INFORMATIQUE THÉORIQUE ET APPLICATIONS

MIREILLE RÉGNIER

A limiting distribution for quicksort

Informatique théorique et applications, tome 23, n° 3 (1989), p. 335-343.

<http://www.numdam.org/item?id=ITA_1989__23_3_335_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 23, n° 3, 1989, p. 335 à 343)

A LIMITING DISTRIBUTION FOR QUICKSORT (*)

by Mireille RÉGNIER (1)

Communicated by P. FLAJOLET

Abstract. – *We establish the existence of a limiting distribution for the number of comparisons performed by quicksort, ot, equivalently, for the external path length of a binary search tree. We assume a uniform distribution of the data and prove a convergence in distribution and in L^p, $p \geq 1$. The proof is based on a martingale argument.*

Résumé. – *Nous prouvons l'existence d'une distribution limite pour le nombre de comparaisons effectuées par quicksort, ou, de manière équivalente, pour la longueur de cheminement externe d'un arbre binaire de recherche. Sous l'hypothèse d'une distribution uniforme des données, nous prouvons la convergence en distribution et dans L^p, $p \geq 1$. La preuve se fonde sur un argument de martingale.*

I. INTRODUCTION

In this note, we establish the existence of limiting distributions for Quicksort and Binary Search Trees. Quicksort, discovered by Hoare [HO62], is a widely used algorithm for internal sorting [KN73], [SE77], [SE83]. Notably, it is the standard sort on Unix systems. Binary search trees are a basic data structure for sorting and searching. By a standard equivalence principle (recalled in section II), evaluating the comparison cost of Quicksort reduces to the study of the *external path length* of a binary search tree. The first two moments, mean and variance, have been known for a long time [KN73], [SE77]. More recently, a study of the moments of any order has been done. In [HE87], these moments are obtained by successive differentiation of a multivariate generating function $C(z, q)$. This function satisfies a simple

(*) Received June 1987.
(1) I.N.R.I.A., 78153 Le Chesnay, France.

Informatique théorique et Applications/Theoretical Informatics and Applications
0296-1598 89/03 335 9/$2.90/© Gauthier-Villars

bivariate non-linear difference-differential equation:

$$\frac{\partial}{\partial z} C(z, q) = C^2(qz, q) \tag{1}$$

Though moments of an arbitrary fixed order can be computed, it appears impossible so far to derive a general formula (a conjecture can be made, *see* [HE87] and our conclusion). Thus, one cannot deduce the existence of a limiting distribution from the asymptotics of the moments. In our approach, we make use of *martingale theory* [FE57], [NE72] to prove the *existence of a limiting distribution* in law and in L^p. As pointed out in [HE87], this distribution is *not gaussian* since its third moment is not zero.

The plan of this note is the following. In section II, we recall briefly the Quicksort algorithm and state the relationship with Binary Search Trees. In section III, we first present some elementary notions on martingales. Then, we associate a martingale to Quicksort and prove our convergence results. Finally, in our conclusion, we discuss the problems that remain open.

II. QUICKSORT ALGORITHM

We first present the algorithm, and then the parameters to be analysed, and our probabilistic hypotheses.

Quicksort is a fast sorting algorithm, widely used for internal sort. The basic idea is the choice of a *partitioning element* K [SE83]. For example, let us consider the integer sequence:

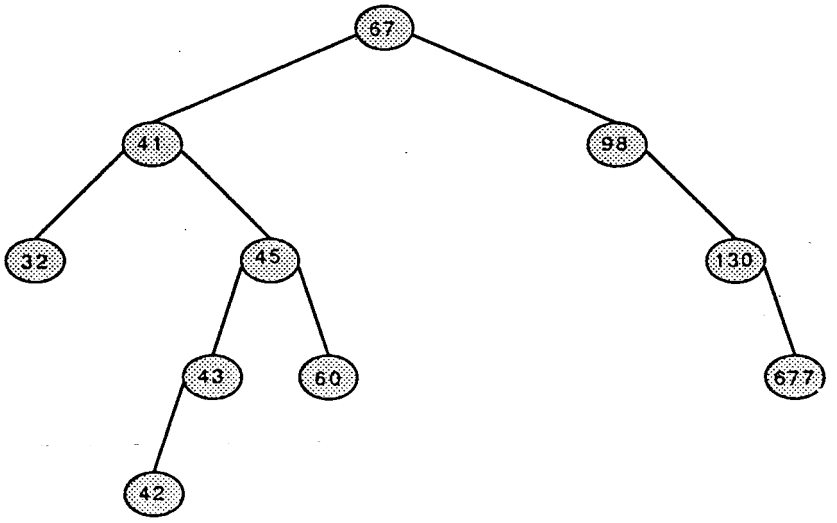
45	677	98	43	42	41	60	130	32	**67**

and choose $K=67$ as the partitioning element. Scanning the left sublist from left to right, one exchanges any key greater than K with a key of the right sublist, scanned from right to left. This builds a list where K has its final position, all the keys at its left (resp. at its right) being smaller then K. The intermediate stages are:

45	32							677	**67**
45	32	60				98	130	677	**67**
45	32	60	43	42	41	98	130	677	**67**
45	32	60	43	42	41	67	130	677	98

Then, the process can be applied recursively to both sublists (45...41) and (130...98).

These successive stages can be represented by a binary search tree. Each key is used as a partioning element in a recursive call and the level of this call is also the depth of the node containing that key.

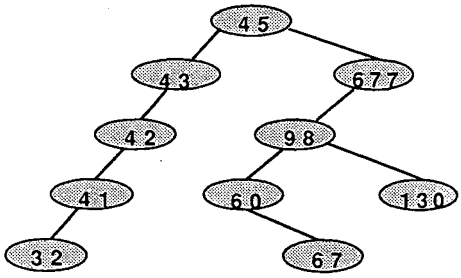


Now, to evaluate the cost running Quicksort on a data set, we count the number of comparisons to be performed. Each key in a node is compared once to all the keys occuring in the path from the root to that node. Thus, our cost is equal to the *External Path Length* of the binary search tree built. As the operations of the algorithm only depend on the relative order of the keys, this data set can also be considered as a permutation in σ_n. We make the standard assumption of a uniform distribution of data [KN73], [SE77], [HE87].

PROBABILISTIC MODEL: *For data sets of size n, the n! relative orders are equally likely*

Under this model, the analysis of Quicksort can be reduced and performed on binary search trees built by *successive insertions.* This standard algorithm for binary search trees is fully described in [KN73], Vol. 3,6.2.2. We only

give here an example: the binary search tree built by successive insertions of the data listed above.



Each order of the keys is associated, for both algorithms, to a binary search tree. Althrough these trees are different, we have the Equivalence Principle:

EQUIVALENCE PRINCIPLE: [KN73], Vol. 3, p. 428 and [SE77]: *The distribution of the binary search trees associated to Quicksort when data are uniformly distributed is the same as the distribution of the binary search trees built by successive insertions of n random keys.*

This allows us to consider only the evaluation of the *External Path Length of a binary search tree built by successive insertions.* We will chose the latter approach in the following. In this case, the definition of a martingale, as well as the proof of the main properties, is more intuitive, hence easier. We shall prove:

THEOREM 1: *There exists a random variable Z such that the internal path length (and the external path length) of a random binary search tree, once centered and normalized, converges, in law and in L^2, to Z.*

III. THE LIMITING DISTRIBUTION

III.1. Elementary notions on martingales

Here, we recall very briefly some notions of probability measure – the σ-fields and the measurability – and the definition and elementary properties of martingales. One can refer to [FE57], Vol. 2, IV.3 and IV.4 for more

details on probability measure and to [FE57], Vol. 2, VII.9 or [NE75] for martingale theory.

In probability theory, one deals with special subsets of the sample space, as defined below:

DEFINITION A: *A σ-field $\mathscr{A}$ is a system of subsets of a set Ω closed under completion and the formation of countable unions and intersections.*

Random variables are real functions on Ω, but one only uses functions for which a distribution function can be defined.

DEFINITION B: *Given a σ-field $\mathscr{A}$ on a set Ω, a real-valued function u on Ω is called $\mathscr{A}$-measurable if for each t the set of all points x where $u(t) \leqq t$ belongs to $\mathscr{A}$.*

The general definitions of martingales involve an increasing sequence of σ-fields (B_n) and a sequence (Z_n) of B_n mesurable random variables. A particular case, that we consider here, occurs when (B_n) are the σ-fields generated by a sequence of r. v. (X_n), *i. e.* generated by the subsets $X_n^{-1}(I)$, $I \subset \mathscr{R}$. The r. v. Z_n are X_n-mesurable or even simply defined as:

$$Z_n = f_n(X_1, \ldots, X_n)$$

where f_n are R^n-mesurable. Thus, we shall use the definition:

DEFINITION C: *Let (X_n) be a sequence of random variables and (Z_n) a sequence of $\mathscr{B}_n$-measurable random variables. The conditional expectation $E(Z/X_1 \ldots X_{n-1})$ is the X_n-measurable random variable satisfying, for every X_n-measurable random variable Y:*

$$E(Z \,.\, Y) = E(E(Z/X_1 \ldots X_{n-1}) \,.\, Y).$$

Two important properties follow from this definition.

(*a*) Z and $E(Z/X_1 \ldots X_{n-1})$ have the same expectation.

(*b*) If Y is $\mathscr{B}_n$-measurable, then:

$$E(YX/X_1 \ldots X_{n-1}) = Y \,.\, E(Z/X_1 \ldots X_{n-1}).$$

Hint for (*a*): Use the definition with $Y \equiv 1$.

DEFINITION D: *Let (X_n) be a sequence of random variables and (Z_n) a sequence of X_n-measurable random variables. (Z_n) is a martingale if it satisfies the property:*

$$E(Z_{n+1}/X_1, \ldots, X_n) = Z_n.$$

From our remark above, such variables have a constant expectation; moreover, simple properties on the variances are sufficient to establish the existence of a limit, as stated in Theorem 2 [FE57]:

THEOREM 2: *Let (Z_n) be an infinite martingale with $E(Z_n^2) < C < +\infty$ for all n. There exists a random variable Z such that $Z_n \to Z$ in law and in L^2. Furthermore $E(Z_n) = E(Z)$ for all n.*

Moreover, if (Z_n) satisfies: $E(|Z_n|^p) < C_p$ for some $p > 2$, then $Z_n \to Z$ in L^p.

III.2. A martingale associated to Quicksort

We prove here the existence of a limiting distribution for the *path length* of a binary search tree. As pointed out in Section II, this is also the number of comparisons performed when running Quicksort.

We first state our notations. The *depth of insertion* of a record is the depth of the node where this record is inserted, where we take the depth of the root to be 0. We note X_n the random variable counting the depth of insertion of a random record in a random binary search tree of size $n-1$. (We have $X_1 \equiv 0$ and $X_2 \equiv 1$.) The *internal path length* is then:

$$IPL_n = \sum_{i=1}^{n} X_i.$$

For a reason of convenience, we also define:

$$Y_n = \sum_i (X_i + 2) = IPL_n + 2n$$

the random variable representing the *external path length* of a binary search tree built on n records. Noting that the path lengths are simple X_n-mesurable functions, we are now ready to define a martingale. This is the purpose of the following Proposition 3. Then, applying Theorem 2 to this result, will yield directly our main result, Theorem 1 above.

PROPOSITION 3: *The random variables*

$$Z_n = \frac{Y_n - 2(n+1)(H_{n+1} - 1)}{n+1} = \frac{IPL_n - 2(n+1)H_n + 4n}{n+1}$$

form a martingale with a null expectation. Their variances satisfy:

$$V_n = \left(7 - \frac{4\pi^2}{6}\right) + O\left(\frac{1}{n}\right).$$

Proof of Proposition 3: The proof is based on the following two lemmas.

LEMMA A: *The conditional expectation of the random variables* X_n *and* Y_n *satisfy:*

$$E(X_n/X_1 \ldots X_{n-1}) = \frac{Y_{n-1}}{n}$$

$$E(Y_n/X_1 \ldots X_{n-1}) = \frac{n+1}{n} Y_{n-1} + 2$$

LEMMA B: *The expectation of the internal and external path length satisfy:*

$$\begin{cases} E(IPL_n) = 2(n+1)H_n - 4n \\ E(Y_n) = 2(n+1)(H_{n+1} - 1) \end{cases}$$

Proof of Lemma A: From the definition, $nE(X_n/X_1 \ldots X_{n-1})$ is the sum of the possible values for X_n, in a tree built by successive insertions at depths $X_1, \ldots, X_{n-1}$. The insertion of the n-1-st key, K_{n-1}, at depth X_{n-1}, modified the binary search tree built on the keys $K_1, \ldots, K_{n-2}$ in the following way. A leaf at depth X_{n-1} was changed to a node, where K_{n-1} was inserted, while two more leaves at depth $X_{n-1}+1$ were created, where the n-th key K_n may now be inserted. Thus:

$$nE(X_n/X_1 \ldots X_{n-1}) = (n-1)E(X_{n-1}/X_1 \ldots X_{n-2}) - X_{n-1} + 2(X_{n-1}+1).$$

Unwinding this recurrence, with $E(X_1) \equiv 0$, yields:

$$nE(X_n, X_1 \ldots X_{n-1}) = \sum_{i=1}^{n-1} (X_i + 2) = Y_{n-1}.$$

Now:

$$E(Y_n/X_1 \ldots X_{n-1}) = E(Y_{n-1} + X_n + 2/X_1 \ldots X_{n-1}) = Y_{n-1} + \frac{1}{n} Y_{n-1} + 2. \quad \blacksquare$$

Proof of Lemma B: It follows directly from Lemma A that:

$$\frac{E(Y_n)}{n+1} = \frac{E(Y_{n-1})}{n} + \frac{2}{n+1} = 2 \sum_{i=2}^{n} \frac{1}{i+1} + 1 = 2(H_{n+1} - 1). \quad \blacksquare$$

As an immediate consequence, one finds the martingale property for Z_n:

$$E(Z_n/X_1 \ldots X_{n-1}) = E(Y_n - E(Y_n)/n+1/X_1 \ldots X_{n-1})$$
$$= \frac{Y_{n-1}}{n} + \frac{2}{n+1} - \frac{E(Y_n)}{n+1} = Z_{n-1}.$$

Now, to derive the expression of the variances, we rewrite:

$$Z_n = \frac{n}{n+1} Z_{n-1} + \frac{X_n - E(X_n)}{n+1}$$

and get:

$$E(Z_n^2) = \frac{n^2}{(n+1)^2} E(Z_{n-1}^2) + \frac{\mathrm{Var}(X_n)}{(n+1)^2} + \frac{2n}{(n+1)^2} E(Z_{n-1}(X_n - E(X_n))).$$

Using $E([X_n - E(X_n)]/X_1 \ldots X_{n-1}) = Z_{n-1}$ and noting $\pi_n = \frac{n+2}{n+1}$, we get:

$$\frac{E(Z_n^2)}{\pi_n} = \frac{E(Z_{n-1}^2)}{\pi_{n-1}} + \frac{\mathrm{Var}(X_n)}{(n+1)(n+2)}.$$

The moments of the depth of insertion X_n are well known [KN73], [LO87] (the limiting distribution being gaussian). In particular:

$$\mathrm{Var}(X_n) = 2(H_n - 1) - 4(H_n^{(2)} - 1),$$

with $H_n^{(2)} = \sum_{i=1}^{n} 1/i^2$. Summing these expressions yields:

$$E(Z_n^2) = \pi_n \left(7 - 4H_{n+1}^{(2)} + O\left(\frac{1}{n}\right)\right) = 7 - 4H_{n+1}^{(2)} + O\left(\frac{1}{n}\right).$$

This completes the proof of Proposition 3 (and hence also Theorem 1). ■

IV. CONCLUSION

We have been able in this note to prove the existence of a limiting distribution for the path length of a binary search tree or, equivalently, the number of comparisons performed by Quicksort. The proof uses a martingale

argument and enables us to rederive, in passing, the first two moments of the distribution.

It is worth pointing out here that in [HE87], the derivation of the equation (1) gave a surprisingly simple formula for the *cumulants* $(k_p)_{p \in N}$ of the distribution, namely:

$$k_p(n) = (-1)^{p+1} 2^p (p-1)! \zeta(p) + C_p + O_p\left(\frac{1}{n}\right)$$

where $\zeta(p) = \sum_{n=1}^{\infty} \frac{1}{n^p}$, and it is *conjectured* that such a formula, verified for $p \leqq 20$, holds true for any p, with a uniform approximation. Solving this open problem and finding the C_p would enable us to explicity characterize the limiting distribution for which we have given here a "non constructive" existence proof.

REFERENCES

[BM85] F. Baccelli and A. M. Makowsky, *Direct martingale arguments for stability: The M/G/1 case*, Systems & Control Letters, Vol. 6, 1985, p. 181-186.

[FE57] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, Wiley, 1957.

[HE87] P. Hennequin, *Combinatorial Analysis of Quicksort Algorithm*, RAIRO, Th. Informatics and Applications (to appear).

[HO62] C. A. Hoare, *Quicksort*, Computer Journal, Vol. 5, N° 1, 1962.

[KN73] D. Knuth, *The Art of Computer Programming*, Vol. 3: *Sorting and Searching*, Addison-Wesley, Reading, Mass., 1973.

[LO87] G. Louchard, *Exact and Asymptotic Distributions in Digital and Binary Seach Trees*, RAIRO, Th. Informatics and Applications (to appear).

[NE75] J. Neveu, *Discrete-Parameter Martingale*, English Translation, North-Holland, Amsterdam, 1975.

[SE77] R. Sedgewick, *The Analysis of Quicksort Programs*, Acta Informatica, Vol. 7, 1977, pp. 327-355, and in *Quicksort*, Garland Pub. Co., New York, 1980.