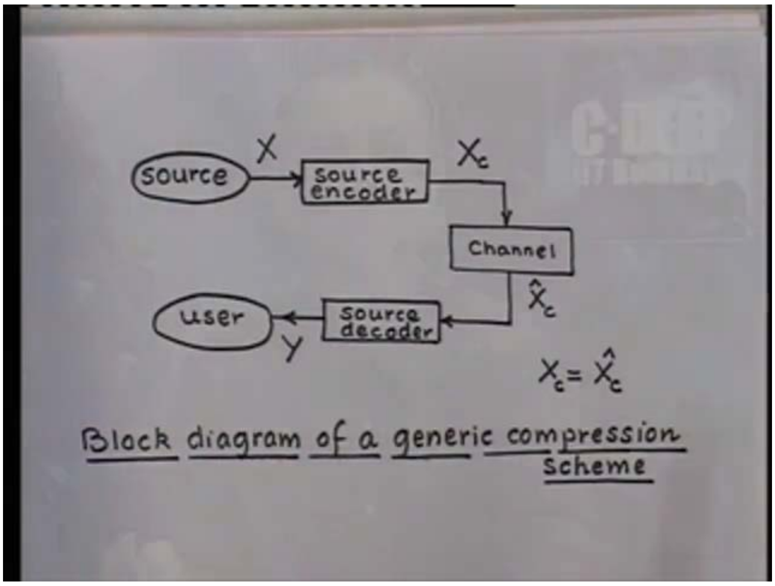
Information Theory and Coding
Prof. S. N. Merchant
Electrical Engineering
Indian Institute of Technology, Bombay

Lecture - 31
Introduction To Rate-Distortion Theory

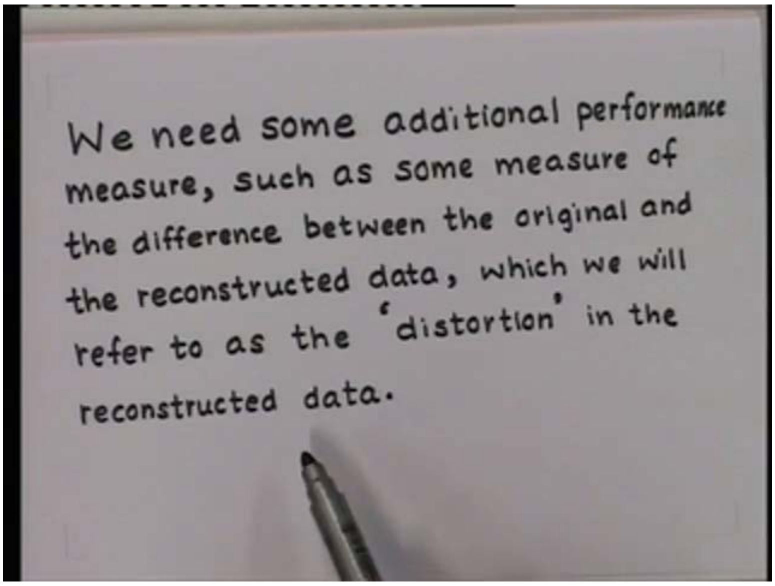
When we speak of a compression algorithm or compression technique, we are actually referring to two algorithms. There is the compression algorithm that takes an input and generates the representation that requires fewer bits and there is the reconstruction algorithm that operates on the compressed representation to generate the reconstruction. These operations are represented in the figure.

(Refer Slide Time: 01:51)



The output of the source is modelled as random variable X, the source coder takes the source output and produces the compressed representation X c. The channel block represents all transformation. The compressed representation undergoes before the source is reconstructed. Usually, we will take the channel to be the identity mapping which means X c is equal to X c cap. The source decoder takes the compressed representation and produces a reconstruction of the source output Y for the user. Now, based on the requirements of reconstruction, data compression algorithms can be divided into two broad classes.

(Refer Slide Time: 03:00)



Lossless compression schemes in which the reconstruction is identical to the original, and lossy compression schemes which generally provide much higher compression than lossless compression, but allow the reconstruction to be different from the original. In our study so far we have investigated source encoding algorithms which were of lossless type. When you are looking at lossless compression algorithms, one thing we never worried about was how the reconstruction sequence would differ from the original sequence.

By definition the reconstruction of a loosely constructed sequence is identical to the original sequence, but the amount of compression available using lossless compression scheme is limited. There is a hard floor defined by the entropy of the source below which we cannot drive the size of the compressed sequence. So, as long as we wish to preserve all the information in the source, the entropy of the source like the speed of light is a fundamental limit.

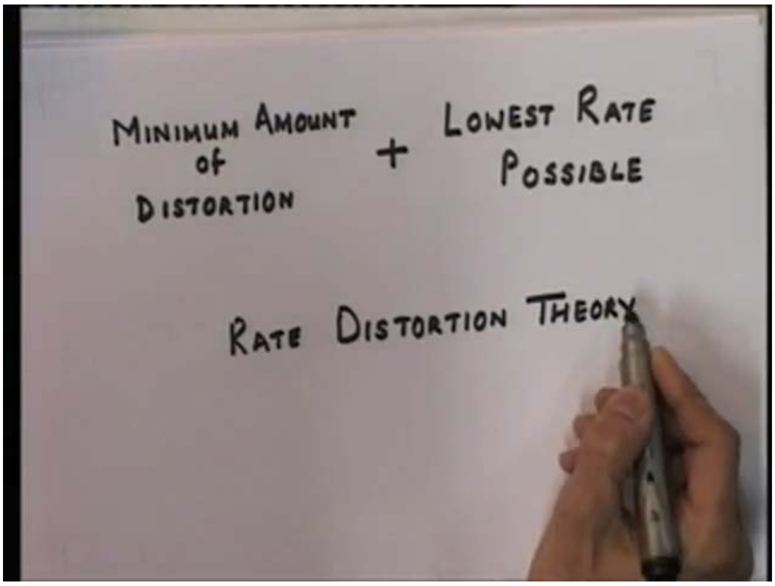
The storage or transmission resources available to us may be sufficient to handle the data requirements of a lossless compression scheme or the consequences of a loss in the information may be much more expensive than the cost of additional storage and on transmission resources. This would be the case for example, when we are concerned with the storage and archiving of bank records. An error in the bank record would be much more expensive than the cost of additional storage media.

Now, if neither of these conditions hold that is resources are limited and we are not concerned with the absolute integrity of the data, then we can achieve a higher amount of compression by using lossy compression schemes. Now, we require some kind of a performance measure to determine the efficiency of a lossy compression schemes. For the lossless compression scheme essentially we use the rate as the criterion for finding out the efficiency of the lossless compression scheme.

Now, this is not visible for a lossy compression scheme, if the rate were the only criterion for lossy compression scheme, where the loss of information is permitted. Then the best lossy compression scheme would be the one where we simply throw away all the data. Therefore, we need some additional performance measure, such as some measure of the difference between the original and the reconstructed data which we will refer to as the distortion in the reconstructed data.

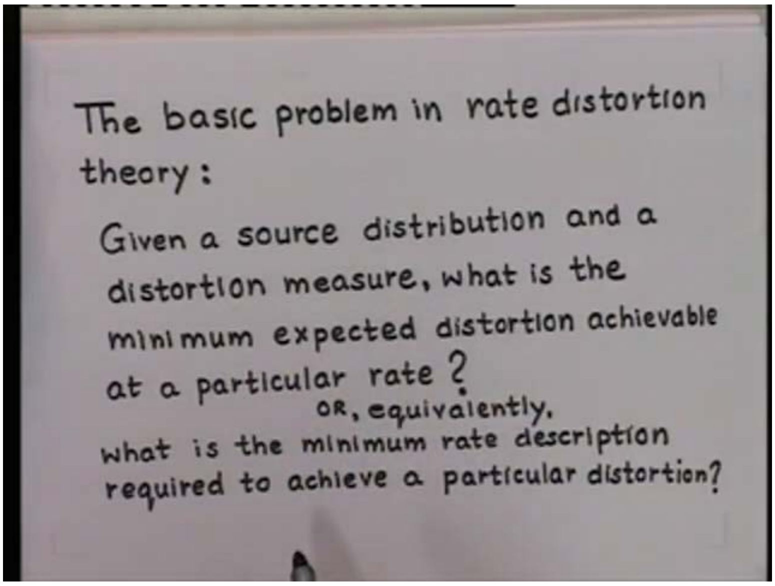
So, in the best of all possible worlds we would like to incur the minimum amount of distortion while compressing to the lowest rate possible. Obviously, there is a trade-off between minimizing the rate and keeping the distortion small. The extreme cases are when we transmit no information in which case the rate is 0 or keep all the information in which case the distortion is 0, and the rate is determined by the entropy for a discreet source. Now, the study of the situations between this two extreme is called rate distortion theory.

(Refer Slide Time: 08:16)



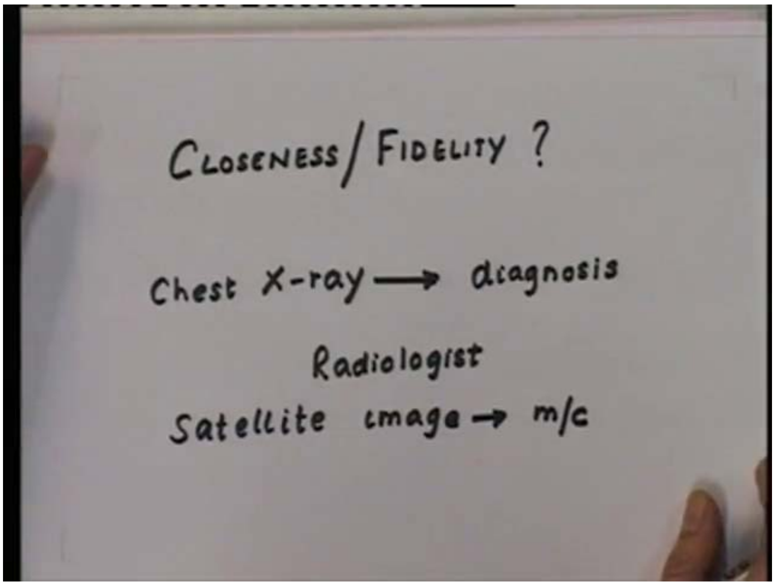
So, the basic problem in rate distortion theory can be stated as follows.

(Refer Slide Time: 08:31)



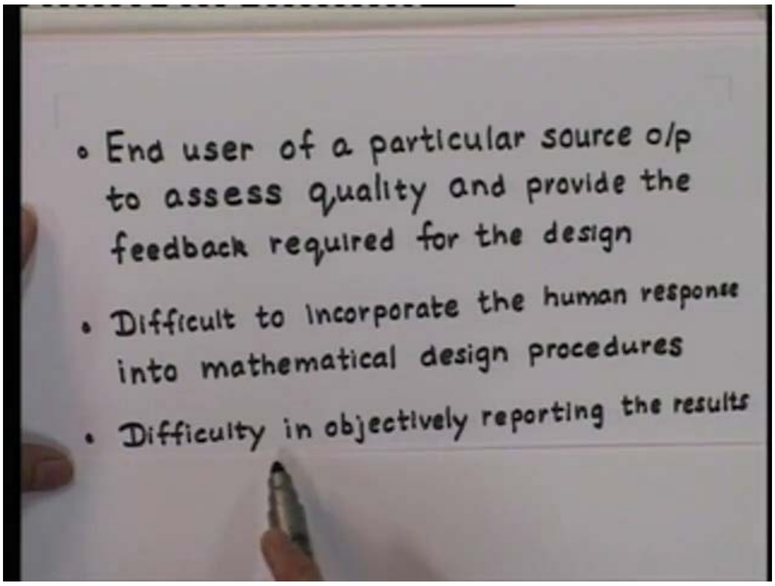
Given a source distribution and a distortion measure, what is the minimum expected distortion achievable at a particular rate? This same problem can be equivalently posed as, what is the minimum rate description required to achieve a particular distortion?

(Refer Slide Time: 09:12)



So, now the next logical question is, how do we measure the closeness of fidelity of a reconstructed source sequence to the original? Now, the answer to this question depends on what is being compressed and who is doing the answering. Suppose, we were to compress and reconstruct images and if the image were digitised, chest x-ray and the resulting reconstruction is to be used for diagnosis. Then the best way to find out how much distortion was introduced and in what manner is to ask a radiologist. But if the image was say a satellite image and this image is to be processed by a machine to obtain information about the objects in the image. Then the best measure of fidelity is to see how the introduced distortion affects the functioning of the machine.

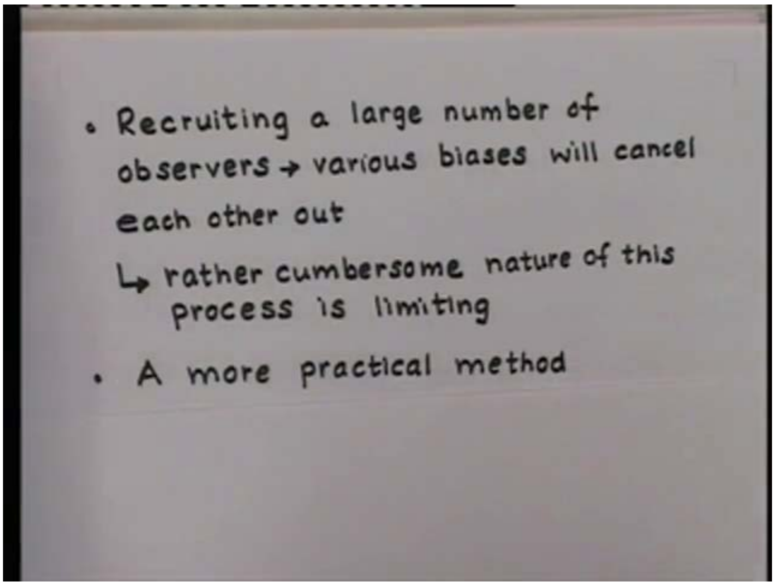
(Refer Slide Time: 11:00)



So, what it implies that in the best of all worlds, we would always use the end user of a particular source output to assess quality and provide the feedback required for the design. Now, in practice this is not often possible, specially when the end user is a human, because it is difficult to incorporate the human response into mathematical design procedures and also there is difficulty in objectively reporting the results. The people who are asked to assess one persons design may be more easy going, then the people who are asked to assessed another person’s design.

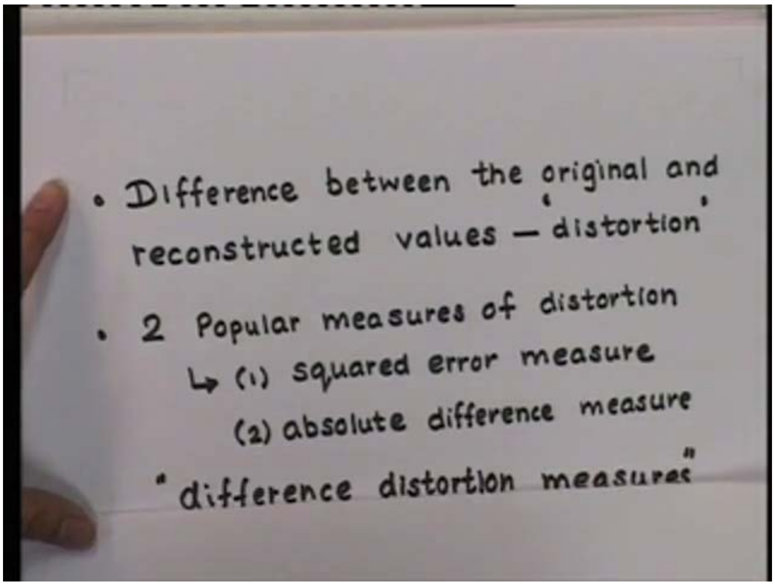
In such a case it is possible that even though the reconstructed output of one persons design is rated as excellent. The reconstruction output of another person’s design is rated as acceptable, switching over the observer may change this ratings. So, a solution to this problem would be to recruit a large number of observers.

(Refer Slide Time: 12:31)



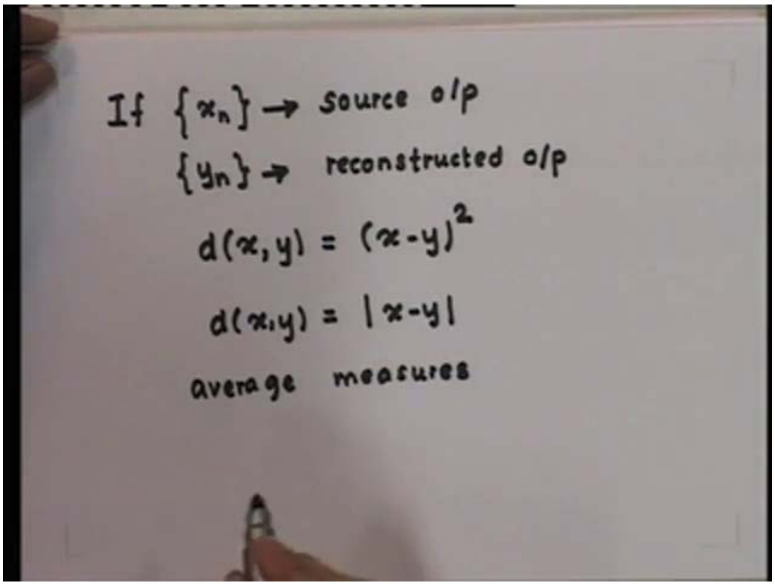
Recruiting a large number of observers in the hope, that this various biases will cancel each other out. Now, this is often the option used specially in the final stages of the design of compression system. However, the rather cumbersome nature of this process is limiting and therefore we generally need a more practical method for looking at how close the reconstruction signal is to the original.

(Refer Slide Time: 13:21)



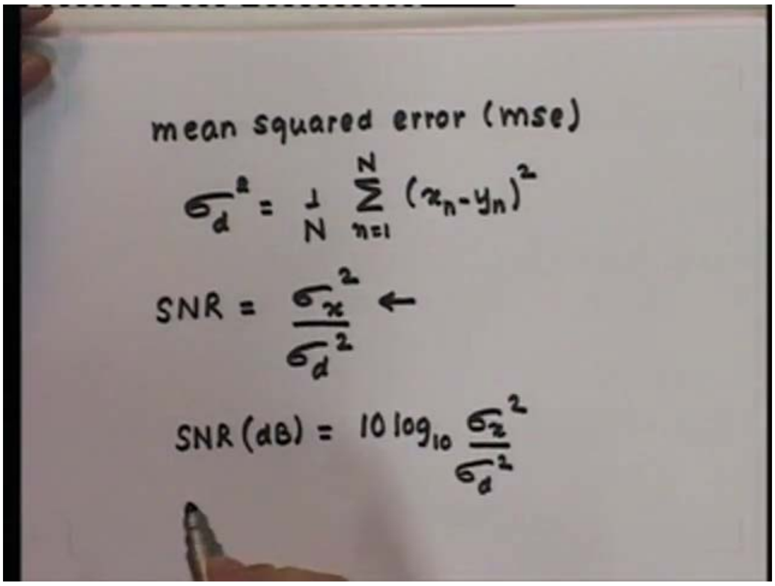
Natural thing to do when looking at the fidelity of the reconstructed sequence is to look at the difference, difference between the original and reconstructed values in other words, the distortion introduce in the compression process. Now, there are two popular majors of distortion or differences between the original and reconstructed sequence and this are squared error measure and absolute difference measure, both of this are called difference distortion measures.

(Refer Slide Time: 14:04)



So, if we denote x n as the source output and y n as the reconstructed output, then the squared error measure is given by d of x, y is equal to x minus y squared and the absolute difference measure is given by mode of difference between x and y. Now, in general it is difficult to examine the difference on a term by term basis. Therefore, a number of average measures are used to summarise the information in the difference sequence. The most often used average measure is the average of the squared error measure and this is called the mean squared error.

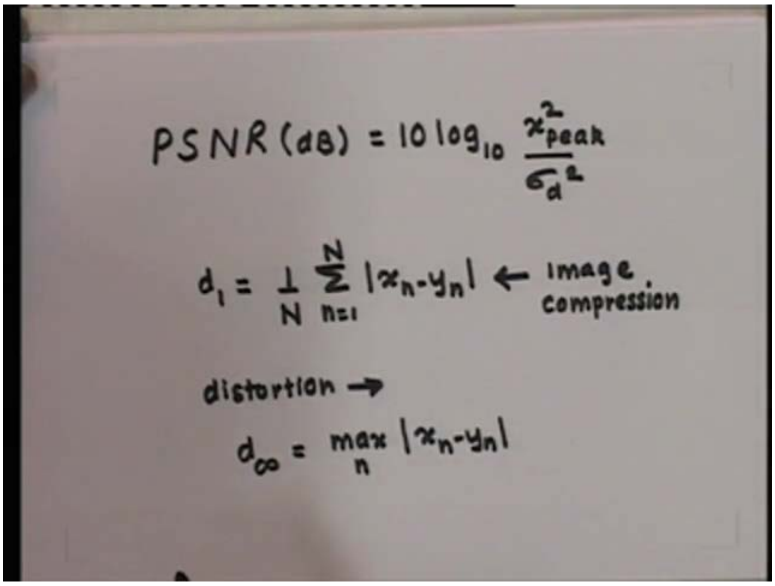
(Refer Slide Time: 15:39)



This is often represented by the symbol sigma d squared is equal to 1 by n summation of the difference x n minus y n squared sum over the length of the sequence. Now, if you are interested in the size of the error relative to the signal, then we can find the ratio of the average squared value of the source output and the mean squared error. This is called the signal to noise ratio and denoted as sigma x squared divided by sigma d squared, where sigma x squared is the average squared value of the source output or signal and sigma d squared is the mean squared error.

Now, the signal to noise ratio is often measured on a logarithmic scale and the units of measurement are decibel. So, it is abbreviated as d B and written as 10 log to the base 10 of sigma x squared over sigma d squared. Now, sometimes we are more interested in the size of the error relative to the peak value of the signal, then with the size of the error relative to the average squared value of the signal.

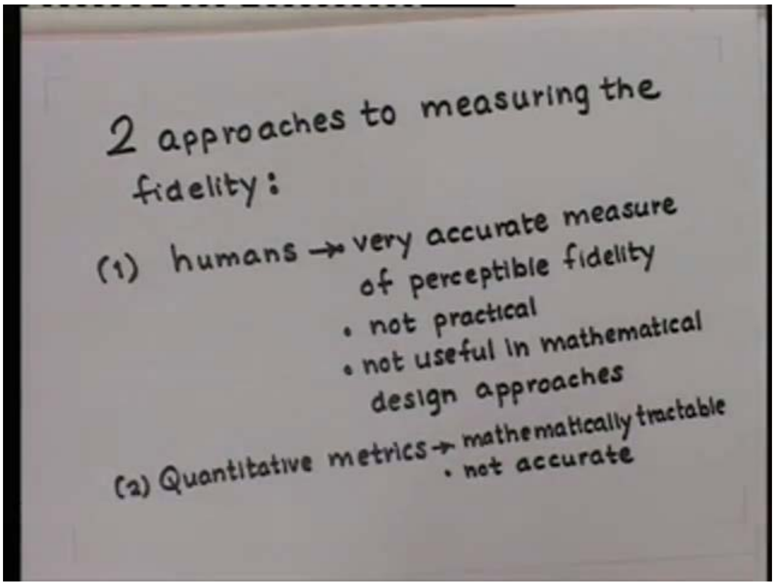
(Refer Slide Time: 17:44)



In such a case this ratio is called as peak signal to noise ratio in d B is equal to 10 log to the base 10 of x peak squared divided by sigma d squared. Now, another difference distortion measure that is used quite often although not as often as the mean squared error is the average of the absolute difference and denoted as d 1 is equal to 1 by n summation of the absolute difference between the two sequence. Now, this measure seems specially useful for evaluating image compression algorithms.

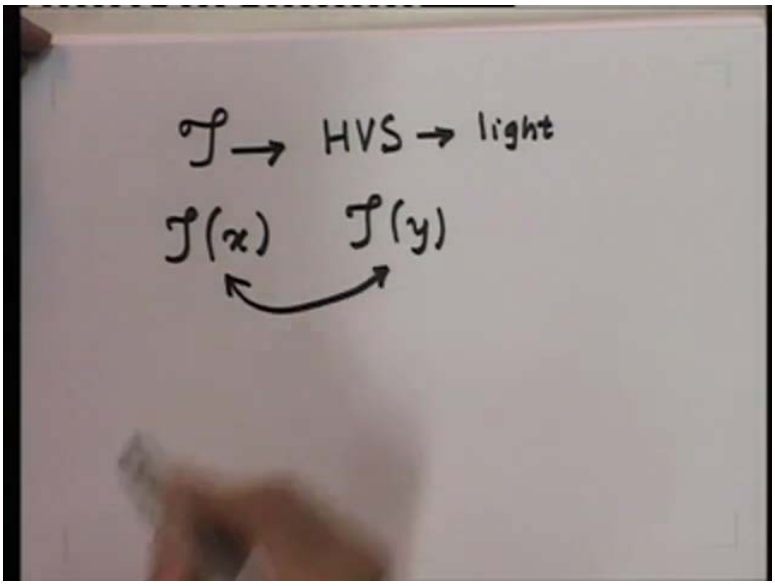
Now, in some applications the distortion is not perceptible as long as it is below some threshold. So, in this situation we might be interested in the maximum value of the error magnitude and we define another distortion measure as d infinity is equal to maximum difference over the sequence. So, we have looked at two approaches to measuring the fidelity of a reconstruction.

(Refer Slide Time: 19:42)



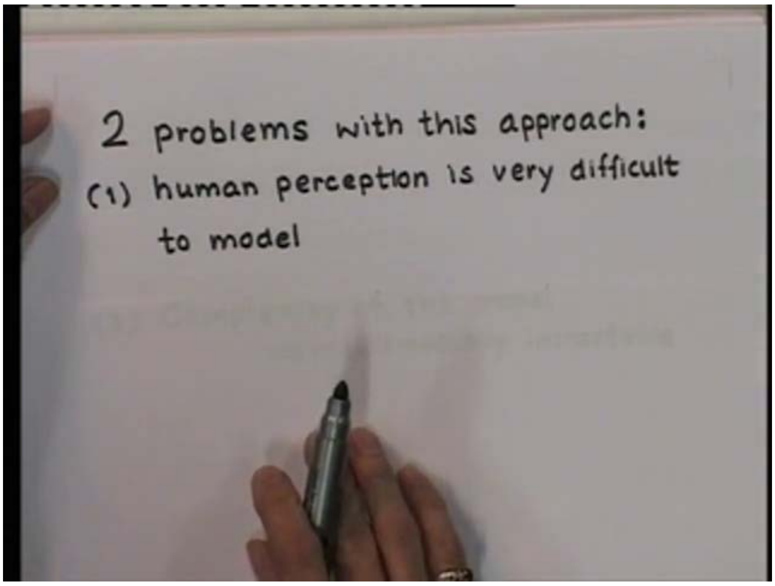
The first method involves human, this is very accurate measure of perceptible fidelity, but it is not practical and not useful in mathematical design approaches. The second approach is based on quantitative metrics, which is mathematical tractable, but it usually does not provide a very accurate indication of the perceptible fidelity of the reconstruction. Now, in these cases, now a middle ground would be to find a mathematical model for human perception and then transform both the source output and the reconstructed output to this perceptual space and then measure the difference between the two in this perceptual space.

(Refer Slide Time: 20:53)



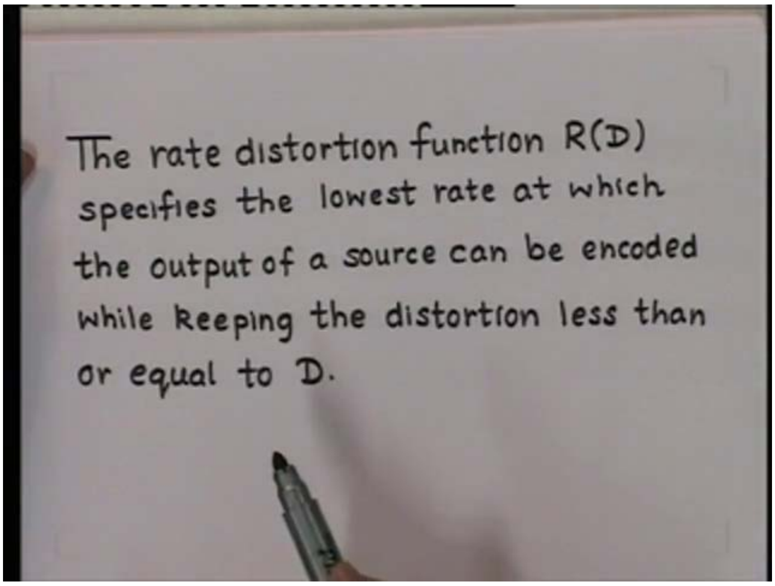
So, for example if we could find a transformation T, that represented the actions performed by the human visual system on the light intensity impinging on the retina, before it is perceived by the cortex. Then we could find the transformation on x, transformation on the reconstructed output y and then examine the difference between this two.

(Refer Slide Time: 21:37)



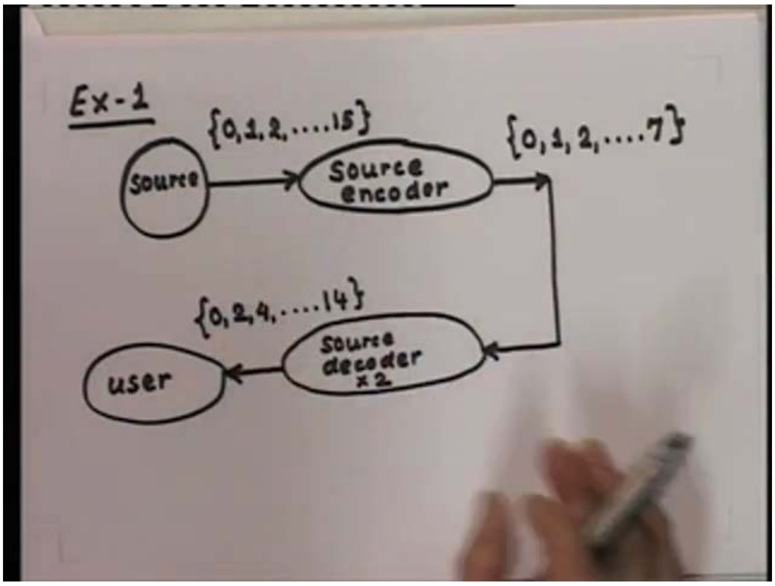
Now, there are two problems associated with this approach. First the process of human perception is very difficult to model and accurate models of perception are yet to be discovered. Second even if you could find a mathematical model for perception, the odds are there it would be so complex, that it would be mathematically intractable. Now, having looked at some of the distortion measures, let us return back to our study of rate distortion theory. As defined earlier, rate distortion theory is concerned with the trade-off between the distortion and the rate in the lossy compression scheme. Rate is defined as the average number of bits used to represent the source output, one way of representing this trade-off between the rate and distortion is a rate distortion function, which is defined as follows.

(Refer Slide Time: 22:55)



The rate distortion function given by R of D specifies the lowest rate at which the output of a source can be encoded while keeping the distortion less than or equal to D. Now, before we mathematically define the rate distortion function, let us look at the evaluation of rate and distortion for some different lossy compression schemes.

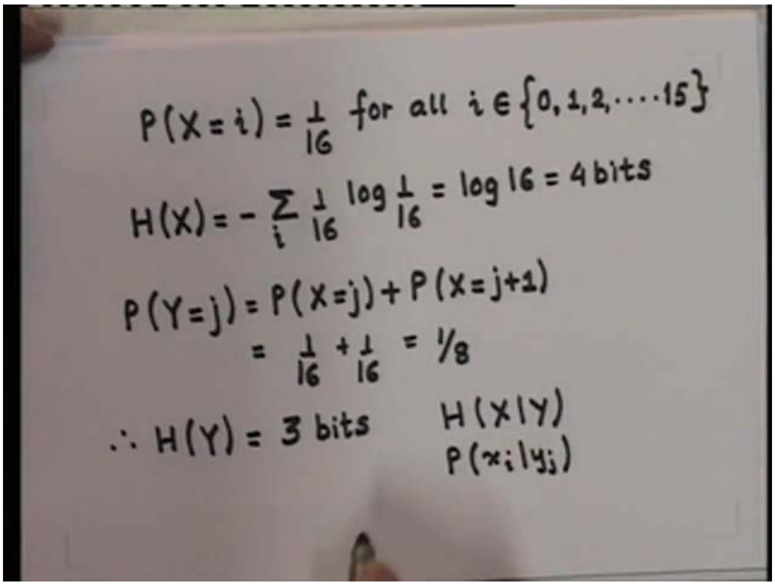
(Refer Slide Time: 23:37)



So, let us consider, example 1, suppose we have a source whose output consists of four bit words 0, 1, 2 up to 15. Now, this is fed to a source encoder, the source encoder encodes each value by shifting out the least significant bit. So, the output alphabet for the source encoder is 0, 1, 2, up to 7, we assume that a channel is an identity mapping. So, at the receiver we have a source decoder and the job of the source decoder is to take this and the job of the source decoder is to reconstruct the output by shifting in a 0 as the L S B or in other words multiplying the source encoder output by 2.

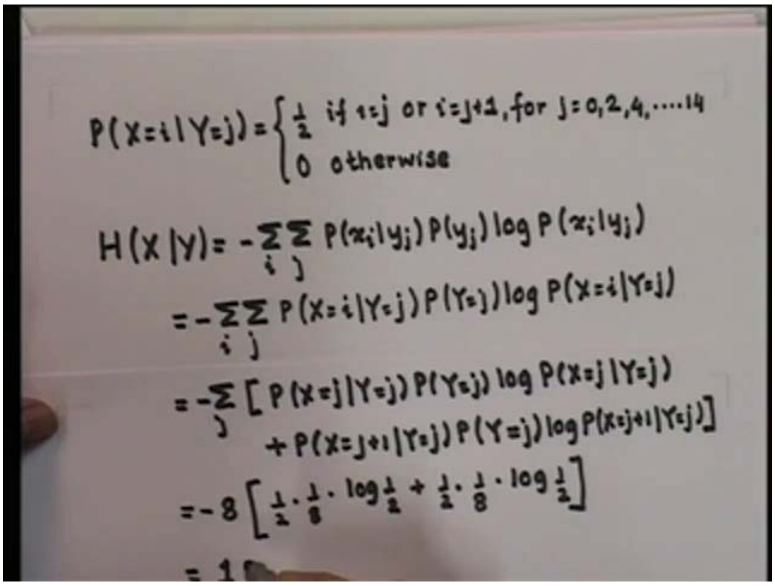
Thus the reconstructed alphabet is 0, 2, 4 up to 14. So, in this case the source alphabet and the reconstruction alphabet are distinct. Now, we need to talk about the information relationship between the random variables that take on values from two different alphabets. So, let us calculate the various entropies for this source and compression scheme.

(Refer Slide Time: 26:37)



Now, as the source outputs are all equally likely probability of X equal to i is equal to 1 by 16 for all i belonging to the set 0, 1, 2 up to 15 and therefore, the entropy for this source is given as follows p x log of 1 by p x to the base 2 is equal to log to the base 2 of 16 is equal to 4 bits. Now, we can calculate the probabilities of the reconstruction alphabet as follows, probability of Y equal to j is equal to probability of X equal to j plus probability of X equal to j plus 1. So, this is equal to 1 by 16 plus 1 by 16 is equal to 1 by 8 and therefore, entropy for the reconstructed output that is H of y is equal to 3 bits. Now, to calculate the conditional entropy that is H of X given Y, we need the conditional probabilities that is P of x i given y j.

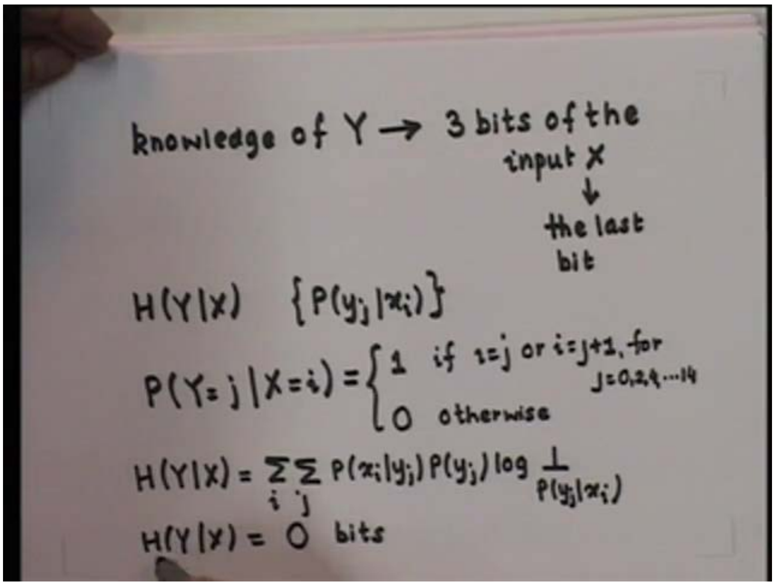
(Refer Slide Time: 29:02)



Now, from our construction of our source encoder, probability of X is equal to i given Y is equal to j is equal to half if i is equal to j or i is equal to j plus 1 for j equal to 0, 1, 4, 14 and it is equal to 0 otherwise. Substituting this in the expression for H of X given Y equal to double summation over i j probability of x i given y j probability of y j log of probability of x i given y j, this is equal to double summation over i j probability of X is equal to i given Y is equal to j multiplied by probability of Y is equal to j log of probability X is equal to i given Y is equal to j.

Probability of X is equal to j given Y is equal to j, probability of Y is equal to j log of probability X is equal to j given Y is equal to j plus another term, which is probability of X is equal to j plus 1 given Y is equal to j multiplied by probability of Y is equal to j log of probability X is equal to j plus 1 given Y is equal to j. This can be simplified as minus 8 times half into 1 8 into log of half plus half times 1 8 times log of half, which is equal to 1. Now, let us compare this answer to what we would have intuitively expected the uncertainty to be based on our knowledge of the compression scheme.

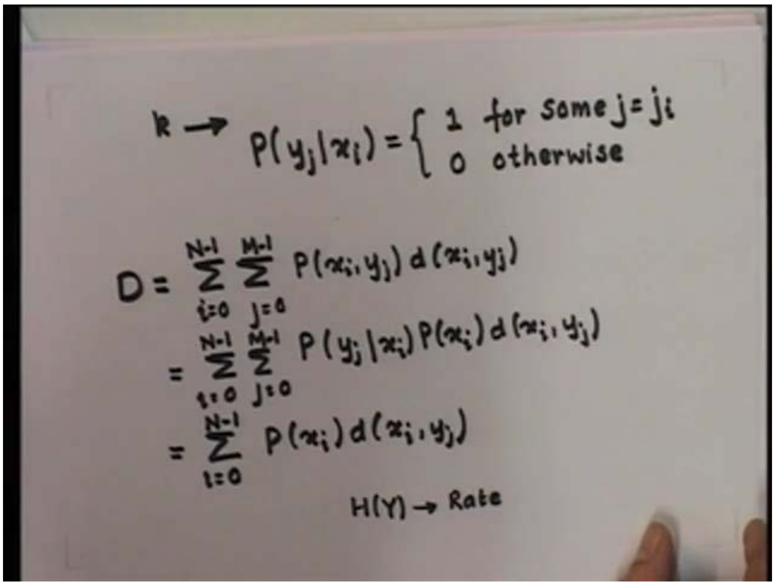
(Refer Slide Time: 32:49)



Now, with the coding scheme describe knowledge of Y means, we know the first 3 bits of the input x. The only thing about the input that we are uncertain about is the value of the last bit, in other words if we know the value of the reconstruction of uncertainty about the source output is 1 bit. Therefore, at least in this case our intuition matches the mathematical definition. Now, to obtain H of Y given X, we need the conditional probabilities, probability of y j given x i now, from our knowledge of the compression scheme we see that, probability of Y equal to j given X is equal to i is equal to 1, if i is equal to j or i is equal to j plus 1 for j equal to 0, 2, 4 up to 14 and this conditional probability is equal to 0 otherwise.

Now, if we substitute this values into the expression for H of Y given X which is equal to double summation i j probability of x i given y j probability of y j log of 1 by probability y j given x i, then we find that H of Y given X is equal to 0 bits. Now, this also makes sense, for the compression scheme described here, if we know the source output we know 4 bits, the first three of which are the reconstruction. Therefore, in this example the knowledge of the source output at a specific time completely specifies the corresponding reconstruction. Now, having looked at the rate for this compression scheme, let us try to evaluate the distortion for this compression scheme.

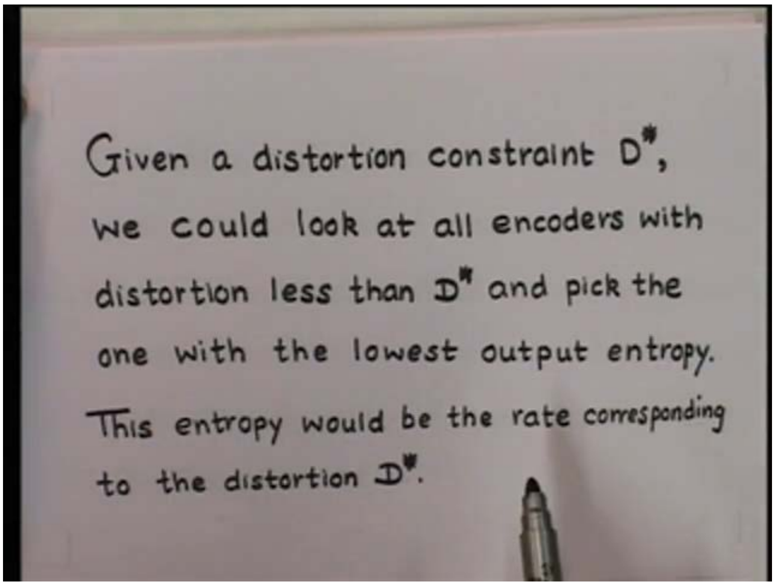
(Refer Slide Time: 36:18)



Now, the knowledge of the value of the input at time k completely specifies the reconstructed value at time k. So, in this situation probability of y j given x i is equal to 1 for some j equal to j i and it is equal to 0 otherwise. Therefore, in this case we can evaluate the distortion D, which is given in general as double summation over i j, probability of x i, y j multiplied by distortion between x i, y j, which can be rewritten as probability of y j given x i multiplied by probability of x i multiplied by distortion measure between x i, y j.

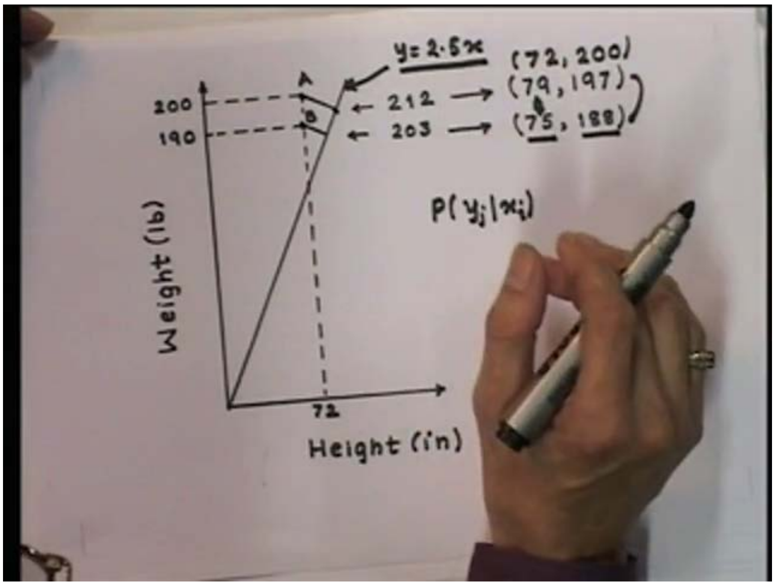
This can be further simplified using this relationship as P of x i multiplied by distortion measure between x i and y j. Now, the rate for this source coder is the output entropy H Y. Now, if this were always the case, then the task of obtaining a rate distortion function would be relatively simple.

(Refer Slide Time: 39:00)



Given a distortion constraint D star, we could look at all encoders with distortion less than D star and pick the one with the lowest output entropy, this entropy would be the rate corresponding to the distortion D star. Now, the requirement that knowledge of the input at time k completely specifies the reconstruction at time k is very restrictive and there are many efficient compression techniques that would have to be excluded under this requirement. So, let us consider another example.

(Refer Slide Time: 39:47)



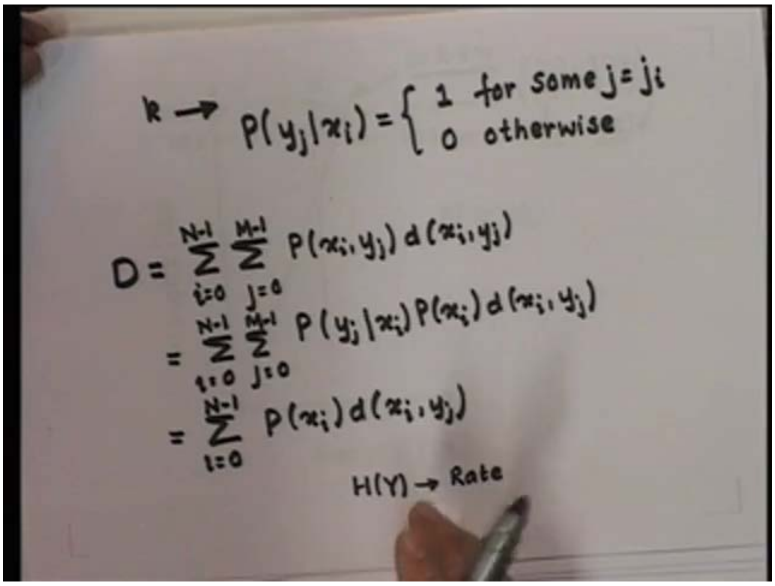
Now, if you have a data sequence that consists of height and weight measurements, it has been observed that height and weight are quiet heavily correlated. So, if we plot the height along the x axis and the weight along the y axis, the data points cluster along the line which is given by y is equal to 2.5 x. Now, in order to take advantage of this correlation, we devise the following compression scheme.

For a given pair of weight and height, we find the orthogonal projection on the y equal to 2.5 x line as shown in the figure here. The point on this line can be represented as the distance to the nearest integer from the original. Thus we encode a pair of values into a single value; at the time of reconstruction we simply map this value back into a pair of height and weight measurements.

For instance, suppose somebody is 72 inches tall and 200 pounds in weight that is corresponding to point A in the figure. Now, this corresponds if we take the projection on the line y is equal to 2.5 x, then this corresponds to a point at a distance of 212 along this line. The reconstructed values of the height and weight corresponding to this value are obtained by projecting this point on the x axis and y axis. So, for this we get as 79 and 197, notice that the reconstructed values differ from the original values.

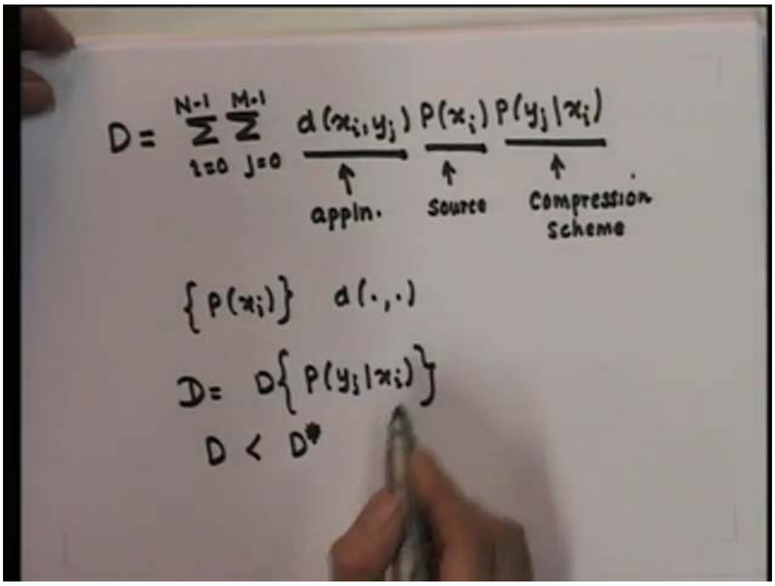
Suppose, we now have another individual who is also 72 inches tall, but weighs 190 points that is corresponding to point B in the figure. The source coder output for this pair would be 203 and the reconstructed values for height and weight are 75 and 188 respectively. Notice that while the height value in both cases was the same, the reconstructed value is different. The reason for this is that, the reconstructed value for the height depends on the weight. Now, for this particular source coder, we do not have conditional probability density function of the form shown earlier.

(Refer Slide Time: 43:31)



Therefore, in this case the calculation of distortion is not the same as what we had done in the earlier example. So, let us examine the distortion for this scheme a little more closely.

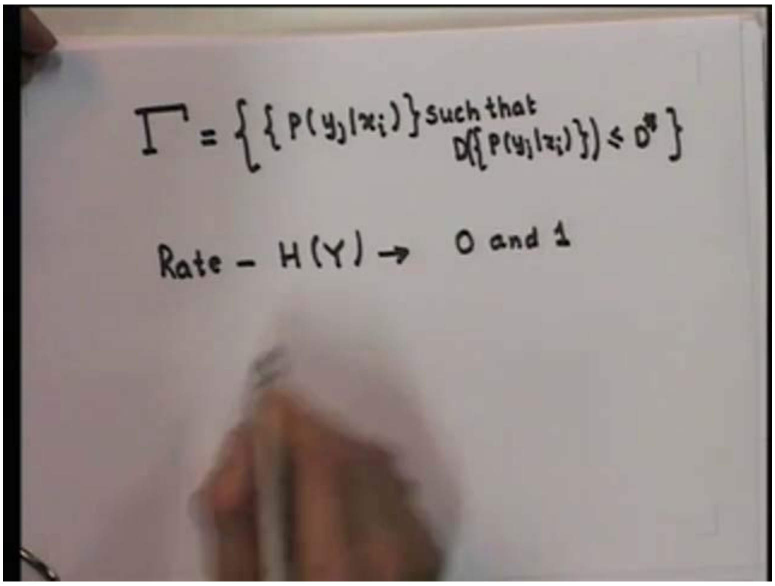
(Refer Slide Time: 44:00)



Now, in general the distortion D is equal to double summation over i and j. Now, each term in this summation consist of three factors, the distortion measure, the source density and the conditional probability. The distortion measure is a measure of closeness to the original and reconstruction versions of the signal and is generally determined by the particular application. The source probabilities are solely determined by the source. The third factor which is the set of conditional probabilities can be seen as a description of the compression scheme.

Therefore, for a given source with some P D F and a specified distortion measure, the distortion is a function only of the conditional probabilities. That is distortion is a function of conditional probabilities P of y j given x i. Therefore, we can write the constraint that the distortion D be less than some value D star as a requirement that the conditional probabilities for the compression scheme, belong to a set of conditional probabilities that have the properties as follows.

(Refer Slide Time: 46:28)



Probability of y j given x i that set such that, distortion which is function of conditional probabilities be less than or equal to D star. So, once we know the set of compression schemes to which we have to confine our self, we can start to look at the rate of these schemes. In the earlier example we saw that the rate was determined by the entropy of the reconstructed output.

However, that was a result of the fact that the conditional probabilities describing that particular source coder took on the values 0 and 1. Now, this did not be the case in a general application. So, in the next class we will mathematically define the rate distortion function and look at the properties of this rate distortion function. We will also look at some specific examples and calculate the rate distortion functions for those examples.