Journal de Théorie des Nombres
de Bordeaux **14** (2002), 415–424

Uniform distribution modulo one and binary search trees

par Michel DEKKING et Peter van der WAL

On the occasion of the 65th birthday of Michel Mendès France

Résumé. On peut construire à partir d'une suite $x = (x_k)_{k=1}^{\infty}$ de nombres distincts de l'intervalle $[0,1]$ un arbre binaire en plaçant successivement ces nombres sur les nœuds selon un algorithme "gauche-droite" (cela revient à classer les nombres selon l'algorithme Quicksort). On note $H_n(x)$ la hauteur de l'arbre obtenu à partir des nombres $x_1, \dots, x_n$. Il est évident que

$$\frac{\log n}{\log 2} - 1 \leq H_n(x) \leq n - 1.$$

Si la suite x est obtenue comme valeurs de variables aléatoires indépendantes uniformes sur $[0,1]$, alors on sait qu'il existe $c > 0$ tel que $H_n(x) \sim c \log n$, $(n \to \infty)$, presque-sûrement Récemment, Devroye et Goudjil ont montré que si les x sont les suites de Weyl, i.e., $x_k = \{\alpha k\}$, $k = 1, 2, \dots,$ où α est une variable aléatoire uniforme sur $[0,1]$, alors

$$H_n(x) \sim (12/\pi^2) \log n \log\log n, \quad n \to \infty,$$

en probabilité.
Dans ce papier nous considérons la classe de *toutes* les suites x uniformément réparties pour lesquelles nous montrons que l'on a nécessairement $H_n(x) = o(n)$ quand $n \to \infty$.

Abstract. Any sequence $x = (x_k)_{k=1}^{\infty}$ of distinct numbers from $[0,1]$ generates a binary tree by storing the numbers consecutively at the nodes according to a left-right algorithm (or equivalently by sorting the numbers according to the Quicksort algorithm).
Let $H_n(x)$ be the height of the tree generated by $x_1, \dots, x_n$. Obviously

$$\frac{\log n}{\log 2} - 1 \leq H_n(x) \leq n - 1.$$

If the sequences x are generated by independent random variables having the uniform distribution on $[0,1]$, then it is well known that there exists $c > 0$ such that $H_n(x) \sim c \log n$ as $n \to \infty$ for almost all sequences x. Recently Devroye and Goudjil have shown that

Manuscrit reçu le 27 décembre 2000.

if the sequences x are Weyl sequences, i.e., defined by $x_k = \{\alpha k\}$, $k = 1, 2, \dots$, and α is distributed uniformly at random on $[0, 1]$ then

$$H_n(x) \sim (12/\pi^2) \log n \log\log n$$

as $n \to \infty$ in probability.
In this paper we consider the class of *all* uniformly distributed sequences x, and we show that the only behaviour that is excluded by the equidistribution of x is that of the worst case, i.e., for these x we necessarily have $H_n(x) = o(n)$ as $n \to \infty$.

1. Introduction

The starting point for the present work is a paper published in 1981, written by Michel Mendès France and the first author titled "Uniform distribution modulo one: a geometric viewpoint" ([3]). The central idea of that paper is to associate to any sequence (x_k) of real numbers a curve $\Gamma(x)$ in the (complex) plane by putting $\Gamma(x) = \gamma([0, \infty))$, where

$$\gamma(0) = 0,\ \gamma(n) = \sum_{k=1}^{n} \exp(2\pi i x_k),$$

and $\gamma(t)$ is linear for $n \leq t \leq n+1$. The curve $\Gamma(x)$ gives finer information on the sequence x than if one considers x as a subset $\{x_1, x_2, \dots\}$ of $\mathbb{R}$. See e.g. Figure 1 which displays a part of $\Gamma((\pi k^2))$.

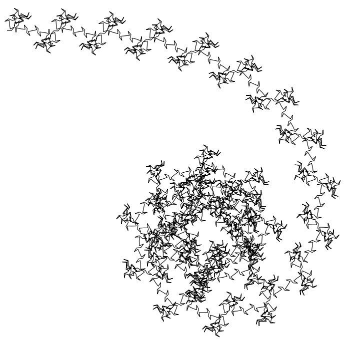

FIGURE 1. $\Gamma((\pi k^2))$.

As noted in [3] the subpatterns of Γ which appear on the spiral consist of 113 segments, the numerator of one of the continued fraction convergents of π. This has been further explained in [4] and [1].
The first author has been involved the last few years in the study of trees generated by algorithms (see e.g. [2]). Two important algorithms, Quicksort ([7]) and the binary search tree algorithm ([8]) generate essentially the same

trees. In this paper we shall consider the question what one can say about the shape of trees generated by the binary search tree algorithm when the input sequence x is equidistributed, in other words we vary on the title above: uniform distribution modulo one - a viewpoint from a tree.

2. Binary search trees

A binary search tree is constructed from a sequence of distinct real numbers, called the *keys*, as follows. The first key is placed at the top of the tree, called the *root*. The next key is directed to the left subtree if it is smaller than the root key, and to the right subtree if it is larger. In general, keys enter the left or right subtree through the root node and then the process of turning left and right is repeated until the end of a branch is reached. We illustrate this with an example.

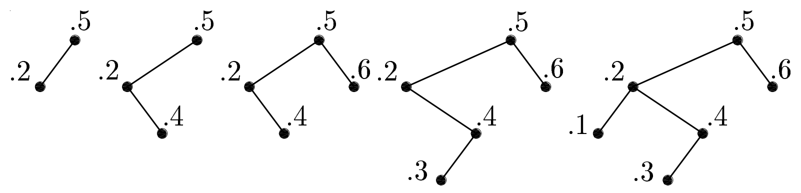

FIGURE 2. Building the tree from the sequence .5, .2, .4, .6, .3, .1.

Consider the sequence of keys $x_1 = .5, x_2 = .2, x_3 = .4, x_4 = .6, x_5 = .3, x_6 = .1$: the first key is .5, and it will be placed at the root of the tree. The second key, .2, goes to the left subtree because $.2 < .5$. The third key, .4, first goes to the left subtree, because $.4 < .5$. After this it gets compared to .2, and because $.4 > .2$, goes to the right. This process continues until all keys are placed in the tree, and we have reached the final tree in Figure 2. We shall denote $\mathcal{T}(x)$ the tree generated by a sequence x with elements from $[0,1]$, and $\mathcal{T}_n(x)$ the tree generated by the first n elements $x_1, \ldots, x_n$. Since the nodes at each level double in number, the ordinary representation of $\mathcal{T}_n(x)$ will quickly get messy. We therefore choose a special embedding of the binary tree in the (complex) plane, where the branches are scaled by a factor $\sqrt{2}$ at each level. More precisely, if we code a level n node ν by the vector $(d_1, \ldots, d_n)$, where $d_k = 1$ when the level k node on the unique path from ν to the root is a right son of his father, and $d_k = -1$ if it is a left son, then the embedding is given by

$$(d_1, \ldots, d_n) \mapsto \sum_{k=1}^{n} \prod_{j=1}^{k} d_j \left(-i\sqrt{2}\right)^{-k+1}$$

(cf. Figure 3).

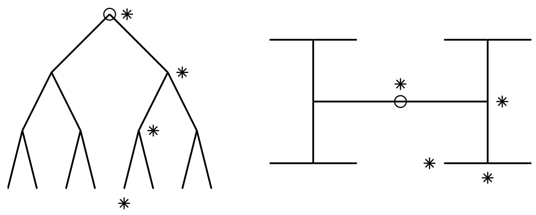

FIGURE 3. Embedding a binary search tree in the plane.

When we apply this algorithm to the first 1500 elements of the sequence $x = (\{\pi k^2\})$ (here $\{y\}$ denotes the fractional part of a number y), and use this embedding we obtain Figure 4.

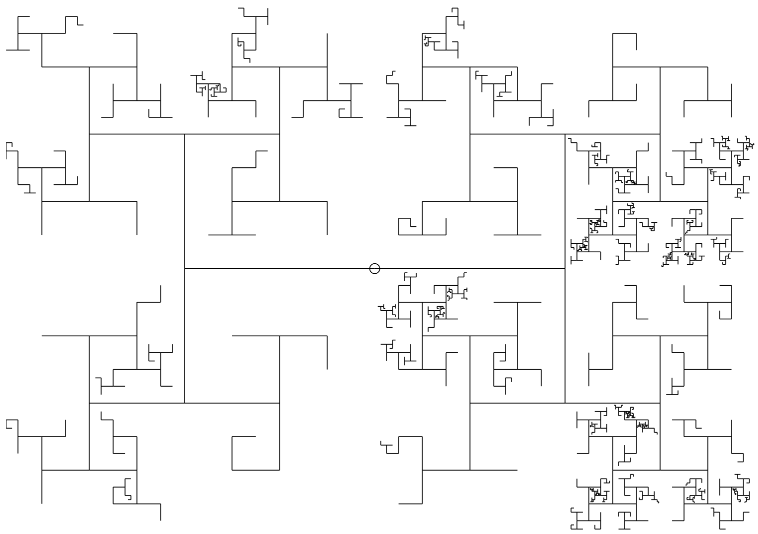

FIGURE 4. $\mathcal{T}_{1500}(\{\pi k^2\})$.

Apparently we do not see now the structure imposed by the continued fraction expansion of π as in Figure 1. However, we *will* see this structure in $\mathcal{T}(\{\pi k\})$ (see Figure 5), and this will be explained in the next section.

An important characteristic of a binary search tree is the *height* $H_n(x)$ of $\mathcal{T}_n(x)$. Here the height is defined as the largest level that is occupied by a key (the level of the root is 0). Obviously we have for all x and all n

$$\frac{\log n}{\log 2} - 1 \leq H_n(x) \leq n - 1.$$

It is customary for the analysis of the algorithm to assume that the keys are generated by independent random variables having the uniform distribution on $[0, 1]$. For this so called *random permutation model* almost all sequences behave essentially in the same way:

Theorem 1 (cf. [5, 9]). *Under the random permutation model* $H_n(x) \sim c \log n$ *as* $n \to \infty$ *almost surely for some* $c > 0$ $(c = 4.31107\cdots)$.

Now typically, in fact almost surely, the sequences x generated by the random permutation model are uniformly distributed. A natural question is therefore what can be said about $H_n(x)$ for an arbitrary equidistributed sequence x.

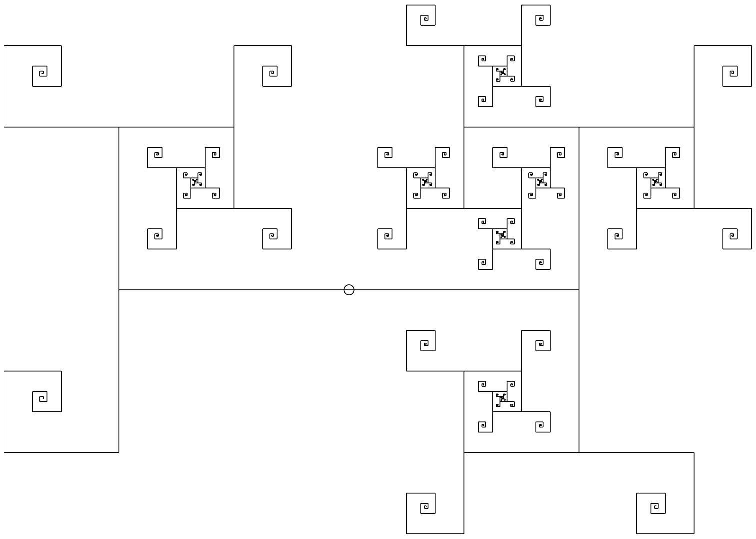

FIGURE 5. $\mathcal{T}_{1500}(\{\pi k\})$.

3. Binary search trees generated by Weyl sequences

To answer the previous question we can profit from a nice analysis performed by Devroye and Goudjil ([6]) of a subclass of equidistributed sequences the Weyl sequences

$$x_\alpha = (x_k) = (\{\alpha k\})_{k=1}^{\infty}.$$

They are able to give a precise description of the height of the tree generated by such sequences.

Theorem 2 ([6]). *Let α be an irrational number in $[0,1]$ with continued fraction expansion $\alpha = [a_1, a_2, \dots]$, and convergents p_n/q_n. Then*

$$H_{q_n-1}(x_\alpha) = \sum_{k=1}^{n} a_k - 2$$

and

$$q_n \leq m < q_{n+1} \Rightarrow \sum_{k=1}^{n} a_k \leq H_m(x_\alpha) + 2 \leq \sum_{k=1}^{n+1} a_k. \tag{1}$$

Devroye and Goudjil deduce from this that for most (i.e., almost all if α is chosen according to the uniform distribution on $[0,1]$) sequences the height behaves as

$$H_n(x_\alpha) \sim (12/\pi^2) \log n \log\log n$$

where the convergence is in probability as $n \to \infty$.
They furthermore deduce the following result on "large height" behaviour.

Theorem 3 ([6]). *Let (h_n) be a monotone sequence of real numbers decreasing from 1 to 0 at any slow rate. Then there exists α such that*

$$H_n(x_\alpha) \geq n h_n \text{ for infinitely many } n. \tag{2}$$

Clearly (1) above implies that "small height" behaviour is obtained for $\alpha = \tau := \frac{1}{2}(\sqrt{5} - 1)$. We then have

$$H_n(x_\tau) \sim \frac{\log n}{\log(1+\tau)}.$$

See Figure 6 for the embedding of this tree.

Our goal will be to show that if one takes arbitrary equidistributed sequences x, one can not do better than in (2), i.e., $H_n(x) = o(n)$, but it is possible to fill the gap between $\frac{\log n}{\log 2}$ and $\frac{\log n}{\log(1+\tau)}$. This will be done in the next two sections.

4. The height of trees generated by uniformly distributed sequences

We first show that denseness of a sequence x already forces some regularity on $\mathcal{T}(x)$.

Proposition 4. *Let $(x_k)_{k\geq 1}$ be a dense sequence of distinct numbers in $[0,1]$. Then for each $\varepsilon > 0$ there exists N such that all level N nodes in $\mathcal{T}(x)$ are occupied, and such that any two neighbours ν and $\tilde{\nu}$ on level N have keys x_k and $x_{\tilde{k}}$ which differ less than ε.*

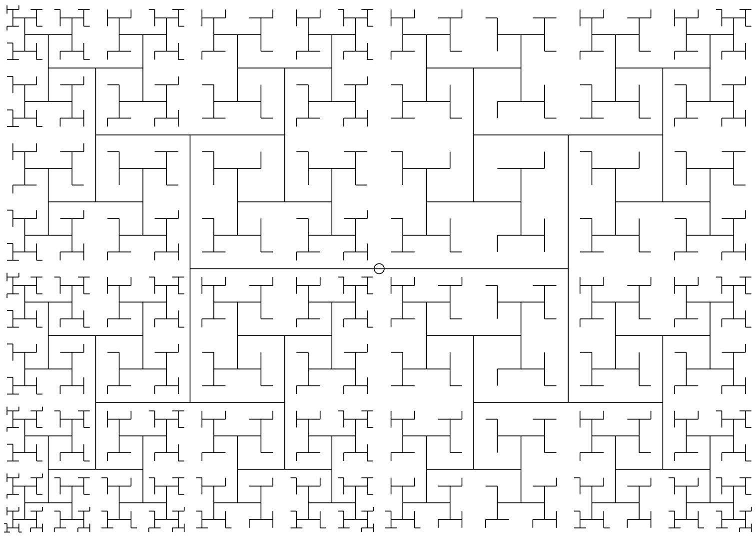

FIGURE 6. $\mathcal{T}_{1500}(\{\frac{1}{2}(1+\sqrt{5})k\})$.

Proof. Fix an $\varepsilon > 0$. By the denseness of (x_k), between any two x_n and x_m an x_k will appear for some $k \geq n, m$ (and numbers arbitrarily close to 0 and 1 will appear). Clearly this implies that all levels will be filled out in the end. Now let n_0 be so large that for any x_k in $\{x_1, x_2, \dots, x_{n_0}\}$ there exists x_ℓ such that $\mid x_k - x_\ell \mid < \frac{\varepsilon}{2}$. Next, choose N such that $N > H_{n_0}$, then obviously level N is yet completely empty. According to the observation at the beginning of this proof, there exists n_1 such that the complete level N is occupied by keys from $(x_k)_{k=n_0+1}^{n_1}$ We claim that for any two neighbours ν and $\tilde{\nu}$ of this level (say ν to the left of $\tilde{\nu}$) their keys x_k and $x_{\tilde{k}}$ differ less than ε. Suppose not, i.e., suppose that $|x_k - x_{\tilde{k}}| > \varepsilon$. Let $\nu \wedge \tilde{\nu}$ be the closest common ancestor of ν and $\tilde{\nu}$, i.e., $\nu \wedge \tilde{\nu}$ is on the path from ν and that from $\tilde{\nu}$ to the root, and its level is maximal with this property. Let x_j be the key at $\nu \wedge \tilde{\nu}$. Then clearly $x_k < x_j < x_{\tilde{k}}$. Let x_ℓ be the smallest number so that $1 \leq \ell \leq n_0$ and

$$x_k < x_\ell < x_j$$

(since $|x_k - x_{\tilde{k}}| > \varepsilon$, either x_ℓ exists, or there will exist a *largest* number x_r with $x_j < x_r < x_{\tilde{k}}$ and the proof will proceed similarly). Since $\ell \leq n_0 < k$, and since there are no other x_m between x_k and x_ℓ, x_ℓ has been assigned to a node ν^* on the path from the root to ν. Then there are two possibilities (writing $\nu' < \nu''$ if two nodes ν' and ν'' are on the same path to the root and the level of ν' is smaller than the level of ν'') : $\nu^* < \nu \wedge \tilde{\nu}$ or $\nu \wedge \tilde{\nu} < \nu^*$. Case 1: $\nu^* < \nu \wedge \tilde{\nu}$. Then we get a contradiction with the fact that $\nu^*, \nu \wedge \tilde{\nu}$

and ν are on the same path: since $x_\ell < x_j, x_j$ has gone to the *right* subtree of ν^*, but since $x_k < x_\ell$, x_k has gone to the *left* subtree of ν^*.
Case 2: $\nu \wedge \tilde{\nu} < \nu^*$. Then we get a contradiction with the fact that ν and $\tilde{\nu}$ are neighbouring nodes: now x_k must have gone to the left subtree both at $\nu \wedge \tilde{\nu}$ and at ν^*. □

We next show that equidistribution of x gives still more structure on $\mathcal{T}(x)$.

Theorem 5. *If $x = (x_k)_{k\geq 1}$ is a uniformly distributed sequence in $[0,1]$, then $H_n(x) = o(n)$ as $n \to \infty$.*

Proof. Suppose $H_n \geq c \cdot n$ infinitely often for some $c > 0$.
Choose N' such that the first node of level N' has a key which is smaller than $c/2$, and the last node of level N' has a key which is larger than $1 - c/2$. By the proposition above there is a level $N > N'$, such that the keys coming from x at level N, which we denote by $y_1 < y_2 < \ldots < y_{2^N}$ have the property that

$$y_2 < c,\ y_{i+1} - y_{i-1} < c,\ 1 - y_{2^N-1} < c.$$

For $\mathcal{T}_n(x)$, where n is large, we define

$$H_n(i) = \text{height of the subtree rooted at the node with key } y_i.$$

Then there is a j such that

$$N + H_n(j) \geq c \cdot n \quad \text{for infinitely many } n.$$

Keys x_k which will be moved to the subtree rooted at the jth node (with label y_j) should at least satisfy $y_{j-1} < x_k < y_{j+1}$, hence this implies that infinitely often

$$\frac{|\,\{k : x_k \in [y_{j-1}, y_{j+1}], k \leq n\}\,|}{n} \geq \frac{H_n(j)}{n} \geq \frac{cn - N}{n} = c - \frac{N}{n}.$$

This contradicts the fact that the left hand side converges to $y_{j+1} - y_{j-1}$ which is strictly smaller than c.

□

5. Small height

Let $x = (x_1, x_2, \ldots) = (\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \ldots)$ denote the van der Corput sequence. Fix c in the interval $(1,2]$ and define an index sequence $i(c) = (i_1, i_2, \ldots)$ by

$$i_n = \begin{cases} n & \text{if } n \geq \lceil c^{n-1} \rceil \\ \lceil c^{n-1} \rceil & \text{otherwise.} \end{cases}$$

Note that the sequence $i(c)$ is strictly increasing. We obtain a sequence $y(c) = (y_1, y_2, \ldots)$ by permuting the van der Corput sequence x in the

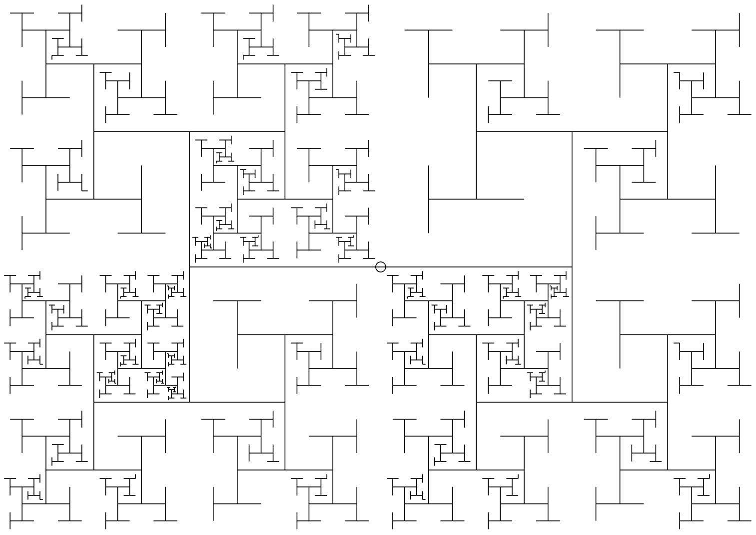

FIGURE 7. $\mathcal{T}_{1500}(\{\sqrt{7}k\})$.

following way. We first place the entry $x_{2^{n-1}} = \frac{1}{2^n}$ at position i_n in $y(c)$ for $n = 1, 2, \ldots$ and then place the remaining entries of x along the open positions, in the same order they appeared in x. Note that $y(2)$ is again equal to the van der Corput sequence.

Lemma 6. *For $c \in (1, 2]$, the sequence $y(c)$ is equidistributed on the unit interval and*

$$H_n(y(c)) = \begin{cases} n-1 & \text{if} \quad n \geq \lceil c^{n-1} \rceil \\ \lfloor \frac{\log n}{\log c} \rfloor & \text{otherwise} \end{cases}$$

for $n = 1, 2, \ldots$.

From this lemma it immediately follows that

$$H_n(y(c))/\log(n) \to 1/\log(c) \quad \text{as} \quad n \to \infty.$$

Proof. Note that all sequences $y(c)$ generate the same binary search tree and that at time n, the height of the tree is equal to the number of powers of $\frac{1}{2}$ stored in the tree minus 1. Hence, if $n \geq \lceil c^{n-1} \rceil$

$$\begin{aligned} H_n(y(c)) &= |\{k : i_k \leq n\}| - 1 \\ &= n - 1, \end{aligned}$$

and if $n < \lceil c^{n-1} \rceil$

$$\begin{aligned} H_n(y(c)) &= |\{k : i_k \le n\}| - 1 \\ &= |\{k : \lceil c^{k-1} \rceil \le n\}| - 1 \\ &= |\{k : \lceil c^k \rceil \le n\}| \\ &= |\{k : c^k \le n\}| \\ &= \lfloor \frac{\log n}{\log c} \rfloor . \end{aligned}$$

To see that the sequence $y(c)$ is equidistributed, note that the number of entries in $(x_1, \ldots, x_n)$ that do not appear in $(y_1, \ldots, y_n)$ is at most

$$|\{1 \le k \le n : i_k \le n\}| - 1 = H_n(y(c)),$$

where we used that $x_1 = y_1$. Hence, for $0 \le a < b \le 1$

$$\begin{aligned} &|\{k \le n : a < x_k < b\}| - H_n(y(c)) \\ &\qquad \le |\{k \le n : a < y_k < b\}| \\ &\qquad\qquad \le |\{k \le n : a < x_k < b\}| + H_n(y(c)). \end{aligned}$$

Since $H_n(y(c))/n \to 0$ and $\frac{1}{n}|\{k \le n : a < x_k < b\}| \to b - a$ as $n \to \infty$, it follows that $\frac{1}{n}|\{k \le n : a < y_k < b\}|$ converges to $b - a$ and hence the sequence $y(c)$ is equidistributed. □

References

[1] M. V. BERRY, J. GOLDBERG, *Renormalisation of curlicues.* Nonlinearity **1** (1988), 1–26.

[2] F. M. DEKKING, S. DE GRAAF, L. E. MEESTER, *On the node structure of binary search trees.* In *Mathematics and Computer Science - Algorithms, Trees, Combinatorics and Probabilities* (Eds. Gardy, D. and Mokkadem, A.), pages 31–40. Birkhauser, Basel, 2000.

[3] F. M. DEKKING, M. MENDÈS FRANCE, *Uniform distribution modulo one: a geometrical viewpoint.* J. Reine Angew. Math. **329** (1981), 143–153.

[4] J.-M. DESHOUILLERS, *Geometric aspect of Weyl sums.* In Elementary and analytic theory of numbers (Warsaw, 1982), pages 75–82. PWN, Warsaw, 1985.

[5] L. DEVROYE, *A note on the height of binary search trees.* J. Assoc. Comput. Mach. **33** (1986), 489–498.

[6] L. DEVROYE, A. GOUDJIL, *A study of random Weyl trees.* Random Structures Algorithms **12** (1998), 271–295.

[7] C. A. R. HOARE, *Quicksort.* Comput. J. **5** (1962), 10–15.

[8] H. M. MAHMOUD, *Evolution of random search trees.* John Wiley & Sons Inc., New York, 1992. Wiley-Interscience Series in Discrete Mathematics and Optimization.

[9] B. PITTEL, *Asymptotical growth of a class of random trees.* Ann. Probab. **13** (1985), 414–427.

Michel DEKKING, Peter VAN DER WAL
Thomas Stieltjes Institute for Mathematics
and Delft University of Technology
Faculty ITS, Department CROSS
Mekelweg 4, 2628 CD Delft
The Netherlands
E-mail : F.M.Dekking@its.tudelft.nl, vanderwal@eurandom.tue.nl