On Variable to Fixed length codes, Source Coding and Rényi's Entropy

Vaneet Aggarwal

Y1385

April 15, 2005

Project Supervisor

Dr. R.K. Bansal

Department of Electrical Engineering,

Indian Institute of Technology,

Kanpur, India.

Table of Contents

Introduction

Shannon information measure has a very concrete operational interpretation: it roughly equals the minimum number of binary digits needed, on the average, to encode the message in question. The coding theorems of information theory provide such overwhelming evidence for the adequateness of the Shannon information measure that to look for essentially different measures of information might appear to make no sense at all. Moreover, it has been shown by several authors, starting with Shannon (1948), that the measure of amount of information is uniquely determined by some rather natural postulates. Still, all the evidence that the *Shannon information measure* is the only possible one is valid only within restricted scope of coding problems considered by Shannon. As pointed out by Rényi (1960)[5] in his fundamental paper on generalized information measures, in other sort of problems other quantities may serve just as well, or even better, as measures of information. This should be supported either by their operational significance or by a set of natural postulates characterizing them, or, preferably, by both. Thus the idea of generalized entropies arises in the literature. It started with Rényi (1960) [5] who characterized scalar parametric entropy as *entropy of order alpha,* which includes Shannon entropy as a limiting case. Later, Campbell [4] proved a coding theorem by using Rényi's entropy. Most of the known properties of Rényi's entropy and its characterization have been summarized by Aczel and Daroczy[7]. Rényi also emphasized the need for an operational characterization of his information measures, and in [6] he gave one for H_α (p) with an integer $\alpha \geq 2$, in terms of a search problem.

We propose a theorem to generalize the Tunstall codes using the Rényi's entropy which will increase the significance of this measure, and give an algorithm to minimize the redundancy. We will also see the consequences of two different definitions of mutual information to the generalization of capacity and rate distortion function. We will then extend the Tunstall theorem to countable alphabet.

Background

Rényi's Entropy, a generalization of Shannon's Entropy

Rényi[5] introduced a one-parameter family of information measures. His entropy is given by

$H_\alpha(P) = \frac{1}{1-\alpha}\log_2 \sum_{i=1}^{m} p_i^{\alpha}$ where $P = (p_1, p_2, ...p_m)$, $\alpha \neq 1$

Here, $0^\alpha = 0$ for all real α

In the limit as $\alpha \to 0$, $H_\alpha(P)$ goes to the Hartley's case and as $\alpha \to 1$, $H_\alpha(P)$ tends to the Shannon's entropy H(P)

$H(P) = -\sum_i p_i \log_2 p_i$

From now on, **log will mean log to the base 2** unless otherwise specified.

$0 \le H_\alpha(P) \le H_h(P)$ =log(m) where $H_h(P)$ is the Hartley's Entropy.

Rényi's entropy is symmetric, normalized, expansible, decisive, additive, nonnegative and measurable like Shannon's entropy.[7]

If $\alpha \ge 0$, Rényi's entropy is maximal, bounded, and monotonic like Shannon's entropy.

If α >0, Rényi's entropy is continuous, and small for small probabilities ($\lim_{q\to 0+} H_\alpha(1-q,q) = 0$) like Shannon's entropy.

$H_\alpha(P)$ is **subadditive(**H_α(X,Y) $\le$ H_α(X)+H_α(Y)) **if, and only if,** α **=0 or** α **=1** [12].

Recall that above holds $\forall\ \alpha$ if X and Y are independent. This leads us to the following counter example $\forall\ \alpha$ not equal to 0 or 1:

P(X=0,Y=0) = pq+e

P(X=0,Y=1) = p(1-q)-e

P(X=1,Y=0) = (1-p)q-e

P(X=1,Y=1) = (1-p)(1-q)+e

0<p<1, 0<q<1, p≠1/2, q≠1/2 and $|e| \le$ min(pq,(1-p)q,(1-q)p, (1-p)(1-q))

If H_α(X,Y) $\le$ H_α(X)+H_α(Y) were true, the function

$f(e) = (pq+e)^\alpha + (p(1-q)-e)^\alpha + ((1-p)q-e)^\alpha + ((1-p)(1-q)+e)^\alpha$

Would have an extremum for e=0. But this is not true as derivative of f(e) wrt e is not 0 at e=0.

$H_\alpha(P)$ is strictly concave with respect to P for $0 < \alpha \le 1$

For the case m=2, $H_\alpha(P)$ is strictly concave w.r.t. P for $0 < \alpha \le 2$

For $\alpha > 2$ and $m \ge 2$, $H_\alpha(P)$ is neither convex nor concave w.r.t. P. [9]

Generalized Huffman Codes

Shannon gave the entropy concept operational significance by proving the first noiseless source coding theorem.

***Shannon's Noiseless Coding Theorem*:** In the absence of noise, it is always possible to encode or transmit a message with a number of bits arbitrarily close to the entropy of the message, but never less than the entropy

A problem left open by Shannon in his 1948 paper was the determination of a noiseless prefix code for a random variable taking finitely many values which is optimal in the sense of minimizing the expected codeword length which was first done in 1952 by David Huffman.

***Campbell's Coding Theorem*:** In 1965, Campbell generalized the Shannon's Noiseless coding theorem using Rényi's entropy.He showed that Rényi's entropy is a lower bound on the exponentially weighted average codeword length

L= $\frac{1}{s}\log(\sum_{i=1}^{m} p_i 2^{sl_i})$, s>0

And the Rényi's entropy has parameter α =1/(s+1) $H_s(P) = \frac{s+1}{s}\log(\sum_{i=1}^{m} pi^{1/(s+1)})$ s>0

$$\frac{1}{s}\log(\sum_{i=1}^{m} p_i 2^{sl_i}) \geq \frac{s+1}{s}\log(\sum_{i=1}^{m} p_i^{1/(s+1)})$$

When s $\rightarrow 0^+$, we have the Shannon's version.

$$\sum_{i=1}^{m} p_i l_i \geq -\sum_{i=1}^{m} p_i \log p_i$$

This can be proved easily using the condition $\sum_{i=1}^{m} 2^{-l_i} \leq 1$ (Kraft-McMillan Inequality)

The redundancy of the code given by $\frac{1}{s}\log(\sum_{i=1}^{m} p_i 2^{sl_i})$ -H_s(P) is minimized by generalized Huffman algorithm. This algorithm is similar to Huffman algorithm except that the new node is assigned the weight $2^s(p_i+p_j)$ where p_i and p_j are the lowest weights on the available nodes.

The generalized Huffman Algorithm finds the optimal solution of the non-linear integer problem.
The equality holds if

$l_i = \log(\sum_{j=1}^{m} p_j^{1/(s+1)}) - \frac{1}{s+1}\log(p_i)$ gives integer lengths. In any case, the inequality is satisfied by letting $l_i = \lceil \log(\sum_{j=1}^{m} p_j^{1/(s+1)}) - \frac{1}{s+1}\log(p_i) \rceil$

Section 1: Generalized Tunstall Codes

Variable-length-to-block coding is a technique of data compression that seems especially attractive for a skew source (where the frequency of some output letters very much exceeds that of others) or for retrieval situations that require blocks formatting of data. Variable-length-to-block coding was considered by Tunstall[15] who described an encoding construction and proved it optimal in a sense that the compression ratio cannot be decreased below the Shannon's entropy and the Tunstall Algorithm minimize the compression ratio.

Tunstall Theorem:

Assign a codeword of length $\lceil \log(m) \rceil$ bits to the ordinal of a source string identified with a leaf in a tree with m leaves. Let the leaf depths and probabilities be l_i's and p_i's respectively. There is an efficient algorithm to minimize the redundancy. This algorithm is called Tunstall Algorithm. In this, we form a tree by splitting the leaf with maximum probability at each step into two till we have m leaves starting from the tree having just the root node. During splitting, assign the probability to the left leaf as p times the probability of node that is split, and the probability of the right leaf as q times the probability of node that is split, where p+q=1.

$$\frac{\lceil \log(m) \rceil}{\sum_{i=1}^{m} p_i l_i} \geq H(p) \text{ where } \mathrm{H(p)} = -p\log(p) - q\log(q)$$

Motivated by Campbell's generalization of the Shannon's noiseless coding theorem, we consider the Rényi's entropy to establish the following generalization of Tunstall Theorem.

Generalized Tunstall Theorem:

$$\frac{\log(m)}{\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})} \geq H_s(p) \text{ ----------(1)}$$

$$H_s(p) = \frac{s+1}{s}\log(p^{1/(s+1)} + q^{1/(s+1)}) \text{ where q=1-p and s>0 ----------(2)}$$

When $s \to 0^+$, we have the Tunstall theorem.

The Left hand Side of eqn 1 is referred to as *generalized compression ratio* and the denominator as *generalized average length* at the input.We will now give some lemmas that will help in the proof of Generalized Tunstall Theorem, and then give an algorithm (Generalized Tunstall Algorithm) to minimize the redundancy.

Remark: Both sides of inequation (1) increase with s as is shown in Lemma 1 and in [8].

Lemma 1: For any given full binary tree with m leaves, $\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})$ decreases as s increases where s>0.

Proof:

Let $l(s) \triangleq \frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})$.

$$l'(s) = \frac{s.\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}} - \log(\sum p_i 2^{-sl_i})}{s^2}$$ (The sum unless specified otherwise runs from 1 to m)

To prove $l'(s) \le 0$, we need to prove $s.\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}} - \log(\sum p_i 2^{-sl_i}) \le 0$ or

$s.\sum p_i l_i 2^{-sl_i} - \sum p_i 2^{-sl_i} \ . \ \log(\sum p_i 2^{-sl_i}) \le 0$

To prove this, use induction wrt l_i's . For one point set sl_1 2^{-sl_1} -sl_1 2^{-sl_1} =0

Let it be true for a k point set, so, $s.\sum_{i=1}^{k} p_i l_i 2^{-sl_i} + \sum_{i=1}^{k} p_i 2^{-sl_i} \ . \ \log(\sum_{i=1}^{k} p_i 2^{-sl_i}) \le 0$

We can assume $l_1 < l_2 < ..... < l_k$ without loss of generality as all the three sums above can be broken into inner and outer sums where the inner sums are over all i's for which l_i's are same. Now one additional point is introduced say l_{k+1}. When $l_{k+1} = l_k$, we have a k point set. If we prove that while keeping rest of tree fixed, if we increase l_{k+1},

$s.\sum_{i=1}^{k+1} p_i l_i 2^{-sl_i} + \sum_{i=1}^{k+1} p_i 2^{-sl_i} \ . \ \log(\sum_{i=1}^{k+1} p_i 2^{-sl_i})$ decreases, we are done as it will hold for all l_{k+1}. To prove this, we show that $d'(l_{k+1}) \le 0$ where

d(l_{k+1})=$s.\sum_{i=1}^{k+1} p_i l_i 2^{-sl_i} + \sum_{i=1}^{k+1} p_i 2^{-sl_i} \ . \ \log(\sum_{i=1}^{k+1} p_i 2^{-sl_i})$.

Now, $d'(l_{k+1}) =$

$sp_{k+1}(2^{-sl_{k+1}})(1 - sl_{k+1} \ . \ \ln(2)) - sp_{k+1}(2^{-sl_{k+1}}) \ . \ \ln(2) \ . \ \log(\sum_{i=1}^{k+1} p_i 2^{-sl_i}) - sp_{k+1}(2^{-sl_{k+1}}) =$

$sp_{k+1} \ . \ 2^{-sl_{k+1}}(1 - sl_{k+1} \ . \ \ln(2) - \ln(2) \ . \ \log(\sum_{i=1}^{k+1} p_i 2^{-sl_i}) - 1)$

To show $d'(l_{k+1}) \le 0$, we need to show that $-sl_{k+1} - \log(\sum_{i=1}^{k+1} p_i 2^{-sl_i}) \le 0$ which is true as l_{k+1} is the maximum depth. $l_1 < l_2 < ..... < l_{k+1}$.

Remark: Note that l(s)$\rightarrow$ $\sum p_i l_i$ as s$\rightarrow 0^+$ and l(s) $\rightarrow l_{min}$ as s$\rightarrow \infty$

Lemma 2: For any given full binary tree with m leaves, (1) holds for s $\rightarrow$ 0^+ and for s $\rightarrow \infty$

Proof: For s $\rightarrow$ 0^+, (1) reduces to Tunstall theorem which uses Shannon's entropy.

For s $\rightarrow \infty$, $\frac{\log(m)}{l_m} \ge 1$ where l_m is the minimum of all the depths. This is obviously true.

Lemma 3: $$\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i} \ (\frac{-1}{s}\log(\ \sum p_i 2^{-sl_i})\ )} \le 1$$

Proof: This is equivalent to proving $\sum p_i l_i 2^{-sl_i} + \frac{1}{s}\sum p_i 2^{-sl_i} \ (\log(\sum p_i 2^{-sl_i})) \le 0$

Use induction wrt l_k 's as in Lemma 1.

For a one point set, $\sum p_i l_i 2^{-sl_i} + \frac{1}{s}\sum p_i 2^{-sl_i}\ (\log(\sum p_i 2^{-sl_i}))=0$

Let Lemma 3 hold for a k point set, i.e.,

$\sum_{i=1}^{k} p_i l_i 2^{-sl_i} + \frac{1}{s}\sum_{i=1}^{k} p_i 2^{-sl_i}(\log(\sum_{i=1}^{k} p_i 2^{-sl_i})) \le 0$ for $l_1 < l_2 < ..... < l_k$

Let one additional point l_{k+1} be introduced. When $l_{k+1} = l_k$, we have a k point set for which result is assumed to be true. Now, keeping the rest of the tree fixed, if by increasing l_{k+1} by an arbitrary amount, if

u(l_{k+1})$\triangleq \sum_{i=1}^{k+1} p_i l_i 2^{-sl_i} + \frac{1}{s}\sum_{i=1}^{k+1} p_i 2^{-sl_i}(\log(\sum_{i=1}^{k+1} p_i 2^{-sl_i}))$ decreases, we are done.

How, u'(l_{k+1})= $p_{k+1}2^{-sl_{k+1}}(1 - sl_{k+1}\ln(2)) + \frac{1}{s\ln(2)}(p_{k+1}2^{-sl_{k+1}}(-s\ln(2)(\ln(\sum_{i=1}^{k+1} p_i 2^{-sl_i}) + 1) \le 0$

is equivalent to $p_{k+1}2^{-sl_{k+1}}[1 - sl_{k+1}\ln(2) - \ln(\sum_{i=1}^{k+1} p_i 2^{-sl_i}) - 1] \le 0$. This translates to

$-l_{k+1} - \frac{1}{s}\log(\sum_{i=1}^{k+1} p_i 2^{-sl_i}) \le 0$ which is true as $-\frac{1}{s}\log(\sum_{i=1}^{k+1} p_i 2^{-sl_i})$ decreases with s from av. length to min. length as seen in Lemma 1, and so is always less than the maximum length l_{k+1}. This completes the proof of Lemma 3.

Lemma 4: $\dfrac{\sum p_i l_i 2^{sl_i}}{\sum p_i 2^{sl_i}\ (\frac{1}{s}\log(\ \sum p_i 2^{sl_i})\ )} \ge 1$

Proof: The proof is similar to Lemma 3, and hence is omitted here.

Lemma 5: For a given distribution (p,q) on a two point set, a parameter s>0, let h(s)$\triangleq \frac{s+1}{s}\log(p^{1/(s+1)} + q^{1/(s+1)})$, the Rényi's entropy and $h_{Sh}(s)$ be the Shannon's entropy for the distribution ($\frac{p^{1/(s+1)}}{p^{1/(s+1)} + q^{1/(s+1)}}, \frac{q^{1/(s+1)}}{p^{1/(s+1)} + q^{1/(s+1)}}$). Then, we have $\frac{h_{Sh}(s)}{h(s)} \ge 1$

Proof: To prove $h_{Sh}(s) - h(s) \ge 0$

Or, $\frac{p^{1/(s+1)}}{p^{1/(s+1)} + q^{1/(s+1)}}\log(\frac{p^{1/(s+1)} + q^{1/(s+1)}}{p^{1/(s+1)}}) + \frac{q^{1/(s+1)}}{p^{1/(s+1)} + q^{1/(s+1)}}\log(\frac{p^{1/(s+1)} + q^{1/(s+1)}}{q^{1/(s+1)}}) - \frac{s+1}{s}\log(p^{1/(s+1)} + q^{1/(s+1)}) \ge 0$

Or, $-\frac{1}{s+1}[p^{1/(s+1)}\log(p) + q^{1/(s+1)}\log(q)] - \frac{1}{s}\log(p^{1/(s+1)} + q^{1/(s+1)}) \ge 0$

Or. $[p^{1/(s+1)}\log(p^{-\frac{s}{s+1}}) + q^{1/(s+1)}\log(q^{-\frac{s}{s+1}})] \ge (p^{1/(s+1)} + q^{1/(s+1)})\log(p^{1/(s+1)} + q^{1/(s+1)})$.

Define a r.v X which takes value $p^{-\frac{s}{s+1}}$ with probability p and $q^{-\frac{s}{s+1}}$ with probability q. We have to prove, E(X log(X)) $\ge$E(X).log(E(X)), which is true by Jensen's Inequality.

Lemma 6: For a given full binary tree with m leaves, let

g(s) $\triangleq$ $\dfrac{\log(m)}{\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})}$, so called generalized compression ratio and h(s) $\triangleq$ $\dfrac{s+1}{s}\log(p^{1/(s+1)}+q^{1/(s+1)})$. When g(s) $\geq$ h(s), g'(s)-h'(s) ≤ 0 for s>0.

Proof:

$$g'(s)-h'(s)=\frac{\ln(m)\ (\frac{s\ln(2)\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}}+\ln(\sum p_i 2^{-sl_i})}{s^2})}{(\frac{-1}{s}\ln(\sum p_i 2^{-sl_i}))^2}+\frac{\log(p^{1/(s+1)}+q^{1/(s+1)})}{s^2}+\frac{s+1}{s\ln(2)(s+1)^2}\frac{p^{1/(s+1)}\ln(p)+q^{1/(s+1)}\ln(q)}{p^{1/(s+1)}+q^{1/(s+1)}}$$

$$=\frac{1}{s^2}[\frac{\log(m)\ (s\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}}+\log(\sum p_i 2^{-sl_i}))}{(\frac{-1}{s}\log(\sum p_i 2^{-sl_i}))^2}+\log(p^{1/(s+1)}+q^{1/(s+1)})+\frac{s}{(s+1)}\frac{p^{1/(s+1)}\log(p)+q^{1/(s+1)}\log(q)}{p^{1/(s+1)}+q^{1/(s+1)}}]$$

$$=\frac{1}{s^2}[\frac{\log(m)}{\frac{-1}{s}\log(\sum p_i 2^{-sl_i})}\ .\ \frac{s\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}}+\log(\sum p_i 2^{-sl_i})}{\frac{-1}{s}\log(\sum p_i 2^{-sl_i})}+\log(p^{1/(s+1)}+q^{1/(s+1)})+\frac{s}{(s+1)}\frac{p^{1/(s+1)}\log(p)+q^{1/(s+1)}\log(q)}{p^{1/(s+1)}+q^{1/(s+1)}}]$$

Thus, if g(s) $\geq$ h(s), we have

$$g'(s)-h'(s)\leq\frac{1}{s^2}[\frac{s+1}{s}\log(p^{1/(s+1)}+q^{1/(s+1)})\ .\ \frac{s\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}}+\log(\sum p_i 2^{-sl_i})}{\frac{-1}{s}\log(\sum p_i 2^{-sl_i})}+$$

$$\log(p^{1/(s+1)}+q^{1/(s+1)})+\frac{s}{(s+1)}\frac{p^{1/(s+1)}\log(p)+q^{1/(s+1)}\log(q)}{p^{1/(s+1)}+q^{1/(s+1)}}]\ \text{(using Lemma 1)}$$

$$=\frac{1}{s^2}[(s+1)\log(p^{1/(s+1)}+q^{1/(s+1)})\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}\ (\frac{-1}{s}\log(\sum p_i 2^{-sl_i}))}-(s+1)\log(p^{1/(s+1)}+q^{1/(s+1)})+$$

$$\log(p^{1/(s+1)}+q^{1/(s+1)})+s\frac{p^{1/(s+1)}\log(p^{1/(s+1)})+q^{1/(s+1)}\log(q^{1/(s+1)})}{p^{1/(s+1)}+q^{1/(s+1)}}]$$

$$=\frac{1}{s^2}[sh(s)\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}\ (\frac{-1}{s}\log(\sum p_i 2^{-sl_i}))}-s\log(p^{1/(s+1)}+q^{1/(s+1)})+s\frac{p^{1/(s+1)}\log(p^{1/(s+1)})+q^{1/(s+1)}\log(q^{1/(s+1)})}{p^{1/(s+1)}+q^{1/(s+1)}}]$$

$=\frac{1}{s}[h(s)\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}\ (\frac{-1}{s}\log(\sum p_i 2^{-sl_i}))}-h_{Sh}(s)]$ (where $h_{Sh}(s)$ is defined in Lemma 5).

$=\frac{h(s)}{s}[\frac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}\ (\frac{-1}{s}\log(\sum p_i 2^{-sl_i}))}-\frac{h_{Sh}(s)}{h(s)}]\leq 0$(by Lemma 3,5). This completes the proof of Lemma 6.

Theorem 1: For a given full binary tree with m leaves, $\dfrac{\log(m)}{\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})} \geq H_s(p)$. This is the generalized Tunstall Theorem.

Proof: $\dfrac{\log(m)}{\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})} \geq H_s(p)$ holds for s $\rightarrow$ 0^+ and for s $\rightarrow \infty$ as proved in Lemma 2, so it is enough to prove that the two curves g(s) and h(s) do not cross each other. If the two curves intersect at s_0, the difference at s_0 decreases (by Lemma 6), and thus can never attain a positive value. Since the difference is ≥ 0 as s $\rightarrow \infty$, if the two curves intersect at s_0, then the difference is zero for all s$\geq$ s_0, which will mean equality in (1) as s $\rightarrow \infty$, which is equivalent to saying that the minimum length of the tree=log(m) which means that all lengths of the tree are log(m) for m=2^k, for some k. So, the two curves can't intersect when all the lengths are not identical in which case also (1) holds. This completes the proof of **Generalized Tunstall Theorem.**

Theorem 2: $\dfrac{\log(m)}{\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})} \xrightarrow{m\rightarrow\infty} H_s(p)$

Proof: Since g(s)$\geq$h(s) by Theorem 1, we have g'(s)-h'(s) ≤ 0 by Lemma 6. So, g(s)-h(s) decrease with s. Since for a tree, g(0)-h(0) $\xrightarrow{m\rightarrow\infty}$ 0 by Tunstall Convergence, we have g(s)-h(s) $\xrightarrow{m\rightarrow\infty}$ 0 **.**

Lemma 7: Let T_m be an arbitrary full binary tree with m leaves. Then the generalized average depth of T_m , $L(T_m) \triangleq \dfrac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i}) = \dfrac{-1}{s}\log(\ 1-(1-2^{-s})\ .\ \sum_{\substack{\text{internal}\\ \text{nodes}}} W(x)\ )$

where W(x)= p(x) . $2^{-sl(x)}$

Proof: $L(T_{m+1}) = \dfrac{-1}{s}\log(2^{-sL(T_m)} + p(x)p2^{-s(l(x)+1)} + p(x)q2^{-s(l(x)+1)} - p(x)2^{-s(l(x))})$ where p(x) and l(x) are the probability and the length (depth) of the node split to obtain T_{m+1} from T_m .

$2^{-sL(T_{m+1})} = 2^{-sL(T_m)} - W(x)\ .\ (1-2^{-s})$ where W(x)= p(x) . $2^{-sl(x)}$

$2^{-sL(T_m)} = C-(1-2^{-s})\sum_{\substack{\text{internal}\\ \text{nodes}}} W(x)$, where C is a constant to be determined later.

Note that for m=2, $L(T_m)$=1 and $\sum_{\substack{\text{internal}\\ \text{nodes}}} W(x)$=1, we have C=1

$$L(T_m) = \frac{-1}{s}\log[\ 1-(1-2^{-s})\sum_{\substack{\text{internal}\\ \text{nodes}}} W(x)\ ]$$

Generalized Tunstall Algorithm (GTA)

There is an efficient algorithm to minimize the generalized redundancy which is defined as the difference between generalized compression ratio and the Rényi's entropy. This

algorithm is called Generalized Tunstall Algorithm. In this, we form a tree by splitting the leaf with maximum weight w_i at each step into two till we have m leaves starting from the tree having just the root node where wi= $p_i.2^{-sl_i}$. During splitting, assign the probability to the left leaf as p times the probability of node that is split, and the probability of the right leaf as q times the probability of node that is split, where p+q=1.

Theorem 2: GTA minimizes generalized compression ratio which is bounded below by Rényi's entropy and thus minimizes the generalized redundancy.
Proof:
GTA maximizes $\sum_{\substack{internal \\ nodes}} W(x)$ or equivalently the expression $\frac{-1}{s}\log(1-(1-2^{-s})\sum_{\substack{internal \\ nodes}} W(x))$ the generalized average depth as s>0 as seen in Lemma 7. So, we can not have lesser generalized compression ratio than what is given by GTA.

Remark: Generalized Tunstall Theorem continues to hold for any finite alphabet of size r>2 if we define the generalized average depth l(s) = $\frac{-1}{s}\log_r(\sum_{i=1}^{m} p_i r^{-sl_i})$ and a common base in the definition of Rényi's entropy and log(m).

Further results on generalized compression ratio, tree entropy and generalized average depth.

Lemma 8: Let f(s)= $\frac{1}{s}\log(\sum p_i 2^{sl_i})$, the generalized average word length introduced by Campbell, and g(s)= $-\frac{1}{s}\log(\sum p_i 2^{-sl_i})$, the generalized average word length introduced in this report, then f(s)g(s) ≥ $(\sum p_i l_i)^2$ or f(s)g(s) ≤ $(\sum p_i l_i)^2$ $\forall$ s>0 if l_i 's are not identical where $\sum p_i l_i$ is the average word length used by Shannon.(f(s).g(s)= $(\sum p_i l_i)^2$ as s $\to$ 0^+)
Proof: Consider the function u(s) = f(s) . g(s) - $(\sum p_i l_i)^2$. Now, u(0)=0, so if u(s) changes sign, u' (s)=0 for some s by Rolle's Theorem as u(s) is differentiable. So, (f(s)g(s))'=0 for some s.

Or, $\dfrac{\frac{1}{s}\log(\sum p_i 2^{sl_i})}{\dfrac{\sum p_i l_i 2^{sl_i}}{\sum p_i 2^{sl_i}}} = \dfrac{-\frac{1}{s}\log(\sum p_i 2^{-sl_i})}{\dfrac{\sum p_i l_i 2^{-sl_i}}{\sum p_i 2^{-sl_i}}}$. Here the LHS ≤ 1, and RHS ≥ 1 (Lemma 3, 4)

The equality holds at s=0, s=∞ and in the case when all the l_i 's are identical. All these cases are not possible. So, the theorem is proved. This theorem means that if $l_m l_M < (\sum p_i l_i)^2$, then $f(s)g(s) < (\sum p_i l_i)^2$, if $l_m l_M > (\sum p_i l_i)^2$, then $f(s)g(s) > (\sum p_i l_i)^2$, and if $l_m l_M = (\sum p_i l_i)^2$, then $f(s)g(s) = (\sum p_i l_i)^2$ for all s>0, where l_m and l_M are respectively the minimum and the maximum depths of the tree.

Lemma 9: For any arbitrary full binary tree T_m with m leaves, the Rényi's entropy of the probability distribution of the leaves,

$$H_s(T_m) = \frac{s+1}{s}\log(1+(p^{1/(s+1)}+q^{1/(s+1)}-1)\sum_{\substack{internal\\nodes}}(p(x))^{1/(s+1)})$$

Proof:

$$H_s(T_m) = \frac{s+1}{s}\log(\sum_{i=1}^{m} p_i^{1/(s+1)})$$

Let p(x) be the probability of the node that is split to obtain T_{m+1} from T_m . Then, one can write

$$2^{\frac{sH_s(T_{m+1})}{s+1}} = 2^{\frac{sH_s(T_m)}{s+1}} + (p(x)p)^{1/(s+1)} + (p(x)q)^{1/(s+1)} - (p(x))^{1/(s+1)}$$

$$= 2^{\frac{sH_s(T_m)}{s+1}} + (p^{1/(s+1)}+q^{1/(s+1)}-1)(p(x))^{1/(s+1)} = C + (p^{1/(s+1)}+q^{1/(s+1)}-1)\sum_{\substack{internal\\nodes}}(p(x))^{1/(s+1)}$$

where C is a constant to be determined later.

$$\Rightarrow 2^{\frac{sH_s(T_m)}{s+1}} = C + (p^{1/(s+1)}+q^{1/(s+1)}-1)\sum_{\substack{internal\\nodes}}(p(x))^{1/(s+1)}$$

Putting m=2, we have C=1

$$H_s(T_m) = \frac{s+1}{s}\log(1+(p^{1/(s+1)}+q^{1/(s+1)}-1)\sum_{\substack{internal\\nodes}}(p(x))^{1/(s+1)})$$

Let T_m denote the class of full binary trees with m leaves and

let $r_m \triangleq \min_{T_m \in T_m}[\frac{\lceil \log(m) \rceil}{\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})}]$. Let us consider $\frac{\lceil \log(m) \rceil}{\frac{-1}{s}\log(\sum_{i=1}^{m} p_i 2^{-sl_i})} \geq H_s(p)$ -----(3)

Lemma 10: a) $r_{2^k+1} \geq r_{2^k+2} \geq \ldots\ldots\ldots\ldots \geq r_{2^{k+1}}$

b) $r_{2^{k+1}} \leq r_{2^k} + \frac{1-r_{2^k}}{L(T_{2^k})+1}$

c) $r_{2^k} \leq 1$

Proof:

a) Let us see the proof of $r_{2^k+1} \geq r_{2^k+2}$, rest is similar

To prove this is same as to prove,

$$\frac{k+1}{\frac{-1}{s}\log(\sum_{i=1}^{2^k} p_i 2^{-sl_i} + p(x).2^{-sl(x)})} \geq \frac{k+1}{\frac{-1}{s}\log(\sum_{i=1}^{2^k} p_i 2^{-sl_i} + p(x)p2^{-s(l(x)+1)} + p(x)q2^{-s(l(x)+1)})}$$

where p(x) and l(x) are the leaf probability and depth of the leaf split to form T_{2^k+2} from T_{2^k+1}.

Or $\sum_{i=1}^{2^k} p_i 2^{-sl_i} + p(x)2^{-sl(x)} \geq \sum_{i=1}^{2^k} p_i 2^{-sl_i} + p(x)p2^{-s(l(x)+1)} + p(x)q2^{-s(l(x)+1)}$

$1 \geq 2^{-sl(x)}$ which is true.

Remark: By this theorem, we see that it suffices to prove (3) for m=2^k.

b)Given a generalized Tunstall tree with 2^k leaves, consider the tree with 2^{k+1} leaves obtained by splitting each leaf into two. Then since $r_{2^{k+1}}$ is the min. over all trees, we have

$$r_{2^{k+1}} \leq \frac{k+1}{\frac{-1}{s}\log(\sum_{i=1}^{2^k} p_i 2^{-s(l_i+1)})}$$

Where p_i's and l_i's are the probabilities and depths of a generalized Tunstall tree with 2^k leaves.

So,

$$r_{2^{k+1}} \leq \frac{k+1}{\frac{-1}{s}[\log(\sum_{i=1}^{2^k} p_i 2^{-sl_i}) + (-s)]} = \frac{k+1}{\frac{-1}{s}\log(\sum_{i=1}^{2^k} p_i 2^{-sl_i}) + 1}$$

Let r_{2^k} =x

k=$L(T_{2^k})$ x

$$r_{2^k+1} \leq \frac{L(T_{2^k})x+1}{L(T_{2^k})+1} = x + \frac{1-x}{L(T_{2^k})+1}$$

c) We know $r_{2^k+1} \leq \frac{L(T_{2^k})x+1}{L(T_{2^k})+1}$ where x=r_{2^k} by Lemma 10 b).

If $x \leq 1$ implies $r_{2^{k+1}} \leq 1$, hence proved by induction as the result is true for k=1

Section 2: Mutual Information

Till now, we saw the generalization of Tunstall codes using Rényi's entropy. Now, we will see the mutual information definitions and extend them to definitions of capacity of the channel and rate distortion function.

Shannon's Version

Information is measured as reduction in uncertainty.

$$I(X,Y) = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X,Y)$$

It can also be seen that mutual information is equal to the Kullback-Leibler distance between p(x,y) and p(x).p(y)

$$I(X,Y) = D(p(x,y) \,||\, p(x)p(y))$$

Capacity of the channel is defined as:

$$C = \max_{p(x)} I(X,Y)$$

The rate distortion function R(D) is the infimum of rates R such that (R,D) is in the rate distortion region of the source for a given distortion D. A rate distortion code (R,D) is said to be achievable if there exists a sequence of (2^{nR},n) rate distortion codes with expected value of distortion between the original n-tuple sequence and the n-tuple sequence generated at output from nR bit sequence as n goes to infinity is less than or equal to D. Rate distortion function is the minimum of I(X,Y) over all p(y|x) such that expected distortion $\leq$ D.

Generalization 1: (through Rényi's entropy)

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\sum_x p^\alpha(x) \text{ [8]}$$

$$H_\alpha(X,Y) = \frac{1}{1-\alpha}\log\sum_{x,y} p^\alpha(x,y)$$

Define $2^{(1-\alpha)H_\alpha(Y|X)}$ α th order expectation of $2^{(1-\alpha)H_\alpha(Y|x)}$

Or in other words,

$$2^{(1-\alpha)H_\alpha(Y|X)} \triangleq \frac{\sum_x p^\alpha(x)\ .\ 2^{(1-\alpha)H_\alpha(Y|x)}}{\sum_x p^\alpha(x)} \text{ for } \alpha \text{ not equal to 1}$$

Or $H_\alpha(Y \mid X) \triangleq \frac{1}{1-\alpha}\log[\frac{\sum_x p^\alpha(x)\ .\ 2^{(1-\alpha)H_\alpha(Y|x)}}{\sum_x p^\alpha(x)}]$ ---------(1)

For α=1, all the definitions are as in the Shannon's case.

$$H_\alpha(X,Y) = \frac{1}{1-\alpha}\log\sum_x p^\alpha(x)\sum_y p^\alpha(y \mid x) \text{ -------(2)}$$

$$H_\alpha(Y \mid x) = \frac{1}{1-\alpha}\log\sum_y p^\alpha(y \mid x)$$

$$\sum_y p^\alpha(y \mid x) = 2^{(1-\alpha)H_\alpha(Y|x)} \text{ -----(3)}$$

Put (3) in (2), we get

$$H_\alpha(X,Y) = \frac{1}{1-\alpha}\log\sum_x p^\alpha(x)2^{(1-\alpha)H_\alpha(Y|x)}$$

=> $H_\alpha(X,Y) = \frac{1}{1-\alpha}\log\sum_x p^\alpha(x)\ .\ 2^{(1-\alpha)H_\alpha(Y|X)}$ by definition (1)

=> $H_\alpha(X,Y) = H_\alpha(X) + H_\alpha(Y \mid X)$ (using the definition as in eqn.(1))

Define generalized mutual information

$I_\alpha(X,Y) \triangleq \max(H_\alpha(X) - H_\alpha(X \mid Y), 0)$

Or $I_\alpha(X,Y) = \max(H_\alpha(X) + H_\alpha(Y) - H_\alpha(X,Y), 0)$

Define generalized capacity of the channel

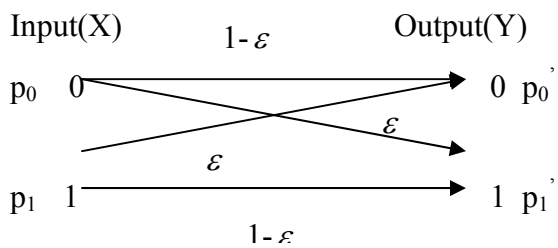
$C_\alpha \triangleq \max_{p(x)} I_\alpha(X,Y)$

Theorem 1: Chain rule for entropy

$H_\alpha(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H_\alpha(X_i \mid X_{i-1}, ..., X_1)$

Proof is obvious using repeated application of the two variable expansion rule for entropies.

Capacity of a Binary Symmetric Channel



$$H_\alpha(X,Y) = \frac{1}{1-\alpha}\log\sum_{x,y} p^\alpha(x,y)$$

$$H_\alpha(X,Y) = \frac{1}{1-\alpha}\log[\,p_0{}^\alpha(1-\varepsilon)^\alpha + p_0{}^\alpha\varepsilon^\alpha + p_1{}^\alpha(1-\varepsilon)^\alpha + p_1{}^\alpha\varepsilon^\alpha\,]$$

$$H_\alpha(X,Y) = H_\alpha(p_0) + H_\alpha(\varepsilon)$$

$$I_\alpha(X,Y) = H_\alpha(X) + H_\alpha(Y) - H_\alpha(X,Y)$$

$$I_\alpha(X,Y) = H_\alpha(p_0) + H_\alpha(p_0{}') - [H_\alpha(p_0) + H_\alpha(\varepsilon)]$$

$$I_\alpha(X,Y) = H_\alpha(p_0{}') - H_\alpha(\varepsilon)$$

$$C_\alpha = \max_{p(x)} I_\alpha(X,Y)$$

$$C_\alpha = \max_{p(x)} H_\alpha(p_0{}') - H_\alpha(\varepsilon) = 1 - H_\alpha(\varepsilon)$$

This maximum is achieved when $p_0{}'$=.5 and this happens when p_0=0.5

So, $C_\alpha = 1 - H_\alpha(\varepsilon)$

Now, we try to generalize the Rate Distortion function. Assume that we have a source that produces a sequence X^n(iid ~ p(x)). The encoder describes this source by an index $f_n(X^n) \in \{1,2,\ldots.2nR\}$. The decoder represents X^n by an estimate Y^n.

Define generalized rate distortion function as the minimum of $I_\alpha(X,Y)$ over all p(y|x) such that expected distortion $\leq$ D, over the region in which $H_\alpha(X,Y) \leq H_\alpha(X) + H_\alpha(Y)$.
First of all let us see that the region is non-empty for each α. For α=0 and α=1, we have the above inequality valid over all distributions. For other alpha, consider binary symmetric channel as above.

$I_\alpha(X,Y) = H_\alpha(p_0^{'}) - H_\alpha(\varepsilon)$

$H_\alpha(X) + H_\alpha(Y) - H_\alpha(X,Y) = H_\alpha(p_0^{'}) - H_\alpha(\varepsilon)$

Choose $p_0^{'}$=0.5(1/m in general case), and we get $H_\alpha(X,Y) \leq H_\alpha(X) + H_\alpha(Y)$

So, the region of minimization is non-empty. In this region, we have interesting result that conditioning reduces entropy.

Theorem 1: Conditioning reduces entropy over the region in which $H_\alpha(X,Y) \leq H_\alpha(X) + H_\alpha(Y)$

Proof: $H_\alpha(X,Y) \leq H_\alpha(X) + H_\alpha(Y)$

$H_\alpha(X,Y) = H_\alpha(X) + H_\alpha(Y \mid X) \leq H_\alpha(X) + H_\alpha(Y)$

So, $H_\alpha(Y \mid X) \leq H_\alpha(Y)$

Hence proved.

Rate Distortion function for Binary Source

Consider a binary source with distortion measure defined as d=1 when x is not equal to its estimate y, 0 otherwise.

$I_\alpha(X,Y) = H_\alpha(X) - H_\alpha(X \mid Y)$

$= H_\alpha(p) - H_\alpha(X \oplus Y \mid Y)$

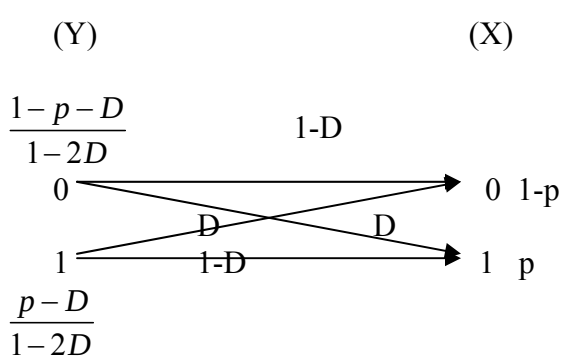
$R_\alpha(D) \geq H_\alpha(p) - H_\alpha(X \oplus Y)$

$R_\alpha(D) \geq H_\alpha(p) - H_\alpha(D)$

(D$\leq$1/2)

To find a case in which $R_\alpha(D) = H_\alpha(p) - H_\alpha(D)$, then we are done

If D$\leq$p$\leq$1/2, take a simple example



$$H_\alpha(X,Y) = \frac{1}{1-\alpha} \log \sum_{x,y} p^\alpha(x,y)$$

$$H_\alpha(X,Y)=\frac{1}{1-\alpha}\log[(\frac{1-p-D}{1-2D})^\alpha(1-D)^\alpha+(\frac{1-p-D}{1-2D})^\alpha D^\alpha+(\frac{p-D}{1-2D})^\alpha(1-D)^\alpha+(\frac{p-D}{1-2D})^\alpha D^\alpha]$$

$$H_\alpha(X,Y)=H_\alpha(Y)+H_\alpha(D)$$

So,

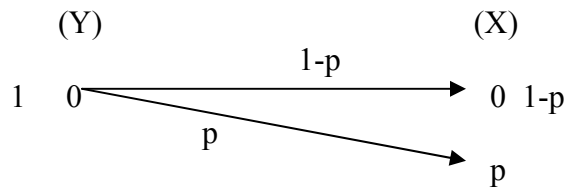
$$I_\alpha(X,Y)=H_\alpha(X)+H_\alpha(Y)-H_\alpha(X,Y)$$

$$I_\alpha(X,Y)=H_\alpha(p)+H_\alpha(Y)-[H_\alpha(Y)+H_\alpha(D)]$$

$$I_\alpha(X,Y)=H_\alpha(p)-H_\alpha(D)$$

If $D\geq p$. we can get R(D)=0 by putting Y=0 with probability 1.

$$I_\alpha(X,Y)=0$$



$$H_\alpha(X,Y)=\frac{1}{1-\alpha}\log[p^\alpha+(1-p)^\alpha]=H_\alpha(p)$$

$$H_\alpha(Y)=0$$

$$I_\alpha(X,Y)=H_\alpha(X)+H_\alpha(Y)-H_\alpha(X,Y)=0$$

So, $R_\alpha(D)=H_\alpha(p)-H_\alpha(D)$ $\quad 0\leq D\leq \min(p,1-p)$

$=0$ $\quad$ otherwise

Generalized Fano's Inequality

Suppose we wish to estimate a random variable X with a distribution p(x). We observe a random variable Y, which is related to X by the conditional distribution p(y|x). From Y, we calculate a function g(Y) =Z, which is an estimate of X. We wish to bound the probability that $Z\neq X$. We observe that $X\rightarrow Y\rightarrow Z$ forms a Markov chain. Define the probability of error

$P_e=pr(Z\neq X)$

Define E=1 if Z=X , and 0 otherwise

In the region in which $H_\alpha(E,Y)\leq H_\alpha(E)+H_\alpha(Y)$, we have

$$H_\alpha(X|Y) \leq H_\alpha(P_e)+\frac{1}{1-\alpha}\log[\frac{(1-p_e)^\alpha+p_e^{\ \alpha}*(m-1)^{(1-\alpha)}}{(1-p_e)^\alpha+p_e^{\ \alpha}}]$$

Proof:

$$H_\alpha(E,X|Y) = H_\alpha(X|Y)+H_\alpha(E|X,Y)$$
$$= H_\alpha(E|Y)+H_\alpha(X|E,Y)$$

$H_\alpha(E|X,Y)=0$ as $H_\alpha(E,X,Y)=H_\alpha(X,Y)$

$H_\alpha(E|Y)\leq H_\alpha(E)$------(imposed condition)

$H_\alpha(X|Y)\leq H_\alpha(E)+H_\alpha(X|E,Y)=H_\alpha(P_e)+H_\alpha(X|E,Y)$

$$2^{(1-\alpha)H_\alpha(X|E,Y)}=\frac{p^\alpha(E=0)*2^{(1-\alpha)H_\alpha(X|Y,E=0)}+p^\alpha(E=1)*2^{(1-\alpha)H_\alpha(X|Y,E=1)}}{\sum_e p^\alpha(e)}$$

$$= \frac{(1-p_e)^{\alpha} * 2^{(1-\alpha)0} + p_e^{\alpha} * 2^{(1-\alpha)H_\alpha(X|Y,E=1)}}{(1-p_e)^{\alpha} + p_e^{\alpha}}$$

$$\leq \frac{(1-p_e)^{\alpha} + p_e^{\alpha} * (m-1)^{(1-\alpha)}}{(1-p_e)^{\alpha} + p_e^{\alpha}} \qquad (0<\alpha<1)\text{(Inequallity changes for } \alpha=1)$$

$$H_\alpha(X \mid E,Y) \leq \frac{1}{1-\alpha}\log[\frac{(1-p_e)^{\alpha} + p_e^{\alpha} * (m-1)^{(1-\alpha)}}{(1-p_e)^{\alpha} + p_e^{\alpha}}] \quad \text{for all } \alpha$$

$$H_\alpha(\mathrm{X|Y}) \leq H_\alpha(\mathrm{P_e}) + H_\alpha(\mathrm{X|E,Y})$$

$$H_\alpha(\mathrm{X|Y}) \leq H_\alpha(\mathrm{P_e}) + \frac{1}{1-\alpha}\log[\frac{(1-p_e)^{\alpha} + p_e^{\alpha} * (m-1)^{(1-\alpha)}}{(1-p_e)^{\alpha} + p_e^{\alpha}}]$$

Conjecture 1: $I_\alpha(X,Y)$ is a convex function of p(y|x) for a fixed p(x)
Proof: This has been tested numerically.

Conjecture 2: Converse of Rate Distortion theorem can be proved easily in the region in which $H_\alpha(\mathrm{X,Y}) \leq H_\alpha(\mathrm{X}) + H_\alpha(\mathrm{Y})$ and using Conjecture 1.

Generalization 2: through Rényi's divergence)

Define generalized mutual information between X and Y as
$I_\alpha(X,Y) = D_\alpha(p(x,y) \,\|\, p(x)p(y))$

$$= \frac{1}{\alpha-1}\log\sum_{x,y} p(x)p^{\alpha}(y \mid x)p^{1-\alpha}(y)$$

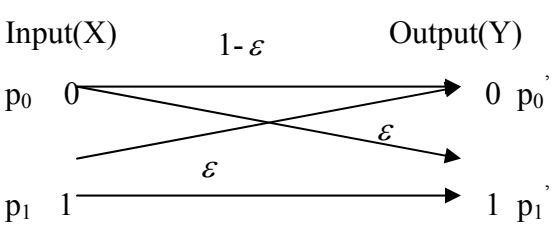
Define $K_\alpha(X,Y) = \min_{p(z)} D_\alpha(p(x,y) \,\|\, p(x)p(z))$

(Where Z is any distribution on output. This is because Y will minimize the above expression in the case of Shannon's entropy. So, this is a possible generalization).

$$= \frac{\alpha}{\alpha-1}\log\sum_{y}(\sum_{x} p(x)p^{\alpha}(y \mid x))^{1/\alpha}$$

Define the generalized capacity as:
$C_\alpha(X,Y) = \max_{p(x)} K_\alpha(X,Y)$ [8]

Capacity of Binary Symmetric channel



$$K_\alpha(X,Y) = \frac{\alpha}{\alpha-1}\log\sum_{y}(\sum_{x} p(x)p^{\alpha}(y \mid x))^{1/\alpha}$$

$$K_\alpha(X,Y) = \frac{\alpha}{\alpha-1}\log[(p_0(1-\varepsilon)^{\alpha} + (1-p_0)\varepsilon^{\alpha})^{1/\alpha} + ((1-p_0)(1-\varepsilon)^{\alpha} + p_0\varepsilon^{\alpha})^{1/\alpha}]$$

To maximize, minimize the term inside the logarithm for α <1and maximize for α >1
So,

$$\frac{d}{dp_0}[(p_0(1-\varepsilon)^\alpha+(1-p_0)\varepsilon^\alpha)^{1/\alpha}+((1-p_0)(1-\varepsilon)^\alpha+p_0\varepsilon^\alpha)^{1/\alpha}]=0$$

It can be easily verified that at p_0=1/2, above term=0, and K_α is maximum. So,

$$C_\alpha(X,Y)=\max_{p(x)} K_\alpha(X,Y)=\frac{\alpha}{\alpha-1}\log[((1-\varepsilon)^\alpha+\varepsilon^\alpha)^{1/\alpha}*\frac{2}{2^{1/\alpha}}]$$

$$=\frac{\alpha}{\alpha-1}\log[((1-\varepsilon)^\alpha+\varepsilon^\alpha)^{1/\alpha}]+\frac{\alpha}{\alpha-1}[1-\frac{1}{\alpha}]=1-H_\alpha(\varepsilon)$$

Theorem: $I_\alpha(X,Y)\geq 0$ for α >0

Proof: For 0<α <1, we need to show

$$\sum_{x,y} p(x)p^\alpha(y\mid x)p^{1-\alpha}(y)=\sum_{x,y}[p(x)p(y)]^{1-\alpha}\,p^\alpha(x,y)\leq 1$$

Or, $E[(\frac{p(x,y)}{p(x)p(y)})^\alpha]\leq 1$. Let Z be a r.v. which takes value $\frac{p(x,y)}{p(x)p(y)}$ with probability $p(x)p(y)$. We need to show $E[Z^\alpha]\leq 1$ which is true as $E[Z^\alpha]\leq E[Z]^\alpha\leq 1$.

For $\alpha\geq 1$, the inequalities reverse and the result follows.

Theorem: $I_\alpha(X,Y)$ is convex w.r.t. p(y|x) for fixed p(x) for 0<α <1

Proof: $I_\alpha(X,Y)=\frac{1}{\alpha-1}\log\sum_{x,y} p(x)p^\alpha(y\mid x)p^{1-\alpha}(y)$

Choose 2 distributions $p_1(y\mid x)$ and $p_2(y\mid x)$

Consider another distribution

$$p_\lambda(y\mid x)=\lambda p_1(y\mid x)+(1-\lambda)p_2(y\mid x)$$

$$p_1(x)=p_2(x)=p_\lambda(x)=p(x)\ \text{(say)}$$

$$p_\lambda(y)=\lambda p_1(y)+(1-\lambda)p_2(y)$$

$$\sum_{x,y} p_\lambda(x)p_\lambda{}^\alpha(y\mid x)p_\lambda{}^{1-\alpha}(y)$$

$$=\sum_{x,y} p(x)(\lambda p_1(y\mid x)+(1-\lambda)p_2(y\mid x))^\alpha(\lambda p_1(y)+(1-\lambda)p_2(y))^{1-\alpha}$$

$$\geq\sum_{x,y} p(x)(\lambda p_1{}^\alpha(y\mid x)p_1{}^{1-\alpha}(y)+(1-\lambda)p_2{}^\alpha(y\mid x)p_2{}^{1-\alpha}(y))\qquad\text{(using Holder's Inequality)}$$

$$=\lambda\sum_{x,y} p_1(x)p_1{}^\alpha(y\mid x)p_1{}^{1-\alpha}(y)+(1-\lambda)\sum_{x,y} p_2(x)p_2{}^\alpha(y\mid x)p_2{}^{1-\alpha}(y)$$

So, term inside log is concave, and so $\log\sum_{x,y} p(x)p^\alpha(y\mid x)p^{1-\alpha}(y)$ is concave

And so $\frac{1}{\alpha-1}\log\sum_{x,y} p(x)p^\alpha(y\mid x)p^{1-\alpha}(y)$ is convex.

Definition: Assume that we have a binary source that produces $X_1,X_2,\ldots,X_n$ i.i.d. ~p(x). The encoder describes the source sequence X^n by an index $f_n(X^n)\in(1,2,\ldots 2^{nR})$. The decoder represents X^n by an estimate $\widehat{X}^n$. The information rate distortion function $R_\alpha{}^{(I)}(D)$ for a source X with distortion measure d(x, $\hat{x}$) is defined as

$$R_\alpha{}^{(I)}(D)\triangleq\min_{p(\hat{x}|x)} I_\alpha(X,\widehat{X}).$$

Theorem: The rate distortion function for an i.i.d. source X with the distortion function d(x, $\hat{x}$) is equal to the maximum of the associated information rate distortion function over $0 \le \alpha < 1$. Thus, $R(D) = \max_{0 \le \alpha < 1} R_\alpha^{(I)}(D) = \max_{0 \le \alpha < 1} \min_{p(\hat{x}|x)} I_\alpha(X, \widehat{X})$.

Proof: $I_\alpha(X, \widehat{X})$ is maximum when $\alpha \to 1$ ($I_\alpha(X, \widehat{X})$ increases with α), and in that case, the result is same as the original rate distortion theorem.

Section 3: Extension of Tunstall Algorithm for countable alphabet.

This extension is on similar lines as version 1 of modified LZ codes for countable alphabet in [16].
Fix a positive integer m and let let the input sequence be $x=(x_1,x_2,\ldots)$.
Define $y=(y_1,y_2,\ldots)$ where
$y_i=x_i$ if $xi \leq m$;
$=m+1$ otherwise.
Apply Tunstall Algorithm on the transformed sequence $y_1,\ldots,y_n$, using the finite alphabet of size m+1. Mark the locations where x_i differs from y_i, i=1,…,n. Let $i_1,\ldots i_{p(n)}$ be the p(n) such instants. Encode the sequence $x_{i_1},\ldots,x_{i_p}$ by encoding each symbol separately by a prefix code with the length function L.
Assume that we find c codewords in n length on which Tunstall coding is being applied.
So, n=c. $\sum_{1}^{k} p_i l_i$ length is encoded by Tunstall coding to c. log(k) binary digits.

So, the total number of binary digits used in encoding = c. log(k) + $\sum_{j=1}^{p(n)} L(x_{ij})$.

The encoded version of the sequence $x_{i_1},\ldots,x_{i_p}$ follows the encoded version of $y_1,\ldots,y_n$, $\lceil \log(n) \rceil$ bits can be used at the outset. While decoding, $y_1,\ldots,y_n$ is decoded first. In this process, the sequence $i_1,\ldots i_{p(n)}$ is recovered. Next the encoded version of $x_{i_1},\ldots,x_{i_p}$ is decoded and the sequence is embedded in $y_1,\ldots,y_n$ to recover the original sequence $x_1,\ldots,x_n$.
So, the compression ratio=

$$\frac{\text{c. log(k)} + \sum_{j=1}^{p(n)} L(x_{ij}) + \lceil \log(n) \rceil}{n} = \frac{\log\ k}{\sum_{1}^{k} p_i l_i} + \frac{\sum_{j=1}^{p(n)} L(x_{ij}) + \lceil \log(n) \rceil}{n}$$

$$\leq (H_m + \in) + \frac{\sum_{j=1}^{p(n)} L(x_{ij}) + \lceil \log(n) \rceil}{n} \leq H + 2 \in$$ (Last step is similar to that in [16] which uses ergodic theorem).

Conclusion and Future work

The Tunstall Theorem has been generalized using Rényi's entropy and the convergence of generalized compression ratio to generalized entropy (Rényi's entropy) has been proved. An algorithm to minimize the redundancy is also given in Section 1. Some possible definitions of rate distortion function were explored in Section 2. Then in Section 3, an extension of Tunstall Algorithm to countable alphabet was proved to be working.

Further work in Section 1 can be for extension of Generalized Tunstall Theorem to Markov setup in the light of work done in [17] and [18]. An operation meaning to Rate Distortion function in the case of second generalization in Section 2 is yet to be found out giving some form of generalized rate distortion theorem.

Bibliography

1. Robert Ash, " Information Theory"
2. Thomas M. Cover, Joy A. Thomas , "Elements of Information Theory"
3. "Lectures On Statistical Modeling Theory" by J. Rissanen http://www.cs.tut.fi/~rissanen/papers/lectures.pdf
4. L.L. Campbell, "A coding theorem and Rényi's entropy."Informat. Contr., vol. 8, pp 423-429, 1965.
5. A. Rényi's, "On measures of entropy and information," in Proc. 4th Berkeley Symp. Math. Stat. and Probability, vol. 1, pp. 547-561,1960.
6. A. Rényi's, "On the foundations of information theory," Rev. Inst. Int. Stat. vol. 33, pp. 1-14, 1965.
7. J. Aczel and Z.Daroczy, "On Measures of Information and their characterizations" Academic Press, 1975.
8. I. Csiszar, "Generalized Cutoff Rates and Rényi's Information Measures" IEEE Trans. Inform. Theory, vol.41, no. 1, Jan. 1995.
9. Moshe Ben-Bassat, "Rényi's entropy and the probability of error", IEEE Trans. Inform. Theory, vol.34, no. 3, May 1978.
10. A.C.Blumer, "The Rényi's redundancy of generalized Huffman codes", IEEE Trans. Inform. Theory, vol.34, no. 5, Sept. 1988.
11. F. Jelinek,"On Variable-Length-to-Block Coding", IEEE Trans. Inform. Theory, vol.18, no. 6, Nov. 1972.
12. A. Rényi, "Probability Theory"
13. Inder Jeet Taneja," Generalized Information Measures and Their Applications"
14. Pompe B. and Heilfort M., "On the Concept of the Generalized Mutual Information Function and Efficient Algorithms for Calculating It" http://www2.physik.uni-greifswald.de/~pompe/
15. B. P. Tunstall, "Synthesis of noiseless compression codes," Ph.D. dissertation. Georgia Inst. Technol., Atlanta. 1968.
16. R.K. Bansal and Jay Deep Sau, "Weakly Universal LZ-extended Codes for Sources with Countable Alphabet," to be presented in ISIT 2005.
17. I. Tabus and J. Rissanen," Asymptotics of Greedy Algorithms for Variable-to-Fixed Length Coding of Markov Sources", *IEEE Trans. Inform. Theory*, vol. 48, pp. 2022–2035, July 2002.
18. S. Savari and R. Gallager, "Generalized Tunstall codes for sources with memory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 658–668, Mar. 1997.