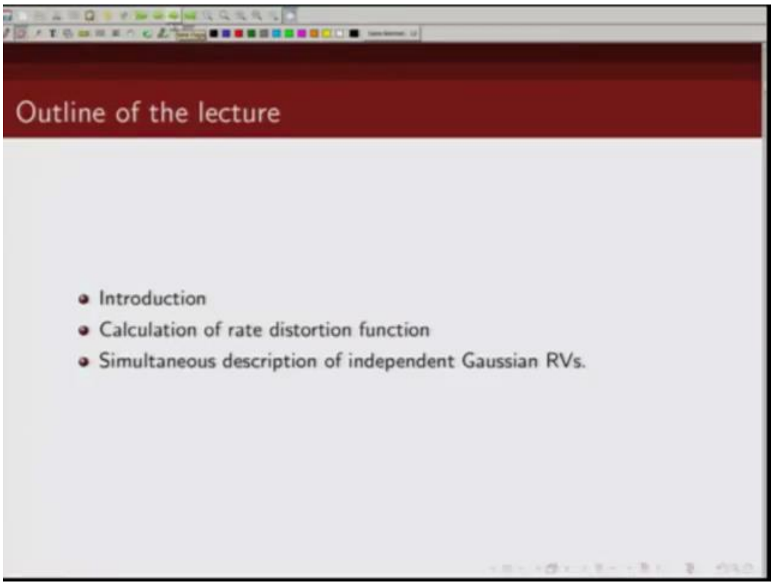
An Introduction to Information Theory
Prof. Adrish Banerjee
Department of Electronics and Communication Engineering
Indian Institute of Technology, Kanpur

Lecture - 13
Rate Distortion Theory

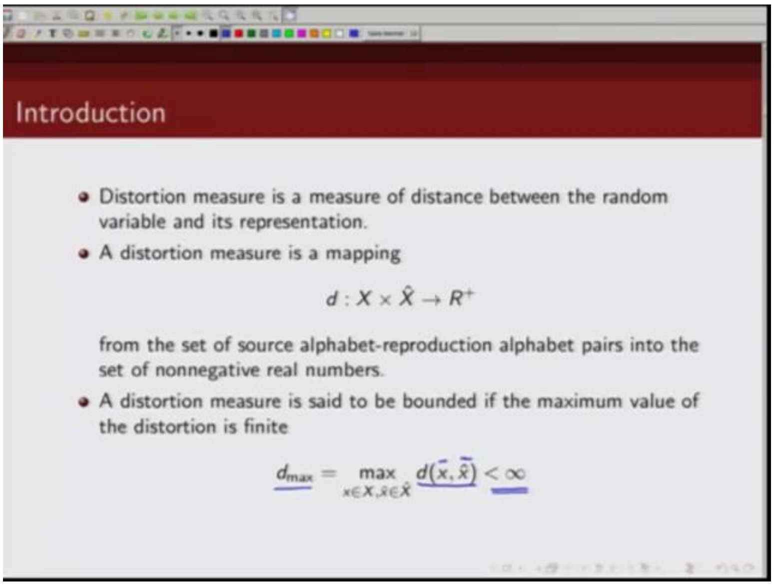
Welcome to the course on An Introduction to Information Theory. So, in this lecture we are going to talk about Rate Distortion Theory. When we try to represent an arbitrary real number to represent it glosslessly, we would require infinite number of bits. Now, whenever we try to represent it using finite number of bits, we are introducing distortion. So, to know that our representation of this real number is good, we need to introduce this notion of distortion measure and in rate distortion theory, we are go to ask questions like for example, if we specify that we can tolerate this much average distortion, then what is the minimum number of bits required to represent a source or let us say if I specify my rate, then what is the minimum distortion that I can achieve.

(Refer Slide Time: 01:28)



So, we will start up this discussion on rate distortion theory with a brief introduction and few definitions and we will define what is a rate distortion code and what do we mean by a rate distortion code is achievable and then we will calculate rate distortion function for some simple examples like Bernoulli source with hamming distortion measure, Gaussian source with squared error distortion measure and we going to abort the description of independent Gaussian random variables, but they are not necessarily identically distributed.

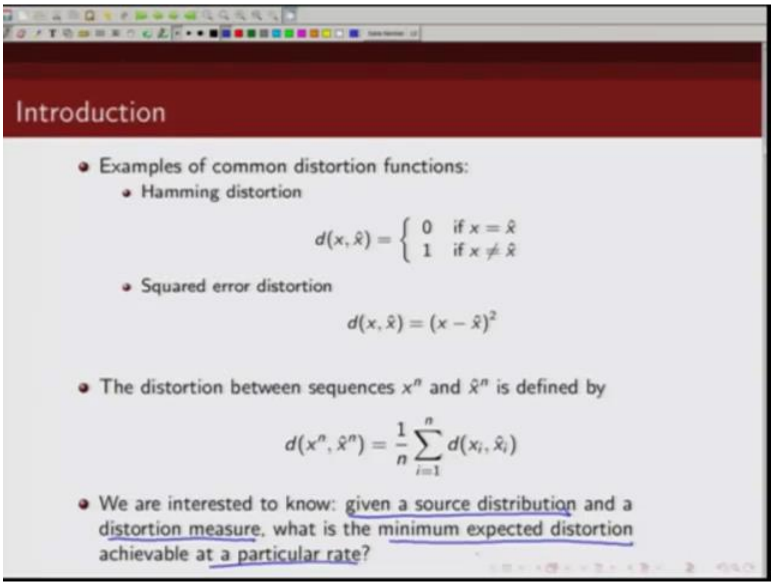
(Refer Slide Time: 02:06)



So, a Distortion measure is a measure of distance between the random variable and its representation. When you are trying to represent an arbitrary real number using fixed number of bits, we are introducing some sort of a distortion. So, distortion measure is a measure of distance between this random variable and its representation. Now, this measure of distance, we will specify that there could be different ways in which we could define this measure of distance between the source alphabet and the re-produce alphabet.

So, a distortion measure is a mapping of set of source alphabet and reproduction alphabet pairs into a set of non-negative real number and this set of non-negative real number will specify how much this source alphabet is differing from this reproduced alphabet. We say a distortion measure is bounded if the maximum value of distortion is finite. So, we define the maximum value of distortion as maximum value of distance between the random variable and its representation. Now, if this is bounded, then we say the distortion measure is bounded.

(Refer Slide Time: 03:48)

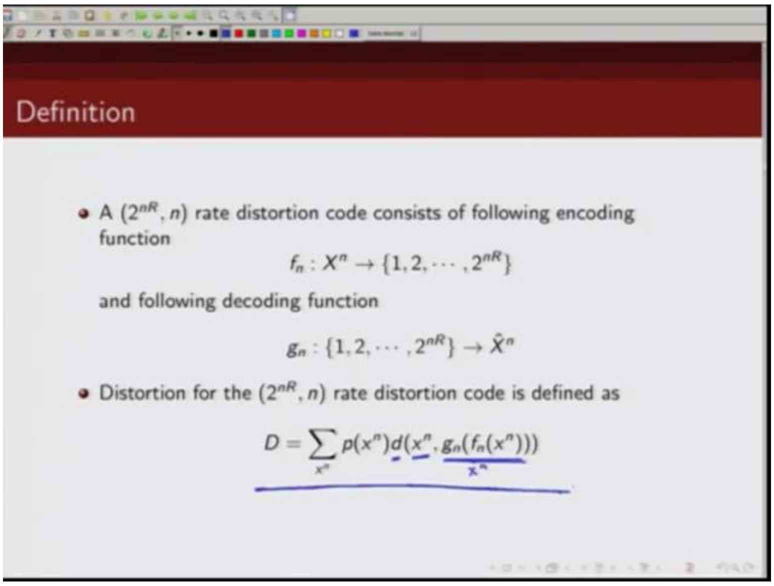


Let us give some example of common distortion function, hamming distortion function. So, hamming distortion measure is defined as follows, when the source alphabet and the reproduce alphabet are same, there is no distortion that is distortion is 0 when the source alphabet and the reproduce alphabet are same. However, if they are different, then the distortion is 1. So, in hamming distortion it is zero, if there is no error and it is 1 if the source alphabet and the reproduce alphabet they are different.

Similarly, we could define squared error distortion. So, squared error distortion is basically the Euclidean distance between the source alphabet and this reproduce alphabet and this is our squared error distortion function. Now, the distortion measure that we have defined so far is on symbol by symbol basis and we could extend this definition for sequences as well. So, to extend this definition of distortion measure for sequences, this is the source sequence and this is the reproduce sequence, then that distortion measure can be written as average distortion measure per symbol. So, this is distortion between symbol source symbol X i and the reproduced symbol X i hat, we sum over all m symbols and then we average it out and that is how we would define distortion measure for sequences.

Now, as I said in the beginning of this lecture in this rate distortion theory given a source distribution and a distortion measure, we are interested in what is the minimum expected distortion for an achievable particular rate. So, given a source distribution and given a distortion measure, what is the minimum expected distortion that is achievable for a particular rate? We can ask this question in a different way also. Given a source distribution and a distortion measure and given average distortion criteria, we are interested in knowing the minimum rate that is achievable.

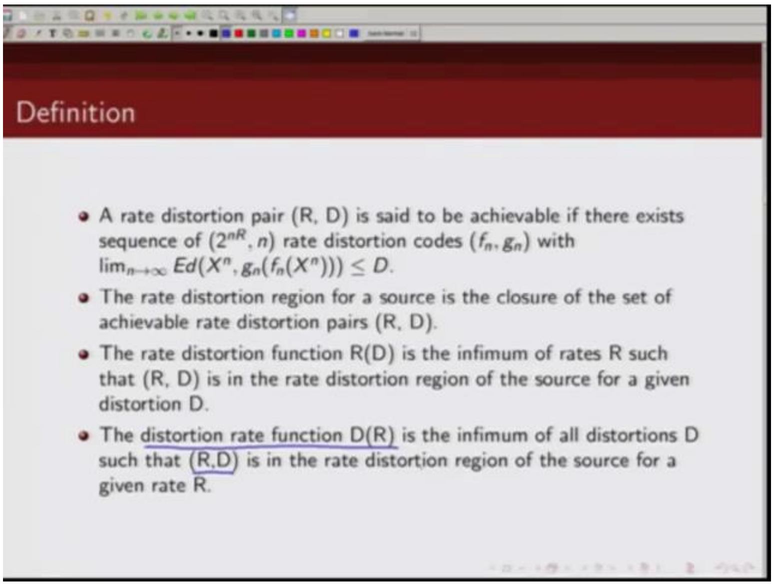
(Refer Slide Time: 06:56)



So, a 2 raise power n r n rate distortion code consists of following encoding function. This is your source sequence of n bit and this is your mapping to 1 of these rate distortion codes and at the decoder which is denoted by this function g of n given that you have received 1 of these indexes, you are interested in getting back than estimate of the sequence.

Now, distortion for a rate distortion code is defined like this. This is the distortion measure between the source sequence and this is the reproduce sequence which I can also write this as X n hat. So, this is basically the average distortion between X of n and X hat of n. We say a rate distortion pair denoted by this rate R and this distortion D is said to be achievable if there exist a 2 raise power n R n rate distortion code with following encoding and decoding function as n tends to infinity, the expected distortion is less than equal to the given distortion D.

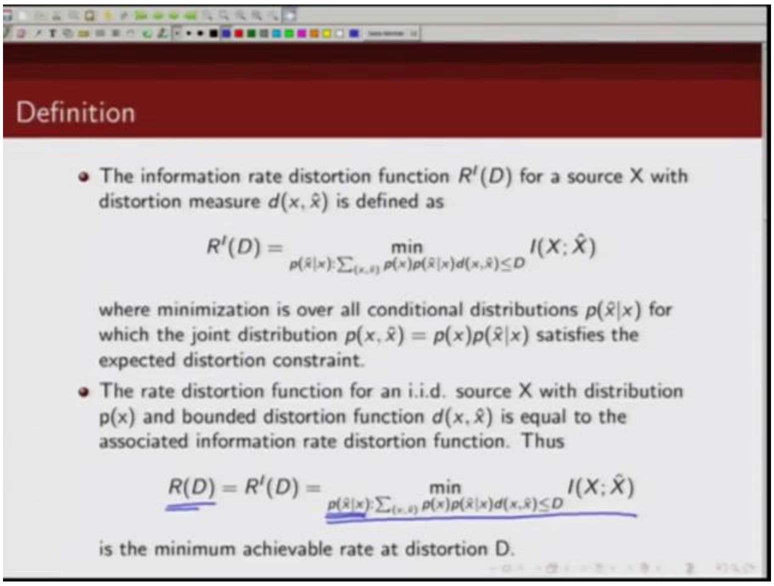
(Refer Slide Time: 08:09)



Now, we could similarly define a rate distortion region. So, a rate distortion region for a source is the closure of set of all achievable rate distortion pair, that is our rate distortion region and we define a rate distortion function as the infimum all rates such that this rate distortion pair is in the rate distortion region of the source for a given distortion.

Similarly, we can also define distortion rate function as the infimum of all distortions D such that this pair rate and distortion is in the rate distortion region of the source for a given rate R.

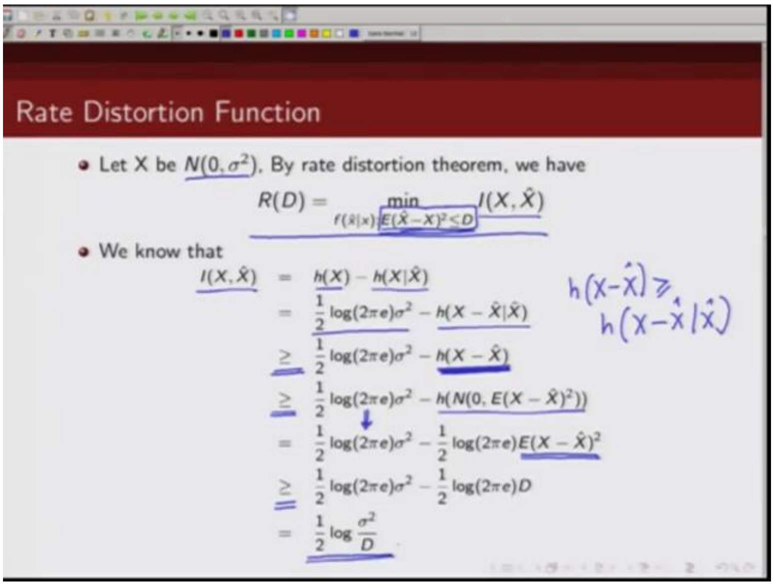
(Refer Slide Time: 09:56)



Now, we define information rate distortion function for a source X with distortion measure given by D of X n X hat. We define information rate distortion as minimum mutual information between this source alphabet X and this reproduced alphabet X hat. So, this is a reproduce source where mutual information between X and X hat, minimum of that and minimization is taken over all conditional distribution p of X hat given X for which the joint distribution p of X x hat satisfies the expected distortion constraint. For this information rate distortion function, we minimize this mutual information between X and X hat and this minimization is over all conditional distribution of p X hat given X such that over this joint distribution p of X x hat, this expected distortion constraint is satisfied.

Now, this information rate distortion function is also our rate distortion function. So, rate distortion function for an IID source with distribution p of X and bounded distortion function given by D of X x hat is equal to its associated information rate distortion function. In other words, the rate distortion function can be found by minimizing the mutual information between X and X hat and we do this minimization over all conditional distribution of p of X hat given X such that this joint distribution p of X x hat satisfies the expected distortion constrain.

(Refer Slide Time: 12:37)

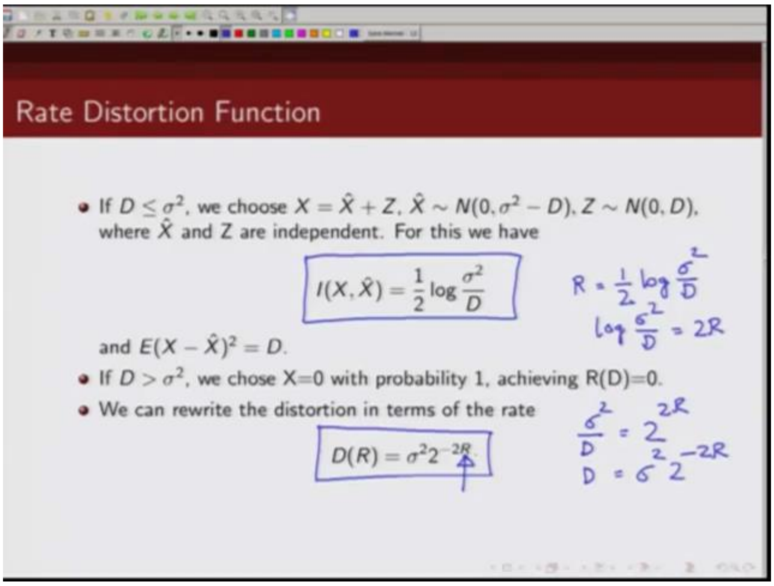


This rate distortion function is given by this and this is the minimum achievable rate for a distortion D. Now, let us illustrate or let us compute this rate distortion function. So, we will first take an example of a Gaussian source. We have a Gaussian source X denoted by X 0 mean variance sigma square. According to the definition, the rate distortion function is given by minimizing the mutual information between X and X hat and this minimization is over all conditional distribution of f X hat given X such that the expected distortion constraint is satisfied. Now, let us compute the mutual information between X and X hat. So, where we are going to compute this rate distortion function is as follows, we are going to lower bound this mutual information between X and X hat and then we are going to show that this lower bound is achievable and that is how basically we are going to show that it is a value of the rate distortion function. So, mutual information between X and X hat, this is given by differential entropy of X minus differential entropy of X given X hat.

Now, X is a Gaussian source 0 mean variance sigma square. So, its differential entropy is given by this expression half log of 2 pi e into sigma square. Now, we know that translation does not change differential entropy. So, differential entropy of X given X hat is same as differential entropy of X minus X hat given X hat. Now, this is differential entropy of X minus X hat condition on X hat. We know that conditioning cannot increase entropy. So, H of X minus X hat is going to be greater than H of X minus X hat given X hat and in this step we are subtracting a larger quantity and that is why I have here lower bound.

Now, we know that given a second order of movement, Gaussian source has the maximum differential entropy and that is why this is upper bounded by differential entropy of a Gaussian source which means 0 and same variance which is given by expected value of X minus X hat square. So, this greater than equal to comes because differential entropy of H X minus X hat is upper bounded by differential entropy of a Gaussian random variable with same second order movement. Now, this is same as this and we know the differential entropy of a Gaussian source. So, that is given by half log 2 pi e variances expected value of X minus X hat whole square, that is this term and this is upper bounded by D. So, then we can write mutual information between X and X hat to be half log of 2 pi e times sigma square minus half log 2 pi e times D, that is because expected value of X minus X square is upper bounded by D and by combining these terms, we get this.

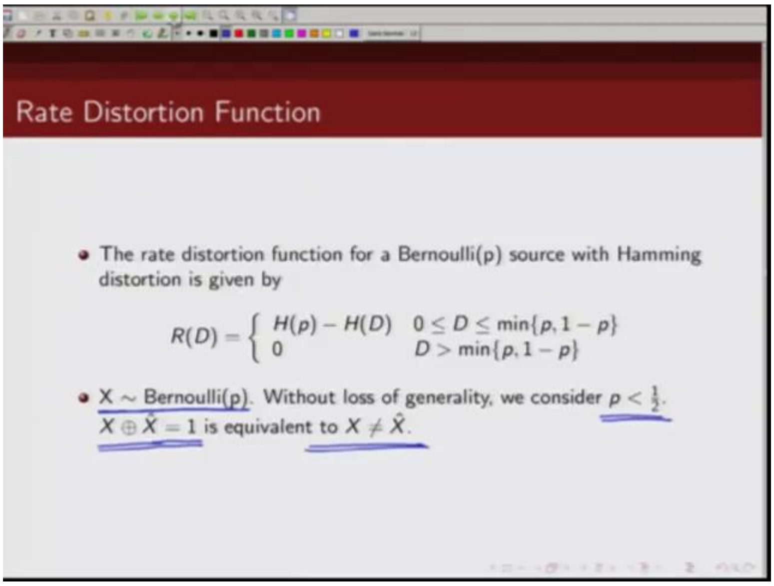
(Refer Slide Time: 17:46)



So, the mutual information between X and X hat is lower bounded by half log of sigma square by D. Now, we are going to show that this lower bound is achievable. So, if D is less than sigma square, we choose X to be X hat plus Z and X hat is Gaussian distributed with 0 mean and variance sigma square minus D and Z is Gaussian distributed with 0 mean and variance D and X is Gaussian distributed with mean 0 and variance sigma square. So, mutual information between X and X hat is given by differential entropy of X minus differential entropy of X given X hat. Now X is Gaussian distributed and X given X hat is also Gaussian distributed and we can compute this mutual information which comes out to be half of log sigma square by D and this is precisely the lower bound that we computed here.

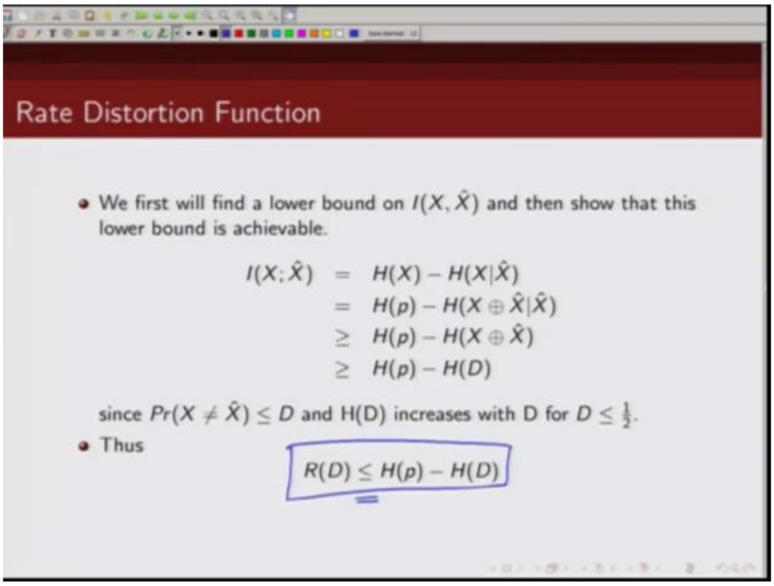
So, this mutual information is actually achievable if we take our X hat and Z in this particular rate and what happens if D is greater than sigma square, in that case we choose X equal to 0 with probability 1 and that gives us rate as 0. Now, this rate can also be written in terms of distortion. So, I have this sigma square by D or log of sigma square by D is 2 R or I can write sigma square by D to be 2 raise to power 2 R or D is sigma square 2 raise to power minus 2 R and I can write distortion in terms of rate and if rate increases, then distortion decreases and that makes sense and if you are using more number of bits to represent a quantity, basically our distortion is going to decrease.

(Refer Slide Time: 20:51)



Now, let us take another example, this time we are considering a Bernoulli source and hamming distortion measure. So, we can show that the rate distortion function for this Bernoulli source is given by this expression and as long as D is within 0 and minimum of p and 1 minus p, it is given by H of p minus H of D; otherwise it is given by 0. Now, without loss of generality less assumed that this p is less than half and we are considering a Bernoulli source, this is a binary source, so X X or X hat. When X X or X hat is one, it means X is not same as X X hat. So, when X is from Bernoulli source, X X or X hat equal to 1 is equivalent of saying X is not same as X hat.

(Refer Slide Time: 22:20)

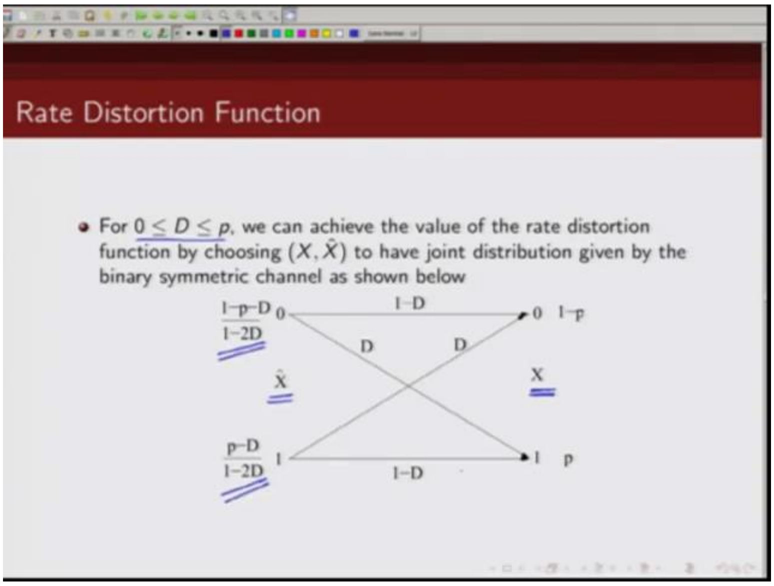


Now, again here to compute the rate distortion function, we will first put a lower bound on mutual information between X and X hat and then we will show that this lower bound is achievable. So, let us compute this mutual information between X and X hat. Now, from the definition of mutual information I can write this as H of X minus H of X given X hat. Now, X is a Bernoulli source with probability b. So, this is basically a binary source and its entropy is given by H of p. This is binary function of p.

Now, what is the uncertainty in X given X hat? So, there is uncertainty in X only when it is not same as X hat. So, uncertainty in X given X hat is basically uncertainty in X not equal to X hat and that is given by this X X or X hat. So, we can write this mutual information between X and X hat as H of p minus H of X X or X hat given X hat. Now, again this is condition on X hat and if you remove this conditioning, we know that conditioning cannot increase entropy. So, this quantity is larger H of X X or X hat, this is more than X. So, we are subtracting a larger quantity and that is why I have here mutual information between X and X hat is greater than equal to H of p minus H of X X or X hat and X X or X hat is basically when there is error and this probability of error is upper bounded by D. So, this is upper bounded and I can write this as greater than equal to H of p minus H of D.

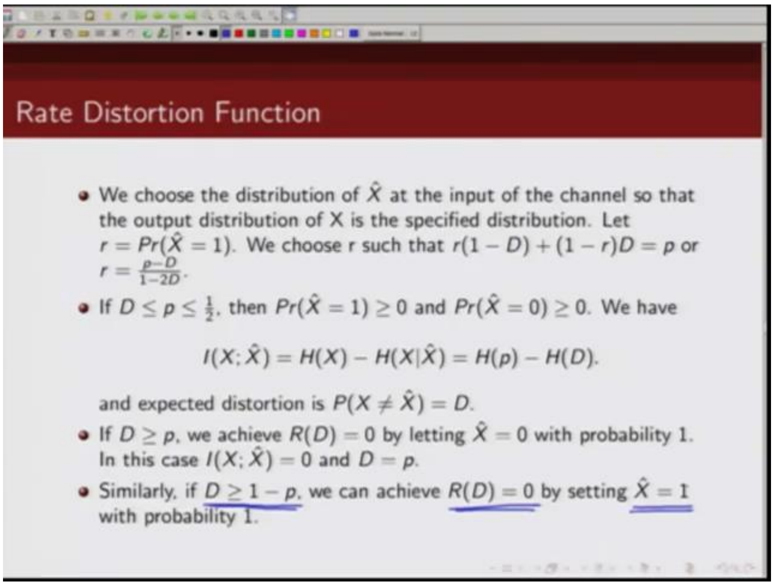
So, the rate distortion function is given by this for the case when D is within minimum of p or 1 minus p. Now, let us consider the case when D lies between 0 to p and remember our objective is to cook up a distribution such that this rate is achievable with equality. So, X is distributed as Bernoulli with p and we want H of X given X hat to be basically H of D. So, we have to cook up a distribution.

(Refer Slide Time: 25:54)



We create a test channel. This is my X and this is my X X hat. Now, I want to create a joint distribution such that my X is distributed as Bernoulli we know with 0 probability 1 minus p 1 with probability p and I need to find out this distribution on X hat keeping in mind we want to keep the distortion less than equal to D.

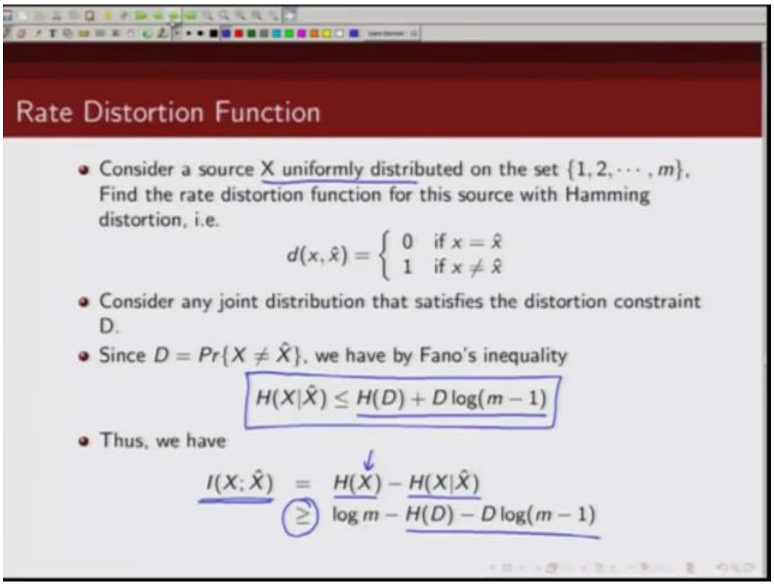
(Refer Slide Time: 26:36)



Now, we choose the distribution of X hat at the input of this test channel such that the output distribution is the specified distribution. So, let r equal to probability of X hat be 1. So, we need to choose r in such a way such that r times 1 minus D plus 1 minus r times D that is equal to p and r times if this is the probability of X hat been 1. So, r times 1 minus D plus 1 minus r times D that has to be p. So, if I do that, I get this input distribution on X hat. So, probability of X hat being 1, I get this as p minus D divided by 1 minus 2 times D. Now, if p lies between D and half, then probability of X hat being 1 is greater than equal to 0 and probability of X hat being 0 is greater than equal to 0 and then in this case the mutual information in X and X hat is equal to H of p minus H of D where the expected distortion is given by D. So, if we take this input distribution on X hat, we are able to achieve this lower bound. So, the rate distortion function is given by H p minus H of D.

Now, for the case when D is greater than equal to p, R D is 0 and we let X hat is equal to 0 with probability 1 and in this case mutual information between X and X hat comes out to be 0 and distortion is basically p and similarly for D greater than equal to 1 minus p, we can achieve R D equal to 0 by setting X hat is equal to 1 with probability 1. In this case also the mutual information between X and X hat will come out to be 0.

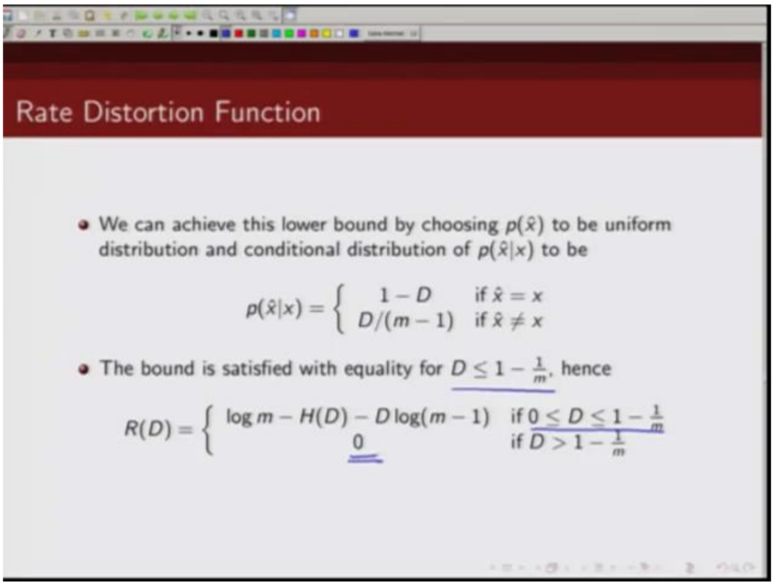
(Refer Slide Time: 29:15)



So, we have a source which is uniformly distributed in this set 1 2 3 and m and let us compute its rate distortion function for hamming distortion. So, if X is equal to X hat, this is 0 and if X is not equal to X hat, distortion is 1.

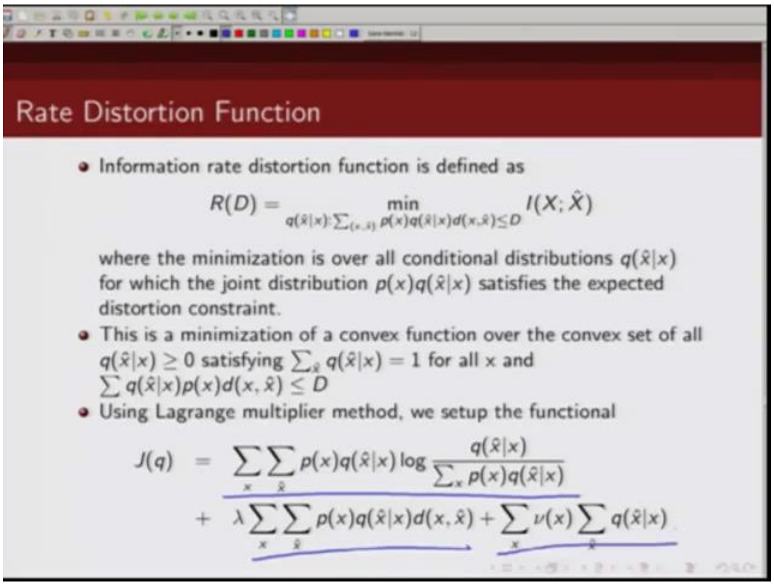
Now, consider any joint distribution that satisfies this distortion constraint of D and since D is probability of X not been equal to X hat, we can invoke Fano's lemma and if we invoke Fano's lemma, we can show the uncertainty in X given X hat is less than equal to H of p, probability of error which is H of D plus D times log of m minus 1, this is p log of number of possibilities minus 1. So, that is our Fano's lemma. So, from Fano's lemma, we get this. Next, we compute the mutual information between X and X hat and this is given by entropy of X minus entropy of X given X hat. Now, from the Fano's lemma, we have an upper bound on H of X given X hat. So, if we subtract this upper bound, since we are subtracting a larger quantity, so we now have a lower bound on mutual information between X and X hat. So, mutual information between X and X hat is lower bounded by log of m minus H of D minus D log of m and of course and this is maximum when X is uniformly distributed. So, H of X will be log of m.

(Refer Slide Time: 31:53)



Now, we can achieve this lower bound by choosing our p of X hat to be uniformly distributed and if we choose our conditional distribution in this way that if X is equal to X hat, probability of X hat given X is 1 minus D, otherwise it is D by m minus 1. Now, this bound is satisfied with equality when D is less than equal to 1 minus 1 by m. So, the rate distortion function for this uniform source under hamming distortion is given by log of m minus H of D minus D log of m minus 1 as long as D is less than equal to 1 minus 1 by m and for D greater than that, this is given by 0.

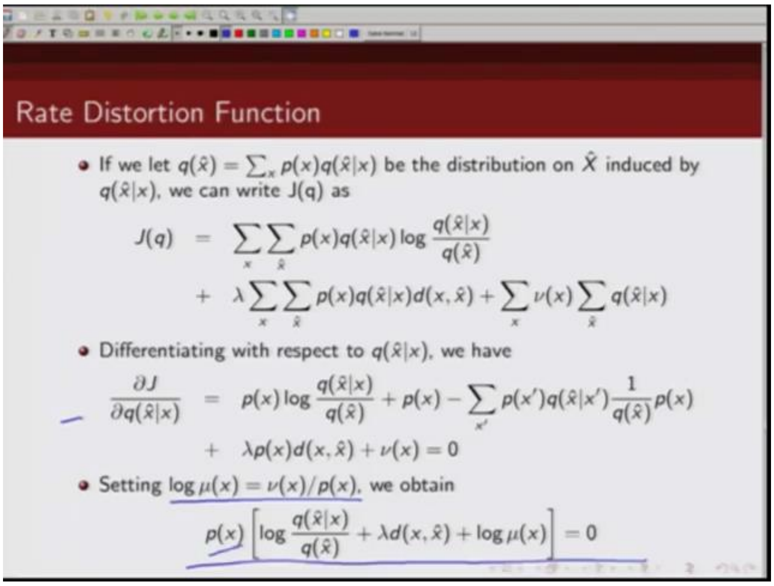
(Refer Slide Time: 32:59)



Now, let us characterize this rate distortion function little bit more. So, we know that the rate distortion function is given by this expression and we are minimizing their mutual information between X and X hat and remember this minimization is over all conditional distribution q of X hat given X for which this joint distribution p x times q of X hat given X, this satisfies the expected distortion constraint. So, this expected value of this distortion is less than equal to D.

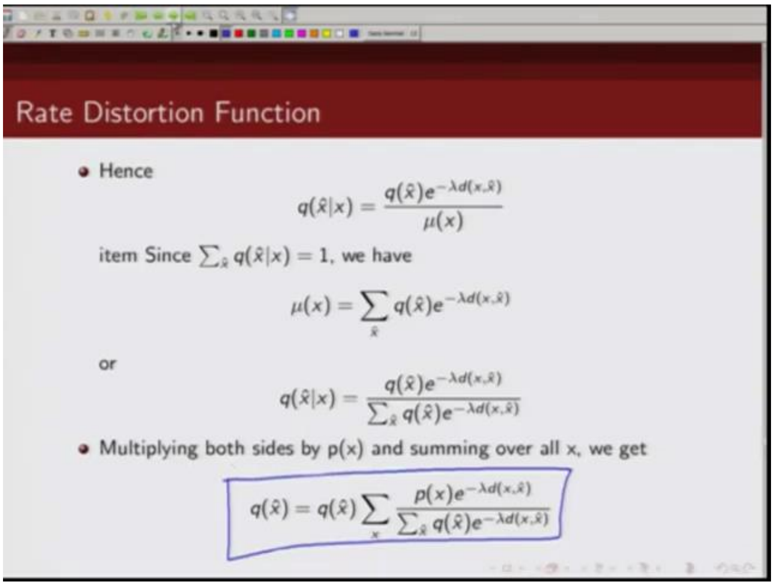
Now, this is a minimization of a convex function over a convex set for all q of X hat given X greater than equal to 0 satisfying this constraint that is sum of q of X hat given X over all X hat, that summation should be 1 and this average distortion constraint should be satisfied. So, again I repeat basically we are minimizing this convex function over this convex set and we have some additional constraint which is sum of probabilities is 1 and this expected distortion constraint should be satisfied. So, then we can set up our Lagrangian, basically using Lagrange multiplier method, we can setup our functional. So, this is our objective function, this is the constraint due to this expected distortion constraint and this is the constraint that comes due to summation of probabilities b equal to 1.

(Refer Slide Time: 35:19)



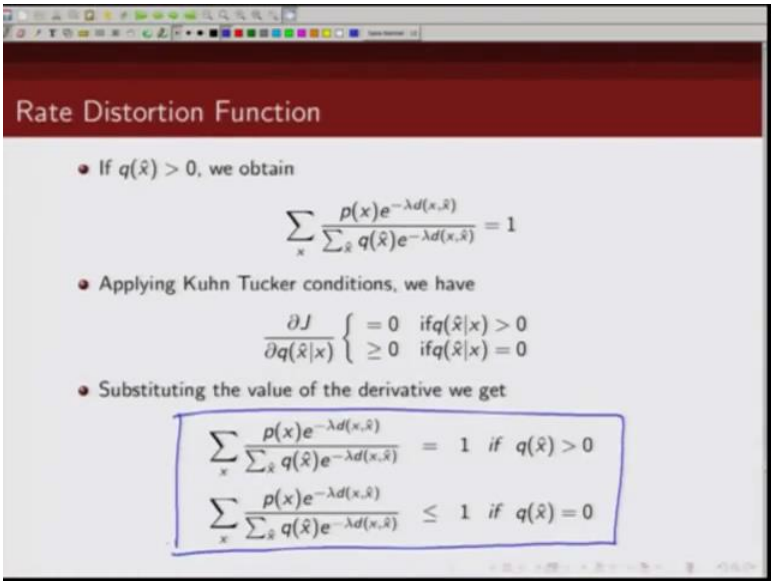
Now, let q of X hat be the distribution on X hat induced by this conditional distribution, then this functional can be written like this, if we differentiate this with respect to q of X hat given X and equate it to 0, we get this. So, p of X log of q of X hat given X by q of X hat plus p of X minus summation of p x prime q X hat given X prime into 1 by q X hat into p x plus lambda times p x and this distortion between X and X hat plus mu of X is equal to 0. Now, setting log of mu X to be v x by p x, from here we get this condition. So, we get p x times log of q of X hat given X by q of X hat plus lambda times distortion measure between X and X hat plus log of mu X that is 0. Now, this is non-zero. So, what we can get is this is 0 from here we can write q of X hat given X and that comes out to be this.

(Refer Slide Time: 37:06)



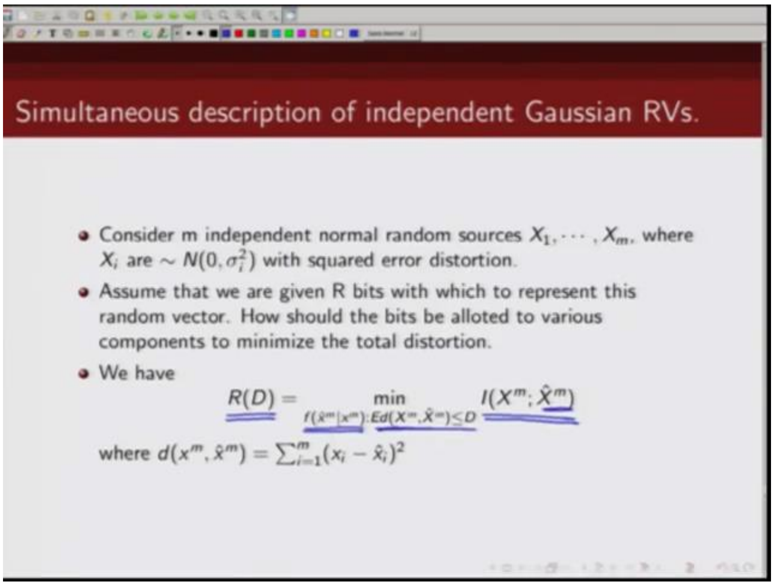
Now, we do not know lambda, but we know that summation of q X hat X is equal to 1. So, mu of X then comes out to be this and q of X hat given X is this quantity. Next, multiplying both side p of X and summing over all X, we get q of X hat is equal to q of X hat into this. Now, if q of X hat is greater than 0, then we get this condition that this summation is equal to 1.

(Refer Slide Time: 38:11)



So, if this is greater than 0, then you can cancel this out and this will be equal to 1. Now, applying Kuhn Tucker condition, a partial derivative of J with respect to this q of X hat given X is 0, if q of X given X hat is greater than 0 and this is greater than equal to 0 if q of X hat given X is 0. Now, substituting this we get this condition. Let us consider a case where we have m independent Gaussian random variables.

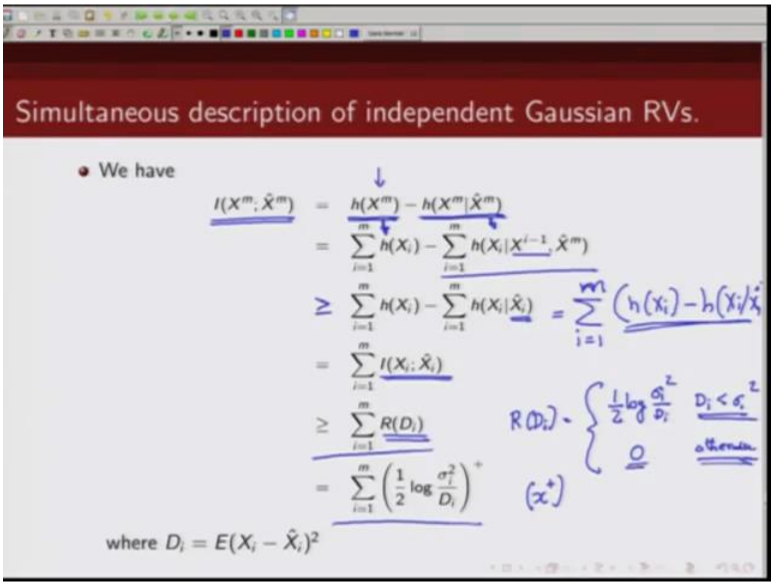
(Refer Slide Time: 39:10)



So, we have m independent Gaussian random variables and these are denoted by X 1 X 2 X 3 X of m. So, these are 0 mean Gaussian with variance sigma i square. They are not identically distributed, but they are independent and we are considering squared error distortion measure. So, assume that we are given R bits and with this R bits, we have to represent this random vector X 1 X 2 X 3 X of m. Now, the question that we are asking is how many bits we should be allocating to each of these components? So, how many bits we should be we should be allocating to X 1, X 2, X m such that we minimize the total distortion. So, again I repeat, we have m independent Gaussian sources which we are denoting by X 1 X 2 X 3 X m, these are independent but not identically distributed and we are considering squared error distortion measure. So, given there are R bits, the question that we are asking is how many bits should be used to represent X 1 X 2 X 3 X 4 X m such that overall distortion is minimized.

Now, from the definition of rate distortion function, we know that rate distortion function can be written as minimum mutual information between the source sequence X of m and this reproduce signal X hat of m and this minimization is over conditional distribution such that the expected distortion constraint is satisfied.

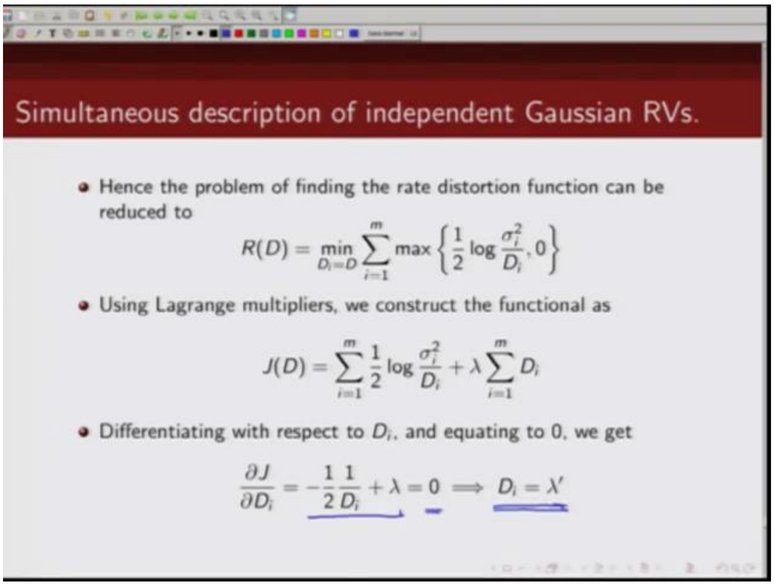
(Refer Slide Time: 41:46)



Now, let us look at mutual information between X m and X hat of m and from the definition of mutual information, we can write mutual information as differential entropy of X of m minus differential entropy of X of m given X hat of m, now X 1 X 2 X 3 are independent right. So, we can write the differential entropy h of X m as h of X 1 plus h of X 2 plus h of X 3 up to h of X m. So, this term can be written like this and similarly I can write this h of X m which is basically X 1 X 2 X 3 X m given X m hat I can write it in this particular fashion.

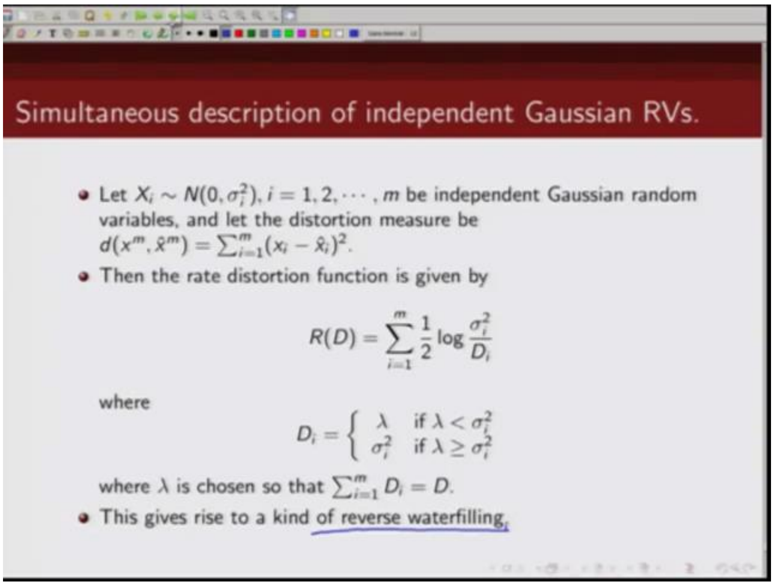
Now, this is conditioned X i is conditioned on this sequence X i minus 1 and X m. So, we know that conditioning cannot increase entropy. So, if I just condition it only on X hat i, then this is a larger quantity, so I can lower bound this mutual information and this can be written as summation i equal to 1 to m h of X i minus h of X i given X i hat. Now, from the definition of mutual information, this is mutual information between X i and X i hat. So, then mutual information between this sequence X of m and X hat of m is lower bounded by mutual information between X i and X i hat sum over i going from 1 to m and since we know rate distortion function is minimum of this, so this can be lower bounded by R of D is summation over i going from 1 to m and we know for the Gaussian source, we know the expression for rate distortion function and if you go back for a Gaussian source this given by half log of sigma square by D. So, we can write this as when D i is less than sigma i square else is 0 otherwise if I am just writing this as this operator X plus.

(Refer Slide Time: 45:35)



So, this is equal to X when this condition is satisfied, otherwise this is 0 and then the problem of finding the rate distortion function has been reduced to maximizing half log of sigma i square by D i or 0 for i going from 1 to m and minimizing this over D i, summation of D i basically is D. Now, using Lagrange multiplier, this is our objective function and this is our constraint, the summation of this D i from i 1 to m cannot exceed D. So, differentiating with respect to D i, we get del J by del D i as this and when we equate it to 0, we get this condition. So, this is interesting, it says that distortion D i should be same, D i is equal to some constant lambda dash.

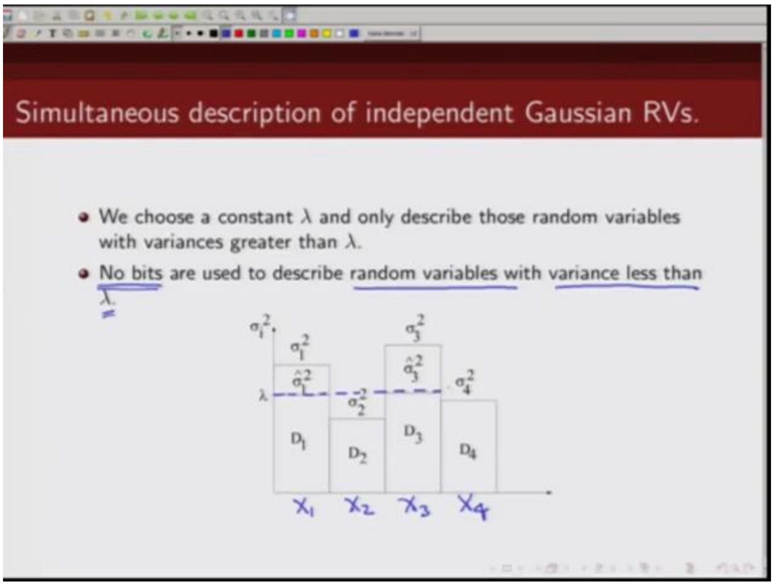
(Refer Slide Time: 47:01)



In fact, if X i is Gaussian distributed with 0 mean and variance sigma i square and if we consider m such independent sources given a square distortion measure, the rate distortion function is given by this expression where D i s are equal to some constant lambda as long as lambda is less than sigma i square, otherwise if lambda is greater than sigma i square D i is equal to sigma i square and we choose our lambda in such a way that sum of this distortion over these m sources that is equal to d.

So, this gives rise to what we call reverse water filling. In case of this panel Gaussian channel, we saw how we are allocating power, it was like putting water when you have various noise levels and here it is lateral difference. So, we choose a constant lambda and we only describe those random variables which have variance greater than lambda.

(Refer Slide Time: 48:34)



So, we do not use any bits to describe those random variables whose variance is less than lambda. Let us just look at let us say this is X 1 X 2 X 3 X 4 and these are the distortions D 1 D 2 D 3 D 4 So, we fix our lambda, this is my lambda. If lambda is less than sigma i square, D i is lambda and in this case lambda is less than sigma 1 square. So, D 1 is lambda. Similarly, D 3 is lambda however, for D 2 and D 4; lambda is greater than sigma square. So, in this case distortion will be given by sigma 2 square and sigma 4 square and as I said this is interesting so, we are only going to describe this random variables whose variance is greater than lambda and no bits are going to be used for those random variables whose variance is less than this threshold lambda.

So, in the next class we are going to talk about how we can compute this rate distortion function in an iterative fashion. Now, this same algorithm can also be used to compute channel capacity and that will be the discussion for next class.

Thank you.