Average Redundancy for Known Sources: Ubiquitous Trees in Source Coding

Wojciech Szpankowski†

Department of Computer Science, Purdue University, West Lafayette, IN, USA. spa@cs.purdue.edu

Analytic information theory aims at studying problems of information theory using analytic techniques of computer science and combinatorics. Following Hadamard's precept, these problems are tackled by complex analysis methods such as generating functions, Mellin transform, Fourier series, saddle point method, analytic poissonization and depoissonization, and singularity analysis. This approach lies at the crossroad of computer science and information theory. In this survey we concentrate on one facet of information theory (i.e., source coding better known as data compression), namely the *redundancy rate* problem. The redundancy rate problem determines by how much the actual code length exceeds the optimal code length. We further restrict our interest to the *average* redundancy for *known* sources, that is, when statistics of information sources are known. We present precise analyses of three types of lossless data compression schemes, namely fixed-to-variable (FV) length codes, variable-to-fixed (VF) length codes, and variable-to-variable (VV) length codes. In particular, we investigate average redundancy of Huffman, Tunstall, and Khodak codes. These codes have succinct representations as *trees*, either as coding or parsing trees, and we analyze here some of their parameters (e.g., the average path from the root to a leaf).

Keywords: Source coding, prefix codes, Kraft's inequality, Shannon lower bound, data compression, Huffman code, Tunstall code, Khodak code, redundancy, distribution modulo 1, Mellin transform, complex asymptotics.

1 Introduction

The basic problem of *source coding* better known as (lossless) *data compression* is to find a binary code that can be unambiguously recovered with shortest possible description either on average or for individual sequences. Thanks to Shannon's work we know that on average the number of binary bits per source symbol cannot be smaller than the source entropy rate. There are many codes achieving the entropy, therefore one turns attention to *redundancy*. The average redundancy of a source code is the amount by which the expected number of binary digits per source symbol for that code exceeds entropy. One of the goals in designing source coding algorithms is to minimize the average redundancy. In this survey, we discuss various classes of source coding and their corresponding average redundancy. It turns out that

†This work was supported in part by the NSF Grants CCF-0513636, DMS-0503742, DMS-0800568, and CCF-0830140, NIH Grant R01 GM068959-01, NSA Grant H98230-08-1-0092, EU Project No. 224218 through Poznan University of Technology, and the AFOSR Grant FA8655-08-1-3018.

such analyses often resort to studying certain intriguing trees such as Huffman, Tunstall and Khodak trees. We study them using tools from analysis of algorithms.

Lossless data compression comes in three flavors: fixed-to-variable (FV) length codes, variable-to-fixed (VF) length codes, and finally variable-to-variable (VV) length codes. The latter includes the previous two families of codes and is the least studied among all data compression schemes. In the fixed-to-variable code the encoder maps fixed length blocks of source symbols into variable-length binary code strings. Two important fixed-to-variable length coding schemes are the Shannon code and the Huffman code. While Huffman has already known that the average code length is asymptotically equal to the entropy of the source, the asymptotic performance of the Huffman code is still not fully understood. In [1] Abrahams summarizes much of the vast literature on fixed-to-variable length codes. In this survey, we present precise analysis from our work [129] of the Huffman average redundancy for memoryless sources. We show that the average redundancy either converges to an explicitly computable constant, as the block length increases, or it exhibits a very erratic behavior fluctuating between 0 and 1.

A VF encoder partitions the source string into variable-length phrases that belong to a given dictionary $\mathcal{D}$. Often a dictionary is represented by a complete tree (i.e., a tree in which every node has maximum degree), also known as the *parsing tree*. The codes assigns a fixed-length word to each dictionary entry. An important example of a variable-to-fixed code is the Tunstall code [133]. Savari and Gallager [112] present an analysis of the dominant term in the asymptotic expansion of the Tunstall code redundancy. In this survey, following [33], we describe a precise analysis of the phrase length (i.e., path from the root to a terminal node in the corresponding parsing tree) for such a code and its average redundancy.

Finally, a variable-to-variable (VV) code is a concatenation of variable-to-fixed and fixed-to-variable codes. A variable-to-variable length encoder consists of a *parser* and a *string encoder*. The parser, as in VF codes, segments the source sequence into a concatenation of phrases from a predetermined dictionary $\mathcal{D}$. Next, the string encoder in a variable-to-variable scheme takes the sequence of dictionary strings and maps each one into its corresponding binary codeword of variable length. Aside from the special cases where either the dictionary strings or the codewords have a fixed length, very little is known about variable-to-variable length codes, even in the case of memoryless sources. Surprisingly, in 1972 Khodak [65] described a VV scheme with small average redundancy that decreases with the growth of phrase length. He did not offer, however, an explicit VV code construction. We will remedy this situation and follow [12] to propose a transparent proof.

Throughout this survey, we study various intriguing trees describing Huffman, Tunstall and Khodak codes. These trees are studied by analytic techniques of analysis of algorithms [42; 70; 71; 72; 130]. The program of applying tools from analysis of algorithms to problems of source coding and in general to information theory lies at the crossroad of computer science and information theory. It is also known as *analytic information theory*. In fact, the interplay between information theory and computer science dates back to the founding father of information theory, Claude E. Shannon. His landmark paper "A Mathematical Theory of Communication" is hailed as the foundation for information theory. Shannon also worked on problems in computer science such as chess-playing machines and computability of different Turing machines. Ever since Shannon's work on both information theory and computer science, the research at the interplay between these two fields has continued and expanded in many exciting ways. In the late 1960s and early 1970s, there were tremendous interdisciplinary research activities, exemplified by the work of Kolmogorov, Chaitin, and Solomonoff, with the aim of establishing algorithmic information theory. Motivated by approaching Kolmogorov complexity algorithmically, A. Lempel (a computer scientist), and J. Ziv (an information theorist) worked together in the late 1970s to develop compression algorithms that are now widely referred to as Lempel-Ziv algorithms. Analytic information theory is a continuation of these efforts.

Finally, we point out that this survey deals only with source coding for *known sources*. The more practical *universal source coding* (in which source distribution is unknown) is left for another time. However, at the end of this survey we provide an extensive bibliography on the redundancy rate problem, including universal source coding. In particular, we note that recent years have seen a resurgence of interest in redundancy rate for *fixed-to-variable* coding (cf. [18; 23; 24; 25; 53; 78; 79; 80; 84; 86; 103; 105; 109; 110; 112; 118; 121; 128; 129; 139; 146; 153; 149; 150]). Surprisingly there are only a handful of results for variable-to-fixed codes (cf. [63; 76; 92; 111; 112; 113; 132; 136; 158]) and an almost non-existing literature on variable-to-variable codes (cf. [36; 44; 65; 76]). While there is some recent work on universal VF codes [132; 136; 158], to the best of our knowledge redundancy for universal VF and VV codes was not studied with the exception of some preliminary work of the Russian school [76; 77] (cf. also [82]).

This survey is organized as follows. In the next section, we present some preliminary results such as Kraft's inequality, Shannon lower bound, and Barron's lemma. In Section 3 we analyze Hufmman's code. Then we turn our attention in Section 4 to the Tunstall and VF Khodak codes. Finally, in Section 5 we the VV code of Khodak and its interesting analysis. We conclude this survey with two remarks concerning average redundancy for sources with unknown parameters and for non-prefix codes.

2 Preliminary Results

Let us start with some definitions and preliminary results. A *source code* is a bijective mapping

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

from the set of all sequences over an alphabet $\mathcal{A}$ to the set $\{0,1\}^*$ of binary sequences. We write $x \in \mathcal{A}^*$ for a sequence of unspecified length, and $x_i^j = x_i \dots x_j \in \mathcal{A}^{j-i+1}$ for a sequence of length $j - i + 1$. We denote by P the probability of the source, and write $L(C, x)$ (or simply $L(x)$) for the code length of the source sequence x over the code C. Finally, the source *entropy* is defined as usual by $H(P) = -\sum_{x\in\mathcal{A}^*} P(x) \lg P(x)$ and the *entropy rate* is denoted by h. We write $\lg := \log_2$ and $\log$ for the logarithm of unspecified base. We often present our results for the binary alphabet $\mathcal{A} = \{0, 1\}$.

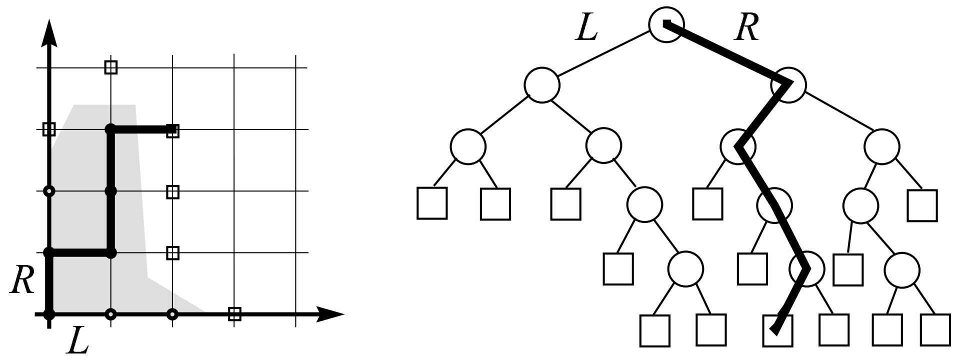


Fig. 1: Lattice paths and binary trees

Throughout this survey (except in Section 6.2) we study *prefix codes* for which no codeword is a prefix of another codeword. For such codes there is a mapping between a prefix code and a path in a tree from the root to a terminal (external) node (e.g., for a binary prefix code move to the left in the tree represents 0 and move to the right represents 1), as shown in Figure 1. We also point out that a prefix code and the corresponding path in a tree defines a lattice path in the first quadrant also shown in Figure 1. If some additional constraints are imposed on the prefix codes, this translates into certain restrictions on the lattice path indicated as the shaded area in Figure 1.

The prefix condition imposes some restrictions on the code length. This fact is knows as Kraft's inequality discussed next.

Theorem 1 (Kraft's Inequality) *Let $|\mathcal{A}| = m$. For any prefix code the codeword lengths $\ell_1, \ell_2, \ldots, \ell_N$ satisfy the inequality*

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

Conversely, if codeword lengths satisfy this inequality, then one can build a prefix code.

Proof. This is an easy exercise on trees. Consider only a binary alphabet $|\mathcal{A}| = 2$. Let $\ell_{\max}$ be the maximum codeword length. Observe that at level $\ell_{\max}$ some nodes are codewords, some are descendants of codewords, and some are neither. Since the number of descendants at level $\ell_{\max}$ of a codeword located at level ℓ_i is $2^{\ell_{\max}-\ell_i}$, we obtain

$$\sum_{i=1}^{N} 2^{\ell_{\max}-\ell_i} \leq 2^{\ell_{\max}},$$

which is the desired inequality. The converse part can also be proved, and is left for the reader. ■

Observe that the Kraft's inequality implies the existence of at least one sequence $\widetilde{x}$ such that

$$L(\widetilde{x}) \geq -\log P(\widetilde{x}).$$

Actually, a stronger statement is due to Barron [5] who proved the following result.

Lemma 1 (Barron) *Let $L(X)$ be the length of a prefix code, where X is generated by a stationary ergodic source over a binary alphabet. For any sequence a_n of positive constants satisfying $\sum_n 2^{-a_n} < \infty$ the following holds*

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

and therefore

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

Proof: We argue as follows:

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

The lemma follows from the Kraft inequality for binary alphabets and the Borel-Cantelli Lemma. ∎

Using Kraft's inequality we can now prove the first theorem of Shannon that bounds from below the average code length.

Theorem 2 *For any prefix code the average code length $\mathbb{E}[L(C, X)]$ cannot be smaller than the entropy of the source $H(P)$, that is,*

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

where the expectation is taken with respect to the distribution P of the source sequence X.

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

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

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

What is the best code length? We are now in a position to answer this question. As long as the expected code length is concerned, one needs to solve the following constrained optimization problem for, say a binary alphabet

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

This optimization problem has an easy solution through Lagrangian multipliers, and one finds that the optimal code length is $L(x) = -\lg P(x)$ provided the *integer character of the length is ignored*.

In general, one needs to round the length to an integer, thereby incurring some cost. This cost is usually known under the name *redundancy*. For *known* distribution P, that we assume throughout this survey, the *pointwise redundancy* $R^C(x)$ for a code C and the *average redundancy* $\overline{R}^C$ are defined as

$$R^C(x) = L(C, x) + \lg P(x), \qquad \overline{R}^C = \mathbb{E}[L(C, X)] - H(P)] \geq 0.$$

The pointwise redundancy can be negative, but the average redundancy cannot due to the Shannon theorem.

3 Redundancy of Huffman's FV Code

We now turn our attention to fixed-to-variable length codes, in particular Shannon and Huffman codes. In this section, we assume that a known source P generates a sequence $x_1^n = x_1 \dots x_n$ of *fixed* length n. The code $C(x_1^n)$ may be of a variable length.

We are interested in constructing an optimal code on average. It is known that the following optimization problem

$$\overline{R}_n^H = \min_{C_n} \mathbb{E}_{X_1^n}[L(C_n, x_1^n) + \lg P(x_1^n)]$$

is solved by the *Huffman code*. Recall that Huffman code is a recursive algorithm built over the associated Huffman tree, in which the two nodes with lowest probabilities are combined into a new node whose probability is the sum of the probabilities of its two children. Huffman coding is still one of the most familiar topics in information theory [1; 45; 46; 124], however, only recently a precise estimate of the average redundancy $\overline{R}_n^H$ of the Huffman code was derived in [129] that we review below.

We study the average redundancy for memoryless sources emitting a binary sequence. Let p denote the probability of generating "0" and $q = 1 - p$ denote the probability of emitting "1". Throughout, we assume that $p < \frac{1}{2}$. We denote by $P(x_1^n) = p^k q^{n-k}$ the probability of generating a binary sequence consisting of k zeros and $n-k$ ones. The expected code length $\mathbb{E}[L_n]$ of the Huffman code is

$$E[L_n] = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} L(k),$$

where

$$L(k) = \frac{1}{\binom{n}{k}} \sum_{j \in \mathcal{S}_k} l_j$$

with $\mathcal{S}_k$ representing the set of all inputs having probability $p^k q^{n-k}$, and l_j being the length of the jth code in $\mathcal{S}_k$. By Gallager's sibling property [46], we know that code lengths in $\mathcal{S}_k$ are either equal to $l(k)$ or $l(k)+1$ for some integer $l(k)$. If n_k denotes the number of code words in $\mathcal{S}_k$ that are equal to $l(k)+1$, then

$$L(k) = l(k) + \frac{n_k}{\binom{n}{k}}.$$

Clearly, $l(k) = \lfloor -\lg(p^k q^{n-k}) \rfloor$. Stubley [124] analyzed carefully n_k and was led to conclude that

$$\begin{aligned} \overline{R}_n^H &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [\lg(p^k q^{n-k}) + \lfloor -\lg(p^k q^{n-k}) \rfloor] \\ &+ 2 \sum_{k=0}^{n-1} \binom{n}{k} p^k q^{n-k} (1 - 2^{(\lg(p^k q^{n-k}) + \lfloor -\lg(p^k q^{n-k}) \rfloor)}) + o(1). \end{aligned}$$

Since

$$\lg(p^k q^{n-k}) + \lfloor -\lg(p^k q^{n-k}) \rfloor = \langle \alpha k + \beta n \rangle$$

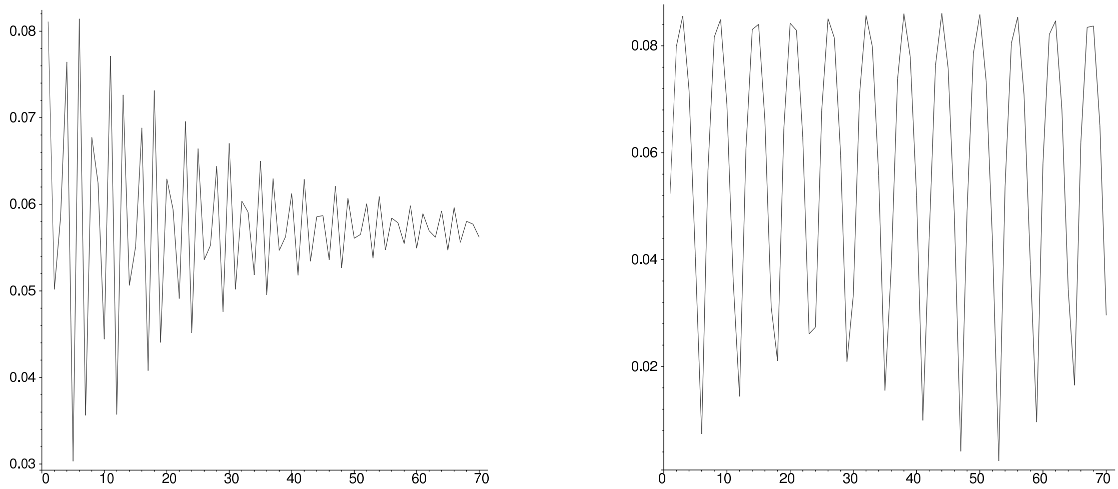
where

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

and $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x, we arrive at the following

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

This is our starting formula for the average Huffman redundancy. In [129] we proved the following result.



(a) (b)

Fig. 2: 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 $p = 1/9$.

Theorem 3 *Consider the Huffman block code of length n over a binary memoryless source. For $p < \frac{1}{2}$ as $n \to \infty$*

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

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

Before we present a sketch of the proof, we plot in Figure 2 the average redundancy $\overline{R}_n^H$ as a function of n for two values of α, one *irrational* and one *rational*. In Figure 2(a) we consider $\alpha = \lg(1-p)/p$ irrational while in Figure 2(b) α is rational. Two modes of behavior are clearly visible. The function in Figure 2(a) converges to a constant (≈ 0.05) for large n as predicted by Theorem 3, while the curve in Figure 2(b) is quite erratic (with the maximum close to Gallager's upper bound 0.086).

We now briefly sketch the proof of Theorem 3. Details can be found in [129]. From the above discussion, it should be clear that in order to evaluate the sums appearing in $\overline{R}_n^H$ we need to understand asymptotics of the following

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

for fixed p and some Riemann integrable function $f : [0,1] \to \mathbb{R}$ uniformly over $y \in \mathbb{R}$ where x_k is a sequence. In our case $x_k = \alpha k$ and $y = \beta n$. We need to consider two cases: α irrational and α rational.

The case when α is rational is relatively elementary. The following lemma taken from [129] is easy to prove. Using below lemma we easily derive (3) for α rational.

Lemma 2 *Let $0 < p < 1$ be a fixed real number and suppose that $\alpha = \frac{N}{M}$ is a rational number with* $\gcd(N, M) = 1$. *Then for every bounded function $f : [0, 1] \to \mathbb{R}$ we have*

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle k\alpha + y\rangle) \;=\; \frac{1}{M}\sum_{l=0}^{M-1} f\left(\frac{l}{M} + \frac{\langle My\rangle}{M}\right) + O(\rho^n) \tag{4}$$

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

The irrational case is more sophisticated and we need to appeal to *theory of sequences modulo* 1 as fully explained in the book by Drmota and Tichy [32]. The following result can be found in [32; 130].

Lemma 3 *Let $0 < p < 1$ be a fixed real number and α be an irrational number. Then for every Riemann integrable function $f : [0, 1] \to \mathbb{R}$ we have*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y\rangle) = \int_0^1 f(t)\,dt, \tag{5}$$

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

In our case we set $f(t) = t$ and $f(t) = 2^{-t}$ in (5) and Theorem 3 follows.

In passing we should point out that the methodology presented here can be used to derive redundancy of other FV codes. For example, Shannon code assigns the code length $\lceil -\lg(p^k q^{n-k})\rceil$ for the probability $p^k q^{n-k}$. Its average redundancy is then

$$\begin{aligned}
\overline{R}_n^S &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left(\lceil -\lg(p^k q^{n-k})\rceil + \lg p^k q^{n-k}\right) \\
&= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle -\lg(p^k q^{n-k})\rangle \qquad (6) \\
&= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \alpha k + \beta n\rangle \qquad (7)
\end{aligned}$$

Using Lemmas 2 and 3 we easily arrive at the following conclusion.

Theorem 4 *Consider the Shannon block code of length n over a binary memoryless source. For $p < \frac{1}{2}$ as $n \to \infty$*

$$\overline{r}_n = \begin{cases} \frac{1}{2} + o(1) & \alpha \quad \text{irrational} \\ \frac{1}{2} - \frac{1}{M}\left(\langle Mn\beta\rangle - \frac{1}{2}\right) + O(\rho^n) & \alpha = \frac{N}{M}, \quad \gcd(N, M) = 1 \end{cases} \tag{8}$$

where $\rho < 1$.

In [129] we also derived the redundancy of Golomb's code which is a Huffman code for unbounded alphabets.

4 Redundancy of Tunstall and Khodak VF Codes

We now study variable-to-fixed (VF) length codes, in particular, the Tunstall and Khodak VF codes. Recall that in the in VF scenario, the source string x, say over m-ary alphabet $\mathcal{A}$, is partitioned into non-overlapping (unique) phrases, each belonging to a given *dictionary* $\mathcal{D}$ represented by a complete *parsing tree* $\mathcal{T}$. The dictionary entries $d \in \mathcal{D}$ correspond to the *leaves* of the associated parsing tree, so that VF codes are prefix codes. The encoder represents each parsed string by the fixed length binary code word corresponding to its dictionary entry. If the dictionary $\mathcal{D}$ 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 the Tunstall code [133]; however, it was independently discovered by Khodak [64].

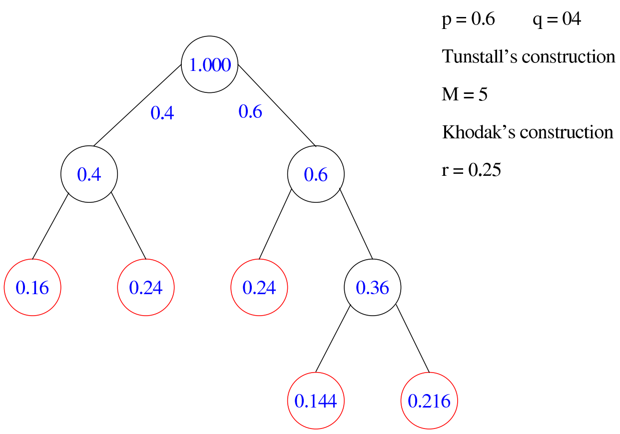


Fig. 3: Tunstall's and Khodak's Codes for $M = 5$ and $r = 0.25$.

Edges in the parsing tree of the Tunstall's code correspond to letters from the source alphabet $\mathcal{A}$ and are labeled by the alphabet probabilities, say $p_1, \ldots, p_m$. Every vertex in such a tree is assigned the probability of the path leading to it from the root, as shown in Figure 3. For memoryless sources, studied here, the probability of a vertex is the product of probabilities of vertices leading to it. More precisely, the root node has m leaves corresponding to all of the symbols in $\mathcal{A}$ and labeled by $p_1, \ldots, p_m$. At each iteration one selects the current leaf corresponding to a string of the *highest probability*, say $P_{\max}$, and grows m children out of it with probabilities $p_1 P_{\max}, \ldots, p_m P_{\max}$. After J iterations, the parsing tree has J non-root *internal nodes* and $M = (m-1)J + m$ leaves, each corresponding to a distinct dictionary entry.

Another version of VF algorithm was proposed by Khodak's [64] who independently discovered the Tunstall code using a rather different approach. 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 until all leaves $d \in \mathcal{D}$ satisfy

$$p_{\min} r \leq P(d) < r. \tag{9}$$

Khodak and Tunstall algorithms are illustrated in Figure 3 with the dictionary $\mathcal{D} = \{00, 01, 10, 110, 111\}$ corresponding to strings represented by the paths from the root to all terminal nodes.

It is known (see, e.g., [112, Lemma 2]) that the parsing trees for Tunstall and Khodak algorithms are exactly the same, however, they react differently to the probability tie when expanding a leaf. More precisely, when there are several leaves with the same probability, the Tunstall algorithm selects *one* leaf and expands it, then selects another leaf of the same probability, and continues doing it until all leaves of the same probability are expanded. The Khodak algorithm expands *all* leaves with the same probability simultaneously, in parallel; thus there are "jumps" in the number of dictionary entries M when the parsing tree grows. For example, in Figure 3 two nodes marked "0.24" will be expanded simultaneously in the

Khodak algorithm, and one after another by the Tunstall algorithm.

Our goal i this section is to present a precise analysis of the Khodak redundancy as well as to provide some insights into the behavior of the parsing tree (i.e., the path length distribution). Let us study first the average redundancy *rate* $\overline{r}$ defined

$$\overline{r} = \lim_{n\to\infty} \frac{\sum_{|x|=n} P_{\mathcal{S}}(x)(L(x) + \log P_{\mathcal{S}}(x))}{n}, \tag{10}$$

where $P_S(x)$ is the probability of the source sequence x. Using renewal theory (i.e., regeneration theory) [9] we find

$$\lim_{n\to\infty} \frac{\sum_{|x|=n} P_{\mathcal{S}}(x)L(x)}{n} = \frac{\sum_{d\in\mathcal{D}} P_{\mathcal{D}}(d)\ell(d)}{\mathbb{E}[D]}, \tag{11}$$

where $\ell(d)$ is the length of the phrase $d \in \mathcal{D}$, and $\mathbb{E}[D] = \sum_{d\in\mathcal{D}} |d|P_{\mathcal{D}}(d)$ is the average phrase length D, known also as the average *delay*, which is actually the average path length from the root to a terminal node in the corresponding parsing tree. In the above $P_{\mathcal{D}}$ represents the distribution of phrases in the dictionary, but from now on we shall write $P := P_{\mathcal{D}}$. Since for the VF codes $\sum_{|x|=n} P_{\mathcal{S}}(x)L(x) = \log_2 M$, we find

$$\overline{r} = \frac{\log M}{\mathbb{E}[D]} - h \tag{12}$$

where $h := h_S$ is the entropy rate of the source. In passing we should observe that by the *Conversation of Entropy Property* [111] the entropy rate of the dictionary $h_{\mathcal{D}}$ is related to the source entropy h as follows

$$h_{\mathcal{D}} = h\mathbb{E}[D]. \tag{13}$$

Tunstall's algorithm has been studied extensively (cf. the survey article [1]). Simple bounds for its redundancy were obtained independently by Khodak [64] and by Jelinek and Schneider [63]. Tjalkens and Willems [132] were the first to look at extensions of this code to sources with memory. Savari and Gallager [112] proposed a generalization of Tunstall's algorithm for Markov sources and used renewal theory for an asymptotic analysis of average code word length and redundancy for memoryless and Markov sources. Our presentation here is based on [33; 34].

In view of (12), we need to study the expected value of the phrase length $\mathbb{E}[D]$. In fact, we find the distribution of D. But, instead of concentrating on the terminal nodes we analyze the behavior of internal nodes. For Khodak's code, it follows from (9) that if y is a proper prefix of one or more entries of $\mathcal{D}_r := \mathcal{D}$, i.e., y corresponds to an internal node of $\mathcal{T}_r := \mathcal{T}$, then

$$P(y) \geq r. \tag{14}$$

Therefore, it is easier to characterize the internal nodes of the parsing tree $\mathcal{T}_r$ rather than its leaves. We shall follow this approach when analyzing the path length D of Khodak's code.

We first derive the moment generating function of the phrase length D and then its moments. Our approach is analytic and we use such tools as the Mellin transform and the Tauberian theorems [42; 130]. Let us define the probability generating function $D(r, z)$ of the phrase length D in the Khodak code with parameter r as

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

However, as mentioned above, it is better to work with another transform describing the probabilities of strings which correspond to *internal nodes* in the parsing tree $\mathcal{T}_r$. Therefore, we also define

$$S(r,z) = \sum_{y:\ P(y)\geq r} P(y) z^{|y|}. \tag{15}$$

In (17) of Lemma 4 below we show that

$$D(r,z) = 1 + (z-1)S(r,z), \tag{16}$$

and therefore,

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

Lemma 4 *Let $\mathcal{D}$ be a uniquely parsable dictionary (i.e., leaves in the corresponding parsing tree) and $\mathcal{Y}$ be the collection of strings which are proper prefixes of one or more dictionary entries (i.e., internal nodes of the parsing tree). Then for all $|z| \leq 1$,*

$$\sum_{d\in\mathcal{D}} P(d)\frac{z^{|d|}-1}{z-1} = \sum_{y\in\mathcal{Y}} P(y) z^{|y|}. \tag{17}$$

We are now in the position to analyze the Khodak algorithm. Let $v = 1/r$ and z be a complex number. Define $\widetilde{S}(v,z) = S(v^{-1},z)$. We restrict our attention here to a binary alphabet $\mathcal{A}$ with $0 < p < q < 1$. Let $A(v)$ devote the number of source strings with probability at least v^{-1} (i.e., number of internal nodes in the corresponding parsing tree), that is,

$$A(v) = \sum_{y: P(y)\geq 1/v} 1. \tag{18}$$

The functions $A(v)$ and $\widetilde{S}(v,z)$ satisfy the following recurrences

$$A(v) = \begin{cases} 0 & v<1, \\ 1 + A(vp) + A(vq) & v \geq 1, \end{cases} \tag{19}$$

and

$$\widetilde{S}(v,z) = \begin{cases} 0 & v<1, \\ 1 + zp\widetilde{S}(vp,z) + zq\widetilde{S}(vq,z) & v \geq 1, \end{cases} \tag{20}$$

since every binary string either is the empty string, a string starting with the first source symbol, or a string starting with the second source symbol. This partition directly leads to the recurrences above. Observe that $A(v)$ represents the number of internal nodes in Khodak's construction with parameter v^{-1} and $M_r = A(v) + 1 = |\mathcal{D}_r|$ is the dictionary size for the binary alphabet. Further, $\mathbb{E}[D_r] = \widetilde{S}(v,1)$ and $\mathbb{E}[D_r(D_r-1)] = \widetilde{S}'(v,1)$.

We illustrate the approach of [33; 34] on distributional results of D. For this we have to analyze (16) which we write in the following form

$$\widetilde{D}(v,z) = D(1/v,z) = 1 + (z-1)\widetilde{S}(v,z)$$

where $\widetilde{S}(v,z)$ satisfies recurrence (20). We study asymptotics of $\widetilde{D}(v,z)$ using the Mellin transform [40; 42; 130]. The Mellin transform $F^*(s)$ of a function $F(v)$ is defined as

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

Using the fact that the Mellin transform of $F(ax)$ is $a^{-s}F^*(s)$, we conclude from recurrence (20) that the Mellin transform $D^*(s,z)$ of $\widetilde{D}(v,z)$ with respect to v becomes

$$\widetilde{D}^*(s,z) = \frac{1-z}{s(1-zp^{1-s}-zq^{1-s})} - \frac{1}{s}, \tag{21}$$

for $\Re(s) < s_0(z)$, where $s_0(z)$ denotes the real solution of $zp^{1-s}+zq^{1-s}=1$. It is easy to see that

$$s_0(z) = -\frac{z-1}{h_e} + \left(\frac{1}{h_e} - \frac{p\ln^2 p + q\ln^2 q}{2h_e^3}\right)(z-1)^2 + O((z-1)^3)$$

as $z \to 1$ where $h_2 = p\ln(1/p) + q\ln(1/q)$ is the natural entropy.

In order to find the asymptotics of $\widetilde{D}(v,z)$ as $v \to \infty$ we proceed to compute the inverse transform of $\widetilde{D}^*(s,z)$, that is (cf. [130])

$$\widetilde{D}(v,z) = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT} \widetilde{D}^*(s,z)v^{-s}\,ds, \tag{22}$$

where $\sigma < s_0(z)$. For this purpose it is usually necessary to determine the polar singularities of the meromorphic continuation of $\widetilde{D}^*(s,z)$ right to the line $\Re(s) = s_0(z)$, that is, we have to analyze the set

$$\mathcal{Z}(z) = \{s \in \mathbb{C} : zp^{1-s} + zq^{1-s} = 1\} \tag{23}$$

of all complex roots of $zp^{1-s}+zq^{1-s}=1$. The next lemma, basically due to Jacquet and Schachinger, summarizes all needed properties of the set $\mathcal{Z}(z)$. Its proof can be found in [34].

Lemma 5 *Suppose that* $0 < p < q < 1$ *and that* z *is a real number with* $|z-1| \le \delta$ *for some* $0 < \delta < 1$. *Let*

$$\mathcal{Z}(z) = \{s \in \mathbb{C} : p^{1-s} + q^{1-s} = 1/z\}.$$

(i) *All* $s \in \mathcal{Z}(z)$ *satisfy*

$$s_0(z) \le \Re(s) \le \sigma_0(z),$$

where $s_0(z) < 1$ *is the (unique) real solution of* $p^{1-s}+q^{1-s} = 1/z$ *and* $\sigma_0(z) > 1$ *is the (unique) real solution of* $1/z + q^{1-s} = p^{1-s}$. *Furthermore, for every integer* k *there uniquely exists* $s_k(z) \in \mathcal{Z}(z)$ *with*

$$(2k-1)\pi/\log p < \Im(s_k(z)) < (2k+1)\pi/\log p$$

and consequently $\mathcal{Z}(z) = \{s_k(z) : k \in \mathbb{Z}\}$.

(ii) *If* $\log q/\log p$ *is irrational, then* $\Re(s_k(z)) > \Re(s_0(z))$ *for all* $k \ne 0$ *and also*

$$\min_{|z-1|\le\delta}(\Re(s_k(z)) - \Re(s_0(z))) > 0. \tag{24}$$

(iii) *If* $\log q/\log p = r/d$ *is rational, where* $\gcd(r,d) = 1$ *for integers* $r, d > 0$*, then we have* $\Re(s_k(z)) = \Re(s_0(z))$ *if and only if* $k \equiv 0 \bmod d$*. In particular* $\Re(s_1(z)), \ldots, \Re(s_{d-1}(z)) > \Re(s_0(z))$ *and*

$$s_k(z) = s_{k \bmod d}(z) + \frac{2(k - k \bmod d)\pi i}{\log p},$$

that is, all $s \in \mathcal{Z}(z)$ *are uniquely determined by* $s_0(z)$ *and by* $s_1(z), s_2(z), \ldots, s_{d-1}(z)$*, and their imaginary parts constitute an arithmetic progression.*

The next step is to use the *residue theorem* of Cauchy (cf. [42; 130]) to estimate the integral in (22), that is, to find $\widetilde{D}(v,z) = \lim_{T\to\infty} F_T(v,z)$ for every $\tau > s_0(z)$ with $\tau \notin \{\Re(s) : s \in Z(z)\}$ where

$$\begin{aligned} F_T(v,z) &= -\sum_{s'\in Z(z),\ \Re(s')<\tau, |\Im(s')|>T} \mathrm{Res}(\widetilde{D}^*(s,z)\, v^{-s},\ s = s') \\ &\quad + \frac{1}{2\pi i}\int_{\tau-iT}^{\tau+iT} \left(\frac{1-z}{s(1-zp^{1-s}-zq^{1-s})} - \frac{1}{s}\right) v^{-s}\, ds \\ &= -\sum_{s'\in Z(z),\ \Re(s')<\tau, |\Im(s')|>T} \frac{(1-z)v^{-s'}}{zs'p^{1-s'}\ln p + zs'q^{1-s'}\ln q} \\ &\quad + \frac{1}{2\pi i}\int_{\tau-iT}^{\tau+iT} \left(\frac{1-z}{s(1-zp^{1-s}-zq^{1-s})} - \frac{1}{s}\right) v^{-s}\, ds \end{aligned}$$

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 $1/s$. To circumvent this problem, we resort to analyze another integral (cf. [134]), namely

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

Clearly, the Mellin transform $\widetilde{D}_1^*(s,z) = -\widetilde{D}^*(s+1,z)/s$, and therefore it is of order $O(1/s^2)$. Then one can estimate its inverse Mellin as described above. However, after obtaining asymptotics of $\widetilde{D}_1(v,z)$ as $v \to \infty$ one must recover the original asymptotics of $\widetilde{D}(v,z)$. This requires a Tauberian theorem of the following form.

Lemma 6 *Suppose that* $f(v,\lambda)$ *is a non-negative increasing function in* $v \ge 0$*, where* λ *is a real parameter with* $|\lambda| \le \delta$ *for some* $0 < \delta < 1$*. Assume that*

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

has the asymptotic expansion

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

as $v \to \infty$ *and uniformly for* $|\lambda| \le \delta$*. Then*

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

as $v \to \infty$ *and again uniformly for* $|\lambda| \le \delta$*.*

Proof. By the assumption

$$\left|F(v,\lambda) - \frac{v^{\lambda+1}}{\lambda+1}\right| \le \varepsilon|\lambda|\frac{v^{\lambda+1}}{\lambda+1}$$

for $v \ge v_0$ and all $|\lambda| \le \delta$. Set $v' = (\varepsilon|\lambda|)^{1/2}v$. By monotonicity we obtain (for $v \ge v_0$)

$$\begin{aligned}
f(v,\lambda) &\le \frac{F(v+v',\lambda) - F(v,\lambda)}{v'} \\
&\le \frac{1}{v'}\left(\frac{(v+v')^{\lambda+1}}{\lambda+1} - \frac{v^{\lambda+1}}{\lambda+1}\right) + \frac{2}{v'}\varepsilon|\lambda|\frac{(v+v')^{\lambda+1}}{\lambda+1} \\
&= \frac{1}{v'(\lambda+1)}\left(v^{\lambda+1} + (\lambda+1)v^\lambda v' + O(v^{\lambda-1}(v')^2) - v^{\lambda+1}\right) + O\left(\frac{\varepsilon|\lambda|v^{\lambda+1}}{v'}\right) \\
&= v^\lambda + O\left(v^\lambda \varepsilon^{\frac{1}{2}}|\lambda|^{\frac{1}{2}}\right) + O\left(\frac{\varepsilon|\lambda|v^{\lambda+1}}{v'}\right) = v^\lambda + O\left(v^\lambda\varepsilon^{\frac{1}{2}}|\lambda|^{\frac{1}{2}}\right).
\end{aligned}$$

In a similar way we find the corresponding lower bound (for $v \ge v_0 + v_0^{1/2}$), the result follows. ■

Combining Mellin transform, Tauberian theorems and singularity analysis allow us to establish our main results that we present next. The reader is referred to [34] for detailed proofs. First, we apply the above approach to recurrence (19) and arrive at the following.

Theorem 5 *Let $v = 1/r$ in the Khodak's construction and assume $v \to \infty$.*
(i) *If $\log q/\log p$ is irrational, then*

$$M_r = \frac{v}{h_e} + o(v) \tag{25}$$

$h_e = p\ln(1/p) + q\ln(1/q)$ is the entropy rate in natural *units (i.e., $h_e = h\ln 2$). Otherwise, when $\log q/\log p$ is rational, let $L > 0$ is the largest real number for which $\log(1/p)$ and $\log(1/q)$ are integer multiples of L. Then*

$$M_r = \frac{Q_1(\ln v)}{h_e}v + O(v^{1-\eta}) \tag{26}$$

for some $\eta > 0$ where

$$Q_1(x) = \frac{L}{1-e^{-L}}e^{-L\langle\frac{x}{L}\rangle}, \tag{27}$$

and, recall, $\langle y\rangle = y - \lfloor y\rfloor$ is the fractional part of the real number y.

(ii) *If $\log q/\log p$ is irrational, then*

$$\mathbb{E}[D_r] = \widetilde{S}(v,1) = \frac{\lg v}{h} + \frac{h_2}{2h^2} + o(1), \tag{28}$$

while in the rational case

$$\mathbb{E}[D_r] = \widetilde{S}(v,1) = \frac{\lg v}{h} + \frac{h_2}{2h^2} + \frac{Q_2(\ln v)}{h\ln 2} + O(v^{-\eta}) \tag{29}$$

for some $\eta > 0$, where

$$Q_2(x) = L\cdot\left(\frac{1}{2} - \left\langle\frac{x}{L}\right\rangle\right) \tag{30}$$

and $h_2 = p\lg^2(1/p) + q\lg^2(1/q)$.

Using these findings and using similar but more sophisticated analysis we obtain out next main result.

Theorem 6 *Let D_r denote the phrase length in Khodak's construction with parameter r of the Tunstall code with a dictionary of size M_r over a biased memoryless source. Then as $M_r \to \infty$*

$$\frac{D_r - \frac{1}{h}\lg M_r}{\sqrt{\left(\frac{h_2}{h^3} - \frac{1}{h}\right)\lg M_r}} \to N(0,1)$$

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

$$\mathbb{V}ar[D_r] = \left(\frac{h_2}{h^3} - \frac{1}{h}\right)\lg M_r + O(1)$$

for large M_r.

By combining (25) and (28) resp. (26) and (29) we can be even more precise. In the irrational case we have

$$\mathbb{E}[D_r] = \frac{\lg M_r}{h} + \frac{\lg(h\ln 2)}{h} + \frac{h_2}{2h^2} + o(1)$$

and in the rational case we find

$$\mathbb{E}[D_r] \quad = \quad \frac{\lg M_r}{h} + \frac{\lg(h\ln 2)}{h} + \frac{h_2}{2h^2} + \frac{-\lg L + \lg(1-e^{-L}) + L\lg(e)/2}{h} + O((M_r^{-\eta}),$$

so that there is actually no oscillation. Recall, $L > 0$ is the largest real number for which $\ln(1/p)$ and $\ln(1/q)$ are integer multiples of L.

As a direct consequence, we can derive a precise asymptotic formula for the average redundancy of the Khodak code, that is,

$$\overline{r}_M^K = \frac{\lg M}{\mathbb{E}[D]} - h\,.$$

The following result is a consequence of the above derivations.

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

$$\overline{r}_{M_r}^K = \frac{h}{\lg M_r}\left(-\frac{h_2\ln 2}{2h} - \lg(h\ln 2)\right) + o\left(\frac{1}{\log M_r}\right).$$

In the rational case we have

$$\overline{r}_{M_r}^K \quad = \quad \frac{h}{\lg M_r}\Big(- \frac{h_2\ln 2}{2h} - \lg(h\ln 2) - \lg\left(\frac{\sinh(L/2)}{L/2}\right)\Big) + O\left(\frac{1}{\log^2 M_r}\right),$$

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.

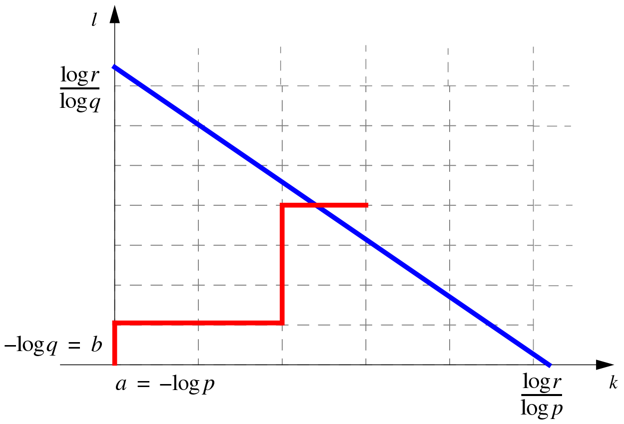


Fig. 4: A random walk with a linear barrier; the exit time is equivalent to the phrase length in the Khodak algorithm (e.g., the exit time $= 7$).

Let us offer some final remarks. We already observed that the parsing trees for the Tunstall and Khodak algorithms are the same except when there is a "tie". In the case of a tie Khodak algorithm develops all nodes with the tie simultaneously while the Tunstall algorithm expends one node after another. This situation can occur both, for the rational case and for the irrational case, and somewhat surprisingly leads to the cancelation of oscillation in the redundancy of the Khodak code for the rational case. As shown in [112] tiny oscillations remain in the Tunstall code redundancy for the rational case. But as easy to see that Central Limit Theorem holds also for the Tunstall code as shown [34].

Finally, we relate our results to certain problems on random walks. As already observed in [112], a path in the parsing tree from the root to a leaf corresponds to a random walk on a lattice in the first quadrant of the plane (cf. Figure 4). Indeed, observe that our analysis of the Khodak code boils down to studying the following sum

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

for some function $f(v)$. Since $P(y) = p^k q^l$ for some nonnegative integers $k, l \geq 0$, we conclude that the summation set of $A(v)$ can be expressed, after setting $v = 2^V$, as

$$k \lg(1/p) + l \lg(1/q) \leq V.$$

This corresponds to a random walk in the first quadrant with the linear boundary condition $ax + by = V$, where $a = \log(1/p)$ and $b = \log(1/q)$ as shown in Figure 4. The phrase length coincides with the exit time of such a random walk (i.e., the last step before the random walk hits the linear boundary). This correspondence is further explored in [31; 62].

5 Redundancy of Khodak VV Code

Recall that a variable-to-variable (VV) length code partitions a source sequence into variable length phrases that are encoded into strings of variable lengths. While it is well known that every VV (prefix) code is a concatenation of a variable-to-fixed length code (e.g., Tunstall code) and a fixed-to-variable length encoding (e.g., Huffman code), an optimal VV code has not yet been found. Fabris [36] proved that greedy, step by step, optimization (that is, a concatenation of Tunstall and Huffman codes) does not lead to an optimal VV code. In this section, we analyze an interesting VV code due to Khodak [65].

Recall that in (10) we define the average redundancy rate as

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

becomes after using renewal theory as in (11)

$$\overline{r} \;=\; \frac{\sum_{d\in\mathcal{D}} P(d)\ell(d) - h_{\mathcal{D}}}{\mathbb{E}[D]} = \frac{\sum_{d\in\mathcal{D}} P(d)(\ell(d) + \lg P(d))}{\mathbb{E}[D]}, \tag{31}$$

where P is the probability law of the dictionary phrases and $\mathbb{E}[D] = \sum_{d\in\mathcal{D}} |d|P(d)$. From now on we shall write $\overline{D} := \mathbb{E}[D]$.

In previous sections we analyzed FV and VF codes. We prove that the average redundancy rate (per block in the case of FV codes) is $O(1/\overline{D})$. It is an intriguing question whether one can construct a code with $\overline{r} = o(1/\overline{D})$. This quest was accomplished by Khodak [65] in 1972 who proved that one can find a VV code with $\overline{r} = O(\overline{D}^{-5/3})$. However, the proof presented in [65] is rather sketchy and complicated. Here we present a transparent proof proposed in [12] of the following main result of this section.

Theorem 7 *For every $D_0 \geq 1$, there exists a VV code with average delay $\overline{D} \geq D_0$ such that its average redundancy rate satisfies*

$$\overline{r} = O(\overline{D}^{-5/3}) \tag{32}$$

and the average code length is $O(\overline{D}\log\overline{D})$.

The rest of this section is devoted to describe a proof of Theorem 7 presented in [12]. We assume an m-ary alphabet $\mathcal{A} = \{a_1,\ldots,a_m\}$ with probability of symbols $p_1,\ldots,p_m$. Let us first give some intuitions. For every $d \in \mathcal{D}$ we can represent $P(d)$ as $P(d) = p_1^{k_1}\cdots p_m^{k_m}$, where $k_i = k_i(d)$ is the number of times symbol a_i appears in d. In what follows we write $\mathrm{type}(d) = (k_1, k_2, \ldots, k_m)$ for all strings with the same probability $P(d) = p_1^{k_1}\cdots p_m^{k_m}$. Furthermore, the string encoder of our VV code uses a slightly modified Shannon code that assigns to $d \in \mathcal{D}$ a binary word of length $\ell(d)$ close to $-\log P(d)$ when $\log P(d)$ is slightly larger or smaller than an integer. (Kraft's inequality will not be automatically satisfied but Lemma 9 below takes care of it.) Observe that the average redundancy of Shannon code is

$$\sum_{d\in\mathcal{D}} P(d)[\lceil -\log P(d)\rceil + \log P(d)] = \sum_{d\in\mathcal{D}} P(d)\cdot\langle k_1(d)\gamma_1 + k_2(d)\gamma_2 + \cdots + k_m(d)\gamma_m\rangle$$

where $\gamma_i = \log p_i$. In order to build a VV code with $\overline{r} = o(1/\overline{D})$, we are to find integers $k_1 = k_1(d), \ldots k_m = k_m(d)$ such that the linear form $k_1\gamma_1 + k_2\gamma_2 + \cdots + k_m\gamma_m$ is close to an integer. In the sequel, we discuss some properties of the distribution of $\langle k_1\gamma_1 + k_2\gamma_2 + \cdots + k_m\gamma_m\rangle$ when at least one of γ_i is irrational (cf. [32]).

Let $\|x\| = \min(\langle x\rangle, \langle -x\rangle) = \min(\langle x\rangle, 1 - \langle x\rangle)$ be the distance to the nearest integer. The *dispersion* $\delta(X)$ of the set $X \subseteq [0,1)$ is defined as

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

that is, for every $y \in [0,1)$ there exists $x \in X$ with $\|y - x\| \leq \delta(X)$. Since $\|y + 1\| = \|y\|$, the same assertion holds for all real y. Dispersion tells us that points of X are at most $2\delta(X)$ apart in $[0,1]$. Therefore, there exist distinct points $x_1, x_2 \in X$ with $\langle y - x_1\rangle \leq 2\delta(X)$ and $\langle y - x_2\rangle \leq 2\delta(X)$.

The following property will be used throughout this paper. This is a standard result following from Dirichlet's approximation theorem, so we leave it for the reader to prove it (cf. [32]).

Lemma 7 (i) *Suppose that θ is an irrational number. There exists an integer N such that*

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

(ii) *In general, let $(\gamma_1, \dots, \gamma_m)$ be an m-vector of real numbers such that at least one of its coordinates is irrational. There exists an integer N such that the dispersion of the set is*

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

is bounded by

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

The central step of all existence results is the observation that a bound on the dispersion of linear forms of $\log_2 p_j$ implies the existence of a VV code with small redundancy. Indeed, our main result of this section follows directly from the below lemma whose proof is presented below.

Lemma 8 *Let $p_j > 0$ $(1 \le j \le m)$ with $p_1 + \cdots + p_m = 1$ be given and suppose that for some $N \ge 1$ and $\eta \ge 1$ the set*

$$X = \{\langle k_1' \log_2 p_1 + \cdots + k_m' \log_2 p_m\rangle : 0 \le k_j' < N \ (1 \le j \le m)\},$$

has dispersion

$$\delta(X) \le \frac{2}{N^\eta}. \tag{33}$$

Then there exists a VV code with the average code length $\overline{D} = \Theta(N^3)$, the maximal length of order $\Theta(N^3 \log N)$, and the average redundancy rate

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

Clearly, Lemma 7 and Lemma 8 directly imply Theorem 7 by setting $\eta = 1$ if one of the $\log_2 p_j$ is irrational. (If all $\log_2 p_j$ are rational, then the construction is simple).

We now concentrate on proving Lemma 8. The main thrust of the proof is to construct a complete prefix free set $\mathcal{D}$ of words (i.e., a dictionary) on an alphabet of size m such that $\log_2 P(d)$ is *very close* to an integer $\ell(d)$ with high probability. This is accomplished by growing an m-ary tree $\mathcal{T}$ in which paths from the root to terminal nodes have $\log P(d)$ close to an integer.

In the first step, we set $k_i^0 := \lfloor p_i N^2 \rfloor$ $(1 \le i \le m)$ and define

$$x = k_1^0 \log_2 p_1 + \cdots + k_m^0 \log_2 p_m.$$

By our assumption (33) of Lemma 8, there exist integers $0 \le k_j^1 < N$ such that

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

Now consider all paths in a (potentially) infinite m-ary tree starting at the root with $k_1^0 + k_1^1$ edges of type $a_1 \in \mathcal{A}$, $k_2^0 + k_2^1$ edges of type $a_2 \in \mathcal{A}$,. . ., and $k_m^0 + k_m^1$ edges of type $a_m \in \mathcal{A}$ (cf. Figure 5). Let $\mathcal{D}_1$

denote the set of such words. (These are the first words of our prefix free set we are going to construct.) By an application of Stirling's formula it follows that there are two positive constants c', c'' such that

$$\frac{c'}{N} \le P(\mathcal{D}_1) = \binom{(k_1^0+k_1^1)+\cdots+(k_m^0+k_m^1)}{k_1^0+k_1^1,\ldots,k_m^0+k_m^1} p_1^{k_1^0+k_1^1}\cdots p_m^{k_m^0+k_m^1} \le \frac{c''}{N} \tag{34}$$

uniformly for all k_j^1 with $0 \le k_j^1 < N$. In summary, by construction all words $d \in \mathcal{D}_1$ have the property that

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

that is, $\log_2 P(d)$ is very close to an integer. Note further that all words in $d \in \mathcal{D}_1$ have about the same length

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

and words in $\mathcal{D}_1$ constitute the first crop of "good words". Finally, let $\mathcal{B}_1 = \mathcal{A}^{n_1} \setminus \mathcal{D}_1$ denote all words of length n_1 not in $\mathcal{D}_1$ (cf. Figure 5). Then

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

In the second step, we consider all words $r \in \mathcal{B}_1$ and concatenate them with appropriately chosen words d_2 of length $\sim N^2$ such that $\log_2 P(rd_2)$ is close to an integer *with high probability*. The construction is almost the same as in the first step. 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 (33) there exist integers $0 \le k_j^2(r) < N$ $(1 \le j \le m)$ such that

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

Now consider all paths (in the infinite tree $\mathcal{T}$) starting at $r \in \mathcal{B}_1$ with $k_1^0 + k_1^2(r)$ edges of type a_1, $k_2^0 + k_2^2(r)$ edges of type a_2, …, and $k_m^0 + k_m^2(r)$ edges of type a_m (that is, we concatenated r with properly chosen words d_2) and denote this set by $\mathcal{D}_2^+(r)$. We again have that the total probability of these words is bounded from below and above by

$$P(r)\frac{c'}{N} \le P(\mathcal{D}_2(r)) = P(r)\binom{(k_1^0+k_1^2(r))+\cdots+(k_m^0+k_m^2(r))}{k_1^0+k_1^2(r),\ldots,k_m^0+k_m^2(r)} p_1^{k_1^0+k_1^2(r)}\cdots p_m^{k_m^0+k_m^2(r)} \le P(r)\frac{c''}{N}.$$

Furthermore, by construction we have $\langle \log_2 P(d)\rangle < \frac{4}{N^\eta}$ for all $d \in \mathcal{D}_2^+(r)$.

Similarly, we can construct a set $\mathcal{D}_2^-(r)$ instead of $\mathcal{D}_2^+(r)$ for which we have $1 - \langle \log_2 P(d) \rangle < 4/N^\eta$. We will indicate in the sequel whether we will use $\mathcal{D}_2^+(r)$ or $\mathcal{D}_2^-(r)$.

Let $\mathcal{D}_2 = \bigcup(\mathcal{D}_2^+(r) : r \in \mathcal{B}_1)$ (or $\mathcal{D}_2 = \bigcup(\mathcal{D}_2^-(r) : r \in \mathcal{B}_1)$). Then all words $d \in \mathcal{D}_2$ have almost the same length $|d| = 2N^2 + O(2N)$, their probabilities satisfy

$$\langle \log_2 P(d) \rangle < \frac{4}{N^\eta} \quad \left(\text{or} \quad 1 - \langle \log_2 P(d) \rangle < \frac{4}{N^\eta}\right)$$

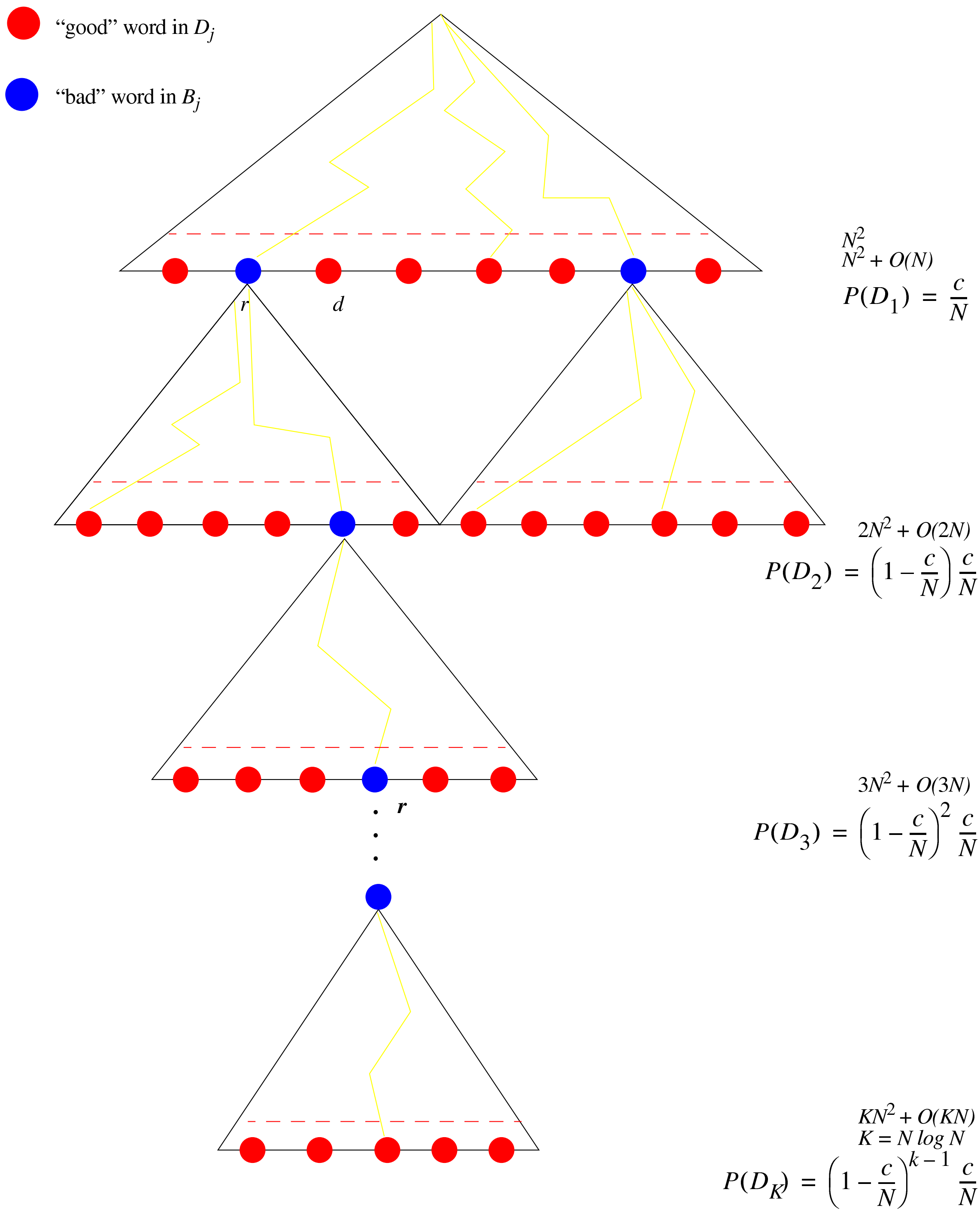


Fig. 5: Illustration to the construction of the VV code.

and the total probability is bounded by

$$\frac{c'}{N}\left(1-\frac{c''}{N}\right) \le P(\mathcal{D}_2) \le \frac{c''}{N}\left(1-\frac{c'}{N}\right).$$

For every $r \in \mathcal{B}_1$, let $\mathcal{B}^+(r)$ (or $\mathcal{B}^-(r)$) denote the set of paths (resp. words) starting with r of length $2(k_1^0+\cdots+k_m^0)+(k_1^1+k_1^2(r)+\cdots+k_m^1+k_m^2(r))$ that are *not* contained in $\mathcal{D}_2^+(r)$ (or $\mathcal{D}_2^-(r)$) and set $\mathcal{B}_2 = \bigcup(\mathcal{B}_2^+(r) : r \in \mathcal{B}_1)$ (or $\mathcal{B}_2 = \bigcup(\mathcal{B}_2^-(r) : r \in \mathcal{B}_1)$). Observe that the probability of $\mathcal{B}_2$ is bounded by

$$\left(1-\frac{c''}{N}\right)^2 \le P(\mathcal{B}_2) \le \left(1-\frac{c'}{N}\right)^2.$$

We continue this construction, as illustrated in Figure 5, and in step j we define sets of words $\mathcal{D}_j$ and $\mathcal{B}_j$ such that all words $d \in \mathcal{D}_j$ satisfy

$$\langle \log_2 P(d)\rangle < \frac{4}{N^\eta} \qquad \left(\text{or} \quad 1-\langle \log_2 P(d)\rangle < \frac{4}{N^\eta}\right)$$

and the length of $d \in \mathcal{D}_j \cup \mathcal{B}_j$ is then given by $|d| = jN^2 + \mathcal{O}(jN)$. The probabilities of $\mathcal{D}_j$ and $\mathcal{B}_j$ are bounded by

$$\frac{c'}{N}\left(1-\frac{c''}{N}\right)^{j-1} \le P(\mathcal{D}_j) \le \frac{c''}{N}\left(1-\frac{c'}{N}\right)^{j-1},$$

and

$$\left(1-\frac{c''}{N}\right)^j \le P(\mathcal{B}_j) \le \left(1-\frac{c'}{N}\right)^j.$$

This construction is terminated 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

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

The complete prefix free set $\mathcal{D}$ on the m-ary alphabet is given by $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K \cup \mathcal{B}_K$.

By the above construction, it is also clear that the average delay is bounded by

$$c_1N^3 \le \overline{D} = \sum_{d\in\mathcal{D}} P(d)\,|d| \le c_2N^3$$

for certain constants $c_1, c_2 > 0$. Notice further that the maximal code length satisfies

$$\max_{d\in\mathcal{D}} |d| = \mathcal{O}\left(N^3\log N\right) = \mathcal{O}\left(\overline{D}\log\overline{D}\right).$$

Now we construct a variant of the Shannon code with $\overline{r} = o(1/\overline{D})$. For every $d \in \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$ we can choose a non-negative integer $\ell(d)$ with

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

In particular, we have

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

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

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

if $1 - \langle \log_2 P(d) \rangle < 2/N^\eta$. For $d \in \mathcal{B}_K$ we simply set $\ell(d) = \lceil -\log_2 P(d) \rceil$. The final problem is now to *adjust* the choices of "$+$" resp. "$-$" in the above construction so that Kraft's inequality is satisfied. For this purpose we use the following easy property (that we adopt from Khodak [65]).

Lemma 9 (Khodak, 1972) *Let $\mathcal{D}$ be a finite set with probability distribution P and suppose that for every $d \in \mathcal{D}$ we have $|\ell(d) + \log_2 P(d)| \le 1$ for a nonnegative integer $\ell(d)$. If*

$$\sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d)) \ge 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d))^2, \tag{35}$$

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

Proof. We use the local expansion $2^{-x} = 1 - x\log 2 + \eta(x)$ for $|x| \le 1$, where $((\log 4)/4)x^2 \le \eta(x) \le (\log 4)x^2$. Hence

$$\begin{aligned}
\sum_{d\in\mathcal{D}} 2^{-\ell(d)} &= \sum_{d\in\mathcal{D}|} P(d) 2^{-(\ell(d)+\log_2 P(d))} \\
&= 1 - \log 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d)) + \sum_{d\in\mathcal{D}} P(d)\eta\left(\ell(d) + \log_2 P(d)\right) \\
&\le 1 - \log 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d)) + 2\log 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d))^2 \\
&\overset{(35)}{\le} 1
\end{aligned}$$

If (35) is satisfied, then Kraft's inequality follows, and there exists an injective mapping $C : \mathcal{D} \to \{0,1\}^*$ such that C is a prefix free set and $|C(d)| = \ell(d)$ for all $d \in \mathcal{D}$. ■

Applying the above lemma, after some tedious algebra, we arrive at the following bound on the average redundancy rate

$$\overline{r} \le \frac{1}{\overline{D}} \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \lg P(d)) \le C \frac{1}{\overline{D} N^{1+\eta}}.$$

Since the average code length $\overline{D}$ is of order N^3 we have

$$\overline{r} = O\left(\overline{D}^{-1-\frac{1+\eta}{3}}\right) = O\left(\overline{D}^{-\frac{4+\eta}{3}}\right).$$

This proves the upper bound for $\overline{r}$ of Lemma 8 and Theorem 7 follows.

6 Generalization and Concluding Remarks

In this concluding section, we address two problems: universal codes and non-prefix codes. In particular, we analyze the average redundancy of Shannon code when the source distribution is *unknown*. Then we construct a one-to-one code whose average length is smaller than the source entropy in defiance of the Shannon lower bound. To focus, we only consider fixed-to-variable codes with block size equal to n over binary $\alpha^* = \{0,1\}^*$ sequences generated by a memoryless source.

6.1 Universal Codes

We study here a FV code over a binary memoryless source with *unknown* parameter θ. The probability of a sequence $x_1^n = x! \dots x_n$ of length n is $P(x_1^n) = \theta^k(1-\theta)^{n-k}$, where k is the number of "1"s. To apply any FV code, say Shannon's code, we need to estimate θ. There are several algorithms to accomplish it. We select θ that minimizes the ***Minimum Description Length*** (MDL) by applying the Krichevsky–Trofimov (KT) estimator [76; 143]. The KT-estimator is defined by the following conditional probability

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

where k is the number of "1"s in the sequence x_1^{n-1}. Thus, the probability $P_e(k, n-k)$ of a sequence of k ones and $n-k$ zeros is

$$P_e(k, n-k) = \frac{1/2 \cdot \ldots \cdot (k-1/2) \cdot 1/2 \cdot \ldots \cdot (n-k-1/2)}{n!},$$

which can be also written as

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

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

Let us now choose a source coding, say the Shannon-Fano code (cf. [19; 49]) which assigns the code length $L_n = \lceil -\log P_e(k, n-k) \rceil + 1$. The average redundancy of such a code is

$$\overline{R}_n^{SF} = 1 + \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \left(\lceil -\log P_e(k, n-k) \rceil + \log \theta^k (1-\theta)^{n-k} \right).$$

Using $\lceil -x \rceil = -x + 1 - \langle -x \rangle$ we reduce the above to the following

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

where

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

The main result of this section is presented next. Its complete proof can be found in [28].

Theorem 8 *Consider the Shannon-Fano code over a memoryless(θ) source. Then*

$$\overline{R}_n^{SF} = \frac{1}{2}\log n - \frac{1}{2}\log\frac{\pi e}{2} + 2 - E_n + O(n^{-1/2}) \tag{36}$$

where E_n behavior depends whether $\alpha = \log\frac{1-\theta}{\theta}$ is rational or not, that is:

(i) *If $\alpha = \log\frac{1-\theta}{\theta}$ is rational, i.e. $\alpha = \frac{N}{M}$ for some positive integers M, N with $\gcd(M,N) = 1$, then*

$$E_n = \frac{1}{2} + G_M\left(-\log(1-\theta)n + \frac{1}{2}\log\frac{\pi n}{2}\right) + o(1) \tag{37}$$

as $n \to \infty$, where

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

is a periodic function with period $\frac{1}{M}$ and maximum $\max|G_M| \le \frac{1}{2M}$.

(ii) *If $\alpha = \log\frac{1-\theta}{\theta}$ is irrational, then*

$$E_n = \frac{1}{2} + o(1) \tag{38}$$

as $n \to \infty$.

Proof. We sketch the proof. We start with the main part of $\overline{R}_n^{SF}$ and then we deal with E_n. Our proof first approximates the binomial distribution by its Gaussian density, and then estimates the sum by the Gaussian integral, coupling with large deviations of the binomial distribution. By Stirling's formula, we have

$$\log\frac{\theta^k(1-\theta)^{n-k}}{P_e(k,n-k)} = \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} - \frac{x^2}{2\ln 2} + O((|x| + |x|^3)n^{-1/2}),$$

for $k = \theta n + x\sqrt{\theta(1-\theta)n}$ and $x = o(n^{1/6})$. Note that the left-hand side is bounded above by $\frac{1}{2}\log n + 1/2$ for $n \ge 2$ and $k \ne 0, n$. This follows easily from the identity

$$\Gamma(n + 1/2) = \frac{(2n)!\sqrt{\pi}}{4^n\, n!} \qquad (n \ge 0),$$

and the inequalities

$$\sqrt{2\pi n}(n/e)^n \le n! \le e^{1/12}\sqrt{2\pi n}(n/e)^n, \qquad (n \ge 1).$$

On the other hand, by using the local limit theorem for the binomial distribution we arrive at

$$\binom{n}{k}\theta^k(1-\theta)^{n-k} = \frac{e^{-x^2/2}}{\sqrt{2\pi\theta(1-\theta)n}}\left(1 + O((1+|x|^3)n^{-1/2})\right), \tag{39}$$

uniformly for $x = o(n^{1/6})$, we deduce that

$$\overline{R}_n^{SF} - E_n = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\left(\frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} - \frac{x^2}{2\ln 2}\right)\mathrm{d}x + O(n^{-1/2}).$$

A straightforward evaluation of the integral leads to (36).

In order to evaluate E_n we need to appeal to theory of sequences modulo 1 as in Section 3. We need to generalize Lemmas 2 and 3 proved in [28].

Lemma 10 *Let $0 < p < 1$ be a fixed real number and $f : [0,1] \to \mathbb{R}$ be a Riemann integrable function.*

(i) *If α is irrational, then*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\rangle\right) = \int_0^1 f(t)\,dt, \tag{40}$$

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

(ii) *Suppose that $\alpha = \frac{N}{M}$ is a rational number with integers N, M such that $\gcd(N, M) = 1$. Then uniformly for all $y \in \mathbb{R}$*

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\rangle\right) = \int_0^1 f(t)\,dt + G_M(y) \tag{41}$$

where

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

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

To estimate E_n we need to set $f(t) = t$ in Lemma 10, and this completes the proof of Theorem 8. ■

Theorem 8 is quite revealing. In previous sections we proved that for *known* sources P the average redundancy of FV codes is $\overline{R} = O(1)$ as the length of the sequence increases. However, if one needs to estimate one parameter, θ in our case, the penalty incurred increases to $\frac{1}{2}\log n + O(1)$. In general, if there are $m-1$ unknown parameters, the average redundancy is

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

as predicted by Rissanen's lower bound [6; 104].

6.2 One-to-One Codes Violating Kraft's Inequality

Finally, we discuss a code known as the *one-to-one* code that is *not* a prefix code, and therefore doesn't satisfy the Kraft's inequality. We show that for such codes the Shannon lower bound doesn't apply.

We consider again a binary memoryless source X over the binary alphabet $\mathcal{A} = \{0,1\}$ generating a sequence $x_1^n = x_1, \ldots, x_n \in \mathcal{A}^n$ with probability $P(x_1^n) = p^k q^{n-k}$, where k is the number of 0's in x_1^n and p is known. We shall assume that $p \le q$. We first list all 2^n probabilities in a nonincreasing order and assign code lengths to them as shown below:

$$\begin{array}{lccccccc} \text{probabilities} & q^n\left(\frac{p}{q}\right)^0 & \ge & q^n\left(\frac{p}{q}\right)^1 & \ge & \cdots & \ge & q^n\left(\frac{p}{q}\right)^n, \\ \text{code lengths} & \lfloor \log_2(1) \rfloor & & \lfloor \log_2(2) \rfloor & & \cdots & & \lfloor \log_2(2^n) \rfloor. \end{array}$$

Observe that there are $\binom{n}{k}$ equal probabilities $p^k q^{n-k}$ that are assigned different code lengths. More precisely, define

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

Starting from the position $A_{k-1}+1$ of the above list , the next $\binom{n}{k}$ probabilities are the same and equal to $p^k q^{n-k}$. For each $j = A_{k-1} + i$, $1 \le i \le \binom{n}{k}$, we assign the code length

$$\lfloor \log_2(j) \rfloor = \lfloor \log_2(A_{k-1}+i) \rfloor$$

to the jth binary string. Thus the average code length is

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

Our goal is to estimate $\mathbb{E}[L_n]$ asymptotically for large n and the average redundancy

$$\overline{R}_n = \mathbb{E}[L_n] - nh(p)$$

where $h(p) = -p \lg p - q \lg q$ is the binary entropy.

Let us first simplify the formula for $\mathbb{E}[L_n]$. We need to handle the inner sum that contains the floor function. To evaluate this sum we apply the following identity (cf. Knuth [70] Ex. 1.2.4-42)

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

for any sequence a_j. Using this, we easily find an explicit formula for the inner sum of (42), namely

$$\begin{aligned} S_{n,k} \quad &= \quad \sum_{j=1}^{\binom{n}{k}} \lfloor \log_2(A_{k-1}+j) \rfloor = \binom{n}{k} \lfloor \log_2 A_k \rfloor - (2^{\lfloor \log_2(A_k) \rfloor + 1} - 2^{\lceil \log_2(A_{k-1}+2) \rceil}) \\ &+ \quad (A_{k-1}+1)(1 + \lfloor \log_2(A_k) \rfloor - \lceil \log_2(A_{k-1}+2) \rceil). \end{aligned}$$

After some algebra, using $\lfloor x \rfloor = x - \langle x \rangle$ and $\lceil x \rceil = x + \langle -x \rangle$, we finally reduce the formula for $\mathbb{E}[L_n]$

to the following

$$\begin{aligned}
\mathbb{E}[L_n] &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log_2 A_k \rfloor && (42)\\
&- 2\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} 2^{-\langle \log_2 A_k \rangle} && (43)\\
&+ \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{1+A_{k-1}}{\binom{n}{k}} \left(1+\log_2\left(\frac{A_k}{A_{k-1}+2}\right) - \langle -\log_2(A_{k-1}+2)\rangle - \langle \log_2 A_k \rangle\right)\\
&- \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{A_{k-1}}{\binom{n}{k}} \left(2^{-\langle \log_2 A_k\rangle+1} - 2^{\langle -\log_2(A_{k-1}+2)\rangle}\right)\\
&+ 2\sum_{k=0}^{n} p^k q^{n-k} 2^{\langle -\log_2(A_{k-1}+2)\rangle}.
\end{aligned}$$

Our main result proved in [131] is presented next.

Theorem 9 *Consider a binary memoryless source and the one-to-one block code described above. Then for* $p < \frac{1}{2}$

$$\begin{aligned}
\overline{R}_n &= -\frac{1}{2}\log_2 n - \frac{3+\ln(2)}{2\ln(2)} + \log_2 \frac{1-p}{1-2p}\frac{1}{\sqrt{2\pi p(1-p)}} + \frac{p}{1-2p}\log_2\left(\frac{2(1-p)}{p}\right)\\
&+ F(n) + o(1) \qquad (44)
\end{aligned}$$

where as before $\alpha = \log_2(1-p)/p$, $\beta = \log_2(1/(1-p)$ *and* $F(n) = 0$ *if* $\log_2 \frac{1-p}{p}$ *is irrational. If* $\log_2 \frac{1-p}{p} = N/M$ *for some integers* M, N *such that* $\gcd(N, M) = 1$*, then*

$$F(n) = -\frac{1-p}{1-2p}H_M(n\beta)[x] - \frac{p}{1-2p}H_M(n\beta-\alpha)[-x] - \frac{2(1-3p)}{1-2p}H_M(n\beta)[2^{-x}] + \frac{p}{1-2p}H_M(n\beta-\alpha)[2^x]$$

where (cf. Lemma 10)

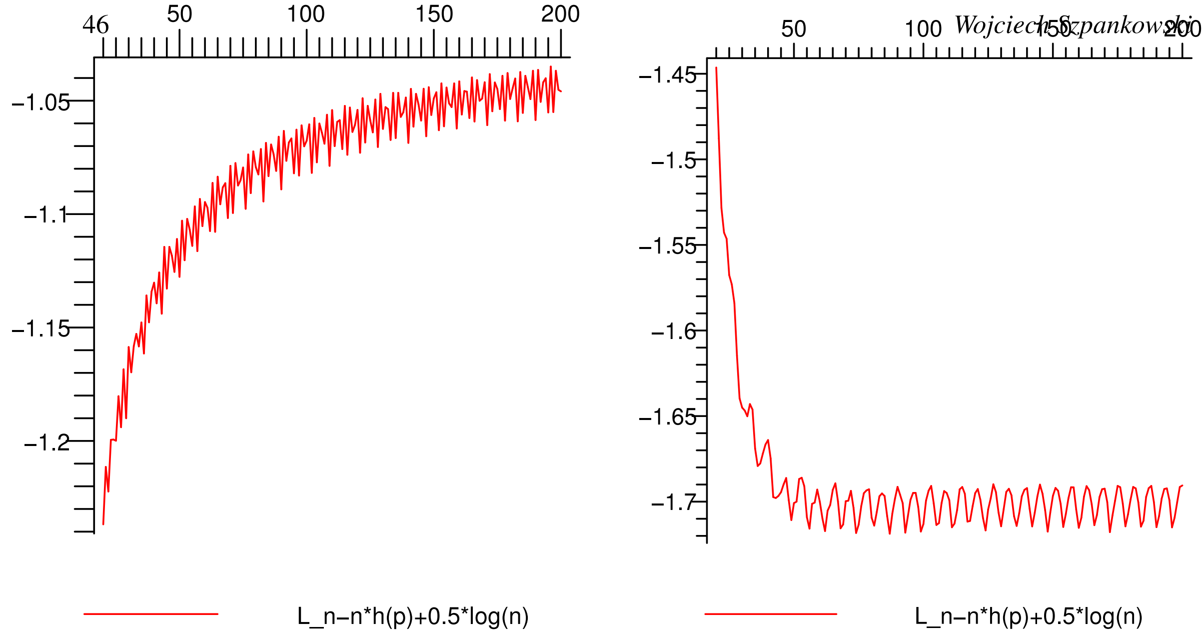
$$H_M(y)[f] := \frac{1}{M\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\left(\left\langle M\left(y - \log_2\left(\frac{1-2p}{1-p}\sqrt{2\pi pqn}\right) - \frac{x^2}{2\ln 2}\right)\right\rangle - \int_0^1 f(t)dt\right)dx$$

for some Riemann function f.
For $p = \frac{1}{2}$*, we have*

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

for every $n \geq 1$.

Some observations are in order. First, the average redundancy $\overline{R}_n$ is *negative*(!) for non-prefix codes such as one-to-one codes. This was already observed by Wyner in 1972 [146] and also discussed in [2]. Second, in view of Theorem 9, we again see that asymptotic behavior of the redundancy depends on the rationality/irrationality of $\alpha = \log_2(1-p)/p$ (cf. [28; 29; 129]). In Figure 6 we plot $\overline{R}_n +$



(a) (b)

Fig. 6: Plots of $L_n - nh(p) + 0.5\log(n)$ (y-axis) versus n (x-axis) for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

$0.5\log_2(n)$ versus n. We observe change of "mode" from a "converging mode" to a "fluctuating mode", when switching from $\alpha = \log_2(1-p)/p$ irrational (cf. Fig. 6(a)) to rational (cf. Fig. 6(b)). Recall that we saw this already for Huffman, Shannon, and Tunstall codes.

We only briefly sketch the proof of Theorem 9. We only analyze here (42) which we re-write as follows

$$\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\lfloor\log_2 A_k\rfloor = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\log_2 A_k - \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\langle\log_2 A_k\rangle,$$

and define

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

We first deal with a_n for which we need to derive a precise asymptotic estimate for A_n. But this is a simple exercise of the saddle point method [42; 130] as presented below.

Lemma 11 *For large n and $p < 1/2$*

$$A_{np} = \frac{1-p}{1-2p}\frac{1}{\sqrt{2\pi np(1-p)}}2^{nh(p)}\left(1 + O(n^{-1/2})\right) \tag{45}$$

where $h(p)$ is the binary entropy. More precisely, for an $\varepsilon > 0$ and $k = np + \Theta(n^{1/2+\varepsilon})$ we have

$$A_k = \frac{1-p}{1-2p}\frac{1}{\sqrt{2\pi np(1-p)}}\left(\frac{1-p}{p}\right)^k \frac{1}{(1-p)^n}\exp\left(-\frac{(k-np)^2}{2p(1-p)n}\right)\left(1+O(n^{-\delta})\right) \quad (46)$$

for some $\delta > 0$.

Proof. We use the saddle point method [130]. Let's first define the generating function of A_k, that is,

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

Thus by Cauchy's formula [130]

$$\begin{aligned} A_k &= \frac{1}{2\pi i}\oint \frac{(1+z)^n - 2^n z^{n+1}}{1-z}\frac{dz}{z^{k+1}} \\ &= \frac{1}{2\pi i}\oint \frac{1}{1-z} 2^{n\log(1+z)-(k+1)\log z} dz. \end{aligned}$$

Define $H(z) = n\log(1+z) - (k+1)\log z$. The saddle point z_0 solves $H'(z_0) = 0$, and one finds $z_0 = (k+1)/(n-k+1) = p/(1-p) + O(1/n)$ for $k = np$ and $H''(z_0) = q^3/p$. Thus by the saddle point method

$$A_k = \frac{1}{1-z_0}\frac{1}{\sqrt{2\pi n H''(z_0)}} 2^{nH(z_0)}(1+O(n^{-1/2})).$$

This proves (45). In a similar manner, as shown in [27], we establish (46). ∎

For b_n we need to appeal to Lemma 10 after observing that for $|k-pn| \le n^{1/2+\varepsilon}$

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

where $\omega = (1-2p)\sqrt{2\pi pq}/(1-p)$. Thus, we need asymptotics of

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

that is discussed in Lemma 10. Details of the proof can be found in [131].

Acknowledgment

This survey couldn't be written without able help of many of my co-authors: M. Drmota (TU Wien, Austria), P. Flajolet (INRIA, France), P. Jacquet (INRIA, France), Y. Reznik (Qualcom Inc.), and S. Savari (Texas A&M, USA).

References

[1] J. Abrahams, Code and parse trees for lossless source encoding, *Communications in Information and Systems* 1,113-146, 2001.

[2] N. Alon and A. Orlitsky, A Lower Bound on the Expected Length of One-to-One Codes, *IEEE Trans. Information Theory*, 40, 1670-1672, 1994.

[3] K. Atteson, The Asymptotic Redundancy of Bayes Rules for Markov Chains, *IEEE Trans. on Information Theory*, 45, 2104–2109, 1999.

[4] R. C. Baker, Dirichlet's Theorem on Diophantine Approximation, *Math. Proc. Cambridge Philos. Soc.* 83, 37–59, 1978.

[5] A. Barron, *Logically Smooth Density Estimation*, Ph.D. Thesis, Stanford University, Stanford, CA, 1985.

[6] A. Barron, J. Rissanen, and B. Yu, The Minimum Description Length Principle in Coding and Modeling, *IEEE Trans. Information Theory*, 44, 2743-2760, 1998.

[7] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, 1971.

[8] J. Bernardo, Reference Posterior Distributions for Bayesian Inference, *J. Roy. Stat. Soc. B.*, 41, 113–147, 1979.

[9] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York 1968.

[10] P. Billingsley, Statistical Methods in Markov Chains, *Ann. Math. Statistics*, 32, 12–40, 1961.

[11] L. Boza, Asymptotically Optimal Tests for Finite Markov Chains, *Ann. Math. Statistics*, 42, 1992-2007, 1971.

[12] Y. Bugeaud, M. Drmota and W. Szpankowski, On the Construction of (Explicit) Khodak's Code and Its Analysis, *IEEE Trans. Information Theory*, 54, 2008.

[13] J. W. S. Cassels, *An Introduction to Diophantine Approximation*, Cambridge University Press, 1957.

[14] B. Clarke and A. Barron, Information-theoretic Asymptotics of Bayes Methods, *IEEE Trans. Information Theory*, 36, 453–471, 1990.

[15] B. Clarke and A. Barron, Jeffrey's Prior is Asymptotically Least Favorable Under Entropy Risk, *J. Stat. Planning Inference*, 41, 37–61, 1994.

[16] R. Corless, G. Gonnet, D. Hare, D. Jeffrey and D. Knuth, On the Lambert W Function, *Adv. Computational Mathematics*, 5, 329–359, 1996.

[17] I. Csiszár, and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[18] I. Csiszàr and P. Shields, Redundancy Rates for Renewal and Other Processes, *IEEE Trans. Information Theory*, 42, 2065–2072, 1996.

[19] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[20] T. Cover and E. Ordentlich, Universal Portfolios with Side Information, *IEEE Trans. Information Theory*, 42, 348–363, 1996.

[21] L. Davisson, Universal Noiseless Coding, *IEEE Trans. Inform. Theory*, 19, 783–795, 1973.

[22] L. Davisson and A. Leon-Garcia, A Source Matching Approach to Finding Minimax Codes, *IEEE Trans. Inform. Theory*, 26, 166–174, 1980.

[23] A. Dembo and I. Kontoyiannis, The Asymptotics of Waiting Times Between Stationary Processes, Allowing Distortion, *Annals of Applied Probability*, 9, 413–429, 1999.

[24] A. Dembo and I. Kontoyiannis, Critical Behavior in Lossy Coding, *IEEE Trans. Inform. Theory*, 47, 1230-1236, 2001.

[25] A. Dembo and I. Kontoyiannis, Source Coding, Large Deviations, and Approximate Pattern Matching, *IEEE Trans. Information*, 48, 1590-1615, 2002.

[26] H. Dickinson and M. M. Dodson, Extremal manifolds and Hausdorff dimension, *Duke Math. J.* 101, 271–281, 2000.

[27] M. Drmota, A Bivariate Asymptotic Expansion of Coefficients of Powers of Generating Functions, *Europ. J. Combinatorics*, 15, 139-152, 1994.

[28] M. Drmota, H-K. Hwang, and W. Szpankowski, Precise Average Redundancy of an Idealized Arithmetic Coding, *Proc. Data Compression Conference*, 222-231, Snowbird, 2002.

[29] M. Drmota and W. Szpankowski, Precise Minimax Redundancy and Regrets, *IEEE Trans. Information Theory*, 50, 2686-2707, 2004.

[30] M. Drmota and W. Szpankowski, Variations on Khodak's Variable-to-Variable Codes, *42-nd Annual Allerton Conference on Communication, Control, and Computing*, Urbana, 2004.

[31] M. Drmota and W. Szpankowski, On the exit time of a random walk with the positive drift, *2007 Conference on Analysis of Algorithms*, Juan-les-Pins, France, *Proc. Discrete Mathematics and Theoretical Computer Science*, 291-302, 2007.

[32] M. Drmota and R. Tichy, *Sequences, Discrepancies, and Applications*, Springer Verlag, Berlin Heidelberg, 1997.

[33] M. Drmota, Y. Reznik, S. Savari and W. Szpankowski, Precise Asymptotic Analysis of the Tunstall Code, *Proc. 2006 International Symposium on Information Theory*, 2334-2337, Seattle, 2006

[34] M. Drmota, Y. Reznik, S. Savari and W. Szpankowski, Tunstall Code, Khodak Variations, and random Walks, preprint Khodak Variations, and random Walks, preprint available on `http://www.cs.purdue.edu/homes/spa`.

[35] Y. Ephraim and N. Merhav, Hidden Markov Processes, *IEEE Trans. Inform. Theory*, 48, 1518–1569, 2002.

[36] F. Fabris, Variable-Length-to-Variable-Length Source Coding: A Greedy Step-by-Step Algorithm, *IEEE Trans. Info. Theory*, 38, 1609 - 1617, 1992.

[37] J. Fan, T. Poo, B. Marcus, Constraint Gain, *IEEE Trans. Information Theory*, 50, 1989-2001, 2004.

[38] M. Feder, N. Merhav, and M. Gutman, Universal Prediction of Individual Sequences, *IEEE Trans. Information Theory*, 38, 1258–1270, 1992.

[39] P. Flajolet, Singularity Analysis and Asymptotics of Bernoulli Sums, *Theoretical Computer Science*, 215, 371–381, 1999.

[40] P. Flajolet, X. Gourdon, and P. Dumas, Mellin transforms and asymptotics: harmonic sums, Special volume on mathematical analysis of algorithms, *Theoretical Computer Science*, 144, 3–58, 1995.

[41] Ph. Flajolet and A. M. Odlyzko, Singularity analysis of generating functions, *SIAM J. Discrete Math.*, 3, 216–240, 1990.

[42] P. Flajolet and R. Sedgewick, Analytic Combinatorics, Cambridge University Press, 2008.

[43] P. Flajolet and W. Szpankowski, Analytic Variations on Redundancy Rates of Renewal Processes, *IEEE Trans. Information Theory*, 48, 2911 -2921, 2002.

[44] Freeman, G.H.; Divergence and the Construction of Variable-to-Variable-length Lossless Codes by Source-word Extensions, Data Compression Conference, 1993. DCC '93., 79-88, 1993

[45] R. Gallager, *Information Theory and Reliable Communications*, New York, Wiley 1968.

[46] R. Gallager, Variations on the Theme by Huffman, *IEEE Trans. Information Theory*, 24, 668-674, 1978.

[47] V. Choi and M. J. Golin, Lopsided trees. I. Analyses. *Algorithmica* 31, 240–290, 2001.

[48] D-K. He and E-H. Yang, Performance Analysis of Grammar-Based Codes Revisited, *IEEE Tans. Information Theory*, 50, 1524-1535, 2004.

[49] P. Howard and J. Vitter, Analysis of Arithmetic Coding for Data Compression, Brown University, Department of Computer Science, *Proc. Data Compression Conference*, 3–12, Snowbird 1991.

[50] H-K. Hwang, Large Deviations for Combinatorial Distributions I: Central Limit Theorems, *Ann. Appl. Probab.*, 6, 297-319, 1996.

[51] P. Jacquet and W. Szpankowski, Analysis of Digital Tries with Markovian Dependency, *IEEE Trans. Information Theory*, 37, 1470-1475, 1991.

[52] P. Jacquet and W. Szpankowski, Autocorrelation on Words and its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combinatorial Theory, Ser. A*, 66, 237-269, 1994.

[53] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161-197, 1995.

[54] P. Jacquet and W. Szpankowski, Entropy Computations via Analytic Depoissonization, *IEEE Trans. Information Theory*, 45, 1072-1081, 1999.

[55] P. Jacquet and W. Szpankowski, Analytical Depoissonization and Its Applications, *Theoretical Computer Science* in "Fundamental Study", 201, No. 1-2, 1–62, 1998.

[56] P. Jacquet and W. Szpankowski, Markov Types and Minimax Redundancy for Markov Sources *IEEE Trans. Information Theory*, 50, 1393-1402, 2004.

[57] P. Jacquet and W. Szpankowski, Analytic Approach to Pattern Matching, Chap. 7 in *Applied Combinatorics on Words* (eds. Lothaire), Cambridge University Press (Encycl. of Mathematics and Its Applications), 2004.

[58] P. Jacquet, G. Seroussi, and W. Szpankowski. On the Entropy of a Hidden Markov Process. In *Data Compression Conference*, Snowbird, 362-371, 2004.

[59] P. Jacquet, G. Seroussi, and W. Szpankowski. On the Entropy of a Hidden Markov Process, *Theoretical Computer Science*, 395, 203-219, 2008.

[60] P. Jacquet, W. Szpankowski, and J. Tang, Average Profile of the Lempel-Ziv Parsing Scheme for a Markovian Source, *Algorithmica*, 31, 318–360, 2001.

[61] P. Jacquet, W. Szpankowski, and I. Apostol, A Universal Predictor Based on Pattern Matching, *IEEE Trans. Information Theory*, 48, 1462-1472, 2002.

[62] S. Janson, Moments for first passage and last exit times, the minimum, and related quantities for random walks with positive drift. *Adv. Appl. Probab.*, 18, 865-879, 1986.

[63] F. Jelinek and K. S. Schneider, On Variable-length-to-block Coding, *Trans. Information Theory* IT-18, 765-774, 1972.

[64] G. L. Khodak, Connection Between Redundancy and Average Delay of Fixed-Length Coding, *All-Union Conference on Problems of Theoretical Cybernetics* (Novosibirsk, USSR, 1969) 12 (in Russian)

[65] G.L. Khodak, Bounds of Redundancy Estimates for Word-based Encoding of Sequences Produced by a Bernoulli Source (Russian), *Problemy Peredachi Informacii* 8, 21–32, 1972.

[66] J.C. Kieffer, A Unified Approach to Weak Universal Source Coding, *IEEE Tans. Information Theory*, 24, 340-360, 1978.

[67] J.C. Kieffer, Strong Converses in Source Coding Relative to a Fidelity Criterion, *IEEE Trans. Information Theory*, 37, 257-262, 1991.

[68] J. C. Kieffer, Sample Converses in Source Coding Theory, *IEEE Trans. Information Theory*, 37, 263-268, 1991.

[69] C. Knessl and W. Szpankowski, Enumeration of Binary Trees, Lempel-Ziv'78 Parsings, and Universal Types, *Proc. the Second Workshop on Analytic Algorithmics and Combinatorics*, Vancouver, 2005.

[70] D. E. Knuth, *The Art of Computer Programming. Fundamental Algorithms,* Vol. 1, Third Edition, Addison-Wesley, Reading, MA, 1997.

[71] D. E. Knuth, *The Art of Computer Programming. Seminumerical Algorithms*. Vol. 2, Third Edition, Addison Wesley, Reading, MA, 1998.

[72] D. E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.

[73] D. E. Knuth, Linear Probing and Graphs, *Algorithmica*, 22, 561–568, 1998.

[74] D. E. Knuth, *Selected Papers on the Analysis of Algorithms*, Cambridge University Press, Cambridge, 2000.

[75] C. Krattenthaler and P. Slater, Asymptotic Redundancies for Universal Quantum Coding, *IEEE Trans. Information Theory*, 46, 801-819, 2000.

[76] R. Krichevsky and V. Trofimov, The Performance of Universal Coding, *IEEE Trans. Information Theory*, 27, 199–207, 1981.

[77] R. Krichevsky, *Universal Compression and Retrieval,* Kluwer, Dordrecht, 1994.

[78] I. Kontoyiannis, An Implementable Lossy Version of the Lempel-Ziv Algorithm — Part I: Optimality for Memoryless Sources, *IEEE Trans. Information Theory*, 45, 2285–2292, 1999.

[79] I. Kontoyiannis, Pointwise Redundancy in Lossy Data Compression and Universal Lossy Data Compression, *IEEE Trans. Inform. Theory*, 46, 136-152, 2000.

[80] I. Kontoyiannis, Sphere-covering, Measure Concentration, and Source Coding, *IEEE Trans. Inform. Theory*, 47, 1544-1552, 2001.

[81] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*. John Wiley & Sons, New York 1974.

[82] J. Lawrence, A New Universal Coding Scheme for the Binary Memoryless Source, *IEEE Trans. Inform. Theory*, 23, 466-472, 1977.

[83] A. Lempel and J. Ziv, On the Complexity of Finite Sequences, *IEEE Information Theory* 22, 1, 75-81, 1976.

[84] T. Linder, G. Lugosi, and K. Zeger, Fixed-Rate Universal Lossy Source Coding and Rates of Convergence for Memoryless Sources, *IEEE Information Theory*, 41, 665-676, 1995.

[85] S. Lonardi, W. Szpankowski, and M. Ward, Error Resilient LZ'77 Data Compression: Algorithms, Analysis, and Experiments, *IEEE Trans. Information Theory*, 53, 1799-1813, 2007.

[86] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, *IEEE Trans. Information Theory*, 41, 478-488, 1995.

[87] G. Louchard and W. Szpankowski, On the Average Redundancy Rate of the Lempel-Ziv Code, *IEEE Trans. Information Theory*, 43, 2–8, 1997.

[88] G. Louchard, W. Szpankowski, and J. Tang, Average Profile for the Generalized Digital Search Trees and the Generalized Lempel-Ziv Algorithms, *SIAM J. Computing*, 28, 935–954, 1999.

[89] T. Luczak and W. Szpankowski, A Suboptimal Lossy Data Compression Based in Approximate Pattern Matching, *IEEE Trans. Information Theory*, 43, 1439–1451, 1997.

[90] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

[91] K. Marton and P. Shields, The Positive-Divergence and Blowing-up Properties, *Israel J. Math*, 80, 331-348, 1994.

[92] N. Merhav and D. Neuhoff, Variable-to-Fixed Length Codes Provided Better Large deviations Performance Than Fixed-to-Variable Codes, *IEEE Trans. Information Theory*, 38, 135-140, 1992.

[93] N. Merhav, M. Feder, and M. Gutman, Some Properties of Sequential Predictors for Binary Markov Sources, *IEEE Trans. Information Theory*, 39, 887–892, 1993.

[94] N. Merhav and M. Feder, A Strong Version of the Redundancy-Capacity Theory of Universal Coding, *IEEE Trans. Information Theory*, 41, 714–722, 1995.

[95] N. Merhav and J. Ziv, On the Amount of Statistical Side Information Required for Lossy Data Compression, *IEEE Trans. Information Theory*, 43, 1112–1121, 1997.

[96] A. Odlyzko, Asymptotic Enumeration, in *Handbook of Combinatorics*, Vol. II, (Eds. R. Graham, M. Götschel and L. Lovász), Elsevier Science, 1063-1229, 1995.

[97] A. Orlitsky, Prasad Santhanam, and J. Zhang Universal Compression of Memoryless Sources over Unknown Alphabets, *IEEE Trans. Information Theory*, 50, 1469-1481, 2004.

[98] A. Orlitsky and P. Santhanam, Speaking of Infinity (i.i.d. Strings), *IEEE Trans. Information Theory*, 50, 2215 - 2230, 2004.

[99] D. Ornstein and P. Shields, Universal Almost Sure Data Compression, *Ann. Probab.*, 18, 441-452, 1990.

[100] D. Ornstein and B. Weiss, Entropy and Data Compression Schemes, *IEEE Information Theory*, 39, 78-83, 1993.

[101] E. Plotnik, M.J. Weinberger, and J. Ziv, Upper Bounds on the Probability of Sequences Emitted by Finite-State Sources and on the Redundancy of the Lempel-Ziv Algorithm, *IEEE Trans. Information Theory*, 38, 66-72, 1992.

[102] Y. Reznik and W. Szpankowski, On Average Redundancy Rate of the Lempel-Ziv Codes with K-error Protocol, *Information Sciences*, 135, 57-70, 2001.

[103] J. Rissanen, Complexity of Strings in the Class of Markov Sources, *IEEE Trans. Information Theory*, 30, 526–532, 1984.

[104] J. Rissanen, Universal Coding, Information, Prediction, and Estimation, *IEEE Trans. Information Theory*, 30, 629–636, 1984.

[105] J. Rissanen, Fisher Information and Stochastic Complexity, *IEEE Trans. Information Theory*, 42, 40–47, 1996.

[106] B. Ryabko, Twice-Universal Coding, *Problems of Information Transmission*, 173–177, 1984.

[107] B. Ryabko, Prediction of Random Sequences and Universal Coding, *Problems of Information Transmission*, 24, 3–14, 1988.

[108] B. Ryabko, The Complexity and Effectiveness of Prediction Algorithms, *J. Complexity*, 10, 281–295, 1994.

[109] S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, *IEEE Trans. Information Theory*, 43, 9–21, 1997.

[110] S. A. Savari, Variable-to-Fixed Length Codes for Predictable Sources, Proc IEEE Data Compression Conference (DCC'98), Snowbird, 481-490, 1998.

[111] S. A. Savari, Variable-to-Fixed Length Codes and the Conservation of Entropy, *Trans. Information Theory* 45, 1612-1620, 1999.

[112] S. Savari and R. Gallager, Generalized Tunstall Codes for Sources with Memory, *IEEE Trans. Information Theory*, 43, 658-668, 1997.

[113] J. Schalkwijk, An Algorithm for Source Coding, *IEEE Information Theory*, 18, 395-399, 1972.

[114] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley Publishing Company, Reading Mass., 1995.

[115] G. Seroussi, On Universal Types, *IEEE Transactions on Information Theory*, 52, 171-189, 2006.

[116] G. Seroussi, On the Number of t-Ary Trees with a Given Path Length, *Algorithmica*, 46, 557-565, 2006.

[117] P. Shields, Universal Redundancy Rates Do Not Exist, *IEEE Information Theory*, 39, 520-524, 1993.

[118] P. Shields, *The Ergodic Theory of Discrete Sample Path*, American Mathematical Society, 1996.

[119] R. Stanley, Enumerative Combinatorics, Wadsworth, Monterey, 1986.

[120] R. Stanley, *Enumerative Combinatorics*, Vol. II, Cambridge University Press, Cambridge, 1999.

[121] Y. Shtarkov, Universal Sequential Coding of Single Messages, *Problems of Information Transmission*, 23, 175–186, 1987.

[122] Y. Shtarkov, T. Tjalkens and F.M. Willems, Multi-alphabet Universal Coding of Memoryless Sources, *Problems of Information Transmission*, 31, 114-127, 1995.

[123] Y. Steinberg and M. Gutman, An Algorithm for Source Coding Subject to a Fidelity Criterion, Based on String Matching, *IEEE Trans. Information Theory*, 39, 877-886, 1993.

[124] P. Stubley, On the Redundancy of Optimum Fixed-to-Variable Length Codes, *Proc. Data Compression Conference*, 90-97, Snowbird 1994.

[125] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, 39, 1647-1659, 1993.

[126] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Compt.*, 22, 1176–1198, 1993.

[127] W. Szpankowski, On Asymptotics of Certain Sums Arising in Coding Theory, *IEEE Trans. Information Theory*, 41, 2087–2090, 1995.

[128] W. Szpankowski, On Asymptotics of Certain Recurrences Arising in Universal Coding, *Problems of Information Transmission*, 34, 55-61, 1998.

[129] W. Szpankowski, Asymptotic Redundancy of Huffman (and Other) Block Codes, *IEEE Trans. Information Theory*, 46, 2434-2443, 2000.

[130] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.

[131] W. Szpankowski, A One-to-One Code and its Anti-redundancy, *IEEE Trans. Information Theory*, 54, 2008.

[132] T. Tjalkens and F. Willems, A Universal Variable-to-Fixed Length Source Code Based on Lawrence's Algorithm, *IEEE Trans. Information Theory*, 38, 247-253, 1992.

[133] B. P. Tunstall, Synthesis of Noiseless Compression Codes, Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1967.

[134] B. Vallée, Dynamics of the Binary Euclidean Algorithm: Functional Analysis and Operators, *Algorithmica*, 22, 660–685, 1998.

[135] B. Vallée, Dynamical Sources in Information Theory : Fundamental intervals and Word Prefixes, *Algorithmica*, 29, 262–306, 2001.

[136] K. Visweswariah, S. Kulkurani, and S. Verdu, Universal Variable-to-Fixed Length Source Codes, *IEEE Trans. Information Theory*, 47, 1461-1472, 2001.

[137] J. Vitter and P. Krishnan, Optimal Prefetching via Data Compression, *J. ACM*, 43, 771–793, 1996.

[138] M. Ward and W. Szpankowski. Analysis of a Randomized Selection Algorithm Motivated by the LZ'77 Scheme. In *The First Workshop on Analytic Algorithmics and Combinatorics*, New Orleans, 153-160, 2004.

[139] M. Weinberger, N. Merhav, and M. Feder, Optimal Sequential Probability Assignments for Individual Sequences, *IEEE Trans. Information Theory*, 40, 384–396, 1994.

[140] M. Weinberger, J. Rissanen, and M. Feder, A Universal Finite Memory Sources, *IEEE Trans. Information Theory*, 41, 643–652, 1995.

[141] M. Weinberger, J. Rissanen, and R. Arps, Applications of Universal Context Modeling to Lossless Compression of Gray-Scale Images, *IEEE Trans. Image Processing*, 5, 575-586, 1996.

[142] M. Weinberger, G. Seroussi, and G. Sapiro, LOCO-I: A Low Complexity, Context-Based Lossless Image Compression Algorithms, *Proc. Data Compression Conference*, 140-149, Snowbird 1996.

[143] F.M. Willems, Y. Shtarkov and T. Tjalkens, The Context-Tree Weighting Method: Basic Properties, *IEEE Trans. Information Theory*, 41, 653–664, 1995.

[144] F.M. Willems, Y. Shtarkov and T. Tjalkens, Context Weighting for General Finite Context Sources, *IEEE Trans. Information Theory*, 42, 1514-1520, 1996.

[145] P. Whittle, Some Distribution and Moment Formulæ for Markov Chain, *J. Roy. Stat. Soc.,* Ser. B., 17, 235–242, 1955.

[146] A. D. Wyner, An Upper Bound on the Entropy Series, *Inform. Control*, 20, 176-181, 1972.

[147] A. J. Wyner, The Redundancy and Distribution of the Phrase Lengths of the Fixed-Database Lempel-Ziv Algorithm, *IEEE Trans. Information Theory*, 43, 1439–1465, 1997.

[148] A. Wyner and J. Ziv, Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression, *IEEE Trans. Information Theory*, 35, 1250-1258, 1989.

[149] Q. Xie, A. Barron, Minimax Redundancy for the Class of Memoryless Sources, *IEEE Trans. Information Theory*, 43, 647-657, 1997.

[150] Q. Xie, A. Barron, Asymptotic Minimax Regret for Data Compression, Gambling, and Prediction, *IEEE Trans. Information Theory*, 46, 431-445, 2000.

[151] E.H. Yang, and J. Kieffer, Simple Universal Lossy Data Compression Schemes Derived From Lempel-Ziv algorithm, *IEEE Trans. Information Theory*, 42, 239-245, 1996.

[152] E.H. Yang, and J. Kieffer, On the Redundancy of the Fixed–Database Lempel-Ziv Algorithm for ϕ-Mixing Sources, *IEEE Trans. Information Theory*, 43, 1101–1111, 1997.

[153] E.H. Yang, and J. Kieffer, On the Performance of Data Compression Algorithms Based upon String Matching, *IEEE Trans. Information Theory*, 44, 1998.

[154] E.H. Yang and Z. Zhang, The Shortest Common Superstring Problem: Average Case Analysis for Both Exact Matching and Approximate Matching, *IEEE Trans. Information Theory*, 45, 1867–1886, 1999.

[155] Y. Yang and A. Barron, Information-Theoretic Determination of Minimax Rates of Convergence, *The Ann. Stat.*, 27, 1564–1599, 1999.

[156] Z. Zhang and V. Wei, An On-Line Universal Lossy Data Compression Algorithm via Continuous Codebook Refinement – Part I: Basic Results, *IEEE Trans. Information Theory*, 42, 803-821, 1996.

[157] J. Ziv, Coding of Source with Unknown statistics – Part II: Distortion Relative to a Fidelity Criterion, *IEEE Trans. Information Theory*, 18, 389-394, 1972.

[158] J. Ziv, Variable-to-Fixed Length Codes are Better than Fixed-to-Variable Length Codes for Markov Sources, *IEEE Trans. Information Theory*, 36, 861-863, 1990.

[159] J. Ziv, Back from Infinity: A Constrained Resources Approach to Information Theory, *IEEE Information Theory Society Newsletter*, 48, 30-33, 1998.

[160] J. Ziv and A. Lempel, A Universal Algorithm for Sequential Data Compression, *IEEE Trans. Information Theory*, 23, 3, 337-343, 1977.

[161] J. Ziv and A. Lempel, Compression of Individual Sequences via Variable-rate Coding, *IEEE Trans. Information Theory*, 24, 530-536, 1978.