Information Theory and Coding

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

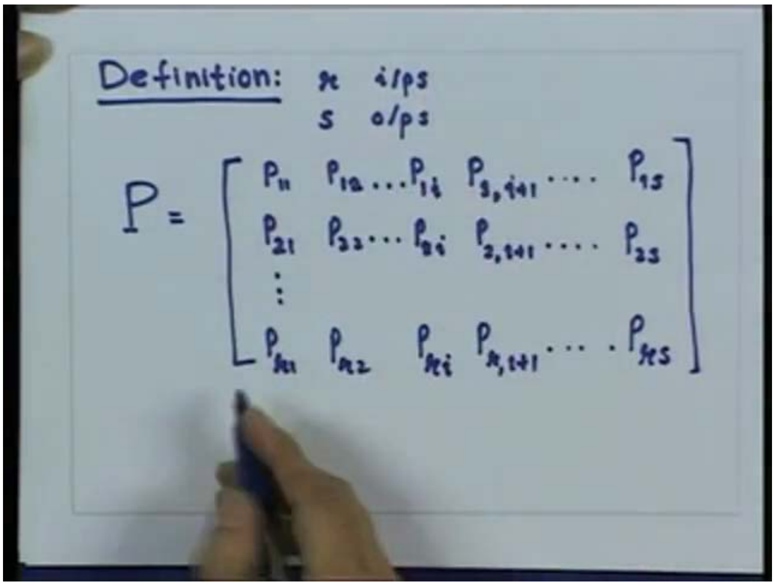
Lecture - 23
Reduction of Information Channels

We have proved mathematically that information tends to leak in a cascade of information channels. We have also found necessarily and sufficient condition that cascade not loose information. We will extend this result to see, how to obtain a radio set of channel outputs without loss in information. For example, in a case of binary data transmission by satellite, the output received at the Earth station often contains irrelevant information.

The antenna on the earth, in a such system might obtain a sequence of pulses of various amplitude. The receiver would take each pulse and compare it with a threshold. If the amplitude of the pulse is larger than the threshold, it converts that pulse to a binary 1 and if it is below the threshold it converts the pulse to a binary 0. We may think of two channels in the situation just described.

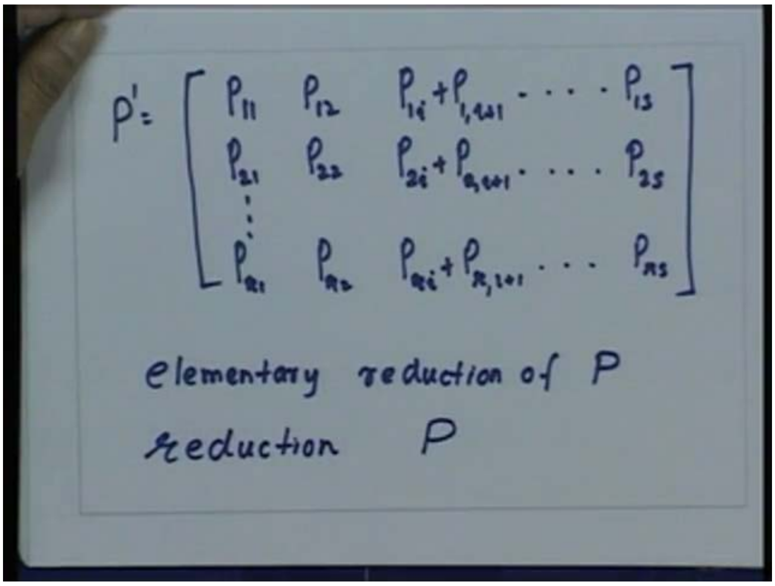
First there is a channel with a binary input that is thus that is sent from the satellite and a large number of outputs, corresponding to the number of distinguishable pulse amplitude. Second we could think of a channel with the binary inputs and binary outputs corresponding to the outputs of a receiver. Now, the second channel is clearly a simplification of the first channel. We call the second channel as a reduction of the first channel so let us define the reduction of a channel.

(Refer Slide Time: 03:01)



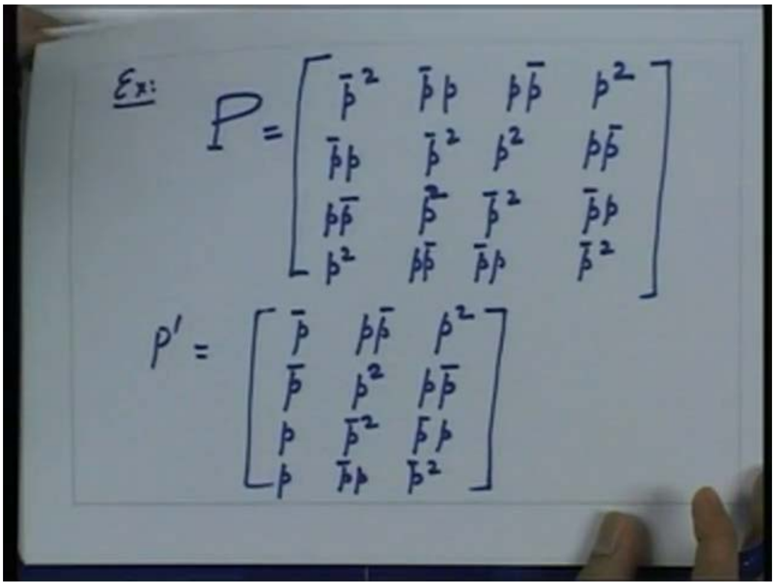
So, consider channel with r inputs and s outputs and this channel is described by a channel matrix P of the form P 11, P 12, P 1i, P 1i plus 1, P 1s. Similarly, we have P 21, P 22, P 2i, P 2i plus 1 P 2s P r1, P r2, P ri, P ri plus 1, P rs. Given this channel matrix, we define a new channel with r inputs, and s minus outs minus 1 output by adding together any two columns of this matrix P. We could combine the i th column and i plus 1, column to get a radio set of output. So, when we do this kind of combination, we call the channel matrix of the new channel P dash.

(Refer Slide Time: 05:06)



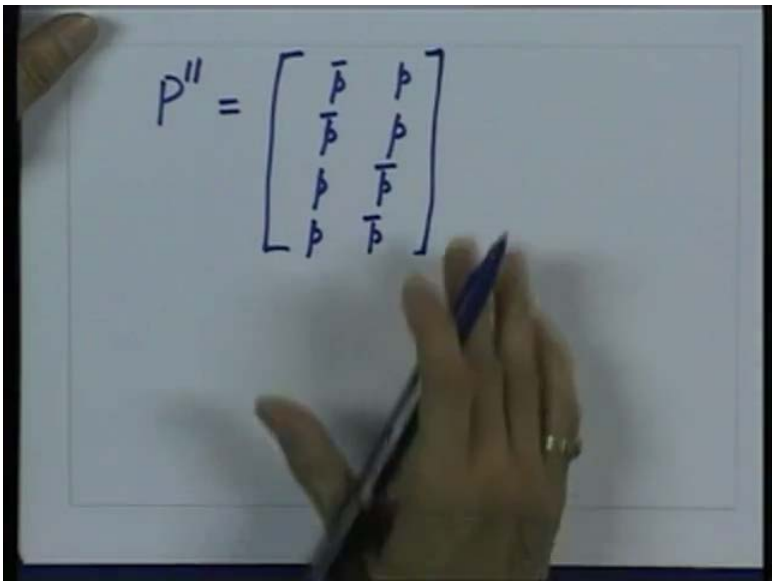
And the is equal to P 11, P 12,P 1i plus P 1i plus 1, P 1s, P 21, P 22, P2i plus P 2i plus 1, P 2s, P r1, P r2, P ri plus P ri plus 1, P rs. So, this channel matrix of the new channel P dash, the new channel P dash is called an elementary reduction of P. Now, we may repeat this process of combining a number of times each time forming a elementary reduction of P dash. So, the end product of more than one elementary reduction will be simply called a reduction of the of the original channel matrix P. So, let us illustrate this with an example. So, let us take a case of a second extension of A binary.

(Refer Slide Time: 06:55)



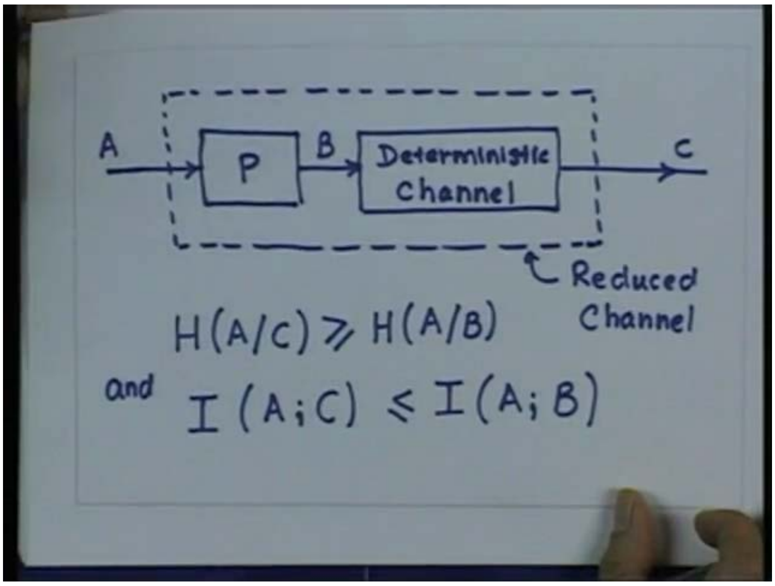
Symmetric channel for which the channel matrix, we have seen earlier is of the form p bar square p bar p p p bar p square p bar p p bar square p square p p bar p p bar p square p bar square p bar p p square p p bar p bar p p bar square. So, this is a channel matrix for the second extension of a binary symmetric channel. So, an elementary reduction of P is form by combining the first and second column of this matrix. And if you do that we will get as P dash is equal to p bar p p bar p square p bar p square p bar p p bar square p bar p. Then finally, p p bar p and p bar square. Now, a reduction of p is formed by combining the second and third column of p dash. So, if we combine second and third column of p dash, we will get a reduction of p.

(Refer Slide Time: 09:20)



So, if we combine second and third column of p dash, we will get a reduction of p which will denote by P double bar, equal to p bar p p bar p p p bar p p bar. A useful way of viewing a reduced channel is as shown in the figure.

(Refer Slide Time: 09:58)

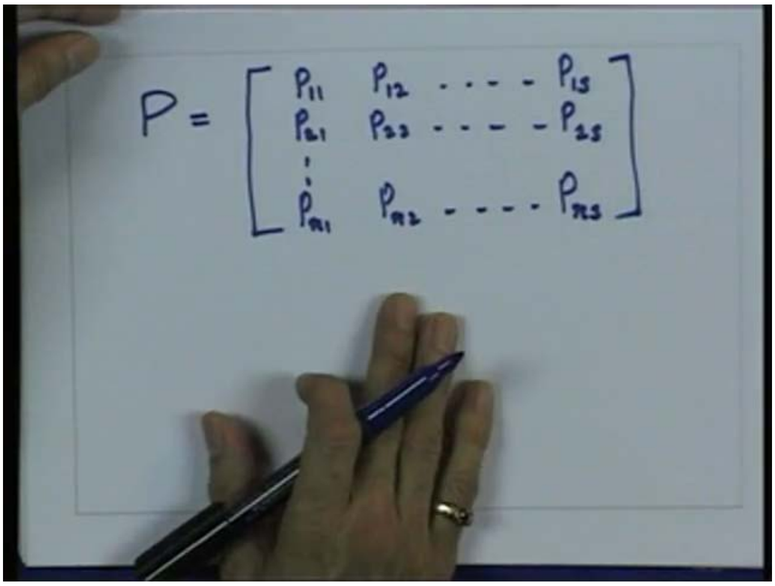


Here, so I have a channel matrix P, I take it is output and feed it to a deterministic channel. The input of this is A, the output is B and the output of this is C, and this is reduced channel. So, the deterministic channel combines symbols of the B alphabet into a smaller number of symbols of the C alphabet. Hence, the channel with the input alphabet A and output alphabet C indicated by dash lines in this figure is a reduction of channel P.

Now, this method of constructing a reduced channel allows us to use the results, which we had discussed in the last class on channel cascades. So, in particular we have with reference to what we had studied in the last class, with reference to this figure, we can say, H of A given C is greater than or equal to H of A given B and I of A semicolon C is less than I of A, B.

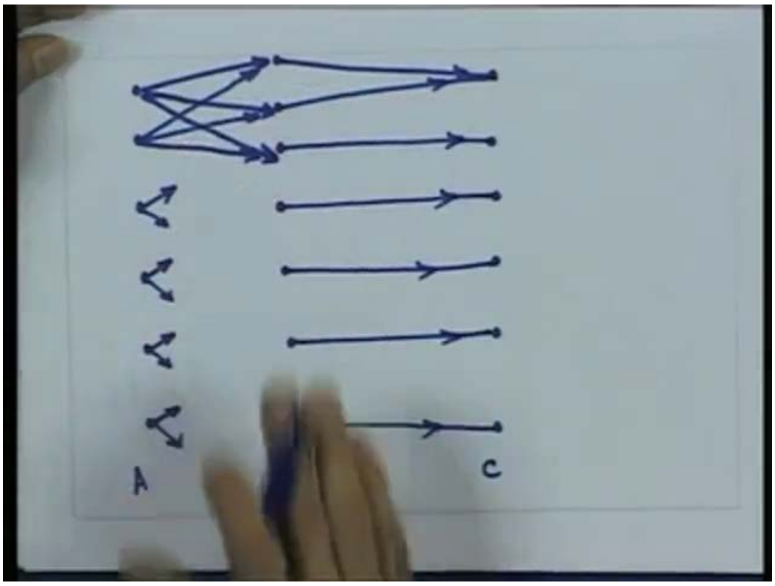
So, forming a reduction of a channel decreases or adversely leaves unchanged the mutual information of the input and output alphabet. Now, this is the price we pay for simplification in the channel. Now, most important question suggested by the above remark is, when can we simplify the channel without paying a penalty in reduced mutual information. That is when is the mutual information of a reduced channel equal to that of the original channel? So, in order to answer this question we need only consider the case of elementary reduction. The question in case of a general reduction may then be answered by induction.

(Refer Slide Time: 13:35)



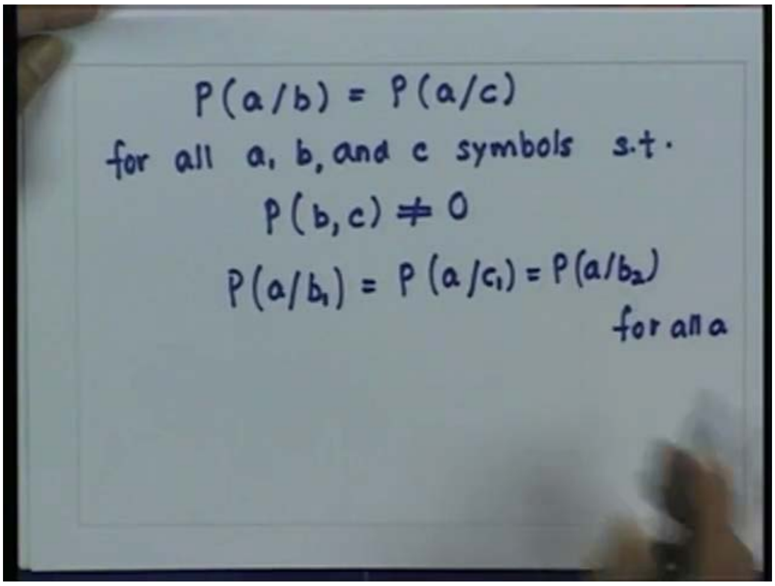
So let us form an elementary reduction of the channel P which is given by P 11, P 12 up to P 1s P 21, P 22, P 2s. P r1, P r2, P rs. Now, without loss of general generality, we may assume the, at the elementary reduction of this channel matrix P is formed by combining the first two columns of P. Now, if we assume this, then the channel diagram.

(Refer Slide Time: 14:17)



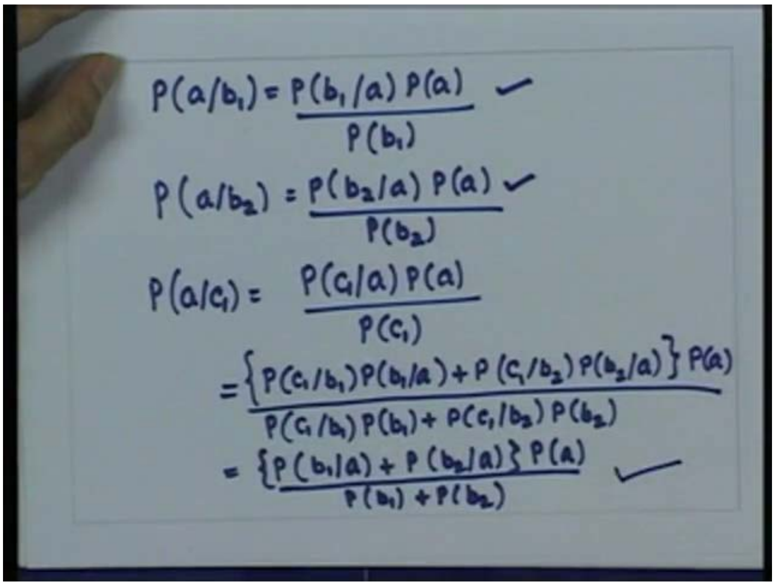
These are the input alphabets and this is the alphabet b and this is the alphabet C. So, A, B and C, so this is the channel reduction by cascade. Now, we found necessary and sufficient condition that A cascade not loose information.

(Refer Slide Time: 16:18)



Let us rewrite that condition, the condition said that probability of a given b is equal to probability of a given c. This is true for all a, b and c symbols, such that probability of b, c is not equal to 0. Now, since we are investigating a elementary reduction, this condition is satisfied truly for all b symbols except the two symbols, we have combined b1 and b2, so these two symbols we have combined here, this is b1 and this is b2. So, in this figure this is your c1 so on. Applying this condition to b1 and b2, we find the necessary and sufficient conditions for cascade not to loose information, are probability of a given b1 is equal to probability of a given c1, is equal to probability of a given b2.This condition should be true for all a.

(Refer Slide Time: 18:16)

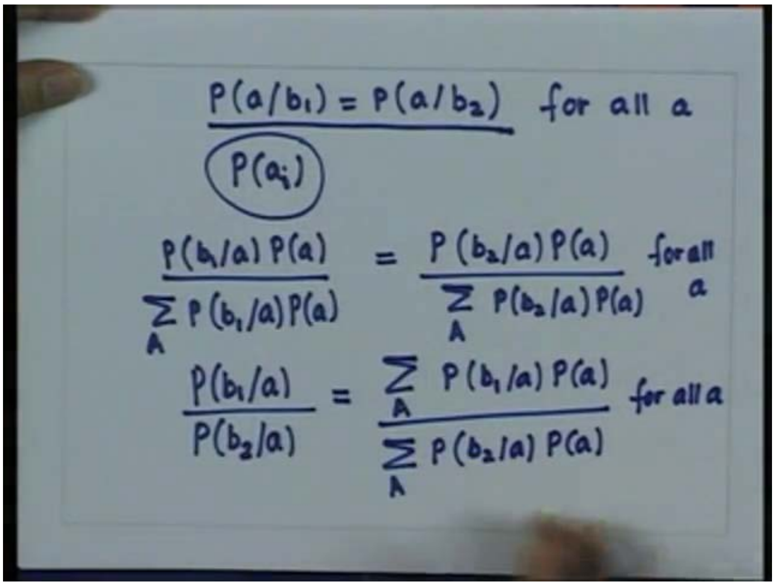


Now, let us look at what is probability of a given b1 and probability of a given b 2. Probability of a given b 1, is equal to probability of b 1 given a, multiplied by probability of a, upon probability of b 1 and probability of a given b 2, is equal to probability of b 2 given a multiplied by a upon probability of b 2, so and probability of a given c1is equal to probability is c1. Given a probability of a, whole over probability of c 1.

Now, this expression can be re written as probability of c 1 given b 1, probability of b 1 given a plus probability of c 1 given b 2 multiplied by probability of b 2 given a probability of c 1 given probability of c 1, is equal to probability of c1 given b1 multiplied by probability of b1 plus probability of c 1 given b 2 multiplied by probability b 2. Now, because the output b is fed to the deterministic channel whose output is c alphabet. What it implies probability of c 1 given b 1 and probability of c 1 given b 2 are both equal to 1.

So, in this that case it simplifies to probability of b 1 given a plus probability of b 2 given a probability of a divided by probability of b 1 plus probability of b 2. Now, what this implies is that, if this two conditions are satisfied, then it implies that probability of a 1 of probability of a given b 1 is equal to probability of a given b 2, is also equal to probability of a given c 1 from this and this condition, this condition is equivalent, probability of a given b 1 equal to probability of a given b 2, for all a.

(Refer Slide Time: 21:47)



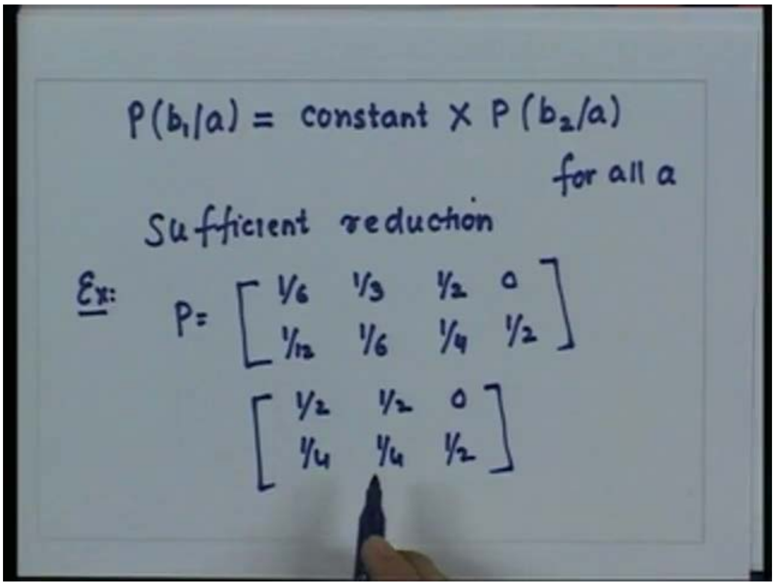
In other words, the two output symbols b 1 and b 2 may be combined, without loss of information if and only if the backward probabilities that is probability of a given b 1 is equal to probability of a given b 2 for all a. Now, this is an important result both in terms of understanding information and from a practical point of view. It provides condition, which a channel may be simplified without paying a penalty.

The backward probabilities however depend upon the a priori probabilities P ai, that is they depend upon how we use our channel. So, it is of even more interest to determine, and then we may combine channel outputs, no matter how we use the channel, that is for any a priori probabilities P of ai. So, in order to get the answer for that, let us use Bayes rule, to rewrite this expression.

So, this expression may be rewritten using Bave rule, Baye's rule as probability of b 1 given a multiplied by probability of a divided by summation of probability of b 1given a probability of a. This will give you probability of b1 is equal to probability of b 2 given a multiplied by probability of a, whole over summation probability of b 2 given a multiplied by probability of a, this condition should be valid for all a.

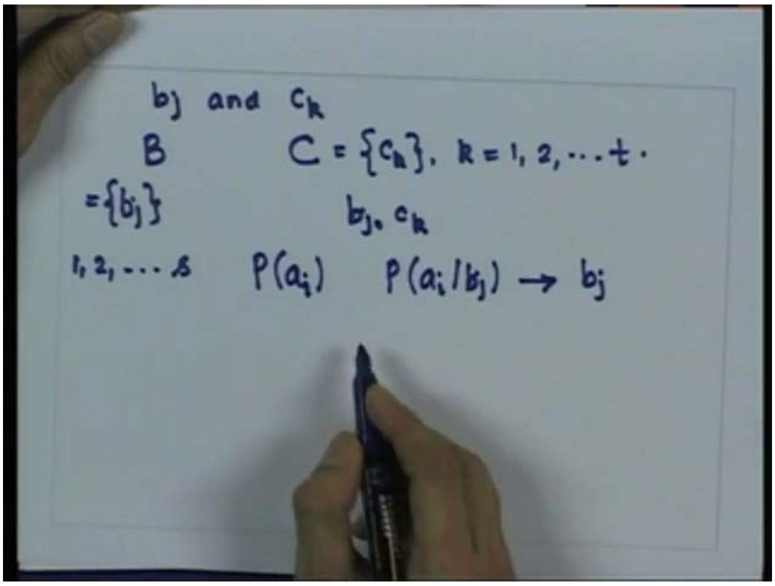
So, this can be simplified as probability of b 1 given a, upon probability of b 2 given a should be equal to summation probability of b 1 given a multiplied by probability of a upon probability of b 2 given a probability of a this should be valid for all a. Now, if this condition is to hold for all possible a priori probabilities P ai, then what it implies that probability of b 1 given a should be equal to constant times probability of b 2 given a for all a, so this is the condition.

(Refer Slide Time: 25:26)



We see, so if you have a channel matrix, satisfying this condition, we may combine two columns of the matrix and the new channel matrix will be as good as the original one. So, more precisely for any set of probabilities over the input alphabet, the mutual information of the channel and the reduced channel will be identical. So, a reduced channel with this property will be called a sufficient reduction. An example of this would be a channel with a channel matrix given as one-sixth, one-third, half, 0, one-twelfth, one-sixth, one-fourth, half. Now, this may be reduced to half, half, 0, one-fourth, one-fourth, half because this column is half the times this column and this may be finally, reduced to by combining this two columns as 1, 0, half, half. So, this is a sufficient reduction of the original channel.

(Refer Slide Time: 27:24)



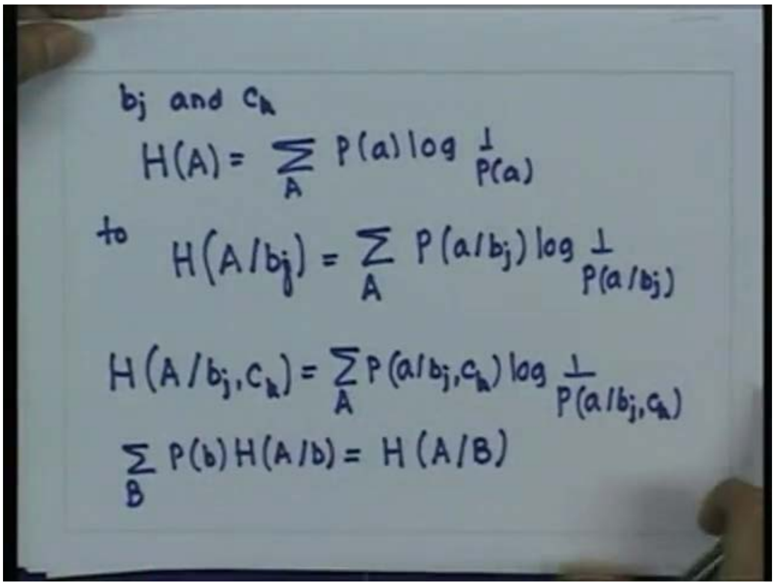
Now, another important property of mutual information is additivity. We will investigate additivity by considering the average amount of information, provide about the set of input symbols by a succession of output symbols. That is, where we consider the case where we may gain information about the input by a number of observations. That is we consider case, where we may gain information about a input by a number of observation instead of a single observation.

An example of this situation occurs when the input symbols of noisy channels are repeated at a number of times, rather than transmitted just once. Such a procedure might be used to improve the reliability of information transmitted through a unreliable channel. Another example is an information channel where the response of a single input is a sequence of output symbols, rather than a single output symbols.

So, let us investigate the additive property of mutual information in the special case where the output of a single input symbol consist of two symbols. The more general case, where the output consist of n symbols, may then be treated by induction. So, let us modify a model of information channel then so that instead of a single output for each input if two symbols say b j and c k, the symbols b j and c k are from the output alphabets B equal to b j, where j goes from 1, 2, s and c k is from the output alphabet C equal to k equal to 1, 2 up to t.

So, without loss of generality, we may assume that the two output symbols are received in the order b j followed by c k, so then the a priori probability of the input symbols P ai change into the posteriori probability P of ai given b j upon reception of the first output symbol b j upon reception of the second output symbols c k. They change into the even more a posteriori probabilities that is, probability of ai given b j, c k.

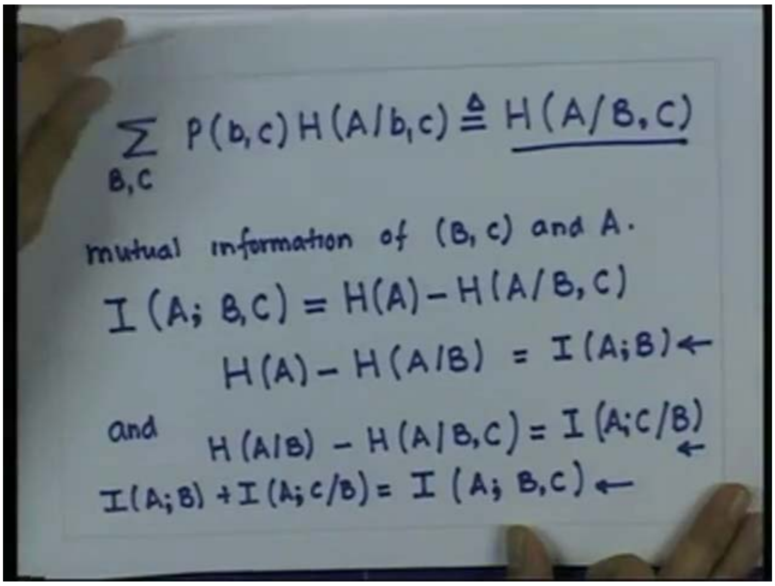
(Refer Slide Time: 31:44)



So, if the two symbols b j and c k are received, the average uncertainty or entropy of the set of input symbols changes from HA equal to Pa log 1 by Pa to the a posteriori entropy given by H of A b j to H of a given b j equal to probability of a given b j log of 1 by probability of a given b j. This summation is over input alphabet A and on receiving c k, this a priori entropy changes to the even more a posteriori entropy which is given by H of A given b j, c k equal to probability of a given b j, c k log of 1 by probability a given b j, ck the summation is over A.

So, now if the average this quantity over the b j to the to find average a posterior entropy, we will get the equivocation of a with respect to b. So, probability of b H of A given b average over B alphabet will give us equivocation of a with respect to B, that is H of A given b, so in the same manner, we may average this term over all b j and c k, in order to find the equivocation of a with respect to the output alphabet b and c.

(Refer Slide Time: 34:15)



So, if we do that we get the quantity as follows, and this is by definition equivocation of A given B and C. So, the results of our generalization of Shannon's first, theorem applies directly to this term. So, this quantity is the average number of binits necessary to encode a symbol from the alphabet A after we are given the corresponding B and C symbols. Now, this equation and this equation out here, suggest two different ways we might measure the amount of information B and C yield about a that is the mutual information of mutual of B, C and A.

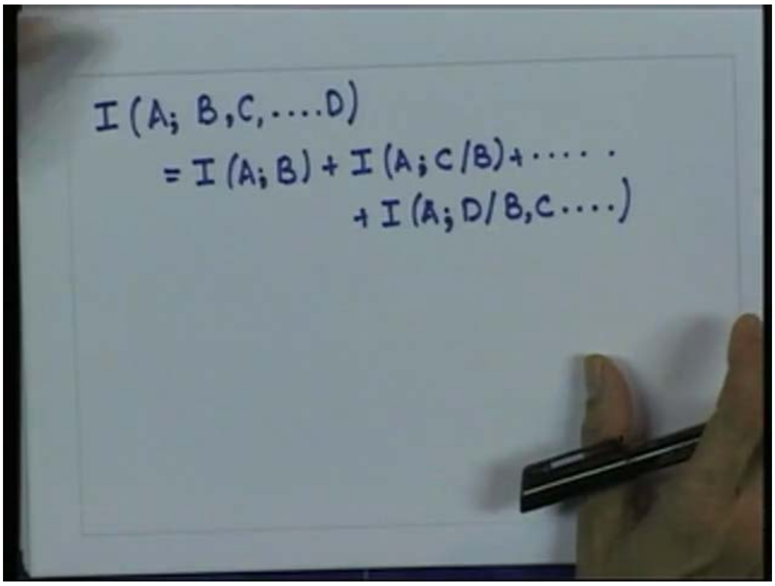
So, we can define the mutual information of A and B, C just as we did the in the case, when the channel output consisted of a single channel, so that is mutual information between A and B, C is by definition we can say equal to H of A minus H of A given B, C. This is one way of defining, second of looking at the same thing is to consider amount of information provided by A. The second way is to consider amount of information provided about A by B alone, and then the amount of information about A provided by C after we have seen B. So, let us look at this quantity. So, we have the quantities of the type H A minus H of A given B and H of A given B minus H of A given B, C.

Now, this first quantity has already been defined and that is nothing but mutual information between A and B. Now, it is natural to define this quantity as mutual information of A and C given B, so if we add this equation and this equation we find I A;

B plus I A; C given B is equal to H A minus H of A given B, C and that is equal to mutual information between A and B, C.

So, what this expression expresses is that, so this equation expresses the additivity property of mutual information. It says that the average information provided by the observation does not depend upon, whether we consider observation in its entire, entirety of broken into component parts. So, this equation maybe generalise immediately to more than one variable.

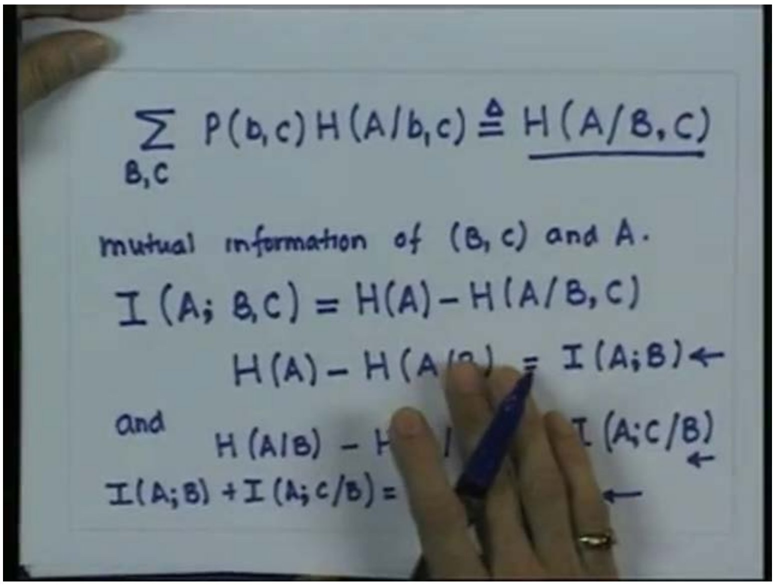
(Refer Slide Time: 39:37)



So, this equation can be generalised as follows I of A and B, C, D equal to I of A, B plus I of A, C given B plus I of A, D given B, C. Now, in this equation the first term on the left is the average amount of information about A provided by an observation. Now, in this equation the term on the left is the average amount of information about A provided by an observation from the alphabets B, C dash, dash up to D.

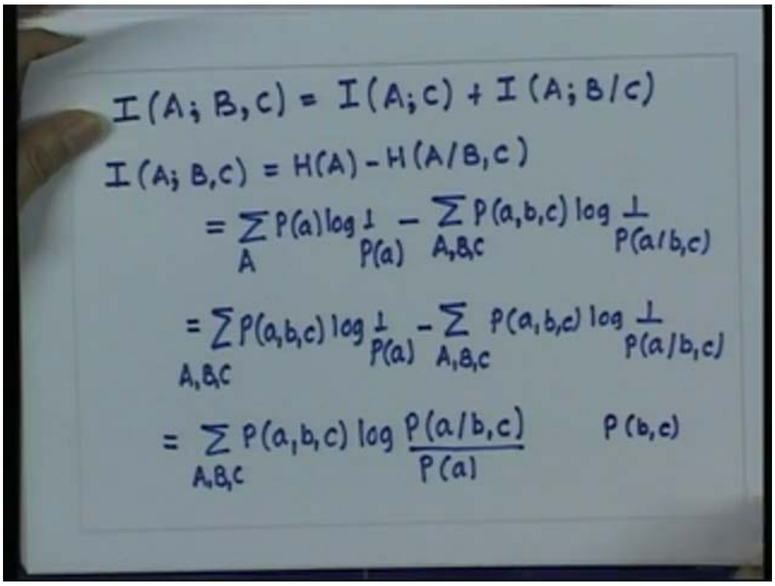
The first term on the right is the average amount of information about a provided by an observation from the alphabet B. The second term on the right is the average amount of information about A provided by an observation from the alphabet C, after an observation from the alphabet B. And this term is the average amount of information about A provided by an observation from the alphabet D after an observation from the alphabets B, C dash, dash.

(Refer Slide Time: 41:39)



The particular order of the information we receive is irrelevant, so it is important to also realise that, if you look at these two equation, it clearly shows that a particular order of the information we receive is irrelevant. For example, we may write this equation as mutual information between A and B, C is equal to mutual information between A and C plus mutual information between A and B given C.

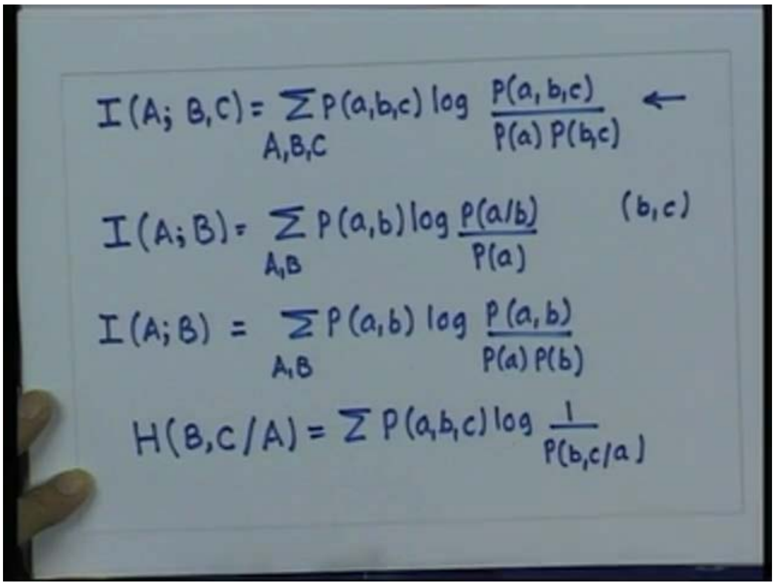
(Refer Slide Time: 42:29)



Now, we may write the information quantities, discussed in several different forms for example, I of A, B, C which is written as H A minus H of A given B, C is equal to probability of a log 1 by probability of a summation over alphabet A minus, this by definition is equal to probability of a, b, c log of 1 by probability a given b, c the summation over A, B, C alphabets.

Now, this can be rewritten as probability of a, b, c log 1 by P a summation over A, B, C alphabet this quantity is again repeated probability of a, b, c log of 1 by probability of a given b, c and this can be simplified as probability a, b, c log of probability a given b, c divided by probability of a summation over A, B, C. Now, another useful form, which is found by multiplying the numerator and denominator of the log term in this equation is by probability b, c. So, if you multiply both terms by probability b, c.

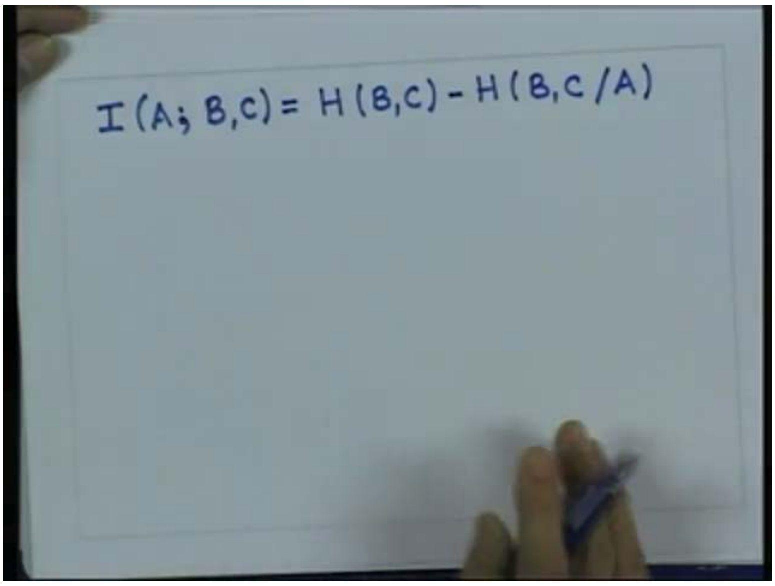
(Refer Slide Time: 45:27)



We get, I of A colon B, C equal to probability of a, b, c log of probability a, b, c divided by probability of a multiplied by probability of b, c summed over A, B, C. Now, the similarity of this equation and this equation is very clear, if we compare to what we had done earlier, I of A, B is equal to probability of a, b log of probability a given b upon probability of a summed over a, b.

So, this expression or this equation is similar to this equation and I of A colon B is equal to probability of a, b log of probability a, b upon probability a probability b summed over A, B alphabet and this is similar to this. So, what it implies that we could have obtained this equation, and this equation by replacing b by b, c. So, if you go by this argument then we can define H of B, C given a equal to probability of a, b, c log of 1 by probability b, c given a summed over A, B , C.

(Refer Slide Time: 48:20)



So, we can using this relationship it is easily verified that mutual information between input alphabet A and output alphabet B, C is equal to H of B, C minus H of B, C. Given A we will illustrate additivity property of mutual information, by examining an example based on a binary symmetric channel, this we will do in the next class. In the next class we will also extend our definition of mutual information, to more than two alphabets.

We will also investigate an interesting property of information channel that is channel capacity. We know to calculate the mutual information of a channel it is necessary for us to know the a priori probabilities of the input alphabets that is P ai. So, what it implies that calculation of mutual information for a channel, not only depends on the channel matrix, but also on the a prior probabilities P ai that is on how we use the channel. So, it is of interest to see how the mutual information varies with this P ai, and what is the maximum value which we can achieve. And this maximum value is defined as channel capacity, we will investigate this in the next class.