Elect. Comm. in Probab. **5** (2000) 95–99

ELECTRONIC COMMUNICATIONS in PROBABILITY

PERFECT SIMULATION FROM THE QUICKSORT LIMIT DISTRIBUTION

LUC DEVROYE[1]
School of Computer Science, McGill University
3480 University Street, Montreal, CANADA H3A 2K6
email: `luc@cs.mcgill.ca`

JAMES ALLEN FILL[2]
Department of Mathematical Sciences, The Johns Hopkins University
34th and Charles Streets, Baltimore, MD 21218-2682 USA
email: `jimfill@jhu.edu`

RALPH NEININGER[1]
Institut für Mathematische Stochastik, Universität Freiburg
Eckerstr. 1, D-79114 Freiburg, GERMANY
email: `rn@stochastik.uni-freiburg.de`

submitted May 11, 2000 *Final version accepted* June 5, 2000

AMS 2000 Subject classification: Primary 65C10; secondary 65C05, 68U20, 11K45.
Quicksort, random variate generation, simulation, perfect simulation, rejection method, Monte Carlo method, fixed-point equation.

Abstract
The weak limit of the normalized number of comparisons needed by the `Quicksort` *algorithm to sort* n *randomly permuted items is known to be determined implicitly by a distributional fixed-point equation. We give an algorithm for perfect random variate generation from this distribution.*

1 Introduction

Let C_n denote the number of key comparisons needed to sort a list of n randomly permuted items by `Quicksort`. It is known that

$$\mathbb{E}C_n = 2(n+1)H_n - 4n \sim 2n\ln n \quad \text{and} \quad \operatorname{Var} C_n \sim (7 - (2\pi^2/3))n^2,$$

where H_n denotes the nth harmonic number. Furthermore,

$$X_n := \frac{C_n - \mathbb{E}C_n}{n} \longrightarrow X$$

[1]Research for the first and third authors supported by NSERC Grant A3456 and FCAR Grant 90-ER-0291.
[2]Research for the second author supported by NSF grant DMS–9803780, and by the Acheson J. Duncan Fund for the Advancement of Research in Statistics.

in distribution. This limit theorem was first obtained by Régnier [8] by an application of the martingale convergence theorem. Rösler [9] gave a different proof of this limit law via the contraction method. Rösler's approach identifies the distribution of X to be the unique solution with zero mean and finite variance of the distributional fixed-point equation

$$X \stackrel{\mathcal{D}}{=} UX^{(1)} + (1-U)X^{(2)} + c(U), \tag{1}$$

where $X^{(1)}$, $X^{(2)}$, and U are independent; $X^{(1)}$ and $X^{(2)}$ are distributed as X; U is uniform $[0,1]$; c is given by $c(u) := 1 + 2u \ln u + 2(1-u)\ln(1-u)$; and $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. The limit random variable X has finite moments of every order which are computable from the fixed point equation (1). Tan and Hadjicostas [10] proved that X has a Lebesgue density. Not much else was known rigorously about this distribution until Fill and Janson recently derived some properties of the limiting density [5] and results about the rate of convergence of the law of X_n to that of X [6]. Some of these results are restated for the reader's convenience in the next section.

We develop an algorithm, based on the results of Fill and Janson, which returns a perfect sample of the limit random variable X. We assume that we have available an infinite sequence of i.i.d. uniform $[0,1]$ random variables. Our solution is based on a modified rejection method, where we use a convergent sequence of approximations for the density to decide the outcome of a rejection test. Such an approach was recently used by Devroye [3] to sample perfectly from perpetuities.

2 Properties of the quicksort density

Our rejection sampling algorithm is based on a simple upper bound and an approximation of (the unique continuous version of) the `Quicksort` limit density f. We use the following properties of f established in [5] and [6]. Let F_n denote the distribution function for X_n.

P1. f is bounded [5]:

$$\sup_{x\in\mathbb{R}} f(x) \le K := 16,$$

P2. f is infinitely differentiable and [5]

$$\sup_{x\in\mathbb{R}} |f'(x)| \le \widetilde{K} := 2466,$$

P3. With $\delta_n := (2\hat{c}/\widetilde{K})^{1/2} n^{-1/6}$, where $\hat{c} := (54cK^2)^{1/3}$, $c := 589$, we have [6]

$$\sup_{x\in\mathbb{R}} \left| \frac{F_n(x+(\delta_n/2)) - F_n(x-(\delta_n/2))}{\delta_n} - f(x) \right| \le R_n,$$

where $R_n := (432cK^2\widetilde{K}^3)^{1/6} n^{-1/6}$.

By property P2, f is Lipschitz continuous with Lipschitz constant $\widetilde{K}$. Therefore, Theorem 3.5 in Devroye [2, p. 320] implies the upper bound

$$f(x) \le \sqrt{2\widetilde{K}\min(F(x), 1-F(x))}.$$

Here, F denotes the distribution function corresponding to f. Markov's inequality yields $F(x) = \mathbb{P}(X \le x) \le (\mathbb{E}X^4)/x^4$ for all $x < 0$. Similarly, $1 - F(x) = \mathbb{P}(X > x) \le (\mathbb{E}X^4)/x^4$ for $x > 0$. The fourth moment of X can be derived explicitly in terms of the zeta function either by Hennequin's formula for the cumulants of X (this formula was conjectured in Hennequin [7] and proved later in his thesis) or through the fixed point equation (1). From (1), Cramer [1] computed $\mathbb{E}X^4 = 0.7379\ldots$ (accurate to the indicated precision), so $\mathbb{E}X^4 < 1$. Therefore, if we define

$$g(x) := \min\left(K, (2\widetilde{K})^{1/2} x^{-2}\right), \qquad x \in \mathbb{R}, \tag{2}$$

we have $f \le g$. The scaled version $\widetilde{g} := \xi g$ is the density of a probability measure for $\xi := 1/\|g\|_{L^1} = [4K^{1/2}(2\widetilde{K})^{1/4}]^{-1}$. A perfect sample from the density $\widetilde{g}$ is given by

$$[(2\widetilde{K})^{1/4}/K^{1/2}] S U_1 / U_2,$$

with S, U_1, and U_2 independent; U_1 and U_2 uniform $[0, 1]$; and S an equiprobable random sign (cf. Theorem 3.3 in Devroye [2, p. 315]).

Remark. According to the results of [5], f enjoys superpolynomial decay at $\pm\infty$, so certainly $f \le g$ for some g of the form $g(x) := \min(K, Cx^{-2})$. One way to obtain an explicit constant C is to use

$$x^2 f(x) \le \frac{1}{2\pi} \int_{-\infty}^{\infty} |\phi''(t)|\, dt, \qquad x \in \mathbb{R},$$

where ϕ is the characteristic function corresponding to f, and to bound $|\phi''(t)|$ [e.g., by $\min(c_1, c_2 t^{-2})$ for suitable constants c_1, c_2] as explained in the proof of Theorem 2.9 in [5]. But we find that our approach is just as straightforward, and gives a smaller value of C (although we have made no attempt to find the best C possible using the Fourier techniques of [5]).

3 The rejection algorithm

We have found an explicit, integrable upper bound on f. Furthermore, an approximation of f with explicit error estimate is given by P3. Let

$$f_n(x) := \frac{F_n(x + (\delta_n/2)) - F_n(x - (\delta_n/2))}{\delta_n}$$

with δ_n given in P3. Then $|f_n(x) - f(x)| \le R_n$ for all $x \in \mathbb{R}$, and $R_n \to 0$ for $n \to \infty$.
To calculate the values of f_n we require knowledge about the probabilities of the events $\{C_n = i\}$. Let $N(n, i)$ denote the number of permutations of n distinct numbers for which `Quicksort` needs exactly i key comparisons to sort. Then

$$\mathbb{P}(C_n = i) = \frac{N(n, i)}{n!}.$$

These probabilities are non-zero only if $n-1 \le i \le n(n-1)/2$. With the initializing conventions $N(0, 0) := 1$ and $N(i, 0) := 0$ for $i \ge 1$ and the obvious values $N(n, i) = 0$ for $i < n - 1$ and for $i > n(n-1)/2$, we have the following recursion for $n \ge 1$ and for $n - 1 \le i \le n(n-1)/2$:

$$N(n, i) = \sum_{k=1}^{n} \sum_{l=0}^{i-(n-1)} N(k-1, l) N(n-k, i-(n-1)-l).$$

This recurrence is well known. To verify it, assume that the first pivot element is the kth largest element out of n. Then the number of permutations leading to i key comparisons is the number $N(k-1,l)$ of permutations of the items less than the pivot element which are sorted with l key comparisons, multiplied by the corresponding number of permutations for the elements greater than the pivot element, summed over all possible values of k and l. Note that $n-1$ key comparisons are used for the splitting procedure. Observe that we also have $\mathbb{E}C_n = \sum_i iN(n,i)/n!$. The table $(N(n,i) : i \le n(n-1)/2)$, and $\mathbb{E}C_n$, can be computed from the previous tables $(N(k,i) : i \le k(k-1)/2)$, $0 \le k < n$, in time $O(n^5)$. Then, observe that, for $y < z$,

$$F_n(z) - F_n(y) = \frac{1}{n!} \sum_{\mathbb{E}C_n + ny < i \le \mathbb{E}C_n + nz} N(n,i),$$

and thus $f_n(x)$ is computable from the table $(N(n,i) : i \le n(n-1)/2)$ and $\mathbb{E}C_n$ in time $O(n(z-y)) = O(n\delta_n) = O(n^{5/6})$. Now, the following rejection algorithm gives a perfect sample X from the `Quicksort` limit distribution F:

```
repeat
  generate U, U_1, U_2 uniform [0,1]
  generate S uniform on {-1,+1}
  X ← ((2K̃)^{1/4}/K^{1/2}) S U_1/U_2
  T ← U g(X)    (where g(x) := min(K, (2K̃)^{1/2}/x^2))
  n ← 0
  repeat
    n ← n+1
    compute the full table of N(n,i) for all i ≤ n(n-1)/2
    Y ← f_n(X)
  until |T - Y| ≥ R_n
  Accept = [T ≤ Y - R_n]
until Accept
return X
```

This algorithm halts with probability one, and produces a perfect sample from the `Quicksort` limit distribution. The expected number of outer loops is $\|g\|_{L^1} = 4K^{1/2}(2\widetilde{K})^{1/4} \doteq 134.1$. Note, however, that the constants K and $\widetilde{K}$ are very crude upper bounds for $\|f\|_\infty$ and $\|f'\|_\infty$, which from the results of numerical calculations reported in [10] appear to be on the order of 1 and 2, respectively.
Moreover, considerable speed-up could be achieved for our algorithm by finding another approximation f_n to f that either is faster to compute or is faster to converge to f (or both). One promising approach, on which we hope to report more fully in future work, is to let $f_1, f_2, \ldots$ be the densities one obtains, starting from a suitably nice density f_0 (say, standard normal), by applying the method of successive substitutions to (1). Indeed, Fill and Janson [6] show that then $f_n \to f$ uniformly at an exponential rate. However, one difficulty is that these computations require repeated numerical integration, but it should be possible to bound the errors in the numerical integrations using calculations similar to those in [5].

Remark. Let $k \equiv k_n := \lfloor \log_2(n+1) \rfloor$. We noted above that if $N(n,i) > 0$, then $n-1 \le i \le n(n-1)/2$. This observation can be refined. In fact, using arguments as in [4], it can be shown that $N(n,i) > 0$ if and only if $m_n \le i \le M_n$, with

$$\begin{aligned} m_n &:= k(n+1) - 2^{k+1} + 2 \sim n\log_2 n = (1/\ln 2)\, n\ln n = (1.44\ldots)\, n\ln n \\ &= \text{the total path length for the complete tree on } n \text{ nodes} \end{aligned}$$

and $M_n := n(n-1)/2$. These extreme values satisfy the initial conditions $m_0 = 0 = M_0$ and, for $n \ge 1$, the simple recurrences

$$m_n = m_{n-1} + \lfloor \log_2 n \rfloor \qquad \text{and} \qquad M_n = M_{n-1} + (n-1).$$

References

[1] Cramer, M. 1996, A note concerning the limit distribution of the quicksort algorithm. *RAIRO, Theoretical Informatics and Applications* **30**, 195–207.

[2] Devroye, L. 1986, *Nonuniform Random Variate Generation.* Springer–Verlag, New York.

[3] Devroye, L. 1999, Simulating perpetuities. Preprint.

[4] Fill, J. A. 1996, On the distribution for binary search trees under the random permutation model. *Random Structures and Algorithms* **8**, 1–25.

[5] Fill, J. A. and Janson, S. 2000, Smoothness and decay properties of the limiting Quicksort density function. Refereed article, to apppear in a book edited by D. Gardy and A. Mokkadem, published in 2000 by Birkhäuser, and based on the Colloquium on Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities (University of Versailles–St. Quentin, Versailles, France, September, 2000). Preprint available from `http://www.mts.jhu.edu/~fill/`.

[6] Fill, J. A. and Janson, S. 2000, `Quicksort` asymptotics. Unpublished manuscript.

[7] Hennequin, P. 1989, Combinatorial analysis of quicksort algorithm. *RAIRO, Theoretical Informatics and Applications* **23**, 317–333.

[8] Régnier, M. 1989, A limiting distribution for quicksort. *RAIRO, Theoretical Informatics and Applications* **23**, 335–343.

[9] Rösler, U. 1991, A limit theorem for “Quicksort”. *RAIRO, Theoretical Informatics and Applications* **25**, 85–100.

[10] Tan, K. H. and Hadjicostas, P. 1995, Some properties of a limiting distribution in Quicksort. *Statistics & Probability Letters* **25**, 87–94.