

Digital Communication
Professor Surendra Prasad
Department of Electrical Engineering
Indian Institute of Technology Delhi
Lecture No 33
Introduction to Information Theory

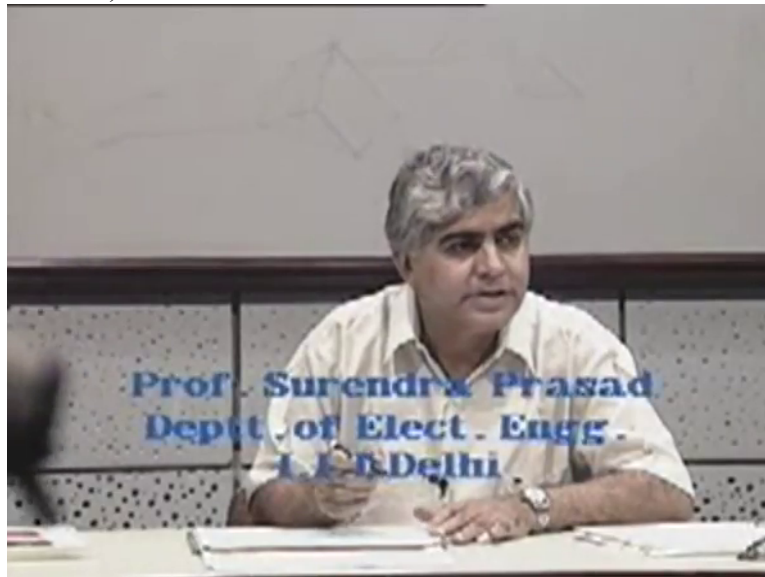
(Refer Slide Time: 01:03)



Information theory, but before we come to that, I would like to motivate why we need to study the information theory to understand better the basic principles of the digital communication. To do that let me first summarize what we have done so far in a nutshell, in terms of a digital communication over a noisy channel, right? Suppose if we were to summarize our knowledge in a nutshell, then what you have learnt is something like this.

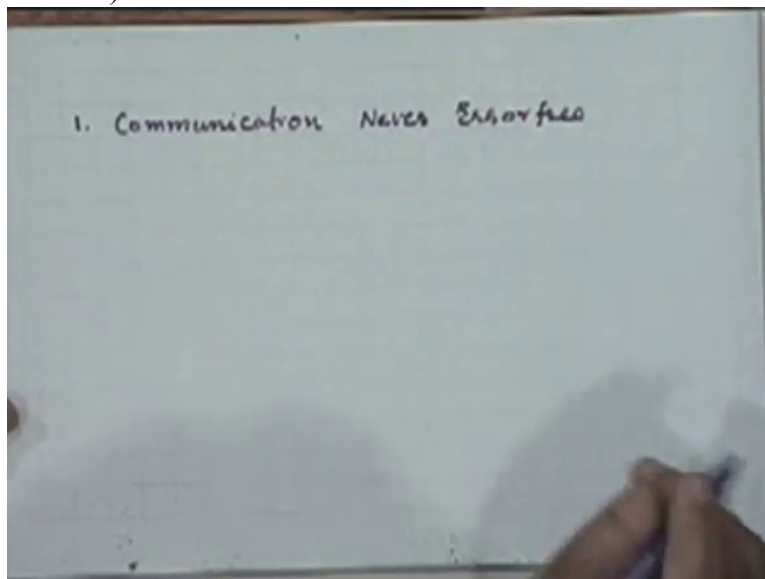
First you have learnt that communication can never be error-free,

(Refer Slide Time: 01:43)



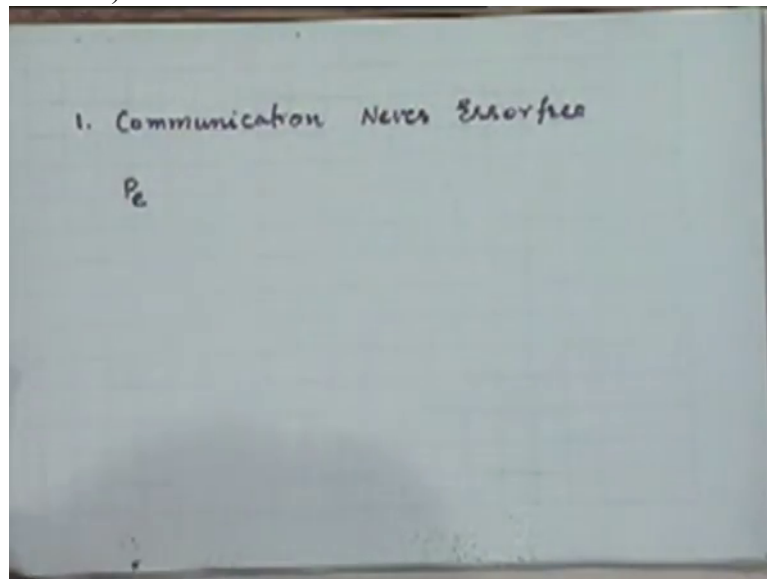
right? As, when we are talking about digital communication, transmission of symbols or bits, in the 0:01:50.6 due to noise and other distortions and at the moment we are only concerning ourselves with noise, there is going to be, there are going to be errors. So that is the first lesson from our

(Refer Slide Time: 02:12)



discussion so far, Ok. There is always

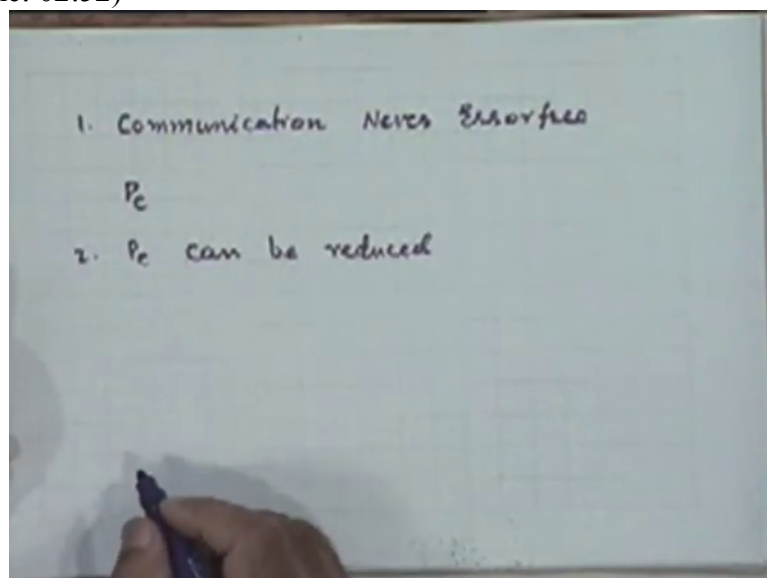
(Refer Slide Time: 02:18)



a certain error probability associated with whatever modulation demodulation scheme or whatever coding scheme you might employ, right?

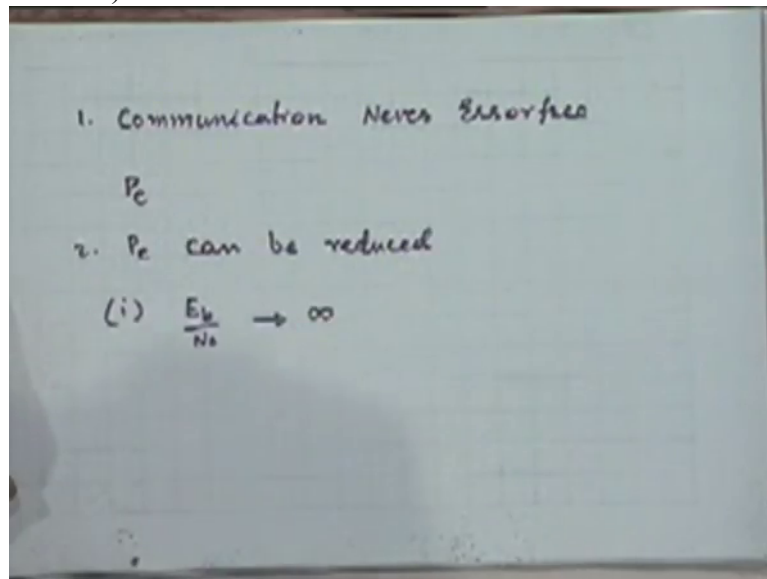
We can reduce this error probability that is the second thing we have learnt, that error probability can be reduced but cannot be made zero. It can be made zero only asymptotically. P_e can be reduced

(Refer Slide Time: 02:52)



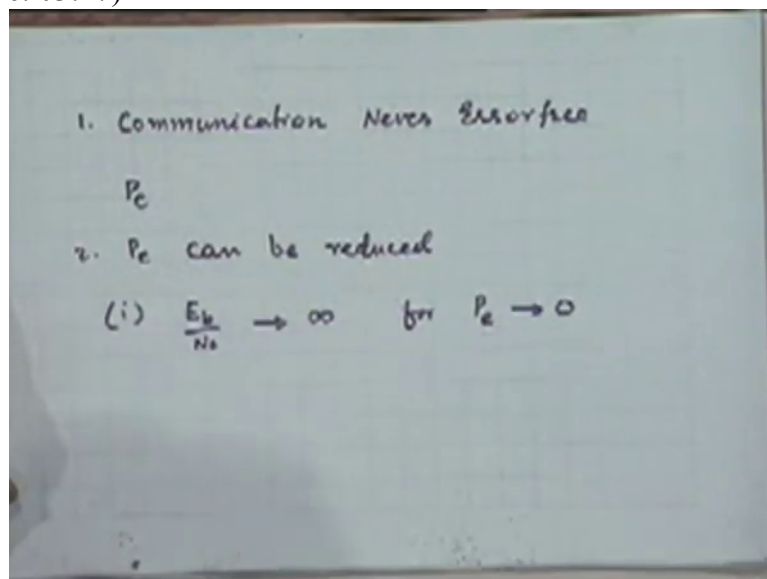
but if you want to really reduce it to zero, that is make the communication really error-free, then you can, you will have to take recourse to one of the two things. That is, increase your signal to noise ratio and you will

(Refer Slide Time: 03:10)



require this increase to go to infinity if you want the error probability to go to zero,

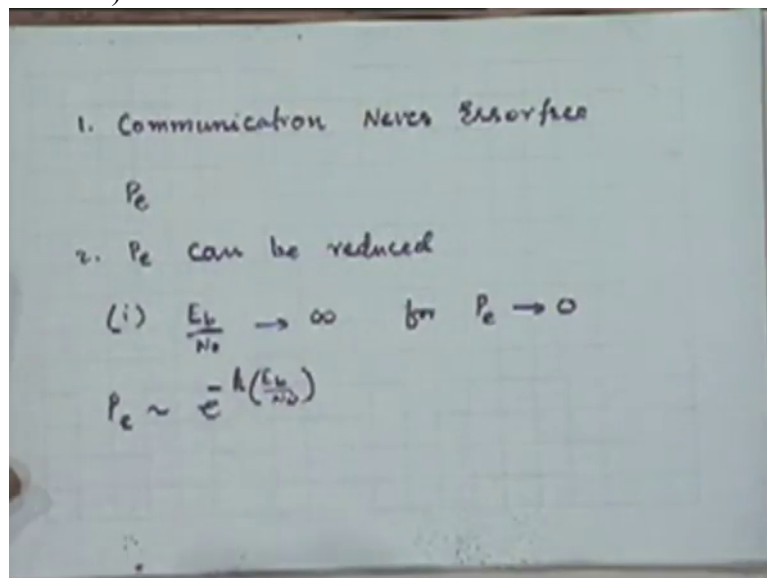
(Refer Slide Time: 03:17)



right?

Because if you look at any expression for the error probability that we have derived, whether it is for binary or M-ary system, whether it is for coherent or a non-coherent systems, you will see that asymptotically each of these expressions become zero as E_b/N_0 is made to go to infinity. Because asymptotically all of them are of this kind.

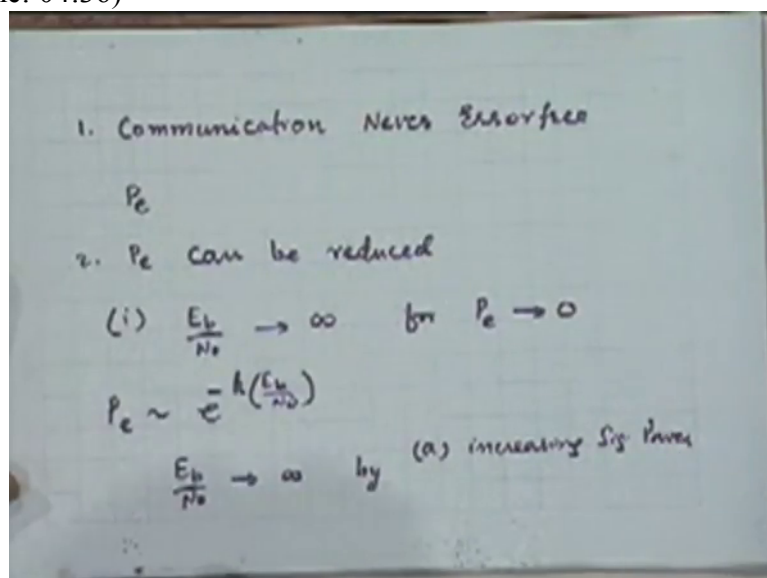
(Refer Slide Time: 03:53)



Asymptotically if you look at any of these expressions, whether it is in terms of Q function or any other kind of function, asymptotically you will see that you can approximate it as, e to the power some constant, minus constant, some constant times e b by N zero, right?

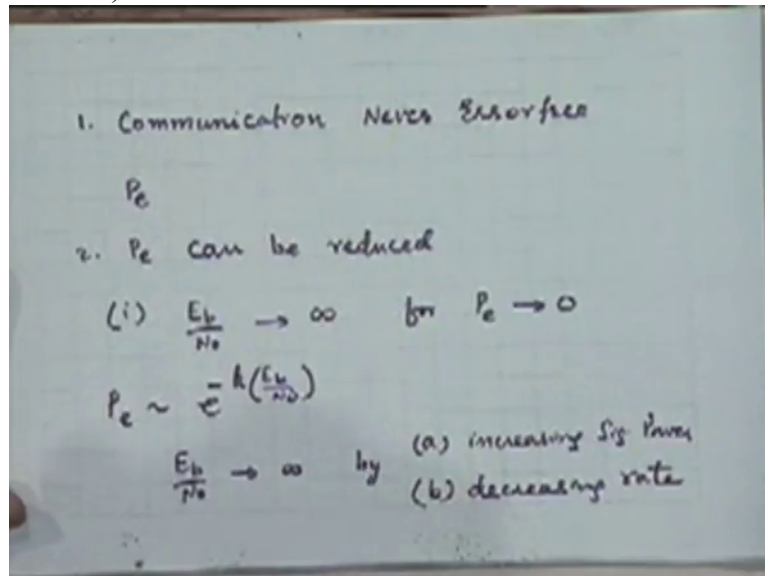
Now this implies that you have to either increase the signal power, this E b by, to increase E b by N zero, again there are 2 ways, right? You can increase your signal power, transmitted signal power.

(Refer Slide Time: 04:36)



Or alternatively by increasing your interval of 1 bit or one symbol, increasing the bit duration, or in other words decreasing the data rate, transmission of rate of data, or decreasing rate, rate of transmission.

(Refer Slide Time: 05:02)

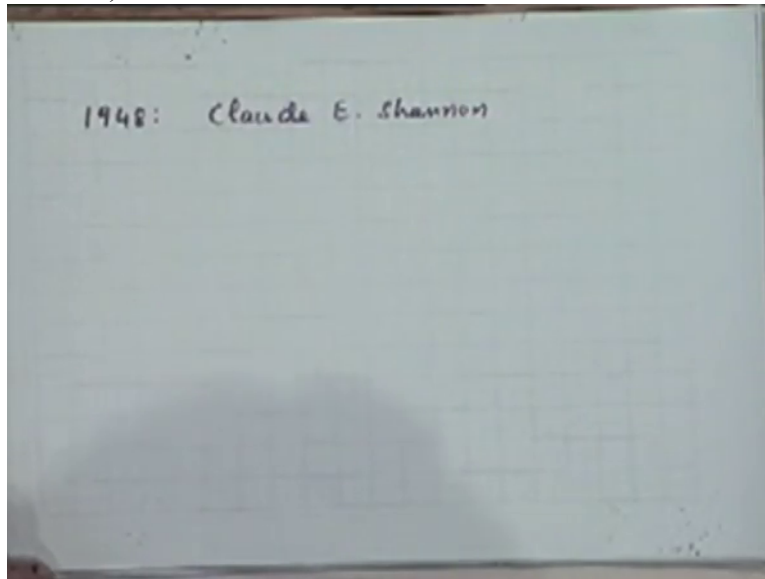


Is that obvious from this expression?

Because E_b is what? It is the energy per symbol or energy per bit actually, right? If you want to increase this to infinity, there are two ways. Either you increase the signal power to infinity because energy per bit is proportional to power times the duration of each bit, right? So either you increase the average power, pump in more and more of power as you can do, or alternatively you reduce your data rate. In fact if you wanted to be infinity with the finite amount of signal power, the data rate has to be reduced to trivial value of zero, right?

So therefore it seems that there is no hope of reducing error rate to a level of zero errors eventually in the technical sense of the world 0:05:57.0, right? Either you go for infinite power or go for trivially low data rates. Now this was the classical view of communication theories till, let us say, a 0:06:12.6 particular gentleman came into picture and that was before, I think in the year 1948. When Claude E. Shannon came to the picture and he changed the view of communication theories

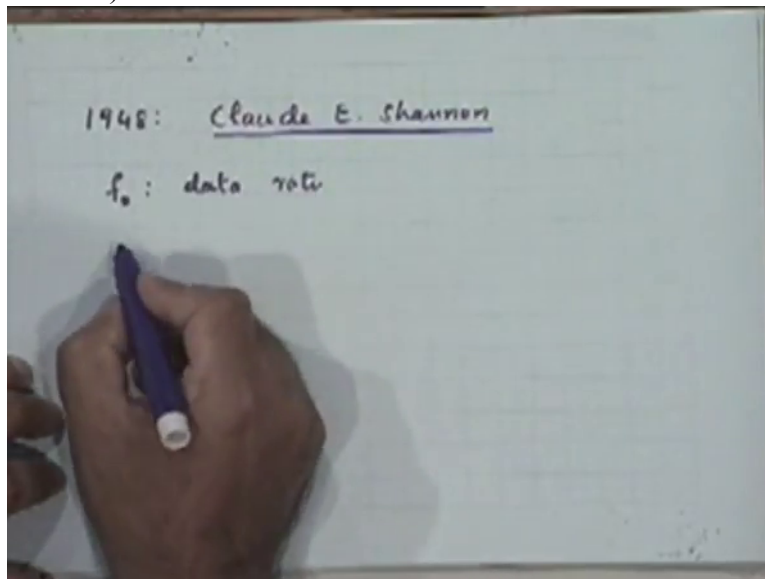
(Refer Slide Time: 06:32)



regarding this perception that you cannot achieve error-free communication in a non-trivial way that is either without increasing signal power to infinity or reducing the data rate to zero in the presence of noise.

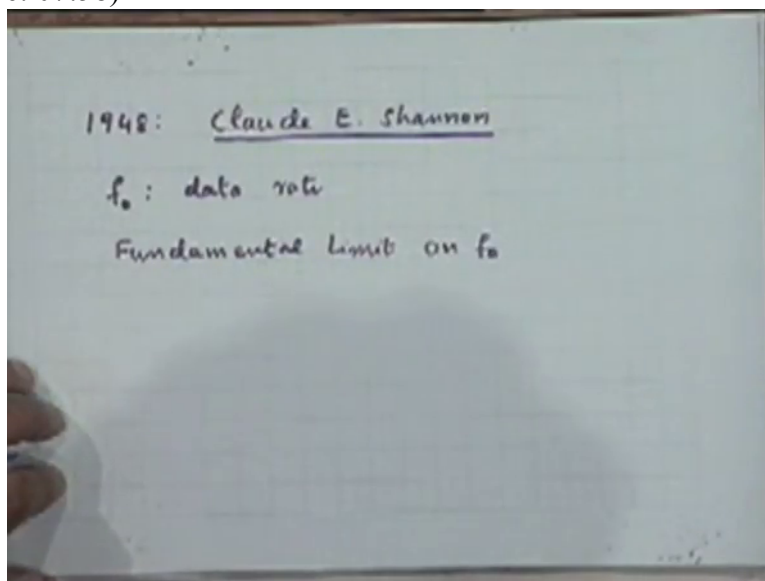
What he said, of course we will have occasion to learn about his theory a little more in detail as we go now, but essentially the revolutionary thinking that he brought about was concerning the fact that in the presence of noise, the noise is not the ultimate limiting factor. The noise is a limiting factor but it is not an ultimate limiting factor for reducing the data rate or it is not a factor which affects, which requires you to put infinite power to get zero error probability, right. What in fact he said was if there is really a fundamental limit on the error, on the data rate, suppose I do not want rate of transmission of data by f zero, so his contention was

(Refer Slide Time: 07:47)



that there is a fundamental limit on f sub zero,

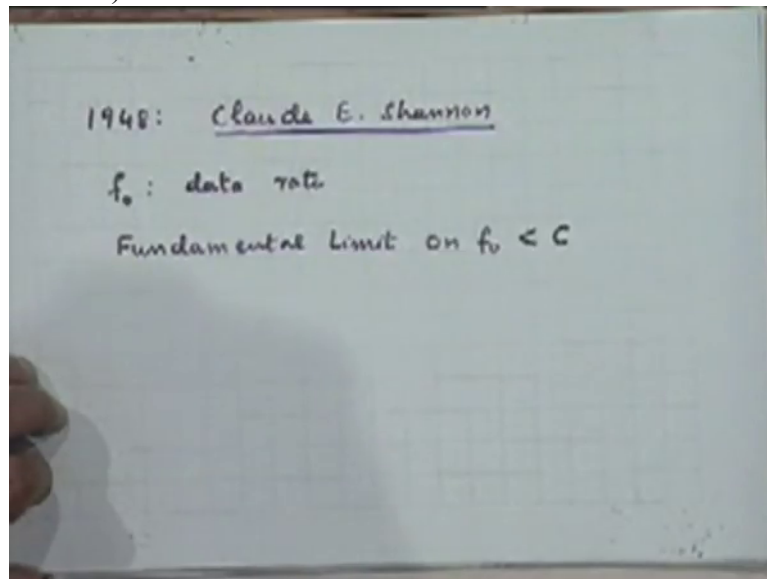
(Refer Slide Time: 07:58)



the data rate at which one can carry out transmission over a noisy channel, let us say a channel with a white Gaussian noise.

Fundamental limit in the sense that if you restrict your transmission of information at a rate less than this, less than a specific value which he called C ,

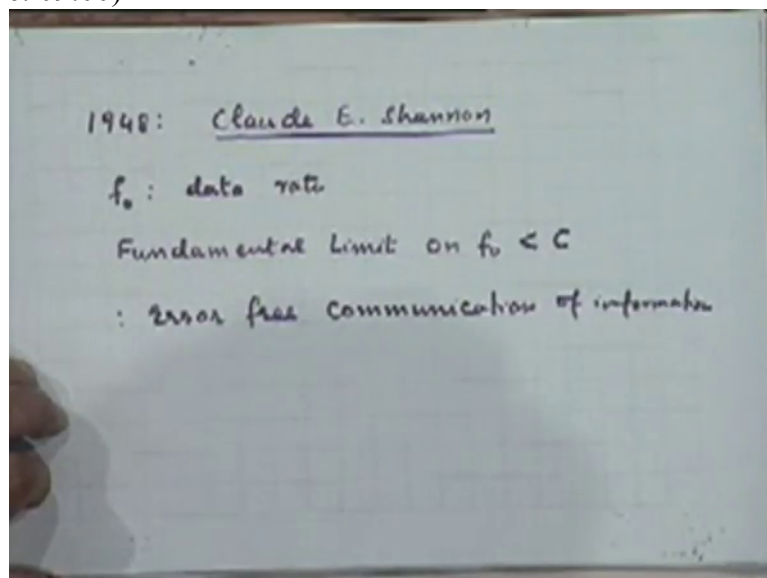
(Refer Slide Time: 08:22)



or which he called the channel capacity. So for a given channel he defined an attribute of the channel which he called the channel capacity and the result that he propagated was, as long as you keep your rate of information transmission to be less than this number which he called the channel capacity, it was possible to achieve error-free communication of information, right?

And this was a very

(Refer Slide Time: 09:00)



revolutionary result and it sounded to be something against common sense. Because for whatever we have developed so far and the theories I have told you are well-understood theories. They are even modern theories, right? And there is nothing wrong with these

theories that we are talking so far. But their error probability becomes zero only exponentially, asymptotically as signal to noise ratio is increased and therefore signal power is increased, or data rate is decreased.

So on the one hand there is nothing wrong with those theories. On the other hand, there is this result. There seems to be something wrong somewhere, right. There is actually nothing wrong, because the emphasis is on this word, transmission of information, Ok. What he said, and this is what is to be understood really, this result is not for transmission per se, error free transmission per se but error free transmission of information in whatever you are transmitting. So the information content of the message can be made to reach the other side without any error provided our information rate, mind you, this is also a limit on, not the data rate but the information rate is less than the so-called channel capacity.

So basically every channel now has an additional attribute by which you can characterize it, that is the rate at which it can support transmission of information. So we have brought in an extra dimension here, a dimension that we have not talked about so far. Intuitively we all know that when we are talking about communication, basically we are talking about communication of information. And that aspect is deservedly being emphasized here. That our interest is not in communication per se but in communication of information

(Professor – student conversation starts)

Student: F naught is information rate

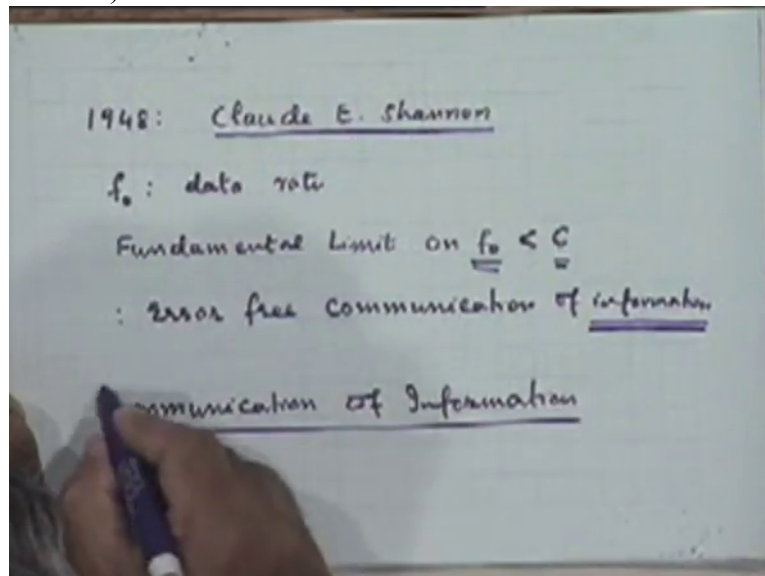
Professor: That is right.

Student: What do you exactly mean by data rate?

Professor: I am coming to these discussions.

(Professor – student conversation ends)

(Refer Slide Time: 11:20)

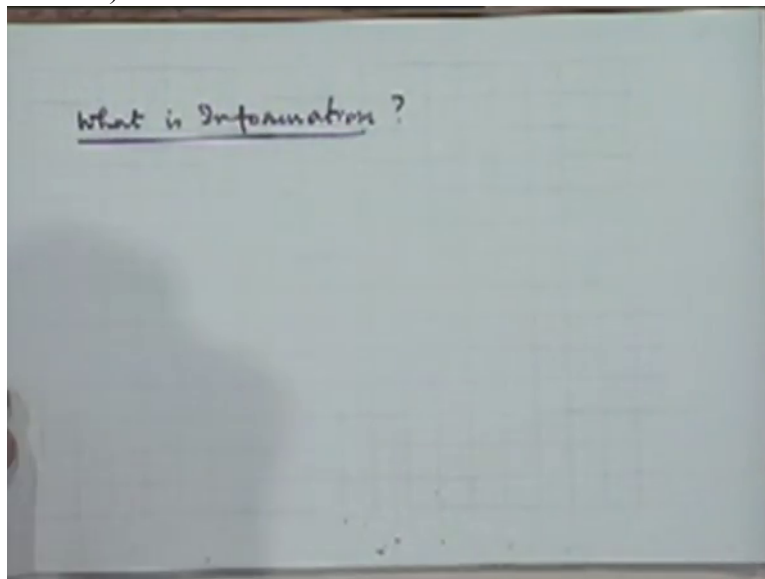


Ok, so basically to take up kind of questions that some of you have immediately put forward as to what is the difference between information rate and data rate and things like that, before we can even take up the answers to these questions, even more important is to be able to define what we really mean by this so-called abstract notion of information, right? How can we make it more precise?

So Shannon's contribution lay in giving a very precise mathematical formulation of information and then based on that mathematical formalism, coming up to result of this kind, right. This was his major contribution. So let us try to understand what we really precisely meant by information. How we can define it or model it so that we can work with it in a more precise way. And that is what information theory is all about. Starting with an understanding of what is meant by information in a given message, how much information a message is conveying and then working with this notion to derive the characterization of communication channels, Ok.

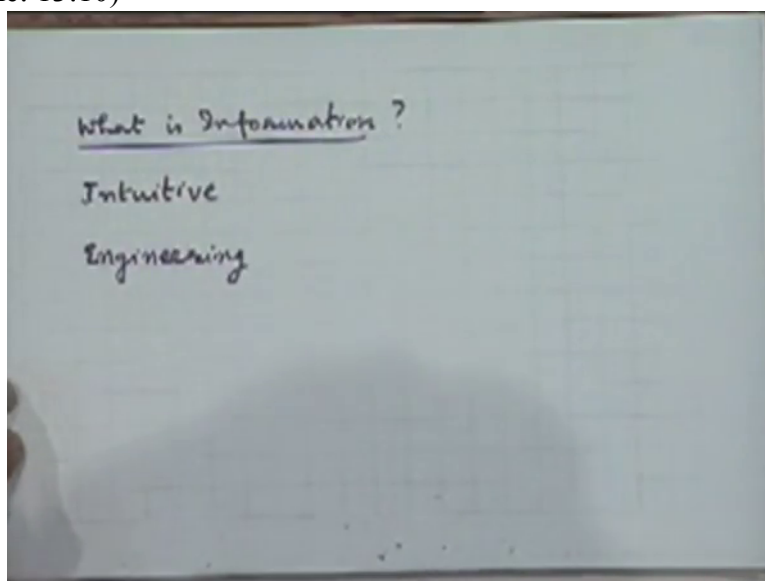
So this is what we will take up for discussion. So let us come to this question of how to define precisely what is meant by information.

(Refer Slide Time: 12:51)



Now there are two ways of looking at this question. One is what I call an intuitive way and the other what we can call an engineering viewpoint,

(Refer Slide Time: 13:10)



an intuitive viewpoint and an engineering viewpoint. Let us take the intuitive viewpoint first. As far as intuitive viewpoint is concerned the information content of the message can be linked to the element of surprise associated with the message, right? Element of new information that you are getting out of it, right? Basically element of surprise. For example let us say if you make statements of the following kind, that every day we have this class at 10 o'clock or every whatever, Monday, Tuesday and Friday. You already know about it, there is nothing surprising about it. It does not convey you any new information except perhaps on the first day of the timetable when you did not know when the class is going to be.

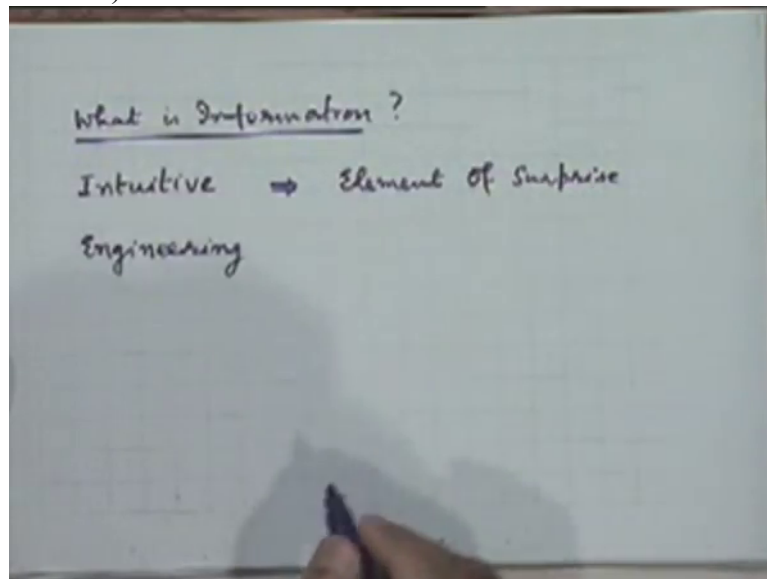
Or if I tell you that let us say, Bangladesh has invaded India, let us say, right? That will be some element of surprise because you don't expect such an event. Or even for that matter, India has invaded Bangladesh or whatever Pakistan or whatever. Although that is more surprising and therefore there is some information content in it, it is not as much as the other way around.

Therefore the intuitive viewpoint looks upon information as something connected with an element of surprise with the message, right? If there is surprise value in the message it conveys more information. If there is no surprise value in the message it conveys less information, right? That is intuitive way of looking at it.

In other words, it is linked with the probability of the occurrence of the message, right? If it is something that is not surprising means it is highly probable and therefore conveys less information. On the other hand, something that is, has a very small probability of occurrence, if it occurs, when it occurs, it conveys a much larger amount of information. If someone tells you that this is happening then you feel that you are getting a lot more information than otherwise.

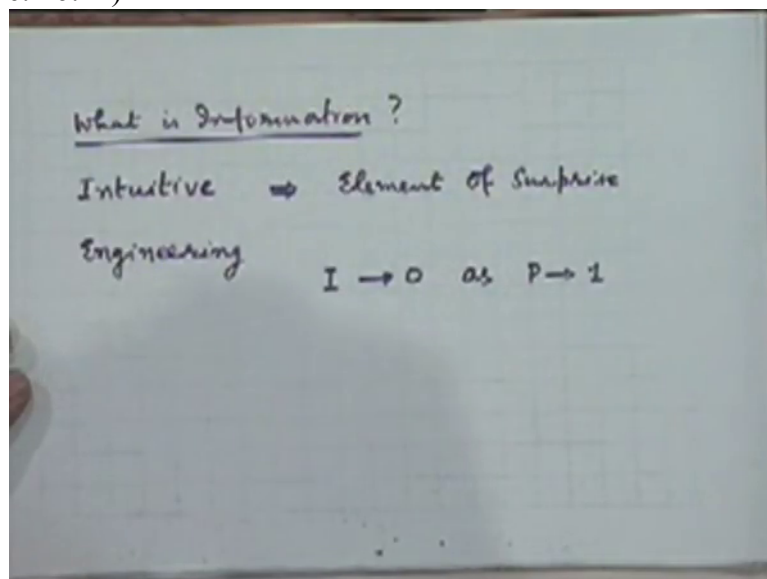
Like for example in stock exchanges, you can talk in the same way. A particular share is known to do well and it does well, no problem. But if on the other hand, suddenly some unknown shares go up without the scam, then of course it is a different matter, and then you could take a clue that something is happening in the market which one should look into, Ok. So this is the clue here, linked with element of surprise.

(Refer Slide Time: 15:49)



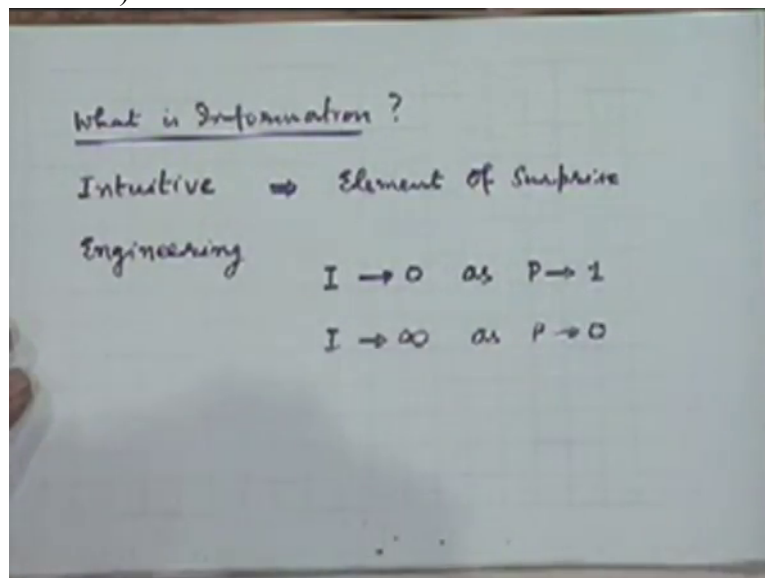
Now let us make it more precise. What we would like to see f o is that if you want to develop a mathematical model of information, call it I , it should be such that, it should, the information content is practically zero if it is, if you are talking of an event which is associated with the large probability, right as p goes to 1.

(Refer Slide Time: 16:21)



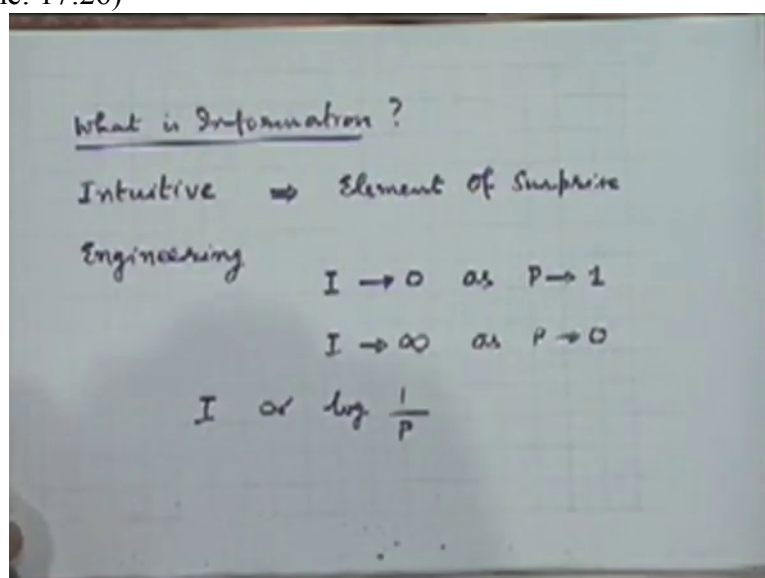
Similarly information content should be very large, let us say going up to infinity as p goes to zero.

(Refer Slide Time: 16:33)



So first we have appreciated that there is perhaps a link between information and probability of occurrence of event and the kind of relation you are looking for is associated with these properties. So if you are thinking of mapping of probability to information which has these properties that would be a suitable candidate for, a model for defining information. And one such model that emerges from these properties is that I is proportional to \log of 1 by p , right? That is, as p tends to 1 , this becomes, this tends to zero and p tends to zero, this also tends to infinity,

(Refer Slide Time: 17:26)



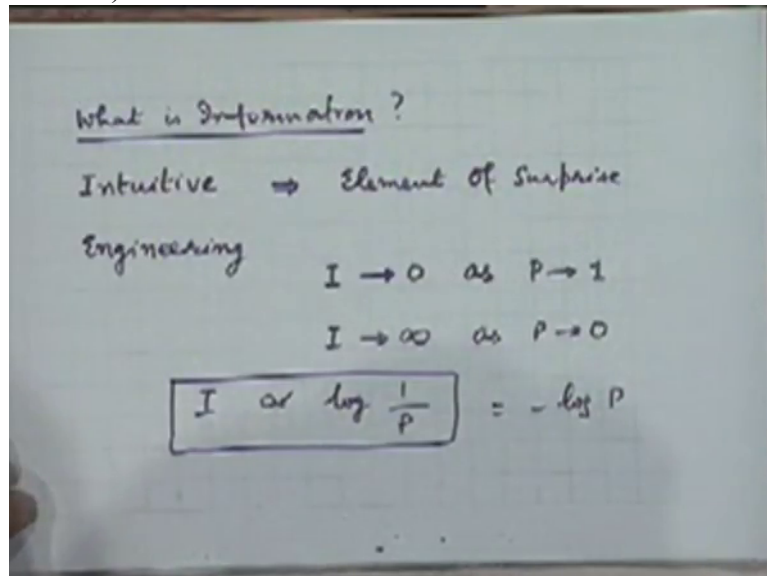
Ok. Now this...

(Professor – student conversation starts)

Student: Log of p 0:17:34.7

Professor: Then you can make it minus log of p, which is of course this,

(Refer Slide Time: 17:42)



right. Think about it, why it should be minus and not plus. So this is, from the intuitive point of view, if you proceed this is what we get, this kind of relationship. If on the other hand, you take the engineering viewpoint, an engineer's viewpoint, suppose you are an Information engineer whose job was to help people pass messages containing information from one place to another. How would you like to look at information?

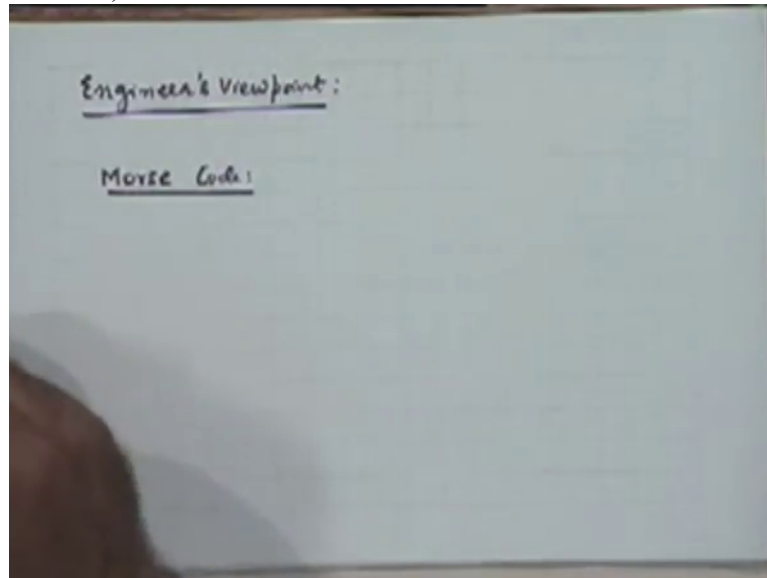
(Professor – student conversation ends)

You would probably link it up with how much time is required to convey a particular piece of information so that if a guy takes too much time to convey his information you could charge him more or if somebody can do it quickly you could charge him less. Somehow the engineering viewpoint could be based on the amount of time it takes to convey a given amount of information, right, makes sense? Because then I can directly, from engineer's point of view I have a way of calculating how much to charge him, right. And the services that I am providing to him for letting him convey this information, right?

If he uses the channel for more time, it must be that he is trying to convey more information. If he is using the channel for less time, it should imply that he is conveying less amount of information, right?

So let us see, if you base the thing on this, this is very satisfying from even some of the earlier examples that you might think of. For example you are all familiar with Morse code, right?

(Refer Slide Time: 19:33)



I am sure some of you are probably radio amateurs and even otherwise you are familiar with Morse code. What kind of code it is? It is full of dots and dashes and if you look at the mapping from the English alphabet to this code, you will see some definite features in this mapping.

(Professor – student conversation starts)

Student: Greater the alphabets the lesser

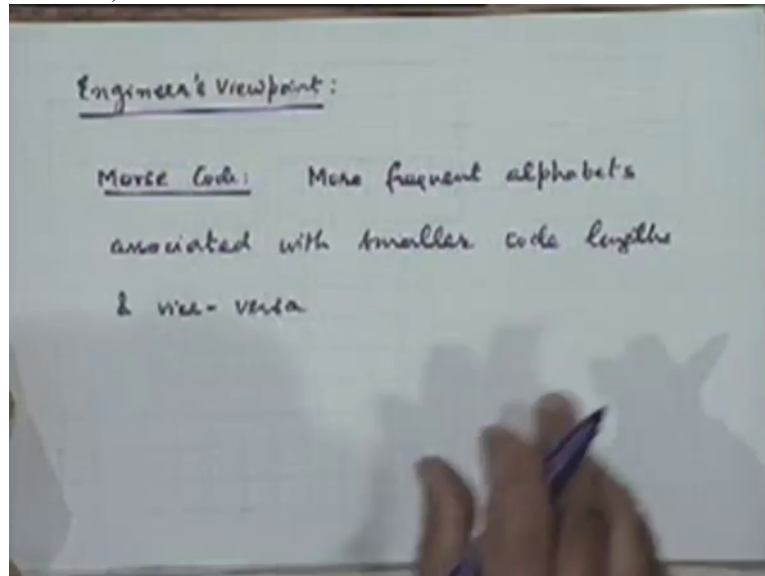
Professor: That is right. Alphabets which are associated with, let us say, very large probability of occurrence like e, that means they are going to occur very frequently in a given text, so you do not want to spend too much time on conveying that, right? You will convey it with a smaller amount of coded message so that it takes less amount of time. The overall text takes less amount of time to transmit. Similarly a message which occurs rather infrequently, we do not mind if it is coded with larger length of the code, right? Because it is going to happen so infrequently that it is not going to affect your average message length very much, right?

(Professor – student conversation ends)

So in Morse code, more frequent alphabets associated with smaller code representations or smaller code lengths and vice versa, right? So this is intuitively appealing to an engineer.

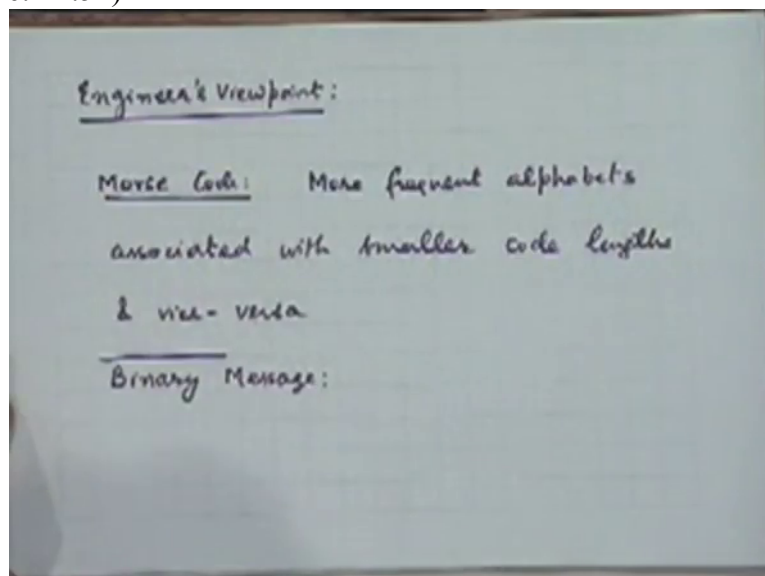
Because he wants to minimize the amount of time that he takes to convey a given amount of information, right. And he thinks this is a good way of doing things. So somehow information is now linked with the time it takes to transmit a message.

(Refer Slide Time: 21:31)



Now to motivate the definition of information from this point of view, let us take a very simple example. Let us say we have, I have to convey a binary message. Morse code is a much more complicated example. I will take a very simple example where one is conveying what is called a binary message,

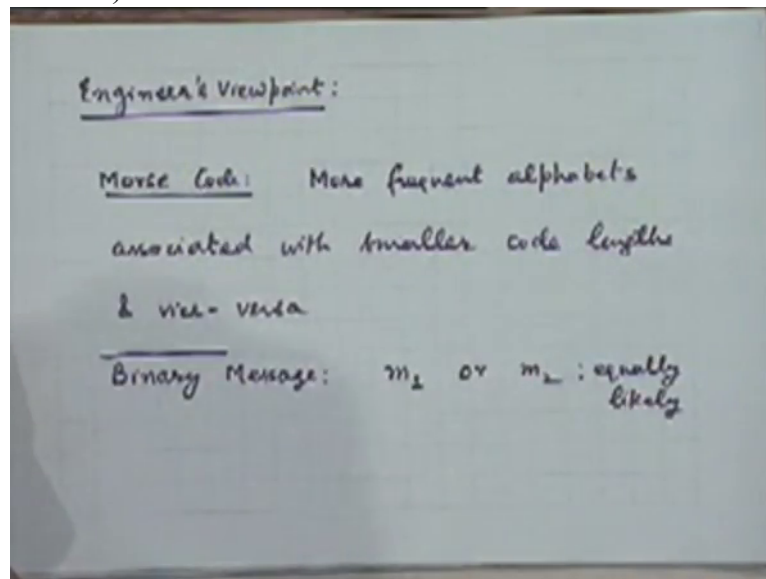
(Refer Slide Time: 21:51)



as a kind of you are dealing with in digital communication, where in every, let us say interval, some unit interval, you are conveying one of two possible messages, say m_1 or m_2 .

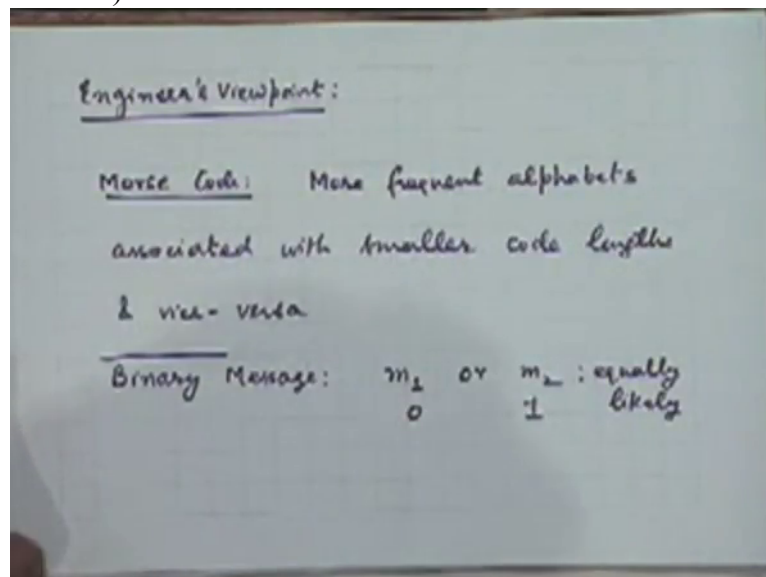
I would have said both are equally likely. They have each a probability of

(Refer Slide Time: 22:15)



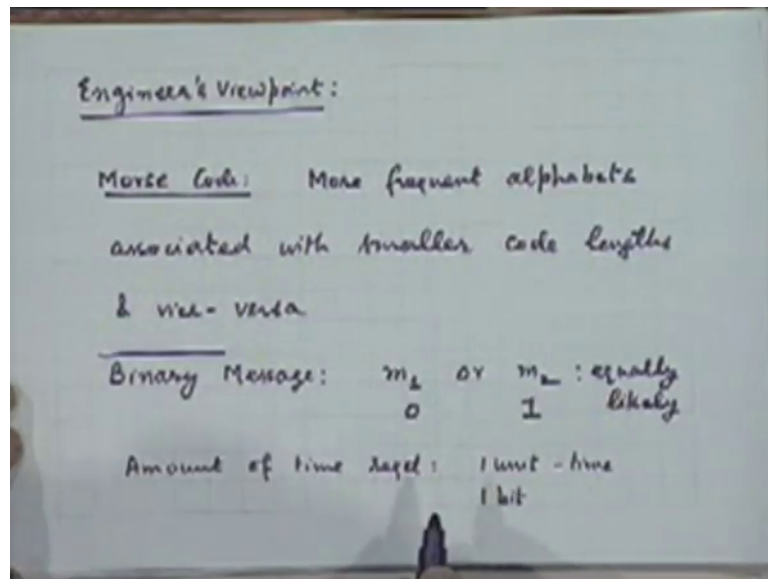
half of occurrence, right? Of course in real life we could represent the message

(Refer Slide Time: 22:22)



1 with, $m_{sub 1}$ as 0 and $m_{sub 2}$ as 1, right? That is 0:22:27.7 representation. You can call them m_1 , m_2 or 0 and 1, it does not really matter, alright. So intuitively now, you know how many symbols in one unit of time, in whatever the unit of time I have selected, I require to transmit this information one bit because depending on what occurs, m_1 or m_2 , I will transmit this or that but basically one bit of information or one, if a unit of information is bit in the usual sense the amount of time required will be one unit time in which we will be convey one bit of information,

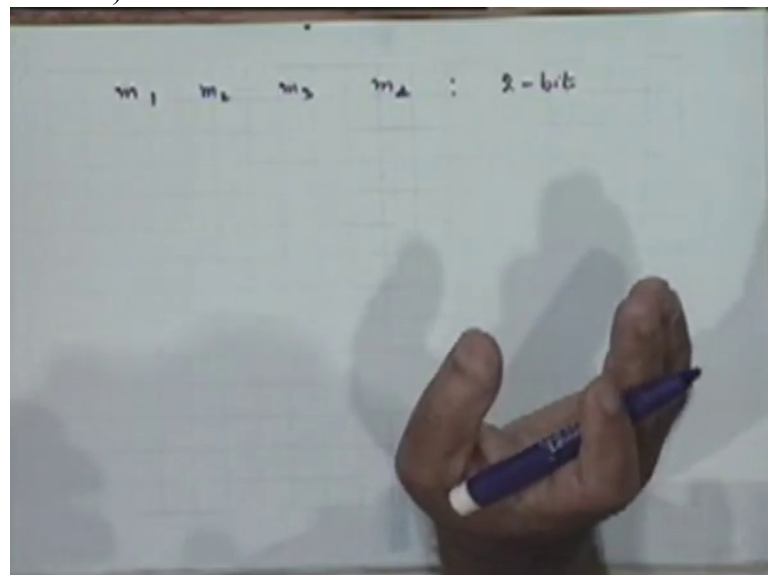
(Refer Slide Time: 23:16)



right?

Suppose I now make into a four level message just for the sake of argument. Let us say now my messages are m_1 , m_2 , m_3 , m_4 , right, four valid message that is, in any, in any unit of time, one of these four possible messages are to occur, right? So then what is the number of bits required to represent it? Two bits, right. So if the time, amount of time required

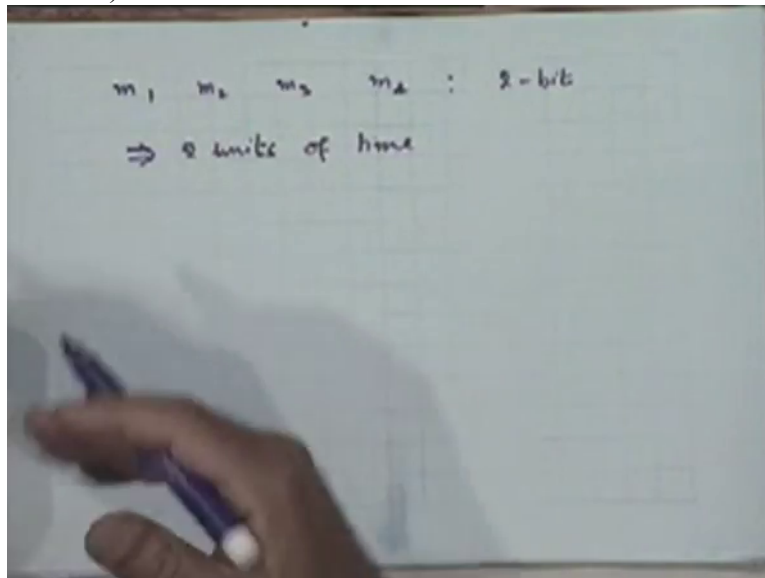
(Refer Slide Time: 24:00)



to transmit 1 bit is fixed, then amount of time required to transfer information contained in this message will be twice as much, right? So if I say 2 bits, essentially I am also implying that 2 units of time are required.

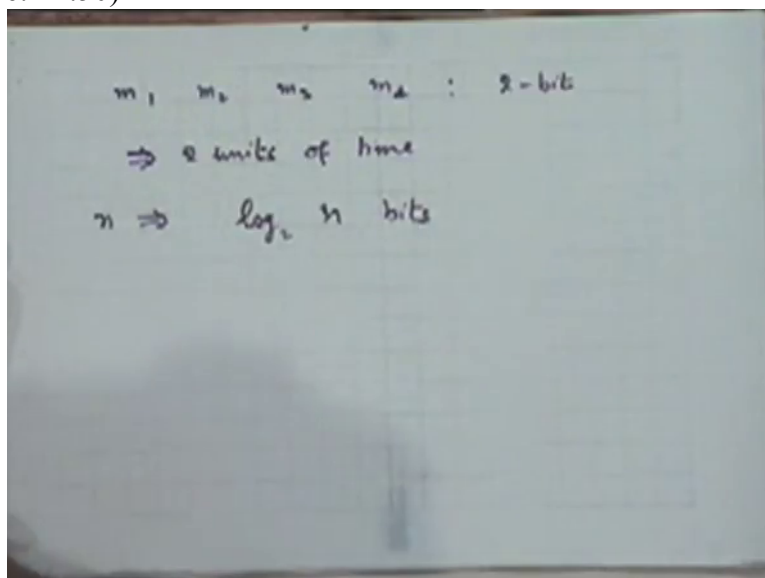
Of course, this is again based on the assumption that each of the four messages are equiprobable,

(Refer Slide Time: 24:25)



and so on. We can increase these arguments so forth, but in any case, suppose I would like to generalize this to n messages and let us say n is the power of 2 for simplicity, like we took 2, 4 and 8, what can we say about the amount of time that will be required, $\log_2 n$ bits, right?

(Refer Slide Time: 24:50)



$\log_2 n$ to the base 2. Or since it requires so many bits and time to transmit every bit is fixed, it requires so many units of time.

(Refer Slide Time: 25:04)

$$\begin{aligned}
 m_1 \quad m_2 \quad m_3 \quad m_4 &: 2\text{-bit} \\
 \Rightarrow 2 \text{ units of time} \\
 n \Rightarrow \log_2 n \text{ units of time}
 \end{aligned}$$

So now we have seen what kind of relationship comes out. And what is the value of n in terms of, let us say, probability? If you take 4-level message what will be, what is p in that case, probability of occurrence of each message, 1 by 4 or 1 by n , right? In general it is 1 by n .

(Refer Slide Time: 25:31)

$$\begin{aligned}
 m_1 \quad m_2 \quad m_3 \quad m_4 &: 2\text{-bit} \\
 \Rightarrow 2 \text{ units of time} \\
 n \Rightarrow \boxed{\log_2 n} \text{ units of time} \\
 p = \frac{1}{4} \Rightarrow \frac{1}{n}
 \end{aligned}$$

So I can substitute n by 1 by p .

(Refer Slide Time: 25:35)

Handwritten notes on a whiteboard:

$$m_1, m_2, m_3, m_4 : 2\text{-bits}$$
$$\Rightarrow 2 \text{ units of time}$$
$$n \Rightarrow \boxed{\log_2 n} \text{ units of time}$$
$$p = \frac{1}{4} \Rightarrow \frac{1}{n} \quad n \Rightarrow \frac{1}{p}$$

So once again what I see is the amount of time that would be needed will somehow depend on

(Refer Slide Time: 25:49)

Handwritten notes on a whiteboard:

$$m_1, m_2, m_3, m_4 : 2\text{-bits}$$
$$\Rightarrow 2 \text{ units of time}$$
$$n \Rightarrow \boxed{\log_2 n} \text{ units of time}$$
$$p = \frac{1}{4} \Rightarrow \frac{1}{n} \quad n \Rightarrow \frac{1}{p}$$
$$\text{Amount of time} = O(\log_2 \frac{1}{p})$$

\log_2 , or \log of $1/p$ to the base 2. Is it Ok?

So no matter what view you take, whether you take the intuitive viewpoint or the engineering viewpoint you find that it seems a reasonable thing to associate a measure of information as some kind of a relation between this and basic probability of occurrence of an event and one could choose an arbitrary constant of proportionality here and one usually chooses that to be unity and defines this as information as measured in terms of bits per message symbol, or per unit, right?

(Refer Slide Time: 26:42)

$$\begin{aligned} m_1 \quad m_2 \quad m_3 \quad m_4 &: 2\text{-bits} \\ \Rightarrow 2 \text{ units of time} \\ n \Rightarrow \boxed{\log_2 n} \text{ units of time} \\ p = \frac{1}{4} \Rightarrow \frac{1}{n} \quad n \Rightarrow \frac{1}{p} \\ \text{Amount of time} \propto \log_2 \frac{1}{p} \\ I = \log_2 \frac{1}{p} \text{ bits/message symbol} \end{aligned}$$

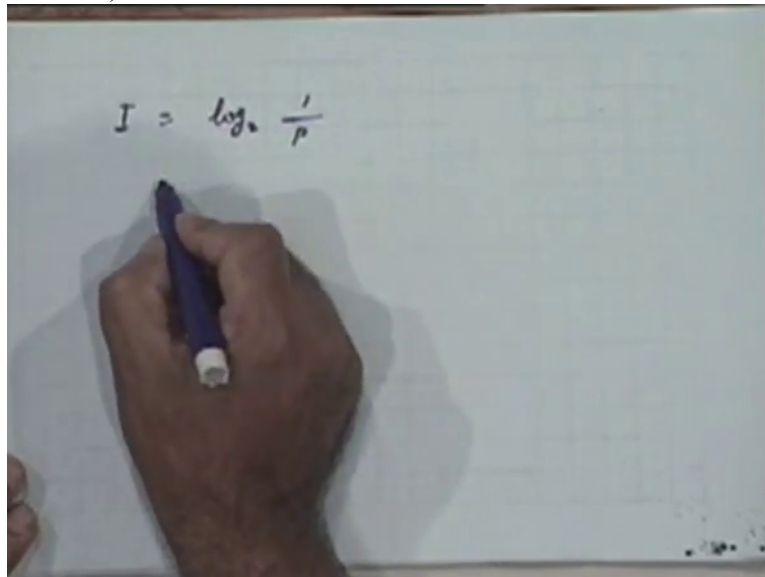
That is the unit that you use based on this example we have just taken.

For example for a message which has four symbols associated with it, right m_1 to m_4 , occurrence of any one of this symbols conveys 2 bits of information in the sense that we will require 2 units of time to transmit it, right? So that is a unit that we will associate with it, bits per message symbol, sometimes it is briefly written as bits per symbol, right or sometimes even per symbol is omitted and we just write bits. I think the per symbol is a very important thing which you should not forget. Even if you do not write it, it should be there in your mind, right.

It is the amount of information as a result of occurrence of a specific event. That event is occurrence of a particular symbol from a dictionary or s m s. So is the motivation for this definition understood by everybody? Is there some question on which you would like to discuss?

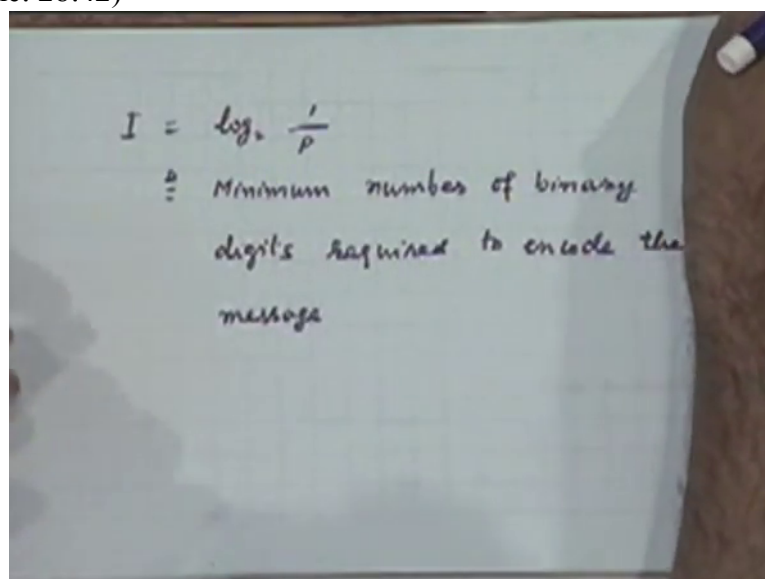
Also you could associate another interpretation to this but I will come back to that in a minute. So I is equal to $\log_2 \frac{1}{p}$ by p and

(Refer Slide Time: 28:12)


$$I = \log_2 \frac{1}{p}$$

we could say that, this we can also interpret as the minimum number of binary digits required to encode the message. This interpretation is based on the same example

(Refer Slide Time: 28:42)


$$I = \log_2 \frac{1}{p}$$

$\hat{=}$ Minimum number of binary digits required to encode the message

that we had just discussed. I have 2 message code, I have 2 message source, a source which emits one of two possible messages. How many bits are required, how many, how many information bits are required to encode it?

(Refer Slide Time: 28:58)



(Professor – student conversation starts)

Student: 1

Professor: 1 bit. For a 4 message source, all messages equi-probable we require 2 bits; 8 message source, 3 bits and so on.

(Professor – student conversation ends)

So in that sense, it seems to be also having this interpretation. Minimum number of binary digits required to encode the message, right? Encode in the sense, represent it by several number of bits, appropriate number of bits.

Now the next related concept is that of entropy. Not the entropy that you learnt in your thermodynamics, but entropy as we define in information theory. Now this is attribute of a source of information, right? It is an attribution of an information source.

(Refer Slide Time: 30:08)

$$I = \log_2 \frac{1}{p}$$

$\hat{=}$ Minimum number of binary digits required to encode the message

Entropy: Attribute of an Information Source

And precisely it is defined as average information associated with any message from that source, average, on an average. So it is really speaking, very briefly average information emitted by the source per message, or per message symbol to be more precise.

(Refer Slide Time: 30:34)

$$I = \log_2 \frac{1}{p}$$

$\hat{=}$ Minimum number of binary digits required to encode the message

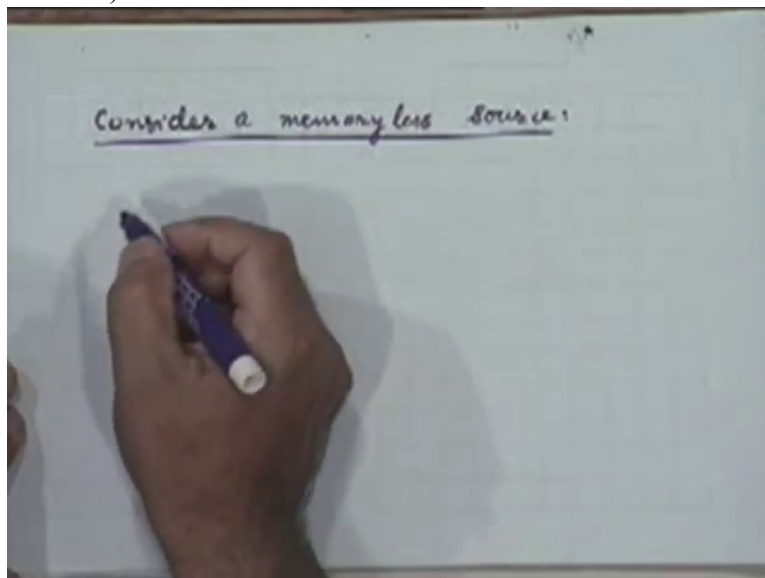
Entropy: Attribute of an Information Source

: Average information / message symbol

I am deliberately taking our source here to be of the discrete kind, right, because it is very convenient to do the discussion for that situation, a source of information to be of a discrete kind, or the kind which we deal with in digital communication, right? But of course a real source of information could also be a continuous valued source of information. Like all of us have continued value source of information because our speech signal which we emit is a continuous signal.

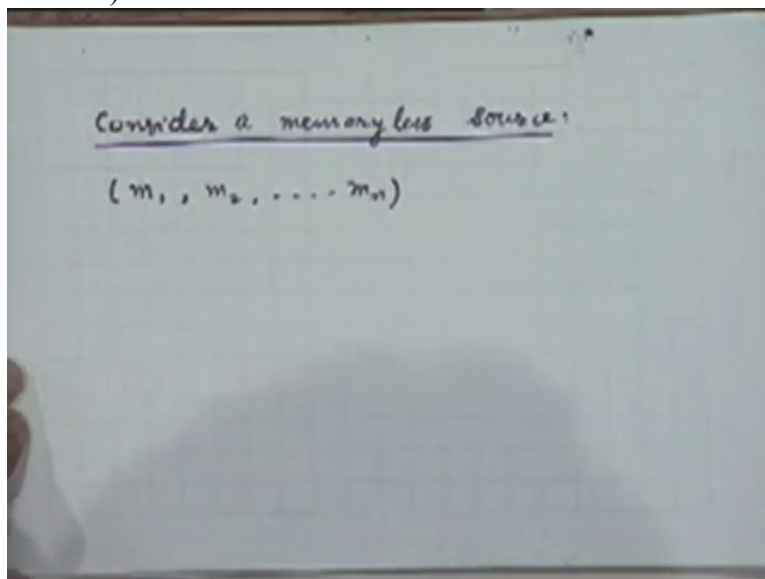
One can also develop a corresponding theory for continuous time and if we have time, we will spend some time on that. But even working purely with discrete sources, one does not lose that much of generality because we can always go from a continuous source to a corresponding discrete source representation through let us say analog to digital conversion or whatever, right? So, in other words, our discussion will be with respect to discrete sources. So well, so entropy is an attribute of an information source and is really defined as average information per message symbol associated with that source. More precisely, let me make a notion very precise. Let me define, or let me consider what we call a memoryless source.

(Refer Slide Time: 32:07)



Memoryless source is characterized by first, a set of symbols;

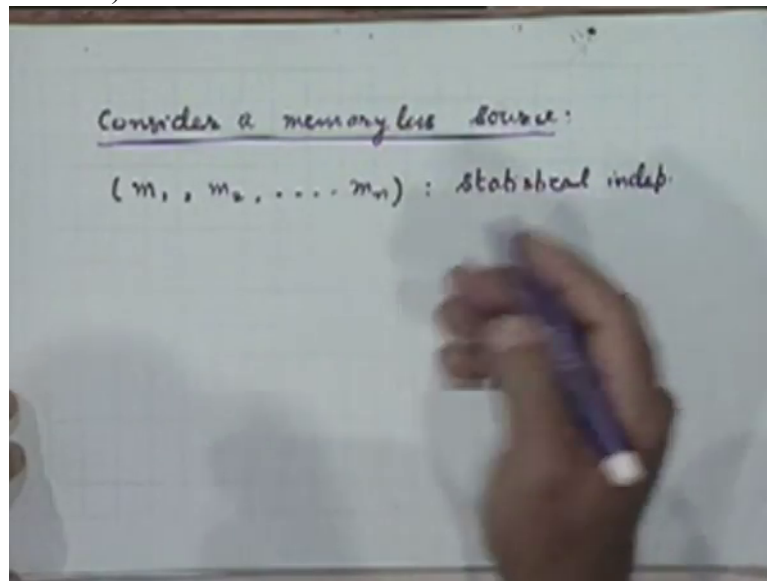
(Refer Slide Time: 32:17)



let us say n symbols which I am designating as m_1, m_2, \dots, m_n . So it comprises of n symbols which it may emit in any given unit of time. And it is called memoryless provided that successive symbols that are emitted by it, in successive units of time are statistically independent, right? They do not depend on what was emitted in the previous interval or will be emitted in the next interval, right?

So memoryless implies statistical independence of, that is,

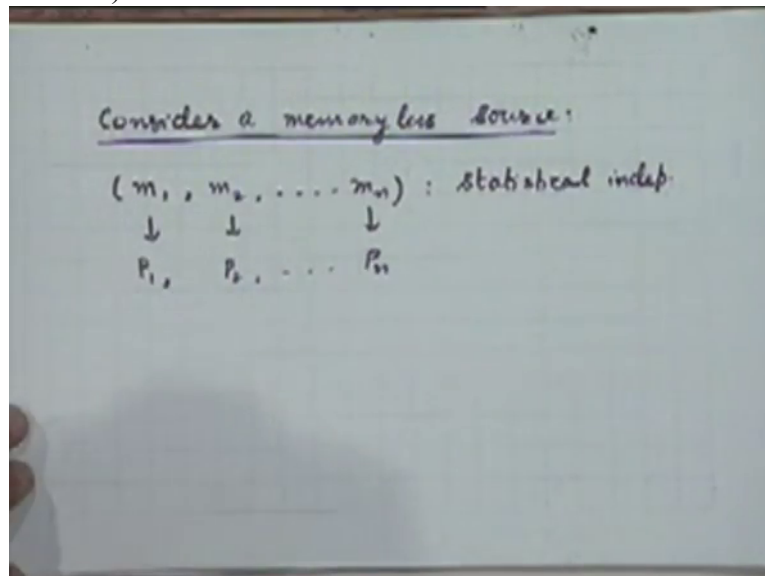
(Refer Slide Time: 32:58)



they do not remember at all what was emitted before or what will be emitted later. Let us associate with each of these, I am going to generalize things a little bit now. So far our discussion has been based on the fact that all of them are equi-probable message symbols. That is the source can emit any of these messages or any of these symbols with the same probability.

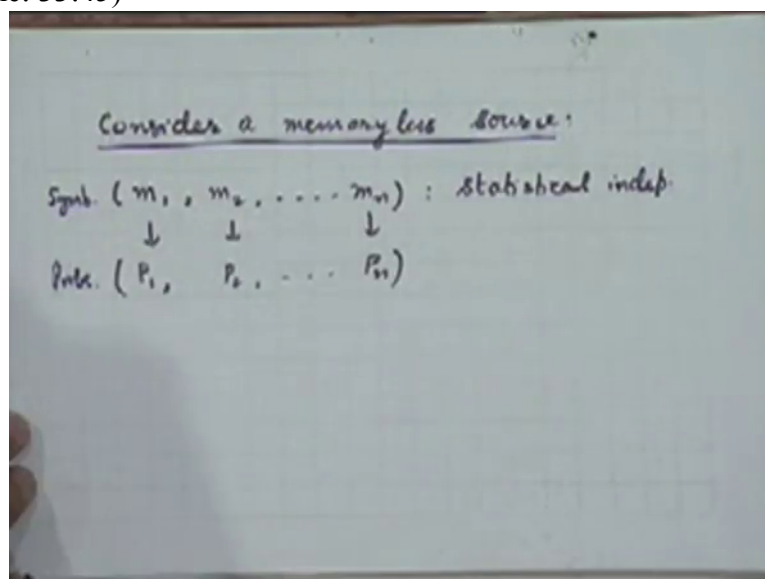
But now let me associate any arbitrary probability

(Refer Slide Time: 33:31)



p_1, p_2, p_n with each of these symbols. So these are the symbols, these are the probabilities.

(Refer Slide Time: 33:45)



Now what will be the information associated with emission of the message m_1 ?

(Professor – student conversation starts)

Student: 1 by

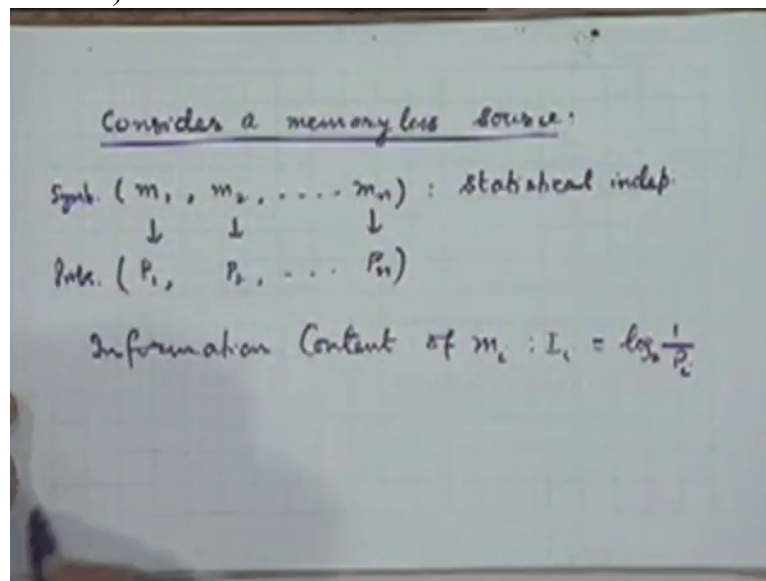
Professor: \log of 1 by p_1 . Similarly information associated with the occurrence of m_2 will be \log of 1 by p_2 and so on, right.

(Professor – student conversation ends)

So you can compute the information as each symbol is emitted. Now if I ask the question what is the average amount of information associated with this source or which this source emits on an average, then obviously it would be the weighted sum of all these information values we individually compute and the weight should be nothing but the corresponding probabilities, right.

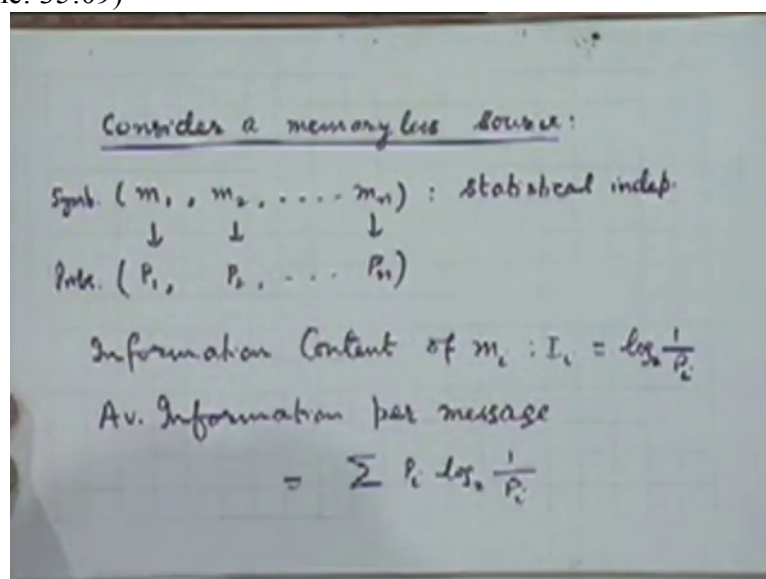
So information content of m_i , let us call this I_i and that is equal to $\log_2 \frac{1}{p_i}$,

(Refer Slide Time: 34:50)



and therefore average information per message symbol is simply $\sum p_i \log_2 \frac{1}{p_i}$

(Refer Slide Time: 35:09)



sub i. Or summed over i minus $\sum p_i \log_2 p_i$,

(Refer Slide Time: 35:21)

Consider a memoryless source:

Symb. (m_1, m_2, \dots, m_n) : statistical indep.
 \downarrow \downarrow \downarrow
Prob. (p_1, p_2, \dots, p_n)

Information Content of m_i : $I_i = \log_2 \frac{1}{p_i}$

Av. Information per message

$$= \sum_i p_i \log_2 \frac{1}{p_i} = - \sum_{i=1}^n p_i \log_2 p_i$$

Ok. The units will be bits because, or bits per symbol, any questions?

So the average information that is associated with the source again therefore depends on the probabilities tend to distribution of the various messages in the symbol, and this is called the entropy of the source, right? This is precisely what you call the entropy of the source or source entropy,

(Refer Slide Time: 36:08)

Consider a memoryless source:

Symb. (m_1, m_2, \dots, m_n) : statistical indep.
 \downarrow \downarrow \downarrow
Prob. (p_1, p_2, \dots, p_n)

Information Content of m_i : $I_i = \log_2 \frac{1}{p_i}$

Av. Information per message

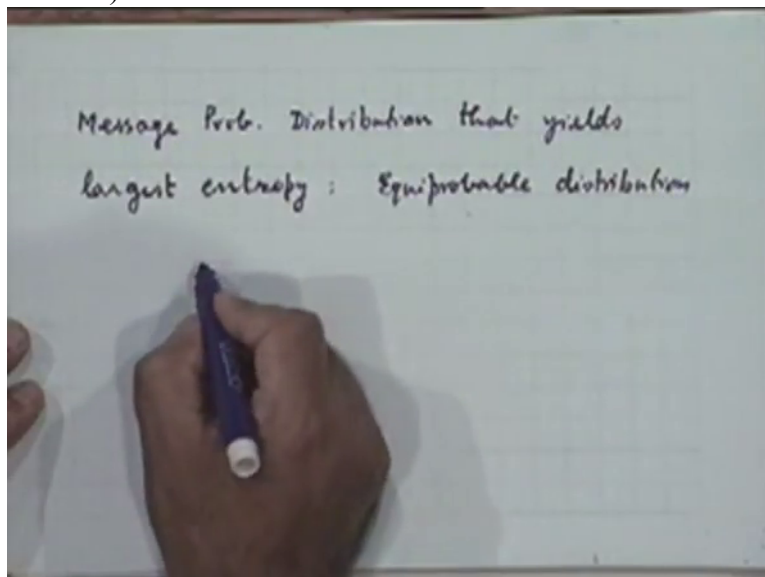
Entropy = $\sum_i p_i \log_2 \frac{1}{p_i} = - \sum_{i=1}^n p_i \log_2 p_i$

Ok. So entropy is an attribute not of a communication channel but of the source of information, remember that, right. It tells us how much information a source is emitting every time it emits a symbol, right? Basically that is what it tells us.

Now suppose I ask a very simple question, since this entropy depends on the probability distribution of various messages in the, in the dictionary of the source, for what kind of message source, discrete message source of this kind, or what distribution will this entropy be maximum? Answer should be obvious. Entropy will be maximized if all the probabilities are equal, right. One can prove this formally. I think I will leave that as an exercise for you to yourself, alright? So the message, I will just write down the result here, I would like you to read the proof yourself or perhaps try to prove it yourself.

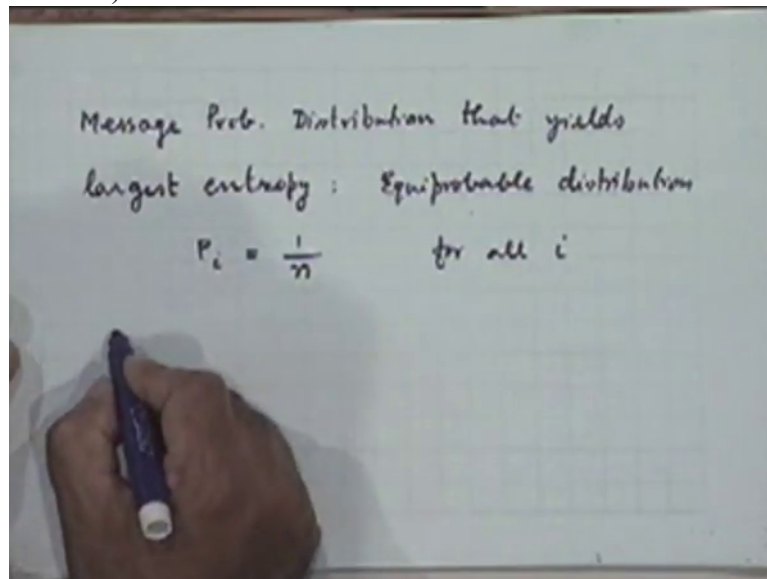
Message probability distribution that yields largest or maximum entropy is the equi-probable distribution, that is

(Refer Slide Time: 37:45)



p_i , they all, for all values of i should be equal to $1/n$ because the total number of symbols or messages in that source, associated with that source is assumed to be n for all i . If I substitute that in the expression

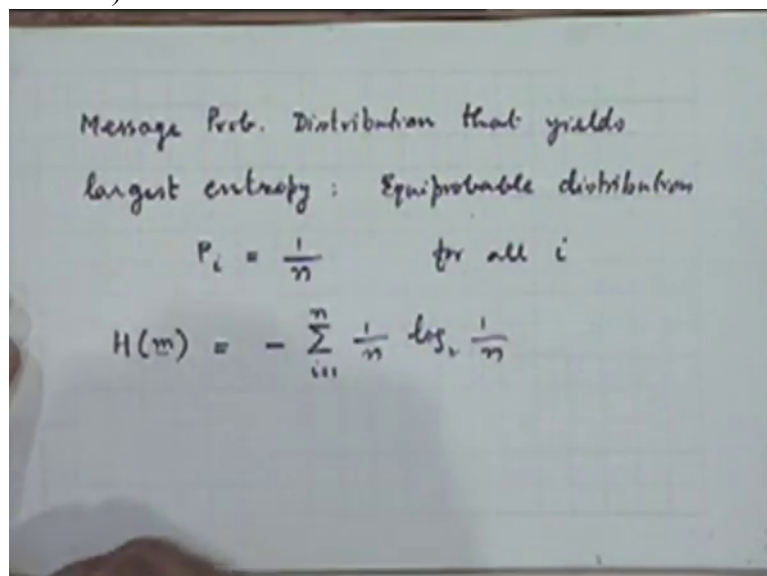
(Refer Slide Time: 38:06)



for entropy incidentally entropy is usually denoted by this notation H and sometimes you write an argument m saying that it is associated with that particular message source m , right? This is just a notation.

So this expression will then become minus 1 by m , right, i going from 1 to n .

(Refer Slide Time: 38:38)



Since this is independent of i this will become simply, right, n times this value

(Refer Slide Time: 38:55)

Message Prob. Distribution that yields
largest entropy: Equiprobable distribution

$$P_i = \frac{1}{n} \quad \text{for all } i$$
$$H(m) = - \sum_{i=1}^n \frac{1}{n} \log_2 \frac{1}{n}$$
$$= - \log_2 \frac{1}{n}$$

or simply $\log_2 n$,

(Refer Slide Time: 39:00)

Message Prob. Distribution that yields
largest entropy: Equiprobable distribution

$$P_i = \frac{1}{n} \quad \text{for all } i$$
$$H(m) = - \sum_{i=1}^n \frac{1}{n} \log_2 \frac{1}{n}$$
$$= - \log_2 \frac{1}{n} = \log_2 n$$

Ok which is interesting because you are familiar with this $\log_2 n$. In the context of equiprobable message source you could interpret $\log_2 n$ as the number of bits that will be required to encode that kind of source, right?

Therefore in general, entropy has, sorry, earlier we discussed in terms of a single symbol, now we are discussing in terms of source on an average irrespective of its probability distribution. We could interpret entropy as some kind of a measure of minimum number of digits required to encode the messages or the symbols of the source.

(Professor – student conversation starts)

Student: Sir why 0:39:57.7 minimum average number?

Professor: Minimum average number.

Student: Minimum average

Professor: Minimum average to encode the source.

(Refer Slide Time: 40:13)

Message Prob. Distribution that yields
largest entropy: Equiprobable distribution

$$P_i = \frac{1}{n} \quad \text{for all } i$$
$$H(m) = - \sum_{i=1}^n \frac{1}{n} \log_2 \frac{1}{n}$$
$$= - \log_2 \frac{1}{n} = \boxed{\log_2 n}$$

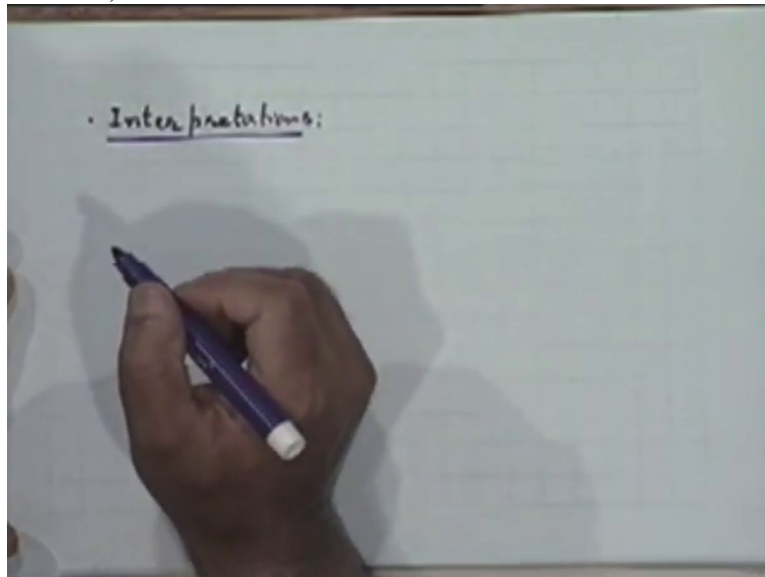
: Min. ^{Av.} no. of digits required to encode the source.

But I will come, return to that, that does not, this minimum aspect of this is something that I am just telling you from an intuitive discussion here. There is a proper 0:40:25.0 information theory really emphasizes the minimum part of it. We will talk about that later. Any questions so far? So what have we, so let us stop for a minute and take stock of what we have discussed so far.

(Professor – student conversation ends)

We have discussed the following interpretations of information and entropy. We have looked at

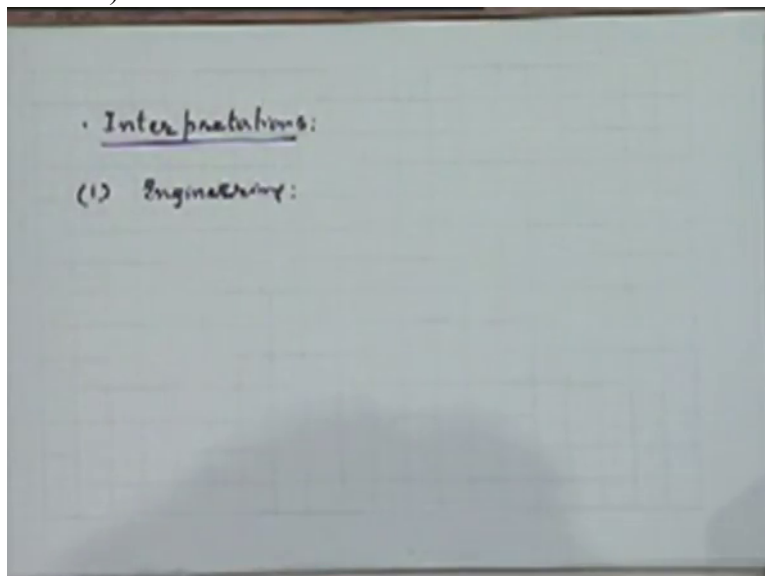
(Refer Slide Time: 40:59)



engineering point of view of information, engineering definition of information that is the information content of the message is the minimum number of digits required to encode the source, to encode the message sorry, not the source. And H_m is the minimum number of bits, average number of bits required to encode the source for each of these messages, right?

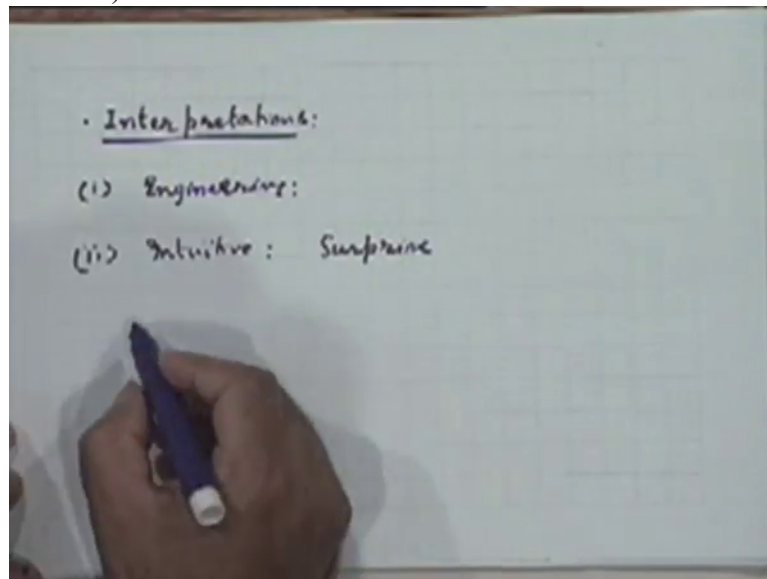
So we looked at the engineering point of view which is this. It is linked with the number of bits that are

(Refer Slide Time: 41:35)



required to represent a message, right? Then the intuitive picture that we have seen is associated with the surprise element, right?

(Refer Slide Time: 41:50)

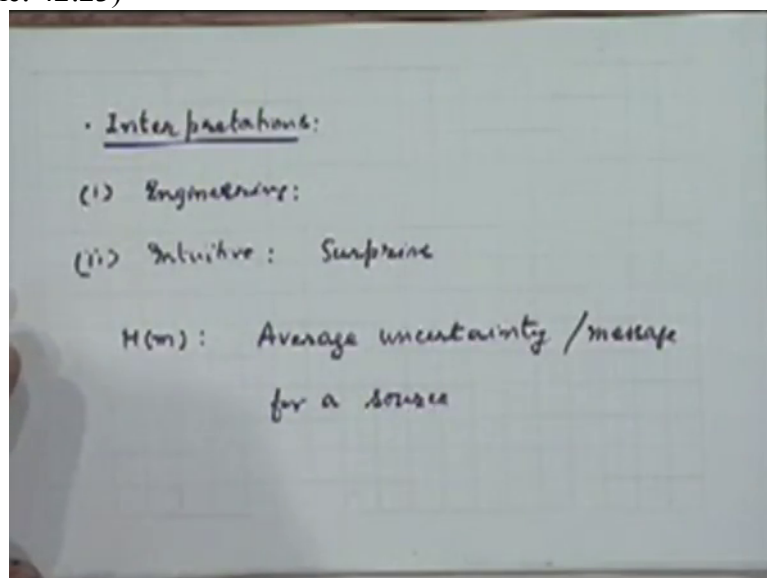


In which case $H(m)$ can be interpreted as the average amount of surprise or the average amount of uncertainty associated with the messages per message of the source, right? Average uncertainty. So these are terms you will come across when you are reading about these things. Average uncertainty associated with the message for a source.

(Professor – student conversation starts)

Student: What do we mean by

(Refer Slide Time: 42:23)



0:42:24.1, why do we say this is average message?

Professor: This is just reinterpretation of our earlier discussion of measure of information where we associated information content with element of surprise, element of uncertainty,

right? The more uncertain the event, the more message it conveys, right, the more information it conveys, right?

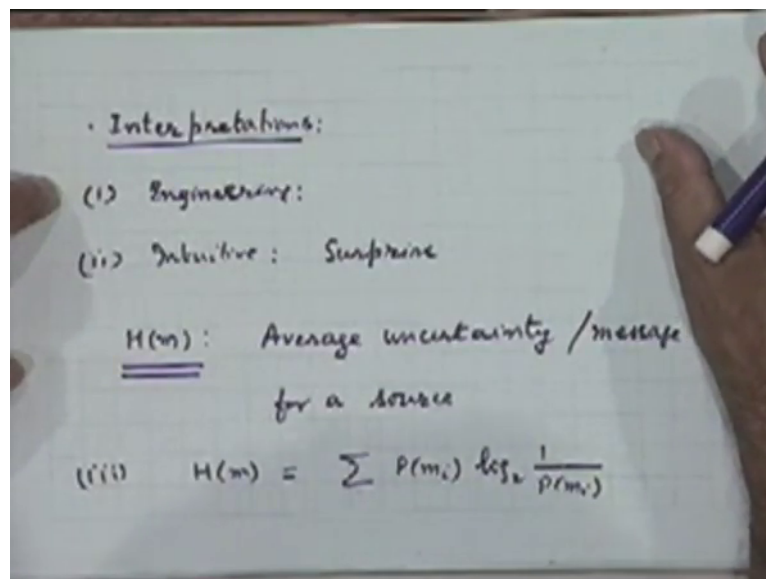
Student: According to the definitions?

Professor: That is all. We had earlier associated information with uncertainty. Now we are just extending the definition to entropy. That was based on event based interpretation. Now it, or single message bit based interpretation, now we are talking about an average over a source for every message that is associated with a source. That is why it is average uncertainty per message for a source.

(Professor – student conversation ends)

And thirdly we had a precise, we can re-affirm, we can reinterpret it just as a mathematical relationship between the probability distribution of the source and like this. So these are 3 different ways of looking at entropy. Here we look upon it as

(Refer Slide Time: 43:45)

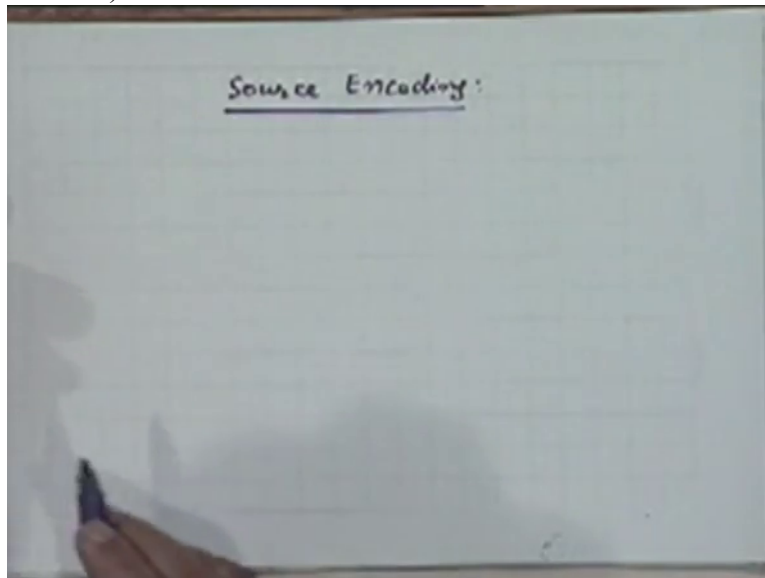


the minimum number of bits required, minimum number of binary digits required to encode messages of the source, on an average. Here we will call it as average uncertainty associated with each message of the source. Here we just regard it as a pure mathematical relation, right? These are 3 different ways of looking at it. Ok.

Next thing we would like to look at is the more precise relationship between what we call, source coding and entropy. We have seen now that entropy is a very important attribute of a

source, information source right and therefore it brings us to subject of source encoding. So in

(Refer Slide Time: 44:50)



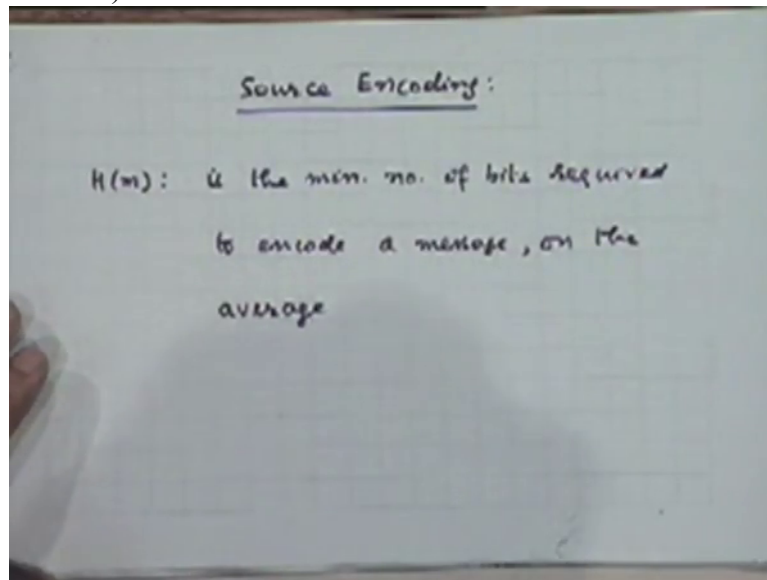
information theories therefore, representation of a source is just about as important as finding a good means for communicating it over the channel. Particularly with regard to the new result that we just discussed which was given by Shannon.

Because what Shannon said was for it to be possible for us to convey information without errors, what is really important is we keep the information rate below some number for a given channel, right? Therefore it is now important that we make sure that we do not use information rate far in excess of what is permitted. It should be in fact less than that. And therefore it is very important that for any information source that you might have, you represent the information coming out of that source efficiently.

It has a particular amount of information coming in but what engineer has to do is, find out what that amount of information is and to represent it, and code it by binary digits, by using an average number of binary digits no more than the average amount of information coming out of the source. If you do not do that, you might be doing a very inefficient job of information representation, right? Efficient information representation therefore is a very important concern of an information theorist, or therefore a communication theorist and that is precisely what we study under the title of source coding, or source encoding, Ok?

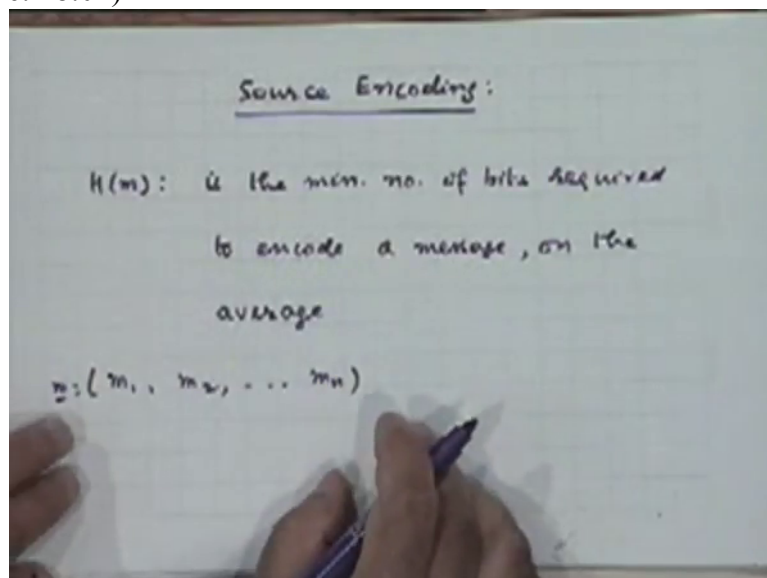
That is representing information coming out of a source with as few bits of encoder, encoding as possible. And we know what is the minimum number that is required. We have intuitively discussed a few times that somehow it is linked with the entropy of the source. Now we have shown that for equi-probable sources we would like to discuss that for other sources, that entropy is the minimum number of bits required to encode a source on an average.

(Refer Slide Time: 47:40)



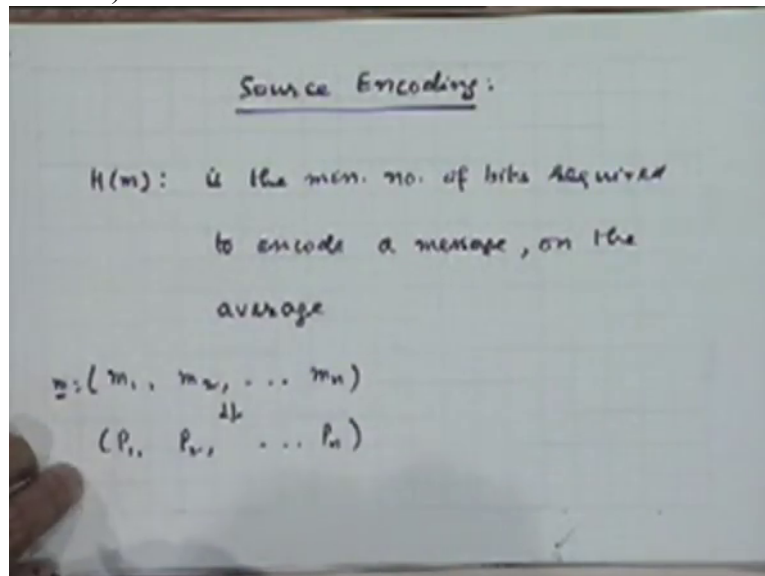
What I like to show is that if I take an arbitrary message source, let us say m_1, m_2 to $m_{sub n}$, that is your m , message source m

(Refer Slide Time: 48:01)



is a vector comprised of an arbitrary set of symbols with corresponding arbitrary probabilities associated with them.

(Refer Slide Time: 48:14)



What I would like to do is show whether it is possible to encode such a message with a number of, using binary digits such that average number of binary digits required is $H m$. Is it obvious to you that it can be done, a scheme of some kind coming out in your mind immediately?

(Professor – student conversation starts)

Student: 0:48:35.7

Professor: Suppose I ask, this is the interpretation you are giving to entropy. That is it is the minimum number of bits required to encode the message on the average, for various messages and source. In other words somehow implied by this is the fact that there is a particular procedure of mapping m_1 to some encoded sequence of binary bits, m_2 to a corresponding coded sequence and so on such that the average length of that set of sequences associated with various messages is no more than $H m$, right? Does such a practical procedure occur to you?

(Professor – student conversation ends)

In fact the first question that bothers us is 0:49:30.1, does such representation exist at all? Because this is what is meant by efficient representation of information. Efficient representation means if this is our feeling that no more than this number of bits are required in the average, then I must have at least 1 possible procedure by which I can carry out a mapping from these message symbols to binary digits such that average length is no more than $H m$. For equi-probable sources it is very easy to do that, isn't it?

(Professor – student conversation starts)

Student: 0:50:04.1

Professor: No, the procedure. Because I can associate, suppose I have a 4 symbol source, I have 4 messages. So I choose, I put any arbitrary set of 2 bits with any of these messages. So the coding procedure is simple. But for an arbitrary source, it is not obvious how we could do that, right?

So first I would like to discuss that and then come to a very important result called the source coding theorem.

Student: Can we assume there be equal probability for all the symbols?

Professor: Yeah we want to remove that assumption now. We want to be able to say that kind of result. I will give a procedure which is general enough for any probability distribution, for any kind of message source in which the probabilities may not be equal, right? We will take that up next time, thank you.

(Professor – student conversation ends)