Strong Concentration for Quicksort*

Colin McDiarmid† Ryan Hayward‡

Proc. 3rd ACM-SIAM SODA (1992) 414-421

Abstract

Let Q_n be the random number of comparisons made by quicksort in sorting n distinct keys, when we assume that all $n!$ possible orderings are equally likely. Known results concerning moments for Q_n do not show how rare it is for Q_n to make large deviations from its mean. Here we give a good approximation to the probability of such a large deviation, and find that this probability is quite small. As well as the basic quicksort we consider the variant in which the partitioning key is chosen as the median of $(2t+1)$ keys.

1 Introduction

In the short history of computer science, Hoare's quicksort has emerged as one of the classic algorithms. There are several reasons for this.

First, the algorithm is efficient. It is an $O(n \log n)$ expected time sorting algorithm (where n is the number of keys to be sorted). Arguably one of the best general purpose computer sorting algorithms from a point of view of space and time efficiency, it is the basis for example of the Unix "sort" feature, which uses the variant of quicksort in which the partitioning key is chosen as the median of three keys.

Second, the algorithm embodies two paradigms that are today considered key ideas in algorithm design, namely divide and conquer, and randomization.

Third, since the introduction of quicksort [Hoa61], an extensive body of literature has been published that is based on the design and analysis of many variants of the original algorithm. (Many of the commonly considered variants were in fact anticipated by Hoare in [Hoa62]. For a discussion of this, see [Sed80].) Indeed, the study of quicksort and its variants has become a model for the analysis of algorithms in general. Examples in point are the work by Knuth [Knu73] and by Sedgewick [Sed80]. (For other recent results on the analysis of quicksort specifically pertinent to this paper, see [Hen89], [Rég89] or [Rös].)

*The support of N.S.E.R.C. and the Alexander von Humboldt-Stiftung is gratefully acknowledged.

†Department of Statistics, Oxford University

‡Department of Computing and Information Science, Queen's University at Kingston, Canada

The point of this paper is to establish a new and rather precise result concerning quicksort's expected performance. We establish much tighter bounds than have been shown previously on

> *the probability that the number of comparisons of a random execution of quicksort will have a large deviation from the expected number of comparisons.*

Part of the appeal of our result is that the ideas behind the proofs are motivated by a recently popularised combinatorial approach known as the "method of bounded differences".

Before introducing the definitions we need in order to state our theorems precisely, let us discuss the two variants of quicksort we shall refer to in this paper. First, by "basic" quicksort (and unless otherwise stated, that is the version we will be referring to) we mean the original, unadorned version:

> *A partitioning key is selected at random from the list of unsorted keys, and used to partition the keys into two sublists. The algorithm is called recursively on remaining unsorted sublists, until sublists have size one or zero.*

One common variant is to use a different sorting algorithm (usually insertion sort) for lists whose size is not greater than a certain threshold value M. We refer to this variant as "cutting at length M". Another common variant is to select the partitioning key as the median of $2t+1$ keys selected from the list of unsorted keys. (Observe that basic quicksort can be considered as the "median of 1" version of this variant.) A comprehensive survey and comparative analysis of common variants is given in [Sed80].

We now introduce some notation. Let Q_n be the number of key comparisons made when (basic) quicksort sorts n keys. We make the usual assumption that the n keys are distinct and that all $n!$ linear orders are equally likely. (Alternatively, our results apply to a suitably randomised version of quicksort acting on any list of n distinct keys.) We further assume that the partitioning phase of the algorithm is performed so that the resulting sublists are also "random". Some care must be taken to ensure this, but it is not difficult to do. See [Sed80] for a description of such a partitioning algorithm, and a proof that the randomness of the sublists is preserved.

A straightforward and well known analysis (in fact, so well known it is likely to appear in an undergraduate algorithms course!) shows that the expected number $q_n = E[Q_n]$ of key comparisons satisfies $q_0 = 0$ and for $n \geq 1$

$$q_n = n - 1 + \frac{1}{n}\sum_{j=1}^{n}(q_{j-1} + q_{n-j}) . \tag{1.1}$$

It follows easily from Equation 1.1 that (as $n \to \infty$)

$$q_n = 2n \ln n + O(n) .$$

Of course one attaches more credibility to an average case result like this if it is known that there is a strong concentration of probability around the mean. In particular, given $\varepsilon > 0$, what bounds can be placed on the quantity

$$\Pr\left[\left|\frac{Q_n}{q_n} - 1\right| > \varepsilon\right] \quad ?$$

The point of this paper is to establish tight bounds for the above probability. There are some previous results of this form. For instance, from the fact that the variance of Q_n is $\Theta(n^2)$ (see [Knu73] or [Sed80]), it follows from Chebyshev's inequality that for $\varepsilon > 0$

$$\Pr\left[\left|\frac{Q_n}{q_n} - 1\right| > \varepsilon\right] = O((\varepsilon \ln n)^{-2}) .$$

Recently Hennequin [Hen89] used Chebyshev's inequality with fourth moments to show that for $\varepsilon > 0$

$$\Pr\left[\left|\frac{Q_n}{q_n} - 1\right| > \varepsilon\right] = O((\varepsilon \ln n)^{-4}) .$$

Observe that even this bound does not approach zero very quickly.

Our main result, whose proof is based on the idea of the "method of bounded differences" [McD89], is the following:

THEOREM 1.1. *For any $\varepsilon > 0$, as $n \to \infty$,*

$$\Pr\left[\left|\frac{Q_n}{q_n} - 1\right| > \varepsilon\right] = n^{-2\varepsilon(\ln\ln n + O(\ln\ln\ln n))} .$$

(Throughout the paper, we use $\ln n$ to denote $\log_e n$. Also, we shall sometimes use the notation $\ln^{(k)} n$, where as usual $\ln^{(k)} n = \ln\ln^{(k-1)} n$ for $k \geq 2$, and $\ln^{(1)} n = \ln n$.)

We shall prove this theorem for basic quicksort. However, the difference between the number of comparisons obtained with basic quicksort as opposed to the the more efficient "cutting at length M" variant is only a linear number in n, the number of keys. This linear term will not affect our results, so the theorem we have just stated (as well as the one that is to follow) also holds for the "cutting at length M" variant.

More care must be taken, however, when considering the "median of $(2t+1)$" variants of quicksorts. Recall that these are the variants in which the partitioning key is chosen as the median of $(2t+1)$ keys. Here t is a fixed non-negative integer, $t = 0$ corresponds to basic quicksort, and $t = 1$ is perhaps most common in practice.

We need more notation before stating our second and final theorem. Let $Q_n^{(t)}$ be the random number of comparisons taken to sort a random list of length n, and let $q_n^{(t)} = E\left[Q_n^{(t)}\right]$. Thus $Q_n^{(0)}$ is Q_n.

For $j = 1, 2, \ldots$ let $H_j = \sum_{i=1}^{j} 1/i$ and let $K_j = (H_{2j+2} - H_{j+1})^{-1}$. Thus for example $K_0 = 2$ and $K_1 = 12/7$. It is well known ([Van70] and [Sed80]) that

$$q_n^{(t)} = K_t\, n \ln n + O(n) \qquad \text{as } n \to \infty .$$

We can now state the "median of $(2t+1)$" analogue of our first theorem:

THEOREM 1.2. *For any $\varepsilon > 0$, as $n \to \infty$,*

$$\Pr\left[\left|\frac{Q_n^{(t)}}{q_n^{(t)}} - 1\right| > \varepsilon\right] = n^{-(t+1)K_t\varepsilon(\ln^{(2)} n + O(\ln^{(3)} n))} .$$

2 Basic quicksort

The purpose of this section is to outline the proof of Theorem 1.1. In §3 we will outline the proof of Theorem 1.2. In this section we will sketch the proofs of the lemmas in considerable detail. As the proofs of the corresponding lemmas in §3 are often similar, the sketches there will be sketchier.

2.1 List lengths in the partition tree. A standard technique in the analysis of randomised quicksort is to associate with an execution of quicksort a binary tree whose nodes contain the sublists obtained by the algorithm, the root corresponding to the original unsorted list, and the children of any node being the sublists obtained by the splitting of the parent node. We now describe this correspondence more precisely.

Consider the infinite binary tree, with nodes numbered 1,2,3,... level by level and left to right in the usual manner. (So for instance, the path from the root down the left side is 1,2,4,8,...). Each execution of (basic) quicksort yields a labelling of a subtree of this tree, corresponding to the recursive structure of quicksort. The root, node 1, is labelled with the unsorted list of n keys, and its "list length" L_1 is n. A partitioning key is chosen, and after partitioning an (unsorted) list of those keys less than the partitioning key forms the label for the left child (node 2) which then acts like the root of a new tree. Similarly, those keys greater than the partitioning key are sent to the right child of the root.

For each $j = 1, 2, \ldots$ let L_j be the length of the list to be sorted at node j. Thus $L_1 = n$ and only n of the L_j are non-zero. Our aim in this section is to show that as we move down the tree the list lengths shorten suitably with high probability. Let M_k^n be the maximum value of the list length L_j over the 2^k nodes j at depth k, that is

$$M_k^n = \max\{L_{2^k+i} : i = 0, 1, \ldots, 2^k - 1\} .$$

LEMMA 2.1. *For any $0 < \alpha < 1$ and any integer $k \geq \ln(1/\alpha)$*

$$\Pr[M_k^n \geq \alpha n] \leq \alpha \left(\frac{2e\ln(1/\alpha)}{k}\right)^k .$$

Proof. (*Sketch.*) The key observation is that we can obtain the exact joint distribution of $(L_1, L_2, \ldots)$ as follows. Here we are developing an idea from [Dev86].

Let the random variables $X_1, X_2, \ldots$ be independent with each uniformly distributed on the interval (0,1). Define random variables $\tilde{L}_1, \tilde{L}_2, \ldots$ as follows. Let $\tilde{L}_1 = n$ and for $i \geq 1$ let $\tilde{L}_{2i} = \lfloor X_i \tilde{L}_i \rfloor$ and $\tilde{L}_{2i+1} = \lfloor (1 - X_i)\tilde{L}_i \rfloor$. Then it is easily seen that $(L_1, L_2, \ldots)$ and $(\tilde{L}_1, \tilde{L}_2, \ldots)$ have the same joint distribution. Also, let $\tilde{M}_k^n$ be the maximum value of $\tilde{L}_j$ over the nodes j at depth k. Then it follows that M_k^n and $\tilde{M}_k^n$ have the same distribution.

Now define further random variables $Z_1, Z_2, \ldots$ from $X_1, X_2, \ldots$ as follows. Let $Z_1 = 1$ and for $i \geq 1$ let $Z_{2i} = X_i Z_i$ and $Z_{2i+1} = (1 - X_i)Z_i$. Then we have $\tilde{L}_i \leq n Z_i$ for each $i = 1, 2, \ldots$. Let Z_k^* be the maximum value of Z_j over the 2^k nodes j at depth k. Then

$$\tilde{M}_k^n \leq n Z_k^* .$$

Now the conclusion follows from a series of routine probability inequalities and arguments involving Z_k^*. □

2.2 The bounded differences approach. In this section we shall use the idea of the method of "bounded differences" to establish some necessary lemmas.

We shall be interested in the comparisons performed by quicksort on each of the levels of the partition tree. For $k = 0, 1, 2, \ldots$ let H_k be the random "history" of the comparisons performed at level k. Thus the vector $(H_0, H_1, \ldots, H_{k-1})$ records the entire history of the process for the first k levels: we call this the *k-history* $\underline{H}^{(k)}$. Observe that the k-history determines all the list lengths at level k. In particular, given a particular k-history $\underline{h}^{(k)}$ we know the value of M_k^n.

The key property of (basic) quicksort that makes our proofs work is given in the following lemma.

LEMMA 2.2. *Let* n *be a positive integer and let $A_n = \{n - 1 + q_{k-1} + q_{n-k} - q_n : k = 1, 2, \ldots, n\}$. Then $|x| \leq n$ for all $x \in A_n$.*

Proof. This follows from straightforward manipulations of Equation 1.1. □

LEMMA 2.3. *Let* n *and* k *be positive integers and let $\underline{h}$ be any possible k-history for Q_n. Then*

$$\left| E\left[Q_n \mid \underline{H}^{(k)} = \underline{h}\right] - q_n \right| \leq kn .$$

Proof. This follows from Lemma 2.2 by an easy induction on k. □

We shall use Lemma 2.3 when considering levels near the top of the partition tree. For levels further down the tree we use another inequality, again based on Lemma 2.2. First we need two preliminary lemmas taken (essentially) from [McD89].

LEMMA 2.4. (see Lemma 5.8 in [McD89]) *Let X be a random variable with $E[X] = 0$, $a \leq X \leq b$. Then for any $s > 0$,*

$$E[\exp\{sX\}] \leq \exp\{s^2(b-a)^2/8\} . \qquad \square$$

From the proof of Theorem 6.7 of [McD89] we may obtain immediately the following variant of that result.

LEMMA 2.5. *Let $(\Phi, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$ be a filter, let X be an integrable random variable, and let $X_0, X_1, \ldots, X_n$ be the martingale obtained by setting $X_k = E[X \mid \mathcal{F}_k]$. Suppose that for each $k = 1, \ldots, n$ there is a constant c_k such that for any $s > 0$ we have*

$$E[\exp\{s(X_k - X_{k-1})\} \mid \mathcal{F}_{k-1}] \leq \exp\{c_k^2 s^2/8\} .$$

Then for any $s > 0$

$$\Pr[\,|\,E[X \mid \mathcal{F}_n] - E[X]\,|\ \geq\ s\,] \leq\ 2\exp\left\{\ -2s^2 / \sum_{k=1}^{n} c_k^2\ \right\}.$$

□

LEMMA 2.6. *Let* $0 < k_1 < k_2$ *be integers, let* $\underline{h}^{(k_1)}$ *be a* k_1*-history such that* $M_{k_1}^n \leq \alpha n$ *and let* A *be the event that* $\underline{H}^{(k_1)} = \underline{h}^{(k_1)}$. *Then for any* $s > 0$

$$\Pr\left[\left(\left|\ E\left[Q_n \mid \underline{H}^{(k_2)}\right]\ -\ E[Q_n \mid A]\ \right| \geq s\right)\ \mid A\right] \leq\ 2\exp\left\{\ \frac{-s^2}{2(k_2-k_1)\alpha n^2}\ \right\}$$

Proof. Let $k_1 \leq k < k_2$ and let $\underline{h}^{(k)}$ be any k-history which extends the history $\underline{h}^{(k_1)}$. Observe that, given $\underline{H}^{(k)} = \underline{h}^{(k)}$, all the list lengths corresponding to the nodes at depth k are determined, say $L_{2^k+i} = l_i$ for $i = 0, 1, \ldots, 2^k - 1$; and further we have $\sum_i l_i^2 \leq \alpha n^2$ since each $l_i \leq \alpha n$ and $\sum_i l_i \leq n$.

Now we consider how the conditional expected value of Q_n given $\underline{H}^{(k)}$ can change when we learn also H_k, the comparisons at level k. Define the random variable T by setting

$$T\ =\ E\left[Q_n \mid \underline{H}^{(k)} = \underline{h}^{(k)}, H_k\right] - E\left[Q_n \mid \underline{H}^{(k)} = \underline{h}^{(k)}\right].$$

Observe that T has the distribution of the sum of 2^k independent random variables T_{l_i} for $i = 0, 1, \ldots, 2^k - 1$, where T_{l_i} is uniformly distributed on the set A_{l_i} (see Lemma 2.2 for the definition of A_n). Hence, by Lemmas 2.2 and 2.4, for any $s > 0$

$$\begin{aligned} E[\exp\{\ sT\ \}] &= \prod_i E[\exp\{\ sT_{l_i}\ \}] \\ &\leq \exp\{\ (4\alpha n^2)s^2/8\ \}\ . \end{aligned}$$

The lemma now follows from Lemma 2.5, with each $c_k^2 = 4\alpha n^2$. □

2.3 Upper bound. We now need only assemble the pieces from §2.1 and §2.2 to give a non-asymptotic upper bound for the probability of a large deviation – this is Lemma 2.7 below – and then choose appropriate values for the parameters to yield the upper bound in Theorem 1.1.

LEMMA 2.7. *Let* n, k_1 *and* s *be positive integers. Then for any real* α *with* $0 < \alpha \leq 1$ *and positive integer* k_2 *such that* $\ln(1/\alpha) \leq k_1$, $k_2 > k_1$, $k_2 \geq \ln(n/2)$ *we have*

$$\begin{aligned} &\Pr[\,|\,Q_n - q_n\,|\ \geq\ k_1 n + s\,]\ \leq \\ &\quad \frac{2}{n}\left(\frac{2e\ln(n/2)}{k_2}\right)^{k_2} + \alpha\left(\frac{2e\ln(1/\alpha)}{k_1}\right)^{k_1} \\ &\quad + 2\exp\left\{\ \frac{-s^2}{2(k_2-k_1)\alpha n^2}\ \right\}. \end{aligned}$$

Proof. Let R_n be the random variable $E\left[Q_n \mid \underline{H}^{(k_2)}\right]$. Recall that a k-history $\underline{h}^{(k_1)}$ determines $M_{k_1}^n$. Let $\mathcal{H}$ be the set of k_1-histories $\underline{h}^{(k_1)}$ such that $M_{k_1}^n \leq \alpha n$. Then

$$\begin{aligned} &\Pr[\,|\,Q_n - q_n\,|\ \geq\ k_1 n + s\,] \\ &\leq\ \Pr[\,Q_n \neq R_n\,]\ +\ \Pr\left[\underline{H}^{(k_1)} \notin \mathcal{H}\right] \\ &\qquad +\ \Pr\left[\,|\,R_n - q_n\,| \geq k_1 n + s \quad\text{and}\quad \underline{H}^{(k_1)} \in \mathcal{H}\right] \\ &=\ \Pr[\,Q_n \neq R_n\,]\ +\ \Pr\left[\underline{H}^{(k_1)} \notin \mathcal{H}\right] \\ &\qquad +\ \sum_{\underline{h}\in\mathcal{H}}\Big(\Pr\left[\,|\,R_n - q_n\,| \geq k_1 n + s \mid \underline{H}^{(k_1)} = \underline{h}\right] \\ &\qquad\qquad \times \Pr\left[\underline{H}^{(k_1)} = \underline{h}\right]\Big) \\ &\leq\ \Pr[\,M_{k_2}^n \geq 2\,]\ +\ \Pr[\,M_{k_1}^n > \alpha n\,] \\ &\qquad +\ \sum_{\underline{h}\in\mathcal{H}}\Pr\left[\left(\left|\ R_n - E\left[R_n \mid \underline{H}^{(k_1)} = \underline{h}\right]\ \right| \geq s\right)\right. \\ &\qquad\qquad \left.\mid\ \underline{H}^{(k_1)} = \underline{h}\right]\ \times\ \Pr\left[\underline{H}^{(k_1)} = \underline{h}\right] \end{aligned}$$

since $\left|\ E\left[R_n \mid \underline{H}^{(k_1)} = \underline{h}\right] - q_n\ \right| \leq k_1 n$ by Lemma 2.3. The result now follows from Lemmas 2.1 and 2.6. □

LEMMA 2.8. *For any* $\varepsilon > 0$

$$\Pr\left[\ \left|\ \frac{Q_n}{q_n} - 1\ \right| > \varepsilon\ \right]\ \leq\ n^{-2\varepsilon(\ln^{(2)} n\, +\, O(\ln^{(4)} n))}\ .$$

Proof. (*Sketch.*) This is where we need to choose parameters appropriately. Let $s = s(n)$ and $k_1 = k_1(n)$ be integers with $s\ =\ n(\ln n)/\ln^{(2)} n\ +\ O(1)$, and $k_1\ =\ 2\varepsilon\ln n - 2s/n\ +\ O(1)$.
Observe that

$$\begin{aligned} k_1 n + s &= 2\varepsilon n\ln n - s\ +\ O(n) \\ &\leq \varepsilon q_n \qquad \text{for } n \text{ sufficiently large,} \end{aligned}$$

and so

$$\Pr\left[\ \left|\ \frac{Q_n}{q_n} - 1\ \right| > \varepsilon\ \right]\ \leq\ \Pr[\,|\,Q_n - q_n\,| \geq k_1 n + s\,]\ .$$

Next let $\alpha = \alpha(n) = (\ln^{(2)} n)^{-5}$ and let $k_2 = k_2(n)$ be an integer with $k_2 = (\ln n)(\ln^{(2)} n) + O(1)$. Now to complete the proof, it remains only to check that each of the three terms in the right hand side of the inequality in Lemma 2.7 is suitably small. □

It was pointed out to us by Joel Spencer that a slightly weaker result may be obtained by considering the levels separately and just using inequalities for independent summands. Another alternative approach still using martingales was pointed out by Harry Kesten.

2.4 The lower bound. In this section we shall prove the lower bound part of Theorem 1.1. We shall see that a few "bad splits" near the top of the partition tree can account for the probability of large deviations.

Lemma 2.9. *Let $\varepsilon > 0$. Then, as $n \to \infty$,*

$$\begin{aligned} &\Pr[\, Q_n > (1+\varepsilon)q_n \,] \\ &\quad \geq \exp\left\{ -2\varepsilon \ln n \left(\ln^{(2)} n + O(\ln^{(3)} n)\right) \right\} . \end{aligned}$$

Proof. (*Sketch.*) Again, we need to choose the appropriate parameters carefully. Let $\mu = \mu(n) = 3\ln^{(3)} n / \ln^{(2)} n$ and let $\kappa = 2+\mu$, $\lambda = \mu/3$ and $\delta = \mu/4$. Note that $\kappa(1-\kappa\lambda/2) \geq 2+\delta$ for n sufficiently large. Now let

$$\begin{aligned} k = k(n) &= \lfloor \kappa\varepsilon \ln n \rfloor , \\ l = l(n) &= \lfloor (\lambda n)/(\varepsilon \ln n) \rfloor , \\ J = J(n) &= \{2^i + 1 : i = 0, 1, \ldots, k-1\} \cup \{2^k\} , \end{aligned}$$

and let $\mathcal{L} = \mathcal{L}(n)$ be the set of vectors $(l_j : j \in J)$ of non-negative integers l_j such that $l_j \leq l$ for each $j \in J \backslash \{2^k\}$.

For each $\underline{l}$ in $\mathcal{L}$ let $A(\underline{l})$ be the event that $L_j = l_j$ for each j in J. Finally let A be the union of the events $A(\underline{l})$ for $\underline{l}$ in $\mathcal{L}$. Now it is routine to show that

$$\begin{aligned} \Pr[\, A \,] &\geq \left(\frac{l+1}{n}\right)^k \\ &= \exp\left\{ -2\varepsilon \ln n(\ln^{(2)} n + O(\ln^{(3)} n)) \right\} . \end{aligned}$$

Let Q'_n be the number of comparisons corresponding to partitioning the leftmost nodes at depth at most $k-1$ (these are the parents of the nodes in the set J) and let $Q''_n = Q_n - Q'_n$. If the event A occurs then

$$\begin{aligned} Q'_n &\geq \sum_{i=0}^{k-1} (n - j(l+1) - 1) \\ &\geq (2+\delta)\varepsilon n \ln n \;+\; O(n) . \end{aligned}$$

Now let $\underline{l} \in \mathcal{L}$ be such that $\Pr[\, A(\underline{l}) \,] > 0$. Observe that $\sum_{j\in J} l_j = n - k$ at least once $k(l+1) < n$. Conditional on $A(\underline{l})$, Q''_n is distributed like $\sum_{j\in J} Q_{l_j}$, where these $k+1$ random variables Q_{l_j} are independent. Hence

$$\begin{aligned} E[Q''_n \mid A(\underline{l})] &= \sum_{j\in J} E[Q_{l_j}] \\ &\geq 2\, l_{2^k} \, \ln l_{2^k} \;+\; 2k\, \hat{l} \ln \hat{l} \;+\; O(n) \end{aligned}$$

where $\hat{l} = \frac{1}{k} \sum_{j \in J\backslash\{2^k\}} l_j$. But now it follows from our choice of parameters that

$$E[Q''_n \mid A(\underline{l})] \;\; \geq \;\; 2n \ln n - (4 + o(1)) n \ln^{(3)} n \, .$$

Also

$$\mathrm{var}\;\, (Q''_n \mid A(\underline{l})) \;=\; \sum_{j\in J} \mathrm{var}\;\, Q_{l_j} \;=\; O(\sum_{j\in J} l_j^2) \;=\; O(n^2) \, .$$

Here we are using the fact that $\mathrm{var}(Q_n) = O(n^2)$, as noted in §1. Hence, by Chebyshev's inequality

$$\Pr\left[Q''_n \geq 2n \ln n - 5n \ln^{(3)} n \mid A(\underline{l}) \right] \to 1 \;\text{ as }\; n \to \infty \, ,$$

and this convergence is uniform over $\underline{l}$ in $\mathcal{L}$.

Now, for n sufficiently large,

$$(2+\delta)\varepsilon n \ln n + 2n \ln n - 5n \ln^{(3)} n \;>\; (1+\varepsilon) q_n \, .$$

Hence we have

$$\begin{aligned} &\Pr[\, Q_n > (1+\varepsilon)q_n \,] \\ &\quad \geq \Pr[\, Q_n > (1+\varepsilon)q_n \mid A \,] \, \Pr[\, A \,] \\ &\quad \geq \sum_{\underline{l}\in\mathcal{L}} \Pr\left[Q''_n \geq 2n \ln n - 5n \ln^{(3)} n \mid A(\underline{l}) \right] \Pr[\, A(\underline{l}) \,] \\ &\quad = \exp\left\{ -2\varepsilon \ln n(\ln^{(2)} n + O(\ln^{(3)} n) \right\} \end{aligned}$$

as required. This completes the (fairly detailed!) sketch of the proof of Lemma 2.9. □

3 Median-of-$(2t+1)$ quicksort

In this section we outline the proof of Theorem 1.2. As many of the ideas and techniques of this section are similar to those explained in the previous section, we shall omit most proofs.

3.1 List lengths in the partition tree. We argue much as in §2.1 though the details are more complicated here.

LEMMA 3.1. *Let n and k be positive integers, let $0 < \alpha < 1$ and suppose that $(2t+1)2tk < \alpha n$.*

Let $X_1, X_2, \ldots, X_k$ be independent random variables each distributed as the median of $2t+1$ independent random variables uniformly distributed on [0,1]. Then

$$\Pr[\, M_k^n \geq \alpha n \,] \leq 2^k \left(1 - \frac{(2t+1)2tk}{\alpha n}\right)^{-1} \Pr\left[\prod_{i=1}^{k} X_i \geq \alpha\right].$$

Proof. (*Sketch.*) Let $U_j^{(i)}$ for positive integers i and j be independent random variables each uniformly distributed on [0,1]. For each $i = 1, \ldots, k$ we may take X_i to be the median of $U_1^{(i)}, \ldots, U_{2t+1}^{(i)}$.

Next we define a decreasing sequence N_0, N_1, …, N_k of random variables corresponding to the list lengths L_0, L_2, …, L_{2^k}. Let $N_0 = n$. For each $i = 1, \ldots, k$ do the following. If $N_{i-1} < \alpha n$ then set $N_i = \ldots = N_k = 0$ and stop. If $N_{i-1} \geq \alpha n$ then consider $U_1^{(i)}, U_2^{(i)}, \ldots$ in turn until we obtain $2t+1$ distinct numbers $\lfloor U_j^{(i)} N_{i-1} \rfloor$. Then let N_i be one less than the median of these $2t+1$ numbers. The key observation is that

$$\Pr[\, L_{2^k} \geq \alpha n \,] = \Pr[\, N_k \geq \alpha n \,].$$

Let A be the event that for each $i = 1, \ldots, k$ with $N_{i-1} \geq \alpha n$ the first $2t+1$ numbers $\lfloor U_1^{(i)} N_{i-1} \rfloor, \ldots, U_{2t+1}^{(i)} N_{i-1} \rfloor$ are distinct. Then

$$\Pr[\, N_k \geq \alpha n \text{ and } A \,] \leq \Pr\left[\prod_{i=1}^{k} X_i \geq \alpha\right]$$

since clearly $N_k \leq n \prod_{i=1}^{k} X_i$ on A. Also

$$\begin{aligned}\Pr[\, A \mid N_k \geq \alpha n \,] &\geq \left(1 - \binom{2t+1}{2} \frac{2}{\alpha n}\right)^k \\ &\geq 1 - \frac{(2t+1)(2t)k}{\alpha n}.\end{aligned}$$

Now the desired conclusion follows from routine probability inequalities. □

LEMMA 3.2. *Let t be a non-negative integer. Let $0 < \alpha < 1$ and let n and k be positive integers such that*

$$k > \frac{2t+1}{t+1} \ln \frac{1}{\alpha}.$$

Let X be distributed like the median of $2t+1$ random variables each uniform on [0,1], and let $X_1, \ldots, X_k$ be independent random variables, each distributed like X. Then

$$\Pr\left[\prod_{i=1}^{k} X_i \geq \alpha\right] \leq \alpha^{2t+1} \left(\frac{(2t+1)e \ln(1/\alpha)}{(t+1)k}\right)^{(t+1)k}.$$

Proof. (*Sketch.*) It is well known and routine to check that X has the β distribution with parameters $t+1$, $t+1$; which has probability density function

$$f(x) = \frac{\Gamma(2t+2)}{(\Gamma(t+1))^2} x^t (1-x)^t \quad \text{for } 0 < x < 1.$$

Thus, for $s > -(t+1)$, it follows from straightforward integration that

$$E[X^s] = \prod_{i=0}^{t} \left(\frac{2t+1-i}{s+2t+1-i}\right).$$

Now it is not hard to show that for any $s > 0$

$$\Pr\left[\prod_{i=1}^{k} X_i \geq \alpha\right] \leq \alpha^{-s} \left(\frac{2t+1}{s+2t+1}\right)^{(t+1)k}.$$

Choosing $s > 0$ to minimise this bound yields the desired result. □

3.2 The bounded differences approach.

Our first result here is the key property for median-of-$(2t+1)$ quicksort and corresponds to Lemma 2.2 for basic quicksort. That result was non-asymptotic and proved from first principles: Here we are not so lucky.

LEMMA 3.3. *Let t be a positive integer. For each positive integer n let* $A_n = \{n - 1 + q_{k-1}^{(t)} + q_{n-k}^{(t)} - q_n^{(t)} : k = 1, 2, \ldots, n\}$. *Then there is an $\eta > 0$ and a function $g(n) > 0$ with $g(n) = O(n^{1-\eta})$ such that for each positive integer n*

$$-(\kappa_t \ln 2 - 1)\, n - g(n) < x < n + g(n) \text{ for all } x \in A_n.$$

Proof. It is known (see [Hen89]) that for some constants $\beta = \beta(t)$ and $\eta = \eta(t)$ with $0 < \eta < 1$ we have

$$q_n^{(t)} = \kappa_t n \ln n + \beta n + \gamma(n),$$

where $\gamma(n) = O(n^{1-\eta})$. Suppose that $|\,\gamma(n)\,| \leq c n^{1-\gamma}$. Let

$$f(n) = \kappa_t n \ln n + \beta n.$$

It is easy to check that

$$f(k-1) + f(n-k) - f(n) \begin{cases} \leq O(1) \\ \geq -\kappa_t (n \ln 2 + \ln n) + O(1). \end{cases}$$

It follows that

$$q_{k-1}^{(t)} + q_{n-k}^{(t)} - q_n^{(t)} \begin{cases} \leq 3 \max_{1 \leq k \leq n} |\,\gamma(k)\,| + O(1) \\ \geq -\kappa_t (n \ln 2 + \ln n) \\ \qquad -3 \max_{1 \leq k \leq n} |\,\gamma(k)\,| + O(1). \end{cases}$$

Thus there is a suitable function $g(n)$ with

$$g(n) = 3cn^{1-\gamma} + 2\ln n + O(1). \qquad \square$$

Consider the variant of median of (2t+1) quicksort which cuts lists at length $\ln(n)$ say. Let $\tilde{Q}_n$ be the corresponding number of comparisons and let $\tilde{q}_n = E\left[\tilde{Q}_n\right]$. Note that

$$\left| Q_n^{(t)} - \tilde{Q}_n \right| = O(n \ln^{(2)} n).$$

LEMMA 3.4. *There is an $\eta > 0$ such that the following holds. Let n and k be positive integers and let $\underline{h}$ be a possible k-history for $\tilde{Q}_n$. Then*

$$\left| E\left[\tilde{Q}_n \mid \underline{H}^{(k)} = \underline{h}\right] - \tilde{q}_n \right| \leq \left(1 + (\ln n)^{-\eta}\right) kn.$$

Proof. This follows from Lemma 3.1 by induction on k. Note that $\kappa_t \ln 2 \leq 2\ln 2 < 2$ for each t. $\square$

3.3 An upper bound.

LEMMA 3.5. *Let $\varepsilon > 0$. Then*

$$\Pr\left[\left| \frac{Q_n^{(t)}}{q_n^{(t)}} - 1 \right| > \varepsilon\right] \leq \exp\left\{ -(t+1)\kappa_t \varepsilon \ln n \left(\ln^{(2)} n + O(\ln^{(4)} n)\right) \right\}.$$

Proof. (*Sketch.*) Consider $\tilde{Q}_n$ as in Lemma 3.4 above. Define s, k_1, α, k_2, R_n, $\mathcal{H}$ as in the proof of Lemma 2.8, except with the term $2n\ln n$ in the definition of k_1 replaced by $\kappa_t\ n\ \ln n$. Now the proof of this lemma is similar to the proofs of Lemmas 2.7 and 2.8. We omit the details. $\square$

3.4 A lower bound. This may be proved just as for the basic case, and is left to the reader.

4 Simulations

Figure 1 shows the frequency distribution obtained when 30,000 trials of quicksort were performed with $n = 30,000$.

Figure 2 shows how the bound established in Theorem 1.1 compares with actual trials of quicksort for a specific value of ε, namely $\varepsilon = 0.1$. In particular, 1000 trials of quicksort were performed for $n = 1000, 1500, 2000, 2500, \ldots, 18000$. For each set of 1000 trials, the empirical value $\Pr\left[\left|\frac{Q_n}{q_n} - 1\right| > \varepsilon\right]$ and the quantity $n^{-2\varepsilon(\ln\ln n - c)}$ are shown, for the value $c = 0.3$ (fitted to the data).

The simulations were performed on a cluster of Suns and Sparcs.

References

[Dev86] L. Devroye. A note on the height of binary search trees. *Journal of the ACM*, 33(3):489–498, July 1986.

[Hen89] P. Hennequin. Combinatorial analysis of quicksort algorithm. *Theoretical Informatics and Applications*, 23(3):317–333, 1989.

[Hoa61] C.A.R. Hoare. Partition (algorithm 63), quicksort (algorithm 64), and find (algorithm 65). *J. ACM*, 7:321–322, 1961.

[Hoa62] C.A.R. Hoare. Quicksort. *Computer J.*, 5:10–15, 1962.

[Knu73] Donald Knuth. *The Art of Computer Programming.* Addison-Wesley, first edition, 1973.

[McD89] Colin McDiarmid. On the method of bounded differences. In J Siemons, editor, *Surveys in Combinatorics.* London Math Society, 1989.

[Rég89] Mireille Régnier. A limiting distribution for quicksort. *Theoretical Informatics and Applications*, 23(3):335–343, 1989.

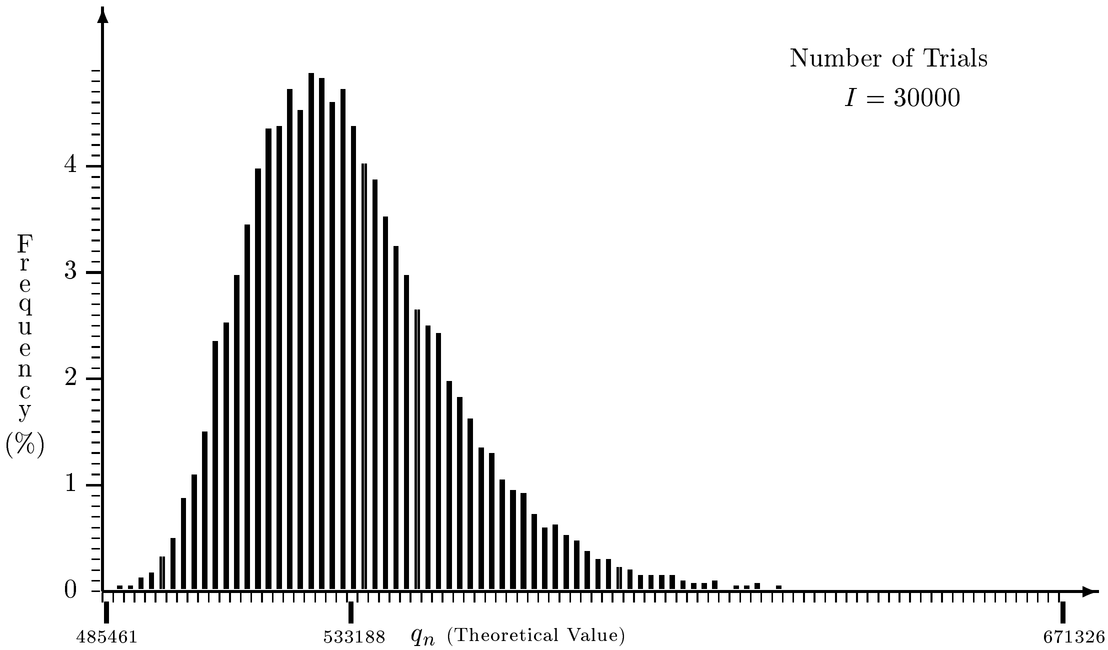
[Rös] Uwe Rösler. A limit theorem for quicksort. manuscript.

[Sed80] Robert Sedgewick. *Quicksort.* Garland Publishing Inc., first edition, 1980.

[Van70] M. H. VanEmden. Increasing the efficiency of quicksort. *Communications of the ACM*, 13(9):563–567, September 1970.

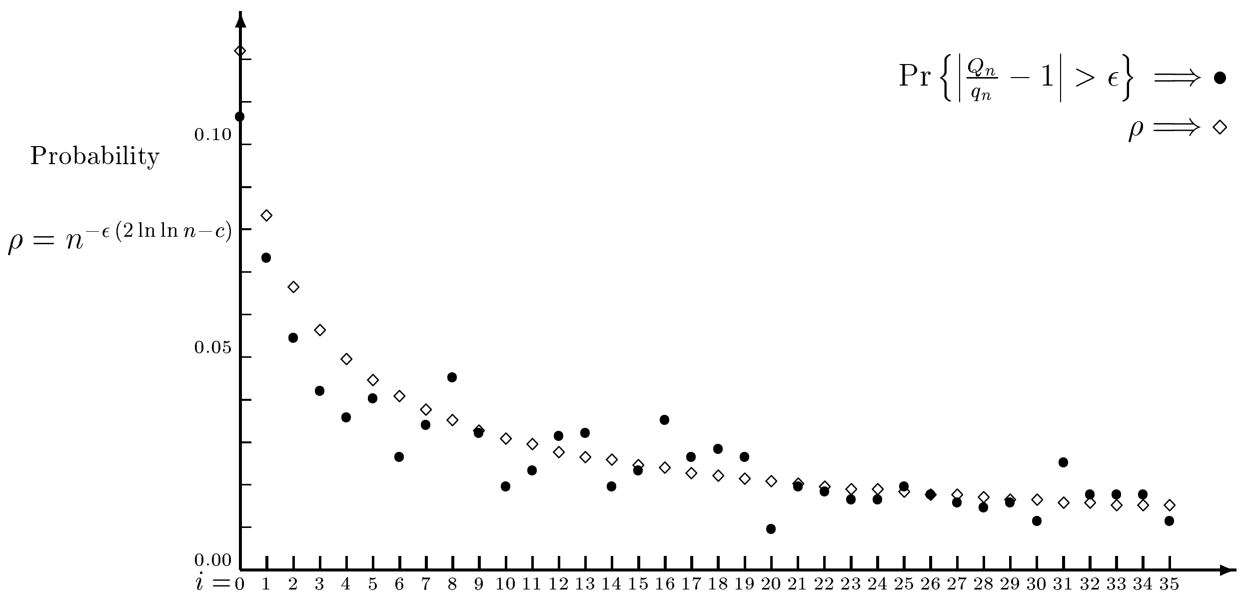
Figure 1.

A Simulation of Basic Quicksort:
Frequency Distribution of Actual
Number of Comparisons that Occurred



Actual Number of Comparisons (Q_n)
Size of Each Simulation: $n = 30000$ Each bar corresponds to a range of 2065

Figure 2.
Basic Quicksort: Empirical and Theoretical Probability that
$\Pr\left\{\left|\frac{Q_n}{q_n} - 1\right| > \epsilon\right\}$ **vs.** n **(# of Keys in each Simulation)**



n (The # of Keys in each Simulation)

$n = (i * 500) + 500$
Number of trials = 1000 (for each n)
$\epsilon = 0.100$
$c = 0.300$