INFORMATIQUE THÉORIQUE ET APPLICATIONS

MICHAEL CRAMER

A note concerning the limit distribution of the quicksort algorithm

Informatique théorique et applications, tome 30, n° 3 (1996), p. 195-207

<http://www.numdam.org/item?id=ITA_1996__30_3_195_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 30, n° 3, 1996, pp. 195-207)

A NOTE CONCERNING THE LIMIT DISTRIBUTION OF THE QUICKSORT ALGORITHM (*)

by Michael CRAMER ([1])

Communicated by Christian CHOFFRUT

Abstract. – *We perform simulations in order to obtain information on the limit distribution of the Quicksort algorithm. This distribution is also correlated to the external path length of a binary search tree. It turns out that the lognormal distribution is a very good approximation for that distribution. However, by exact and numerical calculation of some moments we shall demonstrate that these distributions are not the same.*

Résumé. – *La distribution limite de Quicksort (qui est la même que la distribution limite de la longueur de cheminement externe dans les arbres binaires de recherche) est inconnue. Nous montrons ici, par simulation, que l'approximation de cette distribution par une loi log-normale est en pratique excellente. Cependant prouvons aussi, par le calcul précis de quelques moments, que la distribution limite de Quicksort n'est pas log-normale.*

1. INTRODUCTION

Since the early sixties when Hoare [5] introduced the Quicksort algorithm it has become one of the most applied sorting algorithms. It is, e.g., the standard sorting procedure on Unix systems. The number of comparisons, say X_n, needed by Quicksort to sort a list of n elements given at random order is the essential quantity that determines the performance time in any implementation of Quicksort. Recently several new results on the distribution of X_n have been published.

Régnier [7] has established convergence of the normed centered law of X_n to an unknown limit distribution. The employed method applies the martingale convergence theorem [1] which is "non-constructive". Furthermore, the equivalence of the number of comparisons in Quicksort with the external path length in a binary search tree has been established

(*) Manuscript received July 1994.

([1]) Albert-Ludwigs-Universität Freiburg, Institut für Mathematische Stochastik, Hebelstraβe 27, D-79104 Freiburg, Germany.

Informatique théorique et Applications/Theoretical Informatics and Applications
0988-3754/96/03/$ 4.00/

and used. Independently, Rösler [8] has proved the same limit theorem for Quicksort by the use of a contraction property of an appropriate operator. The limit distribution is not normal (*cf.* [4], [7], [8]) and is characterized as the fixed point of that operator.

In this paper we examine what this limit distribution looks like. We find that a lognormal distribution with appropriate parameters is a very good approximation for the limit distribution of Quicksort.

From the fixed point equation we determine the first four moments of the limit distribution. These can also be deduced from the results of Hennequin [4].

Then we carry out simulations of X_n for $n = 1000$, 5000 and 20000. Comparisons with known distributions suggest that the limit distribution might be lognormal. However, we will demonstrate that this is not the case.

2. MOMENTS

Let $Y_n := (X_n - \mathbb{E}\, X_n)/n$. Régnier [7] proves the weak convergence of $(Y_n)_{n\in\mathbb{N}}$ to a limit distribution as well as the convergence in L^2.

Rösler [8] even derives the convergence of all moments and characterizes the limit distribution as the unique fixed point of the operator $T : D_2 \to D_2$, $T(P) = \mathcal{L}(\tau X + (1-\tau)\bar{X} + C(\tau))$ where D_2 is the space of distributions with mean 0 and finite second moment. τ is uniformly distributed on [0, 1], $\mathcal{L}(X) = P$, $\bar{X} \stackrel{d}{=} X$ and τ, X, $\bar{X}$ are independent ($\stackrel{d}{=}$ / $\stackrel{\mathcal{D}}{\to}$ means equality/convergence in distribution). The function C : $[0, 1] \to \mathbb{R}$ is defined by $C(x) = 2x \ln x + 2(1-x)\ln(1-x) + 1$. So

$$Y_n \stackrel{\mathcal{D}}{\to} Y, \tag{2.1}$$

$$\mathbb{E}\, Y_n^k \to \mathbb{E}\, Y^k \,(\mathbb{E}|Y|^k < \infty) \tag{2.2}$$

and

$$Y \stackrel{d}{=} \tau Y + (1-\tau)\bar{Y} + C(\tau), \tag{2.3}$$

where $\tau \sim \mathcal{U}(0, 1)$, $\bar{Y} \stackrel{d}{=} Y$, and τ, Y, $\bar{Y}$ are independent.

The recursion formula (2.3) provides us with the possibility to calculate moments of any order $k \in \mathbb{N}$ of the random variable (r.v.) Y (resp. its distribution $\mathcal{L}(Y)$). We obtain

$$\mathbb{E}\, Y^k = \mathbb{E}\,(\tau Y + (1-\tau)\bar{Y} + C(\tau))^k \quad \forall k \in \mathbb{N}. \tag{2.4}$$

Clearly for $k = 1$ we know $\mathbb{E}\, Y = 0$ from (2.2) and the definition of Y_n. We will need the moments up to order 4. For $k = 2$ to 4 equation (2.4) becomes

$$\mathbb{E}\, Y^2 = 3\,\mathbb{E}\,(C^2(\tau)) \tag{2.5}$$

$$\mathbb{E}\, Y^3 = 2\,\mathbb{E}\,(C^3(\tau)) + 12\,\mathbb{E}\,(\tau^2\ C^2(\tau)) \cdot \mathbb{E}\, Y^2 \tag{2.6}$$

$$\begin{aligned}\mathbb{E}\, Y^4 = {} & \frac{5}{3}\,\mathbb{E}\,(C^4(\tau)) + 20\,\mathbb{E}\,(\tau^2\ C^2(\tau)) \cdot \mathbb{E}\, Y^2 \\ & + 10\,\mathbb{E}\,(\tau^2 (1-\tau)^2) \cdot (\mathbb{E}\, Y^2)^2 \\ & + \frac{40}{3}\,\mathbb{E}\,(\tau^3\ C(\tau)) \cdot \mathbb{E}\, Y^3,\end{aligned} \tag{2.7}$$

where $\mathbb{E}\, Y = 0$ and the independence of Y, $\bar{Y}$ and τ have been used.

The following integral formulas serve to solve the integrals involved in (2.5)-(2.7).

LEMMA 1: *For j, $k \in \mathbb{N}_0$ holds*

$$\int_0^1 x^j (\ln x)^k\, dx = \frac{(-1)^k\, k!}{(j+1)^{k+1}}.$$

Proof: Fix j and show the above formula by induction on k. Use partial integration. □

LEMMA 2: *For $j \in \mathbb{N}_0$, $k \in \mathbb{N}$ holds*

$$\int_0^1 x^j\, \frac{(\ln x)^k}{1-x}\, dx = (-1)^k\, k! \left[\zeta(k+1) - \sum_{i=1}^{j} \frac{1}{i^{k+1}}\right],$$

where $\zeta(\cdot)$ is the Riemann zeta function.

Proof: Fix k and show the statement by induction on j. [3, p. 549, 4.271, 4.] provides the following formula which we will use for $p = k + 1$:

$$\int_0^1 \frac{(\ln x)^{p-1}}{1-x}\, dx = \exp[\iota\,(p-1)\,\pi]\,\Gamma(p)\,\zeta(p).$$

Note that $\Gamma(k+1) = k!$ and $\exp(\iota\, k\, \pi) = (-1)^k$. This is the basis of the induction. The induction step $j \to j+1$ uses the expansion $\frac{x^{j+1}}{1-x} = -x^j + \frac{x^j}{1-x}$ as well as Lemma 1. □

THEOREM 2.1: *The second moment (which is identical to the variance) and the third moment of the limit distribution of the Quicksort algorithm have*

the following values:

$$\mathbb{E}\,Y^2 = 7 - \frac{2}{3}\pi^2 = 0.42026373\ldots \tag{2.8}$$

$$\mathbb{E}\,Y^3 = 16\,\zeta(3) - 19 = 0.23291044\ldots \tag{2.9}$$

The fourth moment has the numerical value

$$\mathbb{E}\,Y^4 = 0.73794549\ldots \tag{2.10}$$

Remarks

– The accuracy of the numerical values is at least 10^{-8}.

– The evaluation of $\mathbb{E}\,Y^4$ has been performed with the help of a numerical value for an integral calculated by Maple.

– The cumulants of X_n stated in [4, p. 331] may be converted into moments so that (2.8)-(2.10) are implicitly given in [4]. (2.8) is stated in [8]. □

3. SIMULATIONS OF Y_n FOR $n = 1000, 5000$ AND 20000

In this section we investigate on the shape of the limit distribution of Quicksort. In terms of section 2 this is the law of the r.v. Y. Since the sequence (Y_n) of r.vs converges in distribution to Y it is reasonable to expect a good coincidence between the distribution functions of Y and Y_n resp. if n is large enough.

To this end we run a C-program which mimics the decisive steps of the Quicksort algorithm for $n = 1000$, 5000 and 20000.

100000 runs were performed for each n. Therefore, the empirical distribution functions can be regarded as indistinguishable from the distribution functions of Y_n.

If we compare the three empirical distribution functions to each other, we find an excellent agreement of the graphs. Figure 3.1 contains the empirical distribution functions of Y_n for $n = 1000$, 5000 and 20000.

Since the distribution functions change hardly from $n = 1000$ to $n = 20000$ it is natural to suppose that the distribution function of Y looks like those of Y_n for $n \geq 1000$.

More informative than the distribution function is the empirical density function of a simulation. For $n = 5000$ one obtains Figure 3.2. Smoothing this function with help of the "smooth.spline-function" of S-plus (an

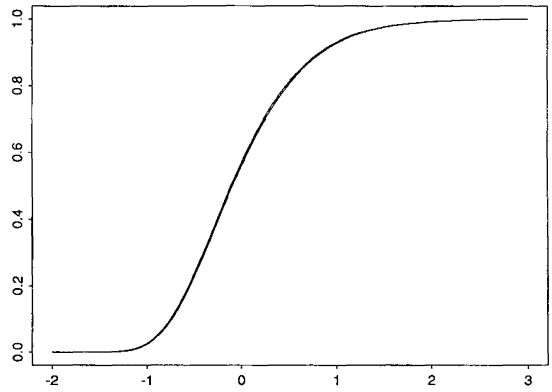

Figure 3.1. – The empirical distribution functions of Y_n for $n = 1000, 5000$ and 20 000.

implementation of the well-known statistical programming environment *S*) yields the graph in Figure 3.3.

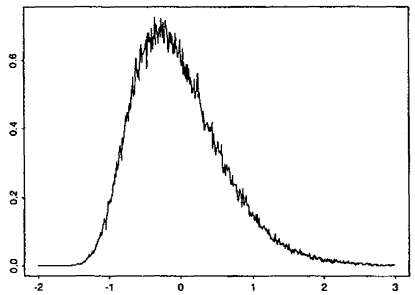

Figure 3.2. – Empirical density for Quicksort 5000.

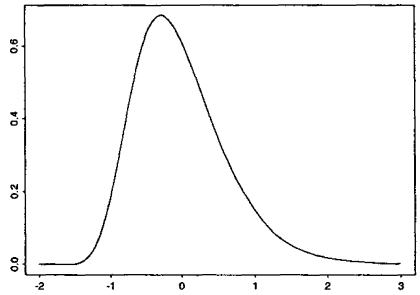

Figure 3.3. – Smoothed empirical density for Quicksort 5000.

This smoothed empirical density shows similarity to a 3-parameter-lognormal distribution density. By choosing five points of support we determine the parameters so that the sum of squares of the difference of the density function of the lognormal distribution and the smoothed empirical density function at the chosen points is minimal. We obtain the following parameters:

$$\left.\begin{aligned} \theta &= -2.272425 \\ \sigma &= 0.2800911 \\ \zeta &= 0.7717708 \end{aligned}\right\} \tag{3.1}$$

A good survey of the lognormal distribution can be found in [6, p. 112 ff]. The density f of the 3-parameter-lognormal distribution is given by

$$f(x) = \frac{1}{(x-\theta)\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}\left[\ln(x-\theta)-\zeta\right]^2\right) \mathbf{1}_{(\theta,\infty)}(x).$$

The results are shown below:

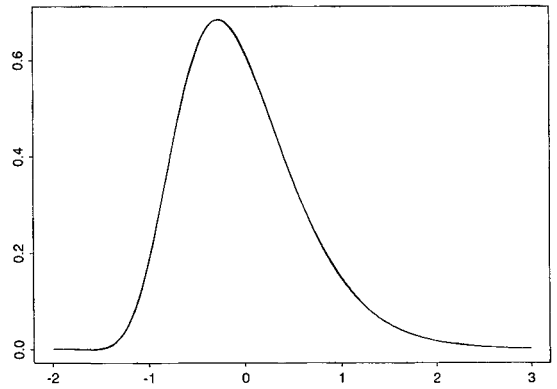

Figure 3.4. – Smoothed empirical density for Quicksort 5000 / Density of a lognormal distribution according to (3.1).

Figure 3.5 shows the corresponding distribution functions which can scarcely be distinguished. Their maximal deviation is about 0.3% (*Fig.* 3.6).

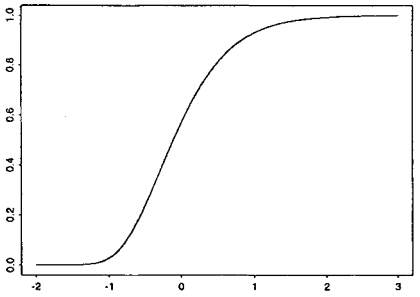

Figure 3.5. – Distribution function for Quicksort 5000 / Lognormal d.f. according to (3.1).

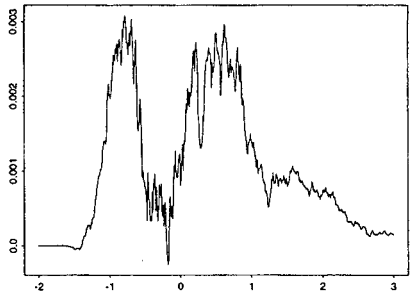

Figure 3.6. – Difference d.f. for Quicksort 5000 – Lognormal d.f. according to (3.1).

4. NON-LOGNORMALITY OF THE LIMIT DISTRIBUTION

In this section we shall prove that the limit distribution of Quicksort is not lognormal. This is a consequence of the fact that no lognormal distribution has the same first four moments as the limit distribution.

THEOREM 4.1: *Let* $X \sim \text{lognorm}(\theta, \sigma, \zeta)$ *such that* $\mathbb{E}X = 0$, $\mathbb{E}X^2 = 7 - \frac{2}{3}\pi^2$ *and* $\mathbb{E}X^3 = 16\zeta(3) - 19$. *Then*

$$\sigma = \sqrt{\ln a} \tag{4.1}$$

$$\zeta = \frac{1}{2}\left[\ln(\mathbb{E}X^2) - \ln(a^2 - a)\right] \tag{4.2}$$

$$\theta = -\exp\left(\zeta + \frac{1}{2}\sigma^2\right) \tag{4.3}$$

where a *is the only real solution of the cubic equation*

$$a^3 - 3a^2 - 4 - d = 0 \qquad \text{with} \quad d := \frac{(\mathbb{E}X^3)^2}{(\mathbb{E}X^2)^3}. \tag{4.4}$$

Proof: For $\tilde{X} := X - \theta$ we have (*cf.* [6, p. 115])

$$\mathbb{E}(\tilde{X}^r) = \exp\left(r\zeta + \frac{1}{2}r^2\sigma^2\right), \qquad r \in \mathbb{N} \tag{4.5}$$

$\mathbb{E}X = 0$ and (4.5) for $r = 1$ lead to formula (4.3). (4.5) for $r = 2$ supplies the second moment of $\tilde{X}$ which on the other hand determines $\mathbb{E}X^2$ if θ is known. So with (4.3) we obtain

$$\mathbb{E}X^2 + \exp(2\zeta + \sigma^2) = \exp(2\zeta + 2\sigma^2).$$

This implies (4.2) if we set $a := \exp(\sigma^2)$ which gives (4.1). So we have to check that σ, ζ, θ as above fulfill $\mathbb{E}X^3 = 16\zeta(3) - 19$ iff a is the real solution of (4.4). The substitution of (4.1)-(4.3) into (4.5) for $r = 3$ yields

$$\sqrt{\frac{\mathbb{E}X^2}{a^2 - a}}^{\,3}\,[a^{9/2} - a^{3/2} - 3a^{5/2} + 3a^{3/2}] = \mathbb{E}X^3,$$

which for $a > 1$ (since $\sigma > 0$) is equivalent to $(a+2)^2(a-1) = d$. Because of $(1 + d/2)^2 - 1 \neq 0$ this cubic equation has only one real solution, namely

$$a = \sqrt[3]{1 + \frac{d}{2} + \sqrt{\left(1 + \frac{d}{2}\right)^2 - 1}} + \sqrt[3]{1 + \frac{d}{2} - \sqrt{\left(1 + \frac{d}{2}\right)^2 - 1}} - 1. \quad \square$$

Remarks

– The proof also yields that a lognormal distribution whose parameters satisfy (4.1)-(4.4) has the same first three moments as the limit distribution of Quicksort.

– Numerical results:

One obtains $a = 1.07718027\ldots$ which implies

$$\left.\begin{aligned} \theta &= -2.333499\ldots \\ \sigma &= 0.27266604\ldots \\ \zeta &= 0.8101958\ldots \end{aligned}\right\} \tag{4.6}$$

□

THEOREM 4.2: $X \sim \text{lognorm}(\theta, \sigma, \zeta)$ *with* θ, σ, ζ *determined by* (4.1)-(4.4) *is not a fixed point of* T.

Proof: From Theorem 2.3 we know that the unique fixed point of T has moments according to (2.8)-(2.10). We calculate the fourth moment of X as above to obtain a contradiction to (2.10). This proves the theorem. Equation (4.5) leads in an analogous way as in the proof of Theorem 4.1 to

$$\mathbb{E}\, X^4 = (\mathbb{E}\, X^2)^2\, [a^4 + 2\, a^3 + 3\, a^2 - 3],$$

where $a \in \mathbb{R}$ stems from (4.4) as in Theorem 4.1. So we obtain

$$\mathbb{E}\, X^4 = 0.7642470\ldots \tag{4.7}$$

which clearly contradicts (2.10). (Note that the accuracy of 7 digits in (4.7) and 8 digits in (2.10) suffices to distinguish numbers which differ even in the second digit behind the decimal point). □

5. APPROXIMATION BY LOGNORMAL DISTRIBUTIONS

5.1. The stability of a lognormal distribution under T

We want to show that a lognormal distribution with parameters according to (4.1)-(4.4), say "limit moments lognormal distribution" (abbreviate: LML distribution), is nearly stable under the operator T. Unfortunately it seems difficult to determine the image of the LML distribution under T analytically. Thus we do so by simulations.

We simulate random variables which are lognormal with parameters as in (4.1)-(4.4) as well as a uniformly distributed random variable so that we can mimic the operation of T on a lognormal distribution. If we compare

the LML distribution function to its simulated picture under T we confirm a good conformity, the maximal deviation being about 0.6%.

In order to estimate this deviation we perform another simulation. We compare the LML distribution function to the empirical distribution function of a simulation of the LML distribution. The deviation in this case is at its maximum about 0.25%. This indicates that the maximal deviation of 0.6% is quite small because it is of the same order of magnitude as in the case where T is not applied.

So we conclude that the LML distribution is a very good approximation for the fixed point of T because the image of the LML distribution under T hardly differs from the LML distribution itself.

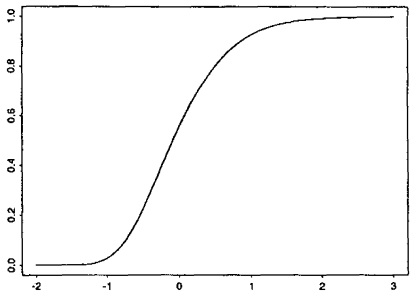

Figure 5.1. – Empirical d.f. after T operation / LML distribution.

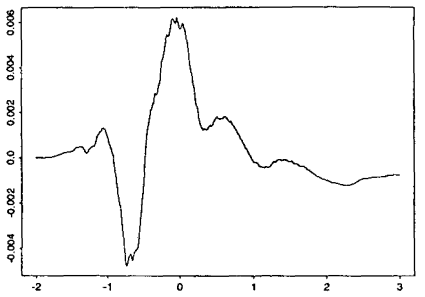

Figure 5.2. – Empirical d.f. after T operation – LML distribution.

5.2. Approximations for Y_n, $n = 1000, 5000, 20000$

For the empirical distributions, already shown in Figure 3.1, we shall now determine lognormal distributions such that the difference is minimal in each case. The corresponding parameters are given in Table 1 at the end of this section.

But first we will look at the difference between the empirical distribution functions of Y_n (for $n = 1000$, 5000 and 20000) and the LML distribution. In all three cases we find a maximum deviation of 0.4%. This is another indication that the LML distribution is a good approximation for the limit distribution of Quicksort.

For finite values of n based on simulations we can obtain even better approximations. From the variety of possibilities to determine appropriate parameters we take one which supplies a uniform approximation of the corresponding distribution functions. That is, we choose the parameters

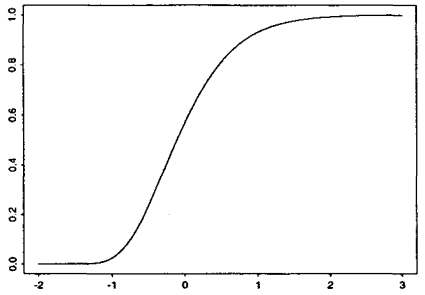

Figure 5.3. – Lognormal distribution function (Table 1)/Empirical distribution function, $n = 1000$.

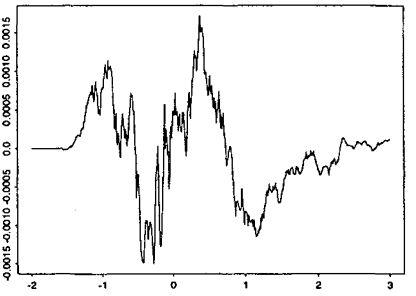

Figure 5.4. – Lognormal distribution function (Table 1) – Empirical distribution function, $n = 1000$.

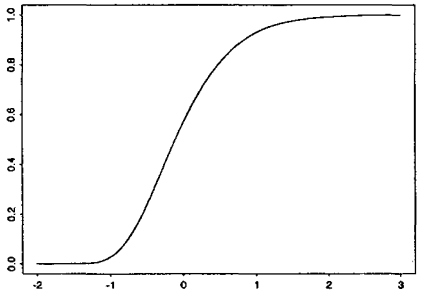

Figure 5.5. – Lognormal distribution function (Table 1)/Empirical distribution function, $n = 5000$.

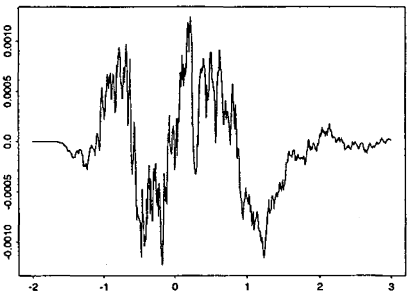

Figure 5.6. – Lognormal distribution function (Table 1) – Empirical distribution function, $n = 5000$.

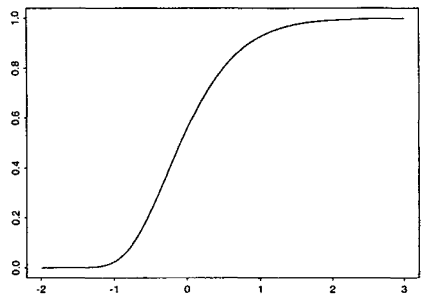

Figure 5.7. – Lognormal distribution function (Table 1)/Empirical distribution function, $n = 20000$.

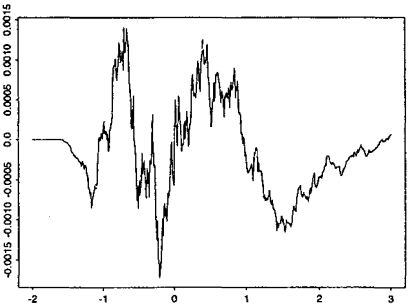

Figure 5.8. – Lognormal distribution function (Table 1) – Empirical distribution function, $n = 20\,000$.

of a lognormal distribution so that the sum of the squares of the differences between the empirical distribution function and the lognormal one is minimized where the sum is taken over all points $-2 + k \cdot 0.01$ ($k \in \{0, \ldots, 500\}$).

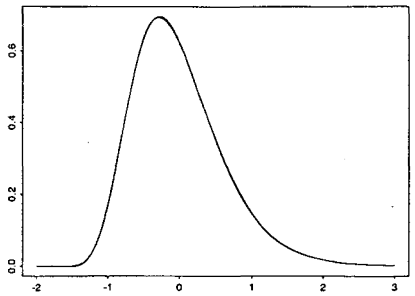

Figure 5.9. – Density of lognormal distribution (Table 1) / Smoothed empirical density, $n = 1000$.

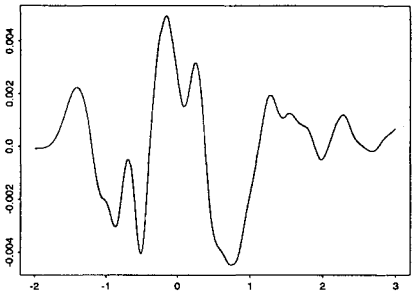

Figure 5.10. – Density of lognormal distribution (Table 1) – Smoothed empirical density, $n = 1000$.

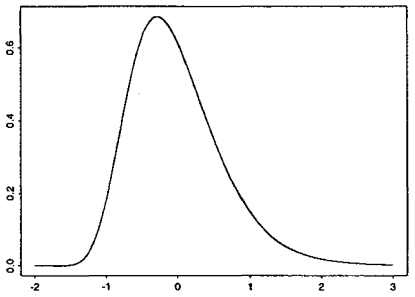

Figure 5.11. – Density of lognormal distribution (Table 1) / Smoothed empirical density, $n = 5000$.

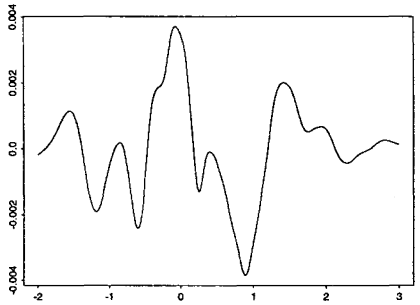

Figure 5.12. – Density of lognormal distribution (Table 1) – Smoothed empirical density, $n = 5000$.

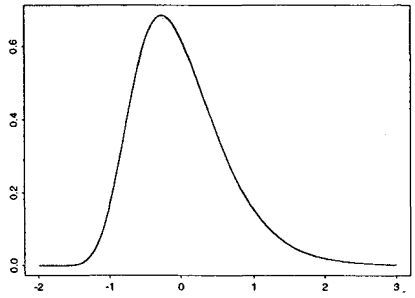

Figure 5.13. – Density of lognormal distribution (Table 1) / Smoothed empirical density, $n = 20\,000$.

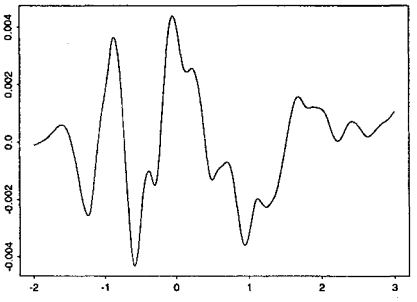

Figure 5.14. – Density of lognormal distribution (Table 1) – Smoothed empirical density, $n = 20\,000$.

The corresponding graphs show that this adaption is even better than the LML distribution for Y_n (n as above) in the sense that the maximum

difference of the distribution functions is now about 0.15%, which is significantly smaller than 0.4%.

Even the density functions fit quite well (smoothed empirical density against density of lognormal distribution with parameters according to Table 1 below).

In Table 1 the corresponding parameter values are given.

TABLE 1
Parameter values for Figures 5.3-5.14.

	Qs1000	Qs5000	Qs20000
θ	− 2.1241330	− 2.1871626	− 2.1651145
σ	0.2967048	0.2917981	0.2928525
ζ	0.7025572	0.7322862	0.7298622

So the distributions of Y_{1000}, Y_{5000} and Y_{20000} can be very properly described by lognormal distributions with parameters according to Table 1. For values of n between 1000 and 20000 one may still expect a rather good description of the law of Y_n by a lognormal distribution if one chooses the parameters according to Table 1 for that number in {1000, 5000, 20000}, which is next to n. For larger values of n parameters of (4.1)-(4.4), *i.e.* the parameters adapted to the moments of the limit distribution, are appropriate.

We should expect a maximum error of about 0.4% in the deviation of the distribution functions.

6. CONCLUSIONS

We have presented simulations of Y_n for $n = 1000$, 5000 and 20000. These hardly differ, which indicates that the limit distribution looks similar.

The main result is the fact that one can use a lognormal distribution with appropriate parameters to approximate the distribution of Y (Y_n respectively). Although we have proved that the limit distribution of Quicksort is not lognormal itself, we have seen that the LML distribution is an extremely good approximation, sufficient for practical purposes.

REFERENCES

1. P. BILLINGSLEY, *Probability and Measure*, Wiley, New York, 1986.
2. W. FELLER, *An Introduction to Probability Theory and its Applications*, Vol. II, Wiley, New York, 1957.

3. I. S. GRADSHTEYN and I. M. RYZHIK, *Table of Integrals, Series and Products*, Academic Press, New York, 1965.
4. P. HENNEQUIN, Combinatorial Analysis of Quicksort Algorithm, *Informatique Théorique et Applications*, 1989, *23*, pp. 317-333.
5. C. A. R. HOARE, Quicksort, *Computer Journal*, 1962, *5*, pp. 10-15.
6. N. L. JOHNSON and S. KOTZ, *Continuous univariate distributions - 1*, Houghton Mifflin, Boston, 1970.
7. M. RÉGNIER, A Limiting Distribution of Quicksort, *Informatique Théorique et Applications*, 1989, *23*, pp. 335-343.
8. U. RÖSLER, A Limit Theorem for Quicksort, *Informatique Théorique et Applications*, 1991, *25*, 85-100.