Tunstall Code, Khodak Variations, and Random Walks

February 5, 2008

Michael Drmota*, Yuryi A. Reznik†, Serap Savari‡, and Wojciech Szpankowski§

Abstract

A variable-to-fixed length encoder partitions the source string into variable-length phrases that belong to a given and fixed dictionary. Tunstall, and independently Khodak, designed variable-to-fixed length codes for memoryless sources that are optimal under certain constraints. In this paper, we study the Tunstall and Khodak codes using analytic information theory, i.e., the machinery from the analysis of algorithms literature. After proposing an algebraic characterization of the Tunstall and Khodak codes, we present new results on the variance and a central limit theorem for dictionary phrase lengths. This analysis also provides a new argument for obtaining asymptotic results about the mean dictionary phrase length and average redundancy rates.

Index Terms: Analytic information theory, Tunstall code, variable-to-fixed length codes.

*Institute of Discrete Mathematics and Geometry, TU Wien, Wiedner Hauptstr. 8–10, A-1040 Wien, Austria michael.drmota@tuwien.ac.at. The work of this author was supported in part by the Austrian Science Foundation FWF Grant No. S9604.

†Qualcomm Inc., 5775 Morehouse Dr., San Diego, CA 92121, yreznik@qualcomm.com.

‡Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, savari@ece.tamu.edu. The work of this author was done while she was with the University of Michigan, Ann Arbor and was supported in part by NSF Grant No. CCF-0430201.

Department of Computer Science, Purdue University, W. Lafayette, IN 47907, spa@cs.purdue.edu. The work of this author was supported in part by the NSF Grants CCF-0513636 and DMS-0503742, NIH Grant R01 GM068959-01, and AFOSR Grant 073071.

1 Introduction

A variable-to-fixed length encoder partitions the source string over an m-ary alphabet $\mathcal{A}$ into a concatenation of variable-length phrases. Each phrase except the last one is constrained to belong to a given dictionary $\mathcal{D}$ of source strings; the last phrase is a non-null prefix of a dictionary entry. One common constraint on a dictionary is that it leads to a *unique* parsing of any string over $\mathcal{A}$ (see [27] for examples of dictionaries without this constraint). For the rest of the paper we will assume that all dictionaries are uniquely parsable. It is convenient to represent a uniquely parsable dictionary by a complete parsing tree $\mathcal{T}$, i.e., a tree in which every internal node has all m children nodes in the tree. The dictionary entries $d \in \mathcal{D}$ correspond to the leaves of the associated parsing tree. The encoder represents each parsed string by the fixed length binary code word corresponding to its dictionary entry. If the dictionary $\mathcal{D}$ has M entries, then the code word for each phrase has $\lceil \log_2 M \rceil$ bits. The best known variable-to-fixed length code is now generally attributed to Tunstall [40]; however, it was independently discovered by Khodak [15], Verhoeff [42], and possibly others. In this paper, we offer a new perspective and asymptotic analysis of the Tunstall and Khodak codes.

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

Tunstall's algorithm has been studied extensively (cf. the survey article [1]). Simple bounds for its redundancy were obtained independently by Khodak [15] and by Jelinek and Schneider [14]. Tjalkens and Willems [37] were the first to look at extensions of this code to sources with memory. Savari and Gallager [24] 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. Savari [25] later published a non-asymptotic analysis of the Tunstall code for binary, memoryless sources with small entropies. *Universal* variable-to-fixed length codes were analyzed in [39, 21, 20, 19, 38, 43]; however, we are unaware of analyses of the minimax redundancy for variable-to-fixed and variable-to-variable length codes, and these problems remain open. Finally Tjalkens has studied constructions of Tunstall codes in [34] and [36] and Kieffer focused on the problem of binary sources in [17]. The paper [3] discusses the use of Tunstall algorithm for the approximation of uniform distributions for random number generation and for related problems.

Our focus will be Khodak's [15] construction of the Tunstall code (see also [19]). Khodak independently discovered the Tunstall code using a rather different approach. Let p_i be the probability of the ith source symbol and let $p_{\min} = \min\{p_1, \ldots, p_m\}$. Khodak suggested choosing a real number $r \in (0, p_{min})$ and growing a complete parsing tree until all leaves d satisfy

$$p_{\min} r \leq P(d) < r \,, d \in \mathcal{D}. \tag{1}$$

It can also be shown (see, e.g., [14, Lemma 6] and [24, Lemma 2]) that the resulting parsing tree is exactly the same as a tree constructed by Tunstall's algorithm (cf. Section 2 for detailed comments). The asymptotic relationship between r and the resulting number of dictionary entries M_r was studied in [24] and will be established here in a different way.

It follows from (1) that if y is a proper prefix of one or more entries of $\mathcal{D} = \mathcal{D}_r$, i.e., y corresponds to an internal node of $\mathcal{T} = \mathcal{T}_r$, then

$$P(y) \geq r. \tag{2}$$

Therefore, it is easier to characterize the internal nodes of the parsing tree $\mathcal{T}_r$ rather than its leaves. We shall approach the analysis of $\mathcal{D} = \mathcal{D}_r$ by representing the moment generating function of the phrase length in terms of the transform of the path lengths to internal nodes in $\mathcal{T}_r$. We will show that the moment generating function of the dictionary phrase length in the parsing tree satisfies certain recurrences that could be analyzed through analytic algorithmic methods (cf. [33]) such as the Mellin transform and the Tauberian theorems. This analysis provides a precise asymptotic characterization of the behavior of the Tunstall and Khodak codes. In passing, we mention that this work directly extends recent analyses of fixed-to-variable codes (cf. [7, 11, 32, 33]) through tools of analytic algorithmics and is hence in the domain of analytic information theory.

Our goal is to establish the limiting distribution of the phrase length and to provide a precise asymptotic analysis of the average redundancy of the Tunstall and Khodak codes. While the average redundancy of the Tunstall code for memoryless and Markov sources has been studied previously by Savari [24, 25, 26], we provide here a new approach that allows us to precisely quantify oscillations involved in the redundancy for a certain class of sources. Our central limit theorem concerning the phrase length is new and has been derived in an analytic way that hopefully will serve as a template for the future analysis of variable-to-fixed length and variable-to-variable length codes.

The paper is organized as follows. In the next section we present our main results and their consequences. In particular, in Theorem 1 we translate the probability generating function of the phrase length (leaf of the associated parsing tree) into the probability generating function of internal nodes. The latter function is next represented by a recurrence that we handle through the Mellin transform and the Tauberian theorems leading to our main findings which are presented in Theorems 2 and 3. At this point we observe that the phrase length can be equivalently described as the length of a random walk on a lattice with a linear barrier. Indeed, observing that $P(y) = p_1^{k_1} \cdots p_m^{k_m}$ condition (2) becomes $a_1 k_1 + \cdots + a_m k_m \leq \log(1/r)$ where $a_i = \log(1/p_i)$. Thus, the phrase length D is the exit time of a random walk on a lattice with steps $a_1, \ldots, a_m$ and the linear barrier $a_1 k_1 + \cdots + a_m k_m = \log(1/r)$, as discussed in [8] (cf. also [12]). Finally, in the last section we prove Theorems 2 and 3 after introducing some technical results.

2 Main Results and Consequences

Assume a memoryless source over an m-ary alphabet $\mathcal{A}$ generates an output sequence. Let $p_i > 0$ be the probability of the i^{th} letter of alphabet $\mathcal{A}$, $i \in \{1, \ldots, m\}$, $p_{\min} =$

$\min\{p_1, \ldots, p_m\}$, and $p_{\max} = \max\{p_1, \ldots, p_m\}$. Given a dictionary $\mathcal{D}$ and corresponding complete parsing tree $\mathcal{T}$, the encoder partitions the source output sequence into a sequence of variable-length phrases. Let $d \in \mathcal{D}$ denote a dictionary entry, $P(d)$ be its probability, and $|d|$ be its length. Our focus will be on the random variable $D = |d|$, the phrase length of a dictionary string. One of our goals is to investigate the moment generating function of the phrase length $D = D_r$ in Khodak's construction of the Tunstall dictionary with parameter r. That is, we consider

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

Towards this end, we next introduce a second transform describing the probabilities of strings which correspond to the internal nodes in the parsing tree $\mathcal{T}_r$. Let

$$S(r, z) = \sum_{y:\ P(y) \geq r} P(y) z^{|y|}. \tag{3}$$

We next find a relation between $D(r, z)$ and $S(r, z)$.

2.1 Connections Between Probabilities of Nodes and Leaves

Our first result concerns *arbitrary* complete parsing trees, i.e., not necessarily Tunstall trees, and relates the transform for the probabilities of internal nodes to a function of the leaf probabilities.

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

$$\sum_{d \in \tilde{\mathcal{D}}} P(d) z^{|d|} = 1 + (z - 1) \sum_{y \in \tilde{\mathcal{Y}}} P(y) z^{|y|}. \tag{4}$$

Proof. Instead of (4) we prove the equivalent statement

$$\sum_{d \in \tilde{\mathcal{D}}} P(d) \left(\frac{z^{|d|} - 1}{z - 1} \right) = \sum_{y \in \tilde{\mathcal{Y}}} P(y) z^{|y|}, \tag{5}$$

and we use induction on the number of internal nodes in the corresponding dictionary tree. For the basis step, (5) is clearly true when $\tilde{\mathcal{D}} = \mathcal{A}$ since the only element of $\tilde{\mathcal{Y}}$ is the null string, which has probability one and length zero.

For the inductive step, suppose that (5) is true for all dictionaries with parsing trees having k internal nodes. Let $\tilde{\mathcal{D}}$ be a dictionary with a corresponding proper prefix set $\tilde{\mathcal{Y}}$ having $k + 1$ elements. Choose $y_0 \in \tilde{\mathcal{Y}}$ to have maximum length so that its single letter extensions correspond to the dictionary entries $d_1, d_2, \ldots, d_m \in \tilde{\mathcal{D}}$. Observe that $P(y_0) = P(d_1) + P(d_2) + \cdots + P(d_m)$. We next define an auxiliary dictionary $\tilde{\mathcal{D}}'$ with $\tilde{\mathcal{D}}' = \tilde{\mathcal{D}} \cup \{y_0\} \setminus \{d_1, \ldots, d_m\}$. Then $\tilde{\mathcal{D}}'$ has a corresponding proper prefix set $\tilde{\mathcal{Y}}' = \tilde{\mathcal{Y}} \setminus \{y_0\}$ with k elements.

Using the inductive hypothesis, we have

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

This completes the proof of the lemma. ∎

Observe that $\mathbf{E}[D] = \sum_{d\in\tilde{\mathcal{D}}} P(d)\,|d|$. Hence by taking the derivative of both sides of (4) with respect to z and setting $z = 1$, we see that Theorem 1 offers a new proof of the well-known result that

$$\mathbf{E}[D] = \sum_{y\in\tilde{\mathcal{Y}}} P(y).$$

By considering the second derivatives of both sides of (4) we obtain a new result for uniquely parsable dictionaries:

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

Furthermore, Theorem 1 and equation (3) imply that for Khodak's construction of the Tunstall codes

$$D(r,z) = 1 + (z-1)S(r,z). \tag{6}$$

Thus we can express the moment generating function for the phrase length of a Tunstall dictionary entry in terms of the transform describing the probabilities of proper prefixes of the dictionary entries. As we will discuss below, this relationship enables us to exploit a recurrence description for our analysis of the Tunstall and Khodak codes.

2.2 Formulation of Main Results

Let $v = 1/r$, and z be a real (more generally, a complex) number. Define $\tilde{S}(v,z) = S(v^{-1},z)$. We restrict our attention here to a binary alphabet $\mathcal{A}$ with $0 < p_1 < p_2 < 1$.

Let $A(v)$ devote the number of source strings with probability at least v^{-1}; i.e.,

$$A(v) = \sum_{y:P(y)\geq 1/v} 1. \tag{7}$$

Observe that $A(v)$ represents the number of internal nodes in Khodak's construction with parameter v^{-1} of a Tunstall tree. Furthermore, $A(v)$ counts the number of strings y with the self-information $-\log P(y) \le \log v$. The functions $A(v)$ and $\tilde{S}(v,z)$ satisfy the following recurrences since every binary string either is the empty string, a string starting with the first source letter $a_1 \in \mathcal{A}$, or a string starting with the letter $a_2 \in \mathcal{A}$:

$$A(v) = \begin{cases} 0 & v < 1, \\ 1 + A(vp_1) + A(vp_2) & v \ge 1 \end{cases} \tag{8}$$

and

$$\tilde{S}(v,z) = \begin{cases} 0 & v < 1, \\ 1 + zp_1\tilde{S}(vp_1,z) + zp_2\tilde{S}(vp_2,z) & v \ge 1. \end{cases} \tag{9}$$

Observe that $A(v)$ represents the number of internal nodes in Khodak's construction with parameter v^{-1} of a Tunstall tree, that is $M_r = A(v)+1 = |\mathcal{D}_r|$ is the dictionary size. Finally, $\mathbf{E}[D_r] = \tilde{S}(v,1)$ is the corresponding expected value of the phrase length ($r = 1/v$).

These recurrences can be studied through the Mellin transform (see, e.g. [4, 33]). In the next section we prove our second main result (note that log denotes the logarithm to base e):

Theorem 2. *Let $v = 1/r$ in the Khodak's construction and assume $v \to \infty$.*

(i) *If $\log p_2/\log p_1$ is irrational, then*

$$M_r = \frac{v}{H} + o(v). \tag{10}$$

Otherwise, (when $\log p_2/\log p_1$ is rational) let $L > 0$ is the largest real number for which $\log(1/p_1)$ and $\log(1/p_2)$ are integer multiples of L. Then

$$M_r = \frac{Q_1(\log v)}{H} v + O(v^{1-\eta}) \tag{11}$$

for some $\eta > 0$ where (cf. Figure 1)

$$Q_1(x) = \frac{L}{1-e^{-L}} e^{-L\langle \frac{x}{L} \rangle}, \tag{12}$$

and $H = p_1\log(1/p_1) + p_2\log(1/p_2)$ is the entropy rate in natural *units while $\langle y \rangle = y - \lfloor y \rfloor$ is the fractional part of the real number y.*

(ii) *If $\log p_2/\log p_1$ is irrational, then*

$$\mathbf{E}[D_r] = \tilde{S}(v,1) = \frac{\log v}{H} + \frac{H_2}{2H^2} + o(1), \tag{13}$$

while in the rational case

$$\mathbf{E}[D_r] = \tilde{S}(v,1) = \frac{\log v}{H} + \frac{H_2}{2H^2} + \frac{Q_2(\log v)}{H} + O(v^{-\eta}) \tag{14}$$

for some $\eta > 0$, where

$$Q_2(x) = L \cdot \left(\frac{1}{2} - \left\langle \frac{x}{L} \right\rangle \right) \tag{15}$$

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

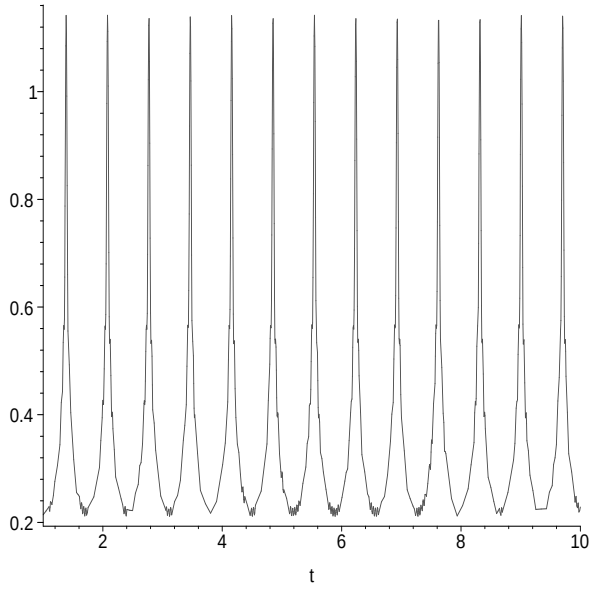

Figure 1: The absolute value of $Q_1(t)$ versus $t = \log v$ for $p_0 = p_1 = -.5$.

In order to obtain distributional results on D we have to analyze $\tilde{D}(v,z) = D(1/v,z) = 1 + (z-1)\tilde{S}(v,z)$ uniformly for z in a neighborhood of $z = 1$. Although the recurrence for $\tilde{S}(v,z)$ looks similar to that of $A(v)$ its analysis is more complex. The main technical problem lies in the slow convergence rates of certain series. We will deal with this problem in the next section by appealing to some Tauberian theorems. Now we present our third main result:

Theorem 3. *Let D_r denote the phrase length in Khodak's construction of the Tunstall code with a dictionary of size M_r over a biased memoryless source. Then as $M_r \to \infty$*

$$\mathbf{E}[D] = \frac{\log M_r}{H} + O(1), \tag{16}$$

$$\mathbf{Var}[D_r] \sim \left(\frac{H_2}{H^3} - \frac{1}{H}\right) \log M_r, \tag{17}$$

and

$$\frac{D_r - \frac{1}{H}\log M_r}{\sqrt{\left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log M_r}} \to N(0,1), \tag{18}$$

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

By combining (10) and (13) resp. (11) and (14) we can be even more precise. In the irrational case we have

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

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

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

Note that (12) and (15) yield

$$Q_2(\log v) - \log\left(Q_1(\log v)\right) = -\log L + \log(1 - e^{-L}) + \frac{L}{2} = \log\left(\frac{\sinh(L/2)}{L/2}\right)$$

so that there is actually no oscillation.

As a direct consequence, we can derive a precise asymptotic formula for the average redundancy of the Tunstall and Khodak codes that is defined in [24] by

$$\mathcal{R}_M = \frac{\log M}{\mathbf{E}[D]} - H\,. \tag{19}$$

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 $\log p_1/\log p_2$ is irrational, then*

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

In the rational case, we have

$$\mathcal{R}_{M_r} = \frac{H}{\log M_r}\Big(-\frac{H_2}{2H} - \log H - \log\left(\frac{\sinh(L/2)}{L/2}\right)\Big) + O\left(\frac{1}{(\log M_r)^2}\right),$$

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

In passing we observe that the Corollary 1 is a special case of Theorems 5 and 12 of [24] for the Tunstall code. Observe that the Tunstall code redundancy has some oscillations for the rational case, that disappear for the Khodak code, as shown above. This is explained in the next section.

2.3 Some Consequences

In this subsection, we provide some comments regarding our results. In particular, we explain the slightly different asymptotic behaviors of the redundancies for the Tunstall and Khodak codes. We also observe that a path in the parsing tree can be viewed as a random walk in the first quadrant of the plane. Finally, we quantify how well the probabilities of the leaves approximate the uniform distribution.

Khodak vs Tunstall Codes Behavior. Tunstall's code and Khodak's code are "almost equivalent." They ultimately produce the same parsing trees, 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 M_r when the parsing tree grows. 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 [24] tiny oscillations remain in the Tunstall code redundancy for the rational case.

Let's be more precise. Suppose that we have $p^k q^l = r$ for some $r > 0$ and integers $k, l \geq 0$. Then there are exactly $\binom{k+l}{k}$ words with k 1's and l 2's. If $\log p_1 / \log p_2$ is irrational then r uniquely determines k and l. Thus, the jump in M_r is exactly $\binom{k+l}{k}$ and by using Stirling's approximation of $n!$ we see that the total probability of these nodes equals

$$\binom{k+l}{k} p_1^k p_2^l = O\left(\frac{1}{\sqrt{k+l}}\right).$$

Thus, if $\tilde{D}$ denotes the path length of any Tunstall code where only some of these $\binom{k+l}{k}$ nodes have been expanded, then $\mathbf{E}[\tilde{D}] = \mathbf{E}[D_r] + o(1)$. Since the words in $\tilde{D} \setminus D_r$ correspond to word in $D_r \setminus \tilde{D}$ that differ in length by 1, the central limit theorem for D_r also directly provides a central limit theorem for $\tilde{D}$.

In the rational case one has to be more careful. If $\log p_1 / \log p_2 = d/b$ for coprime integers $r, d > 0$ then we have $p_1^k p_2^l = p_1^{k'} p_2^{l'}$ if and only if $k' = k + bj$ and $l' = l - dj$ for some integer j. Hence, the jump in M_r equals

$$A = \sum_j \binom{k + l + (b-d)j}{k + bj},$$

and the total probability of these nodes equals $A\, p^k q^l$ and is not $o(1)$. Nevertheless, because of the coupling of word in the Khodak und Tunstall dictionaries, we have $|\tilde{D} - D_r| \leq 1$ and consequently. $\mathbf{E}[\tilde{D}] = \mathbf{E}[D_r] + O(1)$, where $\tilde{D}$ denotes the path length of a Tunstall code where only some of the A nodes have been expanded. Furthermore, since $|\tilde{D} - D_r| / \log M \leq 1/\log M \to 0$ the central limit theorem is not affected by this variation. Thus, if $\tilde{D}_M$ denotes the path length of a Tunstall code word with dictionary size M, then

$$\frac{\tilde{D}_M - \frac{1}{H} \log M}{\sqrt{\left(\frac{H_2}{H^3} - \frac{1}{H}\right) \log M}} \to N(0,1).$$

For expected value and variance we obtain $\mathbf{E}[\tilde{D}_M] = \frac{\log M}{H} + O(1)$ and

$$\begin{aligned} \mathbf{Var}[\tilde{D}_M] &= \mathbf{Var}[D_r] + O\left(\sqrt{\mathbf{Var}[D_r]}\right) \\ &\sim \left(\frac{H_2}{H^3} - \frac{1}{H}\right) \log M. \end{aligned}$$

Indeed, more generally if $Y_n = X_n + Z_n$ and we know that X_n satisfies a central limit theorem of the form $(X_n - \mathbf{E}[X_n])/\sqrt{\mathbf{Var}[X_n]} \to N(0,1)$ and we have $\mathbf{Var}[X_n] \to \infty$ and $\mathbf{Var}[Z_n]/\mathbf{Var}[X_n] \to 0$ (as $n \to \infty$) then Y_n satisfies a central limit theorem, too, i.e. $(Y_n - \mathbf{E}[Y_n])/\sqrt{\mathbf{Var}[Y_n]} \to N(0,1)$, and we also have $\mathbf{Var}[Y_n] = \mathbf{Var}[X_n] + \mathbf{Var}[Z_n] + O(\sqrt{\mathbf{Var}[X_n]\mathbf{Var}[Z_n]}) = \mathbf{Var}[X_n] \cdot (1 + o(1))$, which follows from Cauchy-Schwarz's inequality $|\mathbf{E}[(X_n - \mathbf{E}[X_n])(Z_n - \mathbf{E}[Z_n])| \leq (\mathbf{Var}[X_n])^{1/2} (\mathbf{Var}[Z_n])^{1/2}$.

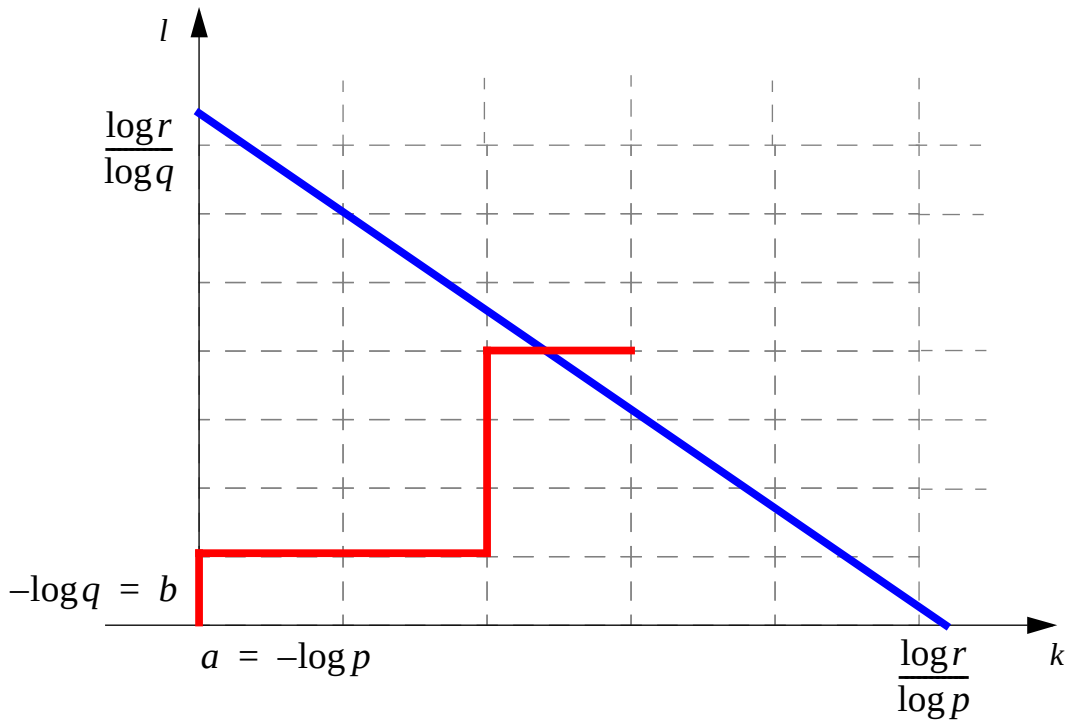

Figure 2: A random walk with a linear barrier; the exit time is equivalent to the phrase length in the Khodak algorithm (e.g., the exit time = 7).

Random Walk with Positive Drift. As already observed in [24], 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 2). 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 $y = p_1^k p_2^l$ for some nonnegative integers $k, l \geq 0$, we conclude that the above summation set can be expressed, after setting $v = 2^V$, as

$$k \log_2 \frac{1}{p_1} + l \log_2 \frac{1}{p_2} \leq V.$$

But this corresponds to a random walk in the first quadrant with the linear boundary condition $ax + by = V$ where $a = \log(1/p_1)$ and $b = \log(1/p_2)$. 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) as illustrated in Figure 2. This correspondence is further explored in [8, 12].

Approximately Uniform Distribution of Leaves. As a consequence of Corollary 1, it is easy to show that probabilities assigned to leaves of the Tunstall parsing trees can be approximated by the uniform distribution, that is, these probabilities are "close" to $1/M$ in the sense that we make more precise below. Observe first that by the *conservation of entropy*

[22, 26] we can rewrite the formula for average redundancy (19) as

$$\mathcal{R}_M \mathbf{E}[D] = \log M - H(\mathcal{D})$$

where $H(\mathcal{D})$ is the entropy of the dictionary phrases (leaves of $\mathcal{T}$). From the above corollary and (16) we conclude that $\mathcal{R}_M \mathbf{E}[D] = O(1)$. Now let $\mathcal{U}_M$ represent the uniform distribution over M and $P_{\mathcal{D}}$ the distribution of the dictionary entries (leaves in $\mathcal{T}$). Clearly,

$$D(P_{\mathcal{D}}||\mathcal{U}_M) = \log M - H(\mathcal{D}).$$

Then by Pinsker's inequality [6] the squared of the L_1 distance $||\cdot||_1$ between $P_{\mathcal{D}}$ and $\mathcal{U}_M$ becomes

$$||P_{\mathcal{D}} - \mathcal{U}_M||_1^2 \leq 2 \log 2 D(P_{\mathcal{D}}||\mathcal{U}_M) = O(\mathcal{R}_M \mathbf{E}[D]) = O(1).$$

The previous observation can be also directly obtained from the fact that $c_1/M \leq P(d) \leq c_2/M$.

In summary, the distribution of leaves $P_{\mathcal{D}}$ can be approximated by the uniform distribution $\mathcal{U}_M$. However, we should point out that it is also known that the ratio of maximum to minimum leaf probability approaches $1/p_{\min}$. Another notion of closeness has recently been considered in [3].

3 Analysis

In this section we prove our main results. We first present an overview of our derivations. Then we discuss some technical lemmas (cf. Section 3.1) that are of their own interest and are instrumental for our analysis. Finally, we prove Theorem 2 in Section 3.2 and Theorem 3 in Section 3.3. Redundancy is discussed in Section 3.4. Throughout this section we write $p := p_1$ and $q := p_2 = 1 - p$.

We should point out that the proof in the rational case of Theorems 2 and 3 is elementary (i.e., complex analysis is not used), while the irrational case requires the use of Mellin transform and Wiener's Tauberian theorem (cf. [4]). In fact, we can uniformly use the Mellin transform for the rational case, however, this makes things more complicated than necessary.

Let us first present an overview of our approach for the irrational case that requires more sophisticated tools. In order to obtain distributional results on D we have to analyze

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

uniformly for z in a neighborhood of $z = 1$, where $\tilde{S}(v, z)$ satisfies the functional equation (9). The main technical problem of analyzing this functional equation lies in slow convergence rates of certain series, as we shall discuss in the sequel.

The Mellin transform $F^*(s)$ of a function $F(v)$ is defined as (cf. [33])

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

if it exists. Using the fact that the Mellin transform of $F(ax)$ is $a^{-s}F^*(s)$ a simple analysis of recurrence (9) reveals that the Mellin transform of $\tilde{D}(v,z)$ with respect to v becomes

$$\tilde{D}^*(s,z) = \frac{1-z}{s(1-zp_1^{1-s}-zp_2^{1-s})} - \frac{1}{s}, \tag{20}$$

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

In order to find the asymptotics of $\tilde{D}(v,z)$ as $v \to \infty$ we proceed to compute the inverse transform of $\tilde{D}^*(s,z)$, that is (cf. [33])

$$\tilde{D}(v,z) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{\sigma-iT}^{\sigma+iT} \tilde{D}^*(s,z) v^{-s}\, ds, \tag{21}$$

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

$$\mathcal{Z}(z) = \{s \in \mathbb{C} : zp^{1-s} + zq^{1-s} = 1\} \tag{22}$$

of all complex roots of $zp^{1-s}+zq^{1-s}=1$. In Lemma 2 below (cf. Section 3.1) we study properties of the set $\mathcal{Z}(z)$.

Furthermore, using the residue theorem we expect to evaluate the integral in (21) as $T \to \infty$. The problem is that this integral is not absolutely convergent since the integrand is only of order $1/s$. To circumvent this problem, we resort to analyze another integral (cf. [41]), namely

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

The Mellin transform of $\tilde{D}_1(v,z)$ is of order $O(1/s^2)$ and then one can estimate its inverse Mellin as described above. However, after obtaining asymptotics of $\tilde{D}_1(v,z)$ as $v \to \infty$ one must recover the original asymptotics of $\tilde{D}(v,z)$. This requires some Tauberian theorems that we discuss in the next section.

3.1 Some Technical Lemmas

In this section we first present some properties of the set of zeros of $zp^{1-s}+zq^{1-s}=1$ and then discuss some Tauberian results that are instrumental to the proof of our main results.

Properties of the set $\mathcal{Z}$. We now deal with the set of zeroes $\mathcal{Z}(z) = \{s \in \mathbb{C} : zp^{1-s} + zq^{1-s} = 1\}$. In the next lemma we first discuss the case $z=1$ and then extend to the general case. The following lemma is basically due to Schachinger [29] and Jacquet [33].

Lemma 1. *Suppose that $0 < p < q < 1$ and let*

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

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

$$-1 \le \Re(s) \le \sigma_0,$$

where σ_0 is a real positive solution of $1+q^{-s}=p^{-s}$. Furthermore, for every integer k there uniquely exists $s_k \in \mathcal{Z}$ with

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

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

(ii) *If $\log q/\log p$ is irrational, then $s_0=-1$ and $\Re(s_k)>-1$ for all $k\neq 0$.*

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

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

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

Proof. We first prove part (i). If $p<q$ and $\sigma=\Re(s)<-1$, then we have $|p^{-s}+q^{-s}|\leq p^{-\sigma}+q^{-\sigma}<p+q=1$. Next, there uniquely exists $\sigma_0>0$ with $1+q^{-\sigma_0}=p^{-\sigma_0}$, and we have $1+q^{-\sigma}>p^{-\sigma}$ if and only if $\sigma<\sigma_0$. Now if $p^{-s}+q^{-s}=1$, then $1+q^{-\Re(s)}\geq|1-q^{-s}|=|p^{-s}|=p^{-\Re(s)}$. Consequently $\Re(s)\leq\sigma_0$.

Let B_k denote the set

$$B_k=\{s\in\mathbb{C} : -2\leq\Re(s)\leq\sigma_0+1,\ (2k-1)\pi/\log p\leq\Im(s)\leq(2k+1)\pi/\log p\}.$$

We will show that B_k contains exactly one point of $\mathcal{Z}$ for each k. First, the function $f(s)=p^{-s}-1$ has exactly one zero $\tilde{s}_k=(2k\pi i)/(\log p)\in B_k$. We shall show that $|q^{-s}|<|p^{-s}-1|$ on the boundary of B_k. Thus, by Rouche's theorem [10] the function $g(s)=p^{-s}-1+q^{-s}$ has also exactly one zero in B_k as required.

We now prove that $|q^{-s}|<|p^{-s}-1|$ to complete the proof of part (i). If $\Im(s)=(2k\pm1)\pi/\log p$, then $p^{-s}-1=-p^{-\Re(s)}-1$. Next observe, that $p<q$ implies that for all real σ we have $q^{-\sigma}<1+p^{-\sigma}$. Consequently, $|q^{-s}|=q^{-\Re(s)}<1+p^{-\Re(s)}=|1-p^{-s}|$. If $\Re(s)=-2$, then it follows from $p^2+q^2<1$ that $|q^{-s}|=q^2<1-p^2\leq|1-p^{-s}|$. Finally, if $\Re(s)=\sigma_0+1$, then we have $q^{-\sigma_0-1}+1<p^{-\sigma_0-1}$ and, thus, $|q^{-s}|=q^{-\sigma_0-1}<p^{-\sigma_0-1}-1=|1-p^{-s}|$. This completes the proof that $|q^{-s}|<|p^{-s}-1|$ on the boundary of B_k.

For part (ii), suppose that $s=-1+it$ for some $t\neq0$ and $p^{-s}+q^{-s}=pe^{-it\log p}+qe^{-it\log q}=1$. Of course, this is only possible if there exist (non-zero) integers $\tilde{d}$, $\tilde{r}$ with $t\log p=2\pi\tilde{d}$ and $t\log q=2\pi\tilde{r}$. This implies that $\log p/\log q=\tilde{d}/\tilde{r}$ is rational. Conversely, if $\log p/\log q$ is irrational, then it follows that all $s\in\mathcal{Z}$ with $s\neq-1$ satisfy $\Re(s)>-1$.

Finally, for part (iii), suppose that $\log p/\log q=d/r$ is rational, where we assume that d are r are coprime positive integers. It is immediately clear that all s of the form $s=-1+2\ell\pi id/\log p$ (where ℓ is arbitrary integer) are contained in $\mathcal{Z}$. This means that

$$s_{\ell d}=-1+2\ell\pi id/\log p=s_0+2\ell\pi id/\log p.$$

Similarly we get

$$s_{\ell d+j}=s_j+2\pi i\ell d/\log p$$

for $j = 1, 2, \ldots, d-1$. It remains to show that $\Re(s_j) > -1$ for $j = 1, 2, \ldots, d-1$. From the proof of (ii) it follows that every $s \in \mathcal{Z}$ with $\Re(s) = -1$ has imaginary part $\Im(s) = 2\pi\tilde{d}/\log p$, where $\tilde{d}$ is an integer with $\log p/\log q = \tilde{d}/\tilde{r}$. Thus, $\tilde{d}$ is an integer multiple of d, that is, $\tilde{d} = \ell d$ and consequently $s = s_{\ell d}$. ∎

The next lemma is a direct generalization of Lemma 1. The important point is the uniformity in z. Note that for $z = 1$ we have $\mathcal{Z}_1(1) = \mathcal{Z} + 1$.

Lemma 2. *Suppose that $0 < p < q < 1$ and that z is a real number with $|z-1| \le \delta$ for some $0 < \delta < 1$. Let*

$$\mathcal{Z}_1(z) = \{s \in \mathbb{C} : p^{1-s} + q^{1-s} = 1/z\}.$$

Then
(i) *All $s \in \mathcal{Z}_1(z)$ 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 $s_k(z) \in \mathcal{Z}_1(z)$ with

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

and consequently $\mathcal{Z}_1(z) = \{s_k(z) : k \in \mathbb{Z}\}$.

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

$$\min_{|z-1|\le\delta} (\Re(s_k(z)) - \Re(s_0(z))) > 0. \tag{23}$$

(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 $s \in \mathcal{Z}_1(z)$ 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 proof of this lemma follows the same steps as that of Lemma 1. We only have to take care of z within $|z-1| \le \delta$. Note that the mapping $z \mapsto s_k(z)$ is continuous (even analytic). Thus, it follows from $\Re(s_k(z)) > \Re(s_0(z))$ that the minimum $\min_{|z-1|\le\delta} (\Re(s_k(z)) - \Re(s_0(z)))$ exists and is certainly positive, i.e. (23).

For our analysis, we need also the Taylor expansion of $s_0(z)$ around $z = 1$. By a local inversion of $p^{1-s} + q^{1-s} = 1/z$ (compare also with [33]) we observe that

$$s_0(z) = -\frac{z-1}{H} + \left(\frac{1}{H} - \frac{H_2}{2H^3}\right)(z-1)^2 + O((z-1)^3) \tag{24}$$

as $z \to 1$. Here $H = p\log(1/p) + q\log(1/q)$ is the entropy rate for memoryless sources, and $H_2 = p\log^2(1/p) + q\log^2(1/q)$; recall that log denotes the logarithm to base e.

Tauberian Theorems. As indicated above, in the proof of our main findings we need to resort to certain Tauberian results. The most classical result in this direction is the following property [10]

Lemma 3. *Suppose that $f(v)$ is a monotone function such that*

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

is asymptotically given by

$$F(v) \sim \frac{v^{a+1}}{(a+1)} \quad (v \to \infty)$$

for some $a > -1$. Then

$$f(v) \sim v^a \quad (v \to \infty).$$

In what follows we consider two variations of this principle. We start with the following simple lemma.

Lemma 4. *Suppose that $f(v)$ is a non-negative increasing function in $v \geq 0$ such that*

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

has the following asymptotic expansion

$$F(v) = C_1 v \log v + C_2 v + o(v) \quad (v \to \infty).$$

Then

$$f(v) = C_1 \log v + C_1 + C_2 + o(1) \quad (v \to \infty).$$

Proof. By the assumption

$$|F(v) - C_1 v \log v + C_2 v| \leq \varepsilon v$$

for $v \geq v_0$. Set $v' = \varepsilon^{1/2} v$, then by monotonicity we obtain (for $v \geq v_0$)

$$\begin{aligned} f(v) &\leq \frac{F(v+v') - F(v)}{v'} \\ &\leq \frac{1}{v'}\left(C_1(v+v')\log(v+v') + C_2(v+v') - C_1 v \log v - C_2 v\right) + \varepsilon \frac{2v+v'}{v'} \\ &= C_1 \log(v+v') + C_2 + C_1 \frac{v}{v'} \log\left(1 + \frac{v'}{v}\right) + \varepsilon \frac{2v+v'}{v'} \\ &= C_1 \log v + C_2 + C_1 + O\left(\varepsilon^{1/2}\right), \end{aligned}$$

where the O-constant is an absolute one. In a similar manner, we obtain the corresponding lower bound (for $v \geq v_0 + v_0^{1/2}$). Hence, it follows that

$$|f(v) - C_1 \log v - C_1 - C_2| \leq C' \varepsilon^{1/2}$$

for $v \geq v_0 + v_0^{1/2}$. This completes the proof of the lemma. ∎

The next lemma shows that we can also transfer certain uniformity properties.

Lemma 5. *Suppose that $f(v,\lambda)$ is a non-negative increasing function in $v \geq 0$, where λ is a real parameter with $|\lambda| \leq \delta$ 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 $|\lambda| \leq \delta$. Then

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

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

Proof. By the assumption

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

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

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

In completely the same way we get a corresponding lower bound (for $v \geq v_0 + v_0^{1/2}$). Hence, the result follows. ∎

3.2 Proof of Theorem 2

We start with the analysis of $A(v)$ defined in (7) and satisfying the functional equation (8). In fact, (8) can be rewritten as

$$F(x) = 1 + F(x - c_1) + F(x - c_2). \tag{25}$$

by setting $A(v) = F(\log v)$, $c_1 = \log \frac{1}{p}$ and $c_2 = \log \frac{1}{q}$. The above recurrence was analyzed in [4] and we can follow the footsteps of [4]. Nevertheless, we provide a complete proof since we cannot apply [4] for $\tilde{S}(v,z)$ (resp. $\tilde{D}(v,z)$), where we need uniformity with respect to z in a neighborhood of $z = 1$. By presenting first a complete proof for $A(v)$, which is easier to understand than that of $\tilde{S}(v,z)$, we gently introduce the reader to our methodology. A more thorough discussion of this methodology can be found in [33].

The following results are different for irrational and rational $\log p / \log q$. If $\log p / \log q$ is rational, then let $L > 0$ denote the largest real number for which $\log(1/p)$ and $\log(1/q)$

are integer multiples of L. We should observe that when $\log p/\log q = d/b$ for some integers b, d such that $\gcd(b,d) = 1$, we can also write $L = \log(1/p)/d = \log(1/q)/b$. Note also that we always have $L \le \log 2$ since $\min\{\log\frac{1}{p}, \log\frac{1}{q}\} \le \log 2$ and $\min\{b,d\} \ge 1$. Furthermore, $L = \log 2$ if and only if $p = q = \frac{1}{2}$.

Let $\langle y\rangle = y - \lfloor y\rfloor$ denote the fractional part of the real number y, and $H = p\log(1/p) + q\log(1/q)$ be the source entropy.

Proposition 1. *If $\log p/\log q$ is irrational then, as $v \to \infty$,*

$$A(v) = \frac{v}{H}(1 + o(1)). \tag{26}$$

If $\frac{\log p}{\log q}$ is rational, then

$$A(v) = \frac{Q_1(\log v)}{H} v + O(v^{1-\eta}) \tag{27}$$

for some $\eta > 0$, where

$$Q_1(x) = \frac{L}{1-e^{-L}} e^{-L\langle \frac{x}{L}\rangle} = \sum_{h\in\mathbb{Z}} \frac{1}{1 + 2\pi i h/L} e^{2\pi i h x/L} \tag{28}$$

and $\mathbb{Z}$ denotes the set of integers.

Proof. First suppose that $\log p/\log q$ is *rational*, that is, $\log(1/p) = Ld$ and $\log(1/q) = Lb$, where b, d are positive coprime integers. Set $G(x) = F(xL)$ with F as in (25). Then

$$G(x) = 1 + G(x-d) + G(x-b) \tag{29}$$

for $x \ge 0$ with $G(x) = 0$ for $x < 0$. Clearly $G(x) = G(\lfloor x\rfloor)$. Thus, it is only necessary to consider integral $x = n$, and with $g(Z) = \sum_{n\ge 0} G(n) Z^n$ the recurrence (29) becomes

$$g(Z) = \frac{1}{(1-Z)(1-Z^d - Z^b)}.$$

The dominating root of the denominator equals $Z_0 = e^{-L} < 1$ and since b and d are coprime there are no other roots on the circle $|Z| = Z_0 = e^{-L}$. Hence

$$G(n) = \frac{1}{(1-e^{-L})(de^{-dL} + be^{-bL})} e^{Ln} + O(e^{Ln(1-\eta)})$$

for some $\eta > 0$. Finally with $F(x) = G(\lfloor x/L\rfloor)$ we obtain

$$\begin{aligned} F(x) &= \frac{L}{(1-e^{-L})H} e^{L\lfloor \frac{x}{L}\rfloor} + O(e^{x(1-\eta)}) \\ &= \frac{L}{(1-e^{-L})H} e^{-L\langle \frac{x}{L}\rangle} e^{x} + O(e^{x(1-\eta)}), \end{aligned}$$

and with $A(v) = F(\log v)$ we finally derive (27). Expressing $\langle x\rangle$ as a Fourier series, we immediately obtain the Fourier expansion for $Q_1(x)$ as in (28).

In the *irrational* case we use the Mellin transform $A^*(s)$ of $A(v)$ defined as (cf. [33])

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

Easy calculation shows

$$A^*(s) = \frac{-1}{s(1-p^{-s}-q^{-s})}, \qquad \Re(s) < -1.$$

In order to find asymptotics of $A(v)$ as $v \to \infty$ we must compute the inverse transform of $A^*(s)$:

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

where $\sigma < -1$. For this purpose one (usually) has to determine the polar singularities of the meromorphic continuation of $A^*(s)$ to the range $\Re(s) \geq -1$, and for this we need the detailed properties of the set

$$\mathcal{Z} = \{s \in \mathbb{C} : p^{-s} + q^{-s} = 1\} \tag{30}$$

that we analyzed in Lemma 1. Recall that $\mathcal{Z} = \{s_k : k \in \mathbb{Z}\}$ with $\Re(s_k) \geq -1$ and $(2k-1)\pi/\log p \leq \Im(s_k) \leq (2k+1)\pi/\log p$; also if $\log p/\log q$ is irrational, then $s_0 = -1$ is the only root of $p^{-s} + q^{-s} = 1$ with $\Re(s) = -1$.

At that stage we can directly apply the Tauberian theorem (for the Mellin transform) by Wiener-Ikehara.[1] However, since we need a generalization of such a Tauberian theorem (where we need uniformity with respect to another parameter z, see Lemma 5) we present the main steps of the proof and then generalize it. In order to demonstrate the convergence problems that (usually) occur in that context we try to apply directly the residue theorem. For every $\sigma > -1$ with $\sigma \notin \{\Re(s) : s \in \mathcal{Z}\}$ we obtain

$$\begin{aligned} A(v) &= -\lim_{T\to\infty} \sum_{s'\in Z,\ \Re(s')<\sigma, |\Im(s')|<T} \operatorname{Res}(A^*(s)\, v^{-s},\ s = s') \\ &\quad -\frac{1}{2\pi i} \lim_{T\to\infty} \int_{\sigma-iT}^{\sigma+iT} \frac{1}{s(1-p^{-s}-q^{-s})} v^{-s}\, ds \\ &= -\lim_{T\to\infty} \sum_{s'\in Z,\ \Re(s')<\sigma, |\Im(s')|<T} \frac{v^{-s'}}{s' H(s')} - \frac{1}{2\pi i} \lim_{T\to\infty} \int_{\sigma-iT}^{\sigma+iT} \frac{1}{s(1-p^{-s}-q^{-s})} v^{-s}\, ds \end{aligned}$$

provided that the series of residues converges and the limit $T \to \infty$ of the last integral exists. Here we have used the notation

$$H(s) = -p^{-s}\log p - q^{-s}\log q.$$

Observe that $H(-1) = -p\log p - q\log q$ equals H.

The problem is that neither the series nor the integral above are absolutely convergent since the integrand is only of order $1/s$. We have to introduce the auxiliary function

$$A_1(v) = \int_0^v A(w) dw$$

[1] One major assumption is that there are no singularities on the line $\Re(s) = -1$ despite $s_0 = -1$. In fact this Tauberian theorem is usually used to prove the prime number theorem. The function $-\zeta'(s)/\zeta(s)$ (where $\zeta(s) = \sum_{n\geq 1} n^{-s}$ denotes the Riemann zeta function) is (almost) the Mellin transform of the the Chebyshev Ψ-function $\Psi(x) = \sum_{p^k \leq x} \log p$. Since $\zeta(s)$ has no zeroes on the line $\Re(s) = 1$, $s \neq 1$, it follows that $\Psi(x) \sim x$ $(x \to \infty)$ which is equivalent to the prime number theorem $\pi(x) = \sum_{p\leq x} 1 \sim x/\log x$.

which is also given by

$$\begin{aligned} A_1(v) &= \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} A^*(s)\frac{v^{-s+1}}{s-1}\,ds \\ &= \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \frac{1}{s(1-p^{-s}-q^{-s})}\frac{v^{-s+1}}{1-s}\,ds, \end{aligned}$$

for $\sigma < -1$. Note that there is no need to consider the limit $T \to \infty$ in this case since the series and the integral are now absolutely convergent. Hence, the above procedure works without any convergence problem. We shift the line of integration to $\Re(s) = \sigma > -1$, in particular we assume that $\sigma > \sigma_0$ (with σ_0 from Lemma 1). Then we have to consider the sum of residues

$$\sum_{s'\in Z,\ \Re(s')<\sigma} \operatorname{Res}\left(A^*(s)\,\frac{v^{-s+1}}{1-s},\ s=s'\right) = \sum_{s'\in\mathcal{Z}} \frac{1}{s'(s'-1)H(s')}v^{-s'+1}$$

and the integral

$$\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} A(s)\frac{v^{-s+1}}{1-s}\,ds = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \frac{1}{s(1-p^{-s}-q^{-s})}\frac{v^{-s+1}}{1-s}\,ds = O(v^{1-\sigma}).$$

Thus, we obtain

$$A_1(v) = \frac{v^2}{2H}(1+Q(\log v)) + O(v^{1-\sigma})$$

where

$$Q(x) = \sum_{s'\in\mathcal{Z}\setminus\{-1\}} \frac{2H}{s'(s'-1)H(s')}e^{-x(s'+1)}.$$

It is easy to show that $Q(x) \to 0$ as $x \to \infty$ (cf. also with [29, Lemma 4] and [33]). Suppose that $\varepsilon > 0$ is given. Then there exists $S_0 = S_0(\varepsilon) > 0$ such that

$$\sum_{s'\in\mathcal{Z},\,|s'|>S_0} \left|\frac{2H}{s'(s'-1)H(s')}\right| < \frac{\varepsilon}{2}.$$

Further, since $\Re(s') > -1$ for all $s' \in \mathcal{Z}\setminus\{-1\}$ it follows that there exists $x_0 = x_0(\varepsilon) > 0$ with

$$\left|\sum_{s'\in\mathcal{Z}\setminus\{-1\},\,|s'|\le S_0} \frac{2H}{s'(s'-1)H(s')}e^{-x(s'+1)}\right| < \frac{\varepsilon}{2}$$

for $x \ge x_0$. Hence $|Q(x)| < \varepsilon$ for $x \ge x_0(\varepsilon)$.

Note that we cannot obtain the rate of convergence for $Q(x)$. This means that $A_1(v) \sim v^2/(2H)$ as $v \to \infty$. Since $A(v)$ is monotonely increasing (by definition) it also follows that $A(v)$ has to satisfy $A(v) \sim v/H$. This proves Proposition 1. ∎

3.3 Proof of Theorem 3

The analysis of $\tilde{S}(v,z)$ is similar to the one just discussed. The Mellin transform of $\tilde{S}(v,z)$ (with respect to v) is

$$\tilde{S}^*(s,z) = \frac{-1}{s(1-zp^{1-s}-zq^{1-s})}, \qquad \Re(s) < 0,$$

and consequently the Mellin of $\tilde{D}(v,z)$[2] becomes

$$\tilde{D}^*(s,z) = \frac{1-z}{s(1-zp^{1-s}-zq^{1-s})} - \frac{1}{s}.$$

In what follows we will first discuss $\tilde{S}(v,1)$, and then $\tilde{D}(v,z)$. Recall that $H_2 = p(\log p)^2 + q(\log q)^2$.

Proposition 2. *If $\log q/\log p$ is irrational, then $\tilde{S}(v,1)$ can be represented as*

$$\tilde{S}(v,1) = \frac{\log v}{H} + \frac{H_2}{2H^2} + o(1) \tag{31}$$

as $v \to \infty$. However, if $\frac{\log q}{\log p}$ is rational, then there exists $\eta > 0$ such that

$$\tilde{S}(v,1) = \frac{\log v}{H} + \frac{H_2}{2H^2} + \frac{Q_2(\log v)}{H} + O(v^{-\eta}), \tag{32}$$

as $v \to \infty$, where

$$Q_2(x) = L \cdot \left(\frac{1}{2} - \left\langle \frac{x}{L} \right\rangle\right) = \sum_{h\neq 0} \frac{1}{2\pi i h/L} e^{2\pi i h x/L}.$$

Proof. As in the proof of Proposition 1 we first consider the *rational* case. With $F(x) = \tilde{S}(e^x,1)$ we find

$$F(x) = 1 + pF(x-c_1) + qF(x-c_2)$$

and with $G(x) = F(Lx)$

$$G(x) = 1 + pG(x-d) + qG(x-b)$$

for $x \geq 0$ and $G(x) = 0$ for $x < 0$. Furthermore, $G(x) = G(\lfloor x \rfloor)$ and $g(Z) = \sum_n G(n)Z^n$ is given by

$$g(Z) = \frac{1}{(1-Z)(1-pZ^d-qZ^b)}.$$

Here the dominating root is $Z_0 = 1$ that is also a double pole. Since b and d are coprime there are no other singularities on the circle $|Z| = 1$. Thus

$$G(n) = \frac{n}{pd+qb} + \left(\frac{pd^2+qb^2}{2(pd+qb)^2} + \frac{1}{2(pd+qb)}\right) + O(e^{-\eta n})$$

[2] In fact, we consider the Mellin transform of $\overline{D}(v,z) = (z-1)\tilde{S}(v,z) + 1_{[v\geq 1]}$ since the constant function 1 has no Mellin transform. Obviously, $\tilde{D}(v,z)$ and $\overline{D}(v,z)$ coincide for $v \geq 1$. Thus this makes no difference for our considerations.

for some $\eta > 0$. With $F(x) = G(\lfloor x/L \rfloor)$ and $\tilde{S}(v,1) = F(\log v)$ we directly obtain the above representation (32).

In the *irrational* case we again use the Mellin transform

$$\tilde{S}^*(s,1) = \frac{-1}{s(1-p^{1-s}-q^{1-s})}, \qquad \Re(s) < 0$$

and consider the polar singularities. Obviously, there is a double pole at $s = 0$ and simple poles at $\mathcal{Z}_1 \setminus \{0\} = \{s \in \mathbb{C} \setminus \{0\} : p^{1-s} + q^{1-s} = 1\}$. (Note that this set is exactly $(\mathcal{Z} \setminus \{-1\}) + 1$, where $\mathcal{Z}$ is defined in (22).) Hence, by Lemma 2 all $s \in \mathcal{Z}_1 \setminus \{0\}$ satisfy $\Re(s) \geq 0$ and for every integer $k \neq 0$ there uniquely exists $s = s_k(1) \in \mathcal{Z}_1$ with $(2k-1)\pi/\log p < \Im(s_k(1)) < (2k+1)\pi/\log p$.

The singular expansion of $\frac{-1}{s(1-p^{1-s}-q^{1-s})}\frac{v^{1-s}}{1-s}$ around $s = 0$ is given by

$$\tilde{S}^*(s,1) = \frac{1}{H}\,\frac{1}{s^2} - \frac{H_2}{2H^2}\,\frac{1}{s} + O(1).$$

Recall that

$$\int_1^\infty v^{s-1}\,dv = -\frac{1}{s} \quad \text{and} \quad \int_1^\infty \log v\, v^{s-1}\,dv = \frac{1}{s^2}.$$

Thus, this singularity contributes to the inverse Mellin transform with

$$\frac{\log v}{H} + \frac{H_2}{2H^2},$$

compare also with [9]. Consequently, by doing just *formal residue calculations* we directly obtain the proposed asymptotic relations for $\tilde{S}(v,1)$.

However, we again have to deal with the convergence problem. We have to introduce the auxiliary function

$$\tilde{S}_1(v,1) = \int_0^v \tilde{S}(w,1)\,dw,$$

that is,

$$\tilde{S}_1(v,1) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \frac{-1}{s(1-p^{1-s}-q^{1-s})}\frac{v^{-s+1}}{1-s}\,ds,$$

for $\sigma < 0$. The singular expansion of $\frac{-1}{s(1-p^{1-s}-q^{1-s})}\frac{v}{1-s}$ at $s = 0$ is given by

$$\frac{v}{Hs^2} - \left(\frac{H_2}{2H^2} - \frac{1}{H}\right)\frac{v}{s}$$

We further have simple poles at $s' \in \mathcal{Z} \setminus \{0\}$. Hence, by shifting the line of integration to $\Re(s) = \sigma' > 0$ and collecting all residues we obtain with help of the absolute convergence for the corresponding series and integral

$$\begin{aligned}\tilde{S}_1(v,1) &= v\frac{\log v}{H} + \left(\frac{H_2}{2H^2} - \frac{1}{H}\right)v + \sum_{s' \in \mathcal{Z}_\infty \setminus \{0\}} \frac{1}{s'(s'-1)H(s'-1)} v^{1-s'} + O(v^{1-\sigma'}) \\ &= v\frac{\log v}{H} + \left(\frac{H_2}{2H^2} - \frac{1}{H}\right)v + o(v)\end{aligned}$$

as $v \to \infty$. Since $\tilde{S}(v,1)$ is monotone (by definition) we can apply Lemma 4 arriving at

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

as required. ∎

Finally, we deal with $\tilde{D}(v,z)$, where we have to take care of the uniformity question with respect to z.

Proposition 3. *Suppose that z is a positive real number with $|z-1| \le \delta$ (for some $0 < \delta < 1$). Then we uniformly have, as $v \to \infty$*

$$\tilde{D}(v,z) = v^{\frac{z-1}{H} - \left(\frac{1}{H} - \frac{H_2}{2H^3}\right)(z-1)^2 + O(|z-1|^3)} \left(1 + O\left(|z-1|^{1/2}\right)\right).$$

Proof. First, let us consider the *rational* case. Writing $F(x) = \tilde{S}(e^x, z)$ and $G(x) = F(Lx)$ we have

$$G(x) = 1 + pzG(x-d) + qzG(x-b)$$

for $x \ge 0$ and $G(x) = 0$ for $x < 0$. Furthermore, as above, $G(x) = G(\lfloor x \rfloor)$. Hence, $g(Z) = \sum_{n \ge 0} G(n) Z^n$ is given by

$$g(Z) = \frac{1}{(1-Z)(1 - pzZ^d - qzZ^b)}.$$

Similarly we can set $U(x) = \tilde{D}(e^x, z) = 1 + (z-1)F(x)$, $V(x) = U(Lx) = 1 + (z-1)G(x)$, and $v(Z) = \sum_n V(n) Z^n$ that is given by

$$v(Z) = \frac{1}{1-Z} + (z-1)g(Z) = \frac{1}{1-Z} - \frac{1-z}{(1-Z)(1 - pzZ^d - qzZ^b)}.$$

Note that $Z = 1$ is no singularity of $v(Z)$.

Let $Z_0(z)$ denote the dominant singularity of $v(Z)$, that is, the real zero of $pzZ^d + qzZ^b = 1$. Then $Z_0(z) = e^{Ls_0(z)}$, where $s_0(z)$ is the real zero of the equation $zp^{1-s_0} + zq^{1-s_0} = 1$ (cf. Lemma 2). As already expressed in (24),

$$s_0(z) = -\frac{z-1}{H} + \left(\frac{1}{H} - \frac{H_2}{2H^3}\right)(z-1)^2 + O(|z-1|^3)$$

if $|z-1| \le \delta$. Again, since b and d are coprime there are no other zeroes on the circle $|Z| = Z_0(z)$. Furthermore, since the zeroes of $pzZ^d + qzZ^b = 1$ vary continuously in z there exist $\eta > 0$ such that there are no other zeros of $pzZ^d + qzZ^b = 1$ within the circle $|Z| \le Z_0(z) + 2\eta \le Z_0(z)e^{\eta}$ for all real z with $|z-1| \le \delta$. Hence, by Cauchy's formula and by the residue theorem we have

$$\begin{aligned} V(n) &= \frac{1}{2\pi i} \int_{|Z| = Z_0(z)e^{\eta}} v(Z) Z^{-n-1}\, dZ \\ &= -\frac{1-z}{(1 - Z_0(z))(pzdZ_0(z)^d + qzbZ_0(z)^b} Z_0(z)^{-n} + O\left(|z-1| Z_0(z)^{-n} e^{-\eta n}\right). \end{aligned}$$

uniformly for all real z with $|z-1| \le \delta$. By using (24) and $Z_0(z) = e^{Ls_0(z)}$ we obtain

$$-\frac{1-z}{(1-Z_0(z))(pzdZ_0(z)^d + qzbZ_0(z)^b)} = 1 + O(|z-1|)$$

and

$$Z_0(z)^{-\left[\frac{\log v}{L}\right]} = v^{-s_0(z)}\left(1 + O(|z-1|)\right).$$

Hence, we directly get

$$\begin{aligned}\tilde{D}(v,z) &= V\left(\left\lfloor \frac{\log v}{L} \right\rfloor\right)\\ &= v^{-s_0(z)}\left(1 + O(|z-1|)\right)\\ &= v^{\frac{z-1}{H} - \left(\frac{1}{H} - \frac{H_2}{2H^3}\right)(z-1)^2 + O(|z-1|^3)}\left(1 + O\left(|z-1|\right)\right).\end{aligned}$$

In the *irrational* case we again use the Mellin transform of $\tilde{D}(v,z)$ which is given by

$$\tilde{D}^*(s,z) = \frac{1-z}{s(1-zp^{1-s}-zq^{1-s})} - \frac{1}{s}.$$

Observe that residue of the *singular value* $s=0$ equals 0 (due to the additional term $-1/s$). Thus, $s=0$ does not contribute.

The unique real polar singularity is $s_0(z)$ (given by the equation $zp^{1-s_0} + zq^{1-s_0} = 1$) As in the previous cases we have to introduce an auxiliary function

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

also given by

$$\tilde{D}_1(v,z) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty}\left(\frac{1-z}{s(1-zp^{1-s}-zq^{1-s})} - \frac{1}{s}\right)\frac{v^{-s+1}}{1-s}\,ds$$

By shifting the integral and collecting residues we obtain (as above)

$$\tilde{D}_1(v,z) = \frac{v^{-s_0(z)+1}}{-s_0(z)+1}(1 + o(|z-1|)$$

as $v \to \infty$ and uniformly for $|z-1| \le \delta$. Note that the uniformity of the error term $o(|z-1|)$ follows from (23) of Lemma 2. Since $\tilde{D}(v,y)$ is monotone in v we finally find (from Lemma 5)

$$\tilde{D}(v,z) = v^{-s_0(z)}\left(1 + o\left(|z-1|^{1/2}\right)\right),$$

where the convergence is again uniform for $|z-1| \le \delta$. ∎

Finally we prove our main result of this paper, a central limit for the phrase length D, that is, Theorem 3.

We first use the local expansion (24) with $z = e^{t/(\log v)^{1/2}}$ to obtain

$$\begin{aligned} v^{-s_0(z)} &= \exp\left(\log v\left(\frac{z-1}{H} - \left(\frac{1}{H} - \frac{H_2}{2H^3}\right)(z-1)^2 + O(|z-1|^3)\right)\right) \\ &= \exp\left(\frac{1}{H}t\sqrt{\log v} + \frac{1}{H}\frac{t^2}{2} - \left(\frac{1}{H} - \frac{H_2}{2H^3}\right)t^2 + O(t^3/\sqrt{\log v})\right) \\ &= \exp\left(\frac{1}{H}t\sqrt{\log v} + \left(\frac{H_2}{H^3} - \frac{1}{H}\right)\frac{t^2}{2} + O(t^3/\sqrt{\log v})\right) \end{aligned}$$

Hence, by Proposition 3 and by the property that $\mathbf{E}[z^D] = \tilde{D}(v, z)$ we get

$$\mathbf{E}\left[e^{t(D-\frac{1}{H}\log v)/\sqrt{\log v}}\right] = e^{-(t/H)\sqrt{\log v}}\mathbf{E}\left[e^{Dt/\sqrt{\log v}}\right] = e^{\frac{t^2}{2}\left(\frac{H_2}{H^3} - \frac{1}{H}\right)} + o(1). \tag{33}$$

By Goncharov's theorem [33] this proves the normal limiting distribution as $v \to \infty$.

The expected value $\mathbf{E}[D]$ has already been determined. In order to get asymptotic information on the variance we just note that (33) implies that

$$\mathbf{Var}[D] \sim \left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log v$$

as $v \to \infty$.

If $\log p/\log q$ is rational it is easy to be more precise. By using an analysis similarly to the above it follows that the second moment of D is asymptotically given by

$$\mathbf{E}[D^2] = \frac{1}{H^2}(\log v)^2 + Q_3(\log v)\log v + Q_4(\log v) + O(v^{-\eta}),$$

where $Q_3(x)$ and $Q_4(x)$ are certain periodic functions. Thus, we get $\mathbf{Var}[D] = Q_5(\log v)\log v + Q_6(\log v) + O(v^{-\eta})$ with periodic function $Q_5(x)$ and $Q_6(x)$. Actually, $Q_5(x)$ is constant, which finally implies

$$\mathbf{Var}[D] = \left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log v + O(1).$$

Recall that $\log v$ and $\log M_r$ are (asymptotically) almost the same, that is, $\log v = \log M_r + O(1)$. Hence, we also have

$$\frac{D - \frac{1}{H}\log M_r}{\left(\left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log M_r\right)^{1/2}} \to N(0,1)$$

and

$$\mathbf{E}[D] = \frac{\log M_r}{H} + O(1), \qquad \mathbf{Var}[D] \sim \left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log M_r.$$

This completes the proof of Theorem 3.

3.4 Redundancy Analysis

We finally collect what is needed to derive asymptotic properties for the redundancy, i.e., we prove Corollary 1.

Recall that the redundancy is defined by

$$R^*_{\mathrm{VB}}(\mathcal{D}, S) = \frac{\log M}{\mathbf{E}[D]} - h(S).$$

The number $M = |\mathcal{D}|$ of phrases is given by

$$M = (m-1)A(v) + 1,$$

where $A(v) = \{y : P(y) \geq 1/v\}$ denotes the number of internal nodes. In particular, for the binary case $m = 2$ we have $M = A(v) + 1$.

In the irrational case we have

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

and

$$M = \frac{v}{H}(1 + o(1)).$$

This directly leads to

$$\mathbf{E}[D] = \frac{\log M}{H} + \frac{\log H}{H} + \frac{H_2}{2H^2} + o(1)$$

and

$$R^*_{\mathrm{VB}}(\mathcal{D}, S) = \frac{H}{\log M}\left(-\frac{H_2}{2H} - \log H\right) + o\left(\frac{1}{\log M_r}\right).$$

In the rational case, we have to be more precise. From

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

and

$$\mathbf{E}[D] = \frac{\log v}{H} + \frac{H_2}{2H^2} + \frac{Q_2(\log v)}{H} + O(v^{\eta})$$

it follows that

$$\mathbf{E}[D] = \frac{\log M}{H} + \frac{\log H}{H} + \frac{H_2}{2H^2} + \frac{Q_2(\log v) - \log Q_1(\log v)}{H} + O(v^{-\eta}).$$

Since

$$Q_2(x) - \log Q_1(x) = \frac{L}{2} - \log L + \log(1 - e^{-L}) = \log\left(\frac{\sinh(L/2)}{L/2}\right)$$

there is no oscillation and

$$\mathbf{E}[D] = \frac{\log M}{H} + \frac{\log H}{H} + \frac{H_2}{2H^2} + \frac{\log\left(\frac{\sinh(L/2)}{L/2}\right)}{H} + O(v^{-\eta}).$$

Thus

$$R^*_{\mathrm{VB}}(\mathcal{D}, S) = \frac{H}{\log M}\left(-\frac{H_2}{2H} - \log H - \log\left(\frac{\sinh(L/2)}{L/2}\right)\right) + O\left(\frac{1}{\log^2 M}\right).$$

This completes the proof of Corollary 1.

References

[1] J. Abrahams, "Code and parse trees for lossless source encoding," *Communications in Information and Systems* 1(2):113-146, April 2001.

[2] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, (Dover, New York, 1972).

[3] F. Cicalese, L. Gargano, and U. Vaccaro, A note on Tunstall codes with applications to optimal approximation of uniform distributions, *IEEE Trans. Inform. Theory* – submitted.

[4] V. Choi and M. J. Golin, Lopsided trees. I. Analyses. *Algorithmica* 31 (2001), no. 3, 240–290.

[5] T. M. Cover, Enumerative Sources Encoding, *IEEE Trans. Inform. Theory*, 19 (1) (1973) 73–77.

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

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

[8] 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.

[9] P. Flajolet, X. Gourdon, amd P. Dumas, Mellin transforms and asymptotics: harmonic sums, Special volume on mathematical analysis of algorithms, *Theoret. Comput. Sci.*, **144**, 3–58, 1995.

[10] P. Henrici, *Applied and Computational Complex Analysis*, Vols. 1–3, John Wiley *&* Sons, New York, 1977.

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

[12] 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.

[13] F. Jelinek, *Probabilistic Information Theory* (McGraw-Hill, New York, 1968).

[14] F. Jelinek and K. S. Schneider, On Variable-Length-to-Block Coding, *IEEE Trans. Inform. Theory*, **18** (6) (1972) 765–774.

[15] 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)

[16] G. L. Khodak, Redundancy Estimates for Word-Based Encoding of Messages Produced by Bernoulli Sources, *Probl. Inform. Trans.*, **8**, (2) (1972) 21–32 (in Russian)

[17] J. Kieffer, Fast generation of Tunstall codes, *Proc. of ISIT 2007*, Nice, France, 2007.

[18] D. Knuth, *The Art of Computer Programming. Sorting and Searching. Vol. 3* (Addison-Wesley, Reading MA, 1973).

[19] R. E. Krichevsky, Universal Data Compression and Retrieval. (Kluwer, Norwell, MA, 1993).

[20] R. E. Krichevsky and V. K. Trofimov, The Performance of Universal Encoding, *IEEE Trans. Information Theory*, **27** (1981) 199–207.

[21] J. C. Lawrence, A New Universal Coding Scheme for the Binary Memoryless Source, *IEEE Trans. Inform. Theory*, 23 (4) (1977) 466–472.

[22] J. L. Massey, "The entropy of a rooted tree with probabilities," *Proc. of ISIT 1983.*

[23] I. Mudrov, An algorithm for enumeration of combinations, *Vyc. Math. and Math. Phys.*, 5 (4) (1965) 776–778 (in Russian).

[24] S. A. Savari and Robert G. Gallager; Generalized Tunstall codes for sources with memory, IEEE Trans. Info. Theory, vol. IT-43, pp. 658 - 668, March 1997.

[25] S. A. Savari, Variable-to-Fixed Length Codes for Predictable Sources, Proc IEEE Data Compression Conference (DCC'98), Snowbird, UT, March 30 - April 1, 1998, pp. 481–490.

[26] S. A. Savari, Variable-to-Fixed Length Codes and the Conservation of Entropy, IEEE Trans. Info. Theory, vol. 45, pp. 1612 - 1620, July 1999.

[27] S. A. Savari, Renewal Theory and Source Coding, *Proceedings of the IEEE*, vol. 88, no. 11, pp. 1692-1702, November 2000.

[28] D. Knuth, *The Art of Computer Programming. Fundamental Algorithms. Vol. 1* (Addison-Wesley, Reading MA, 1968).

[29] W. Schachinger, Limiting distributions for the costs of partial match retrievals in multidimensional tries. *Random Structures and Algorithms* 17 (2000), no. 3-4, 428–459.

[30] J. P. M. Schalkwijk, An Algorithm For Source Coding, *IEEE Trans. Inform. Theory*, 18 (3) (1972) 395–399.

[31] V. M. Sidelnikov, On Statistical Properties of Transformations Carried out by Finite Automata, *Cybernetics*, 6 (1965) 1–14 (in Russian)

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

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

[34] Tj. J. Tjalkens, Efficient and fast data compression codes for discrete sources with memory, Ph.D. Thesis, Eindhoven Univ. Techn., The Netherlands, 1987.

[35] Tj.J. Tjalkens, The Complexity of Minimum Redundancy Coding, *in Proc. 21-th Symp. Inform. Theory in the Benelux* (May 2000) 226-254.

[36] Tj. J. Tjalkens, A Comparative Study of Fixed-to-Variable Length and Variable-to-Fixed Length Source Codes, *in Proc. 25-th Symp. Inform. Theory in the Benelux* (Rolduc, Kerkrade, The Netherlands June 2-4, 2004) 81–88.

[37] T. J. Tjalkens and F. M. J. Willems, "Variable to fixed-length codes for Markov sources," *I.E.E.E. Trans. Inform. Theory* IT-33, 246-257, 1987.

[38] 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.

[39] V. K. Trofimov, Universal Variable-Length to Block Codes for Bernoulli Sources, *Methods of Discrete Analysis*, 29 (Novosibirsk, 1976) 87–100 (in Russian)

[40] B. P. Tunstall, Synthesis of Noiseless Compression Codes, Ph.D. dissertation, (Georgia Inst. Tech., Atlanta, GA, 1967)

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

[42] J. Verhoeff, A new data compression technique, *Annals of Systems Research*, **6** (1977) pp. 139-148.

[43] K. Visweswariah, S. R. Kulkarni, and S. Verdu, "Universal variable-to-fixed length source codes," IEEE Trans. Info. Theory, vol. IT-47, pp. 1461 - 1472, May 2001.