

# Module 3

## LOSSY IMAGE COMPRESSION SYSTEMS

# Lesson 6

## Theory of Quantization

Version 2 ECE IIT, Kharagpur

## Instructional Objectives

At the end of this lesson, the students should be able to:

1. Define quantization.
2. Distinguish between scalar and vector quantization.
3. Define quantization error and optimum scalar quantizer design criteria.
4. Design a Lloyd-Max quantizer.
5. Distinguish between uniform and non-uniform quantization.
6. Define rate-distortion function.
7. State source coding theorem.
8. Determine the minimum possible rate for a given SNR to encode a quantized Gaussian signal.

## 6.0 Introduction

In lesson-3, lesson-4 and lesson-5, we have discussed several lossless compression schemes. Although the lossless compression techniques guarantee exact reconstruction of images after decoding, their compression performance is very often limited. We have seen that with lossless coding schemes, our achievable compression is restricted by the source entropy, as given by Shannon's noiseless coding theorem. In lossless predictive coding, it is the prediction error that is encoded and since the entropy of the prediction error is less due to spatial redundancy, better compression ratios can be achieved. Even then, compression ratios better than 2:1 is often not possible for most of the practical images. For significant bandwidth reductions, lossless techniques are considered to be inadequate and lossy compression techniques are employed, where psycho-visual redundancy is exploited so that the loss in quality is not visually perceptible. The main difference between the lossy and the lossless compression schemes is the introduction of the quantizer. In image compression systems, discussed in lesson-2, we have seen that the quantization is usually applied to the transform-domain image representations. Before discussing the transform coding techniques or the lossy compression techniques in general, we need to have some basic background on the theory of quantization, which is the scope of the present lesson.

In this lesson, we shall first present the definitions of scalar and vector quantization and then consider the design issues of optimum quantizer. In particular, we shall discuss Lloyd-Max quantizer design and then show the relationship between the rate-distortion function and the signal-to-noise ratio.

## 6.1 Quantization

*Quantization* is the process of mapping a set of continuous-valued samples into a smaller, finite number of output levels. Quantization is of two basic types – (a) *scalar quantization* and (b) *vector quantization*.

In *scalar quantization*, each sample is quantized independently. A scalar quantizer  $Q(\cdot)$  is a function that maps a continuous-valued variable  $s$  having a probability density function  $p(s)$  into a discrete set of reconstruction levels  $r_i$  ( $i = 1, 2, \dots, L$ ) by applying a set of the decision levels  $d_i$  ( $i = 1, 2, \dots, L$ ), applied on the continuous-valued samples  $s$ , such that

$$Q(s) = r_i \quad \text{if } s \in (d_{i-1}, d_i], \quad i = 1, 2, \dots, L \dots \dots \dots (6.1)$$

where,  $L$  is the number of output level. In words, we can say that the output of the quantizer is the reconstruction level  $r_i$ , if the value of the sample lies within the range  $(d_{i-1}, d_i]$ .

In vector quantization, each of the samples is not quantized. Instead, a set of continuous-valued samples, expressed collectively as a vector is represented by a limited number of vector states. In this lesson, we shall restrict our discussions to scalar quantization. In particular, we shall concentrate on the scalar quantizer design, i.e., how to design  $d_i$  and  $r_i$  in equation (6.1).

The performance of a quantizer is determined by its distortion measure. Let  $\hat{s} = Q(s)$  be the quantized variable. Then,  $\varepsilon = s - \hat{s}$  is the quantization error and the distortion  $D$  is measured in terms of the expectation of the square of the quantization error (i.e., the mean-square error) and is given by  $D = E[(s - \hat{s})^2]$ . We should design  $d_i$  and  $r_i$  so that the distortion  $D$  is minimized.

There are two different approaches to the optimal quantizer design –

- (a) Minimize  $D = E[(s - \hat{s})^2]$  with respect to  $d_i$  and  $r_i$  ( $i = 1, 2, \dots, L$ ), subject to the constraint that  $L$ , the number of output states in the quantizer is fixed. These quantizers perform non-uniform quantization in general and are known as *Lloyd-Max quantizers*. The design of *Lloyd-Max quantizers* is presented in the next section.
- (b) Minimize  $D = E[(s - \hat{s})^2]$  with respect to  $d_i$  and  $r_i$  ( $i = 1, 2, \dots, L$ ), subject to the constraint that the source entropy  $H(\hat{s}) = C$  is a constant and the number of output states  $L$  may vary. These quantizers are called *entropy-constrained quantizers*.

In case of *fixed-length coding*, the rate  $R$  for quantizers with  $L$  states is given by  $\lceil \log_2 R \rceil$ , while  $R > H(\dot{s})$  in case of *variable-length coding*. Thus, *Lloyd-Max quantizers* are more suited for use with *fixed-length coding*, while *entropy-constrained quantizers* are more suitable for use with *variable-length coding*.

## 6.2 Design of Lloyd-Max Quantizers

The design of Lloyd-Max quantizers requires the minimization of

$$D = E[(s - r_i)^2] = \sum_{i=1}^L \int_{d_{i-1}}^{d_i} (s - r_i)^2 p(s) ds \dots\dots\dots(6.2)$$

Setting the partial derivatives of  $D$  with respect to  $d_i$  and  $r_i$  ( $i = 1, 2, \dots, L$ ) to zero and solving, we obtain the necessary conditions for minimization as

$$r_i = \frac{\int_{d_{i-1}}^{d_i} sp(s) ds}{\int_{d_{i-1}}^{d_i} p(s) ds}, \quad 1 \leq i \leq L \dots\dots\dots(6.3)$$

$$d_i = \frac{r_i + r_{i+1}}{2}, \quad 1 \leq i \leq L \dots\dots\dots(6.4)$$

Mathematically, the decision and the reconstruction levels are solutions to the above set of nonlinear equations. In general, closed form solutions to equations (6.3) and (6.4) do not exist and they need to be solved by numerical techniques. Using numerical techniques, these equations could be solved in an iterative way by first assuming an initial set of values for the decision levels  $\{d_i\}$ . For simplicity, one can start with decision levels corresponding to uniform quantization, where decision levels are equally spaced. Based on the initial set of decision levels, the reconstruction levels can be computed using equation (6.3) if the *pdf* of the input variable to the quantizer is known. These reconstruction levels are used in equation (6.4) to obtain the updated values of  $\{d_i\}$ .

Solutions of equations (6.3) and (6.4) are iteratively repeated until a convergence in the decision and reconstruction levels are achieved. In most of the cases, the convergence is achieved quite fast for a wide range of initial values.

## 6.3 Uniform and non-uniform quantization

*Lloyd-Max quantizers* described above perform *non-uniform quantization* if the *pdf* of the input variable is not uniform. This is expected, since we should perform *finer quantization* (that is, the decision levels more closely packed and consequently more number of reconstruction levels) wherever the *pdf* is large and *coarser quantization* (that is, decision levels widely spaced apart and hence, less number of reconstruction levels), wherever *pdf* is low. In contrast, the reconstruction levels are equally spaced in *uniform quantization*, i.e.,

$$r_{i+1} - r_i = \theta \quad 1 \leq i \leq L-1$$

where  $\theta$  is a constant, that is defined as the *quantization step-size*.  
In case, the *pdf* of the input variable  $s$  is uniform in the interval  $[A, B]$ , i.e.,

$$p(s) = \begin{cases} \frac{1}{B-A} & A \leq s \leq B \\ 0 & \text{otherwise} \end{cases}$$

the design of Lloyd-Max quantizer leads to a uniform quantizer, where

$$\theta = \frac{B-A}{L}$$

$$d_i = A + i\theta \quad 0 \leq i \leq L$$

$$r_i = d_{i-1} + \frac{\theta}{2} \quad 1 \leq i \leq L$$

If the *pdf* exhibits even symmetric properties about its mean, e.g., Gaussian and Laplacian distributions, then the decision and the reconstruction levels have some symmetry relations for both uniform and non-uniform quantizers, as shown in Fig.6.1 and Fig.6.2 for some typical quantizer characteristics (reconstruction levels vs. input variable  $s$ ) for  $L$  even and odd respectively.

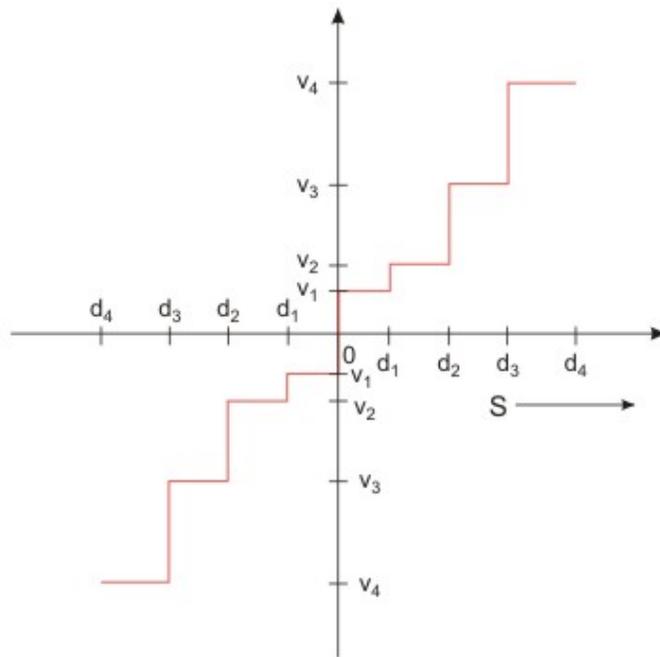


Fig 6.1 Typical quantizer characteristics for L even (=8)

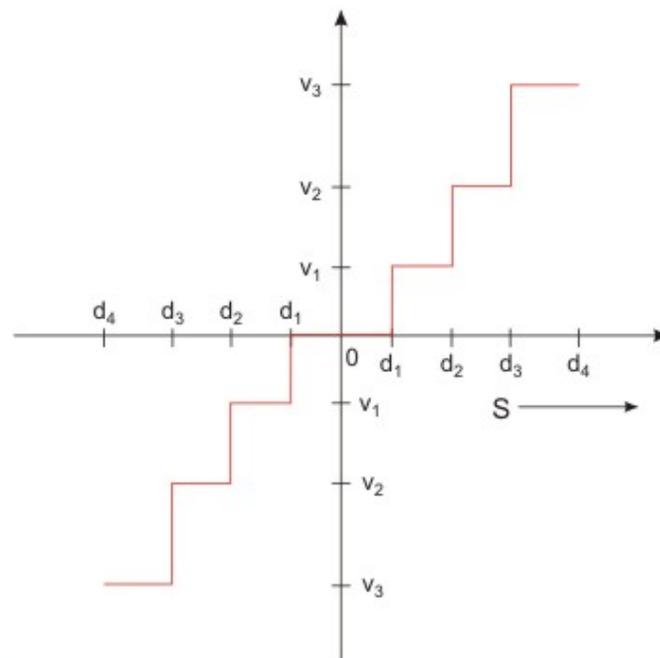


Fig 6.2 Typical quantizer characteristics for L odd (=7)

When *pdf* is even symmetric about its mean, the quantizer is to be designed for only  $L/2$  levels or  $(L-1)/2$  levels, depending upon whether  $L$  is even or odd, respectively.

## 6.4 Rate-Distortion Function and Source Coding Theorem

Shannon's Coding Theorem on noiseless channels considers the channel, as well as the encoding process to be lossless. With the introduction of quantizers, the encoding process becomes lossy, even if the channel remains as lossless. In most cases of lossy compressions, a limit is generally specified on the maximum tolerable distortion  $D$  from fidelity consideration. The question that arises is "Given a distortion measure  $D$ , how to obtain the smallest possible rate?" The answer is provided by a branch of information theory that is known as the *rate-distortion theory*. The corresponding function that relates the smallest possible rate to the distortion, is called the *rate-distortion function*  $R(D)$ . A typical nature of *rate-distortion function* is shown in Fig.6.3.

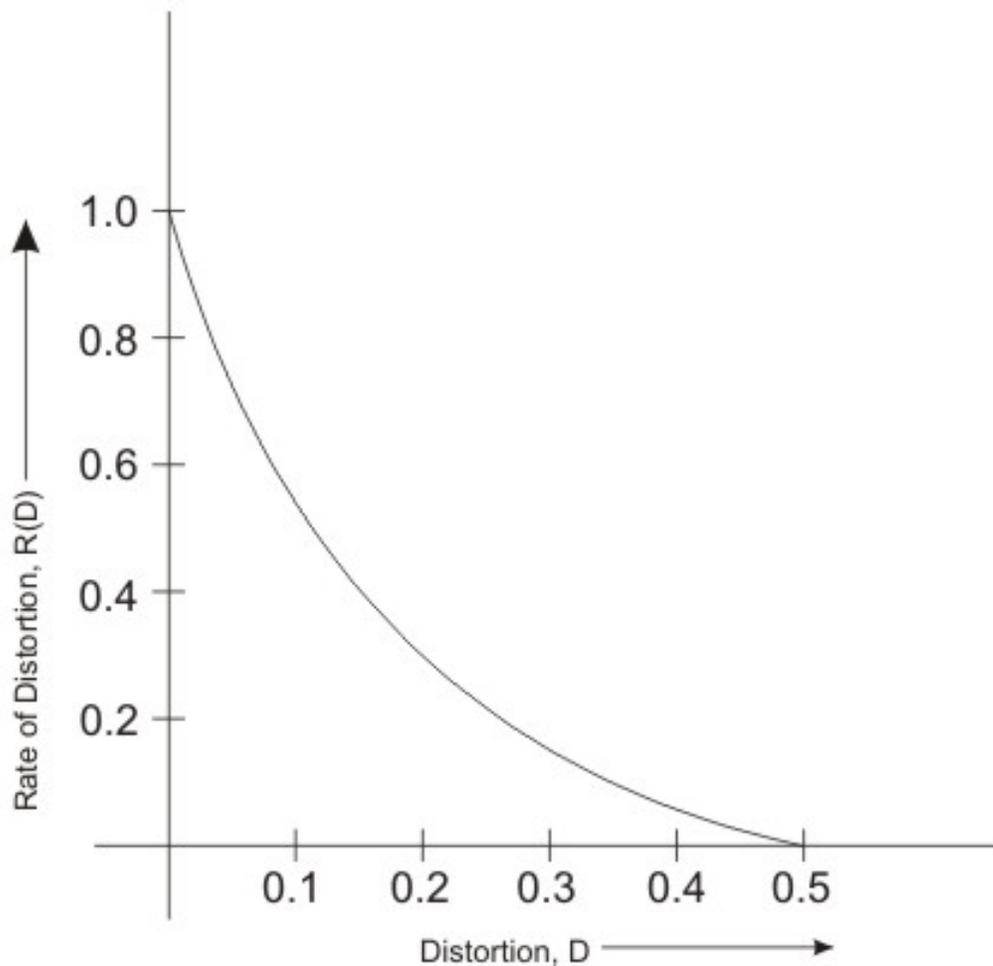


Fig 6.3 Typical rate distortion function

At no distortion ( $D=0$ ), i.e. for lossless encoding, the corresponding rate  $R(0)$  is equal to the entropy, as per Shannon's coding theorem on noiseless channels. Rate-distortion functions can be computed analytically for simple sources and distortion measures. Computer algorithms exist to compute  $R(D)$  when analytical methods fail or are impractical. In terms of the rate-distortion function, the source coding theorem is presented below.

## Source Coding Theorem

There exists a mapping from the source symbols to codewords such that for a given distortion  $D$ ,  $R(D)$  bits/symbol are sufficient to enable source reconstruction with an average distortion arbitrarily close to  $D$ . The actual bits  $R$  is given by

$$R \geq R(D)$$