Information Channel

Definition : An Information Channel is described by giving an input alphabet $A = a_i; i = 1, 2, 3, \ldots, r$; an output alphabet $B = b_j; j = 1, 2, 3, \ldots, s$; and a set of conditional probabilities $P(b_j/a_i)$ for all i and j .

$P(b_j/a_i)$ is just the probability that the output symbol b_j will be received if the input symbol a_i is sent.

The channel is described in matrix form:

$$\begin{bmatrix} p(b_1/a_1) & p(b_2/a_1) & p(b_3/a_1) & \ldots & p(b_s/a_1) \\ p(b_1/a_2) & p(b_2/a_2) & p(b_3/a_2) & \ldots & p(b_s/a_2) \\ \vdots & & & & \\ p(b_1/a_r) & p(b_2/a_r) & p(b_3/a_r) & \ldots & p(b_s/a_r) \end{bmatrix}$$

Each row of this array corresponds to a fixed input and the terms in this row are just the

probabilities of obtaining the various b_j at the output for a fixed input is sent.

$P_{ij} = P(b_j/a_i)$ notation is used to describe the channel matrix in this text henceforth.

Sum of each row should be 1 i.e. $\sum_i P_{ij} = 1$ for all $i = 1, 2, \ldots r$

Binary Symmetric Channel

A particular channel of great theoretical and practical importance is the binary symmetric channel (BSC) .

The BSC channel has two input symbol (a_1=0 ,a_2=1) and two output symbols (b_1=0,b_2=1).

It is symmetric when the probability of receiving 1 if a 0 is sent is equal to the probability of receiving a 0 if a 1 is sent ;

this probability ,that an error will occur , is p. $p' = 1 - p$.

The channel matrix may be described as

$$P = \begin{bmatrix} p' & p \\ p & p' \end{bmatrix}$$

The 2nd extension of the BSC channel has four input symbols and four output symbols .Then its channel matrix becomes

$$\Pi = \begin{bmatrix} p'P & pP \\ pP & p'P \end{bmatrix}$$

$$= \begin{bmatrix} p'^2 & p'p & pp' & p^2 \\ p'P & p'^2 & p^2 & pp' \\ pp' & p^2 & p'^2 & p'p \\ p^2 & pp' & p'p & p'^2 \end{bmatrix}$$

The above matrix is known as the Kronecker square (or tensor square) of the matrix P.

In general, the channel matrix of the nth extension of a channel is the nth Kronecker power of the original channel matrix.

Probability Relation in Information Channel

From the channel matrix it is clear that there are r ways in which we might receive output symbol $b_j, j = 1, 2, \ldots, s$.

So, $P(b_j) = \sum_{i=1}^{r} P(b_j/a_i)P(a_i)$

According to Bayes law, the conditional probability of an input a_i given that an output b_j has been received, is

$$P(a_i/b_j) = \frac{P(b_j/a_i)P(a_i)}{P(b_j)}$$

$$P(a_i/b_j) = \frac{P(b_j/a_i)}{P(a_i)} \sum_{i=1}^{r} P(b_j/a_i)P(a_i)$$

The probabilities $P(a_i/b_j)$ are called **backward probabilities** and the probabilities $P(b_j/a_i)$ is referred to as **forward probabilities**.

Probability of the joint event (a_i, b_j)

$P(a_i, b_j) = P(b_j/a_i)P(a_i) = P(a_i/b_j)P(b_j)$

A Priori and A Posteriori Entropies

The probabilities $P(a_i)$ are called a priori probabilities of the input symbols.

The probabilities $P(a_i/b_j)$ are called a posteriori probabilities of the input symbols- the probabilities after the reception of b_j.

APriori Entropy $H(A) = \sum_A P(a) log \frac{1}{P(a)}$

APosteriori Entropy $H(A/b_j) = \sum_A P(a/b_j) log \frac{1}{P(a/b_j)}$

Generalization of Shannons First Theorem

According to Shannons first theorem, the entropy of an alphabet may be interpreted as the average number of bits necessary to represent one symbol of that alphabet.

So, the average number of bits necessary to represent a symbol from the input alphabet with a priori statistics is H (A).

And the average number of bits necessary to represent a symbol from the input alphabet with a posteriori statistics is $H(A/b_j)$.

Since the output symbols occur with probabilities $P(b_j)$, it is expected that the average number of bits necessary to represent an input symbol a_i, if we are given an output symbol

i.e. the average codeword length is the average posteriori entropy,

$H(A/B) = \sum_B P(b) H(A/b)$

$$= \sum_j P(b_j) \sum_i P(a_i/b_j) log \frac{1}{P(a_i/b_j)}$$

$$= \sum_j \sum_i P(a_i, b_j) log \frac{1}{P(a_i/b_j)}$$

Similarly, $H(B/A) = \sum_j \sum_i P(a_i, b_j) log \frac{1}{P(b_j/a_i)}$

Joint Entropy

$H(A, B) = \sum_j \sum_i P(a_i, b_j) log \frac{1}{P(a_i, b_j)}$

$H(A, B) = \sum_j \sum_i P(a_i, b_j) log \frac{1}{P(a_i/b_j) P(b_j)}$

$H(A,B) = \sum_j \sum_i P(a_i, b_j) log \frac{1}{P(a_i/b_j)}$

$+ \sum_j \sum_i P(a_i, b_j) log \frac{1}{P(b_j)}$

$H(A,B) = H(A/B) + H(B)$

Similarly, writing $P(a_i, b_j) = P(b_j/a_i)P(a_i)$

$H(A,B) = H(B/A) + H(A)$

Mutual Information

For an information channel with input A and output B, the entropy H (A) of A is the uncertainty about A when B is unknown.

The equivocation H (A/B) is the uncertainty about A when B is known.

Mutual information is the difference between H (A) and H (A/B), i.e. the amount of uncertainty about A resolved by knowing B. $I(A;B) = H(A) - H(A/B)$ which an also be written as $H(B) - H(B/A)$

Properties of Mutual Information

1. $I(A;B) = \sum\sum P(a_i, b_j) log \frac{P(a_i/b_j)}{P(a_i)}$

   $= \sum\sum P(a_i, b_j) log \frac{P(a_i,b_j)}{P(a_i)P(b_j)}$

   This equation is symmetric in the two random variables a_i and b_j.

   Therefore, $I(a_i; b_j) = I(b_j; a_i)$

2. $I(a_i; b_j) \leq I(a_i)$

3. The mutual information $I(a_i; b_j) = 0$ when the input and output symbols are statistically independent.

4. I (A; B) = H (A) +H (B)-H (A, B)

   $I(A;B) = \sum\sum P(a_i, b_j) log \frac{P(a_i,b_j)}{P(ai)P(bj)}$

$$= \sum \sum P(a_i, b_j) log \frac{1}{P(a_i)P(b_j)}$$
$$- \sum \sum P(a_i, b_j) log \frac{1}{P(a_i, b_j)}$$

$$= \sum_i P(a_i) log \frac{1}{P(a_i)} + \sum_j P(b_j) log \frac{1}{P(b_j)}$$
$$- \sum \sum P(a_i, b_j) log \frac{1}{P(a_i, b_j)}$$

$$I(A; B) = H(A) + H(B) - H(A, B)$$

Hence the proof.

Summary of entropy results

From the diagram we obtain different equations which represent channel equivocations, mutual information and joint entropy.

1. Equations for channel equivocation:

   H (A/B) =H (A)-I (A; B)

   H (B/A) =H (B)-I (A; B)

2. Equations for mutual information:

   I (A; B) = H (A) +H (B)-H (A, B)

   = H (A)-H (A/B)

   = H (B)-H (B/A)

3. Equations for joint entropy:

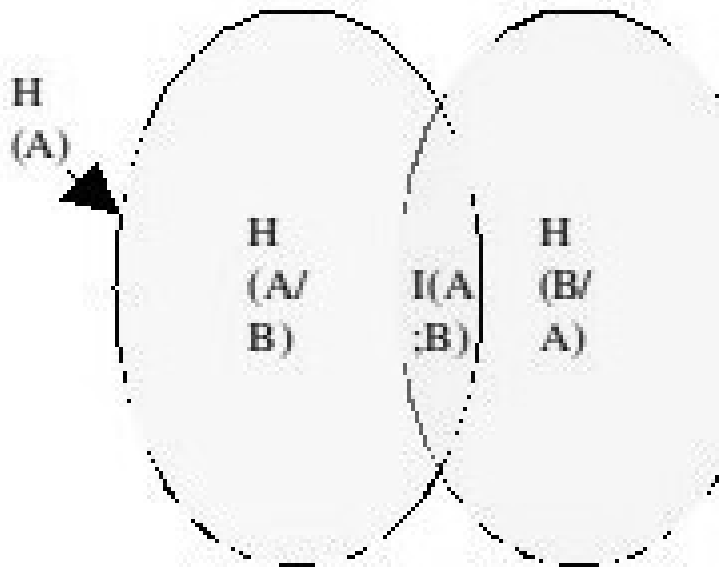
$$H(A, B) = H(A) + H(B) - I(A; B)$$

$$= H(A) + H(B/A)$$

$$= H(B) + H(A/B)$$

H
(A,
B)

(



Special Channels

Noiseless channel: It is specified by the one and only one non zero element in the column of the channel matrix.

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/5 & 3/10 & 1/10 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Deterministic channel: It is specified by the one and only one non zero element in the row of the channel matrix.

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Equivocation of A w.r.t B i.e. H(A/B) =

$$= \sum_j P(b_j) \sum_i P(a_i/b_j) log \frac{1}{P(a_i/b_j)}$$

For noiseless channel

$P(a_i/b_j) = \frac{P(b_j/a_i)P(a_i)}{\sum_i (P(b_j/a_i)P(a_i))}$=0 or 1 due to the column assignment.

So H(A/B)=0.

For noiseless channel I(A;B)=H(A) since H(A/B)=0.

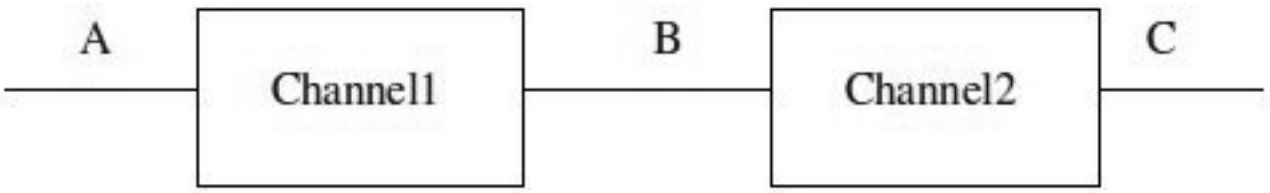
Outputs of a noiseless channel are sufficient by themselves to specify the inputs to the channel. Average number of binits needed to specify the inputs when outputs are known is zero. The uncertainty of the input propagates through the channel.

For deterministic channel $P(b_j/a_i)$=0 or 1.

So H(B/A)=0.

For deterministic channel I(A;B)=H(B)

Cascading of Channels



Suppose channel1 and channel2 are connected in cascade such that

A= $a_1, a_2 \ldots a_r$ and B= $b_1, b_2 \ldots b_s$ be the input and output alphabets for channel1

B= $b_1, b_2 \ldots b_s$ and C= $c_1, c_2 c_t$ be the input and output alphabets for channel2.

$P(c_k/b_j, a_i) = P(c_k/b_j)$

i.e. no dependence on initial symbol in 2nd stage

Similarly, $P(a_i/b_j, c_k) = P(a_i/b_j)$

$$H(A/C) - H(A/B) =$$

$$\sum_{A,C} P(a,c) log \frac{1}{P(a/c)}$$

$$- \sum_{A,B} P(a,b) log \frac{1}{P(a/b)}$$

$$\sum_{A,B,C} P(a,b,c) log \frac{1}{P(a/c)}$$

$$- \sum_{A,B,C} P(a,b,c) log \frac{1}{P(a/b)}$$

$$\text{or, } H(A/C) - H(A/B) = \sum_{A,B,C} P(a,b,c) log \frac{P(a/b)}{P(a/c)}$$

$$\text{or, } H(A/C) - H(A/B) = \sum_{A,B,C} P(a,b,c) log \frac{P(a/b,c)}{P(a/c)}$$

$$= \sum_{B,C} P(b,c)(\sum_{A} P(a/b,c) log \frac{P(a/b,c)}{P(a/c)})$$

or, $H(A/C) - H(A/B) \geq 0$

by taking $x_i = P(a/b,c)$ and $y_i = P(a/c)$

and applying $\sum_{i=1}^{q} x_i log \frac{y_i}{x_i} \leq 0$

Hence, $H(A/C) \geq H(A/B)$ and $I(A;B) \geq I(A;C)$

This means that upon cascading, equivocation increases and channel leaks.

Shannon's theorem for channels

On having recieved b_j, pick for a_i's lengths given by $l_{i,j}$ so that

$$log_r \frac{1}{P(a_i/b_j)} \leq l_{i,j} < log_r \ \frac{1}{P(a_i/b_j)} + 1$$

or, $P(a_i/b_j) \geq \frac{1}{r^{l_{i,j}}} \geq \frac{1}{r} P(a_i/b_j)$

Since $\sum_{i=1}^{q} P(a_i/b_j) = 1$, Kraft inequality is met for j^{th} code.

For each recieved b_j, one instantaneous code exists.

$$H(A/b_j) \leq \sum_{i=1}^{q} P(a_i/b_j) l_{i,j} = L_j < H(A/b_j) + 1$$

upon multiplying by $P(a_i/b_j)$ and summing over i.

L_j is average length of j^{th} code. Averaging using $P(b_j)$,

$H(A/B) = \sum_j p(b_j) H(A/b_j)$

$\leq \sum_i \sum_j p(b_j) P(a_i/b_j) l_{i,j} = L$

or, $H(A/B) \leq L < H(A/B) + 1$

Taking n^{th} extension, we have

$H(A/B) \leq \frac{L_n}{n} < H(A/B) + \frac{1}{n}$

This is Shannon's noiseless theorem of coding extended to a family of instantaneous codes.

Channel Capacity

Consider an information channel with input alphabet A, output alphabet B, and conditional probabilities $P(b_j/a_i)$.

In order to calculate the mutual information it is necessary to know the input symbol probabilities $P(a_i)$.

The mutual information, therefore, depends not only upon the channel, but also upon how we use the channel i.e.,

the probabilities with which we choose the channel inputs.

It is of some interest to examine the variation of $I(A;B)$ as we change the input probabilities.

$$I(A;B) = \sum_{A,B} P(a,b) log \frac{P(a,b)}{P(a)P(b)}$$

For a general information channel, we see that the mutual information can always be made 0 by choosing one of the input symbols with probability 1.

The maximum value of $I(A;B)$ as we vary the input symbol probabilities is called C, the capacity of the channel $C = max[I(A;B)]$

Capacity of an information channel is a function only of the conditional probabilities defining that channel.

It does not depend upon the input probabilitieshow we use the channel.

The capacity of a BSC with error probability p is 1 - H(p) with P(0)=P(1)=0.5.

In certain cases the calculation of the capacity of an information channel can be simplified. The most important class of channels for

which the calculation simplifies is the class of uniform channels.

Uniform channels:

Consider a channel defined by the channel matrix. This channel is said to be uniform if the terms in every row of the channel matrix consist of an arbitrary permutation of the terms in the first row.

We now calculate the capacity of a general uniform channel. The capacity is the maximum of I(A; B) as we vary the input probability distribution.

$$I(A;B) = H(B) - H(B/A)$$

$$= H(B) - \sum_{A} P(a) \sum_{B} P(b/a) log \frac{1}{P(b/a)}$$

Let $\sum_B P(b/a) log \frac{1}{P(b/a)} = W$

No matter which symbol is picked, average is same.

Then above equation becomes

$I(A;B) = H(B) - W \sum P(a)$

For noiseless channel W will be zero.

The summation over B in the last term is a summation, for each a_i, of the terms in the i-th row of the channel matrix.

For a uniform channel, however, this summation is independent of i. Hence,

$I(A;B) = H(B) - W$

The last term above is independent of the input symbol distribution.

To find the maximum of the right side, we need only find the maximum of H(B).

Since the output alphabet consists of r symbols, we know that H(B) cannot exceed $logr$ bits.

H(B) will equal $logr$ bits if and only if all the output symbols occur with equal probability.

In general, it is not true that there exists a distribution over the input symbols such that the output symbols are equiprobable.

For a uniform channel, however, it is easy to check that equiprobable symbols at the input produce equiprobable symbols at the output.

Therefore, the maximum value of RHS, the capacity of the uniform channel, is

$$C = logr + \sum_{B} P(b/a) log P(b/a)$$

Consider n-bit single error detecting binary code.

No of Input symbols= $q = 2^{n-1}$

Let P be the probability of no error

Then Q= 1 P is distortion probability

Then k errors count to $C(n,k)Q^k P^{n-k}$

$$W = \sum_{k=0}^{n} C(n,k) P^{n-k} Q^k log \frac{1}{P^{n-k}Q^k}$$

Simplifying, we get $W = nPlog\frac{1}{P} + nQlog\frac{1}{Q}$

Hence, $I(A;B) = H_2(B) - nH_2(P)$.

Decoding in Noisy Channel

Message rate Message rate is measured in equivalent binary messages per symbol. Sending one of M possible messages using n symbols is equivalent to sending $log\ M$ binary messages in n symbols i.e $log\ M/n$ binary messages per symbol.

Decision rule: Consider a channel with an r-symbol input alphabet $A = a_1, a_2, \ldots a_r$; and an s-symbol output alphabet $B = b_1, b_2, \ldots, a_s$.

A decision rule, $d(b_j)$, is any function specifying a unique input symbol for each output symbol.

Any such rule would involve some error. We calculate the probability of error P_E and hope to minimize it.

This probability may be written as the average of $P(E/b_j)$, the conditional probability of error given that the output of the channel is b_j.

$P_E = \sum_B P(E/b)P(b)$

Equation expresses the error probability as a sum of non-negative terms. Therefore, in order to minimize P_E by choice of a decision rule $d(b_j)$

we may select $d(b_j)$ to minimize each term in the sum separately.

$P(b_j)$ does not depend upon the decision rule we use; so it is equivalent to choose $d(b_j)$ to minimize the conditional probability of error $P(E/b_j)$.

For a fixed decision rule, $d(b_j) = a_i$, $P(E/b_j) = 1 - P[d(b_j)/b_j]$

Where, since our decision rule is fixed, $P[d(b_j)/b_j]$ is the

backward probability $P(a_i/b_j)$.

Finally, in order to minimize for each b_j, we choose $d(b_j) = a^*$

where a^* is defined by $P(a^*/b_j) \geq P(a_i/b_j)$ for all i.

In other words, the channel error probability is minimized if we use that decision rule which chooses for each o/p symbol the corresponding i/p symbol with the highest probability. This decision rule is sometimes called a conditional maximum-likelihood decision rule. The conditional maximum-likelihood decision rule depends upon the priori probabilities $P(a_i)$.

We may use Bayes law to write as

$\frac{p(b_j/a^*)p(a^*)}{p(b_j)} \geq \frac{p(b_j/a_i)p(a_i)}{p(b_j)}$ for all i.

Hence, when the priori probabilities are all equal, the conditional maximum-likelihood decision rule may be written

$d(b_j) = a^*$ where $p(b_j/a^*) \geq P(b_j/a_i)$ for all i

The error probability using any given decision rule is easily obtained as

$P_E = \sum_B P(E/b)P(b)$

$= \sum_B P(b) - \sum_B [P(d(b)/b)]P(b)$

$= 1 - \sum_B P[d(b), b]$

The terms in the summation are just the joint probabilities that a^* is transmitted and b_j is received (for each j).

Hence, defining $\bar{P}_E = 1 - P_E$,

we may write $\bar{P}_E = \sum_B P(a^*, b)$

Since $\sum_{A,B} P(a, b) = 1$ we may write

$P_E = \sum_{B, A-a^*} P(a, b)$

The notation $\sum_{A-a^*}$ is meant to indicate a sum over all members of the A alphabet except $d(b_j) = a^*$.

An alternative way of writing P_E is

$P_E = \sum_{B, A-a^*} P(b/a)P(a)$

If the a priori probabilities P(a) are all equal, then we have

$P_E = \frac{1}{r} \sum_{B, A-a^*} P(b/a)$

The Fano Bound

The error probability has been presented so far without reference to entropy, equivocation, or mutual information. We provide upper and lower bound on the equivocation in terms of error probability.

$$H(P_E) + P_E log(r-1) = P_E log\frac{r-1}{P_E} + \bar{P}_E log\frac{1}{\bar{P}_E}$$

$$= \sum_{B,A-a^*} P(a,b) log\frac{r-1}{P_E} + \sum_B P(a^*,b) log\frac{1}{\bar{P}_E}$$

The equivocation H(A/B) may be written in terms of the same sort of summations:

$$H(A/B) = \sum_{B,A-a^*} P(a,b) log\frac{1}{P(a/b)}$$

$$+ \sum_B P(a^*,b) log\frac{1}{P(a^*/b)}$$

Subtracting yields

$$H(A/B) - H(P_E) - P_E log(r-1)$$

$$= \sum B, A - a^* P(a,b) log \frac{P_E}{(r-1)P(a/b)}$$

$$+ \sum_B P(a^*, b) log \frac{\bar{P}_E}{P(a^*/b)}$$

Now we change the base of the logarithms to get $log\ e(\sum_{B, A-a^*} P(a,b) ln \frac{P_E}{(r-1)P(a/b)}$

$$+ \sum_B P(a*, b) ln \frac{\bar{P}_E}{P(a^*, b)})$$

So that we may use the inequality $ln\ x \leq x - 1$ on each term in the summations.

$$(log\ e)^{-1}(H(A/B) - H(P_E) - P_E log(r-1))$$

$$\leq (\sum_{B,A-a^*} P(a,b)[\frac{P_E}{(r-1)P(a/b)} - 1]$$

$$+ \sum_{B} P(a*,b)[\frac{\bar{P}_E}{P(a^*,b)} - 1])$$

$$= [\frac{P_E}{r-1} \sum_{B,A-a^*} P(b)] - P_E + [\bar{P}_E \sum_B P(b)] - \bar{P}_E$$

Hence, RHS ≤ 0.

And we have the inequality we seek,

$$H(A/B) \leq H(P_E) + P_E log(r-1)$$

$H(P_E)$ bits are needed to specify error and $log\ (r-1)$ bits for $r-1$ remaining bits.

Condition for equality is $x = 1$ i.e. $P(a/b) = \frac{P_E}{r-1}$ for all $b, a \neq a^*$ and $P(a^*/b) = \bar{P}_E$ for all b.

For each b, all input symbols except a^* are equally probable.

The n^{th} extension of a source with r symbols has r^n i/p symbols.

Use only M as valid messages to decrease P_E.

Reliable decoding of unreliable messages

Converse of Shannon's main theorem Shannon says keep $M < 2^{nC}$ to make P_E arbitrarily small. For this M, message rate is $\frac{log\ M}{n} = C$ and channel capacity corresponds to upper limit of error-free message rate.

Try to make $M = 2^{n(C+\epsilon)}$ with $\epsilon -> 0$

For n^{th} extension, $H(A^n)H(A^n/B^n) \leq nC$

Since $p = \frac{1}{M}$, $H(A^n) = \sum \frac{1}{M} log\ M = log\ M = n(C + \epsilon)$

So, rearranging, $n(C+\epsilon)-nC = n\epsilon \leq H(A^n/B^n)$

Applying Fano's result, $n\epsilon \leq H(A^n/B^n) \leq H(P_E)+ P_E log\ (q-1)$

or $n\epsilon \leq 1 + P_E(nC + C\epsilon)$ using $H(P_E) \leq 1$ and $q - 1 < q = M$

or, $P_E \geq \frac{n\epsilon - 1}{nC + n\epsilon} \geq \frac{\epsilon - \frac{1}{n}}{C + \epsilon} \geq \frac{\epsilon}{C}$

As n goes to ∞, P_E is bounded away from zero and does not vanish at all.

This implies that message rate exceeding the channel capacity does not ensure reliable decoding.

Proof of Shannon's main theorem for BSC

Consider a BSC with error probability p, and hence capacity $C = 1 - H(p)$.

Let ϵ be an arbitrarily small positive number,and let $M = 2^{(n*(C-\epsilon))}$.

Then ,for n sufficiently large,it is possible to select a subset of M code words (to represent M equiprobable messages) from the set of 2^n possible inputs to the BSC,such that the probability of error in decoding the channel output can bemade as small as we wish.

In order to send M messages through this channel,we select M of the 2^n possible inputs as codewords. The question we must now answer is, How many messages is it possible to send and still have the message error probability remain small?

If we select the code words so that they cluster together, we can expect a higher probability of error than if we construct a code with the same number of codewords more or less equally spaced from one another. The method of coding will be crucial to the probability of error we obtain and, therefore,to the maximum number

of messages we may use. Let us assume that somehow we have selected a code consisting of M code words of n binits each for use with the BSC.

When one of these code words, say α_0, is sent through the channel, some other binary sequence of length n, say β_j, is received. We know that the maximum likelihood decision rule described earlier will minimize our probability of error.

We have already noted that the average distance between the transmitted sequence α_0 and the received sequence β_j will be np, where n is the order of BSC, and p is the probability of error of the BSC. When we receive a symbol β_j out of our channel, therefore our natural inclination is to hunt for the transmitted code symbol among those code symbols at a distance np or less from β_j.

In geometrical terms we can say that we draw a sphere of radius np about β_j and search for α_0 in this sphere. The quantity np is only the average distance of α_0 from β_j, however, and it might be prudent to enlarge our sphere a bit by ϵ amount, as insurance that α_0 will be inside the sphere with high probability.

Consider a sphere of radius np_ϵ, about β_j, where $p_\epsilon = p + \epsilon$.

Our decision procedure will be to draw a sphere of radius np_ϵ about β_j, and if there is a unique code word inside the sphere we shall decide that this code word was transmitted.

Using this procedure just described ,there are two ways in which an error in decoding a received symbol may occur .

Let $S(np_\epsilon)$ be the sphere. Then the error probability P_E is

$P_E = Pr\alpha_0 \notin S(np_\epsilon)+$

$Pr\alpha_0 \in S(np_\epsilon) \times Pr(atleastanothercodeword \in S(np_\epsilon)$

$P_E \leq Pr\alpha_0 \notin S(np_\epsilon)+Pr(atleastanothercodeword \in S(np_\epsilon)$ since $Pr\alpha_0 \in S(np_\epsilon) \leq 1$

$$P_E \leq Pr\alpha_0 \notin S(np_\epsilon) + \sum_{\alpha_i \neq \alpha_0} Pr\alpha_i \in S(np_\epsilon)$$

Equation expresses a simple bound on the probability of error for a specific set of M code words.

The first term is the probability that the received word and the transmitted word will not be within a Hamming distance of $n(p+\epsilon)$ from each other. The first term is easy to evaluate . It is just the probability that more than $n(p+\epsilon)$ errors were made in the transmission

of n binits through a BSC with probability of error p.

The average number of errors occurring in a block of n binits is np. For any finite value of n, there will be some finite probability that the number of errors will exceed its mean value by $n\epsilon$ or more . As n increases , however ,this becomes less probable.; more precisely the weak law of large numbers tells us that by taking n large enough we may be sure that $Pr\alpha_0 \notin S(np_\epsilon) < \delta$

Then, $P_E \leq \delta + \sum_{\alpha_i \neq \alpha_0} Pr\alpha_i \in S(np_\epsilon)$

The other term is related to the signalling rate involving other $M - 1$ codewords.

Average random code: $n(p + \epsilon) = np_\epsilon = r$ the radius of the sphere.

In all, 2^{nM} possible codes can be selected with probability $\frac{1}{2^{nM}}$. The average probability of error is then

$$\tilde{P}_E \leq \delta + (M-1)\overbrace{P(a \in S(r))} \leq \delta + M\overbrace{P(a \in S(r))}$$

Each of M symbols is selected randomly from 2^n possible words, and the average probability so that $a(\neq a_i)$ is $\overbrace{P(a \in S(r))}$ is $\frac{N(r)}{2^n}$, $N(r)$ being count of codewords inside the sphere.

For binary symmetric channel, $N(r) = 1 + C(n,1) + C(n,2) + \ldots + C(n,r)$

Here, r comes from Hamming distance, upto r terms in n, $N(r) = \sum_{k=0}^{r} C(n,k)$

Since $r = (Q + \epsilon_2)n$ and $Q + \epsilon_2 < 0.5$; $N(r) \leq 2^{nH(\lambda)}$ where $\lambda = Q + \epsilon_2$

Then, $\overbrace{P(a \in S(r))} \leq 2^{-n(1-H(\lambda))}$ with $a \neq a_i$ and $\tilde{P_E} \leq \delta + M2^{-n(1-H(\lambda))}$

Now, for BSC, $1-H(P) = C$ and $H(P) = H(Q)$ so that $1 - H(\lambda) = 1 - H(Q + \epsilon_2)$

$= 1 - H(Q) + H(Q) - H(Q + \epsilon_2)$

$= C - H(Q + \epsilon_2) + H(Q)$

Also, H is convex so that $H(Q+\epsilon_2)) \leq H(Q)+ \epsilon_2 \frac{dH}{dQ}$ and $\frac{dH}{dQ} = log\ \frac{1-Q}{Q} > 0$

Therefore, $1 - H(\lambda) = C - \epsilon_2 log\ (\frac{1-Q}{Q}) = C - \epsilon_3$ bounded by this.

Hence, $\tilde{P_E} \leq \delta + M2^{-n(C-\epsilon_3)}$ Using $M = 2^{n(C-\epsilon_1)}$ we have

$\tilde{P_E} \leq \delta + 2^{-n(\epsilon_1-\epsilon_3)}$

When ϵ_2 is small enough, so that $\epsilon_1 - \epsilon_3 = \epsilon_1 - \epsilon_2 log\,(\frac{1-Q}{Q}) > 0$ the $\tilde{P}_E$ can be made smaller by choosing large n. So, choice of small ϵ_2 ensures reliable transmission for average code.

There exists codes that are arbitrarily reliable and which can signal at rates arbitrarily close to the channel capacity of the BSC using very large messages.

Proof for approximation with entropy

Due to Stirling: $n! \approx n^n e^{-n}\sqrt{2\pi n}$

Take $log\ n! = \sum_{k=1}^{n} log_e k$

Now, $\int_1^n log\ x dx = x log\ x - x = n log\ n - n + 1$

Applying trapezoidal rule, this integration can be approximated to

$\frac{1}{2} log\ 1 + log\ 2 + \ldots + \frac{1}{2} log\ n$

$\geq log\ n! - \frac{1}{2} log\ n$

Hence, $n log\ n - n + 1 \geq log\ n! - \frac{1}{2} log\ n$

Taking antilogs, $n^n e^{-n}\sqrt{n}e \geq n!$

From error due to Trapezoidal rule, we get

$n! \approx n^n e^{-n}\sqrt{2\pi n}$

$C(n,\lambda n) \approx \frac{n^n e^{-n}\sqrt{2\pi n}}{\lambda n^{\lambda n} e^{-\lambda n}\sqrt{2\pi\lambda n}(n-\lambda n)^{(n-\lambda n)} e^{-(n-\lambda n)}\sqrt{2\pi(n-\lambda n)}}$

$C(n,\lambda n) \approx [\frac{n^n}{n^{\lambda n} n^{n-\lambda n}}][\frac{1}{\lambda^{\lambda n}(1-\lambda)^{(1-\lambda)n}}][\frac{e^{-n}}{e^{-\lambda n} e^{-(n-\lambda n)}}][\frac{\sqrt{2\pi n}}{\sqrt{2\pi\lambda n}\sqrt{2\pi n(1-\lambda}}]$

$C(n,\lambda n) \approx [\frac{1}{\lambda^{\lambda n}(1-\lambda)^{(1-\lambda)n}}][\frac{1}{2\pi\lambda(1-\lambda)n}]^{\frac{1}{2}}$

First term $= 2^{nH(\lambda)}$. Second term is $const(\lambda)n^{-\frac{1}{2}}$. Then, $C(n,\lambda n) + C(n,\lambda n - 1) + \ldots + C(n,1) + 1$ is bounded by a GP where the terms are obtained by multiplying $\frac{n}{1}, \frac{n-1}{2}, \frac{n-2}{3}, \ldots, \frac{n-\lambda n+1}{\lambda n}$ or in reverse order the terms are $\frac{\lambda n}{n-\lambda n+1}, \ldots, \frac{3}{n-2}, \frac{2}{n-1}, \frac{1}{n}$.

Upper bound of these terms is $\frac{\lambda}{1-\lambda}$ and upper bound on the sum is $\sum_{m=0}^{\infty}(\frac{\lambda}{1-\lambda})^m$ which is equal to $\frac{1}{1-\frac{\lambda}{1-\lambda}} = \frac{1-\lambda}{1-2\lambda}$

Hence, $\sum_{k=0}^{\lambda n} C(n,k) \leq 2^{nH(\lambda)}[\frac{1}{2\pi\lambda(1-\lambda)}]^{\frac{1}{2}}\frac{1-\lambda}{1-2\lambda}(\frac{1}{n})^{\frac{1}{2}}$

For all n such that $n \geq \frac{1-\lambda}{2\pi\lambda(1-2\lambda)^2}$,

the product of last three terms < 1.

So, $\sum_{k=0}^{\lambda n} C(n,k) \leq 2^{nH(\lambda)}$ for such large n.

Generalization of Shannon's theorem

- Hamming distance metric has to be replaced.
- Counting the number of messages inside sphere is difficult.
- Channel capacity has to be derived from fundamentals.
- If noise is non-white, sphere has to be replaced with oval shape.

Distance can be generalized from maximum likelihood decoding concept.

$logP(\beta_j|\alpha^*) \geq logP(\beta_j|\alpha_i)$

or, $log\frac{1}{P(\beta_j|\alpha^*)} \leq log\frac{1}{P(\beta_j|\alpha_i)}$

The input probabilities can be introduced to give the concept of distance function.

or, $log\frac{P_0(\beta_j)}{P(\beta_j|\alpha^*)} \leq log\frac{P_0(\beta_j)}{P(\beta_j|\alpha_i)}$

This distance can be averaged for a transmitted α_0 to β_j as

$$\sum_{B^n} P(\beta_j|\alpha_0) log\frac{P_0(\beta_j)}{P(\beta_j|\alpha_0)}$$

This sum is $-I(\alpha_0; B^n)$

Now, $I(A^n; B^n) = nC$

$I(A; B) = H(B) - \sum_B P(b|a) log\frac{1}{P(b|a)}$

$I(\alpha_0; B) = \sum P(\beta_j|\alpha_0) log\frac{1}{P_0(\beta_j)} - \sum P(\beta_j|\alpha_0) log\frac{1}{P(\beta_j|\alpha_0)}$

Then, $I(\alpha_0; B) = \sum_{B^n} P(\beta_j|\alpha_0) log \frac{P_0(\beta_j)}{P(\beta_j|\alpha_0)}$

$= -nC$ for each α_0

So, for recd β_j find α_0 such that

$log \frac{P_0(\beta_j)}{P(\beta_j|\alpha_0)} \approx -nC$

Draw a sphere about recd β_j so that

$log \frac{P_0(\beta_j)}{P(\beta_j|\alpha_0)} \leq -nC + n\epsilon$

When there is unique code point inside, there is successful decoding.

Coming back to $\overbrace{Pr(\alpha_i \in S_\epsilon)} = \sum_{B^n, S(\epsilon)} P_0(\beta_j) P_0(\alpha_i)$

This sum is over all pairs of α_i, β_j such that

$log\frac{P_0(\beta_j)}{P(\beta_j|\alpha_i)} \leq -n(C-\epsilon)$

$P_0(\beta_j)P_0(\alpha_i) \leq P(\beta_j|\alpha_i)P_0(\alpha_i)2^{(-n(C-\epsilon))}$ for any such pair.

Then $\sum_{B^n,S(\epsilon)} P_0(\beta_j)P_0(\alpha_i) \leq$

$P(\beta_j|\alpha_i)P_0(\alpha_i)2^{-n(C-\epsilon)}$

Hence, $\tilde{P_E} \leq \delta + M2^{-n(C-\epsilon)}$

So, as long as $M < 2^{n(C-\epsilon)}$ $\tilde{P_E}$ remains arbitrarily small. This leads to Shannon's main theorem.