Chapter 9

Lossy Source Coding

In the source coding problem we have studied, perfect recovery is considered for both zero error and vanishing errors, which is possible only when $R > H$. Note that when $R < H$, the error probability tends to 1, i.e., it is not possible to recover the message perfectly. In many practical cases, we do not need perfect recovery, i.e., we allow recovery with certain *distortion*, for example, images and voices, where we can have $R < H$.

9.1 Quantization

Representing a single sample from a continuous source

- Let X be the random variable to be represented.
- Let $\hat{X}$ the be representation of X (a function f of X).
- $\hat{X}$ can take M values (i.e., $\log_2 M$ bits representation).
- Find the optimal function f in terms of certain distortion measure d, i.e.,

$$\min_{f:\text{range of } f \leq M} E[d(X, f(X))].$$

Normal Distribution with Squared Error

- $X \sim N(0, \sigma^2)$.
- $d(X, f(X)) = (X - f(X))^2$.
- $\min_{f:\text{range of } f \leq M} E[(X - f(X))^2]$.
- What are the solutions for $M = 1$ and $M = 2$, respectively?

When $M = 1$, the reconstruction point is 0, and the distortion is σ^2. When $M = 2$, i.e., the rate is 1, the reproduction point is $\pm\sigma\sqrt{2/\pi}$, and the distoration is $\frac{\pi-2}{\pi}\sigma^2$.

Lloyd Algorithm

1. Initialize a set of reconstruction points $x_1, x_2, \ldots, x_M$.
2. For each point x_i, find its nearest neighbor region V_i w.r.t. the distortion measure.
3. Replace x_i by the optimal reconstruction point of V_i.
4. Repeat 2 and 3.

One of the most intriguing aspects of this theory is that joint descriptions are better than individual descriptions, even for independent random variables.

9.2 Rate Distortion Problem

Rate Distortion Codes

- Consider a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$ each with distribution p and alphabet $\mathcal{X}$.
- Encoder: $f : \mathcal{X}^n \to \{1, 2, \ldots, M\}$.
- Decoder: $g : \{1, 2, \ldots, M\} \to \hat{\mathcal{X}}^n$.
- Distortion measure: $d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$.

- Distortion of the code: $E[d(X^n, g(f(X^n)))]$.

The distortion function is applied on each letter separately. Examples of distortion measures include the Hamming distortion and the squared-error distortion. Why average distortion when applying to a sequence of letters?

Rate Distortion Region

- A rate distortion pair (R, D) is said to be achievable if there exists a sequence of rate distortion codes $(f, g)_{n,M}$ such that for any $\epsilon > 0$ and all sufficiently large n, $\frac{1}{n}\log M \leq R + \epsilon$ and $E[d(X^n, g(f(X^n)))] \leq D + \epsilon$.
- The rate-distortion region is the collection of all the achievable rate distortion pair (R, D), which is closed and convex.
- The rate distortion function $R(D)$ is the minimum of all rates R such that (R, D) is achievable.

Characterize the basic properties of $R(D)$, including 1) monotonity, 2) convexity, 3) case with $R(D) = 0$, and 4) $R(0)$.

Rate Distortion Theorem

Theorem 9.1 For a bounded distortion function,

$$R(D) = R_I(D) \triangleq \min_{Q(\hat{x}|x):E[d(X,\hat{X})]\leq D} I(X;\hat{X}),$$

where the distribution of $(X, \hat{X})$ is given by $p(x)Q(\hat{x}|x)$.

- The rate distortion function for a $N(0, \sigma^2)$ source with the squared-error distortion is

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2. \end{cases}$$

- Note that $R(\sigma^2/4)$ is 1 bit.

With a one bit description, the best distortion is $\sigma^2/4$, achieved by considering long block length. The distortion for block length 1 is $\frac{\pi-2}{\pi}\sigma^2 > \sigma^2/4$.

Consider the case of $n = 2$. The joint distribution of (X_1, X_2) is

$$f(x_1, x_2) = \frac{1}{2\pi\sigma^2}\exp\left(\frac{x_1^2 + x_2^2}{2\sigma^2}\right).$$

If we use the four reproduction points $(\pm\sigma\sqrt{2/\pi}, \pm\sigma\sqrt{2/\pi})$, the coding rate is 1 and the distortion is $\frac{\pi-2}{\pi}\sigma^2$, the same as $n = 1$.

A potentially better set of reproduction points is $\{(0, 0), (a, 0), (-a/2, \pm\sqrt{3}/2)\}$, where a can be obtained by minimizing the distortion.

Example 9.1
- The rate distortion function for the binary Bernoulli(p) source ($p \leq 1/2$) with the Hamming distortion is

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D < p \\ 0, & D \geq p. \end{cases}$$

■

Let $Y = d(X, \hat{X})$, where d is the Hamming distortion. Then, for any $Q(\hat{x}|x)$ such that $E[d(X, \hat{X})] \leq D$,

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(\hat{X}|X) \\ &= H(p) - H(Y|X) \\ &\geq H(p) - H(Y) \\ &= H(p) - H(\Pr\{X \neq \hat{X}\}) \\ &\geq H(p) - H(D). \end{aligned}$$

To achieve the above lower bound of the distortion, we consider the joint distribution $p(x, \hat{x})$ satisfying

$$p(x|\hat{x}) = \begin{cases} 1 - D, & x = \hat{x}, \\ D, & x \neq \hat{x}. \end{cases}$$

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

9.3 Proof of Converse

Converse

- Let (R, D) be any achievable rate-distortion pair.
- For any $\epsilon > 0$ and any sufficiently large n, fix an (n, M) code such that

$$\frac{1}{n} \log M \leq R + \epsilon,$$

and

$$E[d(X^n, g(f(X^n)))] \leq D + \epsilon.$$

Let $\hat{X}^n = g(f(X^n))$. Then,

$$\begin{aligned}
n(R+\epsilon) &\geq \log M && (9.1)\\
&\geq H(f(X^n)) && (9.2)\\
&\geq H(g(f(X^n))) && (9.3)\\
&= H(\hat{X}^n) && (9.4)\\
&= H(\hat{X}^n) - H(\hat{X}^n|X^n) && (9.5)\\
&= I(\hat{X}^n; X^n) && (9.6)\\
&= H(X^n) - H(X^n|\hat{X}^n) && (9.7)\\
&= \sum_{k=1}^{n} H(X_k) - \sum_{k=1}^{n} H(X_k|\hat{X}^n, X_1, \ldots, X_{k-1}) && (9.8)\\
&\geq \sum_{k=1}^{n} H(X_k) - \sum_{k=1}^{n} H(X_k|\hat{X}_k) && (9.9)\\
&= \sum_{k=1}^{n} \Big[H(X_k) - H(X_k|\hat{X}_k) \Big] && (9.10)\\
&= \sum_{k=1}^{n} I(X_k; \hat{X}_k) && (9.11)\\
&\geq \sum_{k=1}^{n} R_I(E[d(X_k; \hat{X}_k)]) && (9.12)\\
&= n \sum_{k=1}^{n} \frac{1}{n} R_I(E[d(X_k, \hat{X}_k)]) && (9.13)\\
&= nR_I\left(1/n \sum_{k=1}^{n} E[d(X_k, \hat{X}_k)] \right) && (9.14)\\
&= nR_I(E[d(X^n, \hat{X}^n)]) && (9.15)\\
&\geq nR_I(D+\epsilon). && (9.16)
\end{aligned}$$

9.4 Conditional Typicality

Conditional Typical Set

Definition 9.1 For any $\mathbf{x} \in T^n_{[X]\delta}$, define

$$T^n_{[Y|X]\delta}(\mathbf{x}) = \{\mathbf{y} : (\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}\}.$$

R $T^n_{[Y|X]\delta}(\mathbf{x})$ may be empty. The point is that $\frac{N(a|\mathbf{x})}{n} - p(a) \le \delta$ implies $\sum_b \left(\frac{N(a,b|\mathbf{x},\mathbf{y})}{n} - p(a,b) \right) \le \delta$, but it is possible that $\sum_b \left| \frac{N(a,b|\mathbf{x},\mathbf{y})}{n} - p(a,b) \right| > \delta$ for all $\mathbf{y}$.

Conditional Strong AEP

Theorem 9.2 For any $\mathbf{x} \in T^n_{[X]\delta}$, if $|T^n_{[Y|X]\delta}(\mathbf{x})| > 0$, then

$$2^{n(H(Y|X)-\nu)} \le |T^n_{[Y|X]\delta}(\mathbf{x})| \le 2^{n(H(Y|X)+\nu)},$$

where $\nu \to 0$ as $n \to \infty$ and $\delta \to 0$.

Proof of Conditional Strong AEP

- Upper bound: Let $\nu = 2\eta$ where η is given as in Strong AEP. Then

$$\begin{aligned} 2^{-n(H(X)-\nu/2)} &\ge p(\mathbf{x}) \\ &= \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \\ &\ge \sum_{\mathbf{y} \in T^n_{[Y|X]\delta}(\mathbf{x})} p(\mathbf{x}, \mathbf{y}) \\ &\ge \sum_{\mathbf{y} \in T^n_{[Y|X]\delta}(\mathbf{x})} 2^{-n(H(X,Y)+\nu/2)} \\ &= |T^n_{[Y|X]\delta}(\mathbf{x})| 2^{-n(H(X,Y)+\nu/2)}. \end{aligned}$$

- Lower bound: Use types (see [**yeung08**]).

A Corollary

- Let $S^n_{[X]\delta}$ be the set of all sequences $\mathbf{x} \in T^n_{[X]\delta}$ such that $T^n_{[Y|X]\delta}$ is nonempty.
- Then

$$|S^n_{[X]\delta}| \ge (1-\delta)2^{n(H(X)-\psi)}$$

  where $\psi \to 0$ as $n \to \infty$ and $\delta \to 0$.
- Moreover, for any $\delta > 0$, $P(X^n \in S^n_{[X]\delta}) > 1 - \delta$ for n sufficiently large.

9.5 Proof of Achievability

Settings

- Fix $Q(x|\hat{x})$ such that $E[d(X, \hat{X})] \le D$.
- Fix $\epsilon > 0$, a sufficiently large integer n, and an integer M such that

$$I(X;\hat{X}) + \epsilon/2 \le \frac{1}{n} M \le I(X;\hat{X}) + \epsilon.$$

- We will show the existence of an (n, M) code such that $E[d(X^n, \hat{X}^n)] \le D + \epsilon$.
- Let $\delta > 0$ to be specified later.

Code Construction

- Generate a random codebook $\mathcal{C}$ of M codewords using the distribution $\prod_{i=1}^n p(\hat{x}_i)$, where the i-th codeword is $\hat{X}^n(i)$, $i = 1, 2, \ldots, M$.
- Each source sequence x^n is encoded by i if i is the lagest index such that $(x^n, \hat{X}^n(i)) \in T^n_{[X\hat{X}]\delta}$. If there is no such an i, x^n is encoded to 1.
- The reproduced sequence for i is $\hat{X}^n(i)$. Denote by $\mathcal{C}(x^n)$ the reproduced sequence for x^n.

Calculation of Distortion

- $E[d(X^n, \mathcal{C}(X^n))] = \sum_{x^n} p(x^n) E[d(x^n, \mathcal{C}(x^n))]$.
- The contribution of non-typical x^n to the distortion can be made arbitrarily small by using a sufficiently large n, as the distortion function is bounded.
- For $x^n \notin S^n_{[X]\delta}$, $\mathcal{C}(x^n) = \hat{X}^n(1)$.
- For $x^n \in S^n_{[X]\delta}$, consider two cases:
  - at least one of the codewords in $\mathcal{C}$ is jointly typical with x^n;
  - no codewords are jointly typical with x^n.