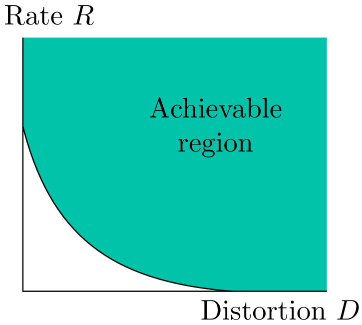
EEE 551 Information Theory (Spring 2022)

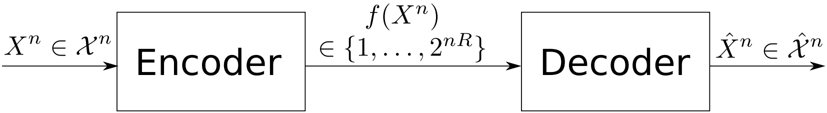
Chapter 10: Rate Distortion Theory

Rate Distortion or Lossy Source Coding

- It is not possible to transmit a source over a noisy channel with arbitrarily small probability of error if $H > C$
- What is the best we can do in such a case?
- Examples:
  - transmission of continuous source over discrete channel
  - storing or transmitting data for human perception, such as audio, images, or video
- **Rate distortion theory** studies the limits of the tradeoff between **rate** (how many bits are sent or stored) and **distortion** (quality of the source reproduction)



Rate Distortion Setup



$X^n \overset{\text{iid}}{\sim} p(x)$

The distortion between X^n and $\hat{X}^n$ is given by a **distortion function** (or **distortion measure**)

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$$

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

Examples

- Squared error distortion: $d(x, \hat{x}) = (x - \hat{x})^2$
  $\mathbb{E}\big[d(X, \hat{X})\big] = \mathbb{E}(X - \hat{X})^2$, i.e. mean square error
- Hamming distortion: $d_H(x, \hat{x}) = \begin{cases} 1, & x \neq \hat{x} \\ 0, & x = \hat{x} \end{cases}$
  $\mathbb{E}\big[d_H(X, \hat{X})\big] = \Pr\{X \neq \hat{X}\}$, i.e. symbol-by-symbol probability of error

Rate Distortion Definitions

- An (M, n) rate distortion code consists of

$$\text{an encoding function} \quad f : \mathcal{X}^n \to \{1, \ldots, M\}$$
$$\text{a decoding function} \quad g : \{1, \ldots, M\} \to \hat{\mathcal{X}}^n$$

- The associated distortion of a code (f, g) is

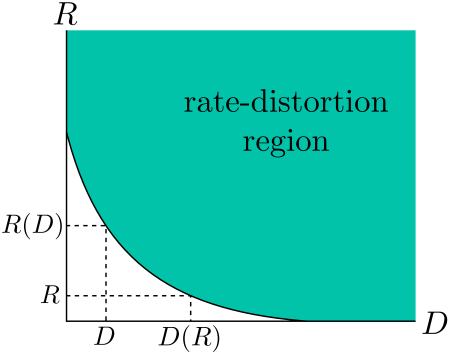
$$D = \mathbb{E}\big[d\big(X^n, g(f(X^n))\big)\big] = \sum_{x^n} p(x^n) d\big(x^n, g(f(x^n))\big)$$

- A rate-distortion pair (R, D) is **achievable** if there exists a sequence of $(2^{nR}, n)$ codes with

$$\lim_{n \to \infty} \mathbb{E}\big[d\big(X^n, g(f(X^n))\big)\big] \le D$$

- The **rate-distortion region** is the closure of all achievable rate-distortion pairs (R, D)

- The **rate-distortion function** $R(D)$ is the infimum of all rates R such that (R, D) is in the rate-distortion region
- The **distortion-rate function** $D(R)$ is the infimum of all distortions D such that (R, D) is in the rate-distortion region



Information Rate-Distortion Function

The **information rate-distortion function** for source $X \sim p(x)$ and distortion function $d(x, \hat{x})$ is

$$R^{(I)}(D) = \min_{p(\hat{x}|x) : \mathbb{E}\left[d(X,\hat{X})\right] \le D} I(X; \hat{X})$$

where

$$\mathbb{E}\left[d(X, \hat{X})\right] = \sum_{x,\hat{x}} p(x)\, p(\hat{x}|x)\, d(x, \hat{x})$$

Theorem

$R(D) = R^{(I)}(D)$

Example 1: Bernoulli Source

$\mathcal{X} = \hat{\mathcal{X}} = \{0, 1\}$, $X \sim$ Bern(p) with Hamming distortion $d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x \neq \hat{x} \end{cases}$

Calculate $R^{(I)}(D) = \min_{p(\hat{x}|x): \mathbb{E}\left[d(X, \hat{X})\right] \leq D} I(X; \hat{X})$

- If $D \geq p$, then setting $\hat{X} = 0$ gives

$$\mathbb{E}d(X, \hat{X}) = \Pr\{X = 1\} = p \leq D$$

  $I(X; \hat{X}) = 0$, so $R^{(I)}(D) = 0$

- If $D \geq 1 - p$, then setting $\hat{X} = 1$ gives

$$\mathbb{E}d(X, \hat{X}) = \Pr\{X = 0\} = 1 - p \leq D$$

  $I(X; \hat{X}) = 0$, so $R^{(I)}(D) = 0$

- If $D = 0$, need $\hat{X} = X$, so $I(X; \hat{X}) = H(X) = H(p)$, so $R^{(I)}(D) = H(p)$

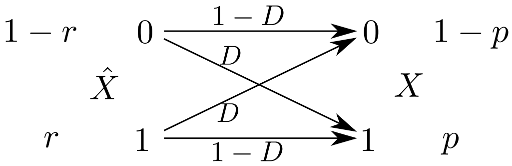
Assume $0 < D < \min\{p, 1-p\}$

- $D < 1/2$
- Assuming $p(\hat{x}|x)$ is such that $\mathbb{E}d(X,\hat{X}) \le D$. Note that $X \oplus \hat{X} = 1$ iff $X \ne \hat{X}$

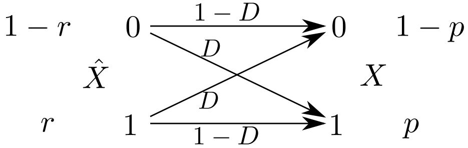
$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= H(p) - H(X \oplus \hat{X}|\hat{X}) \\ &\ge H(p) - H(X \oplus \hat{X}) \\ &\ge H(p) - H(D) \end{aligned}$$

  Thus $R^{(I)}(D) \ge H(p) - H(D)$

- We need to find a distribution $p(\hat{x}|x)$ satisfying $\mathbb{E}d(X,\hat{X}) \le D$ and $H(X|\hat{X}) = H(D)$
- Consider the test channel:



  Satisfies $\mathbb{E}d(X,\hat{X}) = D$, $H(X|\hat{X}) = H(D)$ if we can find valid r

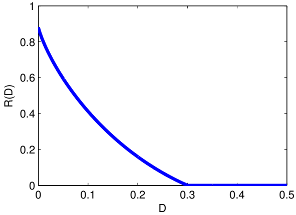


- $p = \Pr\{X = 1\} = r(1-D) + (1-r)D$
- Thus $r = \dfrac{p-D}{1-2D}$, $1-r = \dfrac{1-p-D}{1-2D}$
- Since $D < \min\{p, 1-p\}$, $0 \le r \le 1$
- Hence $R^{(I)}(D) = H(p) - H(D)$

In summary:

$$R^{(I)}(D) = \begin{cases} H(p) - H(D), & 0 \le D < \min\{p, 1-p\} \\ 0, & D \ge \min\{p, 1-p\} \end{cases}$$

e.g. for $p = 0.3$:



Example 2: Gaussian Source

$X \sim \mathcal{N}(0, \sigma^2)$, squared error distortion $d(x, \hat{x}) = (x - \hat{x})^2$

Calculate $R^{(I)}(D) = \min_{f(\hat{x}|x):\mathbb{E}(X-\hat{X})^2 \le D} I(X; \hat{X})$

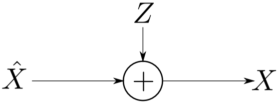
- If $D \geq \sigma^2$, setting $\hat{X} = 0$ gives

$$\mathbb{E}(X - \hat{X})^2 = \sigma^2 \leq D$$

  $I(X; \hat{X}) = 0$, so $R^{(I)}(D) = 0$
- If $D = 0$, would need $\mathbb{E}(X - \hat{X})^2 = 0$, i.e. $\hat{X} = X$, so $I(X; \hat{X}) = \infty$. Thus $R^{(I)}(D) = \infty$
- Assume $0 < D < \sigma^2$, and $f(\hat{x}|x)$ satisfies $\mathbb{E}(X - \hat{X})^2 \leq D$

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

- We need to find a distribution $f(\hat{x}|x)$ satisfying $\mathbb{E}(X-\hat{X})^2 \leq D$ and $h(X|\hat{X}) = \frac{1}{2}\log 2\pi e D$
- Consider the test channel:

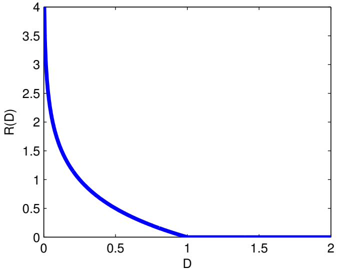


  where $Z \sim \mathcal{N}(0, D)$ independent of $\hat{X}$
- To ensure $X \sim \mathcal{N}(0, \sigma^2)$, choose $\hat{X} \sim \mathcal{N}(0, \sigma^2 - D)$ (possible since $D < \sigma^2$)
- We have $\mathbb{E}(X-\hat{X})^2 = D$ and $h(X|\hat{X}) = \frac{1}{2}\log 2\pi e D$
- Thus $R^{(I)}(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$

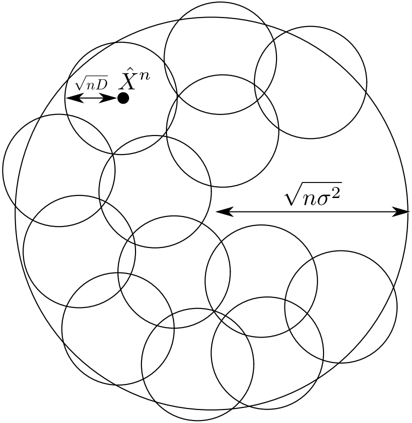
In summary:

$$R^{(I)}(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & D < \sigma^2 \\ 0, & D \geq \sigma^2 \end{cases}$$

e.g. for $\sigma^2 = 1$:



Geometrical Interpretation



How many small spheres needed to **cover** the large sphere? (Dual to **sphere packing** in channel coding)

$$\frac{C_n \left[\sqrt{n\sigma^2}\right]^n}{C_n \left[\sqrt{nD}\right]^n} = \left(\frac{\sigma^2}{D}\right)^{n/2} = 2^{\frac{n}{2}\log(\sigma^2/D)}$$

Proof Sketch of the Rate-Distortion Theorem: Achievability

Fix $p(\hat{x}|x)$ with $\mathbb{E}[d(X,\hat{X})] \leq D$. Let $p(\hat{x}) = \sum_x p(x)\, p(\hat{x}|x)$.

Random codebook generation

- For each $m \in \{1, \ldots, 2^{nR}\}$, generate codeword $\hat{X}^n(m) \overset{\text{iid}}{\sim} p(\hat{x})$
- Codebook $\mathcal{C} = \left(\hat{X}^n(1), \hat{X}^n(2), \ldots, \hat{X}^n(2^{nR})\right)$

Encoding Process

- Given X^n, choose $f(X^n) = m$ if $(X^n, \hat{X}^n(m)) \in A_\epsilon^{(n)}$
- If there is more than one such m, choose the smallest
- If there is no such m, set $f(X^n) = 1$

Decoding Process

- Given $m = f(X^n)$, set $\hat{X}^n = \hat{X}^n(m)$

Distortion Analysis

- Given $(X^n, \hat{X}^n) \sim p(x)p(\hat{x})$, $\Pr\{(X^n, \hat{X}^n) \in A_\epsilon^{(n)}\} \geq 2^{-nI(X;\hat{X})}$
- Thus, if $R > I(X;\hat{X})$, with high probability, at least one m will satisfy $(X^n, \hat{X}^n(m)) \in A_\epsilon^{(n)}$

Proof Sketch of the Rate-Distortion Theorem: Converse

- Given a code achieving distortion D at rate R:

$$\begin{aligned} nR &\geq H(f(X^n)) \\ &\geq I(X^n; f(X^n)) \\ &\geq I(X^n; \hat{X}^n) \\ &= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i | \hat{X}^n, X_1, \ldots, X_{i-1}) \\ &\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i | \hat{X}_i) \\ &= \sum_{i=1}^{n} I(X_i; \hat{X}_i) \\ &\geq \sum_{i=1}^{n} R^{(I)}(D_i) \end{aligned}$$

  where $D_i = \mathbb{E}[d(X_i, \hat{X}_i)]$

- We can show that $R^{(I)}$ is **convex**, so

$$R \geq \frac{1}{n} \sum_{i=1}^{n} R^{(I)}(D_i) \geq R^{(I)}\left(\frac{1}{n} \sum_{i} D_i\right) \geq R^{(I)}(D)$$