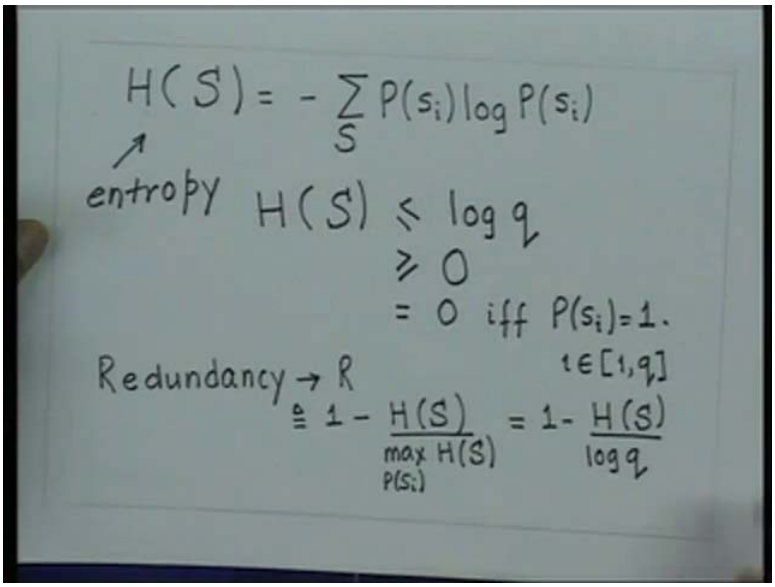
Information Theory and Coding
Prof. S. N. Merchant
Department of Electrical Engineering
Indian Institute of Technology, Bombay

Lecture - 3
Extension of an Information Source and Markov Source

In the previous class, we had a look at the information measure in terms of entropy of a source.

(Refer Slide Time: 00:56)



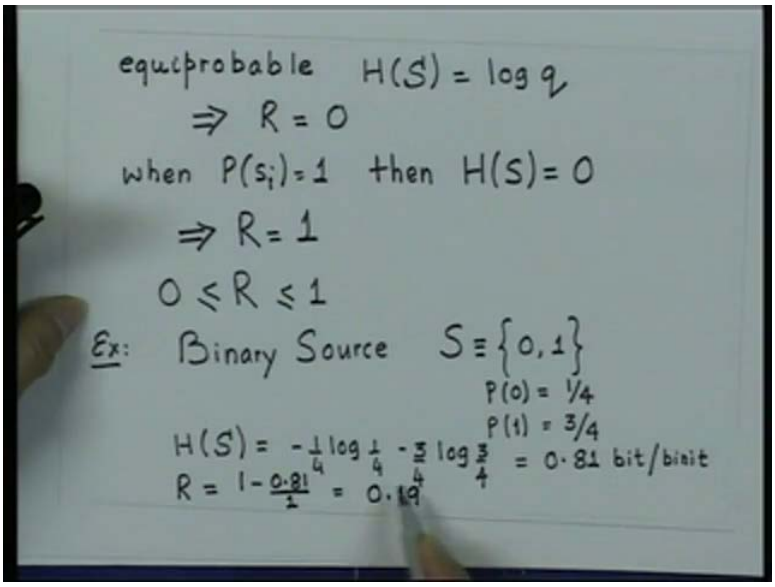
Entropy of the source was given as H S equal to minus summation of P s i log P s i. this is what we had defined as the entropy of a zero memory source. Interpretation of entropy is average information, which I get per symbol of the source S. We can look at the concept of entropy in a slightly different manner. I could say that entropy is also a measure of uncertainty that gets resolved when that event takes place. So, when an event e occurs, I get some information on the occurrence of that event e. A different way of looking at the same problem is to say that when I observe the event e, whatever uncertainty was associated before my observation, that gets resolved on the observation of that event e.

So, entropy of the source S could also be interpreted as uncertainty resolved when I observe a particular symbol be emitted from the source. The concept of uncertainty in terms of, in terms of, the concept of uncertainty will be utilized when we are talking of

mutual information during the course of our study. We also had a look at some properties of entropy and we came to conclusion that entropy of a source is always less than equal to log q, where q is the size of the source alphabet s. And we also saw that H S is always greater than equal to 0, it is equal to 0 if and only if probability of s i is equal to 1 for some i belonging to 1 to q. When this condition is satisfied, then value of entropy I get is equal to 0. For any other case other than this, the value of entropy I get is always greater than equal to 0, but less than log q. Associated with entropy of a source, I can define another quantity that is known as redundancy of a source.

The definition of a redundancy of a source is given as it is 1 minus H S. H S is the actual entropy of that source S and what is the maximum entropy which I can get from the for that source S; that maximum entropy obviously will be dependent upon the probabilities of the symbols of the source alphabet. For the case of a zero memory source, this can be written as 1 minus H of S upon log q because the maximum entropy of zero memory source is given by log q. If you take, let us look at the property of the parameter redundancy.

(Refer Slide Time: 05:48)



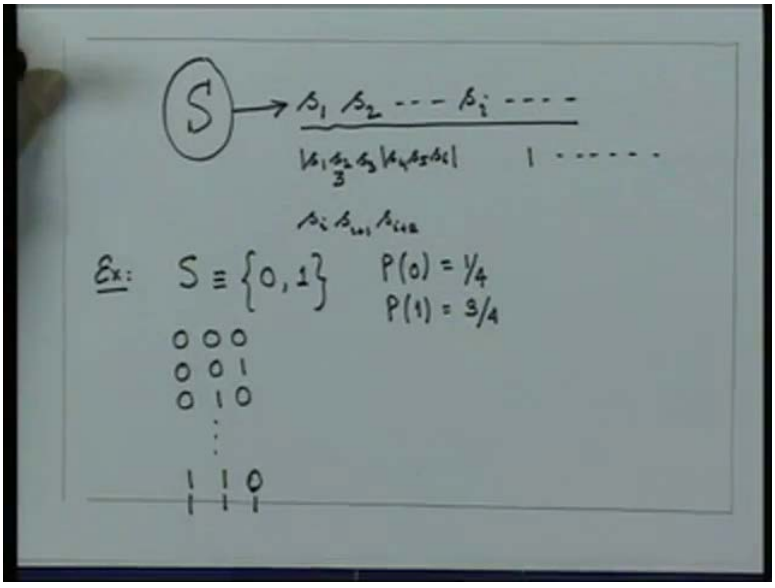
When you have equi probable symbols, when you have equi probable symbols, you have H S, actual H S is equal to log q and this will imply that the value of redundancy R is equal to 0. When P s i is equal to 1 for some symbol s i in the alphabet S, then H S is equal to 0. This implies that R is equal to 1. So, the value of your redundancy will be

always lying in between these two values, 0 and 1. The lower bound is 0 and the upper bound is 1.

Let us take a simple example to get the feel of this factor, which we have defined recently that is redundancy. Let me take a simple case of a binary source. So, I have a binary source. Let me assume that binary source alphabet is 0 and 1 and the probability of symbols is given as one fourth that is the probability for 0 and probability of 1 is given as three fourth. Now, I can simply calculate the entropy of this source as H S equal to minus one fourth log of one fourth minus three fourth log of three fourth and this and turns out to be 0.81 bit per symbol. In my case, the symbols are binary digits 0 and 1.

We will call the binary digit as binits. So, we can say that entropy is 0.81 bit per binits. For this source, my redundancy would be 1 minus 0.81; log q would be equal to 1, so the value which I get is 0.19. So, I can say that roughly there is a redundancy of 19 percent in the source S. Now, all this time, we have been looking at a source, which emits symbol individually. So, what I mean by that.

(Refer Slide Time: 09:44)



If I have a source S, then this source S emits symbol, this symbol belongs to the source alphabet and the probability of occurrence of that particular symbol is also given to me, but I looked at the emission of the symbols from the source individually. So, I had s 1, s 2, s i continuously like this and we found out the average entropy, average information that is nothing but the entropy of the source for a symbol.

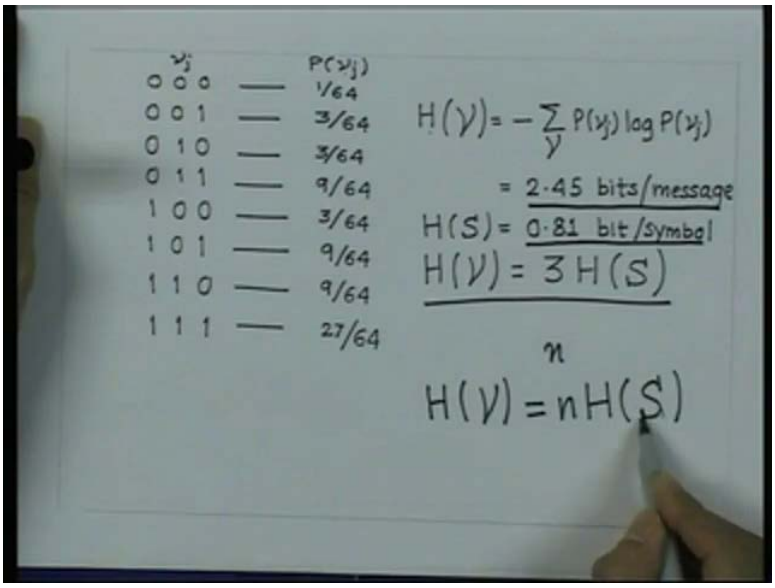
If I assume that this output sequence which I get and I block them in terms of let us consider that this output sequence, which I get out here, we look at the output sequence in terms of blocks. For time being let me assume that I start looking at the output sequence in the blocks of three symbols. So, this would be one block, the second block will be like this, continues like this, I can look at the output sequence from this source in terms of blocks.

Now, when I start looking at this output sequence in terms of block, what I could consider is that I am forming new messages or sub messages out of this string. This sub messages which I have are nothing but they are being formed out of symbols, which are being emitted from this source S. So, in our case, this is block length of 3. So, I have messages of length 3.

Now, if I start looking this in terms of messages and if I were to ask you that, what is the information, the average information which I get per message from this source; let us look at this example little more specifically. I consider the previous example which for which we calculated redundancy, the same example I consider. So, I have a source given by the source alphabet, which is 0, 1, the probability of 0 as one fourth and probability of one as three fourth.

Now, the output of the sequence from this source S will be looked in terms of blocks of length 3. So, in that case, the number of sub messages which I can form from the block of length 3 are nothing but 0 0 0 0 0 1 1 1 0 and finally, I have 1 1 1. So, these are the number of messages, which I can form from the source S if I start looking the sequence of the output in terms of blocks of three. How do I find out the information average information per message for all this eight messages? It is very simple.

(Refer Slide Time: 14:10)



What we can do is these are the number of messages, which I have different messages, I can find out what is the probability of occurrence of each of this sub messages. Now, if I assume that my symbols are independent, then probability of getting 0 0 0 is nothing but one fourth, multiply by one fourth one fourth and that is what I get 1 by 64. Similarly, I can find out the probabilities of occurrence of these messages, which I call by v j, j ranges from 1 to 8.

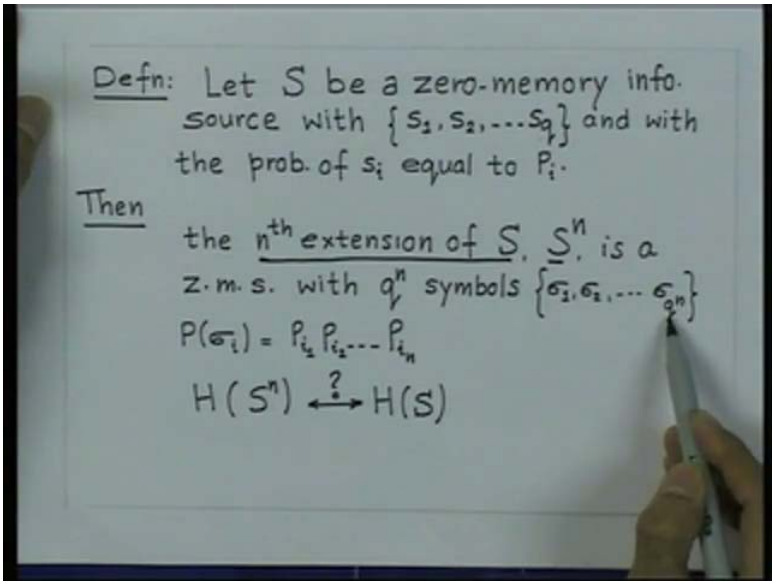
Now, going by the definition of the entropy, I can define the entropy or the average information which I get from the source per message would be nothing but given by this simple formula, which we had seen earlier too. If you calculate, just plug in this values out here into this formula, what value I will get is 2.45 bits per message and we had just looked that the entropy of the binary source, when I look at the sequence in terms of symbol being emitted individually, then I get 0.81 bit per symbol. So, the relationship between H V and H S turns out to be H V is equal to 3 times H S.

This is simple example, which I took to show the relationship between a new source that is V and the old source S, when I start looking at the output of the sequence from the output of the sequence from the source S in terms symbols in block lengths of 3. Instead of looking block lengths of 3, suppose I start looking in block lengths of n. Then what is the relationship which I will get between the new source V and my old source? It is not very difficult to prove and we will do very shortly that what it will turn out to be is

nothing but n times H S. This is valid only when source my original source S is a zero memory source.

What is the advantage of looking at the source in this form? When we do coding, we will see that when we start looking at the original source in terms of blocks of n symbols, then it is possible for me to carry out the coding which is more efficient than when at not looked at the source in this form. So, with this motivation, we will go ahead and try to define this new source generated from the primary source in terms of symbols of length n.

(Refer Slide Time: 18:13)



Let me formally define this. Let me assume that I have a source S, which is a zero memory information source. This zero memory information source will have its source alphabet. I give the source alphabet as s 1, s 2 and s q. In the earlier case, which we saw the example s 1, s 2, s q, we just had 0 and 1. There were only two letters in that alphabet and with each of this symbols in the alphabet or letters in the alphabet, I have the probabilities of s i given and let me assume that probability of s i is equal to P i.

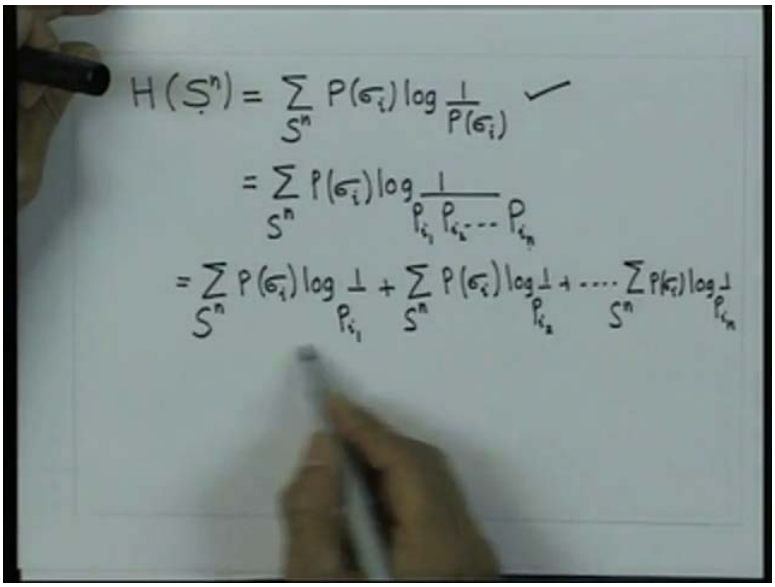
Then, the n th extension of S, which I am going to denote by S n is again a zero memory source with q raised to n symbols. So, the new alphabet which I generate for the nth extension of the source s that is S n will be consisting of q n symbols. I denote those q n symbols as sigma 1, sigma 2 up to sigma q raise to n. Each of these symbols out here in

the new source is nothing but a string of symbols, which come from my old source of primary source S and the length of sigma 1 is n of S size.

Similarly, sigma 2 would be another symbol of the n th extension, which I generate by having a string of symbols from my primary source S. So, I know basically what is my source alphabet for the nth extension of the source S. We have seen in the earlier class that if I want to define my source along with the alphabet, I require the probability of symbols.

So, let me assume that probability of the symbol in the new x source that is S n are given by probability of sigma 1, probability of sigma 2, probability of sigma q n and any one of this sigma i is related to the probabilities of symbols in the original source S. That is not very difficult to show. Now, the question is I have my entropy of the new source, I have the entropy of the old source, how are these two entropies related? We already know the answer. What we expect is it should be n times H S. Let us see whether we can prove this formally.

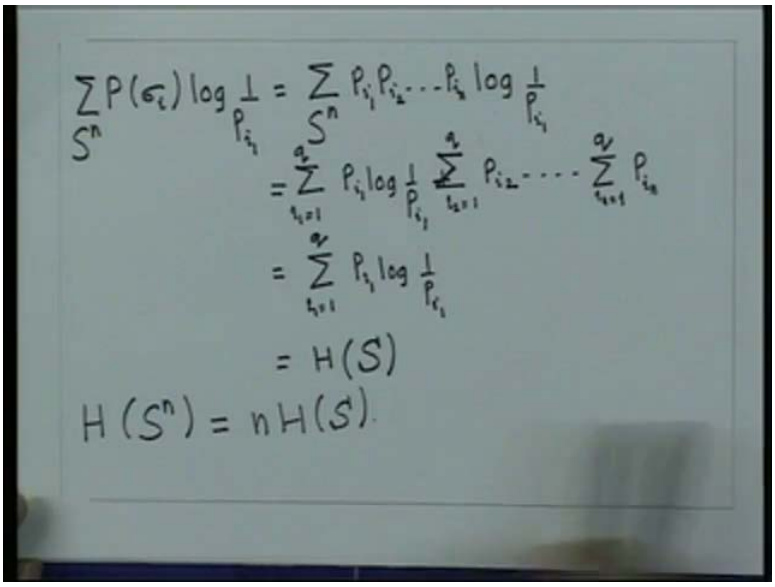
(Refer Slide Time: 21:56)



So, the entropy of my n th extension, which is given by S n H S is nothing but this formula. Now, we can simplify this formula as I write probabilities of sigma i's as nothing but probabilities of i 1, i 2 up to i n. This here when I am writing this, I am assuming that the sequence is such that the symbols in this sequence are independent. Now, this I can simplify as this summation is over source alphabet S n. Now, this I can

simplify as P of sigma i log of, I can break up into n summations. So, finally, the last will be, now let us look at one of this term, let us see whether we can simplify this term.

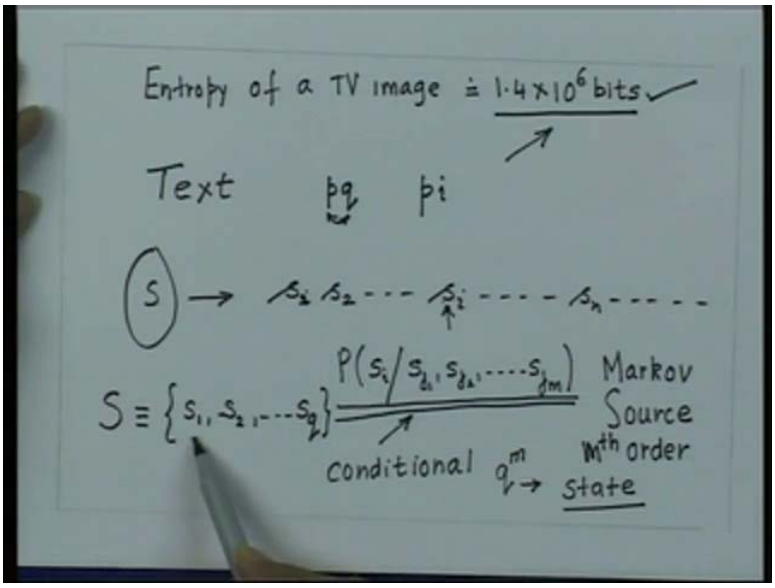
(Refer Slide Time: 24:12)



So, I take the first term in that summation, which is this. I again break up probabilities of sigma i in terms of my probabilities of original symbols. Now, this summation will be done over the alphabet S n. Now, this summation itself can be broken up into n summations as follows, the multiplications and finally, you have and obviously because the summation is out here are all 1, this is nothing but q i 1 equal to 1, P i 1 log of and this is by definition entropy of my primary source or the original source S.

Now, so my final expression for the entropy of nth extension resource will be I have shown that this is the entropy I get for the first term here. So, similarly, I can show that this is H S, this is H S and I have n number of terms. So, finally, I get this value to be equal to n times H of S. This we had seen with an example where I had n equal to 3 and we verified the same thing. As I have said that motivation for studying the nth extension of a source will be when we are trying to code a zero memory source, we want to design efficient codes. We will have a look at this little later in our course.

(Refer Slide Time: 27:45)



In the previous class, we had calculated an entropy of a TV image and entropy of that TV image can be calculated was roughly around 1.4 into 10 raise to 6 bits. At that stage, I had pointed out that the calculation of the entropy, which we have done for the TV image is not really exact. In a practical situation, you will find that the entropy of a TV image is much less than this quantity.

The reason is that when we calculated this value, we assume that each pixel of pel in the TV image was independent. In a practical situation, really this is not the case. This is one example of a source, where you have the symbols or the pels to be very specific in that case. In our case, they are not independent, they are related to each other and because of the inter relationships between this pels, when we calculate the entropy of a real TV image, we will find that this value the real value turns out to be much less than what we had calculated based on the assumption that is a zero memory source.

Another example is if you look at English text, you will find that the occurrence of the characters in the English text is not independent. For example, p followed by q, these combinations will be much less compared to p followed by i. So, if you look at the text string, and if you try to calculate the information based on the assumption that each of the characters are independent and calculate the entropy or the average information, which I will get from that same text string based on the fact that there is a relationship

between the characters. Then you will find that the entropy calculated in the later case will much less than the entropy calculated in the earlier case.

So, let us look at those sources where there is a dependency of symbols in the sequence of strings coming out from the source S. Let us try to look at that more formally. So, if you have a source let us say S, which emits s i, s 1, s 2, s i continuously it emits symbols. Now, so far we have assumed that all these symbols are independent. What it means that probability of a occurrence of a particular symbol at this instant is not dependent on the occurrence of the previous symbols, but in a practical situation, what will happen that the probability of occurrence of the symbol s i at a particular instant i will be dependent upon the presiding symbol.

So, let us take a simple case where I find that the probability of occurrence of a particular symbol at this instant say s i is dependent upon the occurrence of the presiding symbols. So, let me assume that it is dependent upon the previous symbols. In this case, I assume that occurrence of s i is dependent upon previous m symbols, s j m is earlier to s i and s j 1 is farthest away from s i.

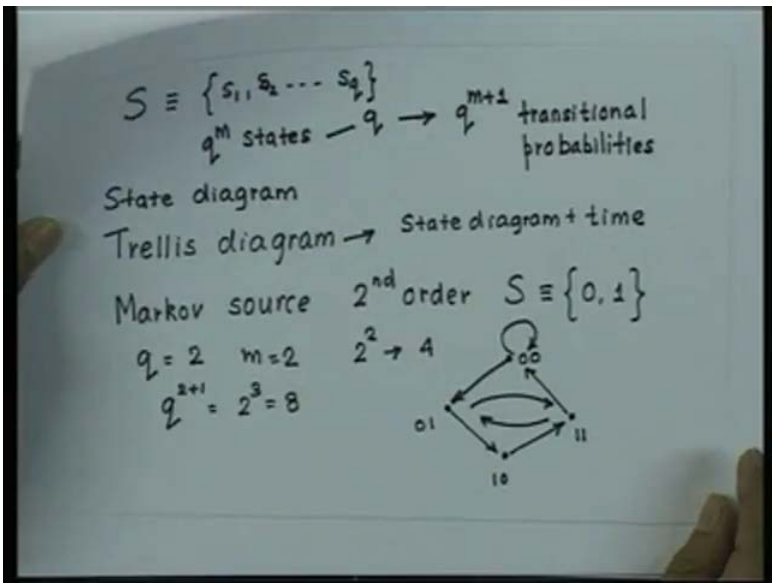
So, in this case, when you have this kind of dependencies, then this is known as a Markov source and for this specific example, since the dependency is over m preceding symbols, I will say that this Markov source is m th order. So, if you have a Markov source of first order, it means the probability of occurrence of a symbol is dependent upon the preceding symbols. If I have a Markov source of order two, then it is dependent upon preceding two symbols. Now, if you want to identify such sources Markov sources, then what is required to specify is you should again specify, what is the source alphabet?

So, Markov source will be identified by the source alphabet. Let me assume this case. Also, the source alphabet consists of few symbols or few letters and since the occurrence of a particular symbol in the sequence is dependent upon m preceding symbols, then just the probability of occurrence of the symbol is not enough for me to characterize this source. To characterize this source, what I need is this kind of probabilities and this are known as conditional probabilities. So, along with the source alphabet, I should provide conditional probabilities.

Now, at any particular instant of time, this symbol can take any of the values for this source alphabet. So, it can take q values. Now, emission of these values is not

independent. It is dependent upon the previous m preceding symbols. Now, each of this m preceding symbols can take the values from this source alphabet. Therefore, the number of possibilities of this m preceding symbols will be q raise to m and each of this possibility is known as state. Once I know the state, with each of the state, I have to specify q conditional probabilities q conditional probabilities associated with the length of the alphabet which I have.

(Refer Slide Time: 36:54)



A Markov source S which is identified now by source alphabet and conditional probabilities since there are q raise to m states and with each state, you have q transition probabilities, therefore you will have totally q raise m plus 1 transitional probabilities. Therefore, to specify a Markov source of m th order, in this case, I will require this alphabet. I will require q raise to m plus 1 transitional probabilities. How do you depict a Markov's source? Is it possible to present or represent this Markov source in a form which describes the source completely?

One way to do that is with the help of what is known as state diagram. The state diagram basically is used to characterize this Markov source. Another way of depicting the Markov source is with source is with the use of what is known as trellis diagram. The difference between trellis diagram and state diagram is that in trellis diagram, the state diagram is augmented with time. So, with trellis diagram, you have state diagram plus

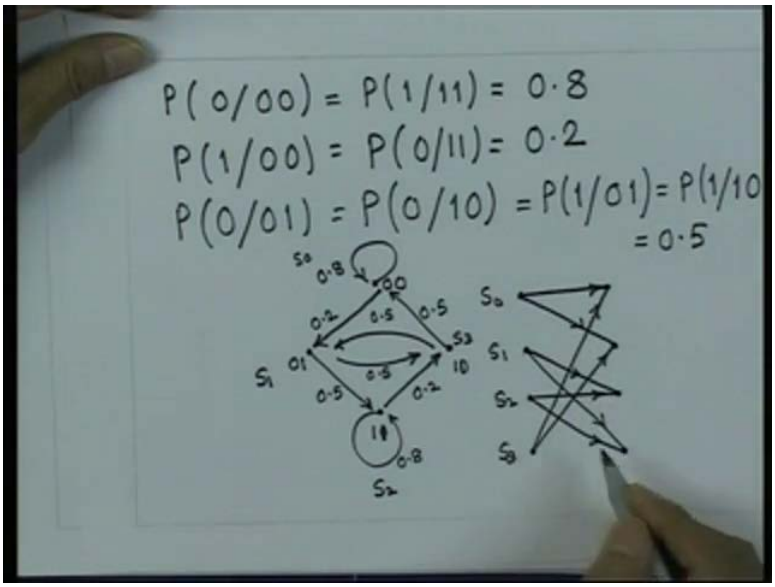
time. Trellis diagram tells me basically at any particular instant what is the state of the source; that is not very clear just from the state diagram.

So, I would say that the state diagram more concise form of representation, whereas trellis diagram is a more elaborate form of representing a Markov source. Let us take a simple example to understand this. If I have a Markov source of second order and let me assume that I have my source as again given by this source alphabet where the binary symbols are there, then if I were to represent this source in terms of s state diagram, then the way to do it is since q in this case is equal to 2, m is equal to 2, the number of states which I have is 2 raise to 2 and that is equal to 4. You represent these states by dots.

So, in our case, I will have four states. I represent them by this four dots and this four states can be identified as 0 0, 0 1, 1 0, 1 1. I will require the conditional probabilities for this source S since we have q is equal to m is equal to 2 we get m is equal to 2 plus 1 that is equal to in our case 8. So, I should specify eight conditional probabilities and these eight conditional probabilities are depicted in this state diagram by arrows. For example, there could be arrows running from one state to another state like this. So, arrows basically indicate what is the conditional probabilities?

Now, to be very specific, let us take an example. Let me assume that probability of 0 given 0 0 is equal to probability of 1 given 1 1 is equal to 0.8, probability of 1 given 0 0 is equal to probability of 0 given 1 1 is equal to 0.2. And probability of 0 given 0 1 is equal to probability of 0 given 1 0.

(Refer Slide Time: 43:10)



Probability of 1 0 1 is equal to 0.5. If I were to depict this in a state diagram form, then since there are four states, I can indicate this four states by simple dots as here. These are nothing but 0 0, 0 1, 1 0, 1 1 make this 1 1, 10 and these are the arrows. This will be 0.8 because the probability of 0 given 0 0 is 0.8 and when 0 occurs, it will again go into the state 0 0.

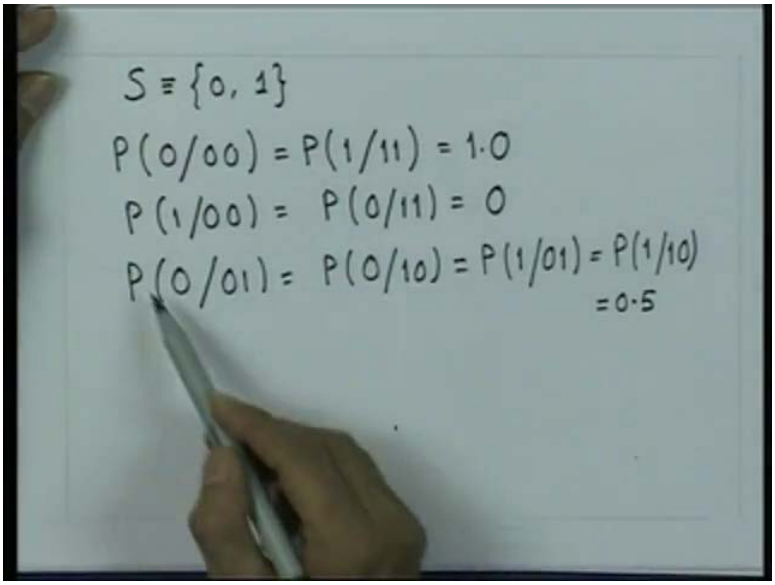
So, this is what it means. Then I have this when it is in state 0 0, when 1 occurs, it will move over to state 0 1. So, this is the arrow that indicates moves from 0 0 to 0 1 and then from this, I have 0.5, I have 0.2 and 0.5 and probability of 1 occurring given 1 1 is 0.8. Now, same thing can be indicated with the help of a trellis diagram. What you have to do is basically at any instant of time, you draw four states. Let us indicate the four states are s 0, s 1, s 3, s 0 corresponding to 0 0, s 1 corresponding to 0 1, s 2 corresponding to 1 1, s 3 corresponding to 1 0.

Now, you look basically at a next instant of time, you can have again four states. So, s 0 can go from s 0 to s 0. So, you can have one arrow going from s 0 to s 0 or s 0 can go to s 1. So, I have like this. These are two branches, which take off from s 0. Similarly, if you look at s 1, this is my s 1; s 1 can go from s 1 to s 2. So, I have s 1 going from s 1 to s 2. This is my s 2 state; this is my s 3 state. It is my s 0 state. There should also be a link between this and this is again 0.5, 0.5. So, I have state from s 1 to s 2 or it can be from s 1 to s 3. So, it is for s 2, I have from s 2 to s 3, s 2 to s 3 or from s 2 to s 2 itself the way.

Finally, for s 3, I can go from s 3 to s 0. I write it down like this and s 3 can go to s 1. So, this is another time instance.

Similarly, for each time instance, I can keep on connecting like this. So, you can specify the exact part with the source follows using this trellis diagram. So, with the help of a trellis diagram, it is possible to find the exact sequence of the symbols being emitted with reference to time. This is another form of representation for the source S. Now, these are important properties of this source S. To understand those important properties, let me take another simple example.

(Refer Slide Time: 48:03)



Suppose, I have a source S, which is given by the same source alphabet 0, 1, but conditional probabilities are given like this, a small difference from the earlier example which we just saw. Now, if I, again this source is a second order source, if I were to depict the source in terms of a state diagram, then what I would get is something like this. Now, there is something very interesting about this source. What this state diagram shows that there is a probability that you will always keep on getting 1s or you will always keep on getting 0s. Actually, this is not complete. I should have something like this.

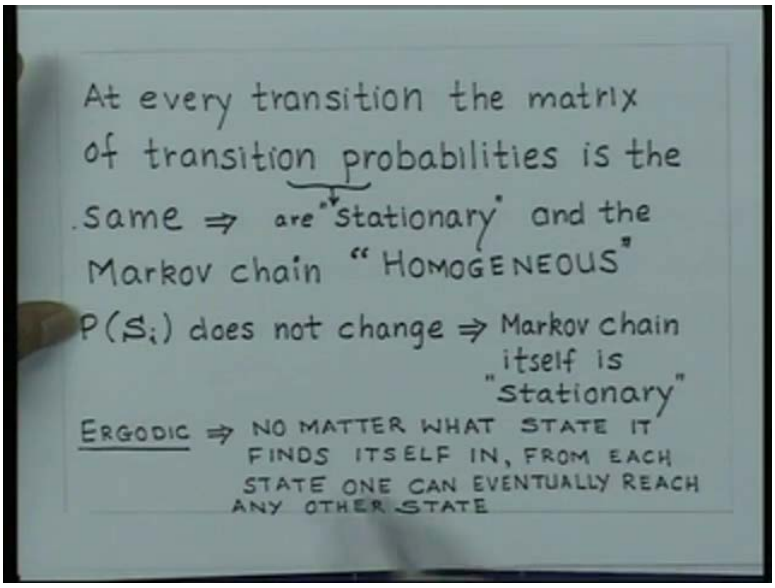
So, initially I start the source at particular state. Let me assume that the source starts at any one of the states 0 0, 0 1, 1 1, and 1 0 and the probability of this happening are equal, so one fourth, one fourth, one fourth, one fourth. Once it is one state in long run, you will

find that this source either emits just all 1s or emits all 0s. So, what is the difference between this source and the earlier source which we saw? We find that in this source, once I am in this state this state, it is not possible for me to come out of the states, whereas that was not true in the previous case.

What is the difference between this? Technically, we would say that this source is non ergodic, whereas so this is I would say state diagram of a non ergodic second order Markov source, whereas this state diagram is for second order Markov source, but this is ergodic. Without going into the mathematical intricacies of the definition for ergodicity, we can simply define as an ergodic Markov source as a source, which observe for a long time. There will be a definite probability of occurrence of each and every state in that source. In this example, I had four states. So, I can start from any state initially.

If I observe this source for a very long time and calculate the states through which it is passing, then those transition probabilities, or the probabilities of the states in the long term will be definite and it will be possible for me to go from one state to any other state. It may not be possible for me to go directly, but indirectly. For example, if I want to go from this state s 0 to s 2, it is not necessary that I will have a directly link between s 0 to s 2 but I can always go to a state s 2 via s 1. So, I go to the state s 1 and then may be directly s 2 or it is possible from s 0 to s 1 and from s 1 to s 3 and s 3 to again s 0, but in the long run, I will be able to reach from one state to another state. This is a crude definition of an ergodic Markov source. To be very specific, there are different definitions. So, just let me look into those definitions.

(Refer Slide Time: 53:21)



At every transition, the matrix of transition probability, if it is the same, then this transition probability is known as stationary. We know that each state, there will be some transition probabilities and if these transition probabilities are stationary, then the Markov's chain is known as homogeneous Markov chain or Markov source. If you calculate the probability of the states, this S i denotes the probability of the states, not the probability of the symbols in the source, this basically denotes the probability of a particular state in a Markov chain, this probability of state will be definite. If it does not change with time, then I will say that that Markov chain is, a Markova source is stationary.

As discussed earlier, ergodic Markov source or Markov chain means that no matter what state it finds itself in, from each state one can eventually reach the other state. That is a crude definition for ergodicity and this understanding is more than sufficient for our course. Now, how do I calculate the probability of the state? Is it possible for me to calculate? If I assume that the Markov source is ergodic, then just with the help of condition symbol probabilities, it is possible for me to calculate probability of state. We will look into the calculation of this in our next lecture, and we will also look at the calculation of entropy for the Markov source.