# Analog to digital conversion

Sampling and reconstruction

# Quantization

Turning continuous amplitudes into discrete ones

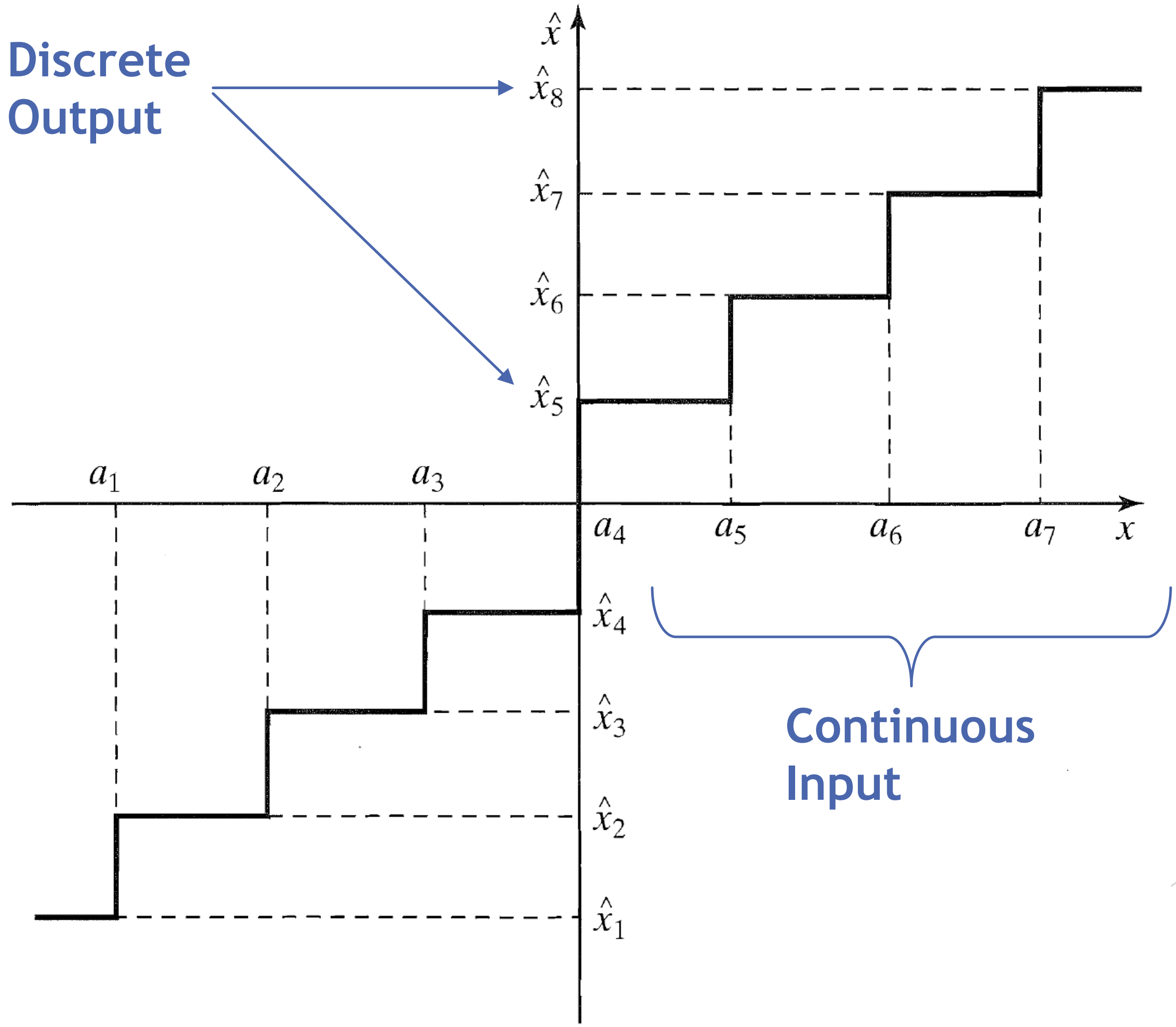
# Quantization

- After we have converted continuous time into discrete time, we can convert continuous amplitudes into discrete amplitudes
- The algorithm that does so is referred to as a **quantizer**

# Quantization

- What is a quantizer ?
  - It is a mapping between a continuous interval to a discrete interval
  - How do we compare quantizers?
    - By the dimension of the continuous set
    - By the dimension of the discrete set

# Quantization



# Quantization

- A uniform 8 level quantizer

Figure 7.3 shows an example of an 8-level quantization scheme. In this scheme, the eight regions are defined as $\mathcal{R}_1 = (-\infty, a_1]$, $\mathcal{R}_2 = (a_1, a_2]$, $\ldots$, $\mathcal{R}_8 = (a_7, +\infty)$. The representation point (or quantized value) in each region is denoted by $\hat{x}_i$ and is shown in the figure. The quantization function $Q$ is defined by

$$Q(x) = \hat{x}_i \quad \text{for all } x \in \mathcal{R}_i. \tag{7.2.1}$$

This function is also shown in the figure.

# Quantization

- Quantization is less easy to analyze than sampling
- In sampling it is reasonable to sample at equidistance interval
- Quantizing at equidistance values is usually not a good idea
  - Think about music or voice: the variation over time are pretty much constant but the amplitudes are distributed for from uniform

# Quantization

- How do we design the mapping between continuous and discrete values?
  - We need to decide the cardinality of the discrete set
  - Then we need to decide a many-to-one mapping
- What design criteria should we use?

# Quantization

- What is a quantizer ?
  - It is a mapping between a continuous interval to a discrete interval
  - How do we compare quantizers?
    - By the dimension of the continuous set
    - By the dimension of the discrete set

# Disclaimer

- Now!
  - This is where first introduce some probability
  - This is very simple thing so I won't review this now
- Most of the probability will be in the next semester
  - Here we have only one random variable and some simple expectation operation

# Quantization

- We make a **critical assumption** now
  - each sample from the sampler is an IID
- With this assumption we can look at the expected value of distortion introduced by quantization

# Quantization

Depending on the measure of distortion employed, we can define the average distortion resulting from quantization. A popular measure of distortion, used widely in practice, is the *squared error distortion* defined as $(x - \hat{x})^2$. In this expression $x$ is the sampled signal value and $\hat{x}$ is the quantized value, i.e., $\hat{x} = Q(x)$. If we are using the squared error distortion measure, then

$$d(x, \hat{x}) = (x - Q(x))^2 = \tilde{x}^2,$$

where $\tilde{x} = x - \hat{x} = x - Q(x)$. Since $X$ is a random variable, so are $\hat{X}$ and $\tilde{X}$; therefore, the average (mean squared error) distortion is given by

$$D = E[d(X, \hat{X})] = E(X - Q(X))^2.$$

# Quantization: an example

- Take X to be a Gaussian random variable with mean zero and variance 400
- Consider the uniform 8 level quantizer of before

$$D = \sum_{i=1}^{8} \int_{\mathcal{R}_i} (x - Q\ (x))^2 f_X(x) dx,$$

or equivalently,

$$D = \int_{-\infty}^{a_1} (x - \hat{x}_1)^2 f_X(x) dx + \sum_{i=2}^{7} \int_{a_{i-1}}^{a_i} (x - \hat{x}_i)^2 f_X(x) dx$$

$$+ \int_{a_7}^{\infty} (x - \hat{x}_8)^2 f_X(x) dx, \qquad (7.2.2)$$

where $f_X(x)$ is $\frac{1}{\sqrt{2\pi 400}} e^{-\frac{x^2}{800}}$. Substituting $\{a_i\}_{i=1}^{7}$ and $\{\hat{x}_i\}_{i=1}^{8}$ in this integral and evaluating the result with the $Q$-function table, we obtain $D \approx 33.38$. Note that if we were to use zero bits per source output, then the best strategy would be to set the reconstructed signal equal to zero. In this case, we would have a distortion of $D = E(X - 0)^2 = \sigma^2 = 400$. This quantization scheme and transmission of 3 bits per source output has enabled us to reduce the distortion to 33.38, which is a factor of 11.98 reduction or 10.78 dB. ∎

# Quantization

- How do we compare across quantizers?
  - It's convenient to have a simple reference number

**Definition 7.2.1.** If the random variable $X$ is quantized to $Q(X)$, the *signal-to-quantization noise ratio* (SQNR) is defined by

$$\text{SQNR} = \frac{E(X^2)}{E(X - Q(X))^2}. \qquad (7.2.3)$$

■

# Quantization

- When we deal with a signal, we have deterministic quantities

When dealing with signals, the quantization noise power is

$$P_{\tilde{X}} = \lim_{T \to \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} E\left(X(t) - Q\left(X(t)\right)\right)^2 dt \tag{7.2.4}$$

and the signal power is

$$P_X = \lim_{T \to \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} E(X^2(t))dt. \tag{7.2.5}$$

# Quantization

Hence, the signal-to-quantization noise ratio is

$$\text{SQNR} = \frac{P_X}{P_{\tilde{X}}}. \tag{7.2.6}$$

If $X(t)$ is stationary, then this relation simplifies to Equation (7.2.3), where $X$ is the random variable representing $X(t)$ at any point.

# Distortion

- How do we optimize a quantizer?
- For a uniform quantizer, we can minimize the distortion as a function of the boundaries and the spacing
- For non uniform quantizers, we can use numerical techniques

# Minimum Distortion

$$D = \int_{-\infty}^{a_1} (x - (a_1 - \Delta/2))^2 f_X(x)dx + \sum_{i=1}^{N-2} \int_{a_1+(i-1)\Delta}^{a_1+i\Delta} (x - (a_1 + i\Delta - \Delta/2))^2 f_X(x)dx + \int_{a_1+(N-2)\Delta}^{\infty} (x - (a_1 + (N-2)\Delta + \Delta/2))^2 f_X(x)dx. \qquad (7.2.7)$$

# Minimum Distortion

- Numerical table are available for different distribution of sources
  - The classic example is the Gaussian one

**TABLE 7.1 OPTIMAL UNIFORM QUANTIZER FOR A GAUSSIAN SOURCE**

| Number Output Levels $N$ | Output-level Spacing $\Delta$ | Mean-squared Error $D$ | Entropy $H(\hat{x})$ |
|---|---|---|---|
| 1 | — | 1.000 | 0.0 |
| 2 | 1.596 | 0.3634 | 1.000 |
| 3 | 1.224 | 0.1902 | 1.536 |
| 4 | 0.9957 | 0.1188 | 1.904 |
| 5 | 0.8430 | 0.08218 | 2.183 |
| 6 | 0.7334 | 0.06065 | 2.409 |
| 7 | 0.6508 | 0.04686 | 2.598 |
| 8 | 0.5860 | 0.03744 | 2.761 |
| 9 | 0.5338 | 0.03069 | 2.904 |
| 10 | 0.4908 | 0.02568 | 3.032 |
| 11 | 0.4546 | 0.02185 | 3.148 |
| 12 | 0.4238 | 0.01885 | 3.253 |
| 13 | 0.3972 | 0.01645 | 3.350 |
| 14 | 0.3739 | 0.01450 | 3.440 |
| 15 | 0.3534 | 0.01289 | 3.524 |
| 16 | 0.3352 | 0.01154 | 3.602 |
| 17 | 0.3189 | 0.01040 | 3.676 |
| 18 | 0.3042 | 0.009430 | 3.746 |
| 19 | 0.2909 | 0.008594 | 3.811 |
| 20 | 0.2788 | 0.007869 | 3.874 |

# Minimum Distortion

$$D = \int_{-\infty}^{a_1} (x - \hat{x}_1)^2 f_X(x)dx + \sum_{i=1}^{N-2} \int_{a_i}^{a_{i+1}} (x - \hat{x}_{i+1})^2 f_X(x)dx$$

$$+ \int_{a_{N-1}}^{\infty} (x - \hat{x}_N)^2 f_X(x)dx. \quad (7.2.8)$$

There exists a total of $2N - 1$ variables in this expression ($a_1, a_2, \ldots, a_{N-1}$ and $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N$) and the minimization of $D$ is to be done with respect to these variables. Differentiating with respect to $a_i$ yields

$$\frac{\partial}{\partial a_i} D = f_X(a_i)[(a_i - \hat{x}_i)^2 - (a_i - \hat{x}_{i+1})^2] = 0, \quad (7.2.9)$$

which results in

$$a_i = \frac{1}{2}(\hat{x}_i + \hat{x}_{i+1}). \quad (7.2.10)$$

# Minimum Distortion

To determine the quantized values $\hat{x}_i$, we differentiate $D$ with respect to $\hat{x}_i$ and define $a_0 = -\infty$ and $a_N = +\infty$. Thus, we obtain

$$\frac{\partial}{\partial \hat{x}_i} = \int_{a_{i-1}}^{a_i} 2(x - \hat{x}_i) f_X(x) dx = 0, \tag{7.2.11}$$

which results in

$$\hat{x}_i = \frac{\int_{a_{i-1}}^{a_i} x f_X(x) dx}{\int_{a_{i-1}}^{a_i} f_X(x) dx}. \tag{7.2.12}$$

# An Example

**Example 7.2.3**

How would the results of Example 7.2.1 change if, instead of the uniform quantizer shown in Figure 7.3, we used an optimal nonuniform quantizer with the same number of levels?

# An Example

**Solution** We can find the quantization regions and the quantized values from Table 7.2 with $N = 8$, and then use the fact that our source is an $\mathcal{N}(0, 400)$ source, i.e., $m = 0$ and $\sigma = 20$. Therefore, all $a_i$ and $\hat{x}_i$ values read from the table should be multiplied by $\sigma = 20$ and the distortion has to be multiplied by 400. This gives us the values $a_1 = -a_7 = -34.96$, $a_2 = -a_6 = -21$, $a_3 = -a_5 = -10.012$, $a_4 = 0$ and $\hat{x}_1 = -\hat{x}_8 = -43.04$, $\hat{x}_2 = -\hat{x}_7 = -26.88$, $\hat{x}_3 = -\hat{x}_6 = -15.12$, $\hat{x}_4 = -\hat{x}_5 = -4.902$ and a distortion of $D = 13.816$. The SQNR is

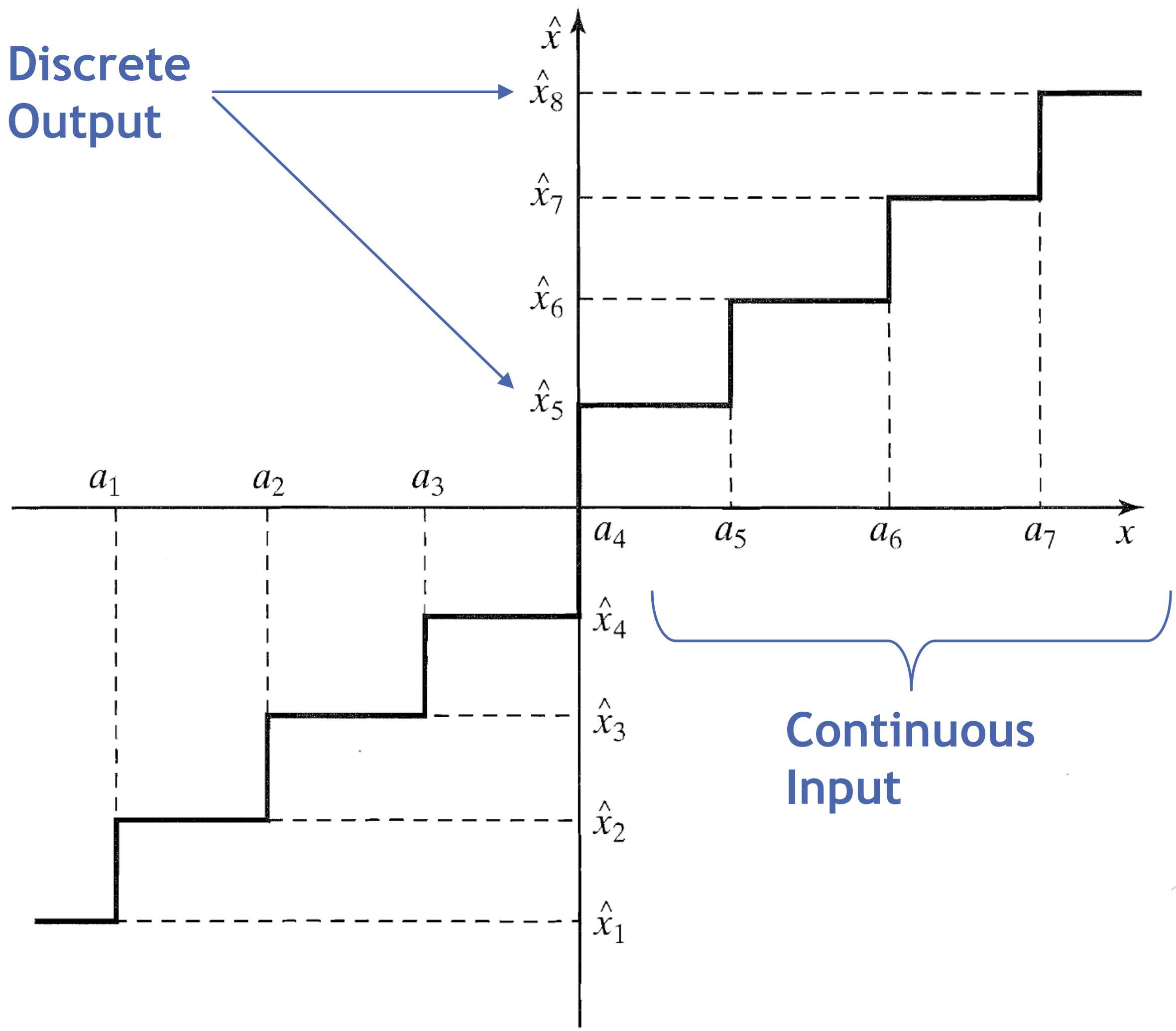
$$\text{SQNR} = \frac{400}{13.816} = 28.95 \approx 14.62 \text{ dB},$$

which is 3.84 dB better than the SQNR of the uniform quantizer. ■

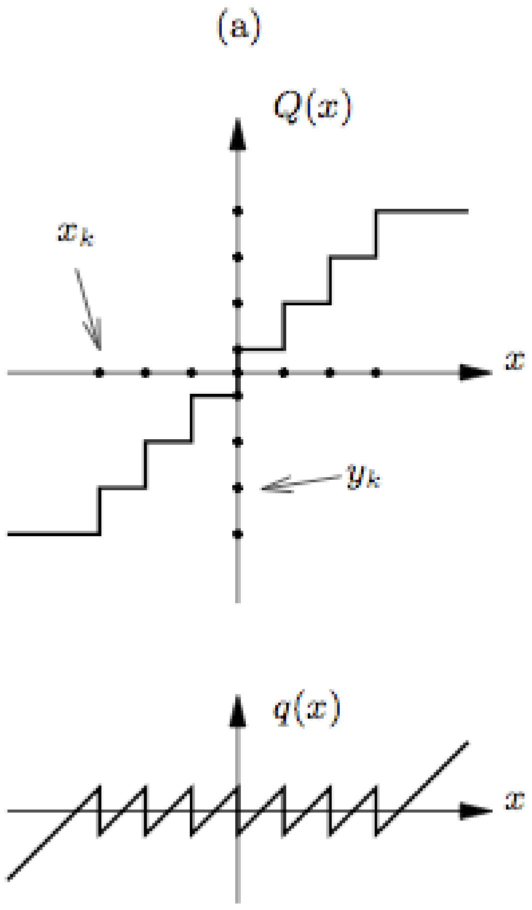
# Linear Quantizers

- Linear quantizers are not particularly good unless the source is uniform
- In general we can numerically optimize the location of the boundary points and the reconstruction points based on the source distribution

# Linear Quantizers



# Non-Linear Quantizers



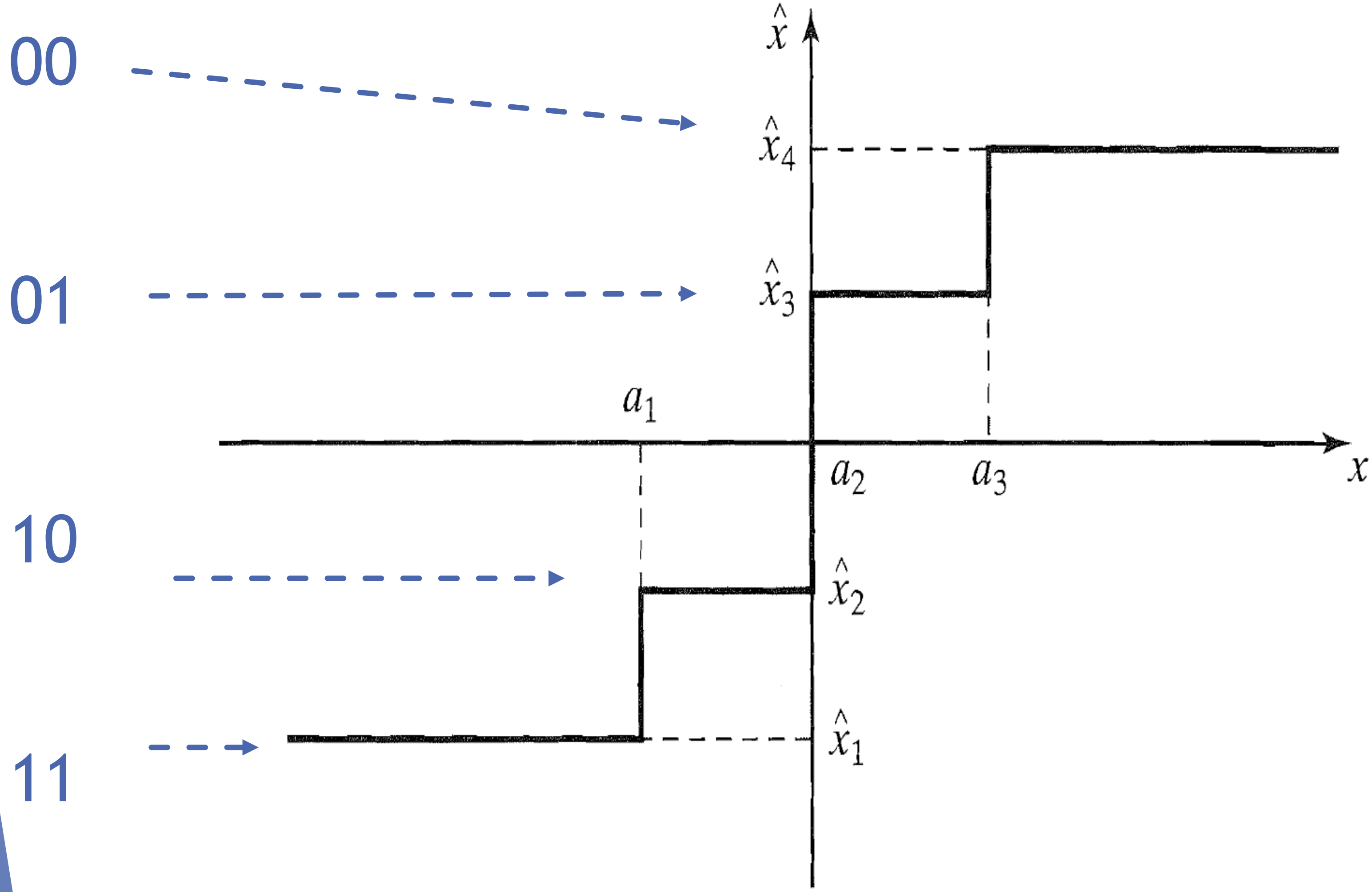
# A question

- So we quantize every sample

# Non-Linear Quantizers

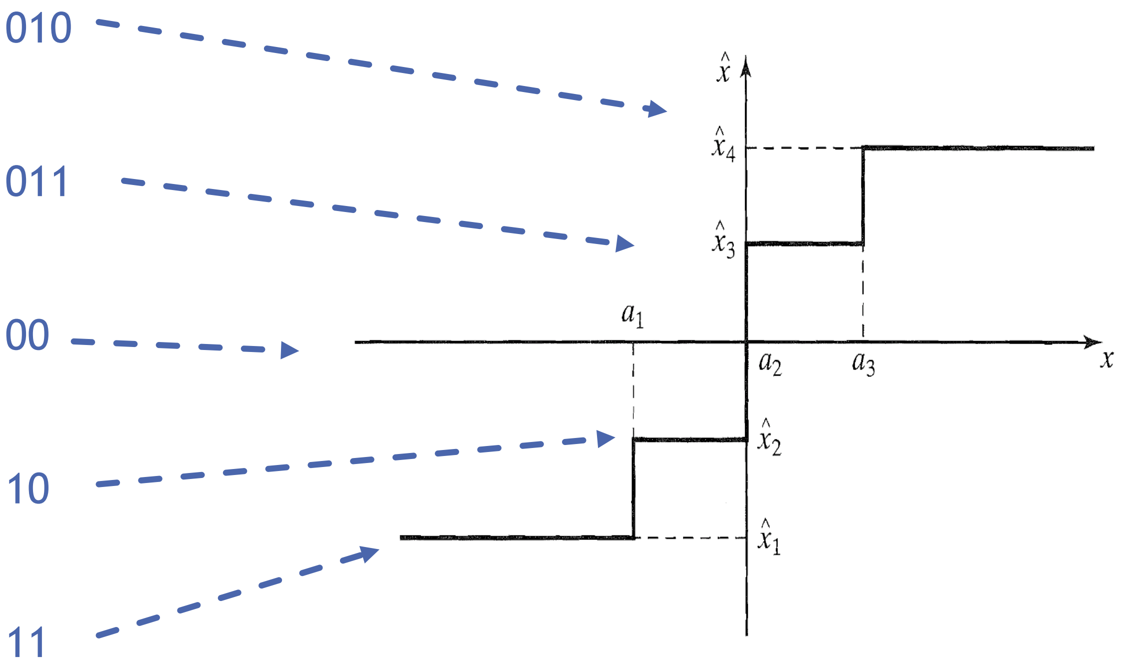
In scalar quantization, each output of the discrete-time source (which is usually the result of sampling of a continuous-time source) is quantized separately and then encoded. For example, if we are using a 4-level scalar quantizer and encoding each level into two bits, we are using two bits per each source output. This quantization scheme is shown in Figure 7.4.

# After quantization we have a bit-stream



# After quantization we have a bit-stream

- Now a question for you:
  - How would things change if we would us different length labels for each point?



# Vector Quantizers

- So we are sampling each sample separately
  - Is there any advantage in sampling more samples at one time

# Vector Quantizers

- Think about it…

# Vector Quantizers

- Think about it some more…

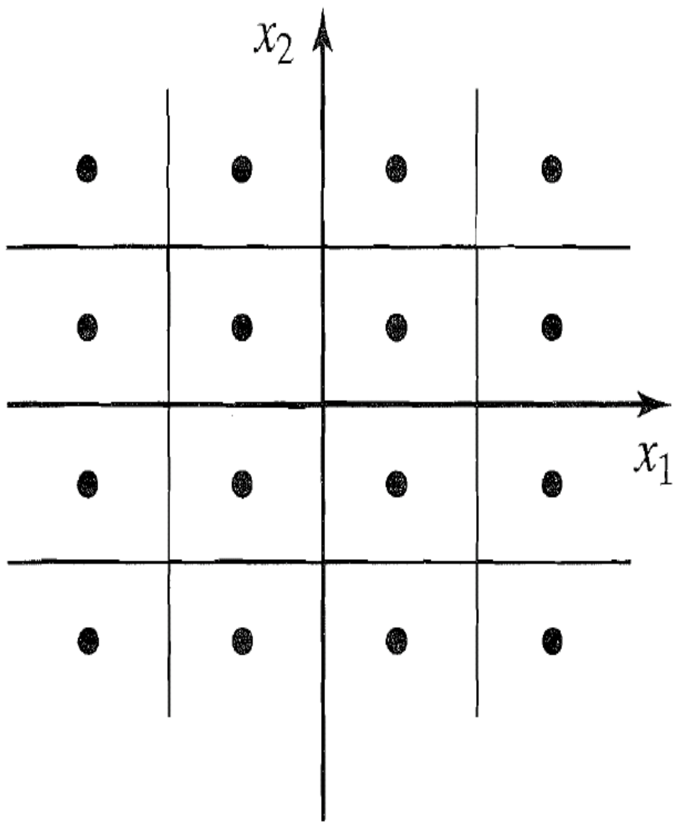
# Vector Quantizers

- Let's find a way of imagining a quantizer for vectors of samples

Now if we consider two samples of the source at each time, and we interpret these two samples as a point in a plane, the quantizer partitions the entire plane into 16 quantization regions, as show in Figure 7.5. We can see that the regions in the two-dimensional space are all of rectangular shape. If we allow 16 regions of any shape in the two-dimensional space, we are capable of obtaining better results.

# Vector Quantizers

Quantization of
the second sample

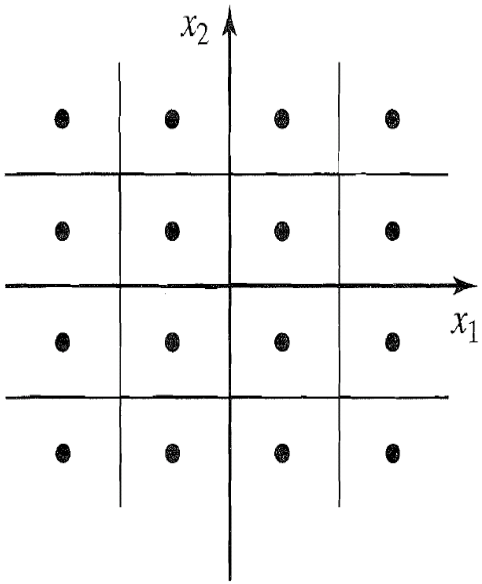



Quantization of
the first sample

# Vector Quantizers

- How do you think we can improve on this quantizer?
  - Remember that the samples have the samp distribution, so that does not play a role

Quantization of
the second sample




Quantization of
the first sample

# Vector Quantizers

- How can we build a vector quantizer then?

Let us assume that the quantization regions in the $n$-dimensional space are denoted by $\mathscr{R}_i$, $1 \leq i \leq K$. These $K$ regions partition the $n$-dimensional space. Each block of source output of length $n$ is denoted by the $n$-dimensional vector $\mathbf{x} \in R^n$; if $\mathbf{x} \in \mathscr{R}_i$, it is quantized to $Q(\mathbf{x}) = \hat{\mathbf{x}}_i$. Figure 7.6 shows this quantization scheme for $n = 2$. Now, since there are a total of $K$ quantized values, $\log K$ bits are enough to represent these values. This means that we require $\log K$ bits per $n$ source outputs, or the rate of the source code is

$$R = \frac{\log K}{n} \quad \text{bits/source output.} \tag{7.2.13}$$

# Vector Quantizers

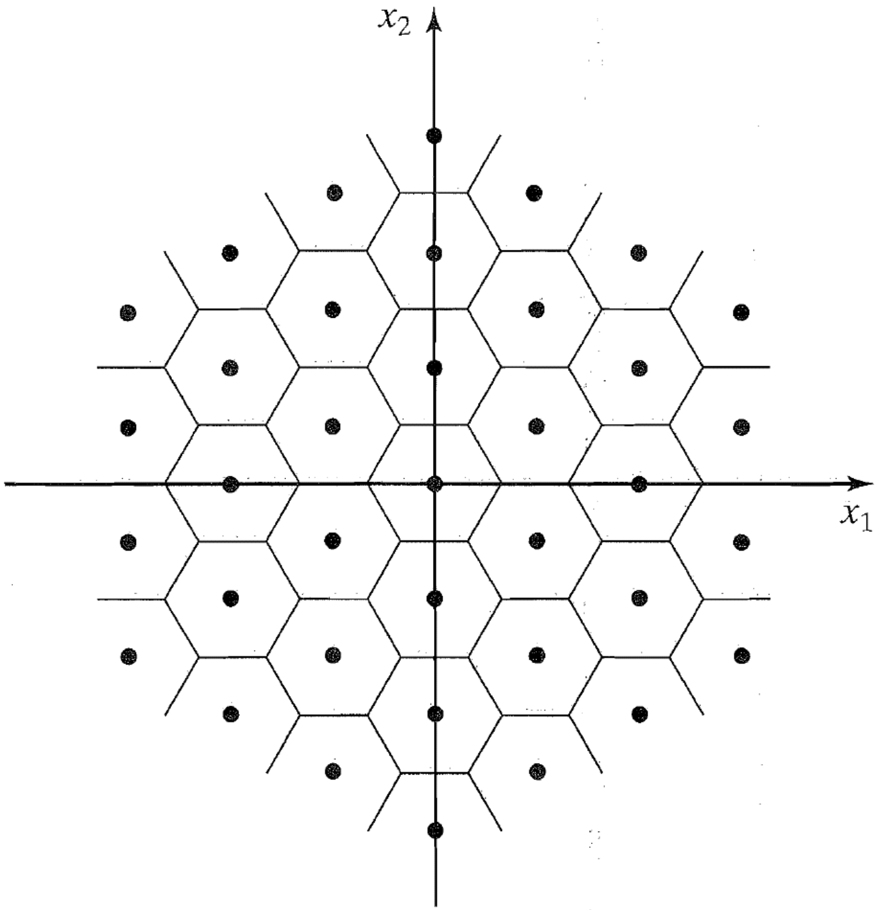



**Figure 7.6** Vector quantization in two dimensions.

# Vector Quantizers

1. Region $\mathcal{R}_i$ is the set of all points in the $n$-dimensional space that are closer to $\hat{\mathbf{x}}_i$ than any other $\hat{\mathbf{x}}_j$, for all $j \neq i$; that is,

$$\mathcal{R}_i = \{\mathbf{x} \in R^n : \|\mathbf{x} - \hat{\mathbf{x}}_i\| < \|\mathbf{x} - \hat{\mathbf{x}}_j\|, \ \forall j \neq i\};$$

2. $\hat{\mathbf{x}}_i$ is the centroid of the region $\mathcal{R}_i$; that is,

$$\hat{\mathbf{x}}_i = \frac{1}{P(\mathbf{X} \in \mathcal{R}_i)} \int \cdots \int_{\mathcal{R}_i} \mathbf{x} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

# That's it for today

I shall see you next time