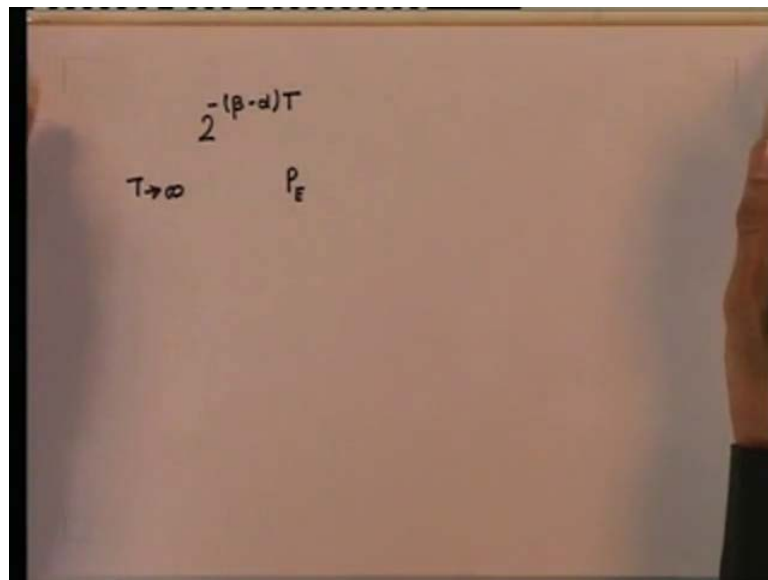


**Information Theory and Coding**  
**Prof. S. N. Merchant**  
**Department of Electrical Engineering**  
**Indian Institute of Technology, Bombay**

**Lecture - 28**  
**Error Free Communication over a Binary Symmetric Channel and Introduction to**  
**Continuous Sources and Channels**

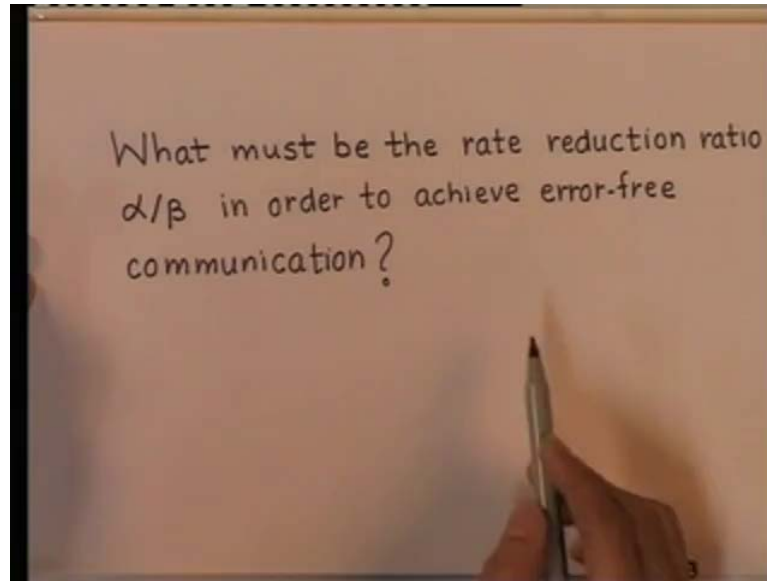
In the last class, we studied that the occupancy factor by transmitted message is  $2$  to the power negative of the term  $\beta$  minus  $\alpha$  times  $T$ .

(Refer Slide Time: 01:32)



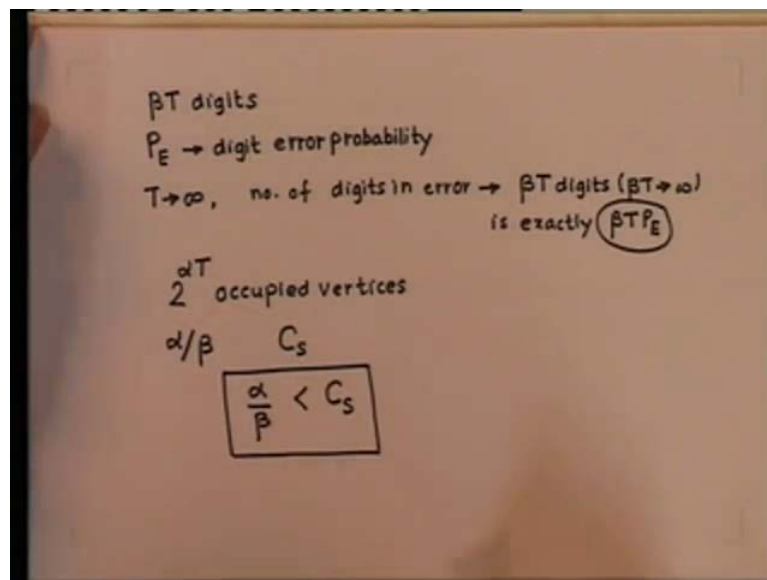
And this can be made as small as possible simply by increasing  $T$ . In the limit as  $T$  tends to infinity, the occupancy factor tends to 0. This will make the error probability  $P_E$  go to 0 and we have the possibility of error free communication. One important question however remains unanswered.

(Refer Slide Time: 02:25)



The question is what must be the rate reduction ratio that is alpha by beta in order to achieve error free communication?

(Refer Slide Time: 02:54)



To answer this question we observe that increasing  $T$  increases the length of the transmitted sequence, which is given by beta times capital  $T$  digits. Now, if  $P_E$  denotes the digit error probability then it can be seen from the relative frequency definition or the law of large numbers that as  $T$  tends to infinity the total number of digits in error in a sequence of beta times  $T$  digits is exactly beta times  $T$  error probability. Hence, the

received sequences will be at a hamming distance of  $\beta T P E$  from the transmitted sequence.

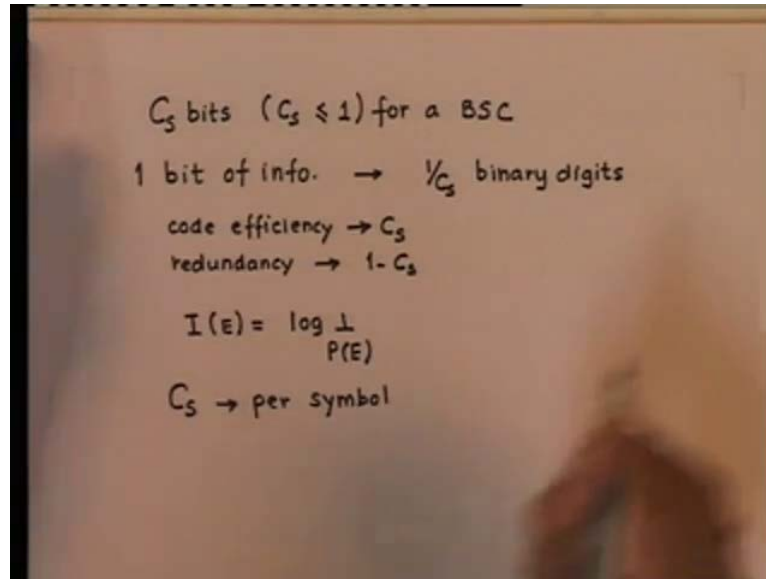
Therefore, for error free communication we must leave all the vertices unoccupied within spheres of radius given by  $\beta T P E$  drawn around each of the  $2^{\alpha T}$  occupied vertices. In short we must be able to pack  $2^{\alpha T}$  non overlapping spheres, each of radius  $\beta T P E$  into the hamming space of  $\beta T$  dimensions. This means that for a given  $\beta$   $\alpha$  cannot be increased beyond some limit without causing overlap in the spheres and the consequent failure of the scheme.

Shannon's theorem states that for this scheme to be successful  $\alpha$  by  $\beta$  ratio must be less than some constant and that constant is denoted by channel capacity  $C S$ , which is a function of the channel noise and the signal power. So, Shannon's theorem says that  $\alpha$  by  $\beta$  should be less than  $C S$  which is the channel capacity. It must be remembered that such perfect error free communication is not practical. In this system, we accumulate the information digits for  $T$  seconds before encoding them and because  $T$  tends to infinity for error free communication we must wait until eternity before we start encoding.

Hence, there will be an infinite delay at the transmitter and an additional delay of the same amount at the receiver. Second, the equipment needed for the storage encoding and decoding of the sequence of infinite digits would be monstrous. Needless to say that in practice the dream of error free communication cannot be achieved. Then the question is, what is the use of Shannon's theorem? First, it indicates the upper limit on the rate of error free communication that can be achieved on a channel.

This result in itself is monumental. Second, it indicates the way to reduce the error probability with only a small reduction in the rate of transmission of information digits. We can therefore, seek or compromise between error free communication with infinite delay and virtually error free communication with a finite delay. Next, let us investigate the problem of error free communication over a binary symmetric channel. We have seen that channel capacity is the property of a physical channel over which the information is transmitted. We have also shown that over a noisy channel  $C S$  bits of information can be transmitted per channel.

(Refer Slide Time: 09:08)

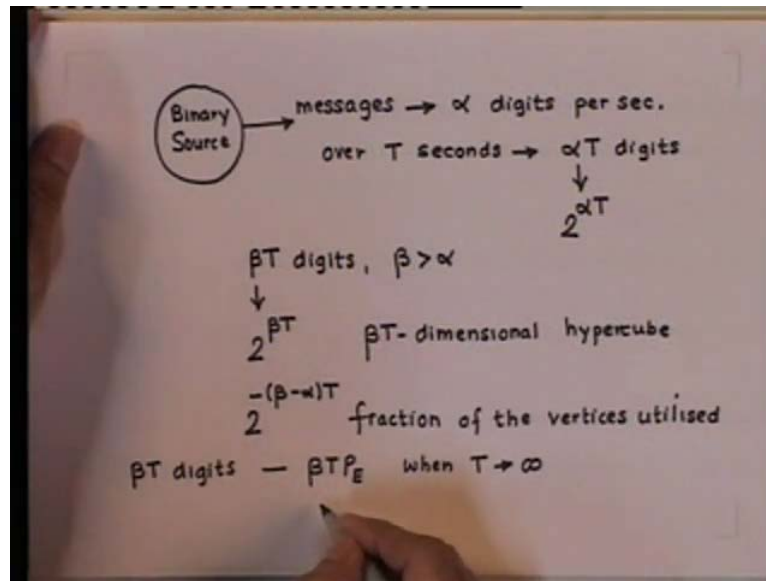


So, if we consider a binary channel what this means is that for each binary digit or symbol transmitted the received information is  $C_S$  bits where  $C_S$  is less than equal to 1 for a binary symmetric channel. Thus, to transmit one bit of information over a binary symmetric channel we need to transmit at least  $\frac{1}{C_S}$  binary digits. This scheme gives us a code efficiency that is  $C_S$  and redundancy as  $1 - C_S$ . When a transmission of information is implied it means error free communication because mutual information was defined as the transmitted information minus the loss of information caused by the channel noise.

The problem with this derivation is that it is based on a speculative definition of information. The problem with this derivation is that it is based on a speculative definition of information. That is information associated with the occurrence of a particular  $E$  is given by  $I$  equal to  $\log$  of  $\frac{1}{P(E)}$  by probability of occurrence of that event. And based on this definition we defined the information lost during the transmission over the channel.

Now, we really have no direct proof that the information lost over the channel will oblige in this way. Hence, the only way to ensure that this whole speculative structure is sound is to verify it. So, if we can show that  $C_S$  bits of error free information can be transmitted per symbol over a channel then the verification will be complete. Here we shall verify the results for a binary symmetric channel.

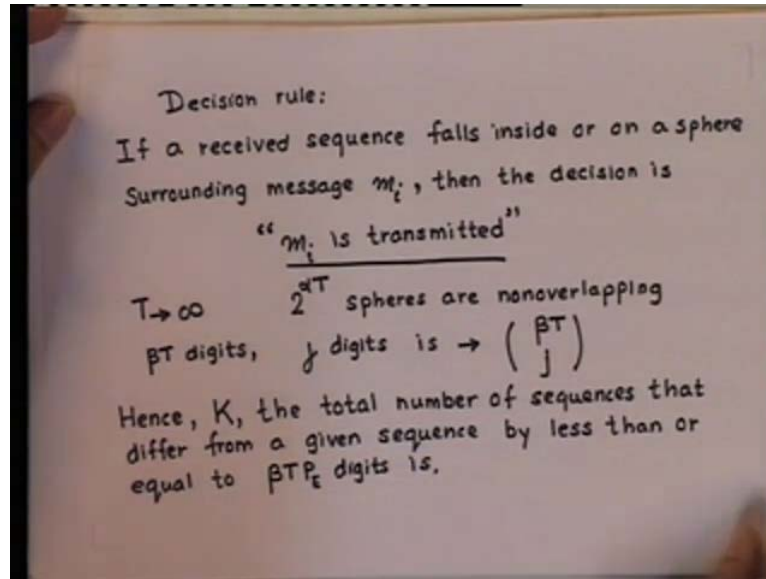
(Refer Slide Time: 12:19)



Let us consider a binary source. This source emits messages at the rate of alpha digits per second. We accumulate these information digits over T seconds to give a total of alpha T digits. Now, because alpha T digits form 2 raised to alpha T possible combinations, our problem now, is to transmit one of this 2 raised to alpha T super messages every T seconds. These super messages are transmitted by a code word of length beta times T digits where beta is greater than alpha to ensure redundancy. Now, because beta times T digits can form 2 raised to beta times T distinct patterns which are the vertices of a beta times T dimensional hypercube and we have only 2 raised to alpha times T super messages.

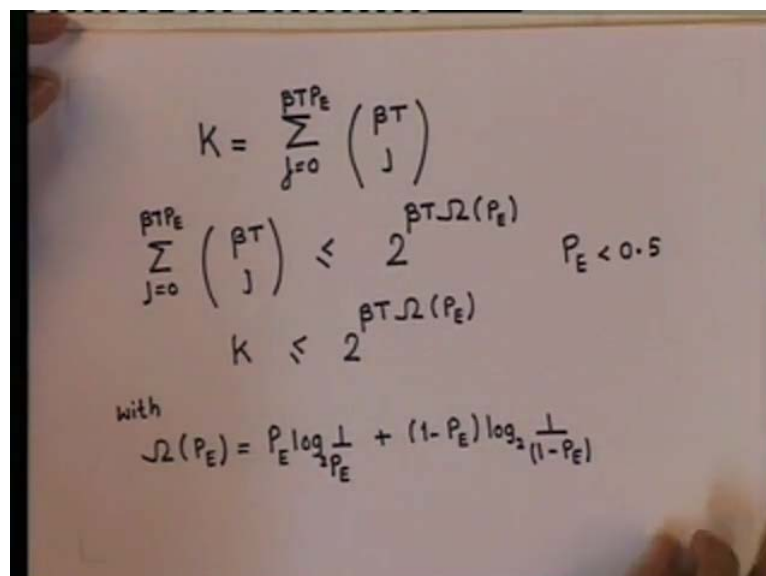
We are utilizing only a 2 raised to minus beta minus alpha times T fraction of the vertices. The remaining vertices are deliberately unused in order to combat noise. Now, if you let T tend to infinity the fraction of the vertices used approaches 0 and because there are beta times T digits in each transmitted sequence the number of digits received in error will be exactly beta times T multiplied by digit error probability which is given by P E when T tends to infinity. We now construct hamming spheres of radius beta times T P E around each of the 2 raised to alpha T vertices which are used for the messages. When any message is transmitted the received message will be on the hamming sphere surrounding the vertex corresponding to that message.

(Refer Slide Time: 16:05)



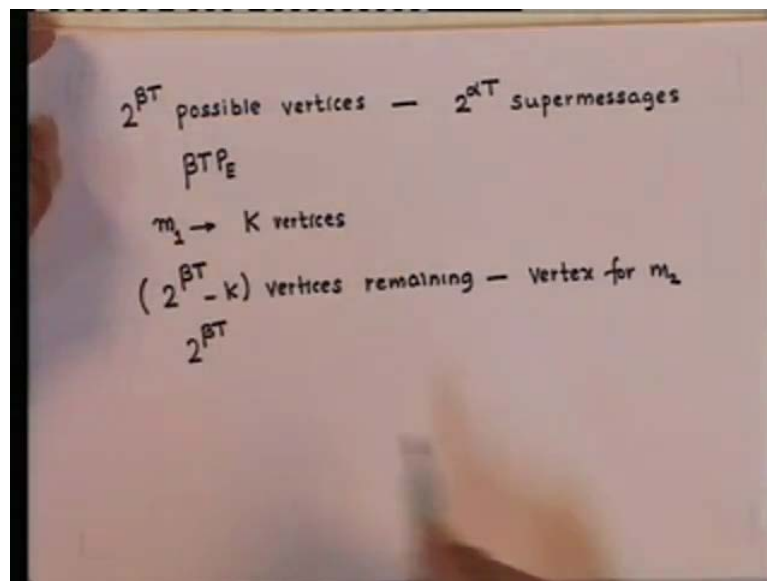
We use the following decision rule. If a received sequence falls inside or on a sphere surrounding message  $m_i$ , then the decision is  $m_i$  is transmitted. Now, if  $T$  tends to infinity the decision will be without error if all the  $2^{\alpha T}$  spheres are non overlapping. Now, of all the possible sequences of  $\beta T$  digits the number of sequences that differ from given sequence by exactly  $j$  digits is given by  $\binom{\beta T}{j}$ , this combination.

(Refer Slide Time: 19:28)



Hence, capital  $K$  which denotes the total number of sequences that differ from a given sequence by less than or equal to  $\beta T P E$  digits is  $K$  is equal to summation over  $j$  is equal to 0 to  $\beta T P E$ . Now, if we use an inequality which is often used in information theory, the inequality is less than equal to 2 to the power  $\beta T$  entropy function which is function of error probability where error probability is less than 0.5. Using this inequality what it implies that  $K$  is less than equal to 2 to the power  $\beta T$  entropy function with entropy function given as  $P E \log$  of 1 by  $P E$  plus 1 minus  $P E \log$  to the base 2 1 minus...

(Refer Slide Time: 21:41)

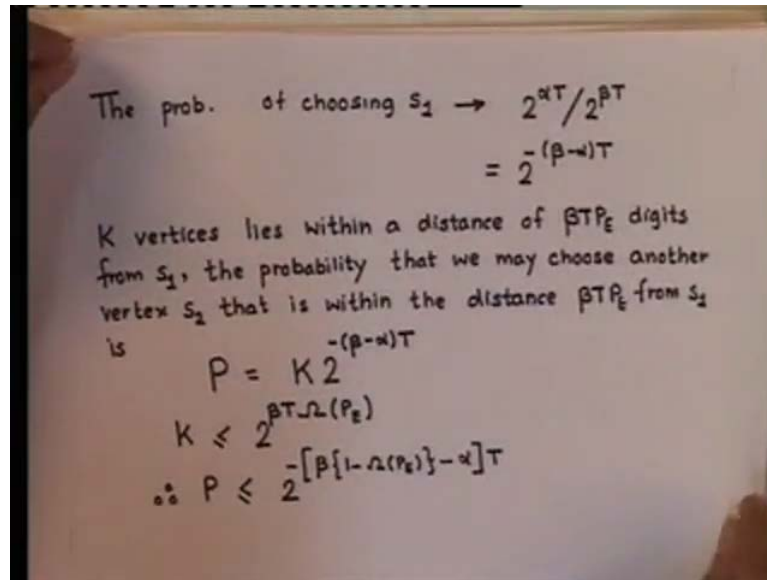


Now, from the 2 raise to  $\beta T$  possible vertices we chose 2 raise to  $\alpha T$  vertices to be assigned to the super messages. How shall we select this vertices is the next question. So, let us look at the decision procedure. From the decision procedure it is clear that if you assign a particular vertex to a super message then none of the other vertices lying within a sphere of radius of  $\beta T P E$  can be assigned to another super message.

Thus, when we choose a vertex for a message say  $m_1$  the corresponding  $K$  vertices become ineligible for consideration. Then from the remaining  $2^{\beta T} - K$  vertices we choose another vertex for  $m_2$ . We proceed in this way until all the  $2^{\beta T}$  vertices are exhausted. Now, this is a rather tedious procedure. So, let us see what happens if we choose the required  $2^{\alpha T}$  vertices randomly from the  $2^{\beta T}$

to beta T vertices. Now, if we adopt this procedure then there is a danger that we may select more than 1 vertex lying within a distance beta T P E. If however alpha by beta is sufficiently small the probability of making such a choice is extremely small as T tends to infinity. Let us look at this in a little more detail.

(Refer Slide Time: 24:40)



Now, the probability of choosing any particular vertex  $S_1$  as one of the  $2^{\alpha T}$  vertices from  $2^{\beta T}$  vertices is given by  $2^{\alpha T} / 2^{\beta T}$  which is equal to  $2^{-(\beta-\alpha)T}$ . Now, remembering that  $K$  vertices lies within a distance of  $\beta T P_E$  digits from the vertex  $S_1$ , the probability that we may choose another vertex  $S_2$  that is within the distance  $\beta T P_E$  from the vertex  $S_1$  is given by the expression  $P$  that is this probability is equal to  $K$  times  $2^{-(\beta-\alpha)T}$ . Now, we have shown earlier that  $K$  is less than equal to  $2^{\beta T \Omega(P_E)}$ . Therefore, from this equation it follows that probability of choosing another vertex  $S_2$  that is within the distance  $\beta T P_E$  from  $S_1$  is less than equal to  $2^{-[\beta\{1-\Omega(P_E)\}-\alpha]T}$ .



(Refer Slide Time: 27:47)

The image shows a whiteboard with handwritten mathematical notes. At the top, it says 'T → ∞, P → 0 if β {1 - H(p\_e)} > α'. Below that, 'i.e. α/β < {1 - H(p\_e)}'. Then, 'But {1 - H(p\_e)} is → C\_s → BSC'. At the bottom, there is a boxed equation 'α/β < C\_s' and another equation 'α/β = C\_s - ε' with a hand pointing to it.

$$T \rightarrow \infty, P \rightarrow 0 \text{ if } \beta \{1 - H(p_e)\} > \alpha$$

i.e.

$$\frac{\alpha}{\beta} < \{1 - H(p_e)\}$$

But  $\{1 - H(p_e)\}$  is  $\rightarrow C_s \rightarrow \text{BSC}$

$$\boxed{\frac{\alpha}{\beta} < C_s} \quad \frac{\alpha}{\beta} = C_s - \epsilon$$

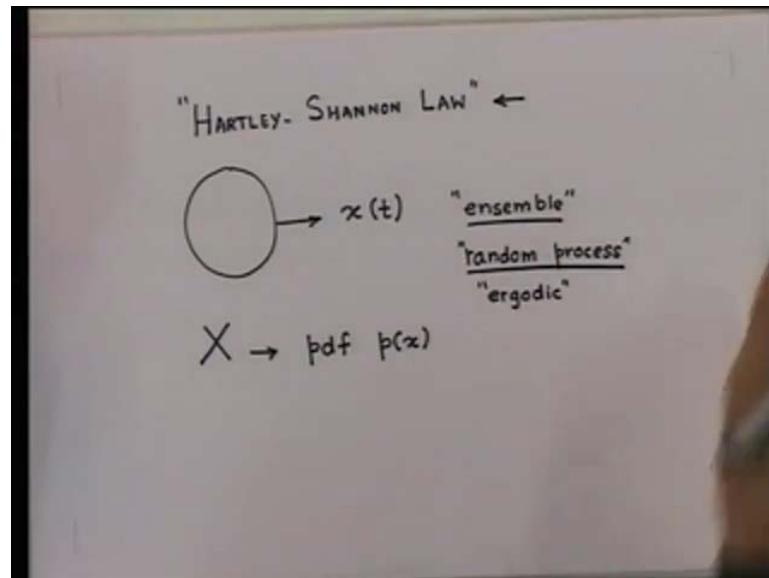
If you look at this expression as  $T$  tends to infinity this probability tends to 0 if the following condition is satisfied  $\beta$  times 1 minus entropy function is greater than  $\alpha$ . That is if  $\alpha$  by  $\beta$  is less than 1 minus entropy. Now, 1 minus entropy function is the channel capacity  $C_s$  for a binary symmetric channel. Therefore, we conclude that this probability of choosing another vertex  $S_2$  that is within the distance of  $\beta T P E$  from  $S_1$  will tend to 0 if  $\alpha$  by  $\beta$  is less than channel capacity  $C_s$ .

Hence, the probability of choosing two sequences randomly within a distance of  $\beta T P E$  approaches 0 as  $T$  tends to infinity, provided  $\alpha$  by  $\beta$  is less than  $C_s$ . And in this case we have error free communication. We can choose  $\alpha$  by  $\beta$  is equal to  $C_s$  minus  $\epsilon$  where  $\epsilon$  is arbitrarily small. So, we have verified the Shannon second theorem of a error free communication for a binary symmetric channel. So, far in our study the sources and channels considered in our discussion of information theoretic concepts have involved and symbols of random variables that are discrete in amplitude.

Next, we will extend some of these concepts to continuous random variables and random vectors. The motivation for doing so is to pave the way for the description of channel capacity in terms of the band width of the channel, channel noise and signal power. Having developed concepts of information transmission for discrete case we are now ready to tackle the more realistic case of a continuous source and channel. We will begin with the measure of information for a source that emits continuous signal. The material

may seem heavy going at first, but we will then make reasonable assumptions about transmission of continuous signals to express the channel capacity in terms of band width and signal to noise ratio.

(Refer Slide Time: 31:56)



This result is known as Hartley Shannon law. This result leads us to the definition of an ideal communication system which serves as a standard for system comparison and a guide to design improved communication systems. A continuous information source produces a time varying signal denoted by  $x(t)$ , we will treat the set of possible signals as an ensemble of wave forms generated by some random process which is assumed to be ergodic.

And by definition ergodic process means that time averages and ensemble averages are the same. We will also assume that the process has a finite band width meaning that the signal  $x(t)$  is completely characterized in terms of periodic sample values. Thus, at any sampling instance the collection of possible sample values constitutes a continuous random variable denoted by capital  $X$  and described by its probability density function  $p(x)$ .

(Refer Slide Time: 34:19)

The image shows handwritten mathematical derivations on a whiteboard. The first line shows a discrete random variable  $X$  taking values  $x_1, x_2, \dots, x_n$  with probabilities  $P(x_1), P(x_2), \dots, P(x_n)$ . The second line defines the entropy  $H(X)$  as the summation  $\sum_{i=1}^n P(x_i) \log \frac{1}{P(x_i)}$ , labeled as equation (1). The third line defines the entropy  $H(X)$  for a continuous random variable as the integral  $\int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx$ , labeled as equation (2). The fourth line shows a small interval  $X$  in the range  $[n\Delta x, (n+1)\Delta x]$  with probability  $p(n\Delta x)\Delta x$ , and notes the limit as  $\Delta x \rightarrow 0$ .

Now, we have seen that for a discrete random variable  $X$  which takes on the values  $x_1, x_2$  up to  $x_n$  with probabilities  $P(x_1), P(x_2), P(x_n)$ . The entropy  $H(X)$  was defined as summation of  $P(x_i) \log \frac{1}{P(x_i)}$ ,  $i$  equal to 1 to  $n$ . Now, we can extend the definition of entropy to continuous random variable by using the integral instead of discrete summation in equation number one. So, if we do that we can define the entropy of a continuous random variable as integral minus infinity to plus infinity of  $p(x) \log \frac{1}{p(x)}$ .

We shall see that equation two is indeed the meaningful definition of entropy for a continuous random variable. However, we cannot accept this definition unless we show that it also has a meaningful interpretation in terms of uncertainty. Our random variable  $X$  takes a value in the range  $n\Delta x, (n+1)\Delta x$  with probability  $p(n\Delta x)\Delta x$  multiplied by  $\Delta x$  in the limit as  $\Delta x$  tends to 0. Now, the error in the approximation will vanish in the limit as  $\Delta x$  tends to 0.

(Refer Slide Time: 37:30)

$$\begin{aligned}H(x) &= \lim_{\Delta x \rightarrow 0} \sum_n p(n\Delta x) \Delta x \log \frac{1}{p(n\Delta x) \Delta x} \\&= \lim_{\Delta x \rightarrow 0} \left[ \sum_n p(n\Delta x) \Delta x \log \frac{1}{p(n\Delta x)} - \sum_n p(n\Delta x) \Delta x \log \Delta x \right] \\&= \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx - \lim_{\Delta x \rightarrow 0} \log \Delta x \int_{-\infty}^{\infty} p(x) dx \\&= \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx - \lim_{\Delta x \rightarrow 0} \log \Delta x \rightarrow \textcircled{3} \leftarrow -\infty\end{aligned}$$

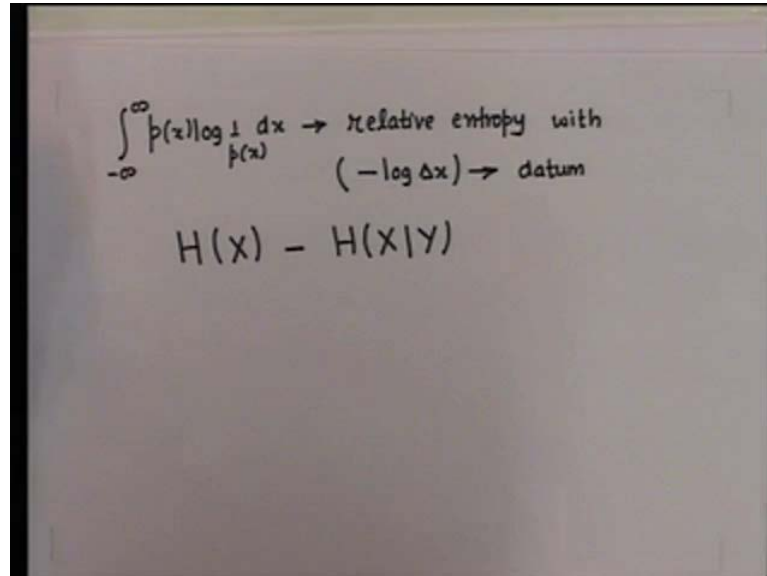
So, using this simplification for a continuous random variable we can define the entropy for the continuous random variable  $X$  as  $H(X)$  is equal to limit of  $\Delta x$  tending to 0 summation of probability of  $n\Delta x \Delta x \log$  of 1 by  $p(n\Delta x)$  multiplied by  $\Delta x$ . This is the summation over  $n$  and this can be simplified as limit of  $\Delta x$  tending to 0. We can break up this summation in two parts as follows, log of minus summation over  $n$ ,  $p(n\Delta x) \Delta x \log$  of  $\Delta x$ .

Now, as  $\Delta x$  tends to 0 this summation can be approximated by the integral which gives the following expression  $\int p(x) \log$  of 1 by  $p(x) dx$  minus limit  $\Delta x$  tending to 0  $\log \Delta x$  integral of probability distribution function from minus infinity to plus infinity. And this can be further simplified as... So finally, we get the expression for entropy for a continuous random variable which is indicated by equation number three. So, in the limit as  $\Delta x$  extends to 0  $\log$  of  $\Delta x$  will tend to minus infinity.

So, it appears that the entropy of a continuous random variable is infinite. Now, this is quite true. The magnitude of uncertainty associated with a continuous random variable is infinite. This fact is also apparent intuitively. Continuous random variable assumes an uncountable infinite number of values and hence the uncertainty is on the order of infinity. So, does this mean that there is no meaningful definition of entropy for a continuous random variable. On the contrary we will see that the first term in the

equation number three serves as a meaningful measure of the entropy that is average information for a continuous random variable  $x$ . Now, this may be argued as follows.

(Refer Slide Time: 41:26)



$$\int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx \rightarrow \text{relative entropy with}$$

$$(-\log \Delta x) \rightarrow \text{datum}$$

$$H(X) - H(X|Y)$$

We can consider integral  $p(x) \log \frac{1}{p(x)}$  as a relative entropy with  $-\log \Delta x$  serving as a datum or a reference. Now, the information transmitted over the channel is actually the difference between the two terms entropy of  $X$  and entropy of  $X$  given  $Y$ . Now, obviously if we have our common datum for both  $H(X)$  and  $H(X|Y)$  then the difference  $H(X) - H(X|Y)$  will have the same difference as the difference between the relative entropies.

(Refer Slide Time: 43:09)

$$\begin{aligned}
 H(X) &= \lim_{\Delta x \rightarrow 0} \sum_n p(n\Delta x)\Delta x \log \frac{1}{p(n\Delta x)\Delta x} \\
 &= \lim_{\Delta x \rightarrow 0} \left[ \sum_n p(n\Delta x)\Delta x \log \frac{1}{p(n\Delta x)} - \sum_n p(n\Delta x)\Delta x \log \Delta x \right] \\
 &= \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx - \lim_{\Delta x \rightarrow 0} \log \Delta x \int_{-\infty}^{\infty} p(x) dx \\
 &= \underbrace{\int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx}_{\text{Differential entropy}} - \lim_{\Delta x \rightarrow 0} \log \Delta x \rightarrow \textcircled{\infty} \\
 &\qquad\qquad\qquad \searrow -\infty
 \end{aligned}$$

So, we are therefore justified in considering the first term in equation three as the relative entropy of X or sometimes it is also known as differential entropy of X. We must however remember that this is relative entropy and not absolute entropy. Failure to realize this crucial point generates many apparent fallacies. Let us try to take an example to understand this.

(Refer Slide Time: 43:42)

$x(t) \quad y(t)$   
 $X \rightarrow [-1, 1]$   
 $p(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & \text{otherwise} \end{cases}$   
 $H(X) = \int_{-1}^1 \frac{1}{2} \log 2 dx = 1 \text{ bit}$

$x(t) \xrightarrow{x \times 2} y(t)$   
 $Y$   
 $[-2, 2]$   
 $p(y) = \begin{cases} \frac{1}{4} & |y| < 2 \\ 0 & \text{otherwise} \end{cases}$   
 $H(Y) = \int_{-2}^2 \frac{1}{4} \log 4 dx = 2 \text{ bits}$   
 $Y = 2X$

In particular consider the two information signals  $x(t)$  and  $y(t)$ . A signal amplitude  $X$  associated with the signal  $x(t)$  is a random variable uniformly distributed in the range

minus 1 to plus 1. This signal  $x(t)$  is passed through an amplifier of gain 2 to obtain the signal  $y(t)$ . Therefore, the output of this process which is a continuous random variable  $y$  is also uniformly distributed in the range minus 2 to plus 2. So, we have probability distribution function for the random variable  $x$  given by half when  $\text{mod } x$  is less than 1 and 0 otherwise. And probability distribution function for  $y$  is given as one-fourth when  $\text{mod } y$  is less than 2 and is equal to 0 otherwise.

Now, using this pdf's we can calculate the relative entropy's of  $x$  and  $y$  as follows. Entropy of  $y$  is equal to 2 bits. Now, the entropy of the random variable  $y$  is twice that of  $x$ . This result may come as a surprise since a knowledge of  $x$  uniquely determines  $y$  and vice versa because  $y$  is equal to twice of  $x$ . Hence, the average uncertainty of  $x$  and  $y$  should be identical. Amplification by itself can neither add or subtract information. So, the question is why there is  $H$  of  $Y$  as twice as large, why there is  $H$  of  $Y$  is twice as large as  $H$  of  $X$ . This becomes clear when we remember that  $H$  of  $X$  and  $H$  of  $Y$  are differential entropy's. And they will be equal only if the datum or references for both the random variable  $x$  and  $y$  are equal.

(Refer Slide Time: 47:10)

The image shows a whiteboard with handwritten mathematical derivations. At the top right, there is a small diagram with 'x' and 'y' connected by a horizontal line. The main text on the whiteboard is as follows:

$$R_1 \text{ (for } X) \rightarrow -\log \Delta x$$

$$R_2 \text{ (for } Y) \rightarrow -\log \Delta y$$

$$R_1 = \lim_{\Delta x \rightarrow 0} -\log \Delta x$$

$$R_2 = \lim_{\Delta y \rightarrow 0} -\log \Delta y$$

$$R_1 - R_2 = \lim_{\Delta x, \Delta y \rightarrow 0} \log \left( \frac{\Delta y}{\Delta x} \right) = \log \frac{dy}{dx} = \log 2 = 1 \text{ bit.}$$

Now, the reference entropy  $R_1$  for random variable  $X$  is minus log delta  $x$  and reference entropy for random variable  $Y$  is minus log delta  $y$  in the limit. So,  $R_1$  is equal to limit delta  $x$  tending to 0 minus log delta  $x$  and  $R_2$  is equal to limit delta  $y$  tending to 0 minus log delta  $y$ . Therefore, the difference  $R_1$  minus  $R_2$  is equal to limit delta  $x$  delta  $y$

tending to 0 of  $\log \frac{\Delta y}{\Delta x}$ , which is equal to  $\log \frac{dy}{dx}$  is equal to  $\log$  of 2 is equal to 1 bit.

Thus,  $R_1$  the reference entropy of  $x$  is higher than the reference entropy  $R_2$  for  $y$  by 1 bit. Hence, if  $x$  and  $y$  have equal absolute entropy's the differential or relative entropy's must differ by 1 bit. Now, we have seen that the entropy for a discrete random variable is maximum, when all the outcomes have equal probabilities. The next question is, is it possible to derive such relationship for a continuous random variable? We will investigate this in the next class.