Communication Theory

Mini-Project

MSc in Mathematics and Foundations of Computer Science
Candidate Number: 445476

Rate Distortion Theory

Introduction

Rate distortion theory is a branch of information theory dealing with lossy data compression. As the name suggests, rate distortion theory is concerned with the relationship between *rate*, that is the number of bits per date sample, and *distortion*, that is the measure of the difference between input and output.

Rate distortion theory focuses on the trade-off between rate and distortion and offers an answer to the following questions [1]:

- Given a constraint on the maximum rate, which is the achievable minimum distortion? (*High fidelity*)
- Given a constraint on the maximum distortion, which is the the achievable minimum rate? (*High compression*)

Lossy data compression is the way in which part of the entropy of an input is sacrificed in order to reduce the rate [9]; in other words, a sequence of symbols (from an input alphabet) is transformed into another sequenence of symbols (from an output alphabet) containing less information. There are two main scenarios in which we could be interested in using lossy data compression:

- Given a digital input, we could be interested in reducing the rate of the input in order to minimize the space required to store it or the time required to transmit it;
- Given an analog input, we need to discard information and make it discrete in order to store it on a digital support or transmit it over a digital medium; by definition, an analog signal can assume values from a set having an infinite cardinality; but a digital representation of it can use only a finite amount of bit. Therefore it is necessary to encode each sampled value of the analog signal with a digital approximation. This process is called *quantization*.

Information Source and Alphabets

In the following analysis, we will always deal with stationary memoryless discrete sources generating random variables X with a probability distribution $p(x)$. We will assume that an input x can assume a value out of the i values of the input alphabet $\mathbb{I} = \{x_1, x_2, ..., x_i\}$ and that an output $\hat{x}$ will be reproduced using one of the o values of the output alphabet $\mathbb{O} = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_o\}$. More in general, we will consider stationary memoryless sources of sequences of random variables X^n which will generate input sequences x^n from the input alphabet $\mathbb{I}^n$ and which will be mapped to output sequences $\hat{x}^n$ from the output alphabet $\mathbb{O}^n$ [3].

Distortion Functions

There are many ways to quantizate an analog signal or to compress a digital signal, that is to map an input symbol x to an output symbol $\hat{x}$. In order to evaluate the goodness of this mapping it is useful to introduce a *distortion function* which measures the distance between the input and the output or computes the cost of representing the input x with the ouput $\hat{x}$.

Given an input alphabet $\mathbb{I}$ and an output alphabet $\mathbb{O}$, a distortion function is a function $d : \mathbb{I} \times \mathbb{O} \to \mathbb{R}^+$, which assigns a non-negative value to every pair of input-output.

There are different functions that can be used as a distortion function. The most common are:

- *Hamming distance*: $d(x, \hat{x}) = \begin{cases} 0 & if\, x = \hat{x} \\ 1 & if\, x \neq \hat{x} \end{cases}$
- *Squared error distortion*: $d(x, \hat{x}) = (x - \hat{x})^2$

A distortion measure is said to be bounded if its maximum value over all the input-output pairs is finite, that is $\forall x \in \mathbb{I}, \forall \hat{x} \in \mathbb{O} : d(x, \hat{x}) < \infty$.

It can also be useful to extend this symbol-based definition of the distortion function to a distortion function evaluating the overall distortion of a sequence of symbols; we can define a function $d : \mathbb{I}^n \times \mathbb{O}^n \to \mathbb{R}^+$ which assing a non-negative value to every pair of sequences of input-output.

The overall distortion of a pair of sequences of input-output can be computed in different ways. The most common are:

- *Average distortion*: $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$
- *Maximal distortion*: $d(x^n, \hat{x}^n) = max_{1 \leq i \leq n} d(x_i, \hat{x}_i)$

In the following analysis we will use the squared error distortion as a distortion function and the average distortion as a distortion functions for sequences of symbols.

Rate Distortion Code

We introduce now some definitions which we will use to describe our problem [2, 4, 1].

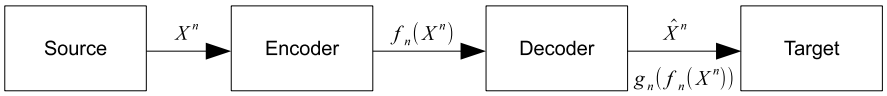

Figure 1: Rate distortion scenario

Definition Given a source producing symbols from an input alphabet $\mathbb{I}$ and R bits to reproduce the input, a $(2^{nR}, n)$-*rate distortion code* is the following pair of functions:

- *encoding function* $f_n : \mathbb{I}^n \rightarrow \{1, 2, ..., 2^{nR}\}$ converting the input to its digital representation;
- *decoding function* $g_n : \{1, 2, ..., 2^{nR}\} \rightarrow \mathbb{O}^n$ reconverting the digital represenation to the output.

Definition An *assignment region* is the region defined by $f_n^{-1}(1), f_n^{-1}(2)$, $..., f_n^{-1}(2^{nR})$; this means that $f_n^{-1}(k)$ is the region associated with the index k.

Definition A *codebook* is the set of elements $g_n(1), g_n(2), ..., g_n(2^{nR})$ denoted by $\hat{X}^n(1), \hat{X}^n(2), ..., \hat{X}^n(2^{nR})$; this means that any X^n in region k is represented as $\hat{X}^n(k)$.

Definition A *distortion associated with a* $(2^{nR}, n)$*-rate distortion code* is the expected value of the distortion over the probability distribution of the input X:

$$D = E[d(X^n, g_n(f_n(X^n)))] = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n)))$$

Rate Distortion Function

We define now explicitly the relationship between rate and distortion [2, 4, 1].

Definition A *rate distortion pair* (R, D) is the pair given by a rate R, expressed in bit, and a distortion D, expressed as the expected value of the distortion over the probability distribution of the input.

Given a rate distortion pair we define:

Definition An *achievable* rate distortion pair is, if it exists, a sequence of $(2^{nR}, n)$-rate distortion codes (f_n, g_n) such that $\lim_{n\rightarrow\infty} E[d(X^n, g_n(f_n(X^n)))] < D$;

An achievable rate distortion pair allows us to define:

Definition A *rate distortion region* C for a source is the closure of set of achievable rate distortion pairs (R, D).

A rate distortion region C is the region of all the feasible rates R given a fixed distortion D or the region of all the feasible distortions D given a fixed rate R.

Using the rate distortion region we can define:

Definition A *rate distortion function* $R(D)$ is a function giving, for any distortion D, the infimum of all the rates R such that the rate distortion pair (R, D) is in the rate distortion region C:

$$R(D) = \inf_{(R,D)\in C} R$$

The rate distortion function allows us to express the rate as a function of the distortion.

Definition A *distortion rate function* $D(R)$ is a function giving, for any rate R, the infimum of all the distortions D such that the rate distortion pair (R, D) is in the rate distortion region C;

$$D(R) = \inf_{(R,D)\in C} D$$

The distortion rate function allows us to express the distortion as a function of the rate.

It is easy to see that the rate distortion function $R(D)$ and the distortion rate function $D(R)$ are equivalent, as both of them describe the boundary of the rate distortion region C [4].

Information Rate Distortion Function

We define now the rate distortion function as a function of the source.

Definition Given a source X and a distortion function $d(x, \hat{x})$, the information rate distortion function $R^I(D)$ is the minimization over all conditional distributions $p(\hat{x} \mid x)$ such that the joint distribution $p(x)p(\hat{x} \mid x)$ satisfies the expected distortion constraint:

$$R^I(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X; \hat{X})$$

The information rate distortion function is then expressed as the minimum of the mutual information between the original random variable X and its rate distortion encoded and decoded representation $\hat{X}$.

$R^I(D)$ is a monotonous decreasing function whose range is $0 \leq R(D) \leq H(X)$. This can be easily proved showing that mutual information $I(X;\hat{X})$ is equal to the difference $H(X) - H(X \mid \hat{X})$ and that:

$$\begin{aligned} 0 &\leq H(X) \leq \log n \\ 0 &\leq H(X \mid \hat{X}) \leq H(X) \end{aligned}$$

It is also easy to see that the minimal distortion $D = 0$ is achieved when $R^I(0) = H(X)$, that is when all the entropy or information in the input X is reproduced in the output $\hat{X}$; on the other hand, the maximal distortion $D = D_{max}$ is achieved when $R^I(D_{max}) = 0$, that is when the output $\hat{X}$ has no information from the source X [9].

The computation of $R(D)$ is a hard problem to solve analytically, since it involves the minimization of a non-linear function over an unknown but constrained set of probabilities, but it can still be solved numerically [8].

Rate Distortion Theorem

We now state and prove a theorem showing the relationship between the rate distortion function and the information distortion function [2].

Theorem (Rate Distortion Theorem) Given a source of independent and identically distributed random variables X with probability distribution $p(x)$ and given a bounded distortion function $d(x, \hat{x})$, the rate distortion function is equal to the associated information rate distortion function:

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

And by the definition of the information rate distortion function we can now state that the minimum achievable rate at distortion D is:

$$\min_{p(\hat{x}|x)\,:\,\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X})$$

Proof (Rate Distortion Theorem) To prove the rate distortion theorem we will prove two different results:

(a) First we will prove the converse of the rate distortion theorem, that is we will prove that given a distribution D, the rate R of any achievable code is lower bounded by the information rate distortion function $R(D)$, that is $\min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X})$; in other words, we will prove that if $R < R(D)$ then the rate distortion pair (R, D) is not achievable [2];

(b) Then we will prove that the minimum achievable rate R given by the rate distortion function $R(D)$, that is $\min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X})$,

is indeed achievable; in other words, we will prove that for any $R > R(D)$ the rate distortion pair (R, D) is achievable [2].

Having proved that the R can be be greater but not smaller than $R(D)$ implies that we have the minimal rate at $R = R(D)$.

Before starting with the first part of the proof, it will be useful to state the following lemma.

Lemma (Convexity of $R(D)$) The rate distortion $R(D)$ is a non-increasing convex function of D [2, 6].

Proof (Convexity of $R(D)$) Recall that $R(D)$ is convex if, for any two points D_1 and D_2 and for any $\lambda \in [0, 1]$ then:

$$R(\lambda D_1 + (1-\lambda)D_2) \leq \lambda R(D_1) + (1-\lambda)R(D_2)$$

For example, we already know that given a pair of random variables $(X, Y) \sim p(x, y) = p(x)p(y \mid x)$, then the mutual information $I(X;Y)$ is a convex function of $p(y \mid x)$ for fixed $p(x)$:

$$I_{p_\lambda}(X;Y) \leq \lambda I_{p_1}(X;Y) + (1-\lambda) I_{p_2}(X;Y)$$

So, let's consider two values of the rate distortion function (R_1, D_1) and (R_2, D_2) with joint distributions $p_1(x, \hat{x}) = p(x)p_1(\hat{x} \mid x)$ and $p_2(x, \hat{x}) = p(x)p_2(\hat{x} \mid x)$. We can consider the distribution:

$$p_\lambda(x, \hat{x}) = \lambda p_1(x, \hat{x}) + (1-\lambda) p_2(x, \hat{x})$$

and, consequently, being the distortion a linear function of the distribution:

$$D(p_\lambda) = \lambda D(p_1) + (1-\lambda) D(p_2)$$

Now, using the definition of rate distortion function:

$$R(D(p_\lambda)) \leq I_{p_\lambda}(X; \hat{X})$$

by the convexity of the mutual information:

$$R(D(p_\lambda)) \leq \lambda I_{p_1}(X; \hat{X}) + (1-\lambda) I_{p_2}(X; \hat{X})$$

and substituting, we obtain:

$$R(D(p_\lambda)) \leq \lambda R(D(p_1)) + (1-\lambda) R(D(p_2))$$

which proves the convexity of the rate distortion function. ■

Proof (Converse of Information Rate Distortion Theorem) Back to the information rate distortion theorem, we want to show that for any rate $R \geq R(D)$.

Let's consider a rate distortion code $(2^{nR}, n)$ with encoding function f_n and decoding function g_n.

Knowing that the codomain of the encoding function f_n contains at most 2^{nR} elements, we can state that:

$$nR \geq H(\hat{X}^n)$$

We also know that $\hat{X}^n = g_n(f_n(X^n))$ is a function of X^n, and therefore the conditional entropy $H(\hat{X}^n \mid X^n) = H(g_n(f_n(X^n)) \mid X^n) = 0$, as once we know the information contained in X^n we know also all the information in a deterministic function of X^n. So we can write:

$$nR \geq H(\hat{X}^n) - H(\hat{X}^n \mid X^n)$$

But $H(\hat{X}^n) - H(\hat{X}^n \mid X^n)$ is the definition of the mutual information $I(\hat{X}^n; X^n)$:

$$nR \geq I(\hat{X}^n; X^n)$$

Using now the other definition of the mutual information:

$$nR \geq H(X^n) - H(X^n \mid \hat{X}^n)$$

By the independence of the independent identically distributed X_i:

$$nR \geq \sum_{i=1}^{n} H(X_i) - H(X^n \mid \hat{X}^n)$$

Remember now that the chain rule for entropy proves that $H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, X_{i-2}, ..., X_1)$; then:

$$nR \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid \hat{X}^n, X_{i-1}, X_{i-2}, ..., X_1)$$

By the principle that conditioning reduces entropy, that is $H(X) \geq H(X \mid Y)$:

$$nR \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid \hat{X}_i)$$

Again, this is the definition of mutual information:

$$nR \geq \sum_{i=1}^{n} I(X_i; \hat{X}_i)$$

and using the definition of rate distortion function, that is the infimum of the rates R such that (R, D) is in the rate distortion region for a given D:

$$\begin{aligned} nR &\geq \sum_{i=1}^{n} R(E[d(X_i; \hat{X}_i)]) \\ nR &\geq n \sum_{i=1}^{n} \frac{1}{n} R(E[d(X_i; \hat{X}_i)]) \end{aligned}$$

Now, knowing that $R(D)$ is a convex function, as we proved in the previous lemma, and knowing, by Jensen's inequality, that given a random variable X and a convex function f then $f(E[X]) \leq E[f(X)]$, we can write:

$$nR \geq nR \left(\frac{1}{n} \sum_{i=1}^{n} E[d(X_i; \hat{X}_i)] \right)$$

and using the definition of average distortion $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$, we have:

$$nR \geq nR \left(E[d(X_i^n; \hat{X}_i^n)] \right)$$

But $E[d(X_i^n; \hat{X}_i^n)]$ is the definition of the distortion associated with the rate distortion code and so:

$$nR \geq nR(D)$$

We proved that given a distribution D, the rate R of any achievable code must be greater than the information rate distortion function. ■

Now, before proving the achievability of the information rate distortion function, we will introduce new definitions and lemmas which will be useful in the next proof [2].

Definition Given an input symbols x and an output symbols $\hat{x}$ with joint probability distribution $p(x, \hat{x})$ and given a distortion measure $d(x, \hat{x})$, a *distortion typical set* $A_{d,\epsilon}^{(n)}$ for $\epsilon > 0$ is the set of all the pairs $(x^n, \hat{x}^n)$ such that:

$$\begin{aligned} \left| -\frac{1}{n} \log p(x^n) - H(X) \right| &< \epsilon \\ \left| -\frac{1}{n} \log p(x^n) - H(\hat{X}) \right| &< \epsilon \\ \left| -\frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| &< \epsilon \\ \left| d(x^n, \hat{x}^n) - E[d(X, \hat{X})] \right| &< \epsilon \end{aligned}$$

Lemma (Limit of the Distortion Typical Set) Given an independent and identically distributed pair $(X_i, \hat{X}_i)$ with distribution $p(x, \hat{x})$, then $\Pr(A_{d,\epsilon}^{(n)}) \to 1$ as $n \to \infty$.

Proof (Limit of the Distortion Typical Set) By the law of large numbers the set $A_{d,\epsilon}^{(n)}$ has probability tending to 1 as n tends to infinity, since all the sums in the definition of $A_{d,\epsilon}^{(n)}$ are normalized sums of independent identically distributed random variables tending to their expected value as n tends to infinity. ■

Lemma (Bound on $p(\hat{x}^n)$) For all $(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}$, $p(\hat{x}^n) \geq p(\hat{x}^n \mid x^n) 2^{-n(I(X;\hat{X})+3\epsilon)}$

Proof (Bound on $p(\hat{x}^n)$) By definition of conditional probability:

$$p(\hat{x}^n \mid x^n) = \frac{p(x^n, \hat{x}^n)}{p(x^n)} = p(\hat{x}^n) \frac{p(x^n, \hat{x}^n)}{p(x^n)p(\hat{x}^n)}$$

and using the property of typical sets:

$$p(\hat{x}^n) \frac{p(x^n, \hat{x}^n)}{p(x^n)p(\hat{x}^n)} \leq p(\hat{x}^n) \frac{2^{-n(H(X,\hat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)} 2^{-n(H(\hat{X})+\epsilon)}}$$

we get:

$$\begin{aligned} p(\hat{x}^n \mid x) &\leq p(\hat{x}^n) 2^{n(I(X,\hat{X})+3\epsilon)} \\ p(\hat{x}^n) &\geq p(\hat{x}^n \mid x) 2^{-n(I(X,\hat{X})+3\epsilon)} \end{aligned}$$ ■

Lemma (Bound on $(1-xy)^n$) For $0 \leq x, y \leq 1$, $n > 0$

$$(1-xy)^n \leq 1 - x + e^{-yn}$$

Proof (Bound on $(1-xy)^n$) Consider the function:

$$f(y) = e^{-y} - 1 + y$$

and its derivative:

$$\frac{d}{dy} f(y) = -e^{-y} + 1$$

since $f(0) = 0$ and $\frac{d}{dy} f(y) > 0$ for $y > 0$ we have that $f(y) > 0$ for $y > 0$. So for $0 \leq y \leq 1$ we have:

$$\begin{aligned} e^{-y} - 1 + y &\geq 0 \\ e^{-y} &\geq 1 - y \\ e^{-ny} &\geq (1-y)^n \end{aligned}$$

The bound $(1-xy)^n \leq 1 - x + e^{-yn}$ then holds for sure for $x = 1$ and $x = 0$.

Now if we consider $g_y(x) = (1-xy)^n$ and differentiate it with respect to x, we can see that $g_y(x)$ is a convex function. And for $0 \leq x \leq 1$:

$$g_y(x) = (1-xy)^n = (1-x)g_y(0) + xg_y(1) = (1-x)1 + x(1-y)^n$$

Bounding with the approximation formula $1 - x \leq e^{-x}$:

$$(1-x)1 + x(1-y)^n \leq 1 - x + xe^{-yn} \leq 1 - x + e^{-yn} \quad \blacksquare$$

Proof (Achievability of Information Rate Distortion Function) We can now prove the achievability of the information rate distortion function.

Let $p(\hat{x} \mid x)$ be chosen so that $R(D) = I(X; \hat{X})$ and compute $p(\hat{x}) = \sum_x p(x)p(\hat{x} \mid x)$. We will prove that for any $\delta > 0$ there is a rate distortion code with rate R and distortion less or equal to $D + \delta$.

Let the rate distortion codebook $\mathcal{C}$ be the set of 2^{nR} codewords w indexed by $\{1, 2, ..., 2^{nR}\}$ and given by sequences $\hat{X}^n$ drawn from the distribution $\prod_{i=1}^n p(\hat{x}_i)$.

Let the encoding function f_n encode X^n with the least w such that $(X^n, \hat{X}^n(w)) \in A_{d,\epsilon}^{(n)}$; otherwise let f_n encode X^n to 1. In this way, nR bits will be enough to encode the index w of a jointly typical word.

Let the distortion be computed as the expected distortion over all the random choices of a codebook C and over X^n, $\overline{D} = E_{X^n,\mathcal{C}}[d(X^n, \hat{X}^n)]$.

Now, given a codebook $\mathcal{C}$ and a $\epsilon > 0$, we have that a sequence x^n which is encoded by a codeword $\hat{X}^n(w)$ and whose distortion is $d(x^n, \hat{x}^n(w)) < D + \epsilon$, contributes to $\overline{D}$ at most by $D + \epsilon$, since the total probability of these sequences is at most 1; on the other hand if x^n is a sequence which is not encoded by a codeword $\hat{X}^n(w)$, then its contribution to $\overline{D}$ will be at most $P_e d_{max}$, where P_e is the probability of these sequences and d_{max} is the maximal distortion. Therefore the total distortion is:

$$E[d(X^n, \hat{X}^n)] \leq D + \epsilon + P_e d_{max}$$

To prove that the total distortion is bounded by $D + \delta$ we have to choose an appropriate ϵ and show that P_e is small.

Let $J(\mathcal{C})$ be the set of source sequences such that at least one codeword in the codebook $\mathcal{C}$ is distortion typical with x^n. Then the probability of all sequences not well represented by a code, averaged over the randomly chosen code is:

$$P_e = \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{x^n : x^n \notin J(\mathcal{C})} p(x^n)$$

which is also the probability of choosing a codebook $\mathcal{C}$ which does not well represent sequence x^n, averaged over $p(x^n)$:

$$P_e = \sum_{x^n} p(x^n) \sum_{\mathcal{C} : x^n \notin J(\mathcal{C})} P(\mathcal{C})$$

Let $K(x^n, \hat{x}^n)$ be an indicator variable assuming value 1 if $(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}$, 0 otherwise. Given a randomly chosen codeword $\hat{X}^n$ we can compute the probability that it does not well represent a fixed x^n as:

$$\Pr((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n)$$

From this we can compute the probability that 2^{nR} randomly independently chosen codewords do non represent x^n, averaged over $p(x^n)$:

$$P_e = \sum_{x^n} p(x^n) \sum_{\mathcal{C} : x^n \notin J(\mathcal{C})} P(\mathcal{C}) = \sum_{x^n} p(x^n) \left[1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \right]^{2^{nR}}$$

Now we use the lemma bounding $p(\hat{x}^n)$:

$$\sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \geq \sum_{\hat{x}^n} p(\hat{x}^n \mid x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} K(x^n, \hat{x}^n)$$

and then:

$$P_e \leq \sum_{x^n} p(x^n) \left[1 - \sum_{\hat{x}^n} p(\hat{x}^n \mid x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} K(x^n, \hat{x}^n) \right]^{2^{nR}}$$

Using now the lemma bounding $(1 - xy)^n$:

$$\left[1 - \sum_{\hat{x}^n} p(\hat{x}^n \mid x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} K(x^n, \hat{x}^n) \right]^{2^{nR}} \leq$$

$$\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n \mid x^n) K(x^n, \hat{x}^n) + e^{-(2^{-n(I(X;\hat{X})+3\epsilon)} 2^{nR})}$$

and then:

$$P_e \leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n \mid x^n) K(x^n, \hat{x}^n) + e^{-2^{-n(R-I(X;\hat{X})-3\epsilon)}}$$

Now, since we chose $p(\hat{x} \mid x)$ to be the conditional distribution achieving the minimum of the rate distortion function, then $R > R(D)$ implies $R > I(X;\hat{X}) + 3\epsilon$. The last term goes to zero exponentially fast with n; for the first two terms, using the lemma on the limit of the distortion typical set we have:

$$1 - \sum_{x^n} \sum_{\hat{x}^n} p(x^n, \hat{x}^n) K(x^n, \hat{x}^n) = \Pr((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)})$$

which, for n sufficiently large, can be made smaller than ϵ.

In conclusion, for every choice of δ, we can find an ϵ and a n such that the expected distortion is less than $D + \delta$ and we can devise a codebook $\mathcal{C}*$ with the required average distortion.

Proving that the choice of δ is arbitrary, we have proved that (R, D) is achievable if $R > R(D)$. ■

Beyond proving the existence of a rate distortion code of rate $R(D)$ with average distortion close to D, it is also possible to prove a stronger statement, that the total probability that the distortion is greater than $D + \delta$ is close to 0 [2].

Continuous Rate Distortion Functions

So far, we have studied a discrete model of the rate distortion function. However, in the real world, we have to deal with continuous signals and it is therefore natural to define continuous rate-distortion functions [9].

Let's consider a stationary memoryless continuous source X with probability distribution $p(x)$.

To evaluate the distortion between an input x and its output $\hat{x}$, we define a distortion function $d(x, \hat{x})$; the most common distortion functions in the continuous case are:

- *Absolute error*: $d(x, \hat{x}) = |x - \hat{x}|$;
- *Squared error distortion*: $d(x, \hat{x}) = (x - \hat{x})^2$;

Given a distortion function and a conditional probability distribution $p(\hat{x} \mid x)$, we can compute the average distortion as:

$$D_{p(\hat{x}|x)} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x) p(\hat{x} \mid x) d(x, \hat{x})\, dx\, d\hat{x}$$

And equally we can define the mutual information as:

$$I_{p(\hat{x}|x)} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x) p(\hat{x} \mid x) \log \frac{p(\hat{x} \mid x)}{p(\hat{x})}\, dx\, d\hat{x}$$

So, if we fix a permissible distortion D, then the rate distortion function $R(D)$ is the minimum of $I_{p(\hat{x}|x)}$ satisfying this constraint:

$$R(D) = \inf_{p(\hat{x}|x)\,:\,D_{p(\hat{x}|x)}<D} I_{p(\hat{x}|x)}$$

Similarly to the discrete case, $R(D)$ is a monotonous decreasing function, but whose range is now $0 < R(D) < \infty$; it is easy to see that if the distortion tends to zero, $D \to 0$, then the rate tends to infinity, $R(D) \to \infty$, as reproducing a continuous source without distortion requires an infinite rate; on the other hand, as the allowed distortion increases, the rate decreases till reaching 0 when all the information of the input x is lost.

Shannon Lower Bound

We are now going to state and prove the *Shannon lower bound*, a lower bound $R_{SLB}(D)$ on the value of $R(D)$, which is helpful whenever computing the rate distortion function happens to be too difficult [2, 7, 5].

Theorem (Shannon Lower Bound) Given a source of random variables X from an input alphabet $\mathbb{I} = \{1, 2, ..., m\}$, given a distortion measure $d(x, \hat{x})$ satisfying the property that all the columns of the distortion matrix are permutations of the set$\{d_1, d_2, ..., d_m\}$, if we define a function $\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \leq D} H(\mathbf{p})$ then:

$$R(D) \geq H(X) - \phi(D)$$

Before proving this theorem, we will introduce the following lemma.

Lemma (Concavity of $\phi(D)$) The function $\phi(D)$ is an increasing concave function of D.

Proof (Concavity of $\phi(D)$) Again, recall that $\phi(D)$ is concave if, for any two points D_1 and D_2 and for any $\lambda \in [0, 1]$ then:

$$\phi(\lambda D_1 + (1-\lambda)D_2) \geq \lambda\phi(D_1) + (1-\lambda)\phi(D_2)$$

So, let's consider two values of the distortion D_1 and D_2 and let $\mathbf{p}_1$ and $\mathbf{p}_2$ be the value corresponding to $\phi(D_1)$ and $\phi(D_2)$. We can consider the distribution:

$$\mathbf{p}_\lambda = \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2$$

and, consequently, being the distortion a linear function of the distribution, the function:

$$D(\mathbf{p}_\lambda) = \lambda D(\mathbf{p}_1) + (1-\lambda)D(\mathbf{p}_2)$$

So, the definition of $\phi(D)$ is:

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \leq D} H(\mathbf{p})$$

and considering it in the case of D_λ we have:

$$\phi(D_\lambda) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \leq D_\lambda} H(\mathbf{p})$$

Now, we defined the value of $\mathbf{p}$ that maximizes $\phi(D_\lambda)$ to be $\mathbf{p}_\lambda$, so:

$$\phi(D_\lambda) \geq H(\mathbf{p}_\lambda)$$

The entropy $H(p)$, too, is a function which is concave in the distribution p:

$$H(\mathbf{p}_\lambda) \geq \lambda H(\mathbf{p}_1) + (1-\lambda) H(\mathbf{p}_2)$$

But $\mathbf{p}_1$ and $\mathbf{p}_2$ were defined as the values corresponding to to $\phi(D_1)$ and $\phi(D_2)$:

$$\phi(D_\lambda) \geq \lambda H(\mathbf{p}_1) + (1-\lambda) H(\mathbf{p}_2) = \lambda \phi(D_1) + (1-\lambda)\phi(D_2)$$

And so we proved that:

$$\phi(\lambda D_1 + (1-\lambda) D_2) \geq \lambda \phi(D_1) + (1-\lambda)\phi(D_2) \ \blacksquare$$

Proof (Shannon Lower Bound) We can now prove the Shannon lower bound. Let's assume $D \geq E[d(X, \hat{X})]$.

Thanks to the rate distortion theorem we know that:

$$R(D) = \min_{p(\hat{x}|x) \,:\, \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X})$$

Let's consider the mutual information $I(X; \hat{X})$; by definition we have:

$$I(X; \hat{X}) = H(X) - H(X \mid \hat{X})$$

Expressing explicitly the conditional entropy $H(X \mid \hat{X})$ we get:

$$I(X; \hat{X}) = H(X) - \sum_{\hat{x}} p(\hat{x}) H(X \mid \hat{X} = \hat{x})$$

Now, by the definition of $\phi(D)$, we have:

$$\phi(D_{\hat{x}}) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \leq D_D} H(\mathbf{p}) \geq H(X \mid \hat{X} = \hat{x})$$

and so:

$$I(X; \hat{X}) \geq H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}})$$

Knowing that $\phi(D)$ is a concave function, as we proved in the previous lemma:

$$I(X;\hat{X}) \geq H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}}\right)$$

Substituting $D_{\hat{x}}$ with its definition we get:

$$\begin{aligned} I(X;\hat{X}) &\geq H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x}) \sum_{x} p(x \mid \hat{x}) d(x,\hat{x})\right) \\ I(X;\hat{X}) &\geq H(X) - \phi\left(\sum_{x}\sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x})\right) \\ I(X;\hat{X}) &\geq H(X) - \phi\left(D\right) \end{aligned}$$

In this way we have proved the Shannon lower bound on the rate distortion:

$$R(D) \geq H(X) - \phi(D) \quad \blacksquare$$

In the discrete case, it is also possible to prove that, if the source has a uniform distribution and the rows of the distortion matrix are permutations of each other, strict equality holds in Shannon lower bound, that is $R(D) = H(X) - \phi(D)$ [2].

References

[1] Chia-Ping Chen. Rate distortion theory. Lecture Notes for the class in Information Theory at National Sun Yat-Sen University.

[2] Thomas M. Cover and Thomas M. Joy. *Elements of Information Theory.* John Wiley & Sons, 1991.

[3] Natasha Devroye. Rate distortion theory. Lecture Notes for the class in Information Theory at the University of Illinois in Chicago.

[4] Markku Juntti. Rate distortion theory. Lecture Notes for the class in Basics of Information Theory at the University of Oulu.

[5] Tamas Linder and Ram Zamir. On the asymptotic tightness of the shannon lower bound. *IEEE Transactions on Information Theory*, 40:2026–2031, 1994.

[6] Robert J. McEliece. *The Theory of Information and Coding.* Cambridge University Press, 2002.

[7] Khalid Sayood. *Introduction to data compression.* Elsevier, 2000.

[8] Roberto Togneri and Cristopher J. S. deSilva. *Fundamentals of Information Theory and Coding Design.* Chapman & Hall, 2002.

[9] Jan C. A. van der Lubbe. *Information Theory.* Cambridge University Press, 1988.