A q–ANALOGUE OF THE PATH LENGTH OF BINARY SEARCH TREES

HELMUT PRODINGER

Abstract. A reformulation of the path length of binary search trees is given in terms of permutations, allowing to extend the definition to the instance of words, where the letters are obtained by independent geometric random variables (with parameter q). In this way, expressions for expectation and variance are obtained which in the limit for $q \to 1$ are the classical expressions.

The path length $\rho(t)$ of a binary search tree t satisfies the recursion $\rho(t) = \rho(t_L) + \rho(t_R) + |t_L| + |t_R|$ where t_L and t_R are the left resp. right subtree of the root. ($|t|$ denotes the size of the tree t, i. e. the number of nodes.)

Binary search trees are obtained from permutations. For some background see [5, 2, 3]. Our aim is to rewrite the definition of the path length in terms of permutations, since then we are able to obtain q–analogues: This is done by considering *words* over the alphabet $\{1, 2, \dots\}$ instead, with probabilities (weights) for the letters $p, pq, pq^2, \dots$, where $p + q = 1$ (geometric probability distribution). In the limit $q \to 1$, this model turns into the model of random permutations, as equal letters appear with probability 0 and each relative ordering is equally likely.

For a permutation $\pi = \pi_1 \dots \pi_n$ we define $\rho(\pi)$ by

$$\rho(\pi) = \Big|\{(j,k) \mid 1 \le j < k \le n,\ \pi_j = \min\{\pi_j, \dots, \pi_k\} \quad \text{or} \quad \pi_k = \min\{\pi_j, \dots, \pi_k\}\}\Big|.$$

Then $\rho(\square) = 0$ and, if $\pi = \sigma 1 \tau$, then $\rho(\pi) = \rho(\sigma) + \rho(\tau) + |\sigma| + |\tau|$, as pairs with the left coordinate in σ and the right coordinate in τ are definitely not counted.

This definition gives the path length of the binary search tree associated to the permutation $\pi = \pi_1 \dots \pi_n$, provided we interpret π_i as the arrival position of key i.

But this definition of π can be taken as it is, where $\pi_1 \dots \pi_n$ now denotes a *word* over the alphabet $\{1, 2, \dots\}$. This will be our starting point.

We want to point out that our previous paper [4] contains easier but related parameters.

In the sequel we want to compute the expectation and the variance of the parameter ρ, for random words of length n. The probability model is as follows: the probability of letter i is pq^{i-1}, and the letters are independent, so that the word $a_1 \dots a_m$ appears with probability $(\frac{p}{q})^m q^{a_1 + \dots + a_m}$. We define random variables

Date: October 14, 2000.

1991 *Mathematics Subject Classification.* 05A15, 05A30, 68P10.

Key words and phrases. Binary search tree, path length, permutations, geometric distribution, q–analogues, harmonic numbers.

$$
\begin{aligned}
L_{jk} &= \begin{cases} 1 & \text{if } \pi_j = \min\{\pi_j, \dots, \pi_k\} \\ 0 & \text{ otherwise,} \end{cases} \\
R_{jk} &= \begin{cases} 1 & \text{if } \pi_k = \min\{\pi_j, \dots, \pi_k\} \\ 0 & \text{ otherwise,} \end{cases} \\
B_{jk} &= L_{jk} \cdot R_{jk}, \\
N_{jk} &= (1 - L_{jk}) \cdot (1 - R_{jk}).
\end{aligned}
$$

(The letters L, R, B, N are chosen to indicate *left, right, both, not.*)

Then the parameter ρ may be described as

$$\rho = \sum_{1 \le j < k \le n} \Big[L_{jk} + R_{jk} - B_{jk} \Big].$$

Now we can introduce the generating function

$$f(v) = \Big(\frac{p}{q}\Big)^n \sum_{i_1, \dots, i_n \ge 1} q^{i_1 + \dots + i_n} \prod_{1 \le j < k \le n} \Big[L_{jk} v + R_{jk} v - B_{jk} v + N_{jk} \Big];$$

the coefficient of v^k in $f(v)$ is the probability that parameter ρ has value k, assuming random words of length n.

As always, the expected value is obtained via $\mathbb{E} = f'(1)$;

$$
\begin{aligned}
\mathbb{E} &= \Big(\frac{p}{q}\Big)^n \sum_{i_1, \dots, i_n \ge 1} q^{i_1 + \dots + i_n} \sum_{1 \le j < k \le n} \big(L_{jk} + R_{jk} - B_{jk} \big) \\
&= \sum_{1 \le j < k \le n} \Big(\frac{p}{q}\Big)^{k+1-j} \Big[\sum_{L_{jk}=1} q^{i_j + \dots + i_k} + \sum_{R_{jk}=1} q^{i_j + \dots + i_k} - \sum_{B_{jk}=1} q^{i_j + \dots + i_k} \Big] \\
&= \sum_{1 \le j < k \le n} \Big(\frac{p}{q}\Big)^{k+1-j} \Big[2 \sum_{i \ge 1} q^{i(k+1-j)} \frac{1}{p^{k-j}} - \sum_{i \ge 1} q^{i(k+1-j)} \frac{1}{p^{k-j-1}} \Big] \\
&= (2p - p^2) \sum_{1 \le j < k \le n} \Big(\frac{1}{q}\Big)^{k+1-j} \sum_{i \ge 1} q^{i(k+1-j)} \\
&= p(2-p) \sum_{1 \le j < k \le n} \frac{1}{1 - q^{k+1-j}} \\
&= p(2-p) \sum_{2 \le i \le n} \frac{n+1-i}{1 - q^i} \\
&= p(2-p) \sum_{1 \le i \le n} \frac{n+1-i}{1 - q^i} - n(2-p).
\end{aligned}
$$

The terms that would survive the limit $q \to 1$ are

$$2p\sum_{1\le i\le n}(n+1-i)\frac{1}{1-q^i}-2n,$$

and the limit is

$$\lim_{q\to 1}\mathbb{E}=2\sum_{1\le i\le n}\frac{n+1-i}{i}-2n=2(n+1)H_n-4n,$$

as is of course well known. (The notations $H_n=\sum_{1\le k\le n}\frac{1}{k}$ and $H_n^{(2)}=\sum_{1\le k\le n}\frac{1}{k^2}$ refer to harmonic numbers of the first and second order, see e. g. [1].)

Now we turn to the variance, and this is much harder, since we must first compute the second factorial moment, which is obtained by a second derivative;

$$\begin{aligned}
\mathbb{E}^{\underline{2}} &= \Big(\frac{p}{q}\Big)^n\sum_{i_1,\dots,i_n\ge 1}q^{i_1+\dots+i_n}\times\\
&\times\sum_{1\le j<k\le n,1\le l<m\le n,(j,k)\ne(l,m)}\big(L_{jk}+R_{jk}-B_{jk}\big)\big(L_{lm}+R_{lm}-B_{lm}\big)\\
&=\Big(\frac{p}{q}\Big)^n\sum_{i_1,\dots,i_n\ge 1}q^{i_1+\dots+i_n}\times\\
&\times\sum_{1\le j<k\le n,1\le l<m\le n,(j,k)\ne(l,m)}\big(2L_{jk}L_{lm}+2L_{jk}R_{lm}-4L_{jk}B_{lm}+B_{jk}B_{lm}\big)\\
&=2\Xi^{\mathrm{LL}}+2\Xi^{\mathrm{LR}}-4\Xi^{\mathrm{LB}}+\Xi^{\mathrm{BB}}
\end{aligned}$$

(using several symmetries). Observe that we use quantities like L_{jk} now also in a deterministic sense. This slight abuse of notation has been chosen in order not to overburden the paper with heavy notation.

The range $\Lambda=\{1\le j<k\le n,1\le l<m\le n,(j,k)\ne(l,m)\}$ must be split into the following 12 disjoint subranges:

$$\begin{aligned}
\Lambda_1&=\{1\le j<k<l<m\le n\},\\
\Lambda_2&=\{1\le j<l<m<k\le n\},\\
\Lambda_3&=\{1\le j<l<k<m\le n\},\\
\Lambda_4&=\{1\le j<k=l<m\le n\},\\
\Lambda_5&=\{1\le j<l<m=k\le n\},\\
\Lambda_6&=\{1\le j=l<k<m\le n\},\\
\Lambda_7&=\{1\le l<m<j<k\le n\},\\
\Lambda_8&=\{1\le l<j<k<m\le n\},\\
\Lambda_9&=\{1\le l<j<m<k\le n\},\\
\Lambda_{10}&=\{1\le l<m=j<k\le n\},\\
\Lambda_{11}&=\{1\le l<j<k=m\le n\},
\end{aligned}$$

$$\Lambda_{12} = \{1 \le l = j < m < k \le n\}.$$

And we will have contributions Θ_i^{LL}, Θ_i^{LR}, Θ_i^{LB}, Θ_i^{BB}, to Ξ^{LL}, Ξ^{LR}, Ξ^{LB}, Ξ^{BB}, for $i = 1, \dots, 12$, according to the 12 ranges Λ_i.

Therefore we must compute 48 (not necessarily) different contributions.

For convenience, we state them as a lemma.

Lemma 1. *The contributions Θ_i^{LL}, Θ_i^{LR}, Θ_i^{LB}, Θ_i^{BB}, for $i = 1, \dots, 12$, are given by*

$$\begin{aligned}
\Theta_1^{LL} = \Theta_7^{LL} &= p^2 \sum_{1\le j<k<l<m\le n} \frac{1}{1-q^{k+1-j}} \frac{1}{1-q^{m+1-l}},\\
\Theta_2^{LL} = \Theta_8^{LL} &= p^2 \sum_{1\le j<l<m<k\le n} \frac{1}{1-q^{k+1-j}} \frac{1}{1-q^{m+1-l}},\\
\Theta_3^{LL} = \Theta_9^{LL} &= p^2 \sum_{1\le j<l<k<m\le n} \frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{m+1-l}},\\
\Theta_4^{LL} = \Theta_{10}^{LL} &= p^2 \sum_{1\le j<k=l<m\le n} \frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{m+1-k}},\\
\Theta_5^{LL} = \Theta_{11}^{LL} &= p^2 \sum_{1\le j<l<m=k\le n} \frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{m+1-l}},\\
\Theta_6^{LL} = \Theta_{12}^{LL} &= p \sum_{1\le j=l<k<m\le n} \frac{1}{1-q^{m+1-j}};
\end{aligned}$$

$$\begin{aligned}
\Theta_1^{LR} = \Theta_7^{LR} &= p^2 \sum_{1\le j<k<l<m\le n} \frac{1}{1-q^{k+1-j}} \frac{1}{1-q^{m+1-l}},\\
\Theta_2^{LR} = \Theta_8^{LR} &= p^2 \sum_{1\le j<l<m<k\le n} \frac{1}{1-q^{k+1-j}} \frac{1}{1-q^{m+1-l}},\\
\Theta_3^{LR} = \Theta_9^{LR} &= p^2 \sum_{1\le j<l<k<m\le n} \left[\frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{m+1-l}} + \frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{k+1-j}} - \frac{1}{1-q^{m+1-j}}\right],\\
\Theta_4^{LR} &= p^2 \sum_{1\le j<k=l<m\le n} \left[\frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{m+1-k}} + \frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{k+1-j}} - \frac{1}{1-q^{m+1-j}}\right],\\
\Theta_5^{LR} &= p^2 \sum_{1\le j<l<m\le n} \frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{m+1-l}},\\
\Theta_6^{LR} &= p^2 \sum_{1\le j<k<m\le n} \frac{1}{1-q^{m+1-j}} \frac{1}{1-q^{k+1-j}},\\
\Theta_{10}^{LR} &= p \sum_{1\le l<m=j<k\le n} \frac{1}{1-q^{k+1-l}},
\end{aligned}$$

$$\Theta_{11}^{LR}=\Theta_{12}^{LR}=p^2\sum_{1\le l=j<m<k\le n}\frac{1}{1-q^{k+1-l}};$$

$$\Theta_{1}^{LB}=\Theta_{7}^{LB}=p^3\sum_{1\le j<k<l<m\le n}\frac{1}{1-q^{k+1-j}}\frac{1}{1-q^{m+1-l}},$$

$$\Theta_{2}^{LB}=\Theta_{8}^{LB}=p^3\sum_{1\le j<l<m<k\le n}\frac{1}{1-q^{k+1-j}}\frac{1}{1-q^{m+1-l}},$$

$$\Theta_{3}^{LB}=\Theta_{9}^{LB}=p^3\sum_{1\le j<l<k<m\le n}\frac{1}{1-q^{m+1-j}}\frac{1}{1-q^{m+1-l}},$$

$$\Theta_{4}^{LB}=p^3\sum_{1\le j<k=l<m\le n}\frac{1}{1-q^{m+1-j}}\frac{1}{1-q^{m+1-k}},$$

$$\Theta_{5}^{LB}=p^3\sum_{1\le j<l<m\le n}\frac{1}{1-q^{m+1-j}}\frac{1}{1-q^{m+1-l}},$$

$$\Theta_{6}^{LB}=p^2\sum_{1\le j<k<m\le n}\frac{1}{1-q^{m+1-j}},$$

$$\Theta_{10}^{LB}=p^2\sum_{1\le l<m=j<k\le n}\frac{1}{1-q^{k+1-l}},$$

$$\Theta_{11}^{LB}=p^3\sum_{1\le l<j<k=m\le n}\frac{1}{1-q^{k+1-l}},$$

$$\Theta_{12}^{LB}=p^2\sum_{1\le l=j<m<k\le n}\frac{1}{1-q^{k+1-l}};$$

$$\Theta_{1}^{BB}=\Theta_{7}^{BB}=p^4\sum_{1\le j<k<l<m\le n}\frac{1}{1-q^{k+1-j}}\frac{1}{1-q^{m+1-l}},$$

$$\Theta_{2}^{BB}=\Theta_{8}^{BB}=p^4\sum_{1\le j<l<m<k\le n}\frac{1}{1-q^{k+1-j}}\frac{1}{1-q^{m+1-l}},$$

$$\Theta_{3}^{BB}=\Theta_{9}^{BB}=p^4\sum_{1\le j<l<k<m\le n}\frac{1}{1-q^{m+1-j}},$$

$$\Theta_{4}^{BB}=\Theta_{10}^{BB}=p^3\sum_{1\le j<k=l<m\le n}\frac{1}{1-q^{m+1-j}},$$

$$\Theta_{5}^{BB}=\Theta_{11}^{BB}=p^3\sum_{1\le j<l<m\le n}\frac{1}{1-q^{m+1-j}}\frac{1}{1-q^{m+1-l}},$$

$$\Theta_6^{BB} = \Theta_{12}^{BB} = p^3 \sum_{1\le j=l<k<m\le n} \frac{1}{1-q^{m+1-j}}.$$

Proof. The computations are as (or slightly more complicated than) the one for the expected value. We don't give more details. ☐

We simplify those sums and write $a_i = \frac{1}{1-q^i}$ for convenience.

Lemma 2.

$$\begin{aligned}
\Xi^{LL} &= 2p^2 \sum_{2\le i,j\le n-2, i+j\le n} a_i a_j \binom{n+2-i-j}{2} \\
&+ 2p^2 \sum_{2\le i<j\le n} a_i a_j (n+1-j)(j-i-1) \\
&+ 2p^2 \sum_{3\le i<j\le n} a_i a_j (n+1-j)(i-2) \\
&+ 4p^2 \sum_{2\le i<j\le n} a_i a_j (n+1-j) \\
&+ 2p \sum_{3\le i\le n} a_i (i-2)(n+1-i),
\end{aligned}$$

$$\begin{aligned}
\Xi^{LR} &= 2p^2 \sum_{2\le i,j\le n-2, i+j\le n} a_i a_j \binom{n+2-i-j}{2} \\
&+ 2p^2 \sum_{2\le i<j\le n} a_i a_j (n+1-j)(j-i-1) \\
&+ 4p^2 \sum_{3\le i<j\le n} a_i a_j (n+1-j)(i-2) \\
&- 2p^2 \sum_{4\le i\le n} a_i \binom{i-2}{2} (n+1-i) \\
&+ 4p^2 \sum_{2\le i<j\le n} a_i a_j (n+1-j) \\
&+ p^2 \sum_{3\le i\le n} a_i (i-2)(n+1-i) \\
&+ p \sum_{3\le i\le n} a_i (i-2)(n+1-i),
\end{aligned}$$

$$\begin{aligned}
\Xi^{LB} &= 2p^3 \sum_{2\le i,j\le n-2, i+j\le n} a_i a_j \binom{n+2-i-j}{2} \\
&+ 2p^3 \sum_{2\le i<j\le n} a_i a_j (n+1-j)(j-i-1)
\end{aligned}$$

$$
\begin{aligned}
&+2p^3 \sum_{3\le i<j\le n} a_i a_j (n+1-j)(i-2)\\
&+2p^3 \sum_{2\le i<j\le n} a_i a_j (n+1-j)\\
&+(3p^2+p^3) \sum_{3\le i\le n} a_i (i-2)(n+1-i),
\end{aligned}
$$

$$
\begin{aligned}
\Xi^{BB} = 2p^4 &\sum_{2\le i,j\le n-2, i+j\le n} a_i a_j \binom{n+2-i-j}{2}\\
&+2p^4 \sum_{2\le i<j\le n} a_i a_j (n+1-j)(j-i-1)\\
&+2p^3 \sum_{2\le i<j\le n} a_i a_j (n+1-j)\\
&+2p^4 \sum_{4\le i\le n} a_i \binom{i-2}{2}(n+1-i)\\
&+4p^3 \sum_{3\le i\le n} a_i (i-2)(n+1-i).
\end{aligned}
$$

$\square$

The variance is given by

$$\mathbb{V} = 2\Xi^{\mathrm{LL}} + 2\Xi^{\mathrm{LR}} - 4\Xi^{\mathrm{LB}} + \Xi^{\mathrm{BB}} + \mathbb{E} - \big(\mathbb{E}\big)^2.$$

In order to simplify this expression, we note the following formula:

Lemma 3.

$$
\begin{aligned}
p^2 &\sum_{2\le i,j\le n-2, i+j\le n} a_i a_j \binom{n+2-i-j}{2}\\
&= 2p^2 \sum_{1\le i<j\le n} a_i a_j \binom{n+2-j}{2} - p^2 \sum_{1\le i\le n} a_i \binom{n+2-i}{2}(i-1)\\
&\quad - 2p \sum_{1\le i\le n} a_i \binom{n+1-i}{2} + \binom{n}{2}.
\end{aligned}
$$

Proof. Note that

$$\frac{1}{1-q^i}\frac{1}{1-q^j} = \frac{1}{1-q^{i+j}}\left(\frac{1}{1-q^i} + \frac{1}{1-q^j} - 1\right)$$

and do some trivial rearrangements. $\square$

Lemma 4.

$$\Big(\sum_{1\le i\le n} a_i(n+1-i)\Big)^2 = 2\sum_{1\le i<j\le n} a_i a_j (n+1-i)(n+1-j) + \sum_{1\le i\le n} a_i^2 (n+1-i)^2.$$

Proof. Obvious. ∎

Using these lemmata and numerous simplifications that were partially supported by Maple, we can state our main result:

Theorem 1. *The expectation and the variance of the q–ified path length in words of length n, generated by n independent geometric random variables are given by*

$$\mathbb{E} = p(2-p)\sum_{1\le i\le n}\frac{n+1-i}{1-q^i} - n(2-p)$$

and

$$\begin{aligned}\mathbb{V} &= 2p^2\sum_{1\le i<j\le n}\frac{(n+1-j)(4i+p(5-4i))}{(1-q^i)(1-q^j)}\\ &- p^2(2-p)^2\sum_{1\le i\le n}\frac{(n+1-i)^2}{(1-q^i)^2}\\ &+ p\sum_{1\le i\le n}\frac{n+1-i}{1-q^i}\Big(6i-2+p(-4ni+4n-19i+3i^2+7)\\ &\qquad + 4p^2(ni-n+3i-1-i^2)+p^3(-ni+n+2i^2-8i+8)\Big)\\ &+ 5pn-3p^2n-2p^3n.\end{aligned}$$

∎

The terms in the variance that would survive the limit $q\to 1$ are these:

$$8p^2\sum_{1\le i<j\le n}\frac{(n+1-j)i}{(1-q^i)(1-q^j)} - 4p^2\sum_{1\le i\le n}\frac{(n+1-i)^2}{(1-q^i)^2} + 2p\sum_{1\le i\le n}\frac{(n+1-i)(3i-1)}{1-q^i}.$$

The limit is

$$\begin{aligned}\lim_{q\to1}\mathbb{V} &= 8\sum_{1\le j\le n}\frac{(n+1-j)(j-1)}{j} - 4\sum_{1\le i\le n}\frac{(n+1-i)^2}{i^2} + 2\sum_{1\le i\le n}\frac{n+1-i}{i}(3i-1)\\ &= 4(n+1)n - 8(n+1)H_n + 8n - 4(n+1)^2H_n^{(2)} + 8(n+1)H_n - 4n\\ &\qquad + 3n(n+1) - 2(n+1)H_n + 2n\\ &= 7n^2 - 4(n+1)^2H_n^{(2)} - 2(n+1)H_n + 13n.\end{aligned}$$

This is (of course!) the variance in the classical case.

Remark. As one referee has pointed out and as was clear to me from the beginning, one could study the alternative version

$$\rho(\pi) = \big|\{(j,k)\mid 1\le j<k\le n,\ \pi_j \text{ or } \pi_k \text{ is a strict minimum of } \{\pi_j,\dots,\pi_k\}\big|.$$

The expected value is then

$$\frac{2p}{q}\sum_{1\le j<k\le n}\sum_{i\ge 1}q^{i(k-j+1)}=2p\sum_{1\le i\le n}(n+1-i)\frac{q^i}{1-q^i}-2qn,$$

which in the limit $q\to 1$ gives the result $2(n+1)H_n-4n$, as it should. The computation of the variance would be a bit easier, since terms like B_{jk} would disappear. Nevertheless, a new computation would be complicated enough, perhaps requiring a full day's work, and so I leave it as an exercise to the interested.

References

[1] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics (Second Edition).* Addison Wesley, 1994.

[2] D. E. Knuth. *The Art of Computer Programming*, volume 3: Sorting and Searching. Addison-Wesley, 1973. Second edition, 1998.

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

[4] H. Prodinger. Combinatorics of geometrically distributed random variables: Inversions and a parameter of Knuth. *submitted*, 2000.

[5] R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithms.* Addison-Wesley, 1996.

Helmut Prodinger, The John Knopfmacher Centre for Applicable Analysis and Number Theory, Department of Mathematics, University of the Witwatersrand, P. O. Wits, 2050 Johannesburg, South Africa, email: helmut@gauss.cam.wits.ac.za.
Homepage: http://www.wits.ac.za/helmut/index.htm