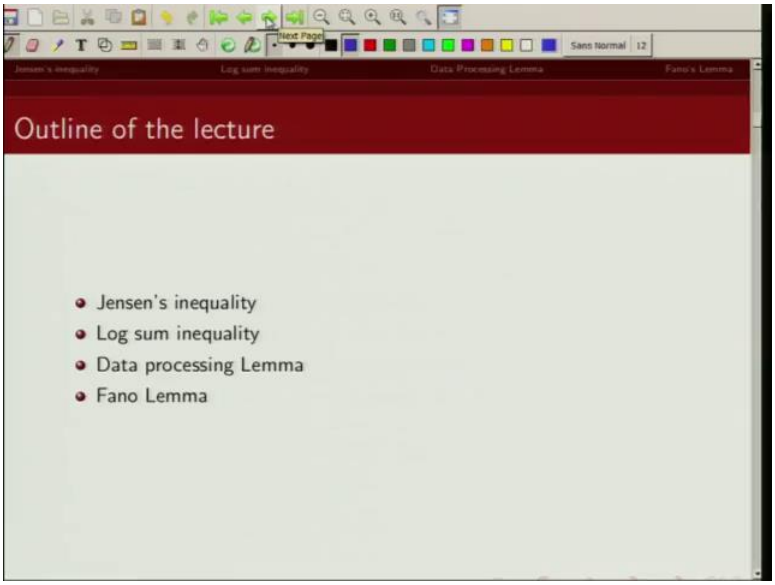
An Introduction to Information Theory
Prof. Adrish Banerjee
Department of Electronics and Communication Engineering
Indian Institute of Technology, Kanpur

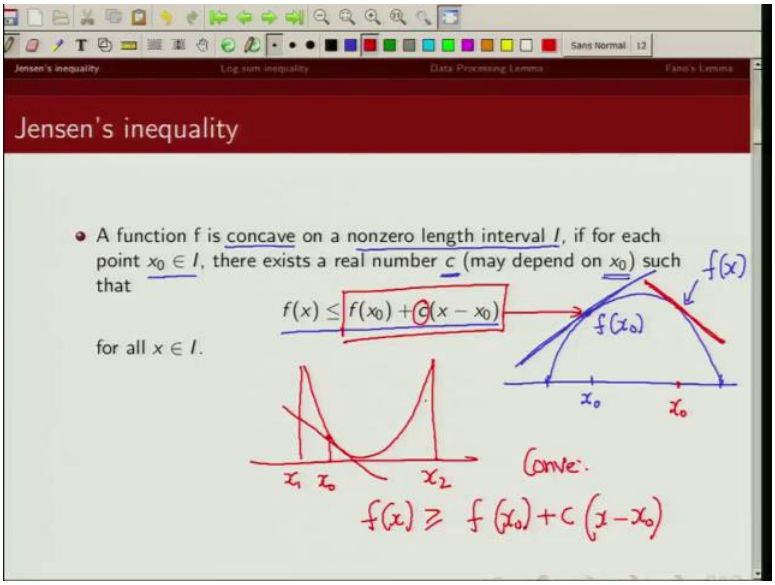
Lecture - 2A
Information Inequalities

(Refer Slide Time: 00:24)



Welcome to the course on An Introduction to Information Theory. I am Adrish Banerjee. And today we are going to talk about information inequalities. In particular, we will talk about what is Jensen's inequality and how this is used to find whether a function is concave or convex. We will talk about log sum inequality and we will show one example of how we can use log sum inequality to prove concavity or convexity of a function. Next, we will talk about data processing lemma and finally, we will talk about Fanos lemma. So, this lecture will we will talk about these inequalities and lemma.

(Refer Slide Time: 01:08)

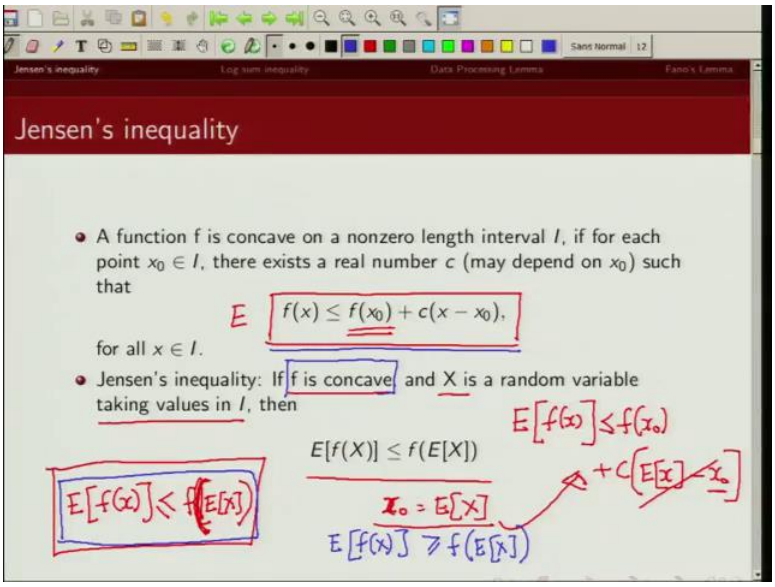


So, let us start with Jensen's inequality. So, before we go to Jensen's inequality, let us just see what do we mean by a function is concave or convex. So, function f is concave over a non zero length interval I, if for each point x naught which belongs to this interval i there exists a real number c such that this condition holds. Now, what is this condition? So, let us just draw, let us say this is a function defined over this interval from here to here. Now take any point belonging to this interval, let us just take we take this point x naught, and this is my function f of x. So, what is f of x 0 that is this point this is f of x 0. So there exists a real number c, of course the value of c may depend on will depend on x naught such that function f of x is below f x 0 plus c x minus x naught.

What is c x minus x naught, it is a line passing through x naught and passing through this point this will be basically a line, this straight line passing through x naught. Now, note that this c the slope of this line depends on what choice of x naught I choose. For example, if I choose this as my x naught then slope is negative, so this is what the line is. Since, the function is concave, I can draw a line passing through x naught and this function will lie below that so that is what I mean by function is concave. Now, if the function is convex then what is going to happen is you will have a function like this, let us say this is a function defined over this interval x 1 to x 2, and take any point, let us say take this point x naught. If I draw a line passing through this thing x naught then for the

function to be convex f of x should be above this line. This is the condition for function to be convex, and this is the condition for the function to be concave.

(Refer Slide Time: 04:04)

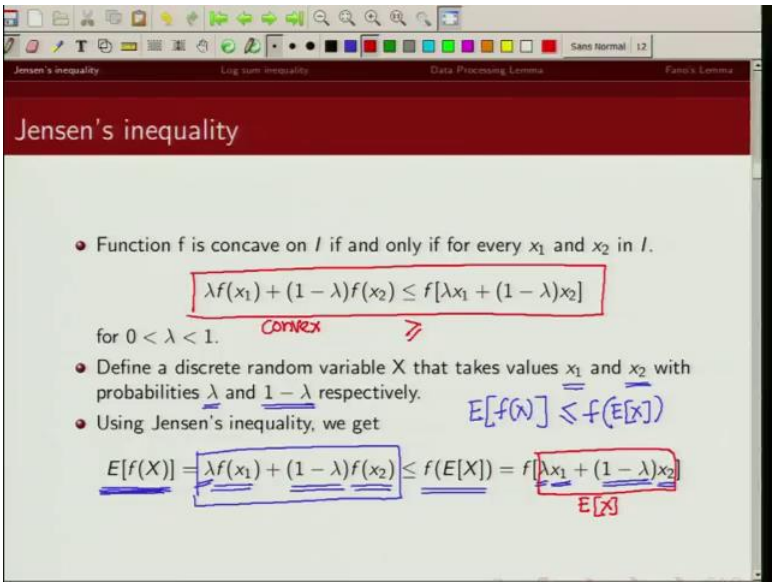


Now, we will use this result to obtain Jensen's inequality which says if the function is concave and we have a random variable x that takes values within the interval over which this function f of x is defined. We are with what the interval where the function is concave then expected value of the function is less than equal to the function evaluated at expected value of x. Now, let us see at the proof of this, we will get the proof from this result. So, if I take expectation both side what do I get f of f x this is less than expected value of f of x naught. Now, what is f of x naught, x naught is the number in the interval over which this function is concave. So, expected value of f of x naught will be f of x naught plus c times expected value of x minus x naught.

Now, if I choose my point x naught such that x naught is expected value of x, which is possible because x naught is belongs to the interval over which this function is concave. So, I can choose my x naught such that x naught is expected value of x. Now, if choose this value of x naught and I plug it in here what do I get I get expected value of f of x if less than equal to function. What is the x naught, x naught is expected value of x and c this is the expected value of x minus expected value of x. So, this term become 0, so

what I will be left with is this and this is precisely my Jensen's inequality. So if I have a function f, which are concave then this relation holds. Now, similarly if the function f is convex, we can similarly show that expected value of f of x will be greater that equal function evaluated at expected value of x. The proof is very similar to how we proved starting from this result, so this is our Jensen's inequality.

(Refer Slide Time: 07:24)



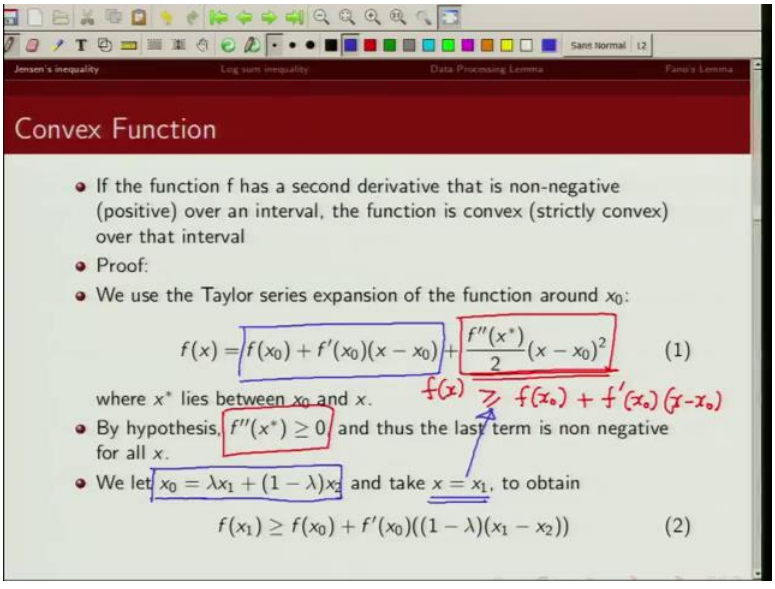
Now, we will show some more content going for proving whether a function is concave or convex. So, if you have a function f which is concave over an interval I then it is concave if and only if for every x 1 and x 2 which belongs to this interval I this relation holds. What is this, this lambda times f of x 1 plus 1 minus lambda times f of x 2 this should be less than equal to function evaluated at lambda time x 1 plus 1 minus lambda time x 2, where lambda lies between 0 and 1. Now, using Jensen's inequality, we can prove this result. So, let us look at the proof.

Let us say we have a random variable x and that takes two values x 1 and x 2 it takes x 1 with probability lambda, so x 2 happens with probability 1 minus lambda. Now, what does Jensen's inequality says that Jensen's inequality says if the function f is concave then expected value of the function is less than equal function evaluated at expected value. So, then let us compute expected value of the function. What is the expected value

of the function, expected value of the function is so x takes the value lambda 1 x 1 and x 2 with probability lambda n 1 minus lambda. So, the expected value of the function will be f of x 1 multiplied by probability of occurrence of x 1 which is lambda. And f of x 2 multiplied by probability of occurrence of x 2 which is 1 minus lambda. So, this is a value for expected value of the function f of x.

Now, Jensen's inequality says if the function is concave then function expected value of the function if less than equal to function evaluated at expected value of x. So, according to Jensen's inequality this term will be less than this. Now, what is expected value of x? So x 1 happens with probability lambda and x 2 happens with probability 1 minus lambda. So, the expected value of x is given by this expression, this is your expected value of x. So, then applying this Jensen's inequality, we get this result on the left hand side is the expected value of the function and on the right hand side is function evaluated at expected value. So, if the function is concave over this interval this relation holds; and vice versa if the function is convex then this relation will be greater than equal to OK.

(Refer Slide Time: 11:10)



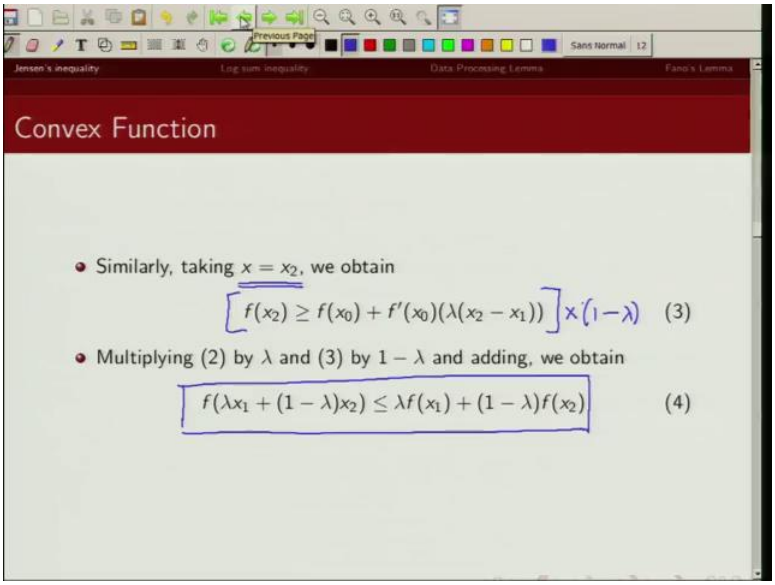
Now, we will show another way of proving whether a function is concave or convex. So, if we have a function f that has a second derivative and the second derivative is non-negative over this interval over which function is defined. When the function is concave

over that interval, so what do we need, we want the second derivative to be non-negative. Now, how can we prove this? So we do a Taylor expansion of this function around x naught, where x naught is point in this interval. So, Taylor's series expansion, I can write this f of x naught plus first derivative of f evaluated at x naught x minus x naught plus second derivative of this function f evaluated at some x star by 2 x by x naught, we are this x star point lies between x 0 and x.

Now, what am I seeing, I am seeing for function to be convex, I want this second derivative to be non-negative. So, what I need to show is if the second derivative is non-negative then this function f of x is convex. So, let us prove it. So, we are seeing that if second derivative is positive then function should be convex. So, if the second derivative is positive then this last term is going to be if this non-negative this is square of x minus x naught so that is also non-negative. So, this whole term will be non-negative term. So, then what I can do is I can write f of x to be so since this term is positive, so this is non-negative f of x will be greater than equal to f of x naught plus f first derivative of f evaluated at x naught x minus x naught.

So, this follows from the fact that since this term is non-negative this whole term will be non negative. So, f of x will be greater than equal to this term, so that is what I have here. Now, let us take the value of x naught to be this which is lambda times x 1 plus 1 minus lambda times x 2 and let us take x to be x 1. So, I plug in this value of x and x naught into this expression. So, I get here f of x 1 is greater than equal to f of x naught plus first derivative of this function f evaluated at x naught 1 minus lambda x minus x naught where x naught is given by f.

(Refer Slide Time: 15:03)

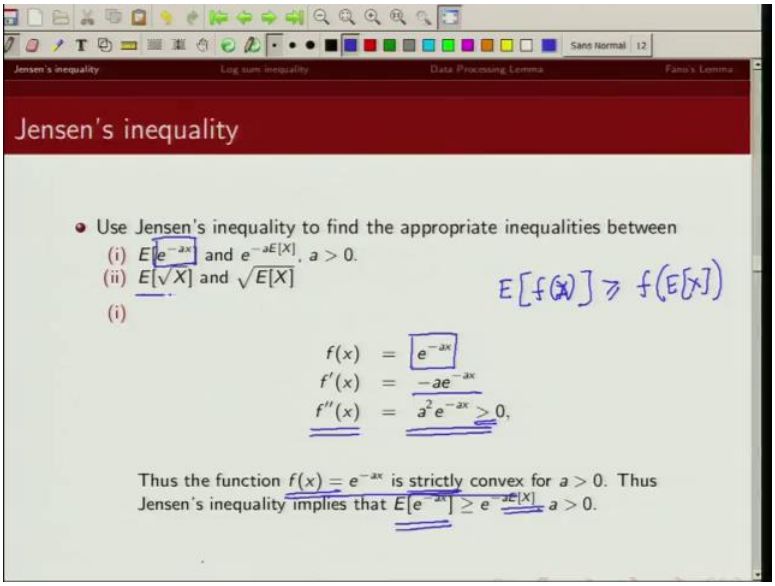


Next, so I put x equal to x 2 in this, this equation. Here, I put x equal to x 2. So, if similarly put x is equal x 2, what I get is f of x 2 is greater equal to f of x naught first derivative of f evaluated at x naught lambda x 2 minus x 1, where x naught is again lambda times x 1 plus 1 minus lambda time x 2. Now, what I am going to do is I am going to multiply this expression by 1 minus lambda and this expression I am going to multiply by lambda. So, equation 2, I am going to multiply by lambda; and equation 3, I am going to multiply by 1 minus lambda. And then I am going to add them up so I do that what I get here is I get f of x naught is less than equal to lambda times f of x 1 plus 1 minus lambda times f of x 2. So, again I repeat if I multiply equation two by lambda and equation three by 1 minus lambda and I add them up I get this condition that f of x naught is less than equal to lambda times f of x 1 plus 1 minus lambda time f of x 2.

Now, what is this, you go back to the previous expression that we have, we have said if the function is concave then lambda f of x 1 plus 1 minus lambda f of x 2 is less than equal to function evaluated at lambda x 2. And if the function is convex we said lambda f of x 1 plus 1 minus lambda f of x 2 would be greater than equal to function evaluated at lambda x 1 plus 1 minus lambda x 2.

Now, go back and see what we have arrived at, we have arrived at the condition where function evaluated at x naught is less than equal to lambda f of x 1 plus 1 minus lambda f of x 2. And this is from the result which I have just shown you, this is the condition for the function to be convex. Hence, we have proved that if the second derivative is positive then the function is convex. If function has second derivative that is non-negative over the interval then the function is convex. And if the second derivative is positive then it is strictly convex. And similarly, we can also show the second derivative is non-positive then function is concave.

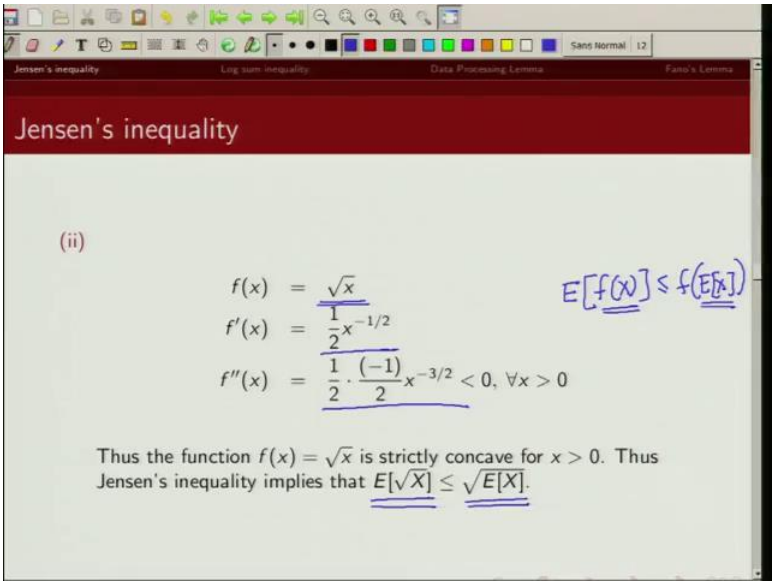
(Refer Slide Time: 18:46)



So, let us take some example to illustrate the Jensen's inequality that we have talked about. So, use Jensen's inequality to find appropriate inequalities between expected value of this function e raised of minus a x and e raised to minus a expected value of x, where a is greater than equal to 0. And the second function that we are considering is expected value of under root X and square root of expected value of X. So, in the first case, this function is given by this. So, we need to first find out whether the function is concave or convex; and depending on whether the function is concave or convex, we can find relationship between expected value of the function and function evaluated at expected value.

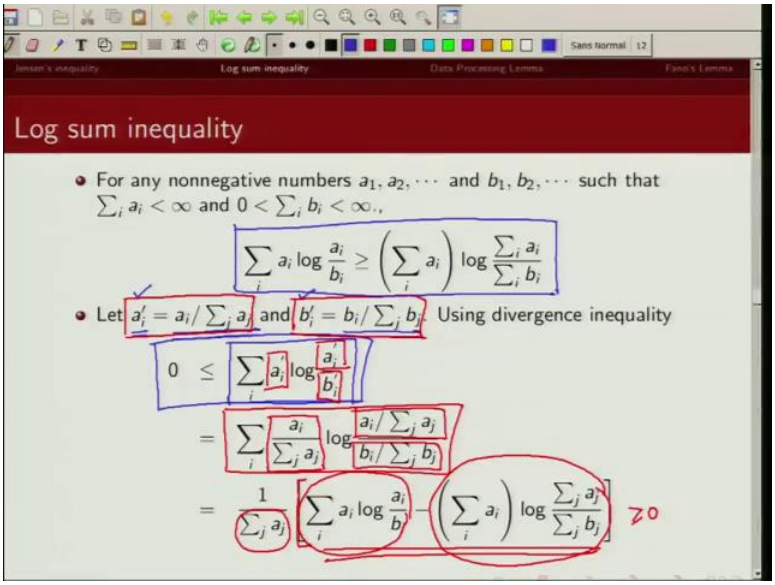
So, in this case, if we look at this function f of x, we take the first derivative that is given by this and the second derivative is given by this. And note a is greater than equal to 0, so a square is basically positive, this quantity e of e minus x is also positive, so this term will be greater than 0. The second derivative is greater than 0, what do we know about the function we know that the function is convex function. So, then our f of x is convex function. And strictly convex function because this is a second derivative is positive. Now, if it is a convex function then we know the expected value of the function is greater than equal to function evaluated at expected value of x. So, for this particular function, we can write expected value of the function is greater than equal to function evaluated at expected value, this is because this function was convex function.

(Refer Slide Time: 21:12)



Now, let us look at this second example. So, here the function is under root of x. Now, we take the first derivative that is given by this, and this is the second derivative. This second derivative is non-positive so that means this is a concave function. And if it is the concave function then we know expected value of the function is less than equal to function evaluated at expected value. So, what is the function here function is under root of x. So, expected value of under root of x will be less than equal to now the function which is under root x so that means, under root and function evaluated expected value of under root x will be less than equal to under root of expected value of x.

(Refer Slide Time: 22:16)



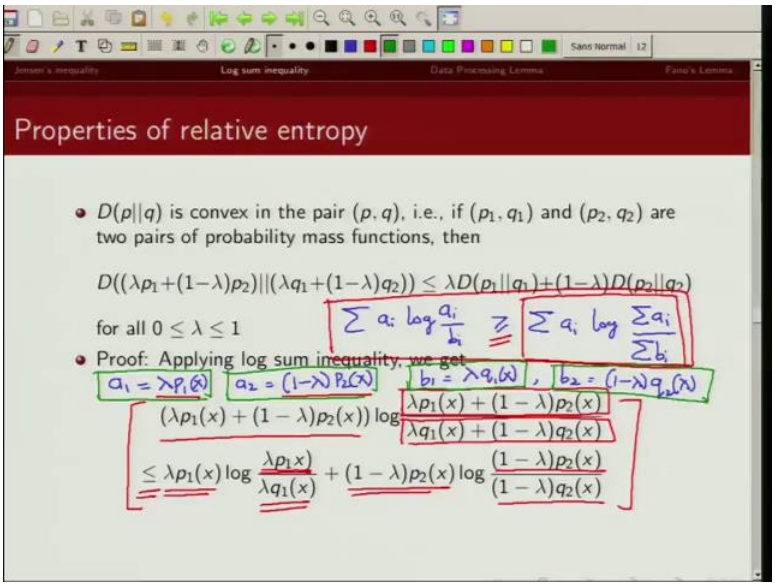
Now, let us just move now to log sum inequality. So, what does log sum inequality says if you have non-negative numbers a 1 a 2, a 3, a n lets say and b 1, b 2, b 3, b n where summation of if a i is bounded and summation of this b i greater than 0. And again its bounded then log sum inequality says that summation a i log a i by b i is greater than equal to summation of a i log of summation of a i by summation b i. So, this basically your log sum inequality.

Now let us prove this. So, let us consider a i dash which is a i by summation of this a i's bi's summation of this bi's. Now, let us find divergence between a i dash b i dash. Now, what do we know about divergence between two that is density function basically we know the divergence is greater than equal to 0. So, divergence between a i dash b i dash would be greater than equal to 0. So, this we are writing it here. So, this is the expression for divergence between ai dash and bi dash and we know that divergence is greater than equal to 0, this we have proved in the earlier lectures.

Now, let us replace a i dash and b i dash in this expression. So, a i dash is given by this and b i dash is given by this. So, we plug in these values of a i dash and bi dash, what we get is this expression. You can see this is my a i dash this is my a i dash and this is my b i dash. Now, further simplifying, I take this summation a i out, so what I get here is

summation of a i log a i by b i minus summation a i log of summation of a i divided by summation of bi's, now I know this is greater than equal to 0. Now, this is sum of non-negative numbers, so that is the positive quantity. So, when we this is greater than equal to 0, when this term is greater than equal to this term. So, what we have proved here is now then this relationship holds and this is known as log sum inequality.

(Refer Slide Time: 25:32)



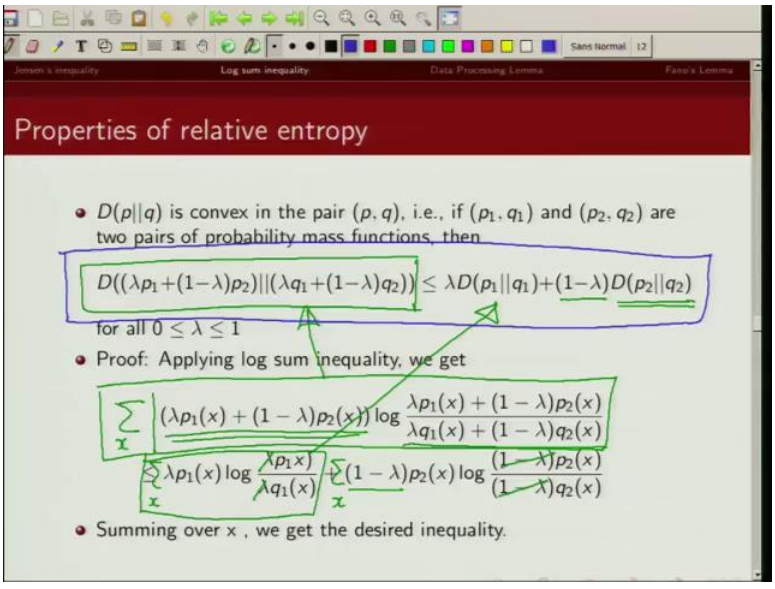
Now, let us take a simple example, where we will prove that divergence is a convex function in the probability period p and q. And we are going to make use of log sum inequality to prove this result. So, we want to show that divergence between p and q is convex function in the pair p and q. So, what do we mean is if p 1, q 2 and p 2, q 2 are pairs of probability mass function then function evaluated at x naught is less than equal to expected value of the function. This is when the function is convex. When the function we have just showed that function at expected value of x is less than equal to expected value of function this is for the case when f of x is convex, this is from the Jensen's inequality. So, to prove that this is convex we have to then from Jensen's inequality we have to show this result where lambda lies between 0 and 1.

So, let us see how we can make use of log sum inequality to prove this result. Now, we are applying log sum inequality. Now, what does log sum inequality says the log sum

inequality says summation a i log a i by b i this is greater than equal to summation a i log of summation a i by summation b i. Now, let us take a 1 to be lambda times b 1 x and a 2 to be 1 minus lambda times b 2 x. Similarly, let us take b 1 to be lambda times q 1 x and b 2 to be 1 minus lambda times q 2 x and plug in this value of a 1, a 2 and b 1, b 2 into our log sum inequality. Then according to log sum inequality summation a i log summation a i by summation b i is less than equal to summation of a i log of a i by b i.

So, what is summation a i this is this plus this, so this is this quantity. What is summation a i, so log of summation a i is this quantity. And what is summation b i, b 1 is lambda q 1 x and b 2 is 1 minus lambda q 2 x. So, this is my summation bi. Now according to log sum inequality this summation is less than equal to summation of a i log a i by b i. So, this summation is less than a 1 log a 1 by b 1 plus a 2 log a 2 by b 2. So, what is a 1 this is my a 1 and this is my b 1, and this is my b 1 this is my a 2, this is my a 2 and this is my b 2. So, this result which I have written so far is just a direct application of log sum inequality where I chose a 1 to be given by this a 2 to be given by this b 1 to be given by this and b 2 to be given by this. So if I do this using log sum inequality, I get this result.

(Refer Slide Time: 30:29)

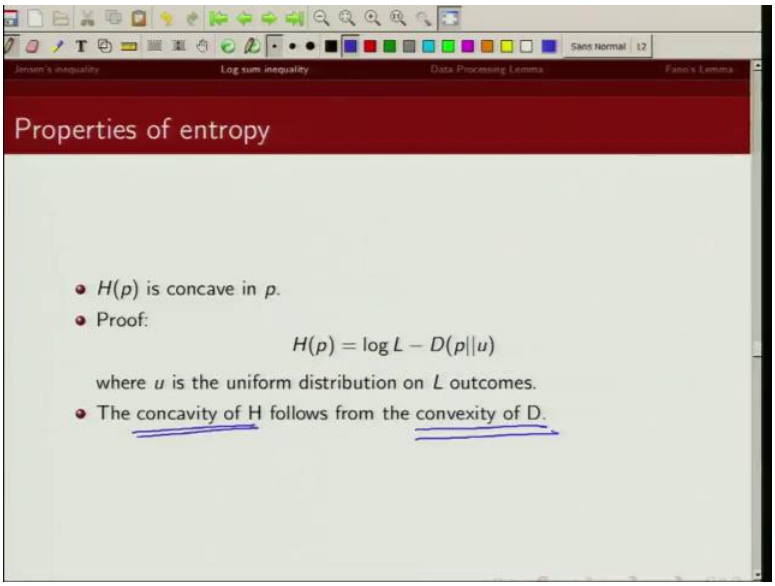


Now, what is the next step I do, I sum it over all x, so I sum it over all x, sum it over all x, sum it over all x, sum it over all x. So, what I get here, you can think of new

probability which is lambda p 1 x plus 1 minus lambda p 2 x, and this is another new probability lambda q 1 x plus 1 minus lambda q 2 x. So, this quantity that you see here this quantity you see here is nothing but this divergence between this probability and this probability. So, the term that you see on the left hand side is nothing but this term.

What about this? This is lambda p 1 x log of this lambda, lambda cancels out. So, this is lambda times p 1 x log p 1 x by q 1 x sum over x. This quantity is nothing but this quantity; lambda times divergence between p 1 and q 1. And what about this quantity second quantity, it is 1 minus lambda times p 2 x log this 1 minus lambda 1 minus lambda cancels out, this p 2 x log of p 2 x q 2 x summation of over x 1 minus lambda times. So, this term will be 1 minus lambda times divergence between p 2 and q 2. So, using log sum inequality we have shown that this relation holds. And from Jensen's inequality, we know this is the condition for this function divergence to be convex and the pair p and q.

(Refer Slide Time: 32:36)

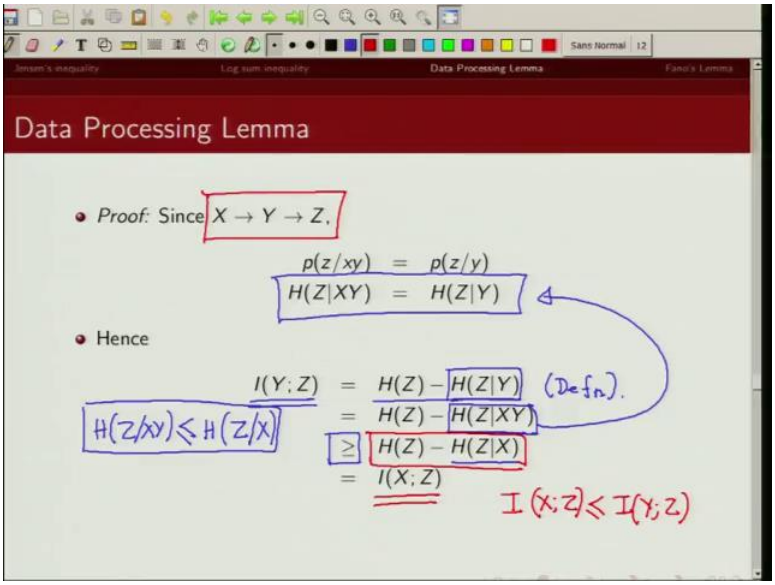


Now, let us use these results to do some more results on concavity and convexity. So, let us prove that entropy is a concave function of p. Now, in the previous lecture, we have shown when we have defined divergence between two random variables x and x hat

where x hat was uniformly distributed, we have shown that entropy of p is basically log of l minus divergence between p and uniform distribution.

Now, we have already shown in the previous slide that divergence function is a convex function, so minus of a convex function will be a concave function. So, then from the previous result it is very clear that entropy is a concave function of p that is because this is just a constant and minus of a convex function will be a concave function. So, as a function of p h of p entropy of p is a concave function. So, this result follows from the previous result that we have shown that divergence is a convex function. So, the concavity of entropy function follows from the convexity of the divergence function.

(Refer Slide Time: 34:13)



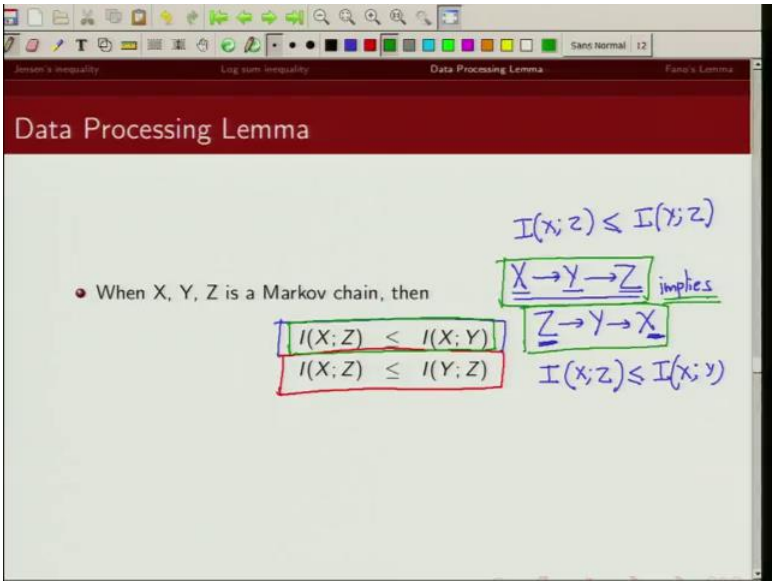
Now, let us prove and state what is known as data processing lemma. So, data processing lemma says that if we have random variable x y and z. And if they form a Markov chain then mutual information between x and z is less than mutual less than mutual information between x and y or mutual information between x and z is less than mutual information between y and z. So, in other words, further processing of data does not increase in information content. So, how do we prove it? So we are first going to prove. So, let us prove, so since x y and z forms the Markov chain then we use the probability of Markov chain, probability of Z given X Y will only dependent probability of Z given

Y because X Y and Z form the Markov chain. Or in terms of entropy we can write uncertainty in z given x y is same as uncertainty in z given y this follows from the point that x y z form the Markov chain.

So, let us write down the mutual information between Y and Z. Now, following the definition of mutual information, we can write this follows from the definition this follows from the definition. So, mutual information, I can write as h of z minus h of z by given y now from this I know H of Z given y is same as H of Z given X and Y. So, I replace this H of Z given by this uncertainty in z given x y so this line follows from here Markov process.

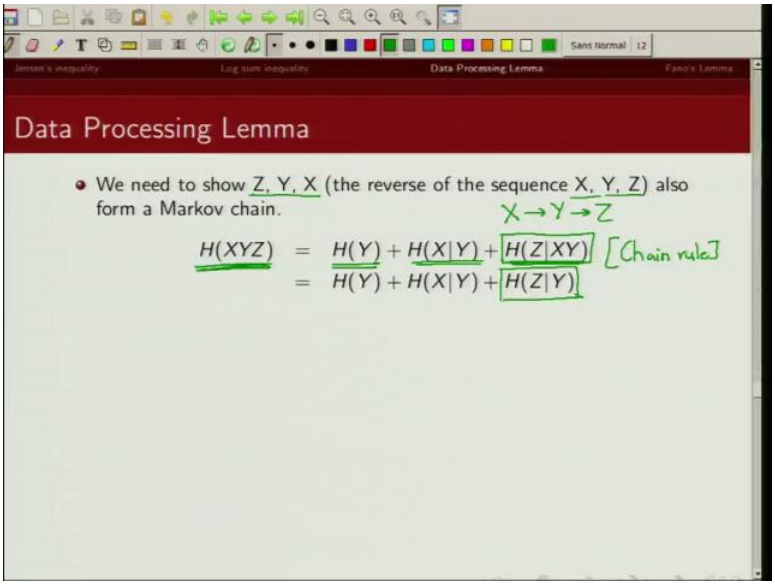
Now, I know that conditioning cannot increase entropy so uncertainty in z given x y is less than equal to uncertainty in Z given X, this follows from the property that conditioning cannot increase entropy, if I use this relationship and replace uncertainty in Z given X Y by larger term H of Z given X. So, what I get here is greater than equal to sine and I am replacing this by H of Z given X. And what is this, this term uncertainty in Z minus uncertainty in Z given X thus mutual information between this is from the definition this is the mutual information between X and Z. So, what I have shown you is mutual information between X and Z is less than equal to mutual information between Y and Z, when X, Y and Z forms a Markov chain.

(Refer Slide Time: 38:01)



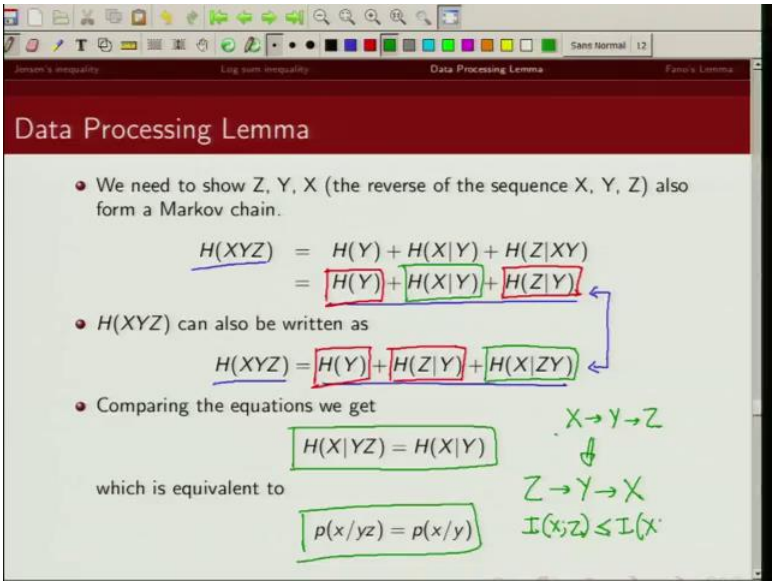
So, I have shown you this, I have proved this result. Now, how do I prove this result? Now if you see closely, if I can show that if x y and z forms a Markov chain. Then it also follows that z y and x forms a Markov chain then I can use the previous derivation to show that mutual information between z and x. If I can show that this implies, if I can show that x y and z the form a Markov chain then z x and z y and x also forms a Markov chain if I can prove this then from the previous result which I just now proved. I can show that mutual information between X and Z will be less than mutual information between X and Y because I just now showed you that if X, Y and Z forms a Markov chain. Then mutual information between X and Z is less than equal to mutual information between Y and Z so to prove this result it is sufficient to prove that if X, Y, Z forms a Markov chain it implies that Z, Y and k also forms a Markov chain.

(Refer Slide Time: 39:46)



So, we will show now that if X, Y and Z they form a Markov chain then Z, Y and X will also form a Markov chain. So, to prove this we will first consider the joint entropy X, Y and Z; and applying chain rule we can write this as uncertainty in y plus uncertainty in X given Y plus uncertainty in Z given X Y. Now, since x y and z they form a Markov chain we know that uncertainty in Z given X Y is same as uncertainty in z given y so we can write this joint entropy between X, Y and Z as H of Y plus H of X given Y plus H of Z given Y.

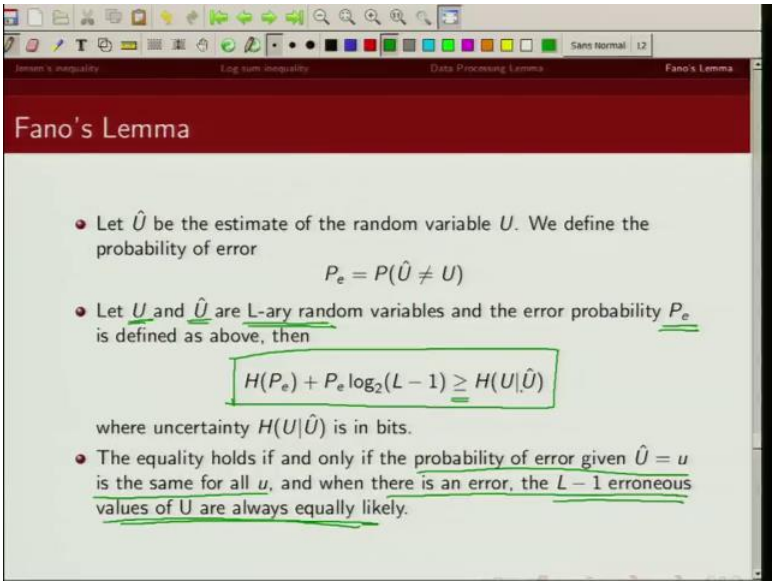
(Refer Slide Time: 41:08)



Now, we are writing this joint entropy again using chain rule in a different way so I can write the same thing as H of Y plus H of Z given Y plus H of X given Z Y. Now, let us compare this is joint entropy given by this expression and this is joint entropy of X, Y, Z given by this expression. So, let us compare these two equations. This is a common term H of Y it is here, it is here; and then I have a common term h of z given y and a common term here H of Z given Y. So, these two are same what do I get, I get this condition that this term must be equal to this. So, in other words, uncertainty in x given y z is same as uncertainty in X given Y and this is true when so in terms of probability I can write probability X given Y, Z is same as probability of X given Y.

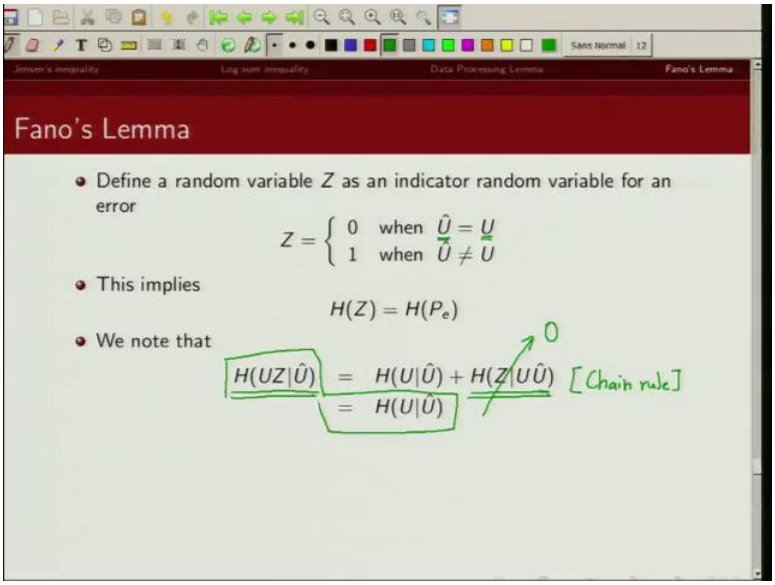
So, when is this true? This is true when Z, Y and x form a Markov chain. So, in that case probability of X given Y Z will only depend on probability of x given y. So, what I have show you is if X, Y and Z forms a Markov chain this also implies that Z Y and X will also form a Markov chain. And hence using the proof that we had just said before we can write that mutual information between X and Z is less than equal to mutual information between X and Y. And this proves our second result that we had this proves our result mutual information between X and Z is less that mutual information between X and Y. So, this is our data processing lemma.

(Refer Slide Time: 43:51)



Now, finally, we are going to state and prove what is known as Fano's lemma. So, we have a random variable U, let us denote U hat is the estimate of this random variable U and we define probability of error. So, when does an error occur, an error occur if our estimated data is not same as the transmitted data. So, if U hat is not same as U error happens and probability of that will give you probability of error. Now, what does Fano's lemma says is as follows. If U and U hat are L-ary random variables and probability of error is given by P sub e then this relation holds. The binary entropy function of well the entropy of this P e plus P e log of L minus 1 is greater than equal to uncertainty in U given U hat. And the equality here happens if and only if probability of error given u hat u is same for all u and if there is an error all L minus 1 erroneous value are equally likely to happen. So, Fano's lemma links probability of error to the estimate of U that we are getting uncertainty in U given U hat.

(Refer Slide Time: 45:37)

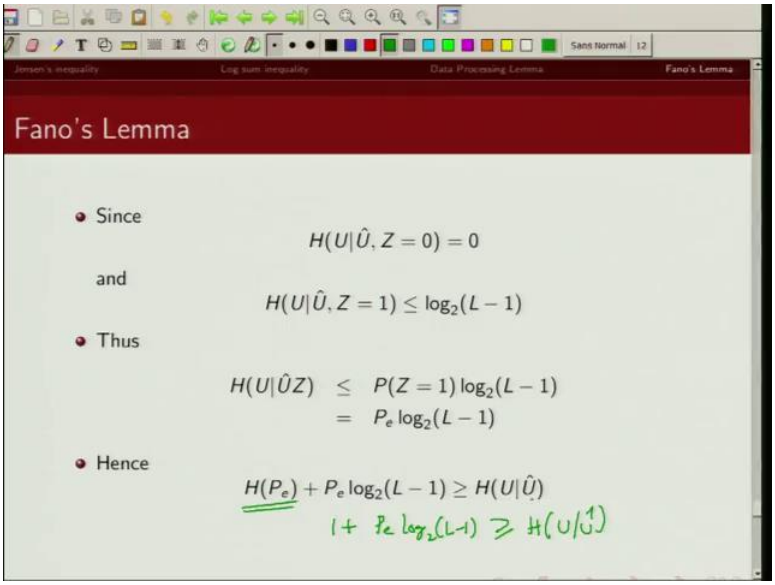


So, let us prove this result. So, let us denote a random variable Z, which basically denotes or indicates whether there is an error. So, this random variable Z which is an indicator a random variable will be 0, if there is no error and then will be 1. If there is an error, so Z take two values either or one 0 when there is no error and 1 when there is an error. So, what is the entropy associated with Z, so we see basically, so Z takes two value so with probability 1 minus P this thing happens there is no error this probability and with this probability there is an error. So, then the entropy of Z will be given by minus of P log P minus 1 minus p log 1 minus P which is nothing but entropy of P.

Now, let us look at this term uncertainty in U Z given the estimate of U. Now, we apply chain rule. So, this if we apply chain rule what we get is uncertainty in u given u hat plus uncertainty Z given U, U hat is equal to this quantity. Now, let us look at each of these quantities what is the uncertainty in Z given U, U hat. Now, what is Z? Z is an indicator random variable which indicates whether there is an error and it depends on what is the value of U and U hat, if you tell me what u and u hat are I can easily find out. What is Z, because if U is same as U hat, Z is 0 and U is not same as u hat and z is one so there is no uncertainty in Z given U hat, so this term is 0 then I can write this term is equal to this fine.

Now, the same thing I am writing, so this expression I got get from here. Now, again I have to write chain rule. So, uncertainty in U Z given U hat can be written as uncertainty in Z given U hat plus uncertainty in U given U hat in U given U hat Z. Now, uncertainty we know that conditioning cannot increase uncertainty so h of z given u hat has to be less then equal h of z. So, then I can write this as this term is less than equal to h of z plus h of u given U hat Z and when is this, this, this, this is same when basically when u hat is independent of Z. So, in other words, when this condition holds when probability of error is same irrespective of what U is then H of Z given u hat will be H of Z. So, from here to here, it follows from the fact that conditioning cannot increase entropy.

(Refer Slide Time: 49:50)



Next, We are going to look into little more detail with this expression. So, let us look at this expression. So, we will first look at this when Z is 0 and then we will look at the case when Z is 1. So, what is the uncertainty in U given U hat and Z is 0. Now, what does Z equal to 0 means, Z equal to 0 means there is no error. So, if I give you U hat and I tell you this no error then you know precisely what is U. So, there is no uncertainty in U given U hat and Z equal to 0. So, then this uncertainty is 0. What about when X is 1, when Z is 1, this corresponds to the case when U is not same as U hat that means the error has happened. So, in this case, you know that U is not equal to U hat, but you do not know which of the remaining l minus 1 possibility. So, we know that U can take any

value except U hat when Z is 1, and we know that if you discreet random then entropy is upper bounded by number of possibilities of x. In this case, number of possibilities for U is L minus 1, because U is not same as U hat and U is a L valued random variable. So, this entropy is upper bounded by log of L minus 1.

So, if I plug that in then the uncertainty in U given U hat and Z can be written as probability of Z equal to 1 uncertainty in U given U hat and Z equal 1 plus probability of z equal to 0, and uncertainty in U given U hat and z equal 0. Now, this uncertainty is 0, so only term that will be left is probability of Z equal to 1 multiplied by uncertainty in u given U hat and Z equal 1. And this quantity we just showed is upper bounded by this, so we plug this value in here what we get is probability of Z equal 1 that is basically probability of error and this is log of L minus 1.

So what we have shown is this quantity is upper bounded by this quantity. Now when is this equal to this? When all the errors are equally likely to happen, we know that H of X is equal to log of number of possibilities of X that is when this is uniformly distributed. So, when all the L minus 1 possibilities erroneous possibilities are equally likely then this happens with this equality. So, in this expression of H of U given U U hat Z, we can plug in this upper bound. If we do that what we get is this expression.

So again let me just go back. We had this expression; H of U given U hat is less than equal to H of Z plus H of U given U hat and Z. We have shown H of Z is given by H of p and we showed that this is this quantity is upper bounded by P log of L minus 1. So, if we plug this values this is H of P e, if we plug this values in we get our Fano's lemma which is this. Note, this is convex function of P this is a linear function of P, so sum of a convex function of P and linear function, so this is what is our Fano's lemma. Now we can also get a weaken version of this Fano's lemma, we know that h of p is upper bounded by 1. So, we can get a weaker version of P this replaced by 1. We can also say 1 plus p e log of L minus 1 is greater than equal to H of U given U hat, so this relates probability of error to the uncertainty in estimation of U.

Thank you.