ECE253/CSE208 Introduction to Information Theory

Lecture 15: Rate-Distortion Theory

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 10 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*.

Lossy Compression

- We want to compress a random source as much as possible.
- The source coding theorem says that the compression limit is the entropy if we want lossless compression (perfect reconstruction).
- However, in many real applications we can tolerate distortion (the mismatch between the source symbol and its reconstruction) to a certain extent.
- For example, JPEG/MPEG for images and MP3/AAC/Ogg Vorbis for audio signals.
- Given a source distribution and a distortion measure, what is the minimum rate description (number of bits per symbol) required to achieve a particular distortion?
- Or, what is the minimum expected distortion achievable at a particular rate?

Rate–distortion Theory

- Rate–distortion theory is the theoretical foundation of *lossy* data compression: It addresses the problem of the optimal tradeoff between data rate and distortion.

- **Rate:** In the context of source coding, rate refers to the average number of bits needed to represent a source symbol.
- **Distortion:** Distortion measures the quality of the reconstructed signal compared to the original source.
- **Rate-Distortion function:** The Rate-Distortion function represents the trade-off between the rate of transmission and the distortion in the reconstructed signal.

- We aim to determine the minimal number of bits per symbol that is communicated over a channel, so that the source can be approximately reconstructed at the receiver without exceeding an expected distortion D.

Distortion Measure

Definition (Distortion measure)

A single-letter distortion measure is a mapping

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

The value $d(x, \hat{x})$ denotes the distortion incurred when a source symbol and its reconstruction is x and $\hat{x}$, respectively.

Bounded distortion: $d_{\max} \triangleq \max_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) < \infty$.

Example (Distortion functions)

1. Hamming distortion: $d(x, \hat{x}) = \mathbb{1}_{\{x \neq \hat{x}\}}$.
2. Absolute-error distortion: $d(x, \hat{x}) = |x - \hat{x}|$.
3. Square-error distortion: $d(x, \hat{x}) = (x - \hat{x})^2$.

The distortion between sequences x^n and $\hat{x}^n$ is given by averaging per-letter distortion:

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

Rate-distortion Code

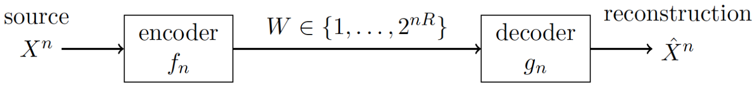


Figure: Rate distortion encoder and decoder: $\hat{X}^n$ is called as vector quantization/reproduction/ reconstruction of X^n.

Definition

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 associated distortion is defined as

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

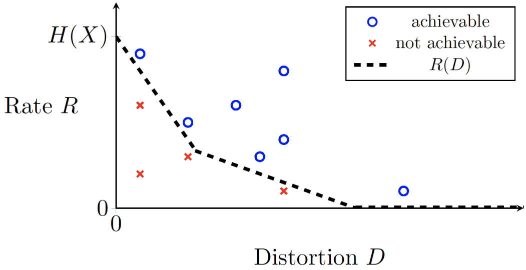
Rate-distortion region

Definition (Achievable rate-distortion pair)

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 $\lim_{n\to\infty} \mathrm{E}[d(X^n, \hat{X}^n)] \le D$.

Definition (Rate distortion region)

The rate distortion region for a source is the closure of the set of achievable rate distortion pairs (R, D).



Rate-distortion Function

Definition (Rate distortion function (from operational perspective))

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.

Definition (Distortion rate function)

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.

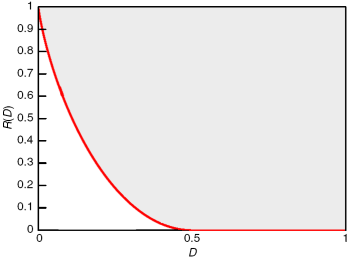


Figure: The rate-distortion function of a Bernoulli random variable with Hamming distortion

Information Rate-distortion Function

Definition

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

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

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

Theorem

The rate-distortion function $R(D)$ is a non-increasing convex function of D.

Proof: The value function of a convex problem is a convex and non-increasing function in the right-hand input parameter, which is the distortion D in our case.

Example 1 of Rate-distortion Functions

Example

The rate-distortion function for a Bernoulli(p) source with Hamming distortion is

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

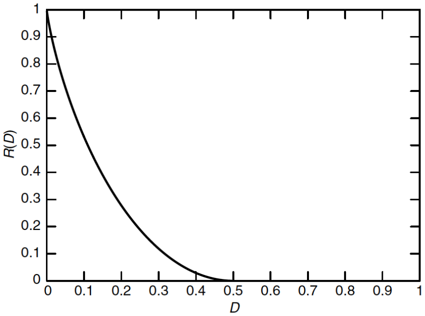


Figure: Rate distortion function for a Bernoulli(0.5) source.

$R(D)$ of Bernoulli Source and Hamming Distortion (Proof)

WLOG, we assume $p \leq \frac{1}{2}$. We will first show a lower bound of $I(X;\hat{X})$, and then find the lower-bound achieving probability.

Let $\oplus$ denote modulo 2 addition, i.e., $X \oplus \hat{X} = 1 \Leftrightarrow X \neq \hat{X}$. Thus, we have

$$
\begin{aligned}
I(X;\hat{X}) &= H(X) - H(X \mid \hat{X}) && \text{(1a)}\\
&= H(X) - H(X \oplus \hat{X} \mid \hat{X}) && \text{(1b)}\\
&\geq H(X) - H(X \oplus \hat{X}) && \text{(1c)}\\
&= H(p) - H(Pr(X \neq \hat{X})) && \text{(1d)}\\
&= H(p) - H(E(d(X,\hat{X}))) && \text{(1e)}\\
&\geq H(p) - H(D), && \text{(1f)}
\end{aligned}
$$

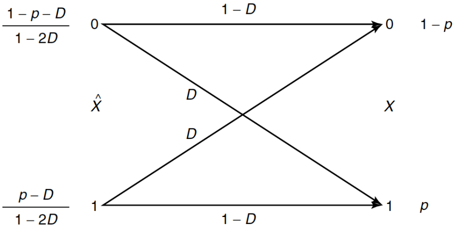
where

- (1c) is because conditioning reduces entropy.
- (1e) is due to Hamming distortion.
- (1f) is because $E(d(X,\hat{X})) \leq D$ and entropy function $H(\cdot)$ is monotonically increasing in $(0, 0.5)$.

$R(D)$ of Bernoulli Source and Hamming Distortion (Proof-cont'd)

To find the lower-bound achieving probability, consider an additive noisy channel (BSC): $X = \hat{X} \oplus N$, where $X \sim \text{Bern}(p)$ and $N \sim \text{Bern}(D)$.

- We must find a distribution of $\hat{X}$ to achieve the derived lower bound (1f).
- Let $\hat{X} \sim \text{Bern}(r)$, where $p = r \times (1 - D) + (1 - r) \times D \implies r = \frac{p-D}{1-2D}$ is a legitimate distribution provided that $D \leq p \leq \frac{1}{2}$.
- Since $X \oplus \hat{X} = N \perp\!\!\!\perp \hat{X}$, so (1c) and (1f) hold with equality. If $D \geq p$, let $\hat{X} \equiv 0$ and no code/description for X. Then, we have $R(D) = 0$ and $D = p$.
- The distortion (expected Hamming distance) between X^n and $\hat{X}^n$ is exactly p because the number of 1's in X^n is p on average.



Example 2 of Rate-distortion Functions

Example

The rate-distortion function for a Gaussian $\mathcal{N}(0, \sigma^2)$ source with square-error distortion is

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

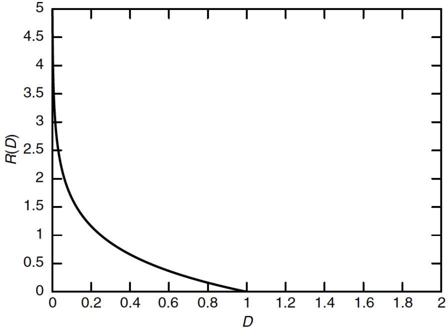


Figure: Rate distortion function for a Gaussian source.

$R(D)$ of Gaussian Source and Square-error Distortion (Proof)

As in the preceding example, we first find a lower bound and then prove that this is achievable.

$$
\begin{aligned}
I(X;\hat{X}) &= h(X) - h(X \mid \hat{X}) && \text{(2a)}\\
&= \frac{1}{2}\log(2\pi e)\sigma^2 - h(X-\hat{X} \mid \hat{X}) && \text{(2b)}\\
&\geq \frac{1}{2}\log(2\pi e)\sigma^2 - h(X-\hat{X}) && \text{(2c)}\\
&\geq \frac{1}{2}\log(2\pi e)\sigma^2 - h\left(\mathcal{N}\left(0, E(X-\hat{X})^2\right)\right) && \text{(2d)}\\
&= \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)E(X-\hat{X})^2 && \text{(2e)}\\
&\geq \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)D && \text{(2f)}\\
&= \frac{1}{2}\log\frac{\sigma^2}{D}, && \text{(2g)}
\end{aligned}
$$

where (2d) is due to the fact that the normal distribution maximizes the entropy for a given second moment.

$R(D)$ of of Gaussian Source and Square-error Distortion (Proof-cont'd)

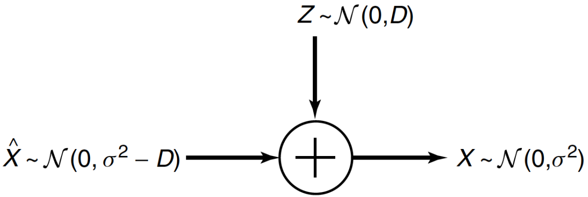


Figure: Joint distribution for Gaussian source. If $D \leq \sigma^2$, then $X = \hat{X} + Z$ with the assigned Gaussian distributions shown above achieves the lower bound.

- If $D \geq \sigma^2$, let $\hat{X} \equiv 0$ and no code/description for X. Then, we have $D = E(X - \hat{X})^2 = EX^2 = \sigma^2$.
- Hence, the distortion-rate function is $D(R) = (\frac{1}{4})^R \sigma^2 \implies$ Each bit of description reduces the expected distortion by a factor of 4.

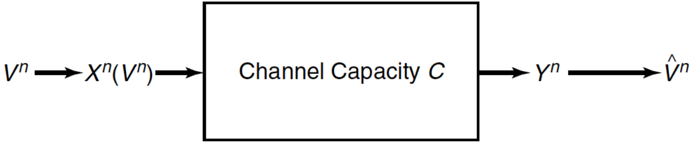
Rate Distortion Coding Theorem

Theorem (Rate Distortion Coding Theorem)

The rate-distortion function for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x, \hat{x})$ is equal to the associated information rate distortion function. That is, $R(D) = R^{(I)}(D)$.

- The theorem asserts that the operational and informational definitions of the rate-distortion function coincide with each other.
- Achievability: $R(D) > R^{(I)}(D) \implies$ the pair (R, D) is achievable. This can be proved by distortion ϵ-typicality.
- Converse: the pair (R, D) is achievable $\implies R(D) \geq R^{(I)}(D)$. This can be proved by DPI and convexity of $R(D)$.
- See detailed proof in Sections 10.5-10.6 of Cover's book.

Source–channel Separation Theorem with Distortion



Theorem (Source–channel separation theorem with distortion)

Let $V_1, V_2, \ldots, V_n$ be a finite alphabet i.i.d. source which is encoded as a sequence of n input symbols X^n of a DMC with capacity C. The output of the channel Y^n is mapped onto the reconstruction alphabet $\hat{V}^n = g\left(Y^n\right)$. Let $D = Ed\left(V^n, \hat{V}^n\right)$ be the average distortion achieved by this combined source and channel coding scheme. Then distortion D is achievable if and only if $C > R(D)$.

Thank You!

Email: <zhangy@ucsc.edu>

Homepage: https://people.ucsc.edu/~yzhan419/