Algorithms, Combinatorics, and Information[*]

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

April 24, 2007

AofA and **IT** logos

---

[*]Research supported by NSF, and NIH.

[†]Joint work with M. Drmota, P. Jacquet, C. Knessl, S. Lonardi, and M. Ward.

Outline

1. Universal Source Coding
2. **Algorithms**: Error-Resilient Lempel-Ziv'77
3. **Combinatorics**: Method of Types
4. **Analytic Information Theory**: One-to-One Codes
5. **Information**: What is it? Today's Challenges

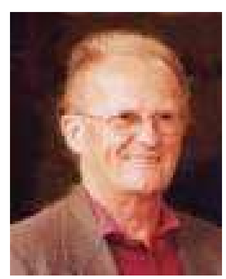
Algorithms:	are at the heart of virtually all computing technologies;
Combinatorics:	provides indispensable tools for finding patterns and structures;
Information:	permeates every corner of our lives and shapes our universe.

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).
- Design efficient algorithms for universal source coding and joint source-channel coding.
- As pointed out 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 a set of 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 $H_n(P) = H(X_1^n) = -\sum_{x_1^n} P(x_1^n) \lg P(x_1^n)$;
  entropy rate $h = \lim H(X_1^n)/n$.

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

Outline Update

1. Universal Source Coding
2. **Algorithms**: Error-Resilient Lempel-Ziv'77
   (a) Redundant Bits in LZ'77
   (b) Design of Encoder and Decoder
   (c) Analysis through the Suffix Tree
3. Combinatorics: Method of Types
4. Analytic Information Theory: One-to-One Codes
5. Information: Today's Challenges

LZ'77 Scheme

The popular Lempel-Ziv'77 scheme works on-line:
It compresses phrases by consecutively replacing the longest prefix of the non-compressed portion of a file with a pointer and the length.

The devastating effect of errors in LZ'77 is a long-standing open problem. Castelli and Lastras in 2004 proved that a single error in LZ'77 corrupts $O(n^{2/3})$ phrases, thus about $O(n^{2/3}\log n)$ symbols, where n is the size the file to be compressed.

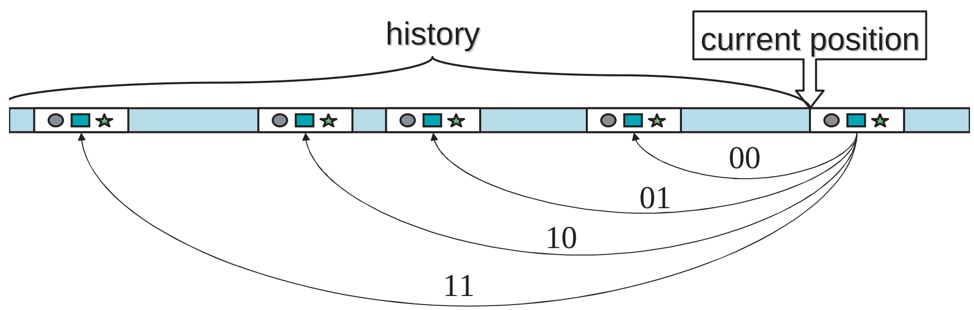


Figure 1: LZ'77 pointers (also for LZRS'77 we have $M_n = 4$).

Our Main Idea of Error Resilient LZ'77

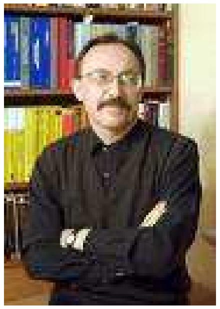
1. We observe that there are usually multiple copies of the longest prefix. By M_n we denote the number of copies of the longest prefix of the uncompressed string that appear in the database.

2. By a judicious choice of pointers in the LZ'77 scheme, we can recover $\lfloor \log_2 M_n \rfloor$ bits **without losing a bit in compression**.

3. Use parity bits recovered from the multiple copies (redundancy) for the Reed-Solomon channel coding.

Note: If the greediness of LZ'77 is relaxed (say, by looking for the 10th largest prefix, for instance), then the number of copies found in the database will increase significantly. This would allow even more errors to be corrected.





Encoder and Decoder of LZRS'77

We use the family of Reed-Solomon codes RS$(255, 255-2e)$ that contains blocks of 255 bytes, of which $255-2e$ are data and $2e$ are parity.

Encoder: The data is broken into blocks of size $255-2e$. Blocks are processed in reverse order, beginning with the very last. When processing block i, the encoder computes first the Reed-Solomon parity bits for the block $i+1$ and then it embeds the extra bits in the pointers of block i.

Decoder: The decoder receives a sequence of pointers, preceded by the parity bits of the first block which are used to correct block B_1. Once block B_1 is correct, it decompresses it using LZS'77. Redundant bits of block B_1 are used as parity bits to correct block B_2, etc.

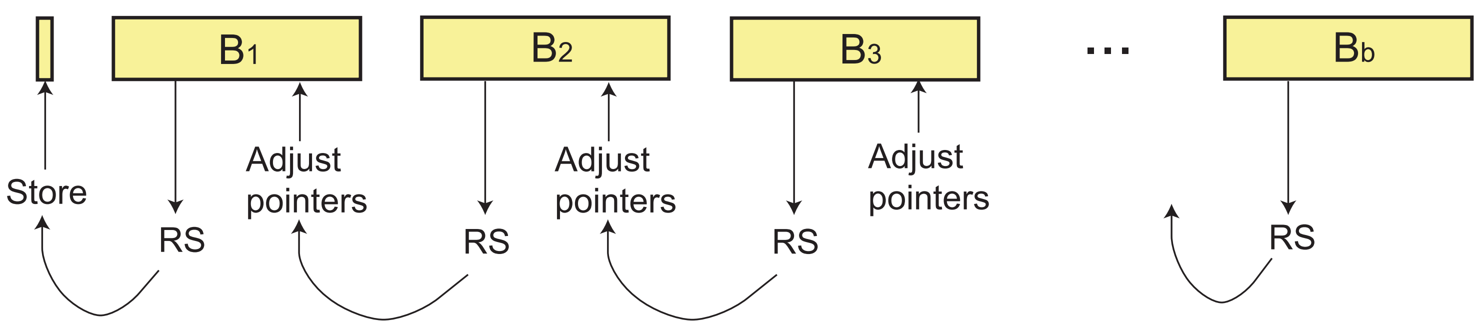


Figure 2: The right-to-left sequence of operations on the blocks.

Experimental Results

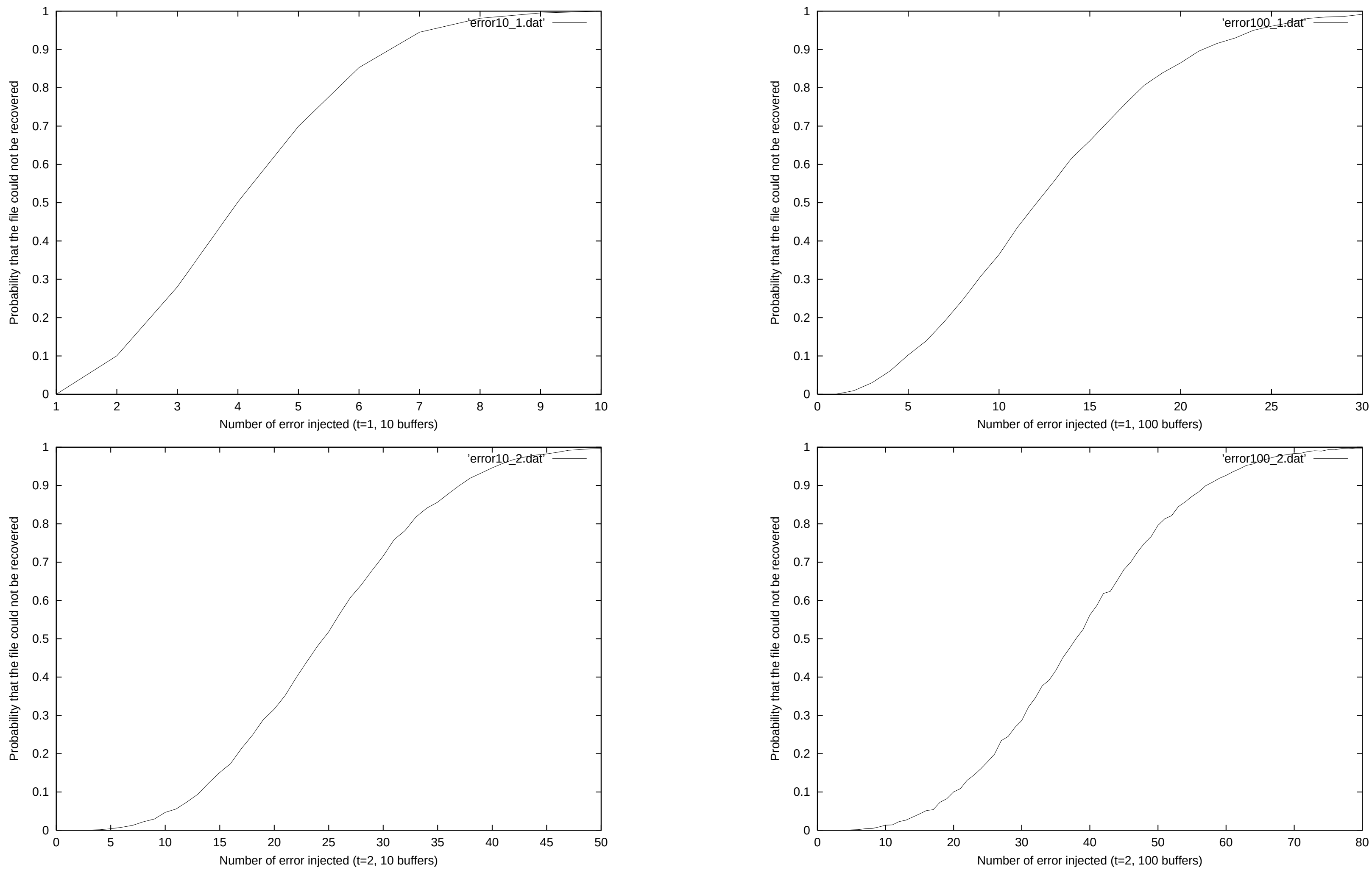


Figure 3: The probability that a file of b blocks could not be recovered correctly vs the number of errors distributed over the blocks. Top-left: $e = 1$ and $b = 10$, top-right: $e = 1$ and $b = 100$, lower-left: $e = 2$ and $b = 10$, lower-right: $e = 2$ and $b = 100$ (e.g., for $e = 2$ and $b = 100$ LZRS'77 can decompress correctly with with 20 uniformly distributed errors 90% of the time).

Analysis of M_n Via Suffix Trees

Performance of LZRS'77 depends on M_n. How does M_n typically behave? Build a **suffix tree** from the first n suffixes of the database X (i.e., $S_1 = X_1^\infty, S_2 = X_2^\infty, \ldots, S_n = X_n^\infty$). Then insert the $(n+1)$st suffix, $S_{n+1} = X_{n+1}^\infty$.

Observe: Depth of insertion of S_{n+1} is the $(n+1)$-st phrase length. Also, M_n is the size of the subtree that starts at the insertion point of the $(n+1)$st suffix.

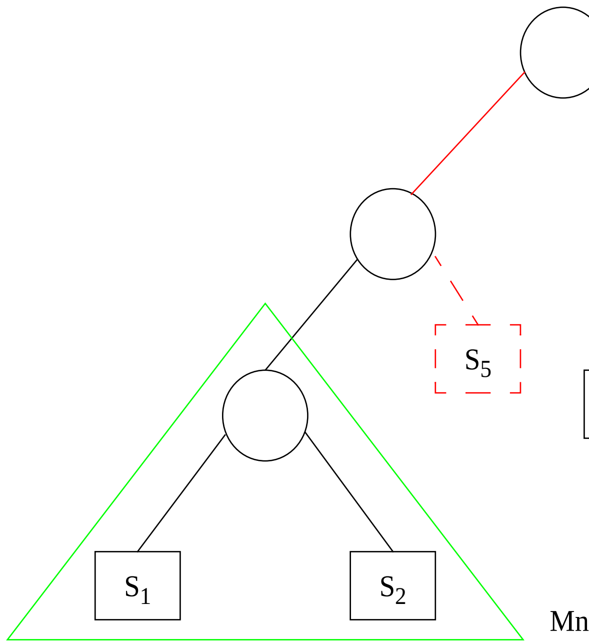


Figure 4: M_4(=2) is the size of the subtree at the insertion point of S_5.

Analytic Information Theory Approach

1. We first consider **digital tries** built over n **independent** strings.
(i) Average $\mathbf{E}[M_n^I]$ satisfies the recurrence
($p = 1 - q$ is the probability of generating a "1"):

$$\mathbf{E}[M_n^I] = p^n(qn+p\mathbf{E}[M_n^I])+q^n(pn+q\mathbf{E}[M_n^I)]+\sum_{k=1}^{n-1}\binom{n}{k}p^k q^{n-k}(p\mathbf{E}[M_k^I]+q\mathbf{E}[M_{n-k}^I]);$$

(ii) The probability generating functions $\mathbf{E}[u^{M_n^I}]$ satisfy

$$\mathbf{E}[u^{M_n^I}] = p^n(qu^n+p\mathbf{E}[u^{M_n^I}])+q^n(pu^n+q\mathbf{E}[u^{M_n^I}])+\sum_{k=1}^{n-1}\binom{n}{k}p^k q^{n-k}(p\mathbf{E}[u^{M_k^I}]+q\mathbf{E}[u^{M_{n-k}^I}])$$

2. Using analytic combinatorics on words we prove that for any $\varepsilon > 0$ there exists $\beta > 1$ such that (all hard analytic work is here!)

$$\Pr(M_n = k) - \Pr(M_n^I = k) = O(n^{-\varepsilon}\beta^{-k})$$

for large n.

Random suffix trees resemble random independent tries.

Main Results

Theorem 1 (Ward, W.S., 2005). *Let* $z_k = \frac{2kr\pi i}{\ln p}$ $\forall k \in \mathbb{Z}$, *where* $\frac{\ln p}{\ln q} = \frac{r}{s}$ *for some relatively prime* $r, s \in \mathbb{Z}$ *(i.e.,* $\frac{\ln p}{\ln q}$ *is rational).*
The j*th factorial moment* $E[(M_n)^{\underline{j}}] = E[M(M-1)\cdots M(-j+1)]$ *is*

$$E[(M_n)^{\underline{j}}] \quad = \quad \Gamma(j)\frac{q(p/q)^j + p(q/p)^j}{h} + \delta_j(\log_{1/p} n) + O(n^{-\eta})$$

where $h = -p\log p - q\log q$ *is the entropy rate,* $\eta > 0$*, and where* Γ *is the Euler gamma function and*

$$\delta_j(t) = \sum_{k\neq 0} - \frac{e^{2kr\pi it}\Gamma(z_k + j)\left(p^j q^{-z_k-j+1} + q^j p^{-z_k-j+1}\right)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q}.$$

δ_j is a periodic function that has a small magnitude and exhibits fluctuation when $\frac{\ln p}{\ln q}$ is rational

Note: On average there are $\mathbf{E}[M_n] \sim 1/h$ additional pointers.

j	$\frac{1}{\ln 2}\sum_{k\neq 0}\left|\Gamma\left(j - \frac{2ki\pi}{\ln 2}\right)\right|$
1	1.4260 $\times 10^{-5}$
3	1.2072 $\times 10^{-3}$
5	1.1421 $\times 10^{-1}$
6	1.1823 $\times 10^{0}$
8	1.4721 $\times 10^{2}$
9	1.7798 $\times 10^{3}$
10	2.2737 $\times 10^{4}$

Distribution of M_n

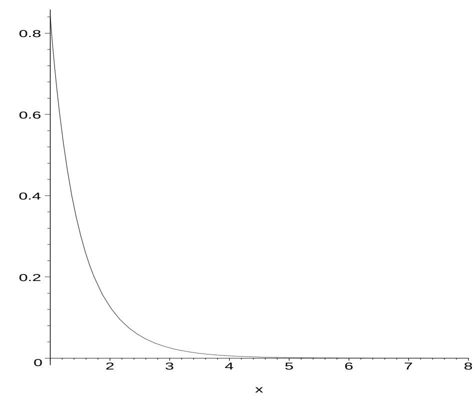
Theorem 2 (Ward, W.S., 2005). *Let* $z_k = \frac{2kr\pi i}{\ln p}$ $\forall k \in \mathbb{Z}$, *where* $\frac{\ln p}{\ln q} = \frac{r}{s}$ *for some relatively prime* $r, s \in \mathbb{Z}$. *Then*

$$P(M_n = j) \quad = \quad \frac{p^j q + q^j p}{jh} + \sum_{k \neq 0} -\frac{e^{2kr\pi i \log_{1/p} n}\Gamma(z_k)(p^j q + q^j p)(z_k)^{\overline{j}}}{j!(p^{-z_k+1}\ln p + q^{-z_k+1}\ln q)} + O(n^{-\eta})$$

where $\eta > 0$ *and* Γ *is the Euler gamma function.*

Therefore, M_n *follows the logarithmic series distribution with mean* $1/h$ *(plus some fluctuations).*

The logarithmic series distribution ($(p^j q + q^j p)/(jh)$)
is well concentrated around its mean $\mathbf{E}M_n \approx 1/h$.



Outline Update

1. Universal Source Coding
2. Algorithms: Error-Resilient Lempel-Ziv'77
3. **Combinatorics**: Method of Types
   (a) Markov Types and Eulerian Paths
   (b) Universal Types and Enumeration of Binary Trees
4. Analytic Information Theory: One-to-One Codes
5. Information: Today's Challenges

Method of Types

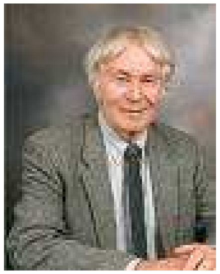
The method of types is a powerful technique in **information theory**; it reduces calculations of the probability of rare events to combinatorics.

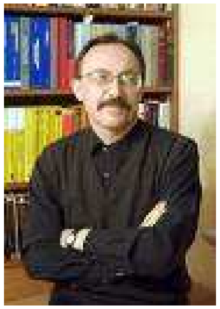
Sequences are of the same type if they have the same empirical distribution.

Warm-up Problem: How many binary strings x_1^n of length n generated by a **memoryless source** have k "1"s (i.e., have the same Bernoulli type)? All such strings have the same probability

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

where p is the probability of generating a 1.
Answer: Certainly, the answer is: $\binom{n}{k}$.





Markov Types

Consider a Markov source over an m-ary alphabet with the transition matrix $P = \{p_{ij}\}_{i,j=1}^{m}$, that is, $P(X_{t+1} = j|X_t = i) = p_{ij}$.
The probability of x_1^n is

$$P(x_1^n) = p_{11}^{k_{11}} \cdots p_{mm}^{k_{mm}}$$

where k_{ij} is the number of pair symbols ij in x_1^n, that is, i followed by j.

Example: Let $x_1^n = 01101$, then

$$P(01101) = p_{01}^2 p_{11} p_{10}.$$

For circular strings (i.e., after the n symbol we re-visit the first symbol of x_1^n), the matrix $[k_{ij}]$ satisfies the following constraints that we denote as $\mathcal{F}_n$

$$\sum_{1 \le i,j \le m} k_{ij} = n; \qquad \sum_{j=1}^{m} k_{ij} = \sum_{j=1}^{m} k_{ji}, \ \forall\ i \ \ (\textbf{balance property})$$

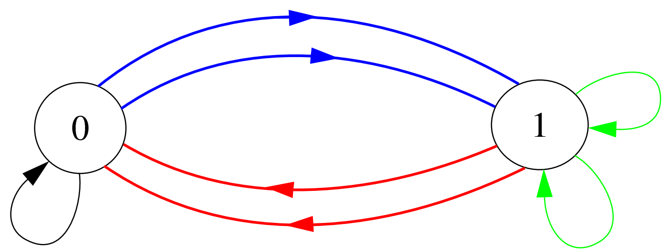
Markov Types and Eulerian Cycles

Problem. Let $\mathbf{k} = [k_{ij}]_{i,j=1}^{m}$ be a given frequency matrix satisfying the balance property.

A: *How many strings of a given frequency matrix* $\mathbf{k}$ (**given type**) *are there*?

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

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



B: *How to enumerate Eulerian paths* (**types**) *in a multigraph* with $|\mathcal{A}|$ vertices and k_{ij} edges between ith and jth vertices?

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 strings x_1^n of type $\mathbf{k}$ and starting with a symbol a.
$N_{\mathbf{k}}^{ab}$ – # strings x_1^n of type $\mathbf{k}$, starting with a symbol a and ending with b.

Enumeration of Eulerian Paths

1. Define for an m-ary alphabet

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

2. $N^a_{\mathbf{k},\mathbf{k}'}$ – # ways $\mathbf{k}$ is transformed into $\mathbf{k}'$ starting from a:

$$N^a_{\mathbf{k},\mathbf{k}'} = N^a_{\mathbf{k}-\mathbf{k}'} \times B_{\mathbf{k}'}, \quad k'_a = 0.$$

Since $\sum_{\mathbf{k}'} N^a_{\mathbf{k},\mathbf{k}'} = B_{\mathbf{k}}$, hence $B_{\mathbf{k}} = \sum_{\mathbf{k}' \in \mathcal{F}, k'_a=0} N^a_{\mathbf{k}-\mathbf{k}'} \times B_{\mathbf{k}'}$.

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

$$N^{b,a}_{\mathbf{k}} = [\mathbf{z}^{\mathbf{k}}] B(\mathbf{z}) z_{ba} \cdot \det_{bb}(\mathbf{I} - \mathbf{z}).$$

where $\mathcal{F}B(\mathbf{z}) = (\det(\mathbf{I} - \mathbf{z}))^{-1}$. Using Cauchy we arrive at

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

where $\mathbf{k}^*$ is the normalized matrix such that $\mathbf{k}^* = [k_{ij}/k_i]$.

4. For example for a binary Markov we have

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

Universal Types

Seroussi introduced in 2003 universal types for stationary ergodic sources:

> 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.

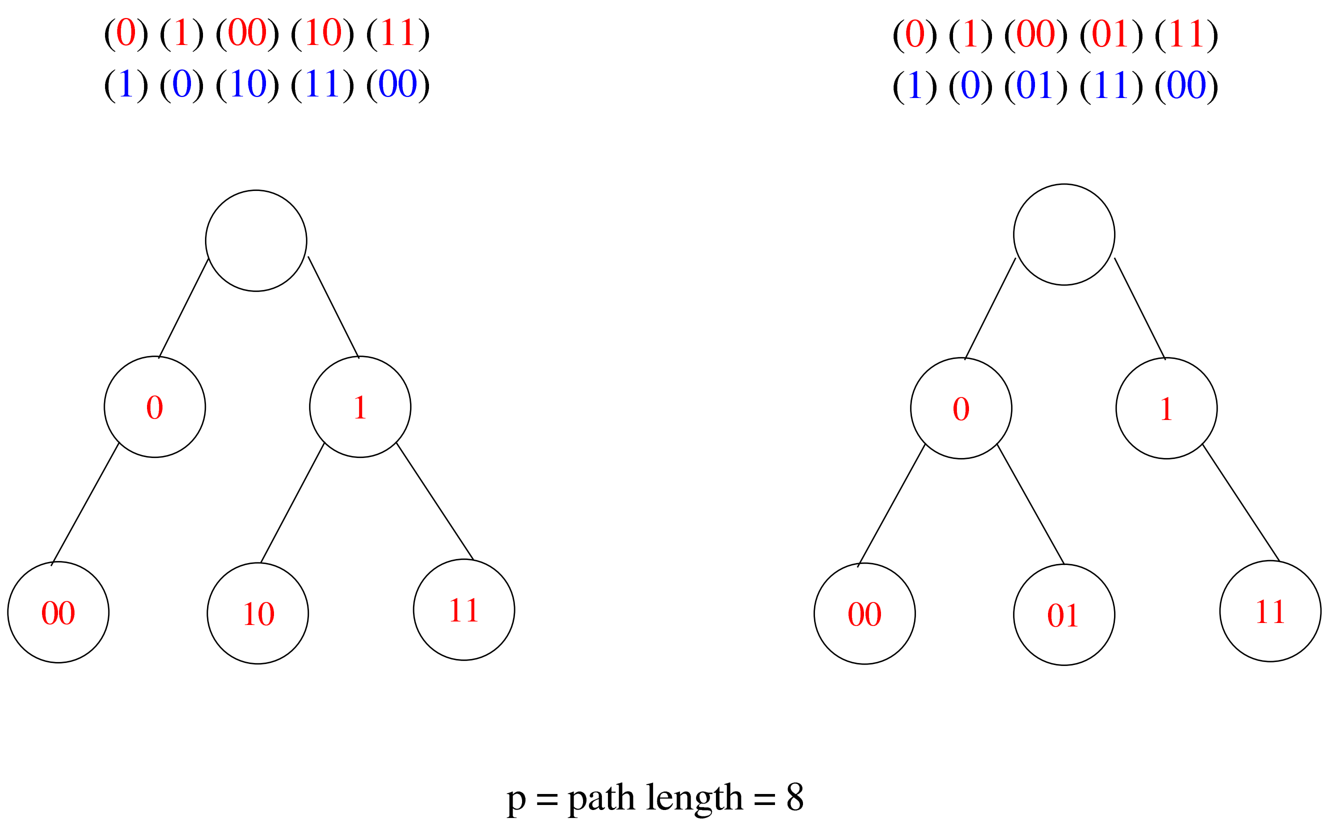


Figure 5: Two universal types and the corresponding binary trees

.

Number of Types and Binary Trees

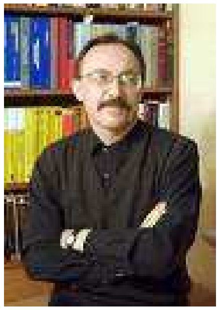
Lempel-Ziv'78 parsing scheme of a sequence of **length** p can be represented by a **binary tree of path length** p. Let

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

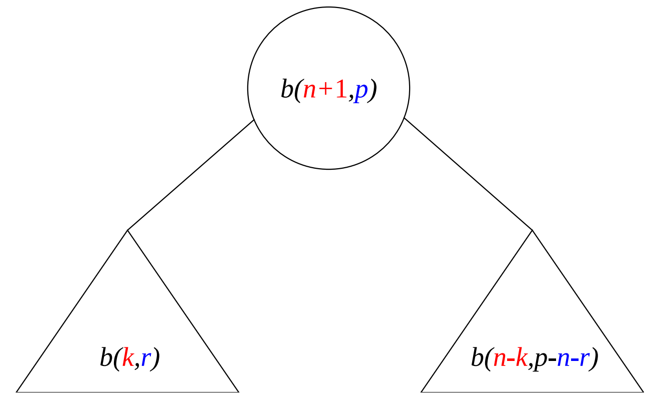
– $\mathcal{T}_p$ be the set of binary trees with **path length** equal to p.

$\boxed{\text{\# \textbf{universal types} over } \mathcal{A}^p \equiv |\mathcal{T}_p|\text{: \# of trees a given path } p.}$

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



Enumeration of Binary Trees



Let $b(n,p)$ be the number of binary trees with n nodes and path length p. It satisfies:

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

Define

$$B_n(w) = \sum_{p=0}^{\infty} b(n,p) w^p, \quad B(z,w) = \sum_{n=0}^{\infty} z^n B_n(w)$$

Then

$$\boxed{B(z,w) = 1 + zB^2(zw,w)}$$

This **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

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

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

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

We want to study the number of trees in $\mathcal{T}_p$. Observe

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

We set $z = 1$ in the functional equation leading to

$$\boxed{B(1,w) = 1 + B^2(w,w)}$$

which is not algebraically solvable.
Seroussi (2004) and Knessl & W.S (2004) prove that (c_1, c_2 are constants)

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

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 $A(\xi)$, $A^{(1)}(\xi)$, ... are unknown functions. These functions must be determined from the equation (asymptotic matching principle).

Outline

1. Universal Source Coding
2. Algorithms: Error-Resilient Lempel-Ziv'77
3. Combinatorics: Method of Types
4. **Analytic Information Theory**: One-to-One Codes
   (a) Lower Bound
   (b) Anti-Redundancy
   (c) Sketch of Proof: Generating Functions and Complex Asymptotics
5. Information: Today's Challenges

Prefix Codes and Lower Bound

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

Kraft's Inequality: Code lengths $\ell_1, \ell_2, \ldots, \ell_m$ satisfy the inequality

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

Lower Bound (Khinchin, 1953): Average code length $\mathbf{E}[L(C_n, X_1^n)]$ satisfies:

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

Proof: Let $K = \sum_{x_1^n} 2^{-L(x_1^n)} \overset{\text{Kraft}}{\leq} 1$.

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

since the divergence cannot be negative (or $\log x \leq x - 1$) and $K \leq 1$.

One-to-One Codes

One-to-One codes are not prefix codes.

In one-to-one codes a distinct codeword is assigned to each source symbol (unique decodability is not required).
Such codes are usually one shot codes and there is one designated an "end of message" channel symbol.

Wyner in 1972 proved that

$$L \leq H(X),$$

further improved by Alon and Orlitsky who showed

$$L \geq H(X) - \log(H(X) + 1) - \log e.$$

We consider a block one-to-one code for $x_1^n = x_1 \dots x_n \in \mathcal{A}^n$ generated by a memoryless source;
p the probability of generating a 0 and $q = 1 - p$.
Throughout: $p \leq q$ so that $P(x_1^n) = p^k q^{n-k}$.

Goal: **More precise bounds for the (anti-)redundancy** $L - H(X)$.

Average Code length

List all 2^n probabilities in a nonincreasing order and assign code lengths:

$$q^n \left(\frac{p}{q}\right)^0 \geq q^n \left(\frac{p}{q}\right)^1 \geq \cdots \geq q^n \left(\frac{p}{q}\right)^n$$

$$\lfloor \log_2(1) \rfloor \qquad \lfloor \log_2(2) \rfloor \qquad \cdots \qquad \lfloor \log_2(2^n) \rfloor$$

There are $\binom{n}{k}$ equal probabilities $p^k q^{n-k}$. Define

$$A_k = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k}, \quad A_{-1} = 0.$$

Since starting from the position A_{k-1} the next $\binom{n}{k}$ probabilities $P(x_1^n)$ are the same, the average code length is

$$\begin{aligned} L_n &= \sum_{k=0}^{n} p^k q^{n-k} \sum_{j=A_{k-1}+1}^{A_k} \lfloor \log_2(j) \rfloor \\ &= \sum_{k=0}^{n} p^k q^{n-k} \sum_{i=1}^{\binom{n}{k}} \lfloor \log_2(A_{k-1} + i) \rfloor. \end{aligned}$$

Main Result

Theorem 3. *For a binary memoryless source, let* $p < \frac{1}{2}$. *Then*

$$\begin{aligned} L_n &= nH(p) - \frac{1}{2}\log_2 n - 1 - \frac{1}{2\ln 2} + \log_2 \frac{1-p}{(1-2p)\sqrt{pq\pi}} \\ &+ \frac{1-p}{1-2p}\log_2\frac{2-3p}{1-p} + \frac{5-4p}{1-2p}\left(\frac{1}{2\ln 2} + G(n)\right) \\ &+ F(n) + o(1) \end{aligned}$$

where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$, *and*

- $\lim G(n) = \lim F(n) = \mathrm{const}$ *if* $\log_2 \frac{1-p}{p}$ *is irrational;*
- $G(n)$ *and* $F(n)$ *are oscillating functions if* $\log_2\frac{1-p}{p} = N/M$ *is rational,* *e.g.,*

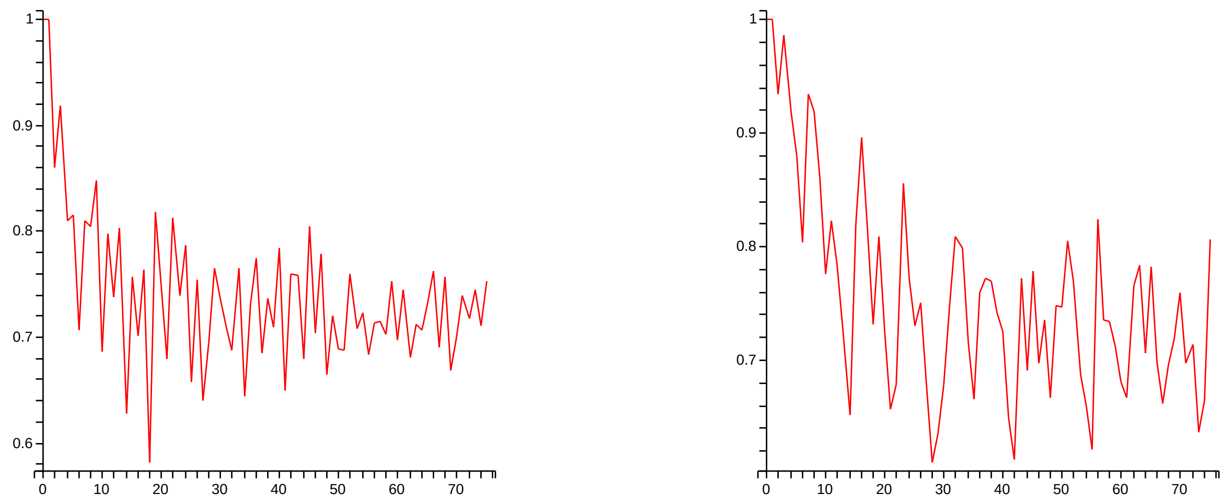
$$F(n) = \frac{1}{M\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\left(\left\langle M\left(n\beta - \log\left(\frac{1-2p}{1-p}\sqrt{2\pi pqn}\right) - \frac{x^2}{2\ln 2}\right)\right\rangle - \frac{1}{2}\right)dx$$

where $\beta = -\log_2(1-p)$ *and* $\langle x\rangle = x - \lfloor x \rfloor$.

For $p = \frac{1}{2}$, *then for all* $n \geq 1$

$$L_n = nH(1/2) - 1 + 2^{-n}(n-2).$$

Oscillations



(a) (b)

Figure 6: The fluctuating part of the average anti-redundancy versus n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

Anti-redundancy $R_n = L_n - nH(p)$ for our one-to-one code is

$$\bar{R}_n = -\frac{1}{2}\log n + O(1)$$

where the $O(1)$ terms contains oscillations, as shown above.

Sketch of Proof

1. Using Knuth's identity (to handle floor functions)

$$\sum_{j=1}^{N} a_j = N a_n - \sum_{j=1}^{N-1} (a_{j+1} - a_j)$$

we can reduce L_n to the sums of the following form

$$\begin{aligned} S_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log_2 A_k \rfloor \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log_2 A_k - \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle \\ &= a_n + b_n \end{aligned}$$

where

$$\begin{aligned} a_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log_2 A_k, \\ b_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle . \end{aligned}$$

Asymptotics of A_n

2. We need the saddle point approximation of A_n.
Lemma 1. *For large* n *and* $p < 1/2$

$$A_{np} = \frac{1-p}{1-2p}\frac{1}{\sqrt{2\pi np(1-p)}}2^{nH(p)}\left(1 + O(n^{-1/2})\right).$$

More precisely, for an $\varepsilon > 0$ *and* $k = np + \Theta(n^{1/2+\varepsilon})$ *we have*

$$\begin{aligned} A_k &= \frac{1-p}{1-2p}\frac{1}{\sqrt{2\pi np(1-p)}}\left(\frac{1-p}{p}\right)^k \frac{1}{(1-p)^n} \\ &\quad \times \exp\left(-\frac{(k-np)^2}{2p(1-p)n}\right)\left(1+O(n^{-\delta})\right) \end{aligned}$$

for some $\delta > 0$.

Proof. Notice that

$$A_n(z) = \sum_{k=0}^{n} A_k z^k = \frac{(1+z)^n - 2^n z^{n+1}}{1-z}.$$

Apply the saddle point method to the Cauchy formula $A_k = [z^n]A_n(z)$.

Returning to b_n

3. We also need asymptotics of

$$b_n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle .$$

From previous lemma we conclude that

$$\log A_k = \alpha k + n\beta - \log_2 \omega\sqrt{n} - \frac{(k-np)^2}{2pqn \ln 2} + O(n^{-\delta})$$

for some $\omega > 0$ and $\alpha = \log p/(1-p)$.

Thus we need asymptotics of the following sum

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left\langle \alpha k + n\beta - \log_2 \omega\sqrt{n} - \frac{(k-np)^2}{2pqn \ln 2} \right\rangle .$$

We must now resort to theory of **Bernoulli sequences modulo** 1.

Final Lemma

Lemma 2. *Let $0 < p < 1$ be a fixed real number and $f : [0,1] \to \mathbf{R}$ be a Riemann integrable function.*

(i) *If α is irrational, then*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\left\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\right\rangle\right) = \int_0^1 f(t)\,dt,$$

where the convergence is uniform for all shifts $y \in \mathbf{R}$.

(ii) If $\alpha = \frac{N}{M}$ (rational) $(\gcd(N,M)=1)$, then uniformly $y \in \mathbf{R}$

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\left\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\right\rangle\right) = \int_0^1 f(t)\,dt + H_M(y)$$

where

$$H_M(y) \quad := \quad \frac{1}{M}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \left(\left\langle M\left(y - \frac{x^2}{2\ln 2}\right)\right\rangle - \int_0^1 f(t)\,dt\right)\,dx$$

is a periodic function with period $\frac{1}{M}$.

Outline

1. Universal Source Coding
2. Algorithms: Error-Resilient Lempel-Ziv'77
3. Combinatorics: Method of Types
4. Analytic Information Theory: Minimax Redundancy
5. **Information**: What is it?
   (a) Computer Science and Information Theory Interplay
   (b) Beyond Shannon
   (c) Today's Challenges
   (d) Information Science Institute?

Information Theory and Computer Science Interface

Although the interplay between **IT** and **CS** dates back to the founding father of information theory, Claude E. Shannon, only in 2003 was the first NSF sponsored Workshop on Information Theory and Computer Science Interface held in Chicago.

Examples of IT and CS Interplay:

Lempel-Ziv schemes (Ziv, Lempel, Louchard, Jacquet, Szpankowski)

LDPC coding, Tornado and Raptor codes (Gallager, Luby, Mitzenmacher, Shokrollahi, Urbanke)

List-decoding algorithms for error-correcting codes (Gallager, Sudan, Guruswami, Koetter, Vardy);

Kolmogorov complexity (Kolmogorov, Cover, Li, Vitanyi, Lempel, Ziv);

Analytic information theory (Jacquet, Flajolet, Drmota, Savari, Szpankowski);

Quantum computing and information (Shor, Grover, Schumacher, Bennett, Deutsch, Calderbank);

Network coding and wireless computing (Kumar, Yang, Effros, Verdu).

Information Beyond Shannon

Participants of the **Information Beyond Shannon** workshop, Orlando, 2005 listed the following research issues:

Delay: In computer networks, delay incurred is a nontrivial issue not yet addressed in information theory (e.g., complete information arriving late maybe useless).

Space: In networks the spatially distributed components raise fundamental issues of limitations in information exchange since the available resources must be shared, allocated and re-used.

Information and Control: Again in networks our objective is to reliably send data with high bit rate and small delay (control).
For example, in wireless/ad-hoc networks, information is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute the basic objective.

Dynamic information: In a complex network in a space-time-control environment (e.g., human brain information is not simply communicated but also processed) how can the consideration of such dynamical sources be incorporated into the Shannon-theoretic model?

Today's Grand Challenges

- We still lack measures and meters to define and appraise the amount of structure and organization embodied in artifacts and natural objects.
- Information accumulates at a rate faster than it can be sifted through, so that the bottleneck, traditionally represented by the medium, is drifting towards the receiving end of the channel.
- Timeliness, space and control are important dimensions of Information. Time and space varying situations are rarely studied in Shannon Information Theory.
- In a growing number of situations, the overhead in accessing Information makes information itself practically unattainable or obsolete.
- Microscopic systems do not seem to obey Shannon's postulates of Information. In the quantum world and on the level of living cells, traditional Information often fails to accurately describe reality.

Science of Information

A Vision

Perhaps it is time to initiate an

Institute for Science of Information

integrating research and teaching activities aimed at investigating the role of **information** from various viewpoints: from the fundamental theoretical underpinnings of information to the science and engineering of novel information substrates, biological pathways, communication networks, economics, and complex social systems.

The specific means and goals for the Center are:

- initiate the Prestige Science Lecture Series on Science of Information to collectively ponder short and long term goals;
- study dynamic information theory that extends information theory to time–space–varying situations;
- advance information algorithmics that develop new algorithms and data structures for the application of information;
- encourage and facilitate interdisciplinary collaborations;
- provide scholarships and fellowships for the best students, and support the development of new interdisciplinary courses.