Compression Fundamentals

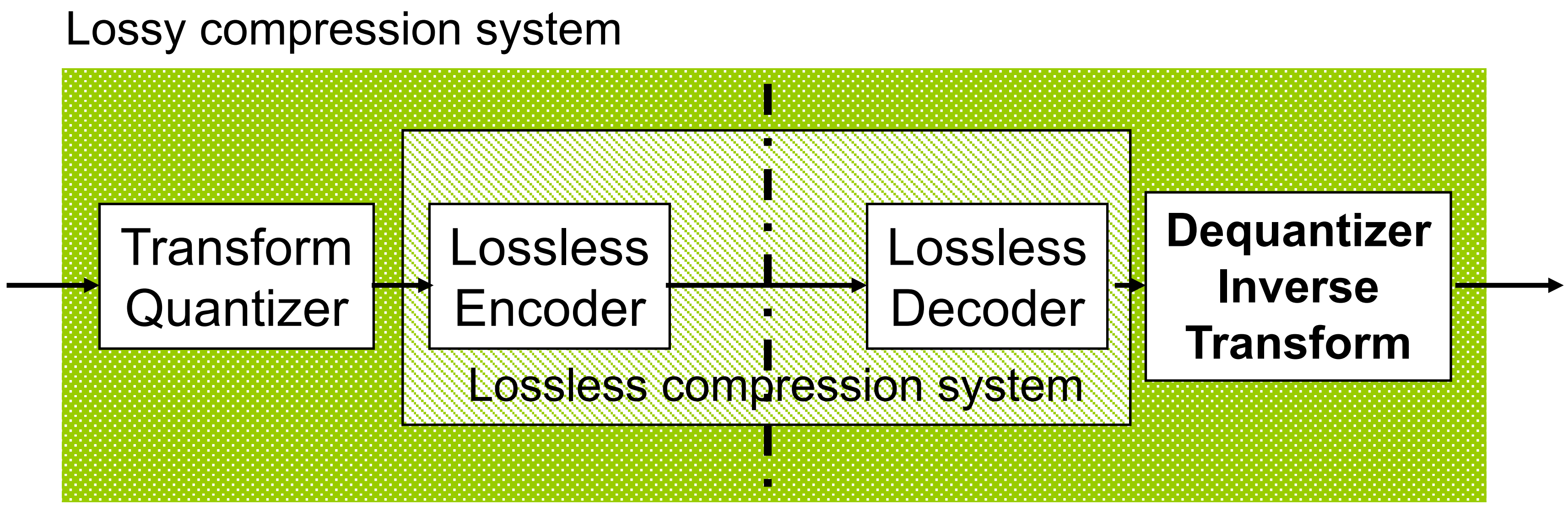
- Lossless compression
  - Information and entropy
  - Noiseless source coding theorem
  - Huffman code
- Lossy compression
  - Rate-distortion theory
  - Noisy source coding theorem
  - Quantization
  - Lloyd-Max quantizer

How does Compression Work?

- Exploit statistical redundancy
  - Take advantage of patterns in the signal
  - Describe frequently occurring events efficiently
  - **Lossless coding:** only statistical redundancy
- Introduce acceptable deviations
  - Omit information that the humans cannot perceive
  - Match the signal resolution (in space, time, amplitude) to the application
  - **Lossy coding:** exploit both visual <u>and</u> statistical redundancy

Lossless Compression in Lossy Compression Systems

- Almost every lossy compression system contains a lossless compression system



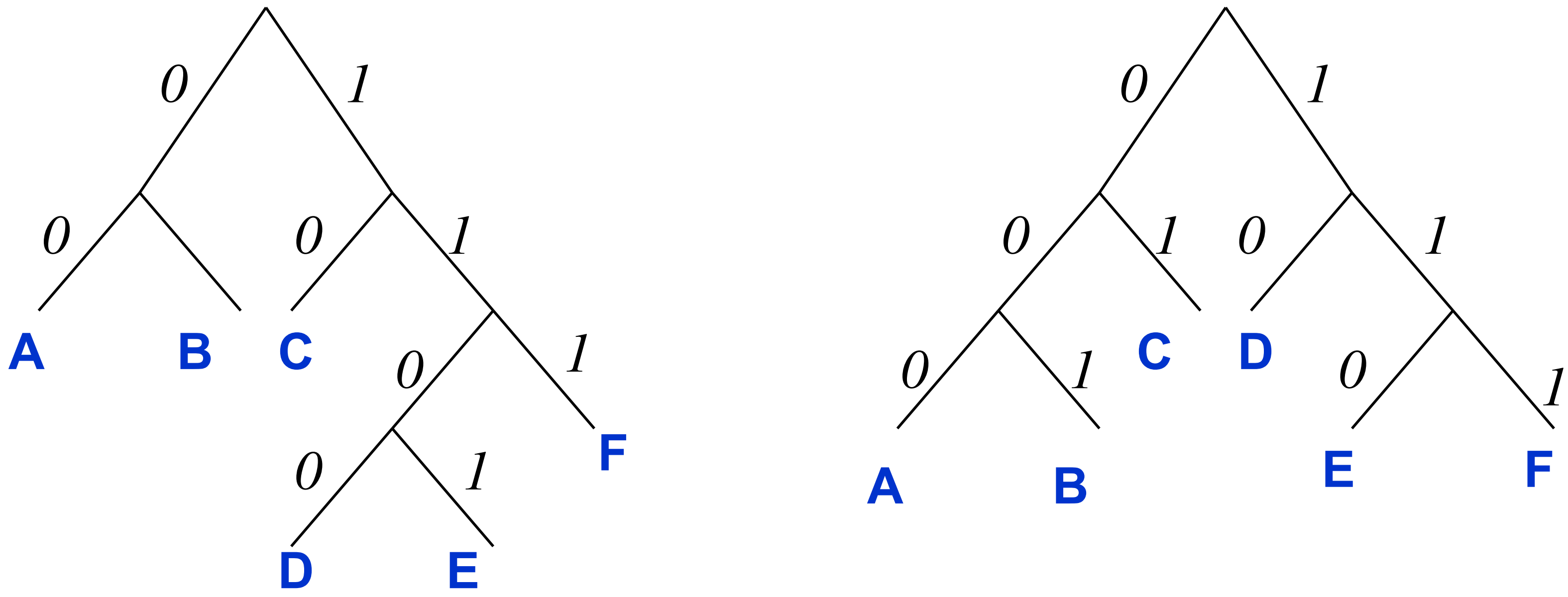
- We will discuss the basics of lossless compression first, then move on to lossy compression

Example: 20 Questions

- *Alice* thinks of an outcome (from a finite set), but does not disclose her selection.
- *Bob* asks a series of yes-no questions to uniquely determine the outcome chosen. The goal of the game is to ask as few questions as possible **on average**.
- **Our goal:** Design the best strategy for *Bob.*

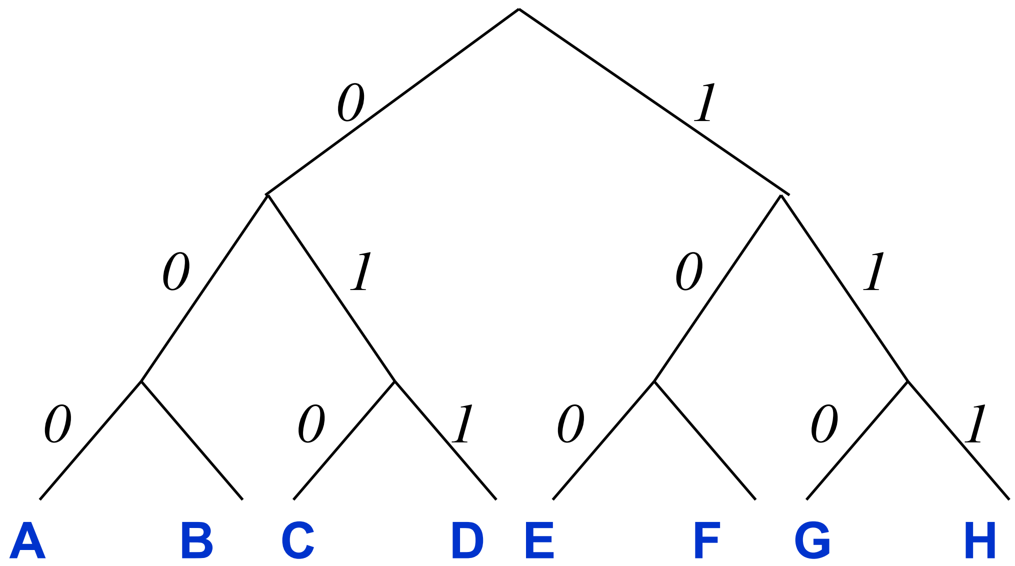
Example: 20 Questions

- Observation: The collection of questions and answers yield a binary code for each outcome.



- Which strategy (=code) is better?

Fixed Length Codes



- Average description length for *K* outcomes $l_{av} = \log_2 K$
- Optimum for equally likely outcomes
- Verify by modifying tree

Variable Length Codes

- If outcomes are NOT equally probable:
  - Use shorter descriptions for likely outcomes
  - Use longer descriptions for less likely outcomes
- Intuition:
  - Optimum balanced code trees, i.e., with equally likely outcomes, can be pruned to yield unbalanced trees with unequal probabilities.
  - The unbalanced code trees such obtained are also optimum.
  - Hence, an outcome of probability *p* should require about

$$\log_2\left(\frac{1}{p}\right) \text{ bits}$$

Entropy of a Random Variable

- Consider a discrete, finite-alphabet random variable *X*

$$\mathcal{A}_X = \{\alpha_0, \alpha_1, \ldots, \alpha_{K-1}\}$$

$$f_X(x) = P(X = x) \quad \forall x \in \mathcal{A}_X$$

- **"Information"** associated with the event *X*=*x*

$$h_X(x) = -\log_2 f_X(x)$$

- **"Entropy of *X*"** is the expected value of that information

$$H(X) = E\{h_X(X)\} = -\sum_{x \in \mathcal{A}_X} f_X(x) \log_2 f_X(x)$$

- Unit: bits

Information and Entropy: Properties

- Information $h_X(x) \geq 0$
- Information $h_X(x)$ strictly increases with decreasing probability $f_X(x)$
- Boundedness of entropy

$$0 \leq H(X) \leq \log_2(|\mathcal{A}_X|)$$

equality if only one outcome can occur

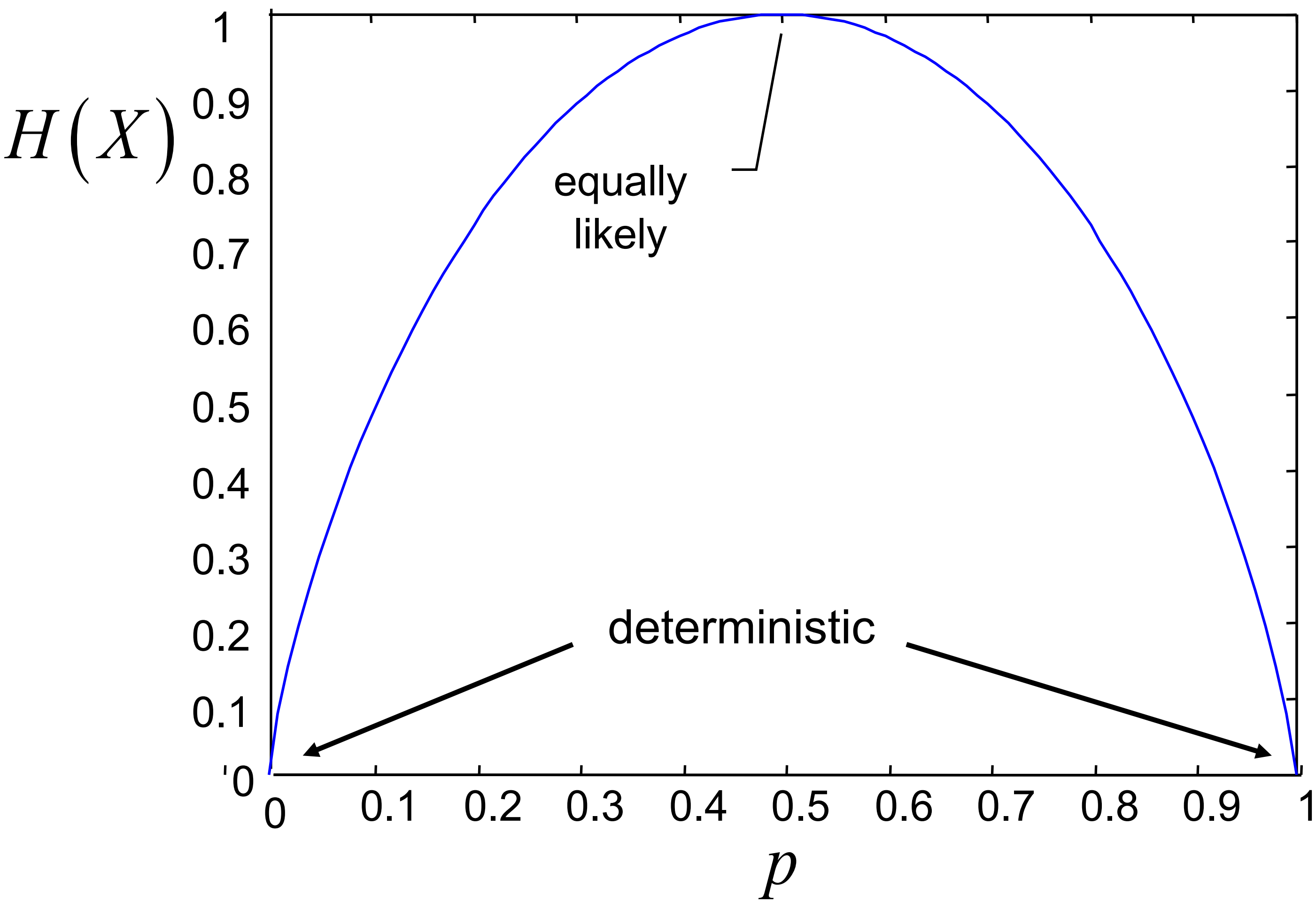
equality if all outcomes are equally likely

- Very likely and very unlikely events do not substantially change entropy

$$-p\log_2 p \to 0 \quad \text{for } p \to 0 \text{ or } p \to 1$$

Example: Binary Random Variable

$$H(X) = -p\log_2 p - (1-p)\log_2(1-p)$$



Entropy and Bit-Rate

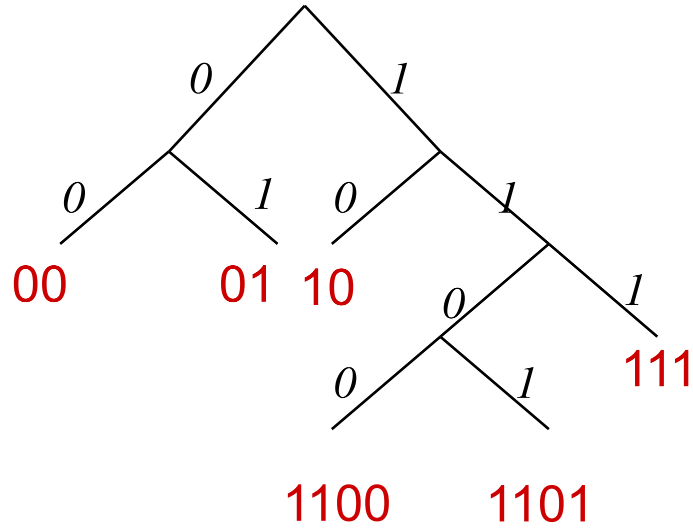
- Consider IID random process $\{X_n\}$ (or “source”) where each sample X_n (or “symbol”) possesses identical entropy *H(X)*
- *H(X)* is called “entropy rate” of the random process.
- **Noiseless Source Coding Theorem (Shannon, 1948):**
  – The entropy *H(X)* is a lower bound for the average word length *R* of a decodable variable-length code for the symbols.
  – Conversely, the average word length *R* can approach *H(X),* if sufficiently large blocks of symbols are encoded jointly.
- Redundancy of a code: $\boxed{\rho = R - H(X) \geq 0}$

Variable Length Codes

- Given IID random process $\{X_n\}$ with alphabet $\mathcal{A}_X$ and PMF $f_X(x)$
- Task: assign a distinct code word, c_x, to each element, $x \in \mathcal{A}_X$, where c_x is a string of $\|c_x\|$ bits, such that each symbol x_n can be determined from a sequence of concatenated codewords c_{x_n}
- Codes with the above property are said to be **"uniquely decodable"**
- Prefix codes
  - No code word is a prefix of any other codeword
  - Uniquely decodable, symbol by symbol, in natural order *0, 1, 2, . . . , n, . . .*

Binary Trees and Prefix Codes

- Each binary tree can be converted into a prefix code by traversing the tree from root to leaves.



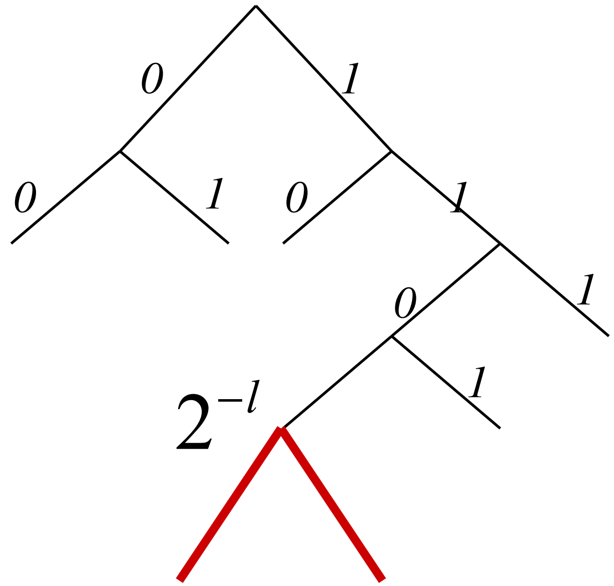
- Each prefix code corresponding to a binary tree meet McMillan condition with equality

$$3 \cdot 2^{-2} + 2 \cdot 2^{-4} + 2^{-3} = 1$$

$$\boxed{\sum_{x \in \mathcal{A}_X} 2^{-\|c_x\|} = 1}$$

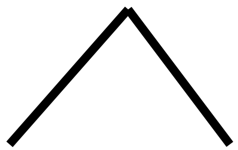
Binary Trees and Prefix Codes

- Augmenting binary tree by two new nodes does not change McMillan sum.
- Pruning binary tree does not change McMillan sum.



$$2^{-(l+1)} + 2^{-(l+1)} = 2^{-l}$$

- McMilllan sum for simplest binary tree



$$2^{-1} + 2^{-1} = 1$$

Instantaneous Variable Length Encoding without Redundancy

- A code without redundancy, i.e.,

$$R = H(X)$$

requires all individual code word lengths

$$l_{\alpha_k} = -\log_2 f_X(\alpha_k)$$

- All probabilities would have to be binary fractions:

$$f_X(\alpha_k) = 2^{-l_{\alpha_k}}$$

Example

α_i	$P(\alpha_i)$	**redundant code**	**optimum code**
α_0	**0.500**	**00**	**0**
α_1	**0.250**	**01**	**10**
α_2	**0.125**	**10**	**110**
α_3	**0.125**	**11**	**111**

$$H(X) = 1.75 \text{ bits}$$

$$R = 1.75 \text{ bits}$$

$$\rho = 0$$

Huffman Code

- Design algorithm for variable length codes proposed by Huffman (1952) always finds a code with minimum redundancy.
- Obtain code tree as follows:

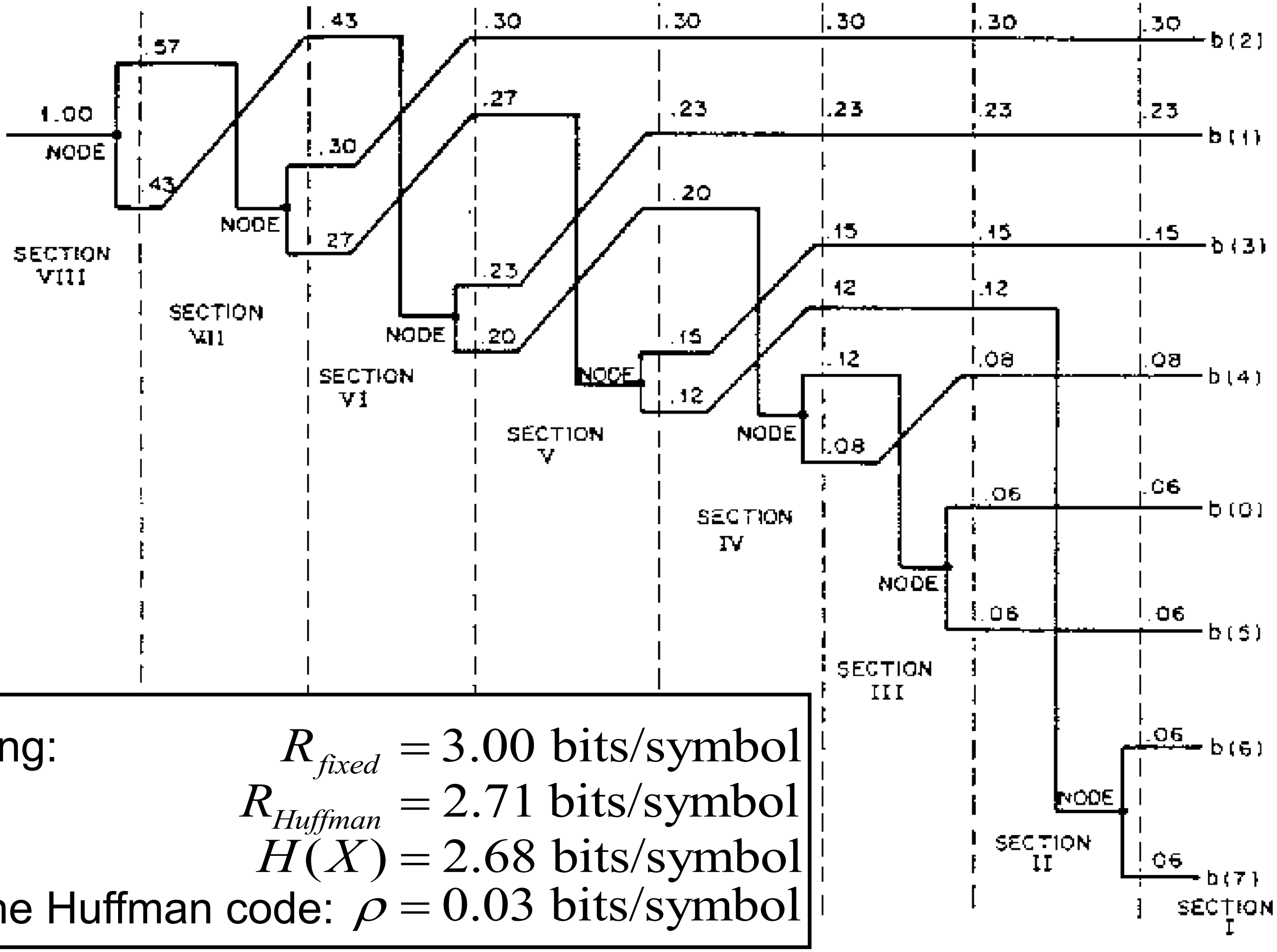
1 Pick the two symbols with lowest probabilities and merge them into a new auxiliary symbol.

2 Calculate the probability of the auxiliary symbol.

3 If more than one symbol remains, repeat steps **1** and **2** for the new auxiliary alphabet.

4 Convert the code tree into a prefix code.

Example: Huffman Code



Fixed length coding: $R_{fixed} = 3.00$ bits/symbol
Huffman code: $R_{Huffman} = 2.71$ bits/symbol
Entropy $H(X) = 2.68$ bits/symbol
Redundancy of the Huffman code: $\rho = 0.03$ bits/symbol

Redundancy of Prefix Code for General Distribution

- Huffman code redundancy $0 \le \rho < 1$ bit/symbol
- **Theorem:** For any distribution f_X, a prefix code may be found, whose rate R satisfies

$$\boxed{H(X) \le R < H(X) + 1}$$

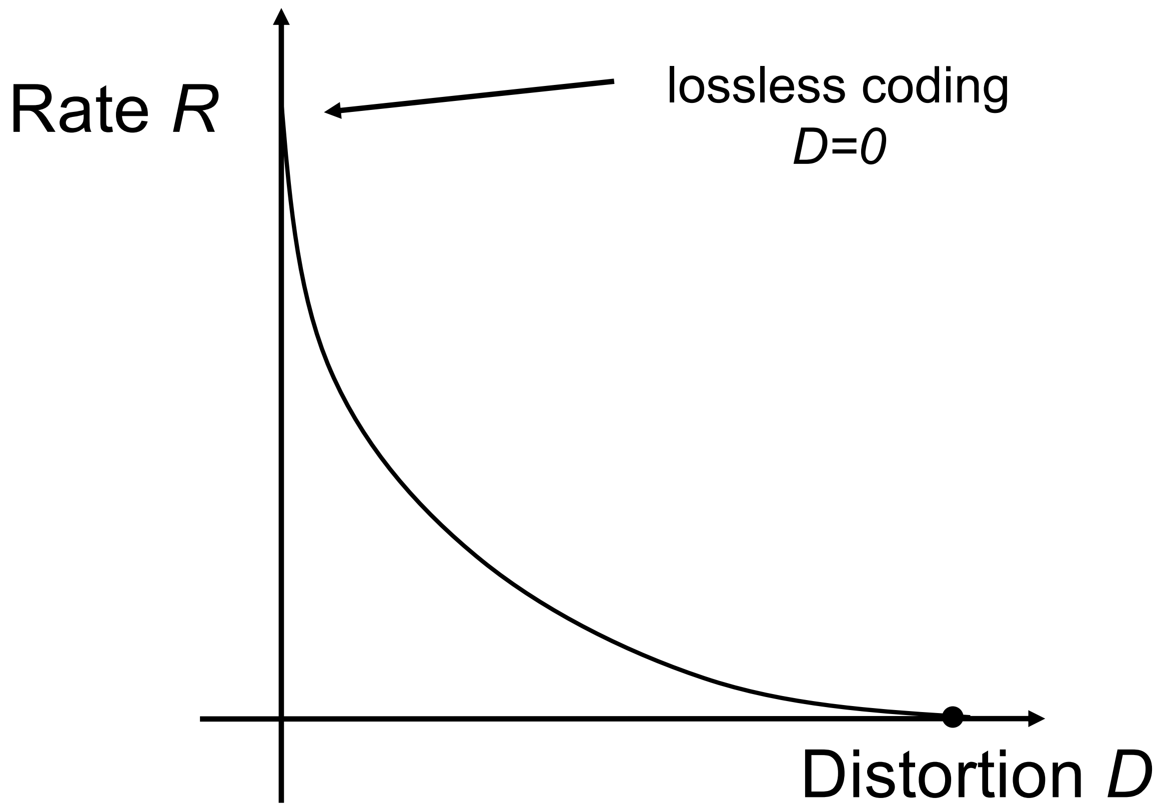
- Proof:
  - Left hand inequality: Shannon's noiseless coding theorem
  - Right hand inequality:

$$\text{Choose code word lengths} \quad \|c_x\| = \left\lceil -\log_2 f_X(x) \right\rceil$$

$$\begin{aligned} \text{Resulting rate} \qquad R &= \sum_{x \in \mathrm{A}_X} f_X(x) \left\lceil -\log_2 f_X(x) \right\rceil \\ &< \sum_{x \in \mathrm{A}_X} f_X(x) \left(1 - \log_2 f_X(x)\right) \\ &= H(X) + 1 \end{aligned}$$

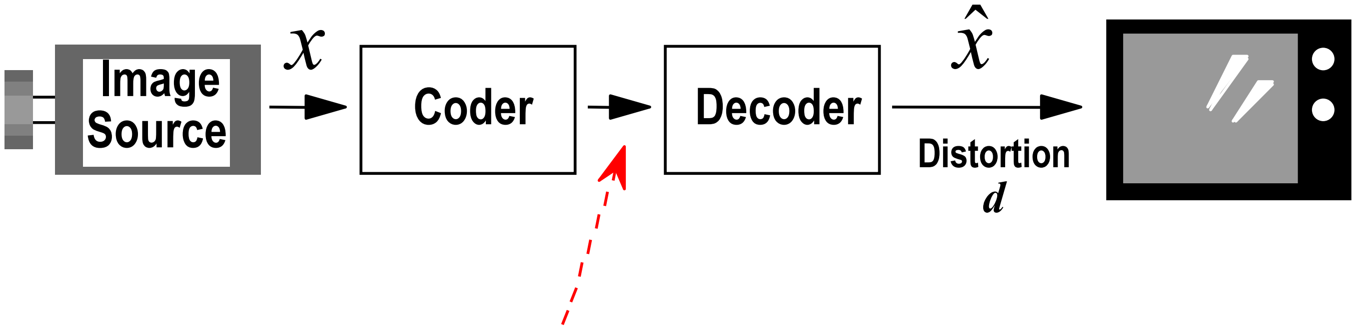
Lossy Compression

- Lower the bit-rate *R* by allowing some acceptable distortion *D* of the signal



Rate Distortion Theory

- Rate distortion theory calculates the minimum transmission bit-rate R for a required picture quality



Bitrate at least R

for distortion $d \leq D$

- Results of rate distortion theory are obtained without consideration of a specific coding method

Distortion

- Symbol (signal, image . . .) x sent, $\hat{x}$ received
- Single-letter distortion measure:

$$\boxed{\begin{array}{l}\rho(x,\hat{x}) \geq 0 \\ \rho(x,\hat{x}) = 0 \quad \text{for } x = \hat{x}\end{array}}$$

- Average distortion:

$$\boxed{d(x,\hat{x}) = E\{\rho(x,\hat{x})\} = \sum_x \sum_{\hat{x}} f_{X,\hat{X}}(x,\hat{x})\,\rho(x,\hat{x})}$$

- Distortion criterion: $\boxed{d(x,\hat{x}) \leq D}$

Maximum permissible average distortion

Joint and Conditional Entropy

- Consider two discrete finite-alphabet r.v. *X* and *Y*

$$\boxed{\begin{aligned} H(X|Y) &= E\left[-\log_2 f_{X|Y}(x,y)\right] = -\sum_y \sum_x f_{X,Y}(x,y) \log_2 f_{X|Y}(x,y) \\ &= -\sum_y f_Y(y) \sum_x f_{X|Y}(x,y) \log_2 f_{X|Y}(x,y) \end{aligned}}$$

- Conditional entropy *H(X|Y)* is average additional information, if *Y* is already known
- Joint entropy: $H(X,Y) = E\left[-\log_2 f_{X,Y}(X,Y)\right]$

$$= E\left[-\log_2 \left(f_Y(y) f_{X|Y}(X,Y)\right)\right]$$

$$= E\left[-\log_2 f_Y(y)\right] + E\left[-\log_2 f_{X|Y}(X,Y)\right]$$

$$= H(Y) + H(X|Y)$$

Mutual Information

- "Mutual information" is the average information that random variables *X* and *Y* convey about each other
  - Reduction in uncertainty about x, if y is observed
  - Reduction in uncertainty about y, if x is observed

$$
\boxed{\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= \sum_x \sum_y f_{X,Y}(x,y) \log_2 \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \end{aligned}}
$$

- Properties $0 \le I(X;Y) = I(Y;X)$

$$I(X;Y) \le H(X)$$

$$I(X;Y) \le H(Y)$$

Rate Distortion Function

- Definition:

$$\boxed{R(D) = \inf_{f_{\hat{X}|X}: d(x,\hat{x}) \leq D} \{I(X;\hat{X})\}}$$

- **Shannon's Noisy Source Coding Theorem:**

  For a given maximum average distortion *D*, the rate distortion function *R(D)* is the (achievable) lower bound for the transmission bit-rate.

- *R(D)* is continuous, monotonically decreasing for *R>0* and convex
- Equivalently use distortion-rate function *D(R)*

Extension to Continuous Random Variables

- Differential entropy

$$h(X) = -E\{\log_2 f_X(X)\} = -\int_x f_X(x) \log_2 f_X(x) dx$$

- Differential conditional entropy

$$h(X|Y) = -E\{\log_2 f_{X|Y}(X,Y)\} = -\iint_{x,y} f_{X,Y}(x,y) \log_2 f_{X|Y}(x,y) dxdy$$

- Mutual information

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$$

- Rate distortion function: $R(D) = \inf_{f_{\hat{X}|X} : d(x,\hat{x}) \le D} \{I(X;\hat{X})\}$

Shannon Lower Bound

- It can be shown that $h(X-\hat{X} \mid \hat{X}) = h(X \mid \hat{X})$
- Thus

$$R(D) = \inf_{d \le D} \{h(X) - h(X \mid \hat{X})\}$$

$$= h(X) - \sup_{d \le D} \{h(X \mid \hat{X})\}$$

$$= h(X) - \sup_{d \le D} \{h(X - \hat{X} \mid \hat{X})\}$$

- Ideally, the source coder would introduce errors $x - \hat{x}$ that are statistically independent from the reconstructed signal $\hat{x}$ (not always possible!).
- Shannon lower bound:

$$\boxed{R(D) \ge h(X) - \sup_{d \le D} h(X - \hat{X})}$$

Shannon Lower Bound

- Mean squared error distortion measure: Gaussian PDF possesses largest entropy for given variance

$$\boxed{\begin{aligned} R(D) &\geq h(X) - \sup_{d \leq D} h(X - \hat{X}) \\ &= h\left(X\right) - \tfrac{1}{2}\log_2 2\pi e D \end{aligned}}$$

- Distortion reduction by 6 dB requires 1 bit/sample

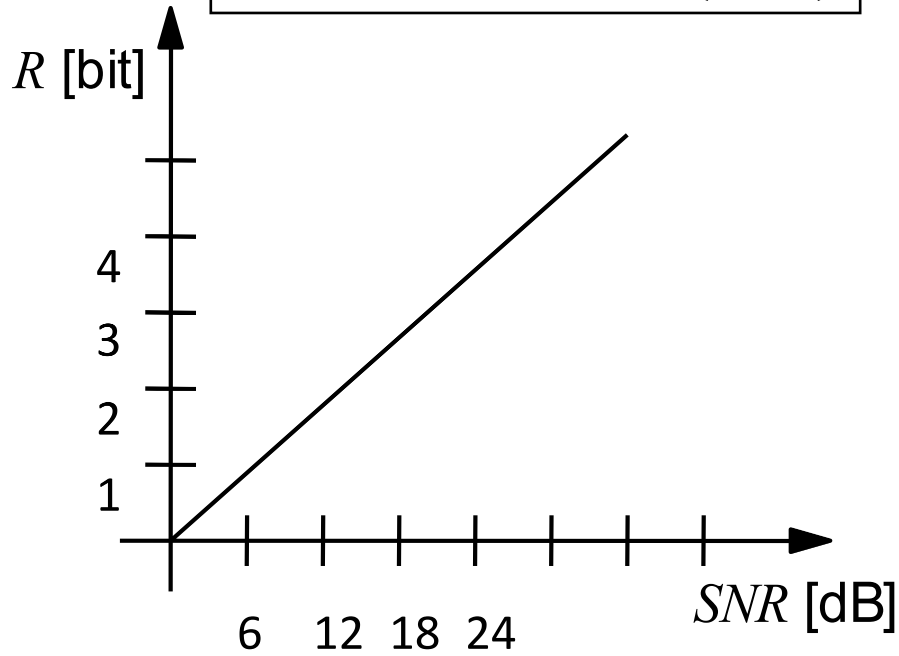
$R(D)$ Function for a Memoryless Gaussian Source and MSE Distortion

- Gaussian source, variance σ^2
- Mean squared error

$$d = E\{(X-\hat{X})^2\} \le D$$

- Rule of thumb: 6 dB $\cong$ 1 bit
- *R(D)* for non-Gaussian sources with the same variance σ^2 is always <u>below</u> this Gaussian *R(D)* curve.

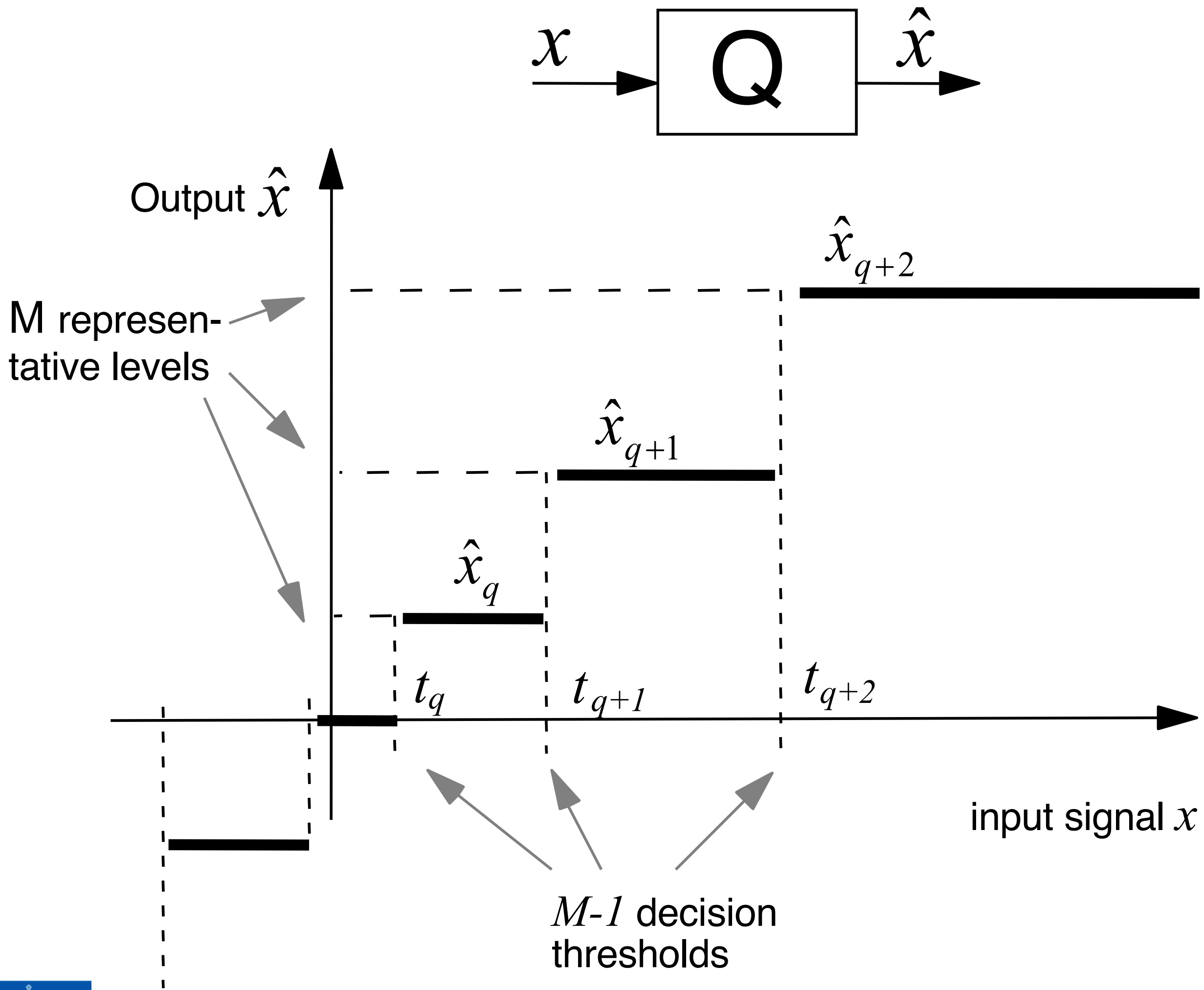
$$\boxed{R(D) = \frac{1}{2}\log_2\left(\frac{\sigma^2}{D}\right)}$$

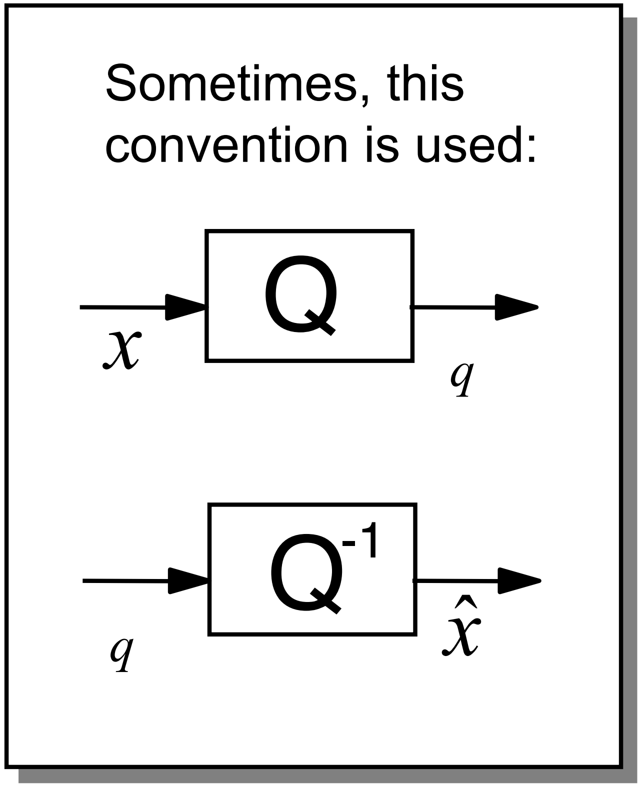


$$SNR = 10\log_{10}\left(\frac{\sigma^2}{D}\right) \text{ [dB]}$$

Quantization

Input-output characteristic of a scalar quantizer





Lloyd-Max Scalar Quantizer

- Problem: For a signal x with given PDF $f_X(x)$ find a quantizer with M representative levels such that

$$\boxed{d = MSE = E\left[\left(X - \hat{X}\right)^2\right] \to \min.}$$

- Solution: Lloyd-Max quantizer

  [Lloyd, 1957][Max, 1960]

  - *M-1* decision thresholds exactly half-way between representative levels.
  - *M* representative levels in the centroid of the PDF between two successive decision thresholds.
  - Necessary condition

$$t_q = \frac{1}{2}\left(\hat{x}_{q-1} + \hat{x}_q\right) \quad q = 1, 2, \ldots, M-1$$

$$\hat{x}_q = \frac{\int_{t_q}^{t_{q+1}} x f_X(x)\,dx}{\int_{t_q}^{t_{q+1}} f_X(x)\,dx} \quad q = 0, 1, \ldots, M-1$$

Iterative Lloyd-Max Quantizer Design

1. Guess initial set of representative levels $\hat{x}_q \quad q = 0,1,2,\ldots,M-1$
2. Calculate decision thresholds

$$t_q = \frac{1}{2}\left(\hat{x}_{q-1} + \hat{x}_q\right) \quad q = 1,2,\ldots,M-1$$

3. Calculate new representative levels

$$\hat{x}_q = \frac{\int_{t_q}^{t_{q+1}} x \cdot f_X(x)dx}{\int_{t_q}^{t_{q+1}} f_X(x)dx} \quad q = 0,1,\ldots,M-1$$

4. Repeat **2.** and **3.** until no further distortion reduction

Lloyd-Max Quantizer Properties

- Zero-mean quantization error

$$E\left[\left(X-\hat{X}\right)\right]=0$$

- Quantization error and reconstruction decorrelated

$$E\left[\left(X-\hat{X}\right)\hat{X}\right]=0$$

- Variance subtraction property

$$\sigma_{\hat{X}}^2=\sigma_X^2-E\left[\left(X-\hat{X}\right)^2\right]$$

- Equal MSE contributions

$$\Pr\left\{t_i \le X < t_{i+1}\right\} E\left[\left(X-\hat{X}\right)^2 \middle| t_i \le X < t_{i+1}\right]$$

$$=\Pr\left\{t_j \le X < t_{j+1}\right\} E\left[\left(X-\hat{X}\right)^2 \middle| t_j \le X < t_{j+1}\right] \quad \text{for all } i, j$$

Deadzone Uniform Quantizer