(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) $h_{1|k+1} \le h_{1|k}$

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.

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

$$= \frac{1}{k}\Big(h(X_1)+h(X_2|X_1)+h(X_3|X_1,X_2) + \ldots + h(X_k|X_1,X_2,\ldots,X_{k-1})\Big) \quad \text{chain rule (8)}$$

$$= \frac{1}{k}\big(h_1+h_{1|1}+h_{1|2}+\ldots+h_{1|k-1}\big) \quad \text{by stationarity}$$

$$\ge \frac{1}{k}\big(h_{1|k-1} + h_{1|k-1} + h_{1|k-1} + \ldots + h_{1|k-1}\big) = h_{1|k-1} \ge h_{1|k} \quad \text{by (10)}$$

(12) $h_{k+1} \le h_k$

It follows that $h_\infty = \lim_{k\to\infty} h_k$ is a well-defined quantity, because the h_k's are nonincreasing and bounded below by zero, they must have a limit.

Derivation: By (11), h_k is the average of the k terms $h_1, h_{1|1}, h_{1|2},\ldots,h_{1|k-1}$.
Similarly, h_{k+1} is the average of the $k+1$ terms $h_1, h_{1|1}, h_{1|2},\ldots,h_{1|k-1}, h_{1|k}$.
Since the extra term in h_{k+1} is no larger than all other terms, $h_{k+1} \le h_k$.

(13) $$\lim_{k\to\infty} h_{1|k} = h_\infty$$

Derivation: Since the h_k's are nonincreasing with k and bounded below by zero, they have a limit.
Since by Prop. 10, h_k is the average of the k terms $h_1, h_{1|1}, h_{1|2},\ldots,h_{1|k-1}$, the limit of the $h_{1|k}$'s equals the limit of the h_k's, which by definition is h_∞.

(14) For an IID source,

$$h_1 = h_k = h_\infty \quad \text{all } k$$

Derivation: This follows from (9).

(15) For a stationary Markov source

$$h_k = \frac{1}{k} h(X_1) + \frac{k-1}{k} h(X_2|X_1)$$

$$h_\infty = h(X_2|X_1) = h_{1|1}$$

Proof: See the proof of (11).

(16) For a stationary, first-order, autoregressive source, $(X_k = a X_{k-1} + W_k$ where the W_k's are IID and independent of past X_k's$)$, (it's Markov, too),

$$h_k = \frac{1}{k} h(X_1) + \frac{k-1}{k} h(W_1)$$

$$h_\infty = h(X_2|X_1) = h(W_1) \quad \text{and} \quad \eta_\infty = \eta_{W,1}.$$

Derivation: Use (15) and the fact that

$$h(X_2|X_1) = h(aX_1+W_2|X_1) = h(W_2|X_1) \quad \text{(because } aX_1 \text{ is a constant)}$$
$$= h(W_2) \quad \text{because } W_2 \text{ and } X_1 \text{ are independent}$$
$$= h(W_1) \quad \text{because } W_2 \text{ and } W_1 \text{ are independent}$$

(17) For a stationary Gaussian source

$$h_\infty = \frac{1}{2} \log_2 2\pi e Q \quad \text{and} \quad \eta_\infty = 2\pi e \frac{Q}{\sigma^2}$$

where $Q = \exp\left\{\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln S(\omega)\, d\omega\right\} =$ one-step prediction error,

and $S(\omega)$ = power spectral density.

Derivation: From (4)

$$h_\infty = \lim_{k\to\infty} h_k = \lim_{k\to\infty} \frac{1}{2} \log_2 2\pi e |K|^{1/k} = \frac{1}{2} \log_2 2\pi e Q$$

where $\lim_{k\to\infty} |K|^{1/k} = Q$ will be demonstrated later in the context of transform coding.

(18) For any stationary source

$$h_\infty \le \frac{1}{2}\log_2 2\pi e Q \quad \text{and} \quad \eta_\infty \le 2\pi e \frac{Q}{\sigma^2}$$

with equality if only if the source is Gaussian.

Derivation: This follows from $h_k \le h_{k,Gauss}$ for all k (see (5)) and the fact that

$$h_\infty = \lim_{k\to\infty} h_k \le \lim_{k\to\infty} h_{k,Gauss} = h_{\infty,Gauss} = \frac{1}{2}\log_2 2\pi e Q.$$

(19) For a stationary, first-order autoregressive, Gaussian source with correlation coefficient ρ

$$h_k = \frac{1}{2}\log_2 2\pi e\sigma^2 + \frac{k-1}{k}\frac{1}{2}\log_2(1-\rho a^2)$$

$$h_\infty = \frac{1}{2}\log_2 2\pi e\sigma^2(1-\rho^2) \quad \text{and} \quad \eta_\infty = 2\pi e(1-\rho^2)$$

Derivation: Uses (16).

(20) For a stationary source, $\eta_\infty = \beta_\infty$

Sketch of Derivation: Uses the following fundamental result of information theory:

Asymptotic Equipartition Property (AEP):

For a stationary source and all sufficiently large k,

$$\Pr(\underline{X} \in T_k) \cong 1$$

where $T_k = \left\{\underline{x} : \frac{1}{k}\log_2 f_k(\underline{x}) \cong h_\infty\right\} = \left\{\underline{x} : f_k(\underline{x}) \cong 2^{-kh_\infty}\right\}$

That is, $f_k(\underline{x}) \cong 2^{-kh_\infty}$ with high probability.

Here's how we use it:

$$\log_2 \sigma^2\beta_\infty \cong \log_2 \sigma^2\beta_k = \log_2\left(\int f_k(\underline{x})^{k/(k+2)} d\underline{x}\right)^{(k+2)/k} \quad \text{for large values of } k$$

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

because $\Pr(\underline{X} \in T_k) \cong 1$

$$\cong \frac{k+2}{k}\log_2 \left(2^{-kh_\infty}\right)^{-2/(k+2)} \quad \text{because } f_k(\underline{x}) \cong 2^{-kh_\infty} \text{ for } \underline{x} \in T_k.$$

$$\cong 2^{2h_\infty} \quad \text{since } k \text{ is large.}$$

$$\Rightarrow \quad \beta_\infty = \frac{2^{2h_\infty}}{\sigma^2} = \eta_\infty$$

WHY DOES VQ-VR ATTAIN THE PERFORMANCE THAT IT DOES?

Compare the point density and cell shapes of

$Q_{k'}$ = optimal k'-dimensional variable-rate VQ

Q_k = optimal k-dimensional variable-rate VQ, where k = large multiple of k'.

$Q_{k'}$ is tesselation of polytope attaining $m^*_{k'}$, with uniform pt. density & inert'l profile.

Q_k is tessel'n of polytope attaining $m^*_k \cong 1/(2\pi e)$. (Polytope ≅ high dimen'l sphere.)

To compare them, consider the k-dimensional product quantizer $Q_{pr,k}$ that is formed by using $Q_{k'}$ k/k' times in succession.

$Q_{pr,k}$ has the same rate, distortion and "shortcomings" as $Q_{k'}$.

$\lambda_{pr,k}(\underline{x})$ = uniform point density in dimension k = best point density

$m_{pr,k}(\underline{x}) = m^*_{k'} > m^*_k$

We see that the only shortcoming of k'-dimensional VQ-VR relative to high-dimensional VQ-VR is the space-filling loss:

$$L_{sp} = \frac{m^*_{k'}}{1/(2\pi e)} .$$

With VQ-VR, there is no point density loss, even for sources with memory. In effect, entropy coding has eliminated the need to compromise between good point density and small oblongitis.

WHY DO VQ-VR AND VQ-FR GIVE SAME PERFORMANCE FOR LARGE k?

Consider their properties on $T_k = \{\underline{x} : f_k(\underline{x}) \cong 2^{-kh_\infty}\}$, which has $\Pr(\underline{X} \in T_k) \cong 1$

Cell shapes: Both have $m(\underline{x}) \cong \frac{1}{2\pi e}$ = NMI of high dim'l sphere

Point density:

VQ-VR: $\lambda(\underline{x})$ = constant

VQ-FR: $\lambda^*_k(\underline{x}) = c\, f_k^{k/(k+2)}(\underline{x}) \cong f_k(\underline{x}) \cong \begin{cases} 2^{kh_\infty}, & \underline{x} \in T_k \\ 0, & \text{else} \end{cases}$

⇒ VQ-VR and VQ-FR have same inertial profile and same point density in T_k, where it matters.

Binary encoders

VQ-FR: all codewords have length kR

VQ-VR: codeword for cell $S_{\underline{x}}$ containing $\underline{x} \in T_k$ has

$$\text{length} \cong -\log_2 \Pr(S_{\underline{x}}) \cong -\log_2 \left(f_k(\underline{x})) \times |S_{\underline{x}}|\right) \cong kh_\infty - \log_2 \frac{1}{\Lambda(\underline{x})} \quad \text{for } \underline{x} \in T_k$$

$$\cong \text{constant} \cong kR$$

⇒ VQ-VR and VQ-FR assign same lengths to $\underline{x}$ in T_k

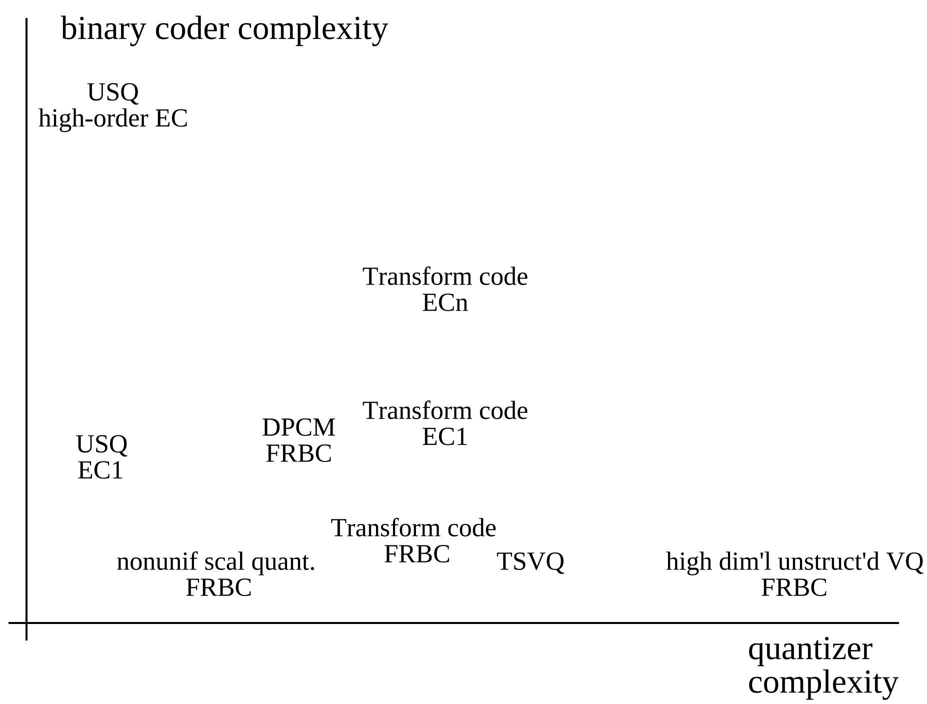
VQ-FR "uses" pt. density to "ignore" $\underline{x} \notin T_k$. VQ-VR assigns short codewords to T_k.

Second view:

VQ-FR: $\Pr(\text{cell containing } \underline{x}) = \Pr(S_{\underline{x}}) \cong f_k(\underline{x})\,|S_{\underline{x}}| \cong f_k(\underline{x}) \dfrac{1}{M\lambda_k^*(\underline{x})} \cong \dfrac{1}{M}$

Since all cells have the same probability, variable-rate coding gains nothing.

Qualitative plot of quantizer and entropy coder complexity



FRBC = Fixed-Rate Binary Code

ECn = nth-order Entropy Code

Positions of codes are subjective. Please don't quote me.