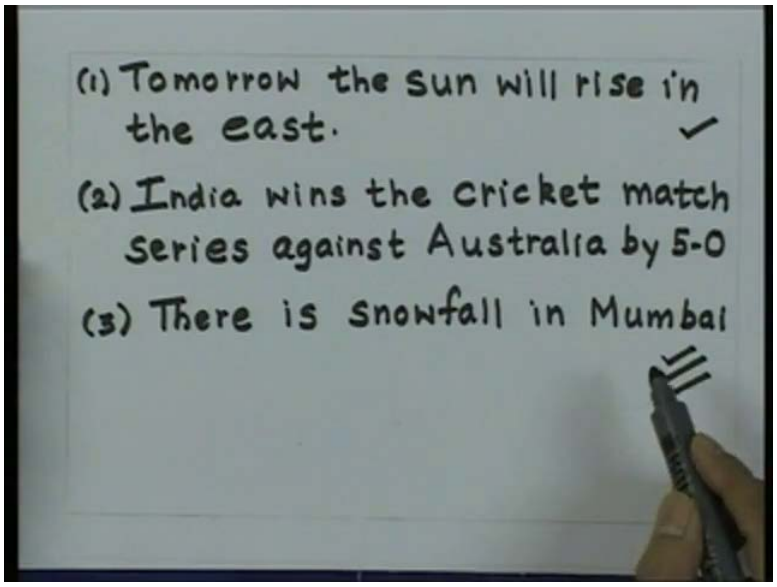
Information Theory and Coding
Prof. S. N. Merchant
Department of Electrical Engineering
Indian Institute of Technology, Bombay

Lecture - 2
Definition of Information Measure and Entropy

In the previous lecture, we had a look at Hartley's definition for the measure of information. This definition was based on two assumptions; first assumption was that e symbols which build up a message can take any value from q possibilities, but all this q possibilities where equiprobable, that was the first assumption which went into the definition of Hartley's measure for information. The second assumption was that all the symbols which build up the message are independent.

Today, we will have a look at the definition as provided by Shannon. This definition of information is more generic, and it overwrites the deficiencies of Hartley's definition for information. To arrive at the definition provide by Shannon, let us approach this definition from the understanding of common sense, let us take a simple example.

(Refer Slide Time: 02:25)



Suppose, if I were to read the following three headlines in a morning newspaper, the first headline says, tomorrow the sun will rise in the east. The second headline says, India wins the cricket match series against Australia by 5 0, and the third headline says, there is snow fall in Mumbai. Now, if I look at all these three headlines then based on my

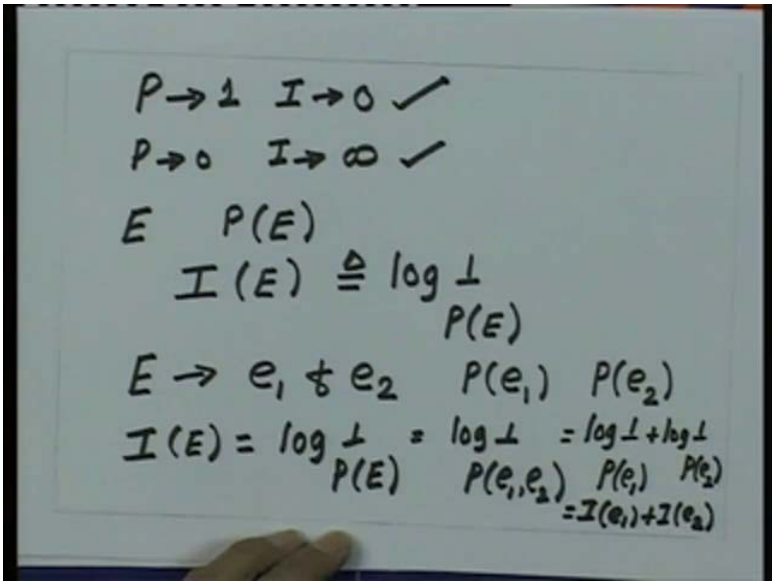
common sense understanding of the word information, I can say the first headline provides the least information, whereas the third headline provides the maximum information.

Let us look at the probability of occurrence of the events associate with each of this headline; the probability of occurrence of the event associate with the headline number one is almost 1. It is almost certain correct, whereas the probability of occurrence associate with the event number three which is says which says that there is a snowfall in Mumbai is almost 0, it is not very certain. So, what it means that if the probability of occurrence of the events is lower then there is a higher surprise.

Therefore, there is more information what it implies that information is connected with an element of surprise which is a result of uncertainty or unexpectance of the occurrence of the event. The more the unexpectedness or uncertainty of an event higher is the surprise and more is the information. The probability of occurrence of an event is a measure of uncertainty or unexpectedness of that event.

So, based on these discussions the common sense understanding of information would measure, would be that information is directly related to uncertainty or inversely related to the probability of occurrence of that event. So, what we should do, when we define the information measure, we should define in a manner which takes into consideration this common sense understanding.

(Refer Slide Time: 06:53)



For example, if the probability of occurrence of an event P tends to 1 which was the case for the statement number one or headline number one. Then the information which I should get from that event should tend to 0, whereas when the probability of occurrence like in the case of the statement number three or headline number three tends to 0. Information should tend to infinity with this understanding of expectations for the measure of information; we will try to define information more formally as defined by Shannon.

So, to define information more formally I would say, let me assume that I have event E and the probability of occurrence of this event is given by P E. Then when the event E occurs the amount of information which I get on the occurrence of the event E will be defined as I of E is by definition log of 1 by probability of E, I define using this relationship. Let us look into little more depth as far as the definition is concerned, if you look at this definition it satisfies my criterion number one out here which says that if the probability tends to 1 then information tends to 0. Another requirement is that if probability tends to 0 information tends to infinity. So, this is the reason for choosing the measure which is inversely proportional to 1 by P E.

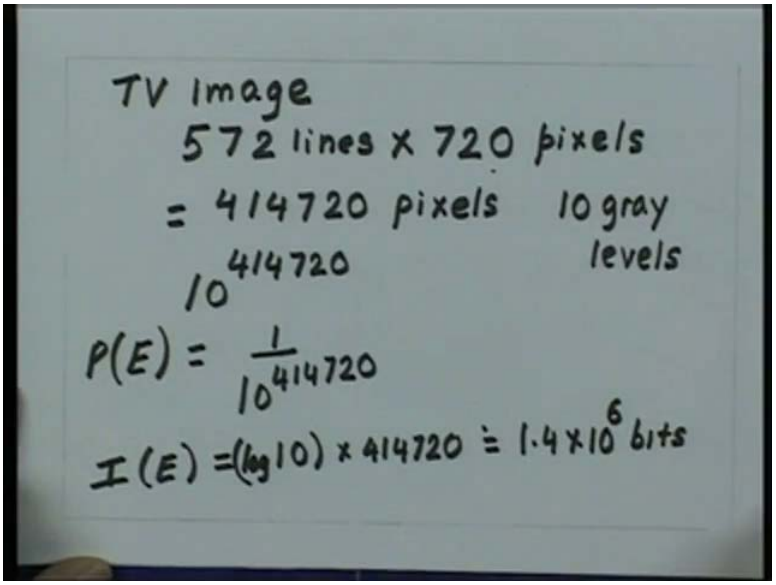
Now, the question that comes to my mind is, why should I choose a function log and why not something other than log, the reason can be explained again based on our common sense understanding of the word information. Let me take simple example, suppose if a highway event E this event E consist of two sub events, let me say e 1 and e 2. Let me assume that both event e 1 and e 2 are independent, so probability of occurrence of e 1 is independent from the probability of occurrence of e 2.

Now, if I were to ask you what is the information which I get when the event E occurs which consist of two sub events. Then I would say by this definition I am supposed to find out information E associate with the event E is nothing but log of 1 by P of this event E. Now, this I can write as log of 1 by probability of event e 1 e 2, this I can simplify as log of 1 by probability of e 1 plus log of 1 by P of e 2 based on the assumption that e 1 and e 2 are independent events. So, what I get is, I get E of information from e 1 plus information from e 2.

Now, this kind of additions of information from two sub events to get the final information in the event E can happen with only log functions it is not difficult to show

this, therefore the choice of log function is there for the measure of information. With this let us go ahead and let us try to calculate the amount of information for a specific case of say a TV image.

(Refer Slide Time: 12:04)



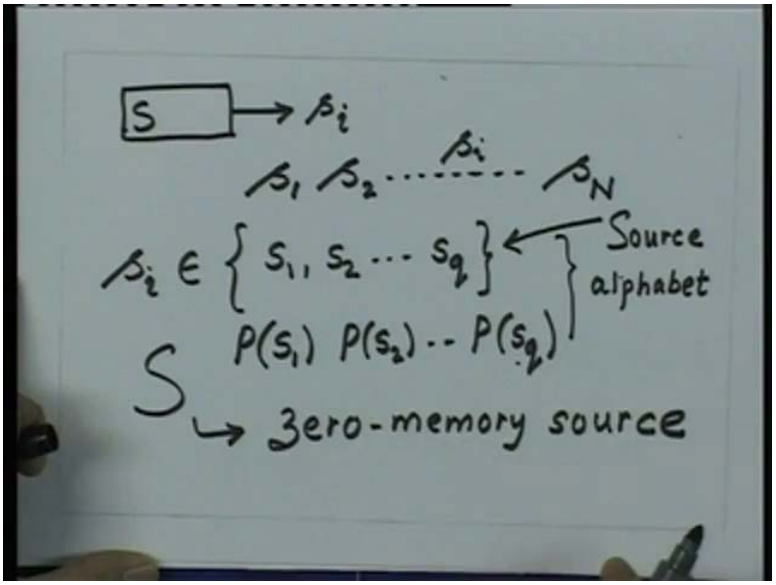
Suppose, if I have a TV image and using this definition for information, if I ask you to calculate the information which I get when I view one TV image then this can be done as follows. Let me assume that each TV image consists of 572 lines and each line consists of 720 pixels that are picture elements. So, one TV image consist of 414720 pixels, if I presume that each pixel can take values from 0 to 9 that means it is allowed to take only ten grey levels. Then the number of TV images which I can formed based on this specification and each grey level consisting of only ten values the number of pictures that I can form is 10 raised to 414720.

These are the total number of pictures TV images which I can form assuming there are ten grey levels, if I assume that any of this picture can occur randomly. Then the probability of occurrence of any particular image will be given by let us call that event as E then the probability associated with that event E would be 10 raised to. Now, to calculation of the information for this TV image is straight forward that will come out to be this comes out to be approximately bits, one thing is important to note that in my definition of for the information I am assuming that the base out here is 2.

It is not necessary for me to restrict my base to 2, I can choose some any other base and it is very easy using the simple conversion formulas between the bases to convert from one base to another base. But as far as this course is concerned most of the time we will be restricting our self to the base 2. So, whenever log is written for the information it is understood that it is to the base 2. So, what I get the information for one TV image turns out to be 1.4 into 10 raised to 6 bits.

We will see later on that is the calculation of information from one image really correct or not we will find that what we have got this value is on much higher side. The information contained in a real TV image would be much less than this we will look into that little later in our lecture. So, far what we have done is that I have tried to define the information associated with one particular event.

(Refer Slide Time: 16:31)



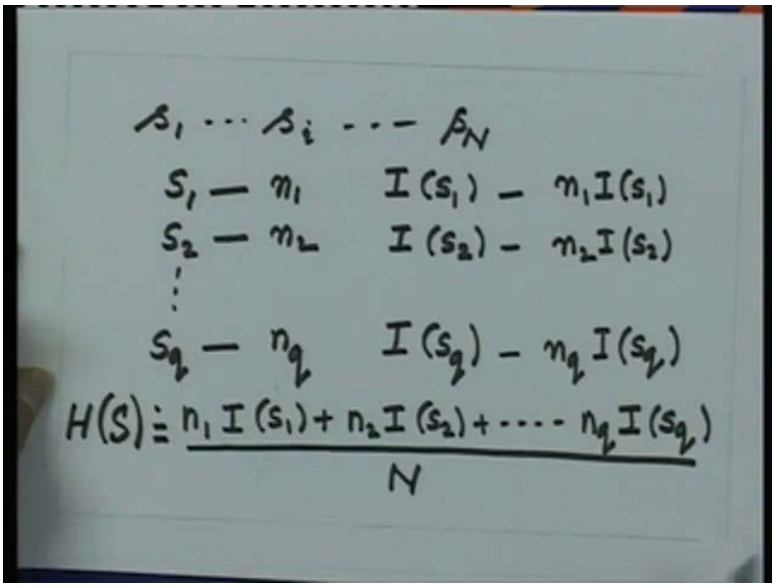
Let us take another case, suppose I have a source S. Let me indicate that source S by S capital S and this source s emits symbols S i it keeps on continuously emitting the symbols, so starting S 1, S 2, S 3 like that continuously it keeps on emitting. Now, the question comes to my mind is, is it possible for me to associate some kind of measure of information to this source S, how do I associate the measure of information to the source S. We have seen how to associate the measure of information to a particular event to solve this problem.

Let us look at the mission of the symbols from the source and put them in the form of a string. So, I assume that first time that I get, I call that symbol which is emitted from the source S as S 1 then I get S 2 and like this. Let me assume that I have end number of symbols which are emitting from the sources, now each of the symbol which I have got S i this itself could take any of the q values from the set.

So, S i belong to a set S 1 S 2 and S q, so these are the only values which S i can take with each of this S 1 S 2 s q there is probability of occurrence associated with it, let me call it as P of S 1 P of S 2 and P of S q. Let me make one more assumption for the time being I assume that all this symbols which are being emitted be the source S they are independent. The occurrence of a particular symbol in this string say S i is not dependent upon the occurrence of the previous symbols that is this assumption which I made to start with. So, if I look at this source, now the source emits a symbol S i each of S i can take the values from the set given by this q possibilities and with each of this q possibilities there is associated probability of occurrence.

This specification is both, this specifications are more than sufficient for me to identify the source S this set is called as source alphabet and with each alphabet in this set there is a probability of occurrence of that alpha of that symbol in that alphabet given by this value is P S 1, P S 2 and P S q. Now, if I am assuming these symbols emission to be independent then I will call this S as a zero memory source. So, let us see whether we can assign some kind of a measure of information to a zero memory source which is given by the source alphabet and the probability of occurrence of each symbol in that source alphabet as I have said.

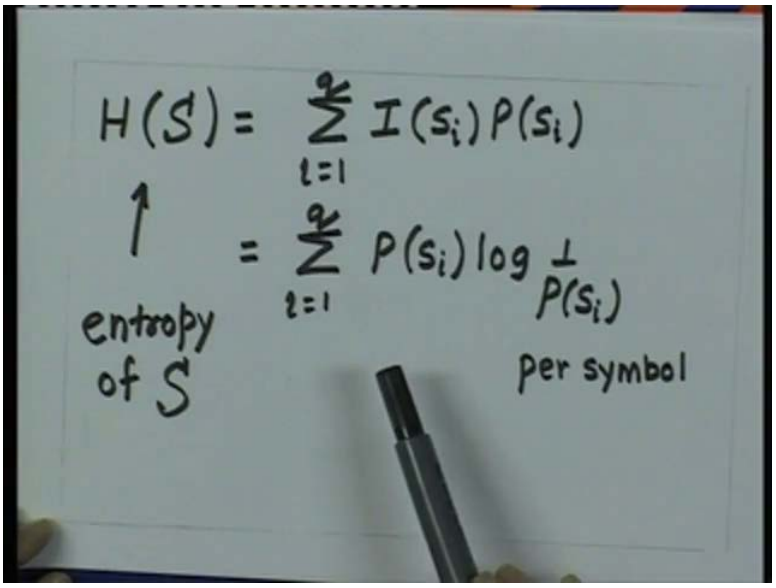
(Refer Slide Time: 21:16)



Let me assume that the string goes from s 1, s i up to s n, n is the lent of the string which I get from the source. Let me presume that a particular symbol in the alphabet S 1 occurs in the string n 1 times S 2 occur n 2 times. Similarly, S q symbol occur n q times with each occurrence of symbol S 1 the information associated to that S 1 would be I of S 1. Similarly, whenever S 2 occurs I can find out what is the information which I get from the occurrence of S 2 and similarly, for I s q these are the information which I get when a particular symbol occurs.

Now, if S 1 occurs n 1 time and if I assume that all the symbols are independent then the amount of information which I get from n 1 symbols S 1 would be n 1 times I of S 1 and similarly, I will get for S 2 and finally, for s q. So, total information which I get from this string S 1 to S n would be the summation of all this. So, the total information which I would get would be nothing but the summation of all this values. Now, if I am interested to find out average value I can simply divide this by N, if I divide this by N then I can write this expression which I have this out here. Let me just denote it for time being I will denote it by H of S, I will denote it approximately equal to this, this is the total information, this is the abbreviation which I am going to use for this quantity on the right hand side.

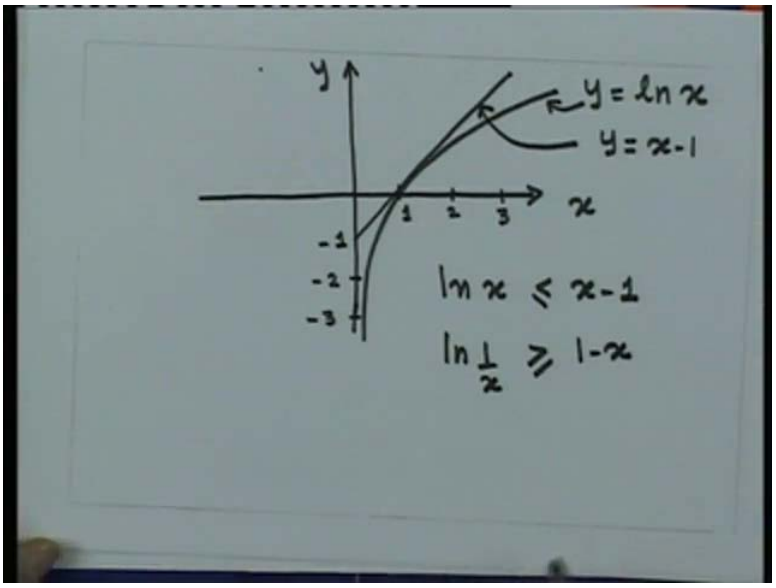
(Refer Slide Time: 24:11)



Then, I can simplify this quantity to H S is nothing but summation. So, going back to the previous expression n 1 by N, n 2 by n and N, q by N, if I use the law of large number then what I get then noting but P S i and this is the summation which I will get I substitute the value for i S i as log of 1 by P S i and what value I get is this. So, this is the average amount of information from that source S per symbol this average amount of symbol per symbol is for the source S is termed as entropy of S.

Now, this entropy of S is dependent on the distribution P S i is this entropy of s bounded from the lower side and upper side. Let us try to look into this if you look at H S definition is nothing but P S i log 1 by P S i, P S i values can go from 0 to 1. So, P S i is always positive log of 1 by P S i is always positive, so what it implies that each of this term is always positive. So, the summation will be always positive, so H S always has to be greater than 0. It cannot be negative that is the first conclusion which we can make about the property of H S. What is the minimum value and what is the maximum value that question can be answered very easily if you follow this following approach.

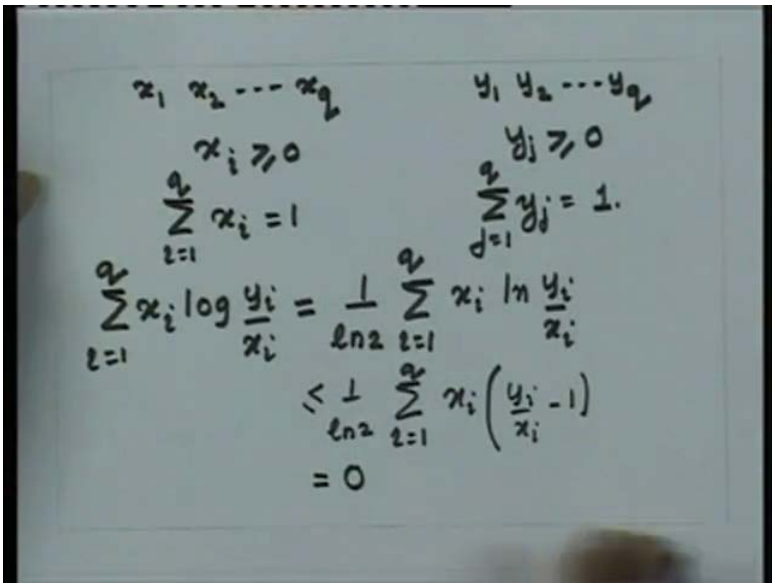
(Refer Slide Time: 27:16)



Let me assume before we try to find out the minimum and maximum value for H S. Let me take a simple function and let me look at the property of that function because that property is going to be utilized for the derivation of the maximum value for H S. If I have a function which is given by y is equal to log x then I can plot this function y log x and the plot would look something like this, this will be my plot for y is equal to log x this value will be 1, 2 this is approximately. If I look at the plot of y is equal to x minus 1 and I draw the same plot on this graph then what I will get would be something like this, this would be the plot for y is equal to x minus 1.

So, graphically I can say that log x is always less than equal to x minus 1 and both the quantities are equal only at 1 and one point and that is x is equal to 1. So, if I just multiply this quantity by minus 1 on both the sides what I will get is log 1 by x is always greater than equal to 1 minus 1 minus x. So, with this small inequality, we will go ahead and find out what is the maximum information, what is the maximum value which we can get for the entropy of the source or what is the maximum information contained in a particular source S. Before I go ahead with this, derivation let me try to define, let me derive one more inequality which will be used later on.

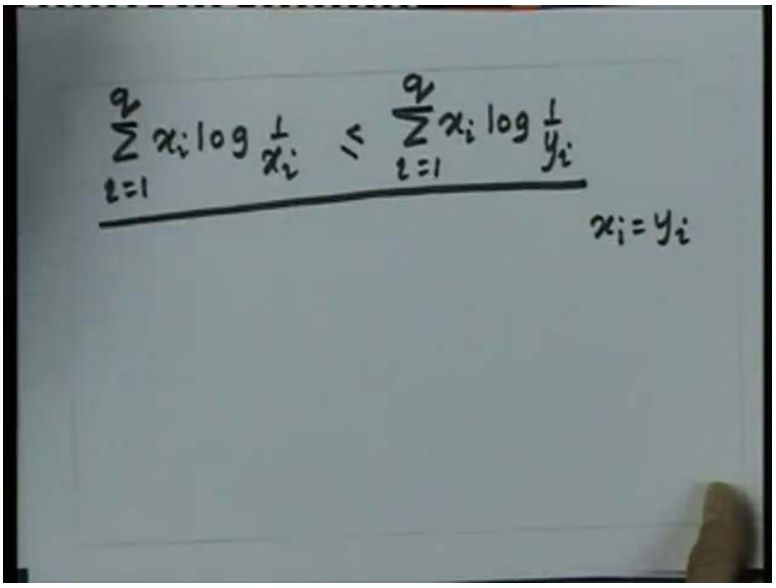
(Refer Slide Time: 30:11)



Let us assume that I have two sets of probabilities given by x 1 x 2 up to x q and another set of probabilities as y 1 y 2 y q these are both two different sets of probability. So, it means x i is always greater than equal to 0 and y j is always greater than equal to 0 for all i and js i and j obviously ranging from 1 to q. So, since this our probability this summation over i 1 to q will be always equal to 1 and this summation out here will be equal to 1.

Now, let me just take a simple expression if I have an expression like this log I can write this as I just substitute the value for log of y i upon x i, we have seen from this relationship that log x is always less than x minus 1. So, if I use this relationship I can write this relationship as less than equal to y i x i minus 1 and this can be shown is nothing but equal to 0.

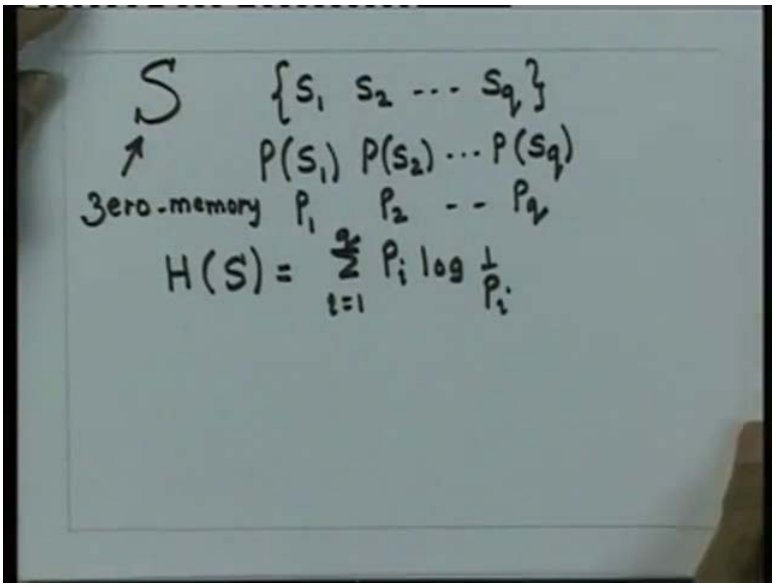
(Refer Slide Time: 32:38)



So, what I get is finally, x i log of 1 by x i i is equal to 1 to q is less than or equal to x i log of 1 by y i i is equal to 1 to q. So, this relationship we will be using it later on in our course. So, it is a very important relationship which we get when x i and y i are two sets of probabilities. So, this inequality will be equal only if x i is equal to y i because e have seen that this two equations are equal only for x is equal to 1. So, correspondingly if you want equality out here what I should get is y i is equal to x i and from there I get this relationship.

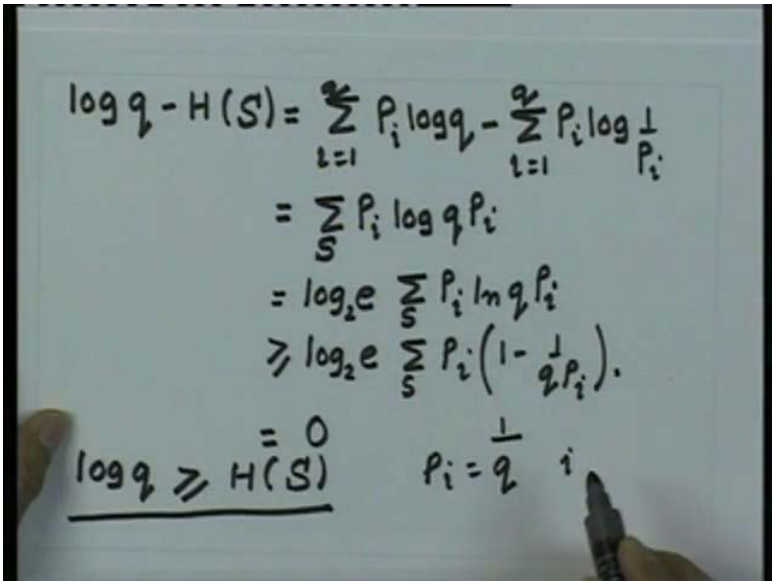
So, we have seen that entropy of the source is given by this expression, let us try to find out what is the maximum value if it exist for what probability distribution function of S i, will I get the maximum value for H S? Let us try to find out that this we will try to do it for a zero memory source.

(Refer Slide Time: 34:23)



So, if I have a zero memory source then that source will be identified S given by source alphabet, I will assume the source alphabet S 1, S 2 and s q are the symbols of the source alphabet along with that I will get the probabilities for the symbols or sometime these symbols in the source alphabet are also known as letters. So, the probability associated with these letters of the alphabet is also given to me and I assume that this source S is a zero memory source. Let us I will try to simplify this P S i, I will write it as P 1, P 2 up to P q, so my entropy for the source H S is nothing but P i log of 1 by P i i is equal to 1 to q.

(Refer Slide Time: 35:49)

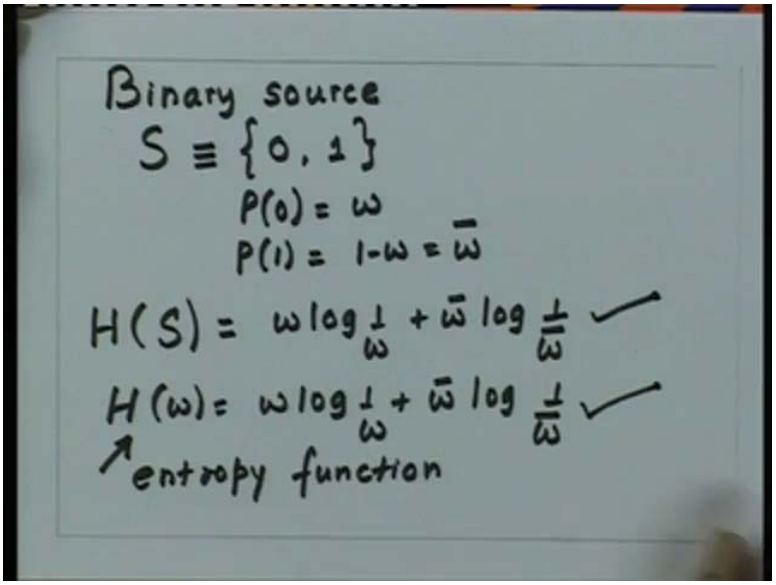


Let us try to evaluate, this expression log q minus H of S, I can write this very simply as P i log q i equal to 1 to q minus P i log of 1 by P i, since P i are the probabilities summation of P i is is equal to 1. So, for log q i can write this expression this I can simplify as P i log of q P i and as a simplification instead of writing i is equal to 1 to q, I can say that a summation over is the source alphabet. So, I just write the source S this implies that I am summing up I from 1 to q that is denoting the alphabet.

So, this I can simplify as log of this, now at this instant I can use the previously derived inequality which is given by graphically log 1 by x is always greater than equal to 1 minus x. So, I can use this relationship to write as greater than equal to this, you can show very easily is nothing but equal to 0. So, from this I get the relationship as log of q is greater than equal to H of S is a very important relationship which we have derived the first important relationship which we have derived. Information theory is these both are equal only when I look at the point whether inequality was introduced; it was at this instant of the time.

So, if you want the thing to be equal what it should happen is 1 should be equal to 1 by q P i, this implies that P i should be equal to q for all I if I can satisfy P i is equal to q for all I then I can write log q is equal to H S. So, what it means that entropy H S is always less than or equal to log q and it will take the maximum value of log q when all P i is equal to 1 by q, sorry it is not P i is not q, but 1 by q. So, what it means that all the symbols or the letters in the alphabet occur in an equiprobable fashion.

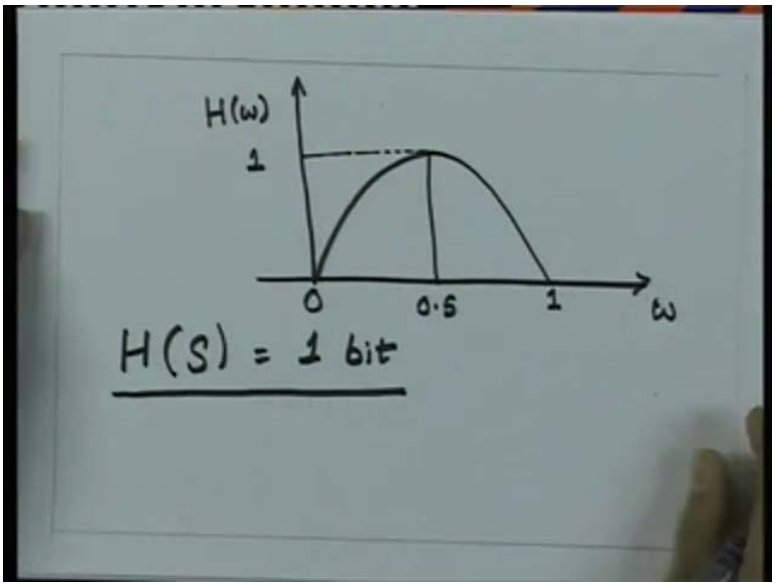
(Refer Slide Time: 39:53)



Let us take a simple example of a binary source I have a binary source which is zero memory. So, if I have a binary source let me denote the alphabet for this binary source S as 0 and 1. Let me also associate the probabilities with this letters of this source as probability of 0 is equal to w and probability of 1 is equal to 1 minus w is equal to w bar. If I assume that this binary source is a zero memory source then I can write the entropy for this binary source H S is equal to this. Now, in the information theory literature, you will find that the expression on the right hand side this is a function of w and it occurs very frequently.

So, if I denote this as a function of w, then I write H w as w log 1 by w plus if I write like this. Then H w is basically termed as entropy function it is important to realize the difference between these two expression. This is an expression for the entropy of zero memory binary source, where this is an expression for entropy function which like any other function is the variable is w out here whereas, here given w this is the entropy function entropy which I will get for the binary source. Let us look at the property of this entropy function.

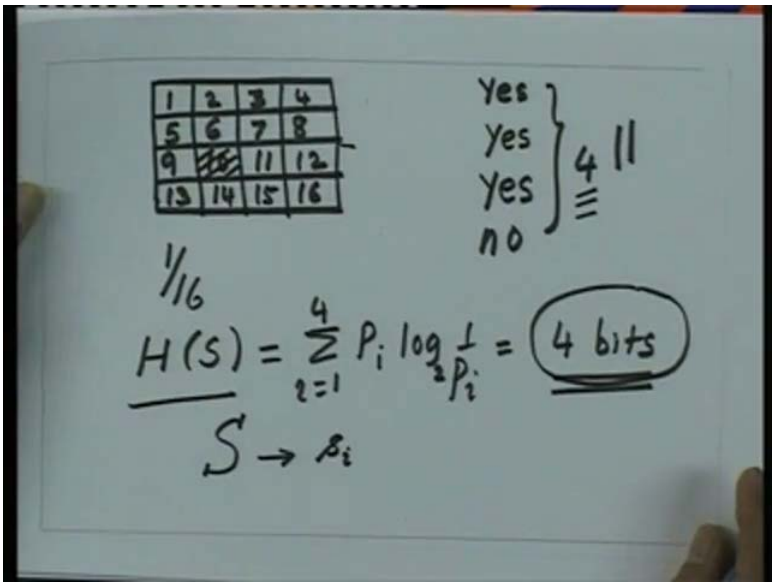
(Refer Slide Time: 42:22)



If we plot the entropy function H w versus w, you will get something like this ,this is my H w and this is my w 0 1. What it shows is that when w is equal to 0, H w becomes 0 and when w is equal to 1 H w again become 0, but when w is equal to 0.5 my H w becomes the maximum value and that is equal to 1. So, for a zero memory binary source if I have

probability of 0, and probability of 1 both equiprobable equal to 0.5 then the entropy of that 0 memory binary source would be nothing but equal to 1 bit. So, this is the maximum entropy I have of a zero memory binary source, we know that if we have binary source, then physically if I want to assign some symbols to those to symbols. Let us look at some of the physical implications of entropy.

(Refer Slide Time: 44:21)



Let me take a simple example. Suppose, I have a graphical grade which is 4 by 4 given like this and I number them from 1 to 16 and let us say that 1 of this grade is marked I mark this grade. My problem is to find out ask questions and find out which grade has been marked my answers will be only in terms of yes and no. So, I can make a guess and ask someone is the grade number seven marked the answer which I expect should be in terms of yes and no, the answer I will get obviously is no. So, if I use this kind of procedure to find out which grade is marked the worst case I can ask from 16 questions, I can say whether grade number 1 is marked or not marked and I expect the answers.

So, in this case if I go in a serial fashion at the tenth question I will get my answer. So, I will require ten questions to be asked before I get the answer which grade is marked, is it possible for us to find out what is the minimum number of questions which you should ask whose answers. We expect in terms of yes or no in such a way I can find out which grade is marked. So, in this case what I can do a clever way of finding out which grade is marked I can ask a question saying which of the I can ask a question simply like this is

the marked grade lies in the bottom eighth grade. So, what I am saying is that the marked grade whether it lies on the upper side or the lower side of this line.

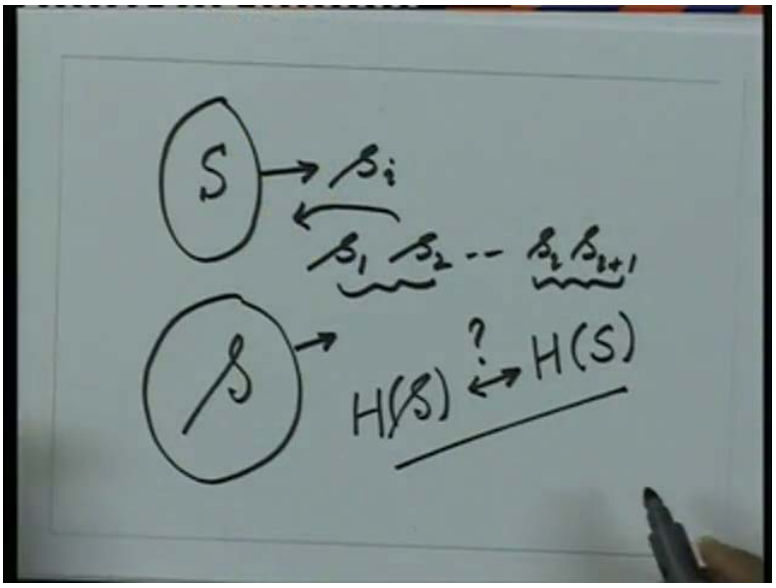
So, the obviously the answer I would get is yes, so I have asked the question once whether the grade lies on the top or the bottom, once I get that answer I know the grade that marked grade lies somewhere from 9 to 16. Now, the next question which I can ask is the marked grade lying on the left of the remaining eight regions, the answer would again be yes. So, I know that it is between 9 ten 13, 14 either one of this again I can ask a next question is my grade lying on the top or the bottom this case he will say it is lying on the top.

So, if I say it is if I ask the question it is lying on top the answer would be yes, so I get a third answer yes. Now, finally I would ask the question is the grade lying on the left the answer would be say no, so I know the grade is basically 10. So, I required four questions to be asked before I could find out which grade was actually marked in this graphical picture. Is it possible for me to relate this minimum number of questions to the concept of entropy which we have just seen? If you look this from the information from information theory point of view I can say that the probability of marking any grade would be 1 by 16 because I have sixteen values out here and any grade can be marked randomly are equiprobable.

Then, I can find out what is the entropy associated with this grade and what I will get is summation of P i log 1 by P i, i is equal to 1 to 4 and in this case it will turn out to be nothing but four bits always the logarithm is to the base 2. So, what it means that and since we are asking in terms the answers we are expecting is in terms of 0s and 1 what I get here entropy also in terms of four bits.

This gives a relationship between a mathematical concept and the physical experiment, which we have carried out to find out which one of the grade is marked. So, this is a simple application of information theory we could have many more applications in this course, we will restrict our application to only communication systems. Now, the way we have defined entropy here was for a single source S emitting single symbols S i.

(Refer Slide Time: 50:54)



The question that comes to my mind is, is it possible to assume that if I have a source S and if it emits symbol S i instead of considering the symbol S i individually if I start blocking this symbols in some length. Let me assume in length of two that means what I do is that when I have S 1, S 2, S i, S i plus 1, I start blocking them in terms of two, so I get like this.

Now, if I start blocking in terms of like this then what I can assume is that I am forming a new source whose symbols are coming out in group of two, but each of the symbols in that group of two is coming from the primary source S. If I start interpreting the output of the source S in terms of another source which I have written here like this. Then is it possible for me to relate the entropy of this new source to the entropy of my primary source; is there some kind of relationship? If there is a relationship existing, what is the usefulness of such relationship? We will look into this matter in the next lecture.