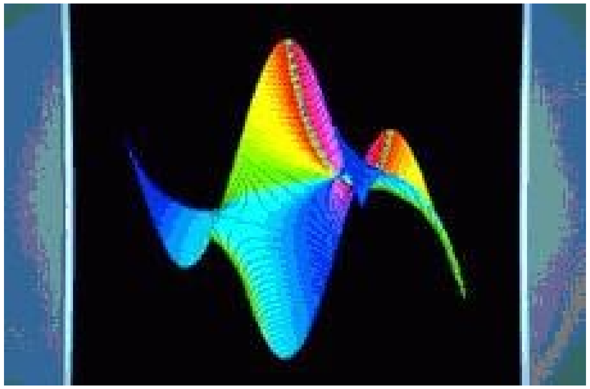
# Analysis of Some Variable-to-Fixed Codes by Analytic Methods

M. Drmota*, Y. Reznik† S. Savari‡, and W. Szpankowski§

February 4, 2006



*Institute of Discrete Mathematics and Geometry, TU Wien, Austria
†Qualcomm Inc., 5775 Morehouse Dr., San Diego
‡Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor
§Department of Computer Science, Purdue University

# Outline of the Talk

1. Tunstall and Khodak VF Codes
2. A Useful Lemma on Trees
3. Some Recurrences
4. Main Results on Tunstall's Code
5. Redundancy of Tunstall's Code
6. Oscillations in Redundancy Formulas

# Variable-to-Fixed Codes

Throughout, we consider an $m$-ary alphabet
$\mathcal{A} = \{1, 2, \ldots, m\}$

**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 $\mathcal{T}$, i.e., a tree in which every internal node has all $m$ children.
The dictionary entries $d \in \mathcal{D}$ correspond to the leaves of parsing tree.

**4**. The encoder represents each parsed string by the fixed length binary code word.
If the dictionary $\mathcal{D}$ is has $M$ entries, then the code word for each phrase has $\lceil \log_2 M \rceil$ bits.

# Tunstall's Code

Tunstall's construction works as follows (cf. Tunstall, Synthesis of Noiseless Compression Codes, Ph.D. 1968):

**1**. Start with a tree consisting of a root node and $m$ leaves.

**2**. In the $J$'s iteration we select the current leaf corresponding to a string of the **highest probability** and grow $m$ children out it.

**3**. After these $J$ steps, the parsing tree has $J$ non-root internal nodes and

$$M = (m - 1)J + m$$

leaves which correspond to distinct dictionary entries.

# Example

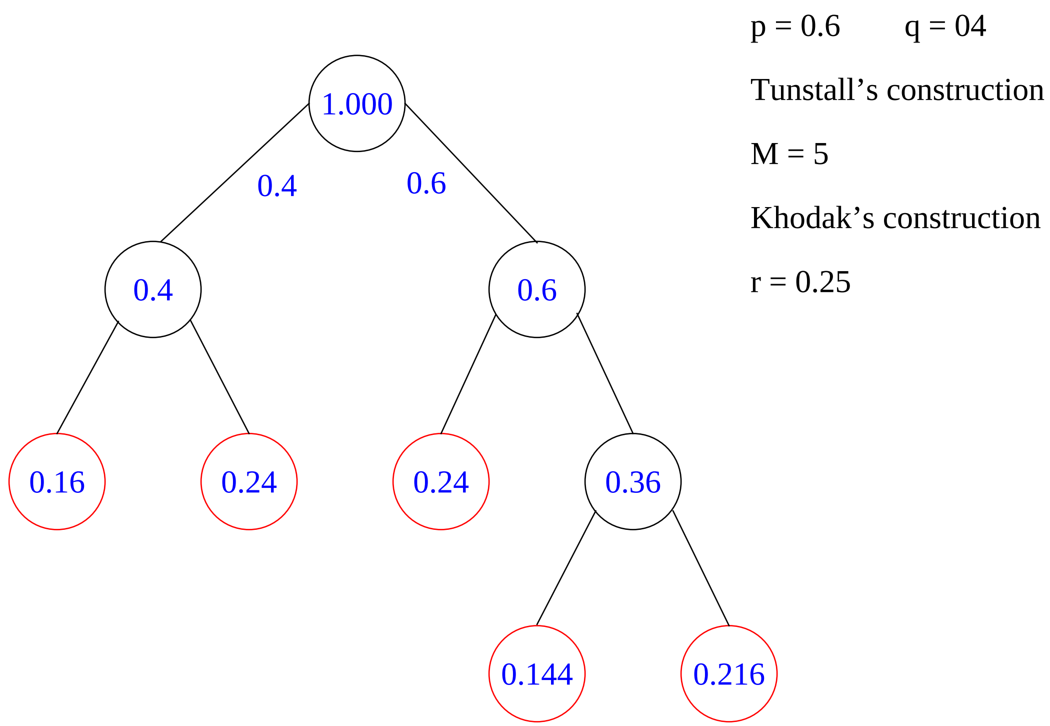



Figure 1: Tunstall's Code for $M = 5$ and Khodak's Construction for $r = 0.25$.

# Khodak's Construction

Khodak's construction of Tunstall's code (cf. Khodak, "Connection Between Redundancy and Average Delay of Fixed-Length Coding", 1969):

**1**. Let $p_i$ be the probability of the $i$th source symbol and let

$$p_{\min} = \min\{p_1, \ldots, p_m\}.$$

**2**. Pick a real number $r \in (0, p_{min})$ and grow a complete parsing tree satisfying

$$p_{\min} r \leq P(d) < r\,, \quad d \in \mathcal{D}.$$

**3**. The resulting parsing tree is exactly the same as a tree constructed by Tunstall's algorithm.

**4**. Observe that if $y$ is a proper prefix of entries of $\mathcal{D} = \mathcal{D}_r$, i.e., $y$ corresponds to an internal node of $\mathcal{T} = \mathcal{T}_r$, then

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

We represent phrases (leaves of $\mathcal{T}$) in terms of the internal nodes of the parsing tree $\mathcal{T}_r$.

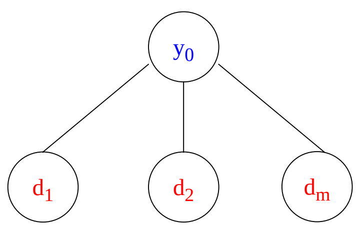
# A Useful Lemma on Trees

**Theorem 1.** *Let $\tilde{\mathcal{D}}$ be a uniquely parsable dictionary (leaves of $\mathcal{T}$) and $\tilde{\mathcal{Y}}$ be the collection of strings which are proper prefixes of one or more dictionary entries (internal nodes of $\mathcal{T}$). Then for all $|z| \leq 1$,*

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

**Proof** By induction with respect to $\tilde{\mathcal{Y}}$:

**1**. For the inductive step, let $\tilde{\mathcal{Y}}$ have $k+1$ elements.




Choose $y_0 \in \tilde{\mathcal{Y}}$ (of maximum length) so that its single letter extensions correspond to the dictionary entries $d_1, d_2, \ldots, d_m \in \tilde{\mathcal{D}}$.

Observe that $P(y_0) = P(d_1) + P(d_2) + \cdots + P(d_m)$.

**2**. Define an auxiliary dictionary $\tilde{\mathcal{D}}'$ with

$$\tilde{\mathcal{D}}' = \tilde{\mathcal{D}} \cup \{y_0\} \setminus \{d_1, \ldots, d_m\}.$$

Then $\tilde{\mathcal{D}}'$ has a corresponding proper prefix set with $k$ elements

$$\tilde{\mathcal{Y}}' = \tilde{\mathcal{Y}} \setminus \{y_0\}.$$

# Final Step of the Induction

3. By induction we have

$$\begin{aligned}
\sum_{y\in\tilde{\mathcal{Y}}} P(y)z^{|y|} &= \sum_{y\in\tilde{\mathcal{Y}}'} P(y)z^{|y|} + P(y_0)z^{|y_0|} \\
&= \sum_{d\in\tilde{\mathcal{D}}'} P(d)\frac{z^{|d|}-1}{z-1} + P(y_0)z^{|y_0|} \\
&= \sum_{d\in\tilde{\mathcal{D}}'\setminus\{y_0\}} P(d)\frac{z^{|d|}-1}{z-1} \\
&\quad +P(y_0)\left(z^{|y_0|} + \frac{z^{|y_0|}-1}{z-1}\right) \\
&= \sum_{d\in\tilde{\mathcal{D}}'\setminus\{y_0\}} P(d)\frac{z^{|d|}-1}{z-1} \\
&\quad +(P(d_1)+\cdots+P(d_m))\left(\frac{z^{|y_0|+1}-1}{z-1}\right) \\
&= \sum_{d\in\tilde{\mathcal{D}}} P(d)\frac{z^{|d|}-1}{z-1}.
\end{aligned}$$

# Phrase Length

Let $D = |d|$ for $d \in \mathcal{D}$.

Assume now that source strings are generated by a memoryless source over an alphabet $\mathcal{A} = \{1, \ldots, m\}$.

Then we can talk about moments of $D$. We have from previous result

$$\begin{aligned} \mathbf{E}[D] &= \sum_{y \in \tilde{\mathcal{Y}}} P(y), \\ \mathbf{E}[D(D-1)] &= 2 \sum_{y \in \tilde{\mathcal{Y}}} P(y)|y|. \end{aligned}$$

**Moment Generating Functions**: Let

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

and its corresponding internal nodes generating function

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

From previous result we also conclude that

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

# Recurrences

Define for a binary alphabet $\{1,2\}$ with $p_1 < p_2$:
$v = 1/r$, $z$ complex: $\tilde{S}(v,z) = S(v^{-1}, z)$.

Let also

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

denote the number of strings with probability at least $v^{-1}$.

We have

$$A(v) = \begin{cases} 0 & v < 1, \\ 1 + A(vp_1) + A(vp_2) & v \geq 1 \end{cases}$$

and

$$\tilde{S}(v,z) = \begin{cases} 0 & v < 1, \\ 1 + zp_1\tilde{S}(vp_1, z) + zp_2\tilde{S}(vp_2, z) & v \geq 1, \end{cases}$$

since every binary string either is:

- – empty string,
- – string starting with $1$
- – string starting with $2$.

# Some Intermediate Results

$A(v)$ represents the number of internal nodes in Khodak's construction:

$$\begin{aligned} M_r &= A(v) + 1 = |\mathcal{D}_r| \\ \mathbf{E}[D_r] &= \tilde{S}(v, 1) \end{aligned}$$

We shall prove the following.
**Lemma 1.** *If $\ln p_2 / \lg p_1$ is irrational, then*

$$M_r = A(v) + 1 = \frac{v}{H} + o(v),$$

*otherwise (i.e., $\ln p_2 / \lg p_1$ is rational)*

$$M_r = A(v) + 1 = \frac{Q_1(\log v)}{H} v + O(v^{1-\eta})$$

*for some $\eta > 0$, where*

$$Q_1(x) = \frac{L}{1 - e^{-L}} e^{-L \langle \frac{x}{L} \rangle}$$

*$L > 0$ is the largest real number for which*
*$\ln(1/p_1)$ and $\ln(1/p_2)$ are integer multiples of $L$;*
*$H = p_1 \ln(1/p_1) + p_2 \ln(1/p_2)$ is the entropy,*
*$\langle y \rangle = y - \lfloor y \rfloor$ is the fractional part of $y$.*

*Furthermore, if* $\ln p_2/\ln p_1$ *is irrational, then*

$$\mathbf{E}[D_r] = \tilde{S}(v,1) = \frac{\log v}{H} + \frac{H_2}{2H^2} + o(1)$$

*otherwise (i.e.,* $\ln p_2/\lg p_1$ *is rational)*

$$\mathbf{E}[D_r] = \tilde{S}(v,1) = \frac{\log v}{H} + \frac{H_2}{2H^2} + \frac{Q_2(\log v)}{H} + O(v^{-\eta})$$

*for some* $\eta > 0$*, where*

$$Q_2(x) = L \cdot \left(\frac{1}{2} - \left\langle \frac{x}{L} \right\rangle\right)$$

*and* $H_2 = p_1 \ln(1/p_1)^2 + p_2 \ln(1/p_2)^2$.

# Idea of the Proof

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

$$\tilde{D}^*(s, z) = \frac{1 - z}{s(1 - z p_1^{1-s} - z p_2^{1-s})} - \frac{1}{s}, \qquad \Re(s) < s_0(z),$$

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

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

of all complex roots of $zp^{1-s} + zq^{1-s} = 1$.

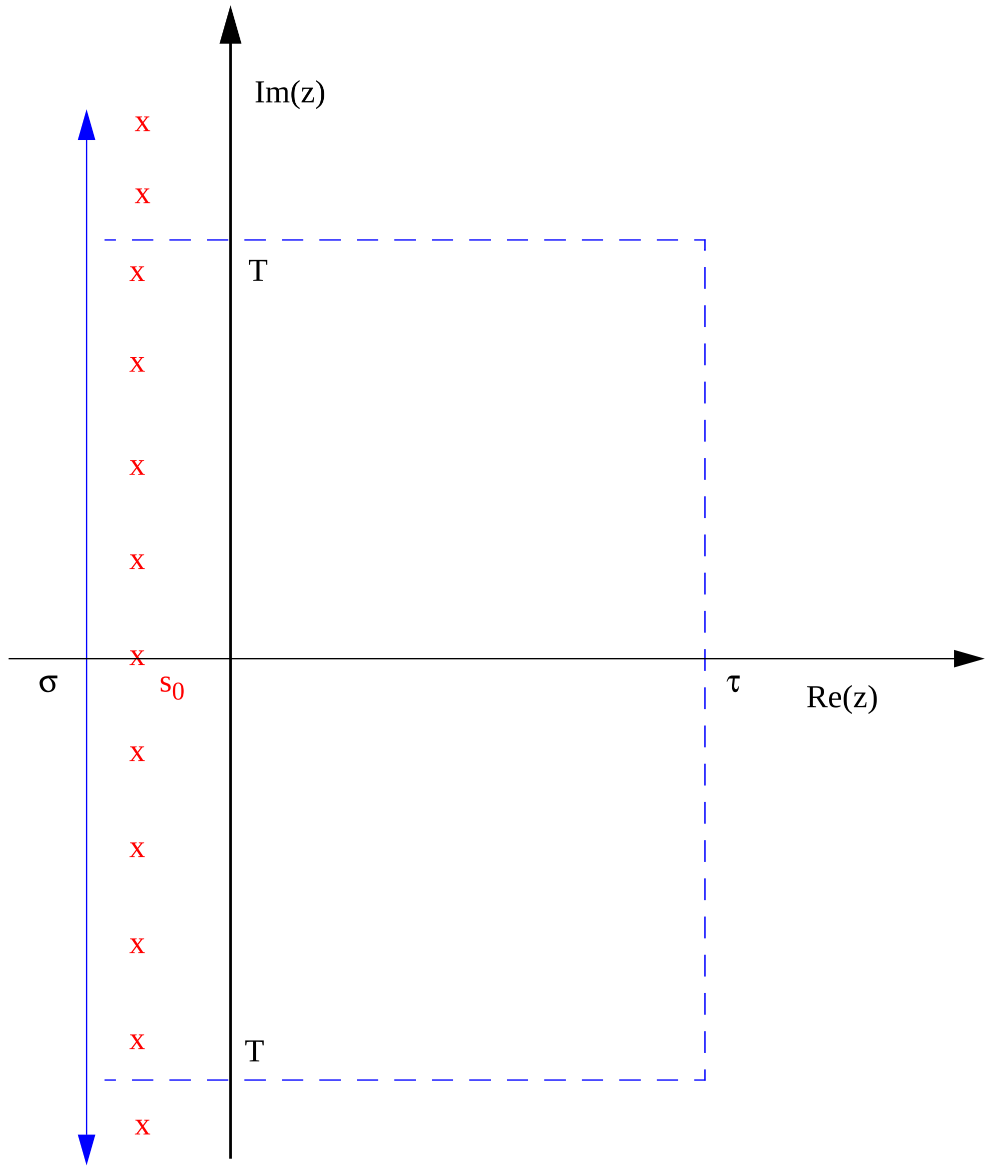



Figure 2: Illustration to the computations of the Mellin transform.

# Difficulties

From Cauchy's residue theorem we obtain
$\tilde{D}(v,z) = \lim_{T\to\infty} F_T(v,z)$ for $\Re(s) < \tau$, where

$$\begin{aligned}
&F_T(v,z)\\
&= - \sum_{s' \in Z(z),\ \Re(s')<\tau, |\Im(s')|>T} \operatorname{Res}(\tilde{D}^*(s,z)\, v^{-s},\ s = s')\\
&+ \frac{1}{2\pi i} \int_{\tau - iT}^{\tau + iT} \left( \frac{1-z}{s(1 - z p_1^{1-s} - z p_2^{1-s})} - \frac{1}{s} \right) v^{-s}\, ds\\
&= - \sum_{s' \in Z(z),\ \Re(s')<\tau, |\Im(s')|>T} \frac{(1-z) v^{-s'}}{z s' p_1^{1-s'} \ln p_1 + z s' p_2^{1-s'} \ln p_2}\\
&+ \frac{1}{2\pi i} \int_{\tau - iT}^{\tau + iT} \left( \frac{1-z}{s(1 - z p_1^{1-s} - z p_2^{1-s})} - \frac{1}{s} \right) v^{-s}\, ds
\end{aligned}$$

provided that the series of residues converges and the limit as $T \to \infty$ of the last integral exists.

The problem is that neither the series nor the integral above are absolutely convergent since the integrand is only of order $1/s$.

## Main Results

**Theorem 2.** *Let $D_r$ and $M_r$ denote the phrase length and the dictionary size in Khodak's construction.*

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

*where $N(0,1)$ denotes the standard normal distribution. Furthermore,*

$$\mathbf{Var}[D_r] = \left(\frac{H_2}{H^3} - \frac{1}{H}\right)\ln M_r + O(1)$$

*for large $M_r$.*
*In the irrational case*

$$\mathbf{E}[D_r] = \frac{\ln M_r}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + o(1)$$

*and in the rational case (i.e., for $\log p_2/\log p_1 = b/d$)*

$$\mathbf{E}[D_r] = \frac{\ln M_r}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + \frac{Q_2(nv) - \ln Q_1(\ln v)}{H} + O(M_r^{-\eta}).$$

*The above yields (miracle: no oscillations)*

$$Q_2(\ln v) - \log Q_1(\ln v) = -\ln L + \ln(1 - e^{-L}) + \frac{L}{2}.$$

# Redundancy

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

$$\mathcal{R}_{M_r} = \frac{\ln M_r}{\mathbf{E}[D]} - H$$

As a consequence of our main result we can prove the following.

**Corollary 1.** *Let $\mathcal{D}_r$ denote the dictionary in Khodak's construction of the Tunstall code of size $M_r$. If $\ln p_1 / \ln p_2$ is irrational then*

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

*In the rational case we have*

$$\begin{aligned}\mathcal{R}_{M_r} \quad = \quad & \frac{H}{\ln M_r}\Big( - \frac{H_2}{2H} - \ln H \\ & + \ln L - \ln(e^L - 1) + \frac{L}{2}\Big) + O\left(M_r^{-\eta}\right),\end{aligned}$$

*for some $\eta > 0$, where $L > 0$ is the largest real number for which $\ln(1/p_1)$ and $\ln(1/p_2)$ are integer multiples of $L$ (no oscillation!)*

See Savari and Gallager, Generalized Tunstall codes for sources with memory, IT-43, 1997 (tool: renewal theory).

# Miracles Happen – Oscillations

Oscillations usually occur in redundancy rates.

Lemple-Ziv-78. In this case:

$$r_n = \frac{2h - h\gamma - \frac{1}{2}h_2 + h\beta - h\delta_0(n)}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right) ,$$

where

$$\begin{aligned} h &= -p\log p - q\log q \\ h_2 &= p\log^2 p + q\log^2 q , \end{aligned}$$

$\gamma = 0.577\ldots$ is the Euler constant,
and $\delta_0(x)$ is a fluctuating functions with small amplitude for $\log p/\log q$ rational, and

$$\lim_{x\to\infty} \delta_0(x) = 0$$

otherwise. Finally, the constant $\beta$ is defined as:

$$\beta = -\sum_{k=1}^{\infty} \frac{p^{k+1}\log p + q^{k+1}\log q}{1 - p^{k+1} - q^{k+1}} .$$

# Shannon and Huffman Codes

**Shannon's Code**. In this case:

$$\bar{R}_n^{SF} = \begin{cases} \frac{1}{2} + o(1) & \alpha \quad \text{irrational} \\ \frac{1}{2} - \frac{1}{M}\left(\langle Mn\beta\rangle - \frac{1}{2}\right) + O(\rho^n) & \alpha = \frac{N}{M}, \end{cases}$$

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

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

**Huffman's Code**. In this case:

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

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