2$. We apply Lemma 3 to the present situation by setting $\|\mathcal{X}\| = n$, $\mathcal{X} = \{1,2,\cdots,n\}$, and by choosing $\mathcal{P}_n$ as the set of all probability distributions on $\mathcal{X}$. Suppose that $\Pr (X = x \mid U = u)$ and $\Pr (U = u)$ are given and that, for every $x \in \mathcal{X}$, $\sum_{u\in U} \Pr (X = x \mid U = u) \cdot \Pr (U = u) = \Pr (X = x)$, the given marginal distribution. We can interpret $\{\Pr (X = x \mid U = u)\}_{x\in\mathcal{X}}$ as an element of $\mathcal{P}_n$ and $\{\Pr (U = u)\}_{u\in U}$ as a Borel measure on $\mathcal{P}_n$. Consider the following continuous functions on $\mathcal{P}_n$:

a) For $\boldsymbol{p} = (p(1),\cdots,p(n)) \in \mathcal{P}_n$, set

$$f_j(\boldsymbol{p}) = p(j), j = 1, 2,\cdots, n;$$

b) $$f_{n+1}(\boldsymbol{p}) = -\sum_{y\in\mathcal{Y}} \left(\sum_{x\in\mathcal{X}} P(y \mid x)p(x)\right) \cdot \log \sum_{x\in\mathcal{X}} P(y \mid x)p(x);$$

c) $$f_{n+2}(\boldsymbol{p}) = \sum_{x\in\mathcal{X}} p(x) \log p(x) + H(X).$$

Clearly, for $j = 1,\cdots, n - 1$,

$$\sum_{u\in U} f_j(\Pr (\cdot \mid U = u)) \Pr (U = u) = p(j)$$

$$\sum_u f_{n+1}(\Pr (\cdot \mid U = u)) \Pr (U = u) = H(Y \mid U)$$

and

$$\sum_u f_{n+2}(\Pr (\cdot \mid U = u)) \Pr (U = u) = I(U \wedge X).$$

Lemma 3 implies that there exists a U^* with $\|U^*\| \leq \|X\| + 2$ such that $H(Y \mid U^*) = H(Y \mid U)$ and $I(Y \wedge U) = H(Y \wedge U^*)$.

E. Remarks

A closer look at the proof of the converse part of our coding theorem shows that we never used the condition $Z = f_n(X^n)$. What we used instead was the conditions that Z, X^n, Y^n form a Markov chain in this order. In other words, this means that if we consider Z as a random code of X^n where, for every fixed value of X^n, the randomization in the encoding is independent of Y^n (this is what the markovity means) we can state that *randomization in the encoding of X^n does not help*. There are no more rates achievable by a randomized encoding of X^n than the ones achievable by deterministic codes. Since deterministic codes are a special case of random codes, it is clear that the region of achievable codes remains the same if the deterministic codes of X^n can be exchanged for random ones.

Finally, we would like to outline the answer to a related source coding problem raised by G. Tusnády generalizing an earlier problem of J. Körner [1]. This is again about source coding with side information. The only difference from our original problem is that, while constructing the encoding function for Y^n, the encoder of Y^n has available the codeword $f_n(X^n)$.

The special case of this theorem, where there is no limitation on the rate of $f_n(X^n)$, was settled in [1]. The answer to the general case follows from our Theorems 1 and 2.

It is clear that, for the same memoryless correlated source, the rates achievable by the original coding scheme are *a fortiori* achievable by the present one. Thus only a converse result is needed. Without going into the details, we mention that upon replacing $g_n(Y^n)$ by a new encoding function $\tilde{g}_n(Y^n, f_n(X^n))$ the converse proof of Theorem 1 in Section II-C literally applies and gives the same result.

III. Converse to Coding Theorem for Degraded Broadcast Channel

A. A Few Lemmas

In this Section we shall provide the tools needed for the proof of the Theorem. The stating of the problem is postponed until the next section. We shall need some properties of the function $t(x)$ defined by the following.

Definition 5.1: Given the finite sets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ and the transition probability matrices $\{P_1(y \mid x) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$ and $\{P_2(z \mid y) \mid y \in \mathcal{Y}, z \in \mathcal{Z}\}$, consider all the quadruplets U,X,Y,Z of DRV's such that U,X,Y,Z form a Markov chain with the given transition probabilities

$$\Pr (Y = y \mid X = x) = P_1(y \mid x)$$

$$\Pr (Z = z \mid Y = y) = P_2(z \mid y),$$

$$\text{for all } x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}. \quad (5.1)$$

We define

$$t(x) \triangleq \sup_{I(U \wedge Z) \geq x} I(X \wedge Y \mid U) \tag{5.2}$$

where the supremum is taken over all the Markov chains U,X,Y,Z satisfying (5.1).

Later on we shall see that the range of U can be bounded, and the supremum can, therefore, be replaced by a maximum.

The main tool of our subsequent proof is given by the following lemma.

Lemma 4: For every $x \geq 0$, $t(\cdot)$ is concave ($\cap$).

Proof: We can arbitrarily vary the joint distribution of the pair U,X. Let us fix an $\varepsilon > 0$. For $i = 1,2$ let U_i,X_i,Y_i,Z_i be a quadruple yielding at least $t(x_i) - \varepsilon$. Let us put $x = \alpha x_1 + (1 - \alpha)x_2$, for some $0 \leq \alpha \leq 1$. We introduce the new RV's T and $\tilde{U},\tilde{X},\tilde{Y},\tilde{Z}$ by the following definition:

$$\Pr(T = 1) = \alpha$$

$$\Pr(T = 2) = 1 - \alpha \tag{5.3a}$$

$$\left.\begin{aligned} \tilde{U} &= U_i \\ \tilde{X} &= X_i \\ \tilde{Y} &= Y_i \\ \tilde{Z} &= Z_i \end{aligned}\right\} \quad \text{for } T = i. \tag{5.3b}$$

It is clear that $(T,\tilde{U}),\tilde{X},\tilde{Y},\tilde{Z}$ again form a Markov chain in this order. Furthermore,

$$\begin{aligned} &\alpha t(x_1) + (1 - \alpha)t(x_2) - \varepsilon \\ &\quad \leq \alpha I(X_1 \wedge Y_1 \mid U_1) + (1 - \alpha)I(X_2 \wedge Y_2 \mid U_2) \\ &\quad = \Pr(T = 1) \cdot I(X_1 \wedge Y_1 \mid U_1) \\ &\qquad + \Pr(T = 2)I(X_2 \wedge Y_2 \mid U_2) \\ &\quad = I(\tilde{X} \wedge \tilde{Y} \mid \tilde{U},T) \end{aligned} \tag{5.4}$$

by the definitions (5.3). On the other hand, we have

$$I(U_i \wedge Z_i) \geq x_i, \qquad \text{for } i = 1,2$$

and hence

$$\alpha I(U_1 \wedge Z_1) + (1 - \alpha)I(U_2 \wedge Z_2) \geq \alpha x_1 + (1 - \alpha)x_2 = x. \tag{5.5}$$

However,

$$\begin{aligned} &\alpha I(U_1 \wedge Z_1) + (1 - \alpha)I(U_2 \wedge Z_2) \\ &\quad = \Pr(T = 1)I(U_1 \wedge Z_1) + \Pr(T = 2)I(U_2 \wedge Z_2) \\ &\quad = E(I(U_i \wedge Z_i \mid T = i)) = I(\tilde{U} \wedge \tilde{Z} \mid T). \end{aligned} \tag{5.6}$$

Thus, combining (5.5) and (5.6), we obtain

$$I(\tilde{U} \wedge \tilde{Z} \mid T) \geq x. \tag{5.7}$$

We need the following well-known identity (see, e.g., [13, formula (2.2.29), p. 22]):

$$I((\tilde{U},T) \wedge \tilde{Z}) = I(\tilde{U} \wedge \tilde{Z} \mid T) + I(T \wedge \tilde{Z}).$$

Because the average mutual information, $I(T \wedge \tilde{Z})$, is nonnegative, this identity gives $I((\tilde{U},T) \wedge \tilde{Z}) \geq I(\tilde{U} \wedge \tilde{Z} \mid T)$. Comparing the last inequality with (5.6), we have $I((\tilde{U},T) \wedge \tilde{Z}) \geq x$. This inequality shows that the Markov chain $(T,\tilde{U}),\tilde{X},\tilde{Y},\tilde{Z}$ satisfies our constraint of maximization for x, therefore, we conclude that $t(x) \geq I(\tilde{X} \wedge \tilde{Y} \mid T,\tilde{U})$. Hence, using (5.4), we get $t(x) \geq I(\tilde{X} \wedge \tilde{Y} \mid T,\tilde{U}) = \alpha t(x_1) + (1 - \alpha)t(x_2) - \varepsilon$. Since $t(x) \geq \alpha t(x_1) + (1 - \alpha)t(x_2) - \varepsilon$ holds for every $\varepsilon > 0$, the statement of the lemma is established. Now we prove two short and technical lemmas.

Lemma 5: Let the RV's U,X^n,Y^n,Z^n form a Markov chain. We suppose that $\Pr(Z^n = z^n \mid Y^n = y^n) = \prod_{i=1}^n P_2(z_i \mid y_i)$ and $P_1(Y^n = y^n \mid X^n = x^n) = \prod_{i=1}^n P_1(y_i \mid x_i)$. Define $U_i \triangleq U,Y_1,Y_2,\cdots,Y_{i-1}$, for $i = 2,3,\cdots,n$. This yields

$$I(U \wedge Z^n) \leq \sum_{i=1}^n I(U_i \wedge Z_i).$$

Proof:

$$I(U \wedge Z^n) = H(Z^n) - H(Z^n \mid U)$$

$$\begin{aligned} H(Z^n \mid U) &= \sum_{i=1}^n H(Z_i \mid UZ_1,Z_2,\cdots,Z_{i-1}) \\ &\geq \sum_{i=1}^n H(Z_i \mid UZ_1,Z_2,\cdots,Z_{i-1},Y_1,Y_2,\cdots,Y_{i-1}). \end{aligned} \tag{5.8}$$

By the memoryless character of the transmission channel from Y^n to Z^n, the rightmost sum equals

$$\sum_{i=1}^n H(Z_i \mid UY_1,\cdots,Y_{i-1}) = \sum_{i=1}^n H(Z_i \mid U_i)$$

where the last equality holds by the definition of the U_i. Using this, one gets from (5.8)

$$\begin{aligned} I(U \wedge Z^n) &\leq \sum_{i=1}^n H(Z_i) - H(Z^n \mid U) \\ &\leq \sum_{i=1}^n [H(Z_i) - H(Z_i \mid U_i)] = \sum_{i=1}^n I(U_i \wedge Z_i). \end{aligned}$$

Lemma 6: With the notation of Lemma 5 we have

$$I(X^n \wedge Y^n \mid U) \leq \sum_{i=1}^n I(X_i \wedge Y_i \mid U_i).$$

Proof:

$$I(X^n \wedge Y^n \mid U) = H(Y^n \mid U) - H(Y^n \mid X^nU). \tag{5.9}$$

However,

$$\begin{aligned} H(Y^n \mid U) &= \sum_{i=1}^n H(Y_i \mid UY_1,\cdots,Y_{i-1}) \\ &= \sum_{i=1}^n H(Y_i \mid U_i) \end{aligned} \tag{5.10}$$

where the last equality follows by definition of U_i.

Since the transmission channel from X^n to Y^n is memoryless, we also have

$$H(Y^n \mid X^nU) = \sum_{i=1}^n H(Y_i \mid X_iU).$$

However, $H(Y_i \mid X_iU) = H(Y_i \mid X_i,Y_1,\cdots,Y_{i-1}U)$ because Y_i and $Y_1,\cdots,Y_{i-1}$ are statistically independent given X_iU. In our notation, this means that $H(Y_i \mid X_iU) = H(Y_i \mid X_iU_i)$. Combining this with (5.10) and substituting in (5.9) the statement of the lemma follows.

We shall prove now that passing to product spaces does not increase the maximum in (5.2).

Lemma 7: Using the notation of Lemma 5, suppose that the DRV's U,X^n,Y^n,Z^n form a Markov chain and vary over all the Markov chains with fixed transmission probabilities from $\mathcal{X}^n$ to $\mathcal{Y}^n$ and from $\mathcal{Y}^n$ to $\mathcal{Z}^n$ where these probabilities are defined in the statement of Lemma 5. Now introducing the function

$$t_n(x) \triangleq \sup \left\{ \frac{1}{n} I(X^n \wedge Y^n \mid U) \mid \frac{1}{n} I(U \wedge Z^n) \geq x \right\},$$

we state that

$$t_n(x) \leq t(x).$$

Proof: Combining Lemmas 5 and 6 we get that

$$\sup \left\{ \frac{1}{n} I(X^n \wedge Y^n \mid U) \mid \frac{1}{n} I(U \wedge Z^n) \leq x \right\}$$

$$\leq \sup \left\{ \frac{1}{n} \sum_{i=1}^{n} I(X_i \wedge Y_i \mid U_i) \mid \frac{1}{n} \sum_{i=1}^{n} I(U_i \wedge Z_i) \geq x \right\}.$$

Let us choose an arbitrary $\varepsilon > 0$. Let $U,\overline{X}^n,\overline{Y}^n,\overline{Z}^n$ yield at least $t_n(x) - \varepsilon$ for a fixed x. Then, taking the supremum in each component, we get

$$t_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(\overline{X}_i \wedge \overline{Y}_i \mid \overline{U}_i) \leq \frac{1}{n} \sum_{i=1}^{n} \cdot \sup \{ I(X \wedge Y \mid U) \mid I(U \wedge Z) \geq I(\overline{U}_i \wedge \overline{Z}_i) \}. \tag{5.11}$$

Now by the definition of the function $t(\cdot)$ for the rightmost expression in (5.11), we get

$$t_n(x) \leq \frac{1}{n} \sum_{i=1}^{n} t(I(\overline{U}_i \wedge \overline{Z}_i)) \leq t\left(\frac{1}{n} \sum_{i=1}^{n} I(\overline{U}_i \wedge \overline{Z}_i) \right) \tag{5.12}$$

where the last inequality follows from the concavity of $t(\cdot)$. Again using the fact that $t(\cdot)$ is decreasing and considering

$$\frac{1}{n} \sum_{i=1}^{n} I(\overline{U}_i \wedge \overline{Z}_i) \geq x$$

by our supposition (5.12) yields $t_n(x - \varepsilon) \leq t(x)$. Since this holds for every $\varepsilon > 0$, we have established the lemma.

We conclude with a lemma which is a slightly modified version of Lemma 1 in [12].

Lemma 8:

$$t(x) = \max \{ I(X \wedge Y \mid U) \mid I(U \wedge Z) \geq x, \quad \|U\| \leq \min (\|X\|,\|Y\|,\|Z\|) \}.$$

The proof is based on [11, lemma 1]. The following is a short outline of this proof.

Since $t(x)$ is concave in x, it follows that it equals its upper convex envelope. This is expressed by the equality

$$t(x) = \inf_{\lambda \geq 0} [\sup_{(U,X)} [\lambda I(U \wedge Z) + I(X \wedge Y \mid U)] - \lambda x]$$

i.e., $t(x)$ is the lower envelope of a family of straight lines parametrized by $(\lambda, D(\lambda) - \lambda x)$ where

$$D(\lambda) = \sup_{(U,X)} \lambda I(U \wedge Z) + I(X \wedge Y \mid U).$$

Therefore, the statement of the lemma follows if we show that, for every $\lambda \geq 0$,

$$\sup_{(U,X)} [\lambda I(U \wedge Z) + I(X \wedge Y \mid U)] = \max [\lambda I(U \wedge Z) + I(X \wedge Y \mid U) \mid (U,X); \|U\| \leq \min (\|X\|,\|Y\|,\|Z\|)] \tag{5.13}$$

where the maximization goes over all the Markov chains U,X,Y,Z with prescribed transition probabilities from $\mathcal{X}$ to $\mathcal{Y}$ and $\mathcal{Y}$ to $\mathcal{Z}$. The statement of (5.13) holds by Gallager [12, lemma 1].

It is worthwhile mentioning that since, for every fixed DRV X, we have $\|X\| + 1$ constraint functions, an argument similar to the one used in the proof of Theorem 2 (using Lemma 3) shows rather easily that for the fixed X one can add the constraint $\|U\| \leq \|X\| + 1$ and still get the same supremum. Therefore, the operation sup can be replaced by taking the maximum.

B. Converse to Coding Theorem

We adopt the usual terminology for the DBC with two components. The DBC is defined by the stochastic matrices describing the noisy channels $\{P_1(y \mid x) \mid x \in \mathcal{X}, Y \in \mathcal{Y}\}$ and $\{P_2(z \mid y) \mid y \in \mathcal{Y}, z \in \mathcal{Z}\}$ where $\mathcal{X},\mathcal{Y},\mathcal{Z}$ are finite sets. We put

$$P_2 * P_1(z \mid x) \triangleq \sum_{y \in \mathcal{Y}} P_2(z \mid y) \cdot P_1(y \mid x)$$

for every $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Let $P_2^n * P_1^n$ denote the nth memoryless extension of the DBC $\{P_1,P_2\}$, i.e., for every $x^n = x_1,x_2,\cdots,x_n$ and $z^n = z_1,z_2,\cdots,z_n$

$$P_2^n * P_1^n(z^n \mid x^n) = \prod_{i=1}^{n} P_2 * P_1(z_i \mid x_i).$$

An element of the set $\{(i,j) \mid 1 \leq i \leq M_1, 1 \leq j \leq M_2\}$ has to be sent over the DBC. The first index of the pair (i,j) is a message for decoder 1 that observes the output of the channel P_1, and the second index is a message for decoder 2 observing the output of P_2. The two sets of messages are encoded by a common encoder that assigns a codeword x_{ij} to every message pair (i,j). If the DBC is used n times during the communication, the concept of a code is the following.

Definition 7.1: Let M_1,M_2, and n be natural numbers. A set of triples $\{x_{ij},A_i,B_j\}$, $i \leq i \leq M_1$, $1 \leq j \leq M_2$, is a *code* for the DBC, if $x_{ij} \in \mathcal{X}^n$, $A_i \subseteq \mathcal{Y}^n$, $B_j \subseteq \mathcal{Z}^n$. Both the A_i and B_j are disjoint sets. At the output of channel, P_1^n i has to be decoded, thus a code serving only the communications between the encoder and decoder 1 is a set $\{x_{ij},\mathcal{A}_i\}$, $1 \leq i \leq M_1$, $1 \leq j \leq M_2$ such that, in case of no error, a sequence $y^n \in \mathcal{A}_i$ has to be received whatever x_{ij}, $1 \leq j \leq M_2$ was sent. Otherwise, we shall say that a decoding error occurred.

The *error probability* of our code is (see also Cover [7])

$$\max_{1 \leq j \leq M_2} \max_{1 \leq i \leq M_1} P_1^n(\overline{\mathcal{A}}_i \mid x_{ij}), \qquad \text{where } \overline{\mathcal{A}}_i = \mathcal{Y}^n - \mathcal{A}_i.$$

This is what can be called *maximal error*. Similarly, for the communication between the common encoder and decoder

2, we are given a code

$$\{x_{ij},\mathscr{B}_j\}, \qquad \text{where } 1 \le i \le M_1,\ 1 \le j \le M_2,\ \mathscr{B}_j \subseteq \mathscr{X}^n.$$

The probability of error for this communication is

$$\max_{1\le i\le M_1} \max_{1\le j\le M_2} P_2^n * P_1^n(\bar{\mathscr{B}}_j \mid x_{ij}),$$

where $\bar{\mathscr{B}}_j = \mathscr{X}^n - \mathscr{B}_j$.

A pair (R_1,R_2) of nonnegative reals is called an *achievable rate* for the DBC $\{P_1,P_2\}$ if, for any $0 \le \lambda_1 < \frac{1}{2}$, $0 \le \lambda_2 < \frac{1}{2}$, $\delta > 0$, and any sufficiently large n, there exists a code $\{x_{ij},\mathscr{A}_i,\mathscr{B}_j\}$ such that

$$M_1 \ge \exp[n(R_1 - \delta)] \qquad M_2 \ge \exp[n(R_2 - \delta)]. \tag{6.1a}$$

$$\max_{1\le j\le M_2} \max_{1\le i\le M_1} P_1^n(\bar{\mathscr{A}}_i \mid x_{ij}) \le \lambda_1 \tag{6.1b}$$

$$\max_{1\le j\le M_2} \max_{1\le i\le M_1} P_2^n * P_1^n(\bar{\mathscr{B}}_j \mid x_{ij}) \le \lambda_2. \tag{6.1c}$$

We shall prove the following.

Weak Converse Theorem: If (R_1,R_2) is an achievable pair of rates for the DBC $\{P_1,P_2\}$, then $R_1 \le t(R_2)$, where $t(x)$ is the function introduced in Definition 5.1 and redefined in a more explicit way in (5.13).

Our proof uses Fano's lemma and follows easily from Lemmas 4–7. Suppose that we are given a code achieving R_1,R_2, for some $(n_1,\lambda_1,\lambda_2)$. Let us consider all the codewords x_{ij} with a fixed second coordinate j. We write

$$T(j) \triangleq \{x_{ij} \mid 1 \le i \le M_1\}.$$

Let us introduce a random variable U ranging over $1 \le j \le M_2$. We suppose that U takes all its values with equal probability. For every fixed value j of U, a codeword x_{ij} is chosen in $T(j)$ at random with equal probability. The codeword finally sent is, therefore, the actual value of a random variable X^n which, conditional on any fixed value j of U, has uniform distribution on $T(j)$.

It is clear that

$$nR_2 = H(U) = I(U \wedge Z^n) + H(U \mid Z^n). \tag{6.2}$$

Using Fano's inequality, the right side of (6.2) can be upperbounded by

$$I(U \wedge Z^n) + \lambda_2 \log (M_2 - 1) + h(\lambda_2).$$

Hence

$$I(U \wedge Z^n) \ge nR_2 - \lambda_2 \cdot \log (M_2 - 1) - h(\lambda_2)$$

$$\ge nR_2(1 - \lambda_2) - h(\lambda_2). \tag{6.3}$$

The last inequality follows by (6.1a). On the other hand, one has

$$I(X^n \wedge Y^n \mid U) = H(X^n \mid U) - H(X^n \mid U,Y^n)$$

$$= nR_1 - H(X^n \mid U,Y^n) \tag{6.4}$$

where the rightmost inequality follows from the fact that conditional on every value of U, X^n is equally distributed. Using Fano's inequality for upper bounding the conditional entropy on the right side of (6.4), we get

$$I(X^n \wedge Y^n \mid U) \ge nR_1 - \lambda_1 \log (M_1 - 1) - h(\lambda_1)$$

$$\ge nR_1(1 - \lambda_1) - h(\lambda_1)$$

where the last inequality follows from condition (6.1a). Thus, for R_1, we obtain the upper bound

$$R_1 \le \frac{1}{n(1 - \lambda_1)}[h(\lambda_1) + I(X^n \wedge Y^n \mid U)]$$

$$\le \frac{1}{n(1 - \lambda_1)} h(\lambda_1) + \frac{1}{1 - \lambda_1} \cdot t\left(\frac{1}{n} I(U \wedge Z^n)\right) \tag{6.5}$$

where the last inequality follows from Lemma 7.

Since $t(\cdot)$ is monotonically decreasing, substituting (6.3) in (6.5) for $(1/n)I(U \wedge Z^n)$, we get

$$R_1 \le \frac{1}{n(1 - \lambda_1)} + \frac{1}{1 - \lambda_1} \cdot t\left(R_2(1 - \lambda_2) + \frac{1}{n} h(\lambda_2)\right). \tag{6.6}$$

If n gets arbitrarily large while λ_1 and λ_2 remain fixed, (6.6) yields

$$R_1 \le \frac{1}{1 - \lambda_1} \cdot t(R_2(1 - \lambda_2)).$$

Since the last inequality holds for every $\lambda_1 > 0$ and $\lambda_2 > 0$, we finally obtain that $R_1 \le t(R_2)$.

Acknowledgment

The authors are indebted to I. Csiszár for valuable remarks and especially for a simple proof of Lemma 3.

References

[1] J. Körner, "A property of conditional entropy," Ph.D. dissertation, Eötvös Loránd Univ., Budapest, Hungary, 1970; and *Studia Sci. Math. Hung.*, vol. 6, pp. 355–359, 1971.
[2] P. Gács and J. Körner, "Common information is much less than mutual information," *Probl. Contr. Inform. Theory*, vol. 2, pp. 149–162, 1973.
[3] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM J. Appl. Math.*, vol. 28, pp. 100–113, Jan. 1975.
[4] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, July 1973.
[5] R. Ahlswede and J. Körner, "On the connection between the entropies of input and output distributions of discrete memoryless channels," Brásor Conf. on Probl. Theory, 1974.
[6] R. Ahlswede, "Universal coding for correlated sources," presented at the 7th Hawaii Int. Conf. System Sciences, Jan. 1974.
[7] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inform. Theory*, vol. PGIT-4, pp. 2–22, Sept. 1954.
[8] T. M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 2–14, Jan. 1972.
[9] A. D. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications: Part I," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 769–772, Nov. 1973.
[10] A. D. Wyner, "A theorem on the entropy of certain binary sequences and applications: Part II," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 772–777, Nov. 1973.
[11] P. P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 197–207, Mar. 1973.
[12] R. G. Gallager, "Coding for degraded broadcast channels," to appear.
[13] ———, *Information Theory and Reliable Communication.* New York: Wiley, 1968.
[14] J. Wolfowitz, *Coding Theorems of Information Theory*, 2nd ed. Springer-Verlag: Berlin, 1964.
[15] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," to appear.