Information Processing Letters 27 (1988) 103–109
North-Holland

29 February 1988

ON THE RANDOM CONSTRUCTION OF HEAPS

Alan M. FRIEZE

Department of Computer Science and Statistics, Queen Mary College, University of London, Mile End Road, London, United Kingdom E1 4NS

Communicated by P. Henderson
Received 1 April 1987
Revised 1 June 1987

We give a simple proof of the result of Bollobás and Simon (1985) on the linear expected time construction of a binary heap. We also give bounds on the probability that the construction time is much more than this and generalise the result to d-heaps, $d \geqslant 2$.

Keywords: Binary heap, d-heap, upper bound

1. Introduction

The binary heap is one of the basic data structures of Computer Science. An array $A[0] = -\infty$, $A[1]$, $A[2], \ldots, A[n]$ is in heap order if $A[i] \geqslant A[\lfloor \frac{1}{2}i \rfloor]$ for $i \geqslant 1$. It is well known (see, for example, [3]) that an originally unsorted array A can be put into heap order in $O(n)$ time. On the other hand, it is also known that the 'repeated insertion' Algorithm RI below requires $\Omega(n \log n)$ time in the worst case.

Algorithm RI. The input, real array $A[1..n]$, is transformed into a binary heap by repeated insertion *from the bottom*.

```
1.   begin
2.        $A[0] := -\infty$;
3.        for $t = 1$ to $n$ do
4.        begin
5.             $p := t;\ q := \lfloor \frac{1}{2}t \rfloor;\ a := A[t]$;
6.             while $A[q] > a$ do
7.             begin
8.                  $A[p] := a[q];\ p := q;\ q := \lfloor \frac{1}{2}p \rfloor$
9.             end
10.            $A[p] := a$
11.       end
12.  end
```

Bollobás and Simon [1] analysed the average case time complexity of Algorithm RI under the assumption that the input array $A[1..n]$ contains a random permutation of $\{1, 2, \ldots, n\}$. If $T_{\mathrm{RI}}(A)$ denotes the (random) number of exchanges (= executions of line **8**) for input A, then they proved the following theorem.

0020-0190/88/$3.50 © 1988, Elsevier Science Publishers B.V. (North-Holland)

Theorem A

$$E(T_{\mathrm{RI}}(A)) \leqslant (1+\phi+\mathrm{o}(1))n, \quad \text{where } \phi=\sum_{t=1}^{\infty}\frac{1}{1+2^t}\approx 0.7645. \tag{1}$$

(Strictly speaking they prove the result for $n+1$ a power of 2. We do the same. For arbitrary n one can only double the bound *from what is proved*.)

This is a nice result but its proof is rather involved and the main purpose of this paper is to simplify the proof. (It also removes the o(1) term in inequality (1)). By using a different model of randomness we are able to cut out most of the 'combinatorial computation'. We are also able to prove some related results. For example, we prove that Algorithm RI is unlikely to require much more than $(1+\phi)n$ exchanges $(n+1=2^k)$.

Theorem B. *Let* $0<\varepsilon<1$ *be fixed. Then, for some constant* $c>0$,

$$\Pr(T_{\mathrm{RI}}(A)\geqslant(1+\phi+\varepsilon)n)=\mathrm{o}\left(\mathrm{e}^{-c\varepsilon^2 n^{1/10}}\right).$$

We also consider d-heaps, $d\geqslant 2$, where all we require is $A[i]\geqslant A[\lfloor(i+d-2)/d\rfloor]$ and analyse the obvious generalisation of Algorithm RI_d, in which $\lfloor\frac{1}{2}t\rfloor$ in line **5** is replaced by $\lfloor(t+d-2)/d\rfloor$ and $\lfloor\frac{1}{2}p\rfloor$ in line **8** is replaced by $\lfloor(p+d-2)/d\rfloor$. We prove the following.

Theorem C

$$E\left(T_{\mathrm{RI}_d}(A)\right)\leqslant\left(\frac{d}{2(d-1)}+\phi_d\right)n, \quad \text{where } \phi_d=\sum_{t=1}^{\infty}\frac{1}{1+d^t}.$$

2. Model of randomness and notation

The main device for simplifying the proof of Bollobás and Simon is a change of model. Instead of considering $A[1..n]$ to contain a random permutation of $\{1,2,\ldots,n\}$ we assume that $A[1], A[2],\ldots, A[n]$ are independent uniform [0, 1] random variables. All we have to observe is that the number of exchanges used by Algorithm RI only depends on $i_1, i_2,\ldots, i_n$, where $A[i_1]<A[i_2]<\cdots<A[i_n]$ and that $i_1, i_2,\ldots, i_n$ is equally likely to be any permutation of $\{1,2,\ldots,n\}$. (Note that $\Pr(\exists i\neq j\colon A[i]=A[j])=0$ and such null events can safely be ignored.)

We obtain a marginal simplification in our description if we imagine the process of heap construction continuing indefinitely (putting $n=\infty$ in line 3) and allowing the input to be an infinite sequence $u_1, u_2,\ldots, u_n,\ldots$ of independent uniform [0, 1] random variables.

The execution of lines **4–11** with a particular value of t takes place *during time period* t and time t denotes the end of time period t.

We imagine that, at time 0, $u_1, u_2,\ldots, u_n,\ldots$ occupy the vertices of an infinite binary tree T_∞, where vertex i has children $2i$ and $2i+1$. We let $u_{i,t}$ denote the *contents* of vertex i at time t, so that $u_{i,0}=u_i$ and $u_{1,t}, u_{2,t},\ldots, u_{t,t}$ is heap ordered.

For $k=0,1,\ldots,$ level k of T_∞ consists of the vertices $L_k=\{2^k, 2^k+1,\ldots,2^{k+1}-1\}$ and $\tau_k=2^{k+1}-1$, the time when Algorithm RI has reached the end of level k.

Next, let

$$A_{i,k}=\sum_{j\in L_i}u_{j,\tau_k} \quad\text{and}\quad B_{i,k}=\sum_{j=0}^{i}a_{j,k} \qquad \text{for } 0\leqslant i\leqslant k.$$

3. Proof of Theorem A

Fix $k>0$ and consider the insertion of the elements in level k. For $t \in L_k$ let $P(t)=\{\lfloor t/2^j \rfloor : j = 1, 2, \ldots, k-1\}$ be the vertices on the path from vertex t to the root of T_∞.

Let

$$\begin{aligned} X_t &= |\{ i \in P(t) : u_{i,t-1} > u_t \}| \\ &= \text{the number of exchanges in the insertion of } u_t \text{ by Algorithm RI} \\ &\leqslant |\{ i \in P(t) \mid u_{i,\tau_{k-1}} > u_t \}| = Y_t, \quad \text{say,} \end{aligned}$$

since $u_{i,s+1} \leqslant u_{i,s}$ for $s \geqslant i$.

We prove that

$$E\left(\sum_{t \in L_k} Y_t \right) \leqslant (1+\phi)\,|L_k|, \quad k = 1, 2, \ldots, \tag{2}$$

and then inequality (1) follows from $X_t \leqslant Y_t$, as observed in [1].

Now,

$$E\left(Y_t \mid u_{i,\tau_{k-1}}, i \in P(t)\right) = \sum_{i \in P(t)} \Pr\left(u_t < u_{i,\tau_{k-1}} \mid u_{i,\tau_{k-1}}, i \in P(t)\right) = \sum_{i \in P(t)} u_{i,\tau_{k-1}}.$$

Hence,

$$\begin{aligned} E\left(\sum_{t \in L_k} Y_t \mid u_{i,\tau_{k-1}}, 1 \leqslant i \leqslant \tau_{k-1} \right) &= \sum_{t \in L_k} \sum_{i \in P(t)} u_{i,\tau_{k-1}} = \sum_{j=0}^{k-1} 2^{k-j} A_{j,k-1} \\ &= 2^k \left(\sum_{j=0}^{k-2} 2^{-(j+1)} B_{j,k-1} + 2^{-(k-1)} B_{k-1,k-1} \right) \end{aligned}$$

by straightforward algebra.

Removing the conditioning yields

$$E\left(\sum_{t \in L_k} Y_t \right) = 2^k \left(\sum_{j=0}^{k-2} 2^{-(j+1)} E(B_{j,k-1}) + 2^{-(k-1)} E(B_{k-1,k-1}) \right). \tag{3}$$

All that is needed are upper bounds to $E(B_{j,i})$ for $0 \leqslant j \leqslant i$.

$$E(B_{0,i}) = 1/2^{i+1} \tag{4}$$

since $B_{0,i}$ is the minimum of $2^{i+1}-1$ independent uniform [0, 1] random variables.

$$E(B_{i,i}) = 2^i - \tfrac{1}{2} \tag{5}$$

since $B_{i,i}$ is the sum of $2^{i+1}-1$ independent uniform [0, 1] random variables.

To estimate $E(B_{j,i})$ for $j<i$ we observe (see [1, Lemma 5]) that

$$B_{j,i} \leqslant B_{j,i-1} - \sum_{t \in L_j} \max\{0, u_{t,\tau_{i-1}} - v_t\},$$

where $v_t = \min\{u_s : s \in L_i \text{ and } t \in P(s)\}$.

To see this, note that if $v_t < u_{t,\tau_{i-1}}$, then the corresponding value u_s will rise to level j at time s and push down the contents of t. Thus, by this time, when looking at the sum of the contents of the first j levels, v_t has replaced $u_{t,\tau_{i-1}}$ and some other values may have been reduced. Hence,

$$B_{j,i} \leqslant B_{j,i-1} - \sum_{t \in L_j} (u_{t,\tau_{i-1}} - v_t) = B_{j-1,i-1} + \sum_{t \in L_j} v_t \tag{6}$$

and so

$$E(B_{j,i}) \leqslant E(B_{j-1,i-1}) + 2^j/(2^{i-j}+1)$$

since each v_t is the minimum of 2^{i-j} independent uniform [0, 1] random variables. Thus,

$$E(B_{j,i}) \leqslant E(B_{0,i-j}) + \frac{2^j + 2^{j-1} + \cdots + 2}{2^{i-j}+1} \leqslant \frac{2^{j+1}-1}{2^{i-j}+1}, \tag{7}$$

on using (4). Using (5) and (7) in (3) yields

$$E\left(\sum_{t \in L_k} Y_t\right) \leqslant 2^k \left(\sum_{j=0}^{k-2} 2^{-(j+1)} \frac{2^{j+1}-1}{2^{k-1-j}+1} + 1 - 1/2^k \right) < 2^k(1+\phi).$$

As $|L_k| = 2^k$ we have (2) and Theorem A.

4. Proof of Theorem B

Again we consider $n+1 = 2^k$ and the insertion of level k and the random variables Y_t for $t \in L_k$. Let $\mathscr{E}_1$ be the event $\{\Sigma_{t \in L_k} Y_t \geqslant (1+\varepsilon)(1+\phi)\,|L_k|\}$. We shall prove that

$$\Pr(\mathscr{E}_1) = o\left(e^{-\varepsilon^2 n^{1/5}/25}\right). \tag{8}$$

Since $\Sigma_{t \in L_j} Y_t \leqslant j\,|L_j|$ always, we then will have

$$\Pr\left(\sum_{j=0}^{k} \sum_{t \in L_j} Y_t \geqslant (1+\varepsilon)(1+\phi)n \right) \leqslant \Pr\left(\sum_{j=\lceil k/2 \rceil}^{k} \sum_{t \in L_j} Y_t \geqslant (1+\varepsilon)(1+\phi)n - \sqrt{n}\log n \right)$$
$$= o\left(k\, e^{-\varepsilon^2 n^{1/10}/25}\right)$$

and this will prove the theorem.

Let $\mathscr{E}_2$ be the event

$$\left\{ \sum_{j=0}^{k-2} 2^{-(j+1)} B_{j,k-1} + 2^{-(k-1)} B_{k-1,k-1} \leqslant (1 + \tfrac{1}{2}\varepsilon)(1+\phi) \right\}.$$

Then,

$$\Pr(\mathscr{E}_1) \leqslant \Pr(\mathscr{E}_1 \mid \mathscr{E}_2) + \Pr(\bar{\mathscr{E}}_2).$$

We shall show later on that

$$\Pr(\bar{\mathscr{E}}_2) = o\left(e^{-\varepsilon^2 n^{1/5}/25}\right) \tag{9}$$

and we quickly dispose of $\Pr(\mathscr{E}_1 \mid \mathscr{E}_2)$.

Consider the following part of Theorem 1 by Hoeffding [2, inequality (2.3)]: let $Z_1, Z_2, \ldots, Z_m$ be independent random variables where $0 \leqslant Z_i \leqslant 1$ for $i = 1, 2, \ldots, m$ and $E((Z_1 + Z_2 + \cdots + Z_m)/m) = \mu$. Then, for $t > 0$,

$$\Pr((Z_1 + Z_2 + \cdots + Z_m)/m \geqslant \mu + t) \leqslant e^{-2mt^2}. \tag{10}$$

We now observe that, given $u_{1,\tau_{k-1}}, \ldots, u_{\tau_{k-1},\tau_{k-1}}$, the variables Y_t, $t \in L_k$, are independent and $0 \leqslant Y_t \leqslant k$. Also,

$$E\left(2^{-k} \sum_{t \in L_k} Y_t \mid \mathscr{E}_2\right) \leqslant (1 + \tfrac{1}{2}\varepsilon)(1 + \phi).$$

Applying inequality (10) with $m = 2^k$, $Z_i = Y_i/k$, $i \in L_k$, $\mu \leqslant (1 + \frac{1}{2}\varepsilon)(1 + \phi)/k$ and $t = \varepsilon(1 + \phi)/(2k)$ yields

$$\Pr(\mathscr{E}_1 \mid \mathscr{E}_2) \leqslant e^{-n(\varepsilon(1+\phi)/2\log n)^2}$$

and the theorem follows once we have dealt with equation (9).

Consider $r \in L_j$ and $j < i$. We show that

$$u_{r,\tau_i} \leqslant (j+1)\text{st smallest of } \left\{ u_s : s \in \bigcup_{u=j+1}^{i} L_u \text{ and } r \in P(s) \right\},$$

i.e., u_{r,τ_i} is at most the $(j+1)$st smallest of $2^{i-j+1} - 2 \geqslant 2^{i-j}$ independent uniform [0, 1] random variables.

To prove this we inductively show that

$$u_{r,t} \leqslant (j+1)\text{st smallest of } \{ v_s : r \in P(s) \}, \quad j \geqslant 0,\ r \in L_j,\ t \geqslant 0, \tag{11}$$

where $v_s = u_s$ if $t \geqslant s$ and $v_s = \infty$ otherwise.

Now, clearly (11) holds for $j = 0$ and $t \geqslant 0$; assume it holds for all $r \in L_{j-1}$ and $t \geqslant 0$ for some $j > 0$. Now, (11) trivially holds while $|\{s : r \in P(s) \text{ and } v_s < \infty\}| < j + 1$ and so we have a basis for assuming (11) true up to some value of t. Let $S = \{s : r \in P(s) \text{ and } v_s < \infty\}$. We must confirm (11) for the next case $t' > t$ such that $r \in P(t')$. Let $S' = S \cup \{t'\}$. We must show that $u_{r,t'} \leqslant (j+1)$st smallest of $\{u_s : s \in S'\}$. This is clearly true if $u_{t'} \geqslant u_{r,t}$ and if $u_{t'} < u_{r,t}$, then $u_{r,t'} = u_{\lfloor r/2 \rfloor, t}$ which is at most the jth smallest of a set which includes S. Hence, (11) continues to hold and the inductive step is complete.

It then follows that

$$\Pr\left(A_{j,i} \geqslant \alpha 2^{2j-i}\right) \leqslant \Pr\left(\exists r \in L_j : u_{r,\tau_i} \geqslant \alpha 2^{j-i}\right) \leqslant 2^j \binom{2^{i-j}}{j} (1 - \alpha 2^{j-i})^{2^{i-j}-j} \quad \text{if } \alpha \leqslant 2^{i-j},$$

$$\leqslant 2^j\, e^{-(\alpha - \alpha j 2^{j-i} - j(i-j)\log_e 2)} \leqslant 2^j\, e^{-(\alpha - \alpha j 2^{j-i} - k^2)} \quad \text{if } \alpha \leqslant 2^{i-j}.$$

Since $A_{j,i} \leqslant 2^j$ always, we can remove the restriction on α in the above inequality. Hence,

$$\Pr\left(B_{j,i} \geqslant \tfrac{4}{3}\alpha 2^{2j-i}\right) \leqslant \Pr\left(\exists \hat{j} \leqslant j : A_{\hat{j},i} \geqslant \alpha 2^{2\hat{j}-i}\right)$$

$$\leqslant e^{-(\alpha - \alpha j 2^{j-i} - k^2)} \sum_{\hat{j}=0}^{j} 2^{\hat{j}} \leqslant 2^{j+1}\, e^{-(\alpha - \alpha j 2^{j-i} - k^2)}. \tag{12}$$

Putting $\alpha = 2^{k/5}$ we see that

$$\Pr\left(\sum_{j=0}^{\lfloor 3k/5 \rfloor} 2^{-(j+1)} B_{j,k-1} \geqslant \tfrac{8}{3} 2^{-k/5} \right) \leqslant \Pr\left(\exists j \leqslant \tfrac{3}{5}k : 2^{-(j+1)} B_{j,k-1} \geqslant \tfrac{4}{3} \times 2^{j-4k/5}\right)$$

$$\leqslant 2n^{3/5}\, e^{-2^{k/5-1}}. \tag{13}$$

Next we give a simple consequence of Hoeffding's Theorem 1 [2] (easily derivable from inequality (2.1) of that paper): if $Z_1, Z_2, \ldots, Z_m$ are as in (10) and $0 < \varepsilon < 1$, then

$$\Pr((Z_1 + Z_2 + \cdots + Z_m)/m \geqslant (1+\varepsilon)\mu) \leqslant \mathrm{e}^{-\varepsilon^2 m\mu/3}. \tag{14}$$

Now suppose that $i > \lfloor \frac{3}{5}k \rfloor$. It follows from (6) and (14) that

$$\Pr\left(B_{i-r,k-1-r} - B_{i-r-1,k-2-r} \geqslant (1+\varepsilon)\frac{2^{i-r}}{2^{k-1-i}+1} \right) \leqslant \mathrm{e}^{-\varepsilon^2 2^{2i-k-r}/3} \tag{15}$$

for $0 \leqslant r \leqslant s = i - \lfloor \frac{3}{5}k \rfloor$. Thus,

$$\begin{aligned} &\Pr\left(2^{-(i+1)}B_{i,k-1} - 2^{-(i+1)}B_{i-s,k-1-s} \geqslant (1+\varepsilon) \sum_{r=0}^{s-1} \frac{2^{-(r+1)}}{2^{k-i-1}+1} \right) \\ &\quad \leqslant \sum_{r=0}^{s-1} \mathrm{e}^{-\varepsilon^2 2^{2i-k-r}/3} < s\,\mathrm{e}^{-\varepsilon^2 2^{k/5}/3}. \end{aligned} \tag{16}$$

If $\lfloor \frac{3}{5}k \rfloor < i \leqslant \lfloor \frac{4}{5}k \rfloor$, then we can use (12) with $\alpha = 2^{k/5}$ to show that

$$\Pr\left(B_{i-s,k-1-s}/2^{i+1} \geqslant \tfrac{4}{3}2^{-k/5} \right) \leqslant 2^{i-s+1}\mathrm{e}^{-2^{k/5-1}}.$$

If $i > \lfloor \frac{4}{5}k \rfloor$, then we use $B_{i-s,k-1-s}/2^{i+1} \leqslant 2^{-s} \leqslant 2^{-k/5}$.

In conjunction with (16) we get that, for $i > \lfloor \frac{3}{5}k \rfloor$,

$$\Pr\left(B_{i,k-1}/2^{i+1} \geqslant (1+\varepsilon)/(2^{k-i-1}+1) + \tfrac{4}{3} \times 2^{-k/5} \right) = \mathrm{o}\left(n\,\mathrm{e}^{-\varepsilon^2 n^{1/5}/6} \right). \tag{17}$$

It follows next from (14) that

$$\Pr\left(2^{-(k-1)}B_{0,k-1} \geqslant 1+\varepsilon \right) \leqslant \mathrm{e}^{\varepsilon^2(2^{k-1}-1/2)/3} \tag{18}$$

as $B_{0,k-1}$ is the sum of $2^k - 1$ independent uniform [0, 1] random variables. (In)equalities (13), (17), and (18) then imply that

$$\Pr\left(\sum_{j=0}^{k-2} 2^{-(j+1)}B_{j,k-1} + 2^{-(k-1)}B_{k-1,k-1} \geqslant (1+\varepsilon)(1+\phi) + 4 \times 2^{-k/5} \right) = \mathrm{o}\left(n\,\mathrm{e}^{-\varepsilon^2 n^{1/5}/6} \right). \tag{19}$$

Replacing ε in (19) by ε' satisfying $(1+\varepsilon')(1+\phi) + 4 \times 2^{-k/5} = (1+\frac{1}{2}\varepsilon)(1+\phi)$ yields (9).

5. Proof of Theorem C

The idea of the proof is the same as that of Theorem A. This time T_∞ denotes an infinite d-ary tree where vertex i has children $(i-1)d+j$, $j = 2, 3, d+1$. The level sets L_k and the $A_{i,k}$, $B_{i,k}$ are defined analogously, as are the variables X_t, Y_t, $t \in L_k$. In place of (3) we have

$$E\left(\sum_{t \in L_k} Y_t \right) = d^k \left(\sum_{j=0}^{k-2} d^{-(j+1)}(d-1)E(B_{j,k-1}) + d^{-(k-1)}E(B_{k-1,k-1}) \right). \tag{20}$$

In place of (4) we have

$$E(B_{0,i}) = (d-1)/(d^{i+1}+d-2). \tag{21}$$

In place of (5) we have

$$E(B_{i,i}) = (d^{i+1}-1)/[2(d-1)]. \tag{22}$$

In place of (6) we have

$$E(B_{j,i}) < E(B_{j-1,i-1}) + d^j/(d^{i-j}+1) < \frac{d^{j+1}-1}{(d-1)(d^{i-j}+1)}. \tag{23}$$

It then follows from (20)–(23) that

$$E\left(\sum_{t\in L_k} Y_t\right) \leqslant d^k\left(\sum_{j=0}^{k-2} \frac{d^{j+1}-1}{d^{j+1}} \frac{1}{d^{k-1-j}+1} + \frac{d^k-1}{2d^{k-1}(d-1)}\right)$$

and Theorem C follows.

6. Final comments

We do not believe that the upper bound in Theorem A is tight, although it is not clear how to improve it significantly. We have not mentioned lower bounds. We can obtain a rather weak bound of approximately $\frac{3}{4}n$ on the expected number of exchanges as follows: the number of exchanges needed to insert u_t, $t \in L_k$, is at least what would the required assuming that levels $L_0, L_1, \ldots, L_{k-1}$ have their contents at time τ_k. Thus,

$$E\left(\sum_{t\in L_k} X_t\right) \geqslant 2^k\left(\sum_{j=0}^{k-2} 2^{-(j+1)}E(B_{j,k}) + 2^{-(k-1)}E(B_{k-1,k})\right). \tag{24}$$

Now, $B_{j,k}$ is at least the sum of the $2^{j+1}-1$ smallest values in $u_1, u_2, \ldots, u_{2^{k+1}-1}$, and so $E(B_{j,k}) \geqslant 2^{j+1}(2^{j+1}-1)/2^{k+2}$. Substitution of this into (24) yields $E(\sum_{t\in L_k} X_t) \geqslant (\frac{3}{4} - o(1))2^k$.

Acknowledgment

I am particularly grateful to Trevor Fenner for his careful reading of the manuscript, and to Martin Dyer for his helpful comments.

References

[1] B. Bollobás and I. Simon, Repeated random insertion into a priority queue, *J. Algorithms* 6 (1985) 466–477.

[2] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58** (1963) 13–30.

[3] D.E. Knuth, *The Art of Computer Programming, Vol. 3, Sorting and Searching* (Addison-Wesley, Reading, MA, 1973).