Elements of Information Theory

George Voutsadakis[1]

[1]Mathematics and Computer Science
Lake Superior State University

LSSU Math 500

Objective of Rate Distortion Theory

- The description of a real number requires an infinite number of bits.
- So a finite representation of a continuous random variable can never be perfect.
- To estimate how good a representation is, it is necessary to define a "goodness" of a representation of a source.
- We define a distortion measure which is a measure of distance between the random variable and its representation.
- The basic problem in rate distortion theory can then be stated as follows:
  - Given a source distribution and a distortion measure, what is the minimum expected distortion achievable at a particular rate?
  - Equivalently, what is the minimum rate description required to achieve a particular distortion?

Results of Rate Distortion Theory

- One of the most intriguing aspects of this theory is that joint descriptions are more efficient than individual descriptions.
- This is true even for independent random variables.
- It is simpler to describe X_1 and X_2 together (at a given distortion for each) than to describe each by itself.
- The reason why independent problems do not have independent solutions is found in the geometry.
- Apparently, rectangular grid points (arising from independent descriptions) do not fill up the space efficiently.

Subsection 1

Quantization

Problem and Notation

- Consider the problem of representing a single sample from the source.
- Let the random variable be represented be X.
- Let the representation of X be denoted as $\widehat{X}(X)$.
- If we are given R bits to represent X, the function $\widehat{X}$ can take on 2^R values.
- The problem is to find the optimum set of values for $\widehat{X}$ (called the **reproduction points** or **code points**) and the regions that are associated with each value $\widehat{X}$.

Example

- Let $X \sim \mathcal{N}(0, \sigma^2)$.

  Assume a squared-error distortion measure.

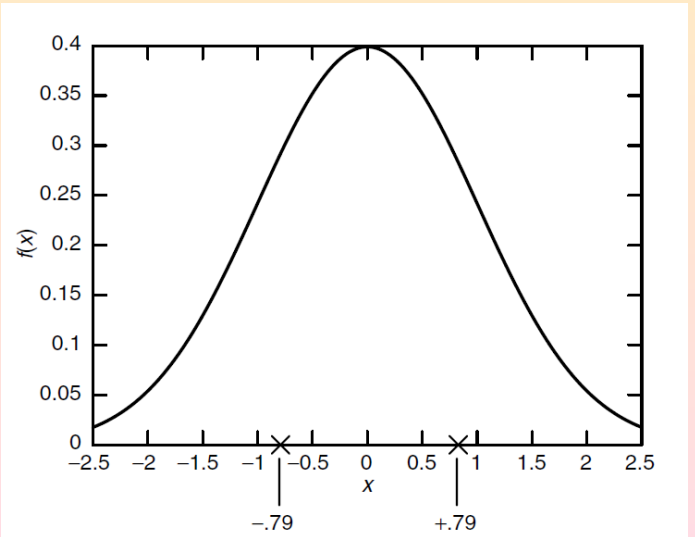
  In this case we wish to find the function $\widehat{X}(X)$, that:
  - Takes on at most 2^R values;
  - Minimizes $E(X - \widehat{X}(X))^2$.

  If we are given one bit to represent X, it is clear that the bit should distinguish whether or not $X > 0$.

  To minimize squared error, each reproduced symbol should be the conditional mean of its region.

  Thus,

$$\widehat{X}(x) = \begin{cases} \sqrt{\frac{2}{\pi}}\sigma, & \text{if } x \geq 0 \\ -\sqrt{\frac{2}{\pi}}\sigma, & \text{if } x < 0 \end{cases}$$



Desired Properties

- Suppose we are given 2 bits to represent the sample.
  - We want to divide the real line into four regions;
  - We want to use a point within each region to represent the sample.
- It is no longer immediately obvious what the representation regions and the reconstruction points should be.
- We give two simple properties of optimal regions and reconstruction points for the quantization of a single random variable.
  - Given a set $\{\widehat{X}(w)\}$ of reconstruction points, the distortion is minimized by mapping a source random variable X to the representation $\widehat{X}(w)$ that is closest to it.
    The set of regions of X defined by this mapping is called a **Voronoi** or **Dirichlet partition** defined by the reconstruction points.
  - The reconstruction points should minimize the conditional expected distortion over their respective assignment regions.

The Lloyd Algorithm

- These two properties enable us to construct a simple algorithm to find a "good" quantizer:
  - We start with a set of reconstruction points.
  - Find the optimal set of reconstruction regions (which are the nearest-neighbor regions with respect to the distortion measure).
  - Find the optimal reconstruction points for these regions (the centroids of these regions if the distortion is squared error).
  - Repeat the iteration for this new set of reconstruction points.
- The expected distortion is decreased at each stage in the algorithm.
- So the algorithm will converge to a local minimum of the distortion.
- This is called the **Lloyd algorithm** (for real-valued random variables).
- It is called the **generalized Lloyd algorithm** (for vector valued random variables).

Representing Multiple Variables

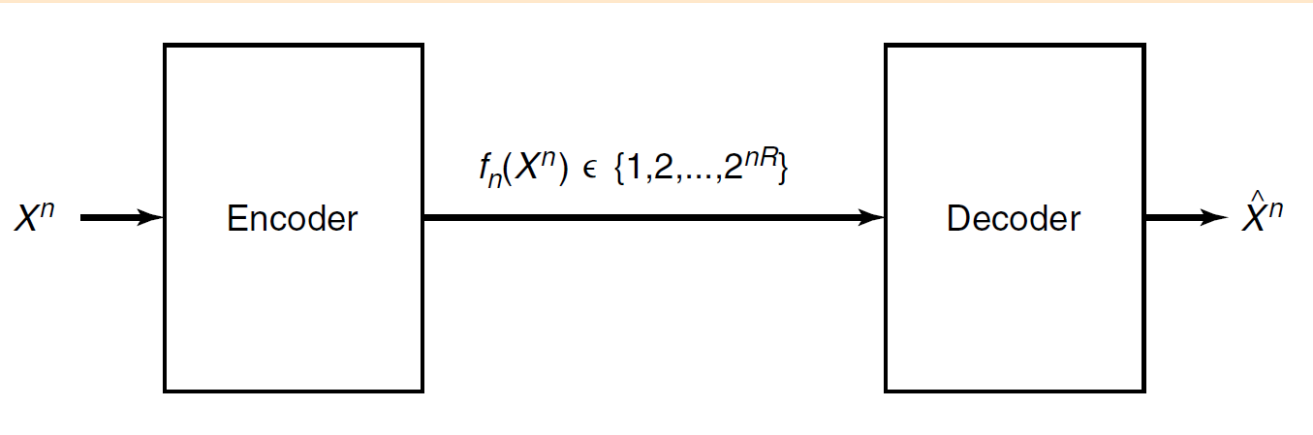
- Assume that we are given a set of n i.i.d. random variables drawn according to a Gaussian distribution.
- These random variables are to be represented using nR bits.
- Since the source is i.i.d., the symbols are independent.
- So it may appear that the representation of each element is an independent problem to be treated separately.
- But this is not true, as the results on rate distortion theory will show.
- We represent the entire sequence by a single index taking 2^{nR} values.
- We will see that this treatment of entire sequences at once achieves a lower distortion for the same rate than independent quantization of the individual samples.

Subsection 2

Definitions

Setup

- Assume that we have a source that produces a sequence $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x)$, $x \in \mathcal{X}$.
- For the proofs we assume that the alphabet is finite, but most of the proofs can be extended to continuous random variables.
- The encoder describes the source sequence X^n by an index $f_n(X^n) \in \{1, 2, \ldots, 2^{nR}\}$.
- The decoder represents X^n by an estimate $\hat{X}^n \in \hat{\mathcal{X}}$.



Distortion

Definition

A **distortion function** or **distortion measure** is a mapping

$$d : \mathcal{X} \times \widehat{\mathcal{X}} \to \mathbb{R}^+$$

from the set of source alphabet-reproduction alphabet pairs into the set of nonnegative real numbers. The distortion $d(x, \widehat{x})$ is a measure of the cost of representing the symbol x by the symbol $\widehat{x}$.

Definition

A distortion measure is said to be **bounded** if the maximum value of the distortion is finite:

$$d_{\max} = \max_{x \in \mathcal{X}, \widehat{x} \in \widehat{\mathcal{X}}} d(x, \widehat{x}) < \infty.$$

The Hamming Distortion

- In most cases, the reproduction alphabet $\widehat{\mathcal{X}}$ is the same as the source alphabet $\mathcal{X}$.
- **Hamming (Probability of Error) Distortion**

  The Hamming distortion is given by

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

  We have

$$Ed(X,\widehat{X}) = \Pr(X \neq \widehat{X}).$$

  So the Hamming distortion results in a probability of error distortion.

The Squared-Error Distortion

- **Squared-Error Distortion**

  The squared-error distortion is given by

  $$d(x, \widehat{x}) = (x - \widehat{x})^2.$$

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

  Advantages:
  - Simplicity;
  - Relationship to least-squares prediction.

  Drawbacks:
  - Inappropriate in applications such as image and speech coding as a measure of distortion for human observers.

Distortion Between Sequences

- For sequences, the distortion measure is defined on a symbol-by-symbol basis.

Definition

The **distortion between sequences** x^n and $\widehat{x}^n$ is defined by

$$d(x^n, \widehat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \widehat{x}_i).$$

- So the distortion for a sequence is the average of the per symbol distortion of the elements of the sequence.

Distortion Codes

Definition

A $(2^{nR}, n)$-**rate distortion code** consists of:

- An encoding function $f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$;
- A decoding (reproduction) function, $g_n : \{1, 2, \ldots, 2^{nR}\} \to \widehat{\mathcal{X}}^n$.

The **distortion** associated with the $(2^{nR}, n)$ code is defined as

$$D = Ed(X^n, g_n(f_n(X^n))),$$

the expectation being with respect to the probability distribution on X:

$$D = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))).$$

Distortion Codes (Cont'd)

Definition (Cont'd)

The set of n-tuples $g_n(1), g_n(2), \ldots, g_n(2^{nR})$, denoted by

$$\widehat{X}^n(1), \ldots, \widehat{X}^n(2^{nR}),$$

constitutes the **codebook**.
The sequence

$$f_n^{-1}(1), \ldots, f_n^{-1}(2^{nR})$$

consists of the associated **assignment regions**.

- $\widehat{X}^n$ is called the **vector quantization**, **reproduction**, **reconstruction**, **representation**, **source code**, or **estimate** of X^n.

Achievability and Rate Distortion Region

Definition

A rate distortion pair (R, D) is said to be **achievable** if there exists a sequence of $(2^{nR}, n)$-rate distortion codes (f_n, g_n) with

$$\lim_{n\to\infty} Ed(X^n, g_n(f_n(X^n))) \leq D.$$

Definition

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

Rate Distortion and Distortion Rate Functions

Definition

The **rate distortion function** $R(D)$ is the infimum of rates R such that (R, D) is in the rate distortion region of the source for a given distortion D.

Definition

The **distortion rate function** $D(R)$ is the infimum of all distortions D such that (R, D) is in the rate distortion region of the source for a given rate R.

The Information Rate Distortion Function

- We now define a mathematical function of the source, which we call the information rate distortion function.
- The main result of the chapter is the proof that the information rate distortion function is equal to the rate distortion function.

Definition

The **information rate distortion function** $R^{(I)}(D)$ for a source X with distortion measure $d(x, \widehat{x})$ is defined as

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

where the minimization is over all conditional distributions $p(\widehat{x}|x)$ for which the joint distribution $p(x, \widehat{x}) = p(x)p(\widehat{x}|x)$ satisfies the expected distortion constraint.

Preview of Developments

- We consider the properties of the information rate distortion function.
- We calculate it for some simple sources and distortion measures.
- We then prove that we can actually achieve this function (i.e., there exist codes with rate $R^{(I)}(D)$ with distortion D).
- Finally, we prove a converse establishing that $R \leq R^{(I)}(D)$ for any code that achieves distortion D.

The Main Theorem

- The main theorem of rate distortion theory can be stated as follows:

Theorem

The rate distortion function for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x, \widehat{x})$ is equal to the associated information rate distortion function:

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

is the minimum achievable rate at distortion D.

- This theorem shows that the operational definition of the rate distortion function is equal to the information definition.

Subsection 3

Calculation of the Rate Distortion Function

Binary Source

Theorem

The rate distortion function for a Bernoulli(p) source with Hamming distortion is given by

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\} \end{cases}.$$

- Consider a binary source $X \sim$ Bernoulli(p).

  Adopt a Hamming distortion measure.

  Without loss of generality, we may assume that $p < \frac{1}{2}$.

  We wish to calculate the rate distortion function,

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

The Lower Bound

- Let $\oplus$ denote modulo 2 addition.
  Thus, $X \oplus \widehat{X} = 1$ is equivalent to $X \neq \widehat{X}$.
  Instead of minimizing $I(X;\widehat{X})$ directly:
  - We find a lower bound;
  - We show that this lower bound is achievable.

  For any joint distribution satisfying the distortion constraint, we have

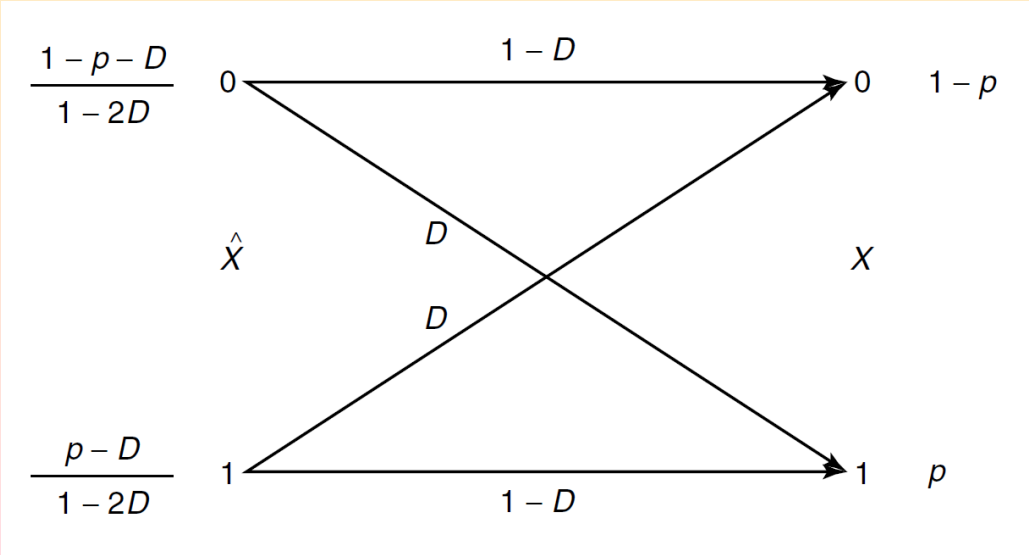
$$\begin{aligned} I(X;\widehat{X}) &= H(X) - H(X|\widehat{X}) \\ &= H(p) - H(X \oplus \widehat{X}|\widehat{X}) \\ &\geq H(p) - H(X \oplus \widehat{X}) \\ &\geq H(p) - H(D), \end{aligned}$$

  since $\Pr(X \neq \widehat{X}) \leq D$ and $H(D)$ increases with D for $D \leq \frac{1}{2}$.
  Thus,

$$R(D) \geq H(p) - H(D).$$

Achieving the Lower Bound

- We now show that the lower bound is actually the rate distortion function by finding a joint distribution that meets the distortion constraint and has $I(X;\widehat{X}) = R(D)$.
- For $0 \leq D \leq p$, we can achieve the value of the rate distortion function $R(D) = H(p) - H(D)$ by choosing $(X, \widehat{X})$ to have the joint distribution given by the binary symmetric channel shown below.



Achieving the Lower Bound (Cont'd)

- We choose the distribution of $\widehat{X}$ at the input of the channel so that the output distribution of X is the specified distribution.

  Let $r = \Pr(\widehat{X} = 1)$.

  Then choose r so that $r(1-D) + (1-r)D = p$.

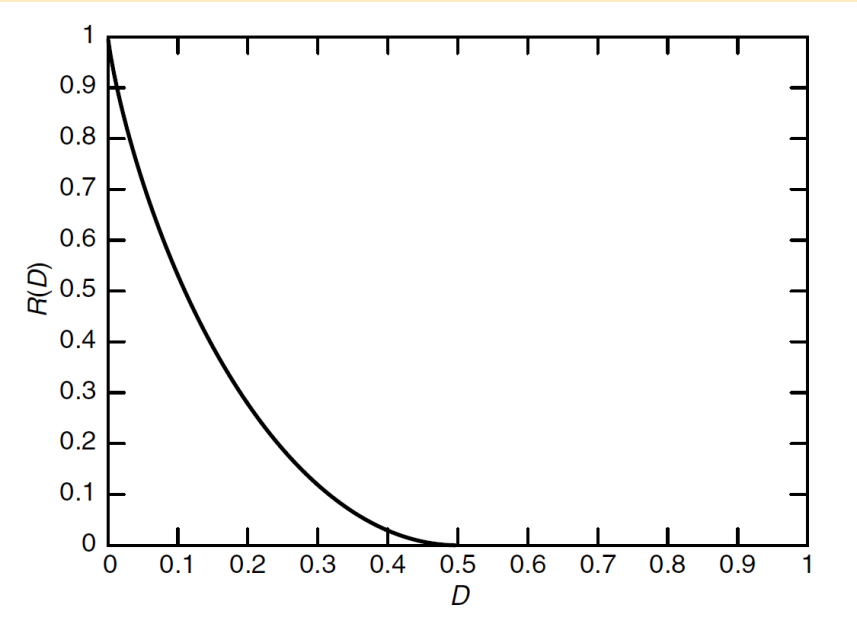
  This gives $r = \frac{p-D}{1-2D}$.

  - Suppose $D \leq p \leq \frac{1}{2}$. Then $\Pr(\widehat{X} = 1) \geq 0$ and $\Pr(\widehat{X} = 0) \geq 0$.
    So $I(X;\widehat{X}) = H(X) - H(X|\widehat{X}) = H(p) - H(D)$.
    Moreover, the expected distortion is $\Pr(X \neq \widehat{X}) = D$.
  - Suppose $D \geq p$. Let $\widehat{X} = 0$ with probability 1.
    Then $R(D) = 0$. In this case, $I(X;\widehat{X}) = 0$ and $D = p$.
  - Suppose $D \geq 1 - p$. Let $\widehat{X} = 1$ with probability 1.
    Then $R(D) = 0$. In this case, $I(X;\widehat{X}) = 0$ and $D = 1 - p$.

The Function $R(D)$

- The rate distortion function for a binary source is

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\} \end{cases}.$$



Gaussian Source

Theorem

The rate distortion function for a $\mathcal{N}(0, \sigma^2)$ source with squared-error distortion is

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2 \\ 0, & D > \sigma^2 \end{cases}$$

- Let X be $\sim \mathcal{N}(0, \sigma^2)$. By the rate distortion theorem extended to continuous alphabets, we have

$$R(D) = \min_{f(\widehat{x}|x): E(\widehat{X}-X)^2 \leq D} I(X; \widehat{X}).$$

Following a similar strategy:

- We first find a lower bound for the rate distortion function;
- We then prove that it is achievable.

Lower Bound

- We observe that

$$
\begin{aligned}
I(X;\widehat{X}) &= h(X) - h(X|\widehat{X}) \\
&= \tfrac{1}{2}\log(2\pi e)\sigma^2 - h(X-\widehat{X}|\widehat{X}) \\
&\geq \tfrac{1}{2}\log(2\pi e)\sigma^2 - h(X-\widehat{X}) \\
&\quad \text{(conditioning reduces entropy)} \\
&\geq \tfrac{1}{2}\log(2\pi e)\sigma^2 - h(\mathcal{N}(0, E(X-\widehat{X})^2)) \\
&\quad \text{(normal distribution maximizes the entropy)} \\
&= \tfrac{1}{2}\log(2\pi e)\sigma^2 - \tfrac{1}{2}\log(2\pi e)E(X-\widehat{X})^2 \\
&\geq \tfrac{1}{2}\log(2\pi e)\sigma^2 - \tfrac{1}{2}\log(2\pi e)D \\
&= \tfrac{1}{2}\log\tfrac{\sigma^2}{D}.
\end{aligned}
$$

Hence, $R(D) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$.

Gaussian Source

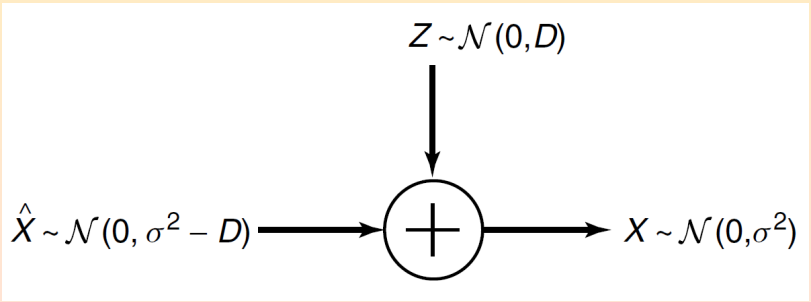
- We want to find $f(\widehat{x}|x)$ that achieves this lower bound.

  It is usually more convenient to look instead at $f(x|\widehat{x})$.

  This is sometimes called the **test channel**.

  We construct $f(x|\widehat{x})$ to achieve equality in the bound.

  - Suppose $D \leq \sigma^2$. Choose $X = \widehat{X} + Z$, with $\widehat{X}$ and Z independent, such that:
    - $\widehat{X} \sim \mathcal{N}(0, \sigma^2 - D)$;
    - $Z \sim \mathcal{N}(0, D)$.

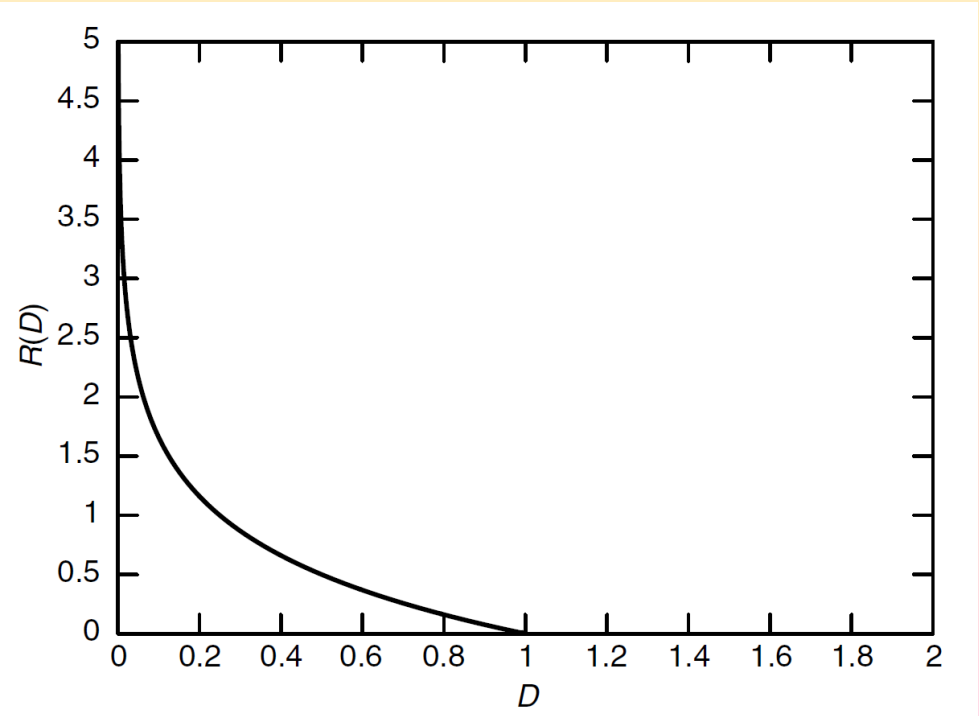


    We have $I(X;\widehat{X}) = \frac{1}{2}\log\left(1 + \frac{\sigma^2 - D}{D}\right) = \frac{1}{2}\log\frac{\sigma^2}{D}$ and $E(X - \widehat{X})^2 = D$.
    This achieves the bound.
  - Suppose $D > \sigma^2$. We choose $\widehat{X} = 0$ with probability 1.
    This achieves $R(D) = 0$.

Gaussian Source

- Hence, the rate distortion function for the Gaussian source with squared-error distortion is

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2 \\ 0, & D > \sigma^2 \end{cases}.$$



Independent Gaussian Random Variables

- Consider the case of representing m independent (but not identically distributed) normal random sources $X_1, \ldots, X_m$, where X_i are $\sim \mathcal{N}(0, \sigma_i^2)$, with squared-error distortion.
- Assume that we are given R bits with which to represent this random vector.
- Extending the definition of the information rate distortion function to the vector case, we have

$$R(D) = \min_{f(\widehat{x}^m|x^m): Ed(X^m, \widehat{X}^m) \leq D} I(X^m; \widehat{X}^m),$$

where $d(x^m, \widehat{x}^m) = \sum_{i=1}^m (x_i - \widehat{x}_i)^2$.

Lower Bound

- Let $D_i = E(X_i - \widehat{X}_i)^2$.

  Now using the arguments in the preceding example, we have

$$\begin{aligned} I(X^m; \widehat{X}^m) &= h(X^m) - h(X^m|\widehat{X}^m) \\ &= \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i|X^{i-1}, \widehat{X}^m) \\ &\geq \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i|\widehat{X}_i) \\ &\quad \text{(conditioning reduces entropy)} \\ &= \sum_{i=1}^m I(X_i; \widehat{X}_i) \\ &\geq \sum_{i=1}^m R(D_i) \\ &= \sum_{i=1}^m (\tfrac{1}{2} \log \tfrac{\sigma_i^2}{D_i})^+. \end{aligned}$$

  For equality in the first inequality, choose $f(x^m|\widehat{x}^m) = \prod_{i=1}^m f(x_i|\widehat{x}_i)$.

  For equality in the second inequality choose $\widehat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i)$.

Rewriting and Minimization

- The problem of finding the rate distortion function can be reduced to the following optimization (using nats for convenience):

$$R(D) = \min_{\sum D_i = D} \sum_{i=1}^{m} \max\left\{\frac{1}{2}\ln\frac{\sigma_i^2}{D_i}, 0\right\}.$$

Using Lagrange multipliers, we construct the functional

$$J(D) = \sum_{i=1}^{m} \frac{1}{2}\ln\frac{\sigma_i^2}{D_i} + \lambda\sum_{i=1}^{m} D_i.$$

Differentiating with respect to D_i and setting equal to 0, we have

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D_i} + \lambda = 0 \quad \text{or} \quad D_i = \lambda'.$$

Hence, the optimum allotment of the bits to the various descriptions results in an equal distortion for each random variable.
This is possible if the constant λ' is less than σ_i^2, for all i.

Rewriting and Minimization (Large Distortion)

- As the total allowable distortion D is increased, the constant λ' increases until it exceeds σ_i^2 for some i.
- At this point the solution is on the boundary of the allowable region of distortions.
- If we increase the total distortion, we must use the Kuhn-Tucker conditions to find the minimum in $J(D)$.
- In this case the Kuhn-Tucker conditions yield

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D_i} + \lambda,$$

where λ is chosen so that

$$\frac{\partial J}{\partial D_i} \begin{cases} = 0, & \text{if } D_i < \sigma_i^2 \\ \leq 0, & \text{if } D_i \geq \sigma_i^2 \end{cases}.$$

Rewriting and Minimization (Solution Theorem)

- The solution to the Kuhn-Tucker equations is given by

Theorem (Rate Distortion for a Parallel Gaussian Source)

Let $X_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, 2, \ldots, m$, be independent Gaussian random variables, and let the distortion measure be $d(x^m, \hat{x}^m) = \sum_{i=1}^m (x_i - \hat{x}_i)^2$. Then the rate distortion function is given by

$$R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{\sigma_i^2}{D_i},$$

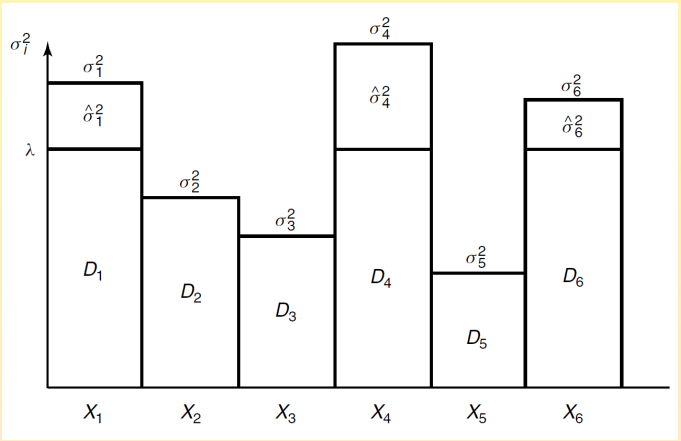
where

$$D_i = \begin{cases} \lambda, & \text{if } \lambda < \sigma_i^2 \\ \sigma_i^2, & \text{if } \lambda \geq \sigma_i^2 \end{cases},$$

where λ is chosen so that $\sum_{i=1}^m D_i = D$.

Reverse Water Filling

- This gives rise to a kind of reverse water-filling.



- We choose a constant λ and only describe those random variables with variances greater than λ.
- No bits are used to describe random variables with variance less than λ.
- If $X \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m^2 \end{bmatrix}\right)$, then:
  - $\widehat{X} \sim \mathcal{N}\left(0, \begin{bmatrix} \widehat{\sigma}_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \widehat{\sigma}_m^2 \end{bmatrix}\right)$;
  - $E(X_i - \widehat{X}_i)^2 = D_i$, where $D_i = \min\{\lambda, \sigma_i^2\}$.

Subsection 4

Converse to the Rate Distortion Theorem

Convexity of $R(D)$

Lemma (Convexity of $R(D)$)

The rate distortion function

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

is a nonincreasing convex function of D.

- $R(D)$ is the minimum of the mutual information over increasingly larger sets as D increases. Thus, $R(D)$ is nonincreasing in D.

  To prove that $R(D)$ is convex, consider two rate distortion pairs, (R_1, D_1) and (R_2, D_2), which lie on the rate distortion curve.
  Let the joint distributions that achieve these pairs be, respectively:
  - $p_1(x, \widehat{x}) = p(x)p_1(\widehat{x}|x)$;
  - $p_2(x, \widehat{x}) = p(x)p_2(\widehat{x}|x)$.

  Consider the distribution $p_\lambda = \lambda p_1 + (1 - \lambda)p_2$.

Convexity of $R(D)$ (Cont'd)

- Since the distortion is a linear function of the distribution, we have

$$D(p_\lambda) = \lambda D_1 + (1-\lambda) D_2.$$

Mutual information, on the other hand, is a convex function of the conditional distribution, whence

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

Hence, by the definition of the rate distortion function,

$$\begin{aligned} R(D_\lambda) &\leq I_{p_\lambda}(X;\widehat{X}) \\ &\leq \lambda I_{p_1}(X;\widehat{X}) + (1-\lambda) I_{p_2}(X;\widehat{X}) \\ &= \lambda R(D_1) + (1-\lambda) R(D_2). \end{aligned}$$

This proves that $R(D)$ is a convex function of D.

Converse to the Rate Distortion Theorem

- We must show for any source X drawn i.i.d. $\sim p(x)$, with distortion measure $d(x, \widehat{x})$, and any $(2^{nR}, n)$ rate distortion code, with distortion $\leq D$, the rate R of the code satisfies $R \geq R(D)$.

  In fact, we prove that $R \geq R(D)$ even for randomized mappings f_n and g_n, as long as f_n takes on at most 2^{nR} values.

  Consider any $(2^{nR}, n)$ rate distortion code defined by functions $f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$ and $g_n : \{1, 2, \ldots, 2^{nR}\} \to \widehat{\mathcal{X}}^n$.

  Let

$$\widehat{X}^n = \widehat{X}^n(X^n) = g_n(f_n(X^n))$$

  be the reproduced sequence corresponding to X^n.

  Assume that $Ed(X^n, \widehat{X}^n) \leq D$ for this code.

Converse to the Rate Distortion Theorem (Cont'd)

- Then we have the following chain of inequalities:

$$
\begin{aligned}
nR &\geq H(f_n(X^n)) \\
&\quad \text{(the range of } f_n \text{ is at most } 2^{nR}\text{)} \\
&\geq H(f_n(X^n)) - H(f_n(X^n)|X^n) \\
&\quad (H(f_n(X^n)|X^n) \geq 0) \\
&= I(X^n; f_n(X^n)) \\
&\geq I(X^n; \widehat{X}^n) \quad \text{(data-processing inequality)} \\
&= H(X^n) - H(X^n|\widehat{X}^n) \\
&= \textstyle\sum_{i=1}^n H(X_i) - H(X^n|\widehat{X}^n) \\
&\quad \text{(the } X_i \text{ are independent)} \\
&= \textstyle\sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\widehat{X}^n, X_{i-1}, \ldots, X_1) \\
&\quad \text{(chain rule for entropy)} \\
&\geq \textstyle\sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\widehat{X}_i). \\
&\quad \text{(conditioning reduces entropy)}
\end{aligned}
$$

Converse to the Rate Distortion Theorem (Cont'd)

- We continue from

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\widehat{X}_i) \\
&= \sum_{i=1}^{n} I(X_i;\widehat{X}_i) \\
&\geq \sum_{i=1}^{n} R(Ed(X_i,\widehat{X}_i)) \\
&\quad \text{(definition of the rate distortion function)} \\
&= n\left(\frac{1}{n}\sum_{i=1}^{n} R(Ed(X_i,\widehat{X}_i))\right) \\
&\geq nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i,\widehat{X}_i)\right) \\
&\quad \text{(convexity of the rate distortion function} \\
&\quad \text{and Jensen's inequality)} \\
&= nR(Ed(X^n,\widehat{X}^n)) \text{ (distortion for blocks of length } n) \\
&\geq nR(D). \text{ (R(D) nonincreasing and } Ed(X^n,\widehat{X}^n) \leq D)
\end{aligned}
$$

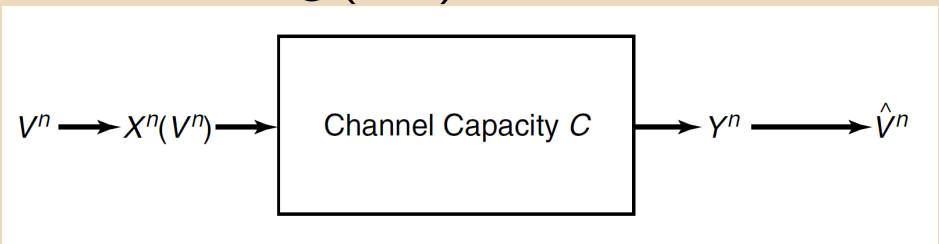
So the rate R of any rate distortion code exceeds the rate distortion function $R(D)$ at the level $D = Ed(X^n,\widehat{X}^n)$ achieved by that code.

Source-Channel Separation Theorem with Distortion

- A similar argument can be applied when the encoded source is passed through a noisy channel, which gives the equivalent of the source channel separation theorem with distortion.

Theorem (Source-Channel Separation Theorem with Distortion)

Let $V_1, V_2, \ldots, V_n$ be a finite alphabet i.i.d. source which is encoded as a sequence of n input symbols X^n of a discrete memoryless channel with capacity C. The output of the channel Y^n is mapped onto the reconstruction alphabet $\widehat{V}^n = g(Y^n)$.



Let $D = Ed(V^n, \widehat{V}^n) = \frac{1}{n}\sum_{i=1}^{n} Ed(V_i, \widehat{V}_i)$ be the average distortion achieved by this combined source and channel coding scheme. Then distortion D is achievable if and only if $C > R(D)$.

Subsection 5

Achievability of the Rate Distortion Function

Distortion Typical Sets

Definition

Let $p(x, \widehat{x})$ be a joint probability distribution on $\mathcal{X} \times \widehat{\mathcal{X}}$. Let $d(x, \widehat{x})$ be a distortion measure on $\mathcal{X} \times \widehat{\mathcal{X}}$. For any $\epsilon > 0$, a pair of sequences $(x^n, \widehat{x}^n)$ is said to be **distortion ϵ-typical** or simply **distortion typical** if

$$\begin{aligned}
\left| -\frac{1}{n} \log p(x^n) - H(X) \right| &< \epsilon, \\
\left| -\frac{1}{n} \log p(\widehat{x}^n) - H(\widehat{X}) \right| &< \epsilon, \\
\left| -\frac{1}{n} \log p(x^n, \widehat{x}^n) - H(X, \widehat{X}) \right| &< \epsilon, \\
\left| d(x^n, \widehat{x}^n) - Ed(X, \widehat{X}) \right| &< \epsilon.
\end{aligned}$$

The set of distortion typical sequences is called the **distortion typical set** and is denoted $A_{d,\epsilon}^{(n)}$.

Distortion Typical versus Jointly Typical

- The definition of distortion typical is the definition of the jointly typical set with the additional constraint that the distortion be close to the expected value.
- So the distortion typical set is a subset of the jointly typical set,

$$A_{d,\epsilon}^{(n)} \subseteq A_{\epsilon}^{(n)}.$$

- If $(X_i, \widehat{X}_i)$ are drawn i.i.d $\sim p(x, \widehat{x})$, the distortion between two random sequences

$$d(X^n, \widehat{X}^n) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, \widehat{X}_i)$$

  is an average of i.i.d. random variables.
- By the Law of Large Numbers, it is close to its expected value with high probability.

Asymptotic Probability of Distortion Typicality

Lemma

Let $(X_i, \widehat{X}_i)$ be drawn i.i.d. $\sim p(x, \widehat{x})$. Then $\Pr(A_{d,\epsilon}^{(n)}) \to 1$ as $n \to \infty$.

- The sums in the four conditions in the definition of $A_{d,\epsilon}^{(n)}$ are all normalized sums of i.i.d random variables.

  Hence, by the Law of Large Numbers, they tend to their respective expected values with probability 1.

  So the set of sequences satisfying all four conditions has probability tending to 1 as $n \to \infty$.

Probabilities for Pairs in Distortion Typical Sets

Lemma

For all $(x^n, \widehat{x}^n) \in A_{d,\epsilon}^{(n)}$,

$$p(\widehat{x}^n) = p(\widehat{x}^n|x^n)2^{-n(I(X;\widehat{X})+3\epsilon)}.$$

- Using the definition of $A_{d,\epsilon}^{(n)}$, we can bound the probabilities $p(x^n)$, $p(\widehat{x}^n)$ and $p(x^n, \widehat{x}^n)$, for all $(x^n, \widehat{x}^n) \in A_{d,\epsilon}^{(n)}$. Hence,

$$\begin{aligned} p(\widehat{x}^n|x^n) &= \frac{p(x^n,\widehat{x}^n)}{p(x^n)} \\ &= p(\widehat{x}^n)\frac{p(x^n,\widehat{x}^n)}{p(x^n)p(\widehat{x}^n)} \\ &\leq p(\widehat{x}^n)\frac{2^{-n(H(X,\widehat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)}2^{-n(H(\widehat{X})+\epsilon)}} \\ &= p(\widehat{x}^n)2^{n(I(X;\widehat{X})+3\epsilon)}. \end{aligned}$$

An Inequality

Lemma

For $0 \leq x, y \leq 1$, $n > 0$,

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

- Let $f(y) = e^{-y} - 1 + y$. Then:
  - $f(0) = 0$;
  - $f'(y) = -e^{-y} + 1 > 0$ for $y > 0$.

  Hence, $f(y) > 0$, for $y > 0$.

  So, for $0 \leq y \leq 1$, we have $1 - y \leq e^{-y}$.

  Raising this to the n-th power, we obtain $(1 - y)^n \leq e^{-yn}$.

  Thus, the lemma is satisfied for $x = 1$.

  It is clear that the inequality is also satisfied for $x = 0$.

An Inequality (Cont'd)

- Consider
$$g_y(x) = (1 - xy)^n.$$
Differentiating, we get:
  - $g_y'(x) = -ny(1 - xy)^{n-1}$;
  - $g_y''(x) = n(n-1)y^2(1 - xy)^{n-2}$.

So $g_y(x) = (1 - xy)^n$ is a convex function of x.

Hence, for $0 \leq x \leq 1$, we have

$$\begin{aligned}
(1 - xy)^n &= g_y(x) \\
&= g_y((1 - x)0 + x1) \\
&\leq (1 - x)g_y(0) + xg_y(1) \\
&= (1 - x)1 + x(1 - y)^n \\
&\leq 1 - x + xe^{-yn} \\
&\leq 1 - x + e^{-yn}.
\end{aligned}$$

Proof of Achievability

- Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim p(x)$.

  Let $d(x, \widehat{x})$ be a bounded distortion measure for this source.

  Let the rate distortion function for this source be $R(D)$.

  For any D, and any $R > R(D)$, we show that the rate distortion pair (R, D) is achievable by proving the existence of a sequence of rate distortion codes with rate R and asymptotic distortion D.

  Fix $p(\widehat{x}|x)$, where $p(\widehat{x}|x)$ achieves equality in the definition of $R(D)$.

  Thus, we have $I(X; \widehat{X}) = R(D)$.

  Calculate $p(\widehat{x}) = \sum_x p(x)p(\widehat{x}|x)$.

  Choose $\delta > 0$.

  We demonstrte the existence of a rate distortion code with rate R and distortion less than or equal to $D + \delta$.

Codebook, Encoding and Decoding

- **Generation of Codebook**:

  Randomly generate a rate distortion codebook $\mathcal{C}$ consisting of 2^{nR} sequences $\widehat{X}^n$ drawn i.i.d. $\sim \prod_{i=1}^{n} p(\widehat{x}_i)$.

  Index these codewords by $w \in \{1, 2, \ldots, 2^{nR}\}$.

  Reveal this codebook to the encoder and decoder.

- **Encoding**:

  Encode X^n by w if there exists a w such that $(X^n, \widehat{X}^n(w)) \in A_{d,\epsilon}^{(n)}$, the distortion typical set.

  If there is more than one such w, send the least.

  If there is no such w, let $w = 1$.

  Thus, nR bits suffice to describe the index w of the jointly typical codeword.

- **Decoding**:

  The reproduced sequence is $\widehat{X}^n(w)$.

Calculation of Distortion

- We calculate the expected distortion over the random choice of codebooks $\mathcal{C}$ as $\overline{D} = E_{X^n,\mathcal{C}} d(X^n, \widehat{X}^n)$, where the expectation is over the random choice of codebooks and over X^n.

  For a fixed codebook $\mathcal{C}$ and choice of $\epsilon > 0$, we divide the sequences $x^n \in \mathcal{X}^n$ into two categories:

  - Sequences x^n such that there exists a codeword $\widehat{X}^n(w)$ that is distortion typical with x^n, i.e., $d(x^n, \widehat{x}^n(w)) < D + \epsilon$.
    The total probability of these sequences is at most 1.
    So they contribute at most $D + \epsilon$ to the expected distortion.
  - Sequences x^n such that there does not exist a codeword $\widehat{X}^n(w)$ that is distortion typical with x^n.
    Let P_e be the total probability of these sequences.
    The distortion for any individual sequence is bounded by $d_{\max}$.
    So these contribute at most $P_e d_{\max}$ to the expected distortion.

  So the total distortion $Ed(X^n, \widehat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\max}$. This can be made $< D + \delta$ for an appropriate choice of ϵ if P_e is small enough.

Calculation of P_e (Setup)

- We must bound the probability P_e that, for a random choice of codebook $\mathcal{C}$ and a randomly chosen source sequence, there is no codeword that is distortion typical with the source sequence.
  Let $J(\mathcal{C})$ denote the set of source sequences x^n, such that at least one codeword in $\mathcal{C}$ is distortion typical with x^n. Then

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

  This is the probability of all sequences not well represented by a code, averaged over the randomly chosen code.
  By changing the order of summation, we can also interpret this as the probability of choosing a codebook that does not well represent sequence x^n, averaged with respect to $p(x^n)$. Thus,

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

Calculation of P_e

- Define

$$K(x^n, \widehat{x}^n) = \begin{cases} 1, & \text{if } (x^n, \widehat{x}^n) \in A_{d,\epsilon}^{(n)} \\ 0, & \text{if } (x^n, \widehat{x}^n) \notin A_{d,\epsilon}^{(n)} \end{cases}.$$

The probability that a single randomly chosen codeword $\widehat{X}^n$ does not well represent a fixed x^n is

$$\begin{aligned} \Pr((x^n, \widehat{X}^n) \notin A_{d,\epsilon}^{(n)}) &= \Pr(K(x^n, \widehat{X}^n) = 0) \\ &= 1 - \textstyle\sum_{\widehat{x}^n} p(\widehat{x}^n) K(x^n, \widehat{x}^n). \end{aligned}$$

Therefore, the probability that 2^{nR} independently chosen codewords do not represent x^n, averaged over $p(x^n)$, is

$$\begin{aligned} P_e &= \textstyle\sum_{x^n} p(x^n) \sum_{\mathcal{C}: x^n \notin J(\mathcal{C})} p(\mathcal{C}) \\ &= \textstyle\sum_{x^n} p(x^n) [1 - \sum_{x^n} p(\widehat{x}^n) K(x^n, \widehat{x}^n)]^{2^{nR}}. \end{aligned}$$

Calculation of P_e (Cont'd)

- By one of the preceding lemmas,

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

Hence,

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

By the preceding lemma, we get

$$\begin{aligned}&(1 - 2^{-n(I(X;\widehat{X})+3\epsilon)} \textstyle\sum_{\widehat{x}^n} p(\widehat{x}^n|x^n) K(x^n \widehat{x}^n))^{2^{nR}} \\ &\qquad \leq 1 - \textstyle\sum_{\widehat{x}^n} p(\widehat{x}^n|x^n) K(x^n, \widehat{x}^n) + e^{-(2^{-n(I(X;\widehat{X})+3\epsilon)} 2^{nR})}.\end{aligned}$$

So we obtain

$$P_e = 1 - \sum_{x^n} \sum_{\widehat{x}^n} p(x^n) p(\widehat{x}^n|x^n) K(x^n, \widehat{x}^n) + e^{-2^{-n(I(X;\widehat{X})+3\epsilon)} 2^{nR}}.$$

Calculation of P_e (Cont'd)

- We got

$$P_e = 1 - \sum_{x^n} \sum_{\widehat{x}^n} p(x^n) p(\widehat{x}^n | x^n) K(x^n, \widehat{x}^n) + e^{-2^{-n(I(X;\widehat{X})+3\epsilon)} 2^{nR}}.$$

The last term in the bound is equal to $e^{-2n(R-I(X;\widehat{X})-3\epsilon)}$.

This goes to zero exponentially fast with n if $R > I(X;\widehat{X}) + 3\epsilon$.

Hence, with our choice of $p(\widehat{x}|x)$, $R > R(D)$ implies $R > I(X;\widehat{X})$.

So we can choose ϵ small enough so that the last term goes to 0.

Calculation of P_e (Conclusion)

- The first two terms in

$$P_e = 1 - \sum_{x^n} \sum_{\widehat{x}^n} p(x^n) p(\widehat{x}^n | x^n) K(x^n, \widehat{x}^n) + e^{-2^{-n(I(X;\widehat{X})+3\epsilon)} 2^{nR}}$$

give the probability under the joint distribution $p(x^n, \widehat{x}^n)$ that the pair of sequences is not distortion typical.

Hence, by a previous lemma, we obtain, for n sufficiently large,

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

By choosing ϵ and n wisely, we can make P_e as small as we like.

In summary, for any choice of $\delta > 0$, there exists an ϵ and n, such that, over all randomly chosen rate R codes of block length n, the expected distortion is less than $D + \delta$.

Hence, there must exist at least one code $\mathcal{C}^*$ with this rate and block length with average distortion less than $D + \delta$.

Channel Coding for the Gaussian Channel

- Consider a Gaussian channel, $Y_i = X_i + Z_i$, where the Z_i are i.i.d. $\sim \mathcal{N}(0, N)$ and there is a power constraint P on the power per symbol of the transmitted codeword.

  Consider a sequence of n transmissions.

  The power constraint implies that the transmitted sequence lies within a sphere of radius $\sqrt{nP}$ in $\mathbb{R}^n$.

  The coding problem is equivalent to finding a set of 2^{nR} sequences within this sphere such that the probability of any of them being mistaken for any other is small.

  In other words, we would like the spheres of radius $\sqrt{nN}$ around each of the codewords to be almost disjoint.

  This corresponds to filling a sphere of radius $\sqrt{n(P+N)}$ with spheres of radius $\sqrt{nN}$.

Channel Coding for the Gaussian Channel (Cont'd)

- One would expect that the largest number of spheres that could be fit would be the ratio of their volumes, or, equivalently, the n-th power of the ratio of their radii.

  Thus, if M is the number of codewords that can be transmitted efficiently, we have

$$M \leq \frac{(\sqrt{n(P+N)})^n}{(\sqrt{nN})^n} = \left(\frac{P+N}{N}\right)^{\frac{n}{2}}.$$

  The results of the channel coding theorem show that it is possible to do this efficiently for large n.

  It is possible to find approximately $2^{nC} = \left(\frac{P+N}{N}\right)^{\frac{n}{2}}$ codewords, such that the noise spheres around them are almost disjoint (i.e., the total volume of their intersection is arbitrarily small).

Rate Distortion for the Gaussian Source

- Consider a Gaussian source of variance σ^2.

  A $(2^{nR}, n)$ rate distortion code for this source, with distortion D, is a set of 2^{nR} sequences in $\mathbb{R}^n$, such that most source sequences of length n (all those that lie within a sphere of radius $\sqrt{n\sigma^2}$) are within a distance $\sqrt{nD}$ of some codeword.

  By the sphere-packing argument, it is clear that the minimum number of codewords required is

$$2^{nR(D)} = \left(\frac{\sigma^2}{D}\right)^{\frac{n}{2}}.$$

  The rate distortion theorem shows that this minimum rate is asymptotically achievable.

  That is, there exists a collection of spheres of radius $\sqrt{nD}$ that cover the space except for a set of arbitrarily small probability.

Channel Transmission versus Rate Distortion

- The above geometric arguments also enable us to transform a good code for channel transmission into a good code for rate distortion.
- In both cases, the essential idea is to fill the space of source sequences.
  - In channel transmission, we want to find the largest set of codewords that have a large minimum distance between codewords.
  - In rate distortion, we wish to find the smallest set of codewords that covers the entire space.
- If we have any set that meets the sphere packing bound for one, it will meet the sphere packing bound for the other.
- In the Gaussian case, choosing the codewords to be Gaussian with the appropriate variance is asymptotically optimal for both rate distortion and channel coding.

Subsection 6

Strongly Typical Sequences and Rate Distortion

Strongly Typical Sequences

- Let $N(a|x^n)$ be the number of occurrences of the symbol a in the sequence x^n.

Definition

A sequence $x^n \in \mathcal{X}^n$ is said to be ϵ-**strongly typical** with respect to a distribution $p(x)$ on $\mathcal{X}$ if:

1. For all $a \in \mathcal{X}$ with $p(a) > 0$, we have

$$\left| \frac{1}{n} N(a|x^n) - p(a) \right| < \frac{\epsilon}{|\mathcal{X}|};$$

2. For all $a \in \mathcal{X}$, with $p(a) = 0$, $N(a|x^n) = 0$.

The set of sequences $x^n \in \mathcal{X}^n$, such that x^n is strongly typical is called the **strongly typical set** and is denoted $A_\epsilon^{*(n)}(X)$ or $A_\epsilon^{*(n)}$, when the random variable is understood from the context.

Strongly Typical Pairs

- Let $N(a, b|x^n, y^n)$ be the number of occurrences of the pair (a, b) in the pair of sequences (x^n, y^n).

Definition

A pair of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is said to be ϵ-**strongly typical** with respect to a distribution $p(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ if:

1. For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) > 0$, we have

$$\left| \frac{1}{n} N(a, b|x^n, y^n) - p(a, b) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|};$$

2. For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$, with $p(a, b) = 0$, $N(a, b|x^n, y^n) = 0$.

The set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$, such that (x^n, y^n) is strongly typical is called the **strongly typical set** and is denoted $A_\epsilon^{*(n)}(X, Y)$ or $A_\epsilon^{*(n)}$.

Some Consequences

- From the definition, it follows that if $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$, then $x^n \in A_\epsilon^{*(n)}(X)$.
- By the Strong Law of Large Numbers, we have the following

Lemma

Let (X_i, Y_i) be drawn i.i.d. $\sim p(x, y)$. Then $\Pr(A_\epsilon^{*(n)}) \to 1$ as $n \to \infty$.

Lemma

Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim p(y)$. For $x^n \in A_\epsilon^{*(n)}(X)$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by

$$2^{-n(I(X;Y)+\epsilon_1)} \leq \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_1)},$$

where ϵ_1 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Achievability of the Rate Distortion Function

- Fix $p(\widehat{x}|x)$. Calculate $p(\widehat{x}) = \sum_x p(x)p(\widehat{x}|x)$.

  Fix $\epsilon > 0$. Later we will choose ϵ appropriately to achieve an expected distortion less than $D + \delta$.
- **Generation of Codebook**: Generate a rate distortion codebook $\mathcal{C}$ consisting of 2^{nR} sequences $\widehat{X}^n$ drawn i.i.d. $\sim \prod_i p(\widehat{x}_i)$. Denote the sequences $\widehat{X}^n(1), \ldots, \widehat{X}^n(2^{nR})$.
- **Encoding**: Given a sequence X^n, index it by w if there exists a w such that $(X^n, \widehat{X}^n(w)) \in A_\epsilon^{*(n)}$, the strongly jointly typical set.

  If there is more than one such w, send the first in lexicographic order.

  If there is no such w, let $w = 1$.
- **Decoding**: Let the reproduced sequence be $\widehat{X}^n(w)$.

Calculation of Distortion

- We calculate the expected distortion over the random choice of codebook:

$$\begin{aligned} D &= E_{X^n,\mathcal{C}}\, d(X^n, \widehat{X}^n) \\ &= E_{\mathcal{C}} \sum_{x^n} p(x^n) d(x^n, \widehat{X}^n(x^n)) \\ &= \sum_{x^n} p(x^n) E_{\mathcal{C}}\, d(x^n, \widehat{X}^n), \end{aligned}$$

  where the expectation $E_{\mathcal{C}}$ is over the random choice of codebook.

  For a fixed codebook $\mathcal{C}$, we divide the sequences $x^n \in \mathcal{X}^n$ into three categories.

  - **Nontypical sequences** $x^n \notin A_\epsilon^{*(n)}$. The total probability of these sequences can be made less than ϵ by choosing n large enough. The individual distortion between two sequences is bounded by $d_{\max}$. So the nontypical sequences can contribute at most $\epsilon d_{\max}$ to the expected distortion.

Calculation of Distortion (Cont'd)

- **Typical sequences $x^n \in A_\epsilon^{*(n)}$, such that there exists a codeword $\widehat{X}^n(w)$ that is jointly typical with x^n.** In this case, the source sequence and the codeword are strongly jointly typical.
  Also, distortion is continuous as a function of the joint distribution.
  So the source sequence and the codeword are also distortion typical.
  Hence, the distortion between them is bounded by $D + \epsilon d_{\max}$.
  The total probability of these sequences is at most 1.
  So they contribute at most $D + \epsilon d_{\max}$ to the expected distortion.
- **Typical sequences $x^n \in A_\epsilon^{*(n)}$, such that there does not exist a codeword $\widehat{X}^n$ that is jointly typical with x^n.**
  Let P_e be the total probability of these sequences.
  The distortion for any individual sequence is bounded by $d_{\max}$.
  So these sequences contribute at most $P_e d_{\max}$ to the expected distortion.

The probability of the first category of sequences is less than ϵ for sufficiently large n. The probability of the last category is P_e, which we will show can be made small.

Calculation of P_e

- We must bound the probability that there is no codeword that is jointly typical with the given sequence X^n.

  From the joint AEP, we know that the probability that X^n and any $\widehat{X}^n$ are jointly typical is $\doteq 2^{-nI(X;\widehat{X})}$.

  Hence, the expected number of jointly typical $\widehat{X}^n(w)$ is $2^{nR}2^{-nI(X;\widehat{X})}$.

  This is exponentially large if $R > I(X;\widehat{X})$.

  But this is not sufficient to show that $P_e \to 0$.

  We must show that the probability that there is no codeword that is jointly typical with X^n goes to zero.

  The fact that the expected number of jointly typical codewords is exponentially large does not ensure that there will at least one with high probability.

Calculation of P_e (Cont'd)

- We can expand the probability of error as

$$P_e = \sum_{x^n \in A_\epsilon^{*(n)}} p(x^n)[1 - \Pr((x^n, \widehat{X}^n) \in A_\epsilon^{*(n)})]^{2^{nR}}.$$

By a previous lemma, we have $\Pr((x^n, \widehat{X}^n) \in A_\epsilon^{*(n)}) \geq 2^{-n(I(X;\widehat{X})+\epsilon_1)}$.
Substituting this in and using the inequality $(1-x)^n \leq e^{-nx}$, we have

$$P_e \leq e^{-(2^{nR}2^{-n(I(X;\widehat{X})+\epsilon_1)})}.$$

This goes to 0 as $n \to \infty$ if $R > I(X;\widehat{X}) + \epsilon_1$.
Hence, for an appropriate choice of ϵ and n, we can get the total probability of all badly represented sequences to be as small as we want.
Not only is the expected distortion close to D, but with probability going to 1, we will find a codeword whose distortion with respect to the given sequence is less than $D + \delta$.

Subsection 7

Characterization of the Rate Distortion Function

Minimization Problem

- We have defined the information rate distortion function as

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

  where the minimization is over all conditional distributions $q(\widehat{x}|x)$ for which the joint distribution $p(x)q(\widehat{x}|x)$ satisfies the expected distortion constraint.
- This is a standard minimization problem of a convex function over the convex set of all $q(\widehat{x}|x) \geq 0$ satisfying

$$\sum_{\widehat{x}} q(\widehat{x}|x) = 1, \quad \text{for all } x,$$

  and

$$\sum q(\widehat{x}|x)p(x)d(x,\widehat{x}) \leq D.$$

Lagrange Multipliers

- We can use the method of Lagrange multipliers to find the solution.
  We set up the functional

$$\begin{aligned} J(q) &= \sum_x \sum_{\widehat{x}} p(x) q(\widehat{x}|x) \log \frac{q(\widehat{x}|x)}{\sum_x p(x) q(\widehat{x}|x)} \\ &\quad + \lambda \sum_x \sum_{\widehat{x}} p(x) q(\widehat{x}|x) d(x, \widehat{x}) \\ &\quad + \sum_x \nu(x) \sum_{\widehat{x}} q(\widehat{x}|x), \end{aligned}$$

  where the last term corresponds to the constraint that $q(\widehat{x}|x)$ is a conditional probability mass function.
  If we let $q(\widehat{x}) = \sum_x p(x) q(\widehat{x}|x)$ be the distribution on $\widehat{X}$ induced by $q(\widehat{x}|x)$, we can rewrite $J(q)$ as

$$\begin{aligned} J(q) &= \sum_x \sum_{\widehat{x}} p(x) q(\widehat{x}|x) \log \frac{q(\widehat{x}|x)}{q(\widehat{x})} \\ &\quad + \lambda \sum_x \sum_{\widehat{x}} p(x) q(\widehat{x}|x) d(x, \widehat{x}) \\ &\quad + \sum_x \nu(x) \sum_{\widehat{x}} q(\widehat{x}|x). \end{aligned}$$

Minimization

- Differentiating with respect to $q(\widehat{x}|x)$, we have

$$\begin{aligned}\frac{\partial J}{\partial q(\widehat{x}|x)} &= p(x)\log\frac{q(\widehat{x}|x)}{q(\widehat{x})} + p(x) - \sum_{x'} p(x')q(\widehat{x}|x')\frac{1}{q(\widehat{x})}p(x) \\ &\quad +\lambda p(x)d(x,\widehat{x}) + \nu(x) = 0.\end{aligned}$$

Setting $\log\mu(x) = \frac{\nu(x)}{p(x)}$, we obtain

$$p(x)\left[\log\frac{q(\widehat{x}|x)}{q(\widehat{x})} + \lambda d(x,\widehat{x}) + \log\mu(x)\right] = 0.$$

So

$$q(\widehat{x}|x) = \frac{q(\widehat{x})e^{-\lambda d(x,\widehat{x})}}{\mu(x)}.$$

Minimization (Cont'd)

- We found $q(\widehat{x}|x) = \frac{q(\widehat{x})e^{-\lambda d(x,\widehat{x})}}{\mu(x)}$.

  Now $\sum_{\widehat{x}} q(\widehat{x}|x) = 1$.

  So we must have $\mu(x) = \sum_{\widehat{x}} q(\widehat{x})e^{-\lambda d(x,\widehat{x})}$.

  Therefore, $q(\widehat{x}|x) = \frac{q(\widehat{x})e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}} q(\widehat{x})e^{-\lambda d(x,\widehat{x})}}$.

  Multiply this by $p(x)$ and sum over all x,

$$q(\widehat{x}) = q(\widehat{x}) \sum_x \frac{p(x)e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}'} q(\widehat{x}')e^{-\lambda d(x,\widehat{x}')}}.$$

  If $q(\widehat{x}) > 0$, divide both sides by $q(\widehat{x})$ to get, for all $\widehat{x} \in \widehat{\mathcal{X}}$,

$$\sum_x \frac{p(x)e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}'} q(\widehat{x}')e^{-\lambda d(x,\widehat{x}')}} = 1.$$

Minimization (Conclusion)

- We can combine the $|\widehat{\mathcal{X}}|$ equations

$$\sum_x \frac{p(x)e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}'} q(\widehat{x}')e^{-\lambda d(x,\widehat{x}')}} = 1, \quad \widehat{x} \in \widehat{\mathcal{X}},$$

with the equation defining the distortion and calculate λ and the $|\widehat{\mathcal{X}}|$ unknowns $q(\widehat{x})$.

We can use this and

$$q(\widehat{x}|x) = \frac{q(\widehat{x})e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}} q(\widehat{x})e^{-\lambda d(x,\widehat{x})}}$$

to find the optimum conditional distribution.

The Case $q(\widehat{x}) = 0$

- The inequality condition $q(\widehat{x}) > 0$ is covered by the Kuhn-Tucker conditions, which reduce to

$$\begin{aligned} \frac{\partial J}{\partial q(\widehat{x}|x)} &= 0 \quad \text{if } q(\widehat{x}|x) > 0 \\ &\geq 0 \quad \text{if } q(\widehat{x}|x) = 0. \end{aligned}$$

  Substituting the value of the derivative, we obtain the conditions for the minimum as

$$\begin{aligned} \sum_x \frac{p(x)e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}'} q(\widehat{x}')e^{-\lambda d(x,\widehat{x}')}} &= 1 \quad \text{if } q(\widehat{x}) > 0 \\ &\leq 1 \quad \text{if } q(\widehat{x}) = 0. \end{aligned}$$

- This characterization will enable us to check if a given $q(\widehat{x})$ is a solution to the minimization problem.
- However, it is not easy to solve for the optimum output distribution from these equations.

Subsection 8

Computation of Channel Capacity and Rate Distortion Function

A Minimization Problem

- Consider the problem: Given two convex sets A and B in $\mathbb{R}^n$, find the minimum distance between them: $d_{\min} = \min_{a \in A, b \in B} d(a, b)$, where $d(a, b)$ is the Euclidean distance between a and b.
- An intuitively obvious algorithm to do this would be to:
  - Take any point $x \in A$, and find the $y \in B$ that is closest to it.
  - Fix this y and find the closest point in A.
  - Repeating, it is clear that the distance decreases at each stage.
- Csiszár and Tusnády have shown that if the sets are convex and if the distance satisfies certain conditions, this alternating minimization algorithm will indeed converge to the minimum.
- In particular, if the sets are sets of probability distributions and the distance measure is the relative entropy, the algorithm does converge to the minimum relative entropy between the two sets of distributions.

Distribution Minimizing Relative Entropy

Lemma

Let $p(x)p(y|x)$ be a given joint distribution. Then the distribution $r(y)$ that minimizes the relative entropy $D(p(x)p(y|x)||p(x)r(y))$ is the marginal distribution $r^*(y)$ corresponding to $p(y|x)$:

$$D(p(x)p(y|x)||p(x)r^*(y)) = \min_{r(y)} D(p(x)p(y|x)||p(x)r(y)),$$

where $r^*(y) = \sum_x p(x)p(y|x)$. Also,

$$\max_{r(x|y)} \sum_{x,y} p(x)p(y|x) \log \frac{r(x|y)}{p(x)} = \sum_{x,y} p(x)p(y|x) \log \frac{r^*(x|y)}{p(x)},$$

where

$$r^*(x|y) = \frac{p(x)p(y|x)}{\sum_x p(x)p(y|x)}.$$

Proof of the Lemma

- We have

$$
\begin{aligned}
& D(p(x)p(y|x)||p(x)r(y)) - D(p(x)p(y|x)||p(x)r^*(y)) \\
&= \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)r(y)} - \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)r^*(y)} \\
&= \sum_{x,y} p(x)p(y|x) \log \frac{r^*(y)}{r(y)} \\
&= \sum_{y} r^*(y) \log \frac{r^*(y)}{r(y)} \\
&= D(r^*||r) \geq 0.
\end{aligned}
$$

The proof of the second part of the lemma is similar.

Rewriting Rate Distortion Function

- We can use the lemma to rewrite the minimization in the definition of the rate distortion function as a double minimization,

$$R(D) = \min_{r(\widehat{x})} \min_{q(\widehat{x}|x):\sum p(x)q(\widehat{x}|x)d(x,\widehat{x})\leq D} \sum_{x}\sum_{\widehat{x}} p(x)q(\widehat{x}|x)\log\frac{q(\widehat{x}|x)}{r(\widehat{x})}.$$

- Now let:
  - A be the set of all joint distributions with marginal $p(x)$ that satisfy the distortion constraints;
  - B be the set of product distributions $p(x)r(\widehat{x})$ with arbitrary $r(\widehat{x})$.

  Then we can write

$$R(D) = \min_{q\in B}\min_{p\in A} D(p\|q).$$

Alternating Minimization (Blahut-Arimoto Algorithm)

- We apply alternating minimization (Blahut-Arimoto algorithm):
  - We begin with a choice of λ and an initial output distribution $r(\widehat{x})$ and calculate the $q(\widehat{x}|x)$ that minimizes the mutual information subject to the distortion constraint.
    Using Lagrange multipliers for this minimization, we obtain
    $$q(\widehat{x}|x) = \frac{r(\widehat{x})e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}} r(\widehat{x})e^{-\lambda d(x,\widehat{x})}}.$$
  - For this conditional distribution $q(\widehat{x}|x)$, we calculate the output distribution $r(\widehat{x})$ that minimizes the mutual information.
    By the lemma, this is $r(\widehat{x}) = \sum_x p(x)q(\widehat{x}|x)$.
  - We use this distribution as the starting point of the next iteration.
- Each step, minimizing over $q(\cdot|\cdot)$ and then over $r(\cdot)$, reduces the right-hand side of $R(D)$. Thus, there is a limit.
- This limit has been shown to be $R(D)$ by Csiszár, where the value of D and $R(D)$ depends on λ.
- Choosing λ appropriately sweeps out the $R(D)$ curve.

Alternate Minimization for Channel Capacity

- We rewrite the definition of channel capacity,

$$C = \max_{r(x)} I(X;Y) = \max_{r(x)} \sum_x \sum_y r(x)p(y|x) \log \frac{r(x)p(y|x)}{r(x)\sum_{x'} r(x')p(y|x')}$$

as a double maximization using the lemma

$$C = \max_{q(x|y)} \max_{r(x)} \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)}.$$

- The Csiszár-Tusnády algorithm is one of alternating maximization:
  - We start with a guess of the maximizing distribution $r(x)$.
    We find the best conditional distribution $q(x|y) = \frac{r(x)p(y|x)}{\sum_x r(x)p(y|x)}$.
  - For this distribution, we find the best input distribution $r(x)$ by solving the constrained maximization problem with Lagrange multipliers.
    The optimum input distribution is $r(x) = \frac{\prod_y (q(x|y))^{p(y|x)}}{\sum_x \prod_y (q(x|y))^{p(y|x)}}$.
  - This is used as the basis for the next iteration.