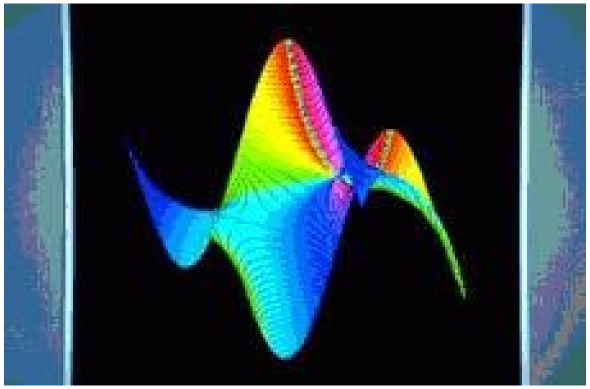
Variable-to-Variable Codes with Small Redundancy Rates*

M. Drmota[†] W. Szpankowski[‡]

September 25, 2004



*This research is supported by NSF, NSA and NIH.

[†]Institut f. Diskrete Mathematik und Geometrie, TU Wien, Austria

[‡]Department of Computer Science, Purdue University, U.S.A.

Outline of the Talk

1. Types of Codes
2. Redundancy and Redundancy Rates
3. Main Results
4. Sketch of Proof
5. Algorithm

Types of Codes

Fixed-to-Variable Code:

Fixed length blocks are mapped into variable-length binary code strings.
Example: Shannon and Huffman codes.

Variable-to-Fixed Code:

- Encoder partitions strings into phrases.
- Phrases belong to dictionary $\mathcal{D}$ (complete tree).
- The encoder represents dictionary string by fixed-length code word.

Example: Tunstall code.

Variable-to-variable (VV) code:

- Encoder consists of a parser and a string encoder.
- The parser works as in the VF code (it partitions a sequence into phrases).
- The string encoder encodes dictionary strings into codeword $C(d)$ of length $|C(d)| = \ell(d)$.

Redundancy and Redundancy Rates

Let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be the input alphabet of $m \geq 2$ symbols with probabilities $p_1, \ldots, p_m$.

A source $\mathcal{S}$ generates a sequence x with the underlying probability $P_\mathcal{S}$.

The average and worst case (maximal) **redundancy** of a fixed-to-variable code are defined, respectively, as

$$\begin{aligned}\bar{R} &= \sum_{x \in \mathcal{A}^n} P_\mathcal{S}[L(x) + \log P_\mathcal{S}(x)] \\ R^* &= \max_{x \in \mathcal{A}^n} [L(x) + \log P_\mathcal{S}(x)]\end{aligned}$$

where $L(x)$ is the code length assigned to $x \in \mathcal{A}^n$.

Redundancy rates are respectively

$$\begin{aligned}\bar{r} &= \frac{\bar{R}}{n} \\ r^* &= \frac{R^*}{n}.\end{aligned}$$

Redundancy Rates for VV Codes

Let $P := P_{\mathcal{D}}$ be the probability induced by the dictionary $\mathcal{D}$.

Define the average delay $\bar{D}$ as

$$\bar{D} = \sum_{d \in \mathcal{D}} P_{\mathcal{D}}(d)|d|.$$

The (asymptotic) average redundancy rate $\bar{r}$ is

$$\bar{r} = \lim_{n\to\infty} \frac{\sum_{|x|=n} P_{\mathcal{S}}(x)(L(x) + \log P_{\mathcal{S}}(x))}{n}.$$

Using renewal theory (regeneration theory) we find

$$\lim_{n\to\infty} \frac{\sum_{|x|=n} P_{\mathcal{S}}(x)L(x)}{n} = \frac{\sum_{d\in\mathcal{D}} P_{\mathcal{D}}(d)\ell(d)}{\bar{D}}.$$

where $\ell(d)$ is the length of the phrase $d \in \mathcal{D}$.

A New Definition

Denote

- $H_{\mathcal{S}}$ the binary source entropy,
- $H_{\mathcal{D}}$ the dictionary entropy.

Then

$$\frac{\sum_{d\in\mathcal{D}} P_{\mathcal{D}}(d)\ell(d)}{\bar{D}} - H_{\mathcal{S}} \;=\; \frac{H_{\mathcal{D}} + \left(\sum_{d\in\mathcal{D}} P_{\mathcal{D}}(d)\ell(d) - H_{\mathcal{D}}\right)}{\bar{D}} - H_{\mathcal{S}}.$$

By the Conservation of Entropy Theorem (Savari 99)

$$H_{\mathcal{D}} \;=\; H_{\mathcal{S}} \cdot \bar{D},$$

hence

$$\bar{r} \;=\; \frac{\sum_{d\in\mathcal{D}} P_{\mathcal{D}}(d)\ell(d) - H_{\mathcal{D}}}{\bar{D}}$$

which we adopt as our definition of the average redundancy rate.

Maximum Redundancy

Maximum (asymptotic) redundancy rate r^*

$$r^* = \lim_{|x|\to\infty} \frac{\max_x [L(x) + \log P_{\mathcal{S}}(x)]}{|x|}.$$

Assuming the source is sequence is partitioned into phrases $x^1, \ldots, x^k$, $k \longrightarrow \infty$ we have

$$\begin{aligned}
r^* &= \lim_{k\to\infty} \frac{\max_{x^1,\ldots,x^k} \sum_{i=1}^k [\ell(x^i) + \log P_{\mathcal{S}}(x^i)]}{\sum_{i=1}^k |x^i|} \\
&= \lim_{k\to\infty} \frac{\sum_{i=1}^k \max_{x^i} [\ell(x^i) + \log P_{\mathcal{S}}(x^i)]}{\sum_{i=1}^k |x^i|} \\
&= \lim_{k\to\infty} \frac{\sum_{i=1}^k \max_{d\in\mathcal{D}} [\ell(d) + \log P(d)]}{\sum_{i=1}^k |x^i|} \\
&= \lim_{k\to\infty} \frac{k \max_{d\in\mathcal{D}} [\ell(d) + \log P(d)]}{\sum_{i=1}^k |x^i|} \\
&= \frac{\max_{d\in\mathcal{D}} [\ell(d) + \log P(d)]}{\bar{D}} \qquad (a.s.).
\end{aligned}$$

Main Result for Average Redundancy

Theorem 1. *Let $m \geq 2$ and $\mathcal{S}$ a memoryless or a Markov source. There exists a variable-to-variable code such that its average redundancy satisfies*

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

There also exists a variable-to-variable code such that the worst case redundancy satisfies

$$r^* = O(\bar{D}^{-4/3}),$$

however, the maximal code length might be infinite.

The estimate for $\bar{r}$ for the memoryless source is the same as in Khodak's 1972 paper. However, the method presented in Khodak's paper is difficult to follow and it is not clear one can construct a VV code.

Main Result for Maximal Redundancy

Theorem 2. *Under the same assumption as in Theorem 1.*
For almost all source parameters p_j (resp. p_{ij}):

- *There exists a variable-to-variable code such that its average redundancy is bounded by*

$$\bar{r} \le \bar{D}^{-\frac{4}{3}-\frac{m}{3}+\varepsilon},$$

  where $\varepsilon > 0$ and the maximal length is $O(\bar{D} \log \bar{D})$.

- *There also exists a variable-to-variable code with worst case redundancy*

$$r^* \le \bar{D}^{-1-\frac{m}{3}+\varepsilon}$$

  for $\varepsilon > 0$.

Lower Bound

Theorem 3. *For every variable-to-variable code and almost all parameters* p_j, *the following lower bound holds*

$$r^* \geq \bar{r} \geq \bar{D}^{-2m-1-\varepsilon}.$$

for $\varepsilon > 0$

Some comments:

1. Typically the best possible average and worst case redundancy are measured in terms of negative powers of $\bar{D}$ that linearly decrease in term of the alphabet size m.
2. It seems to be very difficult to obtain the optimal exponent (almost surely). Nevertheless the bounds we are obtain are best possible with respect to the methods we use.
3. Note that Theorem 4 of Kodak 1972 paper states a lower bound of for the redundancy of the form $\bar{r} \geq \bar{D}^{-9}(\log \bar{D})^{-8}$ (for almost all memoryless sources). In view of our Theorem 3 this cannot be true for large m.

Rough Idea ...

Let $\mathcal{A} = \{1, 2, \ldots, m\}$ and $p_1 + \cdots p_m = 1$.
By k_i we denote the number of symbols i in a word x.

For the Shannon code the redundancy is

$$\begin{aligned} \bar{R} &= \sum_x \left(\lceil - \log p_1^{k_1} \cdots p_m^{k_m} \rceil + \log p_1^{k_1} \cdots p_m^{k_m} \right) \\ &= 1 - \sum_x \langle -k_1 \log p_1 - \cdots - k_m \log p_m \rangle \\ &= \sum_x \langle k_1 \log p_1 + \cdots + k_m \log p_m \rangle \end{aligned}$$

where $\langle a \rangle = a - \lfloor a \rfloor$ is the fractional part of a.

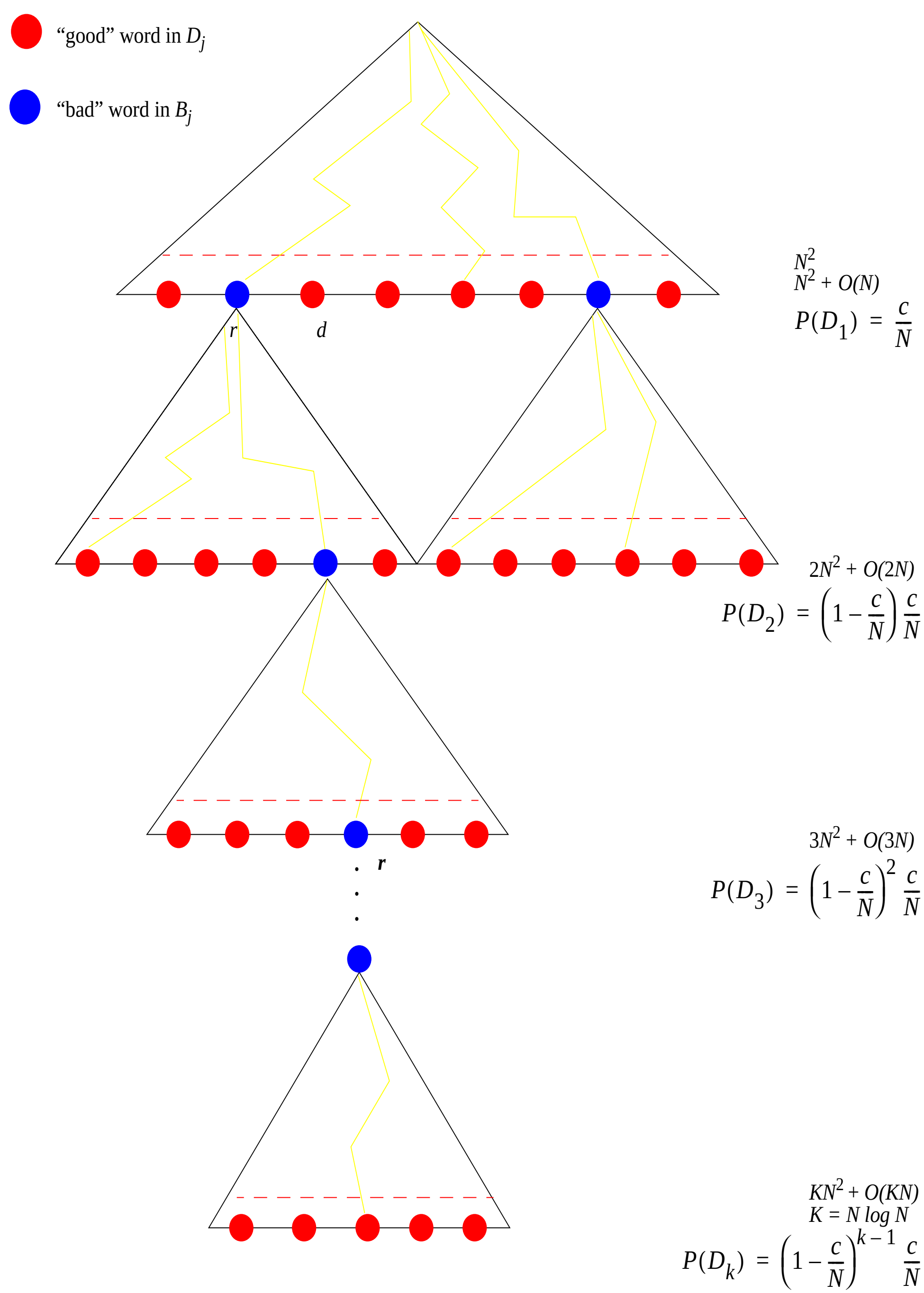
Basic Thrust of our Approach

In order to minimize the redundancy of a code we must find such $k_1, \ldots, k_m$ that

$$\langle k_1 \log p_1 + \cdots + k_m \log p_m \rangle$$

is as close as possible to an integer.

Proof by Picture



Algorithm

Input:

- m, an integer ≥ 2,
- positive rational numbers $p_1, \ldots, p_m$ with $p_1 + \cdots + p_m = 1$, p_m is not a power of 2
- $\varepsilon < 1$, a positive real number.

Output:

- A VV-code (given by a dictionary $\mathcal{D}$, a complete prefix free set on an m-ary alphabet and by a prefix code $C : \mathcal{D} \to \{0,1\}^*$) with the average redundancy
$$\overline{r} \leq \varepsilon / \overline{D}$$
where the average dictionary code length $\overline{D}$ satisfies
$$\overline{D} \geq c(m, p_1, \ldots, p_m)/\varepsilon^3.$$

Algorithm

1. Calculate a convergent $\frac{M}{N} = [c_0, c_1, \ldots, c_n]$ of $\log_2 p_m$ for which $N > 4/\varepsilon$.
2. Set $k_j^0 = \lfloor p_j N^2 \rfloor$, $x = \sum_{j=1}^m k_j^0 \log_2 p_j$, $n_0 = \sum_{j=1}^m k_j^0$.
3. Set $\mathcal{D} = \emptyset$, $\mathcal{B} = \{$empty word$\}$, and $p = 0$

   while $p < 1 - \varepsilon/4$ do

   Take $r \in \mathcal{B}$ of minimal length

   $b \leftarrow \log_2 P(r)$

   Determine $0 \leq k < N$ that solves

   $kM \equiv 1 - \lfloor (x+b)N \rfloor \bmod N$

   (i.e., $1/N \leq \langle kM/N + x + b \rangle \leq 2/N$)

   $n \leftarrow n_0 + k$

   $\mathcal{D}' \leftarrow \{d \in A^n : \mathrm{type}(d) = (k_1^0, \ldots, k_m + k)\}$

   $\mathcal{D} \leftarrow \mathcal{D} \cup r \cdot \mathcal{D}'$

   $\mathcal{B} \leftarrow (\mathcal{B} \setminus \{r\} \cup r \cdot (A^n \setminus \mathcal{D}')$

   $p \leftarrow p + P(r)P(\mathcal{D}')$, where

$$P(\mathcal{D}') = \frac{n!}{k_1^0! \cdots k_{m-1}^0!(k_m^0 + k)!} p_1^{k_1^0} \cdots p_{m-1}^{k_{m-1}^0} p_m^{k_m^0 + k}.$$

   end while

4. $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{B}$
5. Construct a Shannon code $\ell(d) = \lceil -P(d) \rceil$.

Some Definitions Needed in the Proof

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

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

where $\|x\| = \min(\langle x \rangle, \langle 1 - x \rangle)$. That is, for every $y \in [0,1)$ there exists $x \in X$ with $\|y - x\| \le \delta(X)$.

Lemma 1. *Suppose that γ is an irrational number. The there exists N such that*

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

Lemma 2. *Let $(\gamma_1, \ldots, \gamma_m) = (\log p_1, \ldots, \log p_m)$ an m-vector of real numbers such that at least one of its coordinates is irrational. There exists N such that the dispersion of the set*

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

is bounded by

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

Two Important Lemmas

Lemma 3. *Let $\mathcal{D}$ be a finite set with probability distribution P such that that for every $d \in \mathcal{D}$ the length l_d satisfies $|\ell_d + \log_2 P(d)| \leq 1$. If*

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

then there exists an injective mapping $C : \mathcal{D} \to \{0,1\}^*$ such that $C(\mathcal{D})$ is a prefix free set and $|C(d)| = l_d$ for all $d \in \mathcal{D}$.

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

Lemma 4. *Let $\mathcal{D}$ be a finite set with probability distribution P. Then*

$$\bar{r} \geq \frac{1}{2} \frac{1}{\bar{D}} \sum_{d \in \bar{D}} P(d) \| \log_2 P(d) \|^2,$$

for a certain constant $c > 0$.

Main Step of the Proof

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

$$X = \{\langle k_1' \log_2 p_1 + \cdots + k_m' \log_2 p_m \rangle : 0 \leq k_j' < N\}$$

has dispersion

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

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

$$\bar{r} \leq c_m' \cdot \bar{D}^{-\frac{4+\eta}{3}},$$

There exists also another variable-to-variable code with the the worst case redundancy bounded by

$$r^* \leq c_m'' \cdot \bar{D}^{-1-\frac{\eta}{3}},$$

where the constants $c_m', c_m'' > 0$ just depend on m.

Observe that above theorem and previous lemmas immediately implied our main result after setting $\eta = 1$.

Sketch of the Proof of Theorem 3

1. Set $k_i^0 := \lfloor p_i N^2 \rfloor$ $(1 \leq i \leq m)$ and
$x = k_1^0 \log_2 p_1 + \cdots + k_m^0 \log_2 p_m$.
By Theorem 3 there exist integers $0 \leq k_j^1 < N$ such that

$$\begin{aligned} & \left\langle x + k_1^1 \log_2 p_1 + \cdots + k_m^1 \log_2 p_m \right\rangle \\ = \; & \left\langle (k_1^0 + k_1^1) \log_2 p_1 + \cdots + (k_m^0 + k_m^1) \log_2 p_m \right\rangle \\ < \; & \frac{4}{N^\eta}. \end{aligned}$$

2. Build an m-ary tree starting at the root with
$k_1^0 + k_1'$ edges of type 1,
$k_2^0 + k_2'$ edges of type 2, . . ., and
$k_m^0 + k_m'$ edges of type m.
Let $\mathcal{D}_1$ denote the set of the corresponding words. Then

$$\begin{aligned} \frac{c'}{N} \leq P(\mathcal{D}_1) \; & = \; \binom{(k_1^0 + k_1') + \cdots + (k_m^0 + k_m')}{k_1^0 + k_1', \ldots, k_m^0 + k_m'} p_1^{k_1^0 + k_1'} \cdots p_m^{k_m^0 + k_m'} \\ & \leq \; \frac{c''}{N} \end{aligned}$$

for certain positive constants c', c''.

Constructing the Prefix Code

3. By construction, all words $d \in \mathcal{D}_1$ satisfy

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

and have the same length

$$n_1 = (k_1^0 + k_1') + \cdots + (k_m^0 + k_m') = N^2 + O(N).$$

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

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

Second Step ...

5. 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.

For every word $r \in \mathcal{B}_1$ we set

$$x(r) = \log_2 P(r) + k_1^0 \log_2 p_1 + \cdots + k_m^0 \log_2 p_m.$$

By condition of Theorem 3 again there exist integers $0 \leq k_j^2(r) < N$ $(1 \leq j \leq m)$ such that

$$\left\langle x(r) + k_1^2(r) \log_2 p_1 + \cdots + k_m^2(r) \log_2 p_m \right\rangle < \frac{4}{N^\eta}$$

Proving Some Properties ...

6. We continue extending the tree $\mathcal{T}$ by adding a path starting at $r \in \mathcal{B}_1$ with
$k_1^0 + k_1^2(r)$ edges of type 1,
$k_2^0 + k_2^2(r)$ edges of type 2,
. . ., and
$k_m^0 + k_m^2(r)$ edges of type m.
We observe that

$$P(r)\frac{c'}{N} \leq P(\mathcal{D}_2(r)) \leq P(r)\frac{c''}{N}.$$

Furthermore (by construction) we have

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

for all $d \in \mathcal{D}_2(r)$.

Last Step ...

7. 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

$$P(\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K) > 1 - \frac{1}{N^\beta}.$$

8. 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

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

Finishing the Proof ...

9. For every $d \in \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$ we can choose a non-negative integer ℓ_d with

$$|\ell_d + \log_2 P(d)| < \frac{2}{N^\eta}.$$

Indeed:

$$0 \leq \ell_d + \log_2 P(d) < \frac{2}{N^\eta}$$

if $\langle \log_2 P(d) \rangle < 2/N^\eta$ and

$$-\frac{2}{N^\eta} < \ell_d + \log_2 P(d) \leq 0$$

if $1 - \langle \log_2 P(d) \rangle < 2/N^\eta$.

For $d \in \mathcal{B}_K$ we set $\ell_d = \lceil -\log_2 P(d) \rceil$.

10. The final step is to verify that condition of Lemma 3 is satisfied, which is an easy step.