JOURNAL OF COMPUTER AND SYSTEM SCIENCES 29, 1–7 (1984)

Exponential Bounds for the Running Time of a Selection Algorithm

LUC DEVROYE

School of Computer Science, McGill University, 805 Sherbrooke Street West, Montreal, Quebec H3A 2K6, Canada

Received March 20, 1982; revised May 4, 1983

Hoare's selection algorithm for finding the kth-largest element in a set of n elements is shown to use C comparisons where

(i) $E(C^p) \leqslant A_p n^p$ for some constant $A_p > 0$ and all $p \geqslant 1$;

(ii) $P(C/n \geqslant u) \leqslant (\frac{3}{4})^{u(1+o(1))}$ as $u \to \infty$.

Exact values for the "A_p" and "$o(1)$" terms are given.

1. INTRODUCTION

Hoare [7], Aho *et al.* [1, pp. 101–102] and Horowitz and Sahni [9] all consider the following algorithm (with minor modifications) for finding the kth-smallest element in a set S of n elements ($1 \leqslant k \leqslant n$):

procedure FIND (k, S)
if $|S| = 1$ *then return* the single element in S
  else begin choose an element a randomly from S;
    let S_1, S_2 and S_3 be the sequences of elements in S less than, equal to, and greater than a, respectively;
    if $|S_1| \geqslant k$, *then return* FIND (k, S_1)
    else if $|S_1| + |S_2| \geqslant k$ *then return* a
      else return FIND $(k - |S_1| - |S_2|, S_3)$
end.

A nonrecursive version of this algorithm is of course easy to find. The work done here can be measured by the number of comparisons between elements (these occur only in the step in which S is split into S_1, S_2 and S_3). It is known that this algorithm requires $\Omega(n^2)$ comparisons in the worst case. The algorithm runs in average time $O(n)$ (Aho *et al.* [1]). In fact, Knuth [10] has shown that the average number of comparisons is at most

$$2((n+1)H_n - (n+3-i)H_{n-i+1} - (i+2)H_i + n + 3)$$

0022-0000/84 $3.00

where $H_n = \sum_{1 \leqslant j \leqslant n}(1/j)$. Thus, for $k = \lceil n/2 \rceil$, we obtain the bound $2(1+\ln(2))n + o(n) \leqslant 3.39n + o(n)$. For $O(n)$ worst-case selection algorithms, see Blum *et al.* [2] or Schonhage *et al.* [11].

In this paper we give probabilistic bounds for the upper tail of C, the number of comparisons used by FIND. The bounds are reasonably tight, but more importantly, the exponential nature of the bounds shows that deviations from linearity are extremely unlikely. We do not want to challenge the fact that the algorithm of Floyd and Rivest [5, 6] is faster on the average than FIND (it was shown there that the expected number of comparisons is $n + \min(k, n-k) + O(n^{1/2})$). The analysis given here for FIND is ad hoc, and therefore not directly extendible to the Floyd–Rivest algorithm.

Result 1. There exists a random variable T, independent of n and k, such that

$$C \prec nT$$

where $\prec$ denotes "is stochastically smaller than," i.e.,

$$P(C \geqslant u) \leqslant P(nT \geqslant u), \qquad \text{all } u.$$

The random variable T satisfies $E(T^p) < \infty$ for all $p \geqslant 1$.

Result 2.

$$E(C) \leqslant 4n;$$

$$E(C^p) \leqslant A_p n^p, \qquad \text{all integer } p \geqslant 1,$$

where

$$A_p = \frac{16}{3} \frac{p!}{\ln^{p-1}(\frac{4}{3})}.$$

Result 3.

$$P(C/n \geqslant u) \leqslant (1 + Au)\, e(\tfrac{3}{4})^u, \qquad u \geqslant 1/\ln(\tfrac{4}{3}),$$

and

$$P(C/n \geqslant u) \leqslant (\tfrac{3}{4})^{(\sqrt{u} - \sqrt{16/3})^2}, \qquad u \geqslant 16/3,$$

where

$$A = (16/3)\ln^2(\tfrac{4}{3}).$$

2. Analysis

We can and will assume that all elements in S are distinct. We claim that C is stochastically smaller than the outcome of the following algorithm. S, k and n are as defined in the Introduction.

$l \leftarrow 0$, $r \leftarrow n+1$, $C \leftarrow 0$. (C will be the outcome of the algorithm.)
while $r > l+1$ *do*
  begin generate N uniformly and at random in $\{l+1, \ldots, r-1\}$;
    $C \leftarrow C + (r-l-2)$;
    if $N < k$, *then* $(l, r) \leftarrow (N, r)$ *else* $(l, r) \leftarrow (l, N)$
end.

Thus we can define C as the outcome of this algorithm, since we are only interested in upper bounds for C. In our proof, we will construct a probability space in the following manner. Let (U_1, V_1), $(U_2, V_2), \ldots$ be a sequence of independent uniform $[0, 1]^2$ random vectors. We will use the notation (l_i, r_i) for the values of (l, r) in the ith iteration. In particular, $(l_0, r_0) = (0, n+1)$. Our construction is such that the distribution of (l_i, r_i) is completely determined by (U_j, V_j), $j \leqslant i$.

Let (l_{i-1}, r_{i-1}) be given. Then (l_i, r_i) is determined as follows:

$$(l_i, r_i) = \begin{cases} (l_{i-1}, r_{i-1} - 1 - \lfloor (r_{i-1} - k) U_i \rfloor) & \\ \qquad \text{if } V_i < p_{i-1} = (r_{i-1} - k)/(r_{i-1} - l_{i-1} - 1) & \\ \qquad \text{(an event that we shall call } A_i\text{);} & \\ (l_{i-1} + 1 + \lfloor (k - 1 - l_{i-1}) U_i \rfloor, r_{i-1}) & \text{otherwise.} \end{cases}$$

Thus, on A_i, r_i is uniformly distributed on $\{k, \ldots, r_{i-1} - 1\}$, and on its complement, l_i is uniformly distributed on $\{l_{i-1} + 1, \ldots, k-1\}$. Thus, on A_i,

$$\begin{aligned} r_i - l_i - 2 &= r_{i-1} - l_{i-1} - 3 - \lfloor (r_{i-1} - k) U_i \rfloor \\ &\leqslant r_{i-1} - l_{i-1} - 2 - (r_{i-1} - k) U_i \\ &= r_{i-1} - l_{i-1} - 2 - p_{i-1} U_i (r_{i-1} - l_{i-1} - 1), \end{aligned}$$

and on A_i^c, the complement of A_i,

$$\begin{aligned} r_i - l_i - 2 &= r_{i-1} - l_{i-1} - 3 - \lfloor (k - 1 - l_{i-1}) U_i \rfloor \\ &\leqslant r_{i-1} - l_{i-1} - 2 - (k - 1 - l_{i-1}) U_i \\ &= r_{i-1} - l_{i-1} - 2 - (1 - p_{i-1}) U_i (r_{i-1} - l_{i-1} - 1). \end{aligned}$$

Combining this, and using I to denote the indicator function of an event, gives

$$\begin{aligned} r_i - l_i - 2 &\leqslant (r_{i-1} - l_{i-1} - 2)(1 - U_i(p_{i-1} I_{A_i} + (1 - p_{i-1}) I_{A_i^c})) \\ &= (r_{i-1} - l_{i-1} - 2) W_i \qquad \text{(definition of } W_i\text{).} \end{aligned} \tag{1}$$

Inequality (1) is the starting point for all further analysis. Clearly, by recursion,

$$r_i - l_i - 2 \leqslant (n-1) \prod_{j=1}^{i} W_j \leqslant n \prod_{j=1}^{i} W_j \tag{2}$$

and

$$C \leqslant n\left(1+\sum_{i=1}^{\infty}\prod_{j=1}^{i} W_j\right). \tag{3}$$

We will use two Lemmas.

LEMMA 1. *If U is a uniform $[0,1]$ random variable, then $E((1-U/2)^p) < 2/(p+1)$, all $p \geqslant 1$.*

Proof. $E((1-U/2)^p) = \int_0^1 (1-u/2)^p\, du = (2/(p+1))(1-2^{-(p+1)}) < 2/(p+1)$.

LEMMA 2. *Let $W_1, W_2, \ldots$ be a sequence of independent identically distributed nonnegative random variables with* pth *moment $E(W_1^p) = \mu < 1$, and let $X = 1 + \sum_{j=1}^{\infty}\prod_{i=1}^{j} W_i$. For $p \geqslant 1$,*

$$E(X^p) \leqslant 1/(1-\mu^{1/p})^p. \tag{4}$$

For $p=1$, equality is achieved in (4).

Proof. Whenever we have a random variable X that can be written as $\sum_{i=0}^{\infty} X_i$, then for all $\lambda \in (0,1)$,

$$X = \sum_{i=0}^{\infty} \lambda^i(1-\lambda)\frac{X_i}{\lambda^i(1-\lambda)}$$

so that by Jensen's inequality,

$$X^p \leqslant \sum_{i=0}^{\infty} \lambda^i(1-\lambda)\left(\frac{X_i}{\lambda^i(1-\lambda)}\right)^p = (1-\lambda)^{1-p}\sum_{i=0}^{\infty} \lambda^{i(1-p)} X_i^p.$$

If we replace X_i by 1 for $i=0$ and by $W_1 W_2 \cdots W_i$ for $i \neq 0$, and if we note that $E(X_i^p) = \mu^i$, then

$$E(X^p) \leqslant \sum_{i=0}^{\infty} (\mu/\lambda^{p-1})^i (1-\lambda)^{1-p} = (1-\lambda)^{1-p}/(1-\mu/\lambda^{p-1}), \qquad \mu < \lambda^{p-1}.$$

The last expression is minimal for $\lambda = \mu^{1/p}$. Resubstitution gives the bound

$$(1-\mu^{1/p})^{1-p}/(1-\mu^{1-(p-1)/p}) = (1-\mu^{1/p})^{-p}.$$

Proof of Result 1. It is clear that $p_{i-1} I_{A_i} + (1-p_{i-1}) I_{A_i^c} > \frac{1}{2} Z_i$, where Z_i is Bernoulli with parameter $\frac{1}{2}$ (note that $A_i = [V_i < p_{i-1}]$). Thus, if $Z_1, Z_2, \ldots$ are independent Bernoulli $(\frac{1}{2})$ random variables,

$$C < n\left(1+\sum_{i=1}^{\infty}\prod_{j=1}^{i} W_j^*\right) = nT \qquad \text{(definition of } T) \tag{5}$$

where $W_j^* = 1 - \frac{1}{2}Z_j U_j$. Thus, all W_j^*'s are independent and identically distributed. Also, $E(T) = 1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} E(W_j^*) = \sum_{i=0}^{\infty} (\frac{7}{8})^i = 8 < \infty$ so that the right-hand side of (5) is indeed almost surely finite. By Lemma 1,

$$E(W_1^{*p}) = \frac{1}{2}\left(1 + E\left(1 - \frac{U_1}{2}\right)^p\right) \leqslant \frac{1}{2}\left(1 + \frac{2}{p+1}\right) = \frac{1}{2}\frac{p+3}{p+1}.$$

Thus, by Lemma 2, for $p > 1$,

$$E(T^p) \leqslant 1 \Big/ \left(1 - \left(\frac{1}{2}\frac{p+3}{p+1}\right)^{1/p}\right)^p < \infty.$$

Remark 1. The stochastic majorization used in this proof is sloppy. It gives the crude estimate $E(C) \leqslant 8n$. With the more refined majorization

$$p_{i-1} I_{A_i} + (1 - p_{i-1}) I_{A_i^c} > \tfrac{1}{2}(Z_i + (1 - Z_i) V_i)$$

where V_i is uniform $[0, 1]$ and independent of Z_i, and with $W_j^* = 1 - \frac{1}{2}(Z_j + (1 - Z_j)V_j) U_j$ in (5), we obtain the slightly sharper result

$$\frac{E(C)}{n} \leqslant \sum_{i=0}^{\infty} \left(\frac{13}{16}\right)^i = \frac{16}{3}.$$

Proof of Result 2. We take (3) as our starting point, and let $\mathfrak{F}_i$ be the σ-algebra generated by $(U_1, V_1), \ldots, (U_i, V_i)$. $\mathfrak{F}_0$ is the σ-algebra consisting of the empty set and its complement. By well-known properties of conditional expectations (see, e.g., Chow and Teicher [4]),

$$E(W_i \mid \mathfrak{F}_{i-1}) = 1 - \tfrac{1}{2}(p_{i-1}^2 + (1 - p_{i-1})^2) \leqslant 1 - \tfrac{1}{4} = \tfrac{3}{4},$$

and

$$E\left(\prod_{i=1}^{j} W_i\right) = E\left(\prod_{i=1}^{j} E(W_i \mid \mathfrak{F}_{i-1})\right) \leqslant \left(\frac{3}{4}\right)^j, \qquad j \geqslant 1.$$

Since obviously $0 \leqslant W_i \leqslant 1$, we have for $p \geqslant 1$, p integer,

$$\begin{aligned}
\frac{E(C^p)}{n^p} &\leqslant E\left(\left(1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} W_j\right)^p\right) = \sum_{(j_1, \ldots, j_p) \in \{0,1,2,\ldots\}^p} E\left(\prod_{m=1}^{p} \prod_{i \leqslant j_m} W_i\right) \\
&\leqslant \sum_{(j_1, \ldots, j_p) \in \{0,1,2,\ldots\}^p} E\left(\prod_{i \leqslant \max(j_1, \ldots, j_p)} W_i\right) \\
&\leqslant \sum_{(j_1, \ldots, j_p) \in \{0,1,2,\ldots\}^p} \left(\frac{3}{4}\right)^{\max(j_1, \ldots, j_p)} \\
&\leqslant p \sum_{j=0}^{\infty} (j+1)^{p-1} \left(\frac{3}{4}\right)^j = \frac{16}{3} p \sum_{j=0}^{\infty} j^{p-1} \frac{1}{4}\left(\frac{3}{4}\right)^j = \frac{16}{3} pE(X^{p-1}) \qquad (6)
\end{aligned}$$

where X is geometrically distributed: $P(X=j)=\frac{1}{4}(\frac{3}{4})^j$, $j \geqslant 0$. We note that X is distributed as the integer part of $X^*/\ln(\frac{4}{3})$, where X^* is exponentially distributed (i.e., has density e^{-x} on $[0, \infty)$). Thus (6) is bounded from above by

$$\frac{16}{3}\, pE(X^{*p-1})/\ln^{p-1}\left(\frac{4}{3}\right) = \frac{16}{3}\, p(p-1)!/\ln^{p-1}\left(\frac{4}{3}\right) = \frac{16}{3}\, p!/\ln^{p-1}\left(\frac{4}{3}\right).$$

The well-known result $E(C) \leqslant 4n$ follows easily:

$$\frac{E(C)}{n} \leqslant 1 + \sum_{j=1}^{\infty} \prod_{i=1}^{j} E(W_i) \leqslant \sum_{j=0}^{\infty} \left(\frac{3}{4}\right)^j = 4.$$

Proof of Result 3. We start from Result 2. Let t be a real number in $(0, \ln(\frac{4}{3}))$, and let T be C/n. By Result 2,

$$E(e^{tT}) = \sum_{i=0}^{\infty} \frac{t^i}{i!} E(T^i) \leqslant 1 + \frac{16}{3} \sum_{i=1}^{\infty} \frac{t^i}{\ln^{i-1}(\frac{4}{3})} = 1 + \frac{16t}{3}\left(1 - \frac{t}{\ln(\frac{4}{3})}\right)^{-1}. \qquad (7)$$

Thus, by the Bernstein–Chernoff bounding method (see Chernoff [3] or Hoeffding [8]),

$$\begin{aligned} P(T \geqslant u) &\leqslant E(e^{tT})\, e^{-tu} \\ &\leqslant \left(1 + \frac{16t}{3}\left(1 - t/\ln\left(\frac{4}{3}\right)\right)^{-1}\right) e^{-tu}, \qquad 0 < t < \ln\left(\frac{4}{3}\right). \end{aligned} \qquad (8)$$

Result 3 now follows by choosing t carefully. For the first inequality, we take a positive number c, and assume that $u > c/\ln(\frac{4}{3})$, $t = \ln(\frac{4}{3}) - c/u$. The last expression is not greater than

$$(1 + au/c)\, e^c (\tfrac{3}{4})^u$$

where $a = 16\ln^2(\frac{4}{3})/3$. Considered as a function of c, the latter expression is minimal when $c^2 + auc - au = 0$, i.e., when $c = (au/2)(\sqrt{1 + 4/au} - 1) \sim 1$ as $au \to \infty$. Thus the value $c = 1$ is best for large u. This leads to the upper bound

$$(1 + au)\, e(\tfrac{3}{4})^u, \qquad \text{valid for } u > 1/\ln(\tfrac{4}{3}).$$

For the second inequality of result 3, we apply the inequality $1 + u \leqslant e^u$ to (8), and obtain the inequality

$$P(T \geqslant u) \leqslant \exp(-tu + (16t/3)(1 - t/\ln(\tfrac{4}{3}))^{-1}), \qquad 0 < t < \ln(\tfrac{4}{3}), \qquad (9)$$

which has the form $\exp(-tu + at/(1 - bt))$. Such an expression is minimal when $t = (1 - \sqrt{a/u})(1/b)$. Replacement of this value of t in (9) shows that

$$P(T \geqslant u) \leqslant \exp(-(\sqrt{u} - \sqrt{a})^2/b)$$

where $a = 16/3$ and $b = 1/\ln(\frac{4}{3})$. The last inequality is valid for all $u \geqslant a$. This concludes the proof of Result 3.

REFERENCES

1. A .V. AHO, J. E. HOPCROFT, AND J. D. ULLMAN, "The Design and Analysis of Computer Algorithms," Addison–Wesley, Reading, Mass., 1974.
2. M. BLUM, R. W. FLOYD, V. PRATT, R. L. RIVEST, AND R. E. TARJAN, Time bounds for selection, *J. Comput. System Sci.* **7** (1973), 448–461.
3. H. CHERNOFF, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of obervations, *Ann. Math. Statist.* **23** (1952), 493–507.
4. Y. S. CHOW AND H. TEICHER, "Probability Theory," Springer-Verlag, New York/Berlin, 1978.
5. R. W. FLOYD AND R. L. RIVEST, Expected time bounds for selection, *Comm. ACM* **18** (1975), 165–172.
6. R. W. FLOYD AND R. L. RIVEST, Algorithm 489, *Comm. ACM* **18** (1975), 173.
7. C. A. R. HOARE, Find (algorithm 65), *Comm. ACM* **4** (1961), 321–322.
8. W. HOEFFDING, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58** (1963), 13–30.
9. E. HOROWITZ AND S. SAHNI, "Fundamentals of Computer Algorithms," Computer Science Press, Potomac, Md, 1978.
10. D. E. KNUTH, "Mathematical Analysis of Algorithms," Computer Science Dept. Report STAN-CS-71-206, Stanford University, 1971.
11. A. SCHONHAGE, M. PATERSON, AND N. PIPPENGER, Finding the median, *J. Comput. System Sci.* **13** (1976), 184–199.