GI12/4C59: Information Theory

Lectures 25–27: Rate Distortion Theory

Massimiliano Pontil

About these lectures

Theme: We discuss the problem of representing random sequences by a finite number of codewords which minimize an average distortion function. This problem is related to vector quantization and $k-$mean clustering as used in Signal Processing and Machine Learning. Information Theory provides a framework to study this problem, using ideas similar to those encountered in channel coding.

Outline

1. The rate distortion problem
2. Examples
3. Main result
4. Optimization

Statement of the problem (informal)

Problem: given a source distribution and distortion measure, what is the minimum average distortion achievable at a particular coding rate?

We begin with an instance of this problem, where we wish to represent a single random variable X with a finite number R of bits, that is, we wish to find a set of centers $t(1), \ldots, t(M)$, $M = 2^R$ and a function $f : \mathcal{X} \subseteq \mathbb{R} \to \{1, 2, \ldots, M\}$ such that

$$\sum_{x \in \mathcal{X}} p(x)(x - t(f(x)))^2$$

is minimized. We already encountered this problem when we discussed uniform quantization of a continuous r.v..

Example

Let us analyze the case of a Gaussian r.v.: $p(x) = N(0, \sigma^2)$.

- With one bit quantization, by symmetry, we obtain $t(1) = -a$ and $t(2) = a$, with $a = E[X|X > 0]$ (can you say why?), and $f(x) = 2$ if $x > 0$ and $f(x) = 1$ otherwise. We also show below that

$$a = \sigma\sqrt{2/\pi}, \quad D = \int_{-\infty}^{\infty} (x - t(f(x)))^2 N(0, \sigma^2) dx = \sigma^2 \frac{\pi - 2}{\pi}.$$

- If just $R = 2$, the problem is less obvious. Below we discuss an efficient method to find a good solution.

Example (cont.)

Setting $y = x^2/2\sigma^2$, we have

$$a = \int_0^{\infty} x \frac{2}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\} dx \; = \int_0^{\infty} \sigma\sqrt{\frac{2}{\pi}} \exp\{-y\} dy = \sigma\sqrt{\frac{2}{\pi}}.$$

The expected distortion for one bit quantization is

$$\begin{aligned} D &= \int_{-\infty}^{0} \left(x + \sigma\sqrt{\frac{2}{\pi}}\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\} dx + \int_0^{\infty} \left(x - \sigma\sqrt{\frac{2}{\pi}}\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\} dx \\ &= 2\int_{-\infty}^{\infty} \left(x^2 + \sigma^2\frac{2}{\pi}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\} dx - 2\int_0^{\infty} \left(-2x\sigma\sqrt{\frac{2}{\pi}}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\} dx \\ &= \sigma^2 + \frac{2}{\pi}\sigma^2 - 4\frac{1}{\sqrt{2\pi}}\sigma^2\sqrt{\frac{2}{\pi}} = \sigma^2\frac{\pi - 2}{\pi}. \end{aligned}$$

Distortion for i.i.d. sources

More generally, let $X_1, \ldots, X_n$ be i.i.d. r.v. $\sim p(x), x \in \mathcal{X}$. A *rate distortion code* $\mathcal{C}^{(n)} = (2^{nR}, n)$ consists of an encoding function

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$$

and a decoding function

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{T}^n \quad \text{(usually } \mathcal{T} = \mathcal{X}\text{)}.$$

The process of replacing a sequence x_n by one of the possible points $\mathrm{t}(i) = g_n(i), i = 1, \ldots, 2^{nR}$ is called *vector quantization*. The set formed by the points t_i is called the *code-book*.

This way, the space $\mathcal{X}^n$ is divided in the assignment regions $f_n^{-1}(i)$ (or Voronoi partitions).

Distortion function

A distortion function is a mapping $d : \mathcal{X} \times \mathcal{T} \to [0, \infty)$ which we assume to be bounded, that is, $d_{max} = \max_{x,t} d(x, t) < \infty$.

Typically used functions are:

- *Hamming:* $d(x, t) = 0$ if $x = t$ and 1 otherwise.
- *Squared error:* $d(x, t) = (x - t)^2$.

The distortion between sequences $x^n \in \mathcal{X}^n$ and $t^n \in \mathcal{T}^n$ is the average per symbol distortion of the elements of the sequence,

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

Rate distortion function

We focus of the average distortion (reconstruction error) of $\mathcal{C}^{(n)}$)

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

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

$$\lim_{n\to\infty} E[d(X^n, g_n(f_n(X^n)))] \leq D.$$

- The rate distortion function $\rho : [0, \infty) \to [0, \infty)$ is defined by

$$\rho(D) = \inf\{R : (R, D) \text{ is achievable}\}.$$

Note: $\rho(D)$ is non-increasing and convex (skip proof).

Some comments

- We wish to find codes which have small average distortion.

If the code-book is fixed, it is clear that the best encoding function is given by

$$f_n(x^n) = \text{argmin}_i d(x^n, \mathbf{t}(i))$$

because the average distortion is minimize by mapping an x^n to the a codeword $\mathbf{t}(i)$ which is closest to it.

On the other hand if the assignment regions are fixed, the codewords $\mathbf{t}(i)$ are those which minimize the average distortion in each region.

Lloyd algorithm

The above observations suggest the following algorithm for finding a good code-book

- Initialization: choose an initial set of $M = 2^{nR}$ codewords, $\mathbf{t}(i)$, $i = 1, \ldots, M$.
- Repeat until convergence the following two steps
  1. Compute the function f_n by the formula $f_n(x^n) = \text{argmin}_i d(x^n, \mathbf{t}(i))$
  2. Update $\mathbf{t}(i)$ to minimize the average distortion in each region $f^{-1}(i)$.

It can be proved that the algorithm converges to a local minima.

Remark: these ideas are very related to $k-$means clustering as used in Machine Learning.

Main result of rate distortion theory

Theorem: The rate distortion function for an i.i.d. source $X \sim p(x)$ and bounded distortion function is

$$\rho(D) = \min\{I(X;T) : \sum_{x,t} p(x)q(t|x)d(x,t) \leq D\}$$

where the minimum is take w.r.t. $q(t|x)$. Said in other words:

1. If $R > \rho(D)$, there exists a sequence of $(2^{nR}, n)$ distortion codes whose average distortion goes to D as $n \to \infty$.
2. If $R < \rho(D)$ such code sequence does not exist.

Remark: we use the notation $X \sim p(x)$, $T \sim q(t)$, $T|X \sim q(t|x)$, and $X, Y \sim q(x,t) = p(x)q(t|x)$.
Although the result is stated for finite discrete r.v., it also holds in the continuous case.

Main result of rate distortion theory (cont.)

The result provides an optimization method for computing the rate distortion function. We discuss the computation of ρ in two important cases:

- Binary (Bernoulli) sources.
- Gaussian sources.

Later we will give some insights in proving the key part of the theorem (assertion 1) and discuss the relation with the channel coding theorem.

Binary sources

Theorem: The rate distortion function for an i.i.d. source $X \sim p(x)$ and bounded distortion function is

$$\rho(D) = \min\{I(X;T) : \sum_{x,t} p(x)q(t|x)d(x,t) \leq D\}$$

where the minimum is take w.r.t. $q(t|x)$.

Example 1: if $\mathcal{X} = \{0,1\}$, $p(1) = p$, and d the Hamming distortion, then

$$\rho(D) = \begin{cases} H(p) - H(D) & \text{if } D \in [0, \min(p, 1-p)] \\ 0 & \text{otherwise.} \end{cases}$$

Binary sources (cont.)

Proof: assume without loss of generality that $p \leq 1/2$, and consider the binary r.v. $Z = X + T|2$ (i.e., $Z = 1$ if $X \neq T$ an 0 if $X = T$). We then have,

$$H(X|T) = H(Z|T) \leq H(Z) \leq H(D)$$

where the last inequality is because

$$P(Z = 1) = P(X \neq T) = E[d(X, T)] \leq D$$

and $H(D)$ increases with D for $D \leq 1/2$. Consequently,

$$I(X;T) = H(X) - H(X|T) = H(p) - H(X|T) \geq H(p) - H(D)$$

which implies that $\rho(D) \geq H(p) - H(D)$.

Binary sources (cont.)

We now find a specific distribution $q(t|x)$ which achieves the above upper bound and satisfies the distortion constraint.

Consider a binary symmetric channel with input T, output X, and transition probability $(1 - D, D)$. Choose $r = P(T = 1)$ so that $P(X = 1) = p$, that is

$$r(1 - D) + (1 - r)D = p \quad \Rightarrow \quad r = \frac{p - D}{1 - 2D}.$$

and note that, if $D \leq p \leq 1/2$, then r is a well defined probability and

$$I(X;T) = H(X) - H(X|T) = H(p) - H(D)$$

with expected distortion $D = P(X \neq T)$.

On the other hand, if $D \geq p$ and we choose $p(T = 0) = 1$ we have $I(X;T) = 0$ and $D = p$, so also $\rho(D) = 0$.

Gaussian sources

Example 2: if X is a Gaussian source, $p(x) \sim N(0, \sigma^2)$, d the squared distortion, then we have

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

Proof: Since $E[(X-T)^2] \leq D$, we have

$$\begin{aligned} H(X|T) &= H(X-T|T) \leq H(X-T) \\ &\leq H(N(0, E[(X-T)^2]) = \frac{1}{2}\log(2\pi e)E[(X-T)^2]. \end{aligned}$$

where the second inequality is since the normal distribution maximizes the entropy for a fixed value of the second order momentum.

Gaussian sources (cont.)

Using the above inequality and the properties of the normal distribution, we have

$$\begin{aligned} I(X;T) &= H(X) - H(X|T) \geq H(X) - \frac{1}{2}\log(2\pi e)E[(X-T)^2] \\ &\geq \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)D = \frac{1}{2}\log\frac{\sigma^2}{D} \end{aligned}$$

which implies that $\rho(D) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$.

Now, if $D \in [0, \sigma^2]$ and we choose $X = T + Z$ with $p(t) \sim N(0, \sigma^2 - D)$ and $Z \sim N(0, D)$, we have $p(x) = N(0, \sigma^2)$ and $I(X;T) = 1/2\log(\sigma^2/D)$, achieving the bound. On the other hand, if $D > \sigma^2$ and we choose $T = 0$ with probability 1, we get $I(X;T) = 0$ and, so, $\rho(D) = 0$.

An important consequence

We can rewrite the above result as $D(R) = \sigma^2 2^{-2R}$. (this is the infimum of the distortion achieved asymptotic by an $(2^{nR}, n)$ code).

For example if $R = 1$, we get

$$D(1) = \frac{\sigma^2}{4}, \text{ and, in general } D(R+1) = \frac{D(R)}{4}.$$

In contrast, in our opening example (one single r.v.) we found that $D(1) = (1 - 2/\pi)\sigma^2 \approx 0.36\sigma^2$. Thus,

- We can achieve a lower distortion by quantizing a succession of i.i.d. r.v. rather than each r.v. separately!

Proof of the theorem

Theorem: For an i.i.d. source $X \sim p(x)$ with bounded distortion function, we have

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

We sketch the proof of the main part of the theorem:

- for every $R > \rho(D)$, there exists a sequence of $(2^{nR}, n)$ distortion codes whose average distortion goes to D as $n \to \infty$.

For the converse result, see chapters 13.4 of Cover and Thomas's book.

Scheme of the proof

The proof is similar to the proof of the channel coding theorem. It is based on the following steps:

1. Generate a code-book $\mathbf{t}(i)$, $i = 1, \ldots, 2^{nR}$ i.i.d. according to
$$p(\mathbf{t}) = p(t_1)p(t_2)\cdots p(t_n)$$
2. Encode a sequence x^n using a modified notion of jointly - typicality explained below.
3. Show that the average (w.r.t. to code-book generation) of the average distortion tends to D as $n \to \infty$.
4. Extract a good sequence of codes from the set randomly generated codes.

Distortion $\epsilon-$typicality

We define the set of distortion $\epsilon-$typical sequences $\mathcal{A}_\epsilon^{(n)}$ to be the set formed by the pairs $(x^n, t^n) \in \mathcal{X}^n \times \mathcal{T}^n$ (generated $i.i.d.$ with $p(x,t)$) which satisfies the following constraints

$$\left|\frac{1}{n}\log p(x^n) + H(X)\right| < \epsilon \qquad \left|\tfrac{1}{n}\log p(t^n) + H(T)\right| < \epsilon$$

$$\left|\frac{1}{n}\log p(x^n, t^n) + H(X,T)\right| < \epsilon \quad |d(x^n, t^n) - E[d(X^n, T^n)]| < \epsilon$$

This set is similar to the set of $\epsilon-$jointly typical sequences which we used in the proof of the channel coding theorem. Also, by the weak law of large numbers, we have that Prob$((x^n, t^n) \in \mathcal{A}_\epsilon^{(n)}) \to 1$ as $n \to \infty$.

Encoding/decoding

We generate a code-book $\mathcal{C} = \{\mathbf{t}(1), \ldots, \mathbf{t}(2^{nR})\}$ i.i.d. as discussed above.

We then encode x^n as

$$f_n(x^n) = \begin{cases} w & \text{if } w \text{ is the smallest index} : (x^n, \mathbf{t}(w)) \in \mathcal{A}_\epsilon^{(n)} \\ 1 & \text{if, for every } w,\ (x^n, \mathbf{t}(w)) \notin \mathcal{A}_\epsilon^{(n)}. \end{cases}$$

and decode $w \in \{1, 2, \ldots, 2^{nR}\}$ as $g_n(w) = \mathbf{t}(w)$.

We need to calculate $\bar{D} = E_{\mathcal{C}}[E_{X^n}[d(X^n, g_n(f_n(X^n))]]$.

Formula for the average distortion

We decompose $\bar{D}$ as

$$\bar{D} = \sum_{i=1,2} E_{\mathcal{C}}[E_{X^n}[d(X^n, g_n(f_n(X^n)))|Z = i]].$$

$Z(\mathcal{C}, X^n)$ is binary r.v. which is 1 if there exist a codeword which is jointly $\epsilon-$typical with X^n, and zero otherwise.
If $P^{(n)} := Prob(Z = 0)$, we have

$$\bar{D} \leq (1 - P^{(n)})(D + \epsilon) + P^{(n)} d_{max} \leq D + \epsilon + P^{(n)} d_{max}$$

Thus, to prove the result we need to show that $P^{(n)} \to 0$ as $n \to \infty$.

Computation of $P^{(n)}$

The computation of $P^{(n)}$ uses ideas similar to those encountered in proving the channel coding theorem. The conclusion is that, if $R < I(X;T)$, $P^{(n)}$ goes to zero as $n \to \infty$. (See page 355-6 of Cover and Thomas for the details).

This tells us that, if $R < I(X;T)$, for every $\delta > 0$, there exist ϵ and n such that $E_{\mathcal{C}}[E_{X^n}[d(X^n, g_n(f_n(X^n))]] \leq D + \delta$. Consequently, there must exist at least one such code $\mathcal{C}^*$ with

$$E[d(X^n, T^n)|\mathcal{C}^*] \leq D + \delta.$$

This conclude the sketch of the proof.

Remark: It is also possible to use the method of types from last class, to prove a stronger result

Informal argument

We discuss an informal argument for the result in the case of Gaussian sources. We denote by $B(a,r)$ the sphere in $\mathbb{R}^n$ centered at a with radius r. All statements made are meant to be true with "high probability" and asymptotically.

Remember the C.C. Theorem for a Gaussian channel: $Y = X + Z$, with $Z \sim G(0,K)$, with codewords subject to the power constraint $1/n \sum_{i=1}^{n} E[X_i^2] \leq \rho$. The idea in computing the channel capacity is based on the following observations

- Given an input x^n, the output $y^n \in B(x^n, \sqrt{nK})$ (a small sphere)
- The possible outputs lie in $B(0, \sqrt{n(K+\rho)})$ (big sphere)

Consequently, a code has small decoding error if the small spheres do not overlap. The **maximum** number of such spheres (or *packing number*) is

$$M = \frac{Vol(B(0, \sqrt{n(K+\rho)}))}{Vol(B(0, \sqrt{nK}))} = \left(\frac{K+\rho}{K}\right)^{n/2} \Rightarrow C = \frac{\log M}{n} = \frac{1}{2}\log\left(1 + \frac{\rho}{K}\right)$$

Informal argument (cont.)

Consider the rate distortion problem for a Gaussian source, $X \sim N(0, \sigma^2)$, and squared distortion (thus, $\sqrt{d(x^n, t^n)}$ is the Euclidean distance in $\mathbb{R}^n$).

If an $(2^{nR}, n)$ code achieves a distortion D, then, given an input sequence x^n, there exists a codeword which has Euclidean distance $\sqrt{nD}$ to it. Here we wish to find the **minimum** number of such codewords (or *covering number*). As before, this number is

$$M = \frac{Vol(B(0, \sqrt{n(\sigma^2)}))}{Vol(B(0, \sqrt{nD}))} = \left(\frac{\sigma^2}{D}\right)^{n/2} \Rightarrow \rho(D) = \frac{\log M}{n} = \frac{1}{2} \log \frac{\sigma^2}{D}$$

Relation between channel and distortion codes

The observations above suggest that a channel code can be used to build a distortion code and viceversa.

Indeed, there is a general relation between sphere packing and sphere covering. If S is a metric space with distance function $d(\cdot, \cdot)$ (e.g., our big sphere above), we define

- Covering number: $N_c(\delta)$ = size of the smallest set of points $G(\delta)$ such that, for every $x \in S$, $d(x, G(\delta)) = \min\{d(x, t) : t \in G(\delta)\} \leq \delta$.
- Packing number: $N_p(\gamma)$ = size of the largest set of points $F(\gamma)$ such that if $x, z \in F$, $x \neq z$, $d(x, z) \geq \gamma$.

It can be shown that

$$N_p(2\delta) \leq N_c(\delta) \leq N_p(\delta) \quad \text{and, so,} \quad N_c(2\delta) \leq N_p(\delta) \leq N_c(\delta)$$

Characterization of $\rho(D)$

We can use the method of Lagrange multipliers to find

$$\rho(D) = \min_{q(t|x)} \left\{ I(X;T) : \sum_{x,t} p(x)q(t|x)d(x,t) \leq D \right\}.$$

The Lagrangian function is

$$J(q) = \sum_{x,t} p(x)q(t|x) \log \frac{q(t|x)}{\sum_x p(x)q(t|x)} + \lambda \sum_{x,t} p(x)q(t|x)d(x,t) + \sum_x \nu(x) \sum_t q(t|x)$$

where the last term corresponds to the constraint that

$$\sum_t q(t|x) = 1, \quad \text{for every } x \in \mathcal{X}.$$

- In the calculation below we assume, for simplicity that $q(t) > 0$, $t \in \mathcal{T}$.

Characterization of $\rho(D)$ (cont.)

Differentiating w.r.t. $q(t|x)$ and using $\sum_x p(x)q(t|x) = q(t)$, we obtain

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

where $\log \mu(x) := \nu(x)/p(x)$ and, using $\sum_t q(t|x) = 1$ we have

$$\mu(x) = \sum_t q(t)e^{-\lambda d(x,t)}$$

and, so, combining

$$q(t|x) = q(t)\frac{e^{-\lambda d(x,t)}}{\sum_t q(t)e^{-\lambda d(x,t)}} \qquad (*)$$

Multiplying by $p(x)$ and summing over x, we obtain, if $q(t) > 0$,

$$\sum_t \frac{p(x)e^{-\lambda d(x,t)}}{\sum_{t'} q(t')e^{-\lambda d(x,t')}} = 1, \quad t \in \mathcal{T}$$

which combined with the distortion constraint provides a set of $|\mathcal{T}| + 1$ equations to compute λ and $q(t)$.

A method for computing $\rho(D)$

In practice, it is difficult to solve the above system of non-linear equations.

Consider the general problem where we want to find the minimum distance between two convex sets A and B,

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

We can use the following algorithm to compute a solution:

- fixed $\bar{a} \in A$ and look for a closest point $\bar{b} \in B$ to $\bar{a}$. Then, find a closest point $\bar{a} \in A$ to $\bar{b}$, and repeat.

Under some general conditions on the distance d, the algorithm converges to an optimal solution. Let's see how this method can be used to compute the rate distortion function...

A method for computing $\rho(D)$ (cont.)

Lemma: $I(X;T) = D(q(x,t) \parallel p(x)q(t)) = \min_{r(t)} D(q(x,t) \parallel p(x)r(t))$.

Proof (sketch): Remember our notation $q(x,t) = p(x)q(t|x)$. After some algebra we obtain

$$D(q(x,t) \parallel p(x)r(t)) - D(q(x,t) \parallel p(x)q(t)) = D(q(t) \parallel r(t)) \geq 0$$

with inequality if and only if $r = q$.

Using this lemma in the above formula for $\rho(D)$, we see that

$$\rho(D) = \min\{D(a \parallel b) : a \in A, b \in B\}, \quad \text{with}$$

$$A = \{a = q(x,t) : \sum_t q(\cdot,t) = p(\cdot) \geq 0,\ E_q[d(X,T)] \leq D\}$$

$$B = \{b = p(x)r(t) : r(\cdot) \geq 0,\ \sum_{t \in \mathcal{T}} r(t) = 1\}$$

A method for computing $\rho(D)$ (cont.)

We then have the following iterative algorithm to compute $\rho(D)$

- Choose λ and $r(t)$, $t \in \mathcal{T}$ and calculate $q(t|x) \in A$ which minimizes $D(q(t|x)p(x) \parallel p(x)r(t))$. This is given by the above method of Lagrange multipliers. In particular, equation (*) gives

$$q(t|x) = r(t)\frac{e^{-\lambda d(x,t)}}{\sum_t r(t)e^{-\lambda d(x,t)}}.$$

- We fixed this just computed $q(t|x)$ and minimize $D(q(t|x)p(x) \parallel p(x)q(x))$ for $p(x)q(x) \in B$. By the lemma above, the solution is

$$q(t) = \sum_x p(x)q(t|x).$$

A method for computing $\rho(D)$ (cont.)

We summarize the above alternating minimization method

- Repeat until convergence
  1. Choose λ and $r(t)$, $t \in \mathcal{T}$ and calculate:

$$q(t|x) = r(t)\frac{e^{-\lambda d(x,t)}}{\sum_t r(t)e^{-\lambda d(x,t)}}.$$

  2. Update $r(t)$ as

$$r(t) = \sum_x p(x)q(t|x).$$

Remark: This method is called the Blahut-Arimoto algorithm. Csiszar has shown that the converges in the limit to $\rho(D)$.

Bibliography

This lectures are based on Chapter 13 of Cover and Thomas's book.

A review paper is: T. Berger and J. Gibson, "Lossy Source Coding", in *Information Theory: 50 Years of Discoveries*, S. Vertú and S.W. McLaughlin Eds., IEEE Press, 1998.