INFORMATION AND CONTROL 14, 457–473 (1969)

Correlated Decoding for Channels with Arbitrarily Varying Channel Probability Functions[1]

R. AHLSWEDE AND J. WOLFOWITZ

Ohio State University, Columbus, Ohio, and Cornell University, Ithaca, N.Y.

I. PRELIMINARIES

1. CODES AND ERRORS

Let $X = \{1, \cdots, a\}$, $Y = \{1, \cdots, b\}$ be finite sets. A stochastic matrix w with a rows and b columns will be called a channel. X, Y are the input and output alphabets (respectively) of the channel. We denote the set of all channels with input alphabet X and output alphabet Y by $\mathcal{C}(X, Y)$. A channel $w \in \mathcal{C}(X, Y)$ can be used for communication from one person, the sender, to another person, the receiver. There is given in advance a finite set of messages $\mathfrak{N} = \{1, \cdots, N\}$, one of which will be presented to the sender for transmission. We allow the sender a randomized encoding and the receiver a randomized decoding (cf. [4], [5]). More precisely, the sender encodes the message by an encoding channel $E \in \mathcal{C}(\mathfrak{N}, X)$ with $E(\nu, x)$ being the probability that input x is given to channel w when message ν is presented to the sender for transmission. When the receiver observes the output y of the transmission channel w, he decodes it by a decoding channel $D \in \mathcal{C}(Y, \mathfrak{N})$ with $D(y, \mu)$ being the probability that the receiver will decide that message μ is intended.

The matrix $e = e(E, D, w) = E \cdot w \cdot D \in \mathcal{C}(\mathfrak{N}, \mathfrak{N})$ is the error matrix of code (E, D) for channel w. Its element $e(\nu, \mu)$ gives the probability that, when ν is presented to the sender the receiver will decide that message μ is intended, when code (E, D) is used on channel w. The *average* error probability over all messages in the set $\mathfrak{N}$ is therefore

$$\bar{\lambda}_1(E, D, w) = 1 - \frac{1}{N} \text{trace } e(E, D, w).$$

[1] Research of both authors supported by the U.S. Air Force under Grant AF-AFOSR-68-1472 to Cornell University.

One also can define the *maximal* error

$$\lambda_1(E, D, w) = \max_{\nu=1,\cdots,N} (1 - e(\nu, \nu))$$

Of course, the maximal error for a code (E, D) is greater than, or equal to, the average error.

If we restrict the receiver to using nonrandomized decoding only, then $D(y, \mu)$ has only 0 and 1 as elements. Further specialization leads to the definition: A code (E, D) is *pure* if only 0 and 1 occur as elements of E , D. Pure codes usually are defined [*10*] as a system of pairs

$$\{(u_i, A_i) \mid u_i \in X, A_i \subset Y \quad \text{for} \quad i = 1, \cdots, N \qquad \text{and} \quad A_i \cap A_j = \emptyset \quad \text{for} \quad i \neq j\}.$$

The average error of a pure code is given by

$$\bar{\lambda}_3 = 1 - \frac{1}{N} \sum_{i=1}^{n} \sum_{j \in A_i} w(j \mid u_i)$$

and the maximal error by

$$\lambda_3 = \max_{i=1,\cdots,N} (1 - \sum_{j \in A_i} w(j \mid u_i)).$$

Let us denote the set of all pure codes of "length" N by $\mathcal{P}(N, X, Y)$. A probability distribution r over $\mathcal{P}(N, X, Y)$ is a *random* code. The error matrix $e(w, r)$ of a random code r is given by

$$e(w, r) = \sum_{(E,D) \in \mathcal{P}} e(E, D, w) r(E, D)$$

and the *average* error is given by

$$\bar{\lambda}_4(w, r) = 1 - \frac{1}{N} \text{trace } e(w, r).$$

Shannon [*11*] pointed out that every $(\mathfrak{N}, X, Y)$ code (E, D) is equivalent to some random $(\mathfrak{N}, X, Y)$ code r in the sense that $\bar{\lambda}_4(w, r) = \bar{\lambda}_1(w, E, D)$ for ALL $w \in \mathcal{C}(X, Y)$. The converse is not true.

We will call a general (E, D) code a code of type K_1, the special code of type K_1 which uses only nonrandomized decoding one of type K_2 , and a pure code one of type K_3 . Finally a random code will be called one of type K_4 . For each type K_i $(i = 1, 2, 3, 4)$, λ_i will denote the maximum error and $\bar{\lambda}_i$ the average error. For a single channel it is unimportant whether we work with average or with maximal errors (cf. [10], Lemma 3.3.1.) However, for two channels treated simultaneously it already makes a difference—contrary to the belief of many workers in the field—as was shown in [1] Example 1, and the difference becomes even more important for such complex systems as channels with arbitrarily varying channel probability functions (see below). For a detailed discussion see [2].

2. Channels with Arbitrary Varying Channel Probability Functions (a.v.ch.)

Let $X^t = X = \{1, \cdots, a\}$ for $t = 1, 2, \cdots$ and let $Y^t = Y = \{1, \cdots, b\}$ for $t = 1, 2, \cdots$. Also let $\mathcal{C} = \{w(\cdot|\cdot|s) \mid s \in S\}$ be a set of stochastic matrices with a rows and b columns. By $X_n = \prod_{t=1}^{n} X^t$ we denote the set of input n-sequences (words of length n) and by $Y_n = \prod_{t=1}^{n} Y^t$ we denote the set of output n-sequences. Let $S^t = S$, $t = 1, 2, \cdots$. For every n-sequence $s_n = (s^1, \cdots, s^n) \in \prod_{t=1}^{n} S^t$ we can define a discrete memoryless channel $P(\cdot|\cdot|s_n)$ by $P(y_n \mid x_n \mid s_n) = \prod_{t=1}^{n} w(y^t \mid x^t \mid s^t)$ for every $x_n = (x^1, \cdots, x^n) \in X_n$ and every $y_n = (y^1, \cdots, y^n) \in Y_n$. Consider now the class of channels

$$\mathcal{C}_n = \{P(\cdot|\cdot|s_n) \mid s_n \in S_n\}.$$

If we are interested in the simultaneous behavior of all these channels we call this indexed set of channels a "channel with arbitrarily varying channel probability functions" (a.v.ch.). (Sender and receiver communicate without knowing which individual channel actually governs the transmission of any one letter.) The coding problem is completely described when we state which code type and which error the communicators are allowed to use. The combinations $(K_i, \bar{\lambda}_i)$, (K_i, λ_i), $i = 1, 2, \cdots, 4$, are all possible, but not every possible combination corresponds to a problem of practical interest. The errors for codes for $\mathcal{C}_n$ are defined as $\bar{\lambda}_i = \max_{s_n} \bar{\lambda}_i(s_n)$ and $\lambda_i = \max_{s_n} \lambda_i(s_n)$, where $\bar{\lambda}_i(s_n)$(resp. $\lambda_i(s_n)$) is the average (resp. maximal) error for a code of type K_i for channel $P(\cdot|\cdot|s_n)$. The variables of basic interest to us are $N(n, \bar{\lambda}_i)$(resp. $N(n, \lambda_i)$) = maximal cardinality of a set of messages $\mathfrak{N}$ for which we can find a K_i-code with average error not greater than $\bar{\lambda}_i$ (resp. with maximal error not greater than λ_i).

3. The Jammer

For the a.v.ch. as defined above a more intuitive description can be given. Suppose that there is a rational malevolent being, the "jammer" say, who chooses a channel $P(\cdot|\cdot|s_n)$ so as to make communication between sender and receiver as difficult as possible. Sender and receiver want to communicate with small error probability no matter what the choice of the jammer may be. It seems to be realistic that the jammer should be able to randomize over different channels. Let Σ be a σ-algebra of subsets of S which includes all sets which consist of a single element of S. A randomization by the jammer is a probability distribution (p.d.) q on (S_n, Σ_n), where $\Sigma_n = \prod_1^{n*} \Sigma$ is the usual product σ-algebra. We introduce the notation Q_0 for the case where the jammer does not randomize, Q_1 for the case where he randomizes with respect to product probability distributions, and Q_2 for the case where the jammer can randomize with an arbitrary q_n on (S_n, Σ_n). In all these cases the jammer has no knowledge about the sequence the sender is going to send $(\mathfrak{J}^-)$. There are more possibilities for the jammer to randomize in the case $\mathfrak{J}^+$, where the jammer knows the actual sequence being sent *before* it is sent. Then the jammer can choose a p.d. q_{x_n} dependent on the word x_n to be sent. In order to have a short description for the different problems we shall use notation such as $(K_2, \bar{\lambda}_2, Q_1, \mathfrak{J}^+)$. For instance $(\cdot, \cdot, Q_0, \mathfrak{J}^-)$ describes problems introduced in [5]. Not all problems are essentially different.

LEMMA 1. *The problems described by* $(K_3, \lambda_3, \cdot, \cdot)$ *are all equivalent. As long as we use pure codes with maximal error we need not distinguish betwwen* $\mathfrak{J}^+$ *and* $\mathfrak{J}^-$ *and between the different kinds of randomization.*

Proof. A code in the case $(K_3, \lambda_3, Q_0, \mathfrak{J}^-)$ is a set of pairs $\{(u_i, A_i) \mid i = 1, \cdots, N\}$, where $u_i \in X_n, A_i \subset Y_n, A_i \cap A_j = \emptyset$ for $i \neq j$ and

$$P(A_i \mid u_i \mid s_n) \geqq 1 - \lambda_3$$

for all $s_n \in S_n$ and for all $i = 1, \cdots, N$. Therefore we have

$$\int dq_{u_i}(s_n) P(A_i \mid u_i \mid s_n) \geqq 1 - \lambda_3$$

for all p.d. q_{u_i} and all $i = 1, \cdots, N$. This means that we have a code for $(K_3, \lambda_3, Q_2, \mathfrak{J}^+)$ and *a fortiori* a code for

$$(K_3, \lambda_3, Q_2, \mathfrak{J}^-)$$
$$(K_3, \lambda_3, Q_1, \mathfrak{J}^+)$$
$$(K_3, \lambda_3, Q_1, \mathfrak{J}^-)$$
$$(K_3, \lambda_3, Q_0, \mathfrak{J}^+).$$

A code in any one of these cases is obviously a code for

$$(K_3\,, \lambda_3\,, Q_0\,, \mathfrak{S}^-).$$

This proves the lemma.

LEMMA 2. *The problems* $(K_4\,, \bar{\lambda}_4\,, \cdot\,, \mathfrak{S}^-)$ *are all equivalent.*

Proof. A code in the case $(K_4\,, \bar{\lambda}_4\,, Q_0\,, \mathfrak{S}^-)$ is given by a system of pure codes

$$\{({}_\rho u_i\,, {}_\rho A_i), i = 1, \cdots, N \mid \rho \in R\},$$

and a p.d. r on a σ-algebra of subsets of R which includes all sets which contain a single element of R, such that

$$\int_R \frac{1}{N} \sum_{i=1}^{N} P_n({}_\rho A_i \mid {}_\rho u_i \mid s_n)\, dr(\rho) \geqq 1 - \bar{\lambda}_4$$

for all $s_n \in S_n$. Therefore we have

$$\int_{sn} \int_R \frac{1}{N} \sum_{i=1}^{N} P_n({}_\rho A_i \mid {}_\rho u_i \mid s_n)\, dr(\rho) \cdot dq(s_n) \geqq 1 - \bar{\lambda}_4$$

and also

$$\int_R \frac{1}{N} \sum_{i=1}^{N} \int_{sn} P_n({}_\rho A_i \mid {}_\rho u_i \mid s_n)\, dq(s_n)) \cdot dr(\rho) \geqq 1 - \bar{\lambda}_4$$

for all p.d.'s q on s_n. This means that we have a code $(K_4\,, \bar{\lambda}_4\,, Q_2\,, \mathfrak{S}^-)$ and *a fortiori* a code for $(K_4\,, \bar{\lambda}_4\,, Q_1\,, \mathfrak{S}^-)$. A code in any one of these cases is obviously a code for $(K_4\,, \bar{\lambda}_4\,, Q_0\,, \mathfrak{S}^-)$.

4. SIDE INFORMATION

Until now we have assumed that both sender and receiver do not know which individual channel (i.e., s_n) governs the transmission (S^-, R^-). We now adopt the following notation: S^+ shall mean that the sender knows the kth component $(k = 1, \cdots, n)$ of the actual sequence s_n which the jammer will use, only after the first $(k - 1)$ letters of the word have been sent and received but before the kth letter is to be sent. S^{++} shall mean that the sender knows the entire sequence s_n before transmission of the word begins. $S^+(q)$ shall mean that the sender knows the jammer's *distribution* before transmission of the word begins. R^+ shall mean that the receiver knows the entire channel sequence s_n which governs the transmission before he decodes a received code word, and $R^+(q)$ shall mean that the receiver knows the distribution q used by the jammer before he decodes a received code word.

We shall use notation such as $(\lambda_3, Q_0, \mathfrak{J}^-, \mathcal{S}^+, R^-)$ to give a complete description. The type K_i may be omitted because its index is determined by λ_i. Not every expression of this type makes sense. For instance, expressions $(\cdot, \cdot, \mathfrak{J}^+, \mathcal{S}^+, \cdot)$ make no sense.

The following cases have been studied:

1. $(\bar{\lambda}_4, Q_2, \mathfrak{J}^-, \mathcal{S}^-, R^-)$

The coding theorem and weak converse were proved in [4]. In Section II we give a short, perspicuous, and very simple proof of a somewhat stronger result (coding theorem and strong converse). A serious drawback to the use of random codes K_4 is that they require correlated randomization between encoding and decoding. The sender, before "transmitting any message, chooses a code at random, communicates the result of his random experiment to the receiver, and then sends the message according to the code selected. This procedure is repeated at each message. It seems to the writers that this procedure cannot seriously be considered as reflecting anything remotely resembling actual communication. Surely it is vastly more complicated for the sender to transmit to the receiver the designation of the code which is the outcome of the chance experiment than it is to transmit the message itself. Yet a new code must be transmitted with each message. No doubt problems involving correlated encoding and decoding have mathematical interest." [8]

2. The more realistic cases $(i = 0, 1, 2)$

$$(\lambda_3, Q_i, \mathfrak{J}^+, \mathcal{S}^-, R^-)$$

$$(\lambda_3, Q_i, \mathfrak{J}^+, \mathcal{S}^-, R^+)$$

were introduced in [7], and necessary and sufficient conditions for the rate to be positive were given. These conditions have useful applications to several problems. (Compare for instance Section III, examples 1, and 2, and the forthcoming paper [3]).

3. In [5] Dobrushin considered the cases

$$(\bar{\lambda}_1, Q_0, \mathfrak{J}^-, \mathcal{S}^-, R^-)$$

$$(\bar{\lambda}_2, Q_0, \mathfrak{J}^-, \mathcal{S}^-, R^-)$$

Thus he allowed randomized encoding. Randomized decoding seems to provide little advantage (cf. [5]), but randomized encoding sometimes actually helps by either making a longer code possible or by reducing the error (Example 1 of Section III). Communicators interested in giving as much information through the channel as possible should therefore use codes of type K_1 or K_2, whenever feasible. Dobrushin states (without proof) a coding theorem and weak converse, and gives an explicit formula for the capacity. In [6] solutions are given for even more general cases. However, Example 2 in Section III proves that these claims are incorrect and not justified.

II. CORRELATED DECODING

First we prove

THEOREM 2.1. *Let $\mathcal{C}$ consist of the single channel w, whose capacity is $C(w)$ for codes of type K_3 and maximal or average error λ, $0 < \lambda < 1$. The capacity for codes of type K_1, K_2, or K_4 is the same, for maximal or average error.*

Proof. The statement for K_4 follows at once from the strong converse for the discrete memoryless channel, and that for K_2 will follow at once from that for K_1. By Lemma 3.1.1 of [10] the results are the same for both average and maximal error. Our result will therefore follow when we prove that, for average error, randomization in encoding and decoding cannot increase the capacity. Using maximum likelihood decoding we see immediately that randomized decoding cannot increase the capacity. That randomized encoding cannot increase the capacity follows from Lemma 3 of [2]. This proves the theorem.

We now turn our attention to correlated decoding. Let $\bar{\mathcal{C}}$ be the (ordinary) convex hull of $\mathcal{C}$. Let $H(\pi)$ be the entropy of the probability vector $\pi = (\pi_r, \cdots, \pi_a)$. Let the rate $R(\pi, w)$ of the matrix w be defined by

$$R(\pi, w) = H(\pi') - \sum_i \pi_i H(w(\cdot \mid i)),$$

where $\pi' = \pi \cdot w$. Define

$$\gamma = \max_{\pi} \inf_{w \in \bar{\mathcal{C}}} R(\pi, w). \tag{2.1}$$

By a theorem of Stiglitz [12] (see also [3], Lemma 4),

$$\gamma = \inf_{w \in \bar{\mathcal{C}}} \max_{\pi} R(\pi, w). \tag{2.2}$$

We now give a very short and perspicuous proof of a theorem due to Blackwell, Breiman, and Thomasian [4]. Our version is stronger because we prove the strong, not the weak, converse.

THEOREM 2.2. *The capacity of the channel in the case* $(\bar{\lambda}_4, Q_2, \mathfrak{J}^-, S^-, R^-)$ *is* γ.

Proof. The capacity cannot be greater than γ, by (2.2) and Theorem 2.1. It therefore remains only to prove the coding theorem.

Let q_n be any jammer's probability distribution which we temporarily hold fixed. We will prove that, when the jammer employs q_n, there is a pure code, whose average error is not greater than any given λ, $0 < \lambda < 1$, and whose length N satisfies, for any $\epsilon > 0$ and all n larger than a bound independent of q_n,

$$N > \exp\{n(\gamma - \epsilon)\}. \tag{2.3}$$

This is the lemma on page 564 of [4] and constitutes most of the proof of [4].

Let π^* be a value of π such that

$$\gamma = \inf_{w \in \bar{\mathfrak{C}}} R(\pi^*, w). \tag{2.4}$$

Let

$$t = (t_1, \cdots, t_n)$$

be a sequence of independent, identically distributed chance variables with the common distribution π^*. Let

$$t' = (t_1', \cdots, t_n')$$

be a sequence of chance variables, with values in Y, defined on the same sample space as t, and such that t' can be thought of as the chance sequence received when t is sent over the channel. (What this means is obvious.) Of course the conditional distribution of t', given $t = \tau$ (say), depends on τ and q_n. Write

$$t^{(i)} = (t_1, \cdots, t_i), \qquad t'^{(i)} = (t_1', \cdots, t_i').$$

Define the following functions for $i = 1, \cdots, n$:

$$\begin{aligned} I^{(i)}&(j, k \mid j_1, \cdots, j_{i-1}, k_1, \cdots, k_{i-1}) \\ &= \log P\{t_i = j, t_i' = k \mid t^{(i-1)} = j_1, \cdots, j_{i-1}, \\ &\qquad t'^{(i-1)} = k_1, \cdots, k_{i-1}\} \qquad\qquad (2.5) \\ &\quad - \log P\{t_i = j \mid t^{(i-1)} = j_1, \cdots, j_{i-1}, t'^{(i-1)} = k_1, \cdots, k_{i-1}\} \\ &\quad - \log P\{t_i' = k \mid t^{(i-1)} = j_1, \cdots, j_{i-1}, t'^{(i-1)} = k_1, \cdots, k_{i-1}\}. \end{aligned}$$

$(I^{(i)}(\cdot) = 0$ if any of the three expressions $P\{\ \}$ in the right member of (2.5) is zero. Write, for $i = 1, \cdots, n$,

$$\begin{aligned} V_i = I^{(i)}\,(t_i\,, t_i' \mid t^{(i-1)}, t'^{(i-1)}) \\ - E\{I^{(i)}(t_i\,, t_i' \mid t^{(i-1)}, t'^{(i-1)})|\, t^{(i-1)}, t'^{(i-1)}\} \end{aligned} \tag{2.6}$$

Then

$$EV_i = 0 \tag{2.7}$$

$$EV_iV_{i'} = 0, \qquad i \neq i' \tag{2.8}$$

$$EV_i^2 < \text{a constant independent of } i \text{ and } n \tag{2.9}$$

For any set of values of $t^{(i-1)}$ and $t'^{(i-1)}$ we have

$$E\{I^{(i)}(t_i\,, t_i' \mid t^{(i-1)}, t'^{(i-1)})|\, t^{(i-1)}, t'^{(i-1)}\} \geqq \gamma \qquad \text{by (2.4)} \tag{2.10}$$

$$\frac{\sum_{i=1}^{n} V_i}{n} \quad \text{converges stochastically to zero as } n \to \infty. \tag{2.11}$$

For any $\epsilon > 0$ we have, from (2.10) and (2.11),

$$P\left\{\sum_{i=1}^{n} I^{(i)}(t_i\,, t_i' \mid t^{(i-1)}, t'^{(i-1)}) > n(\gamma - \epsilon)\right\} \to 1 \tag{2.12}$$

as $n \to \infty$, uniformly in q_n.
The desired result (2.3) now follows immediately from (2.12) and Shannon's Theorems 7.3.1 and 7.3.2 of [*10*].

We now complete the proof exactly as in [*4*]. Since, as has just been proved, for any jammer's distribution q_n there exists a code of type K_3, with average probability of error at most λ, which satisfies (2.3), it follows from the minimax theorem that there exists a random code, i.e., one of type K_4, whose average probability of error is at most λ for every n-sequence s_n, which satisfies (2.3). This completes the proof of the coding theorem and hence of Theorem 2.2.

We can now very quickly also prove the following

THEOREM 2.3 *The capacity of the channel in each of the cases* $(\bar{\lambda}_4\,, Q_2\,, \mathfrak{J}^-, S^+(q), R^-)$, $(\bar{\lambda}_4\,, Q_2\,, \mathfrak{J}^-, S^-, R^+(q))$, *and* $(\bar{\lambda}_4\,, Q_2\,, \mathfrak{J}^-, S^+(q), R^+(q))$ *is also* γ.

Proof. Obviously the capacity cannot be less than γ, by Theorem 2.2. Let the jammer use the worst channel w^*, i.e., the one such that

$$\gamma = \max_{\pi} R(\pi, w^*),$$

for every letter; we see that the capacity cannot be greater than γ. The jammer can achieve w^* for each letter by a product probability distribution. This proves the theorem.

Obviously we can replace Q_2 by Q_1 in the statement of Theorem 2.4; in fact, the proof of the strong converse was actually given for Q_1 randomization.

We now prove

THEOREM 2.4. *The weak capacity of the channel in the cases* ($\bar{\lambda}_4$, $\cdot$, $\mathfrak{J}^-$, S^-, R^+) *is*

$$\beta = \max_{\pi} \inf_{s \in S} R(\pi, w(\cdot|\cdot| s)).$$

("Weak" capacity means that we prove the coding theorem and *weak* converse, i.e., the converse only for $\bar{\lambda}_4$ sufficiently small.)

The proof of the coding theorem differs so little from the proof of the coding theorem part of Theorem 2.2 that we omit it. As before, we use Shannon's random coding theorem to obtain a code of the required length for any given jammer's strategy q_n. Since the receiver knows the *actual* channel n-sequence which is being used (not only its probability distribution q_n), he uses this fact in the decoding. It is clear then why the rate of transmission can be β. Of course, always $\beta \geqq \gamma$.

For the proof of the weak converse we shall need

LEMMA 3. *For any $\eta > 0$ there exists a finite subset $S(\eta)$ of S such that*

$$\begin{aligned} | \max_{\pi} \inf_{s \in S(\eta)} R(\pi, w(\cdot | \cdot | s)) & \\ - \max_{\pi} \inf_{s \in S} R(\pi, w(\cdot | \cdot | s)) | \leqq \eta. & \end{aligned} \tag{2.13}$$

This is an easy consequence of Lemma 7 of [*9*] or Lemma 4.2.1 of [*10*].

LEMMA 4. *Let X_s, $s = 1, \cdots, d$, be nonnegative chance variables, defined on the same probability space, such that $EX_s \leqq \alpha$, $s = 1, \cdots, d$. For any $\epsilon > 0$ the probability of*

$$B^* = \{X_s \leqq d(\alpha + \epsilon) \quad for \quad s = 1, \cdots, d\}$$

satisfies

$$P(B^*) \geqq \frac{\epsilon}{\alpha + \epsilon}. \tag{2.14}$$

Proof. Define

$$B_s = \{X_s > d(\alpha + \epsilon)\}, \qquad s = 1, \cdots, d.$$

Then

$$P(B_s) \leqq \frac{E(X_s)}{d(\alpha + \epsilon)} \leqq \frac{\alpha}{d(\alpha + \epsilon)}$$

Hence

$$P\left(\bigcup_{s=1}^{d} B_s\right) \leqq \frac{\alpha}{\alpha + \epsilon}$$

and therefore

$$P(B^*) \geqq 1 - \frac{\alpha}{\alpha + \epsilon} = \frac{\epsilon}{\alpha + \epsilon}.$$

This lemma is due to Shannon [*13*].

We now proceed to the proof of the weak converse. Let $\eta > 0$ be arbitrary. We shall prove that, for λ sufficiently small, say $<\lambda_0$, and n sufficiently large, say $>n_0$, any code of the type given in the statement of the theorem must have length N such that

$$N < \exp\{n(\beta + 2\eta)\} \tag{2.15}$$

This is the desired result.

Let d now be the number of c.p.f.'s in $S(\eta)$, and choose λ_0 and ϵ_0 positive and so small that

$$d[d(\lambda_0 + \epsilon_0) + \epsilon_0] < 1. \tag{2.16}$$

Suppose a random code of length N with average error λ_0 is given, then we get, using Q_0 randomization only,

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{\rho \in R} r(\rho) P_n({}_\rho A_i(s_n) \mid {}_\rho u_i \mid s_n) \geqq 1 - \lambda_0 \tag{2.17}$$

for all $s_n \in S_n$. By considering only sequences $s_n = (s, \cdots, s)$ [case of compound channels], where $s \in S(\eta)$, we obtain, from (2.17), that

$$\inf_{\substack{s_n=(s,\cdots,s)\\ s\in S(\eta)}} \frac{1}{N} \sum_{i=1}^{N} \sum_{\rho \in R} r(\rho) P_n({}_\rho A_i(s_n) \mid {}_\rho u_i \mid s_n) \geqq 1 - \lambda_0, \tag{2.18}$$

and therefore

$$\inf_{\substack{s_n=(s,\cdots,s)\\ s\in S(\eta)}} \sum_{\rho \in R} r(\rho) \frac{1}{N} \sum_{i=1}^{N} P_n({}_\rho A_i(s_n) \mid {}_\rho u_i \mid s_n) \geqq 1 - \lambda_0. \tag{2.19}$$

Now apply Lemma 4 to the chance variables

$$X_s(\rho) = 1 - \frac{1}{N}\sum_{i=1}^{N} P_n({}_{\rho}A_i(s_n) \mid {}_{\rho}u_i \mid s_n = (s, \cdots, s)),$$

defined for $s \in S(\eta)$ on the probability space R which has probability measure r defined on it. Clearly,

$$EX_s \leqq \lambda_0 . \tag{2.20}$$

Hence, from Lemma 4 we obtain that there exists an element $\rho^* \in R$ such that

$$1 - \frac{1}{N}\sum_{i=1}^{N} P_n({}_{\rho^*}A_i(s_n) \mid {}_{\rho^*}u_i \mid s_n = (s, \cdots, s)) \leqq d(\lambda_0 + \epsilon_0) \tag{2.21}$$

$$\text{for} \quad s \in S(\eta).$$

We now apply Lemma 4 to the sample space $\{1, \cdots, N\}$, with p.d. $P^*(i) = 1/N$ for $i = 1, \cdots, N$, and chance variables

$$X_s(i) = 1 - P_n({}_{\rho^*}A_i(s_n) \mid {}_{\rho^*}u_i \mid s_n = (s, \cdots, s)), \qquad s \in S(\eta).$$

Then

$$EX_s \leqq d(\lambda_0 + \epsilon_0), \qquad s \in S(\eta).$$

Hence

$$P^*\{X_s \leqq d[d(\lambda_0 + \epsilon_0) + \epsilon_0], s \in S(\eta)\} \geqq \frac{\epsilon_0}{d(\lambda_0 + \epsilon_0) + \epsilon_0} \tag{2.22}$$

and, from the definition of P^*, the number of elements in the set D^* (say) in the left member of (2.22) is not less than

$$\frac{N\epsilon_0}{d(\lambda_0 + \epsilon_0) + \epsilon_0} \geqq N\epsilon_0 = N_1 \quad \text{(say)}.$$

Denote the elements of D^* by i_v, $v = 1, \cdots, N_1$. It follows from the definition of D^* and from (2.22) that

$$P_n({}_{\rho^*}A_{i_v}(s_n) \mid {}_{\rho^*}u_i \mid s_n = (s, \cdots, s)) \geqq 1 - d[d(\lambda_0 + \epsilon_0) + \epsilon_0] \tag{2.23}$$

for $v = 1, \cdots, N_1$. The inequality (2.15) now follows from (2.23) and Theorem 2 of [*9*] (or Theorem 4.4.1 of [*10*]).

III. REMARKS ON PAPERS [*5*] AND [*6*]. A COUNTER-EXAMPLE TO THEOREM 1 OF [*5*] AND THEOREM 2 OF [*6*]

For every fixed $i \in X$ let $T(i)$ denote the minimal convex closed system of probability distributions on Y which contains all distributions $\{w(\cdot \mid i \mid s) \mid s \in S\}$. The set of matrices

$$\overline{\overline{\mathcal{C}}} = \{(w(j \mid i))\underset{j = 1 \cdots b}{i = 1 \cdots a} \mid w(\,\cdot \mid i) \in T(i), i = 1 \cdots a\}$$

is called the row convex closure of the set $\mathcal{C}$. The difference between the row convex closure and the usual convex closure $\overline{\mathcal{C}}$ of a system of matrices lies in the fact that for each row we take a possibly different linear combination of its elements to obtain $\overline{\overline{\mathcal{C}}}$.

Example 1. Randomization in encoding can be an improvement over nonrandomized encoding. Let $a = b = 3$

$$w_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad w_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

$$\mathcal{C} = (w_1, w_2).$$

Obviously $T(i) \cap T(j) \neq \varnothing$ for $i \neq j$. It follows therefore from Theorem 1 of [7] that the capacity is 0 in the case $(\lambda_3, Q_0, \mathfrak{I}^+, S^-, R^-)$.

As a consequence of Lemma 1 the capacity is also 0 in the case $(\lambda_3, Q_0, \mathfrak{I}^-, S^-, R^-)$. Randomization in the encoding can be interpreted as an enlargement of the possible input sequences for a channel. Instead of the set of input n-sequences X_n we have the set $\mathcal{P}(X_n) =$ set of all p.d. on X_n available for the encoding. We shall make use only of the subset $\mathcal{P}^*(X_n) =$ set of all product distributions on X_n. Actually we shall use only all sequences $q_n = q^1 \times q^2 \times \cdots \times q^n$, which have as components q^t either δ or q, where

$$\delta{:}\delta(1) = 1, \qquad \delta(2) = \delta(3) = 0,$$

$$q{:}q(2) = q(3) = \tfrac{1}{2}, \qquad q(1) = 0.$$

This means that we restrict ourselves to special letter by letter randomizations. Randomization per letter means convex combination of rows in our matrices. To find optimal codes using only q_n means therefore to find optimal codes for

$$\mathcal{C}^* = \left\{ w_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \tfrac{1}{2} & \tfrac{1}{2} \end{pmatrix}, \qquad w_2 = \begin{pmatrix} 0 & 1 & 0 \\ \tfrac{1}{2} & 0 & \tfrac{1}{2} \end{pmatrix} \right\}$$

in the case $(\lambda_3, Q_0, \mathfrak{J}^-, S^-, R^-)$. Now $T(1) \cap T(2) = \emptyset$ and we therefore can transmit with rate $R > 0$ (Theorem 1 of [7].) *A fortiori* we can transmit over our original channel with a positive rate if we use randomized encoding.

Define

$$C_D = \max_{\pi} \inf_{w \in \bar{\bar{\mathfrak{C}}}} R(\pi, w). \tag{3.1}$$

By a theorem of Stiglitz [12] (see also [3], Lemma 4),

$$C_D = \inf_{w \in \bar{\bar{\mathfrak{C}}}} \max_{\pi} R(\pi, w). \tag{3.2}$$

Dobrushin asserts without proof that

C_D is the capacity of the channel in the case

$$(\bar{\lambda}_2, Q_0, \mathfrak{J}^-, S^-, R^-) \tag{3.3}$$

([5], Theorem 1, Remark 3), and that

C_D is the capacity of the channel in the case

$$(\bar{\lambda}_1, Q_0, \mathfrak{J}^-, S^-, R^-) \tag{3.4}$$

([5], end of paragraph following Eq. (4)). We shall now prove, by Example 2, that (3.3) and (3.4) are not true.

Example 2. (Counter-example to Theorem 1 in [5] and Theorem 2 in [6]).

Suppose given two matrices w, w' with 3 rows and 3 columns. We denote the ith row vector in w by i and the ith row vector in w' by i'. We represent these vectors as points in E^2. Let w, w' be such that their representation is given by the following figure 1. The point of intersection G of the Lines 1, 1$'$ and 2, 2$'$ is to be both 3 and 3$'$. Computing C_D by (3.2), using as w the matrix all of whose rows are G, we obtain that $C_D = 0$. Thus, according to Dobrushin, the capacity of this channel in the cases $(\bar{\lambda}_2, Q_0, \mathfrak{J}^-, S^-, R^-)$ and $(\bar{\lambda}_1, Q_0, \mathfrak{J}^-, S^-, R^-)$ is zero.

We now randomize over the letters 1, 2 with probability ½ each, and obtain the points L, L'. However, since the line L, L' and the "line" G, G (which is $T(3)$) are disjoint, it follows from Theorem 1 of [7] that, for any λ, $0 < \lambda < 1$, one can transmit at a positive rate with maximal error λ. Hence in the case $(\bar{\lambda}_2, Q_2, \mathfrak{J}^-, S^-, R^-)$, and, *a fortiori*, in the cases $(\bar{\lambda}_2, Q_0, \mathfrak{J}^-, S^-, R^-)$ and $(\bar{\lambda}_1, Q_0, \mathfrak{J}^-, S^-, R^-)$, one can transmit at a positive rate. This contradicts (3.3) and (3.4).

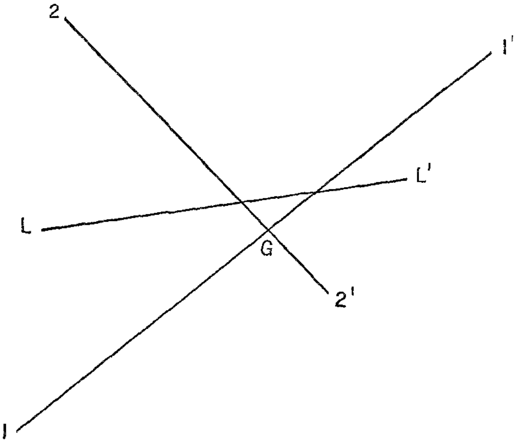

FIG. 1

This example also shows that randomization in encoding can increase the capacity, and therefore could have been used in place of Example 1.

IV. CHANNELS WITH ARBITRARILY VARYING, BINARY SYMMETRIC, CHANNEL PROBABILITY FUNCTIONS

Case $(\lambda_1, Q_0, \mathfrak{I}^-, S^+, R^-)$. Given

$$\mathcal{C} = \{w(\cdot \mid \cdot \mid s) \mid w(\cdot \mid \cdot \mid s) = \begin{pmatrix} 1-s & s \\ s & 1-s \end{pmatrix}$$

for all $s \in S \subset [0, 1]\}$. Define

$$w = \begin{pmatrix} 1-s_0 & s_0 \\ s_0 & 1-s_0 \end{pmatrix},$$

where $1 - 2s_0 = \inf_{s \in S} | 1 - 2s |$.

THEOREM 4.1. *The capacity of this a.v.ch. in the case* $(\lambda_1, Q_0, \mathfrak{I}^-, S^+, R^-)$ *is* $C = 1 + s_0 \log s_0 + (1 - s_0) \log (1 - s_0)$, *that is, the capacity of the discrete memoryless channel defined by* w.

Proof. Let $\{(u_i, A_i) \mid i = 1, \cdots, N\}$ be a λ-code for the d.m.c. w. Write

$$w(\cdot \mid \cdot \mid s_1) = \begin{pmatrix} 1-s_1 & s_1 \\ s_1 & 1-s_1 \end{pmatrix}.$$

Define

$$p^{(1)}(1) = \frac{1 - s_0 - s_1}{1 - 2s_1}, \qquad p^{(1)}(2) = \frac{s_0 - s_1}{1 - 2s_1} \tag{4.1}$$

$$p^{(2)}(1) = \frac{s_0 - s_1}{1 - 2s_1}, \qquad p^{(2)}(2) = \frac{1 - s_0 - s_1}{1 - 2s_1} \tag{4.2}$$

For $s_1 \in S$ both $p^{(1)}(\cdot)$ and $p^{(2)}(\cdot)$ are probability vectors. When the sender wants to transmit message $i(i = 1, \cdots, N)$ he proceeds as follows: When the kth letter $(k = 1, \cdots, n)$ will be transmitted according to $w(\cdot|\cdot|s_1)$ he uses the "random letter" $p^{(1)}(\cdot)$ if the kth letter of u_i is 1 and the "random letter" $p^{(2)}(\cdot)$ if the kth letter of u_i is 2. Thus the probability, of receiving any output n-sequence y_n when the ith message is sent, is the same for all channel n-sequences. We can transmit with rate $R \geqq C$. The strong converse follows from the strong converse for the d.m.c. w and Theorem 2.1.

In Theorem 4.1 we could obviously have replaced S^+ by S^{++}.

THEOREM 4.2. *The capacity of this a.v.ch. in the case* $(\lambda_1, Q_0, \mathcal{S}^-, S^-, R^+)$ *is also* $C = 1 + s_0 \log s_0 + (1 - s_0) \log (1 - s_0)$.

Proof. Suppose the kth letter is sent according to $w(\cdot|\cdot|s_1)$, and this is known to the receiver. When the letter $j(j = 1, 2)$ is actually received, the receiver performs an independent random experiment with probability distribution $p^{(j)}(\cdot)$ from (4.1) or (4.2), and acts as if the outcome of this experiment were the actual letter received. It follows from the computations in Theorem 4.1 that

$$\begin{pmatrix} 1 - s_1 & s_1 \\ s_1 & 1 - s_1 \end{pmatrix} \cdot \begin{pmatrix} p^{(1)}(1) & p^{(2)}(1) \\ p^{(1)}(2) & p^{(2)}(2) \end{pmatrix} = \begin{pmatrix} 1 - s_0 & s_0 \\ s_0 & 1 - s_0 \end{pmatrix}.$$

The second matrix on the left being symmetric we obtain that

$$\begin{pmatrix} 1 - s_1 & s_1 \\ s_1 & 1 - s_1 \end{pmatrix} \cdot \begin{pmatrix} p^{(1)}(1) & p^{(1)}(2) \\ p^{(2)}(1) & p^{(2)}(2) \end{pmatrix} = \begin{pmatrix} 1 - s_0 & s_0 \\ s_0 & 1 - s_0 \end{pmatrix}.$$

Thus, whatever be the channel n-sequence s_n, the receiver randomizes in such a way that the distribution of the virtual received sequence is the same as that for the d.m.c. w. This proves the coding theorem. The converse is obvious.

The method can be extended to a.v.ch. for which one matrix is right included (receiver R^+) resp. left included (sender S^+) by all others (cf. Shannon [11]).

RECEIVED: February 19, 1969

REFERENCES

1. AHLSWEDE, R., "Certain results in coding theory for compound channels," to appear in Colloquium on Information Theory, Debrecen, Hungary, 1967.
2. AHLSWEDE, R. AND WOLFOWITZ, J., The structure of capacity functions for compound channels (to appear).
3. AHLSWEDE, R. AND WOLFOWITZ, J., The capacity of a channel with arbitrarily varying channel probability functions and binary output alphabet (to appear).
4. BLACKWELL, D., BREIMAN, L., AND THOMASIAN, A. J., The capacities of certain channel classes under random coding *Ann. Math. Stat.* **31** (1960), 558–567.
5. DOBRUSHIN, R. L., Individual methods for transmission of information for discrete channels without memory and messages with independent components *Doklady Akad. Nauk SSSR* **148** (1963), 1245–48, *Soviet Math., Doklady* **4** (1963), 253.
6. DOBRUSHIN, R. L., Unified methods for the transmission of information: the general case *Doklady Akad. Nauk SSSR*, **149** (1963), 16–19.
7. KIEFER, J. AND WOLFOWITZ, J., Channels with arbitrarily varying channel probability functions *Inform. Control,* **5** (1962), 44–54.
8. KOTZ, S. AND WOLFOWITZ, J., Twenty-five years of progress in information theory, The Centennial Celebration Volume of the University of California (to appear).
9. WOLFOWITZ, J., Simultaneous Channels, *Arch. Rat. Mech. Anal.* **4** (1960), 371–386.
10. WOLFOWITZ, J., "Coding Theorems of Information Theory." Springer Verlag, Berlin. 1st ed., 1961; 2nd ed., 1964.
11. SHANNON, C. E., A note on a partial ordering for communication channels *Inform. Control* **1** (1958), 390–397.
12. STIGLITZ, I. G., Coding for a class of unknown channels *IEEE Trans. on Info. Theory*, **IT12** (1966), 189–195.
13. SHANNON, C. E., "Two-way communication channels" *Proc. Fourth Berkeley Symposium on Math. Statistics and Probability*, Vol. I, 611–644. Univ. of California Press, Berkeley, and Los Angeles, 1961.