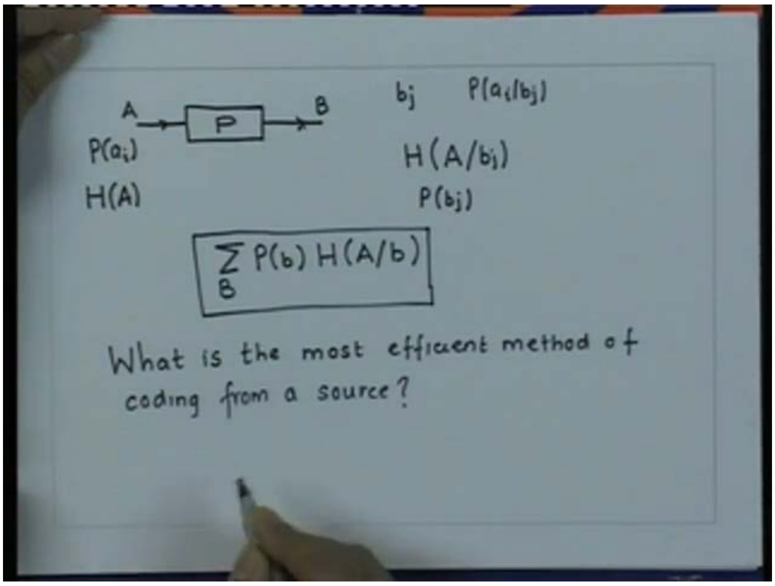
Information Theory and Coding

Prof. S. N. Merchant
Department of Electrical Engineering
Indian Institute of Science, Bombay

Lecture - 21
Equivocation and Mutual Information

Shannon's first theorem tells us the entropy of an alphabet may be interpreted as the average number of binary digits of binits necessary to represent one symbol of that alphabet. Let us extend this interpretation to a priory and a posteriori entropies, which we studied in the earlier class in relationship to information channel.

(Refer Slide Time: 01:20)



So, the information channel was denoted by this figure as shown here. A is the input alphabet. B is the output alphabet. P is the channel matrix. Before reception of the output symbol of the channel, we associate the apriority probabilities P i with the input alphabet a. The average number of binits necessary to represent a symbol from this alphabet is the entropy of the source. That is H A. Now, if you receive a given symbol say b j, we associate the a posteriori probabilities P of a i. Given b j with the input alphabet, the average number of binits necessary to represent a symbol from the alphabet A with this a posteriori statistics is entropy of a given b j.

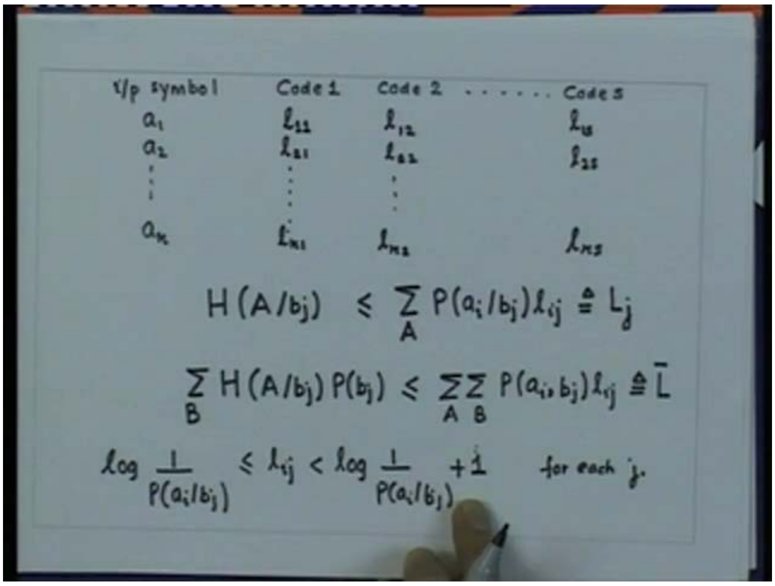
Since, the output symbols occur with probabilities P b j, we might expect that the average number of binits average over b j necessary to represent an input symbol a i. If you are given output signal is the average a posteriori entropy, so what we expect that the average value should be P b H of A b. Now, for the sake of convenience, we will use the short notation for a’s and b’s whenever we are writing in terms of summation. Instead of b j, we will right b and instead of a i, we will just write a.

So, this is the average number of binits required to represent a symbol in input alphabet a given that we have received an output symbol. Now, this result does not follow directly from the Shannon’s first theorem. Shannon’s first theorem deals only with coding for a source with a fixed set of source probabilities and not with coding for a source. It selects a new set of probabilities after each output symbol. So, let us try to generalise Shannon’s first theorem.

In order to obtain such a generalisation, we need to ask ourselves a question. This question is similar to the one, which we asked when we proved Shannon’s first theorem. The question is what most efficient method of coding from a source. In this case, our source is the input alphabet a. Now, this time however, the statistics of the source we wish to quote change from symbol to symbol. Now, the pointer to which set of source statistics we have to use is provided by the output of the channel b j.

Now, since a compact code for 1 set of statistics will not in general be a compact code for another set of source statistics, we take advantage of our knowledge of b j for each transmitted symbol to construct as binary codes as corresponds to the size of the output alphabet. So, we will build s codes, one for each of the possible receives symbols b j. So, when the output of the channel is b j, we use the j th binary code to encode the transmitted symbol a i. So, let the words of our s code be as shown here.

(Refer Slide Time: 06:56)



So, we have input symbols, which are a 1 a 2 up to a r. Depending on what we receive, we will have that many number of codes. Since, the output alphabet is of size s, we will have s codes. Each of these codes will have the code words with the lengths given as follows. So, for the code 1, we have length l 1 1, l 2 1. l 2 1 corresponds to the length of the code word for the input symbol a 2. Similarly, l r 1 corresponds to the code word length for the input symbol a r, when I am using the code 1.

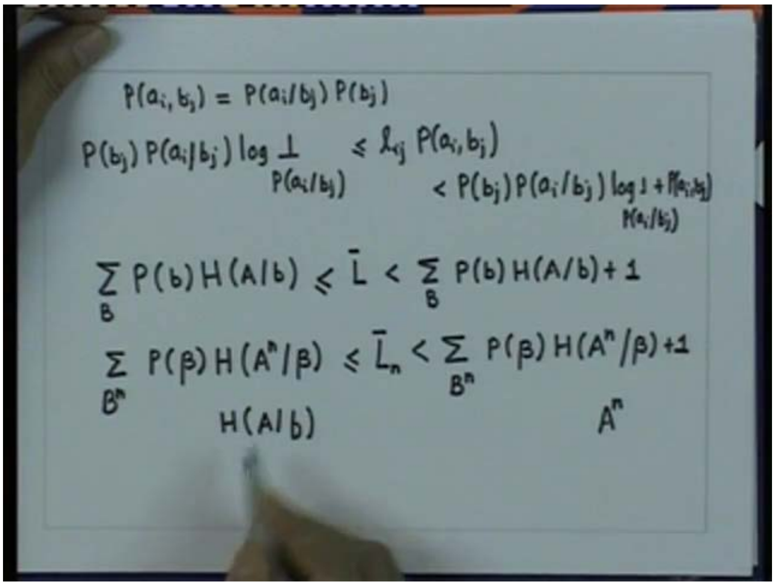
Similarly, for code 2, we have l 1 2, l 2 2 and finally, l r 2. For code s, we will have l 1 s, l 2 s, l r s. So, we have s codes corresponding to the input alphabet. Now, we require that each of these codes to be instantaneous. So, we may apply the Shannon's first theorem, which we had studied earlier to each code separately. We obtained that entropy of a given b j will be always less than or equal to probability of a i given b j l i j average over input alphabet a. This by definition we will call it as l j. This is the average length for the j th code.

Now, here we employ the conditional probabilities P a i b j rather than the marginal probabilities P a i to calculate L j. Since, the j th code is employed only, event b j is the received symbol. So, the average number of binits used for each member of the a alphabet, when we encode in this fashion is obtained by averaging with respect to the receive symbols b j.

So, if we multiply both the sides by probability of b j and sum it over the output alphabet B, we will get the following relationship H of A given b j multiplied by probability b j over B is less than equal to A B probability of a i b j joint probability multiplied by l i j. This by definition, we will call it as the average number of binits per symbol from the a alphabet average with respect to both the input and output symbols.

This relationship is similar to what we had done earlier for a single source. Now, in order to see, in order to show that the bound can be achieved, we next describe a specific coding procedure. So, when b j is the output of a channel, we select an l i j, the word length of the code word corresponding to the input a i as the unique integers. Now, word length defined in this fashion satisfy the craft inequality for each j. The l i j therefore, defines s sets of what length acceptable as the word length of s instantaneous codes.

(Refer Slide Time: 12:46)

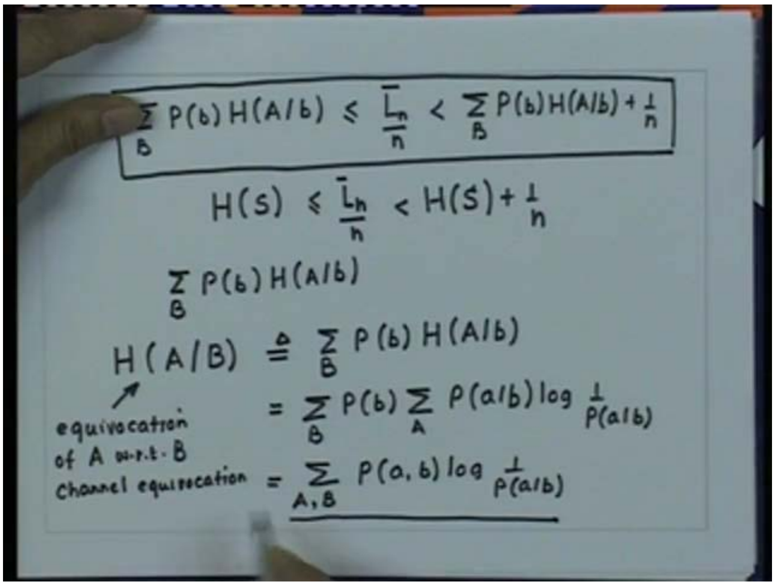


Now, if you multiply this inequality on all the sides by probability of a i b j, which is equal to probability of a i, given b j multiplied by probability b j, then we get a relationship probability. It is b j probability of a i given b j log of probability of a i j given b j less than equal to l i j joint probability of a i b j less than probability of b j multiplied by probability of a i given b j log of 1 by a i given b j plus probability of a i b j. Now, if we sum this equation overall members of a and b alphabets, we will get the relationship as follows.

This should be P b j. But, for convenience, we are just writing as b. It is clear from the context. So, this equation is valid for any channel of the type we have considered. So, in particular, it is valid for the nth extension of the original channel. If you apply this relationship to the nth extension of the original channel, we will get the following relationship.

L n bar is the average word length of a channel from the source A n, that is nth extension of my original source A or equivalently the average word length of n symbols from the original alphabet A. Now, each a posteriori entropy H of A and given beta can be written as the sum of n terns of the form H A given beta. So, this can be broken into n sum of this form. So, if we use this relationship, we can simplify this expression.

(Refer Slide Time: 16:29)

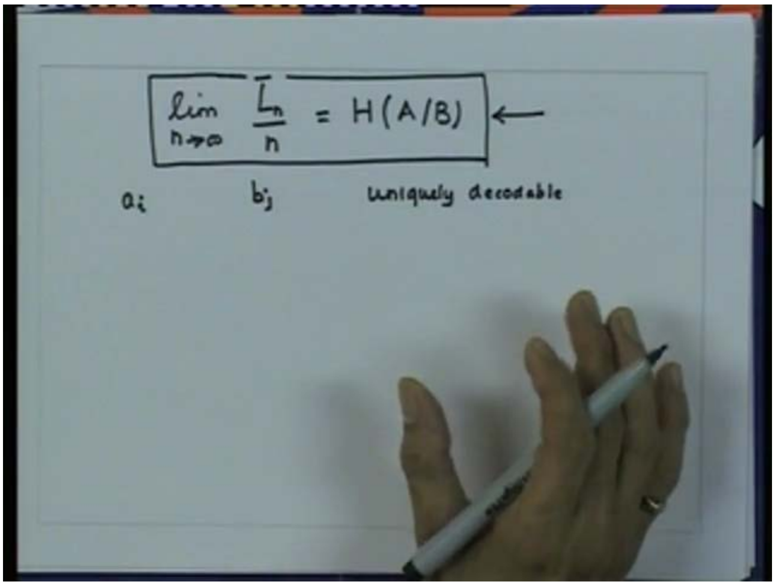


Now, this expression is the generalisation of Shannon’s first theorem. This is very similar to what we had done earlier for a single source, where we had obtained H of S is less than or equal to this, where this was the entropy of the source S. This quantity is the average number of binits needed to encode a symbol from the input alphabet A, if you already have the corresponding symbol from the output alphabet capital B.

So, let us look into this expression given by summation over alphabet B of probability b H of A given b. Now, we will define H of A given capital B. Probability b of H A given b average over output alphabet can be simplified. As once again here, we will not write the subscript a i’s and b j's. Just for the sake of convenience in notation, it is assumed that this a is a i and is summed over the input alphabet A. Similarly, this b is b j. It is summed over the output alphabet capital B.

This can be a simplified as this is double summation over input alphabet and output alphabet, but this is indicated in short by just 1 sigma sign. This is equal to probability of a and b joint log of 1 of by probability of a given b. So, this quantity is by definition equal to H of A given B. That simplifies to the expression shown here. Now, this H of A given B in literature is called the equivocation of A with respect to B or sometimes it is also called as channel equivocation.

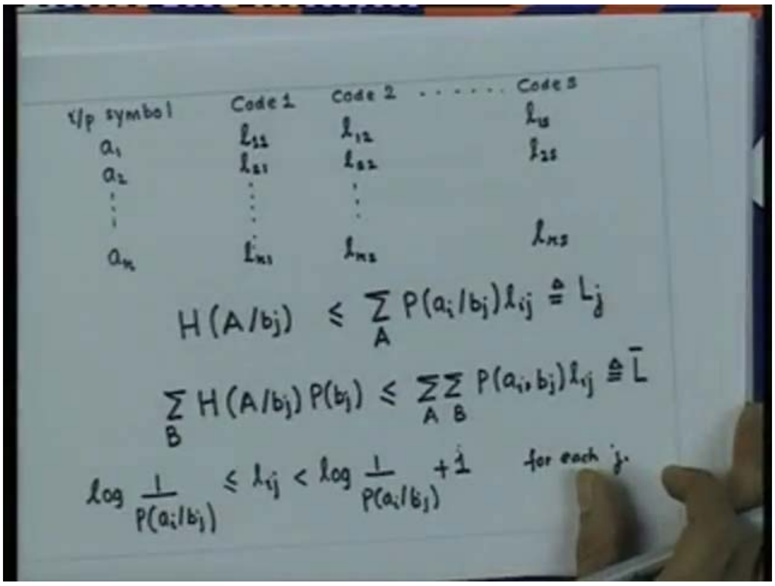
(Refer Slide Time: 20:39)



So, in terms of channel equivocation, we can write the generalised Shannon's first theorem, as limit of n tending to infinity L n bar over n is equal to H of A given B. We had derived a similar relationship earlier while proving the Shannon's first theorem. But there is a major difference between this relationship and the relationship, which we had obtained earlier.

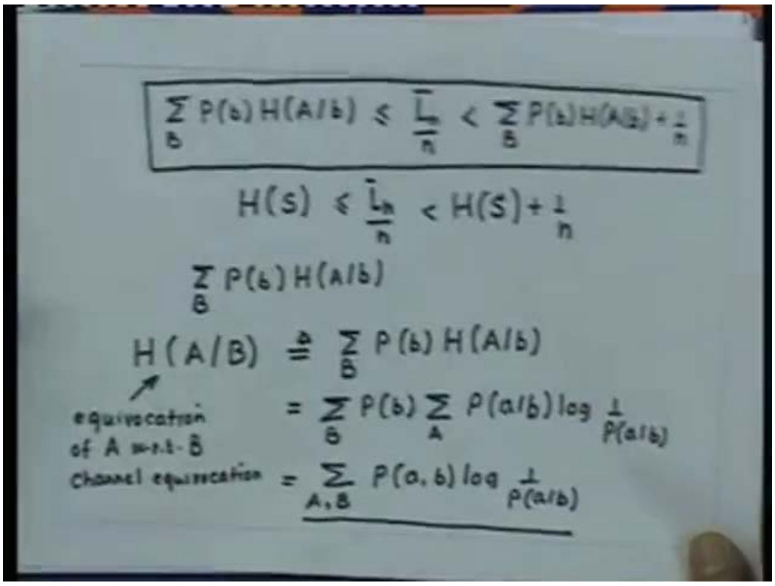
The difference is that that in this case successive input symbols a i of blocks of input symbols are encoded using different codes corresponding to different output symbols b j of blocks of output symbols occurring. Now, even though each of the codes use this uniquely decodable, it is generally not true that a sequence of code words from a known sequence of a uniquely decodable code is uniquely decodable.

(Refer Slide Time: 22:15)



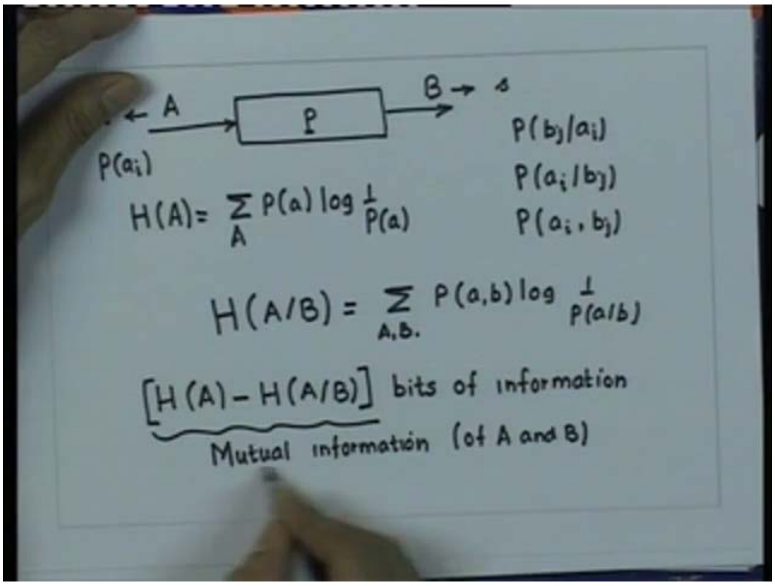
So, in this case, it is not sufficient. Therefore, to select a set of uniquely decodable codes with word length satisfying this condition, than another condition is that codes must be instantaneous. So, what it follows is that the earlier relationship, which we had derived.

(Refer Slide Time: 22:33)



This relationship is valid to only instantaneous code unlike the Shannon's first theorem, which applies to all uniquely decodable code. Let us re-consider the information channel, which we had studied earlier.

(Refer Slide Time: 22:59)

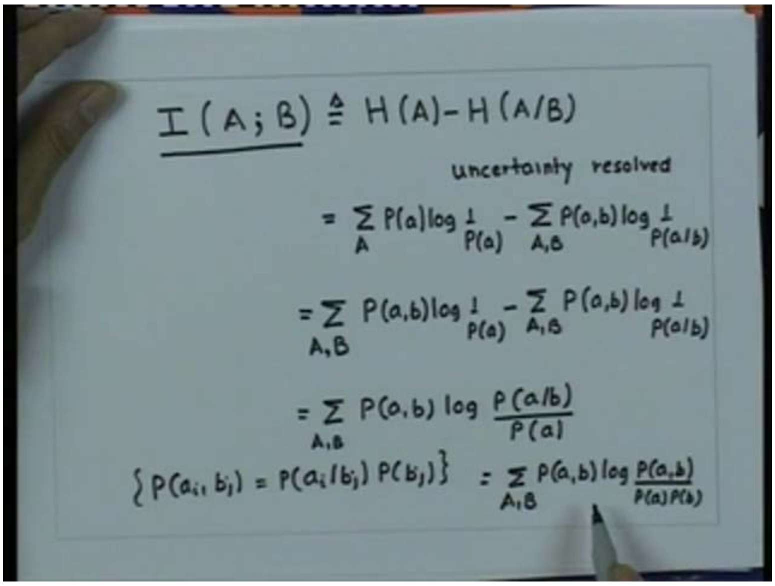


The information channel consists of r inputs and s outputs. So, we have A, we have B. The size of this is s and the size of this is r. Now, the inputs are selected according to the probability P a i for i equal to 1 to r. So, the entropy of the input alphabet can be calculated as P a log of 1 by P a. If you have an input probabilities, forward probabilities is given by the channel matrix P in the form of probability b j given a i. We have seen yesterday that we can calculate the backward probabilities is that probability a i given b j. We can also calculate the joint probabilities P a i b j.

Therefore, we can calculate the equivocation of a channel, which is defined as H of A given B equal to P log of 1 by P of a given b. Now, by Shannon's first theorem, we need an average of H A binits to specify 1 input symbol a i. By the generalisation of the second of the Shannon's first theorem, which we just saw, we need only on average H of A given B binits to specify 1 input symbol, if we are allowed to observe the output symbol produced by that input. Now, what it follows that on the average observation of a single output symbol provides with H A minus H of A given B bits of information.

So, on the average, the observation of single output symbol provides us with this many number of bits of information. So, on the average, observation of single output symbol provides us with these bits of information. This by definition is called as mutual information. That is of A and B. It is also known as the mutual information of the channel.

(Refer Slide Time: 26:43)

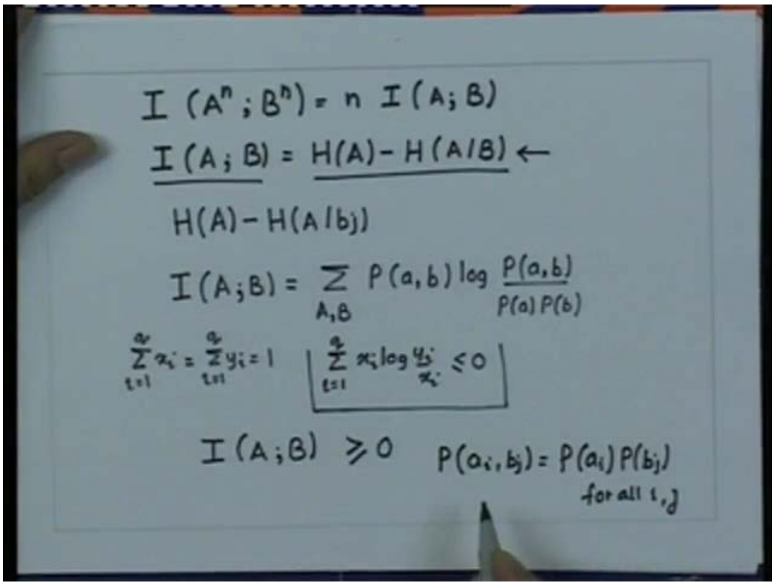


It is written as I of A; B is equal to H A minus H of A given B. So, this by definition is the mutual information. Another way to interpret is this that before I observe the output symbol b j, the uncertainty of the event A was given by the entropy of the source on observing b j output symbol. The entropy on the average of the uncertainty about the input symbol is H of A given B. So, what it means the difference of these 2 entropies should be the uncertainty resolved. So, I can interpret this as uncertainty resolved of the amount of information, which I have received.

Now, with this declaration, let us develop some alternative ways of writing the mutual information. So, if you look at this expression, we can write it as A of P a log of 1 by P a minus double summation over A B probability of a b log of 1 by P a given b. This can be simplified as this. We can write as probability a b log of 1 by P a summed over input and output alphabet because summation over B will give us the marginal probabilities P a and probability a b log of 1 by probability a given b.

This can be simplified as probability of a b log of probability a given b over probability a. Now, since probability a i b j is equal to probability of a i given b j multiplied by probability b j, we can write based on this. We can write this expression equal to probability a b log of probability. So, mutual information of the channel is given by this expression.

(Refer Slide Time: 30:38)



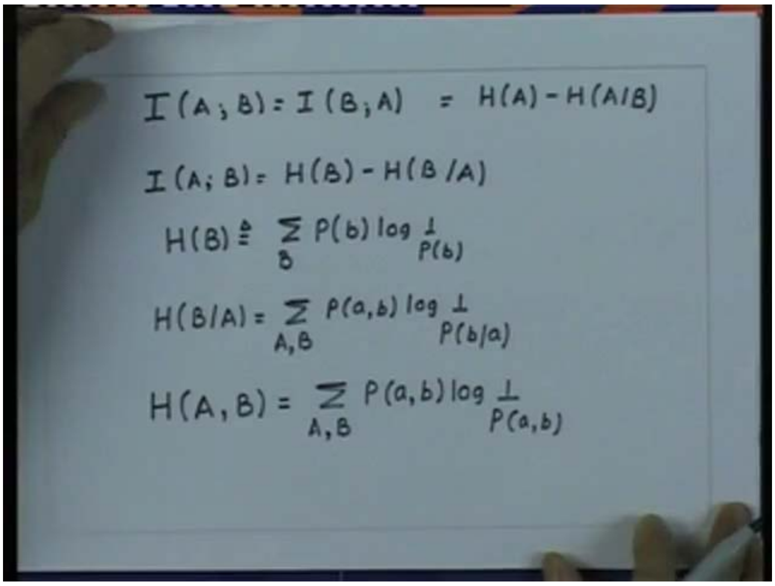
So, if you have the n th extension of a channel, we will find that mutual information can be shown to be equal to n times I of mutual information between A and B. So, we have shown that the mutual information is equal to the average number of binits necessary to specify an input symbol before receiving an output symbol, less the average number of benefits necessary to specify an input symbol after receiving an output symbol. That can be written as shown by this expression. Now, an immediate question that arises is that what about the sign of H A minus H of A given B. Can mutual information be negative?

Now, we have seen yesterday with the help of an example that H A minus H of A given b j may be negative. What it means entropy of the input alphabet may be greater after reception of a particular output symbol b j. But, the mutual information A is just the average over output symbols of H A minus H A given b j. So, it is averages of this quantity were all output symbols. Now, the question is can this average be negative?

So, to answer this question, let us write I A B as given by this expression. Now, we have seen the relationship that I have 2 random variables of 2 sets. Now, we have seen that if you have 2 set of probabilities x i and y i's, then this relationship, which we had derived earlier is valid. So, using this relationship, we can immediately see that I of A B is greater or unequal to 0 where x i's correspond to probability of A B. y i corresponds to P a multiplied by P b. So, this inequality will be equal if and only if y i is equal to x i that is probability of a i b j is equal to probability a i multiplied by probability b j for all i j.

Now, what it says that the average information received through a channel is always non negative. We cannot lose information on the average by observing the output of a channel. For that, the only condition under which the average information is 0 occurs when the input and output symbols are statistically independent, which will happen when this condition is satisfying. Another important property of the mutual information may be seen by inspection of this expression. This equation may take as a definition of I A B is symmetric in the 2 random variables a i and b j. So, interchanging the roles of the input and output symbol leaves I of A B unchanged.

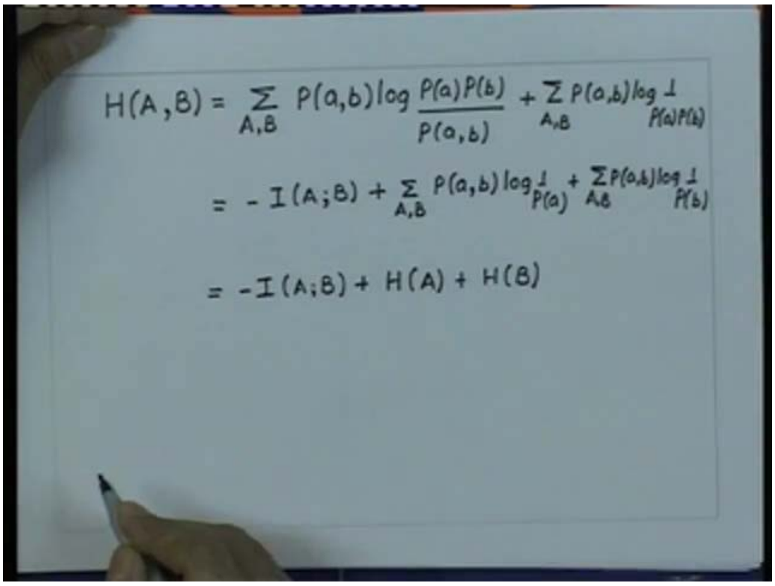
(Refer Slide Time: 35:59)



So, we may write I of A semi colon B is equal to I A of A B is equal to I of B A. So, carrying this argument even further, we may write I A B is equal to H B minus H of B given A because I A B is equal to H of A minus H of A given B. Similarly, I B A is equal to H B minus H of B given A, where H B is by definition equal to entropy of the output alphabet. H of B given A is equal to joint probability of a b log of 1 by probability of b given a summed over input and output alphabets.

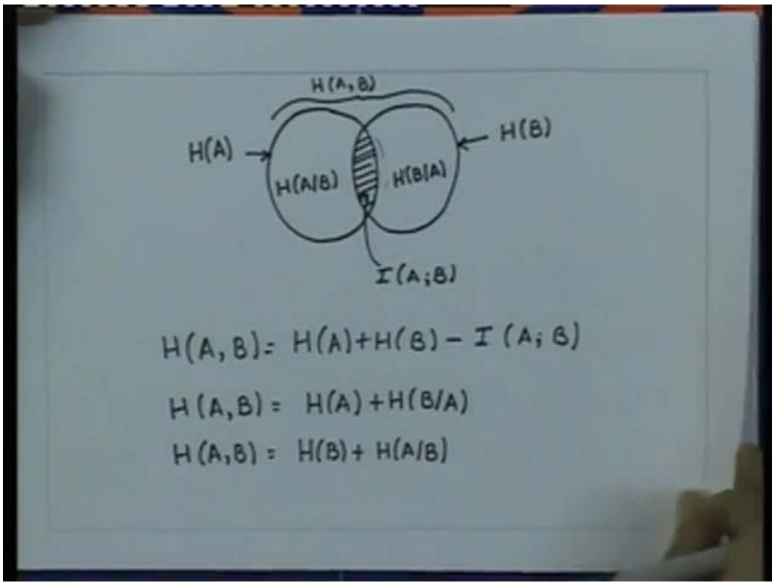
Now, this is called the equivocation of B with respect to A. Now, in addition to the entropies H A and in addition to the entropies H A and H B, it is possible to define a joint entropy, which measures the uncertainty of the joint event a i b j. That will be given as H of A B is equal to summation of probability a b log of 1 by probability a b summed over input alphabet output alphabet. Now, the relationship of H A B to H A and H B can be easily derived as follows; probability a b log of 1 by probability a probability b.

(Refer Slide Time: 39:04)



So, this by definition is equal to negative of mutual information of A and B. This can be split into 2 terms as follows plus log of 1 by p b. This can be simplified as mutual information plus entropy of the input alphabet plus the entropy of the output alphabet. So, this relationship can be easily depicted in the Venn diagram form as follows.

(Refer Slide Time: 41:18)

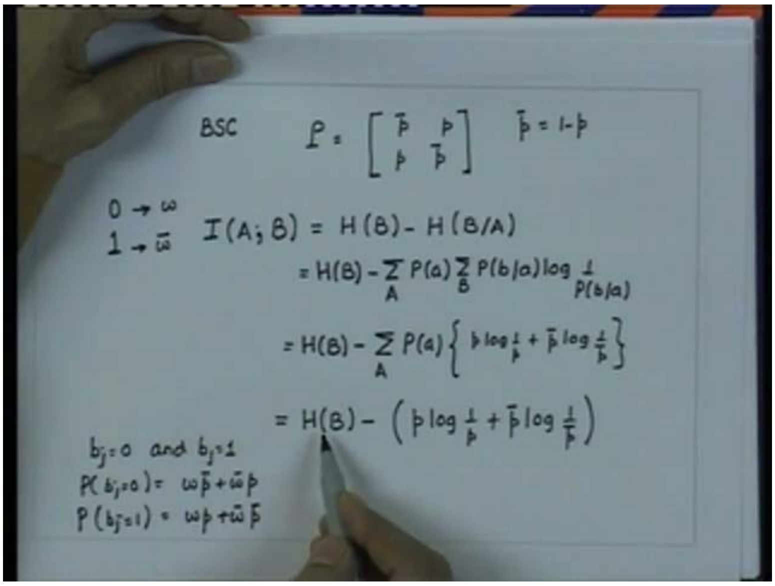


This circle corresponds to entropy of the output alphabet. This circle corresponds to the entropy of the input alphabet. This intersection corresponds to mutual information of A and B. This portion out here without the shaded portion is H of B given A. Similarly, this portion out here is H of A given B. This total is H of A B. So, the entropy of A is represented by the circle on the left and the entropy of B is by circle on the right. The overlap between the 2 circles is the mutual information.

So, the remaining portion of H A H B represent the equivocation H A given B and H B H of B given, A respectively the joint entropy that is H of A B is the sum of H A H B except for the fact that the overlap is included twice. So, we keep looking at this Venn diagram. H of A B is equal to H A plus H B minus A of I of A B. Now, from this Venn diagram, it is very clear that H of A B is equal to H A plus this portion. So, that is H of B given A. Similarly, H of A B is equal to H B plus the remaining portion out here, that is H of A given B.

It is physical interpretation of this. This is very easy to see that uncertainty of the joint event A and B is equal to the sum of uncertainty of B plus the uncertainty of A given B has been observed. It is also equal to the uncertainty of the event A plus the uncertainty of the event B given A has been observed. So, our primary interest is in information channels, but it is important to know that. So, let us compute mutual information for a binary symmetric channel.

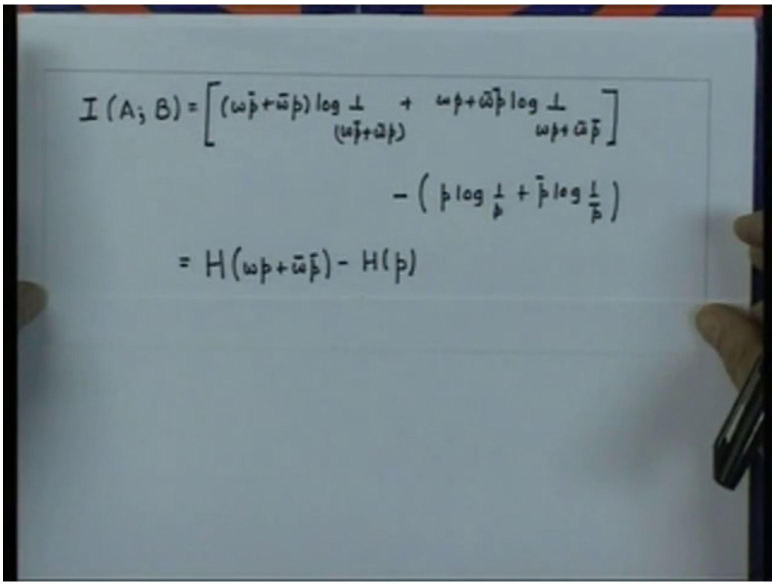
(Refer Slide Time: 45:27)



The channel matrix of the binary symmetric channel is given as p bar p p p bar, where p bar is equal to 1 minus p. Now, assume that probabilities of a 0 and 1 being transmitted are omega and omega bar respectively. So, we can write the mutual information for the binary symmetric channel where the input and output alphabets consist of the binary symbols as H B minus H of B given A.

This is equal to H B minus probability of a probability of b given a log of 1 by probability of b given a. This can be simplified as this quantity is equal to p log 1 by p plus p bar log 1 by p bar. So, this is equal to H of B minus p log 1 by p plus p bar log 1 by p bar. Now, the probability is that b j is equal to 0 and b j equal to one can be easily calculated. It can be shown that these probabilities are omega p bar plus omega bar p. Probability of b j equal to 1 is equal to omega p plus omega bar p bar. From the channel diagram for the binary symmetric channel, it is very easy to show this relationship. Now, using this relationship, we can write the entropy for output alphabet.

(Refer Slide Time: 48:48)



If you write that we get the mutual information as follows plus this. This is the mutual the entropy for the output alphabet, minus p log 1 p plus p log 1 by p bar. This we can write in terms of the entropy function as H omega p plus omega bar p bar minus H of p. In the next class, we will provide geometric interpretation of this relationship. We will also look at the definition of some information channels like noiseless channel and deterministic channels. We will study some interesting properties of entropy and mutual information as revealed by consideration of the cascade of 2 channels.