ON THE PROBABILISTIC WORST-CASE TIME OF "FIND"

Luc Devroye
luc@cs.mcgill.ca

ABSTRACT. We analyze the worst-case number of comparisons T_n of Hoare's selection algorithm FIND when the input is a random permutation, and worst case is measured with respect to the rank k. We give a new short proof that T_n/n tends to a limit distribution, and provide new bounds for the limiting distribution.

KEYWORDS AND PHRASES. Binary search tree, selection problem, order statistics, FIND, expected time analysis.

CR CATEGORIES: 3.74, 5.25, 5.5.

Research of the first author was sponsored by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. Address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6.

Hoare's algorithm FIND.

Most textbooks on algorithms and data structures mention Hoare's algorithm FIND (Hoare, 1961) for the selection of the k-th smallest from a set of n pairwise different elements: one grabs a pivot uniformly at random from the available elements, and compares the $n-1$ remaining elements against it. If the rank of the pivot is ℓ, and $\ell = k$, we return the pivot itself. If $\ell < k$, we recursively return the k-th smallest of the set of smaller elements. If $\ell > k$, we return the $(\ell - k)$-th smallest from the set of larger elements. The randomization is done by initially randomly permuting the data, and keeping the elements in that random order throughout. The pivots are the first elements in the sets searched. Denote the number of comparisons by $T_{n,k}$. This process may be visualized through a binary search tree that is constructed by consecutive insertions of the elements of the random permutation. To find the k-th order statistic, start at the root (the first pivot), assign a cost of $n-1$, go the appropriate subtree, and recurse, always assigning a cost equal to the size of the subtree of a node, minus one. The worst-case cost for this fixed random tree (with respect to k) is obtained by taking that path down the tree that yields the largest sum of individual costs. This way of looking at the problem is explored in the paper.

Many results are known about $T_{n,k}$. In particular, if $k \sim xn$ as $n \to \infty$ and $x \in (0,1)$ is fixed, then $\mathbb{E}\{T_{n,k}\}/n \to 2 - 2t\log t - 2(1-t)\log(1-t)$ (Grübel and Rösler, 1996). This can also easily be derived from a 1972 formula of Knuth,

$$\mathbb{E}\{T_{n,k}\} = 2(n+3+(n+1)H_n - (k+2)H_k - (n+3-k)H_{n+1-k} \ ,$$

where H_n denotes the n-th harmonic number. Furthermore, $T_{n,k}/n \xrightarrow{\mathcal{L}} W(x)$, where $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution, and where the law of $W(x)$ is described in the work of Grübel and Rösler (1996). We also know that

$$\sup_{1\le k\le n} \mathbb{P}\{T_{n,k} \ge tn\} \le C\rho^t$$

for any $\rho > 3/4$ and some constant $C(\rho)$ (Devroye, 1984), so that $T_{n,k}$ is indeed linear in n in a very strong uniform sense. This paper will strengthen that belief. We denote the worst-case time by

$$T_n = \max_{1\le k\le n} T_{n,k} \ .$$

The purpose of this note is to give a short proof of the following.

Theorem 1 (theorem 12 of grübel and rösler, 1996). *We have*

$$\frac{T_n}{n} \xrightarrow{\mathcal{L}} S$$

where S is a random variable supported on $[2,\infty)$, and whose distribution is uniquely described by the sole (proper random variable) solution of the distributional identity

$$S \stackrel{\mathcal{L}}{=} \max(US', (1-U)S'') + 1 \ .$$

Here $\stackrel{\mathcal{L}}{=}$ means identity in distribution, U is a uniform $[0,1]$ random variable, and S', and S'' are distributed as S, and U, S' and S'' are independent.

The random variable S is the supremum of a process also studied by Grübel and Rösler (1996), whose work yields Theorem 1 with trivial modifications. The focus of the work of Grübel and Rösler was the fixed-point method for identifying the limit distribution of the complexity of FIND when one looks for the $\lfloor xn \rfloor$-th order statistic with $x \in (0,1)$ fixed. The fixed point method of analysis was pioneered by Rösler (1991, 1992), and surveyed by Rösler and Rüschendorf (1999), and yielded a characterization of the limit distribution of the complexities of quicksort (Rösler, 1991), FIND (Rösler, 1997), partial match in k-d trees (Neininger, 2000), partial match in quadtrees (Neininger and Rüschendorf, 1999), the internal path length in quadtrees (Neininger and Rüschendorf, 1999), and many other algorithms. Grübel (1997) provides a Markov chain alternative to the analysis of FIND. In the present paper, we merely point out that by using an embedding technique for random binary search trees, the limiting process can be obtained rather routinely by monotone convergence, and that the study of S (but not that of $T_{n,\lfloor xn \rfloor}$) and the proof of Theorem 1 are in fact rather straightforward. This provides an alternative shorter path to Theorem 12 of Grübel and Rösler (1996). Pruned trees (Lent and Mahmoud, 1996) form yet a different way of viewing things. In the last section, we provide more information on the distribution of S and show that it has a density, is Lipschitz continuous, and we also provide bounds on all moments, as well as explicit exponential and asymptotic superexponential tail bounds.

Explanation via binary search trees

We could have solved this problem in a number of ways, but the route followed here is perhaps the most intuitive one. From the random permutation, we construct a random binary search tree by standard insertion. This tree may be used to explain or visualize the complexity of FIND. Indeed, if the k-th smallest element is found by following the path $u_0, u_1, \ldots, u_m$ starting from the root u_0, then the number of comparisons is $(N(u_0) - 1) + (N(u_0, u_1) - 1) + \cdots + (N(u_0, u_1, \ldots, u_m) - 1)$, where $N(u_0, u_1, \ldots, u_m)$ is the size of the subtree rooted at the node whose path is defined by $(u_0, u_1, \ldots, u_m)$. To get a handle on the sum of the $N(u_0, u_1, \ldots, u_j)$'s, we represent this random binary search tree in yet another way, following Devroye (1986). The sizes of the subtrees of the root are distributed jointly as $(\lfloor nU \rfloor, \lfloor n(1-U) \rfloor)$ where U is uniform $[0,1]$. We associate this value U with the root of an infinite binary search tree that describes sizes of subtrees. We recursively associate independent uniform random variables with all nodes in this infinite binary search tree. If we follow the path $u_0, u_1, \ldots, u_m$ in this new tree, and $U_{u_0}, U_{u_0,u_1}, \ldots, U_{u_0,u_1,\ldots,u_m}$ are the uniform $[0,1]$ random variables associated with the nodes on this path, starting with the root u_0, and if we define

$$N(u_0, \ldots, u_m) = \lfloor \cdots \lfloor \lfloor nU_{u_0} \rfloor U_{u_0,u_1} \rfloor \cdots U_{u_0,u_1,\ldots,u_m} \rfloor \ ,$$

then the following property is valid: the infinite collection of N_m's (one per node) is jointly distributed as the N_m's in the original random binary search tree, with the understanding in the original tree that $N_m = 0$ denotes a non-existent node or subtree. Clearly, then,

$$T_n = \sup_{m \geq 1; u_0, u_1, \ldots, u_m} \sum_{j=0}^{m} (N(u_0, \ldots, u_j) - 1)_+$$

where the supremum is over all m and all paths $u_0, u_1, \ldots, u_m$. Observe the following: if $U_{u_0,u_1,\ldots,u_i}$ is the uniform $[0,1]$ random variable associated with u_i, then

$$n \prod_{j=0}^{m} U_{u_0,u_1,\ldots,u_j} - (m+1) \leq N(u_0, \ldots, u_m) \leq n \prod_{j=0}^{m} U_{u_0,u_1,\ldots,u_j} \ .$$

Thus, if $\mathcal{P}$ denotes the collection of all paths, and $\mathcal{P}_m$ the collection of all paths of m edges starting from the root, then taking the supremum over all paths in the previous inequality shows the following:

$$\frac{T_n}{n} \leq \sup_{\mathcal{P}} \sum_{j=0}^{\infty} \prod_{i=0}^{j} U_{u_0,u_1,\ldots,u_i} \ .$$

Furthermore, for any integer $m \geq 1$,

$$\frac{T_n}{n} \geq \sup_{\mathcal{P}_m} \sum_{j=0}^{m} \prod_{i=0}^{j} U_{u_0,u_1,\ldots,u_i} - \frac{(m+1)(m+2)}{2n} \ .$$

So, if we take $m = \lfloor n^{1/4} \rfloor$, and let $n \to \infty$, then, by monotone convergence, we see that almost surely, T_n/n tends to

$$S \stackrel{\text{def}}{=} \sup_{\mathcal{P}} \sum_{j=0}^{\infty} \prod_{i=0}^{j} U_{u_0,u_1,\ldots,u_i} \ .$$

The possibly extended random variable S (an extended random variable is one taking the value ∞ with positive probability) does not involve n, and may thus be studied separately, which is what we will do in the next section. Observe that at this point, we do not know yet whether $S < \infty$ with probability one!

The Hoare process and the proof of theorem 1

The proof of Theorem 1 proceeds as follows: first we give another representation for S in terms of random functions. This representation allows us to prove that S is a proper random variable ($S < \infty$ almost surely). We note that S satisfies the distributional identity, and recall from the last five lines in the proof of Grübel and Rösler's Theorem 12 (1996), that the distributional identity has one and only one random variable as a solution. Therefore, we are done.

For lack of a better name, we now describe the Hoare process on $[0,1]$, defined by an infinite binary tree in which each node has an independent copy of a uniform $[0,1]$ random variable. Let $f_0(x) = 1$, $0 \le x < 1$. Define the interval partition $\mathcal{A}_1 = \{[0,U),[U,1)\} \stackrel{\text{def}}{=} \{A_{1,1}, A_{1,2}\}$, where U is the uniform random variable associated with the root. Let

$$f_1(x) = \begin{cases} f_0(x)U & x \in A_{1,1} \ ; \\ f_0(x)(1-U) & x \in A_{1,2} \ . \end{cases}$$

Each of the sets of $\mathcal{A}_1$ is associated with a child of the root, and has in turn a uniform random variable associated with it. Calling these uniform random variables V and W, we obtain a new partition with 2^2 members, $\mathcal{A}_2 = \{[0,UV),[UV,U),[U,U+(1-U)W),[U+(1-U)W,1)\} \stackrel{\text{def}}{=} \{A_{2,1},\ldots,A_{2,4}\}$, and a new function

$$f_2(x) = \begin{cases} f_1(x)V & x \in A_{2,1} \ ; \\ f_1(x)(1-V) & x \in A_{2,2} \ ; \\ f_1(x)W & x \in A_{2,3} \ ; \\ f_1(x)(1-W) & x \in A_{2,4} \ . \end{cases}$$

This construction, duly iterated, yields functions f_n that are staircase-shaped with 2^n supporting intervals. At each x, $f_n(x)$ is the product of n random variables. Define

$$Z_n = \sup_x f_n(x) \ ,$$

the size of the largest of the 2^n intervals in the collection $\mathcal{A}_n$. By taking logarithms, and noting that when U is uniform $[0,1]$, then $-\log U$ is exponentially distributed, we see that $-\log Z_n$ is distributed as the minimum of 2^n (dependent) sums of n independent exponential random variables, and is in fact the minimum value in a branching random walk in which we have two children per node and all displacements are exponentially distributed. The properties of such minima were studied at length by Biggins (1977) after initial work by Kingman (1975) and Hammersley (1974), and in fact,

$$\frac{-\log Z_n}{n} \to \gamma$$

almost surely, where γ is the solution of $2\gamma e^{1-\gamma} = 1$.

The partial sums

$$g_n(x) = \sum_{j=0}^{n} f_j(x) \ , 0 \leq x < 1,$$

define the n-th Hoare process, and $g(x) = \sum_{j=0}^{\infty} f_j(x)$ is the Hoare process. Note that by positivity of the f_j's, $g(x)$ is well-defined (but possibly infinite) for all $x \in [0, 1)$. We note here that this process goes back to Grübel and Rösler (1996), who use it mainly to analyze the complexity of FIND when searching for the $\lfloor xn \rfloor$-th order statistic, where $x \in (0, 1)$ is fixed. In fact, the complexity of FIND in this case is roughly distributed as $ng(x)$.

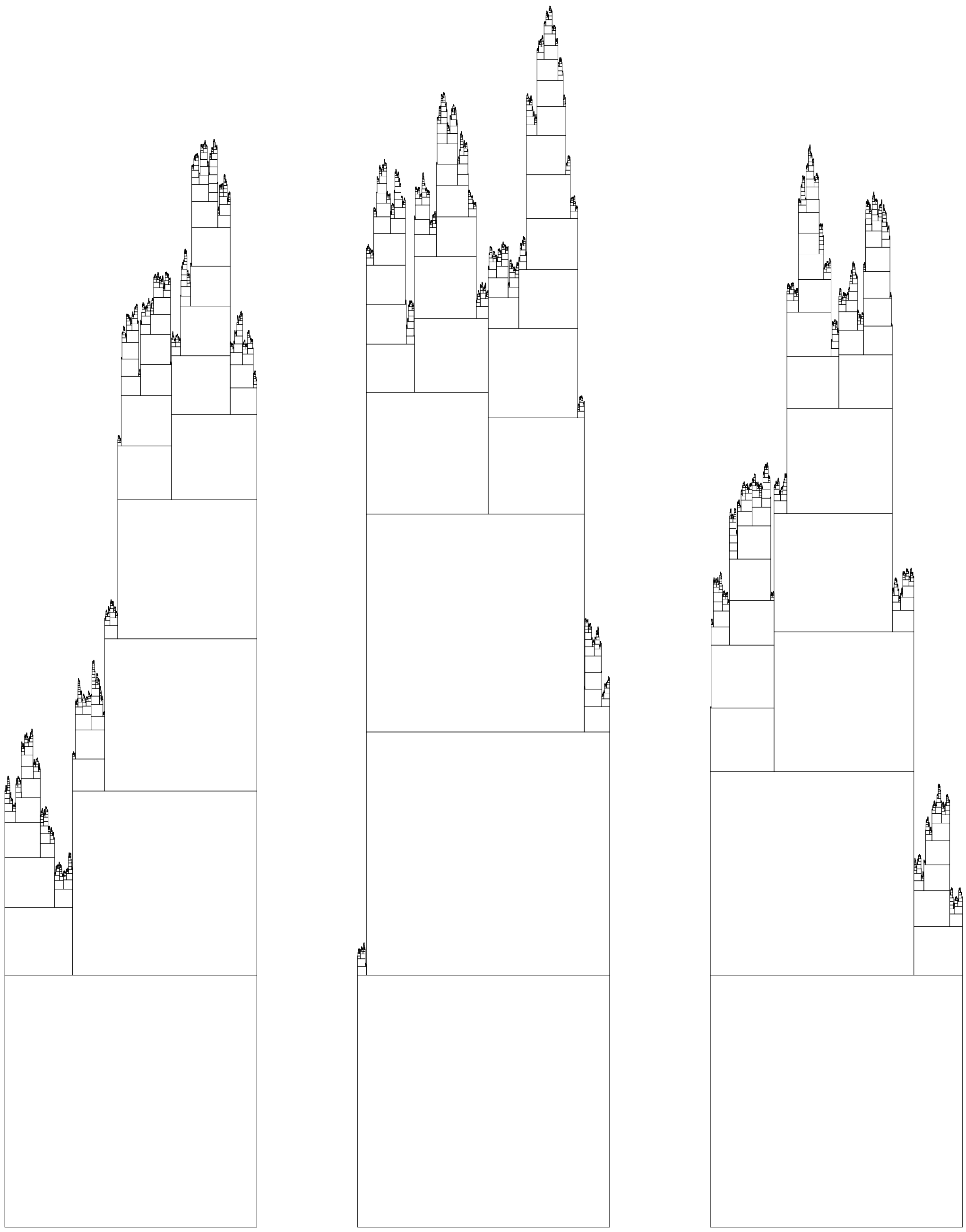

FIGURE 1. The functions g_n and the limit function g are shown. Note that the Hoare supremum S is the supremum of g.

We define the Hoare supremum

$$S = \sup_x g(x) = \lim_{n\to\infty} \sup_x g_n(x)$$

and introduce the sequence of random variables $S_n = \sup_x g_n(x)$. Note that $S_n \le \sum_{j=0}^n Z_j$ so that by Biggins' result cited above, $S_n = O(1)$ almost surely. Furthermore, $S \le \sum_{j=0}^\infty Z_j$ so that $S < \infty$ almost surely. As S is a limit of a monotone supremum of a staircase function, it is easy to see that S is a proper random variable. By the recursive nature of the definition of the Hoare process, we see that S is distributed as $1 + \max(S'U, S''(1-U))$, where S', S'', U are independent, S' and S'' are distributed as S, and U is uniform $[0,1]$. As Grübel and Rösler (1972, theorem 12) showed, there is only one proper random variable that is a solution of this distributional identity. This concludes the proof of Theorem 1.

Additional properties of S

We first show that for our random variable, using only the original definition, $\mathbb{E}\{S\} < \infty$. This will be needed further on for crucial higher moment bounds.

LEMMA 1. $\mathbb{E}\{S-1\} \le 5/\sqrt{2\pi} + 12e/5 < 8.5185$.

PROOF. Note that if the U_i's are independent uniform $[0,1]$ random variables and G_n is a gamma (n) random variable, then, for $t \in (0,1)$,

$$\begin{aligned}\mathbb{P}\{Z_n > t\} &\le 2^n \mathbb{P}\left\{\prod_{i=1}^n U_i > t\right\} = 2^n \mathbb{P}\left\{G_n < -\log t\right\} \\ &= 2^n \int_0^{-\log t} \frac{x^{n-1}e^{-x}}{(n-1)!}\,dx \le \frac{2(2\log(1/t))^{n-1}}{(n-1)!}\ .\end{aligned}$$

Thus, for $c > 0$,

$$\begin{aligned}\mathbb{E}\{Z_n\} &= \int_0^1 \mathbb{P}\{Z_n > t\}\,dt \le \int_0^1 \min\left(\frac{2(2\log(1/t))^{n-1}}{(n-1)!}\ ,\ 1\right)dt \\ &= \int_0^\infty \min\left(\frac{2(2u)^{n-1}}{(n-1)!}\ ,\ 1\right)e^{-u}du \\ &\le \int_0^{cn} \frac{2(2u)^{n-1}}{(n-1)!}du + e^{-cn} \\ &= \frac{(2cn)^n}{n!} + e^{-cn}\ .\end{aligned}$$

Using $n! \ge \sqrt{2\pi n}(n/e)^n$, we have

$$\sum_{n=1}^\infty \mathbb{E}\{Z_n\} \le \sum_{n=1}^\infty \frac{(2cn)^n}{n!} + \frac{1}{e^c-1} \le \sum_{n=1}^\infty \frac{(2ce)^n}{\sqrt{2\pi n}} + \frac{1}{e^c-1} \le \frac{2ce}{(1-2ce)\sqrt{2\pi}} + \frac{1}{c}\ .$$

Recall that $S-1 \leq \sum_{n=1}^{\infty} Z_n$. Therefore, taking $c = 5/(12e)$, we see that $\mathbb{E}\{S-1\} \leq 5/\sqrt{2\pi} + 12e/5 < 8.5185$. $\square$

Introduce $\nu_r = \mathbb{E}\{(S-1)^r\}$.

LEMMA 2. *S is supported on* $[2,\infty)$.

PROOF. Clearly, $S \geq 1$. From the distributional identity,

$$S \geq 1 + \max(U, 1-U) \geq 3/2 \ .$$

But then,

$$S \geq 1 + (3/2)\max(U, 1-U) \geq 7/4 \ .$$

By induction, $S \geq 2$. $\square$

LEMMA 3. *For integer $r \geq 1$,*

$$\mathbb{E}\{(S-1)^r\} \stackrel{\text{def}}{=} \nu_r \leq 3^{r-1} r! \mathbb{E}\{S-1\} < \infty \ .$$

PROOF. The inequality is valid for $r=1$. We prove by induction on r that

$$\nu_r \leq 3^r r! C \ , \ r \geq 1,$$

where $C = \mathbb{E}\{S-1\}/3$. Assume $r \geq 2$ and assume that the inequality is true up to $r-1$. By applying rather crude bounding (a maximum is less than a sum), we note that

$$\nu_r = \mathbb{E}\{(S-1)^r\} \leq 2\mathbb{E}\{S'^r\}\mathbb{E}\{U^r\} = \frac{2}{r+1}\mathbb{E}\{((S-1)+1)^r\} = \frac{2}{r+1}\sum_{j=0}^{r} \nu_j \binom{r}{j} \ .$$

Thus,

$$\nu_r \leq \frac{2}{r-1}\sum_{j=0}^{r-1} \nu_j \binom{r}{j} \leq \frac{2}{r-1}\sum_{j=0}^{r-1} \frac{C 3^j j! r!}{j!(r-j)!} \leq \frac{2C3^r r!}{r-1}\sum_{j=0}^{r-1} \frac{3^{j-r}}{(r-j)!} \leq C 3^r r!$$

provided that

$$2\sum_{j=0}^{r-1} \frac{1}{3^{r-j}(r-j)!} \leq r-1 \ .$$

But the left-hand side is not more than $2\left(e^{1/3}-1\right) \leq 2e^{1/3}/3 < 1 \leq r-1$. Thus, we showed that

$$\nu_r \leq 3^{r-1} r! \mathbb{E}\{S-1\} \ . \square$$

REMARK. From Lemma 3, for some finite constant C,

$$\nu_r \leq C 3^r r! \ , \ r \geq 1.$$

Thus, $(\nu_r)^{1/(2r)} = O(\sqrt{r})$, which implies

$$\sum_{r=2}^{\infty} \frac{1}{(\nu_r)^{1/(2r)}} = \infty \ .$$

This is Carleman's condition (see, e.g., Stoyanov, 1987) for the moments of a positive random variable so that they uniquely define the distribution. Thus, we showed that any solution of the distributional identity having $\mathbb{E}\{S\} < \infty$ is uniquely determined by its moments.

REMARK. By carefully checking the last part of the proof of Lemma 3, we see that if we define $D = 1/W(1/2) = 2.843059872\ldots$, where $W(\cdot)$ is Lambert's W-function (defined as the solution on $W(x)\exp(W(x)) = x$), then

$$\mathbb{E}\left\{(S-1)^r\right\} \leq \frac{r!\mathbb{E}\left\{S-1\right\}}{(W(1/2))^{r-1}} \ .$$

REMARK. Note that the solution X of $X \stackrel{\mathcal{L}}{=} 1 + X \max(U, 1-U)$ is a minorant of S in stochastic order, and may thus be used to obtain lower bounds for moments and tails for S.

THEOREM 2. *The density of the limit random variable S is supported on $[2,\infty)$, has bounded variation not exceeding 2, is bounded by 1, and is in fact Lipschitz with Lipschitz constant not exceeding* 2. *Finally,* $3.3862 < 2 + 2\log 2 \leq \mathbb{E}\left\{S\right\} \leq 1 + 5/\sqrt{2\pi} + 12e/5 < 9.5185$.

PROOF. We first prove that there is a density and that it is bounded by 1. To see this, we look at the density of $S - 1 = \max(S'U, S''(1-U))$ in the notation of Theorem 1. Condition on $[S' = a, S'' = b]$, where $a, b \geq 2$ are arbitrary numbers. Then $\max(aU, b(1-U))$ has the following density:

$$f_{a,b}(x) = \frac{1}{a} I_{[ab/(a+b),a]}(x) + \frac{1}{b} I_{[ab/(a+b),b]}(x) \ .$$

The density is bounded by $1/a + 1/b \leq 1$. As the unconditional density is obtained by replacing a and b by S' and S'' and taking expectations, we see that it too must be bounded by 1. Furthermore, the conditional density is unimodal and of bounded variation equal to $2(1/a + 1/b) \leq 2$. While this does not guarantee that the unconditional density is unimodal, it suffices to conclude that the unconditional density is of bounded variation not exceeding

2. By symmetry, $\mathbb{E}\{f_{S',S''}(x)\} = \mathbb{E}\left\{(2/S')I_{S'S''/(S'+S'')\le x\le S'}\right\}$, and thus we need only look at $g_{a,b}(x) = \frac{2}{a}I_{[ab/(a+b),a]}(x)$. Now, for $x < y$,

$$g_{a,b}(x) - g_{a,b}(y) = \begin{cases} \frac{2}{a} & \text{if } ab/(a+b) \le x \le a < y; \\ -\frac{2}{a} & \text{if } x < ab/(a+b) \le y \le a; \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $|g_{a,b}(x) - g_{a,b}(y)| \le (2/a)I_{x\le a<y} + (2/a)I_{x<ab/(a+b)\le y}$. Unconditioning, we see that

$$\begin{aligned} |\mathbb{E}\{f_{S',S''}(x)\} - \mathbb{E}\{f_{S',S''}(y)\}| &= |\mathbb{E}\{g_{S',S''}(x)\} - \mathbb{E}\{g_{S',S''}(y)\}| \\ &\le \mathbb{E}\{|g_{S',S''}(x) - g_{S',S''}(y)|\} \\ &\le \mathbb{P}\{[S' \in [x,y]] \cup [S'S''/(S'+S'') \in [x,y]]\} \\ &\le 2(y-x) \end{aligned}$$

as the density of S' is bounded by 1, and thus also the density of $S'S''/(S'+S'')$. Therefore, the density of $S-1$ is Lipschitz with constant not exceeding 2. The last statement follows from Lemma 1 and the following simple argument: $\mathbb{E}\{S\} \ge \sup_x \mathbb{E}\{g(x)\}$. From classical results (see, e.g., Grübel and Rösler, 1996), $\sup_x \mathbb{E}\{g(x)\} = 2 + 2\log 2$. □

Tail bounds for S

In this section, we derive useful exponential tail bounds for S and T_n (Theorem 3). In Theorem 4, we give non-explicit superexponential tail bounds for S.

THEOREM 3. *For $t \ge 4$,*

$$\sup_{n\ge 1} \mathbb{P}\{T_n \ge tn\} \le \mathbb{P}\{S \ge t\} \le \frac{\mathbb{E}\{(S-1)\}(t-1)e^{-\frac{t-4}{3}}}{9} .$$

PROOF. For $0 < \lambda < 1/3$, by Lemma 3,

$$\mathbb{E}\left\{e^{\lambda(S-1)}\right\} = \sum_{r=0}^{\infty} \frac{\lambda^r \mathbb{E}\{(S-1)^r\}}{r!} \le \frac{\mathbb{E}\{(S-1)\}}{3} \sum_{r=0}^{\infty} (3\lambda)^r = \frac{\mathbb{E}\{(S-1)\}}{3-9\lambda} .$$

Therefore, by Chernoff's bounding method, for $t \ge 3$,

$$\mathbb{P}\{S-1 \ge t\} \le e^{-\lambda t}\mathbb{E}\left\{e^{\lambda(S-1)}\right\} \le \frac{e^{-\lambda t}\mathbb{E}\{(S-1)\}}{3-9\lambda} \le \frac{e^{-\frac{t-3}{3}}\mathbb{E}\{(S-1)\}\,t}{9}$$

where we took $\lambda = (t-3)/(3t)$ in the last step. Theorem 3 then follows by noting that $T_n \le nS$.
□

Considering the remark following Lemma 3, we note that this bound may be slightly tightened. The bound of Theorem 3 is very useful for comparisons with the standard linear worst-case algorithms, that, depending upon the implementation have time bounds only guaranteed to be about $20n$. The version of Blum, Floyd, Pratt, Rivest and Tarjan (1973) has complexity not exceeding $15n - 163$ for $n > 32$. The much more involved algorithm of Schönhage, Paterson and Pippenger (1976) has complexity $3n + O((n \log n)^{3/4})$.

Regarding the stability of FIND, Grübel and Rösler (1996) showed that the tails of $T_{n,xn}/n$ decrease faster than exponentially for $x \in (0,1)$, but no such superexponential behavior of the tail of T_n/n has been established to date. For related work on superexponential tails, see Mahmoud, Modarres and Smythe (1995). We show precisely that:

THEOREM 4. *For any $A > 0$, there exist constants s and C such that*

$$\mathbb{E}\left\{(S-1)^r\right\} \le Ce^{-A(r-s)}r! \ ,$$

for all integer $r \ge s$. For integer $t \ge s$,

$$\sup_{n\ge 1} \mathbb{P}\{T_n \ge (t+1)n\} \le \mathbb{P}\{S-1 \ge t\} \le Ce^{-A(t-s)} \ .$$

PROOF. The second statement follows from the first one by taking $r = t$, and using Markov's inequality, if $t \ge s$:

$$\mathbb{P}\{S-1 \ge t\} \le \frac{\mathbb{E}\left\{(S-1)^t\right\}}{t^t} \le \frac{Ce^{-A(t-s)}t!}{t^t} \le Ce^{-A(t-s)} \ .$$

Introduce $\nu_r = \mathbb{E}\left\{(S-1)^r\right\}$. In Lemma 3, we showed that for integer $r \ge 1$,

$$\nu_r \le 3^{r-1} r! \mathbb{E}\left\{S-1\right\} \ ,$$

and

$$\nu_r \le \frac{2}{r-1} \sum_{j=0}^{r-1} \nu_j \binom{r}{j} \ .$$

Define

$$s = \left\lceil 4\left(e^{e^A} - 1\right)\right\rceil \ .$$

Set

$$B = \mathbb{E}\left\{(S-1)\right\} 3^{s-1}$$

and note that by Lemma 3, $\max(\nu_0, \nu_1/1!, \ldots, \nu_s/s!) \le B$. Take

$$C = \max\left(B, \frac{4Be}{s}, \frac{2Be(2/e)^2}{s} e^{A\left(e^{A+1}-1\right)}\right) \ .$$

We show by induction on r that

$$\nu_r \le \begin{cases} Br! \le Cr! & r \le s \ ; \\ Ce^{-A(r-s)}r! & r \ge s. \end{cases}$$

The former inequality is obvious by definitions of B and C. Fix $r > s$ and assume the claim about ν_r is valid up to index $r-1$. Then, if $r - s + 1 \ge e^{A+1}$,

$$\begin{aligned}
\nu_r &\le \frac{2}{r-1}\sum_{j=0}^{s-1}\nu_j\binom{r}{j} + \frac{2}{r-1}\sum_{j=s}^{r-1}\nu_j\binom{r}{j} \\
&\le \frac{2Br!}{r-1}\sum_{j=r-s+1}^{\infty}\frac{1}{j!} + \frac{2Cr!}{r-1}\sum_{j=s}^{r-1}\frac{e^{-A(j-s)}}{(r-j)!} \\
&\le \frac{2Ber!}{(r-1)(r-s+1)!} + \frac{2Cr!e^{-A(r-s)}}{r-1}\sum_{j=1}^{r-s}\frac{e^{Aj}}{j!} \\
&\le \frac{2Ber!}{s}\left(\frac{e}{r-s+1}\right)^{r-s+1} + \frac{2Cr!e^{-A(r-s)}}{r-1}\left(e^{e^A}-1\right) \\
&\le \frac{Cr!e^{-A(r-s+1)}}{2} + \frac{2Cr!e^{-A(r-s)}s}{4(r-1)} \\
&\le Cr!e^{-A(r-s)} \ .
\end{aligned}$$

If $s < r < s - 1 + e^{A+1}$, then, using the fact that

$$C \ge \frac{2Be(2/e)^2}{s}e^{A\left(e^{A+1}-1\right)} \ ,$$

we have,

$$\frac{2Ber!}{s(e/2)^2}\left(\frac{e}{r-s+1}\right)^{r-s+1} \le Cr!e^{-A\left(e^{A+1}-1\right)} \le Cr!e^{-A(r-s)}$$

so that we can conclude the induction. □

Inverting any staircase function

The following problem is important in a number of fields: given are unsorted numbers $x_1, \ldots, x_n$, together with a number of positive weights $w_1, \ldots, w_n$. Determine the unique index i such that

$$\sum_{j:x_j<x_i} w_j < y \le \sum_{j:x_j\le x_i} w_j \ ,$$

where y is a given number. We return 0 if no such index exists. It takes just a moment to verify that this problem can be solved by a trivial adjustment of the two-split version of FIND: assume that the (x_i, w_i)'s are already randomly permuted. Then take x_1, and compute $v = \sum_{j:x_j<x_1} w_j$. If $v < y \le v + w_1$, then return 1; if $v + w_1 < y$, then recurse, using only those x_j's whose value is $> x_1$, and replace y by $y - v - w_1$. If $y \le v$, then recurse using only the values $< x_1$.

The complexity of this algorithm (in terms of numbers of comparisons) is a random variable $T(y, w_1, \ldots, w_n)$. However, note that

$$\sup_{y,w_1,\ldots,w_n} T(y, w_1, \ldots, w_n) \stackrel{\mathcal{L}}{=} T_n$$

so that our analysis applies to this situation. In particular, uniformly over all values of the vector $(w_1, \ldots, w_n)$, the complexity is bounded by Sn, where S is the supremum of the Hoare process. The adversary may even pick the weight vector after having inspected the permutation! Finding $F^{\mathrm{inv}}(y)$ where

$$F(x) = \sum_{i=1}^{n} w_i I_{[x_i,\infty)}(x)$$

with positive weights w_i, and the x_i's unsorted can thus be done in time bounded above by a random variable distributed as T_n.

A situation where this arises is the generation of random numbers from $\{x_1, \ldots, x_n\}$ with given probability weights $p_1, \ldots, p_n$. If we preprocess and sort the x_i's and compute the cumulative weights, that is, the cumulative distribution function, and if we let y be a uniform $[0, 1]$ random variable, then the returned x_i has indeed the correct distribution: this is known as the inversion method in random variate generation (Devroye, 1986). By keeping two arrays, the expected time can be shown to be $O(1)$, uniformly over all n and all weight vectors—this is known as the alias method (Walker, 1977). However, without preprocessing, inversion may be organized by the FIND-like algorithm in $O(n)$ expected time, uniformly over all weight vectors, as we showed above.

Generalizations of FIND in the hope of obtaining better performance are relatively easy to obtain. For example, if we use the random permutation model, and pick the median of the first $2m + 1$ for splitting, then with some minor work, one can show that T_n/n has a limit distributed as S, where $S \stackrel{\mathcal{L}}{=} 1 + \max(S'X, S''(1 - X))$, and X is beta $(m+1, m+1)$ distributed. For related work, see Kirschenhofer, Prodinger and Martinez (1997).

Finally, to search for several order statistics, one may turn to multiple quickselect, a generalization of Hoare's method. It was analyzed by Mahmoud and Smythe (1998), Panholzer and Prodinger (1998) and Prodinger (1995). Its worst-case properties (over all combinations of m order statistics) should be studied.

Acknowledgment

I would like to thank both referees for a thorough reading of the paper, and for pointing out some interesting references. Thanks are also due to Ralph Neininger for valuable comments.

References

J. D. Biggins, "The first and last-birth problems for a multitype age-dependent branching process," *Advances in Applied Probability*, vol. 8, pp. 446–459, 1976.

J. D. Biggins, "Chernoff's theorem in the branching random walk," *Journal of Applied Probability*, vol. 14, pp. 630–636, 1977.

M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, "Time bounds for selection," *Journal of Computers and System Sciences*, vol. 7, pp. 448–461, 1973.

L. Devroye, "Exponential bounds for the running time of a selection algorithm," *Journal of Computers and System Sciences*, vol. 29, pp. 1–7, 1984.

L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

L. Devroye, "A note on the height of binary search trees," *Journal of the ACM*, vol. 33, pp. 489–498, 1986.

R. W. Floyd and R. L. Rivest, "Expected time bounds for selection," *Communications of the ACM*, vol. 18, pp. 165–172, 1975.

R. Grübel and U. Rösler, "Asymptotic distribution theory for Hoare's selection algorithm," *Advances of Applied Probability*, vol. 28, pp. 252–269, 1996.

J. M. Hammersley, "Postulates for subadditive processes," *Annals of Probability*, vol. 2, pp. 652–680, 1974.

C. A. R. Hoare, "Find (algorithm 65)," *Communications of the ACM*, vol. 4, pp. 321–322, 1961.

J. F. C. Kingman, "The first-birth problem for an age-dependent branching process," *Annals of Probability*, vol. 3, pp. 790–801, 1975.

P. Kirschenhofer, H. Prodinger, and C. Martinez, "Analysis of Hoare's Find algorithm with median-of-three partition," *Random Structures and Algorithms*, vol. 10, pp. 143–156, 1997.

P. Kirschenhofer and H. Prodinger, "Comparisons in Hoare's Find algorithm," *Combinatorics, Probability and Computing*, vol. 7, pp. 111–120, 1998.

D. E. Knuth, "Mathematical analysis of algorithms," in: *Information Processing 71* , pp. 19–27, Proceedings of IFIP Congress Ljubljana 1971, North-Holland, 1972.

J. Lent and H. Mahmoud, "Average-case analysis of multiple Quickselect: an algorithm for finding order statistics," *Statistics and Probability Letters*, vol. 28, pp. 299–310, 1996.

H. Mahmoud, R. Modarres, and R. Smythe, “Analysis of Quickselect: an algorithm for order statistics,” *RAIRO: Theoretical Informatics and its Applications*, vol. 29, pp. 255–276, 1995.

H. Mahmoud and R. Smythe, “Probabilistic analysis of Multiple Quickselect,” *Algorithmica*, vol. 22, pp. 569–584, 1998.

R. Neininger and L. Rüschendorf, “On the internal path length in quadtrees,” *Random Structures and Algorithms*, vol. 15, pp. 25–41, 1999.

R. Neininger and L. Rüschendorf, “Limit laws for partial match queries in quadtrees,” Technical Report 32-1999, University of Freiburg, Germany, 1999.

R. Neininger, “Asymptotic distributions for partial match queries in K-d trees,” *Random Structures and Algorithms*, vol. 17, pp. 403–427, 2000.

A. Panholzer and H. Prodinger, “A generating functions approach for the analysis of grand averages for Multiple Quickselect,” *Random Structures and Algorithms*, vol. 13, pp. 189–209, 1998.

H. Prodinger, “Multiple Quickselect—Hoare’s Find algorithm for several elements,” *Information Processing Letters*, vol. 56, pp. 123–129, 1995.

U. Rösler, “A limit theorem for quicksort,” *Theoretical Informatics and Applications*, vol. 25, pp. 85–100, 1991.

U. Rösler, “A fixed point theorem for distributions,” *Stochastic Processes and their Applications*, vol. 42, pp. 195–214, 1992.

U. Rösler, “The backward view on the algorithm FIND,” Technical Report, Fachbereich Stochastik, Universität Kiel, Germany, 1997.

U. Rösler and L. Rüschendorf, “The contraction method for recursive algorithms,” Technical Report, University of Kiel, 1999.

A. Schönhage, M. Paterson, and N. Pippenger, “Finding the median,” *Journal of Computers and System Sciences*, vol. 13, pp. 184–199, 1976.

J. M. Stoyanov, *Counterexamples in Probability*, John Wiley, Chichester, 1987.

A. J. Walker, “An efficient method for generating discrete random variables with general distributions,” *ACM Transactions on Mathematical Software*, vol. 3, pp. 253–256, 1977.