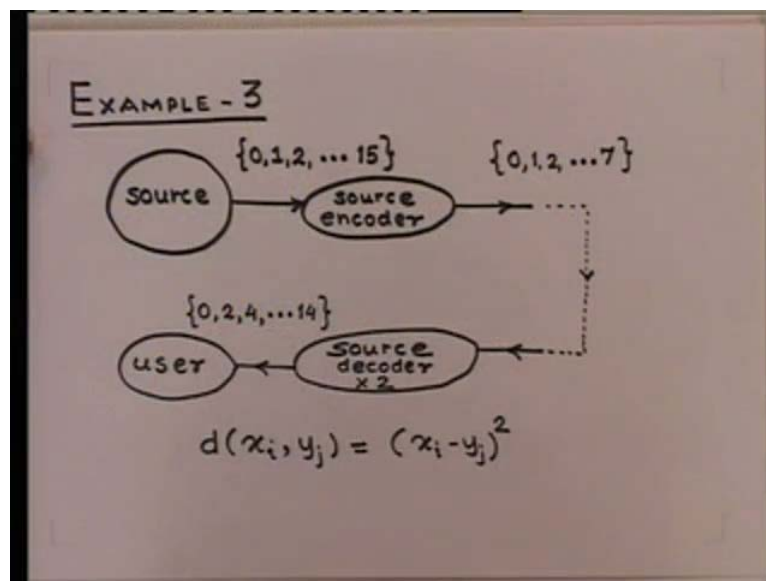


Information Theory and Coding
Prof. S. N. Merchant
Department of Electrical Engineering
Indian Institute of Technology, Bombay

Lecture - 32
Definition and Properties of Rate-Distortion Functions

In the previous class, we provided a qualitative definition for the rate distortion function as the lowest rate at which the output of a source can be encoded, while keeping the distortion less than or equal to some specified distortion level. We also investigated the rate and distortion for two different lossy compression schemes. In example one, we found that the rate was equal to the entropy of the reconstruction alphabet. However, this was the result of the fact that the conditional probabilities for that particular source encoder took on only two values 0 and 1. Now, this need not be the case always, let us take another example to illustrate this concept in a better way.

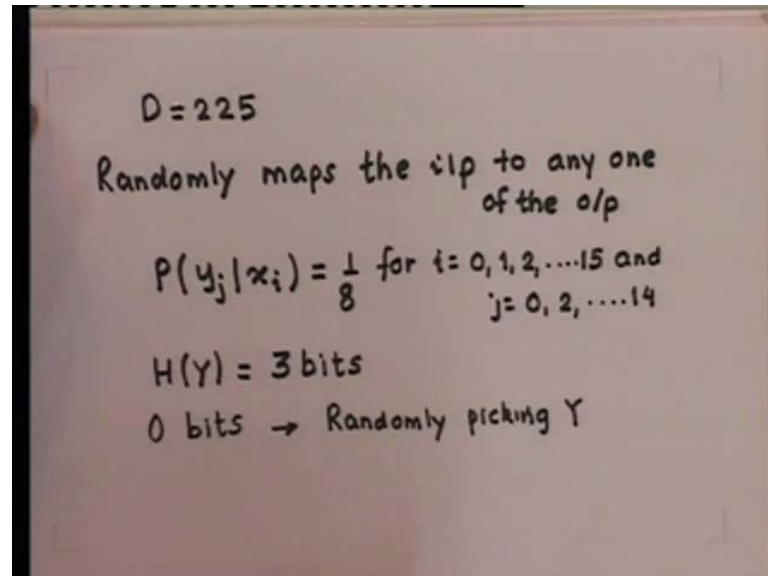
(Refer Slide Time: 02:28)



So, we consider example three which is the same as example one, where the source output consists of four bit words. The source encoder drops the least significant bit of this four bit word to generate three bit word. The source decoder multiplies the output from the source encoder by 2 to generate the reconstruction alphabet given here. So, for this example the source and the reconstruction alphabet is the same as it was in example one. Now, we use the distortion measure based on the square error difference and we

also assume that the distortion level is given to be D equal to 225. The problem is to design a lossy compression scheme that satisfies the distortion constrain.

(Refer Slide Time: 04:01)



So, one such compression scheme that satisfies the distortion constrain would be to randomly map the input to any of the output that is conditional probability of y_j given x_i is equal to $\frac{1}{8}$ for i equal to 0, 1, 2 up to 15 and j equal to 0, 2 up to 14. We can see that this conditional probability assignment satisfies the distortion constrain, as each of the eight reconstruction value is equally likely we get a entropy for the reconstructional alphabet, that is $H(Y)$ is equal to 3 bits. However, for such a glossy compression scheme we are not transmitting any information. So, we could exactly get the same results by transmitting 0 bits and randomly picking y at the receiver. What this example illustrates is that, the entropy of the reconstruction that is $H(Y)$ cannot be the measure of the rate.

(Refer Slide Time: 06:00)

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= \sum_i \sum_j P(x_i) P(y_j|x_i) \log \frac{P(y_j|x_i)}{P(y_j)} \end{aligned}$$

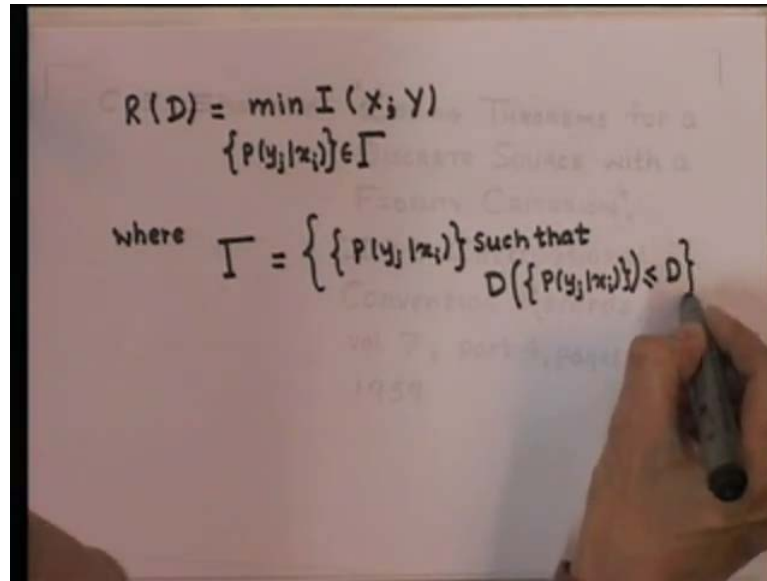
where $P(y_j) = \sum_i P(y_j|x_i) P(x_i)$

$$\{P(x_i)\} \quad \{P(y_j|x_i)\}$$
$$I(X; Y) = I(\{P(y_j|x_i)\})$$

Now, let us look now let us look at average mutual information that is $I(X; Y)$ which is given is $H(X)$ minus entropy of X given Y or this can be rewritten as entropy of Y minus conditional entropy of Y given X . Now, this can be rewritten as double summation over i, j , probability of x_i , conditional probability of y_j given x_i and log of conditional probability y_j given x_i over probability of y_j where probability of y_j is equal to summation over i of conditional probability y_j given x_i multiplied by probability of x_i summed over i .

So, this mutual information is a measure of correlation between the random variable X and Y . Now, in the rate distortion theory, the source that is $P(x_i)$ is given, so that the average mutual information will be regarded as the function of the conditional probabilities y_j given x_i . So, we can write the average mutual information as a function of condition probabilities.

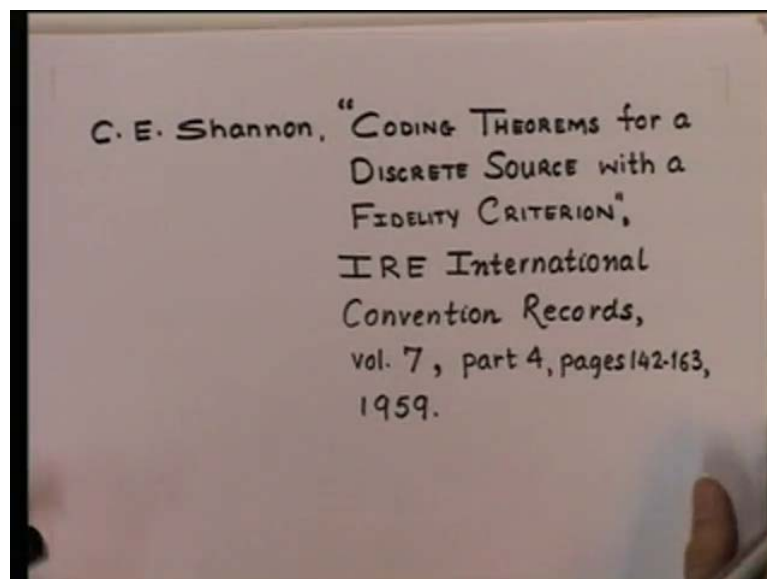
(Refer Slide Time: 08:40)


$$R(D) = \min_{\{P(y_j|x_i)\} \in \Gamma} I(X; Y)$$

where $\Gamma = \left\{ \{P(y_j|x_i)\} \text{ such that } D(\{P(y_j|x_i)\}) \leq D \right\}$

Now, Shannon in his 1959 paper on source coding has found the relationship between the minimum rate and the distortion and this is stated as follows. The rate distortion function is equal to minimum of average mutual information. This minimum is over the conditional probabilities, where this conditional probability belongs to the set given by all those conditional probabilities which satisfies the distortion criteria.

(Refer Slide Time: 10:04)

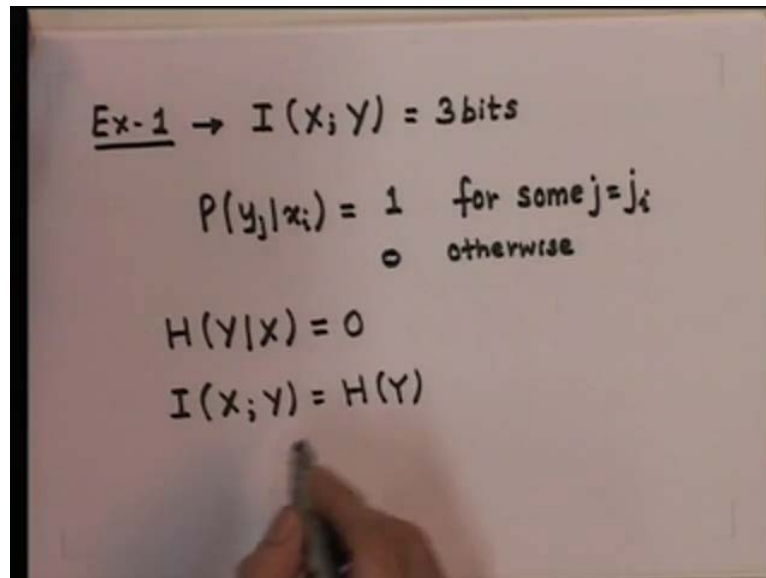


C. E. Shannon, "CODING THEOREMS for a DISCRETE SOURCE with a FIDELITY CRITERION", IRE International Convention Records, vol. 7, part 4, pages 142-163, 1959.

The paper by Shannon is stated here, coding theorems for a discrete source with a fidelity criterion it had appeared in I R E international convention records volume 7 part

4 in 1959. Now, we will not go into the proof of this relationship, however we can at least convince our self that defining the rate as a mutual information gives sensible answers when used for the example discussed.

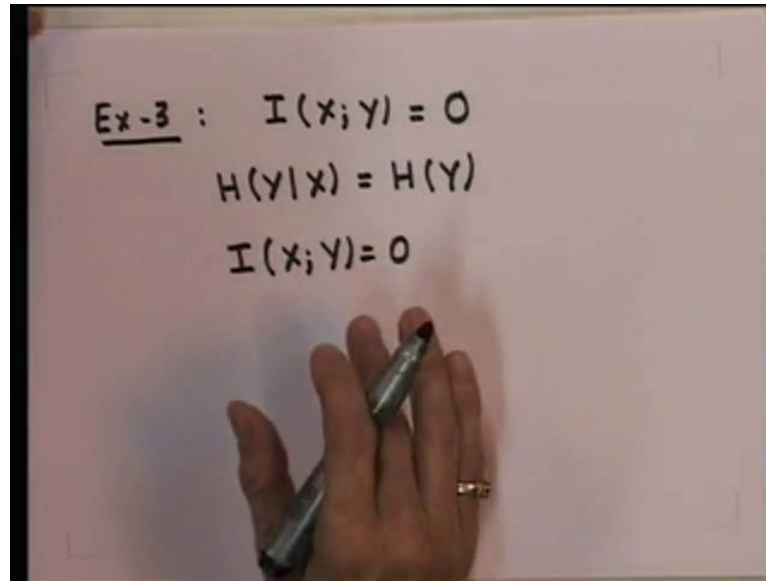
(Refer Slide Time: 10:54)



The image shows a whiteboard with handwritten mathematical expressions. At the top, it says 'Ex-1 → I(x; Y) = 3 bits'. Below that, a conditional probability is defined: $P(y_j | x_i) = 1$ for some $j = j_i$, and 0 otherwise. Then, the conditional entropy is given as $H(Y|X) = 0$. Finally, the mutual information is equated to the entropy of Y: $I(x; Y) = H(Y)$.

So, if we consider example one, which we had discussed in the previous class. We calculated the average mutual information to be 3 bits which is what we said the rate was. In fact notice that whenever the conditional probabilities are constrained to be of the form given equal to 1 for some j equal to j_i is equal to 0, otherwise. Then conditional entropy H of Y given X is equal to 0 and in this case mutual information is equal to entropy of Y which had been a measure of rate.

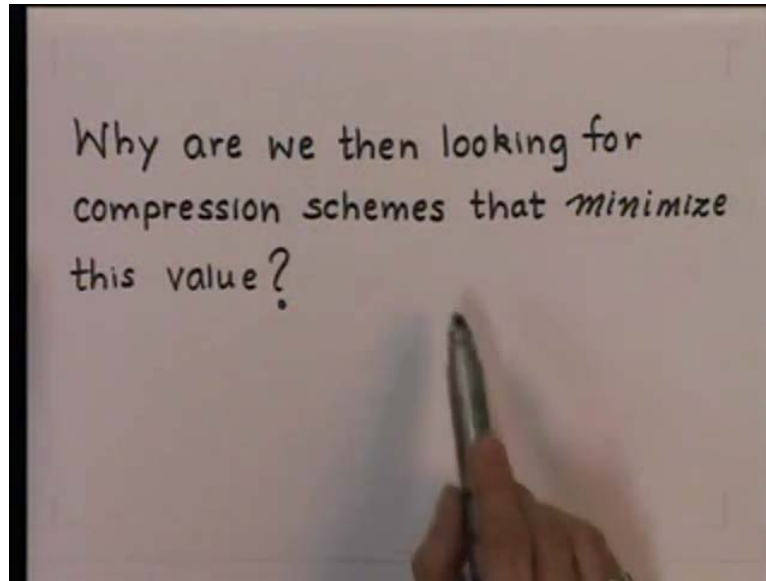
(Refer Slide Time: 12:15)

A photograph of a whiteboard with handwritten mathematical equations. A hand holding a silver marker is visible at the bottom center. The equations are:
$$\underline{\text{Ex-3}} : I(X; Y) = 0$$
$$H(Y|X) = H(Y)$$
$$I(X; Y) = 0$$

In example three, the average mutual information is 0 bits, which accords with our intuitive feeling with what the rate should be. Again whenever conditional entropy H of Y given X is equal to H of Y that is the knowledge of the source gives us no knowledge of the reconstruction. Then mutual information is equal to 0 which seems entirely reasonable.

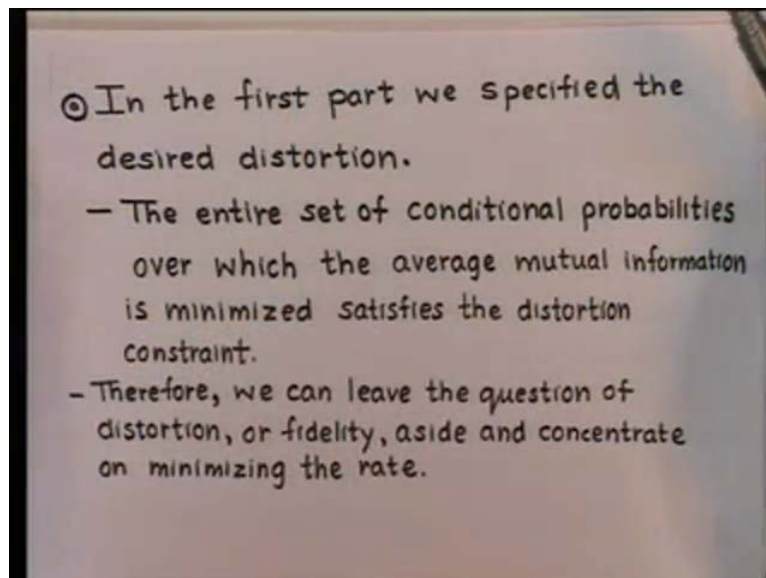
We should not have to transmit any bits when we are not sending any information. At least for the examples discussed here, it seems that the average mutual information represents the rate. However, based on our earlier study we can say that the average mutual information between the source output and the reconstruction is the measure of the information conveyed by the reconstruction about the source.

(Refer Slide Time: 13:45)



So, the natural question is why are we then looking for compression schemes that minimize this value? Now, to understand this we have to remember that the process of finding the performance of the optimum compression scheme had two parts.

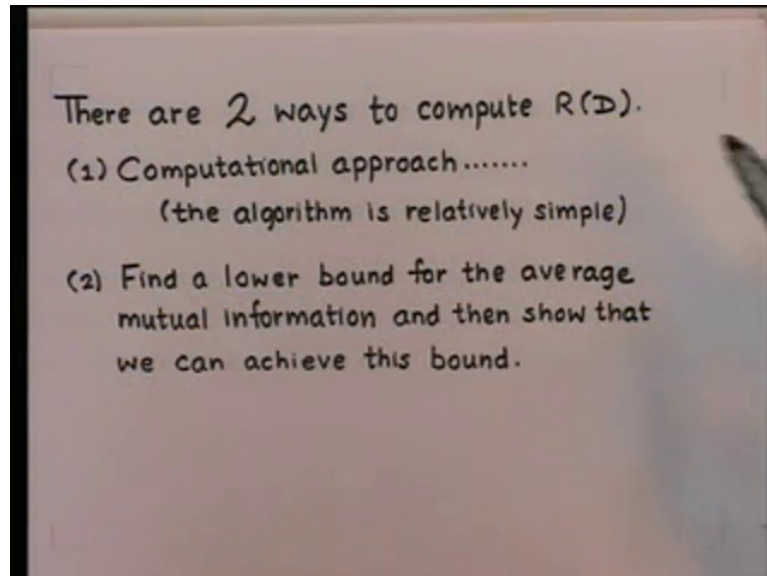
(Refer Slide Time: 14:10)



In the first part we specified the desired distortion. so the entire set of conditional probabilities over which the average mutual information is minimized satisfies the distortion constraint. Therefore, we can leave the question of distortion or fidelity, aside and concentrate on minimizing the rate. So, this provides some kind of qualitative

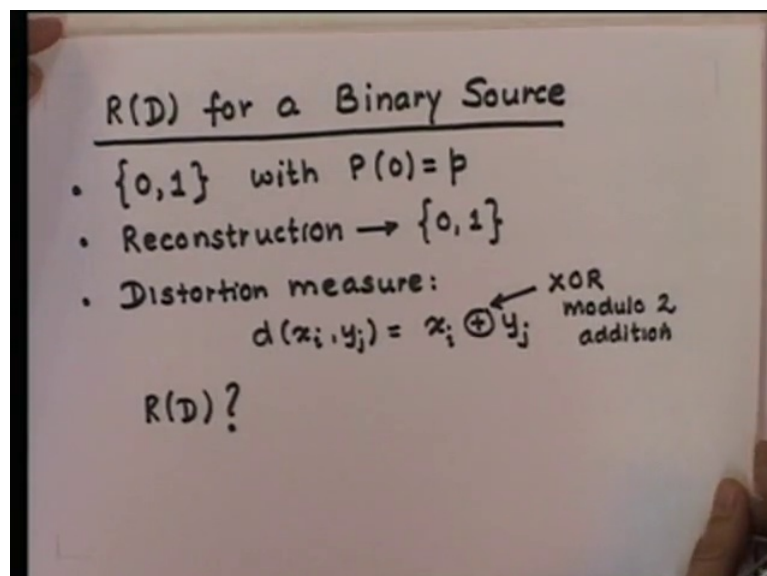
justification for the rate distortion formula. Finally, how do we find the rate distortion function?

(Refer Slide Time: 14:55)



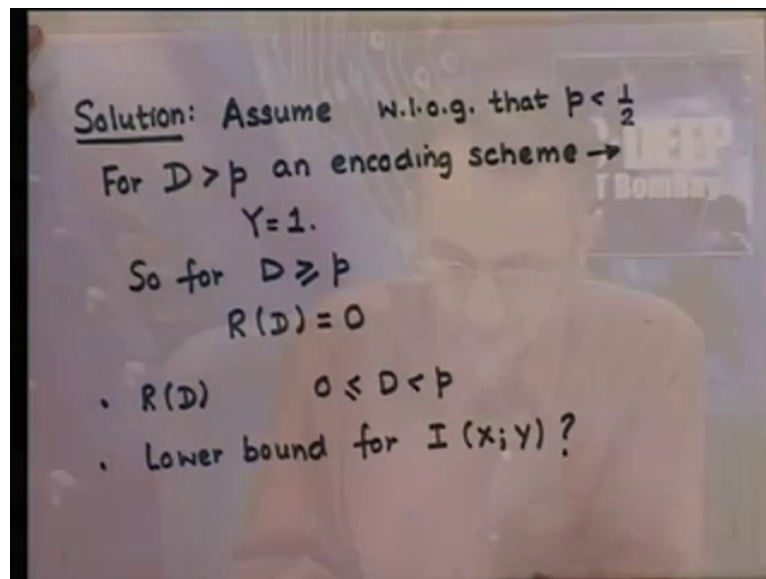
There are two ways to compute the rate distortion function, one approach is basically a computational approach, the algorithm is relatively simple, we will have a look at this algorithm little later. The other approach is based on finding a lower bound for a average mutual information and then show that we can achieve this bound. We will use this approach to find the rate distortion functions for two important sources.

(Refer Slide Time: 15:40)



So, the first important source is a binary source, for which we will calculate the rate distortion function as follows. Suppose we have a source alphabet consisting of 0, 1 with probability of 0 is equal to small p . Let us assume that the reconstruction alphabet is also binary, we assume that the distortion measure is based on the hamming distance that is d of x_i, y_j is equal to x_i modulo two addition or exclusive or operation between x_i and y_j . Given this let us find the rate distortion function for this source, the solution to this problem follows.

(Refer Slide Time: 17:41)

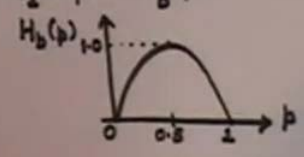


Assume without loss of generality that p is less than $\frac{1}{2}$, so for D that is the specified distortion level greater than p an encoding scheme that would satisfy the distortion criterion would be not to transmit anything and fix Y equal to 1. So, for D greater than equal to p we can say $R(D)$ is equal to 0. Now, we will find the rate distortion function for the distortion range between 0 and less than p . In order to do this let us first find the lower bound for the average mutual information.

(Refer Slide Time: 19:33)

$$\begin{aligned} I(x; y) &= H(x) - H(x|y) \\ &= H(x) - H(x \oplus y | y) \\ (x \oplus y \oplus y &= x) \\ &\geq H(x) - H(x \oplus y) \end{aligned}$$

$\because H(x|y) \leq H(x)$

$$H(x) = -p \log_2 p - (1-p) \log_2 (1-p) = H_b(p)$$
$$H_b(p) = H_b(1-p)$$


Average mutual information is equal to entropy of the source minus conditional entropy of the source given reconstruction, this can be rewritten as entropy of source minus conditional entropy of $X \oplus Y$ given Y . Now, we have used the fact in writing this step, that if we know Y , then knowing X we can obtain $X \oplus Y$ and vice versa as $X \oplus (X \oplus Y) \oplus Y$ is equal to X . Now, this can be written as entropy of X minus entropy of $X \oplus Y$, from here to here we use the relationship that the conditional entropy is always less than the entropy.

From this step to this step we use the fact that conditioning always reduce the entropy that is because H of X given Y is always less than equal to H of X . Now, H of X is equal to minus $p \log$ to the base 2 p minus $(1-p) \log$ to the base 2 of $1-p$ and it is indicated as $H_b(p)$. This is a binary entropy function which we had seen earlier and it is plotted as shown here. Note that $H_b(p)$ is equal to $H_b(1-p)$. So, given that H_X is completely specified by the source probabilities.

(Refer Slide Time: 22:52)

$\{P(y_j|x_i)\}$ such that $H(X \oplus Y) \rightarrow \text{maximized}$
 $E[d(x_i, y_j)] \leq D$ $H_b(P(X \oplus Y)=1)$
 $P(X \oplus Y=1) = P(X=0, Y=1) + P(X=1, Y=0)$
 \hookrightarrow as close as possible to $1/2$

Our task is now to find the conditional probabilities such that entropy of $X \oplus Y$ is maximized while the average distortion is less than the specified distortion level. Now, entropy $H(X \oplus Y)$ is simply the binary entropy function H_b probability of $X \oplus Y$ is equal to 1, where probability of $X \oplus Y$ equal to 1 is equal to joint probability of X is equal to 1, Y equal to 1 plus joint probability of X is equal to 1, Y equal to 0. Therefore to maximize this quantity we would like the probability of $X \oplus Y$ equal to 1 to be as close as possible to half based on this plot.

(Refer Slide Time: 25:05)

$P(X \oplus Y) \rightarrow$ satisfy the distortion constraint
 $E[d(x_i, y_j)] = 0 \times P(X=0, Y=0) + 1 \times P(X=0, Y=1)$
 $\quad + 1 \times P(X=1, Y=0) + 0 \times P(X=1, Y=1)$
 $= P(X=0, Y=1) + P(X=1, Y=0)$
 $= P(Y=1|X=0)p + P(Y=0|X=1)(1-p)$
 $X \oplus Y = 1$
 $P(X \oplus Y = 1) \rightarrow D \quad \therefore E[d(x_i, y_j)] \leq D$

However, the selection of probability of $X \oplus Y$ has to also satisfy the distortion criterion. The distortion is given by 0 multiplied by probability of X equal to 0, Y equal to 0 plus 1 into probability of X equal to 0, Y equal to 1 plus 1 into probability of X is equal to 1, Y equal to 0 plus 0 multiplied by the probability of X equal to 1 and Y equal to 1. This can be written as probability of X equal to 0, Y equal to 1 plus probability of X equal to 1, Y equal to 0.

Which can be rewritten as probability of Y equal to 1 given X equal to 0 multiplied by probability of X equals to 0 plus probability of Y equal to 0 given X is equal to 1 multiplied by probability of X equal to 1 that is $1 - p$, but this is simply the probability that $X \oplus Y$ is equal to 1. Therefore, the maximum value the probability of $X \oplus Y$ equal to 1 can have is D , because average distortion is less than equal to D .

(Refer Slide Time: 27:48)

Handwritten mathematical derivation on a whiteboard:

$$D < p \text{ and } p \leq \frac{1}{2}$$

$$D < \frac{1}{2}$$

$$\therefore P(X \oplus Y = 1) \rightarrow \frac{1}{2} \rightarrow \leq D$$

$$P(X \oplus Y = 1) = D$$

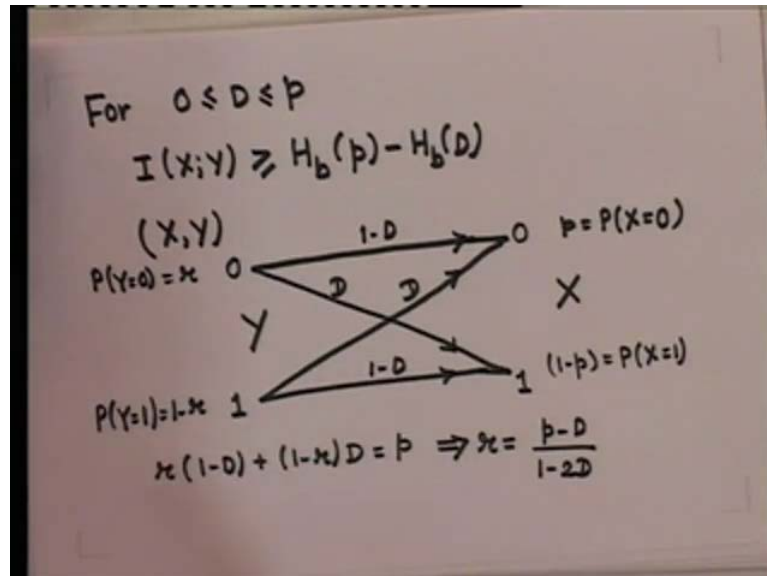
$$\therefore I(X; Y) \geq H_b(p) - H_b(D)$$

$$I(X; Y) = R(D)$$

Now, our assumptions were that D is less than p and p is less than or equal to half which means that D is less than half, therefore the probability of $X \oplus Y$ equal to 1 is closest to half, while being less than or equal to D when $X \oplus Y$ is equal to 1 is equal to D . Therefore, we get mutual information less or equal to binary function $H_b(p)$ minus binary function $H_b(D)$. Now, we will show that this lower bound is actually the rate distortion function and can be achieved by finding joint distribution or conditional

probability distribution for that given source distribution that meets the distortion criterion, and has mutual information equal to the rate distortion function.

(Refer Slide Time: 29:45)



Now, for specified distortion levels between 0 and p we can achieve the minimum value of average mutual information in this relationship by choosing X Y to have the joint distribution given by the binary symmetric channel as shown here. Let us assume that the probability of Y equal to 0 equal to r and probability of Y equal to 1 is equal to 1 minus r . So, we choose the distribution of Y at the input of this channel so the output distribution of X is the specified distribution, that is p and 1 minus p . So, if we assume r to be the probability of Y equal to 0, then choosing r as follows. So, we get the value of r given by this expression, now if D is less than or equal to p and which is less than equal to $1/2$, then probability of Y equal to 0 is greater than equal to 0 and probability of Y is equal to 0 is also greater than or equal to 0.

(Refer Slide Time: 32:40)

$$I(X; Y) = H_b(p) - H_b(D)$$

∴ for $D < p$ and $p \leq \frac{1}{2}$

$$R(D) = H_b(p) - H_b(D)$$

$p > \frac{1}{2}$ p and $(1-p)$

We can have mutual information equal to $H_b(p) - H_b(D)$, therefore for D less than p and p less than equal to half. The rate distortion function is equal to binary entropy function $H_b(p) - H_b(D)$. Finally, if p is greater than half then we simply switch the roles of p and $1 - p$.

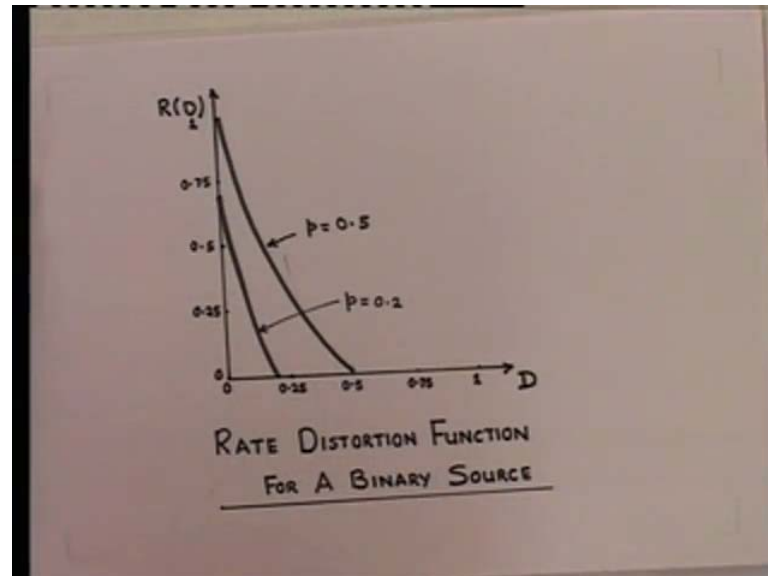
(Refer Slide Time: 33:51)

$$R(D) = \begin{cases} H_b(p) - H_b(D) & \text{for } 0 \leq D \leq \min\{p, 1-p\} \\ 0 & \text{for } D > \min\{p, 1-p\} \end{cases}$$

So, putting all this together the rate distortion function for a binary source is binary entropy function as function of p , and binary entropy function as the function of

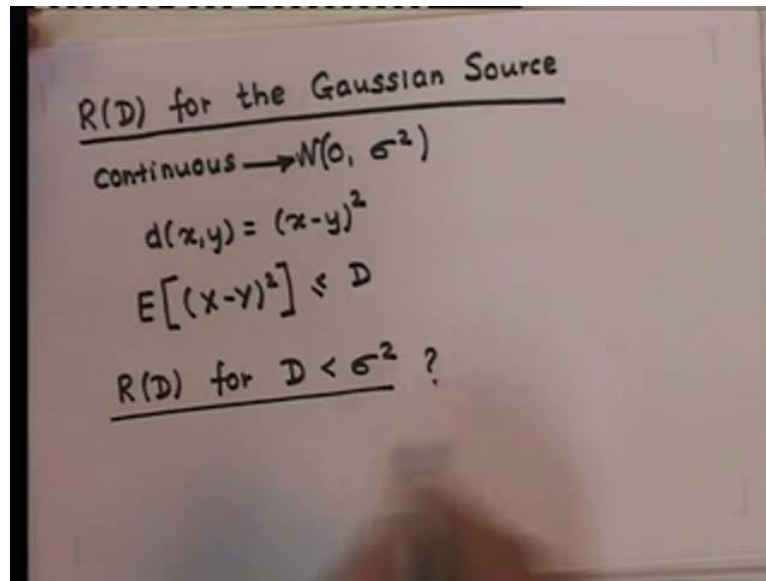
specified distortion level D for is equal to 0 for D greater than minimum of p 1 minus p .
Now, this function is illustrated in the figure.

(Refer Slide Time: 34:42)



So, it is monotonically decreasing function for p equal to 0.5 the maximum value of distortion for 0 rate is equal to 0.5 and maximum value of the rate for 0 distortion is given by the entropy of the source of p equal to 0.5 which is equal to 1. For p equal to 0.2 the maximum value of the distortion is 0.2 for the rate equal to 0 and for distortion equal to 0 the maximum value of the rate is given by the entropy of the source p equal to 0.2, so having calculated the rate distortion function for a binary source. Let us proceed to calculate the rate distortion function for another important source that is Gaussian source.

(Refer Slide Time: 35:58)



$R(D)$ for the Gaussian Source
Continuous $\rightarrow N(0, \sigma^2)$
 $d(x, y) = (x - y)^2$
 $E[(x - y)^2] \leq D$
 $R(D)$ for $D < \sigma^2$?

Again we will follow the same approach, we will write down the average mutual information for a Gaussian source and show that the lower bound for the mutual information is achievable. So, we have a continuous amplitude source that has a 0 mean and Gaussian P D F with variance equal to sigma squared. So, we have normal distribution that is Gaussian distribution with 0 mean and variance is equal to sigma squared. We assume that our distortion measure is again square error measure, so our distortion constrain is given by average square error less than equal to some specified distortion level D. So, first we calculate the rate distortion function for specified distortion level less than sigma squared.

(Refer Slide Time: 38:07)

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(X) - h(X-Y|Y) \\ &\geq h(X) - \underbrace{h(X-Y)}_{\text{Gaussian}} \quad E[(X-Y)^2] \leq D \\ &\quad E[(X-Y)^2] = D \end{aligned}$$

Mutual information equal to differential entropy minus conditional entropy of X given Y this is equal to differential entropy minus conditional entropy of X minus Y given Y which is greater than equal to h of X minus h of X minus Y, from the second step to the third step, the reason is the conditioning always reduces entropy. So, in order to minimize the right hand side of this equation, we have to maximize the second term, subject to the constrain given by mean square error to be less than some specified distortion level D. Now, the second term is maximized if X minus Y is Gaussian based on our earlier study and the constrain can be satisfied if expectation of X minus Y squared is equal to D.

(Refer Slide Time: 39:57)

$$\begin{aligned} \therefore h(X-Y) &\rightarrow \text{Gaussian R.V with } D \\ I(X;Y) &\geq \frac{1}{2} \log(2\pi e \sigma^2) - \frac{1}{2} \log(2\pi e D) \\ &= \frac{1}{2} \log \frac{\sigma^2}{D} \\ Y &\text{ is zero mean Gaussian } \sigma^2 - D \end{aligned}$$

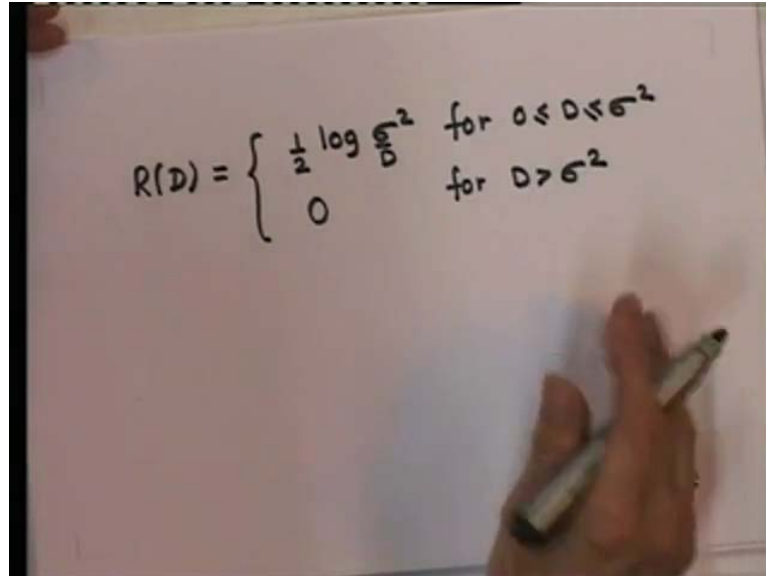
Therefore, h of X minus Y is the differential entropy of Gaussian random variable with variance D and the lower bound becomes I of X semicolon Y greater than equal to halflog 2 pie e sigma squared minus halflog 2 pie e D equal to halflog sigma squared by D . Now, this average mutual information can be achieved if Y is 0 mean Gaussian with variance sigma squared minus D .

(Refer Slide Time: 41:27)

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{1}{\sqrt{2\pi D}} \exp\left(-\frac{x^2}{2D}\right) \\ D > \sigma^2, &\text{ if we set } Y=0, \text{ then} \\ I(X;Y) &= 0 \\ E[(X-Y)^2] &= \sigma^2 < D \end{aligned}$$

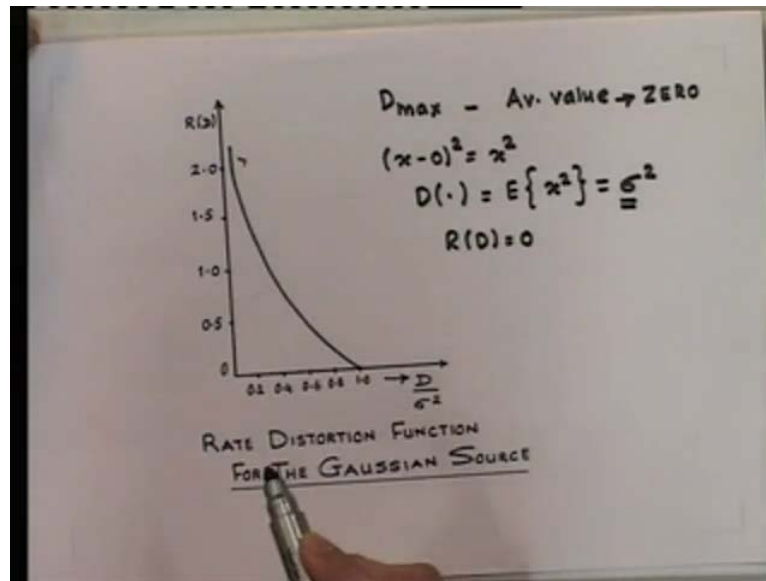
Conditional P D F is given as follows, now for D greater than sigma squared if we set Y equal to 0, then mutual information between X and Y is equal to 0 and the mean squared error is equal to sigma squared, which is less than D.

(Refer Slide Time: 42:34)

A photograph of a hand holding a white marker, writing the rate distortion function R(D) on a whiteboard. The equation is written as a piecewise function: R(D) = { 1/2 log(sigma^2/D) for 0 <= D <= sigma^2, 0 for D > sigma^2. The handwriting is in black ink on a white background.
$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D} & \text{for } 0 \leq D \leq \sigma^2 \\ 0 & \text{for } D > \sigma^2 \end{cases}$$

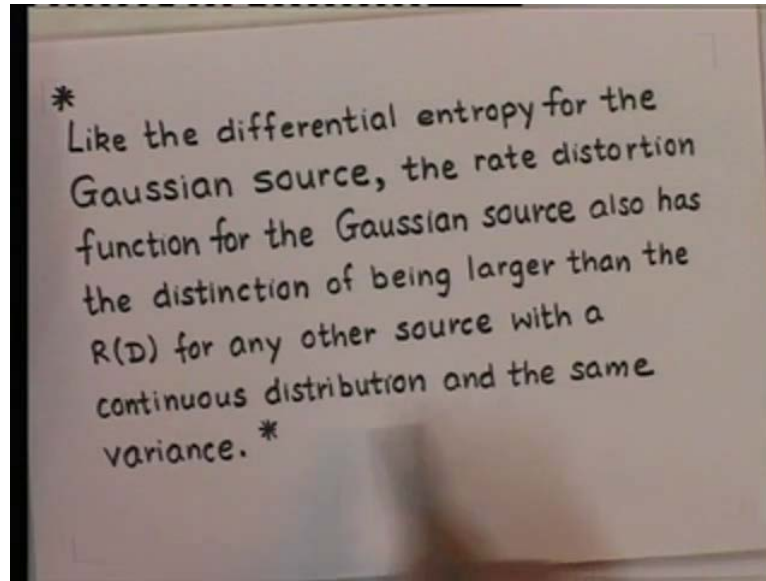
Therefore, to summarize we can say the rate distortion function for the Gaussian source can be written as half log sigma squared by D for D less than equal to sigma squared, and is equal to 0 for D greater than sigma square. This relationship is illustrated in figure, again we see that the rate distortion function is monotonically decreasing function with increasing D.

(Refer Slide Time: 43:24)



Now, we can find out the maximum allowable distortion that is D_{\max} as follows, if no information is transmitted the best choice with the receiver can make consist of choosing the average value, which in this case is equal to 0. Because this has the largest probability density and is therefore the most probable transmitted value. The error which is now made is x minus 0 squared is equal to x squared and the average error is equal to expectation of x squared which is equal to σ^2 . From this it then follows that D_{\max} which is the minimum distortion for $R(D) = 0$ is equal to σ^2 . We further see that $R(D)$ increases to infinity if D is made smaller and smaller, this is of course a mathematical abstraction in practical systems D will be bounded by the smallest measurable signal value.

(Refer Slide Time: 45:16)



So, in conclusion we can say that, like the differential entropy for the Gaussian source, the rate distortion function for the Gaussian source also, has the distinction of being larger than the rate distortion function for any other source with a continuous distribution and the same variance. Now, this result is specially very valuable because for many sources it is very difficult to calculate the rate distortion function. In this situation it is helpful to have the upper bound on the rate distortion function. Now, it would be also nice to have a lower bound on the rate distortion function for a continuous random variable. We will investigate this case in the next class, and will also look at the computation approach to calculate the rate distortion function.