Complexity, Information and Geometry (Module 1)

Peyresque

Alfred Hero

Digiteo and University of Michigan

July, 2008

Outline of Module 1

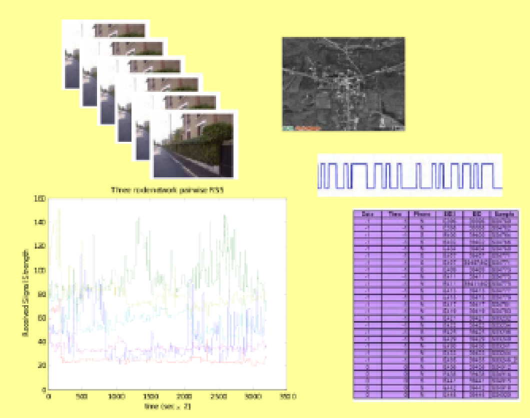
1. Complexity
   - Coding complexity
2. Information
   - Shannon entropy
   - Rényi entropy
3. Entropy estimation
   - Density estimation
   - Entropy estimation in high dimensions

Acknowledgements

- Arvind Rao
- Kumar Sricharan
- Kevin Carter
- Olivier Michel, U. Nice

Complexity

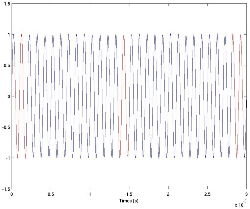
What is complexity of a signal or image?

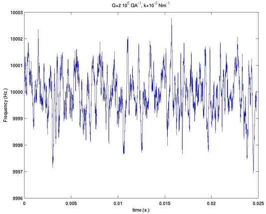


$s_t : t \in \mathcal{S}$: a signal evolving over time $\mathcal{S}=[0, T]$ or space $\mathcal{S} = [0, T] \times [0, T]$

Complexity of a signal

Two signals s_t - which is more complex?





Algorithmic Complexity of a String

`abababababababababababababababababababababababababababababababab`

`4c1j5b2p0cv4w18rx2y39umgw5q85s7urqbjfdppa0q7nieieqe9noc4cvafzf`

The algorithmic complexity (or algorithmic complexity) of a string s is the length of its shortest description p on a universal Turing machine U

$$K(s) = \min\{l(p) : U(p) = s\}$$

- AC satisfies chain rule $K(X, Y) = K(X) + K(Y|X) + O(\log(K(X, Y))$
- However, while $K(s)$ can always be bounded ($|\texttt{gzip s}|$), $K(s)$ is not a computable function
- Algorithmic complexity captures the complexity of a single instance of a string.

Information: complexity of an ensemble

An alternative is to try to capture complexity of an ensemble of strings or signals.

$\Rightarrow$ Information theoretic measures of complexity

Introduced by Weaver, Shannon, Kolmogorov

Probabilistic framework

Probability Model

$(\mathcal{X}, \mathcal{A}, P)$: outcomes, events, probability function.

Sometimes it makes sense to assume a parameteric probability model:

$P = P_\theta$ belongs to a family $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$.

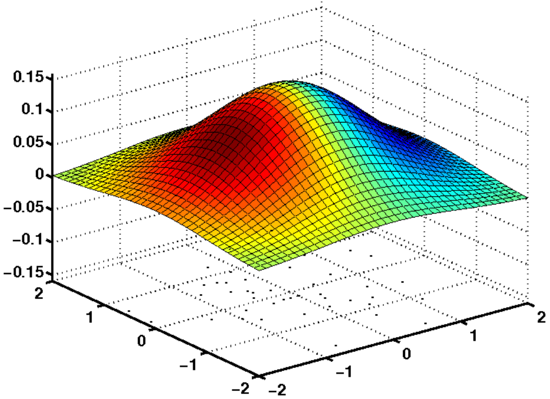
Distingush between discrete and continuous random variables

$$P(X \in B) = \left[\begin{array}{ll} \sum_{x \in B} p(x), & X \textit{ discrete} \\ \int_{x \in B} f(x) dx & , X \textit{ cts.} \end{array} \right.$$

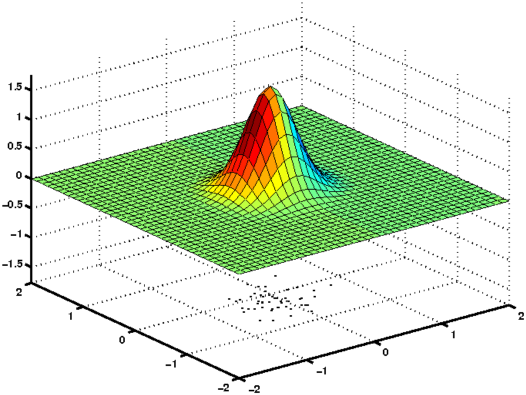
Expectation Operator: For any function $Z = Z(x)$:

$$E_\theta[Z] \stackrel{\text{def}}{=} \int_{\text{supp} dP_\theta(\bullet)} Z(x) dP_\theta(x) = \int Z(x) f_\theta(x) dx$$

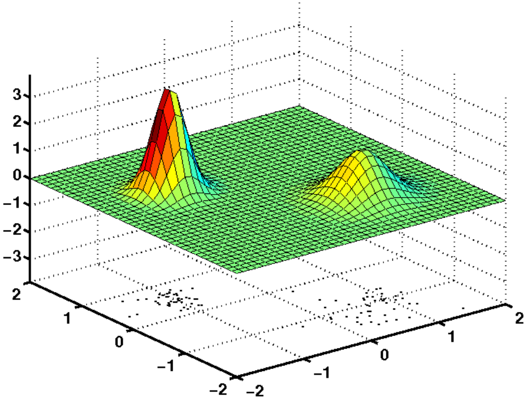
High Entropy Feature Density



Low Entropy Feature Density



Mixture Feature Density



Information: Shannon Entropy

Shannon entropy for a discrete r.v. X with pmf $p(x)$

$$H(X) = H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = E\left[\log \frac{1}{p(X)}\right]$$

Shannon entropy for a continuous r.v. X with pdf $f(x)$

$$H(X) = H(f) = -\int f(x) \log f(x) dx = E\left[\log \frac{1}{f(X)}\right]$$

Relative entropy

$$D(f \| g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

The relative entropy, also called the information (Kullback-Liebler) divergence of pdf's f and g, is non-negative

Conditional entropy and mutual information

Important cases of relative entropy

Conditional entropy between r.v.s X and Y

$$H(Y|X) = -\int f(x,y)\log\frac{f(x,y)}{f(x)}dxdy = -E[\log f(Y|X)]$$

Mutual information between r.v.s X and Y

$$I(X;Y) = \int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy$$

Relation

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

Some simple properties of discrete Shannon Entropy

- Non-negativity

$$H(X) \geq 0, \; ; \; \text{``} =\text{''} \;\; iff \; \exists \mu : f(x) = \delta(x - \mu)$$

  μ fixed

- Concavity

$$H(\epsilon f + (1-\epsilon) g) \leq \epsilon H(f) + (1-\epsilon) H(g), \quad \text{''} = \text{''} \;\; iff \;\; f = g \;\; or \;\; \epsilon \in \{0, 1\}$$

- Chain rule

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

- Sub-additivity

$$H([X, Y]) \leq H(X) + H(Y), \quad \text{''} = \text{''} \;\; iff \;\; f(X, Y) = f(X) f(Y)$$

Continuous Shannon entropy satisfies all but the first property.

Extremal properties of Shannon entropy

- If X is discrete with finite alphabet $\mathcal{X}=\{x_1, \ldots, x_Q\}$

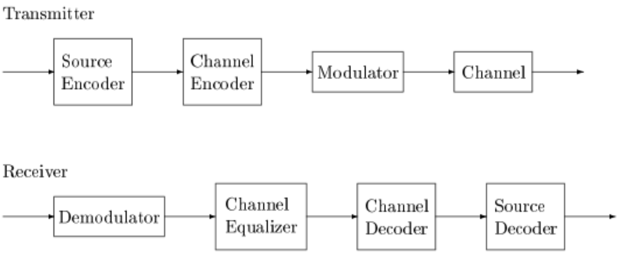
$$H(X) \leq \log|\mathcal{X}| = \log Q, \quad " = " \textit{ iff } p(x_i) = \frac{1}{Q} \; \forall i$$

- If X is continuous on $\mathcal{X}=\mathbb{R}$ with given finite variance $\mathrm{var}(X) = E[X^2] - E^2[X]$

$$H(X) \leq \frac{1}{2}\log(2\pi\sigma^2), \quad " = " \textit{ iff } f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- For X in $\mathbb{R}^d$ with given finite covariance matrix Σ Shannon entropy is maximized by multivariate Gaussian density with given covariance.

Shannon entropy and source coding



Digital communication system (Gupta 2001)

Let X be a discrete random variable with finite alphabet $\mathcal{X}=\{a_1, \ldots, a_Q\}$ where $Q = 2^n$.

For each a_i define binary codeword c_i of length l_i, e.g., $c_i = 010$, $l_i = 3$.

Average length of code is defined as

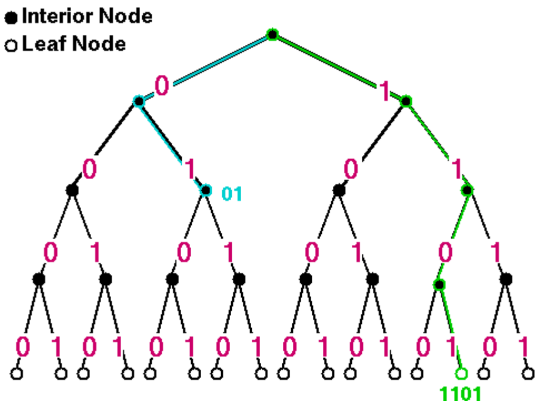
$$L = E[l_i] = \sum_{i=1}^{Q} p_i l_i$$

Enumerative encoding strategy (Coolen 2004)

message :	$n-$bit string :	corresponding number :
a_1	$000\ldots00$	0
a_2	$000\ldots01$	1
a_3	$000\ldots10$	2
a_4	$000\ldots11$	3
$\vdots$	$\vdots$	$\vdots$
a_{2^n}	$111\ldots11$	2^n-1

- Codewords have identical lengths and $L = n = \log Q$
- $\log Q$ might be taken as a natural measure of complexity
- $H(X) = \log Q$ when $p_i = 1/Q$, i.e., symbols are equally likely to occur

Enumerative codewords are at leaves of the depth $\log Q$ tree



If symbols are not equally likely a better (lower average length) code can be obtained

Example (Coolen 2004)

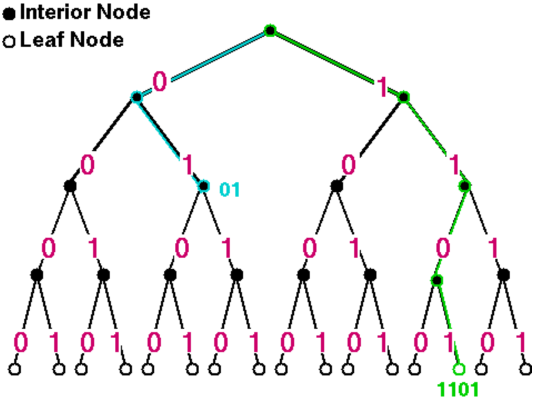
$$A = \{a_1, a_2, a_3, a_4\} \qquad p(a_1) = \frac{1}{2}, \quad p(a_2) = \frac{1}{4}, \quad p(a_3) = \frac{1}{8}, \quad p(a_4) = \frac{1}{8}$$

message :	enumerative code :		prefix code :	
a_1	00	$\ell(a_1) = 2$	1	$\ell(a_1) = 1$
a_2	01	$\ell(a_2) = 2$	01	$\ell(a_2) = 2$
a_3	10	$\ell(a_3) = 2$	001	$\ell(a_3) = 3$
a_4	11	$\ell(a_4) = 2$	000	$\ell(a_4) = 3$

For this example: $L = \frac{1}{2} + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 = 1.75$

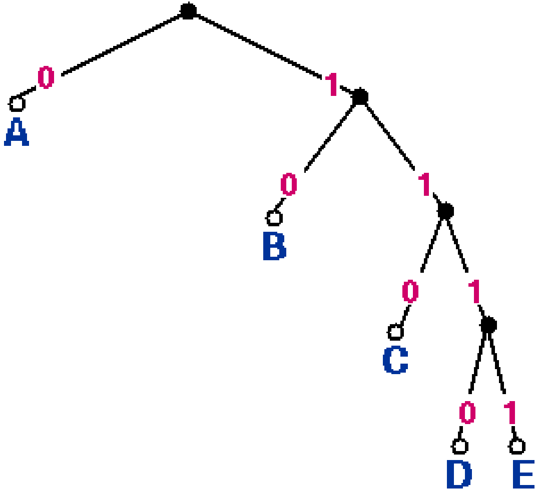
This code is represented by a subtree of the enumerative code tree of depth 4

Example (Coolen 2004)



Shannon entropy and lossless coding

Prefix code



Shannon entropy and lossless coding

This *Shannon entropy coding* strategy is due to Huffman [3]

Codewords assigned to symbols $\{a_1, \ldots, a_Q\}$ in such a way that

$$2^{-l_i} = \mathrm{lub}\,(p_i)$$

Huffman coding minimizes the average code length over all prefix codes.

Fundamental result:

$$H(X) \leq L_{Huffman} \leq H(X) + 1$$

Conclude: Shannon entropy is average coding complexity for lossless encoding of discrete source X

Rényi Entropy

Rényi entropy for a discrete r.v. X with pmf $p(x)$ (here $\alpha > 0$)

$$H_\alpha(X) = H_\alpha(p) = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} p^\alpha(x) = \frac{1}{1-\alpha} E\left[p^{\alpha-1}(X)\right]$$

Rényi entropy for a continuous r.v. X with pdf $f(x)$

$$H_\alpha(X) = H_\alpha(f) = \frac{1}{1-\alpha} \log \int f^\alpha(x) dx = \frac{1}{1-\alpha} E\left[f^{\alpha-1}(X)\right]$$

Conditional Rényi entropy

$$H_\alpha(X|Y) = \int f_Y(y) \underbrace{\left(\frac{1}{1-\alpha} \log \int f^\alpha_{X|Y}(x|y) dx \right)}_{H_\alpha(X|Y=y)} dy$$

Some simple properties of Rényi Entropy

- Non-negativity (discrete X)

$$H_\alpha(X) \geq 0, \ ; \ " = " \ \textit{iff} \ \exists \mu : f(x) = \delta(x-\mu)$$

  μ fixed

- Concavity

$$H_\alpha(\epsilon f + (1-\epsilon) g) \leq \epsilon H(f) + (1-\epsilon) H(g), \quad " = " \ \textit{iff} \ f = g \ \textit{or} \ \epsilon \in \{0,1\}$$

- Sub-additivity

$$H_\alpha([X,Y]) \leq H_\alpha(X) + H_\alpha(Y), \quad " = " \ \textit{iff} \ f(X,Y) = f(X) f(Y)$$

- Monotonic decreasing in α

$$H_{\alpha+\Delta}(X) \leq H_\alpha(X), \quad \Delta > 0$$

Unlike Shannon entropy Rényi entropy does not satisfy the chain rule

Extremal properties of Rényi entropy

- If X is discrete with finite alphabet $\mathcal{X}=\{x_1, \ldots, x_Q\}$

$$H_\alpha(X) \leq \log|\mathcal{X}| = \log Q, \;\; " = " \; \textit{iff} \;\; p(x_i) = \frac{1}{Q} \; \forall i$$

- If X is continous on $\mathcal{X}=\mathbb{R}$ with finite variance $\mathrm{var}(X) = E[X^2] - E^2[X]$ then $H(X)$ is maximized by a student-t density w 1 degree of freedom and identical variance.
- For X in $\mathbb{R}^d$ with given finite covariance matrix Σ Rényi entropy is maximized by multivariate Student-t density with given covariance parameter (Vignat etal [7]).

Limiting forms of Rényi entropy

- Shannon entropy limit

$$\lim_{\alpha \to 1} H_\alpha(X) = H(X)$$

- Equally likely entropy limit

$$\lim_{\alpha \to 0} H_\alpha(X) = \log Q$$

- Rarest outcome limit

$$\lim_{\alpha \to \infty} H_\alpha(X) = \log \frac{1}{\min p(x)}$$

Rényi source encoding: "Source coding under siege"

Let X be a discrete random variable with finite alphabet $\mathcal{X}=\{a_1,\ldots,a_Q\}$ where $Q=2^n$.

Baer (Thesis 2002) considers the average exponential length of code

$$C = E[2^{l_i}] = \sum_{i=1}^{Q} p_i 2^{l_i}$$

As compared to the standard avg codelength $E[l_i]$, C emphasizes the longer codewords. Ziad (Thesis 1998) proposes generalized average codeword length $(t>0)$

$$L(t) = \frac{1}{t}\log \sum_{i=1}^{Q} p_i 2^{t l_i}$$

Properties of Ziad's measure:

$$\lim_{t\to 0} L(t) = E[l_i], \quad \lim_{t\to\infty} L(t) = \max l_i, \quad dL(t)/dt \geq 0$$

Rényi source coding theorem

If assign codewords to symbols $\{a_1, \ldots, a_Q\}$ in such a way that

$$2^{-l_i} = \mathrm{lub}\left(\frac{p_i^\alpha}{\sum_{i=1}^Q p_i^\alpha}\right)$$

then

$$H_{1/(1+t)}(X) \leq L(t) < H_{1/(1+t)}(X) + 1$$

NB: Baer (2007) has specified a modified Huffman prefix code construction that satisfies the assignment condition.

Defn: a *discrete (continuous)source* $\{X_i\}_{i=-\infty}^{\infty}$ is a random sequence with discrete (continuous) alphabet.

Joint distribution of a source is described by its joint distributions, e.g. for a discrete source

$$p(x_{-M}, \ldots, x_M), \quad M = 1, 2, \ldots$$

A source is stationary if for any integers l and M

$$p(x_{l+1}, \ldots, x_{l+M}) = p(x_1, \ldots, x_M)$$

Two cases of stationary sources of interest

- i.i.d. source $p(x_1, \ldots, x_M) = \prod_{i=1}^{M} p(x_i)$
- First order Markov source $p(x_1, \ldots, x_M) = p(x_1) \prod_{i=2}^{M} p(x_i | x_{i-1})$

Multivariate extensions

Shannon joint entropy

The joint entropy of an M segment of a stationary discrete source $X_1, \ldots, X_M\}$ is

$$H(X_1, \ldots, X_M) = -\sum p(x_1, \ldots, x_M) \log p(x_1, \ldots, x_M)$$

Example: i.i.d. source

$$H(X_1, \ldots, X_M) = MH(X_1)$$

Example: stationary Markov source

$$H(X_1, \ldots, X_M) = (M-1)H(X_2|X_1) + H(X_1)$$

These relations also hold for stationary continuous sources

$\Rightarrow$ Joint entropy diverges as $M \rightarrow \infty$

The Shannon entropy rate of a stationary source $\mathcal{X} = \{X_i\}$ is defined as

$$H(\mathcal{X}) = \lim_{M\to\infty} \frac{H(X_1, \ldots, X_M)}{M}$$

Example: i.i.d. source with $P(X_1 = i) = p_i$

$$H(\mathcal{X}) = H(X_1) = -\sum_i p_i \log p_i$$

Example: stationary Markov source with $P(X_1 = i, X_2 = j) = p_{j|i} p_i$

$$H(\mathcal{X}) = H(X_2|X_1) = -\sum_{i,j} p_i p_{j|i} \log p_{j|i}$$

Alternative definition of entropy rate

$$H^{'}(\mathcal{X}) = \lim_{M \to \infty} H(X_M | X_{M-1}, \ldots, X_1)$$

Thm: $H(\mathcal{X}) = H^{'}(\mathcal{X})$

The Rényi entropy rate of a stationary source $\mathcal{X} = \{X_i\}$ is defined as

$$H_\alpha(\mathcal{X}) = \lim_{M \to \infty} \frac{H_\alpha(X_1, \ldots, X_M)}{M}$$

Example: i.i.d. source with $P(X_1 = i) = p_i$

$$H_\alpha(\mathcal{X}) = H_\alpha(X_1) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha$$

For Markov sources the Rényi entropy rate is more complicated than in the case of Shannon's entropy rate

Thm (Ziad, 98): if $\mathcal{X}$ is a discrete Markov source with finite alphabet and $p_{i|j} > 0$ for all i, j, Then

$$H_\alpha(\mathcal{X}) = \frac{\log \lambda(\alpha, P)}{1-\alpha}$$

where $\lambda(\alpha, P)$ is the largest eigenvalue of the matrix

$$R = \begin{bmatrix} p^\alpha_{1|1} & p^\alpha_{1|2} & \cdots & p^\alpha_{1|A} \\ p^\alpha_{2|1} & p^\alpha_{2|2} & \cdots & p^\alpha_{2|A} \\ \vdots & \ddots & \ddots & \vdots \\ p^\alpha_{A|1} & \cdots & p^\alpha_{A|A-1} & p^\alpha_{A|A} \end{bmatrix}$$

Note, with previous definition of conditional Rényi entropy, it is not true that

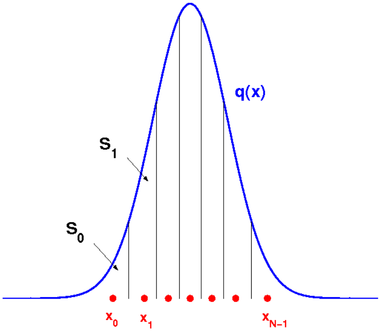
$$H_\alpha(\mathcal{X}) = H_\alpha^{'}(\mathcal{X}) = \lim_{M\to\infty} H(X_M|X_{M-1},\ldots,X_1)$$

However, for discrete finite alphabet stationary sources we could adopt the above as a *definition* of conditional Rényi entropy.

Recap

- Complexity of an ensemble X = average number of bits required to optimally encode X.
- Shannon entropy $H(X)$ is optimal code length that minimizes redundancy
- Rényi entropy $H_\alpha(X)$ is optimal exponentiated code length that minimizes redundancy
- Rényi entropy $H_\alpha(X)$ increasingly sensitive to tail behavior of $f(x)$ as α decreases to zero.

Let X be a 1D source with continuous alphabet in $\mathbb{R}$. A N-level scalar quantizer is defined by a mapping $Q : \mathbb{R} \rightarrow \{x_1, \ldots, x_N\} \subset \mathbb{R}$

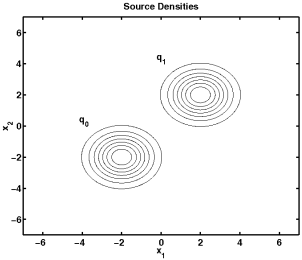


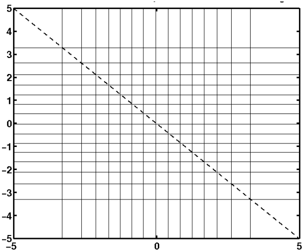
Scalar quantizer of a 1D continuous source X with density $q(x)$

$\mathcal{C}$ is a "codebook consisting" of intervals cells S_i and quantization levels

Let $X = [X_1, X_2]$ be a 2D source with continuous alphabet in $\mathbb{R}$. A N-level vector quantizer Q is defined similarly to before

$$Q(x) = x_i, \quad x_i \in S_i$$





Product vector quantizer of a 2D continuous source X with density $q(x) = [q_0(x) + q_1(x)]/2)$

Obvious observation: any finite-bit encoding of a continous source will necessarily entail some loss in information.

Quantization distortion measures for a given quantizer Q

- Mean squared quantization error (MSQE)

$$\mathrm{MSQE} = E[(X - Q(X))^T(X - Q(X))] = E[\|X - Q(X)\|^2]$$

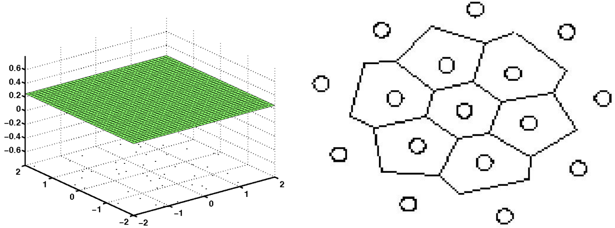
- Increase in minimum probability of decision error (decide q_1 vs q_0)

$$P_e^Q = [P_0(l(Q(X)) > \eta) + P_1(l(Q(X)) < \eta)]/2$$

$l(u) = q_1(u)/q_0(u)$ likelihood ratio

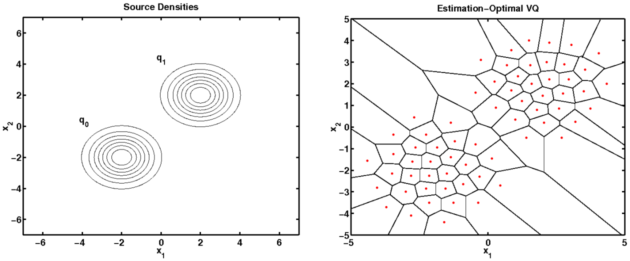
- Linear combinations of the above

Optimal MSQE quantizers produce equally likely codewords $x_1, \ldots, x_N$ for given number of levels N (rate $logN$).



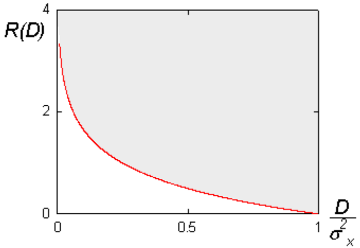
Optimal MSQE vector quantizer for uniform density $q(x) = [0, 1]^d$

Optimal MSQE quantizers produce equally likely codewords $x_1, \ldots, x_N$ for given number of levels N (rate $logN$).



Optimal MSQE vector quantizer for density $q(x) = [q_0(x) + q_1(x)]/2$

Shannon's *rate distortion function*: $R(D) = \min_{E[\rho(X,\hat{X})] \le D} I(X,Y)$



- R is monotonic non-increasing function of distortion D
- R is a theoretical limit (like channel capacity) and cannot generally be achieved exactly
- Practical high rate approximations to VQ can come close to limit

Let $X = [X_1, \ldots, X_d]$ be a d-dimensional continuous source with jpdf. $q(x), x \in \mathbb{R}^d$.

Define $\{Q_N\}_{N=1,2,\ldots}$ a sequence of N-level VQ's

Let the i-th cell of Q_N have the *cell volume*

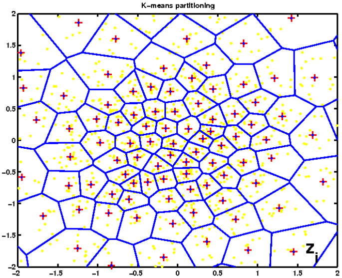
$$V_i = \mathrm{vol}(S_i) = \int_{S_i} dx,$$

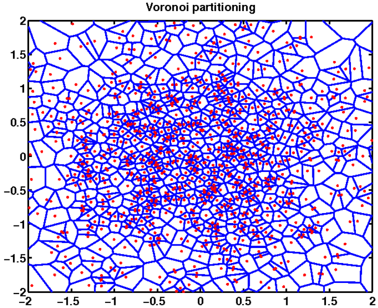
the piecewise constant *point density* function

$$\zeta(x) = \frac{1}{NV_i}, \quad \mathrm{f}or\ x \in S_i$$

and the *specific inertial profile*

$$m(x) = \frac{\int_{S_i} \|y - x_i\|^2 dy}{V_i^{1+2/d}}, \quad \mathrm{f}or\ x \in S_i$$





....

Sequence of high rate VQs of 2D Gaussian source (N=250,500)

Assuming that Q_N converges we have the Bennett integral represention (Na and Neuhoff 1995)

$$\lim_{N\to\infty} N^{2/d} E[\|X - Q_N(X)\|^2] = \int \frac{q(x)m(x)}{\zeta^{2/d}(x)} dx$$

Proof:

I. Facts about spheres $S_i = \mathcal{S}\left(\frac{x-x_i}{r}\right)$ centered at x_i of volume V_i in $\mathbb{R}^d$.

- $V_i = c_1 r^d$, i.e., $r = c_2 V_i^{1/d}$
- $\int_{S_i} (x - x_i) dx = 0$
- $\int_{S_i} \|x - x_i\|^2 dx = c_3 V_i^{\frac{d+2}{d}}$

II. Summation representation of MSQE for smooth $q(x)$

$$
\begin{aligned}
\mathrm{M}SQE &= \sum_i \int_{S_i} \|x - x_i\|^2 q(x) dx \\
&\approx \sum_i q(x_i) \int_{S_i} \|x - x_i\|^2 dx \\
&= \sum_i q(x_i) m(x_i) V_i^{\frac{d+2}{2}}, \qquad \left(m(x_i) \stackrel{\mathrm{def}}{=} \frac{\int_{S_i} \|x - x_i\|^2 dx}{V_i^{\frac{d+2}{2}}} \right) \\
&= \sum_i q(x_i) m(x_i) \frac{1}{(N\zeta(x_i))^{2/d}} V_i, \qquad \left(\zeta(x_i) \stackrel{\mathrm{def}}{=} \frac{1}{NV_i} \right) \\
&= \frac{1}{N^{2/d}} \int \frac{q(x) m(x)}{\zeta(x)^{2/d}} dx
\end{aligned}
$$

Recall Bennett's integral representation

$$\lim_{N\to\infty} N^{2/d} E[\|X - Q_N(X)\|^2] = \int \frac{q(x)m(x)}{\zeta^{2/d}(x)} dx$$

By Hölder's inequality or calculus of variations can easily show

$$\int \frac{q(x)m(x)}{\zeta^{2/d}(x)} dx \leq \left(\int [q(x)m(x)]^{\frac{d}{d+2}} dx \right)^{\frac{d+2}{d}}$$

with equality when "optimal point density" is used

$$\zeta(x) = \frac{[q(x)m(x)]^{\frac{d}{d+2}}}{\int [q(y)m(y)]^{\frac{d}{d+2}} dy}$$

Under Gersho's congruent cell hypothesis, $m(x) = m_d$ independent of x and we obtain Zador-Gersho formula

$$\mathrm{MSQE} = \frac{m_d}{N^{2/d}} \left(\int q^{\frac{d}{d+2}}(x) dx \right)^{\frac{d+2}{d}}$$

Alternative form of Zador-Gersho formula: for fixed encoder complexity $\log N$

$$\frac{d}{2}\log(\mathrm{MSQE}/m_d) = -\log N + \frac{1}{1-\alpha}\log\left(\int q^\alpha(x)dx\right) = -\log N + H_\alpha(X)$$

with $\alpha = \frac{d}{d+2}$ or, for fixed MSQE, the required encoder complexity is

$$\log N = H_\alpha(X) - c$$

Thus: Rényi entropy of source X controls the rate of decrease of the optimal lossy encoder distortion.

Conclude: Rényi entropy captures encoder complexity

- Discrete source: the depth of the lossless Huffman encoder
- Continuous source: the depth of lossy encoder with specified MSQE

Let Y be a random variable representing side information at encoder and decoder for compression of X and define $q(x|y)$ the conditional distribution of X given Y.

Then the depth of the optimal encoder of X given side information $Y = y$ is

Lossless "siege" encoder (Discrete sources with $|\mathcal{X}| = N$)

$$\log N = H_\alpha(X|Y = y)$$

where $\alpha = \frac{1}{1+t}$ and

$$H_\alpha(X|Y = y) = \frac{1}{1-\alpha} \log \sum q^\alpha(x|y)$$

Lossy VQ encoder (Continuous sources encoded with N cell VQ)

$$\log N = H_\alpha(X|Y = y) - c$$

where $\alpha = \frac{d}{d+2}$ and

Side information and conditional Rényi encoding

Average depth over Y for these encoders proportional to conditional Rényi entropy, e.g.,

$$E[\log N] = H_\alpha(X|Y) - c = \int q(y) \left(\frac{1}{1-\alpha} \log \int q^\alpha(x|y) dx \right) dy - c$$

Shannon limits of conditional Rényi encoding complexity:

- Lossless coding: as $t \to 0$ average complexity converges to discrete Shannon conditional entropy $H_1(X|Y) = H(X|Y)$
- Lossy coding: as $d \to \infty$ average complexity converges to cts Shannon conditional entropy $H_1(X|Y) = H(X|Y)$

Shannon entropy and maximum likelihood estimation

Assume measurement X is a realization from a model density $f(X|\mathbf{Y})$ given parameter vector $\mathbf{Y} = Y_1, \ldots, \mathbf{Y}_p$.

Let $\mathbf{X} = X_1, \ldots, X_n$ be i.i.d. sample from $f(X|\mathbf{Y})$ for given $\mathbf{Y}$

Maximum likelihood estimator of $\mathbf{Y}$ given $\mathbf{X}$ maximizes the likelihood function $f(\mathbf{X}|\mathbf{Y})$

$$\hat{\mathbf{Y}} = \mathrm{argmax}_{\mathbf{y}} \prod_{i=1}^{n} f(X_i|\mathbf{y}) = \mathrm{argmax}_{\mathbf{y}} \sum_{i=1}^{n} \ln f(X_i|\mathbf{y})$$

When p is unknown one can try to jointly estimate $\mathbf{Y}, p$.

$$\hat{\mathbf{Y}}, \hat{p} = \mathrm{argmax}_{\mathbf{y},p} \prod_{i=1}^{n} f(X_i|\mathbf{y}) = \mathrm{argmax}_{\mathbf{y},p} \sum_{i=1}^{n} \ln f(X_i|\mathbf{y})$$

Problem: model overfitting - a sufficiently complex model ($p \geq n$) can perfectly fit a finite data sample.

(A) Soln: Penalize the likelihood function for model overcomplexity (Rissanen, Wallace) [6],[8] A lossy source coding derivation of Rissanen's Minimum Description Length penalty

Let $P(\mathbf{Y})$ be a prior distribution on the parameter vector. Assume each of the components of $\mathbf{Y} = [Y_1, \ldots, Y_p]$ is

- continuous valued
- independent identically distributed (iid)

Then the joint complexity of the data and the model is

$$
\begin{aligned}
H([\mathbf{X}, \mathbf{Y}]) &= H(\mathbf{X}|\mathbf{Y}) + H(\mathbf{Y}) \\
&= -E[\log f(\mathbf{X}|\mathbf{Y})] - E[\log f(\mathbf{Y})] \\
&= -\sum_{i=1}^{n} E[\log f(X_i|\mathbf{Y})] - \sum_{j=1}^{p} \underbrace{E[\log f(Y_j)]}_{-H(Y_j)}
\end{aligned}
$$

Assuming large n

$$H([\mathbf{X},\mathbf{Y}]) = -\sum_{i=1}^{n} \log f(X_i|\mathbf{Y}) + \sum_{j=1}^{p} H(Y_i)$$

If discretize Y_i with an N bit quantizer then for N large

$$\log N = H_\alpha(X) - c \geq H(X) - c$$

$\alpha = 1/3$ ($d = 1$)

For quantization loss to be neglible relative to estimation loss: require MSQE on Y_i be of same order as minimum MSE of an optimal estimator of Y_i given $\mathbf{X}$

$$O(N^{-1/2}) = \mathrm{MSQE} = \mathrm{MSEE} = O(n^{-1})$$

or $N = n^2$

Obtain for large n

$$H([\mathbf{X},\mathbf{Y}]) \leq -\sum_{i=1}^{n} \log f(X_i|\mathbf{Y}) + 2p \log n$$

When right hand side is minimized over $\mathbf{Y}$, p obtain equivalent estimator to Rissanen's MDL penalized MLE.

Entropy estimation

Let $h(f)$ be defined as a functional of f for given function ϕ

$$h(f) = \int \phi(f(x))dx$$

Example, $\phi(f) = f^\alpha/(1-\alpha)$

$$h(f) = \frac{1}{1-\alpha}\int f^\alpha(x)dx$$

Question: how to estimate h from empirical data?

Two methods to be discussed here

- Explicit density plug-in estimator
$$\hat{h} = h(\hat{f}), \quad \hat{f} = \hat{f}(X_1, \ldots, X_n)$$
- Estimation without explicit plug-in
$$\hat{h} = \hat{h}(X_1, \ldots, X_n)$$

Entropy estimation in high dimensions

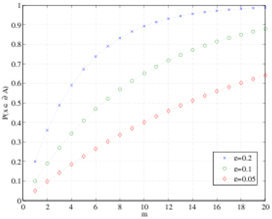
Some peculiarities of high dimensional data (Theorem)

Let $X = [x_1, \ldots, x_d]$ be a random vector uniformly distrbuted in unit cube $[0, 1]^d$

Theorem: for any $\epsilon > 0$

$$P(\epsilon \le x_i \le 1 - \epsilon,\ \forall\ i) \le e^{-2\epsilon d}$$

Thus, as $d \to \infty$, X escapes to the "edge" of cube with overwhelming probability - even though X uniform!



Using the i.i.d. property of components of X

$$\begin{aligned} P(\epsilon \leq x_i \leq 1-\epsilon,\ \forall\ i) &= \prod_{i=1}^{d} P(\epsilon \leq x_i \leq 1-\epsilon) \\ &= (1-2\epsilon)^d \\ &= \exp(d \log(1-2\epsilon)) \\ &\leq \exp(-2\epsilon d) \qquad (\log(1+t) \leq t) \end{aligned}$$

Assume $X_1, \ldots, X_n$ are i.i.d. source symbols uniformly distributed in unit cube $[0,1]^d$.

Theorem: for any $0 < r < 1$

$$P(\min_{j \neq i} \|X_i - X_j\| > r) = (1 - V_d r^d)^{n-1}$$

Thus, as $d \to \infty$ nearest neighbor distances are greater than $1 - \epsilon$ with overwhelming probability.

$\Rightarrow$ the samples X_i become increasingly isolated near the boundaries of $[0,1]^d$!

$$
\begin{aligned}
P(\min_{j\neq i} \|X_i - X_j\| > r) &= \int P(\min_{j\neq i} \|X_i - X_j\| > r | X_i) f(X_i) dX_i \\
&= \int P^{n-1}(\|X_i - X_j\| > r | X_i) f(X_i) dX_i \\
&= \int (1 - V_d r^d)^{n-1} f(X_i) dX_i \\
&= (1 - V_d r^d)^{n-1}
\end{aligned}
$$

For a sample of n i.i.d. realizations from a d-dimensional uniform density over $[0,1]^d$

- As dimension d increases almost all realizations cluster near boundaries of cube
- This phenomenon is due to the increased likelihood of a large deviation in one of components of an X_i.
- Similar phenomenon occurs for non-uniform density supported on $[0,1]^d$.
- Difficult to discriminate between densities differing near the mean but having similar tails.

These pecularities are not mere artifacts for uniform density $f(x) = I_{[0,1]^d}(x)$.

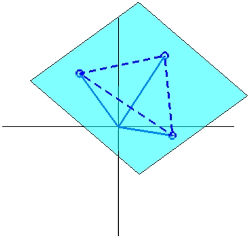
Example: $X_1, \ldots, X_n$ i.i.d. with standard d-variate normal Gaussian density. Then (Marron 2008)

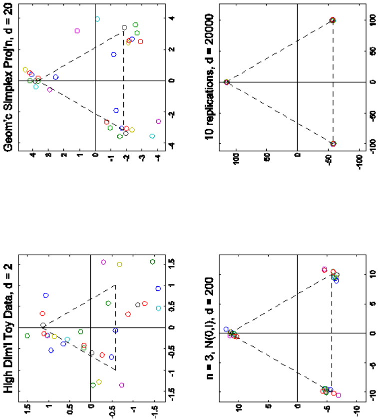
- $\|X_i\| = \sqrt{d} + O(1)$: samples lie on surface of sphere of fixed radius
- $\|X_i - X_j\| = \sqrt{2d} + O(1)$: samples become increasing seperated
- $\cos^{-1}\left(\frac{X_i^T X_j}{\|X_i\|\|X_j\|}\right) = 90^o + O(1/\sqrt{d})$: samples become pairwise equidistant and orthogonal

Entropy estimation in high dimensions

Some peculiarities of high dimensional data

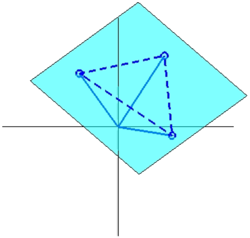
Examples (Marron 2008)

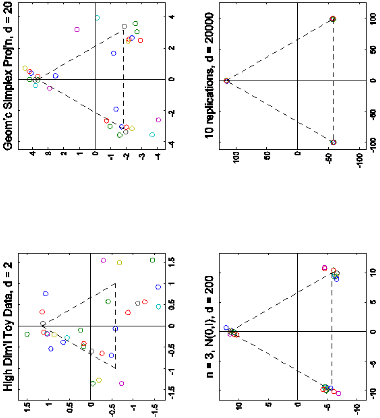




Conclude: Density estimation will become difficult as d increases.

Examples (Marron 2008)





Conclude: Density estimation will become difficult as d increases.

How to estimate the density $f(x)$ of a source X?

Some proposed methods

- Parameteric density estimators
- Histogram estimators
- kNN density estimators
- Kernel density estimators

There exists much theory on density estimation that has been applied to optimize and compare performance Devroye and Lugosi 2001 [1], Devroye 1987 [2], Marron and Hall and Hu [5].

Density estimation
Problem setup

Assume i.i.d. observations: $X_1, \ldots, X_n$ over $\mathbb{R}^d$

Generating density: $X\tilde{f}$, $f : \mathbb{R}^d \to [0, \infty)$

Function class: $f \in \mathcal{F}$ is restricted to be smooth

A density estimator $\hat{f}$ is a function on $\mathbb{R}^d$ indexed by the sample

$$\hat{f}(x) = \hat{f}(x; X_1, \ldots, X_n), \quad x \in \mathbb{R}^d$$

Assume that density class $\mathcal{F} = \mathcal{F}_\Theta = \{g_\theta : \theta \in \Theta\}$ is a family of functions parameterized by a small number of parameters $\theta = [\theta_1, \ldots, \theta_p]$.

Parametric $\hat{\theta}$ estimator of θ provides plug-in estimator of density

$$\hat{f}(x) = g_{\hat{\theta}}(x)$$

Most common approach: maximum likelihood parameter estimator

$$\hat{\theta} = \mathrm{argmax}_{\theta \in \Theta} \prod_{i=1}^{n} g_\theta(X_i) = \mathrm{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \log g_\theta(X_i)$$

Properties of MLE for finite dimensional smooth densities

- Strong consistency: $\hat{\theta} \to \theta$ (w.p.1)
- Asymptotic unbiasedness: $E_\theta[\hat{\theta}] \to \theta$
- Minimum asymptotic covariance: $\mathrm{cov}_\theta(\hat{\theta}) = \frac{1}{n}\mathbf{F}_\theta = \frac{1}{n}E_\theta[-\nabla^2 \log f_\theta(X_1)]$

If $f \in \mathcal{F}_\Theta$ then parametric density estimator has many desirable properties - inherited from finite dimensional MLE $\hat{\theta}$ (Ibragimov and Hasminkii [4])

For all $x \in \mathbb{R}^d$

- $\hat{f} \rightarrow f(x) = g_\theta(x)$ (w.p.1)
- $E[\hat{f}(x)] \rightarrow f(x)$ as $n \rightarrow \infty$, estimator is asymptotically unbiased
- $\mathrm{var}(\hat{f}(x)) = O(1/n)$ for large n
- MSE decreases at rate $1/n$

$$E[(\hat{f}(x) - f(x))^2] = \mathrm{var}(\hat{f}(x)) + (E[\hat{f}(x)] - f(x))^2 = O(1/n)$$

Equivalently:

$$\sqrt{E[(\hat{f}(x) - f(x))^2]} = O(1/\sqrt{n})$$

and we say that the density estimator MSE has "root-n consistency"

It is more customary to use the mean integrated squared error to measure performance of a density estimator

$$\mathrm{MISE} = \int E[(\hat{f}(x) - f(x))^2] dx$$

When f has bounded support, these properties guarantee that MISE also has root-n consistency

If $f \notin \mathcal{F}_\Theta$ then parametric density estimator is not consistent (Ibragimov and Hasminkii [4])

For all $x \in \mathbb{R}^d$

- $\hat{f} \to g_{\theta_o} \neq f(x)$ (w.p.1), where $\theta_o = \mathrm{argmin}_{\theta \in \Theta} D(f \| f_\theta)$.
- $E[\hat{f}(x)] \to g_{\theta_o}$, irreducible bias
- $\mathrm{var}(\hat{f}(x)) = O(1/n)$, dominated by bias

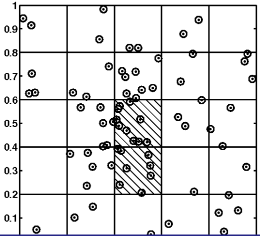
MISE does not converge to zero in limit of large sample size

Assume $f(x)$ has support in $[0,1]^d$ and let $\{S_i\}$ be a uniform partition of $[0,1]^d$ into N cells each of volume $1/N$.

Define $n_j = \sum_{i=1}^n I_{S_j}(X_i)$ the number of observations falling into cell S_j

The histogram density estimator is the peicewise constant function

$$\hat{f}(x) = \sum_{j=1}^{N} \frac{n_j}{n|S_j|} I_{S_j}(x)$$



For large N:

$$\mathrm{MISE} \approx \frac{N}{n} + N^{-2/d} c$$

$c = \frac{1}{4} \int \mathrm{tr}\left(\nabla^2 f(x)\right) dx$

To ensure bounded MISE, assume $\mathcal{F}$ is a set of smooth densities satisfying $c(f) \leq c_{max}$.

- Variance ($\frac{N}{n}$) does not depend on f but bias ($N^{-2/d} c$) does.
- Maximum MISE over $f \in \mathcal{F}$ is worst case MISE

$$\max_f \mathrm{MISE} = \frac{N}{n} + N^{-2/d} c_{max}$$

- Worst case MISE has a bias vs variance tradeoff over N

Using

$$\hat{f}(x) = \sum_{j=1}^{N} \frac{n_j}{n|S_j|} I_{S_j}(x), \quad E[\hat{f}(x)] = \sum_{j=1}^{N} \frac{p_j}{|S_j|} I_{S_j}(x)$$

where $p_j = P(X_i \in S_j)$ and $|S_j| = 1/N$.

$$\begin{aligned} \text{MISE} &= \int \text{var}(\hat{f}(x))dx + \int (E[\hat{f}(x)] - f(x))^2 dx \\ &= \sum_{j=1}^{N} \int_{S_j} \text{var}(\hat{f}(x))dx + \sum_{j=1}^{N} \int_{S_j} (E[\hat{f}(x)] - f(x))^2 dx \\ &= \sum_{j=1}^{N} \int_{S_j} \frac{1}{|S_j|^2} \text{var}\left(\frac{n_j}{n}\right) dx + \sum_{j=1}^{N} \int_{S_j} (p_j/|S_j| - f(x))^2 dx \end{aligned}$$

$$\begin{aligned}\text{MISE} &= \sum_{j=1}^{N} \frac{1}{|S_j|} \frac{1}{n} p_j(1-p_j) + \sum_{j=1}^{N} \int_{S_j} \frac{1}{2}(x-x_j)\nabla^2 f(x_j)(x-x_j)dx \\ &= \frac{N}{n}\sum_{j=1}^{N} p_j(1-p_j) + \sum_{j=1}^{N} \frac{1}{2}\mathrm{tr}\left(\int_{S_j}(x-x_j)^T(x-x_j)^T dx \; \nabla^2 f(x_j)\right)\end{aligned}$$

Note: As S_i is a cube in $\mathbb{R}^d$ with side $N^{-1/d}$

$$\int_{S_j}(x-x_j)(x-x_j)^T dx = N^{-2/d}\frac{|S_j|}{2}\mathbf{I}$$

and $\sum_{j=1}^{N} p_j(1-p_j) = 1 + O(1/N)$

Therefore

$$\mathrm{MISE} = \frac{N}{n} + N^{-2/d} \sum_{i=1}^{N} \mathrm{tr}\left(\nabla^2 f(x_i)\right) \frac{1}{4N}$$

Again, for large N:

$$\mathrm{MISE} \approx \frac{N}{n} + N^{-2/d} c$$

The histogram density estimator bias-variance tradeoff is optimized by choosing N increasing in n at optimal rate that minimizes maximum MISE.

Theorem: Define $N_{opt} = \mathrm{argmin}_N \max_{f \in \mathcal{F}} \mathrm{MISE}$. Then $N_{opt} = (c_{max} n)^{\frac{d}{d+2}}$ and resulting minimax MISE is

$$\mathrm{M}ISE^* = \min_N \max_{f \in \mathcal{F}} \mathrm{M}ISE = a n^{-\frac{2}{d+2}}$$

where $a = (2c_{max}/d)^{\frac{d}{d+2}} + c(2c_{max}/d)^{\frac{-2}{d+2}}$

Recall form of plug-in entropy estimator

$$\hat{h} = h(\hat{f})$$

Define norm $\|\hat{f} - f\|^2 = \int (\hat{f} - f(x))^2 dx$.

Theorem: Assume

1. MISE-consistent $\hat{f}$: $\lim_{n \to \infty} \int E[(\hat{f}(x) - f(x))^2] dx = 0$ (w.p.1))
2. $h(f) = \int \phi(f(x)) dx$ is a smooth functional of f
3. $\int |\phi'(f(x))|^2 dx < \infty$

Then $\hat{h}$ is a consistent estimator of entropy.

Furthermore, if minimax histogram estimator is used then for large n

$$E[(\hat{h} - h)^2] = bn^{-\frac{2}{d+2}}$$

We have

$$\hat{h} = h(\hat{f}) = h(f) + \int \phi'(f(x))(\hat{f}(x) - f(x))dx + O(\|\hat{f}(x) - f(x)\|)$$

By CS inequality

$$\left(\int \phi'(f(x))(\hat{f}(x) - f(x))dx\right)^2 \le \int |\phi'(f(x))|^2 dx \int (\hat{f}(x) - f(x))^2 dx$$

which converges to zero as $n \to \infty$.

Recall that for the minimax histogram estimator

$$MISE^* = \int E[(\hat{f}^* - f)^2] = an^{-\frac{2}{d+2}}$$

which guarantees that MSE of $\hat{h}$ will have the same rate.

Drawbacks of density estimation methods for entropy estimation

- Bandwidth selection $\sigma = N^{-1/d}$ may be difficult
- Datastructures for histograms are impractical in very high dimensions
- Convergence rate becomes logarithmic in N for large d

$$N^{-1/d} = \frac{d}{d + \log N} + O(1/d)$$

- May have few samples (fewer than dimensions) in some cases
- Density estimation in very high dimensions is fraught with difficulties

L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*, Springer, New York, 2001.

Devroye:87, *A course in density estimation*, Birkhäuser-Verlag, Boston, 1987.

D. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, pp. 1098–1102, Sept. 1952.

I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: Asymptotic theory*, Springer-Verlag, New York, 1981.

J. Marron, P. Hall, and T. Hu, "Improved variable window estimators of probabiity density," *Annals of Statistics*, vol. 23, pp. 1–10, 1995.

J. Rissanen, "A universal prior for the integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, pp. 416–431, 1983.

C. Vignat, A. Hero, and J. Costa, "About closedness by convolution of the tsallis maximizers," *Physica A*, vol. 340, no. 1-3, pp. 147–152, 2004.

C. Wallace and D. Dowe, "Minimum message length and kolmogorov complexity," *The Computer*, vol. 42, no. 4, pp. 270–283, 1999.