Renewal theory in analysis of tries and strings: Extended abstract

Svante Janson

Renewal theory in analysis of tries and strings: Extended abstract

Svante Janson

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden

We give a survey of a number of simple applications of renewal theory to problems on random strings, in particular to tries and Khodak and Tunstall codes.

Keywords: tries, random strings, Tunstall code, Khodak code

1 Introduction

Although it long has been realized that renewal theory is a useful tool in the study of random strings and related structures, it has not always been used to its full potential. The purpose of the present paper is to give a survey presenting in a unified way some simple applications of renewal theory to a number of problems involving random strings, in particular several problems on tries, which are tree structures constructed from strings. (Other applications of renewal theory to problems on random trees are given in, e.g., [1], [3], [9], [13], [19], [20].)

Since our purpose is to illustrate a method rather than to prove new results, we present a number of problems in a simple form without trying to be as general as possible. In particular, for simplicity we exclusively consider random strings in the alphabet $\{0,1\}$, and assume that the "letters" (bits) ξ_i in the strings are i.i.d. Note, however, that the methods below are much more widely applicable and extend in a straightforward way to larger alphabets. The methods also, at least in principle, extend to, for example, Markov sources where ξ_i is a Markov chain. (See e.g. Szpankowski [24, Section 2.1] and Clément, Flajolet and Vallée [2] for various interesting probability models of random strings.) Indeed, one of the purposes of this paper is to make propaganda for the use of renewal theory to study e.g. Markov models, even if we do not do this in the present paper. (Some such results may appear elsewhere.)

The results below are (mostly) not new; they have earlier been proved by other methods, in particular Mellin transforms. Indeed, such methods often provide sharper results, with better error bounds or higher order terms. Nevertheless, we believe that renewal theory often is a valuable method that yields the leading terms in a simple and intuitive way, and that it ought to be more widely used for this type of problems. Moreover, as said above, this method may be easier to extend to other situations.

We treat a number of problems on random tries in Sections 3–5 (insertion depth, imbalance, size). Tunstall and Khodak codes are studied in Section 6. A random walk in a region bounded by two crossing lines is studied in Section 7, where we give a (partial) extension of a result by Drmota and Szpankowski [6]. Further results, including related results for b-tries and Patricia tries, and detailed proofs are given in the full-length paper [14].

Notation

We use $\xrightarrow{\mathrm{p}}$ and $\xrightarrow{\mathrm{d}}$ for convergence in probability and in distribution, respectively.

If Z_n is a sequence of random variables and μ_n and σ_n^2 are sequences of real numbers with $\sigma_n^2 > 0$ (for large n, at least), then $Z_n \sim \mathrm{AsN}(\mu_n, \sigma_n^2)$ means that $(Z_n - \mu_n)/\sigma_n \xrightarrow{\mathrm{d}} N(0,1)$.

We denote the fractional part of a real number x by $\{x\} := x - \lfloor x \rfloor$.

Acknowledgement. I thank Allan Gut and Wojciech Szpankowski for inspiration and helpful discussions.

2 Preliminaries

Suppose that $\Xi^{(1)}, \Xi^{(2)}, \dots$ is an i.i.d. sequence of random infinite strings $\Xi^{(n)} = \xi_1^{(n)}\xi_2^{(n)}\cdots$, with letters $\xi_i^{(n)}$ in an alphabet $\mathcal{A}$. (When the superscript n does not matter we drop it; we thus write $\Xi = \xi_1\xi_2\cdots$ for a generic string in the sequence.) For simplicity, we consider only the case $\mathcal{A} = \{0,1\}$, and further assume that the individual letters ξ_i are i.i.d. with $\xi_i \sim \mathrm{Be}(p)$ for some fixed $p \in (0,1)$, i.e., $\mathbb{P}(\xi_i = 1) = p$ and $\mathbb{P}(\xi_i = 0) = q := 1 - p$.

Given a finite string $\alpha_1\cdots\alpha_n \in \mathcal{A}^n$, let $P(\alpha_1\cdots\alpha_n)$ be the probability that the random string Ξ begins with $\alpha_1\cdots\alpha_n$. In particular, for a single letter, $P(0) = q$ and $P(1) = p$, and in general

$$P(\alpha_1\cdots\alpha_n) = \prod_{i=1}^n P(\alpha_i) = \prod_{i=1}^n p^{\alpha_i} q^{1-\alpha_i}. \tag{2.1}$$

Given a random string $\xi_1\xi_2\cdots$, we define

$$X_i := -\ln P(\xi_i) = -\ln\bigl(p^{\xi_i} q^{1-\xi_i}\bigr) = \begin{cases} -\ln q, & \xi_i = 0, \\ -\ln p, & \xi_i = 1. \end{cases} \tag{2.2}$$

Note that $X_1, X_2, \dots$ is an i.i.d. sequence of positive random variables with

$$\mathbb{E}\, X_i = H := -p\ln p - q\ln q, \tag{2.3}$$

the usual *entropy* of each letter ξ_i, and

$$\mathbb{E}\, X_i^2 = H_2 := p\ln^2 p + q\ln^2 q, \tag{2.4}$$

$$\operatorname{Var} X_i = H_2 - H^2 = pq(\ln p - \ln q)^2 = pq\ln^2(p/q). \tag{2.5}$$

The variance (2.5) is in data compression known as the *minimal coding variance*, see [16]. Note that the case $p = q = 1/2$ is special; in this case $X_i = \ln 2$ is deterministic and $\operatorname{Var} X_i = 0$.

By (2.2), X_i is supported on $\{\ln(1/p), \ln(1/q)\}$. It is well-known, both in renewal theory and in the analysis of tries, that one frequently has to distinguish between two cases: the *arithmetic* (or *lattice*) case when the support is a subset of $d\mathbb{Z}$ for some $d > 0$, and the *non-arithmetic* (or *non-lattice*) case when it is not. This yields the following cases:

arithmetic The ratio $\ln p/\ln q$ is rational. More precisely, X_i then is d-arithmetic, where d equals $\gcd(\ln p, \ln q)$, the largest positive real number such that $\ln p$ and $\ln q$ both are integer multiples of d. If $\ln p/\ln q = a/b$, where a and b are relatively prime positive integers, then

$$d = \gcd(\ln p, \ln q) = \frac{|\ln p|}{a} = \frac{|\ln q|}{b}. \tag{2.6}$$

non-arithmetic The ratio $\ln p/\ln q$ is irrational.

We let S_n denote the partial sums of X_i: $S_n := \sum_{i=1}^n X_i$. Thus

$$P(\xi_1 \cdots \xi_n) = \prod_{i=1}^n P(\xi_i) = \prod_{i=1}^n e^{-X_i} = e^{-S_n}. \tag{2.7}$$

(This is a random variable, since it depends on the random string $\xi_1 \cdots \xi_n$; it can be interpreted as the probability that another random string $\Xi^{(j)}$ begins with the same n letters as observed.)

We introduce the standard renewal theory notation (see e.g. Gut [8, Chapter 2]), for $t \geq 0$ and $n \geq 1$,

$$\nu(t) := \min\{n : S_n > t\}. \tag{2.8}$$

We also allow the summation to start with an initial random variable X_0, which is independent of $X_1, X_2, \ldots$, but may have an arbitrary real-valued distribution. We then define

$$\widehat{S}_n := \sum_{i=0}^n X_i = X_0 + \sum_{i=1}^n X_i, \tag{2.9}$$

$$\widehat{\nu}(t) := \min\{n : \widehat{S}_n > t\}. \tag{2.10}$$

3 Insertion depth in a trie

A *trie* is a binary tree structure designed to store a set of strings. It is constructed from the strings by a recursive procedure, see e.g. Knuth [15, Section 6.3], Mahmoud [17, Chapter 5] or Szpankowski [24, Section 1.1]: If there is just one string, it is stored in the root; otherwise, the strings beginning with 0 are passed to the left subtree, and the strings beginning with 1 to the right subtree, and the construction is repeated in the subtrees, with a node at depth k inspecting the $(k+1)$th bit of the strings that are passed to it.

The trie is a finite subtree of the complete infinite binary tree $\mathcal{T}_\infty$, where the nodes can be labelled by finite strings $\boldsymbol{\alpha} = \alpha_1 \cdots \alpha_k \in \mathcal{A}^* := \bigcup_{k=0}^\infty \mathcal{A}^k$ (the root is the empty string). A string Ξ is stored at the node labelled by $\boldsymbol{\alpha}$ if $\boldsymbol{\alpha}$ is the shortest prefix of Ξ that is not a prefix of any other string in the set.

Let D_n be the depth (= path length) of the node containing a given string, for example the first, in the trie constructed from n random strings $\Xi^{(1)}, \ldots, \Xi^{(n)}$. (By symmetry, any of the n strings will have a depth with the same distribution.) Denoting the chosen string by $\Xi = \xi_1\xi_2\cdots$, the depth D_n is thus at most k if and only if no other of the strings begins with $\xi_1 \cdots \xi_k$. Conditioning on the string Ξ, each of the other strings has this beginning with probability $P(\xi_1 \cdots \xi_k)$, and thus by independence,

$$\mathbb{P}(D_n \leq k \mid \Xi) = \bigl(1 - P(\xi_1 \cdots \xi_k)\bigr)^{n-1} = \bigl(1 - e^{-S_k}\bigr)^{n-1}. \tag{3.1}$$

Let $X_0 = X_0^{(n)}$ be a random variable, independent of Ξ, with the distribution

$$\mathbb{P}(X_0^{(n)} > x) = \bigl(1 - e^x/n\bigr)_+^{n-1} = \bigl(1 - e^{x - \ln n}\bigr)_+^{n-1}, \qquad x \in (-\infty, \infty). \tag{3.2}$$

Then, for any $k \geq 1$,

$$\mathbb{P}(D_n \leq k) = \mathbb{P}\bigl(X_0 > \ln n - S_k\bigr) = \mathbb{P}\bigl(\widehat{S}_k > \ln n\bigr) = \mathbb{P}\bigl(\widehat{\nu}(\ln n) \leq k\bigr) \tag{3.3}$$

and thus

$$D_n \xrightarrow{\mathrm{d}} \widehat{\nu}(\ln n). \tag{3.4}$$

Further, as $n \to \infty$, the quantity in (3.2) converges to $\exp(-e^x)$, and thus $X_0^{(n)} \to X_0^*$, where $-X_0^*$ has the Gumbel distribution with $\mathbb{P}(-X_0^* \le x) = \exp(-\exp(-x))$.

We can apply standard renewal theory theorems, and immediately obtain the following. For other, earlier proofs see Knuth [15, Sections 6.3 and 5.2], Pittel [21, 22] and Mahmoud [17, Section 5.5]. The Markov case is treated by Jacquet and Szpankowski [12], ergodic strings by Pittel [21], and a class of general dynamical sources by Clément, Flajolet and Vallée [2].

Theorem 3.1. *For every $p \in (0,1)$,*

$$\frac{D_n}{\ln n} \xrightarrow{\mathrm{p}} \frac{1}{H}, \tag{3.5}$$

with H the entropy given by (2.3)*. Moreover, the convergence holds in every L^r, $r < \infty$, too. Hence, all moments converge in* (3.5) *and*

$$\mathbb{E}\, D_n^r \sim H^{-r}(\ln n)^r, \qquad 0 < r < \infty. \tag{3.6}$$

Theorem 3.2. *More precisely:*

(i) *If $\ln p/\ln q$ is irrational, then, as $n \to \infty$,*

$$\mathbb{E}\, D_n = \frac{\ln n}{H} + \frac{H_2}{2H^2} + \frac{\gamma}{H} + o(1). \tag{3.7}$$

(ii) *If $\ln p/\ln q$ is rational, then, as $n \to \infty$,*

$$\mathbb{E}\, D_n = \frac{\ln n}{H} + \frac{H_2}{2H^2} + \frac{\gamma}{H} + \psi_1(\ln n) + o(1), \tag{3.8}$$

where $\psi_1(t)$ is a small continuous function, with period $d = \gcd(\ln p, \ln q)$ in t, given by

$$\psi_1(t) := -\frac{1}{H} \sum_{k \ne 0} \Gamma(-2\pi \mathrm{i} k/d) e^{2\pi \mathrm{i} k t/d}. \tag{3.9}$$

Theorem 3.3. *Suppose that $p \in (0,1)$. Then, as $n \to \infty$,*

$$\frac{D_n - H^{-1}\ln n}{\sqrt{\ln n}} \xrightarrow{\mathrm{d}} N\Big(0, \frac{\sigma^2}{H^3}\Big),$$

with $\sigma^2 = H_2 - H^2 = pq(\ln p - \ln q)^2$. If $p \ne 1/2$, then $\sigma^2 > 0$ and this can be written as

$$D_n \sim \mathrm{AsN}\big(H^{-1}\ln n, H^{-3}\sigma^2 \ln n\big).$$

Moreover,

$$\operatorname{Var} D_n = \frac{\sigma^2}{H^3}\ln n + o(\ln n).$$

In the argument above, X_0 depends on n. This is a nuisance, although no real problem. An alternative that avoids this problem is to Poissonize by considering a random number of strings. In this case it is simplest to consider $1+\mathrm{Po}(\lambda)$ strings, so that a selected string Ξ is compared to a Poisson number $\mathrm{Po}(\lambda)$ of other strings, for a parameter $\lambda \to \infty$. Conditioned on Ξ, the number of other strings beginning with $\xi_1 \cdots \xi_k$ then has the Poisson distribution $\mathrm{Po}(\lambda P(\xi_1 \cdots \xi_k))$. Thus we obtain instead of (3.4), now denoting the depth by D_λ, $D_\lambda \stackrel{\mathrm{d}}{=} \widehat{\nu}(\ln \lambda)$, where $X_0 := X_0^*$ now is independent of n.

We obtain the same asymptotics as for D_n above. It is in this case easy to depoissonize, by noting that D_n is stochastically monotone in n, and derive the results for D_n from the results for D_λ by choosing $\lambda = n \pm n^{2/3}$.

4 Imbalance in tries

Mahmoud [18] studied the imbalance factor of a string in a trie, defined as the number of steps to the right minus the number of steps to the left in the path from the root to the leaf where the string is stored. We define

$$Y_i := 2\xi_i - 1 = \begin{cases} -1, & \xi_1 = 0, \\ +1, & \xi_1 = 1, \end{cases}$$

and denote the corresponding partial sums by $V_k := \sum_{i=1}^k Y_i$. Thus the imbalance factor Δ_n of the string Ξ in a random trie with n strings is V_{D_n}, with D_n as in Section 3 the depth of the string.

It follows easily that (3.3) holds also conditioned on the sequence $(Y_1, Y_2, \dots)$, and as a consequence

$$(D_n, \Delta_n) = (D_n, V_{D_n}) \stackrel{\mathrm{d}}{=} \bigl(\widehat{\nu}(\ln n), V_{\widehat{\nu}(\ln n)}\bigr).$$

In particular,

$$\Delta_n \stackrel{\mathrm{d}}{=} V_{\widehat{\nu}(\ln n)}.$$

A general renewal theory theorem [8, Section 4.2] applies, and we obtain the central limit theorem by Mahmoud [17]:

Theorem 4.1. *As $n \to \infty$,*

$$\Delta_n \sim \mathrm{AsN}\left(\frac{p-q}{H}\ln n,\ \frac{pq\ln^2(pq)}{H^3}\ln n\right).$$

5 The expected size of a trie

A trie built of n strings as in Section 3 has n external nodes, since each external node contains exactly one string. However, the number of internal nodes, W_n, say, is random. We will study its expectation. For simplicity we Poissonize directly and consider a trie constructed from $\mathrm{Po}(\lambda)$ strings; we let $\widetilde{W}_\lambda$ be the number of internal nodes. The results below have previously been found by other methods, in particular, more precise asymptotics have been found using Mellin transforms; see Knuth [15], Mahmoud [17], Fayolle, Flajolet, Hofri and Jacquet [7], and, in particular, Jacquet and Régnier [10, 11]. The Markov case is studied by Régnier [23] and dynamical sources by Clément, Flajolet and Vallée [2].

If $\boldsymbol{\alpha} = \alpha_1 \cdots \alpha_k$ is a finite string, let $I(\boldsymbol{\alpha})$ be the indicator of the event that $\boldsymbol{\alpha}$ is an internal node in the trie. This event occurs if and only if there are at least two strings beginning with $\boldsymbol{\alpha}$. In our Poisson model, the number of strings beginning with $\boldsymbol{\alpha}$ has a Poisson distribution $\mathrm{Po}(\lambda P(\boldsymbol{\alpha}))$, and thus

$$\mathbb{E}\,\widetilde{W}_\lambda = \sum_{\boldsymbol{\alpha}\in\mathcal{A}^*} \mathbb{E}\,I(\boldsymbol{\alpha}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}^*} \mathbb{P}\bigl(\mathrm{Po}(\lambda P(\boldsymbol{\alpha})) \ge 2\bigr) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}^*} f(\lambda P(\boldsymbol{\alpha})), \tag{5.1}$$

where

$$f(x) := \mathbb{P}\bigl(\mathrm{Po}(x) \ge 2\bigr) = 1 - (1+x)e^{-x}. \tag{5.2}$$

Sums of the type in (5.1) are often studied using Mellin transform inversion and residue calculus. Renewal theory presents an alternative, and the key renewal theorem implies the following. As said in the introduction, this opens the way to straightforward generalizations, e.g. to Markov sources.

Theorem 5.1. *Suppose that f is a non-negative function on $(0,\infty)$, and that $F(\lambda) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}^*} f(\lambda P(\boldsymbol{\alpha}))$, with $P(\boldsymbol{\alpha})$ given by* (2.1)*. Assume further that f is a.e. continuous and satisfies the estimates*

$$f(x) = O(x^2), \quad 0 < x < 1, \qquad \text{and} \qquad f(x) = O(1), \quad 1 < x < \infty. \tag{5.3}$$

Let $g(t) := e^t f(e^{-t})$.

(i) *If $\ln p/\ln q$ is irrational, then, as $\lambda\to\infty$,*

$$\frac{F(\lambda)}{\lambda} \to \frac{1}{H}\int_{-\infty}^{\infty} g(t)\,\mathrm{d}t = \frac{1}{H}\int_0^\infty f(x)x^{-2}\,\mathrm{d}x. \tag{5.4}$$

(ii) *If $\ln p/\ln q$ is rational, then, as $\lambda\to\infty$,*

$$\frac{F(\lambda)}{\lambda} = \frac{1}{H}\psi(\ln\lambda) + o(1), \tag{5.5}$$

where, with $d := \gcd(\ln p, \ln q)$ given by (2.6)*, ψ is a bounded d-periodic function having the Fourier series*

$$\psi(t) \sim \sum_{m=-\infty}^{\infty} \widehat{\psi}(m) e^{2\pi\mathrm{i}mt/d} \tag{5.6}$$

with

$$\widehat{\psi}(m) = \widehat{g}(-2\pi m/d) = \int_{-\infty}^{\infty} e^{2\pi\mathrm{i}mt/d} g(t)\,\mathrm{d}t = \int_0^\infty f(x)x^{-2-2\pi\mathrm{i}m/d}\,\mathrm{d}x. \tag{5.7}$$

Furthermore,

$$\psi(t) = d\sum_{k=-\infty}^{\infty} g(kd-t). \tag{5.8}$$

If f is continuous, then ψ is too.

Returning to $\widetilde{W}_\lambda$, we obtain the following for the expected number of internal nodes in the Poisson trie.

Theorem 5.2. (i) *If* $\ln p/\ln q$ *is irrational, then, as* $\lambda\to\infty$,

$$\frac{\mathbb{E}\,\widetilde{W}_\lambda}{\lambda}\to\frac{1}{H}. \tag{5.9}$$

(ii) *If* $\ln p/\ln q$ *is rational, then, as* $\lambda\to\infty$,

$$\frac{\mathbb{E}\,\widetilde{W}_\lambda}{\lambda}=\frac{1}{H}+\frac{1}{H}\psi_2(\ln\lambda)+o(1), \tag{5.10}$$

where, with $d=\gcd(\ln p,\ln q)$, ψ_2 *is a continuous* d*-periodic function with average* 0 *and Fourier expansion*

$$\psi_2(t)=\sum_{k\neq0}\frac{\Gamma(1-2\pi\mathrm{i}k/d)}{1+2\pi\mathrm{i}k/d}e^{2\pi\mathrm{i}kt/d}=\sum_{k\neq0}\frac{2\pi\mathrm{i}k}{d}\Gamma\Big(-1-\frac{2\pi\mathrm{i}k}{d}\Big)e^{2\pi\mathrm{i}kt/d}.$$

The case of a fixed number n of strings is easily handled by comparison, and (5.9) and (5.10) imply the corresponding results for W_n:

Theorem 5.3. (i) *If* $\ln p/\ln q$ *is irrational, then, as* $n\to\infty$,

$$\frac{\mathbb{E}\,W_n}{n}\to\frac{1}{H}.$$

(ii) *If* $\ln p/\ln q$ *is rational, then, as* $n\to\infty$, *with* ψ_2 *as in Theorem 5.2,*

$$\frac{\mathbb{E}\,W_n}{n}=\frac{1}{H}+\frac{1}{H}\psi_2(\ln n)+o(1).$$

6 Tunstall and Khodak codes

Tunstall and Khodak codes are variable-to-fixed length codes that are used in data compression. See [4], [5] and the survey [25] for details and references, as well as for an analysis using Mellin transforms.

The idea is that an infinite string can be parsed as a unique sequence of nonoverlapping *phrases* belonging to a certain (finite) *dictionary* $\mathcal{D}$.

By a random phrase we mean a phrase distributed as the unique initial phrase in a random infinite string Ξ. Thus a phrase $\boldsymbol{\alpha}$ in the dictionary $\mathcal{D}$ is chosen with probability $P(\boldsymbol{\alpha})$. We let the random variable L be the length of a random phrase.

In Khodak's construction of such a dictionary, we fix a threshold $r\in(0,1)$ and construct a parsing tree as the subtree of the complete infinite binary tree such that the internal nodes are the strings $\boldsymbol{\alpha}=\alpha_1\cdots\alpha_k$ with $P(\boldsymbol{\alpha})\geq r$; the external nodes are thus the strings $\boldsymbol{\alpha}$ such that $P(\boldsymbol{\alpha})<r$ but the parent, $\boldsymbol{\alpha}'$ say, has $P(\boldsymbol{\alpha}')\geq r$. The phrases in the Khodak code are the external nodes in this tree. For convenience, we let $R=1/r>1$. Let $M=M(R)$ be the number of phrases in the Khodak code.

In Tunstall's construction, we are instead given a number M. We start with the empty phrase and then iteratively $M-1$ times replace a phrase $\boldsymbol{\alpha}$ having maximal $P(\boldsymbol{\alpha})$ by its two children $\boldsymbol{\alpha}0$ and $\boldsymbol{\alpha}1$.

It is easily seen that Khodak's construction with some $r>0$ gives the same result as Tunstall's with $M=M(R)$. Conversely, a Tunstall code is almost a Khodak code, with r chosen as the smallest $P(\boldsymbol{\alpha})$ for a proper prefix $\boldsymbol{\alpha}$ of a phrase; the difference is that Tunstall's construction handles ties more flexibly; there

may be some phrases too with $P(\boldsymbol{\alpha}) = r$. Thus, Tunstall's construction may give any desired number M of phrases, while Khodak's does not. We will see that in the non-arithmetic case, this difference is asymptotically negligible, while it is important in the arithmetic case. (This is very obvious if $p = q = 1/2$, when Khodak's code always gives a dictionary size M that is a power of 2.)

Let us first consider the number of phrases, $M = M(R)$, in Khodak's construction with a threshold $r = 1/R$. This is a purely deterministic problem, but we may nevertheless apply our probabilistic renewal theory arguments. In fact, M, the number of leafs in the parsing tree, equals 1 + the number of internal nodes. Thus, $M = 1 + \sum_{\boldsymbol{\alpha}} f(RP(\boldsymbol{\alpha}))$ with $f(x) := \mathbf{1}[x \ge 1]$, and we may apply Theorem 5.1.

Theorem 6.1. *Consider the Khodak code with threshold $r = 1/R$.*

(i) *If $\ln p/\ln q$ is irrational, then, as $R \to \infty$,*

$$\frac{M(R)}{R} \to \frac{1}{H}.$$

(ii) *If $\ln p/\ln q$ is rational, then, as $R \to \infty$,*

$$\frac{M(R)}{R} = \frac{1}{H} \cdot \frac{d}{1-e^{-d}} e^{-d\{(\ln R)/d\}} + o(1).$$

Next, consider the length L of a random phrase. We will use the notation L_M^{T} for a Tunstall code with M phrases and L_R^{K} for a Khodak code with threshold $r = 1/R$.

Consider first the Khodak code. By construction, given a random string $\Xi = \xi_1\xi_2\cdots$, the first phrase in it is $\xi_1\cdots\xi_n$ where n is the smallest integer such that $P(\xi_1\cdots\xi_n) = e^{-S_n} < r = e^{-\ln R}$. Hence, by (2.8),

$$L_R^{\mathsf{K}} = \nu(\ln R). \tag{6.1}$$

Hence, renewal theory immediately yields the following (as well as convergence of higher moments).

Theorem 6.2. *For the Khodak code, the following holds as $R \to \infty$, with $\sigma^2 = H_2 - H^2 = pq\ln^2(p/q)$:*

$$\frac{L_R^{\mathsf{K}}}{\ln R} \xrightarrow{\text{a.s.}} \frac{1}{H}, \tag{6.2}$$

$$L_R^{\mathsf{K}} \sim \mathrm{AsN}\Big(\frac{\ln R}{H},\, \frac{\sigma^2}{H^3}\ln R\Big), \tag{6.3}$$

$$\operatorname{Var} L_R^{\mathsf{K}} \sim \frac{\sigma^2}{H^3}\ln R. \tag{6.4}$$

If $\ln p/\ln q$ is irrational, then

$$\mathbb{E}\, L_R^{\mathsf{K}} = \frac{\ln R}{H} + \frac{H_2}{2H^2} + o(1). \tag{6.5}$$

If $\ln p/\ln q$ is rational, then, with $d := \gcd(\ln p, \ln q)$ given by (2.6)*,*

$$\mathbb{E}\, L_R^{\mathsf{K}} = \frac{\ln R}{H} + \frac{H_2}{2H^2} + \frac{d}{H}\Big(\frac{1}{2} - \Big\{\frac{\ln R}{d}\Big\}\Big) + o(1). \tag{6.6}$$

In the arithmetic case, it suffices to consider thresholds such that $-\ln r = \ln R$ is a multiple of d; in this case (6.6) becomes

$$\mathbb{E}\, L_R^{\mathsf{K}} = \frac{\ln R}{H} + \frac{H_2}{2H^2} + \frac{d}{2H} + o(1). \tag{6.7}$$

We analyze the Tunstall code by comparing it to the Khodak code, which leads to the following result.

Theorem 6.3. *For the Tunstall code, the following holds as $M \to \infty$, with $\sigma^2 = H_2 - H^2 = pq\ln^2(p/q)$:*

$$\frac{L_M^{\mathsf{T}}}{\ln M} \xrightarrow{\text{a.s.}} \frac{1}{H}, \tag{6.8}$$

$$L_M^{\mathsf{T}} \sim \mathrm{AsN}\Big(\frac{\ln M}{H},\ \frac{\sigma^2}{H^3}\ln M\Big), \tag{6.9}$$

$$\operatorname{Var} L_M^{\mathsf{T}} \sim \frac{\sigma^2}{H^3}\ln M. \tag{6.10}$$

If $\ln p/\ln q$ is irrational, then

$$\mathbb{E}\, L_M^{\mathsf{T}} = \frac{\ln M}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + o(1). \tag{6.11}$$

If $\ln p/\ln q$ is rational, then, with $d := \gcd(\ln p, \ln q)$ given by (2.6)*,*

$$\mathbb{E}\, L_M^{\mathsf{T}} = \frac{\ln M}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + \frac{1}{H}\ln\frac{\sinh(d/2)}{d/2} + \frac{d}{H}\psi_4\Big(\Big\{\frac{\ln M + \ln(H(1-e^{-d})/d)}{d}\Big\}\Big) + o(1), \tag{6.12}$$

where

$$\psi_4(x) := \frac{e^{dx}-1}{e^d-1} - x. \tag{6.13}$$

Note that ψ_4 is continuous, with $\psi_4(0) = \psi_4(1) = 0$. ψ_4 is convex and thus $\psi_4 \le 0$ on [0,1]. In the symmetric case $p = q = 1/2$, $d = H = \ln 2$ and $\psi_4(x) = 2^x - 1 - x$, with a minimum $-0.086071\ldots$.

7 A stopped random walk

Drmota and Szpankowski [6] consider walks in a region in the first quadrant bounded by two crossing lines. One of their results is about a random walk in the plane taking only unit steps north or east, which is stopped when it exits the region; the probability of an east step is p each time. Coding steps east by 1 and north by 0, this is the same as taking our random string Ξ. Drmota and Szpankowski [6] study, in our notation, the exit time

$$D_{K,V} := \min\{n : n > K \text{ or } S_n > V\ln 2\}$$

for given numbers K and V, with K integer. We thus have

$$D_{K,V} = (K+1) \wedge \nu(V\ln 2). \tag{7.1}$$

We have here kept the notations K and V from [6], but for convenience we in the sequel write $V_2 := V\ln 2$. We assume $p \neq q$, since otherwise $D_{K,V} = (K \wedge \lfloor V \rfloor) + 1$ is deterministic.

We need a little more notation. Let as usual $\phi(x) := (2\pi)^{-1/2}e^{-x^2/2}$ and $\Phi(x) := \int_{-\infty}^{x} \phi(y)\,\mathrm{d}y$ be the density and distribution functions of the standard normal distribution. Further, let

$$\Psi(x) := \int_{-\infty}^{x} \Phi(y)\,\mathrm{d}y = x\Phi(x) + \phi(x). \tag{7.2}$$

We can now state our version of the result by Drmota and Szpankowski [6]. We do not obtain as sharp error estimates as they do; on the other hand, our result is more general and includes the transition region when $V_2/H \approx K$ and both stopping conditions are important.

Theorem 7.1. *Suppose that $p \neq q$ and that $V, K \to \infty$. Let $V_2 := V\ln 2$ and $\widetilde{\sigma}^2 := (H_2 - H^2)/H^3 > 0$.*

(i) *If $(K - V_2/H)/\sqrt{V_2} \to +\infty$, then $D_{K,V}$ is asymptotically normal:*

$$D_{K,V} \sim \mathrm{AsN}\Big(\frac{V_2}{H}, \widetilde{\sigma}^2 V_2\Big). \tag{7.3}$$

Further, $\mathrm{Var}(D_{K,V}) \sim \widetilde{\sigma}^2 V_2$.

(ii) *If $(K - V_2/H)/\sqrt{V_2} \to -\infty$, then $D_{K,V}$ is asymptotically degenerate:*

$$\mathbb{P}(D_{K,V} = K+1) \to 1. \tag{7.4}$$

Further, $\mathrm{Var}\, D = o(V_2)$.

(iii) *If $(K - V_2/H)/\sqrt{V_2} \to a \in (-\infty, +\infty)$, then $D_{K,V}$ is asymptotically truncated normal:*

$$V_2^{-1/2}(D_{K,V} - V_2/H) \xrightarrow{\mathrm{d}} (\widetilde{\sigma}Z) \wedge a = \widetilde{\sigma}\big(Z \wedge (a/\widetilde{\sigma})\big). \tag{7.5}$$

with $Z \sim N(0,1)$. Further,

$$\mathrm{Var}(D_{K,V}) \sim V_2 \,\mathrm{Var}(\widetilde{\sigma}Z \wedge a) = V_2\widetilde{\sigma}^2 \,\mathrm{Var}(Z \wedge (a/\widetilde{\sigma})).$$

(iv) *In every case,*

$$\mathbb{E}\, D_{K,V} = \frac{V_2}{H} - \widetilde{\sigma}\sqrt{V_2}\Psi\Big(\frac{V_2/H - K}{\widetilde{\sigma}\sqrt{V_2}}\Big) + o(\sqrt{V_2}) \tag{7.6}$$

$$= K - \widetilde{\sigma}\sqrt{V_2}\Psi\Big(\frac{K - V_2/H}{\widetilde{\sigma}\sqrt{V_2}}\Big) + o(\sqrt{V_2}). \tag{7.7}$$

(v) *If $(K - V_2/H)/\sqrt{V_2} \geq \ln V_2$, then*

$$\mathbb{E}\, D_{K,V} = \frac{V_2}{H} + \frac{H_2}{2H^2} + \psi_5(V_2) + o(1), \tag{7.8}$$

where $\psi_5 = 0$ in the non-arithmetic case and $\psi_5(t) = \frac{d}{H}\big(1/2 - \{t/d\}\big)$ in the d-arithmetic case.

(vi) *If $(K - V_2/H)/\sqrt{V_2} \leq -\ln V_2$, then*

$$\mathbb{E}\, D_{K,V} = K + 1 + o(1). \tag{7.9}$$

References

[1] N. Broutin & C. Holmgren, Branching Markov chains: stability and applications. Preprint, 2010.

[2] J. Clément, P. Flajolet & B. Vallée, Dynamical sources in information theory: a general analysis of trie structures. *Algorithmica* **29** (2001), no. 1–2, 307–369.

[3] F. Dennert & R. Grübel, Renewals for exponentially increasing lifetimes, with an application to digital search trees. *Ann. Appl. Probab.* **17** (2007), no. 2, 676–687.

[4] M. Drmota, Y. Reznik, S. Savari & W. Szpankowski, Precise asymptotic analysis of the Tunstall code. *Proc. 2006 International Symposium on Information Theory (Seattle, 2006)*, 2334–2337.

[5] M. Drmota, Y. A. Reznik & W. Szpankowski, Tunstall code, Khodak variations, and random walks. *IEEE Transactions on Information Theory* **56** (2010), no. 6, 2928–2937.

[6] M. Drmota & W. Szpankowski, On the exit time of a random walk with positive drift. *Proceedings, 2007 Conference on Analysis of Algorithms, AofA 07 (Juan-les-Pins, 2007)*, Discrete Math. Theor. Comput. Sci. Proc. **AH** (2007), 291–302.

[7] G. Fayolle, P. Flajolet, M. Hofri & P. Jacquet, Analysis of a stack algorithm for random multiple-access communication. *IEEE Trans. Inform. Theory* **31** (1985), no. 2, 244–254.

[8] A. Gut, *Stopped Random Walks*. 2nd ed., Springer, New York, 2009.

[9] C. Holmgren, Novel characteristics of split trees by use of renewal theory. Preprint, 2010.

[10] P. Jacquet & M. Régnier, Normal limiting distribution of the size of tries. *Performance '87 (Brussels, 1987)*, 209–223, North-Holland, Amsterdam, 1988.

[11] P. Jacquet & M. Régnier, New results on the size of tries. *IEEE Trans. Inform. Theory* **35** (1989), no. 1, 203–205.

[12] P. Jacquet & W. Szpankowski, Analysis of digital tries with Markovian dependency. *IEEE Trans. Inform. Theory*, **37** (1991), no. 5, 1470–1475.

[13] S. Janson, One-sided interval trees. *J. Iranian Statistical Society* **3** (2004), no. 2, 149–164.

[14] S. Janson, Renewal theory in analysis of tries and strings. Tech. Report 2009:28, Uppsala. `http://arxiv.org/abs/0912.2174`

[15] D.E. Knuth, *The Art of Computer Programming. Vol. 3: Sorting and Searching*. 2nd ed., Addison-Wesley, Reading, Mass., 1998.

[16] I. Kontoyiannis, Second-order noiseless source coding theorems. *IEEE Trans. Inform. Theory* **43** (1997), no. 4, 1339–1341.

[17] H. Mahmoud, *Evolution of Random Search Trees*, Wiley, New York, 1992.

[18] H. Mahmoud, Imbalance in random digital trees. *Methodol. Comput. Appl. Probab.* **11** (2009), no. 2, 231–247.

[19] H. Mohamed & P. Robert, A probabilistic analysis of some tree algorithms. *Ann. Appl. Probab.* **15** (2005), no. 4, 2445–2471.

[20] H. Mohamed & P. Robert, Dynamic tree algorithms. *Ann. Appl. Probab.* **20** (2010), no. 1, 26–51.

[21] B. Pittel, Asymptotical growth of a class of random trees. *Ann. Probab.* **13** (1985), no. 2, 414–427.

[22] B. Pittel, Paths in a random digital tree: limiting distributions. *Adv. in Appl. Probab.* **18** (1986), no. 1, 139–155.

[23] M. Régnier, Trie hashing analysis, *Proc. Fourth Int. Conf. Data Engineering (Los Angeles, 1988)*, IEEE, 1988, pp. 377–387.

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

[25] W. Szpankowski, Average redundancy for known sources: ubiquitous trees in source coding. *Proceedings, Fifth Colloquium on Mathematics and Computer Science (Blaubeuren, 2008)*, Discrete Math. Theor. Comput. Sci. Proc. **AI** (2008), 19–58.