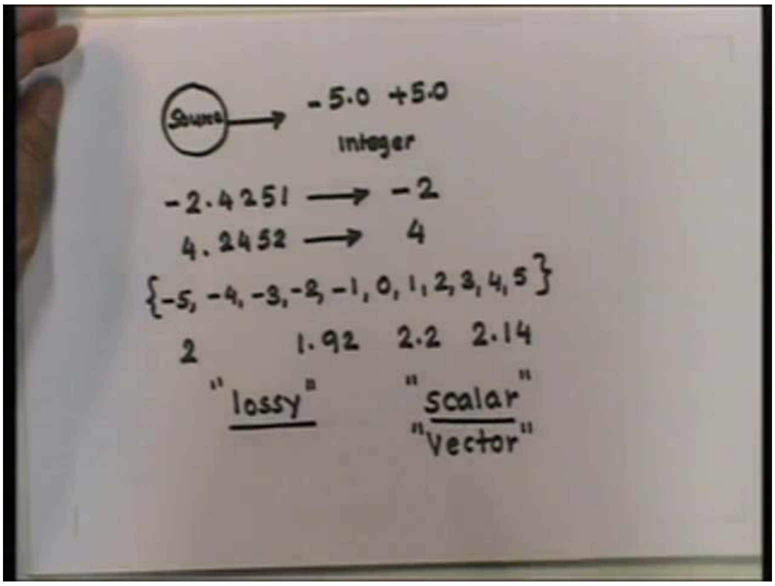
Information Theory and Coding

Prof. S. N. Merchant
Department of Electrical Engineering
Indian Institute of Technology, Bombay

Lecture - 35
Introduction to Quantization

During the next couple of lectures, we will study quantization, one of the simplest and most general ideas in lossy compression. In many lossy compression applications, it is required to represent each source output using one of a small number of code words. The number of possible distinct source output values is generally much larger than the number of code words used to represent them. This process of representing a large possibly infinite set of values with a much smaller set is called quantization. The device which achieves this quantization is called quantizer.

(Refer Slide Time: 02:11)



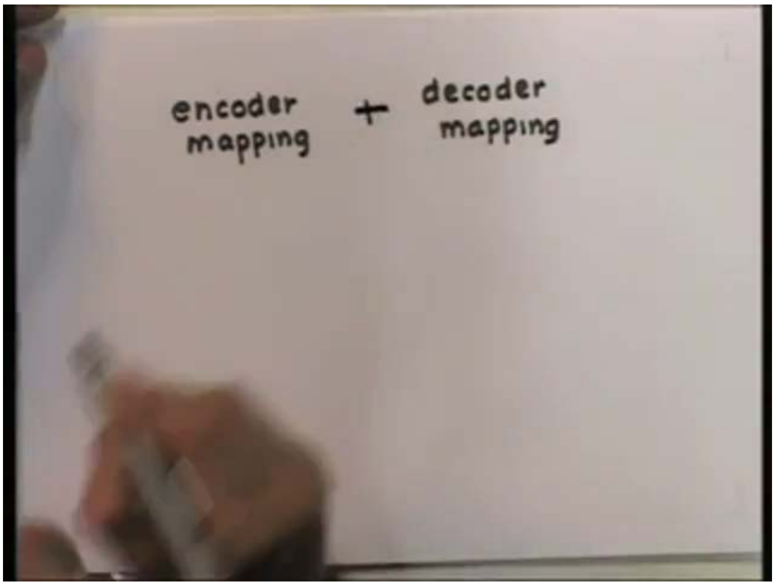
Source → -5.0 +5.0
integer
-2.4251 → -2
4.2452 → 4
{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5}
2 1.92 2.2 2.14
"lossy" "scalar"
"Vector"

Consider a source that generates number between minus 5 and plus 5, a simple quantization scheme would be to represent each output of the source with the integer value closest to it. For example, if the source output is minus 2.4251, we would represent it as minus 2 and if the source output is 4.2452, we would represent it as 4. This approach reduces the size of the alphabet required to represent the source output. The infinite number of values between minus 5 and plus 5 are represented with a set that contains only 11 values minus 5 minus 4 minus 3 minus 2 minus 1.

At the same time, we have also forever lost the original value. If you have told that the reconstruction value is 2, we cannot tell whether source output was 1.92 or 2.2 or 2.14 or any other infinite set of values. In other words, we have lost some information; this loss of information is the reason for the use of the word lossy in many lossy compression schemes. The set of inputs and outputs of a quantizer can be scalars or vectors.

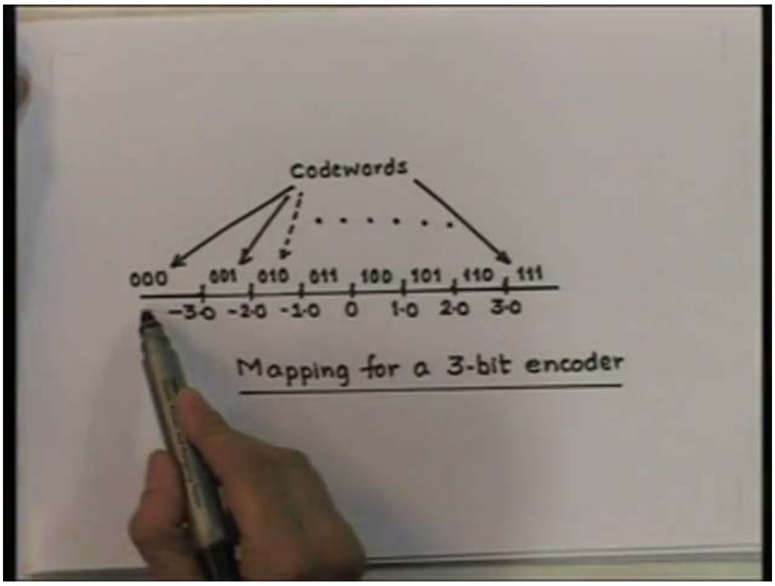
If they are scalars, we call the quantizer as scalar quantizer and if they are vectors, we call the quantizers vector quantizers. Let us begin our study with scalar quantizers, though the process of quantization is very simple, however the design of the quantizer will have significant impact on the amount of compression that can be obtained and the loss incurred in a lossy compression scheme. Let us examine the issues related to the design of the quantizers in little more depth.

(Refer Slide Time: 05:41)



In practice, the quantizer consists of two mappings an encoder mapping and a decoder mapping.

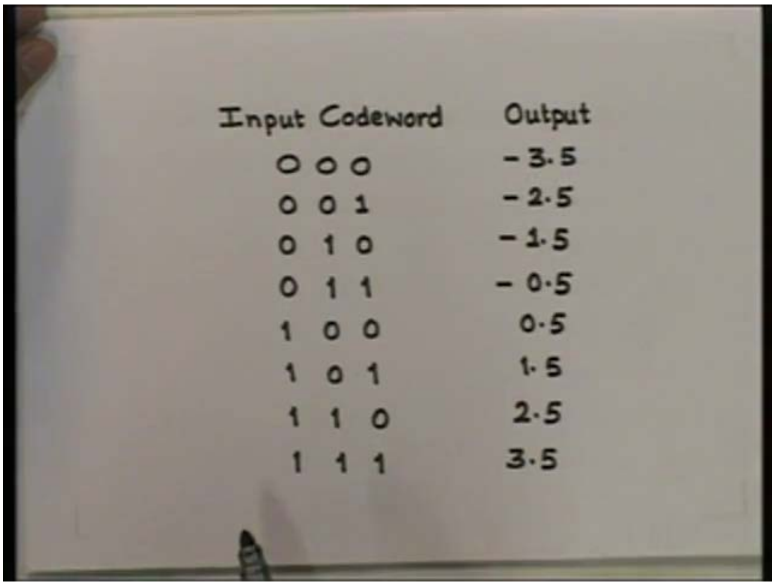
(Refer Slide Time: 06:09)



The task of an encoder is to divide the range of values that the source generates into a number of intervals. Each interval is represented by a distinct code word; the encoder represents all the source outputs that fall into a particular interval by the code word representing that interval. Now, if there could be many possible infinitely many distinct sample values that can fall in any given interval the encoder mapping is irreversible. Knowing the code only tells us the interval to which the sample value belongs. It does not tell which of many values in the interval the actual sample value is; when the sample value comes from an analog source the encoder is called an analog to digital converter.

Now, the task of a decoder is to generate a reconstruction value because a code word represents the entire interval and there is no way of knowing which value in the interval was actually generated by the source. The decoder generates a value that in some sense best represents all the values in the interval. Now, to do this, we can use the information about the distribution of the input in the interval to obtain a representative value. For time being, let us simply use the midpoint of the interval as the represented value generated by the decoder. So, if the reconstruction is analog, the decoder is often referred to as digital to analog converter decoder.

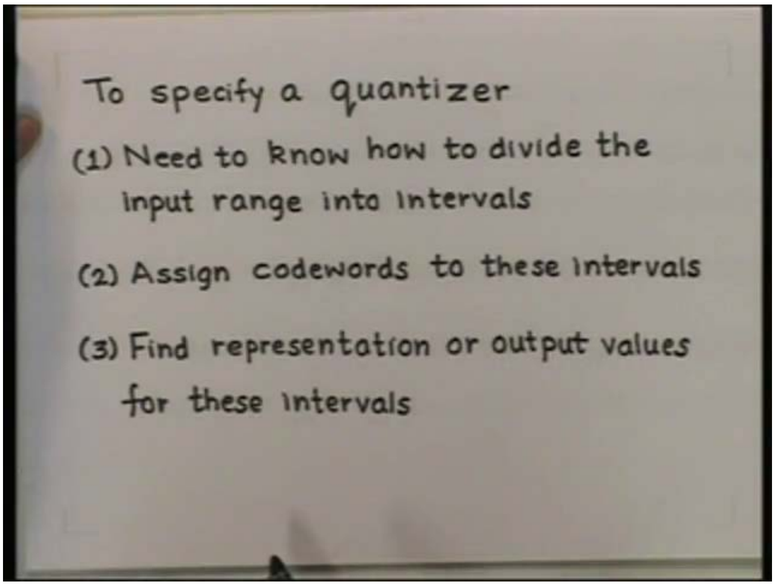
(Refer Slide Time: 08:34)



Input Codeword	Output
0 0 0	− 3.5
0 0 1	− 2.5
0 1 0	− 1.5
0 1 1	− 0.5
1 0 0	0.5
1 0 1	1.5
1 1 0	2.5
1 1 1	3.5

Mapping corresponding to the three bit Encoder discussed earlier is shown here for the input code word 0, 0, 0 the output reconstructed value is minus 3.5 and for the input code word 1, 0, 0, the output of the decoder is 0.5. So, construction of the intervals, the location etceteras can be viewed as part of the design of an encoder selection of reconstruction value is part of a design of a decoder. However, the fidelity or accuracy of the reconstruction depends on both the intervals and the reconstruction values. We call this encoder decoder pair as a quantizer.

(Refer Slide Time: 09:45)



To specify a quantizer

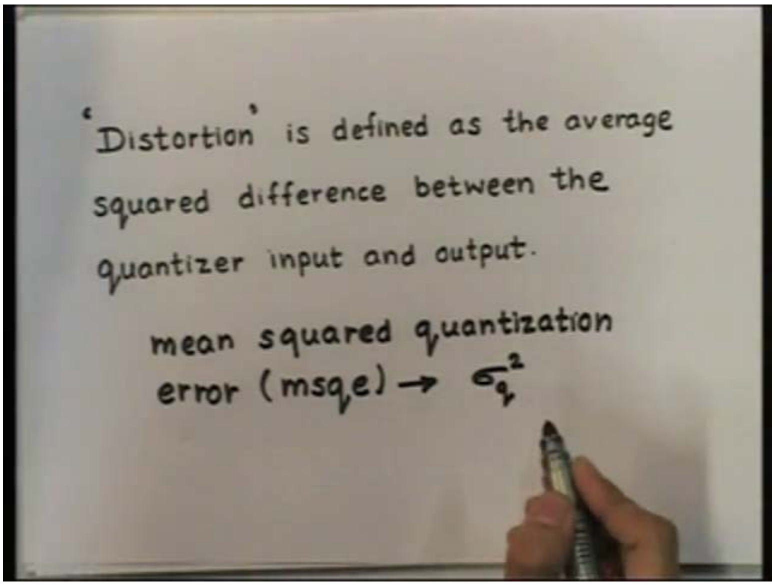
(1) Need to know how to divide the input range into intervals

(2) Assign codewords to these intervals

(3) Find representation or output values for these intervals

In summary, to specify a quantizer, we need to know how to divide the input range into intervals, then assign code words to these intervals and finally, find representation or output values for these intervals. We need to do all these for satisfying distortion and rate criteria.

(Refer Slide Time: 10:14)

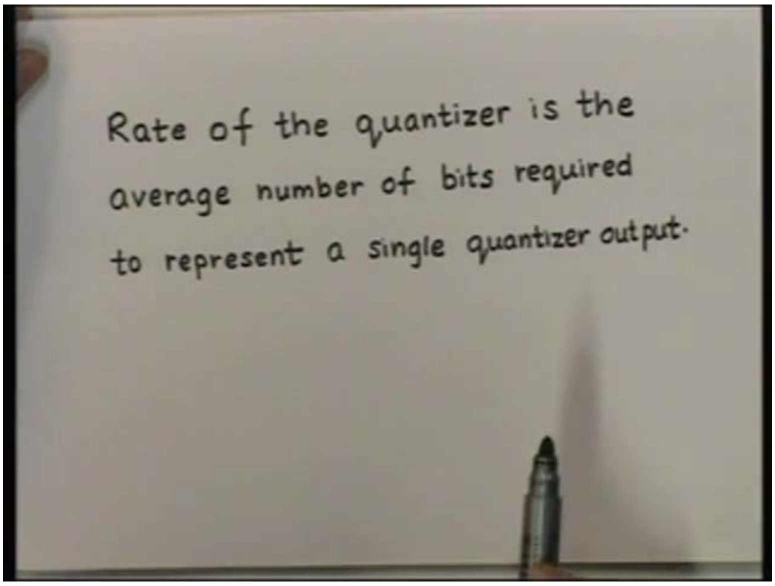


'Distortion' is defined as the average squared difference between the quantizer input and output.

mean squared quantization error (msqe) → σ_q^2

Now, distortion is defined as the average squared difference between the quantizer input and output. We call this mean squared quantization error. In short as m s q e and this is denoted by sigma squared q.

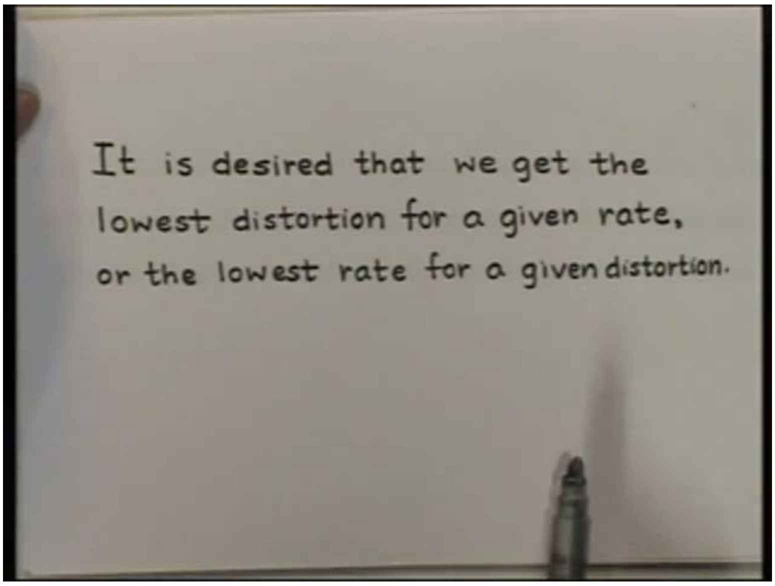
(Refer Slide Time: 11:07)



Rate of the quantizer is the average number of bits required to represent a single quantizer output.

Similarly, the rate of the quantizer is the average number of bits required to represent a single quantizer output.

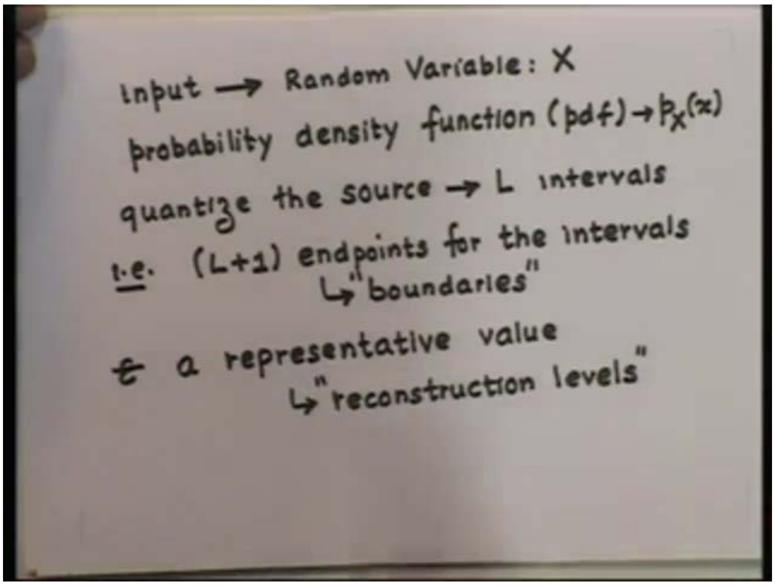
(Refer Slide Time: 11:28)



It is desired that we get the lowest distortion for a given rate, or the lowest rate for a given distortion.

Now, our goal should be to get the lowest distortion for a given rate or the lowest rate for a given distortion. Now, this quantizer design problem can be posed in precise terms as follows.

(Refer Slide Time: 11:51)



input → Random Variable: X

probability density function (pdf) → $p_X(x)$

quantize the source → L intervals

i.e. (L+1) endpoints for the intervals

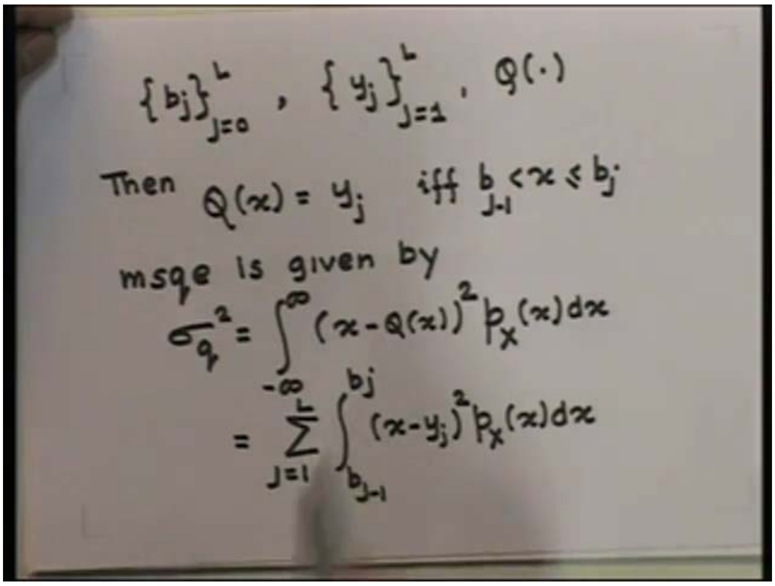
↳ "boundaries"

& a representative value

↳ "reconstruction levels"

Let us take an input model by a random variable X with probability density function that is P D F given by p subscript capital X. Now, the design problem is to quantize the source with L intervals that is we have to specify L plus 1 end point for the intervals and representative to value for each of the intervals. Now, the end points of the intervals are known as boundaries while the representative values are known as reconstruction levels.

(Refer Slide Time: 14:20)



$\{b_j\}_{j=0}^{L}$, $\{y_j\}_{j=1}^{L}$, $Q(\cdot)$

Then $Q(x) = y_j$ iff $b_{j-1} < x \leq b_j$

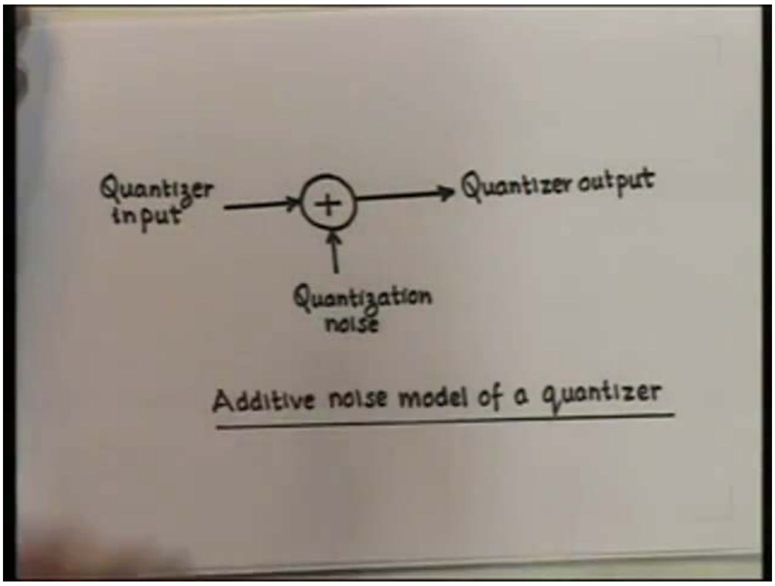
msqe is given by

$$\sigma_q^2 = \int_{-\infty}^{\infty} (x - Q(x))^2 p_X(x)\,dx$$

$$= \sum_{j=1}^{L} \int_{b_{j-1}}^{b_j} (x - y_j)^2 p_X(x)\,dx$$

So, let us denote the decision boundaries of y b j, j goes from 0 to l the reconstruction levels by y j, j is equal to 1 to L and the quantization operation by Q then Q x is equal to y j if and only if x lies in the interval specified as follows. Now, the mean square quantization error is given by integral of minus infinity to plus infinity x minus Q x squared P D F of the random variable x. This can be rewritten as follows: summation over L intervals with quantization error for a particular interval given by the integral the difference between the quantizer input x and output y is equal to Q x is referred to as the quantization error. It is also known as quantization distortion or more popularly known as quantization noise.

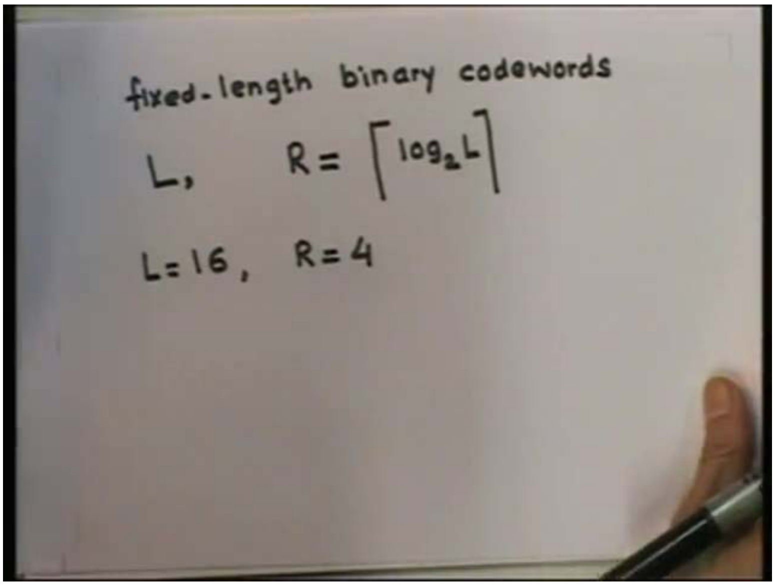
(Refer Slide Time: 16:47)



Additive noise model of a quantizer

So, the quantization process can be modeled as an additive noise process as shown here. We have a quantizer input and quantization noise gets added to it to give the quantizer output, this is additive noise model of a quantizer.

(Refer Slide Time: 17:49)



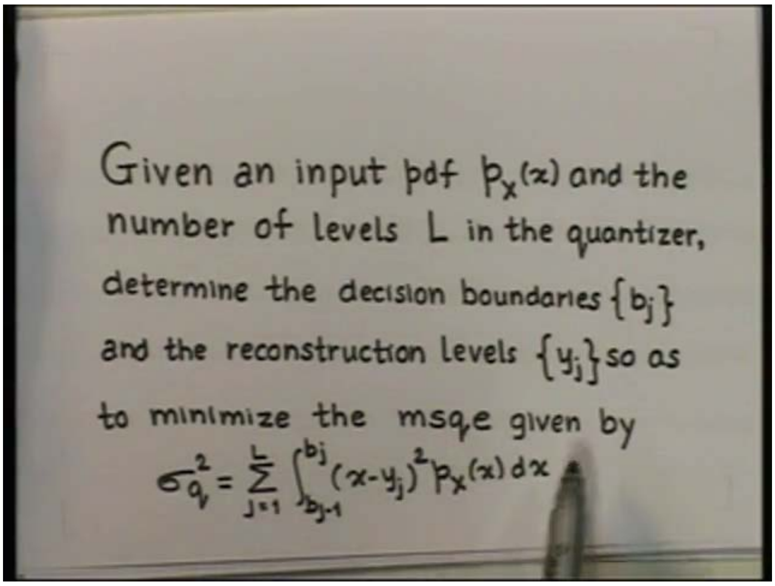
fixed-length binary codewords

$L, \quad R = \lceil \log_2 L \rceil$

$L = 16, \quad R = 4$

Now, if we use fixed length binary code words to represent the quantizer output, then the size of the output alphabet immediately specifies the rate. So, if the number of quantizer outputs is L, then the rate is given by r is equal to the next largest integer of log to the base 2 of L. So, if L is equal to 16, then R is equal to 4.

(Refer Slide Time: 18:53)

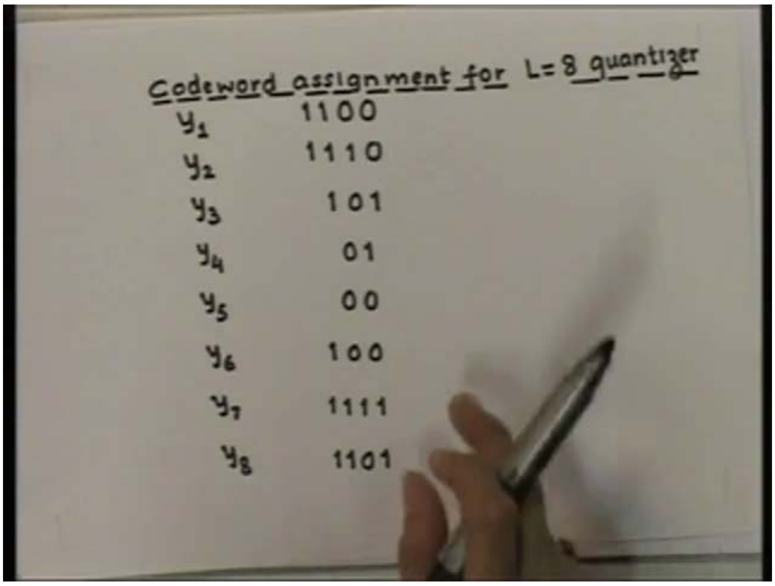


Given an input pdf $p_x(x)$ and the number of levels L in the quantizer, determine the decision boundaries $\{b_j\}$ and the reconstruction levels $\{y_j\}$ so as to minimize the msqe given by

$$\sigma_q^2 = \sum_{j=1}^{L} \int_{b_{j-1}}^{b_j} (x-y_j)^2 \, p_x(x) \, dx$$

Therefore, the quantizer design problem can be posed as follows: given an input P D F and the number of levels l in the quantizer. The problem is to determine the decision boundaries that is b j and the reconstruction levels that is y j, so as to minimize the mean square quatization error given by the following expression. Now, however if we are allowed to use the variable length codes such as Hoffmann codes or arithmetic codes along with the size of the alphabet. The selection of the decision boundaries will also affect the rate of the quantizer. However, if we are allowed to use variable length codes such as Hoffmann code or arithmetic code along with the size of the alphabet, the selection of the decision boundaries will also affect the rate of the quantizer.

(Refer Slide Time: 20:14)

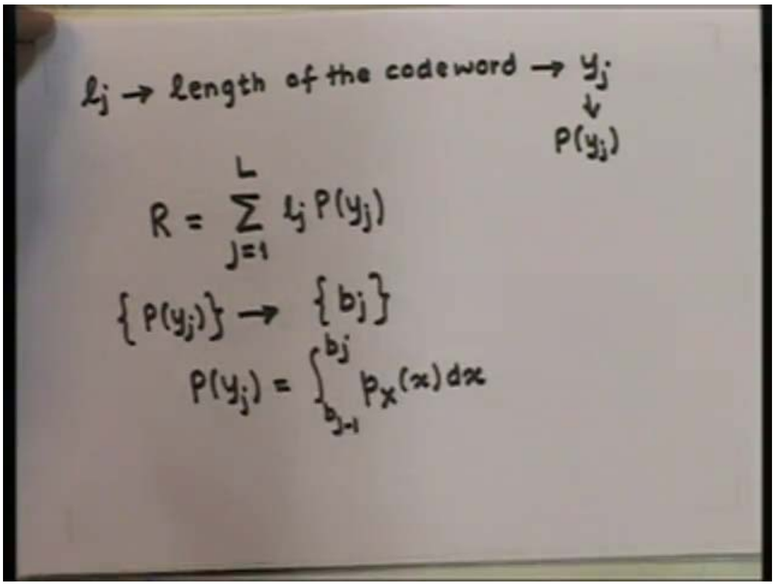


Codeword assignment for L=8 quantizer

y_1	1100
y_2	1110
y_3	101
y_4	01
y_5	00
y_6	100
y_7	1111
y_8	1101

For example, let us look at the code word assignment for L equal to 8 quantizer. According to this, code word assignment if the output y 5 occurs, we used two bits to encode it while if output y 2 occurs we need four bits to encode it. Now, the average rate will depend on how often we have to encode y 5 versus how often we have to encode y 2. This implies that the average rate will depend on the probability of occurrence of the outputs.

(Refer Slide Time: 21:12)



$l_j \to$ length of the codeword $\to y_j$

$\downarrow$

$P(y_j)$

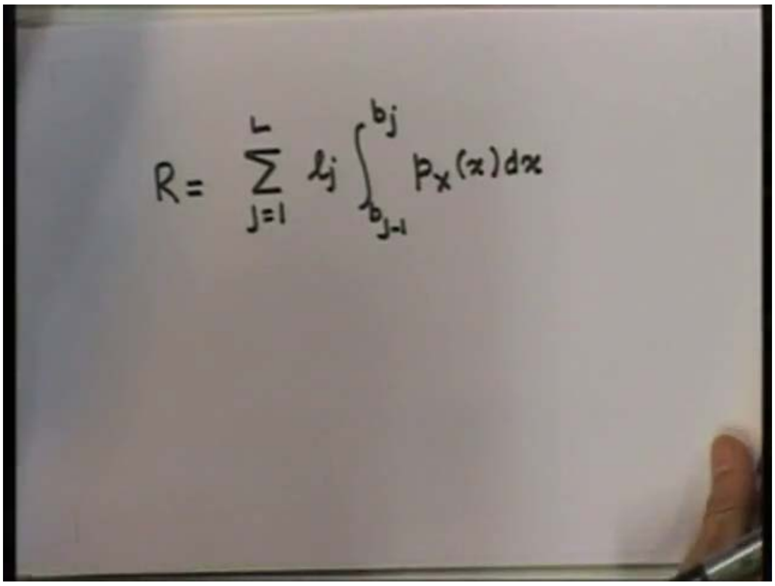
$$R = \sum_{j=1}^{L} l_j P(y_j)$$

$$\{P(y_j)\} \to \{b_j\}$$

$$P(y_j) = \int_{b_{j-1}}^{b_j} p_X(x)\,dx$$

So if l j denotes the length of the code word corresponding to the output y j and probability y j is the probability of occurrence of y j then the rate is given by. However, the probabilities p y j depend on the decision boundaries b j for example, the probability of y j occurring is given by therefore; the rate is a function of the decision boundaries and is given by the expression.

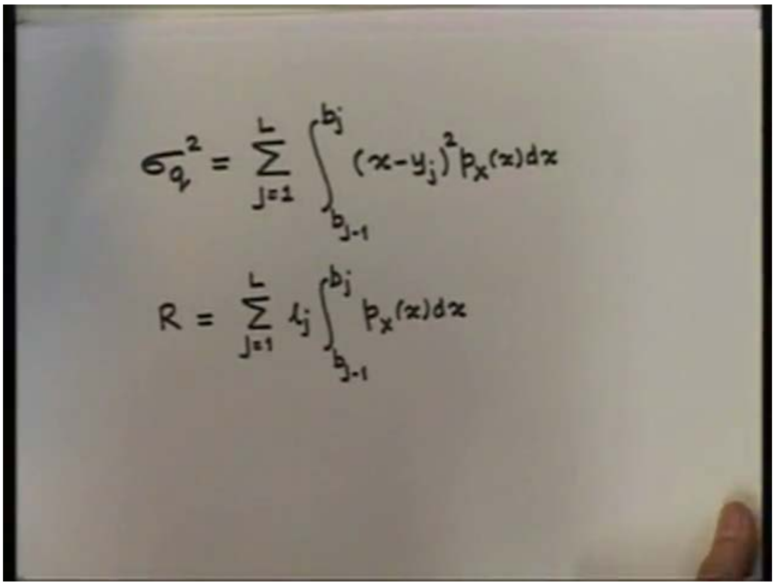
(Refer Slide Time: 22:54)



$$R = \sum_{j=1}^{L} l_j \int_{b_{j-1}}^{b_j} p_X(x)\,dx$$

.So, what follows from the discussions and this equation that for a given source input the partitions, we select and the representation for these partitions will determine the distortion incurred during the quantization process.

(Refer Slide Time: 23:14)

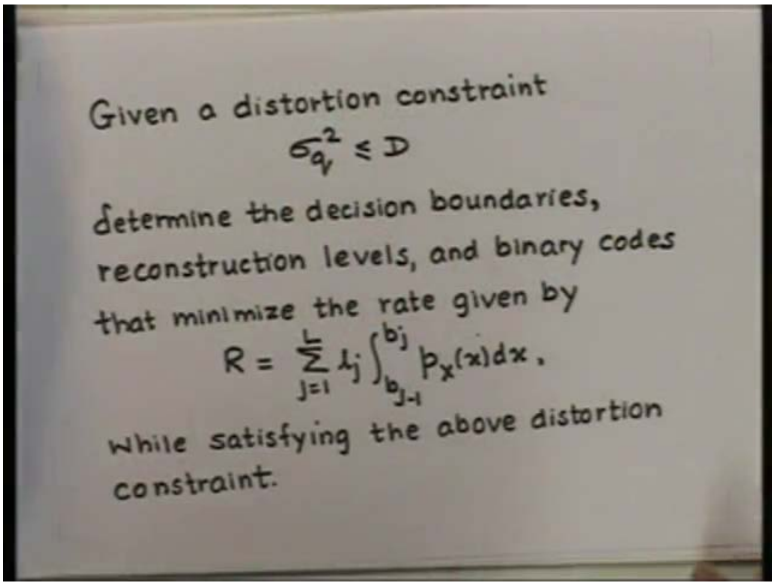


$$\sigma_q^2 = \sum_{j=1}^{L} \int_{b_{j-1}}^{b_j} (x-y_j)^2 p_X(x)\,dx$$

$$R = \sum_{j=1}^{L} l_j \int_{b_{j-1}}^{b_j} p_X(x)\,dx$$

The partitions we select and the binary codes for the partitions will determine the rate for the quantizer. Thus, the problem of finding the optimum partition codes and representation levels are all linked. In light of this information we can restate our problem statement as follows.

(Refer Slide Time: 24:12)



Given a distortion constraint

$$\sigma_q^2 \leq D$$

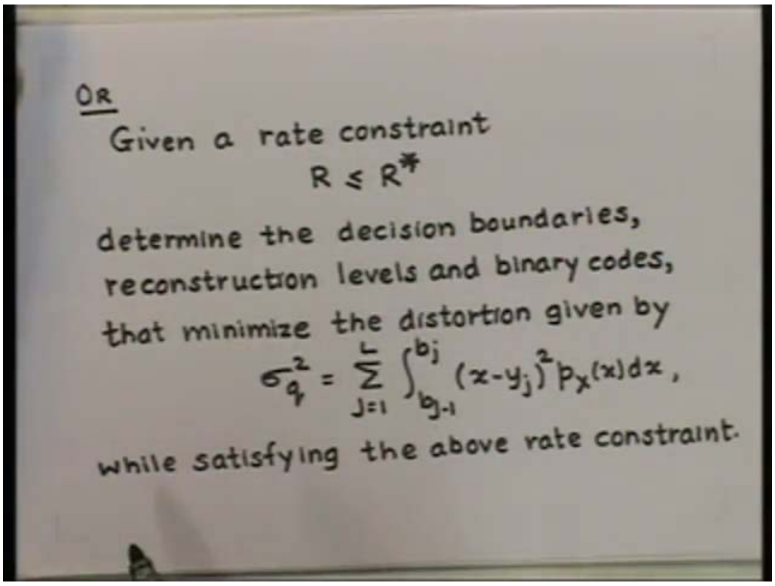
determine the decision boundaries, reconstruction levels, and binary codes that minimize the rate given by

$$R = \sum_{j=1}^{L} l_j \int_{b_{j-1}}^{b_j} p_X(x)\,dx\,,$$

while satisfying the above distortion constraint.

Given a distortion constraint determine the decision boundaries reconstruction levels, and binary codes that minimize the rate given by this expression, while satisfying the above distortion constraint or the same problem can be posed as follows.

(Refer Slide Time: 24:50)



OR

Given a rate constraint

$$R \leq R^*$$

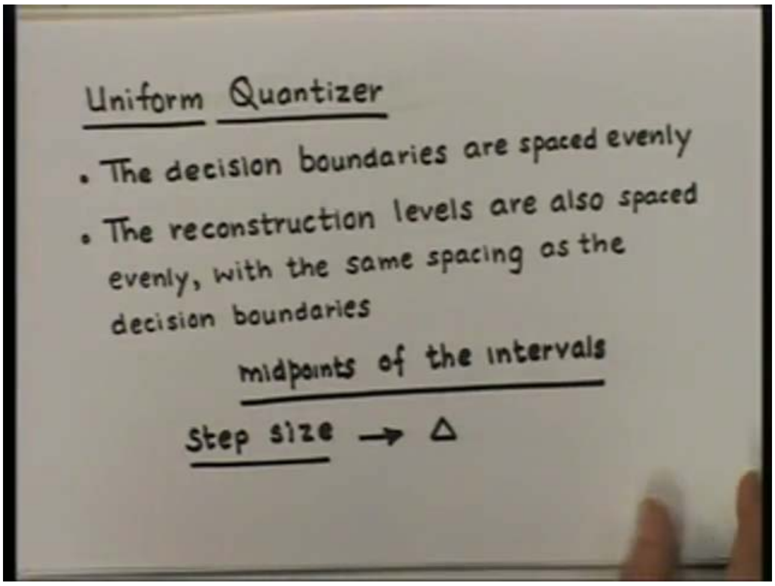
determine the decision boundaries, reconstruction levels and binary codes, that minimize the distortion given by

$$\sigma_q^2 = \sum_{j=1}^{L} \int_{b_{j-1}}^{b_j} (x - y_j)^2 p_X(x)\, dx ,$$

while satisfying the above rate constraint.

Given a rate constraint, determine the decision boundaries reconstruction levels and binary codes that minimize the distortion given by this expression while satisfying the above rate distortion constraint. Now, these problem statements of quantizer design while more general than our initial statement is also substantially more complex. However, in practical applications, there are situations in which we can simplify the problem we often use fixed length code words to encode the quantizer output. In this case, the rate is simply the number of bits used to encode each output and we can use our initial statement of the quantizer design problem. So, let us begin our study of quantizer design problem by looking at the simpler version, then move over to tackle more complex version.

(Refer Slide Time: 26:16)



Uniform Quantizer

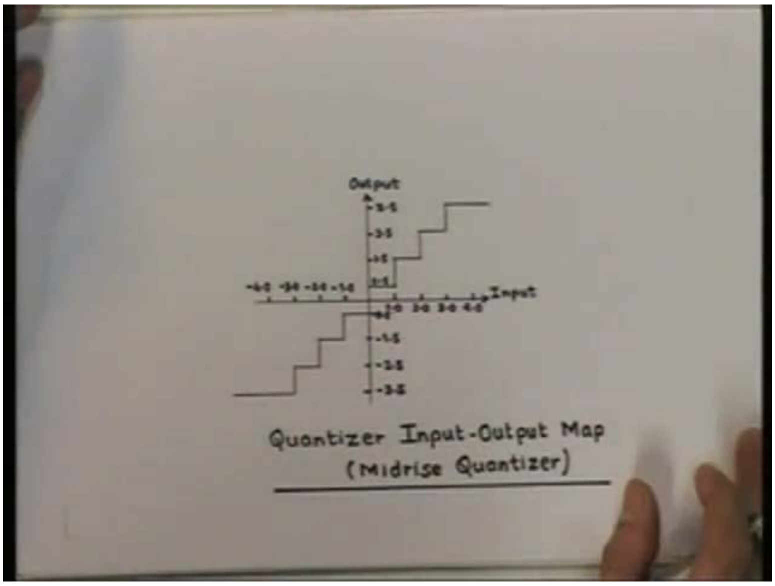
- The decision boundaries are spaced evenly
- The reconstruction levels are also spaced evenly, with the same spacing as the decision boundaries

midpoints of the intervals

Step size → Δ

So, the simplest type of quantizer is the uniform quantizer, all intervals in uniform quantizers are of the same size except possibly of the two outer intervals. So, a uniform quantizer has the following properties, the decision boundaries are spaced evenly the reconstruction level are also spaced evenly with the same spacing as the decision boundaries in the inner intervals. They are the midpoints of the intervals; this constant spacing is usually referred to as the step size and is denoted by delta.

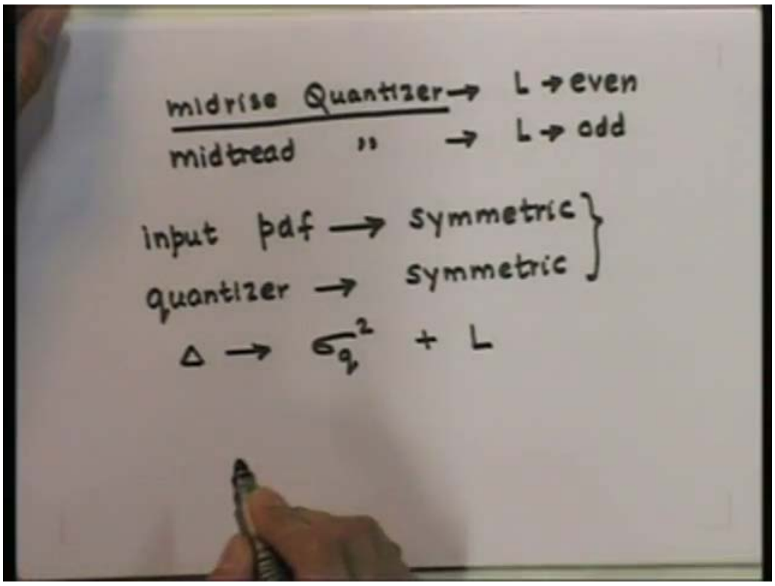
(Refer Slide Time: 27:51)



Quantizer Input-Output Map
(Midrise Quantizer)

Now, the mapping for the quantizer, for the three bit encoder pair discussed earlier can be represented by the input output map shown here. This is a example of a uniform quantizer with delta equal to 1. It does not have 0 has one of its representation level, such a quantizer is called a midrise quantizer and alternative uniform quantizer could be the 1 shown here this is called a midtread quantizer, usually we use a midrise quantizer.

(Refer Slide Time: 28:48)



midrise Quantizer → L → even

midtread " → L → odd

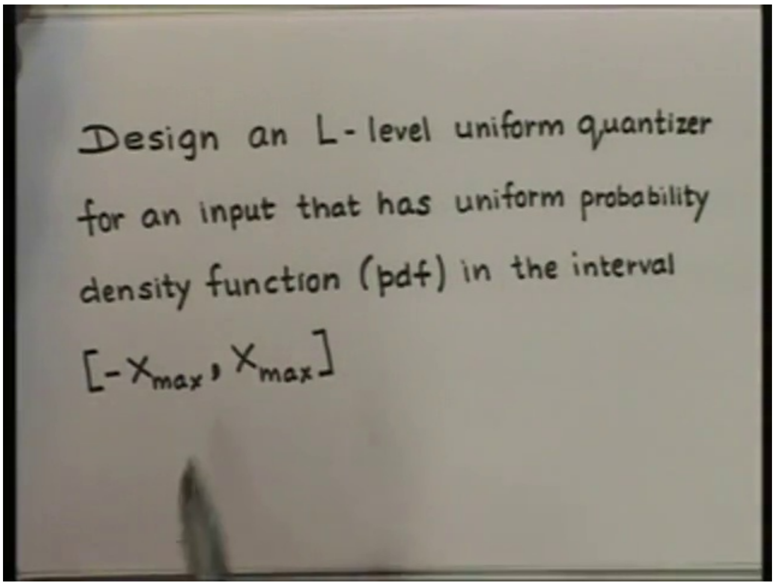
input pdf → symmetric

quantizer → symmetric

$\Delta \rightarrow \sigma_q^2 + L$

If number of construction levels that is L is equal to even then we use a mid tread quantizer. If number of reconstruction level is odd, now for our discussion, we will assume midrise quantizer. We will assume that the input P D F is symmetric around the origin and we also assume that the quantizer is also symmetric, this will simplify our discussions. Now, given all these assumptions, the design of a uniform quantizer consists of finding the step size delta, that minimizes the distortion for a given input process and number of decision levels L. So, we start our study of quantizer design with the simplest of all cases.

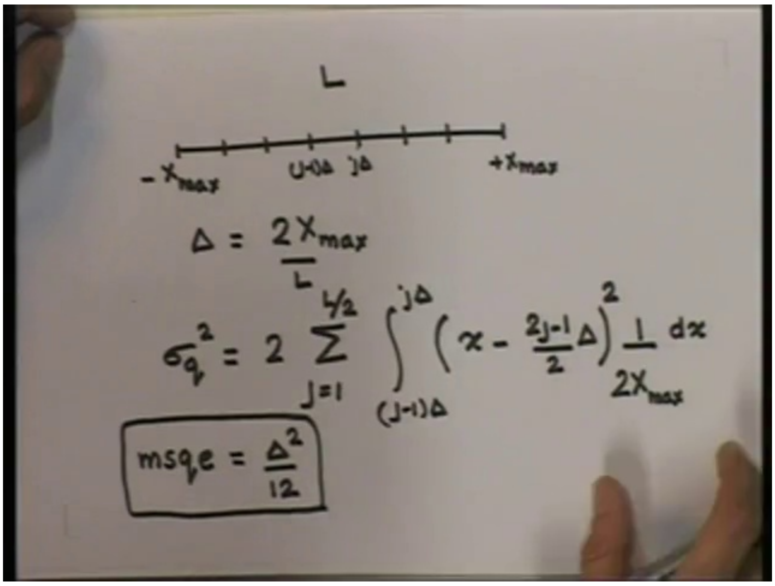
(Refer Slide Time: 30:32)



Design an L - level uniform quantizer for an input that has uniform probability density function (pdf) in the interval $[-X_{max}, X_{max}]$

Let us consider design of a L level uniform quantizer for an input that has uniform probability density function in the interval minus X max to plus X max.

(Refer Slide Time: 30:56)



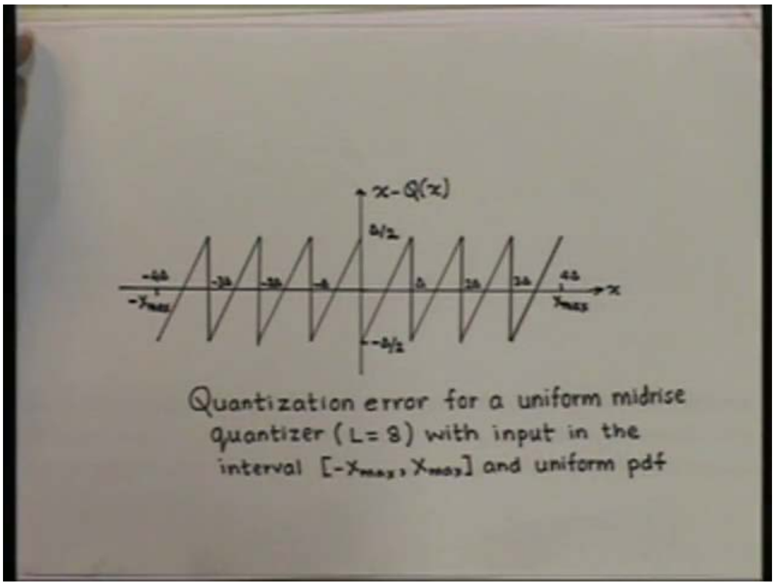
$$\Delta = \frac{2X_{max}}{L}$$

$$\sigma_q^2 = 2\sum_{j=1}^{L/2}\int_{(j-1)\Delta}^{j\Delta}\left(x - \frac{2j-1}{2}\Delta\right)^2 \frac{1}{2X_{max}}\,dx$$

$$\boxed{msqe = \frac{\Delta^2}{12}}$$

Now, this implies that we need to divide the input range minus X max to plus X max into L equally sized intervals and in this case the step sized delta is given by 2 X max upon l. Now, if we assume that the reconstruction level is the midpoint of the boundaries, then for this assumption the distortion becomes we assume L is even and since the P D F of the input is uniform, we can write 1 by 2 X max. The j th interval is given by these boundaries, now if we evaluate this integral we find that the mean squared quantization error is equal to delta squared by 12. Now, this same result can be more easily obtained as follows the quantization error.

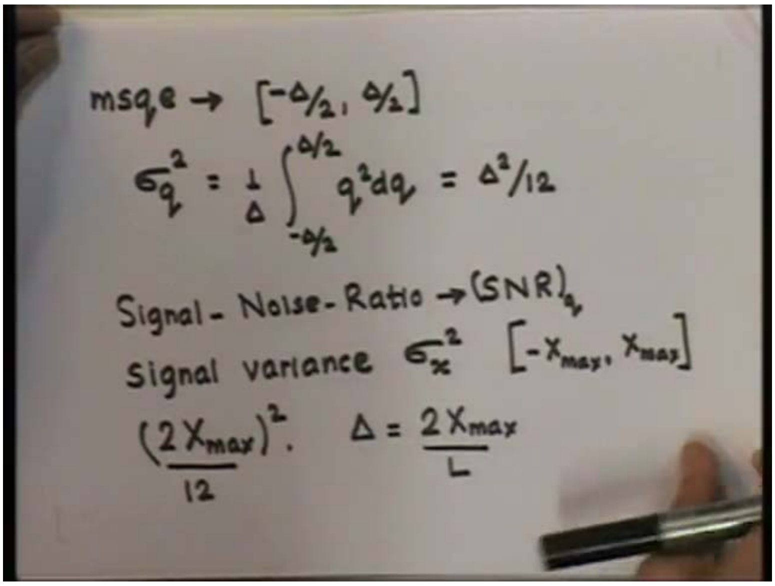
(Refer Slide Time: 33:22)



Quantization error for a uniform midrise quantizer ($L = 8$) with input in the interval $[-X_{max}, X_{max}]$ and uniform pdf

Q is given by x minus Q x and if we plot this quantization error for a uniform midrise quantizer for l equal to 8 with input in the interval minus X max to X max and uniform P D F, we get the following plot therefore. The quantization error lies in the interval minus delta by 2 to plus delta by 2. Now, as the input is of uniform P D F it is not difficult to show that the quantization error q is also uniform over this interval.

(Refer Slide Time: 34:36)



msqe → $[-\Delta/2, \Delta/2]$

$$\sigma_q^2 = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} q^2 dq = \Delta^2/12$$

Signal-Noise-Ratio → $(SNR)_q$

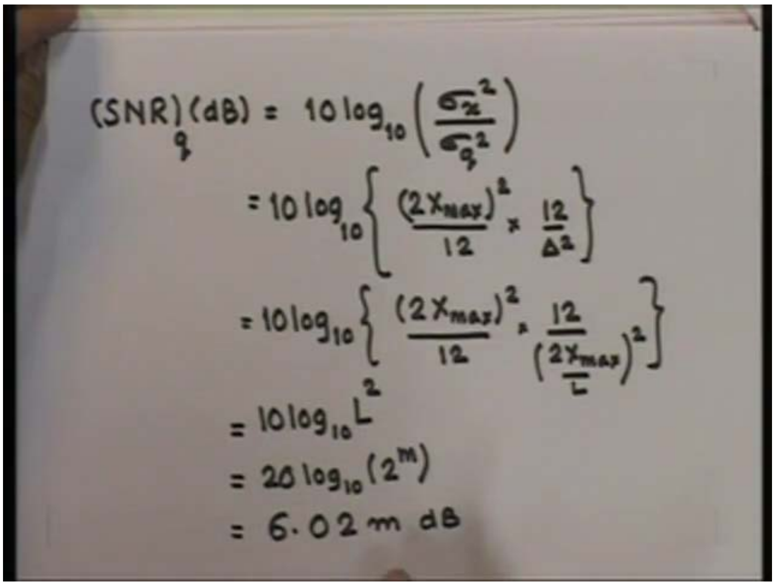
Signal variance σ_x^2 $[-X_{max}, X_{max}]$

$$\frac{(2X_{max})^2}{12}. \quad \Delta = \frac{2X_{max}}{L}$$

Thus, the he mean square quantization error is nothing but the second moment of the random variable with uniform P D F in the interval minus delta by 2 to plus delta by 2. If we assume this, we get mean squared quantization error as 1 by delta, which is the P D F of uniform random variable q integrated over minus delta by 2 to plus delta by 2 of q squared d q.

This is equal to delta squared by 12 let us also calculate the signal to quantization noise ratio which will simply denote as signal to noise ratio this is actually signal to quantization noise ratio, but for simplicity we will denote it as S N R q. Now, the signal variance that is sigma squared x for a uniform random variable X which takes on values in interval minus X max to plus X max can be shown two equal as X max upon and we also know that delta is equal to twice x max upon L. Now, for the case where we have used a fix line codes with each code word be being made of m bits.

(Refer Slide Time: 37:22)



$$(SNR)_q (dB) = 10 \log_{10} \left(\frac{\sigma_x^2}{\sigma_q^2} \right)$$

$$= 10 \log_{10} \left\{ \frac{(2X_{max})^2}{12} \times \frac{12}{\Delta^2} \right\}$$

$$= 10 \log_{10} \left\{ \frac{(2X_{max})^2}{12} \times \frac{12}{\left(\frac{2X_{max}}{L}\right)^2} \right\}$$

$$= 10 \log_{10} L^2$$

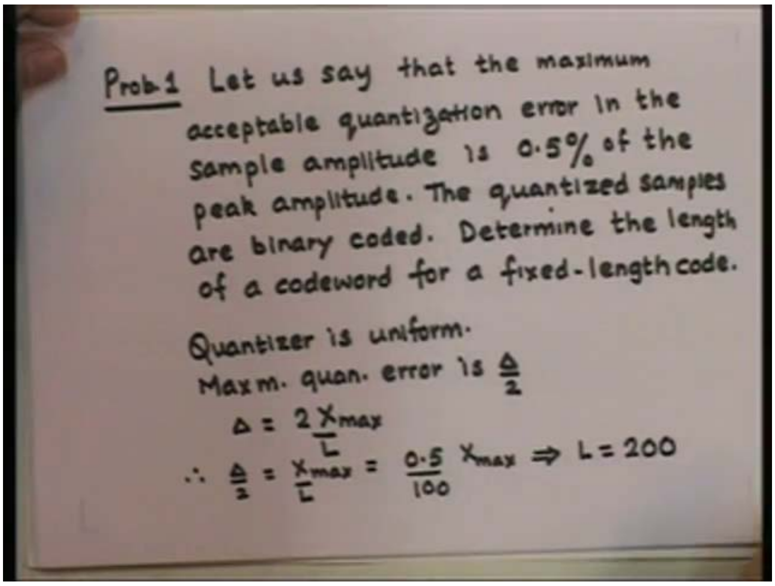
$$= 20 \log_{10} (2^m)$$

$$= 6.02 m \text{ dB}$$

The number of code words or the number of reconstruction level l is equal to 2 raised to m. So, given this we can calculate signal to noise ratio, which is to repeat again signal to quantization noise ratio in d B that is in decibels is equal to 10 log to the base 10 of signal variance divide by quantization noise variance this can be rewritten as 10 log 10. Signal variance is equal to twice X max squared divide by 12 and quantization noise variance is equal to sigma squared by 12.

This can be rewritten as delta is equal to twice X max upon l. So, this can be simplified as and further simplified which is equal to 6.02 m dB. Now, what this equation says that for every additional bit in the quantizer, we get an increase in the signal to noise ratio of approximately 6 d B. Now, this is a well known result and is often used to get an indication of the maximum gain available, if we increase the rate.

(Refer Slide Time: 40:10)



Prob.1 Let us say that the maximum acceptable quantization error in the sample amplitude is 0.5% of the peak amplitude. The quantized samples are binary coded. Determine the length of a codeword for a fixed-length code.

Quantizer is uniform.

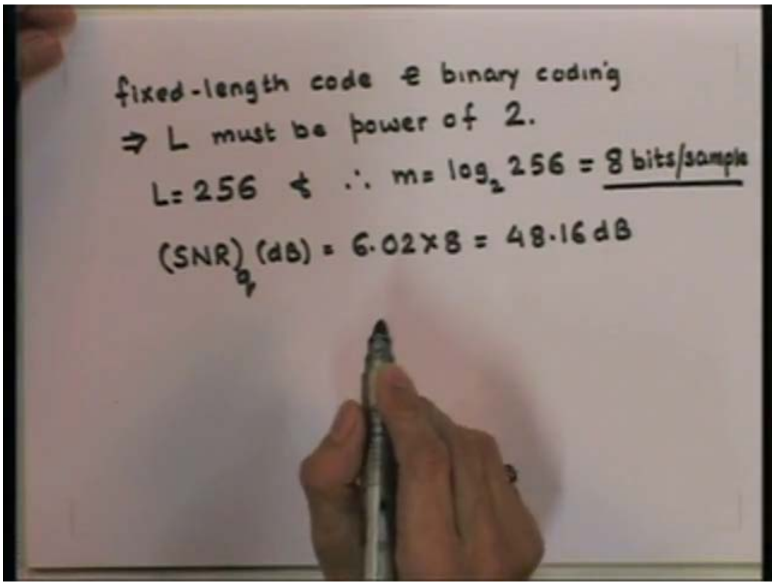
Maxm. quan. error is $\frac{\Delta}{2}$

$$\Delta = \frac{2X_{max}}{L}$$

$$\therefore \frac{\Delta}{2} = \frac{X_{max}}{L} = \frac{0.5}{100} X_{max} \Rightarrow L = 200$$

Now, let us solve a few simple examples based on this concept. Let us say That the maximum acceptable quantization error in the sample amplitude is 0.5 percent of peak amplitude. Let us assume that the quantization samples are binary coded and it is desired that we determine the length of a code word for a fixed length code. Now, the solution to this problem is as follows, so we assume that the quantizer is uniform. So, maximum quantization error in terms of step size delta is delta by 2. Now, we also know that delta is equal to 2 X max by l for a uniform quantizer, therefore delta by 2 is equal to X max by l and this is equal to 0.5 percent of the peak amplitude that is 0.5 x max upon hundred which implies that L is equal to 200 now assuming.

(Refer Slide Time: 43:40)



fixed-length code & binary coding

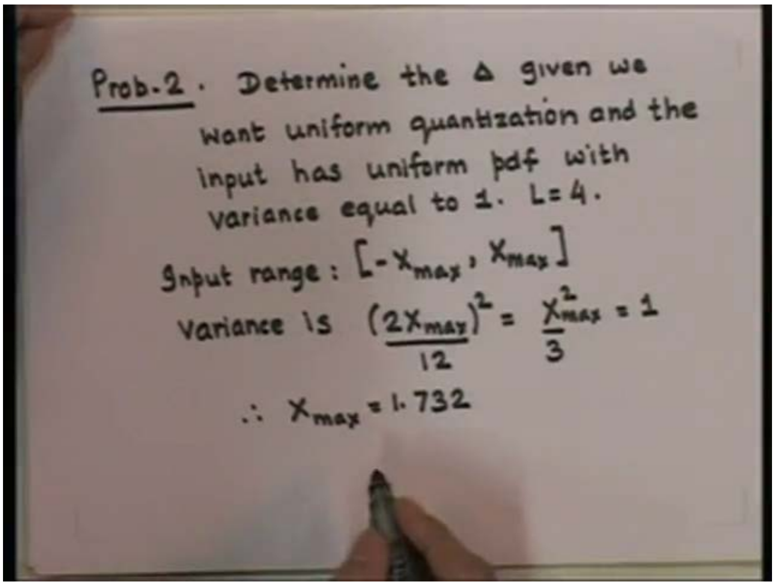
⇒ L must be power of 2.

$L = 256$ & $\therefore m = \log_2 256 = \underline{8 \text{ bits/sample}}$

$(SNR)_q (dB) = 6.02 \times 8 = 48.16 \, dB$

Fixed length code and binary coding implies that L must be power of 2, hence the next higher value of L that is the power of 2 is L is equal to 256 and therefore, m is equal to log to the base 2 of 256 is equal to 8 bits per sample. So, this is the desired solution and the signal to noise ratio that is quantization noise in terms of dB is equal to 6.02 multiplied by 8 is equal to 48.16 dB.

(Refer Slide Time: 45:20)



Prob-2. Determine the Δ given we want uniform quantization and the input has uniform pdf with variance equal to 1. L = 4.

Input range : $[-X_{max}, X_{max}]$

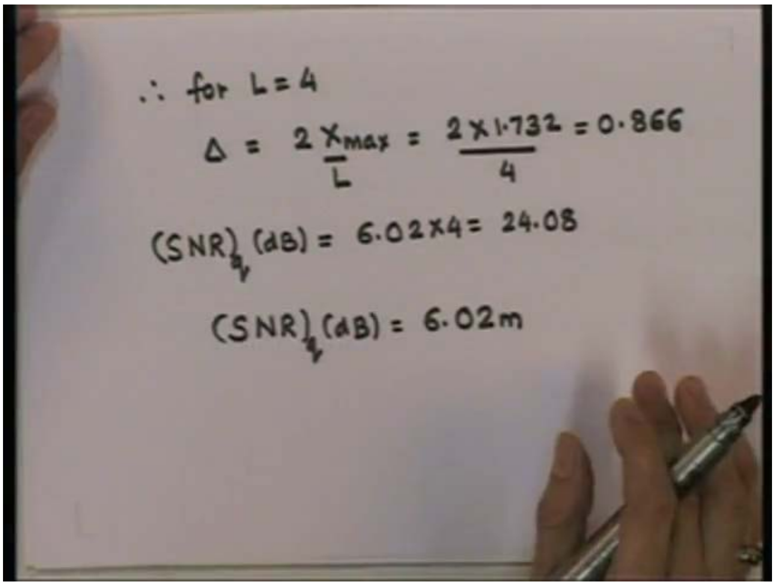
Variance is $\frac{(2X_{max})^2}{12} = \frac{X_{max}^2}{3} = 1$

$\therefore X_{max} = 1.732$

Assuming that the input P D F is uniform let us take another example, let us determine the step size that is delta, given we want uniform quantization and the input has uniform

P D F with variance equal to 1 and let us assume that l is equal to 4. Now, solution is as follows if the input range is minus X max to plus X max and if the input is uniform P D F then its variance is 2 X max square 12 which is equal to X max squared by 3 and this is equal to 1. Therefore, it implies x max is equal to 1.732.

(Refer Slide Time: 47:56)



$\therefore$ for $L = 4$

$$\Delta = \frac{2X_{max}}{L} = \frac{2 \times 1.732}{4} = 0.866$$

$$(SNR)_q (dB) = 6.02 \times 4 = 24.08$$

$$(SNR)_q (dB) = 6.02m$$

Therefore, for L equal to 4, we get the step sized delta equal to twice X max by L is equal to 0.866. Again signal to noise ratio that is quantization noise in dB for this case is equal to 6.02 multiplied by 4 is equal to 24.08. Now, it is important to remember that we derived the signal to noise ratio that is quantization noise db equal to 6.02 m under certain assumptions about the input. If this assumptions are not true, then this result will also not hold, now quite often the sources we deal with does not have a uniform P D F and we still want to have the simplicity of a uniform quantizer.

In this cases even if the input are bounded then simply by dividing the input range by the number of reconstruction level will not provide a very good quantizer design. This approach will become practically impossible when we model the sources that have unbounded P D F such as Gaussian P D F. Now, in the next class we will examine design of uniform quantizer for non-uniform sources.