C. E. Shannon

Channels with Side Information at the Transmitter

Abstract: In certain communication systems where information is to be transmitted from one point to another, additional side information is available at the transmitting point. This side information relates to the state of the transmission channel and can be used to aid in the coding and transmission of information. In this paper a type of channel with side information is studied and its capacity determined.

Introduction

Channels with feedback[1] from the receiving to the transmitting point are a special case of a situation in which there is additional information available at the transmitter which may be used as an aid in the forward transmission system. In Fig. 1 the channel has an input x and an output y.

There is a second output from the channel, u, available at the transmitting point, which may be used in the coding process. Thus the encoder has as inputs the message to be transmitted, m, and the side information u. The sequence of input letters x to the channel will be a function of the available part (that is, the past up to the current time) of these signals.

The signal u might be the received signal y, it might be a noisy version of this signal, or it might not relate to y but be statistically correlated with the general state of the channel. As a practical example, a transmitting station might have available a receiver for testing the current noise conditions at different frequencies. These results would be used to choose the frequency for transmission.

A simple discrete channel with side information is shown in Fig. 2. In this channel, x, y and u are all binary variables; they can be either zero or one. The channel can be used once each second. Immediately after it is used the random device chooses a zero or one independently of previous choices and with probabilities 1/2, 1/2. This value of u then appears at the transmitting point. The next x that is sent is added in the channel modulo 2 to this value of u to give the received y. If the side information u were *not* available at the transmitter, the channel would be that of Fig. 3, a channel in which input 0 has probabilities 1/2 of being received as 0 and 1/2 as 1 and similarly for input 1.

Such a channel has capacity zero. However, with the side information *available*, it is possible to send one bit per second through the channel. The u information is used to compensate for the noise inside by a preliminary reversal of zero and one, as in Fig. 4.

Figure 1

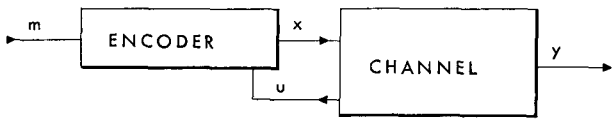


Figure 2

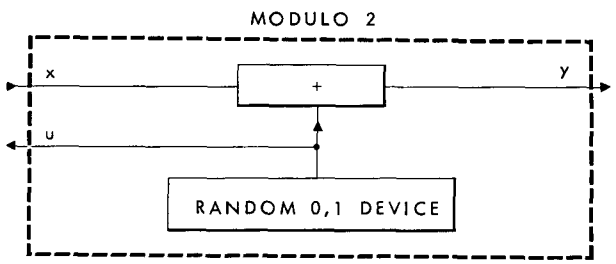


Figure 3

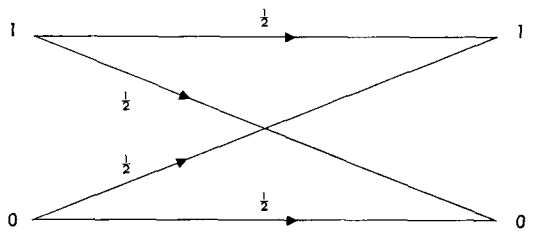
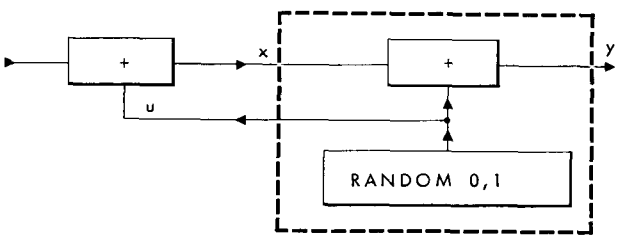


Figure 4



Without studying the problem of side information in its fullest generality, which would involve possible historical effects in the channel, possibly infinite input and output alphabets, et cetera, we shall consider a moderately general case for which a simple solution has been found.

The memoryless discrete channel with side state information

Consider a channel which has a finite number of possible states, $s_1, s_2, \ldots, s_h$. At each use of the channel a new state is chosen, probability g_t for state s_t. This choice is statistically independent of previous states and previous input or output letters in the channel. The state is available as side information u at the transmitting point. When in state s_t the channel acts like a particular discrete channel K_t. Thus, its operation is defined by a set of transition probabilities $P_{ti}(j)$, $t=1, 2, \ldots, h$, $i=1, 2, \ldots, a$, $j=1, 2, \ldots, b$, where a is the number of input letters and b the number of output letters. Thus, abstractly, the channel is described by the set of state probabilities g_t and transition probabilities $p_{ti}(j)$, with g_t the probability of state t and $p_{ti}(j)$ the conditional probability, if in state t and i is transmitted, that j will be received.

A *block code* with M messages (the integers $1, 2, \ldots, M$) may be defined as follows for such a channel with side information. This definition, incidentally, is analogous to that for a channel with feedback given previously.[1] If n is the block length of the code, there are n functions $f_1(m; u_1)$, $f_2(m; u_1, u_2)$, $f_3(m; u_1, u_2, u_3), \ldots, f_n(m; u_1, u_2, \ldots, u_n)$. In these functions m ranges over the set of possible messages. Thus $m=1, 2, \ldots, M$. The u_i all range over the possible side information alphabet. In the particular case here each u_i can take values from 1 to g. Each function f_i takes values in the alphabet of input letters x of the channel. The value $f_i(m; u_1, u_2, \ldots, u_i)$ is the input x_i to be used in the code if the message is m and the side information up to the time corresponding to i consisted of $u_1, u_2, \ldots, u_i$. This is the mathematical equivalent of saying that a code consists of a way of determining, for each message m and each history of side information from the beginning of the block up to the present, the next transmitted letter. The important feature here is that only the data available at the time i, namely $m; u_1, u_2, \ldots, u_i$, may be used in deciding the next transmitted letter x_i, not the side information $u_{i+1}, \ldots, u_n$ yet to appear.

A decoding system for such a code consists of a mapping or function $h(y_1, y_2, \ldots, y_n)$ of received blocks of length n into messages m; thus h takes values from 1 to M. It is a way of deciding on a transmitted message given a complete received block $y_1, y_2, \ldots, y_n$.

For a given set of probabilities of the messages, and for a given channel and coding and decoding system, there will exist a calculable probability of error P_e; the probability of a message being encoded and received in such a way that the function h leads to deciding on a different message. We shall be concerned particularly with cases where the messages are equiprobable, each having probability $1/M$. The *rate* for such a code is $(1/n) \log M$. We are interested in the channel capacity C, that is, the largest rate R such that it is possible to construct codes arbitrarily close to rate R and with probability of error P_e arbitrarily small.

It may be noted that if the state information were *not* available at the transmitting point, the channel would act like a memoryless channel with transition probabilities given by

$$p'_i(j) = \sum_t g_t p_{ti}(j).$$

Thus, the capacity C_1 under this condition could be calculated by the ordinary means for memoryless channels. On the other hand, if the state information were available *both* at transmitting and receiving points, it is easily shown that the capacity is then given by $C_2 = \sum_t g_t C_t$, where C_t is the capacity of the memoryless channel with transmission probabilities $p_{ti}(j)$. The situation we are interested in here is intermediate—the state information is available at the transmitting point but not at the receiving point.

Theorem: *The capacity of a memoryless discrete channel K with side state information, defined by g_t and $p_{ti}(j)$, is equal to the capacity of the memoryless channel K' (without side information) with the same output alphabet and an input alphabet with a^h input letters $X=(x_1, x_2, \ldots, x_h)$ where each $x_i=1, 2, \ldots, a$. The transition probabilities $r_X(y)$ for the channel K' are given by*

$$r_X(y) = r_{x_1, x_2, \ldots, x_h}(y) = \sum g_t p_{tx_t}(y).$$

Any code and decoding system for K' can be translated into an equivalent code and decoding system for K with the same probability of error. Any code for K has an equivocation of message (conditional entropy per letter of the message given the received sequence) at least $R-C$, where C is the capacity of K'. Any code with rate $R>C$ has a probability of error bounded away from zero (whatever the block length n)

$$P_e \geq \frac{R-C}{6\left(R + \frac{1}{n}\ln\frac{R}{R-C}\right)}.$$

It may be noted that this theorem reduces the analysis of the given channel K *with* side information to that for a memoryless channel K' with more input letters but *without* side information. One uses known methods to determine the capacity of this derived channel K' and this gives the capacity of the original channel. Furthermore, codes for the derived channel may be translated into codes for the original channel with identical probability of error. (Indeed, all statistical properties of the codes are identical.)

We first show how codes for K' may be translated into codes for K. A code word for the derived channel K' consists of a sequence of n letters X from the X input alphabet of K'. A particular input *letter* X of this channel may be recognized as a particular *function* from the state alphabet to the input alphabet x of channel K. The full

possible alphabet of X consists of the full set of a^h different possible functions from the state alphabet with h values to the input value with a values. Thus, each letter $X=(x_1, x_2, \ldots, x_h)$ of a code word for K' may be interpreted as a function from state u to input alphabet x. The translation of codes consists merely of using the input x given by this function of the state variable. Thus if the state variable u has the value 1, then x_1 is used in channel K; if it were state k, then x_k. In other words, the translation is a simple letter-by-letter translation without memory effects depending on previous states.

The codes for K' are really just another way of describing certain of the codes for K—namely those where the next input letter x is a function only of the message m and the current state u, and does not depend on the previous states.

It might be pointed out also that a simple physical device could be constructed which, placed ahead of the channel K, makes it look like K'. This device would have the X alphabet for one input and the state alphabet for another (this input connected to the u line of Fig. 1). Its output would range over the x alphabet and be connected to the x line of Fig. 1. Its operation would be to give an x output corresponding to the X function of the state u. It is clear that the statistical situations for K and K' with the translated code are identical. The probability of an input word for K' being received as a particular output word is the same as that for the corresponding operation with K. This gives the first part of the theorem.

To prove the second part of the theorem, we will show that in the original channel K, the change in conditional entropy (equivocation) of the message m at the receiving point when a letter is received cannot exceed C (the capacity of the derived channel K'). In Fig. 1, we let m be the message; x, y, u be the next input letter, output letter and state letter. Let U be the past sequence of u states from the beginning of the block code to the present (just before u), and Y the past sequence of output letters up to the current y. We are assuming here a given block code for encoding messages. The messages are chosen from a set with certain probabilities (not necessarily equal). Given the statistics of the message source, the coding system, and the statistics of the channel, these various entities m, x, y, U, Y all belong to a probability space and the various probabilities involved in the following calculation are meaningful. Thus the equivocation of message when Y has been received, $H(m|Y)$, is given by

$$H(m|Y) = -\sum_{m,Y} P(m, Y) \log P(m|Y)$$

$$= -E\left(\log P(m|Y)\right).$$

(The symbol $E(G)$ here and later means the expectation or average of G over the probability space.) The *change* in equivocation when the next letter y is received is

$$H(m|Y) - H(m|Y, y) = -E\left(\log P(m|Y)\right) + E\left(\log P(m|Y, y)\right)$$

$$= E\left(\log \frac{P(m|Y, y)}{P(m|Y)}\right)$$

$$= E\left(\log \frac{P(m, Y, y)P(Y)}{P(Y, y)P(m, Y)}\right)$$

$$= E\left(\log \frac{P(y|m, Y)P(Y)}{P(Y, y)}\right)$$

$$= E\left(\log \frac{P(y|m, Y)}{P(y)}\right) - E\left(\log \frac{P(Y, y)}{P(Y)P(y)}\right)$$

$$H(m|Y) - H(m|Y, y) \leq E\left(\log \frac{P(y|m, Y)}{P(y)}\right). \qquad (1)$$

The last reduction is true since the term $E\left(\log \frac{P(Y, y)}{P(Y)P(y)}\right)$ is an average mutual information and therefore non-negative. Now note that by the independence requirements of our original system

$$P(y|x) = P(y|x, m, u, U) = P(y|x, m, u, U, Y).$$

Now since x is a strict function of m, u, and U (by the coding system function) we may omit this in the conditioning variables

$$P(y|m, u, U) = P(y|m, u, U, Y),$$

$$\frac{P(y, m, u, U)}{P(m, u, U)} = \frac{P(y, m, u, U, Y)}{P(m, u, U, Y)}.$$

Since the new state u is independent of the past $P(m, u, U) = P(u)P(m, U)$ and $P(m, u, U, Y) = P(u)P(m, U, Y)$. Substituting and simplifying,

$$P(y, u|m, U) = P(y, u|m, U, Y).$$

Summing on u gives

$$P(y|m, U) = P(y|m, U, Y).$$

Hence:

$$H(y|m, U) = H(y|m, U, Y) \leq H(y|m, Y)$$

$$-E\left(\log P(y|m, U)\right) \leq -E\left(\log P(y|m, Y)\right).$$

Using this in (1),

$$H(m|Y) - H(m|Y, y) \leq E\left(\log \frac{P(y|m, U)}{P(y)}\right). \qquad (2)$$

We now wish to show that $P(y|m, U) = P(y|X)$. Here X is a random variable specifying the *function* from u to x imposed by the encoding operation for the next input x to the channel. Equivalently, X corresponds to an input letter in the derived channel K'. We have $P(y|x, u) = P(y|x, u, m, U)$. Furthermore, the coding system used implies a functional relation for determining the next input letter x, given m, U and u. Thus $x = f(m, U, u)$. If $f(m, U, u) = f(m', U', u)$ for two particular pairs (m, U) and (m', U') but for all u, then it follows that $P(y|m, U, u) = P(y|m', U', u)$ for all u and y; since m, U and u lead to the same x as m', U', and u. From this we

obtain $P(y|m, U) = \sum_u P(u)P(y|m, U, u) = \sum_u P(u)P(y|m', U', u) = P(y|m', U')$. In other words, (m, U) pairs which give the same function $f(m, U, u)$ give the same value of $P(y|m, U)$ or, said another way, $P(y|m, U) = P(y|X)$.

Returning now to our inequality (2), we have

$$H(m|Y) - H(m|Y, y) \leq E\left(\log \frac{P(y|X)}{P(y)}\right) \leq \max_{P(X)} E\left(\log \frac{P(y|X)}{P(y)}\right)$$

$$H(m|Y) - H(m|Y, y) \leq C.$$

This is the desired inequality on the equivocation. The equivocation cannot be reduced by more than C, the capacity of the derived channel K', for each received letter. In particular in a block code with M equiprobable messages, $R = 1/n \log M$. If $R > C$, then at the end of the block the equivocation must still be at least $nR - nC$, since it starts at nR and can only reduce at most C for each of the n letters.

It is shown in the Appendix that if the equivocation per letter is at least $R - C$ then the probability of error in decoding is bounded by

$$P_e \geq \frac{R - C}{6\left(R + \frac{1}{n} \ln \frac{R}{R - C}\right)}.$$

Thus the probability of error is bounded away from zero regardless of the block length n, if the code attempts to send at a rate $R > C$. This concludes the proof of the theorem.

As an example of this theorem, consider a channel with two output letters, any number a of input letters and any number h of states. Then the derived channel K' has two output letters and a^h input letters. However, in a channel with just two output letters, only two of the input letters need be used to achieve channel capacity, as shown in (2). Namely, we should use in K' only the two letters with maximum and minimum transition probabilities to one of the output letters. These two may be found as follows. The transition probabilities for a particular letter of K' are averages of the corresponding transitions for a set of letters for K, one for each state. To maximize the transition probability to one of the output letters, it is clear that we should choose in each state the letter with the maximum transition to that output letter. Similarly, to minimize, one chooses in each state the letter with the minimum transition probability to that letter. These two resulting letters in K' are the only ones used, and the corresponding channel gives the desired channel capacity. Formally, then, if the given channel has probabilities $p_{ti}(1)$ in state t for input letter i to output letter 1, and $p_{ti}(2) = 1 - p_{ti}(1)$ to the other output letter 2, we calculate:

$$p_1 = \sum_t g_t \max_i p_{ti}(1),$$

$$p_2 = \sum_t g_t \min_i p_{ti}(1).$$

The channel K' with two input letters having transition probabilities p_1 and $1 - p_1$ and p_2, $1 - p_2$ to the two output letters respectively, has the channel capacity of the original channel K.

Another example, with three output letters, two input letters and three states, is the following. With the states assumed to each have probability 1/3, the probability matrices for the three states are:

State 1			*State 2*			*State 3*		
1	0	0	0	1	0	0	0	1
0	1/2	1/2	1/2	0	1/2	1/2	1/2	0

In this case there are $2^3 = 8$ input letters in the derived channel K'. The matrix of these is as follows:

1/2	1/2	0
0	1/2	1/2
1/2	0	1/2
2/3	1/6	1/6
1/6	2/3	1/6
1/6	1/6	2/3
1/3	1/3	1/3
1/3	1/3	1/3

If there are only three output letters, one need use only three input letters to achieve channel capacity, and in this case it is readily shown that the first three can (and in fact must) be used. Because of the symmetry, these three letters must be used with equal probability and the resulting channel capacity is log (3/2).

In the original channel, it is easily seen that, if the state information were *not* available, the channel would act like one with the transition matrix

1/3	1/3	1/3
1/3	1/3	1/3

This channel clearly has zero capacity. On the other hand, if the state information were available at the *receiving* point or at *both* the receiving point and the transmitting point, the two input letters can be perfectly distinguished and the channel capacity is log 2.

Appendix

Lemma: Suppose there are M possible events with probabilities $p_i (i = 1, 2, \ldots, M)$. Given that the entropy H satisfies

$$H = -\sum p_i \ln p_i \geq \Delta,$$

then the total probability P_e for all possibilities except the most probable satisfies

$$P_e \geq \frac{\Delta}{6 \ln\left(\frac{M \ln M}{\Delta}\right)}.$$

Proof: For a given H, the minimum P_e will occur if all the probabilities except the largest one are equal. This follows from the convexity properties of entropy; equalizing two probabilities increases the entropy. Consequently,

we may assume as the worst case a situation where there are $M-1$ possibilities, each with probability q, and one possibility with probability $1-(M-1)q$. Our given condition is then

$$-(M-1)q\ln q-[1-(M-1)q]\ln[1-(M-1)q]\geqslant\Delta.$$

Since $f(x)=-(1-x)\ln(1-x)$ is concave downward with slope 1 at $x=0$, $(f'(x)=1+\ln(1-x);\ f''(x)=-\frac{1}{1-x}\leqslant 0$ for $0\leqslant x\leqslant 1)$, it follows that $f(x)\leqslant x$ and the second term above is dominated by $(M-1)q$. The given condition then implies

$$-(M-1)q\ln q+(M-1)q\geqslant\Delta$$

or

$$(M-1)q\ln\frac{e}{q}\geqslant\Delta.$$

Now assume in contradiction to the conclusion of the lemma that

$$P_e=(M-1)q<\frac{\Delta}{6\left(\ln M+\ln\frac{\ln M}{\Delta}\right)}.$$

Since $q\ln\frac{e}{q}$ is monotone increasing in q, this would imply that

$$(M-1)q\ln\frac{e}{q}<\frac{\Delta}{6\left(\ln M+\ln\frac{\ln M}{\Delta}\right)}\log\frac{6e(M-1)\left(\ln M+\ln\frac{\ln M}{\Delta}\right)}{\Delta}$$

$$=\frac{\Delta}{6}\left[\frac{\ln\frac{M-1}{\Delta}}{\ln M+\ln\frac{\ln M}{\Delta}}+\frac{\ln 6e}{\ln M+\ln\frac{\ln M}{\Delta}}+\frac{\ln\left(\ln M+\ln\frac{\ln M}{\Delta}\right)}{\ln M+\ln\frac{\ln M}{\Delta}}\right]$$

$$\leqslant\frac{\Delta}{6}\left[1+3+\frac{1}{e}\right]<\Delta\qquad(M>1).$$

The first dominating constant is obtained by writing the corresponding term as $(\ln\ln M-\ln\Delta+\ln(M-1)-\ln\ln M)/(\ln\ln M-\ln\Delta+\ln M)$. Since $\ln M\geqslant\Delta$, this is easily seen to be dominated by 1 for $M\geqslant 2$. (For $M=1$, the lemma is trivially true since then $\Delta=0$.) The term dominated by 3 is obvious. The last term is of the form $\ln\mathcal{Z}/\mathcal{Z}$. By differentiation we find this takes its maximum at $\mathcal{Z}=e$ and the maximum is $1/e$. Since our conclusion contradicts the hypothesis of the lemma, we have proved the desired result.

The chief application of this lemma is in placing a lower bound on probability of error in coding systems. If it is known that in a certain situation the "equivocation," that is, the conditional entropy of the message given a received signal, exceeds Δ, the lemma leads to a lower bound on the probability of error. Actually, the equivocation is an average over a set of received signals. Thus, the $\Delta=\sum P_i\Delta_i$ where P_i is the probability of receiving signal i and Δ_i is the corresponding entropy of message. If $f(\Delta)$ is the lower bound in the lemma, that is,

$$f(\Delta)=\frac{\Delta}{6\ln\left(\frac{M\ln M}{\Delta}\right)},$$

then the lower bound on P_e would be $P_e\geqslant\sum P_i f(\Delta_i)$. Now the function $f(\Delta)$ is convex downward (its second derivative is non-negative in the possible range). Consequently $\sum P_i f(\Delta_i)\geqslant f(\sum P_i\Delta_i)=f(\Delta)$ and we conclude that *the bound of the lemma remains valid even in this more general case by merely substituting the averaged value of* Δ.

A common situation for use of this result is in signaling with a code at a rate R greater than channel capacity C. In many types of situation this results in an equivocation of $\Delta=n(R-C)$ after n letters have been sent. In this case we may say that the probability of error for the block sent is bounded by (substituting these values in the lemma)

$$P_e\geqslant\frac{R-C}{6\left(R+\frac{1}{n}\ln\frac{R}{(R-C)}\right)}=\frac{R-C}{6\left(R-\ln\left(1-\frac{C}{R}\right)\right)}.$$

This then is a lower bound on probability of error for rates greater than capacity under these conditions.

References

1. C. E. Shannon, "Zero Error Capacity of a Noisy Channel," *IRE Transactions on Information Theory, 1956 Symposium*, **IT-2,** No. 3.
2. C. E. Shannon, "Geometrische Deutung einiger Ergebnisse bei der Berechnung der Kanalkapazität," *Nachrichtentechnische Zeitschrift*, **10,** Heft 1, January 1957.
3. C. E. Shannon, "Certain Results in Coding Theory for Noisy Channels," *Information and Control*, **1,** No. 1, September 1957.

Revised manuscript received September 15, 1958