Capacity and Coding for Degraded Broadcast Channels

(Translated into Russian, published in Problemy Peredaci Informacii, 1974)

By

Robert G. Gallager

Abstract

The capacity region of discrete memoryless degraded broadcast channels with two receivers is defined. A converse to the coding theorem is then proved which shows that small error probability is not possible for transmission rates outside of this capacity region. An earlier coding theorem of Bergmans (1973) showing that arbitrarily small error probability is achievable within the capacity region, is strengthened by obtaining bounds on error probability that are exponential in block length and analogous to the results for single receiver channels.

Capacity and Coding for Degraded Broadcast Channels

By

Robert G. Gallager
Massachusetts Institute of Technology

Introduction

The first information theoretic analysis of broadcast channels was done by Cover (1971). He showed, among other things, that if separate messages are to be sent to a number of receivers from a common transmitter, then it is often possible to improve upon the usual engineering solution of time sharing the transmitter between the receivers. Subsequently Bergmans (1973) established a coding theorem for the class of degraded broadcast channels and Wyner and Ziv (1973) and Wyner (1973) established a converse to the coding theorem for the binary symmetric degraded broadcast channel. In this note we establish a converse to the coding theorem for degraded broadcast channels and establish a reliability function bound for coding on the same channels.

The Model and Definition of Capacity

The model we want to analyze is depicted in figure 1. There are two sources, the first producing an integer m, $0\leq m\leq M_1-1$ and the second an integer i, $0\leq i\leq M_2-1$. The encoder maps the pair m,i into a code word $\overline{x}_{m,i}$ which is a sequence of N channel input symbols, $\overline{x}_{m,i}=(x_{m,i,1},\ldots,x_{m,i,N})$. The two channels are discrete memoryless channels, the first with an input alphabet $(0,1,\ldots,K-1)$ and output alphabet $(0,1,\ldots J-1)$, and the second with the output alphabet $(0,1,\ldots,L-1)$. The input to the second channel is of course the output from the first.

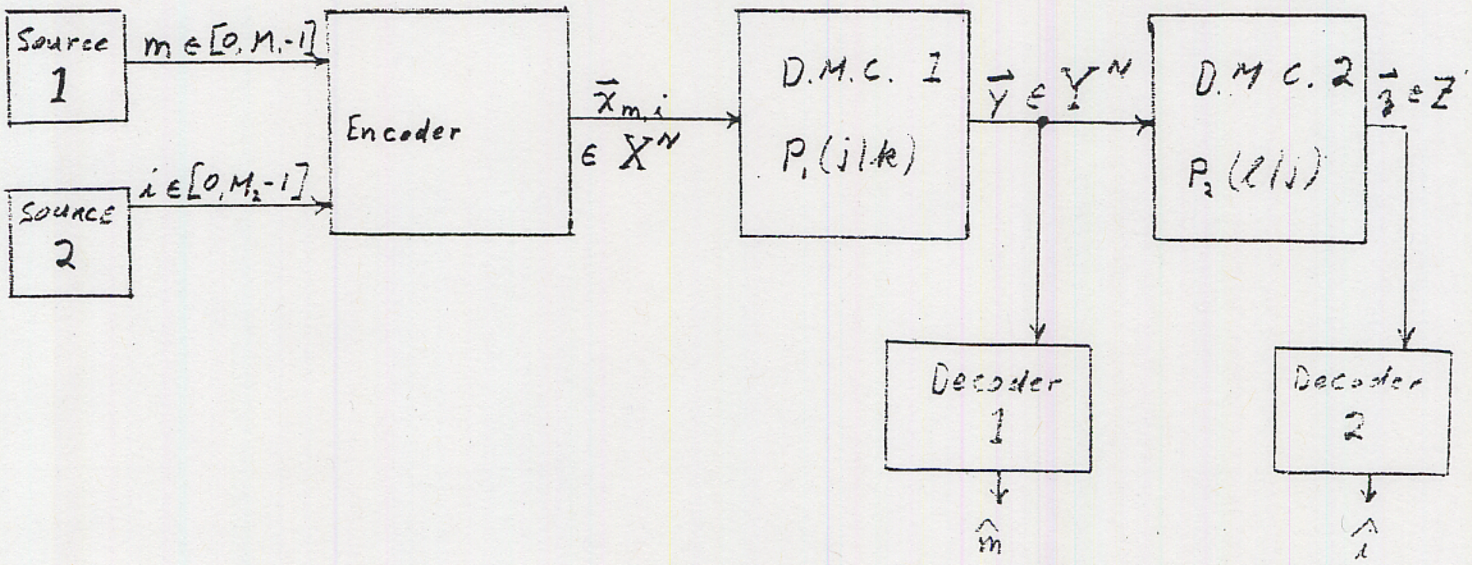
The channel transition probabilities for the two channels are denoted by $P_1(j|k)$ and $P_2(\ell|j)$ respectively where $0 \leq k \leq K-1$, $0 \leq j \leq J-1$, $0 \leq \ell \leq L-1$. Given that $\overline{x}_{m,i}$ is a sequence of N inputs to the first channel, the corresponding sequence of N output symbols, $\overline{y}=(y_1,\ldots,y_N)$ has the probability

$$P_{1,N}(\overline{y}|\overline{x}_{m,i}) = \prod_{n=1}^{N} P_1(y_n|x_{m,i,n}) \tag{1}$$

where P_1 is the transition probability assignment for the first channel described above. Likewise, given the output $\overline{y}=(y_1,\ldots,y_N)$ from the first channel, the probability of output $\overline{z}=(z_1,\ldots,z_N)$ from the second channel is given by

$$P_{2,N}(\overline{z}|\overline{y}) = \prod_{n=1}^{N} P_2(z_n|y_n) \tag{2}$$

The first decoder maps $\overline{y}$ into an estimate $\hat{m}$ of the output of the first source, m, and the second decoder maps $\overline{z}$ into an estimate $\hat{i}$ of the output of the second source i. We say that decoder one makes an error when $\hat{m} \neq m$ and decoder two makes an error when $\hat{i} \neq i$. We shall define the probability of such error events later after we have been more explicit about the probabilistic description of the sources.



Model of Degraded Broadcast Channel

Figure 1

The model described here is restrictive in five ways. First there are only two receivers rather than an arbitrary number; second, separate messages are sent to each receiver, rather than also sending some common information to them; third, the channels are restricted to be discrete; fourth the channels are memoryless; and fifth, the input $\bar{z}$ to the second decoder is restricted to be a degraded version of the input $\bar{y}$ to the first decoder. The first three restrictions are imposed only for simplicity; their removal causes no conceptual problems. The memoryless restriction is added because of the difficulty of dealing cleanly with a single channel with memory rather than because of any new problems brought in by the broadcast situation. The fifth restriction, to degraded channels, is the most serious and is imposed because of conceptual reasons. It is possible to prove a coding theorem of a more general type than the one here for non-degraded broadcast channels, but no way is known of establishing a converse.

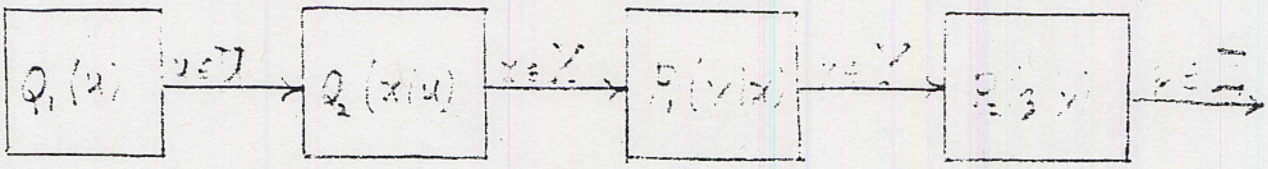
In the situation of figure 1, we would expect a trade-off to exist between the amount of information transmitted from source 1 to decoder 1 and that transmitted from source 2 to decoder 2. Thus, instead of characterizing the system by a single number called capacity, we expect to characterize it by a function $C_2(C_1)$ giving the maximum information from source 2 to decoder 2 as a function of the information from source 1 to decoder 1. Following the usual procedure for single channels, we shall first define $C_2(C_1)$ in terms of mutual information measures for a single use of the channels, and then we will establish a converse to the coding theorem and a coding theorem, giving our definition significance.

Consider the joint ensemble U,X,Y,Z shown in figure 2. The ensembles X,Y,Z, have K,J, and L sample points respectively, corresponding to the input and output alphabets of the broadcast channel. The number of sample points in U is min(K,J,L). The probability of a sample point u,x,y,z in the joint ensemble is given by

$$\Pr(u,x,y,z) = Q_1(u)Q_2(x|u)P_1(y|x)P_2(z|y) \tag{3}$$

Using the usual definitions of average mutual information,

$$I(U;Z) = \sum_{u,z} \Pr(u,z) \ln \frac{\Pr(z|u)}{\Pr(z)} \tag{4}$$



Model for Information Measures used in Defining Capacity

Figure 2

$$I(X;Y|U) = \sum_{x,y,u} \Pr(x,y,u) \ln \frac{\Pr(y|u,x)}{\Pr(y|u\ )} \tag{5}$$

we define, for all $\lambda \geq 0$,

$$C(\lambda) = \max_{Q_1,Q_2} \quad I(U;Z) + \lambda I(X;Y|U) \tag{6}$$

The maximum above is over all probability assignments $Q_1(u)$ and transition probability assignments $Q_2(x|u)$. It is clear that $C(\lambda)$ is non-decreasing in λ and is a function only of λ and of the transition probabilities P_1 and P_2 defining the broadcast channel. Next we define the capacity function

$$C_2(C_1) = \inf_{\lambda \geq 0} \ [C(\lambda) - \lambda C_1] \tag{7}$$

Geometrically, $C_2(C_1)$ is the lower envelope of the set of straight lines parameterized by λ, $C(\lambda) - \lambda C_1$. It then follows that $C_2(C_1)$ is convex $\cap$. An equivalent definition would be to define $C_2{}^*(C_1)$ by

$$C_2^*(C_1) = \max_{Q_1,Q_2:\ I(X;Y|U)\geq C_1} I(U;Z) \tag{8}$$

Then $C_2(C_1)$ is the convex hull of $C_2^*(C_1)$. Intuitively, we associate $I(U;Z)$ with the information transmitted from source 2 to decoder 2, and we associate $I(X;Y|U)$ with the information from source 1 to decoder 1. Naturally it will not be entirely clear why this definition of $C_2(C_1)$ is appropriate until we prove a coding theorem and converse.

A sketch of the function $C_2(C_1)$ is given in figure 3. The upper and lower bounds there are due to Bergmans (1973). The horizontal upper bound arises from the observation that $C(\lambda)$ for $\lambda=0$ is $\max I(X;Z)$. The upper bound of slope minus one arises from considering $C(\lambda)$ at $\lambda=1$. Since $I(U;Z) + I(X;Y|U) \leq I(U;Y) + I(X;Y|U) = I(X;Y)$, the maximum in (6) is achieved with U atomic and $C(1) = \max I(X;Y)$. Using the same argument, we see that for $\lambda \geq 1$, $C(\lambda) = \lambda \max I(X;Y)$, and it then follows from (7) that $C_2(\max I(X;Y)) = 0$. The lower bound follows from applying convexity to the two end points.

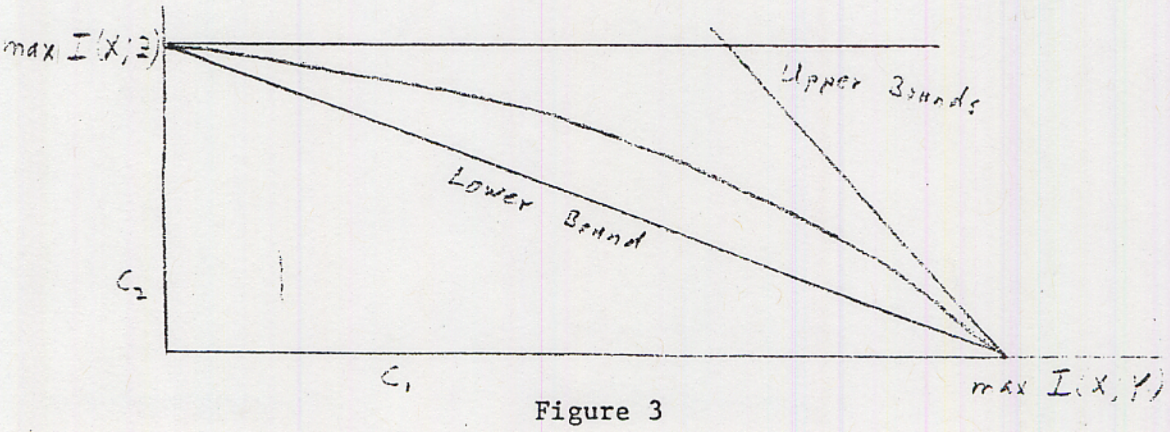


Figure 3

Typical shape of Capacity Region

Before proving a converse to the coding theorem, we need the following lemma.

Lemma 1: Let U,X,Y,Z be a joint ensemble satisfying (3) in which X,Y,Z have K,J, and L sample points respectively and where the number of sample points in U is arbitrary. Then

$$I(U;Z) + \lambda I(X;Y|U) \leq C(\lambda) \tag{9}$$

Furthermore if the number of sample points in U exceeds min(K,J,L), then

$$C(\lambda) = \max_{Q_1,Q_2} \; I(U;Z) + \lambda I(X;Y|U) \tag{10}$$

This lemma is proved in the appendix, but is really just a paraphasing of the well known result that the capacity of a single discrete memoryless channel can be achieved using no more imput symbols than the number of output symbols.

The Converse to the Coding Theorem

In proving a converse to the coding theorem, we must be explicit about the probabilistic structure of the source; the critical parameter turns out to be the entropy of the source rather than the number of letters in the source alphabet. It also costs nothing to allow the encoder to perform a stochastic mapping rather than a deterministic mapping. Thus we let W_1 be the ensemble of outputs from source 1 (with the sample space being the integers 0 to M_1-1) and let W_2 be the ensemble of outputs from source 2 (with the sample space 0 to M_2-1). We let $Q(m,i,\overline{x})$ be an arbitrary probability assignment on $m\varepsilon W_1$, $i\varepsilon W_2$, $\overline{x}\varepsilon X^N$ where $X^N=X_1,X_2,\ldots,X_N$ is the ensemble of N sequences of channel input symbols. We define the source rates R_1 and R_2 (in natural units per channel symbol) by

$$R_1 = \frac{H(W_1|W_2)}{N} = \frac{1}{N} \sum_{m,i} Pr(m,i) \ln \frac{1}{Pr(m|i)} \tag{11}$$

$$R_2 = \frac{H(W_2)}{N} = \frac{1}{N} \sum_{i} Pr(i) \ln \frac{1}{Pr(i)} \tag{12}$$

The notation for probabilities above will be unambiguous so long as we remember to associate the letters i and m with symbols in the appropriate ensembles, namely W_2 and W_1 respectively. For the most important case, in which the sources are independent and produce equi-probable letters, $R_1=(\ln M_1)/N$ and $R_2=(\ln M_2)/N$.

Finally we allow the decoders to be stochastic, the first mapping any $\overline{y}$ into $\hat{m}$ with probability $\mu_1(\hat{m}|\overline{y})$ and the second mapping $\overline{z}$ into $\hat{i}$ with probability $\mu_2(\hat{i}|\overline{z})$. This of course includes deterministic decoders as a special case. We assume then that the probability of a joint sample point $m,i,\overline{x},\overline{y},\overline{z},\hat{m},\hat{i}$ is given by

$$Q(m,i,\overline{x})P_{1,N}(\overline{y}|\overline{x})P_{2,N}(\overline{z}|\overline{y})\mu_1(\hat{m}|\overline{y})\mu_2(\hat{i}|\overline{z}) \tag{13}$$

Heuristically, (13) has the effect of ruling out any side information into the decoders and of ruling out any collusion between the decoders; the first operates only on $\overline{y}$ and the second only on $\overline{z}$. In the above ensemble, the event $m\neq\hat{m}$ is the event that the first decoder makes an error and the event $i\neq\hat{i}$ is the event that the second decoder makes an error. The probability of error for the first and second decoders then are

$$P_{e,1} = \Pr(m\neq\hat{m})$$

$$P_{e,2} = \Pr(i\neq\hat{i})$$

Before establishing a converse to the coding theorem, we need a lemma which relates the mutual information between the sources and decoders to the function $C(\lambda)$ defined in (6).

Lemma 2: For all $\lambda>0$,

$$I(W_2;Z^N) + \lambda I(W_1;Y^N|W_2) \leq NC(\lambda) \tag{15}$$

This lemma is proved in the appendix.

Theorem 1: If for some $\lambda>0$ and some $\varepsilon>0$ we have

$$\lambda R_1 + R_2 \geq C(\lambda) + \varepsilon \tag{16}$$

then

$$\lambda[P_{e,1}\ln(M_1-1)+\mathcal{H}(P_{e,1})] + P_{e,2}\ln(M_2-1)+\mathcal{H}(P_{e,2}) \geq N\varepsilon \tag{17}$$

where the function $\mathcal{H}$ is given by

$$\mathcal{H}(P_e) = -P_e\ln P_e - (1-P_e)\ln(1-P_e) \tag{18}$$

Discussion: The interpretation of this result is almost the same as that of the so called weak converse to the coding theorem for single channels. The condition (16) asserts that the rate pair (R_1,R_2) is outside of the region under the capacity function $C_2(C_1)$, and (17) asserts that $P_{e,1}$ and $P_{e,2}$ cannot both be arbitrarily small in this case.

Proof:

$$I(W_2;Z^N) = H(W_2) - H(W_2|Z^N) \tag{19}$$

$$= NR_2 - H(W_2|Z^N)$$

$$I(W_1;Y^N|W_2) = H(W_1|W_2) - H(W_1|Y^N,W_2)$$

$$\geq NR_1 - H(W_1|Y^N) \tag{20}$$

In (20) we have used the fact that additional conditioning decreases entropy (Gallager (1968), Eq. 2.3.13).

Substituting (19) and (20) into (15), and using (16), we have

$$N\varepsilon \leq \lambda H(W_1|Y^N) + H(W_2|Z^N) \tag{21}$$

From the Fano inequality (Theorem 4.3.1, Gallager (1968)),

$$H(W_1|Y^N) \leq P_{e,1}\ln(M_1-1) + \mathcal{H}(P_{e,1}) \tag{22}$$

$$H(W_2|Z^N) \leq P_{e,2}\ln(M_2-1) + \mathcal{H}(P_{e,2}) \tag{23}$$

Substituting (22) and (23) into (21) completes the proof.

The theorem as stated does not provide much insight into what happens as the block length N is varied. In particular, if the sources were sequences of letters, the lengths of the sequences growing with N, then the theorem asserts that there is appreciable probability that the sequences are not decoded correctly, but doesn't rule out the possibility that the error probability per source digit might go to zero. This in fact cannot happen. The argument is the same as that in Theorem 4.3.4 of Gallager (1968), the only difference being the use of lemma 2 here to upper bound the average mutual informations.

The Coding Theorem

The coding theorem for degraded broadcast channels has been established by Bergmans (1973). The only additional feature here will be to provide upper bounds on the error probabilities which go to zero exponentially in the block length. We consider the same ensemble of codes as Bergmans did. In particular, let $Q_1(u)$ and $Q_2(x|u)$ be arbitrary probability assignments as in figure 2. Consider the ensemble of M_2 code words, $\overline{u}_1,\overline{u}_2,\ldots\overline{u}_{M_2}$ in which each of the N letters in each of the M_2 words is independently selected according to the probability assignment $Q_1(u)$. For each of these code words (called cluster centers), we choose M_1 code words with independent digits according to the assignment $Q_2(x|u)$. That is, conditional on $\overline{u}_i=u_{i,1}\ldots u_{i,N}$ being the ith cluster center, the probability of a sequence $\overline{x}_{m,i}=x_{m,i,1},\ldots,x_{m,i,N}$ being the mth of the M_1 code words selected from $\overline{u}_i$ is given by

$$\Pr(\bar{x}_{m,i}|\bar{u}_i) = \prod_{n=1}^{N} Q_2(x_{m,i,n}|u_{i,n}) \qquad (24)$$

Given a particular code in this ensemble (namely a set of N sequences $\{\bar{x}_{m,i}\}$, $1 \leq m \leq M_1$, $1 \leq i \leq M_2$, the encoding rule is to map m, i (the outputs from the first and second source respectively) into $\bar{x}_{m,i}$, which is then transmitted.

We shall not use maximum likelihood decoding, but rather a slightly simpler decoding rule that is easier to analyze. First define

$$P_3(z|u) = \sum_{x}\sum_{y} Q_2(x|u)P_1(y|x)P_2(z|y) \qquad (25)$$

$$P_{3,N}(\bar{z}|\bar{u}_i) = \prod_{n=1}^{N} P_3(z_n|u_{i,n}) \qquad (26)$$

Note that for a given choice of cluster centers $\bar{u}_1,\ldots\bar{u}_{M_2}$, $P_{3,N}(\bar{z}|\bar{u}_i)$ is the probability, over the ensemble of code words $\{\bar{x}_{m,i}\}$, of receiving sequence $\bar{z}$ at the second decoder. The decoding rule for decoder 2 is taken as: decode that i for which $P_{3,N}(\bar{z}|\bar{u}_i)$ is maximum (if there is a tie for the maximum, we will count the decoding as being in error).
This rule is slightly suboptimal in the sense that decoder 2 is only using its knowledge of the cluster centers and not of the code words. The resulting simplification is that for fixed $\bar{u}_1,\ldots,\bar{u}_{M_2}$, and over the ensemble of code words $\{\bar{x}_{m,i}\}$, $\bar{u}_i$ is stochastically mapped into $\bar{z}$ with the probability assignment $P_{3,N}$, corresponding to a discrete memoryless channel with transition probabilities $P_3(z|u)$. Thus the probability of decoding error

is upper bounded by the usual results for coding on a discrete memoryless channel (see Gallager (1968), Chap. 5)

$$P_{e,2} \leq \exp - N\, E_2(R_2) \tag{27}$$

$$E_2(R_2) = \max_{0\leq\rho\leq 1} \; E_{0,2}(\rho) - \rho R_2 \tag{28}$$

$$E_{0,2}(\rho) = -\ln \sum_z \left(\sum_u Q_1(u) P_3^{\frac{1}{1+\rho}}(z|u)\right)^{1+\rho} \tag{29}$$

$$R_2 = \frac{\ln M_2}{N} \tag{30}$$

Furthermore, $E_2(R_2) > 0$ if $R_2 < I(U;Z)$.

Next consider the first decoder. It will first decode the cluster center $\overline{u}_i$ in the same way as decoder 2 does, and then using its estimate of i, it will choose the m for which $\Pr(\overline{y}|\overline{x}_{m,i})$ is maximum. Let $P_{e,12}$ be the probability that i is incorrectly decoded and let $P_{e,11}$ be the probability (over the entire ensemble, with m and i the randomly chosen source outputs) that for some $m'\neq m$, $\Pr(\overline{y}|\overline{x}_{m',i}) \geq \Pr(\overline{y}|\overline{x}_{m,i})$. We see that

$$P_{e,1} \leq P_{e,11} + P_{e,12} \tag{31}$$

The probability of error on the cluster center is bounded by the same argument that we used for the second decoder. The only difference is that the discrete memoryless channel has transition probabilities

$$P_4(y|u) = \sum_x P_1(y|x) Q_2(x|u) \tag{32}$$

We then have

$$P_{e,12} \leq \exp - N\, E_{12}(R_2) \tag{33}$$

$$E_{12}(R_2) = \max_{0 \leq \rho \leq 1} E_{0,12}(\rho) - \rho R_2 \tag{34}$$

$$E_{0,12}(\rho) = - \ln \sum_{y} \Big(\sum_{u} Q_1(u) P_4^{\frac{1}{1+\rho}}(y\,|u)\Big)^{1+\rho} \tag{35}$$

It is physically plausible, and can be shown analytically by applying Minkowski's inequality to $E_{0,12}(\rho)$, that $E_{12}(R_2) \geq E_2(R_2)$.

Next we must upper bound $P_{e,11}$. We first condition the event of this type of error upon a given pair of source outputs m,i, and a particular choice of cluster center $\overline{u}_i=\overline{u}$. Let $P_{e,11}(\overline{u})$ be the probability of this error event. That is, $P_{e,11}(\overline{u})$ is the probability that $Pr(\overline{y}|\overline{x}_{m',i}) \geq Pr(\overline{y}|\overline{x}_{m,i})$ for some m'≠m in the conditional ensemble where for each m', $1 \leq m' \leq M_1$, $\overline{x}_{m',i}$ is independently chosen with the probability assignment $Pr(\overline{x}_{m',i}|u) = \prod_{n=1}^{N} Q_2(x_{m',i,n}|u_n)$ and $\overline{y}$ has the probability $P_{1,N}(\overline{y}|\overline{x}_{m,i}) = \prod_{n=1}^{N} P_1(y_n|x_{m,i,n})$. The coding theorem (theorem 5.6.1, Gallager (1968)) again applies to this situation, yielding, for any ρ, $0 \leq \rho \leq 1$,

$$P_{e,11}(\overline{u}) \leq (M-1)^{\rho} \sum_{\overline{y}} \Big[\sum_{\overline{x}} Pr(\overline{x}|\overline{u}) P_{1,N}(\overline{y}|\overline{x})^{\frac{1}{1+\rho}}\Big]^{1+\rho} \tag{36}$$

$$= (M-1)^{\rho} \prod_{n=1}^{N} \sum_{j} \Big(\sum_{k} Q_2(k|u_n) P_1^{\frac{1}{1+\rho}}(j|k)\Big)^{1+\rho} \tag{37}$$

Next, $P_{e,11}$ is the expected value of $P_{e,11}(\overline{u})$ over m, i, and $\overline{u}_i=\overline{u}$. Since our bound is independent of m, i , we average only over $\overline{u}$.

$$P_{e,11} = \sum Q_{1,N}(\overline{u})P_{e,11}(\overline{u}) \tag{38}$$

$$\leq (M_1-1)^{\rho}\{\sum_i Q_1(i) \sum_j [\sum_k Q_2(k|i)P_1^{\frac{1}{1+\rho}}(j|k)]^{1+\rho}\}^N \tag{39}$$

This can be rewritten in the form

$$P_{e,11} \leq \exp - NE_{11}(R_1) \tag{40}$$

$$E_{11}(R_1) = \max_{0\leq\rho\leq 1} E_{0,11}(\rho) - \rho R_1 \tag{41}$$

$$E_{0,11}(\rho) = - \ln \sum_i Q_1(i) \sum_j [\sum_k Q_2(k|i)P_1^{\frac{1}{1+\rho}}(j|k)]^{1+\rho} \tag{42}$$

$$R_1 = \frac{\ln M_1}{N} \tag{43}$$

Finally $E_{0,11}(\rho)$ is convex $\cap$ in ρ (see appendix 5B, Gallager) and

$$\left.\frac{\partial E_{0,11}(\rho)}{\partial\rho}\right|_{\rho=0} = I(X;Y|U) \tag{44}$$

It follows as in the ordinary coding theorem that $E_{11}(R_1) > 0$ for $R_1 < I(X;Y|U)$.

We have now established, for any given Q_1 and Q_2, that if $R_1 < I(X;Y|U)$ and $R_2 < I(U;Z)$, then both $P_{e,1}$ and $P_{e,2}$ are bounded by quantities that approach 0 exponentially with block length. It then follows from the definition of $C_2^*(C_1)$ in (8), that if $R_1 < C_1$ and $R_2 < C_2^*(C_1)$ for some C_1, then $P_{e,1}$ and $P_{e,2}$ go to 0 exponentially with N. It is still an open conjecture, however, whether $C_2^*(C_1) = C_2(C_1)$ (where $C_2(C_1)$ is the convex hull of $C_2^*(C_1)$). Assume that for some C_1, $C_2^*(C_1) < C_2(C_1)$. Since $C_2(C_1)$ is the convex hull of $C_2^*(C_1)$ we can find values $C_1^{(a)} < C_1 < C_1^{(b)}$ and λ, $0<\lambda<1$ such that

$$C_1 = \lambda C_1^{(a)} + (1-\lambda)C_1^{(b)} \tag{45}$$

$$C_2(C_1) = \lambda C_2^*(C_1^{(a)}) + (1-\lambda)C_2^*(C_1^{(b)}) \tag{46}$$

Let $Q_1^{(a)}$, $Q_2^{(a)}$ and $Q_1^{(b)}$, $Q_2^{(b)}$ be the probability assignments that maximize (8) for $C_1^{(a)}$ and $C_1^{(b)}$ respectively. Consider an ensemble of codes in which the first $\lfloor N\lambda \rfloor$ letters in each code word and cluster center are chosen according to $Q_1^{(a)}, Q_2^{(a)}$ and the rest of the letters are chosen according to $Q_1^{(b)}$, $Q_2^{(b)}$. Reviewing our previous derivations we see that for i=2, 11, 12

$$E_{0,i}(\rho) = \frac{\lfloor N\lambda \rfloor}{N} E_{0,i}^{(a)}(\rho) + (1 - \frac{\lfloor N\lambda \rfloor}{N})E_{0,i}^{(b)}(\rho) \tag{47}$$

where the superscripts a and b refer to the assignments Q_1 and Q_2 used. There is an awkward diophantine dependence on N in (47), but this vanishes as N increases. In this limit, $C_2(C_1)$ is the derivative of $E_{0,2}(\rho)$ at $\rho=0$ and lower bounds the derivative of $E_{0,12}(\rho)$ at $\rho=0$. Similarly C_1 is the derivative of $E_{0,11}(\rho)$ at $\rho=0$. From this and the convexity of $E_{0,i}(\rho)$ for $i=2,11,12$ we see that the following theorem has been proven.

Theorem 2: For arbitrary N, let sources 1 and 2 have alphabet sizes $\lfloor e^{R_1N} \rfloor$ and $\lfloor e^{R_2N} \rfloor$ respectively for fixed R_1, R_2. If for some C_1, we have $R_1<C_1$ and $R_2<C_2(C_1)$, then block codes of block length N exist for each N such that $P_{e,1}$ and $P_{e,2}$ approach zero exponentially in N.

We have exhibited the exponential dependence for the case where $C_2(C_1) = C_2^*(C_1)$ and outlined how to find it for the hypothetical case where $C_2(C_1)\neq C_2^*(C_1)$. We could also consider the problem of optimizing Q_1 and Q_2 to get the largest exponents, but that seems to be a little rococo. The important point is that Theorems 1 and 2 give meaning to the definition of $C_2(C_1)$ as the capacity function of a degraded broadcast channel (actually theroem 1 plus Bergman's earlier work do this), and further, theorem 2 shows that degraded broadcast channels are more similar to single channels than was formerly appreciated.

Looking at the proofs of theorems 1 and 2, we see that the discrete amplitude assumption was not used (except in limiting the alphabet size in the ensemble U). Thus our results carry over to continuous amplitude channels, and, from them to simple continuous channels such as additive Gaussian noise channels.

Next consider the problem of sending some common information to both decoders plus private information to each. We see that if R_1 is the private information to be sent to decoder 1, then $C_2(R_1)$ is a limit on the private information to decoder 2 plus the broadcast information to both. Since decoder 1 decodes everything sent to decoder 2 anyway, we can come arbitrarily close to these rates with error probabilities vanishing exponentially with block length. Also, observing the way that R_1 and R_2 were defined in theorem 1, we see that we can do no better than this.

Finally, we have defined error probability here as an average over all message pairs and the question arises whether we can achieve a small error probability for every message pair. The answer is yes, and the reader is referred to Bergmans (1973) for a way to bound maximum error probability in terms of average error probability.

Appendix

Proof of Lemma 1

First observe that

$$\max_{Q_1,Q_2} \; I(U;Z) + \lambda I(X;Y|U) \tag{A1}$$

is nondecreasing with the number of sample points in U. If U is an ensemble with M sample points and U' an ensemble with M'>M sample points, then we can map M of the M' points in U' onto the M points of U, assign zero probability to the remaining points, and have identical information measures. With this observation, we see that (10) implies (9). Now assume that U has M sample points, M>min(K,J,L) and assume that Q_1^*, Q_2^* achieve the maximum in (A1). We shall demonstrate the existence of a probability assignment Q_1' such that $Q_1'(i)$ is non zero for at most min(K,J,L) values of i and such that Q_1', Q_2^* also achieves the maximum in (A1). This will complete the proof, since the values of i for which $Q_1'(i)=0$ can be dropped from the set of sample points for U without changing (A1).

For an arbitrary Q_1 and the given Q_2^*, we have

$$I(U;Z) + \lambda I(X;Y|U) = I(U;Z) + \lambda \sum_{i=0}^{M-1} Q_1(i)f(i) \tag{A2}$$

where

$$f(i) = \sum_{k,j} Q_2^*(k|i)P_1(j|k)\ln \frac{P_1(j|k)}{\sum_{k'} Q_2^*(k'|i)P_1(j|k')} \tag{A3}$$

Note that f(i) is independent of Q_1 and therefore the expression in (A2) is convex $\cap$ in Q_1. Necessary and sufficient conditions for maximizing such a function over probability assignments Q_1 are well known (see, for example, Gallager (1968), Theorem 4.4.1) and are given by

$$\sum_{\ell} P_3(\ell|i)\ln \frac{P_3(\ell|i)}{\omega(\ell)} + \lambda f(i) \leq C \tag{A4}$$

for all i with equality for i such that $Q_1(i)>0$. In the above equation,

$$\omega(\ell) = \sum_{i} Q_1(i)P_3(\ell|i) \tag{A5}$$

$$P_3(\ell|i) = \sum_{k,j} Q_2^*(k|i)P_1(j|k)P_2(\ell|j) \tag{A6}$$

and C is whatever constant is necessary to satisfy the equation. By its definition Q_1^* must satisfy (A4), but it then follows that any other choice of Q_1 which gives rise to the same function $\omega(\ell)$ in (A5) must also satisfy (A4), and thus yield a maximum of (A1). For this fixed $\omega(\ell)$, $0\leq\ell\leq L-1$, (A5) is a set of L linear equations in the M unknowns $Q_1(0),\ldots Q_1(M-1)$. There is at least one solution, Q_1^*, and thus if M>L, there is an M-L dimensional subspace of solutions (all of which satisfy $\sum Q_1(i)=1$). Thus if there are more than L values of i for which $Q_1^*(i)>0$, there are other solutions in which the set of values of i for which $Q_1(i)>0$ is deminished to size L.

Next, we rewrite the equation (A5) and (A6) as follows:

$$\omega(\ell) = \sum_j p(j)P_2(\ell|j) \tag{A7}$$

$$p(j) = \sum_i Q_1(i) \sum_k Q_2^*(k|i)P_1(j|k) \tag{A8}$$

Again Q_1^* satisfies (A7) and (A8), as well as any other Q_1 that leads to the same p(j), $0 \leq j \leq J-1$. By the same argument as before, if there are more than J values of i for which $Q_1^*(i) > 0$, then there is another solution with this set reduced to size J. In the same way there is a solution with the set reduced to size K by considering (A5) and (A6) rewritten as

$$\omega(\ell) = \sum_{k,j} r(k)P_1(j|k)P_2(\ell|j) \tag{A9}$$

$$r(k) = \sum_i Q_1(i)Q_2^*(k|i) \tag{A10}$$

This completes the proof.

Proof of Lemma 2

The proof relies heavily on standard inequalities between entropies and average mutual informations. The reader is referred to chapter 2 of Gallager (1968) for derivations of these inequalities using the same notation as employed here.

$$I(W_2;Z^N) = H(Z^N) - H(Z^N|W_2) \tag{A11}$$

$$H(Z^N|W_2) = H(Z_1,Z_2,\ldots Z_N|W_2) \tag{A12}$$

$$= \sum_{n=1}^{N} H(Z_n|W_2,Z_1,\ldots,Z_{n-1}) \tag{A13}$$

$$\geq \sum_{n=1}^{N} H(Z_n|W_2,Y_1,\ldots,Y_{n-1},Z_1,\ldots Z_{n-1}) \tag{A14}$$

$$= \sum_{n=1}^{N} H(Z_n|W_2Y_1,\ldots,Y_{n-1}) \tag{A15}$$

$$H(Z^N) = H(Z_1,\ldots Z_N) \leq \sum_{n=1}^{N} H(Z_n) \tag{A16}$$

Equation (A12) comes from $Z^N=Z_1,\ldots Z_N$; that is the ensemble of sequences of N output symbols is the same as the joint ensemble of each of the N outputs. Eqs. (A13) and (A14) are standard relations (see 2.3.10 and 2.3.13 of Gallager (1968)). Eq. (A15) arises from the fact that Z_n and $Z_1,\ldots,Z_{n-1}$ are statistically independent conditional on $Y_1,\ldots,Y_{n-1}$, as can be seen from (13) and (2). This in fact is the crucial place where the assumption of a degraded broadcast channel is used. Substituting (A15) and (A16) into (A11), we get

$$I(W_2;Z^N) \leq \sum_{n=1}^{N} I(W_2,Y_1,\ldots,Y_{n-1};Z_n) \tag{A17}$$

$$= \sum_{n=1}^{N} I(U_n;Z_n) \tag{A18}$$

where

$$U_n = W_2,Y_1,\ldots,Y_{n-1} \tag{A19}$$

Observe from (13), (1), and (2) that conditional on X_n, U_n is statistically independent of Y_n and Z_n, and therefore $Pr(u_n,x_n,y_n,z_n)$ satisfies (3) for some choice of $Q_1(u_n)$ and $Q_2(x_n|u_n)$. This choice depends on n, of course, but that will not be of any concern.

Next consider $I(W_1;Y^N|W_2)$.

$$I(W_1;Y^N|W_2) = \sum_{n=1}^{N} I(W_1;Y_n|W_2Y_1,\ldots,Y_{n-1}) \tag{A20}$$

$$= \sum_{n=1}^{N} I(W_1;Y_n|U_n) \tag{A21}$$

We have used the chain rule in (A20) and the definition of U_n in (A21).

$$I(W_1;Y_n|U_n) = H(Y_n|U_n) - H(Y_n|U_nW_1) \tag{A22}$$

$$\leq H(Y_n|U_n) - H(Y_n|U_nX_nW_1) \tag{A23}$$

$$= H(Y_n|U_n) - H(Y_n|U_nX_n) \tag{A24}$$

$$= I(X_n;Y_n|U_n) \tag{A25}$$

In (A24) we have used the fact that conditional on X_n, Y_n is statistically independent of U_nW_1. Combining (A25) and (A21) we have

$$I(W_1;Y^N|W_2) \leq \sum_{n=1}^{N} I(X_n;Y_n|U_n) \tag{A26}$$

Combining (A18) and (A26), we have

$$I(W_2;Z^N) + \lambda I(W_1;Y^N|W_2) \leq \sum_{n=1}^{N} [I(U_n;Z_n) + \lambda I(X_n;Y_n|U_n)] \tag{A27}$$

Finally, using lemma 1,

$$I(U_n;Z_n) + \lambda I(X_n;Y_n|U_n) \leq C(\lambda) \tag{A28}$$

and substituting (A28) into (A27) completes the proof.

References

1. T.M. Cover, "Broadcast Channels", IEEE Trans. IT, Jan. 1972, pp. 2-13.

2. P.P. Bergmans, "Random Coding Theorem for Broadcast Channels with Degraded Components", IEEE Trans IT, March 1973, pp. 197-207.

3. A. Wyner & J. Ziv, "A Theorem on the Entropy of Certain Binary Sequences & Applications, Part I", IEEE Trans. I.T., to be published.

4. A. Wyner & J. Ziv, "A Theorem on the Entropy of Certain Binary Sequences & Applications, Part II", IEEE Trans. I.T., to be published.

5. R.G. Gallager, Information Theory and Reliable Communication, J.Wiley & Sons, 1968.

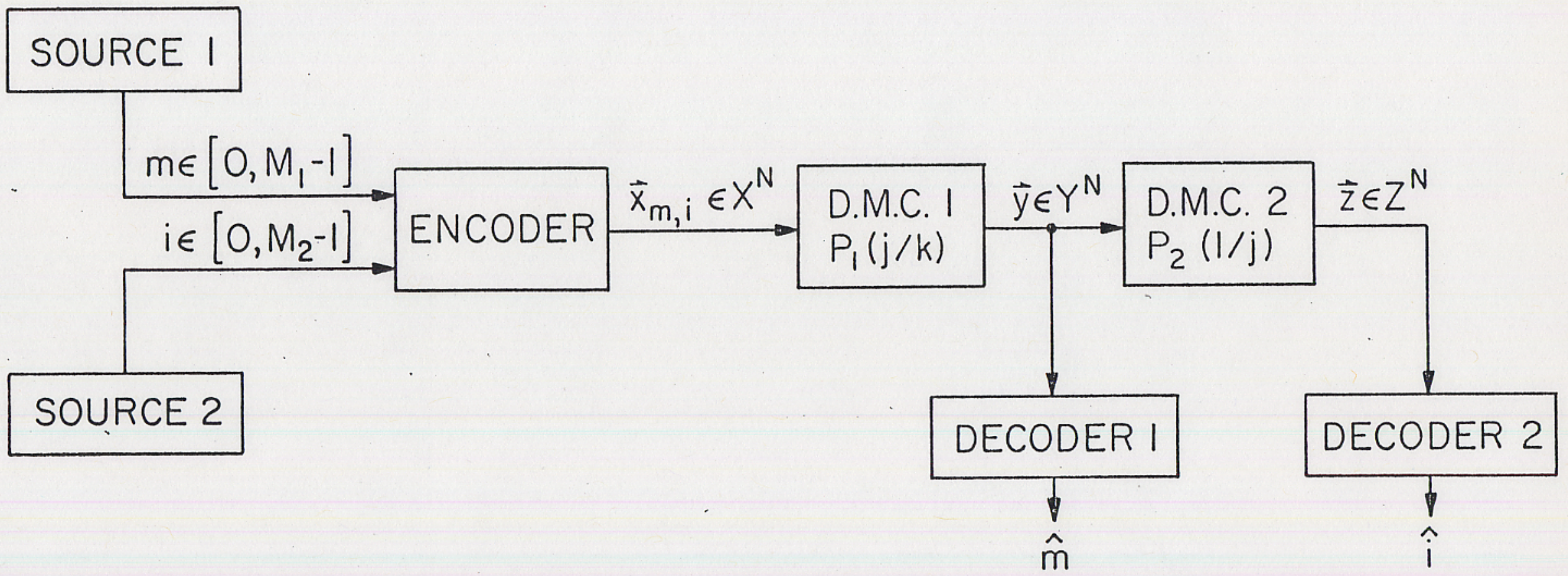


Figure 1

Model of Degraded Broadcast Channel

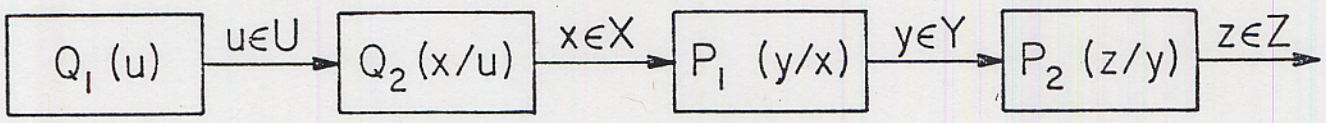


Figure 2

Model for Information Measures used in Defining Capacity

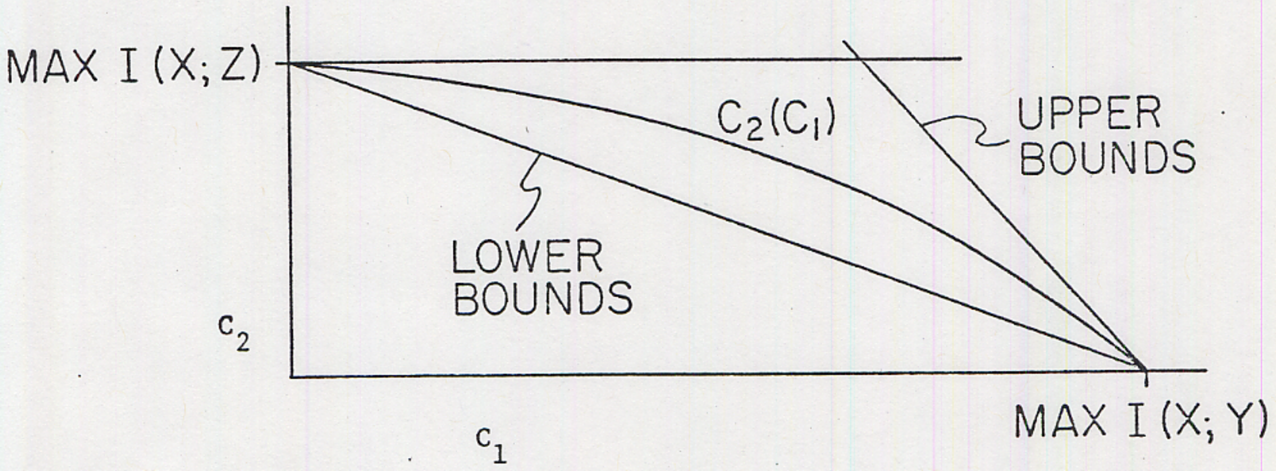


Figure 3

Typical Shape of Capacity Region