# Rate Distortion Theory

# Introduction

- A **distortion measure** is a measure of distance between a random variable and its representation.
- The basic problem in rate-distortion theory is this: *Given a random variable and a distortion measure, what is the minimum expected distortion achievable at a particular rate?*
- Equivalently, *What is the minimum rate required to achieve a distortion?*
- For zero-distortion discrete case, we have the data compression theory.

# Engineering Aspects

- Want low rate (high compression)
  Given the constraint on maximum distortion, what is the minimum rate?
- Want low distortion (high fidelity)
  Given the constraint on maximum rate (such as data channel capacity), what is the minimum distortion?
- Examples
  - speech/audio coding
  - video coding/image compression

# Quantization

- Representing a continuous random variable by a finite number of bits.
- Let the random variable be $X$ and the representation be $\hat{X}$. We want to find the optimal set of values and the associated regions for each $\hat{X}$.
- For $R$ bits to represent $X$, there are $2^R$ values for $\hat{X}$, which are called the reproduction points or code points.
- For example, let $R = 1$ and $X \sim N(0, \sigma^2)$. The answer is given by (13.1). How about $R = 2$?

# Lloyd Algorithm

- Two properties for optimum regions and reconstruction points
  - Given a set of reproduction points, the distortion is minimized by assigning a value $x$ to the closest reproduction point. The set of regions thus defined is called a Voronoi partition.
  - The reproduction points should minimize the conditional expected distortion over the assigned regions.
- The Lloyd algorithm finds a local minimum with these two properties.

# Distortion Functions

- A distortion function or distortion measure is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to R^{+} \cup \{0\},$$

from the set of source-reproduction pairs into the set of non-negative real numbers. For examples,

  - Hamming distortion: $d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x \neq \hat{x} \end{cases}$
  - squared error distortion: $d(x, \hat{x}) = (x - \hat{x})^2$

# Distortion between Sequences

- The distortion between sequences $x^n$ and $\hat{x}^n$ is defined by

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

  In other words, it is the distortion per symbol.

- This is not the only reasonable definition. For example, one can use the maximal instead of the average.

- The analysis here is based on this average distortion measure between sequences.

# Rate Distortion Code

- A $(2^{nR}, n)$ rate distortion code consists of an encoding function

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

and a decoding function

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

- The distortion associated with this code is

$$\begin{aligned} \mathcal{D} &= E\left[d(X^n, g_n(f_n(X^n)))\right] \\ &= \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))). \end{aligned}$$

## Codebook and Vector Quantization

- $f_n^{-1}(1), \ldots, f_n^{-1}(2^{nR})$ are called the assignment regions. I.e., $f_n^{-1}(k)$ is the region associated with index $k$.
- $\hat{X}^n(1), \ldots, \hat{X}^n(2^{nR})$ constitute the codebook. Any $X$ in region $k$ is represented by the code point $\hat{X}^n(k)$.
- It is common to refer to $\hat{X}^n$ as the vector quantization, reproduction, reconstruction, representation, source code, or estimate of $X^n$.
- $\mathcal{D}$ is the averaged distortion between $X^n$ and $\hat{X}^n$.

# Rate-Distortion Region

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

$$\lim_{n\to\infty} E\left[d(X^n, g_n(f_n(X^n)))\right] \le D.$$

- The rate-distortion region for a source is the closure of the set of achievable rate distortion pairs.
- The rate distortion function $R(D)$ is the infimum of $R$ for given $D$ in the rate distortion region.
- The distortion rate function $D(R)$ is the infimum of $D$ for given $R$ in the rate distortion region.

## More Precisely (Mathematical)

- Let $C$ be the rate distortion region, the closure of the set of achievable rate-distortion pairs.
- The rate distortion function is

$$R(D) = \inf_{(R,D)\in C} R$$

- The distortion rate function is

$$D(R) = \inf_{(R,D)\in C} D$$

# Information Rate-Distortion Function

- The information rate distortion function $R^{(I)}(D)$ for a source $X$ with distortion $d(x, \hat{x})$ is defined by

$$R^{(I)}(D) = \min_{p(\hat{x}|x): Ed(X,\hat{X}) \leq D} I(X; \hat{X}).$$

where the minimization is over all conditional distributions $p(\hat{x}|x)$ for which the joint distribution $p(x, \hat{x})$ satisfies the expected distortion constraint.

# $R^{(I)}(D)$ for a Bernoulli Source

- For a Bernoulli($p$) source with Hamming distortion, the information rate distortion function is

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

- This follows from

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

since $p(X \neq \hat{X}) = E(d(X, \hat{X})) \leq D$. Furthermore, this lower bound is achievable. See Figure 13.3.

# $R^{(I)}(D)$ for a Gaussian Source

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

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

- It follows that each bit reduces the expected distortion by a factor of $4$. The 1-bit quantization scheme mentioned earlier has an expected distortion of $\frac{\pi-2}{\pi}\sigma^2$. Why?

# Proof

- The mutual information is lower-bounded by

$$
\begin{aligned}
I(X;\hat{X}) = h(X) - h(X|\hat{X}) &= h(X) - h(X-\hat{X}|\hat{X}) \\
&\geq h(X) - h(X-\hat{X}) \\
&\geq h(X) - h(N(0, E(X-\hat{X})^2)) \\
&= \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log(2\pi e E(X-\hat{X})^2)) \\
&\geq \frac{1}{2}\log\frac{\sigma^2}{D}.
\end{aligned}
$$

- To see this lower bound is achievable, it is more convenient to look at the test channel as in Figure 13.5.

## Rate Distortion Theorem

- The rate distortion function for an i.i.d. source $X$ with distribution $p(x)$ and distortion function $d(x, \hat{x})$ is equal to information rate distortion function. I.e.

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

- Specifically, it can be shown that

$$R(D) \geq R^{(I)}(D)$$

and

$$R^{(I)}(D) \geq R(D)$$

## Rate Distortion Theorem I

If $R < R^{(I)}(D)$, then $(R, D)$ is not an achievable rate-distortion pair.

- Since $R(D)$ is the infimum of all achievable rates given $D$, it follows that

$$R(D) \geq R^{(I)}(D).$$

  Otherwise, there exists $R$ with $R(D) < R < R^{(I)}(D)$ and $(R, D)$ is not achievable, a contradiction to the definition of $R(D)$.

# Proof

Suppose that $(R, D)$ is an achievable pair, then

$$
\begin{aligned}
nR &\geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \\
&= H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \\
&= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{1:i-1}) \geq \sum_{i=1}^{n} I(X_i; \hat{X}_i) \\
&\geq \sum_{i=1}^{n} R^{(I)}(Ed(X_i, \hat{X}_i)) = n \sum_{i=1}^{n} \frac{1}{n} R^{(I)}(Ed(X_i, \hat{X}_i)) \\
&\geq nR^{(I)}(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) = nR^{(I)}(Ed(X^n, \hat{X}^n)) \\
&= nR^{(I)}(D).
\end{aligned}
$$

## Rate Distortion Theorem II

If $R > R^{(I)}(D)$, then $(R, D)$ is an achievable pair.

- Since $R(D)$ is the infimum of all achievable rates given $D$, it follows that

$$R^{(I)}(D) \geq R(D).$$

  Otherwise, there exists $R$ with $R(D) > R > R^{(I)}(D)$ and $(R, D)$ is achievable, again a contradiction.

Before we give the proof, we need a few lemmas.

# Distortion Typical Set

- A pair of sequence $(x^n, \hat{x}^n)$ is distortion $\epsilon$-typical if the four conditions (13.71)-(13.74) is satisfied. The set of distortion typical sequences is the distortion typical set.
- Lemma 13.5.1: The probability of distortion typical set approaches 1 as $n \to \infty$ for any $\epsilon > 0$.
- Lemma 13.5.2: For all $(x^n, \hat{x}^n)$ in distortion typical set,

$$p(\hat{x}^n) \geq p(\hat{x}^n | x^n) 2^{-n(I(X;\hat{X})+3\epsilon)}$$

- Lemma 13.5.3:

$$(1 - xy)^n \leq 1 - x + e^{-yn} \text{ for } 0 \leq x, y \leq 1, n > 0.$$

## Proof of Earlier Theorem

- We will prove the theorem by the existence of a rate-distortion code with rate $R > R^{(I)}(D)$ that satisfies the distortion constraint.
- Let $X_1, \ldots, X_n$ be i.i.d. $\sim p(x)$ and $d(x, \hat{x})$ be a bounded distortion measure. Let $p(\hat{x}|x)$ be the conditional distribution that achieves the minimum. Compute $p(\hat{x})$.

## Rate-Distortion Coding Scheme

- Randomly generate a codebook, say $\mathcal{C}$, of $2^{nR}$ $\hat{x}^n$-sequences drawn i.i.d. from $\prod_{i=1}^{n} p(\hat{x}_i)$.
- Encode $x^n$ by $w$ if $w$ is the only integer such that $(x^n, \hat{x}^n(w))$ is distortion $\epsilon$-typical. If there are more than one $w$, send the least. Else, let $w = 1$.
- Decode: The reproduction point for $x^n$ is $\hat{x}^n(w)$.

## Analysis of Distortion

- Recall that the distortion is defined by

$$E\left[d(X^n, g_n(f_n(X^n)))\right] = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))).$$

- For $x^n$ that is jointly distortion $\epsilon$-typical with a codeword $\hat{x}^n(w)$ in the codebook, $d(x^n, \hat{x}^n(w)) \leq D + \epsilon$. The probability of all such sequences is at most $1$, so the contribution of these $x^n$-sequences to the distortion is at most $D + \epsilon$.

- Let $P_e$ be the probability that $X^n$ is not jointly distortion $\epsilon$-typical with any codeword. The contribution of these sequences to the distortion is at most $P_e d_m$.

# Calculation of $P_e$

- For a codebook $\mathcal{C}$, let $J(\mathcal{C})$ be the set of sequences $x^n$ such that it is jointly distortion $\epsilon$-typical with at least one codeword. Then

$$P_e = \sum_{\mathcal{C}} p(\mathcal{C}) \sum_{x^n : x^n \notin J(\mathcal{C})} p(x^n) = \sum_{x^n} p(x^n) \sum_{\mathcal{C}: x^n \notin J(\mathcal{C})} p(\mathcal{C})$$

- The term $\sum_{\mathcal{C}: x^n \notin J(\mathcal{C})} p(\mathcal{C})$ is the probability of choosing a codebook that does not represent $x^n$ well.
- Define $K(x^n, \hat{x}^n)$ to be the indicator function of the distortion $\epsilon$-typical set $A_{d,\epsilon}^{(n)}$.

# Calculation of $P_e$ (continued)

- The probability that a single randomly chosen codeword $\hat{X}^n$ does not represent a fixed $x^n$ well is

$$p((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n)$$

- So the probability that $2^{nR}$ independently chosen codewords do not represent $x^n$ well, average over $x^n$ is,

$$P_e = \sum_{x^n} p(x^n) \left[ 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \right]^{2nR}$$

- With Lemma 13.5.2, it follows that

$$P_e \leq \sum_{x^n} p(x^n) \left[ 1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}^n|x^n) K(x^n, \hat{x}^n) \right]^{2nR}$$

# Calculation of $P_e$ (continued)

- We now use Lemma 13.5.3,

$$\left[1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n,\hat{x}^n)\right]^{2nR}$$

$$\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n,\hat{x}^n) + e^{-2^{-n(I(X;\hat{X})+3\epsilon)}2^{nR}}$$

$$= 1 - \sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n,\hat{x}^n) + e^{-2^{n(R-I(X;\hat{X})-3\epsilon)}}.$$

  So the last term of $P_e$ goes to 0 if $R > I$.

- Since the sum of the first two terms is the probability that $(X^n, \hat{X}^n)$ are not distortion typical, it goes to 0 as well.