Rate Distortion Theory

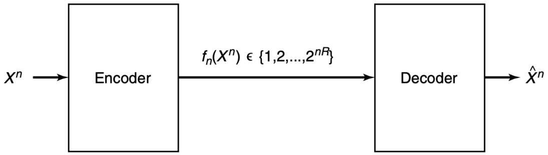
Youlong Wu

ShanghaiTech University

wuyl1@shanghaitech.edu.cn

- Quantization
- Definitions
- Calculation of the rate-distortion function
- Example: Gaussian Source with MSE distortion

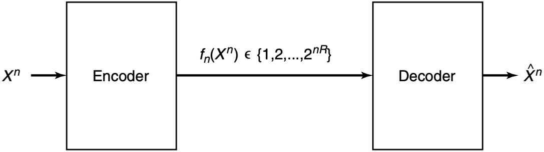
Rate Distortion Theory



Rate-distortion theory describes the trade-off between lossy compression rate and the resulting distortion.

- Lossless Source coding: Recover source data X without error
- Lossy source coding: Recover source with some error and distortion

Quantization

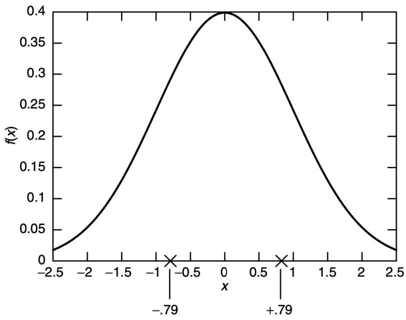


- Given a continuous RV X, if it is lossless source coding, we need infinite bits to perfectly recover X
- Question: What is the best possible representation of X for a given data rate?
- X: random variable to be represented
- $\hat{X}(X)$: representation of X
- R bits for the representation $\rightarrow |\hat{X}| = 2^{nR}$
- Want to find the optimal set of values for $\hat{X}$ and associated regions

Quantization Example: 1-bit Gaussian

Given $X \sim \mathcal{N}(0, \sigma^2)$ and $MSE = \|X - \hat{X}\|^2$, find $\hat{X}$ such that 1) $\hat{X}$ takes on two values; 2) minimizes MSE

$$X = \begin{cases} \sqrt{\frac{2}{\pi}\sigma}, & \text{if } X \geq 0 \\ -\sqrt{\frac{2}{\pi}\sigma}, & \text{if } X < 0 \end{cases} \tag{1}$$



Quantization

Objective: Map the incoming sequence $U_1, U_2, \ldots$ into a sequence of discrete RVs $V_1, V_2, \ldots$, where V_m should represent U_m with as little distortion as possible.

- Scalar Quantization: Each analog RV in the sequence is quantized independently of the other RVs.

- Vector Quantization: The analog sequence is first segmented into blocks of n RVs each; then each n-tuple is quantized as a unit.

Scalar Quantization

- Partition the region $\mathbb{R}$ into M regions $\mathcal{R}_1, \ldots, \mathcal{R}_M$.
- Each region R_j is mapped to a symbol a_j called the representation point for $\mathcal{R}_j$.

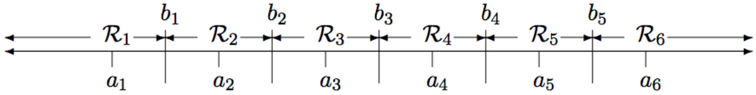
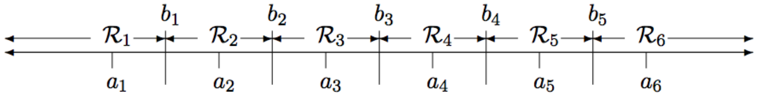


Figure: Quantization Region [Gallagar'Book]

- Each source value $u \in \mathcal{R}_j$ is mapped into the representation point a_j.
- After discrete coding and channel transmission, the receiver sees a_j and the distortion is $u - a_j$.

Scalar Quantization

- View the source value u as a sample value of a RV U.
- The representation a_j is a sample value of the RV V, where V is the quantization of U. If $U \in \mathcal{R}_j$, then $V = a_j$
- The source sequence is $U_1, U_2, \ldots$. The representation is $V_1, V_2, \ldots,$, where if $U_k \in \mathcal{R}_j$, then $V_k = a_j$
- Aussme that $U_1, U_2, \ldots$ is DMS
- For a scalar quantizer, we can look at just a single U and a single V



Mean Square Distortion of a Scalar Quantizer

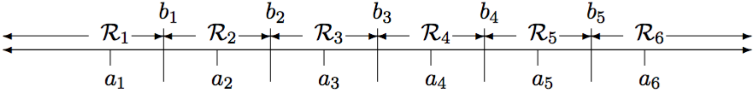
$$\boxed{\text{MSE} = E[(U - V)^2]}$$

Aim: Given pdf $f_U(u)$ and alphabet size M, choose $\{\mathcal{R}_j, 1 \leq j \leq M\}$ and $\{a_j, 1 \leq j \leq M\}$ to miminize MSE.

Explore it into two ways:

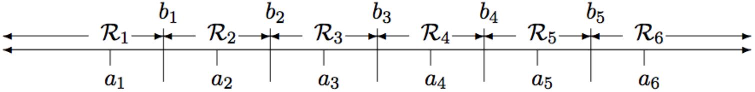
- Given a set of $\{a_j\}$, how should the intervals $\{\mathcal{R}_j\}$ be chosen?
- Given a set of intervals $\{\mathcal{R}_j\}$, how to choose $\{a_j\}$?

Choice of $\{\mathcal{R}_j\}$ for given $\{a_j\}$



Given a_j, choose b_j such that $E[(U-V)^2]$ is minimized.
Solution:

Choice of $\{\mathcal{R}_j\}$ for given $\{a_j\}$



Given a_j, choose b_j such that $E[(U-V)^2]$ is minimized.
Solution:

$$E[(U-V)^2] = \sum_{j=1}^{M} \int_{\mathbf{R_j}} f_U(u)(u-a_j)^2 du = \sum_{j=1}^{M} \int_{b_{j-1}}^{b_j} f_U(u)(u-a_j)^2 du$$

$$= \cdots + \int_{b_{j-1}}^{b_j} f_U(u)(u-a_j)^2 du + \int_{b_j}^{b_{j+1}} f_U(u)(u-a_{j+1})^2 du + \cdots$$

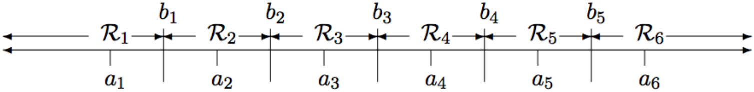
Let $\frac{\partial E[(U-V)^2]}{\partial b_j} = 0$, we have $\left(\frac{\partial \int_{q(x)}^{g(x)} f(u)du}{\partial x} = f(g(x))\frac{\partial g(x)}{\partial x} - f(q(x))\frac{\partial q(x)}{\partial x}\right)$

$$f_U(b_j)(b_j-a_j)^2 - f_U(b_j)(b_j-a_{j+1})^2 = 0$$

$$2b_j(a_{j+1}-a_j) = a_{j+1}^2 - a_j^2$$

$$b_j = \frac{a_j + a_{j+1}}{2}$$

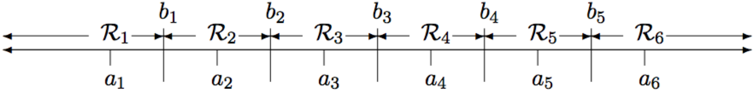
Choice of $\{\mathcal{R}_j\}$ for given $\{a_j\}$



- For source output u, squared error to a_j is $|u - a_j|^2$
- Minimize by choosing closest a_j.
- Thus $\mathcal{R}_j$ is a region closer to a_j than any a_i, for all $i \neq j$.
  (The boundary of b_j is between $\mathcal{R}_j$ and $\mathcal{R}_{j+1}$ must lie hafway between a_j and a_{j+1})
- $\mathcal{R}_j$ is bounded by

$$\boxed{b_j = \frac{a_j + a_{j+1}}{2}}$$

Choice of $\{a_j\}$ for given $\mathcal{R}_j$



$$\text{MSE} = E[(U-V)^2] = \int_{-\infty}^{\infty} f(u)(u-v)^2 du = \sum_{j=1}^{M} \int_{\mathcal{R}_j} f(u)(u-a_j)^2 du$$

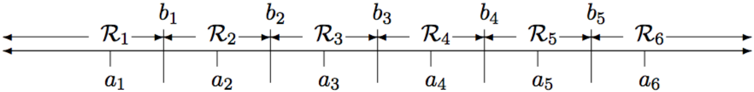
$$= \sum_{j=1}^{M} \int_{\mathcal{R}_j} f(u)(u^2 - 2a_j u + a_j^2) du$$

Let $\frac{\partial E[(U-V)^2]}{\partial a_j} = 0$, we have

$$-2\int_{\mathcal{R}_j} f(u) u du + 2\int_{\mathcal{R}_j} f(u) a_j du = 0$$

$$\Longrightarrow a_j = \frac{\int_{R_j} u f(u) du}{\int_{R_j} f(u) du}$$

Lloyd-Max Algorithm



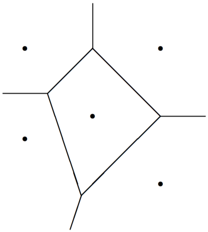
An optimal scalar quantizer must satisfy both $b_j = (a_j + a_{j+1})/2$ and $a_j = E[U_{(j)}]$.

1. Choose $a_1 < a_2 < \cdots < a_M$
2. set $b_j = (a_j + a_{j+1})/2$ for $1 \leq j \leq M - 1$
3. Set $a_j = \frac{\int_{\mathcal{R}_j} u f(u) du}{\int_{\mathcal{R}_j} f(u) du}$ where $\mathcal{R}_j = (b_{j-1}, b_j]$ for $1 \leq j \leq M - 1$
4. Iterate on 2 and 3 until improvement is negligible.

It find local min, not necessarily global min.

Vector Quantization

Quantize n source variables at a time. (In scalar quantization, $n = 1$)



Given $\{(a_j, a'_j)\}$, how to choose $\{\mathcal{R}_j\}$

- The square error is $(u - a_j)^2 + (u' - a'_j)^2$, the point $\{a_j, a'_j\}$ which is the closest to (u, u') in Euclidean distance should be chosen.
- $\{\mathcal{R}_j\}$ contains points that are closer to (a_j, a'_j) than any other representation points, i.e., Voronoi regions.

Given a set of Voronoi region, how to find the $\{a_j, a'_j\}$?.

- Choose $\{a_j, a'_j\}$ to be the conditional means within those regions.

2D Quantization

The symbol $+$ represents the updated point, and $\bullet$ the original point

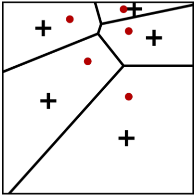


Figure: Iteration 1

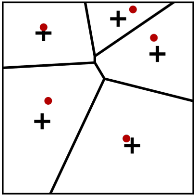


Figure: Iteration 2

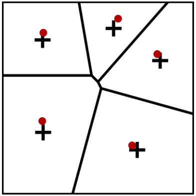


Figure: Iteration 3

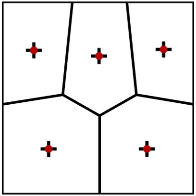
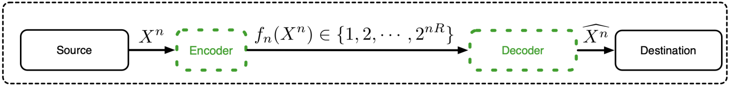


Figure: Iteration 4

More about Vector Quantization

- Popular research topic, related to deep learning algorithm
- Quantizing complexity goes up exponentially with n
- Reduction in MSE with increasing n is quite modest
- Application: Video, Audio

Definitions



- $X_1, X_2, \cdots, X_n$ i.i.d. $\sim p(x), x \in \mathcal{X}$
- A **distortion function** or **distortion measure** is a mapping

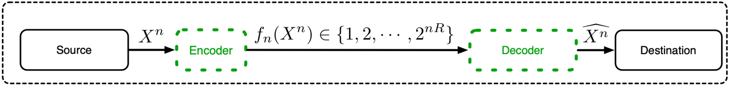
$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$$

  from the set of source alphabet-reproduction alphabet pairs into the set of non-negative real numbers. Measures the "cost" of representing symbol x by $\hat{x}$.

- A distortion measure is said to be **bounded** if the maximum value of the distortion is finite,

$$d_{max} := \max_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) < \infty$$

Definitions



- Two most common distortion functions:
  - Hamming distortion:

$$d(x, \hat{x}) = \begin{cases} 0 \text{ if } x = \hat{x} \\ 1 \text{ if } x \neq \hat{x} \end{cases}$$

  - Squared-error distortion:

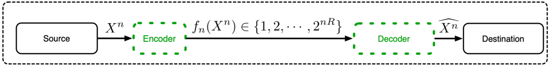
$$d(x, \hat{x}) = (x - \hat{x})^2$$

- We *define* the *distortion between sequences* x^n and $\hat{x}^n$ as

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

Definitions

Definitions



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

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

and a decoding (reproduction) function,

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

- The distortion associated with the $(2^{nR}, n)$ code is defined as

$$D = E[d(X^n, g_n(f_n(X^n))),$$

where the expectation is with respect to the probability distribution on $\mathcal{X}$,

$$D = \sum p(x^n) d(x^n, g_n(f_n(x^n))).$$

- The set of n-tuples $g_n(1), g_n(2), \cdots, g_n(2^{nR})$, denoted by $\hat{X}^n(1), \hat{X}^n(2), \cdots, \hat{X}^n(2^{nR})$ constitutes the *codebook* and $f_n^{-1}(1), \cdots f_n^{-1}(2^{nR})$ are the associated *assignment regions.*

Definitions

- A rate-distortion pair (R, D) is said to be *achievable* if there exists a sequence of $(2^{nR}, n)$-rate distortion codes (f_n, g_n) with $\lim_{n\to\infty} E[d(X^n, g_n(f_n(X^n))] \leq D$.
- The *rate-distortion region* for a source is the closure of the set of achievable rate distortion pairs (R, D).
- The *rate-distortion function* $R(D)$ is the **infimum** of rates R such that (R, D) is in the rate distortion region of the source for a given distortion D.
- The *distortion-rate function* $D(R)$ is the **infimum** of all distortions D such that (R, D) is in the rate distortion region of the source for a given rate R.

Main Results

Theorem: The rate distortion function for an i.i.d. source X with distributed $p(x)$ and bounded distortion function $d(x, \hat{x})$ is equal to the associated information rate distortion function. Thus,

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

is the minimum achievable rate at distortion D.

Theorem: The rate distortion function for a Bernoulli(p) source with Hamming distortion is given by

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

Theorem: The rate distortion function for a $\mathcal{N}(0, \sigma^2)$ source with squared-error distortion is given by

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

The key idea of the Proof:

- Converse: Find a lower bound on $I(X; \hat{X})$
- Achievability: Show that the lower bound is achievable

$R(D)$ of Gaussian and Binary Sources

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

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

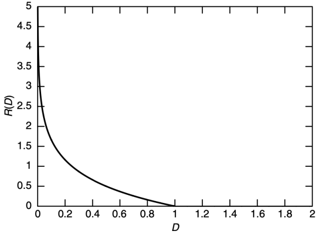


Figure: Gaussian Source $X \sim \mathcal{N}(0, \sigma^2)$ with MSE distortion D

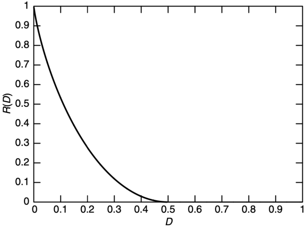


Figure: Binary Source $X \sim \text{Bern}(p)$ with Hamming distortion D

Converse Proof: $R(D)$ of Gaussian

Converse:

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

Achievability Proof: $R(D)$ of Gaussian

Achievability: Find a $f(x|\hat{x})$ to achieve the lower bound

- If $D < \sigma^2$
  - Choose $X = \hat{X} + Z$, where $\hat{X} \sim \mathcal{N}(0, \sigma^2 - D)$, $Z \sim \mathcal{N}(0, D)$ are independent
  - With the above choice, $I(X; \hat{X}) = \frac{1}{2} \log \frac{\sigma^2}{D}$, and $E(X - \hat{X})^2 = D$
- If $D > \sigma^2$, we choose $\hat{X} = 0$ with probability 1, then $R(D) = 0$