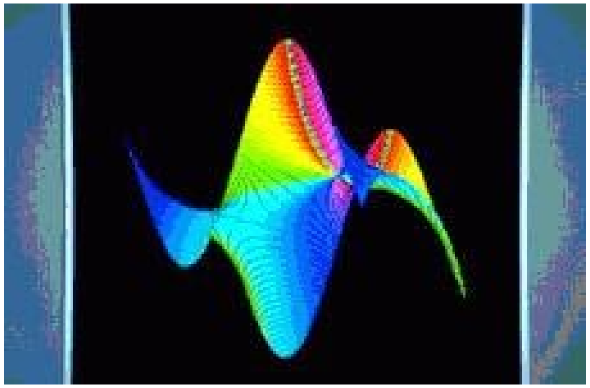
Analytic Algorithmics, Combinatorics, and Information Theory[*]

W. Szpankowski[†]
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

January 18, 2005

AofA LOGO



[*]Research supported by NSF and NIH.

[†]Joint work with M. Drmota, P. Flajolet, and P. Jacquet, C. Knessl.

Outline

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem
   (a) Known Sources (Sequences mod 1)
   (b) Universal Memoryless Sources (Tree-like gen. func.)
   (c) Universal Markov Sources (Balance matrices)
   (d) Universal Renewal Sources (Combinatorial calculus)
4. Method of Types
   (a) Markov Types (Eulerian paths)
   (b) Universal Types (Binary trees of given path)
5. Conclusions

Goals of Source Coding

The basic problem of source coding (i.e., *data compression*) is to find codes with shortest descriptions (lengths) either on *average* or for *individual sequences* when the source (i.e., statistics of the underlying probability distribution) is unknown (the so called **universal source coding**.

Goals:

- Find universal lower bound on compression ratio (bit rate).
- Construct universal source codes that achieve this lower bound up to the second order asymptotics (i.e., match redundancy which is basically a measure of the second term asymptotics).
- As pointed by Rissanen, universal coding evolved into universal modeling where the purpose is no longer restricted to just coding but rather to finding optimal models for data.

Some Definitions

Definition: A block-to-variable (BV) length code

$$C_n : \mathcal{A}^n \to \{0,1\}^*$$

is a bijective mapping from all sequences of length n over the alphabet $\mathcal{A}$ to the set $\{0,1\}^*$ of binary sequences.

For a probabilistic source model $\mathcal{S}$ and a code C_n we let:

- $P(x_1^n)$ be the probability of $x_1^n = x_1 \dots x_n$;
- $L(C_n, x_1^n)$ be the code length for x_1^n;
- Entropy $H_n(P) = -\sum_{x_1^n} P(x_1^n) \lg P(x_1^n)$.

Information-theoretic quantities are expressed in binary logarithms written $\lg := \log_2$.

Two Facts

Prefix code is such that no codeword is a prefix of another codeword.

Kraft's Inequality
A code is a prefix code iff $\ell_1, \ell_2, \ldots, \ell_m$ satisfy the inequality

$$\sum_{i=1}^{m} 2^{-\ell_i} \leq 1.$$

whee $\ell_1, \ell_2, \ldots, \ell_m$ are codeword length.

Proof. An easy exercise on trees.

Shannon First Theorem
For any prefix code the average code length $\mathbf{E}[L(C_n, X_1^n)]$ cannot be smaller than the entropy of the source $H_n(P)$, that is,

$$\mathbf{E}[L(C_n, X_1^n)] \geq H_n(P).$$

Proof: Use $\log x \leq x - 1$ for $0 < x \leq 1$ and Kraft's inequality.

By Kraft's inequality there exists at least one source sequence $\tilde{x}_1^n$ such that

$$L(\tilde{x}_1^n) \geq -\log_2 P(\tilde{x}_1^n).$$

Proof of Shannon Lower Bound

Let $K = \sum_{x_1^n} 2^{-L(x_1^n)} \leq 1$, and $L(C_n, x_1^n) := L(C_n)$. Then

$$
\begin{aligned}
\mathbf{E}[L(C_n, X_1^n)] \quad - \quad & H_n(P) = \\
= \quad & \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) L(x_1^n) + \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \log P(x_1^n) \\
= \quad & \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \log \frac{P(x_1^n)}{2^{-L(x_1^n)}/K} - \log K \\
\geq \quad & 0
\end{aligned}
$$

since $\log x \leq x - 1$ for $0 < x \leq 1$ or the divergence is nonnegative, while $K \leq 1$ by Kraft's inequality.

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem
   (a) Known Sources (Sequences mod 1)
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources
   (d) Universal Renewal Sources
4. Method of Types
5. Conclusions

Redundancy

Known Source P

The pointwise redundancy $R_n(C_n, P; x_1^n)$ and the average redundancy $\bar{R}_n(C_n, P)$ are defined as

$$\begin{aligned} R_n(C_n, P; x_1^n) &= L(C_n, x_1^n) + \lg P(x_1^n) \\ \bar{R}_n(C_n) &= \mathbf{E}[L(C_n, X_1^n)] - H_n(P) \geq 0 \end{aligned}$$

The maximal or worst case redundancy is

$$R^*(C_n, P) = \max_{x_1^n}\{R_n(C_n, P; x_1^n)\}(\geq 0).$$

The pointwise redundancy can be negative, maximal and average redundancy cannot.

The smaller the redundancy is, the better (closer to the optimal) the code is.

Redundancy for Known Sources

Huffman Code

The following optimization problem

$$\bar{R}_n(P) = \min_{C_n \in \mathcal{C}} \mathbf{E}_{x_1^n}[L(C_n, x_1^n) + \log_2 P(x_1^n)].$$

is solved by Huffman's code

Generalized Shannon Code

Drmota and W.S. (2001) proved that

$$R_n^*(P) = \min_{C_n} \max_{x_1^n}[L(C_n, x_1^n) + \lg P(x_1^n)].$$

is solved by the generalized Shannon code C_n^{GS} which is defined as

$$L(C_n^{GS}, x_1^n) = \begin{cases} \lfloor \lg 1/P(x_1^n) \rfloor & \text{if} \quad \langle P(x_1^n) \rangle \leq s_0 \\ \lceil \lg 1/P(x_1^n) \rceil & \text{if} \quad \langle P(x_1^n) \rangle > s_0 \end{cases}$$

where s_0 is a constant, and $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x;

Two Modes of $\bar{R}_n^H(P)$ Behavior

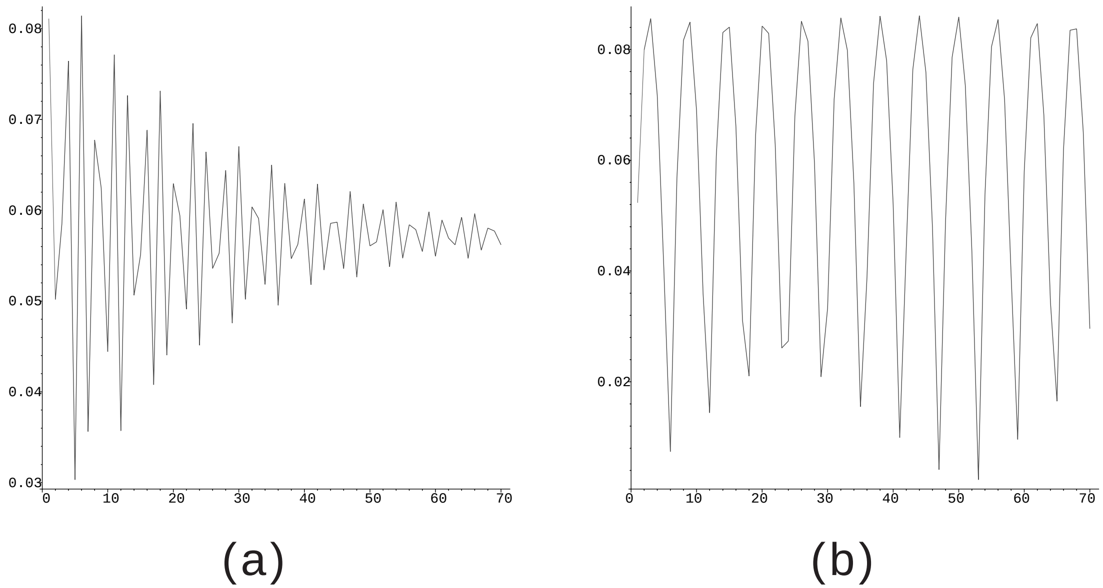


Figure 1: The average redundancy of Huffman codes versus block size n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

WHY?

Why Two Modes: Shannon Code

Consider the Shannon code that assigns the length

$$L(C_n^S, x_1^n) = \lceil -\lg P(x_1^n) \rceil$$

to the source sequence x_1^n. Observe that

$$P(x_1^n) = p^k (1-p)^{n-k}$$

where p is **known** probability of generating 0 and k is the number of 0s.

The Shannon code redundancy is

$$\begin{aligned}
\bar{R}_n^S &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \Big(\lceil -\log_2(p^k (1-p)^{n-k}) \rceil \\
&+ \log_2(p^k (1-p)^{n-k}) \Big) \\
&= 1 - \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \langle \alpha k + \beta n \rangle
\end{aligned}$$

where $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x, and

$$\alpha = \log_2\left(\frac{1-p}{p}\right), \quad \beta = \log_2\left(\frac{1}{1-p}\right).$$

Two Recent Results

To analyze redundancy for known sources one needs to understand asymptotic behavior of the following sum

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle x_k + y \rangle)$$

for fixed p and some Riemann integrable function $f : [0,1] \to \mathbf{R}$.

For the **Huffman code** the following is known [W.S. (2000)]

$$\bar{R}_n^H = \begin{cases} \frac{3}{2} - \frac{1}{\ln 2} + o(1) \approx 0.057304 & \alpha \quad \text{irrational} \\ \frac{3}{2} - \frac{1}{M}\left(\langle \beta M n \rangle - \frac{1}{2}\right) - \frac{1}{M(1-2^{-1/M})} 2^{-\langle n\beta M \rangle / M} + O(\rho^n) & \alpha = \frac{N}{M} \end{cases}$$

N, M are such integers that $\gcd(N, M) = 1$ and $\rho < 1$.

For the **generalized Shannon code** Drmota and W.S. (2004) proved

$$R_n^*(P_p) = -\frac{\log \log 2}{\log 2} + o(1) = 0.5287\ldots + o(1).$$

when α is irrational.

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem
   (a) Known Sources
   (b) Unknown Sources (memoryless, Markov, renewal)
4. Method of Types
5. Conclusions

Minimax Redundancy

Unknown Source P

In practice, one can only hope to have some knowledge about a family of sources $\mathcal{S}$ that generates real data.

Following Davisson we define the average minimax redundancy $\bar{R}_n(\mathcal{S})$ and the worst case (maximal) minimax redundancy $R_n^*(\mathcal{S})$ for a family of sources $\mathcal{S}$ as

$$\begin{aligned}
\bar{R}_n(\mathcal{S}) &= \min_{C_n} \sup_{P \in \mathcal{S}} \mathbf{E}[L(C_n, x_1^n) + \lg P(x_1^n)] \\
R_n^*(\mathcal{S}) &= \min_{C_n} \sup_{P \in \mathcal{S}} \max_{x_1^n} [L(C_n, x_1^n) + \lg P(x_1^n)].
\end{aligned}$$

In the minimax scenario we look for the best code for the the worst source.

Source Coding Goal:
Find data compression algorithms that **match optimal redundancy rates** either on average or for individual sequences.

Maximal Minimax Redundancy

We consider the following **classes of sources** $\mathcal{S}$:

- **Memoryless sources** $\mathcal{M}_0$ over an m-ary (finite) alphabet, that is,

$$P(x_1^n) = p_1^{k_1} \cdots p_m^{k_m}$$

  with $k_1 + \cdots + k_m = n$, where p_i are unknown!

- **Markov sources** $\mathcal{M}_r$ over a binary alphabet of order r. Observe that for $r = 1$

$$P(x_1^n) = p_{x_1} p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}},$$

  where k_{ij} is such that $(x_k, x_{k+1}) = (i, j) \in \{0, 1\}^2$ and

$$k_{00} + k_{01} + k_{10} + k_{11} = n - 1,$$

  and that $k_{01} = k_{10}$ if $x_1 = x_n$ and $k_{01} = k_{10} \pm 1$ if $x_1 \neq x_n$.

- **Renewal Sources** $\mathcal{R}_0$ where an 1 is introduced after a run of 0s distributed according to some distribution.

Improved Shtarkov Bounds

For the maximal minimax redundancy define

$$Q^*(x_1^n) := \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{\sum_{y_1^n\in\mathcal{A}^n}\sup_{P\in\mathcal{S}} P(y_1^n)}.$$

the maximum likelihood distribution. Observe that

$$\begin{aligned}
R_n^*(\mathcal{S}) &= \min_{C_n\in\mathcal{C}} \sup_{P\in\mathcal{S}} \max_{x_1^n} (L(C_n, x_1^n) + \lg P(x_1^n)) \\
&= \min_{C_n\in\mathcal{C}} \max_{x_1^n} \left(L(C_n, x_1^n) + \sup_{P\in\mathcal{S}} \lg P(x_1^n) \right) \\
&= \min_{C_n\in\mathcal{C}} \max_{x_1^n} [L(C_n, x_1^n) + \lg Q^*(x_1^n) + \lg \sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)] \\
&= R_n^{GS}(Q^*) + \lg \sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)
\end{aligned}$$

where $R_n^{GS}(Q^*)$ is the maximal redundancy of a generalized Shannon code built for the (known) distribution Q^*. We also write

$$D_n(\mathcal{S}) = \lg \left(\sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(x_1^n) \right) := \lg d_n(\mathcal{S}).$$

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem
   (a) Known Sources
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources
   (d) Universal Renewal Sources
4. Method of Types
5. Conclusions

Maximal Minimax for Memoryless Sources

We first consider the maximal minimax redundancy $R_n^*(\mathcal{M}_0)$ for a class of memoryless sources over a finite m-ary alphabet. Observe that

$$\begin{aligned}
d_n(\mathcal{M}_0) &= \sum_{x_1^n} \sup_{p_1,\ldots,p_m} p_1^{k_1}\cdots p_m^{k_m} \\
&= \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \sup_{p_1,\ldots,p_m} p_1^{k_1}\cdots p_m^{k_m} \\
&= \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \left(\frac{k_1}{n}\right)^{k_1}\cdots\left(\frac{k_m}{n}\right)^{k_m}.
\end{aligned}$$

The **summation set** is

$$I(k_1,\ldots,k_m) = \{(k_1,\ldots,k_m) :\ k_1+\cdots+k_m = n\}.$$

The number $N_{\mathbf{k}}$ of **types** $\mathbf{k} = (k_1,\ldots,k_m)$ is

$$N_{\mathbf{k}} = \binom{n}{k_1,\ldots,k_m}$$

The (unnormalized) **likelihood distribution** is

$$\sup_{p_1,\ldots,p_m} p_1^{k_1}\cdots p_m^{k_m} = \left(\frac{k_1}{n}\right)^{k_1}\cdots\left(\frac{k_m}{n}\right)^{k_m}$$

Generating Function for $d_n(\mathcal{M}_0)$

We write

$$d_n(\mathcal{M}_0) = \frac{n!}{n^n} \sum_{k_1+\cdots+k_m=n} \frac{k_1^{k_1}}{k_1!} \cdots \frac{k_m^{k_m}}{k_m!}$$

Let us introduce a tree-generating function

$$B(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} z^k = \frac{1}{1-T(z)},$$

where $T(z)$ satisfies $T(z) = ze^{T(z)}$ (= $-W(-z)$, **Lambert's** W-function) and also

$$T(z) = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} z^k$$

enumerates all rooted labeled trees. Let now

$$D_m(z) = \sum_{n=0}^{\infty} \frac{n^n}{n!} d_n(\mathcal{M}_0).$$

Then by the convolution formula

$$D_m(z) = [B(z)]^m.$$

Asymptotics

The function $B(z)$ has an algebraic singularity at $z = e^{-1}$ (it becomes a multi-valued function) and one finds

$$B(z) = \frac{1}{\sqrt{2(1-ez)}} + \frac{1}{3} + O(\sqrt{(1-ez)}.$$

The singularity analysis yields (cf. Clarke & Barron, 1990, W.S., 1998)

$$\begin{aligned} d_n(\mathcal{M}_0) \quad &= \quad \frac{m-1}{2} \log\left(\frac{n}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + \frac{\Gamma(\frac{m}{2})m}{3\Gamma(\frac{m}{2}-\frac{1}{2})} \cdot \frac{\sqrt{2}}{\sqrt{n}} \\ &+ \quad \left(\frac{3+m(m-2)(2m+1)}{36} - \frac{\Gamma^2(\frac{m}{2})m^2}{9\Gamma^2(\frac{m}{2}-\frac{1}{2})}\right) \cdot \frac{1}{n} + \cdots \end{aligned}$$

To complete the analysis, we need $\bar{R}_n^{GS}(Q^*)$. Drmota & W.S., 2001 proved

$$R_n^{GS}(Q^*) = -\frac{\ln \frac{1}{m-1} \ln m}{\ln m} + o(1),$$

In general, the term $o(1)$ can not be improved. Thus

$$R_n^*(\mathcal{M}_0) = \frac{m-1}{2} \log\left(\frac{n}{2}\right) - \frac{\ln \frac{1}{m-1} \ln m}{\ln m} + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + o(1).$$

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem
   (a) Known Source
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources
   (d) Universal Renewal Sources
4. Method of Types
5. Conclusions

Maximal Minimax for Markov Sources

(i) $\mathcal{M}_1$ is a Markov source of order $r = 1$,
(ii) the transition matrix $P = \{p_{ij}\}_{i,j=1}^{m}$
(iii) **circular** sequences (the first symbols follows the last).

$$d_n(\mathcal{M}_1) = \sum_{x_1^n} \sup_P p_{11}^{k_{11}} \cdots p_{mm}^{k_{mm}}$$

$$= \sum_{\mathbf{k} \in \mathcal{F}_n} \mathbf{M}_{\mathbf{k}} \left(\frac{k_{11}}{k_1}\right)^{k_{11}} \cdots \left(\frac{k_{mm}}{k_m}\right)^{k_{mm}},$$

k_{ij} is the number of pairs $ij \in \mathcal{A}^2$ in x_1^n, $k_i = \sum_{j=1}^{m} k_{ij}$,
$\mathbf{k} = \{k_{ij}\}_{i,j=1}^{m}$ is an integer matrix such that

$$\mathcal{F}_n : \sum_{i,j=1}^{m} k_{ij} = n, \quad \text{and} \quad \sum_{j=1}^{m} k_{ij} = \sum_{j=0}^{m-1} k_{ji},$$

Matrix $\mathbf{k}$ satisfying the above conditions is called the frequency matrix or **Markov type**.

$M_{\mathbf{k}}$ represents the numbers of strings x_1^n of **type** $\mathbf{k}$.

We come back to Markov types soon.

Main Technical Tool

Let $g_{\mathbf{k}}$ be a sequence of scalars indexed by matrices $\mathbf{k}$ and

$$g(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$$

be its regular generating function, and

$$\mathcal{F} g(\mathbf{z}) = \sum_{\mathbf{k} \in \mathcal{F}} g_{\mathbf{k}} z^{\mathbf{k}} = \sum_{n \geq 0} \sum_{\mathbf{k} \in \mathcal{F}_n} g_{\mathbf{k}} z^{\mathbf{k}}$$

the $\mathcal{F}$-generating function of $g_{\mathbf{k}}$ for which $\mathbf{k} \in \mathcal{F}$.

Lemma 1. *Let* $g(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$. *Then*

$$\mathcal{F} g(\mathbf{z}) := \sum_{n \geq 0} \sum_{\mathbf{k} \in \mathcal{F}_n} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \left(\frac{1}{2\jmath\pi}\right)^m \oint \frac{dx_1}{x_1} \cdots \oint \frac{dx_m}{x_m} g([z_{ij} \frac{x_j}{x_i}])$$

with the ij*-th coefficient of* $[z_{ij} \frac{x_j}{x_i}]$ *is* $z_{ij} \frac{x_j}{x_i}$.

Proof. It suffices to observe

$$g([z_{ij} \frac{x_j}{x_i}]) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} \prod_{i=1}^{m} x_i^{\sum_i k_{ij} - \sum_j k_{ij}}$$

Thus $\mathcal{F} g(\mathbf{z})$ is the coefficient of $g([z_{ij} \frac{x_j}{x_i}])$ at $x_1^0 x_2^0 \cdots x_m^0$.

Main Results

Theorem 1. *Let $\mathcal{M}_1$ be a Markov source over an m-ry alphabet. Then*

$$d_n(\mathcal{M}_1) = \left(\frac{n}{2\pi}\right)^{m(m-1)/2} A_m \times \left(1 + O\left(\frac{1}{n}\right)\right)$$

with

$$A_m = \int_{\mathcal{K}(1)} m F_m(y_{ij}) \prod_i \frac{\sqrt{\sum_j y_{ij}}}{\prod_j \sqrt{y_{ij}}} d[y_{ij}]$$

where $\mathcal{K}(1) = \{y_{ij} : \ \sum_{ij} y_{ij} = 1\}$ and $F_m(\cdot)$ is a polynomial expression of degree $m-1$.

In particular, for $m = 2$ $A_2 = 16 \times$ Catalan where Catalan is Catalan's constant $\sum_i \frac{(-1)^i}{(2i+1)^2} \approx 0.915965594$.

Theorem 2. *Let $\mathcal{M}_r$ be a Markov source of order r. Then*

$$d_n(\mathcal{M}_r) = \left(\frac{n}{2\pi}\right)^{m^r(m-1)/2} A_m^r \times \left(1 + O\left(\frac{1}{n}\right)\right)$$

where A_m^r is a constant defined in a similar fashion as A_m above.

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding

3. The Redundancy Rate Problem
   (a) Known Sources
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources

   (d) Universal Renewal Sources
4. Method of Types
5. Conclusions

Renewal Sources

The **renewal process** defined as follows:

- Let $T_1, T_2 \ldots$ be a sequence of i.i.d. positive-valued random variables with distribution $Q(j) = \Pr\{T_i = j\}$.
- The process $T_0, T_0 + T_1, T_0 + T_1 + T_2, \ldots$ is called the renewal process.
- With a renewal process we associate a **binary renewal sequence** in which the positions of the 1's are at the renewal epochs (runs of zeros) $T_0, T_0 + T_1, \ldots$.
- We start with $x_0 = 1$.

Csiszár and Shields (1996) proved that $R_n(\mathcal{R}_0) = \Theta(\sqrt{n})$.

Maximal Minimax Redundancy

For a sequence

$$x_0^n = 10^{\alpha_1}10^{\alpha_2}1\cdots 10^{\alpha_n}1\underbrace{0\cdots 0}_{k^*}$$

k_m is the number of i such that $\alpha_i = m$. Then

$$P(x_1^n) = Q^{k_0}(0)Q^{k_1}(1)\cdots Q^{k_{n-1}}(n-1)\Pr\{T_1 > k^*\}.$$

It can be proved that

$$r_{n+1} - 1 \le d_n(\mathcal{R}_0) \le \sum_{m=0}^{n} r_m$$

where

$$\begin{aligned} r_n &= \sum_{k=0}^{n} r_{n,k} \\ r_{n,k} &= \sum_{\mathcal{P}(n,k)} \binom{k}{k_0 \ldots k_{n-1}} \left(\frac{k_0}{k}\right)^{k_0} \left(\frac{k_1}{k}\right)^{k_1} \cdots \left(\frac{k_{n-1}}{k}\right)^{k_{n-1}} \end{aligned}$$

where $\mathcal{P}(n,k)$ is the summation set which happens to be the partition of n into k terms, i.e.,

$$\begin{aligned} n &= k_0 + 2k_1 + \cdots + nk_{n-1}, \\ k &= k_0 + \cdots + k_{n-1}. \end{aligned}$$

Main Results

Theorem 3 (Flajolet and WS, 1998). *Consider the class of renewal processes as defined above. The quantity* r_n *attains the following asymptotics*

$$r_n = \frac{2}{\log 2}\sqrt{cn} - \frac{5}{8}\lg n + \frac{1}{2}\lg\log n + O(1)$$

where $c = \frac{\pi^2}{6} - 1 \approx 0.645$. *Moreover,*

$$R_n^*(\mathcal{R}_0) = \frac{2}{\log 2}\sqrt{cn} + O(\log n).$$

Asymptotics: Overview

Asymptotic analysis is sophisticated and follows these steps:

- first, we transform r_n into another quantity s_n that we know how to handle and (using a probabilistic technique) we know how to read back results for r_n from s_n;
- use combinatorial calculus to find the generating function of s_n, which turns out to be an infinite product of tree-functions $B(z)$ defined above;
- transform this product into a harmonic sum that can be analyzed asymptotically by the Mellin transform;
- obtain an asymptotic expansion of the generating function around $z = 1$ which is the starting point for extracting the asymptotics of the coefficients;
- finally, estimate $R_n^*(\mathcal{R}_0)$ by the saddle point method.

Asymptotics: The Main Idea

The quantity r_n is to hard to analyze due to the factor $k!/k^k$, hence we define a new quantity s_n defined as

$$\begin{cases} s_n & = & \sum_{k=0}^{n} s_{n,k} \\ s_{n,k} & = & e^{-k} \sum_{\mathcal{P}(n,k)} \frac{k^{k_0}}{k_0!} \cdots \frac{k^{k_{n-1}}}{k_{n-1}!}. \end{cases}$$

To analyze it, we introduce the random variable K_n as follows

$$\Pr\{K_n = k\} = \frac{s_{n,k}}{s_n}.$$

Stirling's formula yields

$$\begin{aligned} \frac{r_n}{s_n} &= \sum_{k=0}^{n} \frac{r_{n,k}}{s_{n,k}} \frac{s_{n,k}}{s_n} = \mathbf{E}[(K_n)! K_n^{-K_n} e^{-K_n}] \\ &= \mathbf{E}[\sqrt{2\pi K_n}] + O(\mathbf{E}[K_n^{-\frac{1}{2}}]). \end{aligned}$$

Fundamental Lemmas

Lemma 2. *Let* $\mu_n = \mathbf{E}[K_n]$ *and* $\sigma_n^2 = \mathbf{Var}(K_n)$.

$$\begin{aligned} s_n &\sim \exp\left(2\sqrt{cn} - \frac{7}{8}\log n + d + o(1)\right) \\ \mu_n &= \frac{1}{4}\sqrt{\frac{n}{c}}\log\frac{n}{c} + o(\sqrt{n}) \\ \sigma_n^2 &= O(n\log n) = o(\mu_n^2), \end{aligned}$$

where $c = \pi^2/6 - 1$, $d = -\log 2 - \frac{3}{8}\log c - \frac{3}{4}\log\pi$.

Lemma 3. *For large* n

$$\begin{aligned} \mathbf{E}[\sqrt{K_n}] &= \mu_n^{1/2}(1 + o(1)) \\ \mathbf{E}[K_n^{-\frac{1}{2}}] &= o(1). \end{aligned}$$

where $\mu_n = \mathbf{E}[K_n]$.

Thus

$$\begin{aligned} r_n &= s_n\mathbf{E}[\sqrt{2\pi K_n}](1 + o(1)) \\ &= s_n\sqrt{2\pi\mu_n}(1 + o(1)). \end{aligned}$$

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem

4. Method of Types
   (a) Markov Types (Eulerian paths)
   (b) Universal Type
5. Conclusions

Method of Types

The method of types is a powerful technique in information theory, large deviations, and analysis of algorithms. It reduces calculations of the probability of rare events to a combinatorial analysis.

Two sequences are of the same type if they have the same empirical distribution.

For memoryless sources the type is measured by the relative frequency of symbol occurrences.

For example, for a binary sequences of length n there are

$$\binom{n}{k}$$

sequences of type k (i.e., with k ones).

Markov Types

For (a binary) Markov sources

$$P(x_1^n) = p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}},$$

thus all strings x_1^n with the same matrix $\mathbf{k} = \{k_{ij}\}_{i,j \in \mathcal{A}}$ have the same empirical distribution.
These sequences belong to the same type $\mathbf{k}$.

Consider only cyclic strings in which the last symbol is followed by the first one. The frequency matrix $\mathbf{k}$ satisfies

$$\mathcal{F}_n : \sum_{i,j=1}^{m} k_{ij} = n, \quad \text{and} \quad \sum_{j=1}^{m} k_{ij} = \sum_{j=0}^{m-1} k_{ji},$$

We are interested in:

$N_{\mathbf{k}}$ – number of (cyclic) strings x_1^n belonging to the same type $\mathbf{k}$.
$N_{\mathbf{k}}^{a}$ – number of (cyclic) strings x_1^n belonging to the same type $\mathbf{k}$ and starting with a symbol a.
$N_{\mathbf{k}}^{ab}$ – number of (cyclic) strings x_1^n belonging to the same type $\mathbf{k}$, starting with a symbol a and ending with b.

Eulerian Cycles

Define a directed multigraph $\mathcal{G}_m$ over $m = |\mathcal{A}|$ vertices:
– labeled by symbols from the alphabet $\mathcal{A}$;
– edge multiplicity between ith and jth vertices is k_{ij}.

The number of Eulerian paths starting from vertex $a \in \mathcal{A}$ in a such multigraph is equal to $N_{\mathbf{k}}^{a}$.

Example: Let $\mathcal{A} = \{0, 1\}$ and

$$\mathbf{k} = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$

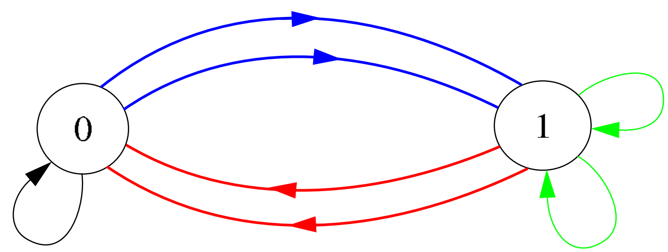


Figure 2: The directed multigraph for a binary alphabet $\mathcal{A} = \{0, 1\}$ with the matrix $\mathbf{k}$ as above.

Enumeration of Eulerian Paths

The enumeration of Eulerian paths in a graph is a classical problem and can be computed through the so called Matrix-Tree Theorem.

We propose a novel approach through multidimensional generating functions.

1. One may try to guess that $N_{\mathbf{k}}^{a}$ is equal to

$$B_{\mathbf{k}} = \binom{k_1}{k_{11} \cdots k_{1m}} \cdots \binom{k_m}{k_{m1} \cdots k_{mm}}$$

with the i factor representing the number of ways departing from vertex i. But this is not true since when tracing the graph $\mathcal{G}_m$ we may get stuck at a node **not** visiting all edges.

2. Let $N_{\mathbf{k},\mathbf{k}'}^{a}$ be the number of ways matrix $\mathbf{k}$ is transformed into another matrix $\mathbf{k}'$ when the Eulerian path starts with symbol a.

$$N_{\mathbf{k},\mathbf{k}'}^{a} = N_{\mathbf{k}-\mathbf{k}'}^{a} \times B_{\mathbf{k}'}, \quad k_a' = 0.$$

Since $\sum_{\mathbf{k}'} N_{\mathbf{k},\mathbf{k}'}^{a} = B_{\mathbf{k}}$,

$$B_{\mathbf{k}} = \sum_{\mathbf{k}' \in \mathcal{F}, k_a' = 0} N_{\mathbf{k}-\mathbf{k}'}^{a} \times B_{\mathbf{k}'}.$$

Generating Functions over Matrices

3. Multiplying by $\mathbf{z}^{\mathbf{k}}$ and summing now over all $\mathbf{z}^{\mathbf{k}}$ such that $k_a \neq 0$ it yields

$$\sum_{\mathbf{k}\in\mathcal{F},k_a\neq 0} B_{\mathbf{k}}\mathbf{z}^{\mathbf{k}} = \left(\sum_{\mathbf{k}\in\mathcal{F}} N^a_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}\right) . \times \left(\sum_{\mathbf{k}\in\mathcal{F},k_a=0} B_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}\right) .$$

4. Let now $B_{\mathcal{A}}(\mathbf{z}) = \sum_{\mathbf{k}\in\mathcal{F}} B_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}$. It is easy to see that

$$B(\mathbf{z}) = \sum_{\mathbf{k}} B_{\mathbf{k}}\mathbf{z}^{\mathbf{k}} = \prod_{a\in\mathcal{A}}(1 - \sum_{b\in\mathcal{A}} z_{a,b})^{-1},$$

but, one must use our Lemma about $\mathcal{F}g(\mathbf{z})$ to find out that

$$B_{\mathcal{A}}(\mathbf{z}) := \mathcal{F}B(\mathbf{z}) = (\det(\mathbf{I} - \mathbf{z}))^{-1}$$

where $\mathbf{I}$ is the identity $m \times m$ matrix.
Above we use the following short-hand notation

$$B_{\mathcal{A}}(\mathbf{z}) := \mathcal{F}B(\mathbf{z})$$

to denote the generating function of $B_{\mathbf{k}}$ over $k \in \mathcal{F}$.

Finishing Up ...

5. We find

$$B_{\mathcal{A}}(\mathbf{z}) - B_{\mathcal{A}-\{a\}}(\mathbf{z}) = N^a(\mathbf{z}) B_{\mathcal{A}-\{a\}}(\mathbf{z}).$$

from which we finally conclude that

$$N^a_{\mathbf{k}} = [\mathbf{z}^{\mathbf{k}}] \frac{B(\mathbf{z})}{B_{\mathcal{A}-\{a\}}(\mathbf{z})} = [\mathbf{z}^{\mathbf{k}}] B(\mathbf{z}) \cdot \det_{aa}(\mathbf{I} - \mathbf{z}).$$

6. Using Cauchy's formula we can prove that

$$N^{b,a}_{\mathbf{k}} = \frac{k_{ba}}{k_b} B_{\mathbf{k}} \cdot \det_{bb}(\mathbf{I} - \mathbf{k}^*) \left(1 + O\left(\frac{1}{n}\right)\right),$$

where $\mathbf{k}^*$ is the matrix whose ij-th element is k_{ij}/k_i, that is, $\mathbf{k}^* = [k_{ij}/k_i]$.

7. For example for a binary Markov we have

$$N^{0,0}_{\mathbf{k}} \sim \frac{k_{10}}{k_{10} + k_{11}} \binom{k_{00} + k_{01}}{k_{00}} \binom{k_{10} + k_{11}}{k_{10}}.$$

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem

4. Method of Types
   (a) Markov Types

   (b) Universal Type
5. Conclusions

Universal Types

Seroussi introduced in 2003 universal types for stationary ergodic sources:

Two sequences of the same length p are said to be of the same universal type if they generate the same set of phrases in the Lempel-Ziv'78 scheme.

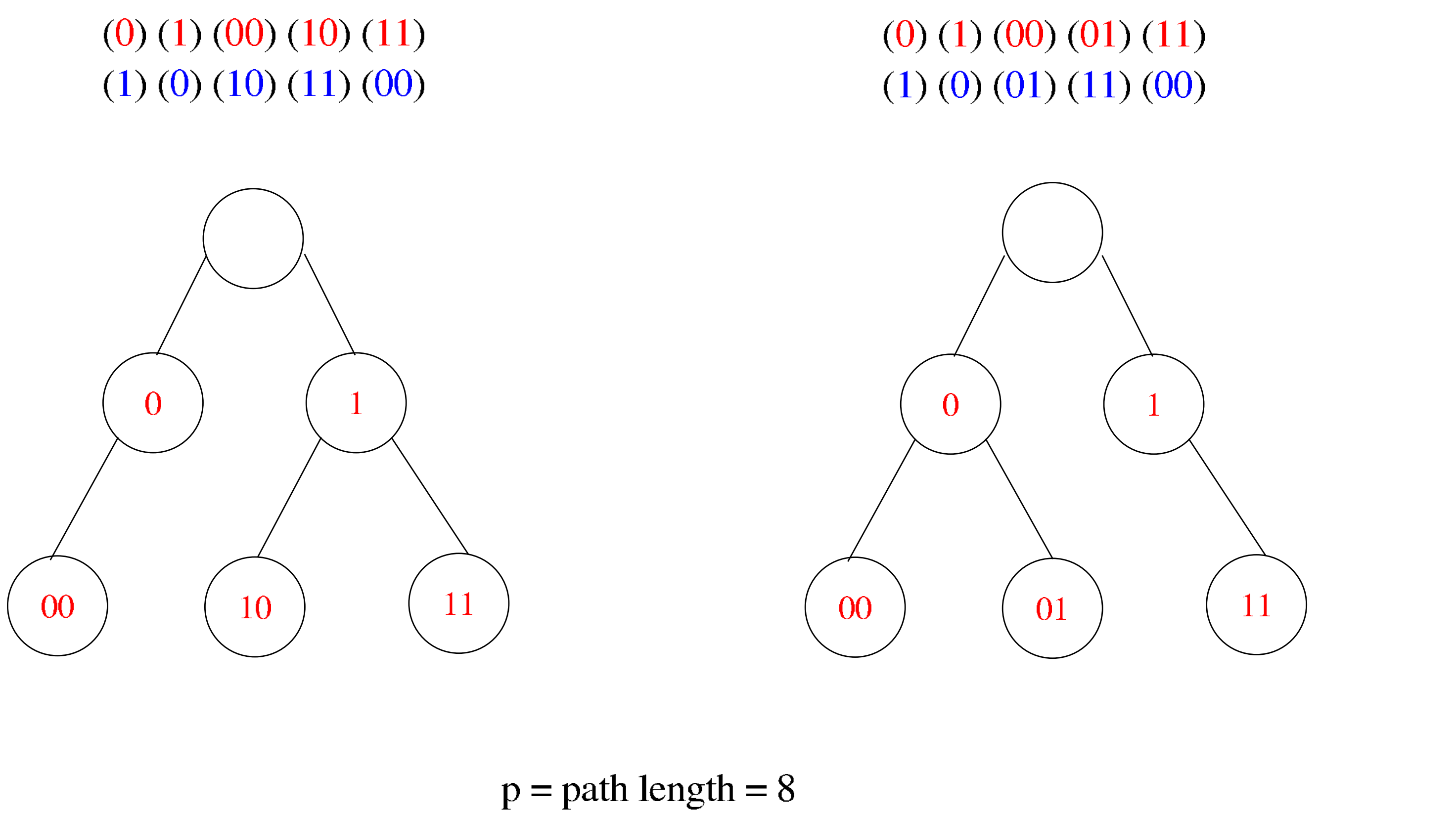


Figure 3: Two universal types and the corresponding binary trees

.

Number of Types and Binary Trees

Lempel-Ziv'78 parsing scheme of a sequence of length p can be represented by a binary tree of path length p.

Let

– $\mathcal{T}_n$ be the set of binary trees built on n nodes.

– $\mathcal{T}_p$ be the set of binary trees with **path length** equal to p.

$$\boxed{\#\ \textbf{universal types} \text{ over } \mathcal{A}^p \quad \equiv \quad |\mathcal{T}_p|.}$$

How to enumerate binary trees of a given path length p?

Enumeration of Binary Trees

$b(n,p)$ – denote the number of binary trees with n nodes and path length p.

$$b(n,p) = \sum_{k+\ell=n-1} \sum_{r+s+n-1=p} b(k,r)b(\ell,s), \quad n \geq 1$$

Define

$$\begin{aligned} B_n(w) &= \sum_{p=0}^{\infty} b(n,p)w^p \\ B(z,w) &= \sum_{n=0}^{\infty}\sum_{p=0}^{\infty} b(n,p)w^p z^n = \sum_{n=0}^{\infty} z^n B_n(w) \end{aligned}$$

Then

$$\begin{aligned} B_{n+1}(w) &= w^n \sum_{\ell=0}^{n} B_\ell(w) B_{n-\ell}(w) \\ B(z,w) &= 1 + zB^2(zw,w) \end{aligned}$$

Enumeration $\mathcal{T}_n$ vs $\mathcal{T}_p$

The functional equation is asymmetric with respect to z and w.

Set $w = 1$, then $B(z,1) = 1 + zB^2(z,1)$, and we find

$$B(z,1) \equiv a(z) = \frac{1}{2z}\left[1 - \sqrt{1-4z}\right].$$

with – $a_n = B_n(1) = \sum_{p\geq 0} b(n,p)$ being the Catalan Number.

We want to study the number of trees in $\mathcal{T}_p$. We set $z = 1$ in the functional equation leading to

$$B(1,w) = 1 + B^2(w,w)$$

which is not algebraically solvable. Observe that

$$|\mathcal{T}_p| = \sum_{n\geq 0} b(n,p) = [w^p]B(1,w).$$

Seroussi (2004) and Knessl & W.S (2004) prove that

$$|\mathcal{T}_p| = \frac{1}{(\log_2 p)\sqrt{\pi p}} 2^{\frac{2p}{\log_2 p}\left(1+c_1\log^{-2/3}p+c_2\log^{-1}p+O(\log^{-4/3}p)\right)}$$

where c_1 and c_2 are constants.

Information-Theoretic Upper Bound

Seroussi's proof is combinatorial. His upper bound uses information-theoretic approach:

Seroussi's Upper Bound:
How many bits are needed to specify a tree with path length p?:
– you must specify p;
– you must tell which one out of $a_n = \frac{1}{n+1}\binom{2n}{n}$ trees of path length p you select.

Thus, the maximum number of bits $B_{\max}$ necessary to describe such a tree is

$$B_{\max} = \log_2 p + \log_2 \frac{1}{n+1}\binom{2n}{n} \sim \frac{2p}{\log_2 p}$$

and clearly

$$|\mathcal{T}_p| \leq 2^{B_{\max}} \sim 2^{2p/\log_2 p}.$$

Analytic Approach – WKB Method

Knessl and W.S. use methods of applied probability called the WKB method.

The WKB method **assumes** that the solution, $B(\xi; n)$, to a functional equation has the following asymptotic form

$$B(\xi; n) \sim e^{n\varphi(\xi)} \left[A(\xi) + \frac{1}{n} A^{(1)}(\xi) + \frac{1}{n^2} A^{(2)}(\xi) + \cdots \right],$$

where $\varphi(\xi)$ and $A(\xi), A^{(1)}(\xi), \ldots$ are unknown functions.
These functions must be determined from the equation itself, often in conjunction with another tool known as the asymptotic matching principle.

For example, for $w = 1 + a/n^{3/2}$ with $a = -Y^{3/2}$ we found that

$$B_n(w) \sim \frac{4^{n+1}}{n^{3/2}}(-a) \sum_{j=0}^{\infty} \exp(-|r_j| 4^{1/3} Y)$$

where $0 > r_0 > r_1 > r_2 > \ldots$ and r_j are the roots of the Airy's function $\mathrm{Ai}(z) = 0$.

Distribution of Number of Nodes

When selecting randomly a tree from $\mathcal{T}_p$ we may define
– N_p number of nodes.

Surprisingly, we can prove that N_p is asymptotically normal, that is,

$$\Pr\{N_p = n\} = \frac{b(n,p)}{\sum_{n=0}^{\infty} b(n,p)} \sim \frac{1}{\sqrt{2\pi\mathcal{V}(p)}} \exp\left[-\frac{(n-\mathcal{N}(p))^2}{2\mathcal{V}(p)}\right]$$

where

$$\mathcal{N}(p) \sim \frac{p}{\log_2 p}, \qquad \mathcal{V}(p) \sim \frac{p}{\log_2 p^{5/3}} \frac{(\log 2)A_0}{6(2^{1/3}}$$

where A_0 is a constant.

Outline Update

1. Some Definitions
2. Two Basic Facts of Source Coding
3. The Redundancy Rate Problem
4. Method of Types
   (a) Markov Types
   (b) Universal Type
5. Conclusions

Conclusions

- In the **1997 Shannon Lecture** Jacob Ziv presented compelling arguments for “backing off” from first-order asymptotics in order to predict the behavior of real systems with finite lengths description.
- To **overcome** these difficulties we propose to replace first-order analyses by full asymptotic expansions and more accurate analyses (e.g., large deviations, central limit laws).
- Following **Knuth and Hadamard’s precept**[1], we study information theory problems using techniques of complex analysis such as generating functions, combinatorial calculus, Rice’s formula, Mellin transform, Fourier series, sequences distributed modulo 1, saddle point methods, analytic poissonization and depoissonization, and singularity analysis.
- This program, which applies complex-analytic tools to information theory, constitutes **analytic information theory**.

---

[1] The shortest paths between two truths on the real line passes through the complex plane.