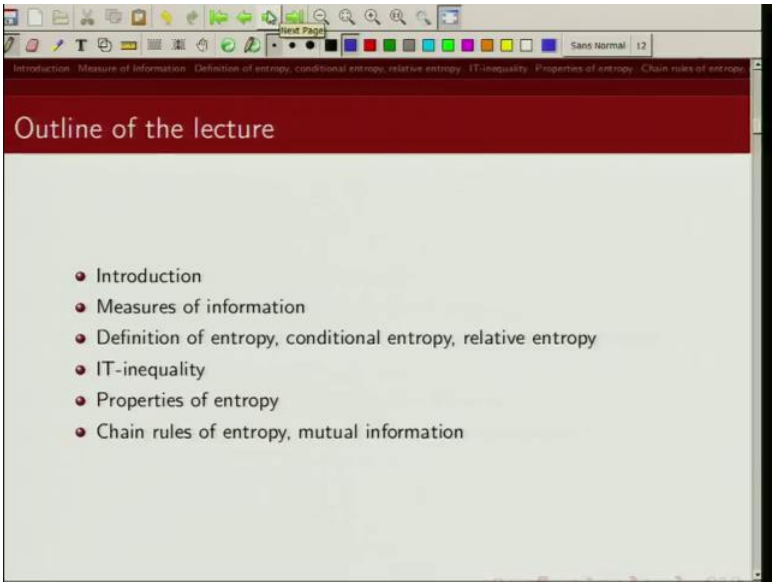
An Introduction to Information Theory
Prof. Adrish Banerjee
Department of Electronics and Communication Engineering
Indian Institute of Technology, Kanpur

Lecture – 1B
Measure of Information

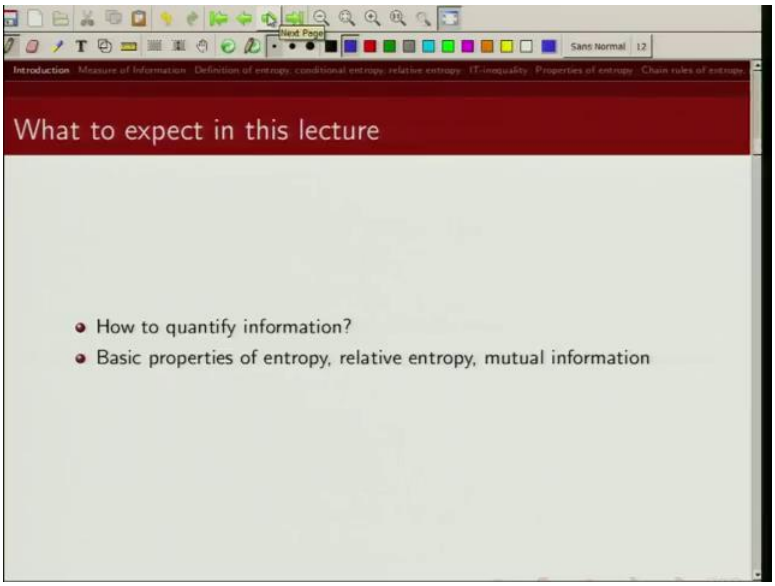
Welcome to the course on an introduction to information theory today in this lecture we are going to talk about how to quantify information. So, we will talk about the channels measure of information and then we will talk about entropy, we will define entropy condition entropy, joint entropy, relative entropy and we will prove some properties of them.

(Refer Slide Time: 00:44)



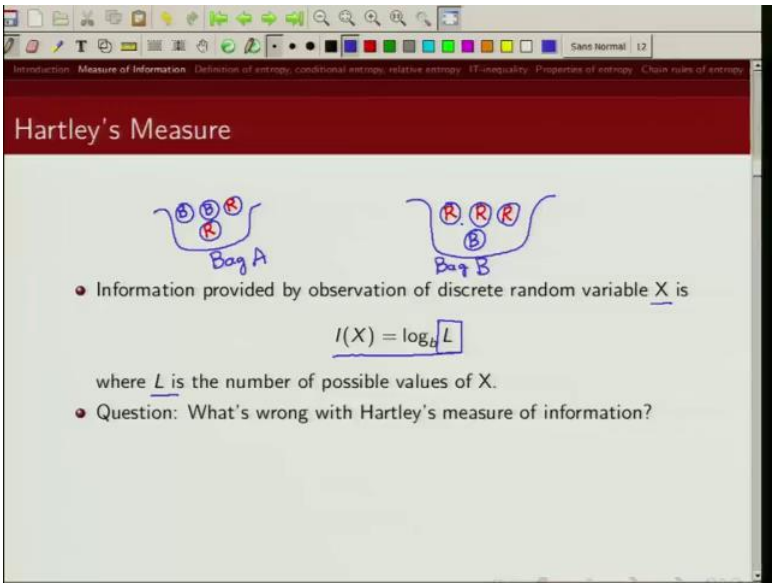
We will also talk about what is known as IT-inequality its and inequality which is used to prove lot of results. We will prove IT-inequalities, and then we will talk about some properties of entropy such as chain rule and will define mutual information.

(Refer Slide Time: 01:06)



So, after this lecture, you should be able to quantify information and you should be able to know what is entropy, conditional entropy, joint entropy, mutual information and their properties.

(Refer Slide Time: 01:25)



So, Hartley was the first person who was able to quantify information and how did he quantify information. So, Hartley said that, so if you have a random variable x and let us say there are X takes L different values then Hartley define measure of information as log of L the unit b determines the units of information. If b is 2, we define information in

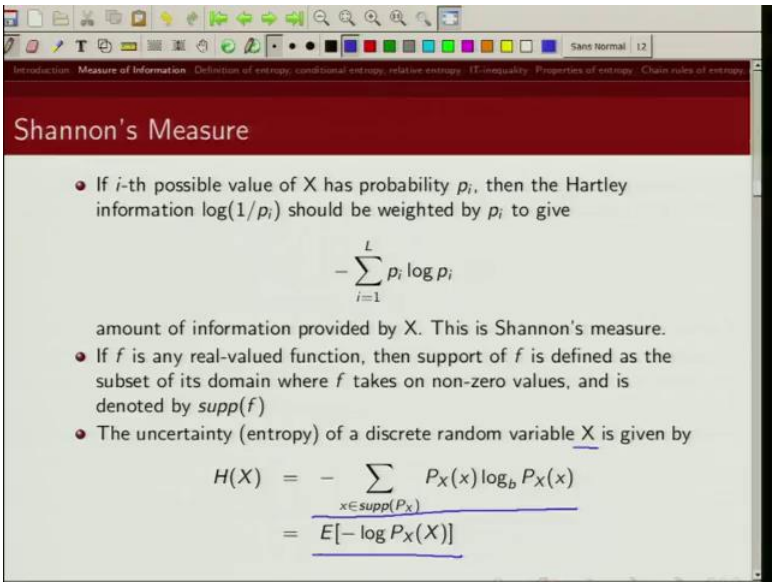
terms of bits if it is Maxwell law we define it in terms of hertz. Now, note here Hartley was able to recognize that to have information basically random variable we should have multiple possible values. So, something which is splittable something which has known values does not convey any information.

So, let us consider tossing of coin. So, if you are considering unbiased coin then with equal probability, we will get head and tail. So, according to ho Hartley then the information content here is log of 2. Now, there is a problem with Hartley's measure of information and what is that. To illustrate, let us consider a bag containing two types of balls - black balls and red balls. So, let us consider let us say there are two bags just call this it as bag A and bag B; each bag, let us say have 4 balls. So, there are 4 balls in each bag. Now, in bag A let us say 2 of the balls are of blue colour and 2 of the balls are of red colour whereas, in bag B lets say there are 3 balls of red colour and 1 ball of blue colour.

Now according to Hartley's measure of information in bag A, there are two possible outcomes. So, what I am doing is I have a bag of these balls and I am picking up ball and I am showing you the colour of the ball. So, according to Hartley in bag A, there are two possibilities, you can either get blue ball or you can get red ball. Similarly, in bag B, you have two possibilities, you can either get blue ball or you can get red ball. So, according to Hartley's measure of information, the information content in both these bags are same, because both have only two possibilities, but is this information contains same.

Look at this; in bag A, there is a 50 percent possibility of getting a blue ball and 50 percent possibility of getting a red ball, whereas, in bag B, there is 75 percent possibility of getting a red ball, and 25 percent possibility of getting a blue ball. So, in bag B if I pick up a ball it is more likely to be a red ball than a blue ball. So, clearly bag B has less information because the most likely balls that I can guess coming out of bag B is red ball because there are more red balls than blue balls whereas, in bag A, blue balls and red balls are equally likely to happen. So, clearly the event of picking a ball in bag from bag A has more information, because it is difficult to predict the colour of the ball, because blue balls and red balls are equally likely whereas, in bag B red balls are more likely. So, this feature of information was not captured by Hartley's measure of information. So, we can see that Hartley missed this crucial piece of information which is how frequently this event is basically how frequently these events are happening, so that was taken care of in Shannon's measure of information.

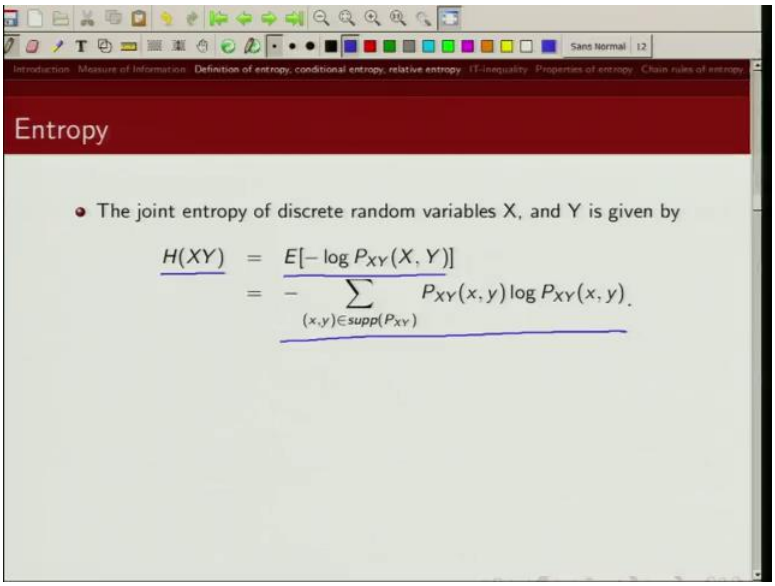
(Refer Slide Time: 06:16)



If i-th possible value of X standard variable X of probability p i then if we weight the Hartley's measure which in this case would be log of 1 by p i if he weighed by p i and we sum it over all possible value for x we get what is known as Shannon's measure of information. So, you can see here, the Hartley's information in some sense is weighted by the probability of occurring of that particular value of x. So, let us define what we mean by a support of a function. So, if we have a real value function f then its support is defined as subset of its domain where the function f takes non-zero value because this probability basically is defined over where P I is greater than 0.

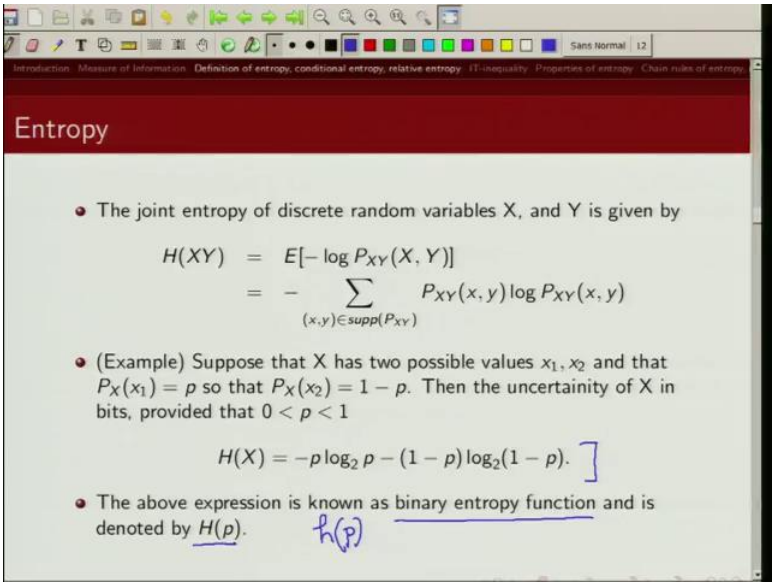
So, we define a support of a function f as a subset of its domain where f takes non-zero values. So, the information content of the uncertainty associated with the random valuable discrete random variable x can be written as follows. So, it is the expected value of minus log of P i P x. So, we define this as minus P x log P x sum over the support set of P x. So, this is because log of P x is defined over those values of P x which is greater than 0. So, we can define the entropy or the uncertainty associated with random variable as expected value of minus log of P x.

(Refer Slide Time: 08:41)



Now, similarly, we can define joint entropy of discrete random variable of X and Y. So, joint entropy of discrete random variable X and Y is defined as minus expectation of minus log of joint probability distribution of X and Y. So, this can be written as minus summation over the support set of joint distribution of P x y log P x y.

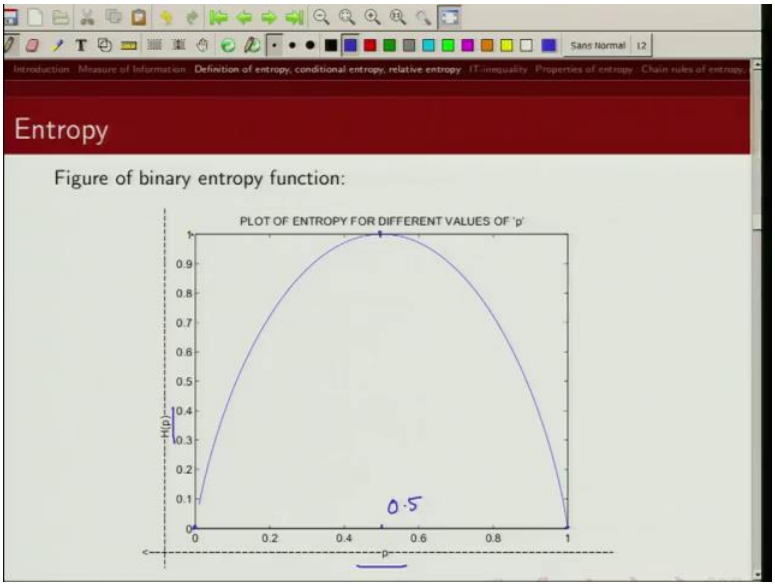
(Refer Slide Time: 09:21)



Now, let us take an example. Let us assume x takes 2 possible value x 1 and x 2 and probability of x 1 is p, so the probability of x 2 is 1 minus p. So, what is the uncertainty or what is the entropy of x according to definition, it is summation minus probability of x

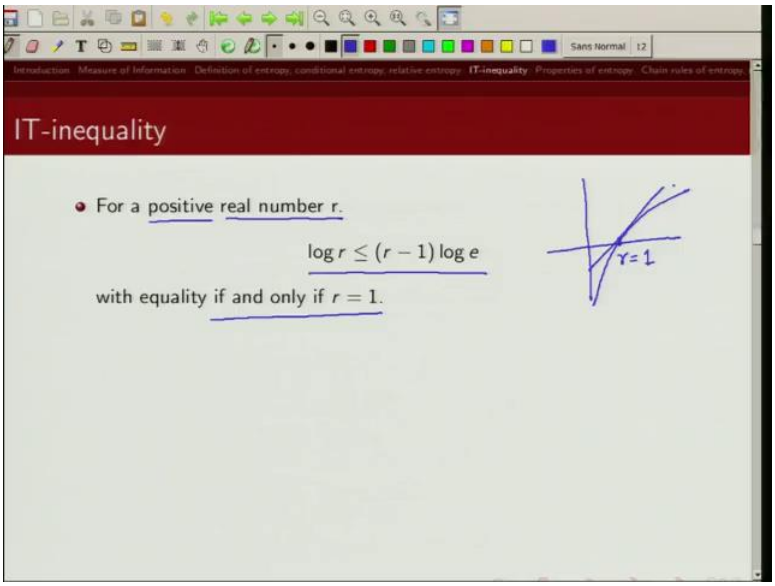
1 that is minus p log p and minus probability of x 2 which is minus log P log of 1 minus p. Now, this expression has a special name it is called binary entropy function. So, this is known as binary entropy function, and it is denoted by H of P; in many books they use this small notation this small h of p. I am just avoiding this small notation to avoid confusion with differential entropy, which we will define when we will talk about continuous random variable. So, I am defined I am denoting this binary entropy function by capital H of p.

(Refer Slide Time: 10:43)



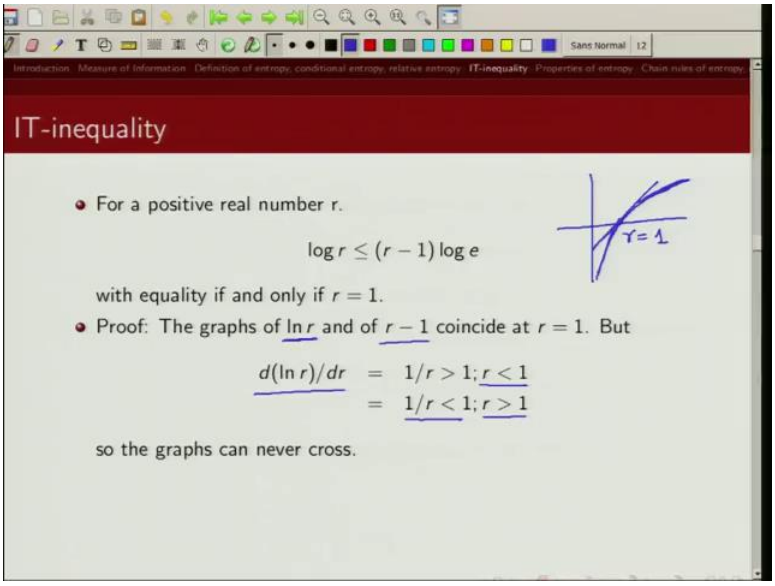
And we plot the entropy binary function its looks like this. So, on the x-axis, you have this probability; and on the y-axis, I have plotted this binary entropy function you can see that P equal to 0 and close to it this is basically here it is 0 and its maximum value is p 0.5 and this corresponds to H of p b 1.

(Refer Slide Time: 11:15)



Now, let us prove an inequality, which will be very handy in proving lot of results related to entropy. So, let us consider a positive real number r then we can show that log of r is less than equal to r minus 1 log of e and this equality happens if and only if r is equal to 1. If we plot this to see, my log of r is defined for r positive real number. So, this would be something like this at r equal to 1, this log of r is 0. And how will this function look like, it is basically this r minus 1 and this slope here would be log of e.

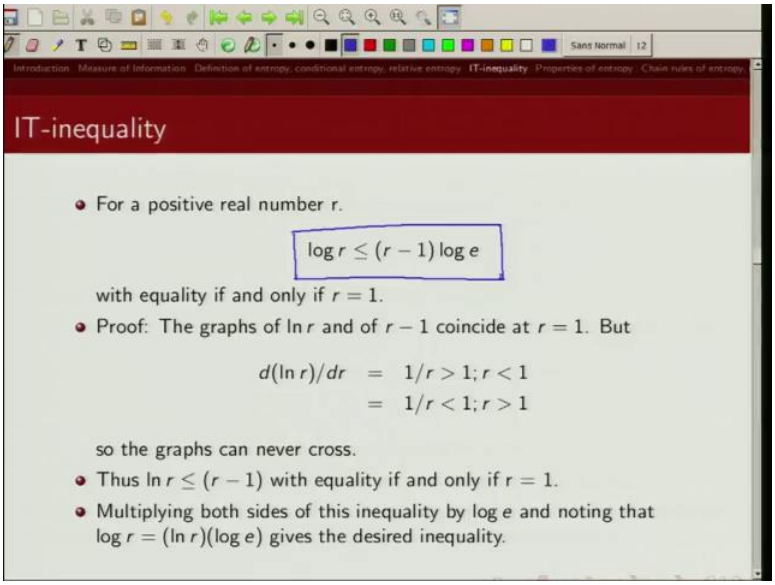
(Refer Slide Time: 12:21)



Now, note that the graph of natural log of r and r minus 1 they coincide at r equal to 1, I

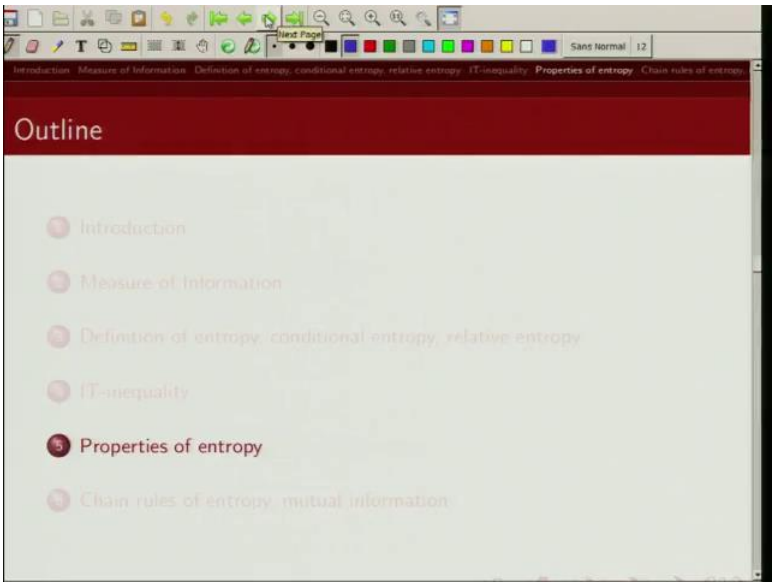
just plotted and this would be plotting again. So, this will be like my log function and this will be my r minus 1. So, they intersect at r equal to 1. And if I consider the slope of this natural log, you can see here this slope here for r less than 1, this slope is greater than 1 and for r greater than 1, and this slope is less than 1. So, clearly, these two graphs will not intersect each other; and as I said natural log of r is below r minus 1.

(Refer Slide Time: 13:19)



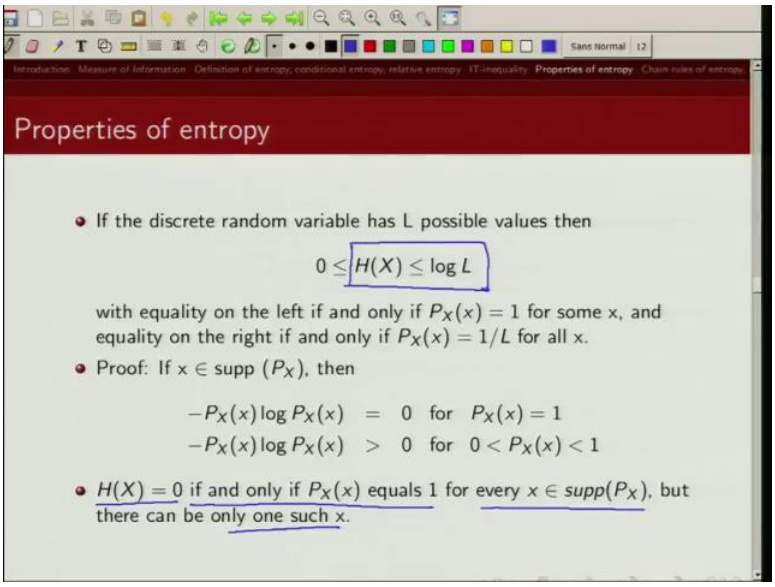
So, this condition holds with equality only if r is equal to 1; if I multiply those sides by log of e, I get this result which is log of r is less than equal to r minus 1 log of e, this is known as IT-inequality.

(Refer Slide Time: 13:47)



Now, let us prove some properties of entropy using this IT-inequality.

(Refer Slide Time: 13:56)

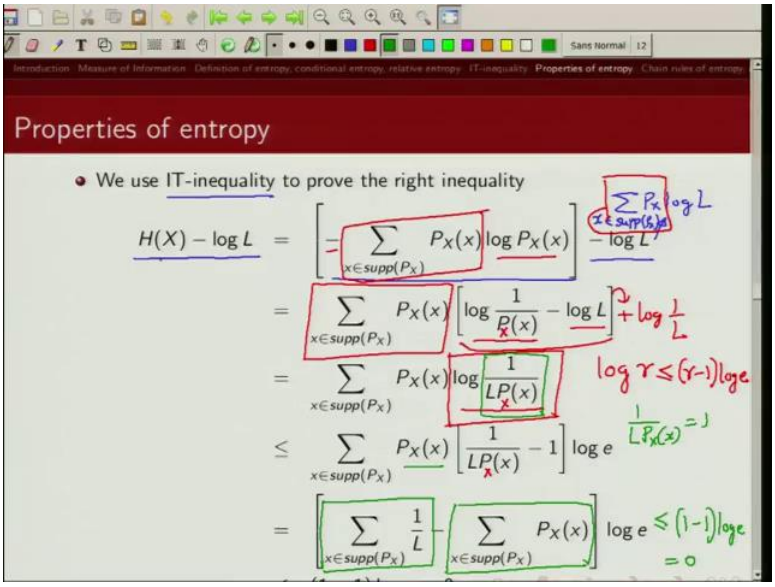


So, the first property of a discrete random variable x that takes L possible values is as follows. Entropy is lower bounded by 0 and upper bounded by log of L. The equality on the left hand side happens if and only if P of x is 1 for some particular x and for all other x it is 0; and equality on the right, which is this H of X is equal to log r this happens if your source is uniformly distributed. So, P of x is 1 by L for all x. So, let us prove this. So, we will first show this result that entropy is greater than equal to 0. So, how do we

prove it lets write we know from the definition of entropy it is minus P x log of P x summation over all values of x.

So, let us look at these quantities minus P x log of P x over those support set of P x. Now, clearly when P of x is 1 log of 1 will be 0. So, this quantity will be 0, and we know that probability lies between 0 and 1. So, when P of x lies between 0 and 1, this would be a fraction. So, log of a fraction will be minus also number and minus, minus becomes positive. So, since P of x is positive. So, minus log of a fraction will be a positive number. So, this will be greater than equal to 0. So, what we have shown here then is for if profile value of x P of x is 1 and for other values P of x is 0 then the entropy will be 0; otherwise, entropy will be greater than equal to 0. So, entropy will be 0 if and only if P x is equal to 1 for every x belongs to support set of P x but there can be only one such x for all other values of x P of x should be 0.

(Refer Slide Time: 17:00)



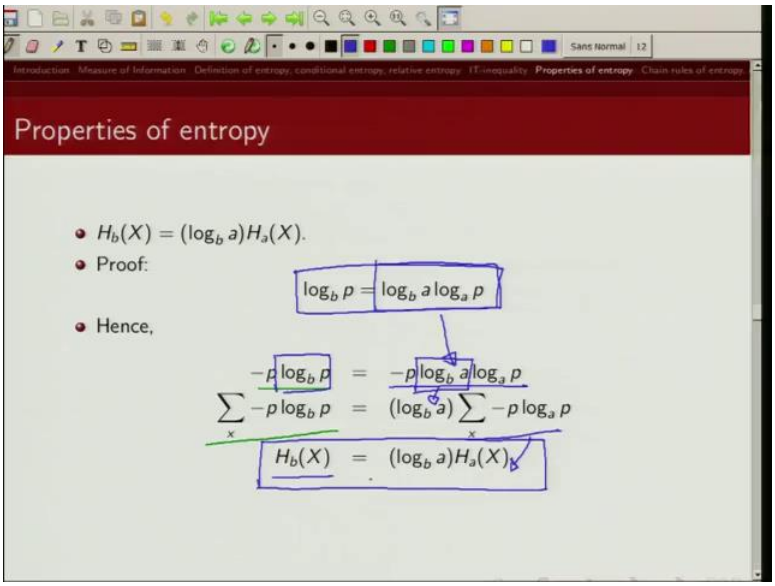
Now, let us prove the other part, which is H X is less than equal to log of L. So, how do we prove this? So, we have to show that H of X is less than equal to log of L. In other words, we can show if we can show that this quantity H of X minus log of L is less than equal to 0, then we would have shown that H of X is greater than H if H of X is less than equal to log of L. And to prove this results, we are going to make use of IT-inequality which we just proved that log of r is less than equal to r minus 1 whole multiplied by log of e.

So, let us look at the proof. So, first thing I did was from definition of entropy, I wrote down the expression for entropies. So, this is minus summation over support set of P x P log of P x; and the second term is minus log of L, I write it like this. Now minus log of L is can be written as so I can write this as summation over P of x log of L the summation of P of x to log of two. So, I can write this term like this. So, note I have a common term and of course, x is defined over the support set of P x. If I write log of L in this particular fashion, you can see in this term, this particular term and this particular term is common. So, I take this term out, now I have minus log of P of x or minus log of P of x that I can write as log of 1 by P of x. So, this is how I am writing this and I have minus log of L that is this term. Now, minus log of L can be written as plus this term can be written as plus log of 1 plus L. And if I have log a plus log b that is log a b, so I can combine these two terms and write it like this log of 1 by L times P of x.

Now, this quantity can be simplified using IT- inequality now what does IT-inequality says IT-inequality says that log of r is less than equal to r minus 1 log of E with equality happening when r is equal to one. So, what is r were, this is my r here. So, I can then write log of r as r minus 1 times log of e. So, this can be written as r minus 1 log of e. So, then simplifying it; I can write the first term is summation over P X 1 by L P x, so that is this term. And then the second term is minus summation over support set of P x P f i. So, this is less than equal to this term will be 1 minus 1 log of E which is basically 0. So, then what have shown here is H of X minus log of L is less than equal to 0. So, in other words H of X is less than equal to log of l.

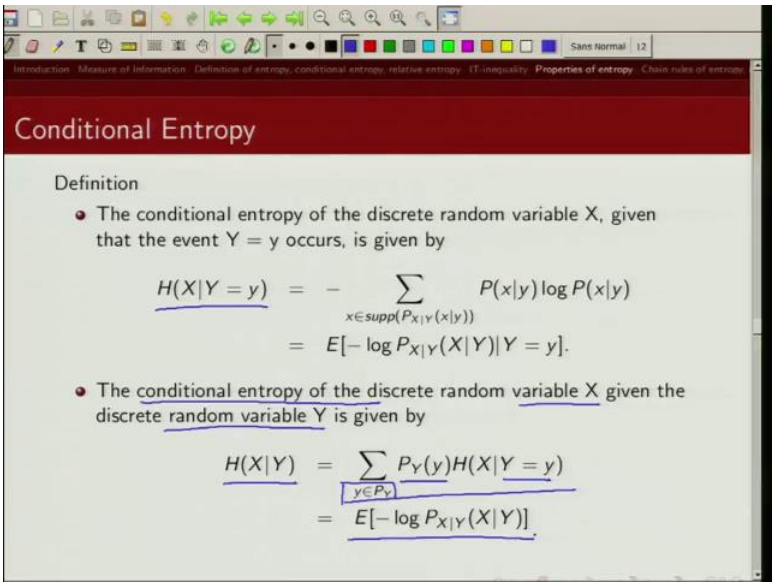
So, then if x is a discrete random variable that can take L different values then the maximum entropy or the maximum uncertainty associated with respect to x is equal to log of L and this happens when r is 1. So, when is r 1 means 1 by L P x is basically going to 1, P of x could be 1 by L for all x belongs to the support set of P. In other words, P x should be uniformly distributed. So, when my discrete random variable is uniformly distributed, I will have H of X less than H of X equal to log of L and this is the maximum uncertainty associated with x. So, you can go back and look at our example of tossing of a coin. So, when it is an unbiased coin, probability of occurrence of head and tail is same half in that case we have the maximum uncertainty, we have the maximum information.

(Refer Slide Time: 23:00)



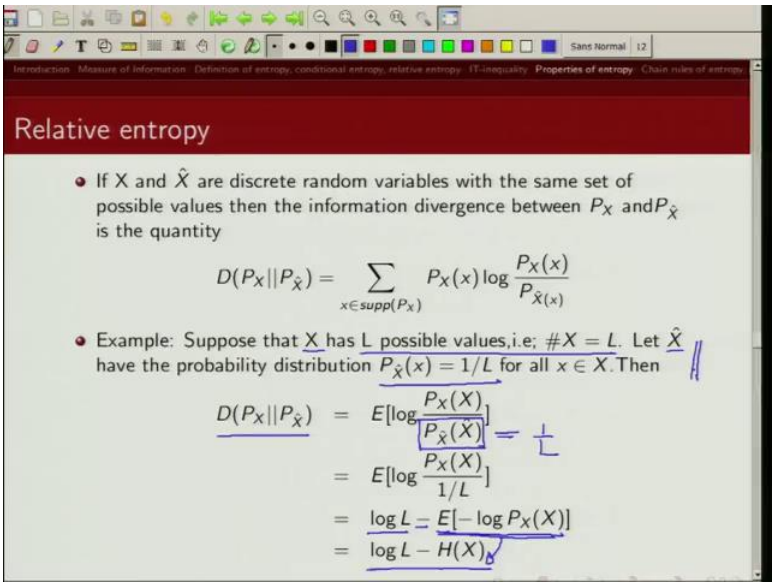
Now, let us prove some properties of entropy. So, entropy computed to the base b can be related to entropy computed to the base a by this particular relation. This is straight forward to prove you can write log of p to the base b of log of a to the base b multiplied by log of p to the base a. Now, what is the definition of entropy? So, we have to compute entropy to the base b. So, we have to compute minus p log of p and we have to sum it over the support set of p. Now, this we know that this quantity is equal to this. So, we replace this by this in this expression. So, we get this expression. Now we are summing over the support set of sets. Now, please note this quantity log of a to the base b does not depend on x. So, I can take this out. So, what I get here is minus p log to the base log of p to the base a. So, this quantity will be entropy computed to the base a and this is computed entropy of x computed to the base b. So, I can write then, so this is the relation governing entropy computed to base b and a.

(Refer Slide Time: 24:46)



Now, let us define what we mean by conditional entropy. So, a conditional entropy of a discrete random variable x given an event Y equal to some y which has occurred given by as follows. So, the conditional entropy of X given an event Y equal to y has happened is given by this expression, which is minus summation over the support set of this conditional probable distribution of x given y this P of x given by log of P of x given y. So, this is basically expectation of minus log of P of x given y equal to y a particular event has occurred. Now, if I have to compute conditional entropy of a discrete random variable x given another discrete random variable y then this can be computed from this conditional entropy, which we just defined now as follows. So, conditional entropy of x given y can be given as follows. It is conditional entropy of x given a particular event y has occurred multiplied by the probability of occurrence of that particular event and we sum it over all y's. So, this can be written as expectation of minus conditional distribution of X given Y.

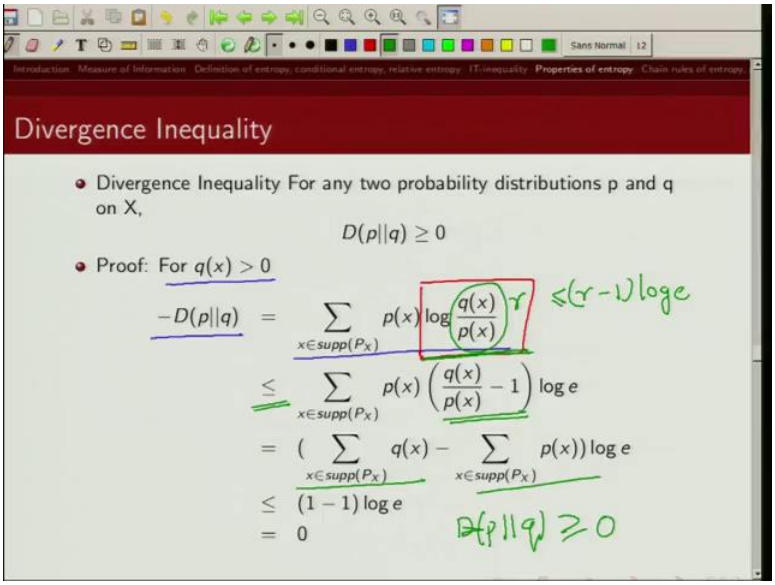
(Refer Slide Time: 26:45)



Now, we define what we mean by relative entropy or divergence. So, if x and x hat are 2 different discrete random variables with same set of possible values then the information divergence or relative entropy between P x and P of x hat is defined as follows. So, you can look at this is expectation with respect to x of log of P of x divided by P of x hat. So, let us take a simple example to compute relative entropy. So, we have a random variable x that takes L possible values we have another random variable x hat which is uniformly distributed. So, P of x hat is 1 by L for all x belonging to x.

Now how do we compute the relative entropy between P of x and P of x hat that is expected value of log of P x divided by P of x hat. Now, what is P of x hat this is uniformly distributed. So, this is basically equal to 1 by L. So, if we plug that in here, we get expected value of log of L times P of x. So, log of L does not depend on x. So, I can take it out. So, log of L and then what I am left with is expected value of log of P x. This I can write as minus of expected values of minus of log of P of x. And what is this quantity, this is our entropy. So, then I can write down the divergence between P of x and P of x hat is equal to log of L minus H of X for this particular random variable X and X hat.

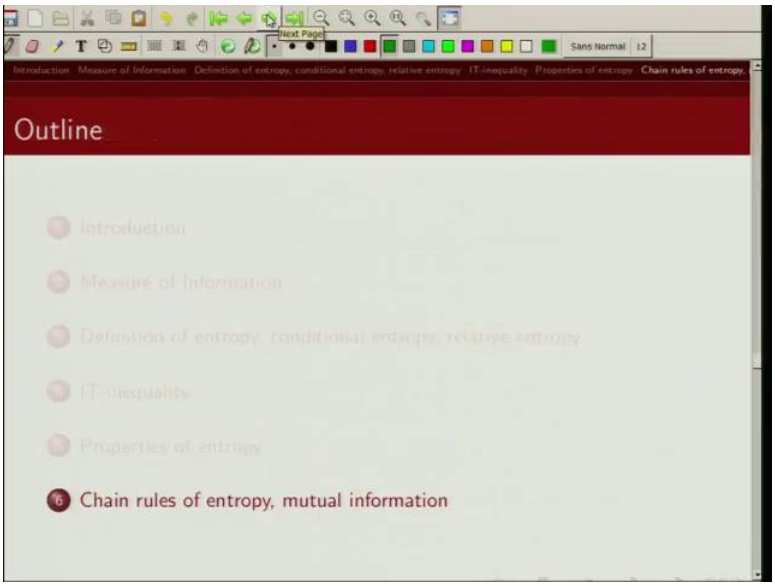
(Refer Slide Time: 29:06)



Now, divergence between any two probability distribution p and q is always greater than equal to 0. So, divergence to its (Refer Time: 29:22) measure of closeness between two distribution, if p is very close to q the value of divergence will be close to 0; otherwise, divergence will be if P hat is substantially different from q divergence will be large. So, for q x greater than 0 let us compute minus of divergence of p and q. So, I will have to show that this quantity is less than equal to 0. So, from definition of divergence I know divergence between p and q is expected value of log of P by q. So, minus of that would be expected value of log of q by p. So, this is minus of divergence between p and q. Now I will again make use of I t inequality. So, this is my r and log of r now log of r is less than equal to r minus 1 times log of e. So, I will make use of IT-inequality.

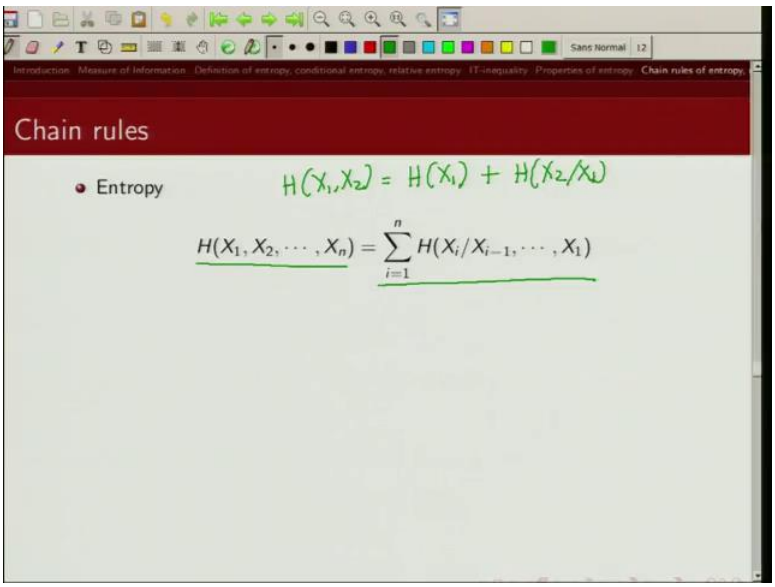
So, then this particular term is less than equal to r minus 1 log of E. Now simplifying, so I will get this is p x multiplied by q x by p x. So, I will get this term summation over 1 x and here I will get summation over p x this is less than equal to 1 minus 1 and which is basically zero. So, what I have shown is minus of divergence between this two probability distribution p and q is less than equal to 0 or in other words divergence between p and q is greater than equal to 0.

(Refer Slide Time: 31:38)



Now, we have defined so far entropy, joint entropy, conditional entropy and relative entropy. And we have proved some properties of entropy like if x is a discrete random variable we have shown that minimum value of entropy is 0 and maximum value is log of L, where x takes L possible values. Now, let us prove some more properties entropy and define what we mean by mutual information.

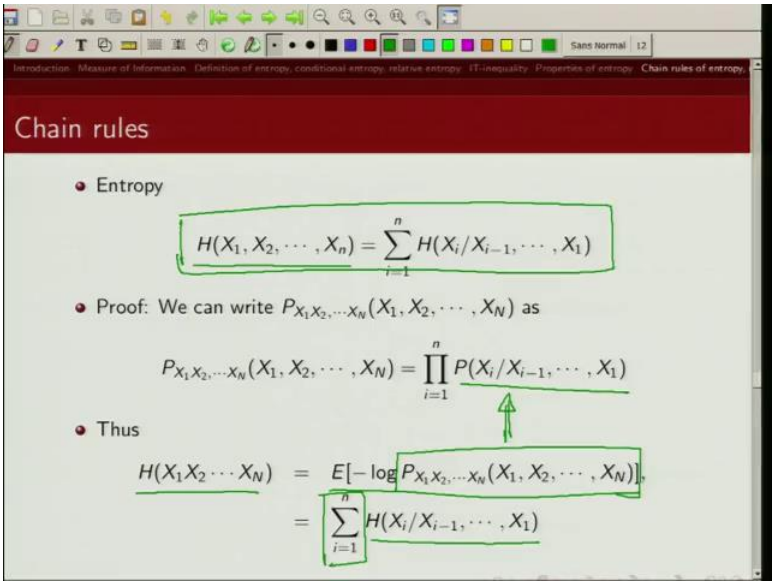
(Refer Slide Time: 32:13)



So, joint entropy of random variable X 1, X 2, X 3, X n can be written in terms of their conditional entropy in this particular fashion. For example if you have H of X 1, X 2 this

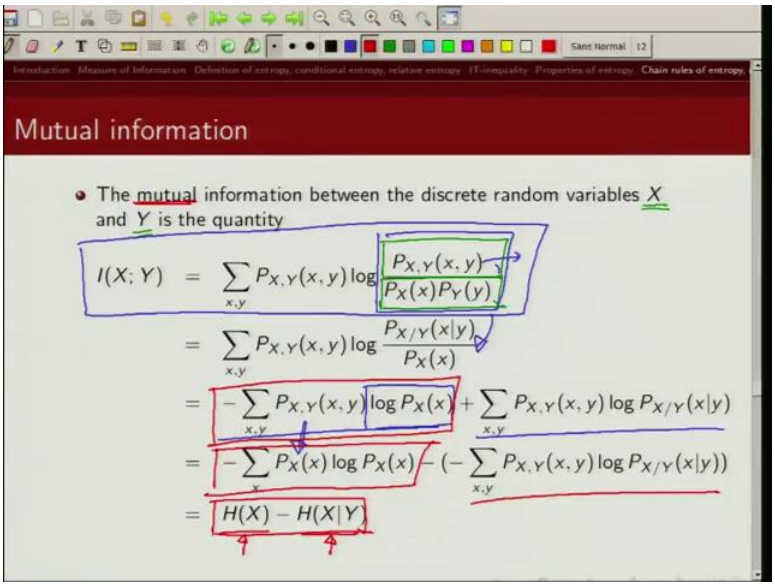
can be written as H of X 1 plus H of X 2 given X 1. So, we can write this joint entropy in terms of summation of conditional entropy.

(Refer Slide Time: 32:46)



So, how do we prove it, we will write this joint distribution in terms of conditional distribution. And now we will invoke the definition of joint entropy. What is the definition of joint entropy; it is an expected value of minus log of this joint distribution. And what is this; this is nothing this is given by this quantity product of these conditional probabilities. So, we take log of product terms, what we will get is summation here and what we will get then is expected value of minus log of these conditional distribution which is nothing but conditional entropy. Hence, we can write this joint entropy in terms of this conditional entropy. This is very important result, we are going to use chain rule repeatedly to prove different properties of entropy and other things.

(Refer Slide Time: 34:07)

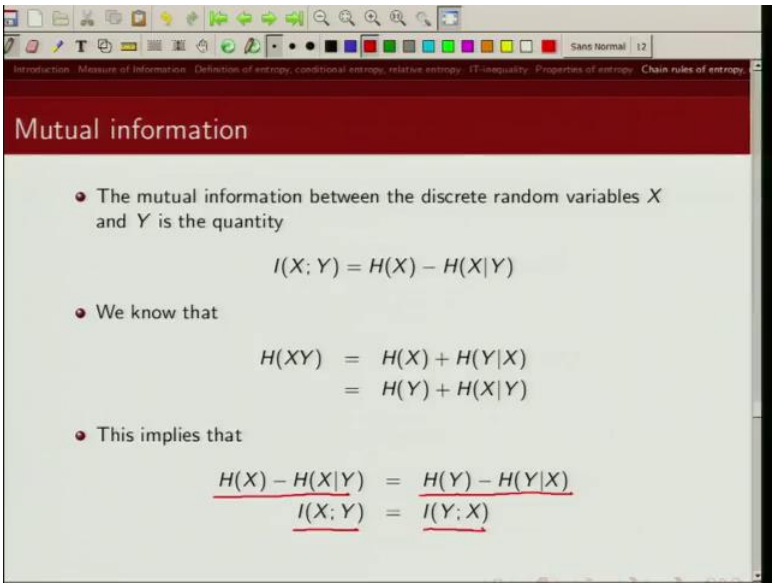


Now, let us define what we mean by mutual information mutual information between two random variables X and Y. So, mutual information between two random variables X and Y is defined as divergence between the joint distribution and these marginal. So, it is defined as expected value of log of joint distribution of X and Y divided by marginal distribution P of x and P of y. So, this is the definition this is how mutual information is defined. Now, this we can write as so we can write this joint distribution in terms of conditional distribution we can write this as probability of x given y into probability of y. So, then this term this particular term can be written like this. Now, log of a by b this can be written as log of a minus log of b. So, then we can write this as so summation P of P x y log of 1 by P x that is this term and then summation over P x y log of x given y that is this term.

Now, what is this term, summation now does this term depend on y, no. So, if you sum it over y what will we get we will get P of x so that is what we will get. So, this summation over x y is sum over y this does not depend on y. So, if I sum it over y, what I will get is P of x. So, this particular term that I have here can be simplified to this and plus this term can be written as minus, minus of this term. Now, what is this, this is nothing but our entropy of x. And what is this term; this is the conditional entropy of x given y. So, what is mutual information this, this is the uncertainty associated with source x, this is the uncertainty in source x given y, then what is the difference telling me. So, this is telling me the information that y is coming about x.

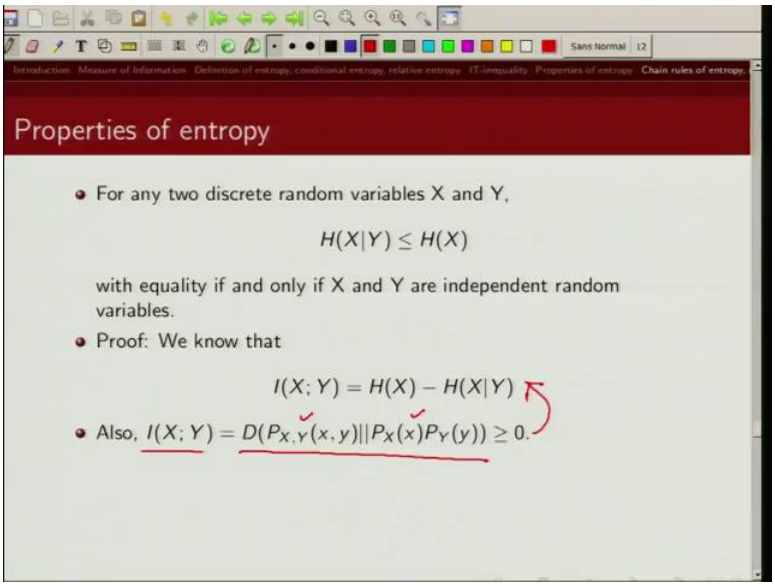
Now, y is this mutual term coming here. We will show that this mutual information can also be written as uncertainty in y minus uncertainty of y given x. So, whatever information y is giving about x the same information is also provided by x about y.

(Refer Slide Time: 37:27)



So, let us do that. So, we have just shown that mutual information can be written like this. Now, we know that we can write this joint entropy using chain rule we can write it in this particular fashion. This is H of X plus H of Y given X. Now, I again apply chain rule in a different fashion, so I can write this joint entropy as H of Y plus H of X given Y. So, if I compare these two equations, what I get is as follows H of X minus H of X given Y is same as H of Y minus H of Y given X. And what is this, this is nothing but mutual information between X and Y and this term is mutual information between Y and X. So, you can see what all information Y is giving about expressing information conveyed by X about Y. So, this is our mutual information.

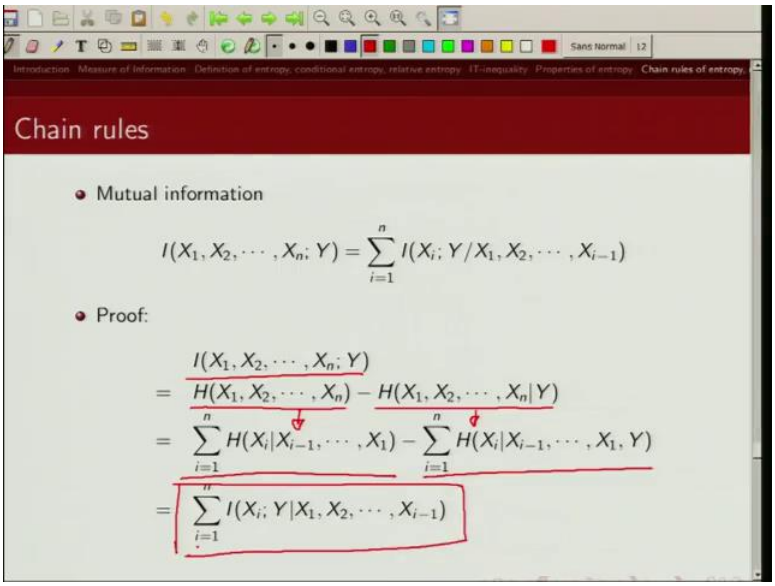
(Refer Slide Time: 38:38)



Now, let us prove some more properties of entropy function. So, if you have two discrete random variables X and Y then conditioning cannot increase entropy. So, entropy of X given Y is always less than equal to entropy of X, and this equality happens only if X and Y are independent. In that case, Y does not provide any information or reduction in uncertainty of X. So, how do we prove this result, there are number of ways you can prove it.

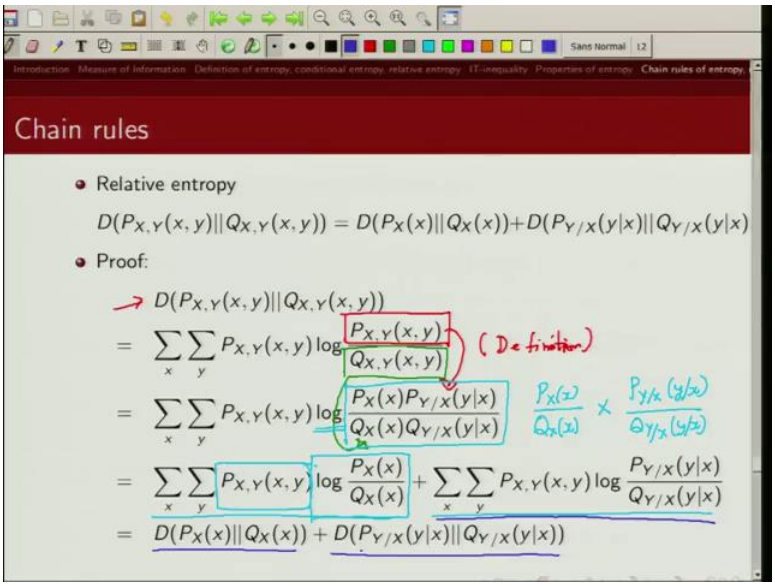
Now, mutual information can be written like this; it is uncertainty in x minus uncertainty in X given Y. Now, we also know that mutual information is nothing, but this relative entropy or divergence between this joint distribution of X and Y and product of this marginal's P of x P of y. And we have shown that divergence is all between two distributions P and Q it is always greater than equal to zero. So, using those results then we have we can say that mutual information because mutual information is divergence between these two distributions. So, mutual information is also greater than equal to zero. Now, we plug that in here, we have proved that H of X is greater than equal to H of X given Y.

(Refer Slide Time: 40:29)



Now, similar to the chain rule for entropy, we can also define chain rule for mutual information. So, mutual summation will X 1, X 2, X 3, X n and Y can be written in terms of conditional mutual information of X i and Y given X 1, X 2, X i minus 1. So, proof follows like this. So, first we write the definition of mutual information in terms of entropy and conditional entropy. So, this mutual information given X 1, X 2, X n and Y can be written as joint entropy of H minus joint entropy of this X 1, X 2, X n given Y. Now, each of this quantity can be written in terms of conditional entropy using chain rule. So, this joint entropy of X 1, X 2, X 3, X n can be written like this. Similarly, this joint entropy given Y can also be written in terms of using chain rule in terms of its conditional entropy. So, if we combine this, we can see that this is nothing, but mutual information between X i and Y given X 1, X 2, X 3, X i minus 1.

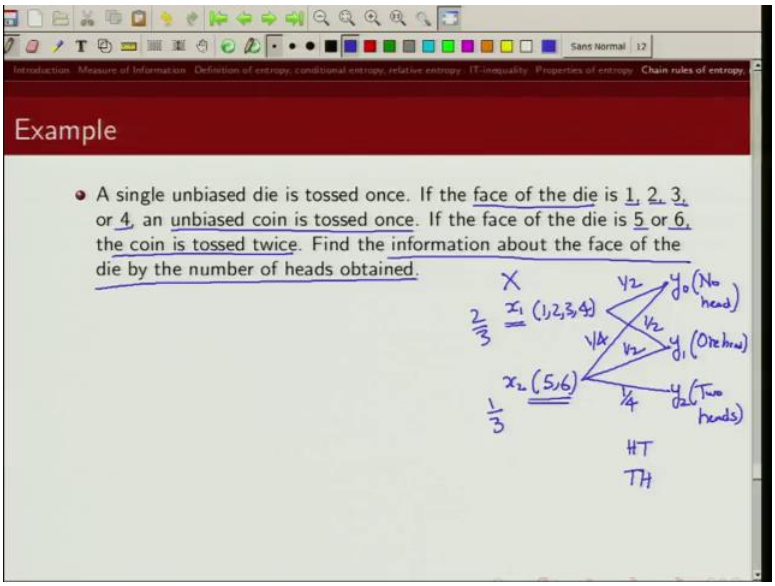
(Refer Slide Time: 42:07)



We can also similarly show chain rule for divergence. So, if we have divergence between joint distribution of x and y P of x of y and Q of x y, this can be written as divergence between P x and Q x plus divergence of conditional distribution of y given x and Q of y given x. Again the proof is straight forward, what we do first is from definition of divergence, we write the expression for divergence between these two probability distribution P of x y and Q of x y. So, this follows from definition. So, from definition, we get this. Now, these joint distributions can be written in terms of conditional distribution. So, I can write this like this and similarly I can write this in this particular fashion.

Now, I will separate these two terms. So, I can write this, this particular term, this particular term, I can write this as P of x Q of x multiplied by P of y given x and Q of y given x and there is a log here. So, it is log of a into b. So, this will be log of a plus log of b. So, then this is my log of a term and this is my log of b term. Now, again this particular quantity does not depend on y. So, when I sum it over Y, I am essentially summing up this over Y. So, if I do that what I will get is P of x. So, this is summation of summation over x P of x log of P of x by Q of x which is nothing but divergence between P of x and Q of x. And similarly, this particular term can be written as divergence between the conditional distribution of y of P or y given x of P and Q y given x. So, this is the chain rule for entropy function.

(Refer Slide Time: 45:03)



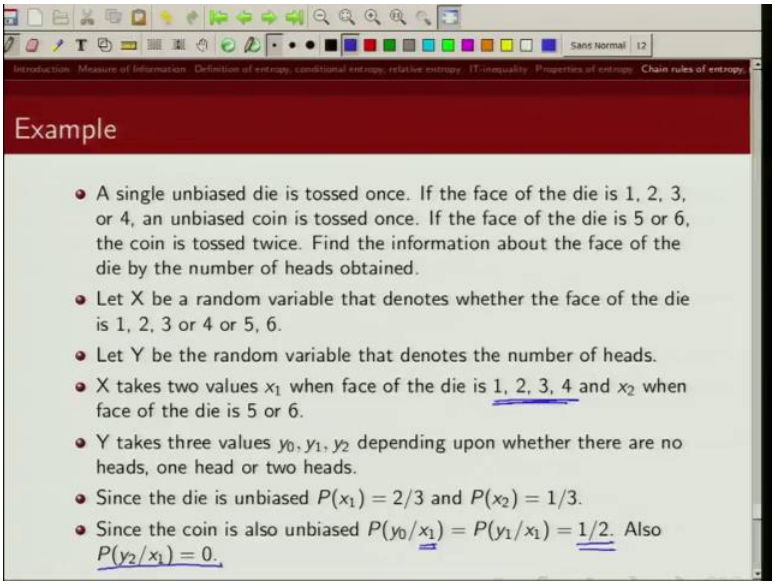
Now, we will just conclude with one example to illustrate how to compute mutual information. So, you have a single unbiased dice which is tossed. So, this is a dice. So, this has numbers from 1 to 6 and you are tossing this coin. Now if the faces of the dice is 1, 2, 3 or 4 what you are doing is you are tossing a unbiased coin. What is a unbiased coin, an unbiased coin is a coin which has head and tail. So, if the outcome of the dice is 1, 2, 3 or 4, I am tossing a coin. And if the outcome of the dice is 5 or 6 then I am tossing the coin twice. And this is the fair coin you can toss tossing unbiased coin. I want to find out what is the information about the face of the dice that is conveyed by number of heads obtained. So, again what I am doing is I am just rolling a dice and depending on what the number is I am tossing a coin. So, you can think of it, basically for input x is the outcome of the dice, so there are two possible outcome x 1 and x 2. So, if the dice outcome is 1, 2, 3 and 4 that is one of possible outcome and if the dice outcome is 5 or 6 that is one possible this thing.

Now, what I do is if I get 1, 2 or 3, I toss a coin once. If I toss a coin once, what can I get, I can get a case when there is no head. If I am tossing a coin once, I can get a condition when there is no head or I can get a condition where there is only one head correct. Now, what am I doing when I am getting 5 or 6, I am tossing the coin twice. So, what are the possible outcomes as far as head is concerned, I can get no head in none of the those two twice I got a head or I can get one head or I can get 2 heads. Now, since this dice, this a dice is a fair dice and this coin is fair dice, then this will happen with

probability 2 by 3, and this will happen with probability 1 by 3. So, given that I get X 1 probability of getting y 0, which is no head is half because my coin is a fair coin, and this probability of getting 1 head is also half.

Now, if I get X 2 which is the face of the dice is 5 or 6, then I am tossing the coin twice. So, what is the probability of getting no head that is half into half? So, this will be 1 by 4. So, what is the probability of getting two heads? So I have to get head in the first toss as well as head in the second toss, so that probability is half into half that is again 4. And what is the probability of getting one head, I can either get head in the first toss and tail in the second toss or I can get tail in the first toss and head in the second toss. So, probability of this and probability of this will add up to half.

(Refer Slide Time: 49:10)

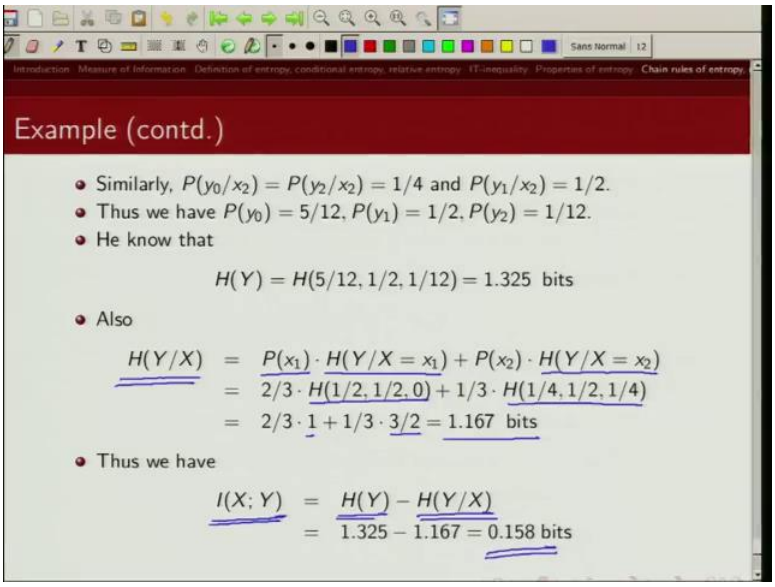


So, that is what I am I am written here. So, X is a random variable that denotes the outcome of the throwing of the dice. So, this can be either 1, 2, 3, 4 and one set of actions I am taking based on whether the outcome is 1, 2, 3, 4 and I have other set of actions which I am taking depending on whether the outcome is 5 or 6. And let Y is the output of my tossing of the coin which basically I am interested in counting number of heads. Now, as I said that there are two possible actions I am taking based on what I get as a result of throwing of the dice. So, I am calling x 1 if I get face of the dice as 1, 2, 3, 4, and I have x 2 and if I get face of the dice as 5 or 6.

Similarly, in outcomes, I am writing it as y 0, y 1, y 2 denoting no head, one head, and

two head. I have already mentioned because probability of getting x 1 which is probability of getting face of 1, 2, 3, 4 that probability is 2 by 3 because this is a fair dice; and probability of getting 5 or 6 is 1 by 3. Similarly, if I get x 1, which is dice value face value of 1, 2, 3, 4, I am tossing the coin once. So, in that case probability of getting no head is same as probability of getting one head which is half and I can never get two heads because I am tossing the coin only once.

(Refer Slide Time: 51:04)



Similarly, if I am getting 5 or 6, I am tossing the coin twice. So, probability of getting no head or probability of getting two heads is 1 by 4, this we have just shown and probability of getting one head is half. Now, that we know what is the probability of getting x 1, what is the probability of getting y i given x i, we can find out what is the probability of getting y 0, y 1 and y 2. What is the probability of getting y 2, there is a probability of getting x 1 into probability of y 0 given x 1 and plus probability of y 0 given x 2 into probability of x 2. Similarly, we can compute this probability of P of y 1 and P of y 2.

In this particular example, this turns out to be P of y 0 turns out to be 5 by 12, P of y 1 turns out to be half and P of y 2 turns out to be 1 by 12. So, then uncertainty in y can be given as minus 5 by 12 log of 5 by 12 minus half log of half minus 1 by 12 log of 1 by 12 and that comes out to be 1.325 bits. Similarly, we can compute conditional entropy of y given x this is given as conditional entropy of y given x 1 into probability of x 1 plus

conditional entropy of Y given x 2 multiplied by probability of x 2.

So, in the similar fashion, we can find out what is a entropy of y given x 1. This is basically minus half minus half log of half and uncertainty in Y given x 2 is minus 1 by 4 log of 1 by 4 plus minus half log of half and minus 1 by 4 log of 1 by 4, this is just solving the definition of entropy. So, if we do that this terms comes out to be 1 and this terms come out to be 3 by 2. So, computing this we find out that uncertainty in Y given X is given by 1.167. So, then the mutual information between X and Y or Y of X is basically given by uncertainty in Y minus uncertainty in Y given X, and this is comes out to be 0.158 bits. So, this is one example to illustrate how we can compute mutual information. So, with this, we conclude this lecture.

Thank you.