Discrete Mathematics and Theoretical Computer Science **(subm.)**, by the authors, 1–rev

On the Exit Time of a Random Walk with Positive Drift†

Michael Drmota[1] and Wojciech Szpankowski[2]

[1] *Institute of Discrete Mathematics and Geometry, TU Wien, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria*
[2] *Department of Computer Science, Purdue University, W. Lafayette, IN 47907, USA*

We study a random walk with positive drift in the first quadrant of the plane. For a given connected region C of the first quadrant, we analyze the number of paths contained in C and the first exit time from C. In our case, region C is bounded by two crossing lines. It is noted that such a walk is equivalent to a path in a tree from the root to a leaf not exceeding a given height. If this tree is the parsing tree of the Tunstall or Khodak fixed-to-variable code, then the exit time of the underlying random walk corresponds to the phrase length not exceeding a given length. We derive precise asymptotics of the number of paths and the asymptotic distribution of the exit time. Even for such a simple walk, the analysis turns out to be quite sophisticated and it involves Mellin transforms, Tauberian theorems, and infinite number of saddle points.

Keywords: Random walk in the plane, exit time, number of paths, Tunstall's code, Khodak code, Mellin transform, Tauberian theorems.

1 Introduction

Let $C \subseteq \mathbb{R}^2_{\geq 0}$ be a bounded connected region of the first quadrant of the plane with the property that if an integer lattice point $(k_1,k_2) \neq (0,0)$ with non-negative integers k_1,k_2 is contained in C, then either (k_1-1,k_2y) or (k_1,k_2-1) is in C, too. Figure 1 illustrates such a walk and region C (grey area). Let also $\mathcal{L}(C)$ denote the set of lattice paths starting at the origin $(0,0)$ with steps of the form $L=(1,0)$ and $R=(0,1)$ such that they exit region C at the last step D (exit time). We shall study both the exit time distribution and the number of paths.

In this paper, we are particularly interested in regions C that are bounded by two lines of the form $ax_1+bx_2=c_1$ (with $a,b>0$ and $a\neq b$) and $x_1+x_2=c_2$ (cf. Figure 2). For later use we will assume (w.l.o.g.) that $a=\log_2\frac{1}{p}$ and $b=\log_2\frac{1}{q}$, where $0<p<q<1$ and $p+q=1$; $\log_2$ denotes the logarithm to base 2.

We should point out that there is an obvious bijection between $\mathcal{L}$ and a binary tree T, where every path in $\mathcal{L}$ corresponds to an external node $\square$ as illustrated in Figure 1. Clearly, the binary tree T has the property that if an w is an internal node of T that is connected to the root by k_1 steps to the right (R) and

† The first author was supported by the Austrian Science Foundation (FWF) Project S9604. The work of the second author was supported in part by the NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742, NIH Grant R01 GM068959-01.

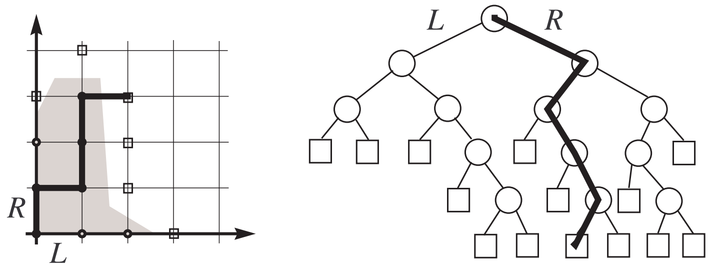

Fig. 1: Lattice paths and binary trees

k_2 steps to the left (L), then T contains all $\binom{k_1+k_2}{k_1}$ possible internal nodes that are connected to the root with (in total) k_1 steps to the right and k_2 steps to the left.

This correspondence was in fact the starting point of our analysis. In our recent work (1) we studied Tunstall and Khodak fixed-to-variable codes, see also (8) for a related result. Briefly, a variable-to-fixed length encoder partitions the source string over a binary alphabet into a concatenation of variable-length phrases. Each phrase belongs to a given dictionary $\mathcal{D}$. We represent this dictionary by a complete parsing tree T, i.e., a tree in which every internal node has both children. The dictionary entries $d \in \mathcal{D}$ correspond to the leaves of parsing tree. If the dictionary $\mathcal{D}$ is has M entries, then the code word for each phrase has $\lceil \log_2 M \rceil$ bits.

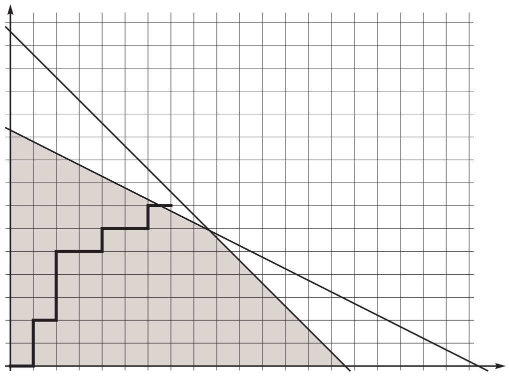

Fig. 2: Lattice paths in a bounded region

Tunstall's code (11) is the best known variable-to-fixed length code; however, it was independently discovered by Khodak (5). Khodak's construction is particularly simple: Let p and $q = 1 - p > p$ be the probability of the binary symbol and let growing a complete parsing tree satisfying $pr \leq P(d) < r$ for $d \in \mathcal{D}$. It follows that if y is a proper prefix of one or more entries of $\mathcal{D}$, i.e., y corresponds to an internal node of T, then $P(y) \geq r$. Let $v = 1/r$. Then, it is shown in (1) that in order to analyze the Khodak code, one needs to investigate the following sums

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

for some function $f(v)$. Since $y = p^{k_1}q^{k_2}$ for some nonnegative integers $k_1, k_2 \geq 0$, we conclude that the above summation set can be expressed, after setting $v = 2^V$, as

$$k_1 \log_2 \frac{1}{p} + k_2 \log_2 \frac{1}{q} \leq V$$

that corresponds to the first line of the boundary of region C of our random walk. Imposing another condition on the phrase length (path in the parsing tree), namely, that it cannot exceed, say K, the above sum becomes

$$A_K(v) = \sum_{y:P(y)\geq 1/v,\,|y|\leq K} f(v)$$

with the second boundary line becoming $k_1 + k_2 = K$ as in our random walk.

While there is a substantial literature on random walks in the first quadrant of the plane (2; 4), the problem we analyze here seems to be unique and only some partial results were reported this far; see Janson (6). Our methodology belongs to analytic algorithmics and is rather sophisticated. After translating the above sums into a recurrence, we apply the Mellin transform and Tauberian theorem to discover that we need to handle infinite saddle points on a line (incidently, already encountered in (7)). This leads to some oscillations in the leading term for the number of paths. We also prove the central limit theorem for the exit time.

2 Main Results

We will discuss two problems. The first one is a counting problem. Set

$$C_{K,V} := \{(x_1, x_2) \in \mathbb{R}^2_{\geq 0} : x_1 + x_2 \leq K,\ x_1 \log_2 \frac{1}{p} + x_2 \log_2 \frac{1}{q} \leq V\}$$

Let $\mathcal{L}_{K,V}$ be the corresponding set of lattice paths and $T_{K,V}$ be the associated binary tree. The first result concerns the number of paths $|\mathcal{L}_{K,V}|$. In this context it is natural to let K be an integer variable and V a continuous variable.

In the formulation of the theorem we will make use of $s_{\mathrm{sp}} = s_{\mathrm{sp}}(K,V)$ defined as

$$R(s_{\mathrm{sp}}) = \frac{p^{-s_{\mathrm{sp}}} + q^{-s_{\mathrm{sp}}}}{p^{-s_{\mathrm{sp}}} \log \frac{1}{p} + q^{-s_{\mathrm{sp}}} \log \frac{1}{q}} = \frac{K}{V \log 2}$$

Note that $s_{\mathrm{sp}} > -1$ if and only if $K/V < (\log 2)/H$.

We further set

$$T(s) = \frac{p^{-s} \log^2 \frac{1}{p} + q^{-s} \log^2 \frac{1}{q}}{p^{-s} + q^{-s}} - \left(\frac{p^{-s} \log \frac{1}{p} + q^{-s} \log \frac{1}{q}}{p^{-s} + q^{-s}} \right)^2$$

and will use the periodic function

$$Q_L(s,x) = \frac{L}{1-e^{sL}} e^{sL\langle \frac{x}{L}\rangle} = \sum_{m\in\mathbf{Z}} \frac{1}{(-s)+\frac{2\pi i m}{L}} e^{\frac{2\pi i m}{L}x},$$

where $s \in \mathbb{C}$ and $x, L \in \mathbb{R}$; $\langle y\rangle = y - \lfloor y \rfloor$ denotes the fractional part of a real number y.

Finally, we set $H = p\log(1/p) + q\log(1/q)$ (that can be interpreted as the entropy of the distribution p,q) and and for later use we set $H_2 = p\log^2(1/p) + q\log^2(1/q)$.

Theorem 1. *Suppose that $\delta > 0$ is given.*

1. *Assume that K and V satisfy the constraints*

$$\frac{V\log 2}{H}\cdot(1+\delta) \le K \le \frac{V\log 2}{\min\{\log(1/p),\log(1/q)\}}\cdot(1-\delta). \tag{1}$$

If $\log p/\log q$ is irrational, then as $K,V\to\infty$

$$|\mathcal{L}_{K,V}| = \frac{2^V}{H}(1+o(1)). \tag{2}$$

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

$$|\mathcal{L}_{K,V}| = \frac{Q_L(-1,V\log 2)}{H}2^V + O(2^{V(1-\eta)}) \tag{3}$$

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

2. *Next, if*

$$\frac{2V\log 2}{\log(1/p)+\log(1/q)}\cdot(1+\delta) \le K \le \frac{V\log 2}{H}\cdot(1-\delta), \tag{4}$$

then

$$|\mathcal{L}_{K,V}| \sim \sum_{\ell\ge 0} \frac{Q_\delta\left(s_{\mathrm{sp}},(K-\ell)\log p - V\log 2\right)}{(p^{-s_{\mathrm{sp}}}+q^{-s_{\mathrm{sp}}})^\ell}\cdot\frac{(p^{-s_{\mathrm{sp}}}+q^{-s_{\mathrm{sp}}})^K 2^{-Vs_{\mathrm{sp}}}}{\sqrt{2\pi K T(s_{\mathrm{sp}})}}, \tag{5}$$

where $\Delta = \log q - \log p$. If $\log p/\log q = d/r$ is rational, then (5) simplifies to

$$|\mathcal{L}_{K,V}| \sim \sum_{j=0}^{d-r-1} \frac{e^{2\pi i\frac{j}{\delta}} Q_L\left(s_{\mathrm{sp}} - \frac{2\pi i j}{\delta}, K\log p - V\log 2\right)}{1-\frac{e^{2\pi i \frac{jd}{d-r}}}{p^{-s_{\mathrm{sp}}}+q^{-s_{\mathrm{sp}}}}}\cdot\frac{(p^{-s_{\mathrm{sp}}}+q^{-s_{\mathrm{sp}}})^K 2^{-Vs_{\mathrm{sp}}}}{\sqrt{2\pi K T(s_{\mathrm{sp}})}}. \tag{6}$$

3. *If*

$$\frac{V\log 2}{\min\{\log(1/p),\log(1/q)\}}\cdot(1+\delta) \le K \le \frac{2V\log 2}{\log(1/p)+\log(1/q)}\cdot(1-\delta). \tag{7}$$

then (for some $\eta > 0$)

$$|\mathcal{L}_{K,V}| = 2^{K+1} - O(2^{K(1-\eta)}). \tag{8}$$

For the second problem we adjust the lattice paths in $\mathcal{L}_{K,V}$ with a natural probability distribution. If $y \in \mathcal{L}_{K,V}$ consists of k_1 steps of the form R and k_2 steps of the form L then we set $P(y) := p^{k_1} q^{k_2}$. This is exactly the probability distribution that is induced by a random walk that starts at $(0,0)$ and is generated by independent steps R and L with probabilities p and q. Since every path y eventually leaves $C_{K,V}$ we surely have $\sum_{y\in\mathcal{L}_{K,V}} P(y) = 1$. Of course, we can also think of the corresponding trees $T_{K,V}$ and its external nodes. Our second result concerns the exit time $D_{K,V}$ of this random walk, that is, the number of steps $|y| = k_1 + k_2$ of $y \in \mathcal{L}_{K,V}$.

Theorem 2. *Let $D_{K,V}$ denote the exit time of the above described random walk and fix $\delta > 0$.*

1. *If (1) holds, then we have, as $K,V \to \infty$,*

$$\frac{D_{K,V} - \frac{1}{H}\log|\mathcal{L}_{K,V}|}{\left(\left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log|\mathcal{L}_{K,V}|\right)^{1/2}} \xrightarrow{\text{d}} N(0,1),$$

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

$$\mathbf{E}\,D_{K,V} = \frac{\log|\mathcal{L}_{K,V}|}{H} + \frac{\log H}{H} + \frac{H_2}{2H^2} + \frac{-\log L + \log(1-e^{-L}) + \frac{L}{2}}{H} + o\left(\frac{1}{\log|\mathcal{L}_{K,V}|}\right),$$

where $L = 0$ if $\log p/\log q$ is irrational and $L > 0$ is defined as in Theorem 1 if $\log p/\log q$ is rational. Further

$$\mathbf{Var}\,D_{K,V} = \left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log|\mathcal{L}_{K,V}| + O(1).$$

2. *If (4) or (7) holds, then the distribution of $D_{K,V}$ is asymptotically concentrated at $K+1$, that is,*

$$\mathbf{Pr}\{D_{K,V} \neq K+1\} = O(e^{-\eta K})$$

as $K,V \to \infty$ for some $\eta > 0$. We also have $\mathbf{E}\,D_{K,V} = K+1+O(e^{-\eta K})$ and $\mathbf{Var}\,D_{K,V} = O(e^{-\eta K})$.

In passing we observe that a random walk (that starts at $(0,0)$ and is generated by independent steps R and L with probabilities p and q) has an average position (pm,qm) after m steps. Further by approximating this random walk by a Brownian motion it is clear that the deviation from the mean is (almost surely) bounded by $O(\sqrt{m\log\log m})$. Thus, if (1) holds then the Brownian motion approximation can be used to derive the central limit theorem, (see, for example, (6)). The bound coming from $k_1 + k_2 \le K$ has practically no influence. However, in the second and third case ((4) and (7)), the bound $k_1 \log\frac{1}{p} + k_2 \log\frac{1}{q} \le V$ is negligible and, thus, the exit time is concentrated at $K+1$.

The special cases $K/V \sim (\log 2)/H$ and $K/V \sim (2\log 2)/(\log\frac{1}{p} + \log\frac{1}{q})$ are not covered by Theorems 1 and 2. In fact it is possible to characterize the limiting behaviour of $|\mathcal{L}_{K,V}|$ and $D_{K,V}$ but the statements (and also the derivations) are very involved and not discussed here.

3 Analysis of a Recurrence

As above, for any lattice path y we set $P(y) = p^{k_1} q^{k_2}$ if y consists of k_1 steps R and k_2 steps L. We further set $v = 2^V$. Then $k_1 \log_2\frac{1}{p} + k_2 \log_2\frac{1}{q} \le V$ is equivalent to $P(y) \ge 1/v$. Observe that

$$A_K(v) = \sum_{y: P(y)\ge 1/v,\, |y|\le K} 1$$

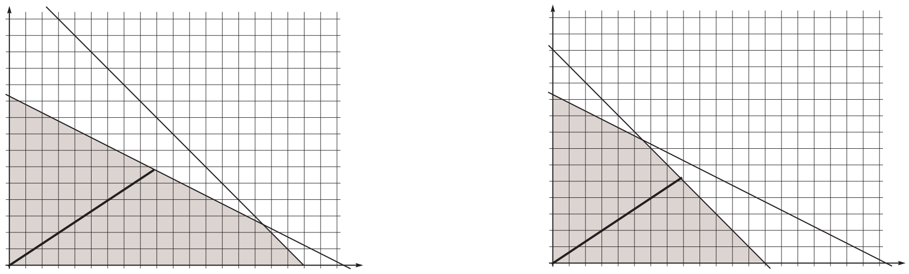

Fig. 3: The drift in the first and second case of Theorem 2

is the number of lattice paths with endpoints contained in $C_{K,V}$. Due to the binary tree interpretation of these lattice paths we have

$$|\mathcal{L}_{K,V}| = A_K(v) + 1 = A_K(2^V) + 1$$

since the number of external nodes of a binary tree exceeds the number of internal nodes by exactly 1.

For the proof of the limit laws of the exit time we will also make use of the following similar sum

$$S_K(v,z) = \sum_{y: P(y) \geq 1/v, |y| \leq K} P(y) z^{|y|}.$$

that will be analyzed as $A_K(v)$.

First, by definition it is clear that $A_K(v) = 0$ and $S_K(v,z) = 0$ for $v < 1$ and all $K \geq 0$, however, for $v \geq 1$ we recursively have

$$A_{K+1}(v) = 1 + A_K(vp) + A_K(vq) \quad \text{and} \quad S_{K+1}(v,z) = 1 + pzS_K(vp,z) + qzS_K(vq,z).$$

From this recursive description we immediately obtain the corresponding relations for the Mellin transforms, namely

$$A^*_{K+1}(s) = -\frac{1}{s} + (p^{-s} + q^{-s}) A^*_K(s) \qquad (\Re(s) < -1)$$

and

$$S^*_{K+1}(s,z) = -\frac{1}{s} + (zp^{1-s} + zq^{1-s}) S^*_K(s,z) \qquad (\Re(s) < 0).$$

Recall that the Mellin transform $f^*(s)$ of a function $f(v)$ is defined by

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

Since $A^*_0(v) = S^*_0(v,z) = -\frac{1}{s}$ we explicitly find

$$A^*_K(s) = -\frac{1 - (p^{-s} + q^{-s})^{K+1}}{s(1 - (p^{-s} + q^{-s}))} \quad \text{and} \quad S^*_K(s,z) = -\frac{1 - (z(p^{1-s} + q^{1-s}))^{K+1}}{s(1 - z(p^{1-s} + q^{1-s}))}.$$

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

$$A_K(v) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{\sigma - iT}^{\sigma + iT} A^*_K(s) v^{-s} \, ds, \tag{9}$$

where $\sigma < -1$.

3.1 First case of Theorem 1

We first assume that we are in the first case of Theorem 1, that is, the relation (1) holds. Then we know that $s_{\mathrm{sp}} < -1$. We split up the integral (9) into two parts:

$$I_1 = -\frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT} -\frac{1}{s(1-(p^{-s}+q^{-s}))}v^{-s}\,ds, \quad I_2 = -\frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT} \frac{(p^{-s}+q^{-s})^{K+1}}{s(1-(p^{-s}+q^{-s}))}v^{-s}\,ds.$$

In order to handle I_1 we have to know something about the set of zeros of the denominator. The following lemma is probably due to Schachinger (9) and independently due to Jacquet, see (10).

Lemma 1. *Suppose that $0 < p < q < 1$ and set $Z = \{s \in \mathbf{C} : p^{-s} + q^{-s} = 1\}$. Then we have*

(i) All $s \in Z$ satisfy $-1 \le \Re(s) \le \sigma_0$, where σ_0 is a positive solution of $1 + q^{-s} = p^{-s}$. Furthermore, for every integer k there uniquely exists $s_k \in Z$ with $(2k-1)\pi/\log p < \Im(s_k) < (2k+1)\pi/\log p$ and consequently $Z = \{s_k : k \in \mathbb{Z}\}$.

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

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

$$s_k = s_{k \bmod d} + \frac{2k\pi i}{\log p}, \qquad (k \notin \{0,1,\ldots,d-1\})$$

that is, all $s \in Z$ 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.

This means that if $\log p/\log q$ is irrational, then there is only one *dominating zero* on the *critical line* $\Re(s) = -1$ and we directly get $I_1 \sim 2^V/H$ by an application of the Tauberian theorem of Wiener-Ikehara (formulated for the Mellin transform) as discussed in (1).

If $\log p/\log q$ is rational, then we have to be more careful. We shift the integral to $\sigma > -1$ with $\sigma < \min\{\Re s_1, \ldots, \Re s_{d-1}\}$ and obtain

$$\begin{aligned}
I_1 &= -\lim_{T\to\infty}\sum_{s'\in Z,\ \Re(s')<-1, |\Im(s')|<T} \mathrm{Res}(A^*(s)\,v^{-s},\ s=s') - \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT}\frac{1}{s(1-p^{-s}-q^{-s})}v^{-s}\,ds \\
&= -\lim_{T\to\infty}\sum_{s'\in Z,\ \Re(s')=-1, |\Im(s')|<T} \frac{v^{-s'}}{s'H(s')} - \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT}\frac{1}{s(1-p^{-s}-q^{-s})}v^{-s}\,ds
\end{aligned}$$

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. We have to face the problem that, both, the series and the integral are not absolutely convergent since the integrand is only of order $1/s$. Nevertheless it is possible to show that these limits exists – we omit the technical details – and we find (with help of Lemma 1)

$$I_1 = \sum_{m\in\mathbb{Z}} \frac{v^{1-2\pi i\frac{m}{L}}}{(1-2\pi i\frac{m}{L})H} + O(v^{1-\eta_1}) = 2^V\frac{S(-1, V\log 2)}{H} + O(v^{1-\eta_1}),$$

where $0 < \eta_1 = \sigma + 1 < \min\{\Re s_1, \ldots, \Re s_{d-1}\} + 1$.

The integral I_2 has to be treated in a completely different way. First of all, we shift the integral to the line $\sigma = s_{\rm sp} < -1$ and observe that

$$(p^{-s_{\rm sp}} + q^{-s_{\rm sp}})^K v^{-s_{\rm sp}} = v^{R(s_{\rm sp})\log(p^{-s_{\rm sp}}+q^{-s_{\rm sp}})-s_{\rm sp}}$$

and that $R(s_{\rm sp})\log(p^{-s_{\rm sp}} + q^{-s_{\rm sp}}) - s_{\rm sp} < 1$ if $s_{\rm sp} \neq -1$. Hence, we expect that I_2 can be estimated by

$$I_2 = O\left(v^{R(s_{\rm sp})\log(p^{-s_{\rm sp}}+q^{-s_{\rm sp}})-s_{\rm sp}}\right) = O\left(2^{V(1-\eta_2)}\right)$$

(for some $\eta_2 > 0$) which is actually true. Again we have to overcome the technical problem that the integral is not absolutely convergent.

3.2 Second case of Theorem 1

Next we assume that we are in the second case of Theorem 1, that is, the relation (4) holds. Here $s_{\rm sp} > -1$ and we do not split the integral (9) into two parts. Of course, we again shift the integral to a line $\sigma > -1$, namely to $\sigma = s_{\rm sp} > -1$. Note that the zeros Z of the denominator are no singularities of the function $A_K^*(s)$ since the numerator has the same zeros. Nevertheless, the integral

$$A_K(v) = -\frac{1}{2\pi i}\lim_{T\to\infty}\int_{s_{\rm sp}-iT}^{s_{\rm sp}+iT}\frac{1-(p^{-s}+q^{-s})^{K+1}}{s(1-(p^{-s}+q^{-s}))}v^{-s}\,ds, \tag{10}$$

needs a delicate analysis. It is again not absolutely convergent but this is just a technical question. The second problem comes from the fact that on the line of integration there are infinitely many *saddle points*. First note that $s = s_{\rm sp}$ is a saddle point of the mapping $s \mapsto (p^{-s}+q^{-s})^K v^{-s} = e^{K\log(p^{-s}+q^{-s})-sV\log 2}$ and, thus the integral from $s_{\rm sp} - iK^{\frac12-\varepsilon}$ to $s_{\rm sp} + iK^{\frac12-\varepsilon}$ (for some $\varepsilon > 0$) is asymptotically given by

$$\frac{1}{\sqrt{2\pi K T(s_{\rm sp})}}\frac{(p^{-s_{\rm sp}}+q^{-s_{\rm sp}})^{K+1}2^{-s_{\rm sp}V}}{s_{\rm sp}(1-(p^{-s_{\rm sp}}+q^{-s_{\rm sp}}))}$$

However, as already noted this is not the only saddle point on this line of integration. Set $t_h = 2\pi h/(\log p - \log q) = -2\pi i h/\Delta$. Then all points $s = s_{\rm sp} + it_h$, $h \in \mathbb{Z}$, are saddle points, as already observed in (7). Consequently, the total contribution of the integral is asymptotically given by

$$\sum_{h\in\mathbb{Z}}\frac{1}{\sqrt{2\pi K T(s_{\rm sp})}}\frac{(p^{-s_{\rm sp}}+q^{-s_{\rm sp}})^{K+1}2^{-s_{\rm sp}V}p^{-it_h(K+1)}2^{-iVt_h}}{(s_{\rm sp}+it_h)(1-(p^{-s_{\rm sp}}+q^{-s_{\rm sp}})p^{-it_h})}.$$

The representations (5) and (6) follow after a few lines of computation by using the Fourier expansion of $Q_L(s,x)$.

4 Exit Time

In order to treat the exit time $D_{K,V}$ we make again use of the corresponding tree $T_{K,V}$ and the following useful lemma (cf. (1)).

Lemma 2. *Let T be an m-ary tree, let X denote the set of leaves and Y the set of internal nodes. Furthermore, we assume a probability distribution $p_1,\dots,p_m$ on an m-ary alphabet A and identify a node in T with a word over A in the usual way. Then we have*

$$\sum_{x\in X} P(x)\frac{z^{|x|}-1}{z-1} = \sum_{y\in Y} P(y) z^{|y|}. \tag{11}$$

This lemma directly implies that the probability generating function of $D_K(v,z) = \mathbb{E}z^{D_{K,V}}$ (where $v = 2^V$) is given by

$$D_K(v,z) = (z-1)S_K(v,z)+1 \qquad (v\ge 1)$$

and consequently its Mellin transform has the following representation (for $\Re(s)<0$):

$$D_K^*(s,z) = (z-1)S_K^*(s,z) - \frac{1}{s} = \frac{(1-z)(1-z(p^{1-s}+q^{1-s})^{K+1})}{s(1-z(p^{1-s}+q^{1-s}))} - \frac{1}{s}.$$

Hence, we have for any $\sigma<0$

$$\mathbb{E}z^{D_{K,V}} = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT}\left(\frac{(1-z)(1-z(p^{1-s}+q^{1-s})^{K+1})}{s(1-z(p^{1-s}+q^{1-s}))} - \frac{1}{s}\right) v^{-s}\,ds.$$

In the first case of Theorem 2, that is, if (1) holds then we split up the integral into two parts $I_1(z)+I_2(z)$:

$$I_1(z) = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT}\left(\frac{(1-z)}{s(1-z(p^{1-s}+q^{1-s}))} - \frac{1}{s}\right) v^{-s}\,ds,$$
$$I_2(z) = -\frac{1}{2\pi i}\lim_{T\to\infty}\int_{\sigma-iT}^{\sigma+iT}\frac{(1-z)(z(p^{1-s}+q^{1-s})^{K+1}}{s(1-z(p^{1-s}+q^{1-s}))} v^{-s}\,ds.$$

Observe that the residue of the *singular value* $s=0$ in the integrand of $I_1(z)$ equals 0 (due to the additional term $-1/s$). Thus, $s=0$ does not contribute if we shift the line of integration to $\sigma>0$. The only polar singularity on the real line of the integrand of $I_1(z_0)$ is $s_0(z)$ that is given by the equation $z(p^{1-s_0}+q^{1-s_0}) = 1$ and has an asymptotic representation of the form

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

The residue of the integrand in $I_1(z_0)$ at $s=s_0(z)$ equals $v^{-s_0(z)}$. Of course, there is an analogue to Lemma 1 for all zeros of the equation $z(p^{1-s}+q^{1-s}) = 1$. By shifting the line of integration in $I_1(z)$ to the right (and taking care that there is no absolute convergence) one finds

$$I_1(z) = v^{-s_0(z)}\left(1+O\left(|z-1|\right)\right).$$

In the second integral $I_2(z)$ we shift the line of integration to $\sigma = s_{\mathrm{sp}}+1<0$ and get an negligible exponentially small error term. Consequently we have

$$\mathbb{E}z^{D_{K,V}} = v^{-s_0(z)}\left(1+O(|z-1|)\right) = v^{\frac{z-1}{H} - \left(\frac{1}{H} - \frac{H_2}{2H^3}\right)(z-1)^2 + O(|z-1|^3)}\left(1+O\left(|z-1|\right)\right)$$

which proves asymptotic normality by setting $z = e^t$ (together with a proper scaling). For the derivation of the mean value (and variance) we refer to (1).

In the second case of Theorem 2 (where (4) or (7) holds) we do not split up the integral into two parts, which implies that the integrand has no singular points other than $s = 0$. We shift the line of integration to $\sigma = s_{\mathrm{sp}} + 1 > 0$ and obtain (again by taking care that there is no absolute convergence)

$$\mathbb{E} z^{D_{K,V}} = z^{K+1} + O(|z-1|v^{-\eta}) \tag{12}$$

where $\eta = s_{\mathrm{sp}} + 1 > 0$. From the construction we know that $D_{K,V} \le K+1$. From (12) we can easily deduce that $D_{K,V}$ is in fact concentrated at $K+1$. By Markov's inequality (for $z < 1$) we directly obtain

$$\mathbf{Pr}\{D_{K,V} \le K\} \le z^{-K}\mathbb{E}\left(z^{D_{K,V}}\mathbf{1}_{\{D_{K,V}\le K\}}\right) = z^{-K}\left(\mathbb{E}z^{D_{K,V}} - z^{K+1}\right) + z\mathbf{Pr}\{D_{K,V} \le K\}$$

which implies (with $z = 1 - \frac{1}{K}$) the estimate $\mathbf{Pr}\{D_{K,V} \le k\} = O(v^{-\eta})$. This proves concentration. We have $v = 2^V$ and, thus, $v^{-\eta} = 2^{-\eta V}$ is exponentially small. By using the corresponding tail estimate of the form $\mathbf{Pr}\{D_{K,V} \le K - r\} = O(e^{-r/K}v^{-\eta})$, we can also deal with moments and obtain $\mathbb{E}D_{K,V} = K + 1 + O(K^2 v^{-\eta})$ and $\mathbf{Var}\, D_{K,V} = O(K^3 v^{-\eta})$.

References

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

[2] G. Fayolle, R. Iasnogorodski, V. Malyshev, *Random Walks in the Quarter-Plane: Algebraic Methods, Boundary Value Problems and Applications*, Springer, 1999.

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

[4] P. Flajolet and R. Sedgewick, Ânalytic Combinatorics, Cambridge University Press, 2007.

[5] 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)

[6] 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.

[7] G. Park, H-K. Hwang, P. Nicodeme, and W. Szpankowski, Profile in Tries, preprint 2006.

[8] 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.

[9] 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.

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

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