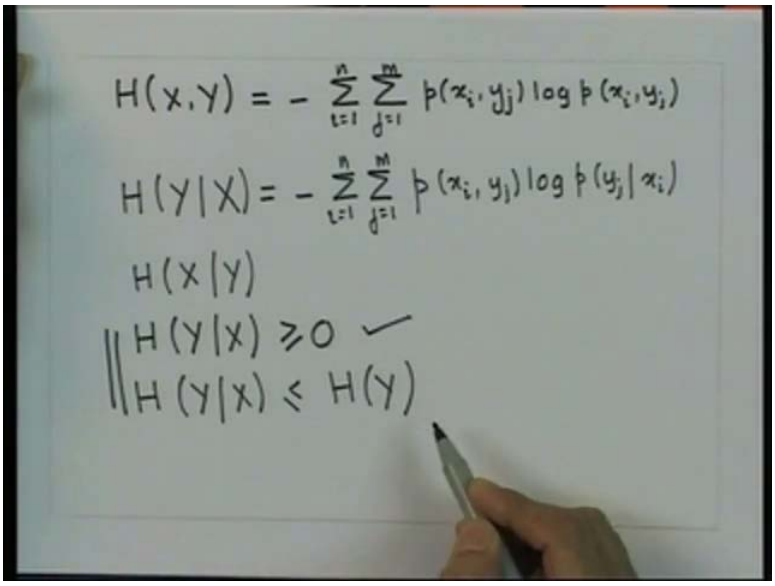
Information Theory and Coding
Prof. S. N. Merchant
Electrical Engineering
Indian institute of Technology, Bombay

Lecture - 5
Properties of Joint and Conditional Information Measures and A Markov Source

In the earlier class, we defined two new information measures. These were joint information measure, and the other was conditional information measure.

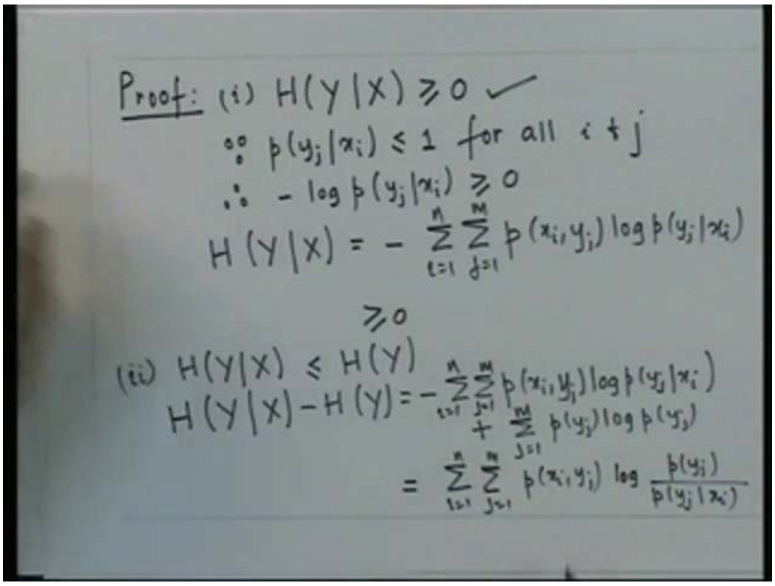
(Refer Slide Time: 00:57)



Joint information measure was given as H of X, Y; i is equal to 1 to n, j is equal to 1 to m. This was the joint information measure, which we define when we have two events taking place simultaneously and we observe them as 1. Another information measure which we define was conditional information measure, and that was defined as H of Y given X was equal to minus, summation over xi yj log of probability, j is equal to 1 to m.

This is a conditional information measure with regard to experiment Y given X, and similarly we can define H of X given Y. Let us look into the properties of H Y given X, little more into depth before we proceed ahead. So, the two important properties of H Y given X would be, one properties H Y given X is always greater than equal to 0. And other property is that H of Y given X is always less than equal to H of Y, with equality if and only if X and Y are statistically independent. Let us try to prove these two properties pertaining to H of Y given X. Now to prove this property, the first property it is not very difficult.

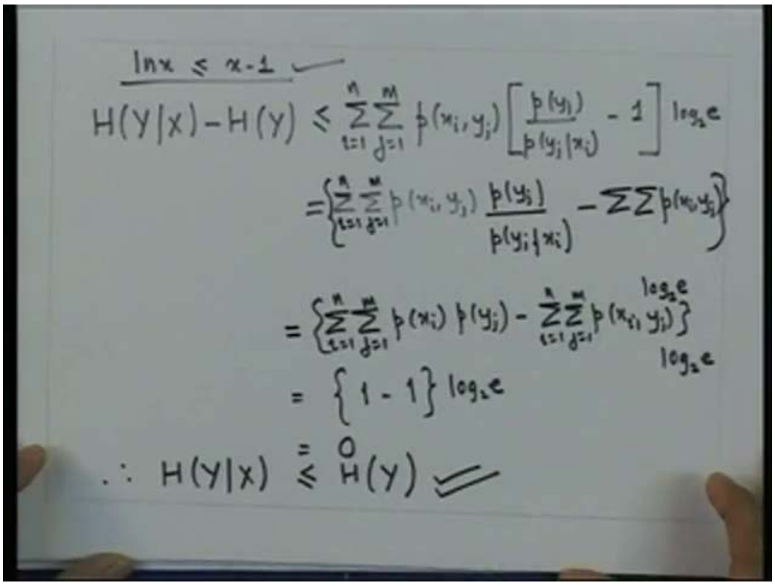
(Refer Slide Time: 03:28)



So, proof the first one we have to prove is H of Y given X is always greater than equal to 0. Now, because probability of yj given xi is less than equal to 1, for all i and j therefore, this implies that minus log of probability, yj given xi is always greater than equal to 0. And so from the definition of H Y X which is nothing but summation over probability of xi yj log probability of, yj given xi. This quantity is positive, this quantity is positive so for totally this quantity is always greater than equal to 0. So, the first property has been proved.

Now to prove the second property, that is H of Y given X is always less than equal to H of Y, it is not very difficult. We write of equation minus H Y is equal to minus summation over probability of xi yj log of p of yj given xi minus H Y, is probability of yj log probability of, j is equal to 1 to m. Here also j is equal to 1 to m, i is equal to 1 to n now, this is very simply can be written as, this quantity out here summation j is equal to 1 to m, p yj can be substituted by double summation and once you do that double summation, this expression can be written as this.

(Refer Slide Time: 06:36)



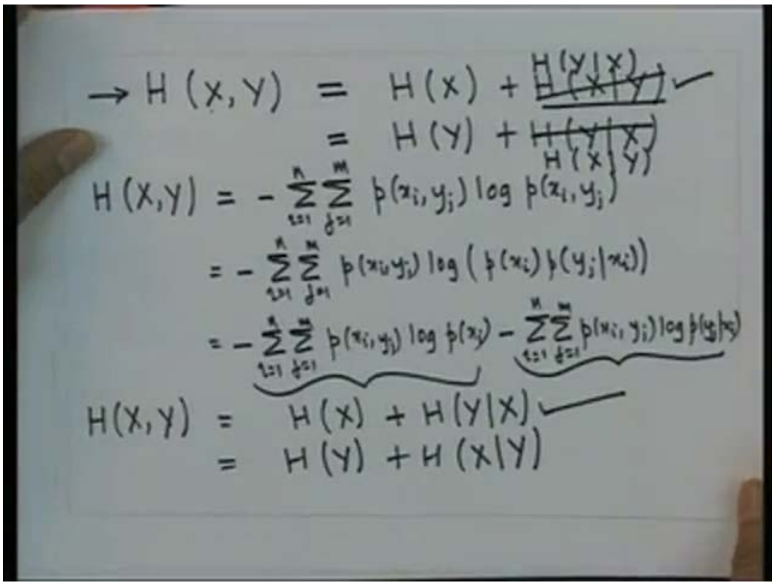
Now, we also have seen that log x is always less than equal to x minus 1. If I use this property then i can write this expression out here, x is equivalent to this quantity out here then I can write that H of Y given, minus H of Y is always less than equal to double summation probability of xi j, probability of yj, i is equal to 1 to n, j is equal to 1 to m. Multiplied by, because this expression is a natural base whereas, we have here log to the base 2 therefore, when we convert from log to the base 1 to the, log to the base e we will get this factor out here.

Now, this can be a simply simplified as minus double summation of this whole thing, gets multiplied by log to the e. Now, writing this expression using the Bayes theorem we will get summation of probability of xi probability of yj minus, probability of xi yj, i is equal to 1 to n, j is equal to 1 to m. Similarly, i will have i is equal to 1 to n, j 1 to m now, this summations are 1 minus 1 and this is equal to 0.

Therefore, what we get is H of Y given X is always less than equal to H of Y, what this expression says is that the uncertainty, which is there in the event Y or experiment Y, after the event X has been observed, will be always less than the uncertainty, which is there initially when I do not observe X. So, when I have the full knowledge about the event X, the uncertainty about the event Y will be always less than the, uncertainty of the event Y, when I do not observe event X. And therefore, the information of H Y given X, will be always less than equal to H of Y.

Now, let us have a look at the relationships between the joint information, conditional information measures and the marginal information. So, I want to find out the relationship, this is my joint information, I have my marginal information measure. And then I have my conditional information measure.

(Refer Slide Time: 11:07)

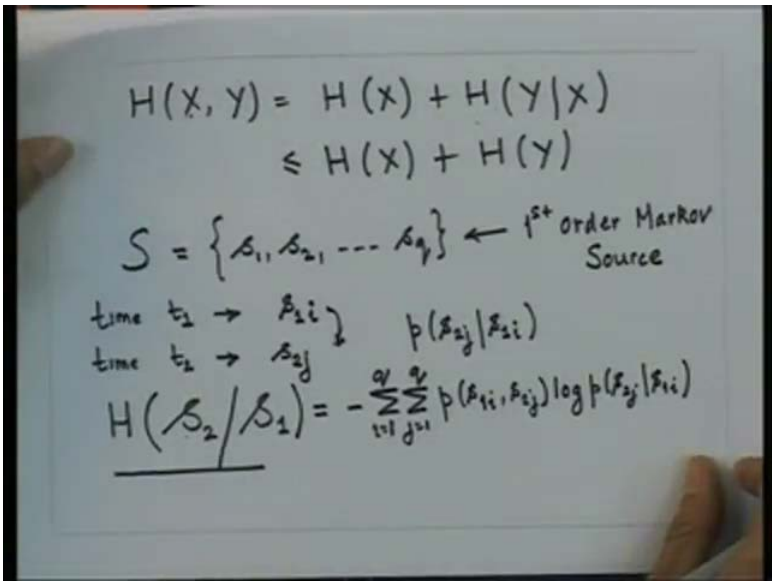


Now, I want to find out the relationship between these three quantities. Now, we will show very shortly that H of X given Y, is nothing but equal to H X plus H of X given Y or I can also write this H of Y plus H of Y given X. So, this is another important relationship which will be using during the course of our lecture today. So, let us try to prove this relationship, let us try to prove the first relationship. Let us look at the definition of H of X given Y, which we just started in the morning.

So, this is nothing but probability of xi j log of p of xi yj. So, let us start with H X Y which by definition is given by this term expression. So, this if I simplify it, I can write it as probability of xi yj log of p of xi probability of yj given xi. This I am writing using the Bayes rule so i is equal to 1 to n, j is equal to 1 to m. And this, I can simply as probability of xi yj log of p of xi minus summation, double summation probability of xi yj log of probability of yj given xi, i is equal to 1 to n, j is equal to 1 to m and similarly, out here.

Now, this quantity out here is nothing but your H X and this quantity is nothing by definition H Y given X. This quantity out here should not be H X given Y, but should be H of Y given X and similarly, this quantity should be H of X given Y. So, finally I get the joint information which I get from two events X and Y is equal to the information, which I get from the event X alone, plus the information, additional information, which I will get from event Y, after the event X has occurred. So, similarly I can show that this is nothing but H of Y plus H of X given Y. Now, H of X given Y is this quantity out here.

(Refer Slide Time: 14:54)



If you look at H of XY is equal to H of X plus, H of Y given X. We just proved that this quantity is always less than, the information in Y alone. So, this quantity out here will be always less than, H X plus H of Y so the joint information in X and Y is always less than, the sum of the information in X and Y. And they only equal when X and Y are independent, with this little background we will move ahead, where we had the left last time. And we were studying the properties of Markov source so let me revisit the Markov source and let us look at, depth into the properties of this Markov source.

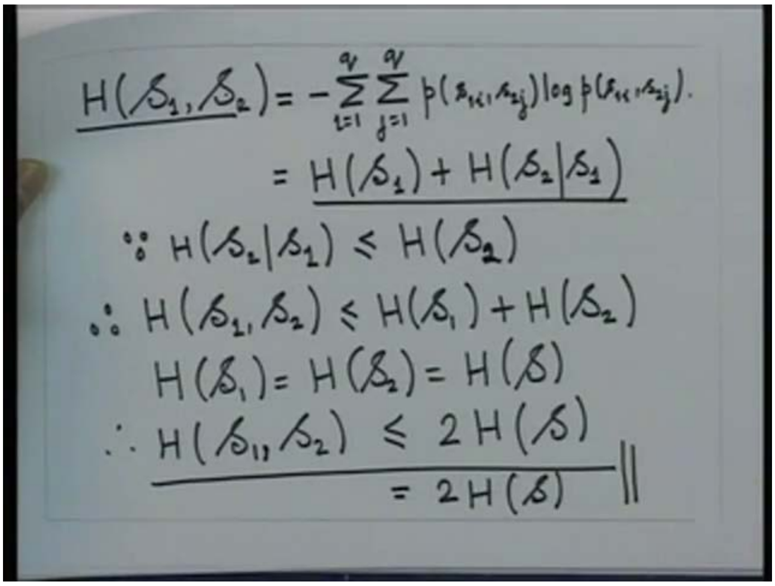
So, I will start again with a first order Markov process so first order Markov source will have its source alphabet. Let me assume as, s1, s2 up to sq the size of the alphabet of this source is q and this is the first order Markov source. What I mean by first order Markov source is that, the occurrence of any particular symbol will be dependent upon, the occurrence of the previous symbol. That is what we mean by a first order Markov source, let me assume that I have a time instant t1, at this time instant t1, let us assume some symbol occurs at this time instant 1.

And let us call that symbol which occurs is i so s1i is one of the symbols, from this source alphabet. And let me assume that, I have another time instance t 2 and at that instant another symbol occurs at time instant t 2 and let us call this s 2 j. s 2 j is again one of the symbols from this source alphabet.

Now, if I were to find out what is the information which I gain, when I transfer, when I go from, when I translate from s1i to s 2 j. Then to find out that information, I will require the conditional probabilities of s 2 j given, s1i. So, if I have this conditional probabilities available, then I can calculate what is the information, which I get when I transit from s1i to s 2 j. Now, considering the time instant at t1 and time instant t 2, as two different experiments with relation, which are related to X and Y. Similar to what we define HYX, we can define the information which I get, additional information which I get, when I transit from s1i to s 2 j.

So, that is very easy to calculate and I can say that H of s 2 to s 1, would be given by this expression. So, this is the amount of information and which I get when there is an arbitrary transition from one state to another state. In the light of what we have done earlier, with relation to the experiment X and Y, I can write that the additional amount of information which I get, when arbitrary transition like this from, t1 to t 2 is given by this expression out here. So, if I were to find out, what is the joint information between s 2 and s 1.

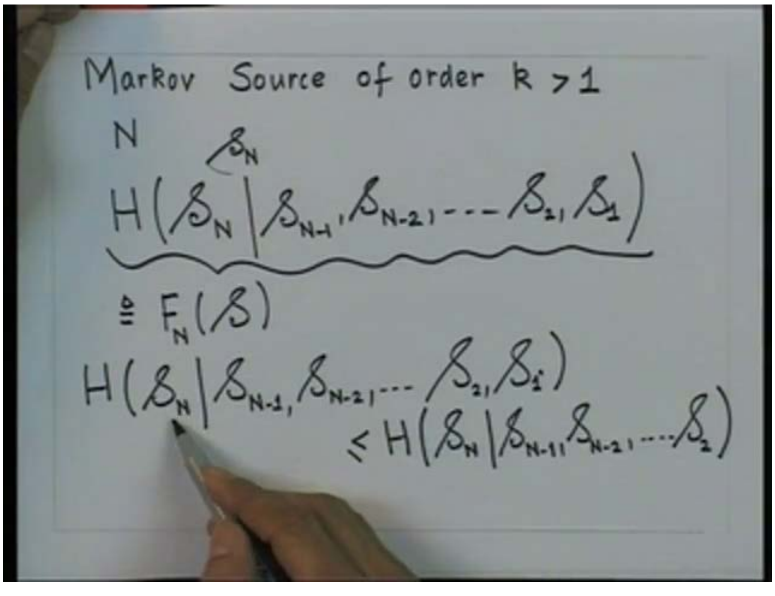
(Refer Slide Time: 20:14)



We can find out very easily as H of s1 s 2, this is the joint information, which I will get from messages of length 2 would be, i is equal to 1 to q, j is equal to 1 to q, probability of s1i, s 2 j log of probability of s1i s 2 j. Now, this can be easily shown to be as H of s1 plus H of s 2 to given s 1. This relationship we have just seen instead of s1 and s 2, we had seen in terms of X and Y. So, it is not very difficult to derive this relationship now, we have also seen that because H of s 2 given s 1, is always less than equal to H of s 1. Therefore, H of s1 s2 is always less than equal to H of s 1. This quantity out here, it should be H of s 2 is this so H of s 1 s 2 is always less than equal to H s1 plus H s 2, which we have derived and if you assume this Markov process as stationary and ergodic, then H of s1 is equal to H of s 2 is equal to H of S. And therefore, H of s 1 s 2 will be always less than equal to twice of H of S.

So, what we have derived now is that, when I look at the entropy of messages which are of length two symbols. And the symbols come from a Markov process of first order then the total information in the messages of symbols of length two, turns out to be less than equal to twice of H of S, HS. Now, if the process, if this Markov process of 0 order then I would have got is equal to twice of HS. So, the conclusion is that whenever the dependency among the symbols then the messages of length L, the total information will be there in that, will be always less than equal to 8 times the entropy of the source. We had seen this result earlier, where we had proved that if the symbols are independent then H of s 1 s 2 turns out to be twice HS. Now, this we had proved it for the first order Markov process. Now, let us move to the sum higher order Markov process, and let us make the things little more generic.

(Refer Slide Time: 24:29)

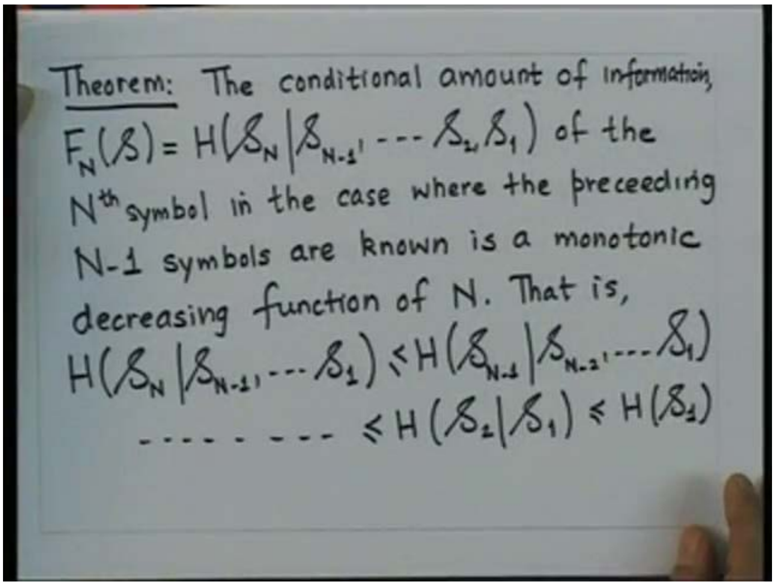


So, I will consider Markov source of order k, which is greater than 1. And at the moment I am interested in the occurrence of a particular symbol at the time instant say, capital N. So, let me say that the symbol which occurred a time instant capital n, let me denoted it as sN. And if I assume that this symbol, which occurs from the source S is coming from a Markov process. Then what I am interested is, I am interested in the additional information, the average additional information, which I get on the occurrence of a symbol, at a time instant sN, given that I have observed all the preceding symbols, right from time instant 1 up to time instant N minus 1.

So, if I use the general properties of entropy then I can define the average additional information, which am going to get on the observation of sN, having observed the preceding symbols will be, nothing but this quantity sN minus 1, sN minus 2 this will continue up to Y. So, this is the additional information which I get when I observe, symbol sN at time instant N now, this quantity I will define it to be as by definition I will call it as FN of S. Now, before we go ahead there is some interesting properties of these FNs. One interesting property would be that H of sN, given s of N minus 1, N minus 2, s2, s1 is always less than equal to H of sN given sN minus 1, sN minus 2 up to s2.

Now, it is not very difficult to prove this ((Refer time: 27:33)) it is very satisfying that, the occurrence of s1, cannot increase the uncertainty of the occurrence of sN. So, what it, what this implies, the relationship implies is that, what this relationship implies is that, the knowledge that is delivered by the first symbol s1, cannot lead to an increase in the uncertainty, about the NH symbol. But it will always decrease or leave it unchanged so this is the significance of this expression. Now, there is a very important theorem, which will try to derive from this.

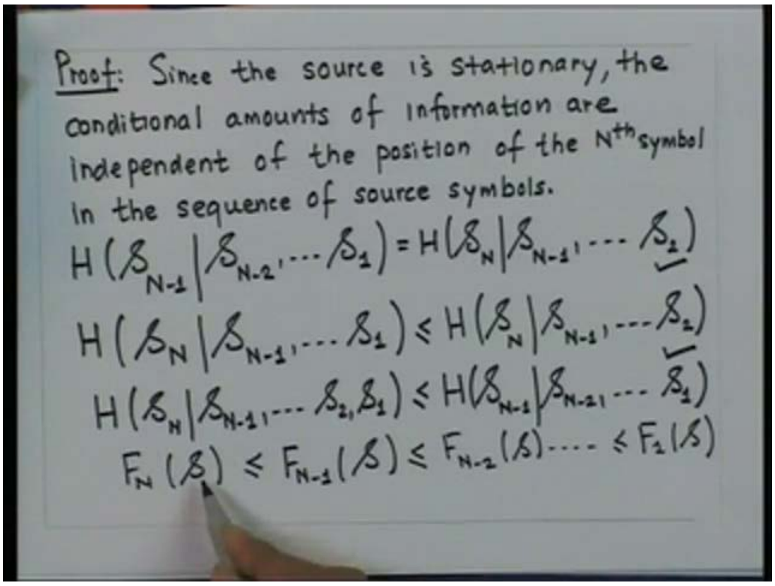
(Refer Slide Time: 28:49)



Theorem says, the conditional amount of information that is FNs, which is by definition equal to H of sN given sN minus 1, up to s1 of the Nth symbol. In the case where, the preceeding N minus 1 symbols are known, is a monotonic decreasing function of N. What we mean by that is, H of sN given N minus 1 preceeding symbols, will be always less than equal to H of sN minus 1 given sN minus 2 to sN1 and this way we can. So, this is a very important theorem, which is associated with a general Markov process.

So, this is what we had defined as additional information, which I get on the occurrence of the Nth symbol, when I know the preceeding N minus 1 symbol so what it says that information which I get from here, will be always less than or equal to the information which I get, when I go back to the time, instant N minus 1. And observe the symbol and given that, at that time instance N minus 2 preceeding symbols have been observed. And if I continue like this finally, I land up with the first symbol. So, the uncertainty which I have, when I observe the first symbol is the maximum compared to the uncertainty, which I have at the Nth instant of time. We will try to prove this theorem.

(Refer Slide Time: 32:40)

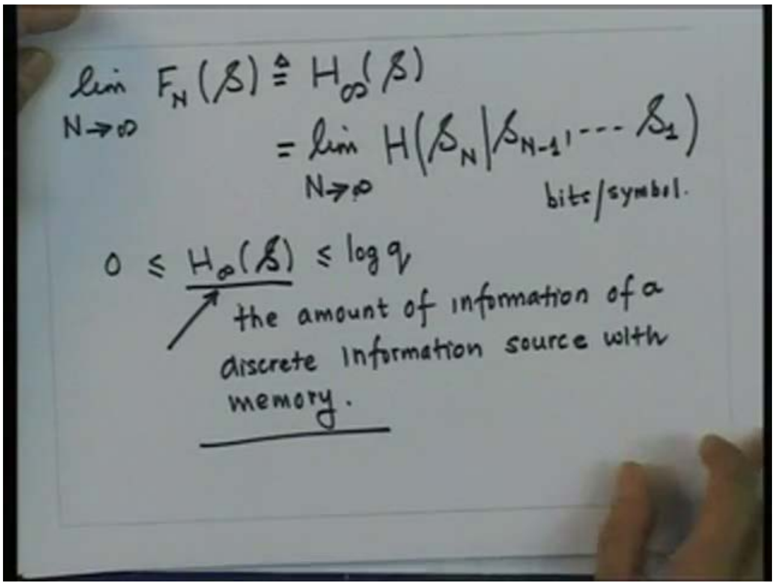


So, let us look into the proof of this theorem since, the source the Markov source which we are considering is stationary, the sources stationary. The conditional amount of information is independent of the position of the N th symbol, in the sequence of source symbols, which are being emitted. So, what it means is that H of sN minus 1 given sN minus 2, s1 is equal to H of sN given S of N minus 1, but this will go up to s2. I can write this expression because the sources is stationary.

And we have just seen that, the property of a Markov source is conditional information, will be always this will be always less than the quantity on the right hand side. This I can write because the source is stationary and this is the property of the additional information measure and from these two it directly follows that, H of sN given sN minus 1 up to s1 is less than equal to H of sN minus 1 given sN minus 2. So, using these two properties I get this so this is by definition, nothing but FN of s is less than equal to F of N minus 1 s.

So, similarly I can extend H of sN minus 1 given sN minus 2 s1 is less than equal to H of sN minus 2 given sN minus 3 up to sN 1. So, I can extend like this and simply show that this condition is true. Now, what it follows that as I keep on increasing N, the value of FNs keeps on decreasing now, since FNs is always greater than equal to 0.

(Refer Slide Time: 37:14)

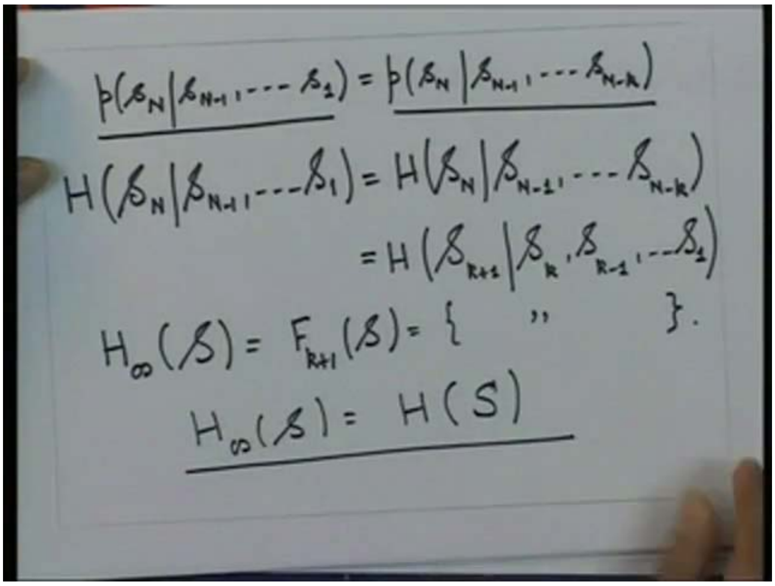


What this implies is that, limit of N tending to infinity of FNs, should converge and let me call that limit as by definition H infinity s. And this is nothing but limit of N tending to infinity of additional information of sN given N minus 1 preceeding symbols have been observed. This is bits per symbol and from this expression out here, because this is expression from these two expressions I can write, 0 is greater than, this expression is F1s

Now, this quantity out here, the quantity which I get when, N tends to infinity of FNs is by definition is known as, this quantity is the amount of information of a discrete information source with memory. Now, we have formally defined the information measure for a Markov source with memory. We have considered the value of k to be arbitrary, the k stands for the order of the Markov process. So, what we get is that, if I want to calculate the entropy of a Markov process then entropy of the Markov process is nothing but limit N tending infinity of FNs.

And that is nothing by definition limit N tending to infinity, the additional amount of information. So, this by definition is the, definition for the information measure of a Markov source memory. Now, if you assume a Markov source of order k then what this implies is that as I keep on increasing the value of N, beyond certain value of N, this FNs will not fall. This is very easy to appreciate.

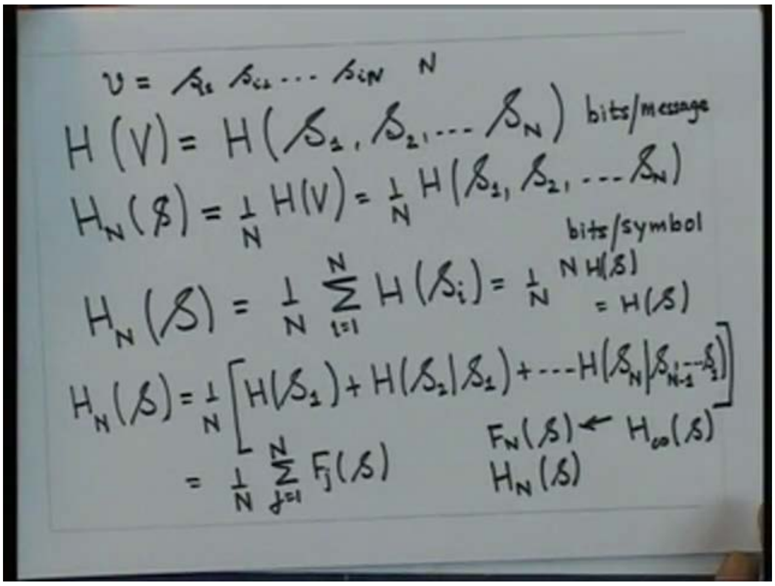
(Refer Slide Time: 40:40)



Because probability of sN given sN minus 1 up to s1, will be equal to probability of sN given sN minus 1 up to sN minus k, when the Markov source is of a kth order. And in this, relationship is valid then H of sN given sN minus 1 up to s1, would be equal to H of sN given sN minus 1 up to sN minus k. Because, only k preceeding symbols will come into picture and because the Markov source is ((Refer Time: 41:36)) and stationary I can write this as H of sk plus 1 given s of k, s of k minus 1 up to s1.

So, for a Markov source of order k H infinity s will be equal to, nothing but F of k plus 1 s is equal to this quantity. Because for N beyond k plus 1, for N beyond k plus 1 this quantity will remain constant and it will not go lower than this value. And then I can write H infinity s is equal to this value, for 0 memory source k is equal to 0 and in that case, H infinity s is nothing but H of S. So far we have considered the symbol, the messages with symbol length of unity. The next question is like, we had done earlier if I look at messages, of length other than 1 then what happens to the entropy of those messages. Let us look into that.

(Refer Slide Time: 43:34)



So, let us assume that I have a message v, which is composed of N symbols of this Markov process. So, I have s1, si1, s12 up to sin capital N. So, what I do basically is that, I assume that I have messages now of N symbols, capital N. I had messages of N symbol now, if I look at these messages, and if I were to find out the entropy of this then how this entropy is related to my original entropy. Let us analyze this so if I want to calculate entropy of this then I can write entropy of H V, is nothing but H of s1 s2 to sN. So, this is the information which I get from messages of symbols consisting of length N, capital N.

Now, if I define another quantity as average information per symbol, which is by definition equal to 1 by N of H V so this is my average information which I get per symbol. So, this is equal to 1 by N H of s1 s2, this will be bits per symbol. Now, obviously the symbols are statistically independent then H of sN would be, nothing but summation of H of si. This expression I can write, for this provided my symbols in this message of symbol length N are independent.

And in that case, I can simplify this to be as H of s now, if the symbols are dependent then I cannot write like this. And then I can go to more fundamental definition, I can say that H Ns is equal to 1 by N. And this information out here, joint information in N symbols can be written as H s1 plus H of s2 given s1 plus H of sN given sN minus 1 up to s1and this is equal to 1 upon N times.

So, this I get, average information per symbol and that is related to 1 by N summation of Fjs. Now, that was the case for FNs we can similarly, show that H of Ns is a mono tonic decreasing function of N. And it will be interesting to find out that, the limiting value of HNs turns out to be the same, as the limiting value of FNs. The limiting value of FNs we have seen, it was H infinity s and this was the entropy of a Markov source, that is what we have defined. Now, we can also show that the limiting value of HNs as N tends to infinity, also turns out to be H infinity s, we will try to prove this in next lecture.