A note on the Horton-Strahler number for random trees

Luc Devroye Paul Kruszewski *
School of Computer Science, McGill University,
3840 University Street, Montreal, Canada H3A 2A7

July 8, 1994

Abstract
We consider the HORTON-STRAHLER number S_n for random equiprobable binary trees with n nodes. We give a simple probabilistic proof of the well-known result that $\mathbf{E}S_n = \log_4 n + O(1)$ and show that for every $x > 0$,

$$\mathbf{P}\{|S_n - \log_4 n| \geq x\} \leq \frac{D}{4^x} ,$$

for some constant $D > 0$.

Keywords: analysis of algorithms; probabilistic analysis; HORTON-STRAHLER number; random binary trees

Introduction

Originally used to classify river systems [4, 12], the Horton-Strahler number has also been applied to binary trees. Let T be a binary tree with n nodes such that each node has at most one left and one right node. For example, with $n = 3$ there are exactly five different trees. Let $|T|$ be the number of nodes in T. Similarly, let $|u|$ be the number of nodes in the subtree rooted at node u in T. For a node u in the binary tree T, let the Horton-Strahler number $S(u)$ be defined as

$$S(u) = \begin{cases} 0 & \text{if } |u| = 0 , \\ \max(S(v), S(w)) + I_{[S(v)=S(w)]} & \text{if } |u| \geq 1 \text{ and} \\ & u \text{ has children } v \text{ and } w , \end{cases}$$

where I_A is the indicator of the event A. We define $S(T)$ as the Horton-Strahler number of the root of tree T. For example, Figure 1 shows a tree with Horton-Strahler number three. At times, we use $S(u)$ and $S(T)$ interchangeably, even though u is a node and T is a tree.

The two extreme values for the Horton-Strahler number are immediately apparent. At the one extreme is a single chain of n nodes and Horton-Strahler number one (see Figure 2).

*Research of the authors was supported by NSERC Grant A3456, and by a 1967 NSERC Scholarship.

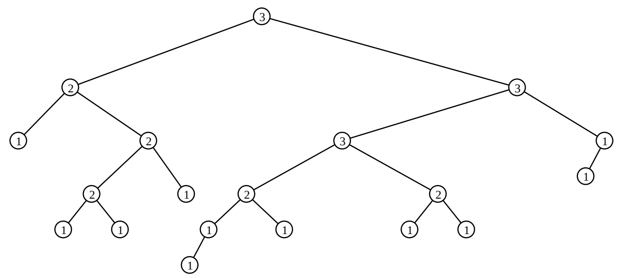

Figure 1: A binary tree with Horton-Strahler number three.

This is sometimes called a *"gourmand de la vigne"* by Viennot [15], because when viewed with its external nodes ($\square$), the chain resembles the bottom part of a vine which is cut to improve the quality and quantity of the wine.

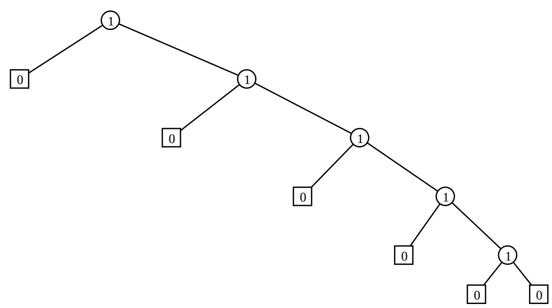

Figure 2: A *"gourmand de la vigne"* with five nodes and Horton-Strahler number one.

At the other extreme is the complete tree with k levels, $2^k - 1$ nodes and Horton-Strahler number k (see Figure 3). Generalizing this, it is clear that, for each binary tree T with n nodes, $S(T) \leq \log_2 n + 1$ [6].

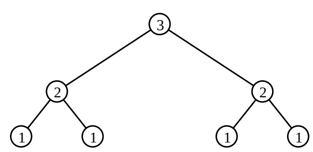

Figure 3: A complete tree with seven nodes and Horton-Strahler number three.

The Horton-Strahler number arises in computer science because of its relationship to expression evaluation. Often in a computer, an arithmetic expression is evaluated by micro-operations using registers. To facilitate this process, the expression is stored as an expression tree with the operators in the internal nodes and the operands in the external nodes. The arithmetic expression is evaluated by traversing the corresponding tree. In 1958, Ershov [1] showed that by always traversing the child node with the lower Horton-Strahler number first, that the corresponding register use is minimal. Furthermore, the minimum number of registers required to evaluate an expression tree T is exactly $S(T) + 1$. As expression evaluation is a specialized type of postorder traversal, this can be generalized that the minimum stack size required for a postorder traversal of binary tree T is $S(T)+1$ [3]. In fact, the Horton-Strahler number occurs in almost every field involving some kind of natural branching pattern. More recently, the Horton-Strahler number has been used to draw trees [6, 16]. Viennot [15] provides a thorough overview. See also Vauchaussade de Chaumont and Viennot [13, 14], and Viennot, Eyrolles, Janey, and Argues [16].

The Horton-Strahler number for equiprobable binary trees

Let an equiprobable binary tree (EBT) with n nodes be a binary tree with n nodes drawn uniformly and at random from all possible binary trees with n nodes. Let S_n be the Horton-Strahler number of a random EBT with n nodes so that $\mathbf{E}S_n$ and $\mathbf{Var}S_n$ are the corresponding expected value and variance.

The result is well-known. Under the assumption that the corresponding expression trees with n internal nodes and $n+1$ external nodes are equiprobable, the expected minimum number of registers needed to evaluate an arithmetic expression with n operators is $\mathbf{E}S_n+1$.

Based on exact computations of $\mathbf{E}S_n$ up to $n = 100$, Shreve [11] conjectured that $\mathbf{E}S_n \sim \log_4 n$. Flajolet, Raoult and Vuillemin [2], Kemp [5], and Meir, Moon and Pounder [7, 8, 9] independently analysed S_n via recurrences and generating functions. Flajolet, Raoult and Vuillemin [2] showed that

$$\mathbf{E}S_n = \log_4 n + D(\log_4 n) + o(1)$$

where $|D(x)| \leq 1$ for $x > 0$. Kemp [5] showed that for all $\varepsilon > 0$,

$$\mathbf{E}S_n = \log_4 n + C + F(n) + O(n^{-0.5+\varepsilon})$$

where $C = 0.82574\ldots$ is a constant and $F(n)$ is a function with $F(n) = F(4n)$ for all $n > 0$ and $-0.574 < F(n) < -0.492$. Meir, Moon and Pounder [8] showed that S_n is very highly concentrated about $\log_4 n$. In fact, for any $s > 0$,

$$\mathbf{E}\left|S_n - \log_4 n\right|^s = O(1) \ .$$

The latter result implies that

$$\mathbf{E}S_n \sim \log_4 n \quad \text{and} \quad \mathbf{Var}S_n = O(1) \ .$$

A Probabilistic Analysis

Almost everything with respect to the Horton-Strahler number for EBTs is known. Furthermore by Chebyshev's inequality, the Meir, Moon and Pounder result [8] implies that if a_n is a sequence tending to infinity, then

$$\mathbf{P}\{|S_n - \log_4 n| > a_n\} \to 0 \ ,$$

as $n \to \infty$. Using probabilistic analysis, we present a stronger result.

Let T be a binary tree with n nodes. Let r be the reduction function from binary trees to binary trees defined recursively as

$$
\left\{
\begin{array}{rcll}
r\left(\square\right) & = & () & (1) \\[6pt]
r\left(\begin{array}{c}\circ\\ /\ \backslash\\ \square\ \ \square\end{array}\right) & = & \square & (2) \\[10pt]
r\left(\begin{array}{c}\circ\\ /\ \backslash\\ \square\ \ T\end{array}\right) = r\left(\begin{array}{c}\circ\\ /\ \backslash\\ T\ \ \square\end{array}\right) & = & r\left(T\right) & (3) \\[10pt]
r\left(\begin{array}{c}\circ\\ /\ \backslash\\ T_L\ \ T_R\end{array}\right) & = & \begin{array}{c}\circ\\ /\ \backslash\\ r\left(T_L\right)\ \ r\left(T_R\right)\end{array} & (4)
\end{array}
\right.
$$

where $()$ is the empty tree, $\square$ is an external node, $\circ$ is an internal node, and T, T_L and T_R are binary trees with at least one internal node each.

We note that

$$S(T) = S(r(T)) + 1 \ .$$

We will show that each reduction reduces the size of the tree by a factor of about four and increases the Horton-Strahler number by one. This observation explains why $\mathbf{E}S_n$ is close to $\log_4 n$.

Let $T' = r(T)$. The number of external nodes in T' is equal to $l(T)$, the number of leaves in T. The number of (internal) nodes in T' is equal to the number of external nodes in T' minus one. Thus, $|T'| = l(T) - 1$. We note the following fact for reductions on EBTs.

Fact 1. *If each binary tree T with n nodes is equally likely, then given $|T'| = k < n$, each tree T' is equally likely.*

Proof. For any tree T', we examine the "expansion" of T' back to T so that $|T| = n$ and $r(T) = T'$. The internal nodes of T' result from Case 4 of r. The external nodes of T' result from Case 2. Therefore, in any "expansion" each external node in T' must expand to a parent node of two external nodes (i.e. $\square \to \begin{array}{c}\circ\\ /\ \backslash\\ \square\ \ \square\end{array}$). The remaining $n - (k + k + 1)$ internal nodes of T result from Case 3. These pairs of single-parents with only-children (external nodes) $\left(\begin{array}{c}\circ\\ /\ \backslash\\ \square\ \ \ \ \end{array} \quad \text{or} \quad \begin{array}{c}\circ\\ /\ \backslash\\ \ \ \ \ \square\end{array}\right)$ can be re-inserted anywhere in the expansion except below the leaves of T. Each combination of insertions results in a different tree T. As this argument is identical for all T', we note that for all T' with k nodes there is an equal number of expansions to trees with n nodes. $\blacksquare$

Before we use reductions to derive the upper and lower converging bounds for $\mathbf{E}S_n$, we need the mean and variance of L_n, the number of leaves in a random binary tree. By mimicking the argument in [10] for the average internal path length of a random equiprobable binary tree, we set up the following double generating function

$$Q(w, z) = \sum_{n \geq 0} \sum_{k \geq 0} Q_{nk} w^n z^k \ ,$$

where Q_{nk} is the number of trees with n nodes and k leaves. This, in turn, may be expressed equivalently as

$$Q(w,z) = \sum_{\text{all trees } T} w^{|T|} z^{l(z)} = \frac{1-\sqrt{1-4w(wz-w+1)}}{2w} .$$

From this, it is straightforward to derive $\mathbf{E}\, L_n = \frac{n(n+1)}{2(2n-1)} \sim \frac{n}{4}$ and $\mathbf{Var}\, L_n = \frac{n(n+1)(n^2-3n+2)}{2(2n-1)^2(2n-3)} \leq \frac{n}{8}$, for $n \geq 3$ [6, 17].

We now can start with the upper bound.

Theorem 1. *For a random* EBT *with n nodes and for every $x > 0$,*

$$\mathbf{P}\{S_n > \lceil \log_4 n + x \rceil\} \leq \frac{1}{4^x} .$$

Proof. Let T_0 be a random EBT with n nodes. Let $T_1 = r(T_0)$, let $T_2 = r(T_1)$, et cetera. Then,

$$\begin{aligned} \mathbf{E}\,|T_{k+1}| &= \mathbf{E}\left\{(l(T_k)-1)I_{[|T_k|\geq 1]}\right\} \\ &= \mathbf{E}\left\{\left(\frac{|T_k|(|T_k|+1)}{2(2|T_k|-1)}-1\right)I_{[|T_k|\geq 1]}\right\} \qquad \text{(by [6])} \\ &\leq \mathbf{E}\left\{\frac{|T_k|}{4}\right\} . \end{aligned}$$

Therefore by this inequality and Fact 1, $\mathbf{E}\,\{|T_k|\} \leq \frac{\mathbf{E}|T_0|}{4^k} = \frac{n}{4^k}$. So by Markov's inequality, $\mathbf{P}\{|T_k| \geq 1\} \leq \mathbf{E}|T_k| \leq \frac{n}{4^k}$. Thus since $[S_n - k > 0] = [|T_k| > 0]$, we have $\mathbf{P}\{S_n > k\} = \mathbf{P}\{|T_k| \geq 1\}$. Consequently, if $k = \lceil \log_4 n + x \rceil$ then $\mathbf{P}\{S_n > k\} \leq \frac{n}{4^k} \leq \frac{1}{4^x}$. ■

Theorem 2. *For a random* EBT *with n nodes and for every $x \geq 1$,*

$$\mathbf{P}\{S_n < \lfloor \log_4 n - x \rfloor\} \leq \frac{C}{4^x} ,$$

where $C > 0$ is a suitable constant.

Proof. Let $T_0, T_1, T_2, \ldots$ be a sequence of random binary trees obtained by successive reductions and $|T_0| = n$. Then by Fact 1 and the bound on the variance of the leaves, $\mathbf{Var}\left\{|T_{k+1}|\,\Big|\, T_k\right\} \leq c|T_k|$, where $c = 1/8$. Also, $\mathbf{E}\left\{|T_{k+1}|\,\Big|\,T_k\right\} \leq \frac{|T_k|}{4}$. Therefore,

$$\mathbf{Var}|T_{k+1}| \leq \mathbf{E}\left\{c|T_k|\right\} + \mathbf{Var}\left\{\frac{|T_k|}{4}\right\} \leq \frac{cn}{4^k} + \frac{1}{16}\mathbf{Var}|T_k| .$$

Iterating the preceding inequality, we have

$$\mathbf{Var}|T_{k+1}| \leq \frac{cn}{4^k} + \frac{1}{16}\left(\frac{cn}{4^{k-1}} + \frac{1}{16}\mathbf{Var}|T_{k-1}|\right)$$

$$
\begin{aligned}
&= \frac{cn}{4^k}(1+\frac{1}{4}) + \frac{1}{16^2}\mathbf{Var}|T_{k-1}| \\
&\ \vdots \\
&\le \frac{cn}{4^k}(1+\frac{1}{4}+\frac{1}{4^2}+\cdots+\frac{1}{4^k}) + \frac{1}{16^{k+1}}\mathbf{Var}|T_0| \\
&= \frac{4}{3}\cdot\frac{cn}{4^k} \qquad\qquad \text{(since } \mathbf{Var}|T_0| = 0\text{)} .
\end{aligned}
$$

We note by inspection $\mathbf{E}|T_{k+1}| \ge \frac{\mathbf{E}|T_k|}{4} - 1$. Iterating this, we obtain

$$
\mathbf{E}|T_{k+1}| \ge \frac{\mathbf{E}|T_0|}{4^{k+1}} - 1 - \frac{1}{4} - \cdots - \frac{1}{4^k} \ge \frac{n}{4^{k+1}} - \sum_{j=0}^{\infty}\frac{1}{4^j} = \frac{n}{4^{k+1}} - \frac{4}{3} .
$$

We have

$$
\begin{aligned}
\mathbf{P}\{S_n \le k\} &= \mathbf{P}\{|T_k| = 0\} \\
&= \mathbf{P}\{|T_k| - \mathbf{E}|T_k| \le -\mathbf{E}|T_k|\} \\
&\le \frac{\mathbf{Var}|T_k|}{\mathbf{E}^2|T_k|} \quad \text{(by Chebyshev's inequality)} \\
&\le \frac{4}{3}\cdot\frac{cn}{4^k}\cdot\frac{1}{(\frac{n}{4^k}-\frac{4}{3})^2} .
\end{aligned}
$$

If $k = \lfloor \log_4 n - x \rfloor$ then $\mathbf{P}\{S_n \le k\} \le \frac{c4^{x+1}}{(4^{x-1}-\frac{4}{3})^2} \le \frac{8c}{9\cdot 4^x} = \frac{1}{9\cdot 4^x}$ when $x \ge 2$. ∎

We combine the upper and lower bounds.

Theorem 3. *For a random* EBT *with n nodes and for every $x > 0$*

$$
\mathbf{P}\{|S_n - \log_4 n| \ge x\} \le \frac{D}{4^x} ,
$$

for some constant $D > 0$.

Proof. This follows directly from Theorems 1 and 2 . ∎

From this theorem, we have the following corollaries.

Corollary 1. *For a random* EBT *with n nodes and for all $s > 0$,*

$$
\mathbf{E}\{|S_n - \log_4 n|^s\} = O(1) .
$$

Corollary 2. *For a random* EBT *with n nodes and for all $\lambda \in (0, \log 4)$,*

$$
\mathbf{E}\left\{e^{\lambda|S_n - \log_4 n|}\right\} < \infty .
$$

Furthermore, Corollary 1 implies that $\mathbf{Var}S_n = O(1)$. In conclusion, we remark that the results from Theorems 1, 2 and 3, and Corollaries 1 and 2 are all non-asymptotic in nature. That is, the results hold for *all* n. Finally, we see that while the trivial upper bound $S_n \leq \log_2 n + 1$ assumed that *every* node in the tree successfully contributed to the Horton-Strahler number, Theorem 3 implies that in EBTs only approximately half the nodes actually do contribute.

Generalizations of the present results to suitably defined m-ary Horton-Strahler numbers for random m-ary trees would be interesting. For tree-drawing purposes, it would also be of interest to introduce new classes of random trees indexed by a real number $c \in (0, 1]$, such that $\mathbf{E}S_n \sim c \log_2 n$. The EBT just corresponds to $c = 1/2$.

Acknowledgements

We thank the referees for their useful comments.

References

[1] A. P. Ershov. On programming of arithmetic operations. *Communications of the ACM 1* (1958), 3–6.

[2] P. Flajolet, J. C. Raoult, and J. Vuillemin. The number of registers required for evaluating arithmetic expressions. *Theoretical Computer Science 9* (1979), 99–125.

[3] J. Françon. Sur le nombre de registres nécessaires à l'évaluation d'une expression arithmétique. *Theoretical Informatics 18* (1984), 355–364.

[4] R. E. Horton. Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Bulletin of the Geological Society of America 56* (1945), 275–370.

[5] R. Kemp. The average number of registers needed to evaluate a binary tree optimally. *Acta Informatica 11* (1979), 363–372.

[6] P. Kruszewski. A probabilistic exploration of the Horton-Strahler number for random binary trees. Master's thesis, School of Computer Science, McGill University, 1993.

[7] A. Meir and J. W. Moon. Stream lengths in random channel networks. *Congressus Numerantium 33* (1980), 25–33.

[8] A. Meir, J. W. Moon, and J. R. Pounder. On the order of random channel networks. *SIAM Journal of Algebraic and Discrete Methods 1* (1980), 25–33.

[9] J. W. Moon. On Horton's law for random channel networks. *Annals of Discrete Mathematics 8* (1980), 117–121.

[10] R. Sedgewick. Part III: Mathematical analysis of combinatorial algorithms. In *Probability Theory and Computer Science*, G. Louchard and G. Latouche, Eds. Academic Press, London, 1983, pp. 145–149.

[11] R. L. Shreve. Statistical law of stream numbers. *Bulletin of the Geological Society of America 74* (1966), 17–37.

[12] A. N. Strahler. Hypsometric (area-altitude) analysis of erosional topology. *Bulletin of the Geological Society of America 63* (1952), 1117–1142.

[13] M. Vauchaussade de Chaumont. *Nombre de Strahler des arbres, langages algébriques et dénombrement des structures secondaires en biologie moléculaire Thèse 3ème cycle.* PhD thesis, Université de Bordeaux 1, 1985.

[14] M. Vauchaussade de Chaumont and X. G. Viennot. Enumeration of RNAs secondary structures by complexity. *Mathematics in Medicine and Biology, Lecture Notes in Biomathematics 57* (1985), 360–365.

[15] X. G. Viennot. Trees everywhere. In *Proceedings of the 15th Colloquium on Trees in Algebra and Programming, Copenhagen, Denmark, May 15-18, 1990, Lecture Notes in Computer Science* (Berlin, 1980), A. Arnold, Ed., vol. 431, Springer-Verlag, pp. 18–41.

[16] X. G. Viennot, G. Eyrolles, N. Janey, and D. Argues. Combinatorial analysis of ramified patterns and computer imagery of trees. In *Proceedings of SIGGRAPH'89, Computer Graphics* (1989), vol. 23, pp. 31–40.

[17] X. Wang and E. Waymire. Central limit theorems for horton ratios. Technical report, Department of Mathematics, Oregon State University, Corvallis, OR 97331, 1989.