EL-GY 6063: Information Theory
Lecture 11

May 14, 2019

1 Rate-Distortion

Reference: Chapter 10, Elements of Information Theory.

Definition 1. *A distortion function is a mapping* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^0$.

Example. Assume that the source X is binary. The distortion function $d_H(x, \hat{x}) = \mathbb{1}(x \neq \hat{x}), x, \hat{x} \in \{0, 1\}$ is called the binary Hamming distortion function.

Example. Assume that X and Y are binary random variables with joint distribution $P_{X,Y}$. Find $\mathbb{E}(d_H(X, Y))$.

$$\mathbb{E}(d_H(X,Y)) = \sum_{x,y} d(x,y) P_{X,Y}(x,y) = \sum_{x,y} \mathbb{1}(x \neq y) P_{X,Y}(x,y) = \sum_{x,y:x\neq y} P_{X,Y}(x,y) = P(X \neq Y).$$

Example. Assume that the source X is m-ary (i.e. $\mathcal{X} = \{1, 2, \cdots, m\}$). The distortion function $d_H(x, \hat{x}) = \mathbb{1}(x \neq \hat{x}), x, \hat{x} \in \{1, 2, \cdots, m\}$ is called the m-ary Hamming distortion function.

Example. Assume that the source X is a real-valued source (i.e. $\mathcal{X} = \mathbb{R}$). Then, the squared error distortion is defined as $d_2(x, \hat{x}) = (x - \hat{x})^2, x, \hat{x} \in \mathbb{R}$.

So far, we have only defined distortion between single variables X and $\hat{X}$. Next, we define distortion between vectors of variables.

Definition 2. *For a given distortion function* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$*, the distortion between the pair of vectors* x^n *and* $\hat{x}^n$ *is defined as:*

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

In the source coding problem, the input alphabet $\mathcal{X}$, the source distribution X and the distortion function d are given.

Definition 3. *A source coding problem is characterized by the triple* $(\mathcal{X}, P_X, d)$*, where* $\mathcal{X}$ *is called the source alphabet,* P_X *is called the source distribution and* d *is called the distortion crieteria.*

Given the source coding problem, we are required to construct the coding strategy.

Definition 4. *For the source coding problem characterized by* $(\mathcal{X}, P_X, d)$*, and natural numbers* $n, k \in \mathbb{N}$*, and* (n, k)*-coding strategy is characterized by a pair of functions* $e : \mathcal{X}^n \to \{0, 1\}^k$ *and* $f : \{0, 1\}^k \to \mathcal{X}^n$ *are called the encoding and decoding functions, respectively. The natural number* n *is called the blocklength of the coding strategy.*

We often write $\hat{X}^n$ instead of $f(e(X^n))$ to denote the reconstruction of X^n at the decoder. The goal is to design coding strategies which have small average distortion.

Definition 5. *For the source coding problem characterized by $(\mathcal{X}, P_X, d)$, and natural numbers $n, k \in \mathbb{N}$, and the coding strategy (e,f), the average distortion is defined as*

$$\mathbb{E}(d(X^n, \hat{X}^n)) = \sum_{x^n} P_{X^n}(x^n) d(x^n, f(e(x^n))).$$

The other objective is to use the minimum number of bits to store the source.

Definition 6. *For the source coding problem characterized by $(\mathcal{X}, P_X, d)$, and natural numbers $n, k \in \mathbb{N}$, and (n, k)-coding strategy, the rate of the coding strategy is defined as $r = \frac{k}{n}$.*

Example 1. *Let $n = 2, k = 1$ and $\mathcal{X} = \{0, 1\}$ and d_H is the Hamming dirstortion. Let $e(00) = e(01) = 0$ and $e(11) = e(10) = 1$. Also, assume that $f(0) = 00$ and $f(1) = 11$. Then, $r = \frac{k}{n} = \frac{1}{2}$ and*

$$D = \mathbb{E}(d(X^2, \hat{X}^2)) = \sum_{x^2 \in \{0,1\}^2} P_{X^2}(x^2) d(x^2, f(e(x^2))) =$$

$$P_{X^2}(0,0) d_H(00, 00) + P_{X^2}(0,1) d_H(01, 00) + P_{X^2}(1,0) d_H(10, 11) + P_{X^2}(1,1) d_H(11, 11) = \frac{1}{2}(P_{X^2}(01) + P_{X^2}(10))$$

Definition 7. *For the source coding problem characterized by $(\mathcal{X}, P_X, d)$, the rate distortion pair (R, D) is said to be achievable if there exists a coding strategy (e, f) with rate at most R and average distortion at most D.*

For a given distortion D, we are interested in finding the minimum rate R which can be used to store the source. This is formalized before:

Definition 8. *For the source coding problem characterized by $(\mathcal{X}, P_X, d)$, the rate-distortion function is defined below:*

$$R(D) = min_{(e,f): \frac{1}{n} \sum_{i=1}^n d(X_i, \hat{X}_i) \leq D} r, \forall D \geq 0.$$

The following theorem states Shannon's source coding theorem:

Theorem 1. *For the source coding problem characterized by $(\mathcal{X}, P_X, d)$, the rate-distortion function is characterized as follows:*

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

2 Proof of Achievability

Let the distribution $P_{\hat{X}|X}$ be the distribution that optimizes the rate-distortion formula. Also, fix $n \in \mathbb{N}$ and $\epsilon > 0$. Let $k = \lceil nR \rceil$.

Codebook Generation: Construct a codebook containing 2^k codewords each of length n where each codeword is chosen independently based on the distribution $\prod_{i=1}^n P_{\hat{X}}(x_i)$.

The matrix $\mathcal{C}$ is generated by choosing each of its entries independently of other entries and based on the distribution $P_{\hat{X}}$.

Encoding: In order to store the sequence X^n, the encoder finds a codeword $\underline{C}_i$ which is jointly typical with X^n with respect to the distribution $P_{X,\hat{X}} = P_X P_{\hat{X}|X}$. If such a codeword exists (not necessarily uniquely), then the index of the codeword is sent using k bits (note that there are a total of 2^k codewords) otherwise an error is declared.

Decoding: The decoder receives the index i and declared the codeword $\underline{C}_i$ as the reconstruction $\hat{X}^n$.

The strategy fails if there is no codeword in the codeword which is jointly typical with X^n. We have:

$$
\begin{aligned}
P(\mathcal{E}) &= P(\nexists i : (X^n, \underline{C}_i) \in \mathcal{T}_\epsilon^n(X, \hat{X})) \\
&= \prod_{i=1}^{2^k} P((X^n, \underline{C}_i) \notin \mathcal{T}_\epsilon^n(X, \hat{X})) \\
&= (1 - P((X^n, \underline{C}_1) \in \mathcal{T}_\epsilon^n(X, \hat{X})))^{2^k}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
P((X^n, \underline{C}_1) \in \mathcal{T}_\epsilon^n(X, \hat{X})) &= \sum_{(x^n, \hat{x}^n) \in \mathcal{T}_\epsilon^n(X, \hat{X})} P(X^n = x^n, \underline{C}_1 = \hat{x}^n) \\
&\overset{(a)}{=} \sum_{(x^n, \hat{x}^n) \in \mathcal{T}_\epsilon^n(X, \hat{X})} P(X^n = x^n) P(\underline{C}_1 = \hat{x}^n) \\
&\geq 2^{n(H(X,\hat{X})-\epsilon)} 2^{-n(H(X)+\epsilon)} 2^{-n(H(\hat{X})+\epsilon)} \\
&= 2^{-n(I(X;\hat{X})+3\epsilon)}.
\end{aligned}
$$

where (a) holds since X^n is produced by the source independently of $\underline{C}_1$. So,

$$
P(\mathcal{E}) \leq (1 - 2^{-n(I(X;\hat{X})+3\epsilon)})^{2^k}.
$$

Next, we use the following well-known inequality: $(1-y)^n \leq 2^{-ny}, \forall y \in [0,1], n \in \mathbb{N}$. We have:

$$
P(\mathcal{E}) \leq 2^{-2^{-n(I(X;\hat{X})+3\epsilon)} \cdot 2^k}.
$$

The exponent can be further simplified as follows:

$$
2^{-n(I(X;\hat{X})+3\epsilon)} \cdot 2^k = 2^{n(-I(X;\hat{X})-3\epsilon+\frac{k}{n})}.
$$

Note that $\frac{k}{n} \geq R = I(X;\hat{X})$. So, $2^{-n(I(X;\hat{X})+3\epsilon+\frac{k}{n})}$ goes to infinity as $n \to \infty$. Define $a_n = 2^{-n(I(X;\hat{X})+3\epsilon+\frac{k}{n})}$. Then, we have shown that $P(\mathcal{E}) \leq 2^{-a_n} \to 0$ as $n \to \infty$.

Next, we investigate the rate and distortion criteria. The rate of the coding strategy is $\frac{k}{n} \approx R$ for large n. As for the distortion, the strategy guarantees that X^n and $\hat{X}^n$ are jointly typical with respect to $P_{X,\hat{X}}$.

$$
\frac{1}{n} \sum_{i=1}^{n} d(X_i, \hat{X}_i) \approx \mathbb{E}_{P_{X,\hat{X}}}(d(X, \hat{X})) \leq D.
$$

Where the last inequality follows from the assumption that $P_{\hat{X}|X}$ is the optimizing distribution in the rate-distortion formula.

3 Proof of converse

We first prove that the rate-distortion function is convex.

Lemma 1. *For the source coding problem $(\mathcal{X}, P_X, d)$, the rate-distortion function defined as $R(D) = \min_{P_{\hat{X}|X}:\mathbb{E}(d(X,\hat{X}) \leq D} I(X;\hat{X})$ is convex. Alternatively, for any positive D_1 and D_2 and $\lambda \in [0,1]$, we have $R(\bar{(D)}) \leq \lambda R(D_1) + (1-\lambda)R(D_2)$, where $\bar{D} = \lambda D_1 + (1-\lambda)D_2$.*

proof. It is enough to show that there exists a distribution $P_{X,\hat{X}}$ for which $\mathbb{E}_{P_{X,\hat{X}}}(d(X,\hat{X}) \leq \bar{D}$ and $I_{P_{X,\hat{X}}}(X;\hat{X}) \leq \lambda R(D_1) + (1-\lambda)R(D_2)$.

Assume that $P^i_{\hat{X}|X}, i \in \{1,2\}$ is the optimizing distribution for $R(D_i)$. Then, define $P^i_{X,\hat{X}} = P_X P^i_{\hat{X}|X}$. Also, define $P_{X,\hat{X}} = \lambda P^1_{X,\hat{X}} + (1-\lambda)P^2_{\hat{X}|X}$. We have:

$$\begin{aligned}
&\mathbb{E}_{P_{X,\hat{X}}}(d(X,\hat{X})) \\
&= \sum_{x,\hat{x}} P_{X,\hat{X}}(x,\hat{x})d(x,\hat{x}) \\
&= \sum_{x,\hat{x}} (\lambda P^1_{X,\hat{X}}(x,\hat{x}) + (1-\lambda)P^2_{\hat{X}|X}(x,\hat{x}))d(x,\hat{x}) \\
&= \lambda \mathbb{E}_{P^1_{X,\hat{X}}}(d(X,\hat{X})) + (1-\lambda)\mathbb{E}_{P^2_{X,\hat{X}}}(d(X,\hat{X})) \\
&\leq \lambda D_1 + (1-\lambda)D_2 = D.
\end{aligned}$$

Also, we know that mutual information $I(X;\hat{X})$ is convex in $P_{\hat{X}|X}$. So,

$$I_{P_{X,\hat{X}}}(X;\hat{X}) \leq \lambda I_{P^1_{X,\hat{X}}}(X;\hat{X}) + (1-\lambda)I_{P^2_{X,\hat{X}}}(X;\hat{X}) = \lambda R(D_1) + (1-\lambda)R(D_2).$$

This completes the proof of the lemma.

We proceed to prove the converse. Let $U^k = e(X^n)$ be the stored bits. Then, clearly $H(U^k) \leq \sum_{i=1}^k H(U_i) \leq k$. So,

$$\begin{aligned}
k &\geq H(e(X^n)) \geq H(f(e(X^n)) = H(\hat{X}^n) \\
&= I(X^n;\hat{X}^n) + H(\hat{X}^n|X^n) \\
&\overset{(a)}{=} I(X^n;\hat{X}^n) \\
&= H(X^n) - H(X^n|\hat{X}^n) \\
&\overset{(b)}{=} \sum_{i=1}^n H(X_i) - H(X^n|\hat{X}^n) \\
&\overset{(c)}{\geq} \sum_{i=1}^n H(X_i) - H(X_i|\hat{X}_i) \\
&= \sum_{i=1}^n I(X_i,\hat{X}_i).
\end{aligned}$$

where (a) follows from the fact that $\hat{X}^n = f(e(X^n))$ is a function of X^n, and (b) follows from the fact that the source is generated i.i.d, (c) follows from the fact that conditioning reduces entropy and the chain rule of entropy (exercise). Let $d_i = \mathbb{E}(d(X_i,\hat{X}_i))$, then, from the definition of the rate-distortion function, we have that $I(X_i;\hat{X}_i) \geq R(d_i)$. So,

$$k \geq \sum_{i=1}^n R(d_i) = n\sum_{i=1}^n \frac{1}{n}R(d_i) \geq nR(\frac{1}{n}\sum_{i=1}^n d_i),$$

where we have used the convexity of the rate-distortion function. On the other hand,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n d_i &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(d(X_i,\hat{X}_i)) \\
&= \mathbb{E}(\frac{1}{n}\sum_{i=1}^n d(X_i,\hat{X}_i)) \leq D,
\end{aligned}$$

where in the last line we have used the fact that by assumption ($\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i) \leq D$.

On the other hand it is straightforward to show that the rate-distortion function is decreasing in D (exercise). So,

$$k \geq nR(\frac{1}{n}\sum_{i=1}^{n} d_i) \geq nR(D) \Rightarrow \frac{k}{n} \geq R(D).$$

This completes the proof of the converse.

4 Calculating the rate-distortion function

Example 2. *Let X be a Be(p) random variable where $p \in [0, \frac{1}{2}]$. Let d_H be the binary Hamming distortion. Then, for the source coding problem $(\mathcal{X}, P_X, d_H)$, the rate-distortion function is given by:*

$$R(D) = \begin{cases} H(p, 1-p) - H(D, 1-D) & \text{if } D < p \\ 0 & \text{Otherwise.} \end{cases}$$

proof. Assume that $D < p$. Note that $\mathbb{E}(d_H(X, \hat{X})) = P(X \neq \hat{X}) = P(X \oplus \hat{X} = 1)$. So,

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) = H(X) - H(X \oplus \hat{X}|\hat{X}) \\ &\geq H(X) - H(X \oplus \hat{X}) \geq H(p, 1-p) - H(D, 1-D). \end{aligned}$$

So, we have shown that $R(D) \geq H(p, 1-p) - H(D, 1-D)$. On the other hand, let $\hat{X} = X \oplus N_D$, where N_D is a Be(D) random variable which is independent of $\hat{X}$ (as an exercise, prove that such a variable always exists, in other words, prove that under these assumptions $\hat{X}$ has a valid distribution and find the distribution $P_{\hat{X}}$). Then, $\mathbb{E}(d_H(X, \hat{X})) = P(X \oplus \hat{X} = 1) = P(N_D = 1) = D$, and

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) = H(p, 1-p) - H(N_D|\hat{X}) = H(p, 1-p) - H(N_D) = H(p, 1-p) - H(D, 1-D).$$

This show that $R(D) \leq H(p, 1-p) - H(D, 1-D)$ which completes the proof for the case when $D < p$. The proof for the case when $D \geq p$ is left as an exercise.

5 Gaussian Sources

Definition 9. *A Gaussian source coding problem is characterized by the triple $(\mathcal{X}, f_X, d)$, where $\mathcal{X} = \mathbb{R}$, f_X is a Gaussian distribution and d_2 is the squared distrotion.*

Definition 10. *For the Gaussian source coding problem characterized by given the natural numbers $n, k \in \mathbb{N}$, and (n, k)-coding strategy is characterized by a pair of functions $e : \mathbb{R}^n \to \{0,1\}^k$ and $f : \{0,1\}^k \to \mathbb{R}^n$ are called the encoding and decoding functions, respectively.*

Theorem 2. *For the Gaussian source coding problem characterized by $(\mathbb{R}, f_X, d_2)$, the rate-distortion function is defined below:*

$$R(D) = min_{(e,f):\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i) \leq D} r, \forall D \geq 0.$$

Example 3. *Let X be a Gaussian random variable with variance σ^2. Then, for the Gaussian source coding problem $(\mathbb{R}, f_X, d_2)$, the rate-distortion function is given by:*

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & \text{if } D < \sigma^2 \\ 0 & \text{Otherwise.} \end{cases}$$

proof. Assume that $D < \sigma^2$. *First, we show that* $R(D) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$.

$$\begin{aligned} I(X;\hat{X}) &= h_d(X) - h_d(X|\hat{X}) = h_d(X) - h_d(X-\hat{X}|\hat{X}) \\ &\geq h_d(X) - h_d(X-\hat{X}) = \frac{1}{2}\log 2\pi e\sigma^2 - h_d(X-\hat{X}). \end{aligned}$$

On the other hand, we know that $h_d(X-\hat{X}) \leq \frac{1}{2}\log 2\pi e\sigma^2_{X-\hat{X}} = \frac{1}{2}\log 2\pi eD$. *So,* $I(X;\hat{X}) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$.

So, we have shown that $R(D) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$. *Next, let* $\hat{X}$ *be a Gaussian random variable with variance* $\sigma^2 - D$ *and* N_D *another Gaussian variable with variance* D *which is independent of* $\hat{X}$. *Let* $X = \hat{X} + N_D$ *(check that this is valid). Then,* $\mathbb{E}(d_2(X,\hat{X})) = Var(X-\hat{X}) = Var(N_D) = D$. *Also,*

$$I(X;\hat{X}) = h_d(X) - h_d(X|\hat{X}) = h_d(X) - h_d(N_D) = \frac{1}{2}\log\frac{\sigma^2}{D}.$$

$R(D) \geq H(p,1-p) - H(D,1-D)$. *On the other hand, let* $\hat{X} = X \oplus N_D$, *where* N_D *is a Be(D) random variable which is independent of* $\hat{X}$ *(as an exercise, prove that such a variable always exists, in other words, prove that under these assumptions* $\hat{X}$ *has a valid distribution and find the distribution* $P_{\hat{X}}$*). Then,* $\mathbb{E}(d_H(X,\hat{X})) = P(X \oplus \hat{X} = 1) = P(N_D = 1) = D$, *and*

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) = H(p,1-p) - H(N_D|\hat{X}) = H(p,1-p) - H(N_D) = H(p,1-p) - H(D,1-D).$$