Precise Asymptotic Analysis of the Tunstall Code

Michael Drmota*, Yuriy Reznik†, Serap A. Savari‡, and Wojciech Szpankowski§
* Institute of Discrete Mathematics and Geometry, TU Wien, Wiedner Hauptstr. 8–10, A-1040 Wien, Austria
† Qualcomm Inc., 5775 Morehouse Dr., San Diego, CA 92121
‡ Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109
§ Department of Computer Science, Purdue University, W. Lafayette, IN 47907

A variable-to-fixed length encoder partitions the source string over an m-ary alphabet $\mathcal{A}$ into a concatenation of variable-length phrases. Each phrase except the last one is constrained to belong to a given dictionary $\mathcal{D}$ of source strings; the last phrase is a non-null prefix of a dictionary entry. One common constraint on a dictionary is that it leads to a *unique* parsing of any string over $\mathcal{A}$. We will assume that all dictionaries are uniquely parsable. It is convenient to represent a uniquely parsable dictionary by a complete parsing tree $\mathcal{T}$, i.e., a tree in which every internal node has all m children nodes in the tree. The dictionary entries $d \in \mathcal{D}$ correspond to the leaves of parsing tree. The encoder represents each parsed string by the fixed length binary code word corresponding to its dictionary entry. If the dictionary $\mathcal{D}$ is has M entries, then the code word for each phrase has $\lceil \log_2 M \rceil$ bits. The best known variable-to-fixed length code is now generally attributed to Tunstall [3]; however, it was independently discovered by Khodak [1], and possibly others.

Tunstall's algorithm is simple to visualize through evolving parsing trees in which every edge corresponds to a letter from the source alphabet $\mathcal{A}$: starting from a tree with a root node and m leaves which together correspond to all of the symbols in $\mathcal{A}$). For J iterations we select the current leaf corresponding to a string of highest probability and grow m children out it, one for each symbol in $\mathcal{A}$. After these J steps, the parsing tree has J non-root internal nodes and $M = (m-1)J+m$ leaves, which each correspond to a distinct dictionary entry. The dictionary entries are prefix-free and can be easily enumerated.

To facilitate our analysis, we will focus upon another construction of the Tunstall code that was invented by Khodak [1]. Khodak independently discovered the Tunstall code using a rather different approach. Let p_i be the probability of the ith source symbol and let $p_{\min} = \min\{p_1, \ldots, p_m\}$. Khodak suggested choosing a real number $r \in (0, p_{min})$ and growing a complete parsing tree satisfying $p_{\min} r \leq P(d) < r$ for $d \in \mathcal{D}$. It can also be shown (cf. [2, Lemma 2]) that the resulting parsing tree is exactly the same as a tree constructed by Tunstall's algorithm. The asymptotic relationship between r and the resulting number of dictionary entries M_r was studied in [2] and will be established here in a different way. It follows that if y is a proper prefix of one or more entries of $\mathcal{D} = \mathcal{D}_r$, i.e., y corresponds to an internal node of $\mathcal{T} = \mathcal{T}_r$, then $P(y) \geq r$.

Assume a memoryless source over an m-ary alphabet $\mathcal{A}$ generates an output sequence. Let $p_i > 0$ be the probability of the i^{th} letter of alphabet $\mathcal{A}$, $i \in \{1, \ldots, m\}$, $p_{\min} = \min\{p_1, \ldots, p_m\}$, and $p_{\max} = \max\{p_1, \ldots, p_m\}$. Given a dictionary $\mathcal{D}$ and corresponding complete parsing tree $\mathcal{T}$, the encoder partitions the source output sequence into a sequence of variable-length phrases. Let $d \in \mathcal{D}$ denote a dictionary entry, $P(d)$ be its probability, and $|d|$ be its length. Our focus will be on the random variable $D = |d|$, the phrase length of a dictionary string. One of our goals is to investigate the moment generating function of the phrase length $D = D_r$ in Khodak's construction of the Tunstall dictionary with parameter r.

Theorem 1: Let D_r denote the phrase length in Khodak's construction of the Tunstall code with a dictionary of size M_r over a biased memoryless source. Then as $M_r \to \infty$

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

where $N(0,1)$ denotes the standard normal distribution. Furthermore, we have $E[D] = \frac{\ln M_r}{H} + O(1)$ and

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

for large M_r, where H is the entropy and $H_2 = \sum_{i=1}^{m} p_i \ln^2 p_i$.

As a direct consequence, we find asymptotics for the average redundancy of the Tunstall code defined by $\mathcal{R}_M = \frac{\ln M}{\mathbf{E}[D]} - H$.

Corollary 1: For a binary alphabet, 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

$$\mathcal{R}_{M_r} \sim \frac{H}{\ln M_r}\Big(-\frac{H_2}{2H} - \ln H + \ln L - \ln(e^L - 1) + \frac{L}{2}\Big)$$

where $L > 0$ is the largest real number for which $\ln(1/p_1)$ and $\ln(1/p_2)$ are integer multiples of L.

References

[1] G. L. Khodak, Connection Between Redundancy and Average Delay of Fixed-Length Coding, *Conf. Problems of Theoretical Cybernetics*, 1969.
[2] S. A. Savari, Robert G. Gallager; Generalized Tunstall codes for sources with memory, IEEE Trans. Info. Theory, vol. IT-43, pp. 658 - 668, March 1997.
[3] B. P. Tunstall, Synthesis of Noiseless Compression Codes, Ph.D. dissertation, (Georgia Inst. Tech., Atlanta, GA, 1967)