Lecture 7: Source Coding
Quantization

• $X \sim p_X$ is a real-valued random variable with range $\mathcal{X}$.

• A $Q$-points scalar quantization function $f$ is a mapping $f : \mathcal{X} \rightarrow \{0, \ldots, Q-1\}$.

• $f$ is defined by decision regions $\mathcal{D}_q$, such that $\mathcal{D}_q = \{x \in \mathcal{X} : f(x) = q\}$.

• Reconstruction function: $g : \{0, \ldots, Q - 1\} \rightarrow \hat{\mathcal{X}} \subseteq \mathbb{R}$.

• Reconstruction criterion (distortion measure): $d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}_+.$

• Average distortion $D = \mathbb{E}[d(X, g(f(X)))]$. 
Example: quadratic distortion

- We let \( d(x, \hat{x}) = |x - \hat{x}|^2 \), then \( D = \mathbb{E} \left[ |X - \hat{X}|^2 \right] \) where \( \hat{X} = g(f(X)) \).

- For a given set of reconstruction points \( \hat{X} = \{\hat{x}_0, \ldots, \hat{x}_{Q-1}\} \) it is well-known that the optimal decision regions are minimum distance regions (Voronoi regions)

\[
D_q = \{ x \in \mathcal{X} : |x - \hat{x}_q| \leq |x - \hat{x}_j| \ \forall \ j \neq q \}
\]

- For a given set of decision regions \( \{D_0, \ldots, D_{Q-1}\} \), the optimal reconstruction points are the conditional mean values:

\[
\hat{x}_q = \frac{\int_{D_q} x p_X(x) dx}{\int_{D_q} p_X(x) dx}
\]
Problem setup (1)

• Let $X^n$ be an i.i.d. discrete random vector with probability distribution $P_X$ on the discrete and finite alphabet $\mathcal{X}$.

• A Lossy Source Code with block length $n$ and rate $R$ is formed by an encoding mapping

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$$

and a decoding mapping

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$$

where $\hat{\mathcal{X}}$ the reconstruction alphabet (not necessarily identical to $\mathcal{X}$).

• Additive distortion measure: let $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$. We define the distortion between two sequences $x \in \mathcal{X}^n$ and $\hat{x} \in \hat{\mathcal{X}}^n$ as

$$d(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$$
Problem setup (2)

- The average distortion is defined by

\[ D = \mathbb{E} \left[ d(X^n, g_n(f_n(X^n))) \right] \]

- In the following, we use the notation \( \hat{X}^n(m) \) to indicate the \( m \)-th reconstruction codeword when the reconstruction codebook is generated at random, and \( M = f_n(X^n) \) denotes the random encoded message, when \( f_n(\cdot) \) is applied to the random vector \( X^n \sim P_{X^n} \).

**Definition 22.** A rate-distortion pair \((R, D)\) is achievable if there exists a sequence of source codes of rate \( R \) and increasing block length \( n \) such that
\[
\lim_{n \to \infty} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D.
\]

**Definition 23.** The rate-distortion region for a source \( \{X_i\} \), distortion measure \( d \) and reconstruction alphabet \( \hat{X} \) is the closure the \((R, D)\) achievable pairs.
Definition 24. The rate-distortion function, denoted by $R(D)$, for a source \( \{X_i\} \), distortion measure \( d \) and reconstruction alphabet \( \hat{\mathcal{X}} \), is defined as

\[
R(D) = \inf \{ R : (R, D) \text{ is achievable} \}
\]

Definition 25. The distortion-rate function, denoted by $D(R)$, for a source \( \{X_i\} \), distortion measure \( d \) and reconstruction alphabet \( \hat{\mathcal{X}} \), is defined as

\[
D(R) = \inf \{ D : (R, D) \text{ is achievable} \}
\]
Theorem 19. Lossy source coding for DMSs: Let \( \{X_i\} \) be a discrete memoryless source with \( X_i \sim P_X \), over \( \mathcal{X} \), let \( \hat{\mathcal{X}} \) be a discrete reconstruction alphabet and \( d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}_+ \) be a bounded distortion measure. Then,

\[
R(D) = \min_{P_{\hat{X} | X} : \mathbb{E}[d(X, \hat{X})] \leq D} I(X; \hat{X})
\]

where \( (X, \hat{X}) \sim P_X P_{\hat{X} | X} \) and \( \hat{\mathcal{X}} \) is defined on \( \hat{\mathcal{X}} \). \( \square \)
Example: Hamming distortion

**Theorem 20.** The rate-distortion function for a bernoulli-$p$ source with Hamming distortion

\[
    d(x, \hat{x}) = \begin{cases} 
    0 & \text{for } x = \hat{x} \\
    1 & \text{for } x \neq \hat{x}
\end{cases}
\]

is given by

\[
    R(D) = \begin{cases} 
    \mathcal{H}(p) - \mathcal{H}(D) & \text{for } 0 \leq D \leq \min\{p, 1-p\} \\
    0 & \text{for } D > \min\{p, 1-p\}
\end{cases}
\]
Example: Quadratic distortion

- Leap of faith: the theorem generalizes to well-behaved continuous-valued sources and unbounded distortion measures.

**Theorem 21.** The rate-distortion function for an i.i.d. Gaussian source $X \sim \mathcal{N}(0, \sigma^2)$ with quadratic distortion

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

is given by

$$R(D) = \begin{cases} 
\frac{1}{2} \log \frac{\sigma^2}{D} & \text{for } 0 \leq D \leq \sigma^2 \\
0 & \text{for } D > \sigma^2 
\end{cases}$$
Covering Lemma

(see Lecture 4)
Source Coding Theorem: Achievability

- **Random codebook generation:** For some $\epsilon > 0$, choose $P_{\hat{X}|X}$ attaining

  $$\mathbb{E}[d(X, \hat{X})] \leq \frac{D}{1 + \epsilon},$$

  let $P_{\hat{X}}(\hat{x}) = \sum_x P_x(x)P_{\hat{X}|X}(\hat{x}|x)$, and generate $2^{nR}$ sequences of length $n$ with i.i.d. elements. These sequences forms a codebook $C_n$, which is revealed to both encoder and decoder.

- **Joint typicality encoding:** Given a sequence $x \in \mathcal{X}^n$, find the index $m$ such that $(x, \hat{x}(m)) \in T_{\epsilon}(X, \hat{X})$.

- **Decoding:** upon receiving the index $m \in \{1, \ldots, 2^{nR}\}$, output the codeword $\hat{x}(m)$.
• **Error probability analysis:** For a given codebook $C_n$, define the encoding error event as

$$\mathcal{E} = \left\{ (X^n, \hat{x}^n(m)) \notin \mathcal{T}_e^{(n)}(X, \hat{X}), \text{ for all } m = 1, \ldots, 2^{nR} \right\}$$

• Via the random coding argument, we analyze the ensemble average encoding error probability, i.e., we are interested in

$$\overline{P}_e^{(n)} = \mathbb{P}\left( (X^n, \hat{X}^n(m)) \notin \mathcal{T}_e^{(n)}(X, \hat{X}), \text{ for all } m = 1, \ldots, 2^{nR} \right)$$

• By the **Covering Lemma** (letting $P_{U,X,\hat{X}} = P_{X,\hat{X}}$, $\tilde{X}^n = X^n$ with $Q_{\tilde{U}^n,\tilde{X}^n}(u, x) = \prod_{i=1}^{n} P_X(x_i)$ in the lemma statement), we have that

$$\lim_{n \to \infty} \overline{P}_e^{(n)} = 0, \quad \text{if } R > I(X; \hat{X}) + \delta(\epsilon)$$
• Denote the reconstruction sequence for $X^n$ by $\hat{X}^n \in C_n$, by the law of total expectation and by the typical average lemma we have

$$\mathbb{E} \left[ d(X^n, \hat{X}^n) \right] = \overline{P}_e^{(n)} \mathbb{E} \left[ d(X^n, \hat{X}^n) \big| \mathcal{E} \right] + (1 - \overline{P}_e^{(n)}) \mathbb{E} \left[ d(X^n, \hat{X}^n) \big| \mathcal{E}^c \right]$$

$$\leq \overline{P}_e^{(n)} d_{\text{max}} + (1 - \overline{P}_e^{(n)}) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} d(X_i, \hat{X}_i) \bigg| \mathcal{E}^c \right]$$

$$\leq \overline{P}_e^{(n)} d_{\text{max}} + (1 - \overline{P}_e^{(n)}) (1 + \epsilon) \mathbb{E}[d(X, \hat{X})]$$

• Since by assumption $\mathbb{E}[d(X, \hat{X})] \leq \frac{D}{1+\epsilon}$, we conclude that

$$\lim_{n \to \infty} \mathbb{E} \left[ d(X^n, \hat{X}^n) \right] \leq D$$
• For any $\epsilon > 0$, we have shown the existence of a sequence of codes $\{C_n^*\}$ for increasing $n$, achieving the rate-distortion point $(R, D)$ with

$$R > I(X; \hat{X}) + \delta(\epsilon)$$

• The proof is completed by minimizing $I(X; \hat{X})$ over all $P_{\hat{X}|X}$ such that $\mathbb{E}[d(X, \hat{X})] \leq \frac{D}{1+\epsilon}$, and letting $\epsilon \downarrow 0$. 


An intuitive explanation

\[ R = \frac{1}{n} \log \frac{2^{nH(X)}}{2^{nH(X|\hat{X})}} = I(X; \hat{X}) \]
Source Coding Theorem: Converse

**Lemma 13.** The rate-distortion function $R(D)$ is convex and non-increasing with $D$.

**Proof:**

- By definition of $R(D)$, if $D$ is increased, the constraint is relaxed and the minimum of $I(X; \hat{X})$ cannot increase. Hence, $R(D)$ is non-increasing with $D$.
- For convexity: consider $(R_1, D_1)$ and $(R_2, D_2)$ on the rate-distortion curve.
- Let $P^{(1)}_{\hat{X}|X}$ and $P^{(2)}_{\hat{X}|X}$ be the attaining pmfs.
- Since $D$ is linear in the joint pmf, we have

$$D_{\lambda} = \lambda D_1 + (1 - \lambda) D_2 = \sum_{x, \hat{x}} d(x, \hat{x}) P_X(x) \left( \lambda P^{(1)}_{\hat{X}|X}(\hat{x}|x) + (1 - \lambda) P^{(2)}_{\hat{X}|X}(\hat{x}|x) \right)$$
• Using convexity of the mutual information we have

\[ I \left( P_X, P_{\hat{X}|X}^{(\lambda)} \right) \leq \lambda I \left( P_X, P_{\hat{X}|X}^{(1)} \right) + (1 - \lambda) I \left( P_X, P_{\hat{X}|X}^{(2)} \right) = \lambda R_1 + (1 - \lambda) R_2 \]

• By definition of rate-distortion function we have

\[ R(D_\lambda) \leq I \left( P_X, P_{\hat{X}|X}^{(\lambda)} \right) \leq \lambda R_1 + (1 - \lambda) R_2 = \lambda R(D_1) + (1 - \lambda) R(D_2) \]
• For any $n$, assume that there exists a code $\{(f_n, g_n)\}$ of rate $R$ satisfying $\mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D$. Then, we will show that it must be $R \geq R(D)$.

• Define the random variable (source-coding index) $M = f_n(X^n)$, and the random vector $\hat{X}^n = g_n(M)$. Then we have

\[
\begin{align*}
nR & \geq H(M) \\
& \geq H(M) - H(M|X^n) \\
& = I(X^n; M) \\
& \geq I(X^n; \hat{X}^n) \\
& = \sum_{i=1}^{n} I(X_i; \hat{X}^n|X^{i-1})
\end{align*}
\]
\[
\begin{align*}
&= \sum_{i=1}^{n} I(X_i; \hat{X}^n, X^{i-1}) - I(X_i; X^{i-1}) \\ &\geq \sum_{i=1}^{n} I(X_i; \hat{X}_i) \\ &\geq \sum_{i=1}^{n} R \left( \mathbb{E}[d(X_i, \hat{X}_i)] \right) \\ &\geq nR \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[d(X_i, \hat{X}_i)] \right) \\ &= nR \left( \mathbb{E}[d(X^n, \hat{X}^n)] \right) \\ &\geq nR(D)
\end{align*}
\]
Source transmissibility

- We wish to transmit a DMS \( \{V_i\} \) over a DMC \( \{\mathcal{X}, P_{Y|X}, \mathcal{Y}\} \) such that:
  1) Each source symbol is transmitted in \( \tau \) channel uses (i.e., \( \tau = n/k \));
  2) The end-to-end average distortion is \( D \) (with respect to a given distortion measure \( d : \mathcal{V} \times \hat{\mathcal{V}} \rightarrow \mathbb{R}_+ \));
  3) The input cost of the source-channel code is \( B \) (with respect to a given cost function \( b : \mathcal{X} \rightarrow \mathbb{R}_+ \)).

- Achievability: use a separated scheme concatenating a source code with block length \( k \) and a channel code with block length \( n = \tau k \). We find

\[
R(D) < \tau C(B)
\]
Can we do better by joint source-channel coding? .... Converse: suppose that a sequence of source-channel codes exist, achieving distortion $D$ with input cost $B$ and $\tau$ channel uses per source symbol. Then

$$nC(B) \geq I(X^n; Y^n) \geq I(V^k; \hat{V}^k) = \sum_{i=1}^{k} I(V_i; \hat{V}^k|V^{i-1})$$
\[
\begin{align*}
&= \sum_{i=1}^{k} I(V_i; \hat{V}^k, V^{i-1}) - I(V_i; V^{i-1}) \\
&\geq \sum_{i=1}^{k} I(V_i; \hat{V}_i) \\
&\geq \sum_{i=1}^{k} R\left(\mathbb{E}[d(V_i, \hat{V}_i)]\right) \\
&\geq kR\left(\frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[d(V_i, \hat{V}_i)]\right) \\
&= kR\left(\mathbb{E}[d(V^k, \hat{V}^k)]\right) \\
&\geq kR(D)
\end{align*}
\]
Blahut-Arimoto Algorithm for Rate-Distortion

- We present a numerical algorithm for computing the rate-distortion function for a DMS with $P_X(r) = p_r$ (a probability vector $p$).
- Let $P = [P_{r,s}]$ denote the transition matrix $P_{\hat{X}|X}(s, r)$ that we wish to optimize, and $d(r, s)$ denote the distortion measure for all pairs $x = r$ and $\hat{x} = s$.
- Our problem is

$$\begin{align*}
\text{minimize} & \quad \sum_r p_r \sum_s P_{r,s} \log \frac{P_{r,s}}{\sum_r' P_{r',s}} \\
\text{subject to} & \quad P \in \mathcal{D}
\end{align*}$$

where

$$\mathcal{D} = \left\{ P : \sum_{r,s} p_r P_{r,s} d(r, s) \leq D, \quad P \text{ is a stochastic matrix}\right\}$$
Theorem 22. Let \( q \) be a pmf on \( \hat{X} \), and define the function

\[
F(p, P, q) = \sum_r \sum_s p_r P_{r,s} \log \frac{P_{r,s}}{q_s}
\]

Then, \( R(D) = \min_{P \in D} \min_q F(p, P, q) \). Furthermore, for fixed \( P \), \( F(p, P, q) \) is minimized by

\[
q_s^* = \sum_r p_r P_{r,s},
\]

and for fixed \( q \), \( F(p, P, q) + \lambda \sum_{r,s} p_r P_{r,s} d(r, s) \) is minimized by

\[
P_{r,s}^* = \frac{q_s \exp(-\lambda d(r, s))}{\sum_{s'} q_{s'} \exp(-\lambda d(r, s'))}
\]

where \( \lambda \) is chosen such that \( \sum_{r,s} p_r P_{r,s}^* d(r, s) = D \) (take log and exp with natural basis).
Blahut-Arimoto Algorithm: Fix $\lambda > 0$, and start from an initial probability vector $q^{(0)}$. Then, for $\ell = 1, 2, 3, \ldots$ iterate the following steps:

$$P_{s,r}^{(\ell)} = \frac{q_s^{(\ell-1)} \exp (-\lambda d(r, s))}{\sum_{s'} q_{s'}^{(\ell-1)} \exp (-\lambda d(r, s'))}$$

and

$$q_s^{(\ell)} = \sum_r p_r P_{r,s}^{(\ell)}$$

It can be proved that the algorithm converges to $R(D) = F(p, P^{(\infty)}, q^{(\infty)})$ where

$$D = \sum_{r,s} p_r P_{r,s}^{(\infty)} d(r, s)$$

Hence, we obtain the rate-distortion function in parametric form with respect to the Lagrange multiplier $\lambda$. 

\[\diamond\]
End of Lecture 7