香港中文大學
The Chinese University of Hong Kong

Chapter 8
Rate-Distortion Theory

© Raymond W. Yeung 2014
The Chinese University of Hong Kong

Information Transmission with Distortion

Information Transmission with Distortion

- Consider compressing an information source with entropy rate H at rate $R < H$.

Information Transmission with Distortion

- Consider compressing an information source with entropy rate H at rate $R < H$.
- By the source coding theorem, $P_e \to 1$ as $n \to \infty$.

Information Transmission with Distortion

- Consider compressing an information source with entropy rate H at rate $R < H$.
- By the source coding theorem, $P_e \to 1$ as $n \to \infty$.
- Under such a situation, information must be transmitted with "distortion".

Information Transmission with Distortion

- Consider compressing an information source with entropy rate H at rate $R < H$.
- By the source coding theorem, $P_e \to 1$ as $n \to \infty$.
- Under such a situation, information must be transmitted with "distortion".
- What is the best possible tradeoff?

In this chapter

In this chapter

- Single-letter distortion measures

In this chapter

- Single-letter distortion measures
- The rate-distortion function $R(D)$

In this chapter

- Single-letter distortion measures
- The rate-distortion function $R(D)$
- The rate-distortion theorem for an i.i.d. information source

香港中文大學
The Chinese University of Hong Kong

8.1 Single-Letter Distortion Measure

Setup

Setup

- Let $\{X_k, k \geq 1\}$ be an i.i.d. information source with generic random variable $X \sim p(x)$, where $|\mathcal{X}| < \infty$.

Setup

- Let $\{X_k, k \geq 1\}$ be an i.i.d. information source with generic random variable $X \sim p(x)$, where $|\mathcal{X}| < \infty$.

- Consider a source sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and a reproduction sequence $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$.

Setup

- Let $\{X_k, k \geq 1\}$ be an i.i.d. information source with generic random variable $X \sim p(x)$, where $|\mathcal{X}| < \infty$.
- Consider a source sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and a reproduction sequence $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$.
- The components of $\hat{\mathbf{x}}$ take values in a reproduction alphabet $\hat{\mathcal{X}}$, where $|\hat{\mathcal{X}}| < \infty$.

Setup

- Let $\{X_k, k \geq 1\}$ be an i.i.d. information source with generic random variable $X \sim p(x)$, where $|\mathcal{X}| < \infty$.

- Consider a source sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and a reproduction sequence $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$.

- The components of $\hat{\mathbf{x}}$ take values in a reproduction alphabet $\hat{\mathcal{X}}$, where $|\hat{\mathcal{X}}| < \infty$.

- In general, $\hat{\mathcal{X}}$ may be different from $\mathcal{X}$.

Setup

- Let $\{X_k, k \geq 1\}$ be an i.i.d. information source with generic random variable $X \sim p(x)$, where $|\mathcal{X}| < \infty$.

- Consider a source sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and a reproduction sequence $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$.

- The components of $\hat{\mathbf{x}}$ take values in a reproduction alphabet $\hat{\mathcal{X}}$, where $|\hat{\mathcal{X}}| < \infty$.

- In general, $\hat{\mathcal{X}}$ may be different from $\mathcal{X}$.

- For example, $\hat{\mathbf{x}}$ can be a quantized version of $\mathbf{x}$.

Definition 8.1 A single-letter distortion measure is a mapping

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

The value $d(x, \hat{x})$ denotes the distortion incurred when a source symbol x is reproduced as $\hat{x}$.

Definition 8.1 A single-letter distortion measure is a mapping

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

The value $d(x, \hat{x})$ denotes the distortion incurred when a source symbol x is reproduced as $\hat{x}$.

Definition 8.2 The average distortion between a source sequence $\mathbf{x} \in \mathcal{X}^n$ and a reproduction sequence $\hat{\mathbf{x}} \in \hat{\mathcal{X}}^n$ induced by a single-letter distortion measure d is defined by

$$d(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{k=1}^{n} d(x_k, \hat{x}_k).$$

Examples of a Distortion Measure

Examples of a Distortion Measure

Let $\hat{\mathcal{X}} = \mathcal{X}$.

Examples of a Distortion Measure

Let $\hat{\mathcal{X}} = \mathcal{X}$.

1. Square-error: $d(x, \hat{x}) = (x - \hat{x})^2$, where $\mathcal{X}$ and $\hat{\mathcal{X}}$ are real.

Examples of a Distortion Measure

Let $\hat{\mathcal{X}} = \mathcal{X}$.

1. Square-error: $d(x, \hat{x}) = (x - \hat{x})^2$, where $\mathcal{X}$ and $\hat{\mathcal{X}}$ are real.

2. Hamming distortion:

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

where the symbols in $\mathcal{X}$ do not carry any particular meaning.

Examples of a Distortion Measure

Let $\hat{\mathcal{X}} = \mathcal{X}$.

1. Square-error: $d(x, \hat{x}) = (x - \hat{x})^2$, where $\mathcal{X}$ and $\hat{\mathcal{X}}$ are real.

2. Hamming distortion:

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

where the symbols in $\mathcal{X}$ do not carry any particular meaning.

Let $\hat{X}$ be an estimate of X.

Examples of a Distortion Measure

Let $\hat{\mathcal{X}} = \mathcal{X}$.

1. Square-error: $d(x, \hat{x}) = (x - \hat{x})^2$, where $\mathcal{X}$ and $\hat{\mathcal{X}}$ are real.

2. Hamming distortion:

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

where the symbols in $\mathcal{X}$ do not carry any particular meaning.

Let $\hat{X}$ be an estimate of X.

1. If d is the square-error distortion measure, $Ed(X, \hat{X})$ is called the mean square error.

Examples of a Distortion Measure

Let $\hat{\mathcal{X}} = \mathcal{X}$.

1. Square-error: $d(x, \hat{x}) = (x - \hat{x})^2$, where $\mathcal{X}$ and $\hat{\mathcal{X}}$ are real.

2. Hamming distortion:

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

where the symbols in $\mathcal{X}$ do not carry any particular meaning.

Let $\hat{X}$ be an estimate of X.

1. If d is the square-error distortion measure, $Ed(X, \hat{X})$ is called the mean square error.

2. If d is the Hamming distortion measure,

$$Ed(X, \hat{X}) = \Pr\{X = \hat{X}\} \cdot 0 + \Pr\{X \neq \hat{X}\} \cdot 1 = \Pr\{X \neq \hat{X}\}$$

is the probability of error. For a source sequence $\mathbf{x}$ and a reproduction sequence $\hat{\mathbf{x}}$, the average distortion $d(\mathbf{x}, \hat{\mathbf{x}})$ gives the frequency of error in $\hat{\mathbf{x}}$.

Normalization of a Distortion Measure

Definition 8.5 For a distortion measure d, for each $x \in \mathcal{X}$, let $\hat{x}^*(x) \in \hat{\mathcal{X}}$ minimize $d(x, \hat{x})$ over all $\hat{x} \in \hat{\mathcal{X}}$. A distortion measure d is said to be normal if

$$c_x \stackrel{\text{def}}{=} d(x, \hat{x}^*(x)) = 0$$

for all $x \in \mathcal{X}$.

Normalization of a Distortion Measure

Definition 8.5 For a distortion measure d, for each $x \in \mathcal{X}$, let $\hat{x}^*(x) \in \hat{\mathcal{X}}$ minimize $d(x, \hat{x})$ over all $\hat{x} \in \hat{\mathcal{X}}$. A distortion measure d is said to be normal if

$$c_x \stackrel{\text{def}}{=} d(x, \hat{x}^*(x)) = 0$$

for all $x \in \mathcal{X}$.

Example 8.6 Let d be a distortion measure defined by

$d(x, \hat{x})$	a	b	c
1	$\underline{2}$	7	5
2	4	$\underline{3}$	8

Then $\hat{x}^*(1) = a$ and $\hat{x}^*(2) = b$.

Normalization of a Distortion Measure

- A normal distortion measure is one which allows a source X to be reproduced with zero distortion.

Normalization of a Distortion Measure

- A normal distortion measure is one which allows a source X to be reproduced with zero distortion.
- The square-error distortion measure and the Hamming distortion measure are normal distortion measures.

Normalization of a Distortion Measure

- A normal distortion measure is one which allows a source X to be reproduced with zero distortion.

- The square-error distortion measure and the Hamming distortion measure are normal distortion measures.

- The normalization of a distortion measure d is the distortion measure $\tilde{d}$ defined by

$$\tilde{d}(x, \hat{x}) = d(x, \hat{x}) - c_x$$

  for all $(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}$.

Normalization of a Distortion Measure

- A normal distortion measure is one which allows a source X to be reproduced with zero distortion.

- The square-error distortion measure and the Hamming distortion measure are normal distortion measures.

- The normalization of a distortion measure d is the distortion measure $\tilde{d}$ defined by
$$\tilde{d}(x, \hat{x}) = d(x, \hat{x}) - c_x$$
for all $(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}$.

- It suffices to consider normal distortion measures as we will see.

Example 8.6 (cont.) Let d be a distortion measure defined by

$d(x, \hat{x})$	a	b	c
1	2	7	5
2	4	3	8

Then $\tilde{d}$, the normalization of d, is given by

$\tilde{d}(x, \hat{x})$	a	b	c
1	0	5	3
2	1	0	5

Example 8.6 (cont.) Let d be a distortion measure defined by

$d(x, \hat{x})$	a	b	c
1	2	7	5
2	4	3	8

$c_1 = 2$

Then $\tilde{d}$, the normalization of d, is given by

$\tilde{d}(x, \hat{x})$	a	b	c
1	0	5	3
2	1	0	5

Example 8.6 (cont.) Let d be a distortion measure defined by

$d(x, \hat{x})$	a	b	c
1	2	7	5
2	4	3	8

$c_1 = 2$

$c_2 = 3$

Then $\tilde{d}$, the normalization of d, is given by

$\tilde{d}(x, \hat{x})$	a	b	c
1	0	5	3
2	1	0	5

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$Ed(X, \hat{X}) = \sum_x \sum_{\hat{x}} p(x, \hat{x}) d(x, \hat{x})$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned} Ed(X,\hat{X}) &= \sum_{x}\sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\ &= \sum_{x}\sum_{\hat{x}} p(x,\hat{x}) \left[\tilde{d}(x,\hat{x}) + c_x\right] \end{aligned}$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned} Ed(X, \hat{X}) &= \sum_{x} \sum_{\hat{x}} p(x, \hat{x}) d(x, \hat{x}) \\ &= \sum_{x} \sum_{\hat{x}} p(x, \hat{x}) \left[\underline{\tilde{d}(x, \hat{x})} + c_x \right] \end{aligned}$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned} Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\ &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) \left[\underline{\tilde{d}(x,\hat{x})} + c_x \right] \\ &= \underline{E\tilde{d}(X,\hat{X})} + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) c_x \end{aligned}$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned} Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\ &= \sum_x \sum_{\hat{x}} \underline{p(x,\hat{x})} \left[\tilde{d}(x,\hat{x}) + c_x \right] \\ &= E\tilde{d}(X,\hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) c_x \end{aligned}$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$
\begin{aligned}
Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\
&= \sum_x \sum_{\hat{x}} \underline{p(x,\hat{x})} \left[\tilde{d}(x,\hat{x}) + c_x \right] \\
&= E\tilde{d}(X,\hat{X}) + \sum_x \underline{p(x)} \sum_{\hat{x}} \underline{p(\hat{x}|x)} c_x
\end{aligned}
$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$
\begin{aligned}
Ed(X, \hat{X}) &= \sum_x \sum_{\hat{x}} p(x, \hat{x}) d(x, \hat{x}) \\
&= \sum_x \sum_{\hat{x}} p(x, \hat{x}) \left[\tilde{d}(x, \hat{x}) + c_x \right] \\
&= E\tilde{d}(X, \hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) c_x
\end{aligned}
$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned} Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\ &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) \left[\tilde{d}(x,\hat{x}) + c_x \right] \\ &= E\tilde{d}(X,\hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) \underline{c_x} \\ &= E\tilde{d}(X,\hat{X}) + \sum_x p(x) \underline{c_x} \left(\sum_{\hat{x}} p(\hat{x}|x) \right) \end{aligned}$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned}
Ed(X, \hat{X}) &= \sum_x \sum_{\hat{x}} p(x, \hat{x}) d(x, \hat{x}) \\
&= \sum_x \sum_{\hat{x}} p(x, \hat{x}) \left[\tilde{d}(x, \hat{x}) + c_x \right] \\
&= E\tilde{d}(X, \hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) \underline{c_x} \\
&= E\tilde{d}(X, \hat{X}) + \sum_x p(x) \underline{c_x} \left(\sum_{\hat{x}} p(\hat{x}|x) \right)
\end{aligned}$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned}
Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\
&= \sum_x \sum_{\hat{x}} p(x,\hat{x}) \left[\tilde{d}(x,\hat{x}) + c_x\right] \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) c_x \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \left(\sum_{\hat{x}} p(\hat{x}|x)\right) \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x
\end{aligned}$$

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned}
Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\
&= \sum_x \sum_{\hat{x}} p(x,\hat{x}) \left[\tilde{d}(x,\hat{x}) + c_x\right] \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) c_x \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \left(\sum_{\hat{x}} p(\hat{x}|x)\right) \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \\
&= E\tilde{d}(X,\hat{X}) + \Delta,
\end{aligned}$$

where

$$\Delta = \sum_x p(x) c_x$$

is a constant which depends only on $p(x)$ and d but not on the conditional distribution $p(\hat{x}|x)$.

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$
\begin{aligned}
Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\
&= \sum_x \sum_{\hat{x}} p(x,\hat{x}) \left[\tilde{d}(x,\hat{x}) + c_x\right] \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) c_x \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \left(\sum_{\hat{x}} p(\hat{x}|x)\right) \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \\
&= E\tilde{d}(X,\hat{X}) + \Delta,
\end{aligned}
$$

where

$$\Delta = \sum_x p(x) c_x$$

is a constant which depends only on $p(x)$ and d but not on the conditional distribution $p(\hat{x}|x)$.

Let $\hat{X}$ be any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned}
Ed(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\
&= \sum_x \sum_{\hat{x}} p(x,\hat{x}) \left[\tilde{d}(x,\hat{x}) + c_x\right] \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) c_x \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \left(\sum_{\hat{x}} p(\hat{x}|x) \right) \\
&= E\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \\
&= E\tilde{d}(X,\hat{X}) + \Delta,
\end{aligned}$$

where

$$\Delta = \sum_x p(x) c_x$$

is a constant which depends only on $p(x)$ and d but not on the conditional distribution $p(\hat{x}|x)$.

Definition 8.7 Let $\hat{x}^*$ minimizes $Ed(X, \hat{x})$ over all $\hat{x} \in \hat{\mathcal{X}}$, and define

$$D_{max} = Ed(X, \hat{x}^*).$$

Note: $\hat{x}^*$ is not the same as $\hat{x}^*(x)$, and it depends on $p(x)$.

Definition 8.7 Let $\hat{x}^*$ minimizes $Ed(X, \hat{x})$ over all $\hat{x} \in \hat{\mathcal{X}}$, and define

$$D_{max} = Ed(X, \hat{x}^*).$$

Note: $\hat{x}^*$ is not the same as $\hat{x}^*(x)$, and it depends on $p(x)$.

- If we know nothing about a source variable X except for $p(x)$, then $\hat{x}^*$ is the best estimate of X, and D_{max} is the minimum expected distortion between X and a constant estimate of X.

Definition 8.7 Let $\hat{x}^*$ minimizes $Ed(X, \hat{x})$ over all $\hat{x} \in \hat{\mathcal{X}}$, and define

$$D_{max} = Ed(X, \hat{x}^*).$$

Note: $\hat{x}^*$ is not the same as $\hat{x}^*(x)$, and it depends on $p(x)$.

- If we know nothing about a source variable X except for $p(x)$, then $\hat{x}^*$ is the best estimate of X, and D_{max} is the minimum expected distortion between X and a constant estimate of X.
- Specifically, by taking $\hat{\mathbf{x}}^* = (\hat{x}^*, \hat{x}^*, \cdots, \hat{x}^*)$ to be the reproduction sequence, D_{max} can be asymptotically achieved, because by WLLN,

$$d(\mathbf{X}, \hat{\mathbf{x}}^*) = \frac{1}{n} \sum_{k=1}^{n} d(X_k, \hat{x}^*) \rightarrow Ed(X, \hat{x}^*) = D_{max}$$

Definition 8.7 Let $\hat{x}^*$ minimizes $Ed(X, \hat{x})$ over all $\hat{x} \in \hat{\mathcal{X}}$, and define

$$D_{max} = Ed(X, \hat{x}^*).$$

Note: $\hat{x}^*$ is not the same as $\hat{x}^*(x)$, and it depends on $p(x)$.

- If we know nothing about a source variable X except for $p(x)$, then $\hat{x}^*$ is the best estimate of X, and D_{max} is the minimum expected distortion between X and a constant estimate of X.
- Specifically, by taking $\hat{\mathbf{x}}^* = (\hat{x}^*, \hat{x}^*, \cdots, \hat{x}^*)$ to be the reproduction sequence, D_{max} can be asymptotically achieved, because by WLLN,

$$d(\mathbf{X}, \hat{\mathbf{x}}^*) = \frac{1}{n} \sum_{k=1}^{n} d(X_k, \hat{x}^*) \rightarrow Ed(X, \hat{x}^*) = D_{max}$$

- Therefore it is not meaningful to impose a constraint $D \geq D_{max}$ on the reproduction sequence.