Rate distortion theory

An introduction

Paulo J S G Ferreira

SPL — Signal Processing Lab / IEETA
Univ. Aveiro, Portugal

April 23, 2010

Contents

Contents

Contents

Contents

Contents

Contents

Entropy

- The entropy of a discrete random variable X with probability density $p(x)$ is

$$H(X) = -\sum_{x} p(x) \log p(x)$$

- Base two logarithms tacitly used throughout the lecture
- The term

$$-\log p(x)$$

  is associated with the uncertainty of X
- The entropy is the average uncertainty:

$$H(X) = E_{p(x)} \log \frac{1}{p(x)}$$

- Exercise: show that $0 \leq H(X) \leq \log n$ and identify the distributions for which the bounds are attained

Relative entropy

- The relative entropy measures the “distance” between two probability functions:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)}$$

- It is nonnegative and zero if and only if $p = q$
- It is not symmetric and it does not satisfy the triangle inequality
- Hence, it is not a metric
- Sometimes called the Kullback-Leibler distance
- Exercise: show that $D(p||q) \geq 0$

Conditional entropy

- $H(X|Y)$ is the expected value of the entropies of the conditional distributions $p(x|y)$, averaged over the conditioning variable Y:

$$\begin{aligned}
H(X|Y) &= \sum_y p(y)\, H(X|Y=y) \\
&= -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \\
&= -\sum_{x,y} p(y)\, p(x|y) \log p(x|y) \\
&= -\sum_{x,y} p(x,y) \log p(x|y) \\
&= -E_{p(x,y)} \log p(x|y)
\end{aligned}$$

Mutual information

- It is the Kullback-Leibler distance between the joint distribution $p(x, y)$ and $p(x)p(y)$:

$$I(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = E_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)}$$

- Meaning:
  - How far from independent X and Y are
  - The information that X contains about Y
- Exercise: show that $I(X, Y) \geq 0$

Mutual information and entropy

$$
\begin{aligned}
I(X,Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)p(y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\
&= -\sum_{x} p(x) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

- Meaning: measures the reduction in the uncertainty of X due to knowing Y

Conditional entropy and entropy

- We have seen that

$$I(X,Y) = H(X) - H(X|Y)$$

- We have also seen that $I(X,Y)$ is nonnegative
- It follows that

$$H(X) \geq H(X|Y)$$

- Meaning: on the average, adding more information does not increase the uncertainty

Chain rules

- The simplest is

$$H(X,Y) = H(X) + H(X|Y)$$

- Compare with $p(x,y) = p(y|x)\,p(x)$
- By repeated application

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) \\ &= H(X) + H(Y|X) + H(Z|X,Y) \end{aligned}$$

- The general case should be apparent:

$$H(X_1, X_2, \ldots X_n) = \sum H(X_i|X_1, \ldots X_{i-1})$$

Differential entropy

- The continuous case is mathematically much more subtle
- The differential entropy of X with probability density $f(x)$ is

$$h(X) = -\int_{-\infty}^{+\infty} f(x) \log f(x)\, dx$$

- Log: base e, unit: nats
- Example: the differential entropy of a uniform random variable is

$$h(X) = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a}\, dx = \log(b-a)$$

- It can be negative!

Gaussian variable

- The differential entropy of a Gaussian variable with density $g(x)$ is

$$H(X) = -\int_{-\infty}^{+\infty} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{g(x)} \log \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{g(x)} \, dx$$

$$= -\log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{2\sigma^2} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{g(x)} \, dx$$

$$= \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} \log e = \frac{1}{2} \log 2\pi e \sigma^2$$

Gaussian variable

- The differential entropy of a Gaussian variable is maximum among all densities with the same variance
- Proof depends on

$$\int_{-\infty}^{+\infty} f(x) \log \frac{f(x)}{g(x)} \, dx \geq 0$$

- Recall relative entropy / Kullback-Leibler distance to see why

Contents

Main ideas

- Typicality: describe "typical" sequences
- Forms of the asymptotic equipartition property
- They are to information theory what the law of large numbers is to probability
- In fact, they depend on the law of large numbers
- Weak typicality requires that "empirical entropy" approaches "true entropy"
- Strong typicality requires that the relative frequency of each outcome approaches the probability

Weak typicality

- Let X^n be an i.i.d. sequence $X_1, X_2, \ldots, X_n$
- It follows that

$$p(X^n) = p(X_1)p(X_2)\cdots p(X_n)$$

- The weakly typical set is formed by sequences such that

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \epsilon$$

- The term on the left is the empirical entropy
- The probability of the typical set approaches 1 as $n \to \infty$
- This does not mean that “most sequences are typical” but rather that the non-typical sequences have small probability
- The most likely sequence may not be weakly typical

Strong typicality

Given $\epsilon > 0$, a sequence is strongly typical with respect to $p(x)$ if the average number of occurrences of each symbol a in the sequence deviates from its probability by less than ϵ:

$$\left| \frac{N(a)}{n} - p(a) \right| < \epsilon$$

Also required: symbols of zero probability do not occur: $N(a) = 0$ if $p(a) = 0$

Strong typicality

- It is a stronger than weak typicality
- It allows better probability bounds
- The probability of a non-typical sequence not only goes to zero, it satisfies an exponential bound:

$$P[X \neq T(\epsilon)] \leq 2^{-n\phi(\epsilon)}$$

  where ϕ is a positive function
- Proof depends on Chernoff-type bound: $u(x-a) \leq 2^{b(x-a)}$ yields

$$E[u(X-a)] = P(X \geq a) \leq E[2^{b(X-a)}] = 2^{-ba}E[2^{bX}]$$

Contents

Rate distortion theory

- Lossy and lossless compression
- Lossy compression implies distortion
- Rate distortion theory describes the *trade-off between lossy compression rate and the corresponding distortion*

Representation of continuous variables

- Shannon wrote in Part V of *A Mathematical Theory of Communication*:

  ...a continuously variable quantity can assume an infinite number of values and requires, therefore, an infinite number of bits for exact specification.

- This means that...

  ...to transmit the output of a continuous source with exact recovery at the receiving point requires, in general, a channel of infinite capacity (in bits per second). Since, ordinarily, channels have a certain amount of noise, and therefore a finite capacity, exact transmission is impossible.

Is everything lost?

- Still quoting Shannon:

  Practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a certain tolerance.

- This leads to the real issue:

  The question is, can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way.

Yet another angle

- Consider a source with entropy rate H
- Source coding theorem: there are good source codes of rate R if $R > H$
- What if the available rate is below H?
- Source coding theorem converse: error probability tends to 1 as n increases – bad news
- What is necessary is a rate distortion code that reproduces the sequence with a certain allowed distortion

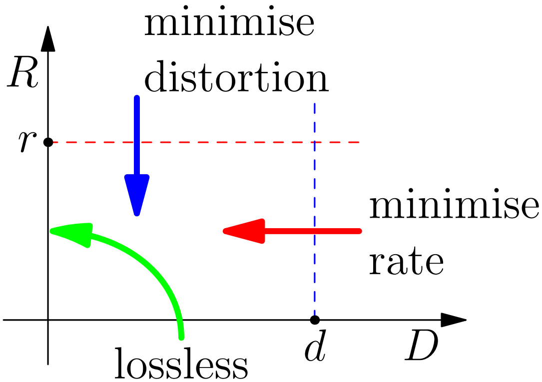
The problem

- Find a way of measuring distortion
- Determine the minimum necessary rate for a certain given distortion
- As the distortion requirements are increased the rate will increase
- It turns out that it is possible to define a rate such that:
    - Proper encoding makes possible transmission over a channel with capacity equal to that rate, at that distortion
    - A channel of smaller capacity is insufficient
- This lecture turns around this problem

Specific cases

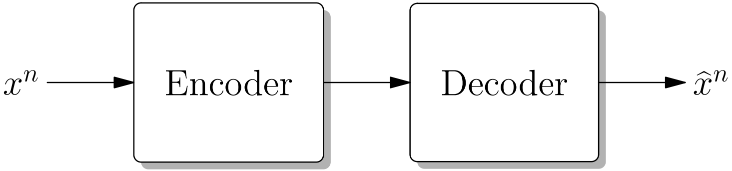
- Lossless case
  - Special case, zero distortion, source coding theorem, etc.
- Low rate / high compression
  - If the maximum distortion has been fixed, what is the minimum rate?
- Low distortion
  - What is the minimum distortion that can be achieved for a certain channel?

Rate distortion plane



Terminology

- There is a source and a reproduction alphabet
- x^n is the source sequence, with n symbols $x_1, x_2, \ldots, x_n$
- $\widehat{x}^n$ is the reproduced sequence, also with n symbols



Distortion measures — symbols

- How far are the data and their representations?
- Distortion measures answer this
- They are functions $d(x, \widehat{x})$ on the pairs of symbols $x, \widehat{x}$
- The functions take nonnegative values
- The are zero on the "diagonal": $d(x, x) = 0$
- They measure the cost of replacing x with $\widehat{x}$

Average distortion

- The average distortion is

$$E_{p(x,\widehat{x})} d(x,\widehat{x})$$

- Note that

$$\sum_{x,\widehat{x}} p(x,\widehat{x}) d(x,\widehat{x}) = \sum_{x,\widehat{x}} p(x)\, p(\widehat{x}|x)\, d(x,\widehat{x})$$

- There are three terms:
  - $d(x,\widehat{x})$, determined by the per-symbol distance
  - $p(x)$, determined by the source
  - $p(\widehat{x}|x)$, determined by the coding procedure
- Corollary: vary $p(\widehat{x}|x)$ to find interesting codes

Distortion measures — example

- The absolute value distortion measure is

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

- The average absolute value distortion is

$$E_{p(x,\hat{x})}|x - \hat{x}|$$

Distortion measures — example

- The squared error distortion measure is

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

- The average squared distortion is

$$E_{p(x,\widehat{x})}(x - \widehat{x})^2$$

Distortion measures — example

- The Hamming distance is

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

- The average Hamming distortion is

$$E_{p(x,\widehat{x})} d(x,\widehat{x})$$

- Easy to simplify: equal to

$$p(x \neq \widehat{x})$$

Distortion measures — sequences

- One way of measuring the distortion between the sequences is

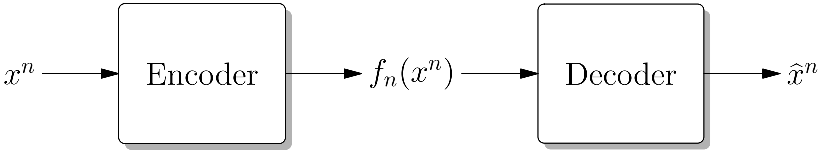
$$d(x^n, \widehat{x}^n) = \frac{1}{n} \sum d(x_i, \widehat{x}_i)$$

- This is simply the distortion per symbol
- Another possibility: the ℓ^∞ (max) norm of the $d(x_i, \widehat{x}_i)$
- Other norms possible, but seldom used

Rate distortion code

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

- An encoding function $f_n : x^n \to \{1, 2, \ldots, 2^{nR}\}$
- A decoding function $g_n : \{1, 2, \ldots, 2^{nR}\} \to \widehat{x}^n$



Why sequences?

- We are representing n symbols — i.i.d. random variables — with nR bits
- The encoder replaces the sequence with an index
- The decoder does the opposite
- There are, of course, 2^{nR} possibilities
- Surprisingly, it is better to represent entire sequences at one go than to treat each symbol separately — even though they are chosen i.i.d.

Code distortion

- The code distortion is

$$D = E\left[d(x^n, \widehat{x}^n)\right] = E\left[d(x^n, g_n(f_n(x^n)))\right] = \sum p(x^n) d(x^n, g_n(f_n(x^n)))$$

- The average is over the probability distribution on the sequences

The rate distortion region

- A point (R, D) in the rate-distortion plane is achievable if there exists a sequence of rate distortion codes (f_n, g_n) such that
$$\lim_{n\to\infty} E\left[d(x^n, g_n(f_n(x^n)))\right] \le D$$
- The rate distortion region is the closure of the set of achievable (R, D)
- Its boundary is important (why?)
- There are two (equivalent) ways of looking at it

Rate distortion function

- Roughly speaking: given D, search for the smallest achievable rate
- This defines a function of D (it yields a rate for every given D)
- This function is the rate distortion function
- More precisely: the rate distortion function $R(D)$ is the infimum of rates R such that (R, D) is in the rate distortion region C for a given distortion D

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

Distortion rate function

- "The rate distortion function, with a twist"
- Given R, search for the smallest achievable distortion
- This defines a function of R (it yields a distortion for every given R)
- More precisely: the distortion rate function $D(R)$ is the infimum of distortions D such that (R, D) is in the rate distortion region C for a given rate R

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

Information rate distortion

- Consider a source X with a distortion measure $d(x, \hat{x})$
- The information rate distortion function $R_I(D)$ for X is

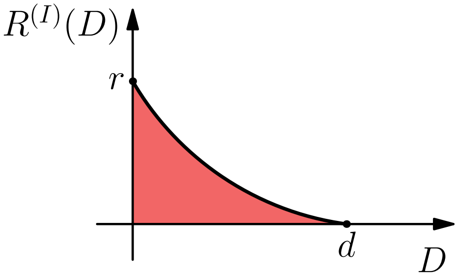
$$R_I(D) = \min I(X, \hat{X})$$

- Minimisation: over all conditional distributions $p(\hat{x}|x)$ such that $E[d(x, \hat{x})] \leq D$
- Recall that $p(x, \hat{x}) = p(x)p(\hat{x}|x)$
- We thus want the minimum over all $p(\hat{x}|x)$ such that

$$\sum_{x,\hat{x}} p(x)\, p(\hat{x}|x)\, d(x, \hat{x}) \leq D$$

Information rate distortion

- $R_I(D)$ is obviously nonnegative
- It is also nonincreasing
- It is convex:



Contents

Rate distortion theorem

- If the rate R is above $R(D)$, there exists a sequence of codes $\widehat{X}^n(X^n)$ with at most 2^{nR} codewords with an average distortion approaching D:

$$E[d(X^n, \widehat{X}^n)] \to D$$

- If the rate R is below $R(D)$, no such codes exist

Rate distortion theorem

- The rate distortion and the information rate distortion functions are equal:

$$R(D) = R_I(D)$$

- More explicitly,

$$R(D) = \min_{p(\widehat{x}|x): E[d(x,\widehat{x})] \le D} I(X, \widehat{X})$$

- Approach: show that $R(D) \ge R_I(D)$ and $R(D) \le R_I(D)$

Example: binary source

- Consider a binary source with

$$P(X = 1) = p, \quad P(X = 0) = q$$

- Distortion: Hamming distance
- Need to find

$$R(D) = \min I(X, \widehat{X})$$

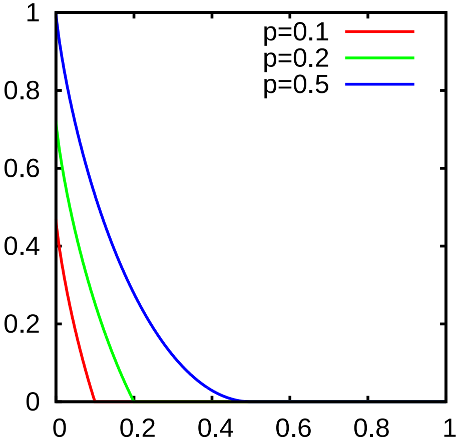
- The minimum is computed with respect to all $p(\widehat{x}|x)$ that satisfy the distortion constraint:

$$\sum_{x,\widehat{x}} p(x)\, p(\widehat{x}|x) d(x, \widehat{x}) \le D$$

$R(D)$ for the binary source

- The rate distortion function for the Hamming distance is

$$R(D) = \begin{cases} H(p) - H(D), & 0 \le D \le \min(p, q) \\ 0, & \text{otherwise} \end{cases}$$



Proof

- The idea is to lower bound $I(X, \widehat{X})$, then show that it is achievable
- We know that
$$I(X, \widehat{X}) = H(X) - H(X|\widehat{X})$$
- The first term $H(X)$ is just the entropy
$$H_b(p) = -p \log p - q \log q$$
- How to handle the second term $H(X|\widehat{X})$?
- Assume without loss of generality that $p \leq 1/2$
- If $D > p$, rate zero is sufficient (why?)
- Therefore assume $D < p \leq 1/2$

Proof (continued)

- Let $\oplus$ denote addition modulo 2 (that is, exclusive-or)
- Note (why?)

$$H(X|\widehat{X}) = H(X \oplus \widehat{X}|\widehat{X})$$

- Conditioning cannot increase entropy:

$$H(X \oplus \widehat{X}|\widehat{X}) \leq H(X \oplus \widehat{X})$$

- $H(X \oplus \widehat{X})$ is $H_b(a)$, where $a =$ probability of $X \oplus \widehat{X} = 1$
- $X \oplus \widehat{X} = 1$ if and only if $X \neq \widehat{X}$
- The probability of $X \neq \widehat{X}$ does not exceed D due to the distortion constraint, thus

$$H(X \oplus \widehat{X}) \leq H(D)$$

Proof (continued)

- Putting it all together, if $D < p \leq 1/2$,

$$I(X, \widehat{X}) = H(X) - H(X|\widehat{X}) \geq H(p) - H(D)$$

- The bound is achieved for

$$P(X = 0|\widehat{X} = 1) = P(X = 1|\widehat{X} = 0) = D$$

Example: Gaussian source

- Source: Gaussian source, zero mean, variance σ^2
- Quadratic distortion:

$$d(x, y) = (x - y)^2$$

- Need to find

$$R(D) = \min I(X, \widehat{X})$$

- The minimum must be computed subject to the constraint

$$E[(X - \widehat{X})^2] \leq D$$

Solution for $D \geq \sigma^2$

- Claim: for $D \geq \sigma^2$ the minimum rate is zero
- How to get rate zero: set $\widehat{X} = 0$
- Then no bits need to be transmitted and $I(X, \widehat{X}) = 0$ (why?)
- This leads to a distortion of $E[(X - \widehat{X})^2] = E[X^2] = \sigma^2$
- This value does not exceed D and so satisfies the constraint
- Still need to consider the case $D < \sigma^2$

Solution for $D < \sigma^2$

$$\begin{aligned} I(X,\widehat{X}) &= h(X) - h(X|\widehat{X}) \\ &= h(X) - h(X-\widehat{X}|\widehat{X}) \\ &\geq h(X) - h(X-\widehat{X}) \end{aligned}$$

- To minimise this, maximise $h(X-\widehat{X})$; thus $X-\widehat{X} =$ Gaussian
- Constraint: $E[(X-\widehat{X})^2] \leq D$, hence variance $X-\widehat{X}$ can be D
- Result:

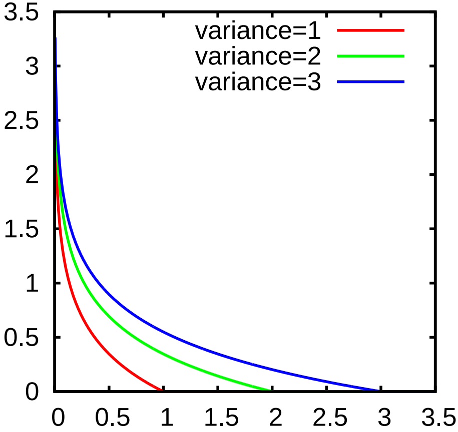
$$I(X,\widehat{X}) \geq \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi eD = \frac{1}{2}\log\frac{\sigma^2}{D}$$

- Bound met if X is Gaussian, zero mean, variance D

$R(D)$ for the Gaussian source

- Thus, the rate distortion for the Gaussian source is

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



Distortion rate for the Gaussian source

- Solve for D in

$$R = \frac{1}{2} \log \frac{\sigma^2}{D}$$

- This leads to

$$D(R) = \sigma^2 \, 2^{-2R}$$

- Increasing R by one bit decreases the distortion by 1/4
- Define SNR as

$$\text{SNR} = 10 \log_{10} \frac{\sigma^2}{D}$$

- Then

$$\text{SNR} = 10 \log_{10} 2^{2R} \approx 6R \text{ dB}$$

- Hence SNR varies at the rate of 6 dB per bit

Example: multiple Gaussian variables

- How to distribute R bits among n variables $\mathcal{N}(0, \sigma_i^2)$?
- In this case

$$R(D) = \min I(X^n, \widehat{X}^n)$$

- The minimum is over all density functions $f(\widehat{x}^n | x^n)$ such that

$$E[d(X^n, \widehat{X}^n)] \leq D$$

- Here, $d(\cdot, \cdot)$ is the sum of the per-symbol distortions:

$$d(x^n, \widehat{x}^n) = \sum (x_i - \widehat{x}_i)^2$$

- Arguments similar to those previously used lead to

$$I(X^n, \widehat{X}^n) \geq \sum_i \left(\frac{1}{2} \log \frac{\sigma_i^2}{D_i} \right)^+$$

Example: multiple Gaussian variables

- It turns out that this lower bound can be met
- We still need to find

$$R(D) = \min_{\sum D_i = D} \sum \left(\frac{1}{2} \ln \frac{\sigma_i^2}{D_i} \right)^+$$

- Simple variational problem, Lagrangian is

$$L = \sum \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda \sum D_i$$

- This leads to

$$\frac{\partial L}{\partial D_i} = -\frac{1}{2D_i} + \lambda = 0$$

- Thus $D_i =$constant, and the optimal bit allocation means "same distortion for each variable"

Example: multiple Gaussian variables

- Based on the previous result it can be shown that the rate distortion function for multiple $N(0, \sigma_i^2)$ variables is

$$R(D) = \sum \frac{1}{2} \log \frac{\sigma_i^2}{D_i}$$

where

$$D_i = \begin{cases} c, & c < \lambda_i^2 \\ \sigma_i^2, & c \geq \lambda_i^2 \end{cases}$$

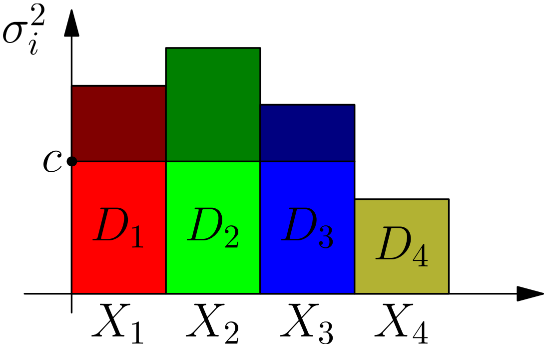
and c must be chosen so that

$$\sum D_i = D$$

is satisfied.

Example: multiple Gaussian variables

- Reverse water-filling: fix c, describe variables such as X_1, X_2, X_3 that have variance greater than c
- Ignore variables such as X_4 of variance smaller than c



Rate distortion theorem

- The proof of the theorem depends on certain typical sequences
- The idea behind typical sequences is that the typical set has high probability
- Thus if the rate distortion code works well for typical sequences it will work well on the average
- These ideas enter the proof of the rate distortion theorem
- Before going on it is necessary to define the typical sequences that will be used

Distortion typical sequences

Given $\epsilon > 0$, a pair of sequences $(x^n, \widehat{x}^n)$ is distortion typical if

$$\left| -\frac{\log p(x^n)}{n} - H(X) \right| < \epsilon$$

$$\left| -\frac{\log p(\widehat{x}^n)}{n} - H(\widehat{X}) \right| < \epsilon$$

$$\left| -\frac{\log p(x^n, \widehat{x}^n)}{n} - H(X, \widehat{X}) \right| < \epsilon$$

$$\left| d(x^n, \widehat{x}^n) - E[d(X, \widehat{X})] \right| < \epsilon$$

Distortion typical sequences

- According to the law of large numbers

$$\frac{-\log p(x^n)}{n} \to -E[\log p(X)] = H(X)$$

  as $n \to \infty$
- The same argument applies to the other conditions
- As $n \to \infty$ the conditions will be met

Strongly typical sequences

Given $\epsilon > 0$, a sequence is strongly typical with respect to $p(x)$ if the average number of occurrences of each alphabet symbol a in the sequence deviates from its probability by less than ϵ (divided by the alphabet size):

$$\left| \frac{N(a)}{n} - p(a) \right| < \frac{\epsilon}{|\mathcal{A}|}$$

Also required: symbols of zero probability do not occur: $N(a) = 0$ if $p(a) = 0$

Strongly typical pairs

- Very similar definition, but for pairs of sequences
- How many symbol pairs (a, b) exist in a sequence pair?
- Definition involves the joint probability $p(a, b)$:

$$\left| \frac{N(a, b)}{n} - p(a, b) \right| < \frac{\epsilon}{|\mathcal{A}|\,|\mathcal{B}|}$$

  for all possible a, b
- Again, pairs of zero probability must not occur

Set size

- A typical sequence with respect to a distribution follows that distribution well
- Intuitively, if n is large enough, this is “very likely”
- This can be made rigorous: due to the law of large numbers, the probability of the strongly typical set approaches 1 as $n \to \infty$
- Thus, roughly speaking, if a system works well for the typical set it will work well on average

Rate distortion theorem

- Select a $p(\widehat{x}|x)$ and $\delta > 0$
- The codebook consists of 2^{nR} sequences $\widehat{X}^n$
- To encode, map X^n to an index k
- To decode, map the index k to $\widehat{X}(k)$
- Pick $k = 1$ if there is no k such that $(X^n, \widehat{X}^n(k))$ is in the strongly jointly typical set
- Otherwise pick, for example, the smallest such k
- In this way, only 2^{nR} values of k are needed

Rate distortion theorem

- It is necessary to compute the average distortion, averaging over the random codebook choice:

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

- The sequences x^n can be divided in three sets:
  - Non-typical sequences
  - Typical sequences for which a $\widehat{X}^n(k)$ exists that is jointly typical with them
  - Typical sequences for which no such $\widehat{X}^n(k)$ exists
- What is the contribution of each set for the distortion?

Rate distortion theorem

- The total probability of the non-typical sequences can be made as small as desired by increasing n
- Hence, they contribute a vanishingly small amount to the distortion if $d(\cdot, \cdot)$ is bounded and if n is sufficiently large
- More precisely, they contribute $\epsilon \, d_{\text{max}}$ where d_{max} is the maximum possible distortion

Rate distortion theorem

- Consider now the case of typical sequences that have codewords jointly typical with them
- This means that the relative frequency of pairs (a, b) in the coded and decoded sequences are "close" to the joint probability
- The distortion is a continuous function of the joint probability, thus it will also be "close" to D
- If $d(\cdot, \cdot)$ is bounded, the distortion will be bounded by $D + \epsilon d_{\text{max}}$
- The total probability of this set is below 1, hence it will contribute at most $D + \epsilon d_{\text{max}}$ to the expected distortion

Rate distortion theorem

- Finally, consider typical sequences without jointly typical codewords
- Let the total probability of this set be P_t
- If $d(\cdot, \cdot)$ is bounded, the contribution due to this set will be at

most $P_t d_{\max}$

- It is possible to derive a bound for P_t that shows that it converges to

zero as $n \to \infty$

Putting it all together

- The three terms contribute differently to the expected distortion
- The non-typical set contributes at most $\epsilon\, d_{\max}$
- The typical / jointly typical set contributes at most $D + \epsilon d_{\max}$
- Finally, the typical / but not jointly typical set contributes a

vanishingly small fraction

- These terms add up to $R + \delta$, where δ can be made arbitrarily small
- This completes the argument

Rate distortion theorem (converse)

- Draw independent samples from a source X with density $p(x)$
- The rate R of any $(2^{nR}, n)$ rate distortion code with distortion $\leq D$ satisfies $R \geq R(D)$
- To show this assume that $E[d(X^n, \widehat{X}^n) \geq D$
- That $R \geq R(D)$ follows from a chain of inequalities

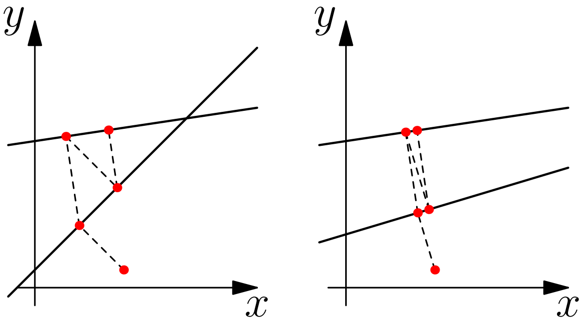
Rate distortion theorem (converse)

$nR \geq H(f_n(X^n))$	f_n assumes 2^{nR} values
$\geq H(f_n(X^n)) - H(f_n(X^n) \mid X^n)$	entropy is nonnegative
$= I(X^n, \widehat{X}^n) = H(X^n) - H(X^n \mid \widehat{X}^n)$	definition
$= \sum H(X_i) - H(X^n \mid \widehat{X}^n)$	independence
$= \sum H(X_i) - \sum H(X_i \mid \widehat{X}^n, X_1 \ldots X_{i-1})$	chain rule
$\geq \sum H(X_i) - \sum H(X_i \mid \widehat{X}_i)$	conditioning reduces entropy
$= \sum I(X_i, \widehat{X}_i)$	
$\geq \sum R\left(E[d(X_i, \widehat{X}_i)]\right)$	rate distortion definition
$= n \frac{1}{n} \sum R\left(E[d(X_i, \widehat{X}_i)]\right)$	
$\geq nR\left(\frac{1}{n} \sum E[d(X_i, \widehat{X}_i)]\right)$	convexity of $R(\cdot)$
$= nR(E[d(X^n, \widehat{X}^n)])$	definition
$\geq nR(D)$	$E(d(X^n, \widehat{X}^n)] \leq D$

Contents

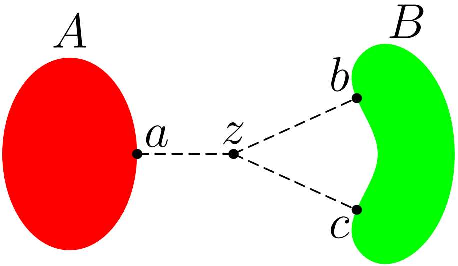
Alternating projections

- Consider two subspaces A and B of a common parent space
- A point in the intersection can be found by alternating projections



POCS

- The method depends on the possibility of defining a "projection"
- Subspaces are obviously convex
- Everything works for convex sets because projections are well defined



Distance between convex sets

- POCS works to find a point in the intersection of two convex sets A, B
- If the sets are disjoint, one can find the distance between them

$$d_{\min} = \min_{\substack{a \in A \\ b \in B}} d(a, b)$$

- Project z_i in A to find z_{i+1}
- Project z_{i+1} in B to find z_{i+2}
- Repeat to obtain $z_1, z_2, z_3, \ldots$
- The distance between the points converges to $d_{\min}$

POCS and rate distortion

- Need to recast rate distortion as convex optimisation
- Mutual information $I(X, \widehat{X})$ is the Kullback-Leibler distance $D(p||q)$
- Hence rate distortion is obtained by minimising this "distance"
- How to proceed?

POCS and rate distortion

- The Kullback-Leibler distance from $p(x)p(y|x)$ to $p(x)q(y)$ is minimised when $q(y)$ is the marginal

$$q_0(y) = \sum_x p(x)p(y|x)$$

- Proof: compute the distances from $p(x)p(y|x)$ to $p(x)q(y)$ and from $p(x)p(y|x)$ to $p(x)q_0(y)$ and show that the first is $\geq$ than the second
- This is an elementary consequence of the nonnegativity of $D(\cdot||\cdot)$

POCS and rate distortion

$$
\begin{aligned}
R(D) &= \min_{p(\widehat{x}|x):\sum p(x)p(\widehat{x}|x)d(x,\widehat{x})\leq D} I(X,\widehat{X}) \\
&= \min_{p(\widehat{x}|x):\sum p(x)p(\widehat{x}|x)d(x,\widehat{x})\leq D} \sum_{x,\widehat{x}} p(x)p(\widehat{x}|x)\log\frac{p(\widehat{x}|x)}{p(\widehat{x})} \\
&= \min_{p(\widehat{x}|x):\sum p(x)p(\widehat{x}|x)d(x,\widehat{x})\leq D} D\Big(\, p(x)p(\widehat{x}|x)\,||\,p(x)p(\widehat{x})\,\Big) \\
&= \min_{q(\widehat{x})} \min_{p(\widehat{x}|x):\sum p(x)p(\widehat{x}|x)d(x,\widehat{x})\leq D} D\Big(\, p(x)p(\widehat{x}|x)\,||\,p(x)q(\widehat{x})\,\Big) \\
&= \min_{a\in A}\min_{b\in B} D(a||b)
\end{aligned}
$$

Blahut-Arimoto

- Alternating minimisation as in POCS
- Start with $q(\widehat{x})$, find $p(\widehat{x}|x)$ that minimises $I(X, \widehat{X})$ (subject to the distortion constraint)
- For this $p(\widehat{x}|x)$, find the $q(\widehat{x})$ that minimises the mutual information, which is simply

$$q_0(\widehat{x}) = \sum p(x)p(\widehat{x}|x)$$