Chapter 8
Rate-Distortion Theory

© Raymond W. Yeung 2012

Department of Information Engineering
The Chinese University of Hong Kong

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?

8.1 Single-Letter Distortion Measure

- 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 8.2 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

- 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 $\hat{\mathbf{x}}$.

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 $x \in \mathcal{X}$.

- 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 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

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 $\hat{x}^*(x)$.

- If we know nothing about a source variable 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, D_{max} can be asymptotically achieved by taking $(\hat{x}^*, \hat{x}^*, \cdots, \hat{x}^*)$ to be the reproduction sequence.
- Therefore it is not meanful to impose a constraint $D \geq D_{max}$ on the reproduction sequence.

8.2 The Rate-Distortion Function

All the discussions are with respect to an i.i.d. information source $\{X_k, k \geq 1\}$ with generic random variable X and a distortion measure d.

Definition 8.8 An (n, M) rate-distortion code is defined by an encoding function

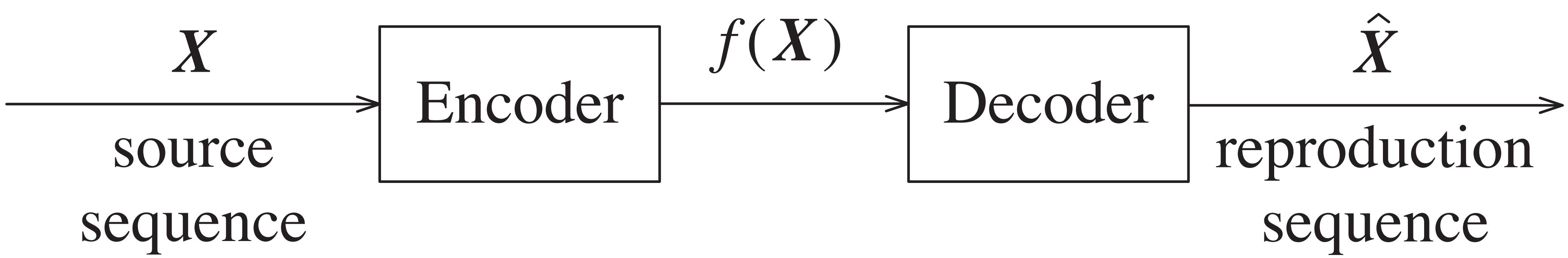
$$f : \mathcal{X}^n \to \{1, 2, \cdots, M\}$$

and a decoding function

$$g : \{1, 2, \cdots, M\} \to \hat{\mathcal{X}}^n.$$

The set $\{1, 2, \cdots, M\}$, denoted by $\mathcal{I}$, is called the index set. The reproduction sequences $g(1), g(2), \cdots, g(M)$ in $\hat{X}^n$ are called codewords, and the set of codewords is called the codebook.

A Rate-Distortion Code



Definition 8.9 The rate of an (n, M) rate-distortion code is $n^{-1} \log M$ in bits per symbol.

Definition 8.10 A rate-distortion pair (R, D) is (asymptotically) achievable if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) rate-distortion code such that

$$\frac{1}{n} \log M \leq R + \epsilon$$

and

$$\Pr\{d(\mathbf{X}, \hat{\mathbf{X}}) > D + \epsilon\} \leq \epsilon,$$

where $\hat{\mathbf{X}} = g(f(\mathbf{X}))$.

Remark If (R, D) is achievable, then (R', D) and (R, D') are achievable for all $R' \geq R$ and $D' \geq D$. This in turn implies that (R', D') are achievable for all $R' \geq R$ and $D' \geq D$.

Definition 8.11 The rate-distortion region is the subset of $\Re^2$ containing all achievable pairs (R, D).

Theorem 8.12 The rate-distortion region is closed and convex.

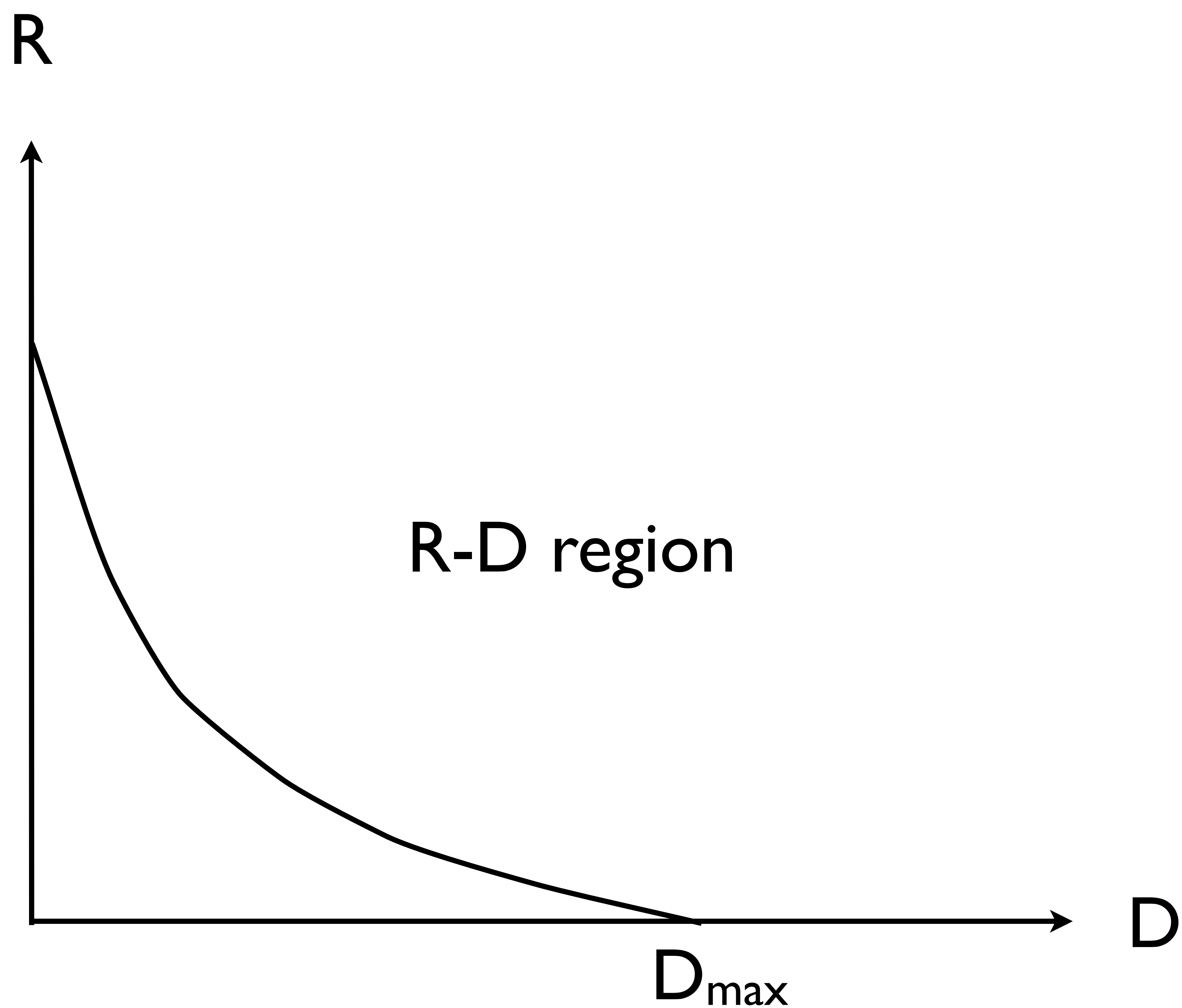
Proof

- The closeness follows from the definition of the achievability of an (R, D) pair.

- The convexity is proved by time-sharing. Specifically, if $(R^{(1)}, D^{(1)})$ and $(R^{(2)}, D^{(2)})$ are achievable, then so is $(R^{(\lambda)}, D^{(\lambda)})$, where

$$
\begin{aligned}
R^{(\lambda)} &= \lambda R^{(1)} + \bar{\lambda} R^{(2)} \\
D^{(\lambda)} &= \lambda D^{(1)} + \bar{\lambda} D^{(2)}
\end{aligned}
$$

  and $\bar{\lambda} = 1 - \lambda$. This can be seen by time-sharing between two codes, one achieving $(R^{(1)}, D^{(1)})$ for λ fraction of the time, and the other one achieving $(R^{(2)}, D^{(2)})$ for $\bar{\lambda}$ fraction of the time.


R
R-D region
D
D_{max}

Definition 8.13 The rate-distortion function $R(D)$ is the minimum of all rates R for a given distortion D such that (R, D) is achievable.

Definition 8.14 The distortion-rate function $D(R)$ is the minimum of all distortions D for a given rate R such that (R, D) is achievable.

Theorem 8.15 The following properties hold for the rate-distortion function $R(D)$:

1. $R(D)$ is non-increasing in D.

2. $R(D)$ is convex.

3. $R(D) = 0$ for $D \geq D_{max}$.

4. $R(0) \leq H(X)$.

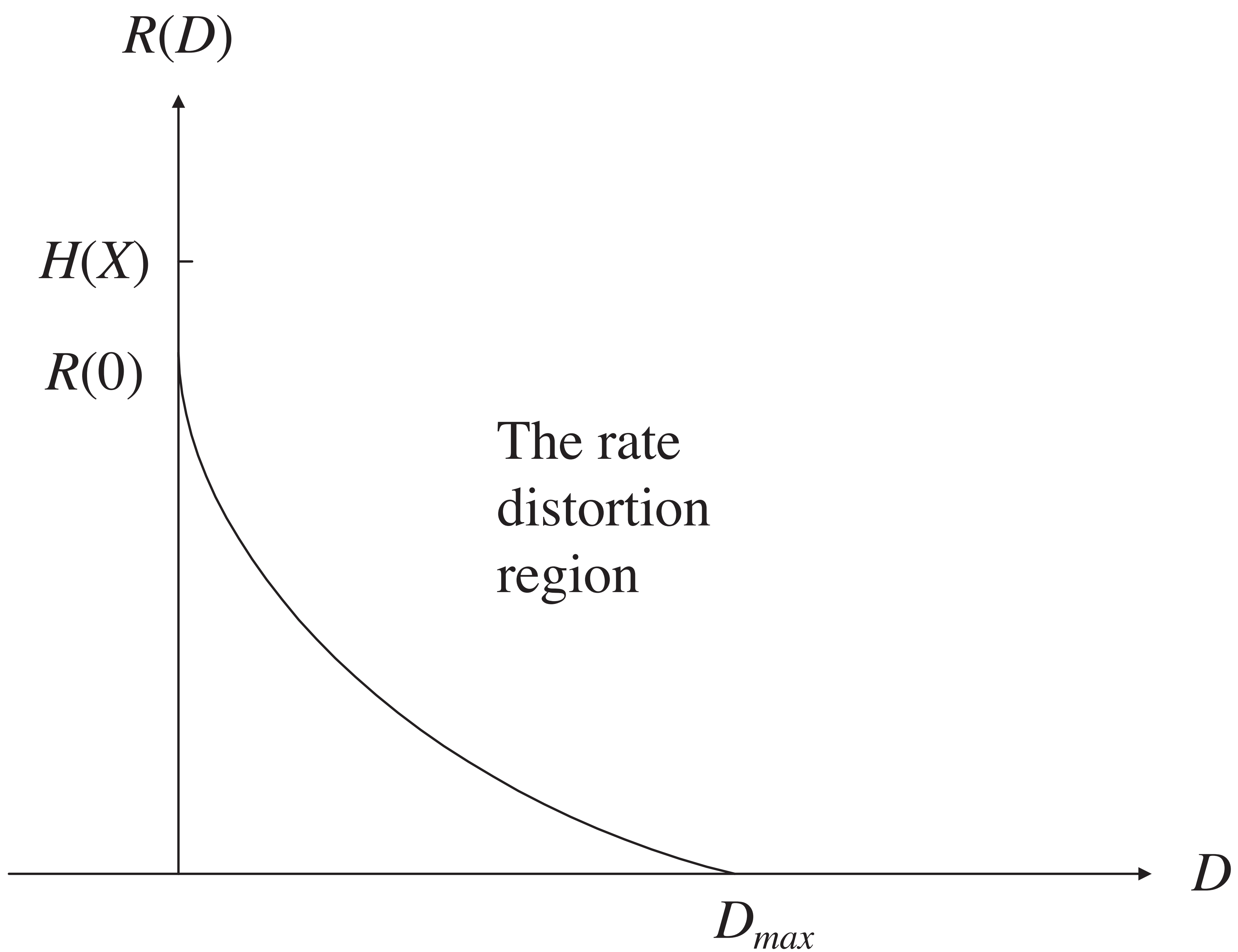
Proof

1. Let $D' \geq D$. $(R(D), D)$ achievable $\Rightarrow$ $(R(D), D')$ achievable. Then $R(D) \geq R(D')$ by definition of $R(\cdot)$.

2. Follows from the convexity of the rate-distortion region.

3. $(0, D_{max})$ is achievable $\Rightarrow$ $R(D) = 0$ for $D \geq D_{max}$.

4. Since d is assumed to be normal, $(H(X), 0)$ is achievable, and hence $R(0) \leq H(X)$.


R(D)
H(X)
R(0)
The rate
distortion
region
D
Dmax

8.3 The Rate Distortion Theorem

Definition 8.16 For $D \geq 0$, the information rate-distortion function is defined by

$$R_I(D) = \min_{\hat{X}: Ed(X,\hat{X}) \leq D} I(X;\hat{X}).$$

- The minimization is taken over the set of all $p(\hat{x}|x)$ such that $Ed(X,\hat{X}) \leq D$ is satisfied, namely the set

$$\left\{ p(\hat{x}|x) : \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D \right\}.$$

- Since this set is compact in $\Re^{|\mathcal{X}||\hat{\mathcal{X}}|}$ and $I(X;\hat{X})$ is a continuous functional of $p(\hat{x}|x)$, the minimum value of $I(X;\hat{X})$ can be attained.

- Since
$$E\tilde{d}(X,\hat{X}) = Ed(X,\hat{X}) - \Delta,$$
where Δ does not depend on $p(\hat{x}|x)$, we can always replace d by $\tilde{d}$ and D by $D-\Delta$ in the definition of $R_I(D)$ without changing the minimization problem.

- Without loss of generality, we can assume d is normal.

Theorem 8.17 (The Rate-Distortion Theorem) $R(D) = R_I(D)$.

Theorem 8.18 The following properties hold for the information rate-distortion function $R_I(D)$:

1. $R_I(D)$ is non-increasing in D.

2. $R_I(D)$ is convex.

3. $R_I(D) = 0$ for $D \geq D_{max}$.

4. $R_I(0) \leq H(X)$.

Proof of Theorem 8.18

1. For a larger D, the minimization is taken over a larger set.

3. Let $\hat{X} = \hat{x}^*$ w.p. 1 to show that $(0, D_{max})$ is achievable. Then for $D \geq D_{max}$, $R_I(D) \leq I(X; \hat{X}) = 0$, which implies $R_I(D) = 0$.

4. Let $\hat{X} = \hat{x}^*(X)$, so that $Ed(X, \hat{X}) = 0$ (since d is normal). Then

$$R_I(0) \leq I(X; \hat{X}) \leq H(X).$$

Proof of Theorem 8.18

2. Consider any $D^{(1)}, D^{(2)} \geq 0$ and $0 \leq \lambda 1$. Let $\hat{X}^{(i)}$ achieves $R_I(D^{(i)})$ for $i = 1, 2$, i.e.,

$$R_I(D^{(i)}) = I(X; \hat{X}^{(i)}),$$

where

$$Ed(X, \hat{X}^{(i)}) \leq D^{(i)},$$

Let $\hat{X}^{(\lambda)}$ be jointly distributed with X defined by

$$p_\lambda(\hat{x}|x) = \lambda p_1(\hat{x}|x) + \bar{\lambda} p_2(\hat{x}|x).$$

Then

$$\begin{aligned}
&Ed(X, \hat{X}^{(\lambda)}) \\
&\quad = \lambda Ed(X, \hat{X}^{(1)}) + \bar{\lambda} Ed(X, \hat{X}^{(2)}) \\
&\quad \leq \lambda D^{(1)} + \bar{\lambda} D^{(2)} \\
&\quad = D^{(\lambda)}.
\end{aligned}$$

Finally consider

$$\begin{aligned}\lambda R_I(D^{(1)}) + \bar{\lambda} R_I(D^{(2)}) &= \lambda I(X;\hat{X}^{(1)}) + \bar{\lambda} I(X;\hat{X}^{(2)}) \\ &\geq I(X;\hat{X}^{(\lambda)}) \\ &\geq R_I(D^{(\lambda)}).\end{aligned}$$

Corollary 8.19 If $R_I(0) > 0$, then $R_I(D)$ is strictly decreasing for $0 \leq D \leq D_{max}$, and the inequality constraint in the definition of $R_I(D)$ can be replaced by an equality constraint.

Proof

1. $R_I(D)$ must be strictly decreasing for $0 \leq D \leq D_{max}$ because $R_I(0) > 0$, $R_I(D_{max}) = 0$, and $R_I(D)$ is non-increasing and convex.

2. Show that $R_I(D) > 0$ for $0 \leq D < D_{max}$ by contradiction.

    - Suppose $R_I(D') = 0$ for some $0 \leq D' < D_{max}$, and let $R_I(D')$ be achieved by some $\hat{X}$. Then

    $$R_I(D') = I(X; \hat{X}) = 0$$

    implies that X and $\hat{X}$ are independent.

    - Show that such an $\hat{X}$ which is independent of X cannot do better than the constant estimate $\hat{x}^*$, i.e., $Ed(X, \hat{X}) \geq Ed(X, \hat{x}^*) = D_{max}$.

    - This leads to a contradiction because

    $$D' \geq Ed(X, \hat{X}) \geq D_{max}.$$

Proof

3. Show that the inequality constraints in $R_I(D)$ can be replaced by an equality constraint by contradiction.

   - Assume that $R_I(D)$ is achieved by some $\hat{X}^*$ such that $Ed(X, \hat{X}^*) = D'' < D$.
   - Then

   $$R_I(D'') = \min_{\hat{X}: Ed(X, \hat{X}) \le D''} I(X; \hat{X}) \le I(X; \hat{X}^*) = R_I(D),$$

   a contradiction because $R_I(D)$ is strictly decreasing for $0 \le D \le D_{max}$.

   - Therefore, $Ed(X, \hat{X}^*) = D$.

Remark In all problems of interest, $R(0) = R_I(0) > 0$. Otherwise, $R(D) = 0$ for all $D \ge 0$ because $R(D)$ is nonnegative and non-increasing.

Example 8.20 (Binary Source)

Let X be a binary random variable with

$$\Pr\{X = 0\} = 1 - \gamma \quad \text{and} \quad \Pr\{X = 1\} = \gamma.$$

Let $\hat{\mathcal{X}} = \{0, 1\}$ and d be the Hamming distortion measure. Show that

$$R_I(D) = \begin{cases} h_b(\gamma) - h_b(D) & \text{if } \ 0 \le D < \min(\gamma, 1-\gamma) \\ 0 & \text{if } D \ge \min(\gamma, 1-\gamma). \end{cases}$$

First consider $0 \le \gamma \le \frac{1}{2}$, and show that

$$R_I(D) = \begin{cases} h_b(\gamma) - h_b(D) & \text{if } \ 0 \le D < \gamma \\ 0 & \text{if } D \ge \gamma. \end{cases}$$

- $\hat{x}^* = 0$ and $D_{max} = Ed(X, 0) = \Pr\{X = 1\} = \gamma$.
- Consider any $\hat{X}$ and let $Y = d(X, \hat{X})$.
- Conditioning on $\hat{X}$, X and Y determine each other, and so, $H(X|\hat{X}) = H(Y|\hat{X})$.
- Then for $D < \gamma = D_{max}$ and any $\hat{X}$ such that $Ed(X, \hat{X}) \le D$,

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= h_b(\gamma) - H(Y|\hat{X}) \\ &\ge h_b(\gamma) - H(Y) \qquad (1) \\ &= h_b(\gamma) - h_b(\Pr\{X \ne \hat{X}\}) \\ &\overset{a)}{\ge} h_b(\gamma) - h_b(D), \qquad (2) \end{aligned}$$

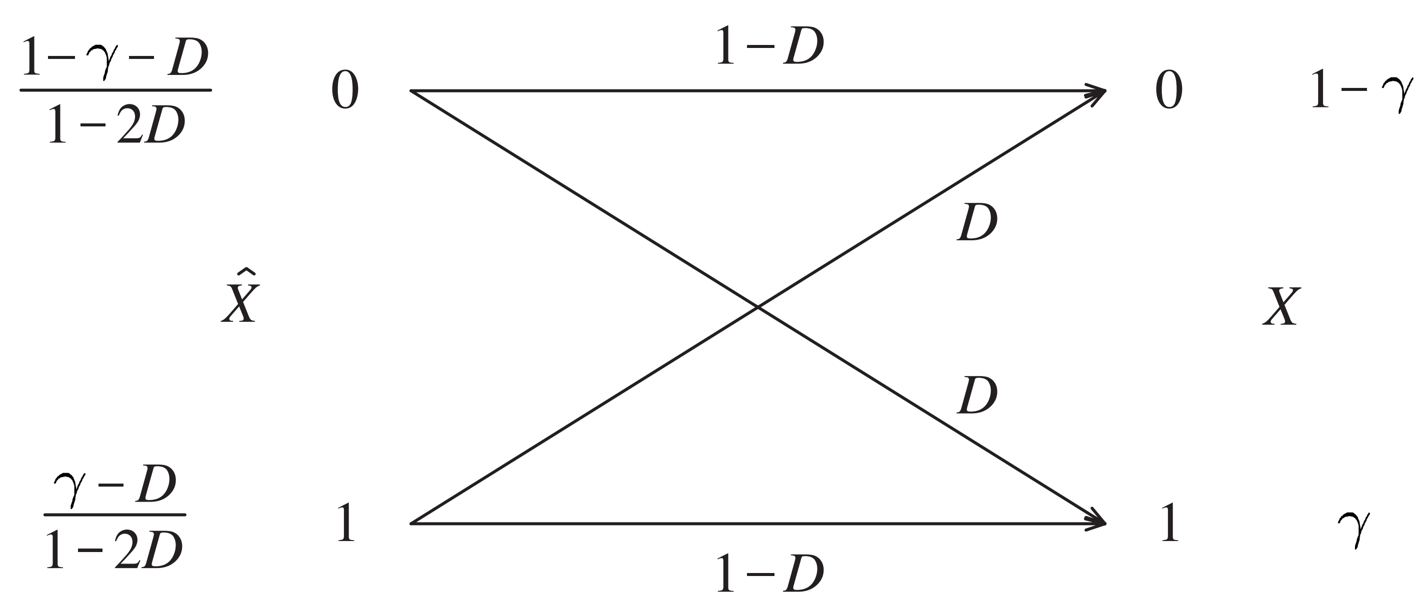
  a) because $\Pr\{X \ne \hat{X}\} = Ed(X, \hat{X}) \le D$ and $h_b(a)$ is increasing for $0 \le a \le \frac{1}{2}$.

- Therefore,

$$R_I(D) = \min_{\hat{X}:Ed(X,\hat{X})\leq D} I(X;\hat{X}) \geq h_b(\gamma) - h_b(D).$$

Now need to construct $\hat{X}$ which is tight for (1) and (2), so that the above bound is achieved.

- (1) tight $\Leftrightarrow$ Y independent of $\hat{X}$
- (2) tight $\Leftrightarrow$ $\Pr\{X \neq \hat{X}\} = D$
- The required $\hat{X}$ can be specified by the following reverse BSC:



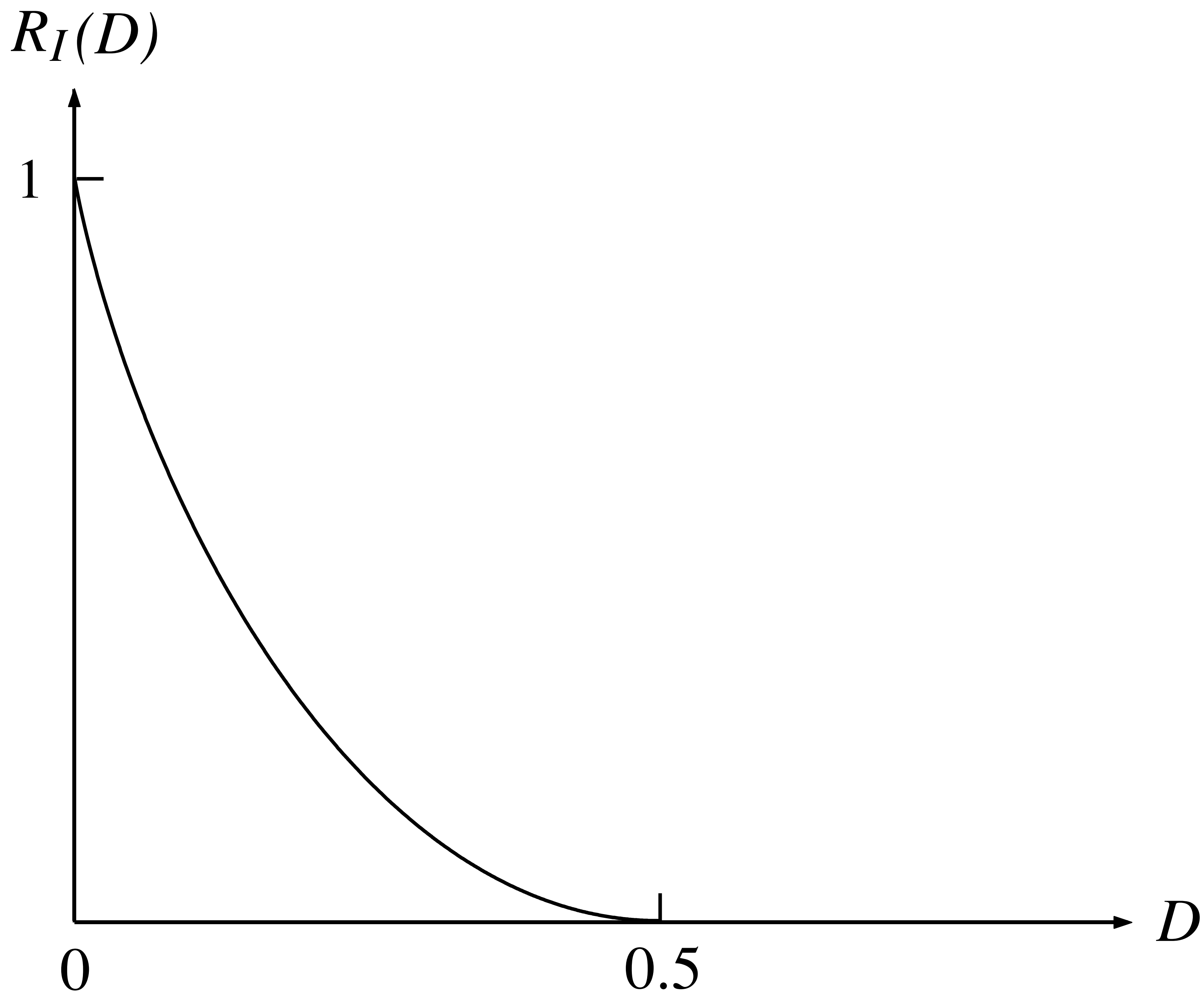
- Therefore, we conclude that

$$R_I(D) = \begin{cases} h_b(\gamma) - h_b(D) & \text{if } \ 0 \le D < \gamma \\ 0 & \text{if } D \ge \gamma. \end{cases}$$

For $1/2 \le \gamma \le 1$, by exchanging the roles of the symbols 0 and 1 and applying the same argument, we obtain $R_I(D)$ as above except that γ is replaced by $1-\gamma$. Combining the two cases, we have

$$R_I(D) = \begin{cases} h_b(\gamma) - h_b(D) & \text{if } \ 0 \le D < \min(\gamma, 1-\gamma) \\ 0 & \text{if } D \ge \min(\gamma, 1-\gamma). \end{cases}$$

for $0 \le \gamma \le 1$.



$R_I(D)$
1
0
0.5
D

A Remark

The rate-distortion theorem does not include the source coding theorem as a special case:

- In Example 8.20, $R_I(0) = h_b(\gamma) = H(X)$.
- By the rate-distortion theorem, if $R > H(X)$, the average Hamming distortion, i.e., the error probability per symbol, can be made arbitrarily small.
- However, by the source coding theorem, if $R > H(X)$, the message error probability can be made arbitrarily small, which is much stronger.

8.4 The Converse

- Prove that for any achievable rate-distortion pair (R, D), $R \geq R_I(D)$.
- Fix D and minimize R over all achievable pairs (R, D) to conclude that $R(D) \geq R_I(D)$.

Proof

1. Let (R, D) be any achievable rate-distortion pair, i.e., for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that

$$\frac{1}{n} \log M \leq R + \epsilon$$

and

$$\Pr\{d(\mathbf{X}, \hat{\mathbf{X}}) > D + \epsilon\} \leq \epsilon,$$

where $\hat{\mathbf{X}} = g(f(\mathbf{X}))$.

2. Then

$$
\begin{aligned}
n(R+\epsilon) &\overset{a)}{\geq} \log M \\
&\geq \cdots \\
&\geq \sum_{k=1}^{n} I(X_k;\hat{X}_k) \\
&\overset{c)}{\geq} \sum_{k=1}^{n} R_I(Ed(X_k,\hat{X}_k)) \\
&= n\left[\frac{1}{n}\sum_{k=1}^{n} R_I(Ed(X_k,\hat{X}_k))\right] \\
&\overset{d)}{\geq} nR_I\left(\frac{1}{n}\sum_{k=1}^{n} Ed(X_k,\hat{X}_k)\right) \\
&= nR_I(Ed(\mathbf{X},\hat{\mathbf{X}})).
\end{aligned}
$$

c) follows from from the definition of $R_I(D)$.

d) follows from the convexity of $R_I(D)$ and Jensen's inequality.

3. Let $d_{max} = \max_{x,\hat{x}} d(x, \hat{x})$. Then

$$\begin{aligned}
&Ed(\mathbf{X}, \hat{\mathbf{X}}) \\
&\quad = \quad E[d(\mathbf{X}, \hat{\mathbf{X}}) | d(\mathbf{X}, \hat{\mathbf{X}}) > D + \epsilon] \Pr\{d(\mathbf{X}, \hat{\mathbf{X}}) > D + \epsilon\} \\
&\quad\qquad + E[d(\mathbf{X}, \hat{\mathbf{X}}) | d(\mathbf{X}, \hat{\mathbf{X}}) \leq D + \epsilon] \Pr\{d(\mathbf{X}, \hat{\mathbf{X}}) \leq D + \epsilon\} \\
&\quad \leq \quad d_{max} \cdot \epsilon + (D + \epsilon) \cdot 1 \\
&\quad = \quad D + (d_{max} + 1)\epsilon.
\end{aligned}$$

That is, if the probability that the average distortion between $\mathbf{X}$ and $\hat{\mathbf{X}}$ exceeds $D + \epsilon$ is small, then the expected average distortion between $\mathbf{X}$ and $\hat{\mathbf{X}}$ can exceed D only by a small amount.

4. Therefore,

$$\begin{aligned}
R + \epsilon &\geq R_I(Ed(\mathbf{X}, \hat{\mathbf{X}})) \\
&\geq R_I(D + (d_{max} + 1)\epsilon),
\end{aligned}$$

because $R_I(D)$ is non-increasing in D.

5. $R_I(D)$ convex implies it is continuous in D. Finally,

$$\begin{aligned} R &\geq \lim_{\epsilon \to 0} R_I(D + (d_{max} + 1)\epsilon) \\ &= R_I\left(D + (d_{max} + 1) \lim_{\epsilon \to 0} \epsilon\right) \\ &= R_I(D). \end{aligned}$$

Minimizing R over all achievable pairs (R, D) for a fixed D to obtain $R(D) \geq R_I(D)$.

8.5 Achievability of $R_I(D)$

- An i.i.d. source $\{X_k : k \geq 1\}$ with generic random variable $X \sim p(x)$ is given.

- For every random variable $\hat{X}$ taking values in $\hat{\mathcal{X}}$ with $Ed(X, \hat{X}) \leq D$, where $0 \leq D \leq D_{max}$, prove that the rate-distortion pair $(I(X; \hat{X}), D)$ is achievable by showing for large n the existence of a rate-distortion code such that

  1. the rate of the code is not more than $I(X; \hat{X}) + \epsilon$;
  2. $d(\mathbf{X}, \hat{\mathbf{X}}) \leq D + \epsilon$ with probability almost 1.

- Minimize $I(X; \hat{X})$ over all such $\hat{X}$ to conclude that $(R_I(D), D)$ is achievable.

- This implies that $R_I(D) \geq R(D)$.

Random Coding Scheme

- Fix $\epsilon > 0$ and $\hat{X}$ with $Ed(X, \hat{X}) \leq D$, where $0 \leq D \leq D_{max}$. Let δ be specified later.

- Let M be an integer satisfying

$$I(X;\hat{X}) + \frac{\epsilon}{2} \leq \frac{1}{n} \log M \leq I(X;\hat{X}) + \epsilon,$$

  where n is sufficiently large.

- The random coding scheme:

  1. Construct a codebook $\mathcal{C}$ of an (n, M) code by randomly generating M codewords in $\hat{\mathcal{X}}^n$ independently and identically according to $p(\hat{x})^n$. Denote these codewords by $\hat{\mathbf{X}}(1), \hat{\mathbf{X}}(2), \cdots, \hat{\mathbf{X}}(M)$.
  2. Reveal the codebook $\mathcal{C}$ to both the encoder and the decoder.
  3. The source sequence $\mathbf{X}$ is generated according to $p(x)^n$.

4. The encoder encodes the source sequence $\mathbf{X}$ into an index K in the set $\mathcal{I} = \{1, 2, \cdots, M\}$. The index K takes the value i if

   (a) $(\mathbf{X}, \hat{\mathbf{X}}(i)) \in T^n_{[X\hat{X}]\delta}$,

   (b) for all $i' \in \mathcal{I}$, if $(\mathbf{X}, \hat{\mathbf{X}}(i')) \in T^n_{[X\hat{X}]\delta}$, then $i' \leq i$;

   i.e., if there exists more than one i satisfying (a), let K be the largest one. Otherwise, K takes the constant value 1.

5. The index K is delivered to the decoder.

6. The decoder outputs $\hat{\mathbf{X}}(K)$ as the reproduction sequence $\hat{\mathbf{X}}$.

Performance Analysis

- The event $\{K = 1\}$ occurs in one of the following two scenarios:

  1. $\hat{X}(1)$ is the only codeword in $\mathcal{C}$ which is jointly typical with $\mathbf{X}$.
  2. No codeword in $\mathcal{C}$ is jointly typical with $\mathbf{X}$.

  In other words, if $K = 1$, then $\mathbf{X}$ is jointly typical with none of the codewords $\hat{X}(2), \hat{X}(3), \cdots, \hat{X}(M)$.

- Define the event

$$E_i = \left\{ (\mathbf{X}, \hat{\mathbf{X}}(i)) \in T^n_{[X\hat{X}]\delta} \right\}$$

- Then

$$\{K = 1\} \subset E_2^c \cap E_3^c \cap \cdots \cap E_M^c.$$

- Since the codewords are generated i.i.d., conditioning on $\{\mathbf{X} = \mathbf{x}\}$ for any $\mathbf{x} \in \mathcal{X}^n$, the events E_i are mutually independent and have the same probability.

- Then for any $\mathbf{x} \in \mathcal{X}^n$,

$$\begin{aligned} \Pr\{K = 1 | \mathbf{X} = \mathbf{x}\} &\leq \Pr\{E_2^c \cap E_3^c \cap \cdots \cap E_M^c | \mathbf{X} = \mathbf{x}\} \\ &= \prod_{i=2}^{M} \Pr\{E_i^c | \mathbf{X} = \mathbf{x}\} \\ &= (\Pr\{E_1^c | \mathbf{X} = \mathbf{x}\})^{M-1} \\ &= (1 - \Pr\{E_1 | \mathbf{X} = \mathbf{x}\})^{M-1}. \end{aligned}$$

- We will focus on $\mathbf{x} \in S^n_{[X]\delta}$ where

$$S^n_{[X]\delta} = \{\mathbf{x} \in T^n_{[X]\delta} : |T^n_{[\hat{X}|X]\delta}(\mathbf{x})| \geq 1\},$$

because $\Pr\{\mathbf{X} \in S^n_{[X]\delta}\} \approx 1$ for large n (Proposition 6.13).

- For $\mathbf{x} \in S^n_{[X]\delta}$, obtain a lower bound on $\Pr\{E_1|\mathbf{X} = \mathbf{x}\}$ as follows:

$$\begin{aligned}
\Pr\{E_1|\mathbf{X} = \mathbf{x}\} &= \Pr\left\{(\mathbf{x}, \hat{\mathbf{X}}(1)) \in T^n_{[X\hat{X}]\delta}\right\} \\
&= \sum_{\hat{\mathbf{x}} \in T^n_{[\hat{X}|X]\delta}(\mathbf{x})} p(\hat{\mathbf{x}}) \\
&\stackrel{a)}{\geq} \sum_{\hat{\mathbf{x}} \in T^n_{[\hat{X}|X]\delta}(\mathbf{x})} 2^{-n(H(\hat{X})+\eta)} \\
&\stackrel{b)}{\geq} 2^{n(H(\hat{X}|X)-\xi)} 2^{-n(H(\hat{X})+\eta)} \\
&= 2^{-n(H(\hat{X})-H(\hat{X}|X)+\xi+\eta)} \\
&= 2^{-n(I(X;\hat{X})+\zeta)},
\end{aligned}$$

where $\zeta = \xi + \eta \to 0$ as $\delta \to 0$. In the above,

a) follows because from the consistency of strong typicality, if $(\mathbf{x}, \hat{\mathbf{x}}) \in T^n_{[X\hat{X}]\delta}$, then $\hat{\mathbf{x}} \in T^n_{[\hat{X}]\delta}$.

b) follows from conditional strong AEP.

- Therefore,

$$
\begin{aligned}
\Pr\{K=1|\mathbf{X}=\mathbf{x}\} &\le \Pr\{E_2^c \cap E_3^c \cap \cdots \cap E_M^c|\mathbf{X}=\mathbf{x}\} \\
&\le \left[1-2^{-n(I(X;\hat{X})+\zeta)}\right]^{M-1}
\end{aligned}
$$

- Then

$$
\begin{aligned}
\ln \Pr\{K=1|\mathbf{X}=\mathbf{x}\} &\le (M-1)\ln\left[1-2^{-n(I(X;\hat{X})+\zeta)}\right] \\
&\overset{a)}{\le} \left(2^{n(I(X;\hat{X})+\frac{\epsilon}{2})}-1\right)\ln\left[1-2^{-n(I(X;\hat{X})+\zeta)}\right] \\
&\overset{b)}{\le} -\left(2^{n(I(X;\hat{X})+\frac{\epsilon}{2})}-1\right)2^{-n(I(X;\hat{X})+\zeta)} \\
&= -\left[2^{n(\frac{\epsilon}{2}-\zeta)}-2^{-n(I(X;\hat{X})+\zeta)}\right]
\end{aligned}
$$

a) follows because the logarithm is negative.
b) follows from the fundamental inequality.

- Let δ be sufficiently small so that

$$\frac{\epsilon}{2} - \zeta > 0. \tag{1}$$

Then the upper bound on $\ln \Pr\{K = 1|\mathbf{X} = \mathbf{x}\}$ tends to $-\infty$ as $n \to \infty$, i.e., $\Pr\{K = 1|\mathbf{X} = \mathbf{x}\} \to 0$ as $n \to \infty$.

- This implies for sufficiently large n,

$$\Pr\{K = 1|\mathbf{X} = \mathbf{x}\} \leq \frac{\epsilon}{2}.$$

- It follows that

$$\begin{aligned}
\Pr\{K=1\} &= \sum_{\mathbf{x}\in S^n_{[X]\delta}} \Pr\{K=1|\mathbf{X}=\mathbf{x}\}\Pr\{\mathbf{X}=\mathbf{x}\} \\
&\quad + \sum_{\mathbf{x}\notin S^n_{[X]\delta}} \Pr\{K=1|\mathbf{X}=\mathbf{x}\}\Pr\{\mathbf{X}=\mathbf{x}\} \\
&\le \sum_{\mathbf{x}\in S^n_{[X]\delta}} \frac{\epsilon}{2}\cdot\Pr\{\mathbf{X}=\mathbf{x}\} + \sum_{\mathbf{x}\notin S^n_{[X]\delta}} 1\cdot\Pr\{\mathbf{X}=\mathbf{x}\} \\
&= \frac{\epsilon}{2}\cdot\Pr\{\mathbf{X}\in S^n_{[X]\delta}\} + \Pr\{\mathbf{X}\notin S^n_{[X]\delta}\} \\
&\le \frac{\epsilon}{2}\cdot 1 + (1-\Pr\{\mathbf{X}\in S^n_{[X]\delta}\}) \\
&< \frac{\epsilon}{2}+\delta,
\end{aligned}$$

  where we have invoked Proposition 6.13 in the last step.

- By letting δ be sufficiently small so that both (1) and $\delta < \frac{\epsilon}{2}$ are satisfied, we obtain

$$\Pr\{K=1\} < \epsilon.$$

Main Idea

- Randomly generate M codewords in $\hat{\mathcal{X}}^n$ according to $p(\hat{x})^n$, where n is large.

- $\mathbf{X} \in S^n_{[X]\delta}$ with high probability.

- For $\mathbf{x} \in S^n_{[X]\delta}$, by conditional strong AEP,

$$\Pr\left\{(\mathbf{X}, \hat{\mathbf{X}}(i)) \in T^n_{[X\hat{X}]\delta} \,|\, \mathbf{X} = \mathbf{x}\right\} \approx 2^{-nI(X;\hat{X})}.$$

- If M grows with n at a rate higher than $I(X;\hat{X})$, then the probability that there exists at least one $\hat{\mathbf{X}}(i)$ which is jointly typical with the source sequence $\mathbf{X}$ with respect to $p(x,\hat{x})$ is high.

- Such an $\hat{\mathbf{X}}(i)$, if exists, would have $d(\mathbf{X}, \hat{\mathbf{X}}) \approx Ed(X, \hat{X}) \leq D$, because the joint relative frequency of $(\mathbf{x}, \hat{\mathbf{X}}(i)) \approx p(x, \hat{x})$.

- Use this $\hat{\mathbf{X}}(i)$ to represent $\mathbf{X}$ to satisfy the distortion constraint.

The Remaining Details

- For sufficiently large n, consider

$$\begin{aligned}\Pr\{d(\mathbf{X},\hat{\mathbf{X}}) > D+\epsilon\} &= \Pr\{d(\mathbf{X},\hat{\mathbf{X}}) > D+\epsilon | K=1\}\Pr\{K=1\} \\ &\quad +\Pr\{d(\mathbf{X},\hat{\mathbf{X}}) > D+\epsilon | K\neq 1\}\Pr\{K\neq 1\} \\ &\leq 1\cdot\epsilon + \Pr\{d(\mathbf{X},\hat{\mathbf{X}}) > D+\epsilon | K\neq 1\}\cdot 1 \\ &= \epsilon + \Pr\{d(\mathbf{X},\hat{\mathbf{X}}) > D+\epsilon | K\neq 1\}.\end{aligned}$$

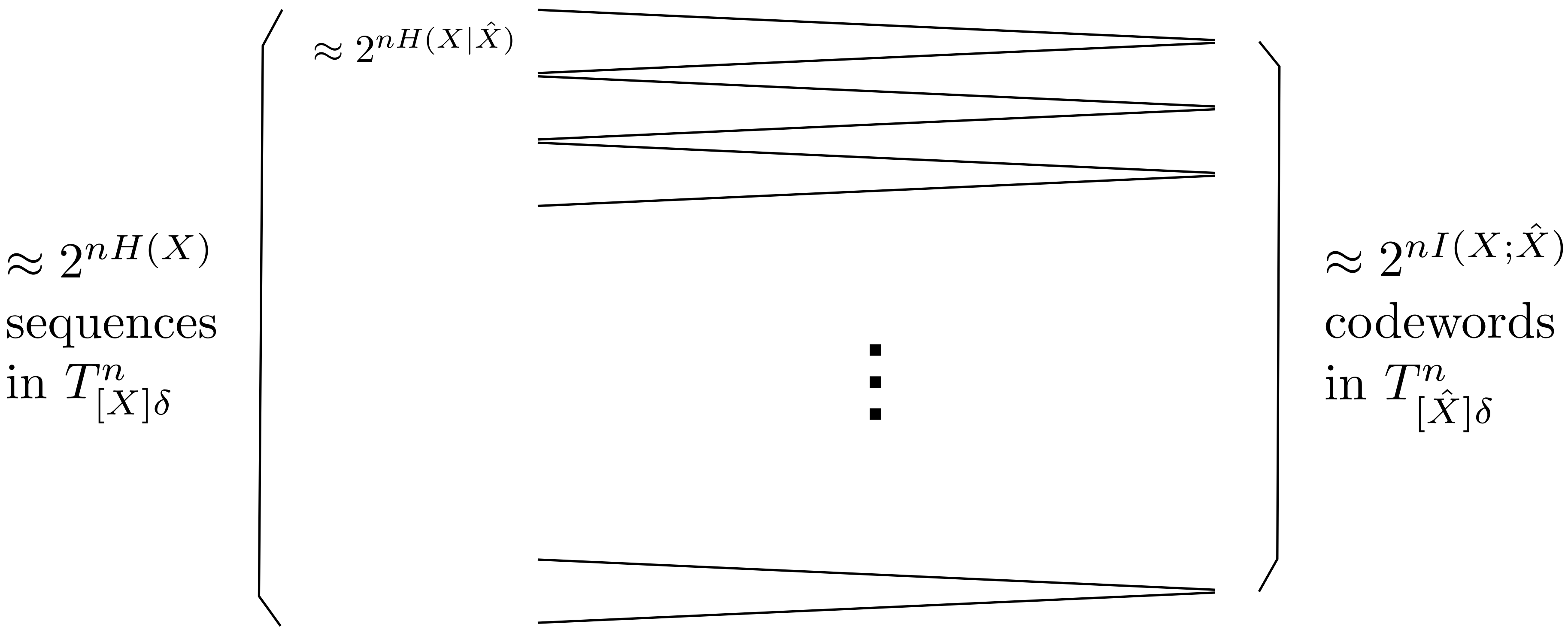
- Conditioning on $\{K \neq 1\}$, we have $(\mathbf{X},\hat{\mathbf{X}}) \in T^n_{[X\hat{X}]\delta}$.

- It can be shown that (see textbook)

$$d(\mathbf{X},\hat{\mathbf{X}}) \leq D + d_{max}\delta.$$

  By taking $\delta \leq \frac{\epsilon}{d_{max}}$, we obtain $d(\mathbf{X},\hat{\mathbf{X}}) \leq D+\epsilon$.

- Therefore, $\Pr\{d(\mathbf{X},\hat{\mathbf{X}}) > D+\epsilon | K\neq 1\} = 0$, which implies $\Pr\{d(\mathbf{X},\hat{\mathbf{X}}) > D+\epsilon\} \leq \epsilon$.



The number of codewords must be at least

$$\frac{2^{nH(X)}}{2^{nH(X|\hat{X})}} \approx 2^{nI(X;\hat{X})}$$