# Chapter 5

# Lossy Data Compression

**Po-Ning Chen, Professor**

**Department of Communications Engineering**

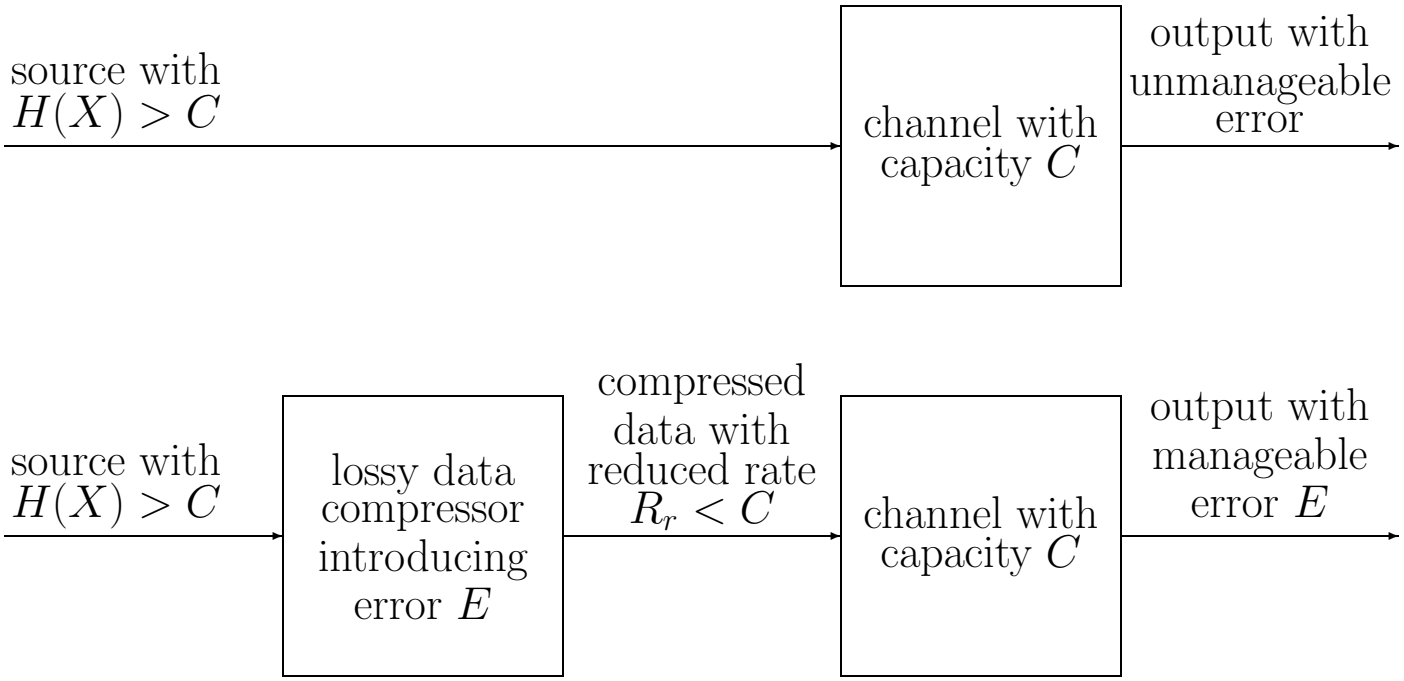
**National Chiao Tung University**

**Hsin Chu, Taiwan 300, R.O.C.**

# Motivations



- Lossy data compression = to compress a source in a rate less than entropy.
  **E.g.** Digitization or quantization of continuous signals, such as voices and multi-dimensional images.




Simple Diagram for applications of lossy data compression codes.

# Motivations



**Another example.** (Extracting useful information) Likelihood ratio test.

- Any two distinct source letters which produce the same likelihood ratio should not be encoded into distinct codewords.
- This is a lossy data compression since the sourcewords may not be reconstructed without distortion.

# Distortion Measures

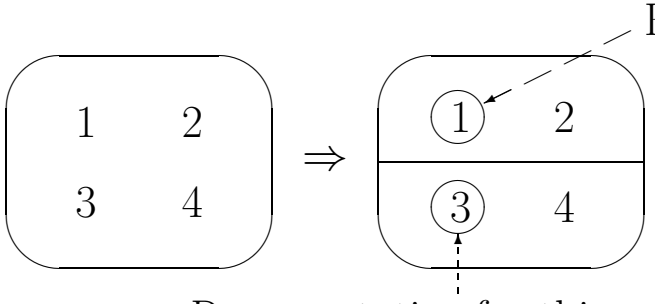


**Definition 5.4 (distortion measure)** A distortion measure is a mapping

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

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 can be viewed as a cost of representing the source symbol $z$ by another source symbol.
  **E.g.** A lossy data compression is similar to "grouping."




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

# 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

$$H(Z) = \log(4) \text{ nats} \quad \Rightarrow \quad H(\hat{Z}) = \log(2) \text{ nats}.$$

# 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

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

  – Suppose only two outcomes are allowed under uniform $Z$.
  Then

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

  is one of the best choice.

  – 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

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

# Important Notes



- It needs to be pointed out that to have

$$|\hat{\mathcal{Z}}| > |\mathcal{Z}| + 1$$

  is usually not advantageous.

- Indeed, it has been proved that under some reasonable assumptions on the distortion measure, to have larger reproduction alphabet that $|\mathcal{Z}| + 1$ will not perform better.

# Distortion Measure for a Single Letter



**Example 5.5 (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

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

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

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

**Example 5.6 (squared error distortion)** Let source alphabet and reproduction alphabet be the same, i.e., $\mathcal{Z} = \hat{\mathcal{Z}}$. The squared error distortion

$$\rho(z, \hat{z}) \stackrel{\triangle}{=} (z - \hat{z})^2,$$

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

# Comments on Squared Error Distortion



- The squared error distortion has the advantages of simplicity and having close form solution for most cases of interest, such as using least squares prediction.
- Yet, such distortion measure has been criticized as an **unhumanized** criterion.
- For example, two speech waveforms in which one is a slightly time-shifted version of the other may have large square error distortion; however, they sound very similar to human.

# Distortion Measure for Sequences



**Definition 5.7 (additive distortion measure)** The additive distortion measure $\rho_n$ between sequence $z^n$ and $\hat{z}^n$ is defined by

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

**Definition 5.8 (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 sequences $z^n$ by sequence $\hat{z}^n$ of the same length is a must or not?
- In other words, can we use $\hat{z}^k$ to represent $z^n$ for $k \neq n$?

**Answer to the second question:** Of course, yes. But it may not be easy to define a distortion measure from $z^n$ to $\hat{z}^k$ based on per-letter distortion, which may cause problems in proving the lossy data compression theorem.

# Distortion Measure for Sequences

**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.

• 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.

# Fixed-Length Lossy Data Compression

**Definition 5.9 (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

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

- Code rate for lossy data compression

$$\frac{1}{n}\log_2 M \text{ bits/sourceword} \quad \text{or} \quad \frac{1}{n}\log M \text{ nats/sourceword}$$

- Ultimate code rate for lossy data compression

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

# Variable-Length Lossy Data Compression



- Note that a parallel definition for variable-length source compression code can certainly be defined.
- Yet, there is no conclusive results for the bound on such code rate up to now, and hence we omit it for the moment.
- This is also an interesting open problem to research on.

# Achievable Rate-Distortion Pair



**Definition 5.10 (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

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

**Definition 5.11 (rate-distortion function)** The rate distortion function, denoted by $R(D)$, is

$$R(D) \stackrel{\triangle}{=} \inf\{\hat{R} \in \Re \ : \ (\hat{R}, D) \text{ is an achievable rate-distortion pair}\}.$$

# Distortion Typical Set



**Definition 5.12 (distortion typical set)** For a memoryless distribution with generic marginal $P_{Z,\hat{Z}}$ and a bounded additive distortion measure $\rho_n(\cdot,\cdot)$, the *distortion $\delta$-typical set* is defined by

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

# AEP for Distortion Typical Set



**Theorem 5.13** If $(Z_1, \hat{Z}_1)$, $(Z_2, \hat{Z}_2)$, ..., $(Z_n, \hat{Z}_n)$, ... are i.i.d., and $\rho_n$ are bounded additive distortion measure, then

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

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

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

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 boundedness assumption, the normalized sum of an i.i.d. sequence is not necessary convergence in probability to its mean.
- That is why an additional condition, "boundedness" on distortion measure, is imposed, which guarantees the required convergence.

# AEP for Distortion Typical Set



**Theorem 5.14 (AEP for distortion measure)** Given a discrete memoryless sources $\boldsymbol{Z}$, a single-letter conditional distribution $P_{\hat{Z}|Z}$, and any $\delta > 0$, the weakly distortion $\delta$-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 $\mathcal{D}_n(\delta)$,

$$P_{\hat{Z}^n}(\hat{z}^n) \geq P_{\hat{Z}^n|Z^n}(\hat{z}^n|z^n)e^{-n[I(Z;\hat{Z})+3\delta]}. \tag{5.3.1}$$

**Proof:** The first one follows from the definition. The second one can be proved by

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

□

# AEP for Distortion Typical Set



**Lemma 5.15** For $0 \le x \le 1$, $0 \le y \le 1$, and $n > 0$,

$$(1-xy)^n \le 1 - x + e^{-yn}, \tag{5.3.2}$$

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

**Proof:** Let $g_y(t) \stackrel{\triangle}{=} (1-yt)^n$, which is strictly convex in $t \in [0,1]$. Hence, for any $x \in [0,1]$,

$$\begin{aligned}
(1-xy)^n &= g_y\big((1-x)\cdot 0 + x \cdot 1\big) \\
&\le (1-x)\cdot g_y(0) + x \cdot g_y(1) \\
&\quad \text{with equality holds if, and only if, } (x=0) \vee (x=1) \vee (y=0) \\
&= (1-x) + x \cdot (1-y)^n \\
&\le (1-x) + x \cdot \big(e^{-y}\big)^n \\
&\quad \text{with equality holds if, and only if, } (x=0) \vee (y=0) \\
&\le (1-x) + e^{-ny} \\
&\quad \text{with equality holds if, and only if, } (x=1).
\end{aligned}$$

From the above derivation, we know that equality holds for (5.3.2) if, and only if,

$$[(x=0) \vee (x=1) \vee (y=0)] \wedge [(x=0) \vee (y=0)] \wedge [x=1] = (x=1, y=0).$$

(Note that $(x=0)$ represents $\{(x,y) \in \Re^2 : x = 0 \text{ and } y \in [0,1]\}$. Similar definition applies to other sets.)

□

# Shannon's Lossy Source Coding Theorem



**Theorem 5.16 (rate distortion theorem)** For DMS and bounded additive distortion measure (namely,

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

the rate-distortion function is

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

**Proof:** Denote $f(D) \stackrel{\triangle}{=} \min_{\left\{P_{\hat{Z}|Z} \,:\, E[\rho(Z,\hat{Z})]\le D\right\}} I(Z;\hat{Z})$. Then we shall show that $R(D) \stackrel{\triangle}{=} \inf\{\hat{R}\in\Re \,:\, (\hat{R},D) \text{ is an achievable rate-distortion pair}\}$ equals $f(D)$.

1. *Achievability* (i.e., $R(D+\varepsilon) \le f(D)+4\varepsilon$ 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

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n \le f(D) + \gamma.$$

# Shannon's Lossy Source Coding Theorem



**Step 1: Optimizer.** Let $P_{\tilde{Z}|Z}$ be the optimizer of $f(D)$, i.e.,

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

Then

$$E[\rho(Z,\tilde{Z})] \le D \quad (\text{and also } \frac{1}{n}E[\rho_n(Z^n,\tilde{Z}^n)] \le D).$$

Choose $M_n$ to satisfy

$$f(D) + \frac{1}{2}\gamma \le \frac{1}{n}\log M_n \le f(D) + \gamma$$

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

$$\delta \stackrel{\triangle}{=} \min\left\{\frac{\gamma}{8}, \frac{\varepsilon}{1+2\rho_{\max}}\right\}.$$

**Step 2: Random coding.** Independently select $M_n$ codewords 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 $\sim\!\mathcal{C}_n$, 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

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

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:

$$h_n(z^n) = \begin{cases} \boldsymbol{c}_m, & \text{if } (z^n, \boldsymbol{c}_m) \in \mathcal{D}_n(\delta); \\ & \text{(when more than one satisfying the requirement,} \\ & \text{ just pick anyone.)} \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

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

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

**Step 4: Calculation of probability outside $\mathcal{J}(\mathscr{C}_n)$.** Let $N_1$ satisfying that for $n > N_1$,

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

Let

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

Then by random coding argument,

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

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 $M_n$ $n$-tuple from $\hat{\mathcal{Z}}^n$ which is not distortion joint typical with $z^n$. Hence,

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

# Shannon's Lossy Source Coding Theorem

For convenient, we let $K(z^n, \tilde{z}^n)$ be 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

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

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

$$\begin{aligned} E[\Omega] &= \sum_{z^n \in \mathcal{Z}^n} P_{Z^n}(z^n) \left(1 - \sum_{\tilde{z}^n \in \hat{\mathcal{Z}}^n} P_{\tilde{Z}^n}(\tilde{z}^n) K(z^n, \tilde{z}^n)\right)^{M_n} \\ &\leq \sum_{z^n \in \mathcal{Z}^n} P_{Z^n}(z^n) \left(1 - \sum_{\tilde{z}^n \in \hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) e^{-n(I(Z;\tilde{Z})+3\delta)} K(z^n, \tilde{z}^n)\right)^{M_n} \\ &\qquad \text{(by (5.3.1))} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{z^n \in \mathcal{Z}^n} P_{Z^n}(z^n) \left(1 - e^{-n(I(Z;\tilde{Z})+3\delta)} \sum_{\tilde{z}^n \in \hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) K(z^n, \tilde{z}^n)\right)^{M_n} \\
&\leq \sum_{z^n \in \mathcal{Z}^n} P_{Z^n}(z^n) \Bigg(1 - \sum_{\tilde{z}^n \in \hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) K(z^n, \tilde{z}^n) \\
&\quad + \exp\left\{-M_n \cdot e^{-n(I(Z;\tilde{Z})+3\delta)}\right\}\Bigg) \quad (\text{from (5.3.2)}) \\
&\leq \sum_{z^n \in \mathcal{Z}^n} P_{Z^n}(z^n) \Bigg(1 - \sum_{\tilde{z}^n \in \hat{\mathcal{Z}}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) K(z^n, \tilde{z}^n) \\
&\quad + \exp\left\{-e^{n(f(D)+\gamma/2)} \cdot e^{-n(I(Z;\tilde{Z})+3\delta)}\right\}\Bigg), \quad (\text{for } f(D)+\gamma/2 < (1/n)\log M_n) \\
&\leq 1 - P_{Z^n,\tilde{Z}^n}(\mathcal{D}_n(\delta)) + \exp\left\{-e^{n\delta}\right\}, \quad (\text{for } f(D) = I(Z;\tilde{Z}) \text{ and } \delta \leq \gamma/8) \\
&= P_{Z^n,\tilde{Z}^n}(\mathcal{D}_n^c(\delta)) + \exp\left\{-e^{n\delta}\right\} \\
&\leq \delta + \delta = 2\delta, \quad \text{for } n > N \stackrel{\triangle}{=} \max\left\{N_0, N_1, \frac{1}{\delta}\log\log\left(\frac{1}{\min\{\delta, 1\}}\right)\right\}.
\end{aligned}
$$

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

**Step 5: Calculation of distortion.** For the optimal codebook $\mathscr{C}_n^*$ (from the previous step) at $n > N$, its distortion is:

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

# Shannon's Lossy Source Coding Theorem

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

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n < f(D),$$

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.

**Step 1: Convexity of mutual information.** By the convexity of mutual information $I(Z;\hat{Z})$ with respective to $P_{\hat{Z}|Z}$,

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

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

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

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

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

we have

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

Therefore, $f(D)$ is a convex function.

# Shannon's Lossy Source Coding Theorem

**Step 3: Strictly decreasingness and continuity of $f(D)$.**

By definition, $f(D)$ is non-increasing in $D$. Also, $f(D) = 0$ for

$$D \geq \min_{\{P_{\hat{Z}}\}} \sum_{z \in \mathcal{Z}} \sum_{\hat{z} \in \mathcal{Z}} P_Z(z) P_{\hat{Z}}(\hat{z}) \rho(z, \hat{z})$$

(which is finite from boundedness of distortion measure). Together with its convexity, the strictly decreasingness and continuity of $f(D)$ over $\{D \geq 0 : f(D) > 0\}$ is proved.

**Step 4: Main proof.**

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

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

# Shannon's Lossy Source Coding Theorem

Finally, $\limsup_{n\to\infty}(1/n)\log M_n < f(D)$ implies the existence of $N$ and $\gamma > 0$ such that $(1/n)\log M_n < f(D) - \gamma$ for all $n > N$.

Therefore, for $n > N$,

$$f\left(\frac{1}{n}E[\rho_n(Z^n, h_n(Z^n))]\right) < f(D) - \gamma,$$

which, together with the strictly decreasing of $f(D)$, implies

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

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

# Shannon's Lossy Source Coding Theorem

3. *Summary*:

- For $D \in \{D \geq 0 : f(D) > 0\}$, the achievability and converse parts jointly imply that

$$f(D) + 4\varepsilon \geq R(D + \varepsilon) \geq f(D)$$

for arbitrarily small $\varepsilon > 0$. Together with the continuity of $f(D)$, we obtain that $R(D) = f(D)$ for $D \in \{D \geq 0 : f(D) > 0\}$.

- For $D \in \{D \geq 0 : f(D) = 0\}$, the achievability part gives us

$$f(D) + 4\varepsilon = 4\varepsilon \geq R(D + \varepsilon) \geq 0$$

for arbitrarily small $\varepsilon > 0$. This immediately implies that $R(D) = 0 (= f(D))$ as desired.

□

# Final Remarks



- The formula of the rate distortion function obtained in the previous theorem is also valid for the squared error distortion over real numbers, even if it is *unbounded*.
- Here, we put the boundedness assumption just to facilitate the exposition of the current proof.
- Readers may refer to volume II of the lecture notes for a more general proof.
- The discussion on lossy data compression, especially on continuous sources, will be continued in the next chapter.
- Examples of the calculation of rate-distortion functions will also be given in the next chapter.

# Final Remarks



- After introducing
  - Shannon's source coding theorem for block codes,
  - Shannon's channel coding theorem for block codes
  - rate-distortion theorem for lossy block compression codes

  for i.i.d. or stationary ergodic system setting, we would like to once again make clear the "key concepts" behind these lengthy proofs, that is:
  - *typical-set*
    - The *typical-set* argument, specifically,
      - $\delta$-typical set for source coding
      - joint $\delta$-typical set for channel coding
      - distortion typical set for rate-distortion

      uses the law-of-large-number or AEP reasoning 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.

# Final Remarks



  - *random-coding.*
    * The *random-coding* argument 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.

- In situations where the two arguments apply, a similar theorem can often be established.
- Question is "Can we extend the theorems to cases where the two arguments fail?"
- It is obvious that only when new proving technique (other than the two arguments) is developed can the answer be affirmative.
- We will further explore this issue in Volume II of the lecture notes.

# Key Notes



- Why lossy data compression (e.g., to transmit a source with entropy larger than capacity)
- Distortion measure
- Why we can focus on cases of source alphabet and reproduction alphabet having the same size?
- 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 $\longrightarrow$ Shannon's first coding theorem;
- Shannon's channel coding theorem $\longrightarrow$ Shannon's second coding theorem;
- Rate distortion theorem $\longrightarrow$ Shannon's third coding theorem.
- Information transmission Theorem $\longrightarrow$ Joint source-channel coding theorem (will be introduced in the next chapter)