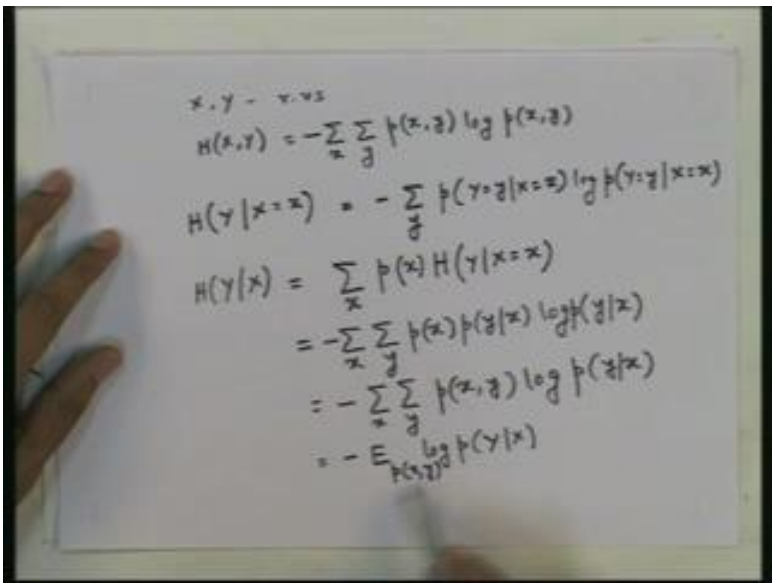
Digital Communication
Prof. Dr. Bikash Kumar Dey
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Lecture - 09
Information Theory
(Part-2)

In the last class we have started discussing information theory and in this class also we will continue some more background on information theory. In the last class we have defined entropy of a random variable discrete random variable. And, we have discussed that that is that entropy that is the average information contained in a random variable that is also the bound for source coding. That is, the random variable can be expressed with a minimum number of bits which is same as the entropy of the random variable. The minimum number of bits required to represent the random variable on average is H X. That is the entropy of the random variable.

Now, in this class we will start by first defining some more information quantities and then we will go towards results related to channel coding. That is, if you want to transmit information through a channel how much maximum how much information can be transmitted? We will try to answer that question after defining some more quantities with some examples. So, in the last class we have also defined joint entropy which is quite straightforward from the definition of entropy itself.

(Refer Slide Time: 02:35)



If we have two random variables X and Y then, we define joint entropy of X Y as, summation over x and summation over y probability that capital X is equal to x, Y is equal to small y times log p x y minus of this. We define this to be the joint entropy of X and Y. And, we also commented that this joint entropy will be equal to the sum of the entropies of the individual random variables, if the individual random variables are independent. And, we also saw examples and we also proved that this is so.

Now, if X and Y are not independent then of course, it is not true that is the joint entropy is not same as the sum of the entropies. And, we will see that it will be less than the sum if X and Y are not independent. So, before going into that let us first define another information quantity. Let us define H Y, given that X is equal to a particular value small x. So, if X is known to be x equal to x then, what is the average information contained in Y. Then of course, that will be clearly determined by the conditional distribution of Y given that X is equal to x.

So, instead of working with the distribution of Y we will work with the distribution of Y given that X is equal to x. So, p y given X equal to x will play a role in this definition. So, this is defined as minus summation over y p Y equal to y, given X equal to x. We will denote this simply by p y given x log p Y equal to y given X equal to x. So, this is the definition. We can see that the simply condition it is the same definition as H Y except that this all the probabilities are conditioned by X equal to x. So, this is the uncertainty or the information contained in Y if you know that capital X is x.

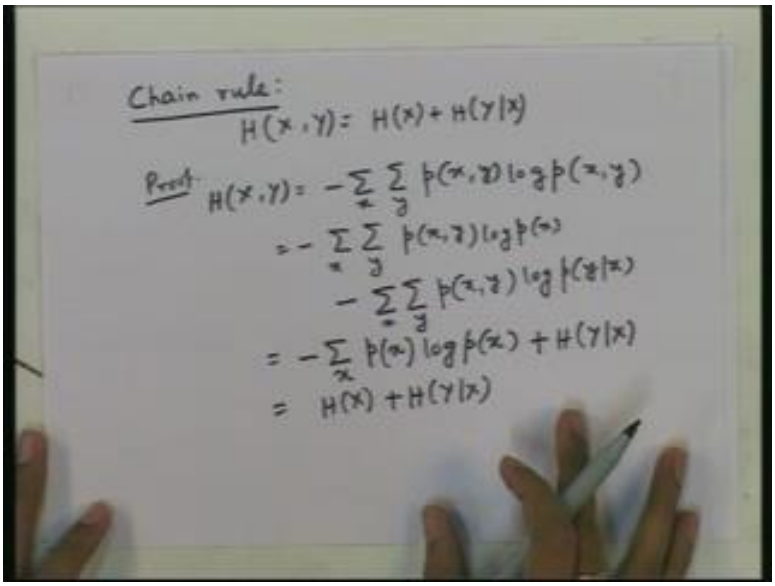
Now, of course X may not be x all the times. So, the probability that X is equal to x is p x. So, we can average this quantity itself over all values of x and that will give us average uncertainty in Y, if we know the value of x. Here this is the average uncertainty in Y if the capital X is equal to x a particular value. But once we average this quantity over all values of x, we get the average information in Y given x. So, H Y given x that is the entropy of conditional entropy of Y given X is defined to be the average. So, x will take the value small x with probability p x. So, that has to be taken into account while averaging. So, this is the definition of conditional entropy of Y given x.

Now, one can intuitively feel that this is kind of the extra information that Y contains once X is told. So, if they are dependent if X and Y are dependent, then if you tell me the

values of X that will also tell me give me some information about Y. Because they are X and Y are dependent on each other. So, if you tell me the value of X we know some information about Y, but even then there is still some uncertainty left in Y. So, this quantifies the average uncertainty left in Y if X is told. So, one can feel that this will be less than H Y itself. Now, we will prove that in a moment.

So, we can write this after substituting this, for this quantity from this definition as summation over x and this summation over y in this minus comes here and the p x and then this quantity p y given x log p y given x. Now, p x times p y given x is nothing, but p x y. So, this is like averaging or expectation of this function. This is the function of x and y because p y given x itself is a function of x and y and this is the joint probability mass function of x and y. So, this quantity is nothing, but the expectation of minus of the expectation of log p Y given x. And, this expectation is computed by using the joint distribution p x y. So, we write this x y to denote that that this expectation is computed using this distribution. Now, we prove a very important result.

(Refer Slide Time: 09:55)



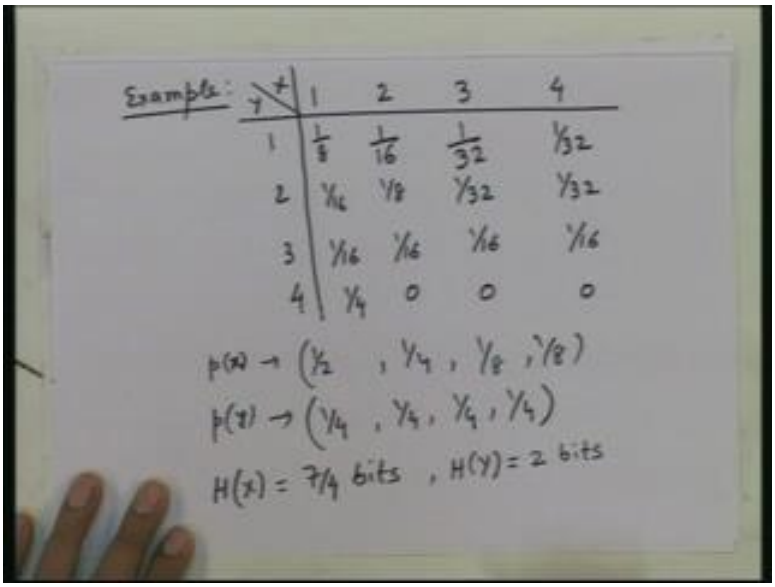
So, chain rule. This is H X Y the joint entropy is H X plus H Y given X. So, what it says is that the total information contained in X and Y is, can be separated into two parts. One is how much information X contains and then given X, what is the extra information that Y contains? So, H X is the total information in X and Y that is the sum of the information contained in X and the extra information in Y given X. So, that that seems quite intuitive, but we need to prove that, using the definitions. So, let us just prove. So,

H Y H X Y is defined this way p x y log p x y, which is same as x y then p x y. Now, this p x y we break into two parts, one is p x times p y given X.

So, we get log p x plus log p y given x. So, log p x and then log p y given X. So, we get p x y log p y with given x. Now, this quantity if you see; if you look at the term inside the summations, we see that this is independent of Y is only p x. Only this part is dependent on Y. So, this summation this quantity can be taken outside the summation on x. So, this is, this quantity is outside the summation of x and the summation p x y on y. If you sum this over all possible values of y what you get is nothing but p x that is the marginal distribution. So, this summation is p x and then this quantity and then this term as it is.

This term we know to be H Y given X this from the definition of H Y given X that is the definition of H Y given X. And what is this quantity is H X. So, we have H X plus H Y given X. Now, this result as we said is very important because it was it is also very intuitive as I explained just now. And, it says many more things it says for example, that that H X Y is H X plus H Y if X and Y are independent. Because, if X and Y are independent H Y given X is same as H Y because X does not have any information about Y because X and Y are independent. So, the H Y given X is same as H Y. So, as a special case we see that H X Y is same as H X plus H Y, if the random variables X and Y are independent of each other. Now, let us see this with a see this by an example.

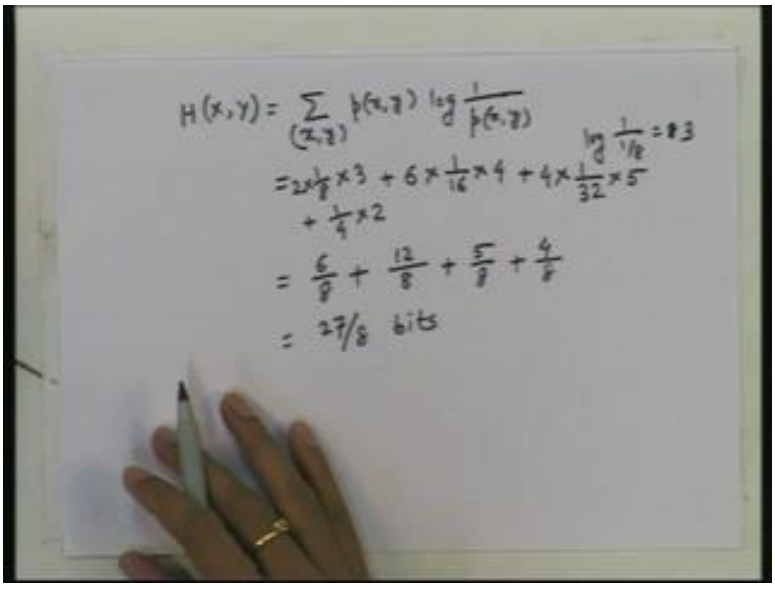
(Refer Slide Time: 14:22)



Let us consider an example of two random variables. And, we take two random variables and we take the joint distribution and we write the probability joint probability mass function as a table 2 dimensional table. X takes 4 values, 1 2 3 4. Y also takes 4 values, 1 2 3 4. And, the probability that X is 1, Y is 1 is one-eighth; probability that X is 2, Y is 1 that is the probability of 2 1 is one-sixteenth. Similarly, all the other probabilities are as following. So, let us take this particular example and compute the important information quantities that we have defined. First of all let us compute the marginal distribution of X and Y. What is p x p x p x? So, for X equal to 1 what is the probability? We will get that by adding all these.

If you add all these we will get half. Then we add all these we will get one-fourth. If you add these we will get one-eight, add these we will get one-eight. Similarly, p y the marginal distribution of Y we will get by probability of 1 is sum of all these which is one-fourth. Probability of 2 is sum of all these which is again one-fourth and one-fourth and one-fourth. So, these are the marginal distributions of X and Y respectively. So, what is H X? This probability distribution may look familiar to you because of we took this as an example in the last class. And, we computed the entropy of this random variable for this distribution and it came out as seven by 4 bits. And, what is the entropy of this random variable H Y this is uniform distribution. So, it is 2 bits. Now, we have H X and H Y. Let us also compute the other quantities H X Y.

(Refer Slide Time: 17:28)

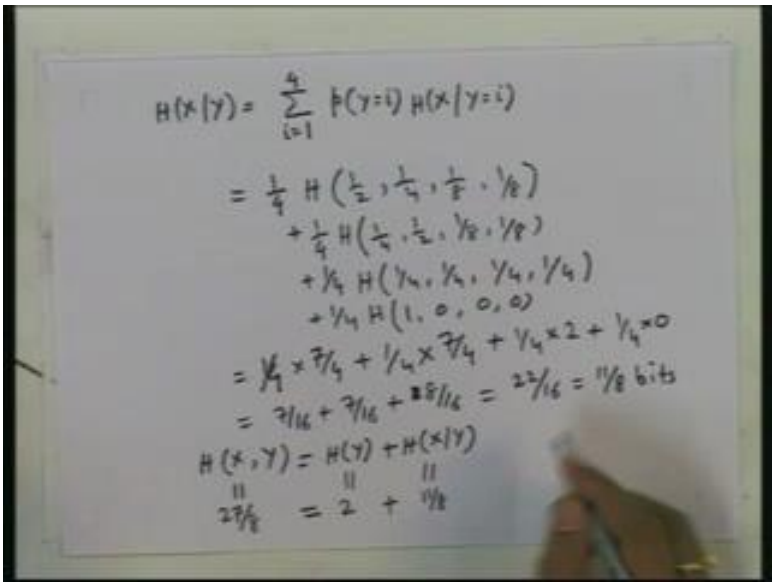


What is H X Y? This is the we have to take all possible values of x and y pair. So, there are 16 possible values 16 entries in the table. And take p x y that is this is the probability times the log of 1 by the probability. So, for this term what do we get? We will get one-eighth times log 1 by one-eighth. So, one-eighth time times log 1 by one-eighth. So, 1 by one-eighth is 8 and log of that is 3 log 8 is 3. So, for this that log is 3 for this the log 1 by this is 4 this is 32. So, this will be 5 and so on. And this will. So, this will give us a term here which is one-eighth times log of 1 by one-eighth. So, that is 3 one-eighth times log 8 that is 3. And, there are the same value appears here. So, this quantity will appear 2 times and there is no other term as one-eighth. So, we will have 2 times this. So, there is the sum of the terms corresponding to this and this.

Similarly, one-sixteen is there one-sixteenth is there 1 2 3 4 5 6 six times. So, there will be 6 terms which are one-sixteenth times log of 16 that is 4 Then, 132 1 by 32 is there 4 times. So, there will be 4 terms with the value 1 by 32 times log 32 that is5 plus there is a term one-fourth. So, this is one-fourth of log 4 that is 2. So, this is the joint entropy of X and Y. And, what is this? This is 2 times 3, that is 6 by 8 plus 6 times 4 by 16 we can. So, 6 times 4 by 16 we can write as 24 by 16 which is 12 by 8, we want to keep the denominator same. So, that we can add. So we are not canceling all the twos here.

So, this is 12 by 8 plus this is 5 by 8 plus this is 2 by 4, which can be written as 4 by 8.If we add 6 plus 12 is 18 18 plus 5 is 23 23 plus 4 is twenty-seven. Twenty-seven by 8 bits. This is the joint entropy of X and Y. Now, now let us compute the conditional entropy of X given Y

(Refer Slide Time: 21:18)



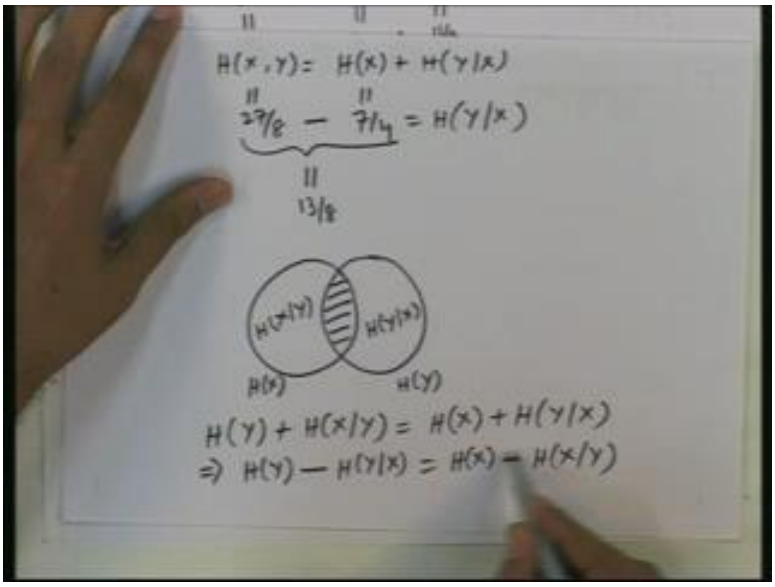
The conditional entropy of X given Y is written as, i equal to 1 to 4 that is the summation over Y. Y takes these values and then probability of Y equal to i times H X, given that Y is equal to i. This is how we defined this quantity. So, what is the probability of Y equal to 1? We can take the marginal distribution marginal distribution is in uniform. So, all the probabilities are one-fourth. So, one-fourth times H X gave Y equal to 1. So, H what is the distribution of X for Y equal to 1? The marginal is one-fourth. So, we have to divide this row by one-fourth that is multiply this row by 4 that is the distribution of X given Y equal to 1. So, divide the joint distribution by the marginal distribution that is the conditional we will get the conditional distribution. So, divide this slope by one-fourth that is multiply by four. So, if you multiply this row by 4, what is the distribution we get? Half one-fourth 4 times 132 is one-eighth one-eight. Then, the entropy of this distribution. So, we write it as H of this.

So, this and then other values of Y also, we have to take Y equal to 2 has the probability one-fourth and the and the conditional distribution is this times 4. Second row times 4 and that is one-fourth half one-eight one-eighth plus one-fourth H. Third row times 4 that is one-fourth because this is all one-sixteenth. Then, one-fourth H fourth row times 4 that is 1 0 0 0 1 0 0 0. Let us now compute these entropies, one-fourth times entropy of this. So, entropy of this we already computed entropy of this is the same as this and entropy of this is seven-fourth bits.

So, this is one-fourth times seven-fourth plus one-fourth times again the same distribution in different order. So, it has the same entropy seven-fourth plus one-fourth times. What is the entropy of the uniform distribution with 4 values? It is 2 plus one-fourth times. What is the entropy of this distribution? This does not have any uncertainty at all. All the probabilities concentrated in one value. So, it has entropy zero. So, we have 7 by 16 plus seven by 16 plus 2 by 4, which can be written as 8 by 16, which is 7 plus 7, 14 plus 8, 14 plus 8 that is 22 by 16 which is 11 by 8 bits. So, the conditional entropy of X given Y is 11 by 8 bits.

Now, let us verify the chain rule that we have proved just now. That is according to the chain rule we should have H X Y to be H Y plus H X given Y. The information contained in Y plus the extra information in X given Y that is the joint entropy of X and Y. So, let us just verify that. This quantity as we have seen as we have computed is 27 by here computed this to be 27 by 8. And, this is 2 bits and this we have just now computed 11 by 8. So, if you add this, so 2 plus 11 by 8 is 16 plus 11 by 8. 16 plus 11 is 27. So, 27 by 8. So, this is really same as the sum of these two. So, this is verified.

(Refer Slide Time: 26:55)



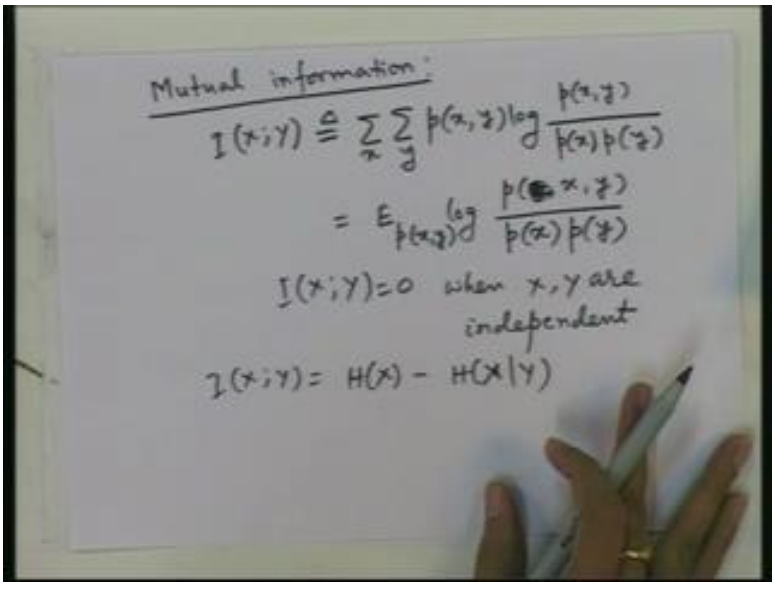
We can also similarly verify that, H X Y is H X plus H Y given X. If we compute this conditional entropy also. We computed here H X given Y we can also compute H Y given X and we can verify that these 2 sides will be same. And, if you if we believe the result that we have proved just now that are this, then we can also compute this using this relation. So, what will be this quantity this we know to be 27 by 8. This we know to be 7

by 4. So, what will be this? This will be this minus this should be H Y given X. And, this is nothing, but 27 minus 14 by 8. So, this is 13 by 8.

So, H Y given X is 13 by 8 which you can which is obtained using this relation. So, this relation can also be expressed in terms of a diagram like Venn diagram. If you denote HX by the circle as this is total information in X and this is total information in Y then there is some common information between X and Y. That is this is part and if Y is told this is known. So, the extra information X is this part and. So, this H X given Y similarly, this is H Y given X. So obviously, one can see that H X Y which is joint entropy is nothing, but this total.

This is H X plus H Y, given X or it can be said to be H Y plus H X given Y. And, what is this part? Then, this is the common information between X and Y, and we will define this quantity also rigorously in terms of the probability distribution in a moment. So, this is called the mutual information between X and Y. The mutual information, it is denoted by X Y. The order does not matter a I X Y is same as H I Y X. This is defined to be summation x, summation y p x y log p x y by p x p y. So, one can see this is the expectation of log p y p x y by p x p y the expectation is taken by using this distribution p x y.

(Refer Slide Time: 29:29)



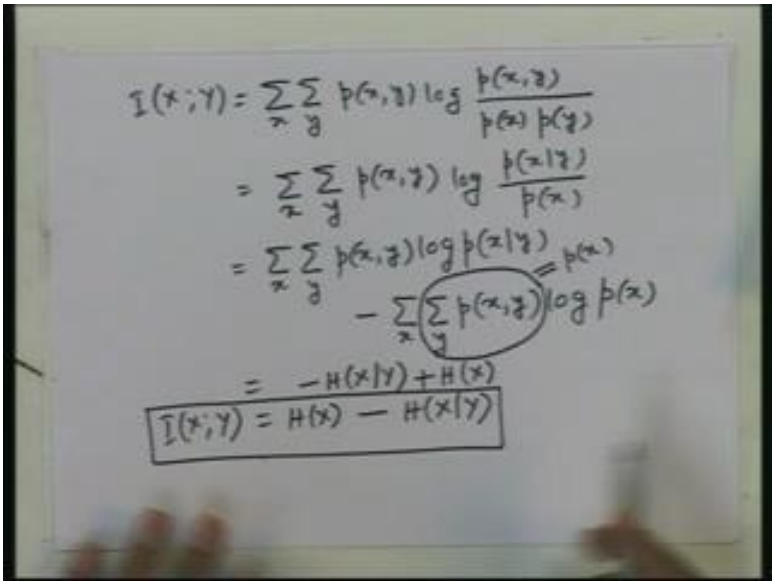
Now, one can if one sees this expression carefully, one observes that if p x y is same as p x times p y that is, if x and y the two random variables x and y independent. Then this quantity is always 1 for all values of x and y this quantity is one. So, log of 1 is zero. So, summation over 0 is nothing, but 0. So, if x and y are independent of each other then, the mutual information is 0 and that is quite expected intuitively also. Because, if x and y are independent you expect no common information between them. If y is told it does not reveal any information about x because they are independent.

So, that is also, so this definition satisfies our expectation in that regard. So, that is mutual information and we can also prove that this is same as this part. This common this common part here denote that is H Y minus H X given Y. The H X minus H X given Y. So, that will be this part already we have seen that this is also true and both these expressions are true that is, H X Y is same as H Y plus H X given Y and which is same as H X plus H Y given X. So, if these 2 quantities are the same we already have that H H Y plus H X given Y which is same as H X Y the same as H X plus H Y given X.

Now, we can this means H Y minus this quantity is taken on the left hand side. And, H Y given X then H X minus this quantity is taken on the right hand side. With negative sign H X given Y. So, that is, what is this quantity H Y minus H Y given X this part? So, this what is left is here this quantity common part. Similarly, H X minus H X given Y is also this common part. So, it is no wonder that this is true because we expect intuitively that this common part is same. So, it is high it is obtained either as H X minus this or H Y minus this. So, that that these two are the same things and these two are the same things as the mutual information between X and Y. So, what we have defined just now the definition of mutual information will actually be same as this and that can proved.

So, first thing we observe is I X Y is 0 when X and Y are independent. Next as you said, we can show that this mutual information is same as either this or this. Then let us just see that. So, we want show that I X Y is H X minus H X give X given Y. Similarly, by interchanging X and Y one can show that this is same as H Y minus HY given X. Let us just see that.

(Refer Slide Time: 34:44)



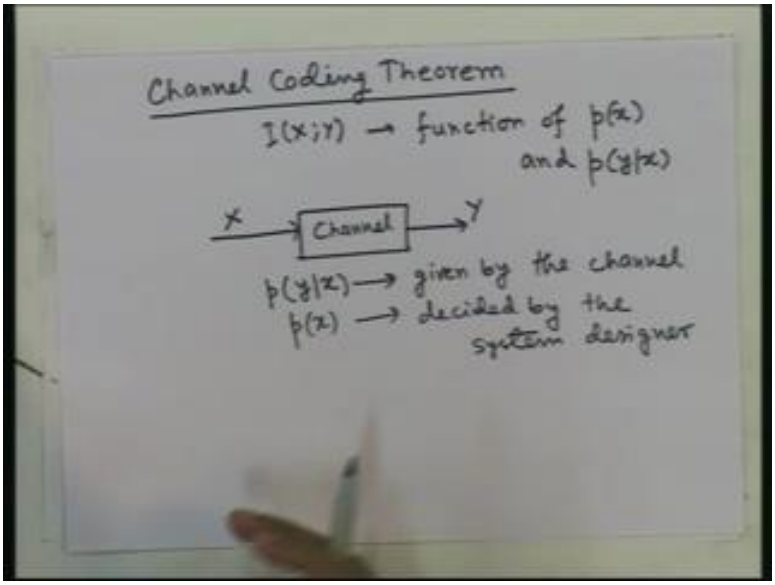
So, let us first write down the definition of I X Y this is summation over x and y p x y log p x y by p x p y this is definition of mutual information. Now, we can write this as summation over x y p x y then log p x y by p x can be written as p x y by p y can be written as p x given y. So, p x given y by p x. So, p x y by p y is written as p x given y. Then, we can type this as log p x given y minus log p x then, we will have 2 terms 1 is log p x given y the other is log p x.

So, what is the first term? First term negative of the first term is H x given y because, this is expectation of log p x given y. So, this is minus H X given Y and this quantity including the minus is H X because this is independent y. So, the summation this quantity is same as p x. So, summation minus summation about x p x log p x is nothing, but H X. So, this is H X minus H X given Y. So, we have I X Y equals H X minus H X given Y. Similarly, one can obtain I X Y equal to H Y minus H Y given X. So, we can now say that this part is really I X Y the mutual information between X and Y. So, this is intuitively also quite nice because if considered any part like H X or H Y.

If we consider H X, for example; it is the sum of 2 parts, one is how much information Y gives about X. That is the mutual information between X and Y and how much extra information X has that is this. So, if Y is revealed it gives some information about X that

quantity is this and there is some extra information in X that is this. The together they form H X. Similarly, for H Y okay, now we will start channel coding theorem.

(Refer Slide Time: 38:09)

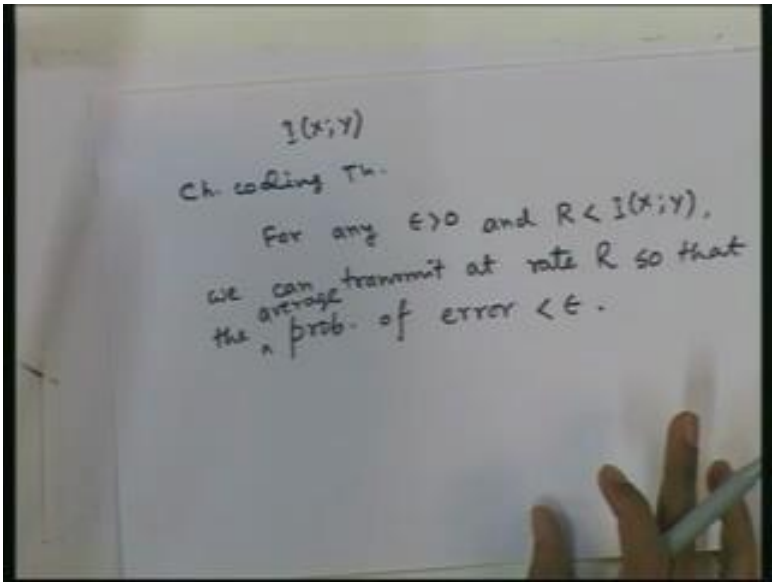


It is very important theorem of important results by Shannon. Before discussing channel coding theorem, let us observe that I X Y is a function p x y, but p x y itself can be expressed as product of p x and p y given X. So, we can say that I X Y is function of p x and p y given x they together give us p x y and I X Y is a function p x y. So, it is function of p x and p y given x. Now, consider the communication setup there is a channel a random variable X is transmitted and Y is received then this p y given x is actually property of the channel. If X is given, that is if the transmitted value is known the density of Y or the distribution of Y is a property of the channel.

We have nothing to do with it. We cannot dictate what this will be. For a particular value of x the distribution of y is given by the channel. So, this is this is given by the channel and we assume it to be known. Now, the other part of which I X Y is also a function that is p x is in our hand, that is this decided by the system designer. So, we have this in our hand we can change p x as we want. This determines how we are doing to transmit the values of x with what distribution. Now, the channel coding theorem says that, for any given value of p x a p x is also told that we want transmit with the x with the distribution p x then the maximum rate at which you can transmit reliably.

We will define what reliably is more precisely now later. But let us accept that there is something called reliable communication. So, if you want to transmit at transmit reliably at what is the maximum rate at which you can transmit for a given p x? That is nothing, but I X Y.

(Refer Slide Time: 41:34)



The mutual information between x y X and Y that is also quite intuitive. Because, we are doing to transmit X with distribution p x. And, we are going to receive Y and p x is also given and p y given X is also given by the channel. So, if we transmit X with the distribution p x then I X Y is fixed. And, then from what we have seen we want to estimate the value of X and that information theoretically. How much information can you get about X from the value of Y? That is nothing but this mutual information. So, this quantity should be the amount of information that can be transmitted through the channel by every use of the channel.

So, by one transmission we should be able to transmit this much information through the channel. And, really the channel coding theorem says that this rate at which once can transmit. More precisely what it says is that for any so, channel coding theorem, so this is actual theorem will come later, but this is the preparation to that. So, it says for any epsilon greater than 0. So, consider any probability of error that we are satisfied with. Let

us say we want to have probability of error that is consider it to be bit error rate or something error rate.

The error rate to be we want the error rate to be less to than ten to be equal minus 6 or 10 to minus 5 10 or minus 10 or whatever. Fix some small quantity that is, that gives us an upper bound on the probability of error that you want. So, given any epsilon greater than zero, which is the upper bound on the probability of error that is desired. We can communicate at a rate R, if R is less than this quantity. So, for any epsilon greater than 0 and R less than I X Y. So, if we choose a rate less then I X Y, then if we choose any small probability of error bound, we can transmit at this rate with less than this probability of error.

So, for any epsilon greater than 0 and or less than this quantity, we can transmit at rate R so that the probability of error, average probability of error is less than epsilon. So, is very important result, but it also very important to understand the result of the statement precisely. What it says is that you fix any probability of error that you are satisfied with. You say that I want probability of error less than 10 to the power minus 100. And, what this result guarantees is that, if you want a probability of error 10 to the power minus 100 or minus 200 or whatever does not matter, as long as you want to communicate at a rate which less than the mutual information between X and Y.

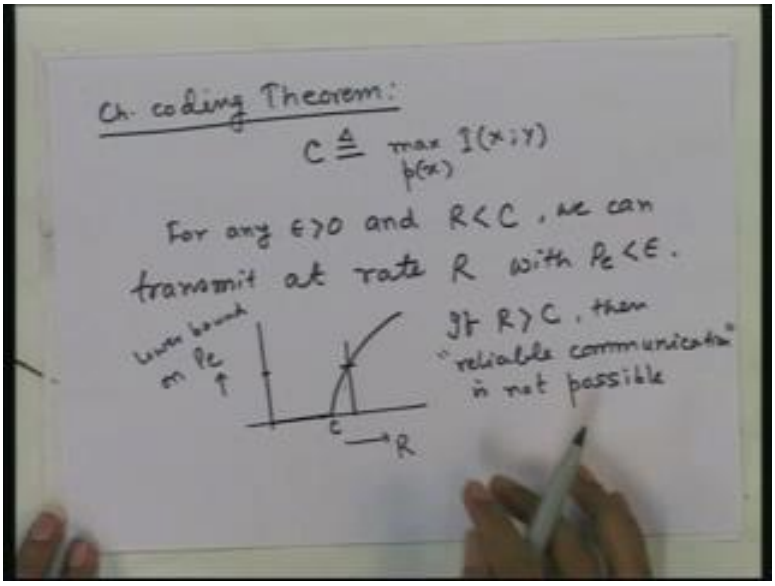
If the rate is less than mutual information between x and y, then there is a transmission scheme by which you can transmit at that rate. And, the probability of error will be less than whatever you say. It cannot be zero, but it can be as small as you want. It can be less than the power minus 10 it can less than ten to the power minus 100. If you want there is scheme for that. So, now here also we fixed p x the distribution of x is fixed. Because, this quantity depends on p x now p x can be still varied by the designer. So, we can chose the p x and try to maximize this quantity also.

So, actually the channel coding theorem says that, you can maximize this quantity by varying p x. Chose the maximum possible quantity I X Y that you can get by dividing p x. And, then if you chose a rate less then that maximum that rate is also achievable with

arbitrarily small probability of error. So, that maximum mutual information between X and Y is called the capacity of the channel. So, let me repeat again that this I X Y this quantity I X Y in this result depends on the probability distribution of X that we choose. So, we can try to maximize this quantity by choosing different p x. So, we can choose p x. So, that this quantity is maximum and then we can assume that p x we can fix that p x and use that for transmission.

Then the same theorem will tell us that, for any rate less then that maximum is also achievable. So, the channel coding theorem says that, first of all we define C that is the capacity of the channel.

(Refer Slide Time: 47:52)



So, this quantity is maximized by choosing a suitable value of p x is maximization is over distributions of x. So, chose that distribution p x and that will give us the maximum value of this and then; obviously, for we can restate the previous result in the following manner. That, for any epsilon greater than 0 and R less then C that is this quantity, if you chose any rate less then C, then you can chose that p x which maximizes this quantity. Then for that p x I X Y will be equal to C and then R will be less then that I X Y. So, this will becomes the same statement as the previous statement.

That for any epsilon greater than 0 and R less then C, we can transmit at rate R with P e less than epsilon. That is the average probability of error less than epsilon. This is the channel coding theorem. And, it says something more it also says that if you chose rate R greater then C then this is not possible. That means; if you chose a rate R which is greater than C then, the probability of error is always greater than sum value. Below if you chose value the probability of error cannot be brought down below that. So, for any rate greater than C, the probability of error cannot be brought down to brought down near 0 as close to 0 as we want we cannot do that.

So, in fact, one can if one plots the upper bound on the probability error the lower bound on the probability of error for different rate, this is rate this P e. Then one can show that it is like this. It is the lower bound on P e. So, it will be something like this. So, and this point is C. So, if R is less than C the lower bound is 0 that is P e can be brought down to and brought down to as close to 0 as we want where as if R is greater than C say here, then probability of error cannot be brought down below this point below this level that is the channel coding. This is the channel coding theorem and the converse this statements is called converse that is, if R greater than C then reliable communication is not possible.

And, here by reliable communication we mean that we cannot bring down the probability of error as close to zero as we want. There is and there is lower bound on the probability of error. So, for R less then C we can bring down the probably of error to as close to 0 as we want where as if it is greater than C it is not possible. So, this is really from practical point of view also this is the capacity of the channel. Because, we reliable communication is possible below this rate, but not above this rate. So, it is defined in terms of information theoretic quantities whereas, channel coding theorem connects this with the practical communication schemes. And, says that this quantity as defined using information theoretic quantities is really the maximum rate at which one can transmit information through a channel.

Okay, so in this class we have we have seen that the relation between the joint entropy of two random variables the entropy of the individual random variables and conditional entropies. And, also we defined the mutual information between two random variables

and seen the relation of this with the other information quantities that we have defined. And, then we have discussed channel coding theorem in terms of mutual information between the transmitter random variables and the received random variables. And, we have said that channel coding theorem says that, reliable communication is possible if the rate is below the mutual information, maximum mutual information that is the capacity and is not possible above that rate.

Thank you.