Chapter 6

Lossy Data Compression and Transmission

Po-Ning Chen, Professor

Institute of Communications Engineering

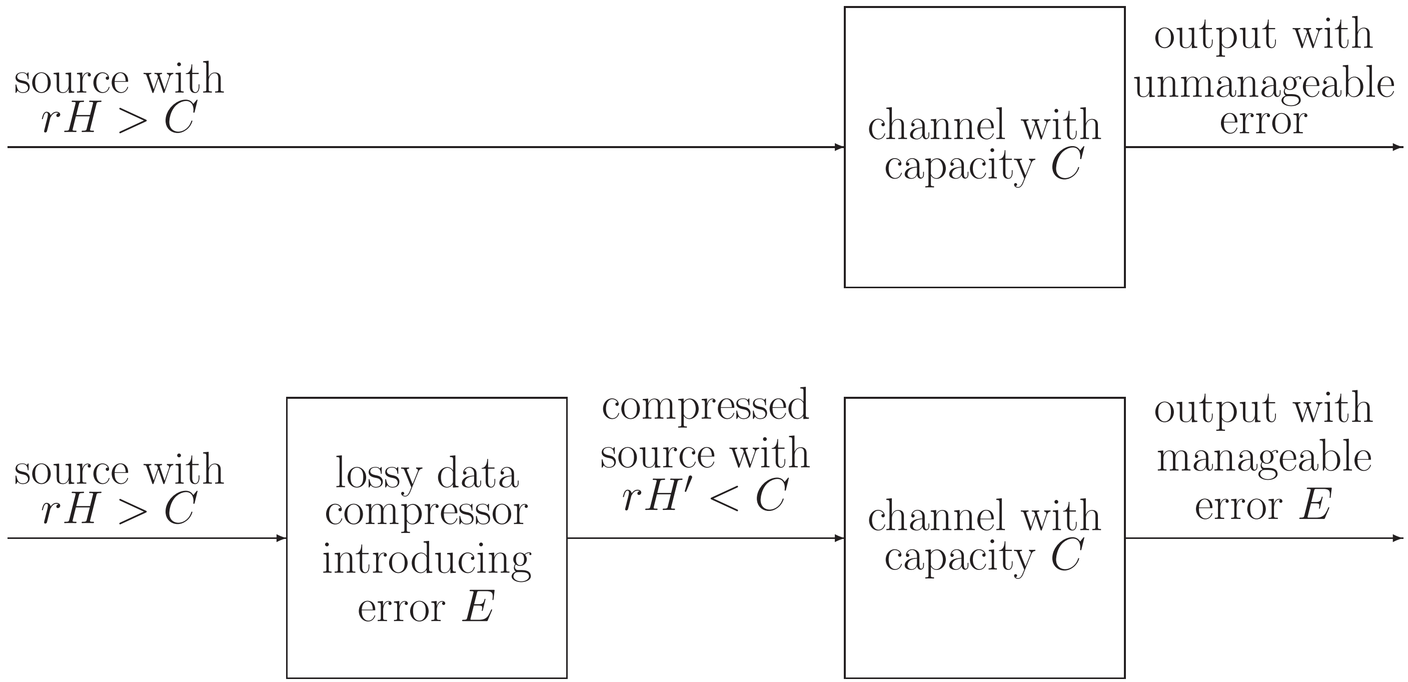
National Chiao Tung University

Hsin Chu, Taiwan 30010, R.O.C.

6.1.1 Motivation

Motivations

- Lossy data compression = to compress a source to a rate less than the source entropy.



Example for the application of lossy data compression

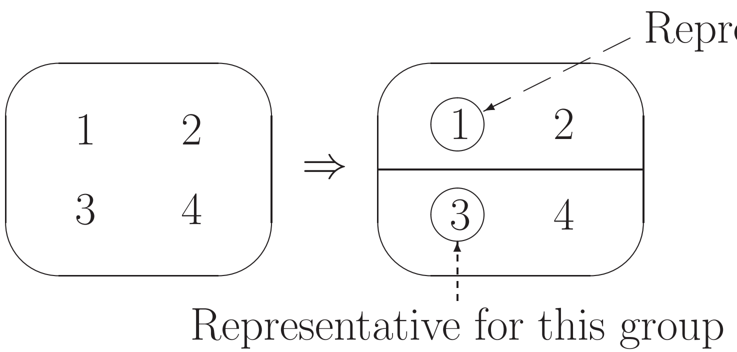
6.1.2 Distortion measures

Definition 6.4 (Distortion measure) A distortion measure is a mapping

$$\rho : \mathcal{Z} \times \hat{\mathcal{Z}} \to \Re^{+},$$

where $\mathcal{Z}$ is the source alphabet, $\hat{\mathcal{Z}}$ is the reproduction alphabet for compressed code, and $\Re^{+}$ is the set of non-negative real numbers.

- The distortion measure $\rho(z, \hat{z})$ can be viewed as the cost of representing the source symbol $z \in \mathcal{Z}$ by a reproduction symbol $\hat{z} \in \hat{\mathcal{Z}}$.
  E.g. A lossy data compression is similar to "grouping."



$$[\rho(i,j)] := \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 2 & 2 \\ 2 & 2 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{bmatrix},$$

6.1.2 Distortion measures

– Average distortion under uniform source distribution

$$\frac{1}{4}\rho(1,1) + \frac{1}{4}\rho(2,1) + \frac{1}{4}\rho(3,3) + \frac{1}{4}\rho(4,3) = \frac{1}{2}.$$

– Resultant entropy

$$H(Z) = \log 2(4) = 2 \text{ bits} \quad \Rightarrow \quad H(\hat{Z}) = \log_2(2) = 1 \text{ bit}.$$

6.1.2 Distortion measures

- The above example presumes $|\hat{\mathcal{Z}}| = |\mathcal{Z}|$.
- Sometimes, it is convenient to have $|\hat{\mathcal{Z}}| = |\mathcal{Z}| + 1$.
  E.g. $|\mathcal{Z} = \{1, 2, 3\}| = 3$ and $|\hat{\mathcal{Z}} = \{1, 2, 3, E\}| = 4$
  and the distortion measure is defined by

$$[\rho(i, j)] := \begin{bmatrix} 0 & 2 & 2 & 0.5 \\ 2 & 0 & 2 & 0.5 \\ 2 & 2 & 0 & 0.5 \end{bmatrix}.$$

  - Suppose only two outcomes are allowed under uniform Z.
    Then

$$(1) \to 1 \quad \text{and} \quad (2, 3) \to E$$

    is an optimal choice (that minimizes the average distortion measure for a given compression rate).
  - Average distortion

$$\frac{1}{3}\rho(1, 1) + \frac{1}{3}\rho(2, E) + \frac{1}{3}\rho(3, E) = \frac{1}{3}.$$

  - Resultant entropy

$$H(Z) = \log_2(3) \text{ bits} \quad \Rightarrow \quad H(\hat{Z}) = [\log_2(3) - 2/3] \text{ bits}.$$

6.1.3 Frequently used distortion measures

Example 6.5 (Hamming distortion measure) Let source alphabet and reproduction alphabet be the same, i.e., $\mathcal{Z} = \hat{\mathcal{Z}}$. Then the Hamming distribution measure is given by

$$\rho(z, \hat{z}) := \begin{cases} 0, & \text{if } z = \hat{z}; \\ 1, & \text{if } z \neq \hat{z}. \end{cases}$$

This is also named the *probability-of-error distortion measure* because

$$E[\rho(Z, \hat{Z})] = \Pr(Z \neq \hat{Z}).$$

Example 6.6 (Absolute error distortion measure) Assuming that $\mathcal{Z} = \hat{\mathcal{Z}} = \mathbb{R}$, the absolute error distortion measure is given by

$$\rho(z, \hat{z}) := |z - \hat{z}|.$$

Example 6.7 (Squared error distortion) Assuming that $\mathcal{Z} = \hat{\mathcal{Z}} = \mathbb{R}$, the squared error distortion measure is given by

$$\rho(z, \hat{z}) := (z - \hat{z})^2.$$

The squared error distortion measure is perhaps the most popular distortion measure used for continuous alphabets.

Comments on squared error distortion

- The squared error distortion measure has the advantages of simplicity and having a closed-form solution for most cases of interest, such as when using least squares prediction.
- Yet, this measure is not ideal for practical situations involving data operated by human observers (such as image and speech data) as it is inadequate in measuring perceptual quality.
- For example, two speech waveforms in which one is a marginally time-shifted version of the other may have large square error distortion; however, they sound quite similar to the human ear.

Distortion measure for sequences

Definition 6.8 (Additive distortion measure between vectors) The additive distortion measure ρ_n between vectors z^n and $\hat{z}^n$ of size n (or n-sequences or n-tuples) is defined by

$$\rho_n(z^n, \hat{z}^n) = \sum_{i=1}^{n} \rho(z_i, \hat{z}_i).$$

Definition 6.9 (Maximum distortion measure)

$$\rho_n(z^n, \hat{z}^n) = \max_{1 \le i \le n} \rho(z_i, \hat{z}_i).$$

Question raised due to distortion measures for sequences

- Whether to reproduce source sequence z^n by sequence $\hat{z}^n$ of the same length is a must or not.
- In other words, can we use $\tilde{z}^k$ to represent z^n for $k \neq n$?

Answer: The answer is certainly *yes* if a distortion measure for z^n and $\tilde{z}^k$ is defined.

Distortion measure for sequences

Problem: A problem for taking $k \neq n$ is that the distortion measure for sequences can no longer be defined based on per-letter distortions, and hence a per-letter formula for the best lossy data compression rate cannot be rendered.

Solution: To view the lossy data compression in two steps.

Step 1 : Find the data compression code

$$h : \mathcal{Z}^n \to \hat{\mathcal{Z}}^n$$

for which the pre-specified **distortion constraint** and **rate constraint** are both satisfied.

Step 2 : Derive the (asymptotically) lossless data compression block code for source $h(Z^n)$. The existence of such code with block length

$$k > H(h(Z^n)) \text{ bits}$$

is guaranteed by Shannon's lossless source coding theorem.

- Therefore, a lossy data compression code from

$$\mathcal{Z}^n \left(\to \hat{\mathcal{Z}}^n\right) \to \{0,1\}^k$$

is established.

Distortion measure for sequences

• Since the second step is already discussed in lossless data compression, we can say that the theorem regarding the lossy data compression is basically a theorem on the first step.

6.2 Fixed-length lossy data compression

Definition 6.10 (Fixed-length lossy data compression code subject to average distortion constraint) An (n, M, D) fixed-length lossy data compression code for source alphabet $\mathcal{Z}^n$ and reproduction alphabet $\hat{\mathcal{Z}}^n$ consists of a compression function

$$h \ : \ \mathcal{Z}^n \to \hat{\mathcal{Z}}^n$$

with the size of the codebook (i.e., the image $h(\mathcal{Z}^n)$) being $|h(\mathcal{Z}^n)| = M$, and the average distortion satisfying

$$E\left[\frac{1}{n}\rho_n(Z^n, h(Z^n))\right] \le D.$$

- Code rate for lossy data compression

$$\frac{1}{n}\log_2 M \text{ bits/sourceword}$$

- Asymptotic code rate for lossy data compression

$$\limsup_{n\to\infty}\frac{1}{n}\log_2 M \text{ bits/sourceword}$$

Achievable Rate-Distortion Pair

Definition 6.11 (Achievable rate-distortion pair) For a given sequence of distortion measures $\{\rho_n\}_{n\geq 1}$, a rate distortion pair (R, D) is *achievable* if there exists a sequence of fixed-length lossy data compression codes (n, M_n, D) with ultimate code rate

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 M_n \leq R.$$

Definition 6.12 (Rate-distortion region) The rate-distortion region R of a source $\{Z_n\}$ is the closure of the set of all achievable rate-distortion pair (R, D).

Achievable Rate-Distortion Pair

Lemma 6.13 (Time-sharing principle) Under an additive distortion measure ρ_n, the rate-distortion region $\mathcal{R}$ is a convex set; i.e., if $(R_1, D_1) \in \mathcal{R}$ and $(R_2, D_2) \in \mathcal{R}$, then $(\lambda R_1 + (1-\lambda)R_2, \lambda D_1 + (1-\lambda)D_2) \in \mathcal{R}$ for all $0 \le \lambda \le 1$.

Proof:

- *time-sharing* argument:
  - If we can use an (n, M_1, D_1) code $\mathscr{C}_1$ to achieve (R_1, D_1) and an (n, M_2, D_2) code $\mathscr{C}_2$ to achieve (R_2, D_2), then for any rational number $0 < \lambda < 1$, we can use $\mathscr{C}_1$ for a fraction λ of the time and use $\mathscr{C}_2$ for a fraction $1-\lambda$ of the time to achieve (R_λ, D_λ), where $R_\lambda = \lambda R_1 + (1-\lambda)R_2$ and $D_\lambda = \lambda D_1 + (1-\lambda)D_2$;
  - hence the result holds for any real number $0 < \lambda < 1$ by the density of the rational numbers in $\mathbb{R}$ and the continuity of R_λ and D_λ in λ.
- Let r and s be positive integers and let $\lambda = \frac{r}{r+s}$; then $0 < \lambda < 1$.

Achievable Rate-Distortion Pair

- Assume that the pairs (R_1, D_1) and (R_2, D_2) are achievable. Then there exist a sequence of (n, M_1, D_1) codes $\mathscr{C}_1$ and a sequence of (n, M_2, D_2) codes $\mathscr{C}_2$ such that for n sufficiently large,

$$\frac{1}{n}\log_2 M_1 \le R_1$$

and

$$\frac{1}{n}\log_2 M_2 \le R_2.$$

- Construct a sequence of new codes $\mathscr{C}$ of blocklength $n_\lambda = (r+s)n$, codebook size $M = M_1^r \times M_2^s$ and compression function $h\colon \mathcal{Z}^{(r+s)n} \to \hat{\mathcal{Z}}^{(r+s)n}$ such that

$$h(z^{(r+s)n}) = (h_1(z_1^n), \ldots, h_1(z_r^n), h_2(z_{r+1}^n), \ldots, h_2(z_{r+s}^n))$$

where

$$z^{(r+s)n} = (z_1^n, \ldots, z_r^n, z_{r+1}^n, \ldots, z_{r+s}^n)$$

and h_1 and h_2 are the compression functions of $\mathscr{C}_1$ and $\mathscr{C}_2$, respectively.

Achievable Rate-Distortion Pair

- The average (or expected) distortion under the additive distortion measure ρ_n and the rate of code $\mathscr{C}$ are given by

$$
\begin{aligned}
E\left[\frac{\rho_{(r+s)n}(z^{(r+s)n}, h(z^{(r+s)n}))}{(r+s)n}\right] &= \frac{1}{r+s}\left(E\left[\frac{\rho_n(z_1^n, h_1(z_1^n))}{n}\right] + \cdots + E\left[\frac{\rho_n(z_r^n, h_1(z_r^n))}{n}\right]\right. \\
&\quad \left. + E\left[\frac{\rho_n(z_{r+1}^n, h_2(z_{r+1}^n))}{n}\right] + \cdots + E\left[\frac{\rho_n(z_{r+s}^n, h_2(z_{r+s}^n))}{n}\right]\right) \\
&\le \frac{1}{r+s}(rD_1 + sD_2) \\
&= \lambda D_1 + (1-\lambda)D_2 = D_\lambda
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{(r+s)n}\log_2 M &= \frac{1}{(r+s)n}\log_2(M_1^r \times M_2^s) \\
&= \frac{r}{(r+s)}\frac{1}{n}\log_2 M_1 + \frac{s}{(r+s)}\frac{1}{n}\log_2 M_2 \\
&\le \lambda R_1 + (1-\lambda)R_2 = R_\lambda,
\end{aligned}
$$

respectively, for n sufficiently large. Thus, (R_λ, D_λ) is achievable by $\mathscr{C}$. □

Achievable Rate-Distortion Pair

Definition 6.14 (Rate-distortion function) The rate-distortion function, denoted by $R(D)$, of source $\{Z_n\}$ is the smallest $\hat{R}$ for a given distortion threshold D such that $(\hat{R}, D)$ is an achievable rate-distortion pair; i.e.,

$$R(D) := \inf\{\hat{R} \geq 0 \colon (\hat{R}, D) \in \mathcal{R}\}.$$

Observation 6.15 (Monotonicity and convexity of $\boldsymbol{R(D)}$) Note that, under an additive distortion measure ρ_n, the rate-distortion function $R(D)$ is non-increasing and convex in D (the proof is left as an exercise).

6.3 Rate-distortion theorem

Definition 6.16 (Distortion typical set) The *distortion δ-typical set* with respect to the memoryless (product) distribution $P_{Z,\hat{Z}}$ on $\mathcal{Z}^n \times \hat{\mathcal{Z}}^n$ and a bounded additive distortion measure $\rho_n(\cdot,\cdot)$ is defined by

$$\begin{aligned}
\mathcal{D}_n(\delta) \; := \; \Bigg\{ & (z^n, \hat{z}^n) \in \mathcal{Z}^n \times \hat{\mathcal{Z}}^n : \\
& \left| -\frac{1}{n} \log_2 P_{Z^n}(z^n) - H(Z) \right| < \delta, \\
& \left| -\frac{1}{n} \log_2 P_{\hat{Z}^n}(\hat{z}^n) - H(\hat{Z}) \right| < \delta, \\
& \left| -\frac{1}{n} \log_2 P_{Z^n,\hat{Z}^n}(z^n, \hat{z}^n) - H(Z, \hat{Z}) \right| < \delta, \\
& \text{and } \left| \frac{1}{n} \rho_n(z^n, \hat{z}^n) - E[\rho(Z, \hat{Z})] \right| < \delta \Bigg\}.
\end{aligned}$$

AEP for distortion typical set

Theorem 6.17 If $(Z_1, \hat{Z}_1)$, $(Z_2, \hat{Z}_2)$, ..., $(Z_n, \hat{Z}_n)$, ... are i.i.d., and ρ_n are bounded additive distortion measure, then

$$-\frac{1}{n}\log_2 P_{Z^n}(Z_1, Z_2, \ldots, Z_n) \to H(Z) \quad \text{in probability;}$$

$$-\frac{1}{n}\log_2 P_{\hat{Z}^n}(\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_n) \to H(\hat{Z}) \quad \text{in probability;}$$

$$-\frac{1}{n}\log_2 P_{Z^n,\hat{Z}^n}((Z_1, \hat{Z}_1), \ldots, (Z_n, \hat{Z}_n)) \to H(Z, \hat{Z}) \quad \text{in probability;}$$

and

$$\frac{1}{n}\rho_n(Z^n, \hat{Z}^n) \to E[\rho(Z, \hat{Z})] \quad \text{in probability.}$$

Proof: Functions of independent random variables are also independent random variables. Thus by the weak law of large numbers, we have the desired result. □

- It needs to be pointed out that without the bounded property assumption, the normalized sum of an i.i.d. sequence does not necessarily converge in probability to a finite mean, hence the need for requiring that ρ be bounded.

AEP for distortion typical set

Theorem 6.18 (AEP for distortion measure) Given a DMS $\{(Z_n, \hat{Z}_n)\}$ with generic joint distribution $P_{Z,\hat{Z}}$ and any $\delta > 0$, the distortion δ-typical set satisfies

1. $P_{Z^n,\hat{Z}^n}(\mathcal{D}_n^c(\delta)) < \delta$ for n sufficiently large.
2. For all $(z^n, \hat{z}^n)$ in $\mathcal{D}_n(\delta)$,

$$P_{\hat{Z}^n}(\hat{z}^n) \geq P_{\hat{Z}^n|Z^n}(\hat{z}^n|z^n) 2^{-n[I(Z;\hat{Z})+3\delta]}. \tag{6.3.1}$$

Proof: The first result follows directly from Theorem 6.17 and the definition of the distortion typical set $\mathcal{D}_n(\delta)$. The second result can be proved as follows:

$$\begin{aligned}
P_{\hat{Z}^n|Z^n}(\hat{z}^n|z^n) &= \frac{P_{Z^n,\hat{Z}^n}(z^n, \hat{z}^n)}{P_{Z^n}(z^n)} \\
&= P_{\hat{Z}^n}(\hat{z}^n) \frac{P_{Z^n,\hat{Z}^n}(z^n, \hat{z}^n)}{P_{Z^n}(z^n) P_{\hat{Z}^n}(\hat{z}^n)} \\
&\leq P_{\hat{Z}^n}(\hat{z}^n) \frac{2^{-n[H(Z,\hat{Z})-\delta]}}{2^{-n[H(Z)+\delta]} 2^{-n[H(\hat{Z})+\delta]}} \\
&= P_{\hat{Z}^n}(\hat{z}^n) 2^{n[I(Z;\hat{Z})+3\delta]},
\end{aligned}$$

where the inequality follows from the definition of $\mathcal{D}_n(\delta)$. □

AEP for distortion typical set

- Alternative form of (6.3.1):

$$\frac{P_{Z^n,\hat{Z}^n}(z^n,\hat{z}^n)}{P_{Z^n}(z^n)P_{\hat{Z}^n}(\hat{z}^n)} \le 2^{-n[I(Z;\hat{Z})+3\delta]} \quad \text{for all } (z^n,\hat{z}^n)\in\mathcal{D}_n(\delta).$$

Lemma 6.19 For $0 \le x \le 1$, $0 \le y \le 1$, and $n > 0$,

$$(1-xy)^n \le 1 - x + e^{-yn}, \tag{6.3.2}$$

with equality holds if, and only if, $(x, y) = (1, 0)$.

Theorem 6.20 (Shannon's rate-distortion theorem for memoryless sources) Consider a DMS $\{Z_n\}_{n=1}^{\infty}$ with alphabet $\mathcal{Z}$, reproduction alphabet $\hat{\mathcal{Z}}$ and a bounded additive distortion measure $\rho_n(\cdot,\cdot)$; i.e.,

$$\rho_n(z^n, \hat{z}^n) = \sum_{i=1}^{n} \rho(z_i, \hat{z}_i) \quad \text{and} \quad \rho_{\max} := \max_{(z,\hat{z}) \in \mathcal{Z} \times \hat{\mathcal{Z}}} \rho(z, \hat{z}) < \infty,$$

where $\rho(\cdot,\cdot)$ is a given single-letter distortion measure. Then the source's rate-distortion function satisfies the following expression

$$R(D) = \min_{P_{\hat{Z}|Z}:\ E[\rho(Z,\hat{Z})] \le D} I(Z; \hat{Z}).$$

Proof: Define

$$R^{(I)}(D) := \min_{P_{\hat{Z}|Z}:\ E[\rho(Z,\hat{Z})] \le D} I(Z; \hat{Z}); \tag{6.3.3}$$

this quantity is typically called Shannon's *information rate-distortion function*. We will then show that the (operational) rate-distortion function $R(D)$ given in Definition 6.14 equals $R^{(I)}(D)$.

Shannon's lossy source coding theorem

1. *Achievability Part* (i.e., $R(D+\varepsilon) \le R^{(I)}(D) + 4\varepsilon$ for arbitrarily small $\varepsilon > 0$): We need to show that for any $\varepsilon > 0$, there exist $0 < \gamma < 4\varepsilon$ and a sequence of lossy data compression codes $\{(n, M_n, D+\varepsilon)\}_{n=1}^{\infty}$ with

$$\limsup_{n\to\infty} \frac{1}{n}\log_2 M_n \le R^{(I)}(D) + \gamma < R^{(I)}(D) + 4\varepsilon.$$

Step 1: Optimizing conditional distribution. Let $P_{\tilde{Z}|Z}$ be the conditional distribution that achieves $R^{(I)}(D)$, i.e.,

$$R^{(I)}(D) = \min_{P_{\hat{Z}|Z}:\, E[\rho(Z,\hat{Z})]\le D} I(Z;\hat{Z}) = I(Z;\tilde{Z}).$$

Then

$$E[\rho(Z,\tilde{Z})] \le D.$$

Choose M_n to satisfy

$$R^{(I)}(D) + \frac{1}{2}\gamma \le \frac{1}{n}\log_2 M_n \le R^{(I)}(D) + \gamma$$

for some γ in $(0, 4\varepsilon)$, for which the choice should exist for all sufficiently large $n > N_0$ for some N_0. Define

$$\delta := \min\Bigg\{ \underbrace{\frac{\gamma}{8}}_{\text{Required in Step 4}}, \underbrace{\frac{\varepsilon}{1+2\rho_{\max}}}_{\text{Required in Step 5}} \Bigg\}.$$

Step 2: Random coding. Independently select M_n words from $\hat{\mathcal{Z}}^n$ according to

$$P_{\tilde{Z}^n}(\tilde{z}^n) = \prod_{i=1}^{n} P_{\tilde{Z}}(\tilde{z}_i),$$

and denote this random codebook by $\mathscr{C}_n$, where

$$P_{\tilde{Z}}(\tilde{z}) = \sum_{z \in \mathcal{Z}} P_Z(z) P_{\tilde{Z}|Z}(\tilde{z}|z).$$

Shannon's lossy source coding theorem

Step 3: Encoding rule. Define a subset of $\mathcal{Z}^n$ as

$$\mathcal{J}(\mathscr{C}_n) := \{z^n \in \mathcal{Z}^n \colon \exists\ \tilde{z}^n \in \mathscr{C}_n \text{ such that } (z^n, \tilde{z}^n) \in \mathcal{D}_n(\delta)\},$$

where $\mathcal{D}_n(\delta)$ is defined under $P_{\tilde{Z}|Z}$. Based on the codebook

$$\mathscr{C}_n = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{M_n}\},$$

define the encoding rule as:

$$h_n(z^n) = \begin{cases} \boldsymbol{c}_m, & \text{if } (z^n, \boldsymbol{c}_m) \in \mathcal{D}_n(\delta); \\ & \text{(when more than one satisfying the requirement,} \\ & \text{ just pick any.)} \\ \text{any word in } \mathscr{C}_n, & \text{otherwise.} \end{cases}$$

Note that when $z^n \in \mathcal{J}(\mathscr{C}_n)$, we have $(z^n, h_n(z^n)) \in \mathcal{D}_n(\delta)$ and

$$\frac{1}{n}\rho_n(z^n, h_n(z^n)) \le E[\rho(Z, \tilde{Z})] + \delta \le D + \delta.$$

Shannon's lossy source coding theorem

Step 4: Calculation of the probability of the complement of $\mathcal{J}(\mathscr{C}_n)$.

Let N_1 be chosen such that for $n > N_1$,

$$P_{Z^n,\tilde{Z}^n}(\mathcal{D}_n^c(\delta)) < \delta.$$

Let

$$\Omega := P_{Z^n}(\mathcal{J}^c(\mathscr{C}_n)).$$

Then the expected probability of source n-tuples not belonging to $\mathcal{J}(\mathscr{C}_n)$, averaged over all randomly generated codebooks, is given by

$$\begin{aligned} E[\Omega] &= \sum_{\mathscr{C}_n} P_{\tilde{Z}^n}(\mathscr{C}_n)\left(\sum_{z^n \notin \mathcal{J}(\mathscr{C}_n)} P_{Z^n}(z^n)\right) \\ &= \sum_{z^n \in \mathcal{Z}^n} P_{Z^n}(z^n)\left(\sum_{\mathscr{C}_n :\ z^n \notin \mathcal{J}(\mathscr{C}_n)} P_{\tilde{Z}^n}(\mathscr{C}_n)\right). \end{aligned}$$

Shannon's lossy source coding theorem

For any z^n given, to select a codebook $\mathscr{C}_n$ satisfying $z^n \notin \mathcal{J}(\mathscr{C}_n)$ is equivalent to *independently* draw M_n n-tuples from $\hat{\mathcal{Z}}^n$ which are not jointly distortion typical with z^n. Hence,

$$\sum_{\mathscr{C}_n :\, z^n \notin \mathcal{J}(\mathscr{C}_n)} P_{\tilde{Z}^n}(\mathscr{C}_n) = \left(\Pr\left[(z^n, \tilde{Z}^n) \notin \mathcal{D}_n(\delta)\right]\right)^{M_n}.$$

For convenience, we let $K(z^n, \tilde{z}^n)$ denote the indicator function of $\mathcal{D}_n(\delta)$, i.e.,

$$K(z^n, \tilde{z}^n) = \begin{cases} 1, & \text{if } (z^n, \tilde{z}^n) \in \mathcal{D}_n(\delta); \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\sum_{\mathscr{C}_n :\, z^n \notin \mathcal{J}(\mathscr{C}_n)} P_{\tilde{Z}^n}(\mathscr{C}_n) = \left(1 - \sum_{\tilde{z}^n \in \hat{\mathcal{Z}}^n} P_{\tilde{Z}^n}(\tilde{z}^n) K(z^n, \tilde{z}^n)\right)^{M_n}.$$

Shannon's lossy source coding theorem

Continuing the computation of $E[\Omega]$, we get

$$
\begin{aligned}
E[\Omega] &= \sum_{z^n\in\mathcal{Z}^n} P_{Z^n}(z^n)\left(1-\sum_{\tilde{z}^n\in\hat{\mathcal{Z}}^n} P_{\tilde{Z}^n}(\tilde{z}^n)K(z^n,\tilde{z}^n)\right)^{M_n}\\
&\leq \sum_{z^n\in\mathcal{Z}^n} P_{Z^n}(z^n)\left(1-\sum_{\tilde{z}^n\in\hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n)2^{-n(I(Z;\tilde{Z})+3\delta)}K(z^n,\tilde{z}^n)\right)^{M_n} \quad \text{(by (6.3.1))}\\
&= \sum_{z^n\in\mathcal{Z}^n} P_{Z^n}(z^n)\left(1-2^{-n(I(Z;\tilde{Z})+3\delta)}\sum_{\tilde{z}^n\in\hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n)K(z^n,\tilde{z}^n)\right)^{M_n}\\
&\leq \sum_{z^n\in\mathcal{Z}^n} P_{Z^n}(z^n)\left(1-\sum_{\tilde{z}^n\in\hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n)K(z^n,\tilde{z}^n)+\exp\left\{-M_n\cdot 2^{-n(I(Z;\tilde{Z})+3\delta)}\right\}\right) \quad \text{(from (6.3.2))}\\
&\leq \sum_{z^n\in\mathcal{Z}^n} P_{Z^n}(z^n)\left(1-\sum_{\tilde{z}^n\in\hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n)K(z^n,\tilde{z}^n)\right.\\
&\qquad \left.+\exp\left\{-2^{n(R^{(I)}(D)+\gamma/2)}\cdot 2^{-n(I(Z;\tilde{Z})+3\delta)}\right\}\right) \quad \text{(for } R^{(I)}(D)+\gamma/2<(1/n)\log_2 M_n\text{)}\\
&\leq 1-P_{Z^n,\tilde{Z}^n}(\mathcal{D}_n(\delta))+\exp\left\{-2^{n\delta}\right\} \quad \text{(for } R^{(I)}(D)=I(Z;\tilde{Z}) \text{ and } \delta\leq\gamma/8\text{)}\\
&= P_{Z^n,\tilde{Z}^n}(\mathcal{D}_n^c(\delta))+\exp\left\{-2^{n\delta}\right\}\\
&\leq \delta+\delta=2\delta
\end{aligned}
$$

for all $n > N := \max\left\{N_0, N_1, \frac{1}{\delta}\log_2\ln\left(\frac{1}{\min\{\delta,1\}}\right)\right\}$.

Since $E[\Omega] = E\left[P_{Z^n}\left(\mathcal{J}^c(\mathcal{C}_n)\right)\right] \leq 2\delta$, there must exist a codebook $\mathcal{C}_n^*$ such that $P_{Z^n}\left(\mathcal{J}^c(\mathcal{C}_n^*)\right)$ is no greater than 2δ for n sufficiently large.

Step 5: Calculation of distortion. The distortion of the optimal codebook $\mathscr{C}_n^*$ (from the previous step) satisfies for $n > N$:

$$
\begin{aligned}
\frac{1}{n}E[\rho_n(Z^n, h_n(Z^n))] &= \sum_{z^n \in \mathcal{J}(\mathscr{C}_n^*)} P_{Z^n}(z^n)\frac{1}{n}\rho_n(z^n, h_n(z^n)) \\
&\quad + \sum_{z^n \notin \mathcal{J}(\mathscr{C}_n^*)} P_{Z^n}(z^n)\frac{1}{n}\rho_n(z^n, h_n(z^n)) \\
&\leq \sum_{z^n \in \mathcal{J}(\mathscr{C}_n^*)} P_{Z^n}(z^n)(D+\delta) + \sum_{z^n \notin \mathcal{J}(\mathscr{C}_n^*)} P_{Z^n}(z^n)\rho_{max} \\
&\leq (D+\delta) + 2\delta \cdot \rho_{max} \\
&\leq D + \delta(1 + 2\rho_{max}) \\
&\leq D + \varepsilon.
\end{aligned}
$$

This concludes the proof of the achievability part.

2. *Converse Part* (i.e., $R(D+\varepsilon) \geq R^{(I)}(D)$ for arbitrarily small $\varepsilon > 0$ and any $D \in \{D \geq 0\colon R^{(I)}(D) > 0\}$): We need to show that for any sequence of $\{(n, M_n, D_n)\}_{n=1}^{\infty}$ code with

$$\limsup_{n\to\infty} \frac{1}{n}\log_2 M_n < R^{(I)}(D),$$

there exists $\varepsilon > 0$ such that

$$D_n = \frac{1}{n}E[\rho_n(Z^n, h_n(Z^n))] > D + \varepsilon$$

for n sufficiently large. The proof is as follows.

Step 1: Convexity of mutual information. By the convexity of mutual information $I(Z;\hat{Z})$ with respect to $P_{\hat{Z}|Z}$ for a fixed P_Z, we have

$$I(Z;\hat{Z}_\lambda) \leq \lambda \cdot I(Z;\hat{Z}_1) + (1-\lambda)\cdot I(Z;\hat{Z}_2),$$

where $\lambda \in [0,1]$, and

$$P_{\hat{Z}_\lambda|Z}(\hat{z}|z) := \lambda P_{\hat{Z}_1|Z}(\hat{z}|z) + (1-\lambda)P_{\hat{Z}_2|Z}(\hat{z}|z).$$

Step 2: Convexity of $R^{(I)}(D)$. Let $P_{\hat{Z}_1|Z}$ and $P_{\hat{Z}_2|Z}$ be two distributions achieving $R^{(I)}(D_1)$ and $R^{(I)}(D_2)$, respectively. Since

$$\begin{aligned} E[\rho(Z,\hat{Z}_\lambda)] &= \sum_{z\in\mathcal{Z}} P_Z(z) \sum_{\hat{z}\in\hat{\mathcal{Z}}} P_{\hat{Z}_\lambda|Z}(\hat{z}|z)\rho(z,\hat{z}) \\ &= \sum_{z\in\mathcal{Z},\hat{z}\in\hat{\mathcal{Z}}} P_Z(z)\left[\lambda P_{\hat{Z}_1|Z}(\hat{z}|z) + (1-\lambda)P_{\hat{Z}_2|Z}(\hat{z}|z)\right]\rho(z,\hat{z}) \\ &= \lambda D_1 + (1-\lambda)D_2, \end{aligned}$$

we have

$$\begin{aligned} R^{(I)}(\lambda D_1 + (1-\lambda)D_2) &\le I(Z;\hat{Z}_\lambda) \\ &\le \lambda I(Z;\hat{Z}_1) + (1-\lambda)I(Z;\hat{Z}_2) \\ &= \lambda R^{(I)}(D_1) + (1-\lambda)R^{(I)}(D_2). \end{aligned}$$

Therefore, $R^{(I)}(D)$ is a convex function.

Step 3: Strictly decreasing and continuity properties of $R^{(I)}(D)$.

By definition, $R^{(I)}(D)$ is non-increasing in D. Also,

$$
\begin{aligned}
R^{(I)}(D) = 0 \quad \text{iff} \quad D \geq D_{\max} &:= \min_{P_{\hat{Z}}} \sum_{z\in\mathcal{Z}} \sum_{\hat{z}\in\hat{\mathcal{Z}}} P_Z(z) P_{\hat{Z}}(\hat{z}) \rho(z,\hat{z}) \\
&= \min_{P_{\hat{Z}}} \sum_{\hat{z}\in\hat{\mathcal{Z}}} P_{\hat{Z}}(\hat{z}) \sum_{z\in\mathcal{Z}} P_Z(z) \rho(z,\hat{z}) \\
&= \min_{\hat{z}\in\hat{\mathcal{Z}}} \sum_{z\in\mathcal{Z}} P_Z(z) \rho(z,\hat{z}) \qquad (6.3.4)
\end{aligned}
$$

which is finite by the boundedness of the distortion measure. Thus since $R^{(I)}(D)$ is non-increasing and convex, it directly follows that it is strictly decreasing and continuous over $\{D \geq 0\colon R^{(I)}(D) > 0\}$.

Shannon's lossy source coding theorem

Step 4: Main proof.

$$
\begin{aligned}
\log_2 M_n &\geq H(h_n(Z^n)) \\
&= H(h_n(Z^n)) - H(h_n(Z^n)|Z^n), \quad \text{since } H(h_n(Z^n)|Z^n) = 0; \\
&= I(Z^n; h_n(Z^n)) \\
&= H(Z^n) - H(Z^n|h_n(Z^n)) \\
&= \sum_{i=1}^{n} H(Z_i) - \sum_{i=1}^{n} H(Z_i|h_n(Z^n), Z_1, \ldots, Z_{i-1}) \\
&\quad \text{by the independence of } Z^n, \text{ and the chain rule for conditional entropy;} \\
&\geq \sum_{i=1}^{n} H(Z_i) - \sum_{i=1}^{n} H(Z_i|\hat{Z}_i), \text{ where } \hat{Z}_i \text{ is the } i^{\text{th}} \text{ component of } h_n(Z^n); \\
&= \sum_{i=1}^{n} I(Z_i; \hat{Z}_i) \geq \sum_{i=1}^{n} R^{(I)}(D_i), \quad \text{where } D_i := E[\rho(Z_i, \hat{Z}_i)]; \\
&= n \sum_{i=1}^{n} \frac{1}{n} R^{(I)}(D_i) \geq n R^{(I)}\left(\sum_{i=1}^{n} \frac{1}{n} D_i\right), \quad \text{by convexity of } R^{(I)}(D); \\
&= n R^{(I)}\left(\frac{1}{n} E[\rho_n(Z^n, h_n(Z^n))]\right),
\end{aligned}
$$

where the last step follows since the distortion measure is additive.

Shannon's lossy source coding theorem

Finally,

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 M_n < R^{(I)}(D)$$

implies the existence of N and $\gamma > 0$ such that

$$\frac{1}{n} \log_2 M_n < R^{(I)}(D) - \gamma$$

for all $n > N$. Therefore, for $n > N$,

$$R^{(I)}\left(\frac{1}{n}E[\rho_n(Z^n, h_n(Z^n))]\right) \quad \left(\leq \frac{1}{n}\log_2 M_n\right) \quad < R^{(I)}(D) - \gamma,$$

which, together with the fact that $R^{(I)}(D)$ is strictly decreasing, implies that

$$\frac{1}{n}E[\rho_n(Z^n, h_n(Z^n))] > D + \varepsilon$$

for some $\varepsilon = \varepsilon(\gamma) > 0$ and for all $n > N$.

Hence, $(R^{(I)}(D), D + \varepsilon)$ is not achievable and the operational $R(D)$ satisfies

$$R(D + \varepsilon) > R^{(I)}(D) \quad \text{for arbitrarily small } \varepsilon > 0.$$

Shannon's lossy source coding theorem

3. *Summary*:

- For $D \in \{D \geq 0 \colon R^{(I)}(D) > 0\}$, the achievability and converse parts jointly imply that

$$R^{(I)}(D) + 4\varepsilon \geq R(D + \varepsilon) \geq R^{(I)}(D)$$

  for arbitrarily small $\varepsilon > 0$.

- These inequalities together with the continuity of $R^{(I)}(D)$ yield that

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

  for $D \in \{D \geq 0 \colon R^{(I)}(D) > 0\}$.

- For $D \in \{D \geq 0 \colon R^{(I)}(D) = 0\}$, the achievability part gives us

$$R^{(I)}(D) + 4\varepsilon = 4\varepsilon \geq R(D + \varepsilon) \geq 0$$

  for arbitrarily small $\varepsilon > 0$. This immediately implies that

$$R(D) = 0 \quad (= R^{(I)}(D)).$$

$\square$

Notes

- The formula of the rate-distortion function obtained in the previous theorems is also valid for the squared error distortion over the real numbers, even if it is unbounded.
  - For example, the boundedness assumption in the theorems can be replaced with assuming that there exists a reproduction symbol $\hat{z}_0 \in \hat{\mathcal{Z}}$ such that $E[\rho(Z, \hat{z}_0)] < \infty$.
  - This assumption can accommodate the squared error distortion measure and a source with finite second moment (including continuous-alphabet sources such as Gaussian sources).
- Here, we put the boundedness assumption just to facilitate the exposition of the current proof.

Notes

- After introducing
  - Shannon's source coding theorem for block codes
  - Shannon's channel coding theorem for block codes
  - Rate-distortion theorem

  in the memoryless (and stationary ergodic) system setting, we briefly elucidate the "key concepts or techniques" behind these lengthy proofs, in particular:

  - The notion of a typical set
    * The typical set construct – specifically,
      · δ-typical set for source coding
      · joint δ-typical set for channel coding
      · distortion typical set for rate-distortion

      uses a law of large numbers or AEP argument to claim the existence of a set with very high probability; hence, the respective information manipulation can just focus on the set with negligible performance loss.

Notes

- The notion of random coding
  * The random coding technique shows that the expectation of the desired performance over all possible information manipulation schemes (randomly drawn according to some properly chosen statistics) is already acceptably good, and hence the existence of at least one good scheme that fulfills the desired performance index is validated.

- As a result, in situations where the above two techniques apply, a similar theorem can often be established.

Theorem 6.21 (Shannon's rate-distortion theorem for stationary ergodic sources) Consider a stationary ergodic source $\{Z_n\}_{n=1}^{\infty}$ with alphabet $\mathcal{Z}$, reproduction alphabet $\hat{\mathcal{Z}}$ and a bounded additive distortion measure $\rho_n(\cdot,\cdot)$; i.e.,

$$\rho_n(z^n, \hat{z}^n) = \sum_{i=1}^{n} \rho(z_i, \hat{z}_i) \quad \text{and} \quad \rho_{\max} := \max_{(z,\hat{z})\in\mathcal{Z}\times\hat{\mathcal{Z}}} \rho(z,\hat{z}) < \infty,$$

where $\rho(\cdot,\cdot)$ is a given single-letter distortion measure. Then the source's rate-distortion function is given by

$$R(D) = \bar{R}^{(I)}(D),$$

where

$$\bar{R}^{(I)}(D) := \lim_{n\to\infty} R_n^{(I)}(D) \tag{6.3.5}$$

is called the *asymptotic information rate-distortion function.* and

$$R_n^{(I)}(D) := \min_{P_{\hat{Z}^n|Z^n}:\ \frac{1}{n}E[\rho_n(Z^n,\hat{Z}^n)]\le D} \frac{1}{n} I(Z^n;\hat{Z}^n) \tag{6.3.6}$$

is the *n-th order information rate-distortion function.*

Notes

- **Question**: Can we extend the theorems to cases where the two arguments fail?'
- It is obvious that only when new methods (other than the above two) are developed can the question be answered in the affirmative.

6.4 Calculation of the rate-distortion function

Theorem 6.23 Fix a binary DMS $\{Z_n\}_{n=1}^{\infty}$ with marginal distribution $P_Z(0) = 1 - P_Z(1) = p$, where $0 < p < 1$. Then the source's rate-distortion function under the Hamming additive distortion measure is given by:

$$R(D) = \begin{cases} h_{\text{b}}(p) - h_{\text{b}}(D) & \text{if } 0 \le D < \min\{p, 1-p\}; \\ 0 & \text{if } D \ge \min\{p, 1-p\}, \end{cases}$$

where $h_{\text{b}}(p) := -p \cdot \log(p) - (1-p) \cdot \log(1-p)$ is the binary entropy function.

Proof: Assume without loss of generality that $p \le 1/2$.

- We first prove the theorem under $0 \le D < \min\{p, 1-p\} = p$. Observe that for any binary random variable $\hat{Z}$,

$$H(Z|\hat{Z}) = H(Z \oplus \hat{Z}|\hat{Z}).$$

  Also observe that

$$E[\rho(Z, \hat{Z})] \le D \text{ implies } \Pr\{Z \oplus \hat{Z} = 1\} \le D.$$

6.4 Calculation of the rate-distortion function

Then

$$\begin{aligned} I(Z;\hat{Z}) &= H(Z) - H(Z|\hat{Z}) \\ &= h_{\rm b}(p) - H(Z \oplus \hat{Z}|\hat{Z}) \\ &\geq h_{\rm b}(p) - H(Z \oplus \hat{Z}) \quad \text{(conditioning never increase entropy)} \\ &\geq h_{\rm b}(p) - h_{\rm b}(D), \end{aligned}$$

where the last inequality follows since $h_{\rm b}(x)$ is increasing for $x \leq 1/2$, and $\Pr\{Z \oplus \hat{Z} = 1\} \leq D$.

- Since the above derivation is true for any $P_{\hat{Z}|Z}$, we have

$$R(D) \geq h_{\rm b}(p) - h_{\rm b}(D).$$

6.4 Calculation of the rate-distortion function

- It remains to show that the lower bound is achievable by some $P_{\hat{Z}|Z}$, or equivalently, $H(Z|\hat{Z}) = h_{\rm b}(D)$ for some $P_{\hat{Z}|Z}$.

  By defining $P_{Z|\hat{Z}}(0|0) = P_{Z|\hat{Z}}(1|1) = 1-D$, we immediately obtain $H(Z|\hat{Z}) = h_{\rm b}(D)$. The desired $P_{\hat{Z}|Z}$ can be obtained by solving

$$
\begin{aligned}
1 &= P_{\hat{Z}}(0) + P_{\hat{Z}}(1) \\
&= \frac{P_Z(0)}{P_{Z|\hat{Z}}(0|0)} P_{\hat{Z}|Z}(0|0) + \frac{P_Z(0)}{P_{Z|\hat{Z}}(0|1)} P_{\hat{Z}|Z}(1|0) \\
&= \frac{p}{1-D} P_{\hat{Z}|Z}(0|0) + \frac{p}{D}(1 - P_{\hat{Z}|Z}(0|0))
\end{aligned}
$$

  and

$$
\begin{aligned}
1 &= P_{\hat{Z}}(0) + P_{\hat{Z}}(1) \\
&= \frac{P_Z(1)}{P_{Z|\hat{Z}}(1|0)} P_{\hat{Z}|Z}(0|1) + \frac{P_Z(1)}{P_{Z|\hat{Z}}(1|1)} P_{\hat{Z}|Z}(1|1) \\
&= \frac{1-p}{D}(1 - P_{\hat{Z}|Z}(1|1)) + \frac{1-p}{1-D} P_{\hat{Z}|Z}(1|1),
\end{aligned}
$$

  and yield

$$
P_{\hat{Z}|Z}(0|0) = \frac{1-D}{1-2D}\left(1 - \frac{D}{p}\right) \quad \text{and} \quad P_{\hat{Z}|Z}(1|1) = \frac{1-D}{1-2D}\left(1 - \frac{D}{1-p}\right).
$$

6.4 Calculation of the rate-distortion function

- Now in the case of $p \le D < 1-p$, we can let $P_{\hat{Z}|Z}(1|0) = P_{\hat{Z}|Z}(1|1) = 1$ to obtain $I(Z;\hat{Z}) = 0$ and

$$E[\rho(Z,\hat{Z})] = \sum_{z=0}^{1}\sum_{\hat{z}=0}^{1} P_Z(z)P_{\hat{Z}|Z}(\hat{z}|z)\rho(z,\hat{z}) = p \le D.$$

Similarly, in the case of $D \ge 1-p$, we let $P_{\hat{Z}|Z}(0|0) = P_{\hat{Z}|Z}(0|1) = 1$ to obtain $I(Z;\hat{Z}) = 0$ and

$$E[\rho(Z,\hat{Z})] = \sum_{z=0}^{1}\sum_{\hat{z}=0}^{1} P_Z(z)P_{\hat{Z}|Z}(\hat{z}|z)\rho(z,\hat{z}) = 1-p \le D.$$

$\square$

- **Remark:** The Hamming additive distortion measure is defined as:

$$\rho_n(z^n,\hat{z}^n) = \sum_{i=1}^{n} z_i \oplus \hat{z}_i,$$

where "$\oplus$" denotes modulo two addition. In such case, $\rho(z^n,\hat{z}^n)$ is exactly the number of bit changes or bit errors after compression.

Theorem 6.26 (Gaussian sources maximize the rate-distortion function) Under the additive squared error distortion measure, namely

$$\rho_n(z^n, \hat{z}^n) = \sum_{i=1}^{n} (z_i - \hat{z}_i)^2,$$

the rate-distortion function for any continuous memoryless source $\{Z_i\}$ with a pdf of support $\mathbb{R}$, zero mean, variance σ^2 and finite differential entropy satisfies

$$R(D) \le \begin{cases} \dfrac{1}{2}\log_2 \dfrac{\sigma^2}{D}, & \text{for } 0 < D \le \sigma^2 \\ 0, & \text{for } D > \sigma^2 \end{cases}$$

with equality holding when the source is Gaussian.

Proof: By Theorem 6.20 (extended to the "unbounded" squared error distortion measure),

$$R(D) = R^{(I)}(D) = \min_{f_{\hat{Z}|Z}:\ E[(Z-\hat{Z})^2] \le D} I(Z; \hat{Z}).$$

So for any $f_{\hat{Z}|Z}$ satisfying the distortion constraint,

$$R(D) \le I(f_Z, f_{\hat{Z}|Z}).$$

6.4.2 Rate distortion func / the squared error dist

For $0 < D \le \sigma^2$:

- Choose a dummy Gaussian random variable W with zero mean and variance aD, where $a = 1 - D/\sigma^2$, and is independent of Z. Let $\hat{Z} = aZ + W$. Then

$$E[(Z - \hat{Z})^2] = E[(1-a)^2 Z^2] + E[W^2] = (1-a)^2\sigma^2 + aD = D$$

which satisfies the distortion constraint.

- Note that the variance of $\hat{Z}$ is equal to $E[a^2Z^2] + E[W^2] = \sigma^2 - D$.

- Consequently,

$$\begin{aligned}
R(D) &\le I(Z;\hat{Z}) \\
&= h(\hat{Z}) - h(\hat{Z}|Z) \\
&= h(\hat{Z}) - h(W + aZ|Z) \\
&= h(\hat{Z}) - h(W|Z) \\
&= h(\hat{Z}) - h(W) \qquad \text{(by the independence of } W \text{ and } Z\text{)} \\
&= h(\hat{Z}) - \frac{1}{2}\log_2(2\pi e(aD)) \\
&\le \frac{1}{2}\log_2(2\pi e(\sigma^2 - D)) - \frac{1}{2}\log_2(2\pi e(aD)) = \frac{1}{2}\log_2\frac{\sigma^2}{D}.
\end{aligned}$$

6.4.2 Rate distortion func / the squared error dist

For $D > \sigma^2$:

- Let $\hat{Z}$ satisfy $\Pr\{\hat{Z} = 0\} = 1$ (and be independent of Z).
- Then $E[(Z-\hat{Z})^2] = E[Z^2]+E[\hat{Z}^2]-2E[Z]E[\hat{Z}] = \sigma^2 < D$, and $I(Z;\hat{Z}) = 0$. Hence, $R(D) = 0$ for $D > \sigma^2$.

The achievability of this upper bound by a Gaussian source (with zero mean and variance σ^2) can be proved by showing that under the Gaussian source,

$$(1/2)\log_2(\sigma^2/D)$$

is a lower bound to $R(D)$ for $0 < D \le \sigma^2$.

6.4.2 Rate distortion func / the squared error dist

Indeed, when the source Z is Gaussian and for any $f_{\hat{Z}|Z}$ such that $E[(Z-\hat{Z})^2] \leq D$, we have

$$
\begin{aligned}
I(Z;\hat{Z}) &= h(Z) - h(Z|\hat{Z}) \\
&= \frac{1}{2}\log_2(2\pi e\sigma^2) - h(Z-\hat{Z}|\hat{Z}) \\
&\geq \frac{1}{2}\log_2(2\pi e\sigma^2) - h(Z-\hat{Z}) \\
&\geq \frac{1}{2}\log_2(2\pi e\sigma^2) - \frac{1}{2}\log_2\left(2\pi e\,\mathrm{Var}[(Z-\hat{Z})]\right) \\
&\geq \frac{1}{2}\log_2(2\pi e\sigma^2) - \frac{1}{2}\log_2\left(2\pi e\,E[(Z-\hat{Z})^2]\right) \\
&\geq \frac{1}{2}\log_2(2\pi e\sigma^2) - \frac{1}{2}\log_2(2\pi eD) \\
&= \frac{1}{2}\log_2\frac{\sigma^2}{D}.
\end{aligned}
$$

□

6.4.2 Rate distortion func / the squared error dist

Theorem 6.27 (Shannon lower bound on the rate-distortion function: squared error distortion) Consider a continuous memoryless source $\{Z_i\}$ with a pdf of support $\mathbb{R}$ and finite differential entropy under the additive squared error distortion measure. Then its rate-distortion function satisfies

$$R(D) \geq h(Z) - \frac{1}{2}\log_2(2\pi e D).$$

Proof: The proof, which follows similar steps as in the achievability of the upper bound in the proof of the previous theorem, is left as an exercise. □

6.4.2 Rate distortion func / the squared error dist

- In Lemma 5.42, we show that for a discrete-time continuous-alphabet memoryless additive-noise channel with input power constraint P and noise variance σ^2, its capacity satisfies

$$C_G(P) + \underbrace{D(Z\|Z_G)}_{\substack{\text{non-Gaussianness}\\ =h(Z_G)-h(Z)}} \geq C(P) \geq \underbrace{C_G(P)}_{\frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right)}.$$

- Similarly, for a continuous memoryless source $\{Zi\}$ with a pdf of support $\mathbb{R}$ and finite differential entropy under the additive squared error distortion measure its rate-distortion function satisfies

$$\underbrace{R_G(D) - D(Z\|Z_G)}_{\substack{\text{Shannon lower bound}\\ \text{on the rate distortion func}}} \leq R(D) \leq \underbrace{R_G(D)}_{\frac{1}{2}\log_2\frac{\sigma^2}{D}}.$$

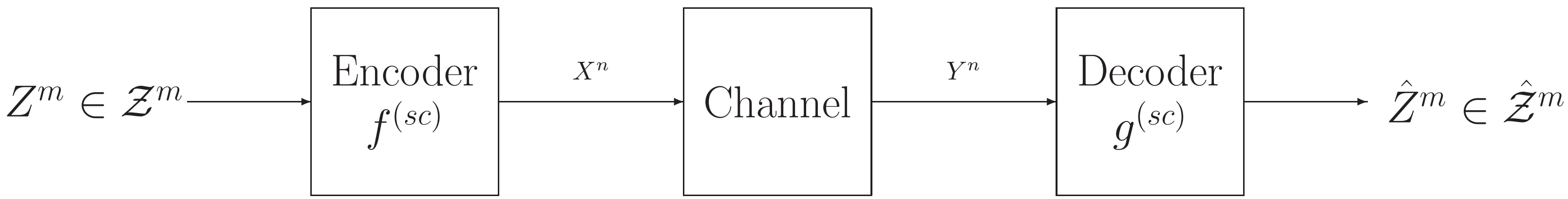
Section 6.4.3 is based on a similar idea but targets for the absolute error distortion; hence, we omit it in our lecture. Notably, a correction has been provided for Theorem 6.29 (See errata for the textbook.)

6.5 Lossy joint source-channel coding theorem

- This is also named *lossy information-transmission theorem*.

Definition 6.32 (Lossy source-channel block code) Given a discrete-time source $\{Z_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{Z}$ and reproduction alphabet $\hat{\mathcal{Z}}$ and a discrete-time channel with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, an m-to-n lossy source-channel block code with rate $\frac{m}{n}$ source symbol/channel symbol is a pair of mappings $(f^{(sc)}, g^{(sc)})$, where

$$f^{(sc)}\colon \mathcal{Z}^m \to \mathcal{X}^n \qquad \text{and} \qquad g^{(sc)}\colon \mathcal{Y}^n \to \hat{\mathcal{Z}}^m.$$



Given an additive distortion measure $\rho_m = \sum_{i=1}^{m} \rho(z_i, \hat{z}_i)$, where ρ is a distortion function on $\mathcal{Z} \times \hat{\mathcal{Z}}$, we say that the m-to-n lossy source-channel block code $(f^{(sc)}, g^{(sc)})$ *satisfies the average distortion fidelity criterion* D, where $D \geq 0$, if

$$\frac{1}{m} E[\rho_m(Z^m, \hat{Z}^m)] \leq D.$$

6.5 Lossy joint source-channel coding theorem

Theorem 6.33 (Lossy joint source-channel coding theorem) Consider a discrete-time stationary ergodic source $\{Z_i\}_{i=1}^{\infty}$ with finite alphabet $\mathcal{Z}$, finite reproduction alphabet $\hat{\mathcal{Z}}$, bounded additive distortion measure $\rho_m(\cdot,\cdot)$ and rate-distortion function $R(D)$, and consider a discrete-time memoryless channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and capacity C. Assuming that both $R(D)$ and C are measured in the same units, the following hold:

- *Forward part (achievability):* For any $D > 0$, there exists a sequence of m-to-n_m lossy source-channel codes $(f^{(sc)}, g^{(sc)})$ satisfying the average distortion fidelity criterion D for sufficiently large m if

$$\left(\limsup_{m\to\infty} \frac{m}{n_m}\right) \cdot R(D) < C.$$

- *Converse part:* On the other hand, for any sequence of m-to-n_m lossy source-channel codes $(f^{(sc)}, g^{(sc)})$ satisfying the average distortion fidelity criterion D, we have

$$\left(\frac{m}{n_m}\right) \cdot R(D) \leq C.$$

6.5 Lossy joint source-channel coding theorem

Observation 6.34 (Lossy joint source-channel coding theorem with signaling rates)

- The above theorem admits another form when the source and channel are described in terms of "signaling rates".
- Let T_s and T_c represent the durations (in seconds) per source letter and per channel input symbol, respectively.
- In this case, $\frac{T_c}{T_s}$ represents the source-channel transmission rate measured in source symbols per channel use (or input symbol).
  - *Forward part:* The source can be reproduced at the output of the channel with distortion less than D (i.e., there exist lossy source-channel codes asymptotically satisfying the average distortion fidelity criterion D) if

$$\left(\frac{T_c}{T_s}\right) \cdot R(D) < C.$$

  - *Converse part:* For any lossy source-channel codes satisfying the average distortion fidelity criterion D, we have

$$\left(\frac{T_c}{T_s}\right) \cdot R(D) \le C.$$

6.6 Shannon limit of communication systems

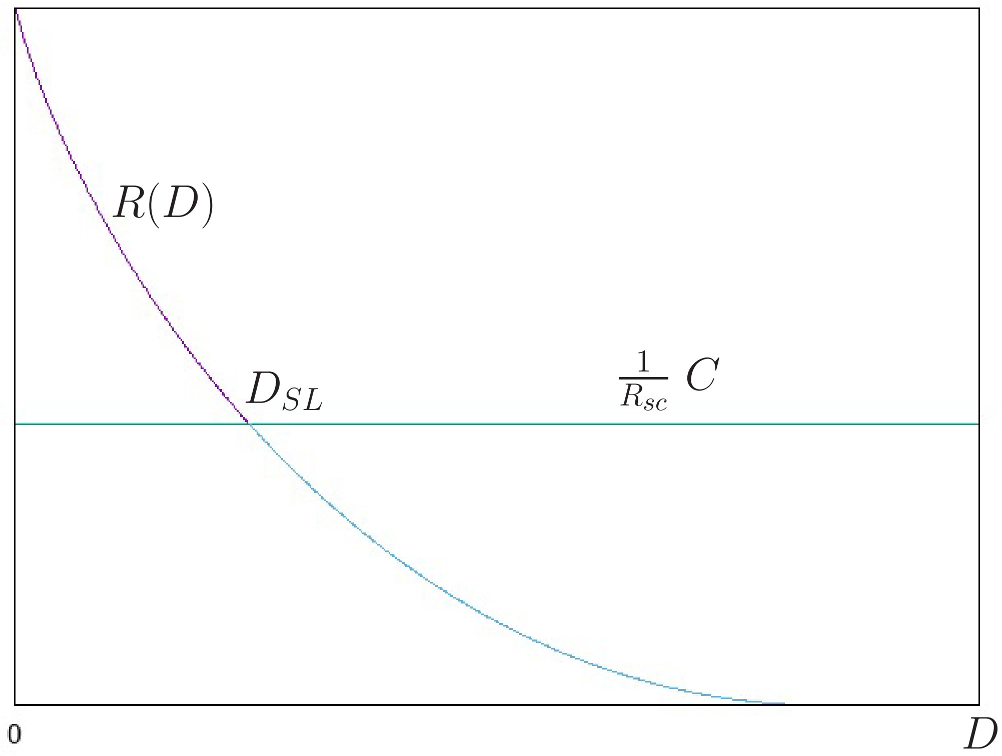
- A bound on the end-to-end distortion of a communication system:
  - If a source with rate-distortion function $R(D)$ can be transmitted over a channel with capacity C via a source-channel block code of rate $R_{sc} > 0$ (in source symbols/channel use) and reproduced at the destination with an average distortion no larger than D, then we must have that

$$R(D) \leq \frac{1}{R_{sc}} C. \tag{6.6.1}$$

  - Shannon limit:

$$D_{SL} := \min \left\{ D : R(D) \leq \frac{1}{R_{sc}} C \right\}.$$

6.6 Shannon limit of communication systems



Example 6.35 (Shannon limit for a binary uniform DMS over a BSC)

- Let $\mathcal{Z} = \hat{\mathcal{Z}} = \{0, 1\}$ and consider a binary uniformly distributed DMS $\{Z_i\}$ (i.e., a Bernoulli(1/2) source) using the additive Hamming distortion measure.
- Note that in this case, $E[\rho(Z, \hat{Z})] = P(Z \neq \hat{Z}) := P_b$.
- We desire to transmit the source over a BSC with crossover probability $\epsilon < 1/2$.
- We then have for $0 \leq D \leq \frac{1}{2}$,

$$R(D) = 1 - h_{\rm b}(D), \quad \text{and} \quad C = 1 - h_{\rm b}(\epsilon).$$

- Hence, for a given ϵ,

$$D_{SL} := \min\left\{D : 1 - h_{\rm b}(D) \leq \frac{1}{R_{sc}}(1 - h_{\rm b}(\epsilon))\right\} = h_{\rm b}^{-1}\left(1 - \frac{1 - h_{\rm b}(\epsilon)}{R_{sc}}\right).$$

- Alternatively, for a given D,

$$\epsilon_{SL} := \max\left\{\epsilon : 1 - h_{\rm b}(D) \leq \frac{1}{R_{sc}}(1 - h_{\rm b}(\epsilon))\right\} = h_{\rm b}^{-1}\Big(1 - R_{sc}\big(1 - h_{\rm b}(D)\big)\Big)$$

6.6 Shannon limit of communication systems

- It is well-known that a BSC with crossover probability ϵ represents a binary-input AWGN channel used with antipodal (BPSK) signaling and hard-decision coherent demodulation.
- With average energy per signal P, noise power $\frac{N_0}{2}$ and signal-to-noise ratio (SNR) $\gamma = P/N_0$, we have

$$\epsilon = Q\left(\sqrt{2\gamma}\right) \tag{6.6.5}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}}\, dt$$

is the Gaussian Q-function.

- If the channel is used with a source-channel code of rate R_{sc} source (or information) bits/channel use, then ϵ can be expressed in terms of a so-called *SNR per source (or information) bit*

$$\gamma_b := \frac{E_b}{N_0} = \frac{1}{R_{sc}}\frac{P}{N_0} = \frac{1}{R_{sc}}\gamma,$$

where E_b is the *average energy per source bit.*

6.6 Shannon limit of communication systems

- Thus,

$$\epsilon = Q\left(\sqrt{2R_{sc}\gamma_b}\right) \tag{6.6.6}$$

- The minimal γ_b (in dB) for a given $P_b = D < \frac{1}{2}$ and a source-channel code rate $R_{sc} < 1$:

$$\gamma_{b,SL} = \frac{1}{2R_{sc}}\big(Q^{-1}(\epsilon_{SL})\big)^2$$

Rate R_{sc}	$P_b = 0$	$P_b = 10^{-5}$	$P_b = 10^{-4}$	$P_b = 10^{-3}$	$P_b = 10^{-3}$
1/3	1.212	1.210	1.202	1.150	0.077
1/2	1.775	1.772	1.763	1.703	1.258
2/3	2.516	2.513	2.503	2.423	1.882
4/5	3.369	3.367	3.354	3.250	2.547

- For $R_{sc} = 1$,

$$\epsilon_{SL} := h_{\rm b}^{-1}\big(1 - R_{sc}\big(1 - h_{\rm b}(D)\big)\big) = D = P_b$$

and

$$\gamma_{b,SL} = \frac{1}{2}\big(Q^{-1}(P_b)\big)^2.$$

6.6 Shannon limit of communication systems

Example 6.37 (Shannon limit for a memoryless Gaussian source over an AWGN channel)

- Let $\mathcal{Z} = \hat{\mathcal{Z}} = \mathbb{R}$ and consider a memoryless Gaussian source $\{Z_i\}$ of mean zero and variance σ^2 and the squared error distortion function.
- The objective is to transmit the source over an AWGN channel with input power constraint P and noise variance $\sigma_N^2 = \frac{N_0}{2}$ and recover it with distortion fidelity no larger than D, for a given threshold $D > 0$.
- The source's rate-distortion function is given by

$$R(D) = \frac{1}{2}\log_2\frac{\sigma^2}{D} \qquad \text{for} \qquad 0 < D < \sigma^2.$$

Furthermore, the capacity (or capacity-cost function) of the AWGN channel is given as

$$C(P) = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma_N^2}\right).$$

6.6 Shannon limit of communication systems

- The Shannon limit D_{SL} for this system with rate R_{sc} is obtained via

$$\begin{aligned} D_{SL} &:= \min\left\{D : R(D) \le \frac{1}{R_{sc}} C(P)\right\} \\ &= \min\left\{D : \frac{1}{2}\log_2\frac{\sigma^2}{D} \le \frac{1}{2R_{sc}}\log_2\left(1 + \frac{P}{\sigma_N^2}\right)\right\} \\ &= \frac{\sigma^2}{\left(1+\frac{P}{\sigma_N^2}\right)^{1/R_{sc}}} \end{aligned} \tag{6.6.10}$$

for $0 < D_{SL} < \sigma^2$.

6.6 Shannon limit of communication systems

Example 6.39 (Shannon limit for a binary uniform DMS over a binary-input AWGN channel)

- Let $\mathcal{Z} = \hat{\mathcal{Z}} = \{0, 1\}$ and consider a binary uniformly distributed DMS $\{Z_i\}$ (i.e., a Bernoulli(1/2) source) using the additive Hamming distortion measure.
- The binary uniform source is sent via a source-channel code over a binary-input AWGN channel used with antipodal (BPSK) signaling of power P and noise variance $\sigma_N^2 = N_0/2$.
- We then have for $0 \le D \le \frac{1}{2}$,

$$R(D) = 1 - h_{\mathrm{b}}(D).$$

6.6 Shannon limit of communication systems

- However, the channel capacity $C(P)$ of the AWGN whose input takes on two possible values $+\sqrt{P}$ or $-\sqrt{P}$, whose output is real-valued and whose noise variance is $\sigma_N^2 = \frac{N_0}{2}$, is given by evaluating the mutual information between the channel input and output under the input distribution $P_X(+\sqrt{P}) = P_X(-\sqrt{P}) = 1/2$:

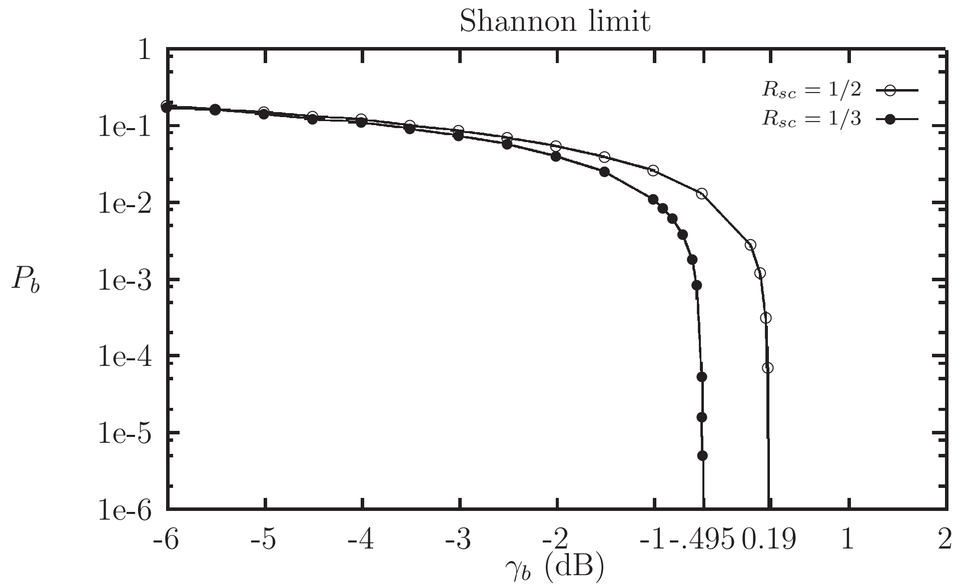
$$\begin{aligned} C(P) &= \frac{P}{\sigma_N^2}\log_2(e) - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\log_2\left[\cosh\left(\frac{P}{\sigma_N^2} + y\sqrt{\frac{P}{\sigma_N^2}}\right)\right]dy \\ &= \frac{R_{sc}E_b}{N_0/2}\log_2(e) - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\log_2\left[\cosh\left(\frac{R_{sc}E_b}{N_0/2} + y\sqrt{\frac{R_{sc}E_b}{N_0/2}}\right)\right]dy \\ &= 2R_{sc}\gamma_b\log_2(e) - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\log_2[\cosh(2R_{sc}\gamma_b + y\sqrt{2R_{sc}\gamma_b})]dy, \end{aligned}$$

  where $P = R_{sc}E_b$ is the channel signal power, E_b is the average energy per source bit, and $\gamma_b = E_b/N_0$ is the SNR per source bit.

- Then, it requires

$$1 - h_b(P_b) \le \frac{1}{R_{sc}}\left[2R_{sc}\gamma_b\log_2(e) - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\log_2[\cosh(2R_{sc}\gamma_b + y\sqrt{2R_{sc}\gamma_b})]dy\right].$$

6.6 Shannon limit of communication systems



The Shannon limits for $(2, 1)$ and $(3, 1)$ codes under binary-input AWGN channel.

- The Shannon limits calculated above are pertinent due to the invention of near-capacity achieving channel codes, such as Turbo or LDPC codes.
- For example, the rate-$1/2$ Turbo coding system proposed in 1993 can approach a bit error rate of 10^{-5} at $\gamma_b = 0.9$ dB, which is only 0.714 dB away from the Shannon limit of 0.186 dB.

6.6 Shannon limit of communication systems

Rate R_{sc}	$P_b = 0$	$P_b = 10^{-5}$	$P_b = 10^{-4}$	$P_b = 10^{-3}$	$P_b = 10^{-3}$
$1/3$	-0.496	-0.496	-0.504	-0.559	-0.960
$1/2$	0.186	0.186	0.177	0.111	-0.357
$2/3$	1.060	1.057	1.047	0.963	0.382
$4/5$	2.040	2.038	2.023	1.909	1.152

Example 6.40 (Shannon limit for a binary uniform DMS over a binary-input Rayleigh fading channel)

- Consider a BPSK modulated Rayleigh fading channel.
- Its input power is $P = R_{sc}E_b$, its noise variance is $\sigma_N^2 = N_0/2$ and the fading distribution is Rayleigh:

$$f_A(a) = 2ae^{-a^2}, \qquad a > 0.$$

- Then,

$$C_{DSI}(\gamma_b) = 1 - \sqrt{\frac{R_{sc}\gamma_b}{\pi}} \int_0^{+\infty} \int_{-\infty}^{+\infty} f_A(a)\, e^{-R_{sc}\gamma_b(y+a)^2} \log_2\left(1 + e^{4R_{sc}\gamma_b ya}\right)\, dy\, da.$$

- We then generate the below table according to:

$$1 - h_b(P_b) \le \frac{1}{R_{sc}} C_{DSI}(\gamma_b),$$

Rate R_{sc}	$P_b = 0$	$P_b = 10^{-5}$	$P_b = 10^{-4}$	$P_b = 10^{-3}$	$P_b = 10^{-3}$
1/3	0.489	0.487	0.479	0.412	−0.066
1/2	1.830	1.829	1.817	1.729	1.107
2/3	3.667	3.664	3.647	3.516	2.627
4/5	5.936	5.932	5.904	5.690	4.331

Key Notes

- Why lossy data compression (e.g., to transmit a source with entropy larger than capacity)
- Distortion measure
- Lossy data compression codes
- Rate-distortion function
- Distortion typical set
- AEP for distortion measure
- Rate distortion theorem

Key Notes

Terminology

- Shannon's source coding theorem → Shannon's first coding theorem;
- Shannon's channel coding theorem → Shannon's second coding theorem;
- Rate distortion theorem → Shannon's third coding theorem.
- Information transmission Theorem → Joint source-channel coding theorem
  - Shannon limit (BER versus $\mathrm{SNR_b}$)