Information Theory and Related Fileds

Lecture 2: Source Coding

Lei Yu

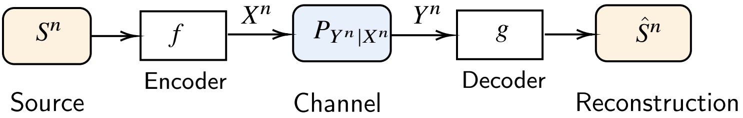
Nankai University

Online Short Course at Beijing Normal University

Outline

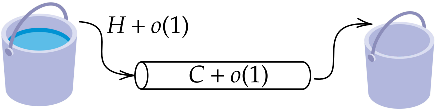
Outline

Recall: Source-Channel Coding Theorem

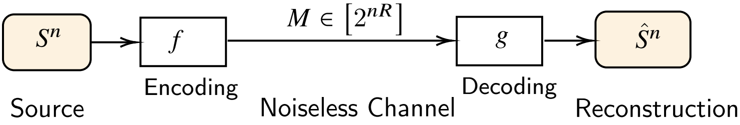


Theorem ([Shannon'48])

Consider discrete memoryless source S and discrete memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < C(P_{Y|X})$, and only if $H(S) \leq C(P_{Y|X})$.

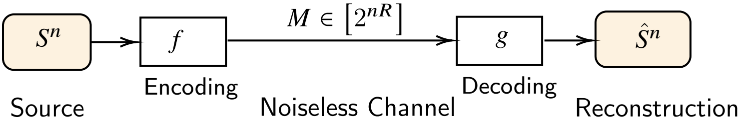


A Special Case: Source Coding



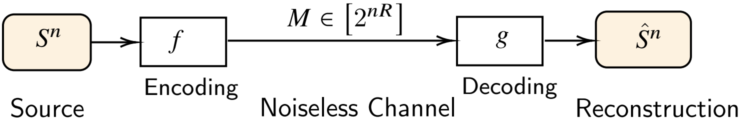
- Throughout this course, $[x] := \{1, 2, ..., \lfloor x \rfloor\}$

A Special Case: Source Coding



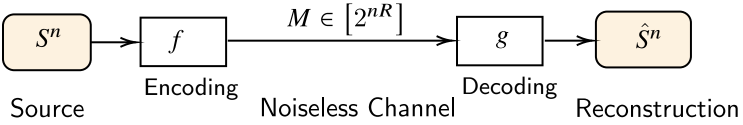
- Throughout this course, $[x] := \{1, 2, ..., \lfloor x \rfloor\}$
- A noiseless rate-R channel is $M \longmapsto M$ for any r.v. $M \in [2^{nR}]$.
  - That is, the output is always identical to the input.
  - The rate is the exponent of the size of the range of the input.

A Special Case: Source Coding



- Throughout this course, $[x] := \{1, 2, ..., \lfloor x \rfloor\}$
- A noiseless rate-R channel is $M \longmapsto M$ for any r.v. $M \in [2^{nR}]$.
  - That is, the output is always identical to the input.
  - The rate is the exponent of the size of the range of the input.
- For this case, $f : \mathcal{S}^n \to [2^{nR}]$ and $g : [2^{nR}] \to \hat{\mathcal{S}}^n$ are respectively also called source encoder and source decoder, and R is also called the rate of (f, g).

A Special Case: Source Coding



- Throughout this course, $[x] := \{1, 2, ..., \lfloor x \rfloor\}$
- A noiseless rate-R channel is $M \longmapsto M$ for any r.v. $M \in [2^{nR}]$.
  - That is, the output is always identical to the input.
  - The rate is the exponent of the size of the range of the input.
- For this case, $f : \mathcal{S}^n \to [2^{nR}]$ and $g : [2^{nR}] \to \hat{\mathcal{S}}^n$ are respectively also called source encoder and source decoder, and R is also called the rate of (f, g).
- Essence of source coding (quantization): Represent a source S^n by another source $\hat{S}^n$ such that the range of $\hat{S}^n$ is no larger than 2^{nR} and moreover, $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$.

As a special case of the source-channel coding theorem, Shannon showed:

Source Coding Theorem

As a special case of the source-channel coding theorem, Shannon showed:

Theorem (Source Coding [Shannon'48])

Consider a discrete memoryless source S and a noiseless rate-R channel. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < R$, and only if $H(S) \leq R$.

Source Coding Theorem

As a special case of the source-channel coding theorem, Shannon showed:

Theorem (Source Coding [Shannon'48])

Consider a discrete memoryless source S and a noiseless rate-R channel. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < R$, and only if $H(S) \leq R$.

We next prove this theorem.

Outline

Asymptotic Equipartition Property (AEP)

- Throughout this course, $\log := \log_2$.

Asymptotic Equipartition Property (AEP)

- Throughout this course, $\log := \log_2$.
- Consider a memoryless source S. The first n symbols $S^n \sim P_S^{\otimes n}$.

Asymptotic Equipartition Property (AEP)

- Throughout this course, $\log := \log_2$.
- Consider a memoryless source S. The first n symbols $S^n \sim P_S^{\otimes n}$.
- The self-information for S^n is the r.v.

$$-\log P_S^{\otimes n}(S^n) = \sum_{i=1}^{n} -\log P_S(S_i)$$

where $\iota(S_i) := -\log P_S(S_i)$ are i.i.d.

Asymptotic Equipartition Property (AEP)

- Throughout this course, $\log := \log_2$.
- Consider a memoryless source S. The first n symbols $S^n \sim P_S^{\otimes n}$.
- The self-information for S^n is the r.v.

$$-\log P_S^{\otimes n}(S^n) = \sum_{i=1}^{n} -\log P_S(S_i)$$

where $\iota(S_i) := -\log P_S(S_i)$ are i.i.d.

- By Law of Large Numbers (LLN), we obtain

Theorem (AEP)

$$-\frac{1}{n} \log P_S^{\otimes n}(S^n) \to H(S) \text{ in probability.}$$

That is, for any $\epsilon > 0$, $\mathbb{P}\left\{\left|-\frac{1}{n} \log P_S^{\otimes n}(S^n) - H(S)\right| \le \epsilon\right\} \to 1$ *as* $n \to \infty$.

Typical Set

Definition

The (weakly) typical set $\mathcal{A}_\epsilon^{(n)}(P_S)$ (or shortly, $\mathcal{A}_\epsilon^{(n)}$) with respect to P_S is the set of sequences $s^n \in \mathcal{S}^n$ such that

$$\left| -\frac{1}{n} \log P_S^{\otimes n}(s^n) - H(S) \right| \leq \epsilon.$$

Properties of Typical Set

Fact

1. *For any* $s^n \in \mathcal{A}_\epsilon^{(n)}$, $2^{-n(H(S)+\epsilon)} \leq P_S^{\otimes n}(s^n) \leq 2^{-n(H(S)-\epsilon)}$.
2. $P_S^{\otimes n}(\mathcal{A}_\epsilon^{(n)}) > 1 - \epsilon$ *for sufficiently large* n.
3. $|\mathcal{A}_\epsilon^{(n)}| \leq 2^{n(H(S)+\epsilon)}$.
4. $|\mathcal{A}_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(S)-\epsilon)}$ *for sufficiently large* n.

Properties of Typical Set

Fact

1. *For any* $s^n \in \mathcal{A}_\epsilon^{(n)}$, $2^{-n(H(S)+\epsilon)} \le P_S^{\otimes n}(s^n) \le 2^{-n(H(S)-\epsilon)}$.
2. $P_S^{\otimes n}(\mathcal{A}_\epsilon^{(n)}) > 1-\epsilon$ *for sufficiently large* n.
3. $|\mathcal{A}_\epsilon^{(n)}| \le 2^{n(H(S)+\epsilon)}$.
4. $|\mathcal{A}_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(S)-\epsilon)}$ *for sufficiently large* n.

Proof: 1. By definition. 2. By the AEP. 3.

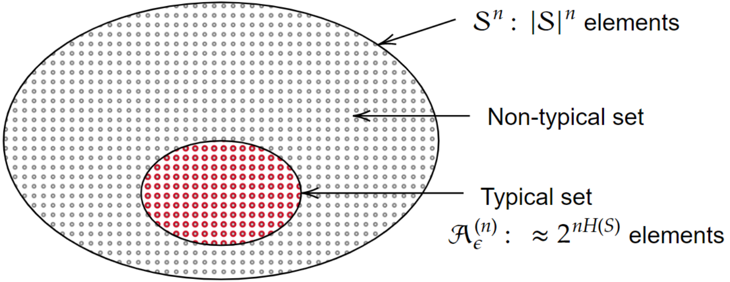
$$\begin{aligned} 1 = \sum_{s^n \in \mathcal{S}^n} P_S^{\otimes n}(s^n) &\ge \sum_{s^n \in \mathcal{A}_\epsilon^{(n)}} P_S^{\otimes n}(s^n) \\ &\ge \sum_{s^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(S)+\epsilon)} = 2^{-n(H(S)+\epsilon)}|\mathcal{A}_\epsilon^{(n)}| \end{aligned}$$

4. By Statement 2,

$$\begin{aligned} 1-\epsilon &< \sum_{s^n \in \mathcal{A}_\epsilon^{(n)}} P_S^{\otimes n}(s^n) \\ &\le \sum_{s^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(S)-\epsilon)} = 2^{-n(H(S)-\epsilon)}|\mathcal{A}_\epsilon^{(n)}| \end{aligned}$$

Concentration of A Memoryless Source

The typical set $\mathcal{A}_\epsilon^{(n)}$ is a high-probability set of size no larger than $2^{n(H(S)+\epsilon)}$.



Is there a high-probability set having size smaller than $2^{n(H(S)-\epsilon)}$ (for some $\epsilon > 0$)?

Concentration of A Memoryless Source (cont.)

Is there a high-probability set having size smaller than $2^{n(H(S)-\epsilon)}$ (for some $\epsilon > 0$)?

Theorem (Smallest High-probability Sets)

Let $S_1, S_2, \ldots$ be i.i.d. $\sim P_S$. If $P_S^{\otimes n}(\mathcal{B}_n) > 1 - \delta$ for $0 < \delta < 1$, then for any $\epsilon > 0$,

$$|\mathcal{B}_n| \geq 2^{n(H(S)-\epsilon)} \text{ for sufficiently large } n.$$

Is there a high-probability set having size smaller than $2^{n(H(S)-\epsilon)}$ (for some $\epsilon > 0$)?

Theorem (Smallest High-probability Sets)

Let $S_1, S_2, \ldots$ be i.i.d. $\sim P_S$. If $P_S^{\otimes n}(\mathcal{B}_n) > 1 - \delta$ for $0 < \delta < 1$, then for any $\epsilon > 0$,

$$|\mathcal{B}_n| \geq 2^{n(H(S)-\epsilon)} \text{ for sufficiently large } n.$$

Fact

Typical sets are smallest high-probability sets. The smallest size is roughly $2^{nH(S)}$.

Proof of Smallest High-probability Sets

By the inclusion-exclusion principle,

$$\begin{aligned} P_S^{\otimes n}(\mathcal{A}_\epsilon^{(n)} \cap \mathcal{B}_n) &= P_S^{\otimes n}(\mathcal{A}_\epsilon^{(n)}) + P_S^{\otimes n}(\mathcal{B}_n) - P_S^{\otimes n}(\mathcal{A}_\epsilon^{(n)} \cup \mathcal{B}_n) \\ &\geq 1 - \epsilon + 1 - \delta - 1 \\ &= 1 - \epsilon - \delta. \end{aligned}$$

On the other hand,

$$\begin{aligned} P_S^{\otimes n}(\mathcal{A}_\epsilon^{(n)} \cap \mathcal{B}_n) = \sum_{s^n \in \mathcal{A}_\epsilon^{(n)} \cap \mathcal{B}_n} P_S^{\otimes n}(s^n) &\leq \sum_{s^n \in \mathcal{A}_\epsilon^{(n)} \cap \mathcal{B}_n} 2^{-n(H(S)-\epsilon)} \\ &= |\mathcal{A}_\epsilon^{(n)} \cap \mathcal{B}_n| 2^{-n(H(S)-\epsilon)} \\ &\leq |\mathcal{B}_n| 2^{-n(H(S)-\epsilon)}. \end{aligned}$$

Therefore,

$$|\mathcal{B}_n| \geq (1 - \epsilon - \delta) 2^{n(H(S)-\epsilon)} = 2^{n(H(S)-\epsilon+o(1))}.$$

AEP for Continuous Sources

Definition

The (weakly) typical set $\mathcal{A}_\epsilon^{(n)}(P_S)$ (or shortly, $\mathcal{A}_\epsilon^{(n)}$) with respect to continuous distribution P_S (with PDF p_S) is the set of sequences $s^n \in \mathcal{S}^n$ such that

$$\left| -\frac{1}{n} \log p_S^{\otimes n}(s^n) - h(S) \right| \leq \epsilon,$$

where $p_S^{\otimes n}(s^n) = \prod_{i=1}^n p_S(s_i)$. Recall that $h(S) = -\int p_S(s) \log p_S(s) \mathrm{d}s$.

Fact: 1. For any $s^n \in \mathcal{A}_\epsilon^{(n)}$, $2^{-n(h(S)+\epsilon)} \leq p_S^{\otimes n}(s^n) \leq 2^{-n(h(S)-\epsilon)}$.

2. (AEP) $P_S^{\otimes n}(\mathcal{A}_\epsilon^{(n)}) \to 1$ as $n \to \infty$.

3. $\mathrm{Vol}(\mathcal{A}_\epsilon^{(n)}) \leq 2^{n(h(S)+\epsilon)}$.

4. $\mathrm{Vol}(\mathcal{A}_\epsilon^{(n)}) \geq (1-\epsilon) 2^{n(h(S)-\epsilon)}$ for sufficiently large n.

5. The set $\mathcal{A}_\epsilon^{(n)}$ is the smallest volume set with probability $\geq 1-\epsilon$, to first order in the exponent.

Joint AEP

Definition

The (weakly) joint typical set $\mathcal{A}_\epsilon^{(n)}(P_{S\hat{S}})$ (or shortly, $\mathcal{A}_\epsilon^{(n)}$) with respect to $P_{S\hat{S}}$ is the set of $(s^n, \hat{s}^n) \in \mathcal{S}^n \times \hat{\mathcal{S}}^n$ such that

$$\left|-\frac{1}{n}\log P_S^{\otimes n}(s^n) - H(S)\right| \le \epsilon$$

$$\left|-\frac{1}{n}\log P_{\hat{S}}^{\otimes n}(\hat{s}^n) - H(\hat{S})\right| \le \epsilon$$

$$\left|-\frac{1}{n}\log P_{S\hat{S}}^{\otimes n}(s^n, \hat{s}^n) - H(S, \hat{S})\right| \le \epsilon.$$

(For continuous distributions, replace P with p and H with h.)

Fact: 1. (Joint AEP) $P_{S\hat{S}}^{\otimes n}(\mathcal{A}_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
2. $|\mathcal{A}_\epsilon^{(n)}| \le 2^{n\left(H(S,\hat{S})+\epsilon\right)}$.
3. $|\mathcal{A}_\epsilon^{(n)}| \ge (1-\epsilon)2^{n\left(H(S,\hat{S})-\epsilon\right)}$ for sufficiently large n.

Outline

Achievability Part ("If" Part)

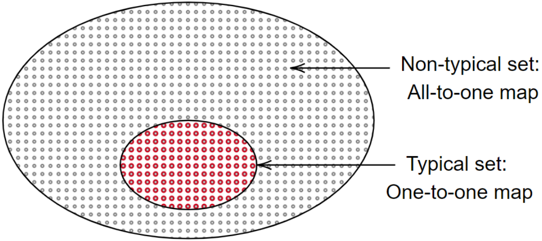
- Partition $\mathcal{S}^n$ into the typical set $\mathcal{A}_\epsilon^{(n)}$ and its complement $\mathcal{A}_\epsilon^{(n)\,c}$.

Achievability Part ("If" Part)

- Partition $\mathcal{S}^n$ into the typical set $\mathcal{A}_\epsilon^{(n)}$ and its complement $\mathcal{A}_\epsilon^{(n)\,c}$.
- Index $\mathcal{A}_\epsilon^{(n)}$ by $1, 2, ..., L$, and hence, $\mathcal{A}_\epsilon^{(n)} = \{s^n(1), s^n(2), ..., s^n(L)\}$, where $L = |\mathcal{A}_\epsilon^{(n)}|$.

Achievability Part ("If" Part)

- Partition $\mathcal{S}^n$ into the typical set $\mathcal{A}_\epsilon^{(n)}$ and its complement $\mathcal{A}_\epsilon^{(n)\,c}$.
- Index $\mathcal{A}_\epsilon^{(n)}$ by $1, 2, \ldots, L$, and hence, $\mathcal{A}_\epsilon^{(n)} = \{s^n(1), s^n(2), \ldots, s^n(L)\}$, where $L = |\mathcal{A}_\epsilon^{(n)}|$.
- Consider the following coding scheme.
  - Encoder: If $s^n \in \mathcal{A}_\epsilon^{(n)}$, send the index i of s^n; otherwise, send 1.
  - Decoder: Reconstruction $s^n(i)$



Analysis of Achievability

Calculation of probability of error:

- If $S^n \in \mathcal{A}_\epsilon^{(n)}$, then reconstruction is exactly S^n (no error).
- Denoting the reconstruction as $\hat{S}^n$ and $\mathrm{error} := \left\{S^n \neq \hat{S}^n\right\}$, we have

$$\begin{aligned}\mathbb{P}(\mathrm{error}) &= \mathbb{P}(S^n \in \mathcal{A}_\epsilon^{(n)})\mathbb{P}(\mathrm{error}|S^n \in \mathcal{A}_\epsilon^{(n)}) \\ &\qquad + \mathbb{P}(S^n \notin \mathcal{A}_\epsilon^{(n)})\mathbb{P}(\mathrm{error}|S^n \notin \mathcal{A}_\epsilon^{(n)}) \\ &\leq 0 + \mathbb{P}(S^n \notin \mathcal{A}_\epsilon^{(n)}) \\ &\to 0\end{aligned}$$

Analysis of Achievability

Calculation of probability of error:

- If $S^n \in \mathcal{A}_\epsilon^{(n)}$, then reconstruction is exactly S^n (no error).
- Denoting the reconstruction as $\hat{S}^n$ and $\mathrm{error} := \left\{S^n \neq \hat{S}^n\right\}$, we have

$$\begin{aligned}\mathbb{P}(\mathrm{error}) &= \mathbb{P}(S^n \in \mathcal{A}_\epsilon^{(n)})\mathbb{P}(\mathrm{error}|S^n \in \mathcal{A}_\epsilon^{(n)}) \\ &\qquad + \mathbb{P}(S^n \notin \mathcal{A}_\epsilon^{(n)})\mathbb{P}(\mathrm{error}|S^n \notin \mathcal{A}_\epsilon^{(n)}) \\ &\le 0 + \mathbb{P}(S^n \notin \mathcal{A}_\epsilon^{(n)}) \\ &\to 0\end{aligned}$$

Calculation of rate: $\frac{1}{n}\log|\mathcal{A}_\epsilon^{(n)}| \le H(S) + \epsilon$ which is arbitrarily close to $H(S)$ (by letting $\epsilon \to 0$)

Converse Part ("Only If" Part)

Lemma (Fano's inequality [Cover–Thomas' book])

Given two random variables X and Y, let $\hat{X} = g(Y)$ be any estimator of X given Y and let $\epsilon = \mathbb{P}(X \neq \hat{X})$ be the probability of error. Then

$$H(X|Y) \leq H(X|\hat{X}) \leq H_2(\epsilon) + \epsilon \log |\mathcal{X}|.$$

This inequality can be weakened to

$$H(X|Y) \leq 1 + \epsilon \log |\mathcal{X}|.$$

Converse Part (cont.)

Proof of Converse: For a pair of rate-R encoder-decoder (f_n, g_n), denote $M_n = f_n(S^n)$ and $\hat{S}^n = g_n(M_n)$. Denote $\epsilon_n = \mathbb{P}(S^n \neq \hat{S}^n)$. We then have

$$
\begin{aligned}
\log 2^{nR} &\geq H(M_n) \qquad \text{maximum entropy} \\
&\geq I(S^n; M_n) \\
&= H(S^n) - H(S^n|M_n) \\
&\geq nH(S) - \left(1 + \epsilon_n \log |\mathcal{S}^n|\right) \qquad \text{Fano's inequality}
\end{aligned}
$$

Proof of Converse: For a pair of rate-R encoder-decoder (f_n, g_n), denote $M_n = f_n(S^n)$ and $\hat{S}^n = g_n(M_n)$. Denote $\epsilon_n = \mathbb{P}(S^n \neq \hat{S}^n)$. We then have

$$\begin{aligned}
\log 2^{nR} &\geq H(M_n) \qquad \text{maximum entropy}\\
&\geq I(S^n; M_n)\\
&= H(S^n) - H(S^n|M_n)\\
&\geq nH(S) - \left(1 + \epsilon_n \log|\mathcal{S}^n|\right) \qquad \text{Fano's inequality}
\end{aligned}$$

So,

$$R \geq H(S) - \frac{1}{n} - \epsilon_n \log|\mathcal{S}|.$$

Since $\epsilon_n \to 0$ as $n \to \infty$, taking $\lim_{n\to\infty}$, we then have

$$R \geq H(S).$$

Converse Part (cont.)

Proof of Converse: For a pair of rate-R encoder-decoder (f_n, g_n), denote $M_n = f_n(S^n)$ and $\hat{S}^n = g_n(M_n)$. Denote $\epsilon_n = \mathbb{P}(S^n \neq \hat{S}^n)$. We then have

$$\begin{aligned}
\log 2^{nR} &\geq H(M_n) \qquad \text{maximum entropy} \\
&\geq I(S^n; M_n) \\
&= H(S^n) - H(S^n|M_n) \\
&\geq nH(S) - \left(1 + \epsilon_n \log|\mathcal{S}^n|\right) \qquad \text{Fano's inequality}
\end{aligned}$$

So,

$$R \geq H(S) - \frac{1}{n} - \epsilon_n \log|\mathcal{S}|.$$

Since $\epsilon_n \to 0$ as $n \to \infty$, taking $\lim_{n\to\infty}$, we then have

$$R \geq H(S).$$

(Here we assume $|\mathcal{S}| < \infty$, but this assumption can be removed by using information-spectral method instead [Han's book])

Outline

Distortion Is Sometimes Allowed

- When transmitting a image/video, a certain level of distortion is usually allowed.

Distortion Is Sometimes Allowed

- When transmitting a image/video, a certain level of distortion is usually allowed.
- A distortion function is a mapping

$$d : \mathcal{S} \times \hat{\mathcal{S}} \to [0, \infty)$$

A distortion function is said to be bounded if $\sup_{s,\hat{s}} d(s, \hat{s}) < \infty$.

Distortion Is Sometimes Allowed

- When transmitting a image/video, a certain level of distortion is usually allowed.
- A distortion function is a mapping

$$d : \mathcal{S} \times \hat{\mathcal{S}} \to [0, \infty)$$

  A distortion function is said to be bounded if $\sup_{s,\hat{s}} d(s, \hat{s}) < \infty$.
- Examples:
  - Hamming distortion: $d(s, \hat{s}) = \begin{cases} 0 & s = \hat{s} \\ 1 & s \neq \hat{s} \end{cases}$
  - Squared-error distortion: $d(s, \hat{s}) = (s - \hat{s})^2$

Distortion Is Sometimes Allowed

- When transmitting a image/video, a certain level of distortion is usually allowed.
- A distortion function is a mapping

$$d : \mathcal{S} \times \hat{\mathcal{S}} \to [0, \infty)$$

  A distortion function is said to be bounded if $\sup_{s,\hat{s}} d(s, \hat{s}) < \infty$.
- Examples:
  - Hamming distortion: $d(s, \hat{s}) = \begin{cases} 0 & s = \hat{s} \\ 1 & s \neq \hat{s} \end{cases}$
  - Squared-error distortion: $d(s, \hat{s}) = (s - \hat{s})^2$
- The distortion between sequences s^n and $\hat{s}^n$ is defined by

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

Rate-Distortion Function

- A rate-distortion pair (R, D) is said to be achievable if there exists a sequence of rate-R encoder and decoder (f_n, g_n) such that

$$\limsup_{n\to\infty} \mathbb{E} d(S^n, g_n(f_n(S^n))) \le D.$$

Rate-Distortion Function

- A rate-distortion pair (R, D) is said to be achievable if there exists a sequence of rate-R encoder and decoder (f_n, g_n) such that

$$\limsup_{n\to\infty} \mathbb{E}d(S^n, g_n(f_n(S^n))) \leq D.$$

- The (operational) rate-distortion function $R_{\text{op}}(D)$ is the infimum of rates R such that (R, D) is achievable.

Lossy Source Coding Theorem

Theorem (Lossy Source Coding [Shannon'48])

Consider a discrete memoryless source S and a bounded distortion function d. Then,

$$R_{\mathrm{op}}(D) = R(D) := \min_{P_{\hat{S}|S}:\mathbb{E}d(S,\hat{S})\leq D} I(S;\hat{S}).$$

Example: Binary Source

Fact

The rate-distortion function for a $\mathrm{Bern}(p)$ *source with Hamming distortion is given by*

$$R(D) = \begin{cases} H_2(p) - H_2(D) & 0 \le D \le \min\{p, 1-p\} \\ 0 & D > \min\{p, 1-p\} \end{cases}.$$

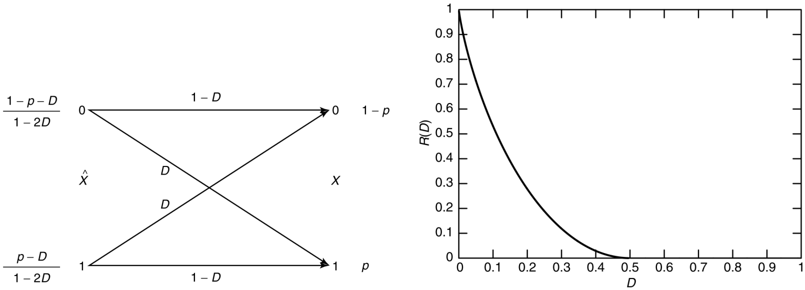


Figure: (left) optimal $P_{S\hat{S}}$ for $\mathrm{Bern}(p)$, and (right) $R(D)$ for $\mathrm{Bern}(\frac{1}{2})$

Example: Gaussian Source

Fact

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

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

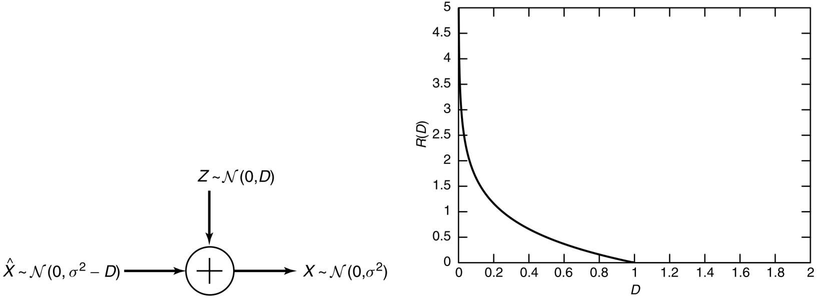


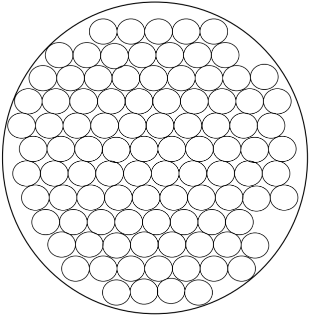
Figure: (left) optimal $P_{S\hat{S}}$ for $\mathcal{N}(0, \sigma^2)$, and (right) $R(D)$ for $\mathcal{N}(0, 1)$

Intuition for Gaussian Source

- For $S \sim \mathcal{N}(0, \sigma^2)$, $h(S) = \frac{1}{2}\log(2\pi e \sigma^2)$. So, $\mathcal{A}_\epsilon^{(n)} = \left\{s^n : \left|\frac{1}{n}\sum_{i=1}^n s_i^2 - \sigma^2\right| \le \epsilon'\right\}$ where $\epsilon' = \frac{2\sigma^2}{\log e}\epsilon$.
- That is, S^n is concentrated on a thin spherical shell (or a ball) of radius around $\sqrt{n}\sigma$

Intuition for Gaussian Source

- For $S \sim \mathcal{N}(0, \sigma^2)$, $h(S) = \frac{1}{2}\log(2\pi e\sigma^2)$. So, $\mathcal{A}_\epsilon^{(n)} = \left\{s^n : \left|\frac{1}{n}\sum_{i=1}^n s_i^2 - \sigma^2\right| \le \epsilon'\right\}$ where $\epsilon' = \frac{2\sigma^2}{\log e}\epsilon$.
- That is, S^n is concentrated on a thin spherical shell (or a ball) of radius around $\sqrt{n}\sigma$
- Covering a radius-$\sqrt{n}\sigma$ ball by radius-$\sqrt{nD}$ balls: The number of small balls is at least $\frac{\mathrm{Vol}(\mathrm{Ball}_{\sqrt{n}\sigma})}{\mathrm{Vol}(\mathrm{Ball}_{\sqrt{nD}})} = \frac{(\sqrt{n}\sigma)^n}{\left(\sqrt{nD}\right)^n} = 2^{n\cdot\frac{1}{2}\log\frac{\sigma^2}{D}}$



Proof of Converse Part (i.e., $R_{\mathrm{op}}(D) \geq R(D)$)

For a pair of rate-R encoder-decoder (f_n, g_n), denote $M_n = f_n(S^n)$ and $\hat{S}^n = g_n(M_n)$. Obviously, $S^n \leftrightarrow M_n \leftrightarrow \hat{S}^n$. We then have

$$
\begin{aligned}
nR &\geq H(M_n) \qquad \text{maximum entropy} \\
&\geq I(S^n; M_n) \geq I(S^n; \hat{S}^n) \qquad \text{DPI for mutual information} \\
&= H(S^n) - H(S^n|\hat{S}^n) \\
&= \sum_{i=1}^{n} H(S_i) - \sum_{i=1}^{n} H(S_i|\hat{S}^n, S^{i-1}) \qquad \text{chain rule} \\
&\geq \sum_{i=1}^{n} H(S_i) - \sum_{i=1}^{n} H(S_i|\hat{S}_i) \qquad \text{conditioning reduces entropy} \\
&= \sum_{i=1}^{n} I(S_i; \hat{S}_i)
\end{aligned}
$$

Proof of Converse Part (cont.)

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} I(S_i; \hat{S}_i) && \text{copy from last slide} \\
&\geq \sum_{i=1}^{n} R\left(\mathbb{E}[d(S_i, \hat{S}_i)]\right) && \text{definition of function } R(D) \\
&= n\left(\frac{1}{n}\sum_{i=1}^{n} R\left(\mathbb{E}[d(S_i, \hat{S}_i)]\right)\right) \\
&\geq nR\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[d(S_i, \hat{S}_i)]\right) && R(D) \text{ is convex} \\
&= nR\left(\mathbb{E}[d(S^n, \hat{S}^n)]\right) \\
&\geq nR(D) && R(D) \text{ is nonincreasing}
\end{aligned}
$$

Proof of Achievability Part (i.e., $R_{\mathrm{op}}(D) \le R(D)$)

Definition

The distortion-typical set $\mathcal{A}_{d,\epsilon}^{(n)}(P_{S\hat{S}})$ (or shortly, $\mathcal{A}_{d,\epsilon}^{(n)}$) with respect to $P_{S\hat{S}}$ is the set of $(s^n, \hat{s}^n) \in \mathcal{S}^n \times \hat{\mathcal{S}}^n$ such that

$$\left| -\frac{1}{n} \log P_S^{\otimes n}(s^n) - H(S) \right| \le \epsilon$$

$$\left| -\frac{1}{n} \log P_{\hat{S}}^{\otimes n}(\hat{s}^n) - H(\hat{S}) \right| \le \epsilon \qquad \text{jointly typical}$$

$$\left| -\frac{1}{n} \log P_{S\hat{S}}^{\otimes n}(s^n, \hat{s}^n) - H(S, \hat{S}) \right| \le \epsilon$$

$$\left| d(s^n, \hat{s}^n) - \mathbb{E} d(S, \hat{S}) \right| \le \epsilon.$$

Proof of Achievability Part (i.e., $R_{\mathrm{op}}(D) \le R(D)$)

Definition

The distortion-typical set $\mathcal{A}_{d,\epsilon}^{(n)}(P_{S\hat{S}})$ (or shortly, $\mathcal{A}_{d,\epsilon}^{(n)}$) with respect to $P_{S\hat{S}}$ is the set of $(s^n, \hat{s}^n) \in \mathcal{S}^n \times \hat{\mathcal{S}}^n$ such that

$$\left| -\frac{1}{n} \log P_S^{\otimes n}(s^n) - H(S) \right| \le \epsilon$$

$$\left| -\frac{1}{n} \log P_{\hat{S}}^{\otimes n}(\hat{s}^n) - H(\hat{S}) \right| \le \epsilon \qquad \text{jointly typical}$$

$$\left| -\frac{1}{n} \log P_{S\hat{S}}^{\otimes n}(s^n, \hat{s}^n) - H(S, \hat{S}) \right| \le \epsilon$$

$$\left| d(s^n, \hat{s}^n) - \mathbb{E} d(S, \hat{S}) \right| \le \epsilon.$$

Fact: 1. (Joint AEP) $P_{S\hat{S}}^{\otimes n}(\mathcal{A}_{d,\epsilon}^{(n)}) \to 1$ as $n \to \infty$.

2. [Cover–Thomas' book] Let $(S'^n, \hat{S}'^n) \sim P_S^{\otimes n} \otimes P_{\hat{S}}^{\otimes n}$. For sufficiently large n,

$$(1-\epsilon)2^{-n\left(I(S;\hat{S})+3\epsilon\right)} \le \mathbb{P}\left\{(S'^n, \hat{S}'^n) \in \mathcal{A}_{d,\epsilon}^{(n)}\right\} \le 2^{-n\left(I(S;\hat{S})-3\epsilon\right)}.$$

Coding Scheme

- Let $P_{\hat{S}|S}$ attain $R(D) = \min_{P_{\hat{S}|S}:\mathbb{E}d(S,\hat{S})\leq D} I(S;\hat{S})$. Let R be any number $> I(S;\hat{S}) + 3\epsilon = R(D) + 3\epsilon$.

Coding Scheme

- Let $P_{\hat{S}|S}$ attain $R(D) = \min_{P_{\hat{S}|S}:\mathbb{E}d(S,\hat{S})\le D} I(S;\hat{S})$. Let R be any number $> I(S;\hat{S}) + 3\epsilon = R(D) + 3\epsilon$.
- **Generation of codebook:** Randomly generate 2^{nR} sequences (codewords) $\hat{S}^n$ drawn i.i.d. $\sim P_{\hat{S}}^{\otimes n}$. Index them by $i \in [2^{nR}]$. Denote $\mathcal{C} := \left\{\hat{S}^n(1), \hat{S}^n(2), ..., \hat{S}^n(2^{nR})\right\}$, which is called (random) codebook. Reveal this codebook to the encoder and decoder.

Coding Scheme

- Let $P_{\hat{S}|S}$ attain $R(D) = \min_{P_{\hat{S}|S}:\mathbb{E}d(S,\hat{S})\leq D} I(S;\hat{S})$. Let R be any number $> I(S;\hat{S}) + 3\epsilon = R(D) + 3\epsilon$.
- **Generation of codebook:** Randomly generate 2^{nR} sequences (codewords) $\hat{S}^n$ drawn i.i.d. $\sim P_{\hat{S}}^{\otimes n}$. Index them by $i \in [2^{nR}]$. Denote $\mathcal{C} := \left\{\hat{S}^n(1), \hat{S}^n(2), ..., \hat{S}^n(2^{nR})\right\}$, which is called (random) codebook. Reveal this codebook to the encoder and decoder.
- **Encoding:** Encode S^n by i if there exists an i such that $(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}$. If there is more than one such i, send any one of them. If there is no such i, let $i = 1$ and declare an error. Thus, nR bits suffice to describe the index i of the jointly typical codeword.

Coding Scheme

- Let $P_{\hat{S}|S}$ attain $R(D) = \min_{P_{\hat{S}|S}:\mathbb{E}d(S,\hat{S})\le D} I(S;\hat{S})$. Let R be any number $> I(S;\hat{S}) + 3\epsilon = R(D) + 3\epsilon$.
- **Generation of codebook:** Randomly generate 2^{nR} sequences (codewords) $\hat{S}^n$ drawn i.i.d. $\sim P_{\hat{S}}^{\otimes n}$. Index them by $i \in [2^{nR}]$. Denote $\mathcal{C} := \left\{\hat{S}^n(1), \hat{S}^n(2), ..., \hat{S}^n(2^{nR})\right\}$, which is called (random) codebook. Reveal this codebook to the encoder and decoder.
- **Encoding:** Encode S^n by i if there exists an i such that $(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}$. If there is more than one such i, send any one of them. If there is no such i, let $i = 1$ and declare an error. Thus, nR bits suffice to describe the index i of the jointly typical codeword.
- **Decoding:** The reconstruction is $\hat{S}^n(i)$.

Calculation of Probability of Error

Lemma ([Cover–Thomas' book])

If $R > I(S;\hat{S}) + 3\epsilon$, then $\mathbb{P}_{S^n,C}(\mathrm{error}) \to 0$ as $n \to 0$.

Intuition behind this lemma:

- Observe that S^n and $\hat{S}^n(1), \hat{S}^n(2), \ldots, \hat{S}^n(2^{nR})$ are independent, and hence, $\mathbb{P}\left\{(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}\right\} \approx 2^{-nI(S;\hat{S})}$ for all $i \in [2^{nR}]$.
- So, the averaged number of codewords $\hat{S}^n(i)$ such that $(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}$ is $2^{n(R-I(S;\hat{S}))}$ which is exponentially large when $R > I(S;\hat{S})$.

- On one hand, by the lemma above, with high probability, no error occurs, i.e., $(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}$.

Calculation of Distortion

- On one hand, by the lemma above, with high probability, no error occurs, i.e., $(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}$.
- On the other hand, if no error, then by definition of distortion-typical sets and by the choice of $P_{\hat{S}|S}$,

$$d(S^n, \hat{S}^n(i)) \le \mathbb{E}d(S, \hat{S}) + \epsilon \le D + \epsilon. \quad (1)$$

Calculation of Distortion

- On one hand, by the lemma above, with high probability, no error occurs, i.e., $(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}$.
- On the other hand, if no error, then by definition of distortion-typical sets and by the choice of $P_{\hat{S}|S}$,

$$d(S^n, \hat{S}^n(i)) \le \mathbb{E}d(S, \hat{S}) + \epsilon \le D + \epsilon. \tag{1}$$

- So, (1) holds with high probability, and hence,

$$\mathbb{E}_{S^n, \mathcal{C}}\left[d(S^n, \hat{S}^n(i))\right] \le D + 2\epsilon. \tag{2}$$

That is, $(R(D) + 3\epsilon, D + 2\epsilon)$ is achievable, i.e., $R_{\mathrm{op}}(D) \le R(D)$ (by letting $\epsilon \to 0$).

Calculation of Distortion

- On one hand, by the lemma above, with high probability, no error occurs, i.e., $(S^n, \hat{S}^n(i)) \in \mathcal{A}_{d,\epsilon}^{(n)}$.
- On the other hand, if no error, then by definition of distortion-typical sets and by the choice of $P_{\hat{S}|S}$,

$$d(S^n, \hat{S}^n(i)) \leq \mathbb{E}d(S, \hat{S}) + \epsilon \leq D + \epsilon. \quad (1)$$

- So, (1) holds with high probability, and hence,

$$\mathbb{E}_{S^n,C}\left[d(S^n, \hat{S}^n(i))\right] \leq D + 2\epsilon. \quad (2)$$

  That is, $(R(D) + 3\epsilon, D + 2\epsilon)$ is achievable, i.e., $R_{\text{op}}(D) \leq R(D)$ (by letting $\epsilon \to 0$).
- Removing randomness of codebook: Since (2) holds on average over C, there must exist a fixed codebook c such that $\mathbb{E}_{S^n}\left[d(S^n, \hat{S}^n(i))|C = c\right] \leq D + 2\epsilon$.

References

1. Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.
2. Cover, T. M. (1999). Elements of information theory. John Wiley & Sons.
3. Han, T. S. (2002). Information-spectrum methods in information theory. Springer Science & Business Media.

Thank you for your attention!