Source Representation: Rate Distortion Characterization 4

This lecture and the next one are concerned with the representation of the message emitted by an information source, as illustrated in Figure 1.1 of Lecture 1. Suppose that the message emitted by an information source is among a finite set $\mathcal{S} = \{a_1, a_2, \ldots, a_{|\mathcal{S}|}\}$. Then, if one wishes to represent the message using a string of binary digits, i.e., bits, in a unambiguous way, clearly the length of such a string should be at least $\lceil \log_2 |\mathcal{S}| \rceil$.

But can we use a shorter string for this task? The answer would be affirmative, if we allow some ambiguity in the representation. For example, consider $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7, 8\}$, and suppose that the destination who wants to reproduce the message does not care as long as the reproduced message does not differ from the source message by more than one. Then, we may safely represent any of $1, 2, 3$ by 2, any of $4, 5, 6$ by 5, and any of $7, 8$ by 8. So a binary string of length $\lceil \log_2 3 \rceil = 2$ suffices, rather than $\lceil \log_2 8 \rceil = 3$ if no ambiguity is tolerated.

From the simple example above, we clearly see that there exists a tradeoff between the amount of resource (e.g., the length of binary string) for representing a source message and the quality (e.g., the degree of ambiguity) of the reproduced message. The tradeoff will become even more interesting if the probabilistic nature of source is taken into consideration. The rate-distortion theory characterizes the fundamental limit of the tradeoff, and is the subject of this lecture. Its extreme case, where the distortion is zero or almost zero, deserves special treatment, and will be investigated in the next lecture.

4.1 Problem Formulation

In the general communication system model in Figure 1.1 of Lecture 1, we model an information source as a probabilistic device generating a stochastic process $S_1, S_2, \ldots$. A message then corresponds to a segment of the stochastic process of a prescribed length n, i.e., $\underline{S} = [S_1, S_2, \ldots, S_n]$.

In this lecture, we focus on the case where the information source is a discrete memoryless source (DMS); that is, $S_1, S_2, \ldots$ are a sequence of i.i.d. random variables, each with pmf $P_S(s)$ and alphabet $\mathcal{S}$.

In order to represent a source message, we need to assign to each possible value of $\underline{S}$, $\underline{s} \in \mathcal{S}^n$, an index selected from a certain finite set, which, without loss of generality, may be fixed as $\{1, 2, \ldots, M_n\}$. This assignment is accomplished by a mapping:

$$f_n^{(s)} : \mathcal{S}^n \mapsto \{1, 2, \ldots, M_n\}, \tag{4.1}$$

which we call a source encoder. Here the subscript n is used to emphasize the dependency upon the length of the source message, and the superscript (s) is used to indicate that the mapping is for source coding, to be distinguished from channel coding in Lecture 6. Given $f_n^{(s)}$, the index $W = f_n^{(s)}(\underline{S})$ is then a random variable induced by the source message random vector $\underline{S}$.

Suppose that the index W is revealed to the destination. The destination then needs to reproduce the source message as $\underline{\hat{S}} = [\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_n]$, which is also a length-n random vector. But we allow $\hat{S}_i$, $i = 1, 2, \ldots, n$, to take values in an alphabet $\hat{\mathcal{S}}$ possibly different from the source alphabet $\mathcal{S}$, in general. This reproduction is accomplished by a mapping:

$$g_n^{(s)} : \{1, 2, \ldots, M_n\} \mapsto \hat{\mathcal{S}}^n, \tag{4.2}$$

which we call a source decoder. So in summary, we have the following Markov chain:

$$\underline{S} \leftrightarrow W = f_n^{(s)}(\underline{S}) \leftrightarrow \underline{\hat{S}} = g_n^{(s)}(W) = g_n^{(s)}(f_n^{(s)}(\underline{S})). \tag{4.3}$$

See Figure 4.1 for an illustration of the source encoding and decoding process.

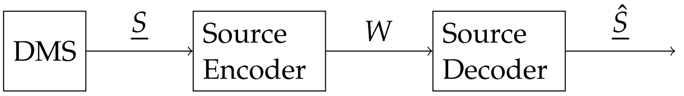

Figure 4.1: Illustration of source encoding and decoding.

Remark 4.1 The key feature of the problem formulation is the Markov chain relationship (4.3). Indeed, one may replace the deterministic mappings ($f_n^{(s)}$ and $g_n^{(s)}$) by some conditional probability distributions $P_{W|\underline{S}}$ and $P_{\underline{\hat{S}}|W}$ respectively, and the central result in this lecture, Shannon's fundamental theorem for source coding, in the next section, still holds.

Remark 4.2 Allowing the alphabet of reproduction $\hat{\mathcal{S}}$ to be different from that of source $\mathcal{S}$ may seem odd at first glance. But this considerably increases the applicability of the problem formulation, by enabling the destination to accomplish different tasks related to the source. For example, in an image classification task, the source is an image, and the destination is interested in deciding which category (e.g., animal, people, or landscape) the image belongs to. In this example, the source message (i.e., image) is an array of pixels, and the reproduced message is simply a label indicating the category of the source message.

A source encoder represents each length-n segment of source message as one of M_n indices, which can be stored as a binary string of length $\lceil \log_2 M_n \rceil$. On average, each source symbol is thus represented by $\lceil \log_2 M_n \rceil / n$ bits. We clarify that the term "bit" here corresponds to a unit of storage in a equipment such as computer, and it should not be confused with the measure of information we have introduced in Lecture 2.

We thus define the rate of a source encoder/decoder pair as:

$$R = \frac{\lceil \log_2 M_n \rceil}{n} \quad \text{bits/source symbol.} \tag{4.4}$$

As we allow a certain degree of ambiguity in the reproduced message at the destination, we introduce the notion of distortion here. A distortion measure d is a mapping that assigns each pair $(s, \hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}$ a non-negative number $d(s, \hat{s})$, that quantifies how much the cost incurred by reproducing s as $\hat{s}$. In our lecture notes, unless otherwise specified, we consider bounded distortion measures; that is, $d_{\max} := \max_{(s,\hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}} d(s, \hat{s}) < \infty$.

Example 4.1 For Hamming distortion, we have $\mathcal{S} = \hat{\mathcal{S}}$, and let $d(s, \hat{s}) = 0$ if $s = \hat{s}$ and 1 otherwise. Note that for Hamming distortion, $\mathbf{E}[d(S, \hat{S})] = P(S \neq \hat{S})$.

For a source encoder/decoder pair $(f_n^{(s)}, g_n^{(s)})$, we further impose an additive structure on the distortion between $\underline{S}$ and $\underline{\hat{S}}$, as

$$d(\underline{s}, \underline{\hat{s}}) = \frac{1}{n} \sum_{i=1}^{n} d(s_i, \hat{s}_i); \tag{4.5}$$

that is, the distortion between a source message and its corresponding reproduced message is the average of the pairwise distortion between each source symbol and its corresponding reproduced symbol. For Hamming distortion in Example 4.1, $d(\underline{s}, \underline{\hat{s}})$ is then the fraction of "errors" in the reproduced message $\underline{\hat{s}}$.

It is certainly desirable to have a source encoder/decoder pair that has a low rate as well as a small distortion. Induced by the Markov chain $\underline{S} \leftrightarrow W \leftrightarrow \underline{\hat{S}}$, the distortion $d(\underline{S}, \underline{\hat{S}})$ is a random variable. This consideration leads to the following definition of an achievable rate-distortion pair.

Definition 4.1 A rate-distortion pair (R, D) is said to be achievable, if there exists a sequence of source encoder/decoder pairs,

$\{(f_n^{(s)}, g_n^{(s)})\}_{n=1,2,\ldots}$, that satisfy

$$\begin{aligned} R &= \frac{\lceil \log_2 M_n \rceil}{n} \quad \text{bits/source symbol}, && (4.6) \\ \lim_{n\to\infty} \mathbf{E}[d(\underline{S}, \hat{\underline{S}})] &\leq D. && (4.7) \end{aligned}$$

Taking the closure of the set* of achievable rate-distortion pairs in the first quadrant of the (R, D)-plane, we obtain the rate-distorion region.

Consequently, the boundary of the rate-distortion region leads to the concepts of rate-distortion function and its inverse, distortion-rate function. See Figure 4.2 for an illustration.

Definition 4.2 The rate-distortion function $R(D)$ is the infimum of rates such that (R, D) is in the rate-distortion region, for each given D; and the distortion-rate function $D(R)$ is the infimum of distortions such that (R, D) is in the rate-distortion region, for each given R.

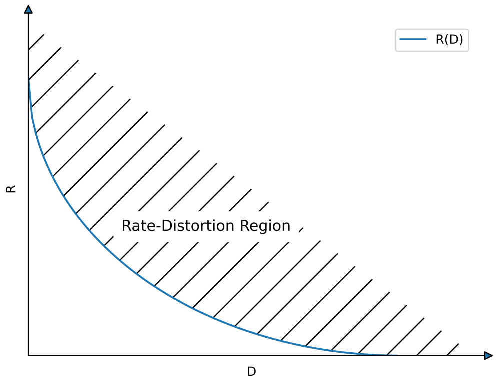

Figure 4.2: Illustration of rate-distortion region and $R(D)$ function.

It is important to note that the rate-distortion region is defined in an asymptotic sense: the source encoder/decoder pair works for arbitrarily large message size as $n \to \infty$. Resorting to asymptotic analysis is a key in information theory, and spectacles often occur in the asymptotic regime.

4.2 Shannon's Fundamental Theorem for Source Coding

In this section, we will introduce a fundamental result, first appearing in Shannon's article [16], that characterizes the rate-distortion function.

* For our purposes, it would suffice to understand the closure of a set as the union of the set and its boundary.

Let us define the information rate-distortion function as follows.

Definition 4.3 For a DMS S with pmf $P_S(s)$ and distortion measure $d(s, \hat{s})$, $(s, \hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}$, the solution of the following constrained optimization problem:

$$\min_{P_{\hat{S}|S}} I(S; \hat{S}), \tag{4.8}$$

$$\text{s.t.} \quad \mathbf{E}[d(S, \hat{S})] \leq D. \tag{4.9}$$

is called the information rate-distortion function $R_{\mathrm{I}}(D)$.

Shannon's fundamental theorem for source coding asserts that the information rate-distortion function is equal to the rate-distortion function.

Theorem 4.1 For a DMS S with pmf $P_S(s)$ and bounded distortion measure $d(s, \hat{s})$, $(s, \hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}$, we have

$$R(D) = R_{\mathrm{I}}(D). \tag{4.10}$$

Note that the rate-distortion function in Theorem 4.1 is not given in closed form. It involes an optimization problem (4.8) (4.9). The objective function is the mutual information $I(S; \hat{S})$ between the source S and its reproduction $\hat{S}$. Since $P_{S,\hat{S}}(s, \hat{s}) = P_S(s)P_{\hat{S}|S}(\hat{s}|s)$ and P_S is given, the minimization of $I(S; \hat{S})$ is conducted with respect to $P_{\hat{S}|S}$. The constraint $\mathbf{E}[d(S, \hat{S})] \leq D$ places restrictions on feasible choices of $P_{\hat{S}|S}$.

According to Definition 4.2, the rate-distortion function $R(D)$ has a coding interpretation as the infimum of rates achieved by a sequence of encoder/decoder pairs, subject to an expected distortion constraint D, over a sufficiently long block of source message. On the other hand, the information rate-distortion function $R_{\mathrm{I}}(D)$ given by Definition 4.3 is only a mathematical object, a function depending upon P_S and d, without any coding interpretation. It is Theorem 4.1 that links up these two, asserting that in fact they are equal.

With Theorem 4.1, when we are asked for the rate-distortion function, we can turn to evaluating the information rate-distortion function.

The proof of Theorem 4.1 will be provided in the next two sections. In the remaining part of this section, we establish some useful properties of the information rate-distortion function $R_{\mathrm{I}}(D)$, and provide some simple illustrative examples for its calculation.

First, since the left side of constraint (4.9) satisfies for any $P_{\hat{S}|S}$,

$$\mathbf{E}[d(S,\hat{S})] \geq D_{\min} := \sum_{s\in\mathcal{S}} P_S(s) \min_{\hat{s}\in\hat{\mathcal{S}}} d(s,\hat{s}), \tag{4.11}$$

the optimization problem (4.8) is infeasible for $D < D_{\min}$ and we may simply define $R_{\mathrm{I}}(D) = \infty$ in that region.

Second, note that $I(S;\hat{S}) = 0$ if and only if S and $\hat{S}$ are independent; — see Corollary 3.3. So if there is a pmf $P_{\hat{S}}$ such that

$$\sum_{(s,\hat{s})\in\mathcal{S}\times\hat{\mathcal{S}}} P_S(s)P_{\hat{S}}(\hat{s})d(s,\hat{s}) \leq D \tag{4.12}$$

holds, we can simply let $P_{\hat{S}|S} = P_{\hat{S}}$ in the optimization problem (4.8), and get $R_{\mathrm{I}}(D) = 0$. Since it always holds that

$$\sum_{(s,\hat{s})\in\mathcal{S}\times\hat{\mathcal{S}}} P_S(s)P_{\hat{S}}(\hat{s})d(s,\hat{s}) \geq D_{\max} := \min_{\hat{s}\in\hat{\mathcal{S}}} \sum_{s\in\mathcal{S}} P_S(s)d(s,\hat{s}), \tag{4.13}$$

it is necessary to have $D \geq D_{\max}$ for $R_{\mathrm{I}}(D) = 0$. On the other hand, this condition is also sufficient. To see this, denote the value of $\hat{s}$ that attains $D_{\max}$ by $\hat{s}^*$, and let $\hat{S} = \hat{s}^*$ with probability one. We then have $I(S;\hat{S}) = 0$, and

$$\begin{aligned} \mathbf{E}[d(S,\hat{S})] &= \mathbf{E}[d(S,\hat{s}^*)] \\ &= \sum_{s\in\mathcal{S}} P_S(s)d(s,\hat{s}^*) = D_{\max}, \end{aligned} \tag{4.14}$$

for any P_S. Therefore, for any $D \geq D_{\max}$ this deterministic choice of $\hat{S} = \hat{s}^*$ is feasible, leading to $R_{\mathrm{I}}(D) = 0$. In summary, we have the following necessary and sufficient condition for $R_{\mathrm{I}}(D)$ to be zero.

Corollary 4.1 The information rate-distortion function $R_{\mathrm{I}}(D)$ is zero if and only if

$$D \geq D_{\max} = \min_{\hat{s}\in\hat{\mathcal{S}}} \sum_{s\in\mathcal{S}} P_S(s)d(s,\hat{s}). \tag{4.15}$$

Some further characteristics of $R_{\mathrm{I}}(D)$ are provided in the next corollary.

Corollary 4.2 The information rate-distortion function $R_{\mathrm{I}}(D)$ is a non-increasing convex function in D. If $R_{\mathrm{I}}(D_{\min}) > 0$, then $R_{\mathrm{I}}(D)$ is strictly decreasing for $D \in [D_{\min}, D_{\max}]$, and the inequality constraint (4.9) can be replaced by an equality constraint.

Proof: Consider any $D_0 < D_1$ no smaller than $D_{\min}$. Denote the set of $P_{\hat{S}|S}$ satisfying $\mathbf{E}[d(S,\hat{S})] \leq D$ as $\mathcal{P}_D$. Then it is obvious that $\mathcal{P}_{D_0} \subseteq \mathcal{P}_{D_1}$, and hence $R(D_0) \geq R(D_1)$. This proves that $R_{\mathrm{I}}(D)$ is non-increasing in D.

For convexity, we need to prove that for any $D_0 < D_1$ no smaller than $D_{\min}$, and any $0 \le \lambda \le 1$,

$$R_{\mathrm{I}}((1-\lambda)D_0 + \lambda D_1) \le (1-\lambda)R_{\mathrm{I}}(D_0) + \lambda R_{\mathrm{I}}(D_1) \tag{4.16}$$

holds.

Denote the conditional probability distribution that achieves $R_{\mathrm{I}}(D_i)$ by $P^{(i)}_{\hat{S}|S}$, $i = 0, 1$. Define

$$P^{(\lambda)}_{\hat{S}|S}(\hat{s}|s) = (1-\lambda)P^{(0)}_{\hat{S}|S}(\hat{s}|s) + \lambda P^{(1)}_{\hat{S}|S}(\hat{s}|s). \tag{4.17}$$

Note that $P^{(\lambda)}_{\hat{S}|S}$ is feasible for the optimization problem of solving $R_{\mathrm{I}}((1-\lambda)D_0 + \lambda D_1)$, because under $P^{(\lambda)}_{\hat{S}|S}$,

$$\begin{aligned}\mathbf{E}[d(S,\hat{S})] &= (1-\lambda)\sum_{(s,\hat{s})\in\mathcal{S}\times\hat{\mathcal{S}}} P_S(s)P^{(0)}_{\hat{S}|S}(\hat{s}|s)d(s,\hat{s}) \\ &\quad +\lambda \sum_{(s,\hat{s})\in\mathcal{S}\times\hat{\mathcal{S}}} P_S(s)P^{(1)}_{\hat{S}|S}(\hat{s}|s)d(s,\hat{s}) \\ &\le (1-\lambda)D_0 + \lambda D_1,\end{aligned} \tag{4.18}$$

where the inequality is from the assumption that $P^{(0)}_{\hat{S}|S}$ and $P^{(1)}_{\hat{S}|S}$ achieve $R_{\mathrm{I}}(D_0)$ and $R_{\mathrm{I}}(D_1)$, respectively. Hence according to the constrained optimization problem formulation (4.8) (4.9), we have

$$R_{\mathrm{I}}((1-\lambda)D_0 + \lambda D_1) \le I^{(\lambda)}(S;\hat{S}), \tag{4.19}$$

the mutual information achieved by letting the conditional probability distribution $P_{\hat{S}|S}$ be $P^{(\lambda)}_{\hat{S}|S}$.

On the other hand, according to Theorem 3.7, $I(S;\hat{S})$ is convex in $P_{\hat{S}|S}$ for fixed P_S. So under $P^{(\lambda)}_{\hat{S}|S}$, we have

$$\begin{aligned} I^{(\lambda)}(S;\hat{S}) &\le (1-\lambda)I^{(0)}(S;\hat{S}) + \lambda I^{(1)}(S;\hat{S}) \\ &= (1-\lambda)R_{\mathrm{I}}(D_0) + \lambda R_{\mathrm{I}}(D_1). \end{aligned} \tag{4.20}$$

Putting together (4.19) and (4.20), we obtain (4.16) and prove the convexity of $R_{\mathrm{I}}(D)$.

We then assume $R_{\mathrm{I}}(D_{\min}) > 0$ and prove that $R_{\mathrm{I}}(D)$ is strictly decreasing over $[D_{\min}, D_{\max}]$. Suppose that this is not the case. Since we have already proved that $R_{\mathrm{I}}(D)$ is non-increasing, the only possibility is that there exist some "plateaus", i.e., $D_{\min} \le D_0 < D_1 \le D_{\max}$ with $R_{\mathrm{I}}(D) = R_{\mathrm{I}}(D_0), \forall D \in [D_0, D_1]$. See Figure 4.3 for illustration. Since we have already proved that $R_{\mathrm{I}}(D)$ is convex, by rewriting

$$D_1 = \frac{D_{\max} - D_1}{D_{\max} - D_0} D_0 + \frac{D_1 - D_0}{D_{\max} - D_0} D_{\max}, \tag{4.21}$$

we have

$$R_{\mathrm{I}}(D_1) \leq \frac{D_{\max} - D_1}{D_{\max} - D_0} R_{\mathrm{I}}(D_0) + \frac{D_1 - D_0}{D_{\max} - D_0} R_{\mathrm{I}}(D_{\max}). \tag{4.22}$$

Noting that $R_{\mathrm{I}}(D_1) = R_{\mathrm{I}}(D_0)$ holds by our assumption, and $R_{\mathrm{I}}(D) = 0, \forall D \geq D_{\max}$, we then arrive at the only possibility of $R_{\mathrm{I}}(D_0) = 0$, and consequently $R_{\mathrm{I}}(D) = 0, \forall D \geq D_0$. But according to Corollary 4.1, for $R_{\mathrm{I}}(D) = 0$ it is necessary to have $D \geq D_{\max}$. This hence leads to a contradiction, and $R_{\mathrm{I}}(D)$ has to be strictly decreasing over $[D_{\min}, D_{\max}]$.

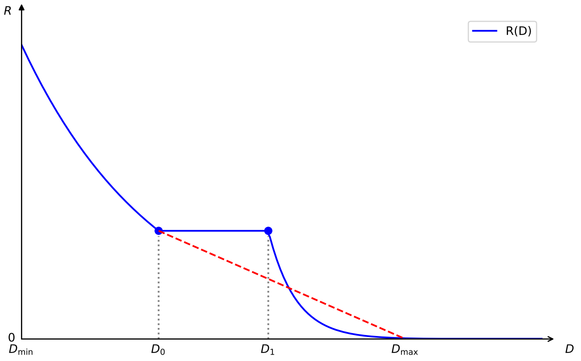

Figure 4.3: Illustration that $R_{\mathrm{I}}(D)$ is strictly decreasing over $[D_{\min}, D_{\max}]$.

To prove our last claim that the inequality constraint (4.9) can be replaced by an equality constraint, let us suppose that there exists some $D \in [D_{\min}, D_{\max}]$ such that when achieving $R_{\mathrm{I}}(D)$, the corresponding $P_{\hat{S}|S}$ attains a strict inequality $\mathbf{E}[d(S, \hat{S})] = D' < D$. This then leads to $R_{\mathrm{I}}(D') = R_{\mathrm{I}}(D)$, in contradiction with the strictly decreasing property of $R_{\mathrm{I}}(D)$ over $[D_{\min}, D_{\max}]$ we have just proved. □

Example 4.2 In this example, we calculate the information rate-distortion function of a source S obeying Bernoulli distribution with parameter δ, with Hamming distortion.

Let us start with finding $D_{\min}$ and $D_{\max}$. It is obvious that $D_{\min} = 0$. According to the definition $D_{\max} = \min_{\hat{s} \in \hat{\mathcal{S}}} \sum_{s \in \mathcal{S}} P_S(s) d(s, \hat{s})$, we have

$$D_{\max} = \min_{\hat{s} \in \{0,1\}} [(1 - \delta) d(0, \hat{s}) + \delta d(1, \hat{s})] = \min\{\delta, 1 - \delta\}. \tag{4.23}$$

To fix ideas, assume $\delta \leq 1/2$. Hence $D_{\max} = \delta$ and $R_{\mathrm{I}}(D) = 0$, $\forall D \geq \delta$. The other case of $\delta > 1/2$ can be treated similarly and is left to the readers.

Now let us investigate the constrained optimization problem (4.8) (4.9) for $D \in [0, \delta]$. For this, we expand $I(S; \hat{S})$ according to

$$I(S; \hat{S}) = H(S) - H(S|\hat{S}), \tag{4.24}$$

in which $H(S) = h_2(\delta)$ is given, and

$$\begin{aligned} H(S|\hat{S}) &\overset{(a)}{=} H(S \oplus \hat{S}|\hat{S}) \\ &\overset{(b)}{\leq} H(S \oplus \hat{S}), \end{aligned} \tag{4.25}$$

where (a) is due to Corollary 3.6 because conditioned upon $\hat{S}$, there is a bijection between S and $S \oplus \hat{S}$, and (b) is due to Theorem 3.6, "conditioning reduces entropy".

But what is $S \oplus \hat{S}$? In fact, $\{S \oplus \hat{S} = 1\}$ is equivalent to $\{S \neq \hat{S}\}$, and is further equivalent to $\{d(S, \hat{S}) = 1\}$ under the Hamming distortion. So the constraint (4.9) (which is an equality constraint now, due to Corollary 4.2) is equivalent to $P(S \oplus \hat{S} = 1) = D$. Consequently, we obtain a lower bound on $I(S; \hat{S})$ as

$$I(S; \hat{S}) \geq h_2(\delta) - h_2(D), \tag{4.26}$$

which holds for any $P_{\hat{S}|S}$.

If we can find some $P_{\hat{S}|S}$ to achieve the lower bound (4.26), then this lower bound would be exactly $R_{\mathrm{I}}(D)$. Inspecting the derivation steps thus far, we see that the sought-after $P_{\hat{S}|S}$ should achieve equality in (b) of (4.25); that is, it should make $\hat{S}$ and $S \oplus \hat{S}$ independent. Inspired by the fact that $S = \hat{S} \oplus (S \oplus \hat{S})$, we view $S \oplus \hat{S}$ as an auxiliary random variable Z which is independent of $\hat{S}$, and is Bernoulli with parameter D, according to the discussion in the previous paragraph. Note that this specifies the "backward" conditional probability distribution $P_{S|\hat{S}}$, instead of the "forward" conditional probability distribution $P_{\hat{S}|S}$ needed in the constrained optimization problem (4.8) (4.9). Since $P_S P_{\hat{S}|S} = P_{\hat{S}} P_{S|\hat{S}}$, in order to specify $P_{\hat{S}|S}$, we still need to find the corresponding $P_{\hat{S}}$. With some calculation, it is not difficult to verify that such $P_{\hat{S}}$ exists, and is given by a Bernoulli distribution with parameter $(\delta - D)/(1 - 2D)$. So the lower bound (4.26) is achievable, and the information rate-distortion function for a Bernoulli source with parameter $\delta \leq 1/2$ and with Hamming distortion is given by

$$R_{\mathrm{I}}(D) = h_2(\delta) - h_2(D) \quad \text{if } 0 \leq D \leq \delta, \text{ and 0 otherwise}, \tag{4.27}$$

as illustrated in Figure 4.4.

In Example 4.2, the procedure of first specifying $P_{S|\hat{S}}$ and then finding a corresponding $P_{\hat{S}}$ to achieve the information rate-distortion function is a useful technique. The thus constructed $P_{S|\hat{S}}$ is usually

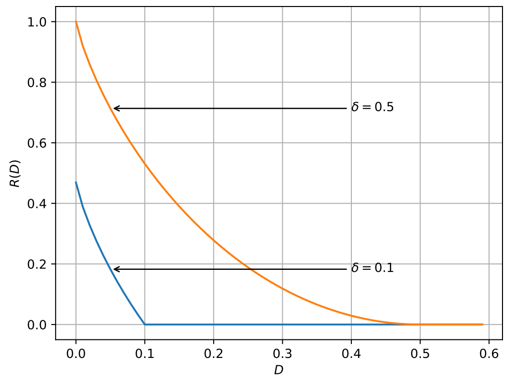

Figure 4.4: Rate-distortion function of Bernoulli DMS with Hamming distortion.

called the "test channel". In retrospect, since we have started with $I(S;\hat{S}) = H(S) - H(S|\hat{S})$, it is natural to consider the construction of suitable $P_{S|\hat{S}}$ and $P_{\hat{S}}$ so as to maximize $H(S|\hat{S})$, subject to the distortion constraint and the constraint on the probability distribution of S.

It should be realized that, except for very few P_S and d with highly symmetric structures (such as Example 4.2 and some of the exercises in this lecture), an explicit construction of test channel is difficult, and we cannot expect the information rate-distortion function to yield a closed-form solution in general. The next example illustrates this, by only slightly modifying the setup in Example 4.2.

Example 4.3 We consider a source S obeying Bernoulli distribution with parameter $1/2$, with the following asymmetric distortion measure:

$$d(0,0) = d(1,1) = 0,\ d(0,1) = 1,\ \text{and}\ d(1,0) = A > 1. \quad (4.28)$$

For this setup, $D_{\min} = 0$, and $D_{\max} = 1/2$. We then focus on $D \in [0, 1/2]$. According to Corollary 4.2, the constraint (4.9) is an equality as

$$\mathbf{E}[d(S,\hat{S})] = \frac{1}{2} P_{\hat{S}|S}(1|0) + \frac{A}{2} P_{\hat{S}|S}(0|1) = D. \quad (4.29)$$

We then let $\alpha = P_{\hat{S}|S}(1|0) \in [0,1]$ and $\beta = P_{\hat{S}|S}(0|1) \in [0,1]$. With them, we can evaluate $I(S;\hat{S})$ to get

$$\begin{aligned} I(S;\hat{S}) \quad &= \quad H(S) - H(S|\hat{S}) = \log 2 - \left[\frac{1-\alpha+\beta}{2} h_2\left(\frac{\beta}{1-\alpha+\beta} \right) \right. \\ &\qquad \left. + \frac{1+\alpha-\beta}{2} h_2\left(\frac{\alpha}{1+\alpha-\beta} \right) \right], \end{aligned} \quad (4.30)$$

which should be minimized over $(\alpha, \beta) \in [0,1]^2$ subject to $\alpha + A\beta = 2D$. Unfortunately, this optimization problem does not appear to possess a closed-form solution, and can only be tackled numerically. Figure 4.5 depicts $R_{\mathrm{I}}(D)$ for $A = 1.5$ and 3. It also includes (4.27) which corresponds to $A = 1$ for reference.

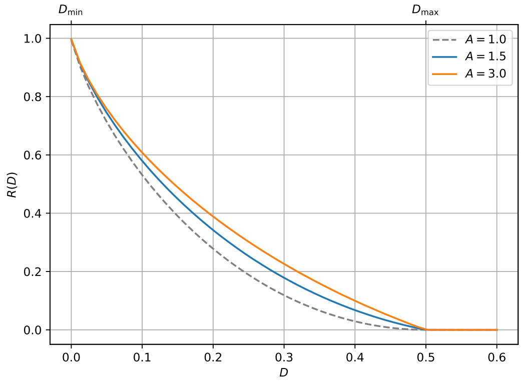

Figure 4.5: Rate-distortion function of Bernoulli(1/2) DMS with asymmetric distortion (4.28).

Nevertheless, there exist efficient numerical algorithms for computing the information rate-distortion function, and we will introduce them in Lecture 9.

4.3 Proof of the Converse Part

In this section, we establish the converse part of Theorem 4.1, i.e., $R(D) \geq R_{\mathrm{I}}(D)$. For this, we will show that, for any source encoder/decoder pair, it is impossible to achieve expected distortion no greater than D for any $R < R_{\mathrm{I}}(D)$.

Let us fix an arbitrary pair of source encoder and decoder $(f_n^{(s)}, g_n^{(s)})$, with source message length n. Since the source message $\underline{S}$ is a random vector, $(f_n^{(s)}, g_n^{(s)})$ induces a joint probability distribution on $(\underline{S}, \hat{\underline{S}})$. Suppose that the resulting expected distortion satisfies $\mathbf{E}[d(\underline{S}, \hat{\underline{S}})] \leq D$. We now inspect the rate of such a source encoder/decoder pair.

We proceed as follows:

$$\begin{aligned} nR &\overset{(a)}{\geq} H(W) \\ &\geq I(\underline{S}; W) \\ &\overset{(b)}{\geq} I(\underline{S}; \hat{\underline{S}}) \\ &= H(\underline{S}) - H(\underline{S}|\hat{\underline{S}}) \\ &\overset{(c)}{=} \sum_{i=1}^{n} H(S_i) - \sum_{i=1}^{n} H(S_i|\hat{\underline{S}}, S_{i-1}, \dots, S_1) \\ &\overset{(d)}{\geq} \sum_{i=1}^{n} H(S_i) - \sum_{i=1}^{n} H(S_i|\hat{S}_i) \\ &= \sum_{i=1}^{n} I(S_i; \hat{S}_i), \end{aligned} \tag{4.31}$$

where, (a) is because W is one of M_n indices according to (4.1), and thus due to Corollary 3.1 its entropy is at most $\log_2 M_n \leq nR$ bits, according to Definition 4.1, (b) is due to Theorem 3.5, the DPI, (c) is due to Theorem 3.1, the chain rule of entropy, and (d) is because conditioning reduces entropy (see Theorem 3.6).

In order to proceed further, let us introduce an auxiliary "time-sharing" random variable Q, which is uniform over $\{1, 2, \ldots, n\}$ and is independent of $\underline{S}$. We can hence continue the preceding bounding steps as

$$\begin{aligned} R &\geq \frac{1}{n}\sum_{i=1}^{n} I(S_i; \hat{S}_i) \\ &\overset{(e)}{=} I(S_Q; \hat{S}_Q|Q) \\ &= H(S_Q|Q) - H(S_Q|\hat{S}_Q, Q), \end{aligned} \tag{4.32}$$

where (e) is exactly the definition of conditional mutual information (see Definition 2.15), noting that $P_Q(i) = 1/n, \forall 1 \leq i \leq n$. Since the probability distribution of the DMS S does not depend on the subscript Q, we have $H(S_Q|Q) = H(S_Q)$. On the other hand, $H(S_Q|\hat{S}_Q, Q) \leq H(S_Q|\hat{S}_Q)$ because conditioning reduces entropy. So we arrive at

$$R \geq H(S_Q) - H(S_Q|\hat{S}_Q) = I(S_Q; \hat{S}_Q). \tag{4.33}$$

Now, noting that the probability distribution of S_Q is nothing but P_S, according to the constrained optimization problem formulation of $R_{\mathrm{I}}(D)$ (4.8) (4.9), we can go one step further to obtain

$$\begin{aligned} R &\geq I(S_Q; \hat{S}_Q) \\ &\geq R_{\mathrm{I}}(\mathbf{E}[d(S_Q, \hat{S}_Q)]). \end{aligned} \tag{4.34}$$

Since Q is uniform over $\{1, 2, \ldots, n\}$, using the law of total expectation (see Theorem 2.3), we have

$$\begin{aligned} \mathbf{E}[d(S_Q, \hat{S}_Q)] &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[d(S_i, \hat{S}_i)] \\ &\overset{(f)}{=} \mathbf{E}[d(\underline{S}, \underline{\hat{S}})] \overset{(g)}{\leq} D, \end{aligned} \tag{4.35}$$

where (f) is due to the additive structure of distortion, i.e., (4.5), and the inequality (g) follows from the assumption regarding $(f_n^{(s)}, g_n^{(s)})$ made at the beginning of the proof. We now conclude that

$$\begin{aligned} R &\geq R_{\mathrm{I}}(\mathbf{E}[d(S_Q, \hat{S}_Q)]) \\ &\geq R_{\mathrm{I}}(D), \end{aligned} \tag{4.36}$$

by using the non-increasing property of $R_{\mathrm{I}}(D)$ (see Corollary 4.2). This completes the proof of the converse part.

Remark 4.3 Inspecting the proof, we see that the only requirement regarding encoding and decoding is the Markov chain $\underline{S} \leftrightarrow W \leftrightarrow \hat{\underline{S}}$. We do not require W to be deterministic given $\underline{S}$ or $\hat{\underline{S}}$ to be deterministic given W. That is, the converse still holds if the source encoder and decoder are stochastic.

4.4 Proof of the Achievability Part

In this section, we establish the achievability part of Theorem 4.1, i.e., $R(D) \le R_{\mathrm{I}}(D)$. For this, we will show that, given D, for any rate $R > R_{\mathrm{I}}(D)$, there exists a sequence of source encoder/decoder pairs, indexed by the source message length $n = 1, 2, \ldots$, such that (R, D) is an achievable rate-distortion pair, in the sense of Definition 4.1.

Before we start, let us make the concept of source encoder/decoder pair more concrete. Note that the index W representing $\underline{S}$ is selected as $f_n^{(s)}(\underline{S})$ among M_n indices, and that each $W = w$ corresponds to a reproduced message vector $g_n^{(s)}(w)$. We call $\{g_n^{(s)}(w), w = 1, \ldots, M_n\}$ the codebook, and denote it by C. Each item of C is called a codeword, and the w-th codeword, i.e., $g_n^{(s)}(w)$, is also denoted by $\mathsf{C}(w)$, whose i-th element is further denoted by $\mathsf{C}_i(w)$, for $i = 1, \ldots, n$.

The encoder $f_n^{(s)}$ specifies a rule that assigns each possible source message vector to a codeword in C, and the decoder $g_n^{(s)}$ simply outputs this assigned codeword as the reproduced message. The dimension parameters of C, n and M_n, determine the rate R as $\lceil \log M_n \rceil / n$; see (4.4). Regarding the expected distortion, we have

$$\mathbf{E}[d(\underline{S}, \hat{\underline{S}})] = \sum_{\underline{s} \in \mathcal{S}^n} P_{\underline{S}}(\underline{s}) d(\underline{s}, \mathsf{C}(f_n^{(s)}(\underline{s}))). \tag{4.37}$$

At this point, very little can be said about $\mathbf{E}[d(\underline{S}, \hat{\underline{S}})]$ for a specific $(f_n^{(s)}, g_n^{(s)})$ pair, in general. Two key ideas make it possible to conduct further analysis. First, let the codebook itself be randomly generated; second, when analyzing the expected distortion, let the expectation be taken with respect to both source message and codebook.

4.4.1 Generation of Codebook

We fix some $P_{\hat{S}|S}$ satisfying $\mathbf{E}[d(S,\hat{S})] \leq D$, and fix a rate $R > I(S;\hat{S})$ which is evaluated using $P_S P_{\hat{S}|S}$. The marginal probability distribution of $\hat{S}$ is $P_{\hat{S}}(\hat{s}) = \sum_{s \in \mathcal{S}} P_S(s) P_{\hat{S}|S}(\hat{s}|s)$. As said, we let the codebook be randomly generated. Specifically, we let all the elements of all the codewords be i.i.d. random variables obeying $P_{\hat{S}}$. We denote such a random codebook by $\mathbf{C}$, to be distinguished from its realization C.

Note that such a random codebook $\mathbf{C}$ is independent of the source message $\underline{S}$. Once $\mathbf{C}$ is generated, it is revealed to both the encoder and the decoder.

Consider $\hat{\mathcal{S}}^n$ as a space, in which each codeword is a point. The generation of $\mathbf{C}$ corresponds to sampling from $\hat{\mathcal{S}}^n$ M_n times, independently with replacement, according to $\prod_{i=1}^n P_{\hat{S}_i}$ where $\{\hat{S}_i, i = 1, \ldots, n\}$ are i.i.d. obeying $P_{\hat{S}}$.

Every codeword is responsible for reproducing some points in the source space $\mathcal{S}^n$, and every point in $\mathcal{S}^n$ needs to be reproduced by a codeword. So the situation is that the space of $\mathcal{S}^n$ should be partitioned into M_n non-overlap subsets, and that each such subset should be associated with a different codeword in a given codebook $\mathsf{C} \subseteq \hat{\mathcal{S}}^n$. It is heuristically clear that, if the size M_n of codebook is small and $\hat{\mathcal{S}}^n$ is only sparsely sampled, each codeword needs to reproduce points in a "large" subset of $\mathcal{S}^n$, leading to a large distortion. Therefore, the codebook should contain sufficiently many codewords, and this consideration translates into the requirement of a minimum rate R.

4.4.2 Encoding

Given a codebook C, how should we encode a source message $\underline{s}$? Clearly, the optimal way of encoding should assign to $\underline{s}$ the codeword that minimizes the distortion between $\underline{s}$ and the codeword; that is, choosing

$$w^* = \arg \min_{w=1,\ldots,M_n} d(\underline{s}, \mathsf{C}(w)), \tag{4.38}$$

and letting $\underline{\hat{s}} = \mathsf{C}(w^*)$. This is essentially a "clustering" operation that clusters the points in $\mathcal{S}^n$ around their respective "nearest" codewords.

The performance of the optimal encoding procedure, however, is somewhat inconvenient to analyze. Therefore we turn to another encoding procedure, which is not optimal in general, but is still sufficient to accomplish our achievability proof. To motivate the encoding procedure, let us conduct the following thought experiment. If $\underline{S}$ and $\mathbf{C}(W)$ were jointly distributed according to $P_{\underline{S},\underline{\hat{S}}} = \prod_{i=1}^{n} P_{S_i,\hat{S}_i}$ where $\{(S_i, \hat{S}_i), i = 1, \ldots, n\}$ are i.i.d. obeying $P_{S,\hat{S}}$, then from the WLLN (see Theorem 2.4), as n grows without bound, with high probability $d(\underline{S}, \mathbf{C}(W))$ would be close to its expectation $\mathbf{E}[d(S, \hat{S})]$ which has been fixed to be no greater than D at the beginning of our proof. However, this is not the reality, because the codebook $\mathbf{C}$ is generated independently of $\underline{S}$. Hence we would like the encoder to choose (whenever possible) some codeword such that the pair of source message and chosen codeword behave as if they indeed obey the joint probability distribution $P_{\underline{S},\underline{\hat{S}}}$.

Motivated by the key insight revealed in the preceding discussion, we devise the following encoding procedure: under $\epsilon > 0$, given $\underline{S} = \underline{s}$, find the smallest index $w \in \{1, 2, \ldots, M_n\}$ such that

$$\begin{aligned} d(\underline{s}, \mathsf{C}(w)) &\leq D + \epsilon, &(4.39)\\ I(S; \hat{S}) - \epsilon \leq i(\underline{s}; \mathsf{C}(w)) &\leq I(S; \hat{S}) + \epsilon, &(4.40) \end{aligned}$$

where

$$\begin{aligned} d(\underline{s}, \mathsf{C}(w)) &= \frac{1}{n}\sum_{i=1}^{n} d(s_i, \mathsf{C}_i(w)), &(4.41)\\ i(\underline{s}; \mathsf{C}(w)) &= \frac{1}{n}\log_2 \frac{P_{\underline{S},\underline{\hat{S}}}(\underline{s}, \mathsf{C}(w))}{P_{\underline{S}}(\underline{s})P_{\underline{\hat{S}}}(\mathsf{C}(w))}\\ &= \frac{1}{n}\sum_{i=1}^{n}\log_2 \frac{P_{S,\hat{S}}(s_i, \mathsf{C}_i(w))}{P_S(s_i)P_{\hat{S}}(\mathsf{C}_i(w))}. &(4.42) \end{aligned}$$

If such an index w exists, we denote it by w^* and use the corresponding codeword $\mathsf{C}(w^*)$ to reproduce $\underline{s}$; otherwise, the encoding procedure fails, and we simply set $w^* = 1$ and use $\mathsf{C}(1)$ to reproduce $\underline{s}$.

The criterion (4.39) is easy to understand. But why do we need the other criterion (4.40)? As we have just discussed in our thought experiment, we would like the encoder to choose (whenever possible) some codeword such that the pair of source message and chosen codeword behave as if they obey the joint probability distribution $P_{\underline{S},\underline{\hat{S}}}$. The criterion (4.40) effectively enforces this, in light of the WLLN. As will be seen in the sequel, the criterion (4.40) facilitates a crucial step in the analysis of expected distortion.

For notational convenience, for a given C, we associate with each $\underline{s} \in \mathcal{S}^n$ an indicator function $E(\mathsf{C}, \underline{s})$ to indicate whether an encoding failure occurs; that is, $E(\mathsf{C}, \underline{s}) = 0$ if there exists some index w to satisfy (4.39) and (4.40) simultaneously, and $E(\mathsf{C}, \underline{s}) = 1$ otherwise.

4.4.3 Analysis of Expected Distortion

For a given codebook C, the expected distortion $\mathbf{E}[d(\underline{S}, \hat{\underline{S}})]$ is still difficult to assess. As said, besides generating the codebook at random, a closely related key idea is studying the expected distortion over the ensemble of codebooks.

Now let us inspect the expected distortion, taken over both $\underline{S}$ and $\mathbf{C}$:

$$\begin{aligned} \mathbf{E}_{\mathbf{C},\underline{S}}[d(\underline{S}, \hat{\underline{S}})] &= \sum_{\mathsf{C},\underline{s}\in\mathcal{S}^n} P_{\mathbf{C}}(\mathsf{C})P_{\underline{S}}(\underline{s})d(\underline{s}, \mathsf{C}(w)) \\ &= \sum_{\mathsf{C},\underline{s}\in\mathcal{S}^n:E(\mathsf{C},\underline{s})=0} P_{\mathbf{C}}(\mathsf{C})P_{\underline{S}}(\underline{s})d(\underline{s}, \mathsf{C}(w)) \\ &+ \sum_{\mathsf{C},\underline{s}\in\mathcal{S}^n:E(\mathsf{C},\underline{s})=1} P_{\mathbf{C}}(\mathsf{C})P_{\underline{S}}(\underline{s})d(\underline{s}, \mathsf{C}(1)).(4.43) \end{aligned}$$

For the first summation in (4.43), since (4.39) is satisfied for each term, $d(\underline{s}, \mathsf{C}(w)) \leq D + \epsilon$ holds, and hence the summation is also upper bounded by $D + \epsilon$. For the second summation in (4.43), we can use $d_{\max}$ to upper bound $d(\underline{s}, \mathsf{C}(1))$ anyway, and hence the summation is upper bounded by $P_{\mathrm{f}} d_{\max}$, where

$$P_{\mathrm{f}} = \sum_{\mathsf{C},\underline{s}\in\mathcal{S}^n:E(\mathsf{C},\underline{s})=1} P_{\mathbf{C}}(\mathsf{C})P_{\underline{S}}(\underline{s}) \qquad (4.44)$$

is the probability of encoding failure.

4.4.4 Estimation of Probability of Encoding Failure

Now let us estimate P_{f}, which can be rewritten as

$$\begin{aligned} P_{\mathrm{f}} &= \sum_{\mathsf{C},\underline{s}\in\mathcal{S}^n} P_{\mathbf{C}}(\mathsf{C})P_{\underline{S}}(\underline{s})E(\mathsf{C}, \underline{s}) \\ &= \sum_{\underline{s}\in\mathcal{S}^n} P_{\underline{S}}(\underline{s}) \sum_{\mathsf{C}} P_{\mathbf{C}}(\mathsf{C})E(\mathsf{C}, \underline{s}). \end{aligned} \qquad (4.45)$$

A key step then is to see that, for any $\underline{s} \in \mathcal{S}^n$,

$$\sum_{\mathsf{C}} P_{\mathbf{C}}(\mathsf{C})E(\mathsf{C}, \underline{s}) = \left[1 - P((\underline{s}, \hat{\underline{S}}) \text{ satisfies } (4.39)(4.40))\right]^{M_n}. (4.46)$$

This is because $E(\mathbf{C}, \underline{s}) = 1$ means that none of the M_n codewords of $\mathbf{C}$ satisfies (4.39) (4.40) simultaneously, and these codewords are i.i.d. with probability distribution $P_{\hat{\underline{S}}}$. At this point we see the benefit of taking expectation with respect to $\mathbf{C}$: the evaluation of P_{f} now involves $P_{\underline{S},\hat{\underline{S}}}$ only, without needing to consider any specific codebook.

To proceed further, let us introduce another indicator function $\tilde{E}(\underline{s}, \underline{\hat{s}})$ as follows: $\tilde{E}(\underline{s}, \underline{\hat{s}}) = 1$ if $(\underline{s}, \underline{\hat{s}})$ satisfies (4.39) and (4.40) simultaneously, and $\tilde{E}(\underline{s}, \underline{\hat{s}}) = 0$ otherwise. So

$$P((\underline{s}, \underline{\hat{S}}) \text{ satisfies (4.39)(4.40)}) \quad = \quad \sum_{\underline{\hat{s}} \in \hat{\mathcal{S}}^n} P_{\underline{\hat{S}}}(\underline{\hat{s}}) \tilde{E}(\underline{s}, \underline{\hat{s}}). \quad (4.47)$$

Now the criterion (4.40) comes into play, from which we have $i(\underline{s}; \underline{\hat{s}}) \leq I(S; \hat{S}) + \epsilon$, which is equivalent to

$$P_{\underline{\hat{S}}}(\underline{\hat{s}}) \geq 2^{-n(I(S;\hat{S})+\epsilon)} P_{\underline{\hat{S}}|\underline{S}}(\underline{\hat{s}}|\underline{s}). \quad (4.48)$$

We hence obtain

$$\begin{aligned} & P((\underline{s}, \underline{\hat{S}}) \text{ satisfies (4.39)(4.40)}) \\ = \;& \sum_{\underline{\hat{s}} \in \hat{\mathcal{S}}^n} P_{\underline{\hat{S}}}(\underline{\hat{s}}) \tilde{E}(\underline{s}, \underline{\hat{s}}) \\ \geq \;& 2^{-n(I(S;\hat{S})+\epsilon)} \sum_{\underline{\hat{s}} \in \hat{\mathcal{S}}^n} P_{\underline{\hat{S}}|\underline{S}}(\underline{\hat{s}}|\underline{s}) \tilde{E}(\underline{s}, \underline{\hat{s}}). \end{aligned} \quad (4.49)$$

Putting (4.49) back to (4.45), we have

$$P_{\mathrm{f}} \leq \sum_{\underline{s} \in \mathcal{S}^n} P_{\underline{S}}(\underline{s}) \left[1 - 2^{-n(I(S;\hat{S})+\epsilon)} \sum_{\underline{\hat{s}} \in \hat{\mathcal{S}}^n} P_{\underline{\hat{S}}|\underline{S}}(\underline{\hat{s}}|\underline{s}) \tilde{E}(\underline{s}, \underline{\hat{s}}) \right]^{M_n}. \quad (4.50)$$

At this point, we apply the inequality $(1 - xy)^n \leq 1 - x + e^{-ny}$, $\forall 0 \leq x, y \leq 1$, to get

$$\begin{aligned} P_{\mathrm{f}} \;&\leq\; \sum_{\underline{s} \in \mathcal{S}^n} P_{\underline{S}}(\underline{s}) \left[1 - \sum_{\underline{\hat{s}} \in \hat{\mathcal{S}}^n} P_{\underline{\hat{S}}|\underline{S}}(\underline{\hat{s}}|\underline{s}) \tilde{E}(\underline{s}, \underline{\hat{s}}) + e^{-M_n 2^{-n(I(S;\hat{S})+\epsilon)}} \right] \\ &=\; 1 - \sum_{(\underline{s},\underline{\hat{s}}) \in \mathcal{S}^n \times \hat{\mathcal{S}}^n} P_{\underline{S},\underline{\hat{S}}}(\underline{s}, \underline{\hat{s}}) \tilde{E}(\underline{s}, \underline{\hat{s}}) + e^{-M_n 2^{-n(I(S;\hat{S})+\epsilon)}}. \end{aligned} \quad (4.51)$$

In (4.51), the second term is nothing but the expectation of $\tilde{E}(\underline{S}, \underline{\hat{S}})$ under $P_{\underline{S},\underline{\hat{S}}}$, or, the probability that $(\underline{S}, \underline{\hat{S}})$ under $P_{\underline{S},\underline{\hat{S}}}$ satisfies (4.39) and (4.40) simultaneously. Due to the WLLN (see Theorem 2.4), for any $\epsilon > 0$, this probability approaches one as $n \longrightarrow \infty$.

In (4.51), the third term is controlled by the asymptotic behavior of $M_n 2^{-n(I(S;\hat{S})+\epsilon)}$ as n grows without bound. Omitting possible rounding-off effect in M_n which becomes irrelevant for large n, we have

$$\begin{aligned} M_n 2^{-n(I(S;\hat{S})+\epsilon)} \;&=\; 2^{nR} 2^{-n(I(S;\hat{S})+\epsilon)} \\ &=\; 2^{n(R - I(S;\hat{S}) - \epsilon)}. \end{aligned} \quad (4.52)$$

Since we have fixed $R > I(S; \hat{S})$ at the beginning of our proof, we can always find a sufficiently small $\epsilon > 0$ such that $R - I(S;\hat{S}) - \epsilon > 0$, and consequently the third term in (4.51) converges to zero as $n \longrightarrow \infty$.

In summary, by considering the ensemble of random codebook $\mathbf{C}$, we have proved that for any sufficiently small $\epsilon > 0$, the probability of encoding failure $P_\mathrm{f} \longrightarrow 0$ as $n \longrightarrow \infty$.

4.4.5 Concluding Steps

Returning to the analysis of expected distortion, we see that for any $\epsilon > 0$, $\mathbf{E}_{\mathbf{C},\underline{S}}[d(\underline{S},\hat{\underline{S}})] \leq D + 2\epsilon$ holds as $n \longrightarrow \infty$. This implies that there exists at least a sequence of codebooks, indexed by n, for which $\mathbf{E}[d(\underline{S},\hat{\underline{S}})] \leq D + 2\epsilon$ holds as $n \longrightarrow \infty$.[†] Letting $\epsilon \longrightarrow 0$, we can then assert that for our chosen $P_{\hat{S}|S}$, any $R > I(S;\hat{S})$ and D constitute an achievable rate-distortion pair. By choosing $P_{\hat{S}|S}$ as the one that attains the information rate-distortion function, i.e., $I(S;\hat{S}) = R_\mathrm{I}(D)$, we complete the proof of the achievability part.

Remark 4.4 In retrospect, the only place where the criterion (4.40) is needed is the analysis of P_f, starting from (4.48). If the encoding procedure ignores (4.40), the corresponding P_f will be even smaller. But the inclusion of (4.40) is essential to proving that $P_\mathrm{f} \longrightarrow 0$ as $n \longrightarrow \infty$.

Notes

The basic idea of representing a source subject to a distortion constraint appeared in Shannon's 1948 paper [1, Part V], and was later systematically developed in his work published in 1959 [16], which laid the foundation of quantization and lossy compression. Quantization and lossy compression have found widespread applications in signal (e.g., image, video, audio, speech, biomedical, financial, ...) processing.

In teaching information theory, it is a bit unusual to place the rate-distortion theory so early, even before channel transmission, but we do so considering that it appears to be more natural to start with the source block in Shannon's general communication system model (see Figure 1.1 of Lecture 1).

Our proof of the achievability part of Shannon's fundamental theorem for source coding is based on the textbook by Robert J. McEliece [17] and does not explicitly rely upon the language

[†] Considering a daily life example, that the average height of a class of students is 170cm, in the class there must be at least one student who is no shorter than 170cm.

of typicality, which has been the "standard" approach in many other textbooks (see, e.g., [18] [8]). Our introduction of the tool of typicality will be postponed until Lecture 8.

Some textbooks (see, e.g., [8] [9]) adopt a different definition of an achievable rate-distortion pair, by replacing the expected distortion constraint $\lim_{n\to\infty} \mathbf{E}[d(\underline{S}, \hat{\underline{S}})] \leq D$ with the mathematically stronger excess distortion constraint $\lim_{n\to\infty} P(d(\underline{S}, \hat{\underline{S}}) \geq D) = 0$.

A classical reference solely devoted to rate-distortion theory is [19] by Toby Berger. The reader is also invited to read the more recent survey [20] by him and Jerry D. Gibson, wherein a historical sketch of the development of rate-distortion theory (until the end of the twentieth century) is provided, including the contribution from the Russian school of information theory. In recent years, with the rapid progress of machine learning, new frontiers of lossy source coding have arised and have been under active research, such as generative coding techniques such as diffusion models (e.g., [21]), and perception oriented lossy compression (e.g., [22]).

Exercises

1. Consider a source with two components, (S_1, S_2), where S_1 and S_2 are independent, and let the distortion measure be of the form of $d((s_1, s_2), (\hat{s}_1, \hat{s}_2)) = d_1(s_1, \hat{s}_1) + d_2(s_2, \hat{s}_2)$. Denote the rate-distortion function of S_i under distortion d_i as $R_i(D)$, $i = 1, 2$. Find the rate-distortion function $R(D)$, expressed in terms of $R_1(D)$ and $R_2(D)$.
2. If the rate-distortion function for a DMS S under distortion measure $d(s, \hat{s})$, $(s, \hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}$, is $R(D)$, what is the rate-distortion function when the distortion measure is changed to $d_{k,b}(s, \hat{s}) = kd(s, \hat{s}) + b$ for some $k > 0, b \geq 0$?
3. We may use a matrix $\mathbf{D}$ to collectively represent the distortion measure for source and reproduction with finite alphabets; that is, the i-th row j-th column of $\mathbf{D}$ corresponds to $d(s_i, \hat{s}_j)$. If $\mathbf{D}$ satisfies that all its columns are permutations of a certain vector $[d_1, d_2, \ldots, d_{|\mathcal{S}|}]$, prove the following lower bound of the rate-distortion function:

$$R(D) \geq H(S) - H(V), \tag{4.53}$$

where V is the random variable that attains the largest entropy among all random variables over $\{1, 2, \ldots, |\mathcal{S}|\}$ satisfying

$$\sum_{i=1}^{|\mathcal{S}|} P_V(i) d_i \leq D.$$

Furthermore, prove that if S is uniform and the rows of $\mathbf{D}$ are permutations of each other, then the lower bound (4.53) is tight, i.e., it is exactly the rate-distortion function $R(D)$.

4. Calculate and plot the rate-distortion functions for the following models.

   a) Source alphabet is $\mathcal{S} = \{-1, 0, 1\}$, reproduction alphabet is $\hat{\mathcal{S}} = \{-1, 1\}$, source distribution is uniform $P_S(s) = \{1/3, 1/3, 1/3\}$, and distortion measure is

   $$\mathbf{D} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

   b) Source alphabet is $\mathcal{S} = \{-1, 1\}$, reproduction alphabet is $\hat{\mathcal{S}} = \{-1, 0, 1\}$, source distribution is uniform $P_S(s) = \{1/2, 1/2\}$, and distortion measure is

   $$\mathbf{D} = \begin{bmatrix} 0 & 1/2 & 1 \\ 1 & 1/2 & 0 \end{bmatrix}.$$

5. If a DMS S is Bernoulli with parameter δ and the distortion measure is Hamming, when $P_{\hat{S}|S}$ achieves $R(D)$, describe the encoding procedures (4.38) and (4.39) (4.40), and find out their differences.
6. Consider a Bernoulli(1/2) DMS S. If we use the following codebook:

   $$\mathsf{C} = \{0000, 0101, 1010, 1111\}, \tag{4.54}$$

   calculate the expected Hamming distortion with the encoding procedure (4.38).
7. Suppose that there are two independent Bernoulli DMSs S_1 and S_2 with parameters $\delta_1 \leq 1/2$ and $\delta_2 \leq 1/2$ respectively. At each time we sample one of the DMSs in a memoryless fashion so as to get a new DMS, denoted by S. Let the probability of sampling S_1 (resp. S_2) be λ (resp. $1 - \lambda$). What is the rate-distortion function of S under Hamming distortion?
8. Consider a DMS S uniform over $\mathcal{S} = \{1, \ldots, m\}$, with reproduction alphabet $\hat{\mathcal{S}} = \mathcal{S}$, and Hamming distortion measure. Calculate the rate-distortion function $R(D)$. This also gives an example where Fano's inequality (Theorem 3.9) attains equality.
9. In the problem formulation in Section 4.1, instead of assuming that the distortion measure d is a mapping from $\mathcal{S} \times \hat{\mathcal{S}}$ to $[0, \infty)$, let it be a random variable $U \in [0, \infty)$ generated according to some conditional probability distribution $P_{U|S,\hat{S}}(u|s, \hat{s})$. Generalize Theorem 4.1 to address this extended problem

formulation, and point out necessary modifications in the proofs of converse and achievability.

10. Our statement of Shannon's fundamental theorem for source coding explicitly makes the assumption that the distortion measure $d(s, \hat{s})$ is bounded, i.e., $d_{\max} < \infty$.

    a) Explain why the proof of the achievability part of the theorem may break down if the distortion measure is unbounded, i.e., if there exists some $(s, \hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}$ such that $d(s, \hat{s}) = \infty$.
    b) Consider the case where there exists $\hat{s}^* \in \hat{\mathcal{S}}$ such that $d(s, \hat{s}^*) < \infty, \forall s \in \mathcal{S}$. Argue that in this case, the rate-distortion function $R(D)$ is still given by Theorem 4.1, i.e., $\min_{P_{\hat{S}|S}} I(S; \hat{S})$ subject to $\mathbf{E}[d(S, \hat{S})] \leq D$.
    c) Argue that there exist some cases of $(\mathcal{S}, \hat{\mathcal{S}}, d(s, \hat{s}))$ such that the rate-distortion function has to be as large as $\log |\mathcal{S}|$; that is, no efficient representation is possible and $R(D)$ may differ from $R_{\mathrm{I}}(D)$.
    d) Calculate the rate-distortion function of the following setup: $\mathcal{S} = \{0, 1\}$, $\hat{\mathcal{S}} = \{0, 1, e\}$, S is Bernoulli(1/2), $d(s, \hat{s}) = 0$ if $\hat{s} = s$, 1 if $\hat{s} = e$, and ∞ if $\hat{s} \neq s$ and $\hat{s} \neq e$.