Average Redundancy for Known Sources: Ubiquitous Trees in Source Coding*

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

September 26, 2008

AofA and **IT** logos

MathInfo, Blaubeuren, 2008

*Research supported by NSF and NIH.

†Partially jointly with M. Drmota, S. Savari and Y. Reznik.

Outline

1. Source Coding (FV, VF, VV Codes)
2. Some Preliminary Facts
3. Redundancy of Huffman Code
4. Redundancy of Tunstall and Khodak VF Codes
5. Redundancy of Khodak VV Code
6. Concluding Remarks
   (a) Universal Shannon Code
   (b) Non-Prefix Codes: One-to-One Codes

Source Coding

A source code is a bijective mapping

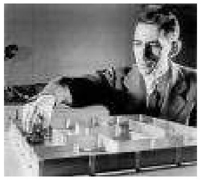
$$C : \mathcal{A}^* \to \{0,1\}^*$$

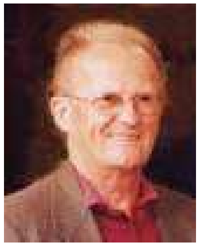
from sequences over the alphabet $\mathcal{A}$ to set $\{0,1\}^*$ of binary sequences.

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*.

Three Basic Types of Source Coding:

- Fixed-to-Variable (**FV**) length codes (e.g., Huffman and Shannon codes).
- Variable-to-Fixed (**VF**) length codes (e.g., Tunstall and Khodak codes).
- Variable-to-Variable (**VV**) length codes (e.g., Khodak VV code).

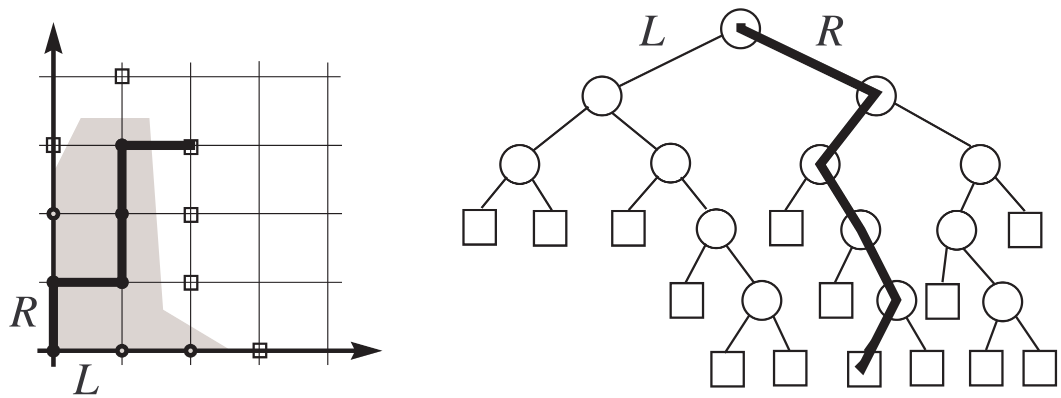




Preliminary Results

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

Tree and lattice representations:



Notation: For a source model $\mathcal{S}$ and a code C we let:

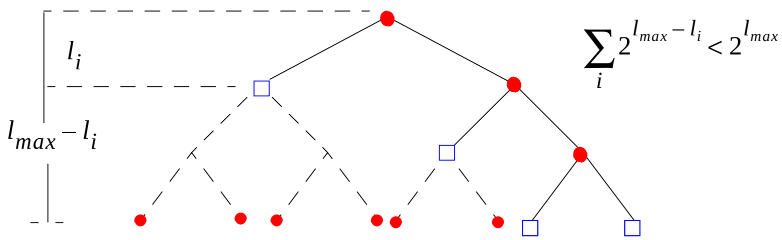
- $P(x)$ be the probability of $x \in \mathcal{A}^*$;
- $L(C, x)$ be the code length for the source sequence $x \in \mathcal{A}^*$;
- Entropy $H(P) = -\sum_{x \in \mathcal{A}^*} P(x) \lg P(x)$.

Quantities are expressed in binary logarithms written $\lg := \log_2$.

Prefix Codes

Kraft's Inequality

A binary code is a prefix code iff the code lengths $\ell_1, \ell_2, \ldots, \ell_N$ satisfy



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

Barron's lemma

For any sequence a_n of positive constants satisfying $\sum_n 2^{-a_n} < \infty$

$$\Pr\{L(X) < -\log P(X) - a_n\} \leq 2^{-a_n},$$

and therefore

$$L(X) \geq -\log P(X) - a_n \quad (\text{a.s}).$$

Proof: We argue as follows:

$$\begin{aligned}
\Pr\{L(X) &< -\log_2 P(X) - a_n\} = \sum_{x : P(x) < 2^{-L(x) - a_n}} P(x) \\
&\leq \sum_{x : P(x) < 2^{-L(x) - a_n}} 2^{-L(x) - a_n} \\
&\leq 2^{-a_n} \sum_x 2^{-L(x)} \leq 2^{-a_n}.
\end{aligned}$$

Shannon Lower Bound

Shannon First Theorem

For any prefix code the average code length $\mathbf{E}[L(C,X)]$ cannot be smaller than the entropy of the source $H(P)$, that is,

$$\mathbf{E}[L(C_n,X)] \geq H(P).$$

Proof: Let $K = \sum_x 2^{-L(x)} \leq 1$, and $L(C,x) := L(C)$. Then

$$\begin{aligned}
\mathbf{E}[L(C,X)] &- H(P) = \\
&= \sum_{x\in\mathcal{A}*} P(x)L(x) + \sum_{x\in\mathcal{A}*} P(x)\log P(x) \\
&= \sum_{x\in\mathcal{A}*} P(x)\log\frac{P(x)}{2^{-L(x)}/K} - \log K \\
&\geq 0
\end{aligned}$$

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

Redundancy

Known Source P (assumed throughout the talk).

The pointwise redundancy $R^C(x)$ and the average redundancy $\bar{R}^C$:

$$\begin{aligned} R^C(x) &= L(C,x) + \lg P(x) \\ \bar{R}^C &= \mathbf{E}[L(C,X)] - H(P) \geq 0 \end{aligned}$$

Optimal Code:

$$\min_L \sum_x L(x)P(x) \quad \text{subject to} \quad \sum_x 2^{-L(x)} \leq 1.$$

Solution: By Lagrangian multipliers we find $L^{opt}(x) = -\lg P(x)$.

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

Outline Update

1. Source Coding
2. Preliminary Facts of Source Coding
3. Redundancy of Huffman Code
4. Redundancy of Tunstall and Khodak VF Codes
5. Redundancy of Khodak VV Code
6. Concluding Remarks

Redundancy for Huffman's Code

We consider **fixed-to-variable length codes**; in particular, Huffman's code.

For a known source P, we consider fixed length sequences $x_1^n = x_1 \dots x_n$.

Huffman Code: The following optimization problem

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

is solved by Huffman's code.

We study the average redundancy for a binary memoryless sources with p denoting the probability of generating "0" and $q = 1 - p$.

In 1994 Stubley proposed the following for Huffman's average redundancy

$$\bar{R}_n^H = 2 - \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \alpha k + \beta n \rangle - 2 \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} 2^{-\langle \alpha k + \beta n \rangle} + o(1).$$

where

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

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

Main Result

Theorem 1 (W.S., 2000). *Consider the Huffman block code of length n over a binary memoryless source with $p < \frac{1}{2}$. Then as $n \to \infty$*

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

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

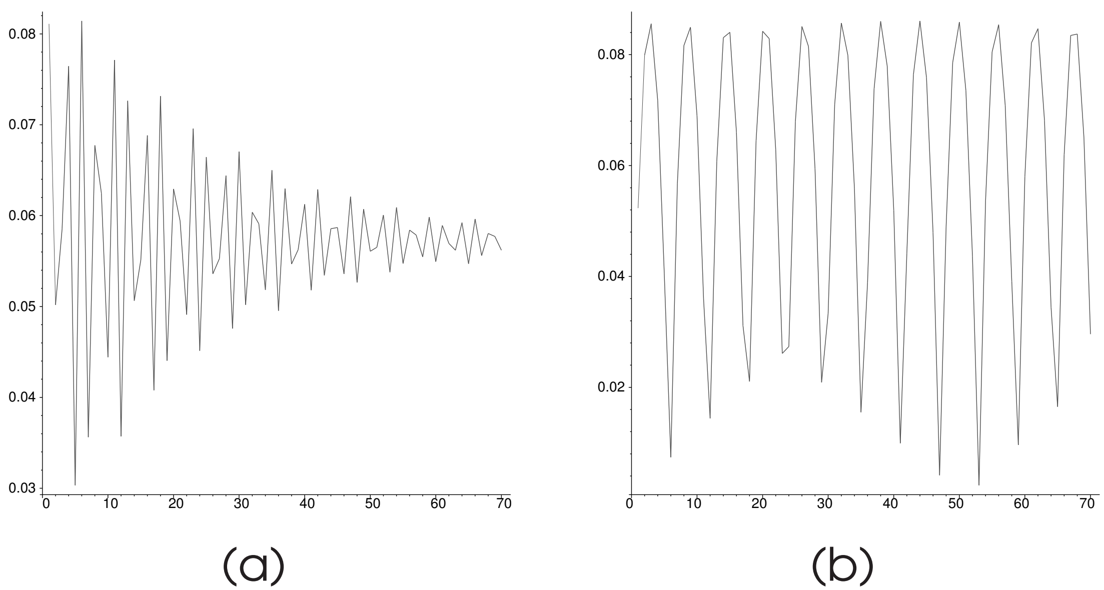


Figure 1: The average redundancy of Huffman codes versus block size n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

Sketch of Proof

We need to understand asymptotic behavior of the following sum (cf. Bernoulli distributed sequences modulo 1)

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

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

Lemma 1. *Let $0 < p < 1$ be a fixed real number and α be an irrational number. Then for every Riemann integrable function $f : [0,1] \to \mathbb{R}$*

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

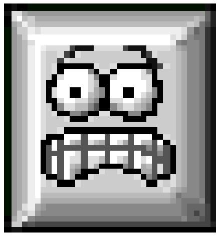
where the convergence is uniform for all shifts $y \in \mathbb{R}$.
Lemma 2. *Let $\alpha = \frac{N}{M}$ be a rational number with $\gcd(N, M) = 1$. Then for bounded function $f : [0,1] \to \mathbb{R}$*

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y \rangle) = \frac{1}{M} \sum_{l=0}^{M-1} f\left(\frac{l}{M} + \frac{\langle My \rangle}{M}\right) + O(\rho^n)$$

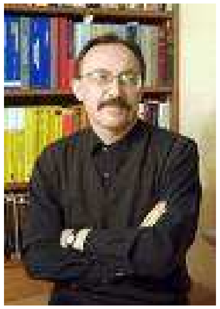
uniformly for all $y \in \mathbb{R}$ and some $\rho < 1$.

Outline Update

1. Source Coding
2. Preliminary Facts of Source Coding
3. Redundancy of Huffman Code
4. Redundancy of Tunstall and Khodak VF Codes
5. Redundancy of Khodak VV Code
6. Concluding Remarks

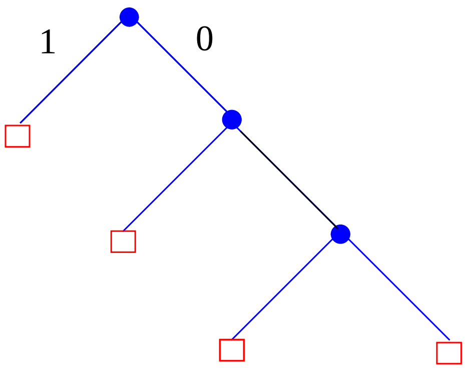
(Y. Reznik)

Variable-to-Fixed Codes

A VF coder consists of a **parser** and a **dictionary**.

Parsing Tree



Dictionary

1
01
001
000

1. A variable-to-fixed length encoder partitions the source string into a concatenation of variable-length phrases.

2. Each phrase belongs to a given dictionary $\mathcal{D}$ of source strings.

3. A dictionary can be represented by a complete parsing tree $\mathcal{T}$.

The dictionary entries $d \in \mathcal{D}$ correspond to the leaves of the parsing tree.

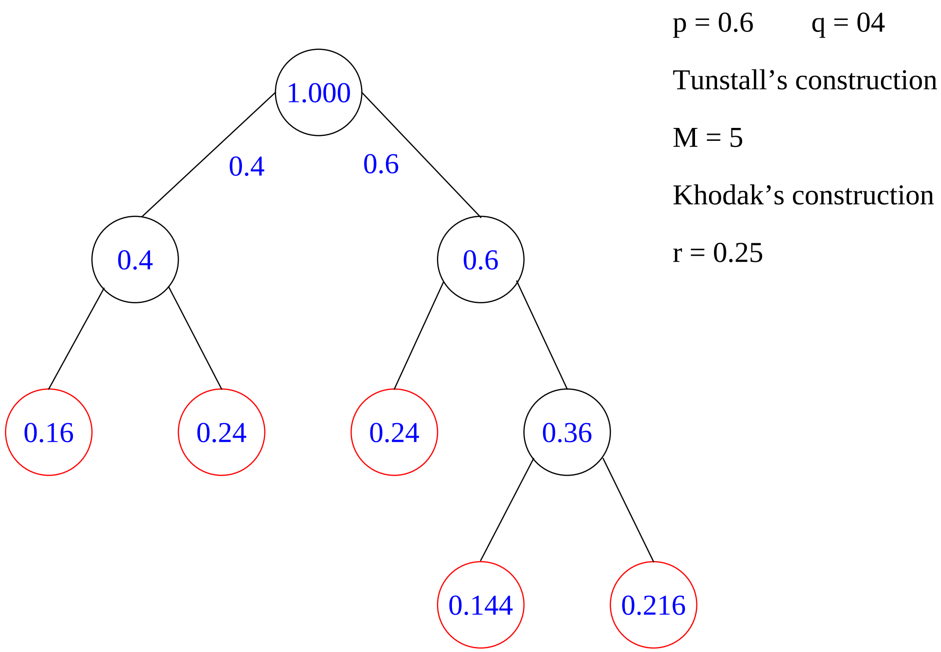
4. The encoder represents phrases by the fixed length binary codewords. i.e., a dictionary $\mathcal{D}$ of M entries requires $\lceil \log_2 M \rceil$ bits to represent entries.

Average Redundancy Rate:

$$\bar{r} = \lim_{n\to\infty} \frac{\sum_{|x|=n} P_S(x)(L(x) + \log P_S(x))}{n} = \frac{\log M}{\mathbf{E}[D]} - h$$

where h is the entropy rate of the source.

Tunstall and Khodak Codes



Tunstall Code:

1. Start with a root and leaves.

2. In the J's iteration select a leaf with the **highest probability** and grow children out it.

3. At Jth step, the parsing tree has J internal nodes and M=J + 1 leaves corresponding to dictionary entries.

Khodak Construction:

1. Pick a real number r and grow a complete parsing tree satisfying

$$\min\{p, 1-p\} \cdot r \leq P(d) < r\,, \quad d \in \mathcal{D}.$$

2. The resulting parsing tree is exactly the same as the Tunstall tree.

3. If y is a proper prefix of entries of $\mathcal{D}_r$, i.e., y is an internal node of $\mathcal{T}_r$, then

$$P(y) \geq r.$$

Phrase Length

We study the phrase length $D = |d|$, i.e., path length in the parsing tree.

Moment Generating Functions: Define

$$D(r, z) := \mathbf{E}[z^D] = \sum_{d \in \mathcal{D}_r} P(d) z^{|d|}.$$

and its corresponding internal nodes generating function

$$S(r, z) = \sum_{y:\ P(y) \geq r} P(y) z^{|y|}.$$

Simple Fact on Trees: Let $\tilde{\mathcal{D}}$ be a dictionary (leaves of $\mathcal{T}$) and $\tilde{\mathcal{Y}}$ be the collection of proper prefixes of dictionary entries (internal nodes of $\mathcal{T}$).

$$\sum_{d \in \tilde{\mathcal{D}}} P(d) \frac{z^{|d|} - 1}{z - 1} = \sum_{y \in \tilde{\mathcal{Y}}} P(y) z^{|y|}.$$

Thus

$$D(r, z) = 1 + (z - 1) S(r, z),$$

and

$$\mathbf{E}[D] = \tilde{S}(v, 1) = \sum_{y \in \tilde{\mathcal{Y}}} P(y), \qquad \mathbf{E}[D(D-1)] = \tilde{S}'(v, 1) = 2 \sum_{y \in \tilde{\mathcal{Y}}} P(y)|y|.$$

Recurrences

Define $v = 1/r$, z complex, and $\tilde{S}(v, z) = S(v^{-1}, z)$.

Let

$$A(v) = \sum_{y: P(y) \geq 1/v} 1$$

be the # of strings of probab. $\leq v^{-1}$ or the # of internal nodes. In fact:

$$M_r = A(v) + 1.$$

We have

$$A(v) = \begin{cases} 0 & v < 1, \\ 1 + A(vp) + A(vq) & v \geq 1 \end{cases}$$

and

$$\tilde{S}(v, z) = \begin{cases} 0 & v < 1, \\ 1 + zp\tilde{S}(vp, z) + zq\tilde{S}(vq, z) & v \geq 1, \end{cases}$$

since every binary string either is:

- – empty string,
- – string starting with the first symbol
- – string starting with second symbol.

Mellin Transform

The Mellin transform $F^*(s)$ of a function $F(v)$ is

$$F^*(s) = \int_0^\infty F(v) v^{s-1} dv.$$

From the recurrence on $S(v, z)$ we find

$$\tilde{D}^*(s, z) = \frac{1 - z}{s(1 - zp^{1-s} - zq^{1-s})} - \frac{1}{s}, \qquad \Re(s) < s_0(z),$$

where $s_0(z)$ denotes the real solution of: $zp^{1-s} + zq^{1-s} = 1$.

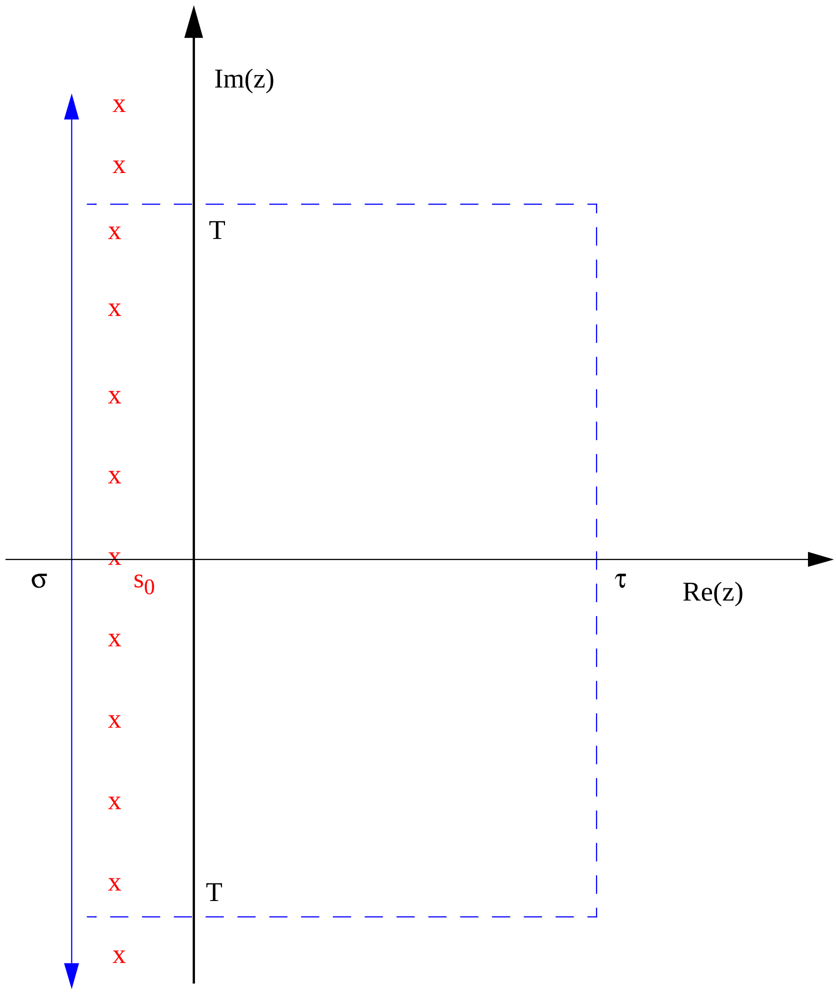
To find the asymptotics of $\tilde{D}(v, z)$ as $v \to \infty$ we compute the inverse transform of $\tilde{D}^*(s, z)$):

$$\tilde{D}(v, z) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{\sigma - iT}^{\sigma + iT} \tilde{D}^*(s, z) v^{-s}\, ds,$$

where $\sigma < s_0(z)$.

To determine the polar singularities of the meromorphic continuation of $\tilde{D}^*(s, z)$, we have to analyze the set

$$\mathbb{Z}(z) = \{s \in \mathbb{C} : zp^{1-s} + zq^{1-s} = 1\}.$$



From

$\tilde{D}(v,$

$F_T(v,$

$= -\sum$

$+\frac{1}{2\pi i}\int$

$= -\sum$

$+\frac{1}{2\pi i}\int$

provided that the series of residues converges and the last integral exists. **But they don't!**.

Tauberian Rescue

Therefore, we analyze (as in analytic number theory; cf. also Vallee)

$$\tilde{D}_1(v,z) = \int_0^v \tilde{D}(w,z)\,dw.$$

whose Mellin transform is

$$\tilde{D}_1^*(s,z) = \frac{-\tilde{D}^*(s+1,z)}{s} = O(1/s^2).$$

Lemma 3 (Tauberian). *Let $f(v,\lambda)$ be a non-negative increasing function such that*

$$F(v,\lambda) = \int_0^v f(w,\lambda)\,dw$$

and has the asymptotic expansion

$$F(v,\lambda) = \frac{v^{\lambda+1}}{\lambda+1}(1+\lambda\cdot o(1))$$

as $v \to \infty$ and uniformly in λ. Then as $v \to \infty$ uniformly in λ

$$f(v,\lambda) = v^{\lambda}(1+|\lambda|^{\frac{1}{2}}\cdot o(1)).$$

Main Results

Theorem 2 (Central Limit Theorem). *For large* M_r

$$\frac{D_r - \frac{1}{H}\ln M_r}{\sqrt{\left(\frac{H_2}{H^3} - \frac{1}{H}\right)\ln M_r}} \to N(0,1) \quad \text{standard normal distribution}$$

where H *is natural entropy and* $H_2 = p\ln^2 + q\ln^2 q$.

If $\ln q/\lg p$ *is irrational, then*

$$\begin{aligned} M_r &= A(v) + 1 = \frac{v}{H} + o(v) \\ \mathbf{E}[D_r] &= \frac{\ln M_r}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + o(1); \end{aligned}$$

if $\ln q/\lg p$ *is rational, then*

$$\begin{aligned} M_r &= \frac{Q_1(\log v)}{H} v + O(v^{1-\eta}), \qquad Q_1(x) = \frac{L}{1-e^{-L}} e^{-L\langle \frac{x}{L} \rangle}, \\ \mathbf{E}[D_r] &= \frac{\ln M_r}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + \frac{-\ln L + \ln(1-e^{-L}) + \frac{L}{2}}{H} + O({M_r}^{-\eta}), \end{aligned}$$

L *largest real number s.t.* $\ln(1/p)$ *and* $\ln(1/q)$ *are integer multiples of* L.

Redundancy Rate

The average redundancy rate of Tunstall/Khodak's VF code is defined as

$$\bar{r}_{M_r} = \frac{\ln M_r}{\mathbf{E}[D]} - h.$$

Case 1: $\ln p/\ln q$ is irrational:

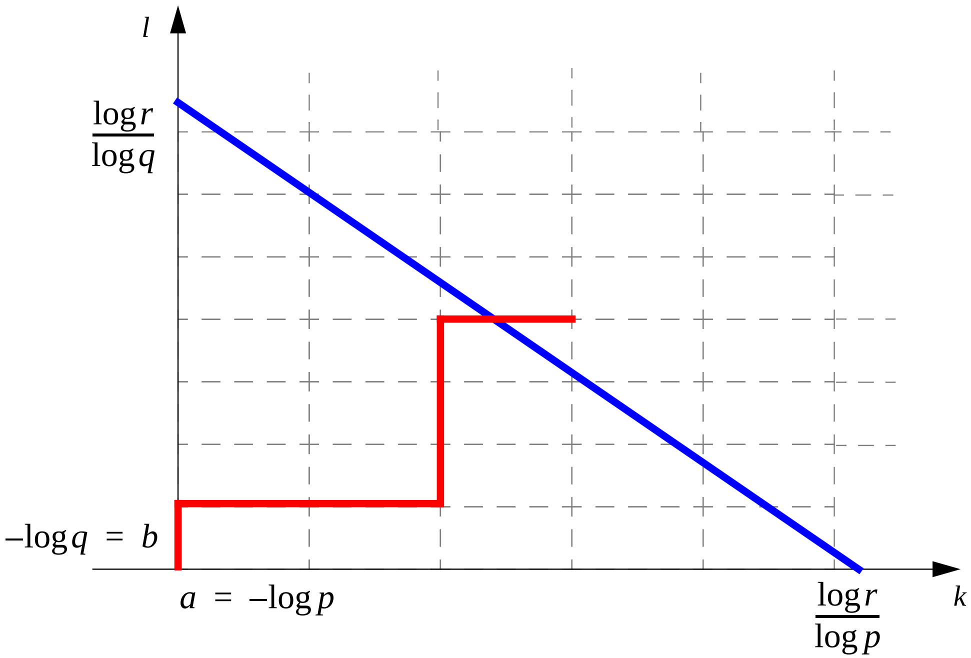
$$\bar{r}_{M_r} = \frac{H}{\ln M_r}\left(-\frac{H_2}{2H} - \ln H\right) + o\left(\frac{1}{\ln M_r}\right).$$

Case 2: $\ln p/\ln q$ is rational:

$$\bar{r}_{M_r} = \frac{H}{\ln M_r}\Big(- \frac{H_2}{2H} - \ln H + \ln L - \ln(e^L - 1) + \frac{L}{2}\Big) + O\left(M_r{}^{-\eta}\right),$$

for some $\eta > 0$, where $L > 0$ is the largest real number for which $\ln(1/p)$ and $\ln(1/q)$ are integer multiples of L. (No oscillation!)

Random Walk



Consider a random walk that corresponds to a path in the associated parsing tree.

For Khodak's code we studied

$$A(v) = \sum_{y:P(y)\geq 1/v} f(v)$$

for some function $f(v)$.

But $P(y) = p^k q^l$ $(k, l \geq 0)$, and set $v = 2^V$ so that

$$\log P(v) = k \lg(1/p) + sl \lg(1/q) \leq V.$$

This corresponds to a random walk in the first quadrant with the linear boundary condition

$$ax + by = V$$

where $a = \log(1/p)$ and $b = \log(1/q)$.

The phrase length coincides with the exit time of such a random walk.

Outline Update

1. Source Coding
2. Preliminary Facts of Source Coding
3. Redundancy of Huffman Code
4. Redundancy of Tunstall and Khodak VF Codes
5. Redundancy of Khodak VV Code
6. Concluding Remarks

VV Codes and Their Redundancy

Variable-to-variable (VV) code:

- A VV encoder consists of a parser and a string encoder.
- Parser works as in the VF code.
- String encoder encodes dictionary into a variable length prefix codes.

Average Redundancy Rate. By renewal theory;

$$\bar{r} = \lim_{n\to\infty} \frac{\sum_{|x|=n} P_{\mathcal{S}}(x)(L(x) + \log P_{\mathcal{S}}(x))}{n} = \frac{\sum_{d\in\mathcal{D}} P_{\mathcal{D}}(d)(\ell(d) + \log P_{\mathcal{D}}(d))}{\mathbf{E}[D]}$$

$$= \frac{\sum_{d\in\mathcal{D}} P_{\mathcal{D}}(d)\ell(d) - h_{\mathcal{D}}}{\mathbf{E}[D]}$$

where $P := P_{\mathcal{D}}$ is the dictionary distribution.

Observe that by the Conversation of Entropy Theorem:

$$h_{\mathcal{D}} = h_{S}\mathbf{E}[D].$$

Main Result

Theorem 3. *For memoryless or Markov sources, there exists a VV code such that its average redundancy satisfies*

$$\bar{r} = O(\bar{D}^{-5/3}).$$

where $\bar{D} := \mathbf{E}[D]$*. The redundancy rate* **decays faster** *than linearly.*

Main Result

Theorem 3. *For memoryless or Markov sources, there exists a VV code such that its average redundancy satisfies*

$$\bar{r} = O(\bar{D}^{-5/3}).$$

where $\bar{D} := \mathbf{E}[D]$. The redundancy rate **decays faster** *than linearly.*

Main Idea Behind the Proof:
An m-ary alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ with probabilities $p_1, \ldots, p_m$:

$$P(d) = p_1^{k_1} \cdots p_m^{k_m}, \quad d \in \mathcal{D}$$

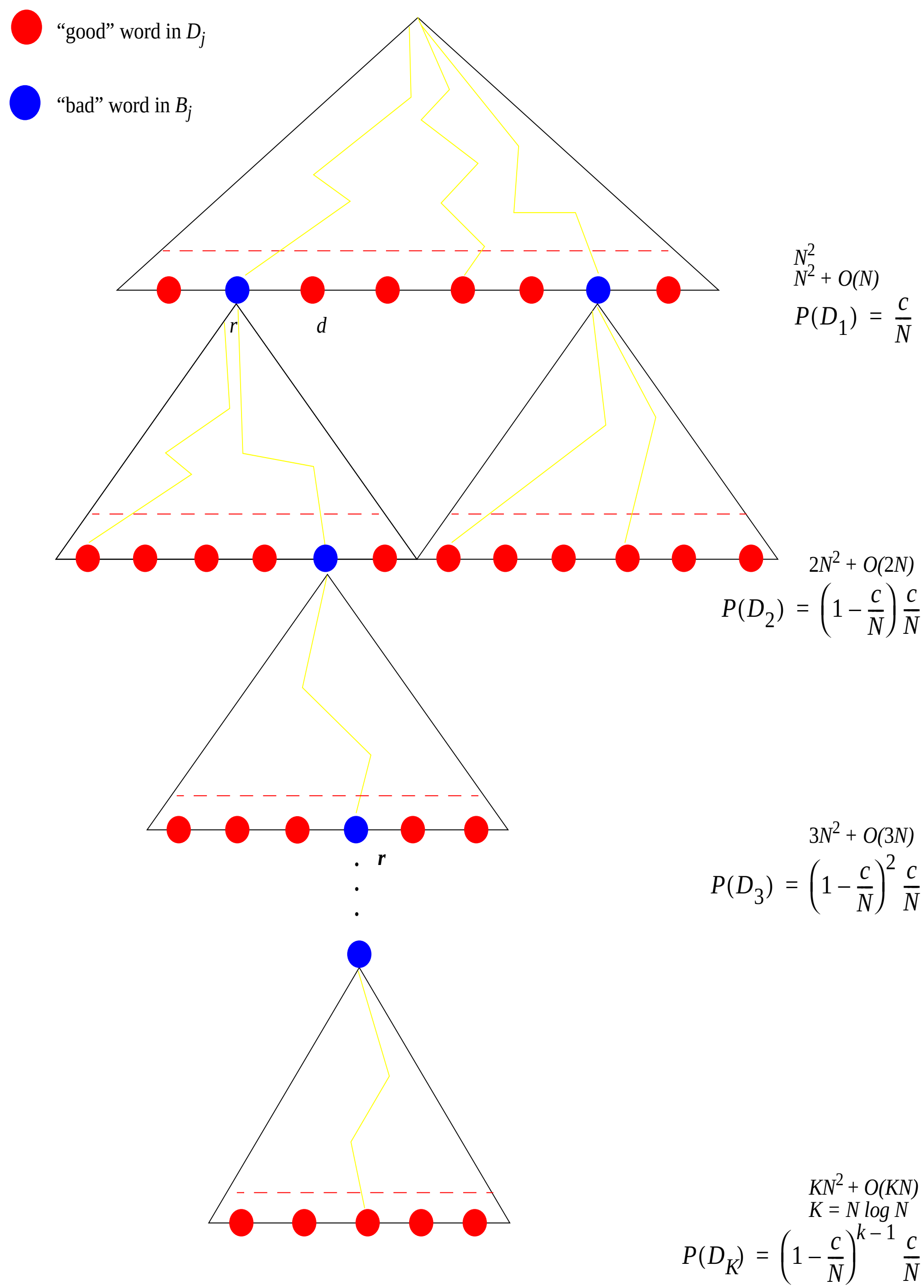
where k_i is the number of a_i in the phrase d.

Modified Shannon code assigns the length $\lceil -\log P(d) \rceil$, when $\log P(d)$ is close to an integer. The average redundancy is

$$\bar{R} = \sum_{d \in \mathcal{D}} P(d)[\lceil -\log P(d) \rceil + \log P(d)] = \sum_{d \in \mathcal{D}} P(d) \cdot \langle k_1 \log p_1 + \cdots + k_m \log p_m \rangle.$$

Khodak code is a prefix code with $\langle k_1 \log p_1 + \cdots + k_m \log p_m \rangle$ as close as possible to an integer.

Proof by Picture



Sketch of Proof

Some Definitions:
Define dispersion of the set $X \subseteq [0, 1)$ as

$$\delta(X) = \sup_{0 \le y < 1} \inf_{x \in X} \|y - x\|, \qquad \|x\| = \min(\langle x \rangle, \langle 1 - x \rangle),$$

that is, for every $y \in [0, 1)$ there is $x \in X$ s.t. $\|y - x\| \le \delta(X)$.
Lemma 4. (i) *Let γ be irrational. There exists an integer N such that*

$$\delta\left(\{\langle k\gamma \rangle : 0 \le k < N\}\right) \le \frac{2}{N}$$

(ii) *In general, there exists an integer N such that the dispersion of*

$$X = \{\langle k_1\gamma_1 + \cdots + k_m\gamma_m \rangle : 0 \le k_j < N\ (1 \le j \le m)\}$$

is bounded by

$$\delta(X) \le \frac{2}{N}$$

provided one of γ_i is irrational.

Two Important Lemmas

Lemma 5 (Khodak, 1972). *For every $d \in \mathcal{D}$ the length ℓ_d satisfies $|\ell_d + \log_2 P(d)| \le 1$. If*

$$\sum_{d \in \mathcal{D}} P(d)(\ell_d + \log_2 P(d)) \ge 2 \sum_{d \in \mathcal{D}} P(d)(\ell_d + \log_2 P(d))^2$$

then there exists prefix code with lengths ℓ_d for $\mathcal{D}$.

Proof: Kraft's inequality and Taylor's expansion.

Main result follows from the below Theorem after setting $\eta = 1$.

Theorem 4. *Suppose that for some $N \ge 1$ and $\eta \ge 1$ the set*

$$X = \{\langle k'_1 \log_2 p_1 + \cdots + k'_m \log_2 p_m \rangle : 0 \le k'_j < N\}$$

has dispersion

$$\delta(X) \le \frac{2}{N^\eta}.$$

Then there exists a variable-to-variable code (with $\bar{D} = \Theta(N^3)$) such that

$$\bar{r} \le c_m \cdot \bar{D}^{-\frac{4+\eta}{3}}.$$

Sketch of Proof for Theorem 4

1. Set $k_i^0 := \lfloor p_i N^2 \rfloor$ $(1 \le i \le m)$ and $x = k_1^0 \log_2 p_1 + \cdots + k_m^0 \log_2 p_m$. By Theorem 4 there exist integers $0 \le k_j^1 < N$ such that

$$\left\langle x + k_1^1 \log_2 p_1 + \cdots + k_m^1 \log_2 p_m \right\rangle < \frac{4}{N^\eta}.$$

2. Build an m-ary tree starting at the root with
$k_1^0 + k_1'$ edges of type 1, and $k_m^0 + k_m'$ edges of type m.
Let $\mathcal{D}_1$ denote the set of the corresponding words whose probability is

$$\frac{c'}{N} \le P(\mathcal{D}_1) \quad = \quad \binom{(k_1^0 + k_1') + \cdots + (k_m^0 + k_m')}{k_1^0 + k_1', \ldots, k_m^0 + k_m'} p_1^{k_1^0 + k_1'} \cdots p_m^{k_m^0 + k_m'} \le \frac{c''}{N}$$

for certain positive constants c', c''. Thus, all words $d \in \mathcal{D}_1$ satisfy

$$\langle \log_2 P(d) \rangle < \frac{4}{N^\eta},$$

and have the same length $N^2 + O(N)$.

3. Consider words not in $\mathcal{D}_1$, that is, $\mathcal{B}_1 = A^{n_1} \setminus \mathcal{D}_1$ that by above satisfy

$$1 - \frac{c''}{N} \le P(\mathcal{B}_1) \le 1 - \frac{c'}{N}.$$

Finishing ...

4. Take a word $r \in \mathcal{B}_1$ and concatenate it with a word d_2 of length $\sim N^2$ such that $\log_2 P(rd_2)$ is close to an integer with high probability.

5. This construction is cut after $K = O(N \log N)$ steps so that

$$P(\mathcal{B}_K) \le c'' \left(1 - \frac{c'}{N}\right)^K \le \frac{1}{N^\beta}$$

for some $\beta > 0$. This also ensures that $P(\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K) > 1 - frac1N^\beta$.

6. The complete prefix free set $\mathcal{D}$ is

$$\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K \cup \mathcal{B}_K.$$

By the construction the average delay of $\mathcal{D}$ is

$$c_1 N^3 \le \bar{D} = \sum_{d \in \mathcal{D}} P(d)\, |d| \le c_2 N^3$$

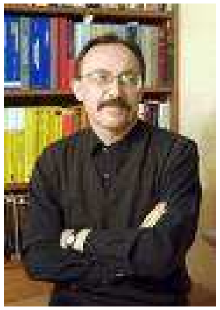
whilethe maximal code length satisfies

$$\max_{d \in \mathcal{D}} |d| = O N^3 \log N = O \bar{D} \log \bar{D}).$$

Outline Update

1. Source Coding
2. Preliminary Facts of Source Coding
3. Redundancy of Huffman Code
4. Redundancy of Tunstall and Khodak VF Codes
5. Redundancy of Khodak VV Code
6. Concluding Remarks
   (a) Universal Shannon Code
   (b) Non-Prefix Codes: One-to-One Codes





Universal Shannon-Fano Code

Let $x_1^n = x_1 \dots x_n$ be a binary sequence generated by an **memoryless**(θ) source where θ is **unknown**. Clearly,

$$P(x_1^n) = \theta^k (1-\theta)^{n-k}, \qquad \theta \in (0,1),$$

where k is the number of "1" and $n-k$ is the number of "0".

Estimating θ:
We need to estimate θ to apply a FV code. In the spirit of the Minimum Description Length, we use the Krichevsky and Trofimov (KT) estimator:

$$P_e(x_n = 1 | x_1^{n-1}) = \frac{k + \frac{1}{2}}{n+1}.$$

which can be written as

$$P_e(k, n-k) = \frac{\frac{1}{2} \cdots (k - \frac{1}{2}) \cdot \frac{1}{2} \cdots (n-k-\frac{1}{2})}{1 \cdot 2 \cdot n} = \frac{\Gamma(k+1/2)\Gamma(n-k+1/2)}{\pi \Gamma(n+1)}$$

where $\Gamma(x)$ is the Euler gamma functions.

Average Redundancy

Shannon-Fano code assigns code length: $\lceil \log P_e(k, n-k) \rceil + 1$. Then

$$\bar{R}_n^{SF} = 2 + \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \log \frac{\theta^k (1-\theta)^{n-k}}{P_e(k, n-k)} - E_n,$$

where $E_n = \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \langle -\log P_e(k, n-k) \rangle$.

Theorem 5. *The average redundancy of the Shannon-Fan code is*

$$\bar{R}_n^{SF} = \frac{1}{2} \log n - \frac{1}{2} \log \frac{\pi e}{2} + 2 - E_n + O(n^{-1/2})$$

where E_n behavior depends whether $\alpha = \log \frac{1-\theta}{\theta}$ is rational or not.
If $\alpha = \frac{N}{M}$ is rational, then

$$E_n = \frac{1}{2} + G_M \left(-\log(1-\theta) n + \frac{1}{2} \log \frac{\pi n}{2} \right) + o(1)$$

where $G_M(y)$ is a periodic function with period $\frac{1}{M}$, and

$$G_M(y) := \frac{1}{M} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \left(\left\langle M \left(y - \frac{x^2}{2 \ln 2} \right) \right\rangle - \frac{1}{2} \right) dx$$

Outline Update

1. Source Coding
2. Preliminary Facts of Source Coding
3. Redundancy of Huffman Code
4. Redundancy of Tunstall and Khodak VF Codes
5. Redundancy of Khodak VV Code
6. Concluding Remarks
   (a) Universal Shannon Code
   (b) Non-Prefix Codes: One-to-One Codes

Non-Prefix Codes: One-to-One Codes

In one-to-one codes a distinct codeword is assigned to each source symbol (unique decodability is not required).

One-to-One codes are **not prefix codes**; Kraft's inequality doesn't hold.

Question: **Does the lower bound still holds?** That is, $\mathbf{E}[L] - H(X) \geq 0$?.

Wyner in 1972 proved that

$$\mathbf{E}[L] \leq H(X),$$

further improved by Alon and Orlitsky who showed

$$\mathbf{E}[L] \geq H(X) - \log(H(X) + 1) - \log e.$$

We consider a FV one-to-one code for $x_1^n = x_1 \ldots x_n \in \mathcal{A}^n$ generated by a memoryless source with p being the probability of generating a 0 and $q = 1 - p$.

Throughout: $p \leq q$ so that $P(x_1^n) = p^k q^{n-k}$.

Average Code length

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

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

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

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

$$\mathbf{E}[L_n] = \sum_{k=0}^{n} p^k q^{n-k} \sum_{i=1}^{\binom{n}{k}} \lfloor \log_2(A_{k-1} + i) \rfloor.$$

We shall study (below $\langle x \rangle = x - \lfloor x \rfloor$)

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

Main Result

Theorem 6. *For a binary memoryless source, let* $p < \frac{1}{2}$. *Then*

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

where • $G(n) = F(n) = 0$ *if* $\log_2 \frac{1-p}{p}$ *irrational;*

• $G(n)$ *and* $F(n)$ *are oscillating functions if* $\log_2 \frac{1-p}{p} = N/M$ *rational:*

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

where $\beta = -\log_2(1-p)$.

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

$$\bar{R}_n = -1 + 2^{-n}(n-2).$$

Analytic Information Theory

- 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 length description.

- To **overcome** these difficulties we propose replacing first-order analyses by full asymptotic expansions and more accurate analyses (e.g., large deviations, central limit laws).

- Following **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 path between two truths on the real line passes through the complex plane.

THANK YOU!