ZADOR-GERSHO THEOREM FOR VARIABLE-RATE VQ

For a stationary source and large R, the least distortion of k-dim'l VQ with nth-order entropy coding and rate R or less is

$$\delta(k,n,R) \cong m_k^* \sigma^2 \eta_{kn} 2^{-2R} \triangleq Z(k,n,R)$$

where

$Z(k,n,R)$ = Zador-Gersho funct. for k-dim'l VQ with nth-order entropy coding

m_k^* = best inert'l profile = least NMI of any tessel'g polytope (Gersho's conj.)

$\eta_k = \frac{1}{\sigma^2} 2^{2h_k}$

Equivalently, $S(k,n,R) \cong 6.02\,R - 10\log_{10} m_k^* \eta_{kn}$. (Again, 6 dB per bit.)

Notes:

- k-dim'l VQ-EC with n = 1 is at least as good as k-dim'l VQ-FR, because the latter is a special case of the former.

  Later we show directly that $\eta_k \le \beta_k$, which implies $Z(k,1,R) \le Z(k,R)$.
- The proof shows that an approximately optimal k-dimensional VQ with nth-order variable-rate coding can be constructed with a partition that is a tesselation of the best k-dimensional polytope, scaled to volume $2^{k(h_{kn}-R)}$. This has

  constant inert'l profile $m(\underline{x}) = m_k^*$, constant pt density $\Lambda(\underline{x}) = \Lambda_k^* = 2^{k(R-h_{kn})}$,

  distortion $D \cong m_k^* (\Lambda_k^*)^{-2/k} = Z(k,nR)$.

PROOF OF ZADOR-GERSHO THEOREM

We begin with

$$\delta(k,n,R) \cong \min_{m(\underline{x}),\Lambda(\underline{x})} \int \frac{m(\underline{x})}{\Lambda^{2/k}(\underline{x})} f(\underline{x})\,d\underline{x}$$

where $\underline{x} = (x_1 \ldots x_k)$ and the minimization is over all inertial profiles $m(\underline{x})$ and all point densities $\Lambda(\underline{x})$ such that

$$h_{kn} + \frac{1}{k}\int f(\underline{x}) \log \Lambda(\underline{x})\,d\underline{x} \le R \qquad (\text{such a } \Lambda \Rightarrow \text{rate} \lessapprox R)$$

- Best inertial profile:

  We assume Gersho's conjecture -- In the high rate, small distortion regime, most cells of an optimal quantizer are, approximately, congruent to the tesselating polytope with least NMI.

  We conclude: The best inertial profile is

  $$m_k^*(x) = m_k^* \triangleq \text{least NMI of k-dimen'l tess'ng polytopes}$$

- It follows that

  $$\delta(k,n,R) \cong m_k^* \min_{\Lambda(\underline{x})} \int \frac{1}{\Lambda^{2/k}(\underline{x})} f(\underline{x})\,d\underline{x}$$

  where the min is taken over functions $\Lambda(\underline{x})$ such that $\Lambda(\underline{x}) \ge 0$ and

  $$h_{kn} + \frac{1}{k}\int f(\underline{x}) \log \Lambda(\underline{x})\,d\underline{x} \le R$$

- Best point density:

Suppose $\Lambda(\underline{x})$ satisfies

$$h_{kn} + \frac{1}{k}\int f(\underline{x}) \log_2 \Lambda(\underline{x})\, d\underline{x} \le R \qquad (*)$$

Then by convexity of the logarithm and Jensen's inequality

$$\log_2 \int \frac{1}{\Lambda^{2/k}(\underline{x})} f(\underline{x})\, d\underline{x} \ \ge\ \int \log_2\left(\frac{1}{\Lambda^{2/k}(\underline{x})}\right) f(\underline{x})\, d\underline{x}$$

equality iff $\Lambda(\underline{x})$ is constant with probability one

$$= -\frac{2}{k}\int f(\underline{x}) \log \Lambda(\underline{x})\, d\underline{x} \ \ge\ 2h_{kn} - 2R$$

Hence,

$$\int \frac{1}{\Lambda^{2/k}(\underline{x})} f(\underline{x})\, d\underline{x} \ \ge\ 2^{2h_{kn}} 2^{-2R} \qquad (**)$$

with equality iff $\Lambda(\underline{x})$ is a constant with probability one.

$(*)$ and $(**)$ $\Rightarrow$ $\delta(k,n,R) \cong m_k^* 2^{2h_{kn}} 2^{-2R} = m_k^* \sigma^2 \eta_{kn}\, 2^{-2R}$

Moreover, we have shown that the optimal point density is a constant. The constant must be such that $(*)$ holds with equality. Therefore,

$$\Lambda(\underline{x}) = 2^{k(R-h_{kn})} \triangleq \Lambda_k^*$$

We see from the proof that an approx'ly optimal VQ can be constructed with a partition that is a tesselation of the best k-dim'l polytope, scaled to volume $2^{k(h_{kn}-R)}$. The tesselation need only cover the region where $f(\underline{x})$ is not small.

SUMMARY OF FIXED- AND VARIABLE-RATE VQ

Let 0th-order entropy coding (n=0) denote fixed-rate coding.

Given a stationary source and large R, the least distortion of VQ with dimension k, nth-order entropy coding, and rate R or less is

$$\delta(k,n,R) \cong m_k^* \sigma^2 \alpha_{k,n}\, 2^{-2R} \quad (= Z(k,n,R))$$

where

$$\alpha_{k,n} = \begin{cases} \beta_k, & n=0 \\ \eta_{kn} = \frac{1}{\sigma^2} 2^{2h_{kn}}, & n \ge 1 \end{cases}$$

- Notice the "," in $\alpha_{k,n}$ but not in η_{kn} or h_{kn} !
- The best k-dimen'l VQ to use with fixed-rate coding (n=0) has
  - has point density $\lambda_k^*(\underline{x}) = c\, f^{k/(k+2)}(\underline{x})$
  - congruent cells with NMI = m_k^* (constant inert'l profile).
- The best k-dimen'l VQ to use with variable-rate binary coding (n≥1) has
  - uniform point density
  - congruent cells with NMI equal to m_k^*. That is, it is simply a tesselation.

WHAT HAPPENS AS k AND n CHANGE?

As usual, consider a stationary source.

Recall:

$$\delta(k,n,R) \cong \sigma^2 m_k^* \alpha_{k,n} 2^{-2R}$$

m_k^* decreases subadditively to $m_\infty^* = \frac{1}{2\pi e}$

$$\alpha_{k,n} = \begin{cases} \beta_k , & n = 0 \\ \frac{1}{\sigma^2} 2^{2h_{kn}} , & n \geq 1 \end{cases}$$

β_k decreases subadditively to β_∞

2^{2h_k} decreases monotonically to 2^{2h_∞}

Therefore,

$\alpha_{k,0}$ decreases monotonically with k to β_∞

$\alpha_{k,n}$ decreases monotonically with k to 2^{2h_∞} for $n \geq 1$

$\alpha_{k,n}$ decreases monotonically with n to 2^{2h_∞}

Key Fact: $2^{2h_\infty} = \beta_\infty$ (proved later)

Therefore,

> $\alpha_{k,n}$ decreases monotonically with n or k to $\beta_\infty = 2^{2h_\infty}$

CONCLUSIONS

(1) The least distortion of vector quantization with rate R or less, with any dimension, and with fixed-rate coding or with variable-rate encoding of any order is

$$\delta(R) \cong \sigma^2 m_\infty^* \beta_\infty 2^{-2R}$$

Among other things, this says that the best possible performance with variable-rate coding is no better than the best possible performance with fixed-rate coding.

(2) Increasing n with k fixed:

$\delta(k,n,R)$ decreases monotonically to the limit

$$\delta(k,\infty,R) \cong \sigma^2 m_k^* \beta_\infty 2^{-2R} \cong \frac{m_k^*}{m_\infty^*} \delta(R) \qquad \text{(space filling loss)}$$

> Therefore, for large n and arbitrary k,
>
> $$\delta(k,n,R) \cong \frac{m_k^*}{m_\infty^*} \delta(R)$$

Among other things, this shows that one needs large k in order to approach the best possible performance.

(3) Increasing k with n fixed:

$\delta(k,n,R)$ "decreases", though not monotonically, to the limit

$$\delta(\infty,n,R) \cong \sigma^2 m^*_\infty \beta_\infty 2^{-2R} = \delta(R) \qquad \text{(no loss)}$$

Therefore, for large k and arbitrary n (even $n = 0$),

$$\delta(k,n,R) \cong \delta(R)$$

Among other things this indicates that for large k, increasing n does not improve the best possible performance attainable with for that k. That is, one can attain the best possible performance, even with $n = 0$ or 1.

(4) To get the best possible performance we must have

(a) k large enough that $m^*_k/m^*_\infty \cong 1$,

i.e. well shaped cells,

(b) k and/or n large enough that $\alpha_{k,n} \cong \beta_\infty$,

i.e. good point density and good exploitation of memory.

Usually (b) is more important (a).

(5) What's the point of variable-rate coding if the best possible performance with variable-rate coding is no better than the best possible performance without it?

From the point of view of VQ, the reason to use variable-rate coding, instead of fixed-rate coding) is to permit a VQ with smaller dimension (less complexity) to work well.

The extreme case -- k=1, i.e. scalar quantization:

- with $n = 0$ (fixed-rate coding)

$$\delta(1,0,R) \cong \sigma^2 m^*_1 \beta_1 2^{-2R} = \frac{m^*_1}{m^*_\infty}\frac{\beta_1}{\beta_\infty}\delta(R)$$

- with large n (high-order variable-rate coding)

$$\delta(1,\infty,R) \cong \sigma^2 m^*_1 \beta_\infty 2^{-2R} = \frac{m^*_1}{m^*_\infty}\delta(R), \quad \text{where } \frac{m^*_1}{m^*_\infty} = 1.42 \text{ or } 1.53 \text{ dB}$$

The variable-coding causes β_1 to be replaced by β_∞.

Moreover, the best scalar quantizer for use with variable-rate coding is a uniform scalar quantizer.

Thus this shows that with variable-rate coding, a uniform scalar quantizer can have performance within only 1.53 dB of the best VQ of any type!

(6) What's the point of vector quantization if uniform scalar quantization plus variable-rate coding can come within 1.53 dB of the best VQ of any type?

From the point of view of the binary code, the purpose of VQ is to permit a lower order variable-rate coder to be used, i.e. it permits a simpler lossless coder. For example, if uniform scalar quantization were used, the variable-rate coder must exploit the memory in the source, i.e. it would have to be complex.

VQ also reduces the space filling loss, i.e. it improves cell shapes.

(7) It's worth re-emphasizing that one can attain

- the best possible performance (D vs. R) by choosing k large and n = 0; i.e. with a complex quantizer and a simple fixed-rate binary encoder.
- the best possible performance minus only 1.53 dB by choosing a uniform scalar quantizer and an entropy coder with large n; i.e. with a simple quantizer and a complex entroopy coder.

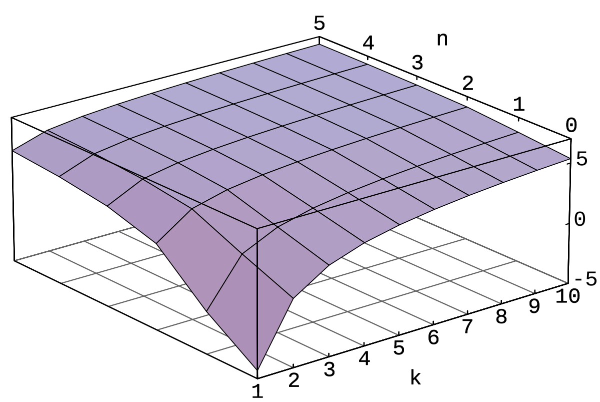
Which is simpler? Hard to say. Both are very complex. Good systems are usually compromises; i.e. a nontrivial quantizer and nontrivial entropy coder.

In some applications variable-rate coding is not an option and so fixed-rate coding must be used.

EXAMPLE: GAUSS-MARKOV SOURCE, $\rho = .9$

$$S(k,n,R) \cong 6.02\,R - 10 \log_{10} m_k^* \alpha_{k,n}$$

Plot of $-\log_{10} \alpha_{k,n}$



Properties of Differential Entropy and the Zador Factor H_K

Most are extensions of properties of the differential entropy of one random variable. Many proofs are similar to those of analogous properties for ordinary entropy.

Definitions:

$$h(X_1,\ldots,X_k) \stackrel{\Delta}{=} -\int f(\underline{x}) \log_2 f(\underline{x})\, d\underline{x}$$

$$h_k \stackrel{\Delta}{=} \frac{1}{k} h(X_1,\ldots,X_k) = \text{kth order differential entropy}$$

$$\eta_k \stackrel{\Delta}{=} \frac{1}{\sigma^2} 2^{2h_k} = \text{Zador factor for variable-rate VQ}$$

Note that h and h_k can be positive or negative!

(1) $h_k \le \frac{1}{2}\log_2 \sigma^2\beta_k$ and $\eta_k \le \beta_k$

where β_k is Zador's factor. Equality holds in each iff $f(\underline{x})$ has the same value wherever it is not zero, e.g. if it is uniform.

Derivation:

$$h_k = -\frac{1}{k}\int f(\underline{x}) \log_2 f(\underline{x})\, d\underline{x} = \frac{k+2}{2k}\int f(\underline{x}) \log_2 f^{-2/(k+2)}(\underline{x})\, d\underline{x}$$

$$= \frac{k+2}{2k} E \log_2 Y, \quad \text{where } Y = f^{-2/(k+2)}(\underline{X})$$

$$\le \frac{k+2}{2k} \log_2 EY \quad \text{by Jensen's ineq. } (\log_2 \text{ is concave})$$

$$= \frac{k+2}{2k} \log_2 \int f(\underline{x})\, f^{-2/(k+2)}(\underline{x})\, d\underline{x}$$

$$= \frac{1}{2}\log_2 \left(\int f^{k/(k+2)}(\underline{x})\, d\underline{x} \right)^{(k+2)/k} = \frac{1}{2}\log_2 \sigma^2 \beta$$

$$\Rightarrow \eta_k = \frac{1}{\sigma^2} 2^{2h_k} \le \beta_k$$

Since $\log_2$ is strictly concave, equality holds if and only if Y is constant w.p.1;
i.e. iff $P(f^{-2/(k+2)}(\underline{X}) = c) = 1$ for some some c,
i.e. iff $f(\underline{x})$ has the same value wherever it is not zero.

(2) If $\underline{Y} = a\,\underline{X} + \underline{b}$, $a \neq 0$, then

$$h_{Y,k} = h_{X,k} + \log_2 |a| \quad \text{and} \quad \eta_{Y,k} = \eta_{X,k}$$

Derivation: Since $f_{\underline{Y}}(\underline{y}) = \frac{1}{|a|^k} f_{\underline{X}}(\frac{\underline{y}-\underline{b}}{a})$,

$$h_{Y,k} = -\frac{1}{k}\int f(\underline{y}) \log_2 f(\underline{y})\, d\underline{y}$$

$$= -\frac{1}{k}\int \frac{1}{|a|^k} f_{\underline{X}}(\tfrac{\underline{y}-\underline{b}}{a}) \log_2\left(\frac{1}{|a|^k} f_{\underline{X}}(\tfrac{\underline{y}-\underline{b}}{a})\right) d\underline{y}$$

$$= -\frac{1}{k}\int \frac{1}{|a|^k} f_{\underline{X}}(\underline{x}) \log_2\left(\frac{1}{|a|^k} f_{\underline{X}}()\right) |a|^k d\underline{x}, \quad \text{letting } \underline{x} = \frac{\underline{y}-\underline{b}}{a}$$

$$= -\frac{1}{k}\int f(\underline{x}) \log_2 f(\underline{x})\, d\underline{x} - \frac{1}{k}\int f(\underline{x}) \log_2 |a|^k\, d\underline{x}$$

$$= h_{X,mk} + \log_2 |a|$$

$$\sigma_Y^2 = a^2\sigma_X^2$$

$$\Rightarrow \eta_{Y,k} = \frac{1}{\sigma_Y^2} 2^{2h_{Y,k}} = \frac{1}{a^2\sigma_X^2} 2^{2h_{X,k}+2\log_2|a|} = \frac{1}{\sigma_X^2} 2^{2h_{X,k}} = \eta_{X,k}$$

(3) (a) If $\underline{Y} = A\,\underline{X} + \underline{b}$ and A is a $k\times k$ nonsingular matrix, then

$$h_{Y,k} = h_{X,k} + \frac{1}{k}\log_2 |A| \quad \text{and} \quad \eta_{Y,k} = \eta_{X,k}\, |A|^{2/k} \frac{\sigma_X^2}{\sigma_Y^2}$$

(b) If A is orthogonal (i.e. $A^{-1} = A^t$), then

$$h_{Y,k} = h_{X,k} \quad \text{and} \quad \eta_{Y,k} = \eta_{X,k}.$$

Derivation:

(a) Since $f_{\underline{Y}}(\underline{y}) = |A|^{-1} f_{\underline{X}}(A^{-1}(\underline{y}-\underline{b}))$,

$$h_{Y,k} = -\frac{1}{k}\int f(\underline{y}) \log_2 f(\underline{y})\, d\underline{y}$$

$$= -\frac{1}{k}\int |A|^{-1} f_{\underline{X}}(A^{-1}(\underline{y}-\underline{b})) \log_2\left(|A|^{-1} f_{\underline{X}}(A^{-1}(\underline{y}-\underline{b}))\right) d\underline{y}$$

$$= -\frac{1}{k}\int |A|^{-1} f_{\underline{X}}(\underline{x}) \log_2\left(|A|^{-1} f_{\underline{X}}(\underline{x})\right) |A|\, d\underline{x}, \quad \text{with } \underline{x} = A^{-1}(\underline{y}-\underline{b})$$

$$= -\frac{1}{k}\int f(\underline{x}) \log_2 f(\underline{x})\, d\underline{x} - \frac{1}{k}\int f(\underline{x}) \log_2 |A|^{-1} d\underline{x} = h_{X,k} + \frac{1}{k}\log_2 |A|$$

$$\Rightarrow \eta_{Y,k} = \frac{1}{\sigma_Y^2} 2^{2h_{Y,k}} = \frac{1}{a^2\sigma_X^2} 2^{2h_{X,k}+\frac{1}{k}\log_2|A|} = \frac{1}{\sigma_X^2} 2^{2h_{X,k}} |A|^{2/k} \frac{\sigma_X^2}{\sigma_Y^2} = \eta_{X,k}\, |A|^{2/k} \frac{\sigma_X^2}{\sigma_Y^2}$$

(b) This follows from (a) and the facts that when A is orthogonal, $|A| = 1$ and $\sigma_Y^2 = \sigma_X^2$.

(4) Gaussian: If $\underline{X} = (X_1,\ldots,X_k)$ is Gaussian with covariance matrix K, then

$$h_k = \frac{1}{2}\log_2 2\pi e|K|^{1/k} \quad\text{and}\quad \eta_k = 2\pi e\frac{|K|^{1/k}}{\sigma^2}$$

This formula for η_k is the same as the formula for β_k except that $((k+2)/k)^{(k+2)/2}$ is replaced by "e". (Note: $((k+2)/k)^{(k+2)/2} \to e$ as $k \to \infty$.)

Derivation:

We may assume $\underline{X}$ has zero mean, since previous properties show the mean has no effect on h_k or η_k. Let $\underline{Y} = A\underline{X}$, where A is the Karhunen-Loeve Transform. Then $\underline{Y}$ is Gaussian with covariance matrix $\Lambda = AKA^t$, which is diagonal, with diagonal elements equal to the eigenvalues $\lambda_1,\ldots,\lambda_k$ of K. Thus $\underline{Y}$ has independent components with variances $\lambda_1,\ldots,\lambda_k$. Note that $|K| = \prod_{i=1}^k \lambda_i$. Since A is orthogonal

$$h_{X,k} = h_{Y,k} = -\frac{1}{k}\int f(\underline{y})\log_2 f(\underline{y})\,d\underline{y} = -\frac{1}{k}\int f(\underline{y})\log_2\left(\left(\prod_{i=1}^k 2\pi\lambda_i\right)^{-1/2}\exp\left\{-\sum_{i=1}^k \frac{y_i^2}{2\lambda_i}\right\}\right)d\underline{y}$$

$$= -\frac{1}{k}\int f(\underline{y})\log_2\left(\left(\prod_{i=1}^k 2\pi\lambda_i\right)^{-1/2}\exp\left\{-\sum_{i=1}^k \frac{y_i^2}{2\lambda_i}\right\}\right)d\underline{y}$$

$$= -\frac{1}{k}\log_2\left(\prod_{i=1}^k 2\pi\lambda_i\right)^{-1/2} + \frac{1}{k}\int f(\underline{y})\sum_{i=1}^k \frac{y_i^2}{2\lambda_i}\,d\underline{y}\,\log_2 e$$

$$= -\frac{1}{k}\log_2 (2\pi)^{-k/2}|K| + \frac{1}{k}\sum_{i=1}^k \frac{EY_i^2}{2\lambda_i}\log_2 e$$

$$= \frac{1}{2}\log_2 2\pi|K|^{1/k} + \frac{1}{2}\log_2 e = \frac{1}{2}\log_2 2\pi e|K|^{1/k}$$

(5) If $\underline{X}$ has covariance matrix K, then

$$h_k \le \frac{1}{2}\log_2 2\pi e|K|^{1/k} \quad\text{and}\quad \eta_k \le 2\pi e\frac{|K|^{1/k}}{\sigma^2}$$

with equality iff $\underline{X}$ is Gaussian

Derivation:

We may assume $\underline{X}$ has zero mean, since the mean has no effect on h_k or η_k. Let $f(\underline{x})$ be the density of $\underline{X}$ with covariance matrix K, and let $g(\underline{x})$ be the Gaussian density with mean zero and covariance matrix K.
We make the proof in two steps.

(a) $h_k(f) \le -\frac{1}{k}\int f(\underline{x})\log_2 g(\underline{x})\,d\underline{x}$:

$$-\frac{1}{k}\int f(\underline{x})\log_2 g(\underline{x})\,d\underline{x} - h_k(f) = -\frac{1}{k}\int f(\underline{x})\log_2\frac{g(\underline{x})}{f(\underline{x})}\,d\underline{x}$$

$$\ge -\frac{1}{k}\int f(\underline{x})\left(\frac{g(\underline{x})}{f(\underline{x})} - 1\right)d\underline{x}\,\frac{1}{\ln 2}, \qquad \left(\log_2 z \le (z-1)\frac{1}{\ln 2}\right)$$

$$= -\frac{1}{k}\int g(\underline{x})\,d\underline{x}\,\frac{1}{\ln 2} + \int f(\underline{x})\,d\underline{x}\,\frac{1}{\ln 2} = -1 + 1 = 0$$

(b) $-\frac{1}{k}\int f(\underline{x}) \log_2 g(\underline{x})\, d\underline{x} = -\frac{1}{k}\int g(\underline{x}) \log_2 g(\underline{x})\, d\underline{x} = \frac{1}{2}\log_2 2\pi e|K|^{1/k}$:

$$-\frac{1}{k}\int f(\underline{x}) \log_2 g(\underline{x})\, d\underline{x} = -\frac{1}{k} E \log_2\left((2\pi)^{-k/2}|K|^{-1/2}\exp\left\{-\frac{1}{2}\underline{X}^t K^{-1}\underline{X}\right\}\right) d\underline{x}$$

$$= \frac{1}{2}\log_2 2\pi|K|^{1/k} + \frac{1}{2k} E_f\left[\underline{X}^t K^{-1}\underline{X}\right] \log_2 e$$

where E_f denotes expecation with respect to f.
Since the expectation is a sum of terms of the form $a_{ij} E[X_i X_j]$, it depends only on the covariance matrix K. Therefore, it will be the same if the expectation is taken with respect to g, because g has the same covariance matrix.
Therefore

$$-\frac{1}{k}\int f(\underline{x}) \log_2 g(\underline{x})\, d\underline{x} = \frac{1}{2}\log_2 2\pi|K|^{1/k} + \frac{1}{2k} E_g\left[\underline{X}^t K^{-1}\underline{X}\right] \log_2 e$$

$$= -\frac{1}{k}\int g(\underline{x}) \log_2 g(\underline{x})\, d\underline{x}$$

$$= \frac{1}{2}\log_2 2\pi e|K|^{1/k} \quad \text{by (4)}$$

Definition:

The *conditional differential entropy* of random variables $X_1,\ldots,X_k$ given random variables $Y_1,\ldots,Y_m$

$$h(X_1,\ldots,X_k|Y_1,\ldots,Y_m) \triangleq -\int f(\underline{x},\underline{y}) \log_2 f(\underline{x}|\underline{y})\, d\underline{x}\, d\underline{y}$$

Most of the following properties are derived in the same way as the corresponding propety for entropy.

(6) $h(X|Y) \le h(X)$ with equality iff X and Y are independent.

Derivation: We'll show $h(X) - h(X|Y) \ge 0$ with equality iff X indep of Y.

$$h(X) - h(X|Y) = -\int f(x,y) \log_2 f(x)\, dx\, dy + \int f(x,y) \log_2 \frac{f(x,y)}{f(y)}\, dx\, dy$$

$$= -\int f(x,y) \ln \frac{f(x)f(y)}{f(x,y)}\, dx\, dy\, \frac{1}{\ln 2}$$

$$\ge -\int f(x,y) \left(\frac{f(x)f(y)}{f(x,y)} - 1\right) dx\, dy\, \frac{1}{\ln 2} \quad \text{since } \ln z \le z-1$$

$$= -\int f(x)f(y)\, dx\, dy\, \frac{1}{\ln 2} + \int f(x,y)\, dx\, dy\, \frac{1}{\ln 2} = 0.$$

Equality holds if and only if $f(x)f(y) = f(x,y)$ for all x,y; i.e. if and only if X and Y are independent.

(7) $h(Y_1,\ldots,Y_n|X_1,\ldots,X_m) \le h(Y_1,\ldots,Y_n|X_1,\ldots,X_{m'})$, $0 \le m' < m$,

with equality iff $Y_1,\ldots,Y_n$ is conditionally independent of $X_{m'+1},\ldots,X_m$ given $X_1,\ldots,X_{m'}$.

Derivation: Similar to that of (6).

(8) Chain rule:

$$h(X_1,\ldots,X_k) = h(X_1)+h(X_2|X_1)+H(X_3|X_1X_2)+ \ldots +h(X_k|X_1\ldots X_{k-1})$$

Derivation: Essentially the same proof as for the chain rule for ordinary entropy, but with H's replace by h's.

(9) $h(X_1,\ldots,X_k) \le h(X_1)+ \ldots + h(X_k)$

with equality if and only if X_i's are independent

Derivation: Essentially the same proof as for the analogous property for ordinary entropy.

STATIONARY SOURCES

Definitions:

$h_k = \frac{1}{k} h(X_1,\ldots,X_k)$

$h_{1|m} = h(X_n|X_{n-m},X_2,\ldots,X_{n-1})$ ($\underline{h}_1 = h(X_1) = h_1$)

$h_{k|m} = \frac{1}{k} h(X_n^{n+k-1}|X_{n-m}^{n-1})$ ($h_{1|k} = \underline{h}_1$)

$h_\infty = \lim_{k\to\infty} h_k$ = differential entropy-rate of X

Properties:

(10) $\underline{h}_{k+1} \le \underline{h}_k$

Derivation: Follows from (7) and stationarity.

(11) $h_k = \frac{1}{k}(h_1+h_{1|1}+h_{1|2}+\ldots+h_{1|k-1}) \ge h_{1|k-1} \ge h_{1|k-1}$

Derivation: Essentially the same as the analogous property for entropy.

$$h_k = \frac{1}{k} h(X_1,\ldots,X_k)$$
$$= \frac{1}{k}\Big(h(X_1)+h(X_2|X_1)+h(X_3|X_1,X_2) +\ldots+ h(X_k|X_1,X_2,\ldots,X_{k-1})\Big) \quad \text{chain rule (8)}$$
$$= \frac{1}{k}\Big(h_1+h_{1|1}+h_{1|2}+\ldots+h_{1|k-1}\Big) \quad \text{by stationarity}$$
$$\ge \frac{1}{k}\Big(h_{1|k-1} + h_{1|k-1} + h_{1|k-1} + \ldots + h_{1|k-1}\Big) = h_{1|k-1} \ge h_{1|k-1} \quad \text{by (10)}$$