JOURNAL OF ALGORITHMS 6, 1–9 (1985)

The Expected Length of the Longest Probe Sequence for Bucket Searching When the Distribution Is Not Uniform*

LUC DEVROYE

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

Received October 21, 1982

We study the expected value of the maximum number of accesses needed to locate an element in a hashing file constructed by using an order-preserving hashing function and with collision resolution by the method of separate chaining. It is assumed that $X_1, \ldots, X_n$ are independent $[0,1]$-valued random variables with common density f, and that X_i is hashed to the $\lfloor nX_i + 1 \rfloor$st bucket (chain). For all densities that are bounded, the expected value of the maximum number of accesses is shown to be asymptotic to $\log n/\log\log n$, and the dependency of this expected value on f is made explicit by exhibiting the first few terms in the asymptotic expansion. For unbounded f, a tight upper bound is given for the expected value.

1. INTRODUCTION

Assume that n elements are stored in a hash table with n locations by the method of separate chaining [7, Sect. 6.4]: for each location we have a linked list with all the elements (keys) that hash to that location. For a perfect hash function (i.e., one that assures that all locations are chosen with equal probability) the average number of probes in successful and unsuccessful search is well known (see Knuth [7]). The expected length of the longest probe sequence increases very slowly with n: Gonnet [6] has shown that this expected length is asymptotic to $\Gamma^{-1}(n)$ where Γ is the gamma function. For example, for $n = 40{,}320$, its value is near 7.35 (Gonnet [6, Table V]). Additional information is given in Larson [8]. There are sometimes reasons to keep the elements in order, or nearly in order, e.g., in the

*Research of the author was supported by NSERC Grant A-3456.

0196-6774/85 $3.00

context of data structures for geometrical problems, or when frequent alphabetical listings of names are required. Order-preserving hash functions lead usually to nonuniform distributions over the locations. For example, when X is a random variable with density f and distribution function F, it is well known that for a monotone function h to give a uniform $[0,1]$ random variable $h(X)$, we must have $h(X) = F(X)$ for almost all $x(F)$. Yet, F is usually not known. It is the purpose of this note to point out how nonuniform distributions affect the expected length of the longest probe sequence.

Our model is the following. Let $X_1, \ldots, X_n$ be a sequence of independent $[0,1]$-valued random variables with common density f. This is the sequence of data points or a suitable transformation of this sequence to force the values to fall into $[0,1]$. We have n locations, or *buckets*: the ith bucket holds all the data points with value in $[(i-1)/n, i/n)$. If f is sufficiently well spread out, the data points are nearly sorted, and a second pass of the buckets is all that is needed to obtain a completely sorted sequence (Dobosiewicz [4], Devroye and Klincsek [3], Meijer and Akl [9]). The structure can be used for searching too (Akl and Meijer [1], Ehrlich [5]). When N_i is the cardinality of the ith bucket, it is easy to see that the longest probe sequence for successful search has length $\max_i N_i$. Gonnet's results are valid for the case of a uniform density on $[0,1]$. We will indicate how the nonuniformity of f influences $E(M_n)$ where $M_n = \max_i N_i$.

In Theorem 1 below, we consider only bounded densities f, and show that

$$E(M_n) \sim \frac{\log n}{\log\log n}$$

for all such densities. Thus, in first approximation, the density does not influence $E(M_n)$. The explanation is due to the fact that the expected value of the maximum of n independent Poisson (λ) random variables is asymptotic to $\log n/\log\log n$, for *any* constant λ. It is thus of some interest to know how f affects $E(M_n)$. We will show that this occurs through the smallest bound on f, and then only in the third term of the asymptotic expansion for $E(M_n)$.

THEOREM 1. *Let* $c = \operatorname{ess\,sup} f$ *(i.e., c is the smallest real number such that the Lebesgue measure of the set $\{x\colon f(x) > c\}$ is 0). Then, if $c < \infty$,*

$$E(M_n) = \frac{\log n}{\log\log n} + \frac{\log n}{(\log\log n)^2}(\log\log\log n + 1 + \log c + o(1)).$$

In particular, $E(M_n) \sim \log n/\log\log n$ *whenever* $c < \infty$. *If* $c = \infty$, *we can* *formally replace c by* ∞ *in the equality: thus,*

$$\lim_{n\to\infty}\left[E(M_n)\frac{\log\log n}{\log n}-1-\frac{\log\log\log n}{\log\log n}\right]\log\log n=\infty.$$

Note. Theorem 1 remains valid when the minimum and the maximum of the X_i's are used to determine an initial interval, and the buckets are defined by dividing this interval into n equal subintervals. The density f is assumed to have support contained in $[0,1]$ but not in $[0,1-\epsilon]$ or $[\epsilon,1]$ for any $\epsilon>0$. The proof of Theorem 1 can be found in Section 2.

When f is unbounded, the theorem gives very little information about $E(M_n)$. Actually, the behavior of $E(M_n)$ depends upon a number of quantities that make a general statement all but impossible. In fact, any slow rate of convergence that is $o(n)$ is achievable for $E(M_n)$. Since N_i is binomial (n,p_i) where p_i is the integral of f over the ith bucket, we have

$$\max_i np_i\leqslant E\Big(\max_i N_i\Big)=E(M_n).$$

When f is monotone nonincreasing, the left-hand side of this inequality is equal to $nF(1/n)$ where F is the distribution function corresponding to f. Thus, since any slow rate of decrease to 0 is possible for F, when $n\to\infty$, any slow rate $o(n)$ is achievable for $E(M_n)$. The rate $\log n/\log\log n$, achieved by all bounded densities, is also a lower bound for $E(M_n)$ for all densities.

This note would not be complete if we did not mention how $E(M_n)$ varies when $\max_i np_i$ diverges. Most of this information can be deduced from the inequalities given in Theorem 2 below. For example, we will see that $E(M_n)\sim\log n/\log\log n$ (the optimal rate achievable) when $\max_i np_i=o(\log n/\log\log n)$, and that $E(M_n)\sim\max_i np_i$ when $\lim_{n\to\infty}\log n/\max_i np_i=0$. What happens when $\max_i np_i$ varies at the critical rate $\log n$ is described in Corollary 3.

THEOREM 2. *Let* $q=\max_i np_i$. *Then*

$$q\leqslant E(M_n)\leqslant q+\frac{1}{t}\left(\log n+q(e^t-1-t)\right),\qquad \textit{all } t>0,\ n\geqslant 3.$$

The proof of Theorem 2 can be found in Section 3.

COROLLARY 1. *The upper bound takes its minimal value for the solution of*

$$e^t(t-1)=\frac{1}{q}(\log n-q).$$

When q is small compared to $\log n$, *an approximate solution is given by*

$$t = \log\frac{1}{q}\log n - \log\log\frac{1}{q}\log n.$$

The upper bound is then not greater than

$$q + \frac{1}{t}\log n + \frac{1}{t}\log n/\log\frac{1}{q}\log n.$$

In particular, when $\lim_{n\to\infty} q/\log n = 0$, *then the upper bound is asymptotic to* $(\sim)$ $q + \log n/\log\log n$. *Thus, combining Theorems* 1 *and* 2, *we see that* $\lim_{n\to\infty} q/\log n = 0$ *implies*

$$\max\left(q, \frac{\log n}{\log\log n}(1+o(1))\right) \leqslant E(M_n) \leqslant \left(q + \frac{\log n}{\log\log n}(1+o(1))\right).$$

In other words, the inequality in Theorem 2 *is strong enough to show that for all f with* $q = o(\log n/\log\log n)$, $E(M_n) \sim \log n/\log\log n$.

COROLLARY 2. *Since* $e^t - 1 - t \leqslant (t^2/2)e^t$, *we see that*

$$E(M_n) \leqslant q + \frac{1}{t}\log n + q\frac{t}{2}e^t, \qquad t > 0,\ n \geqslant 3.$$

Disregarding the contribution of e^t, *this is minimal for* $t = ((2/q)\log n)^{1/2}$. *Thus, we have*

$$E(M_n) \leqslant q + 2(2q\log n)^{1/2}\exp\left(\left(\frac{2}{q}\log n\right)^{1/2}\right),$$

and the upper bound $\sim q$ *when* $\lim_{n\to\infty} q/\log n = \infty$. *In other words,* $\lim_{n\to\infty} q/\log n = \infty$ *implies*

$$E(M_n) \sim q.$$

We can conclude therefore that the inequality of Theorem 2 *is tight when either q is very small compared to* $\log n$, *or q is very large compared to* $\log n$.

COROLLARY 3. *The critical rate of increase for q is a* $\log n$ *for constant* $a > 0$. *In that case, we have*

$$E(M_n) \leqslant \left(a + \frac{1}{t} + \frac{a}{t}(e^t - 1 - t)\right)\log n$$

where t is the solution of $e^t(t-1) = (1-a)/a$. *Since* $E(M_n) \geqslant a\log n = q$, *we see that the ratio of upper bound to lower bound remains absolutely bounded uniformly over n.*

Applications

The entire discussion until this point focused around M_n in the context of searching. It goes without saying that there are numerous places where M_n is an important quantity. It should be noted that the results remain valid for an $n^{1/d} \times \cdots \times n^{1/d}$ grid of n cells on $[0,1]^d$. Such grids are frequently used in computational geometry (for a survey, see Toussaint [11]). For example, a simple algorithm for finding the convex hull of n points in $[0,1]^2$ suggested by Shamos has expected complexity $O(n) + O(E(\sqrt{n}\,M_n \log(\sqrt{n}\,M_n))) = O(n) + O(\sqrt{n}\log n)E(M_n)$ (Devroye, [2]). This is $O(n)$ whenever $E(M_n) = O(\sqrt{n}/\log n)$, i.e., whenever $\max_i np_i = O(\sqrt{n}/\log n)$.

2. Proof of Theorem 1

We will use a Poissonization device. Assume first that we have shown the statement of the theorem for M_n^* where $M_n^* = \max_i N_i^*$ and N_i^* is the number of X_i's in $X_1, \ldots, X_N$ belonging to $[(i-1)/n, i/n)$, where N is a Poisson (n) random variable independent of $X_1, X_2, \ldots$. Now, for all $\epsilon > 0$, we have

$$M_n^* \leqslant M_{n(1+\epsilon)} + nI(N \geqslant n(1+\epsilon))$$

and

$$M_n^* \geqslant M_{n(1-\epsilon)} - nI(N \leqslant n(1-\epsilon))$$

where I is the indicator function, and where $n(1+\epsilon)$ and $n(1-\epsilon)$ should be read as "the smallest integer at least equal to... ." By Chebyshev's inequality and a property of the Poisson distribution,

$$nP(|N-n| \geqslant n\epsilon) \leqslant n\frac{E\big((N-n)^4\big)}{(n\epsilon)^4} = \frac{n(n+3n^2)}{(n\epsilon)^4} \leqslant \frac{4}{n\epsilon^4}.$$

Define $a = \log c$, $b(n) = 1 + a + \log\log n + \log\log\log n$, $c(n) = (\log\log n)^2/\log n$. Thus, by assumption,

$$\begin{aligned} o(1) = E\big(M_n^*\big)c(n) - b(n) &\leqslant E\big(M_{n(1+\epsilon)}\big)c(n) + \frac{4c(n)}{n\epsilon^4} - b(n) \\ &\leqslant E\big(M_{n(1+\epsilon)}\big)c(n(1+\epsilon))\frac{c(n)}{c(n(1+\epsilon))} + o\left(\frac{1}{n}\right) \\ &\quad - b(n(1+\epsilon)) + \big(b(n(1+\epsilon)) - b(n)\big). \end{aligned}$$

Now, $b(n(1+\epsilon)) - b(n) = o(1)$, and, for n large enough, $c(n) \geqslant c(n(1+$

$\epsilon)) \geqslant c(n)\log n/\log(n(1+\epsilon)) \geqslant c(n)/(1+\epsilon/\log n)$. Thus,

$$E\left(M_{n(1+\epsilon)}\right) \geq \frac{b(n(1+\epsilon))+o(1)}{c(n(1+\epsilon))(1+\epsilon/\log n)} = \frac{b(n(1+\epsilon))+o(1)}{c(n(1+\epsilon))}.$$

Similarly, it can be shown that $E(M_n) \leq (b(n)+o(1))/c(n)$, and combining this gives us our theorem.

Lower Bounds for M_n^*

Let $\eta > 0$ be an arbitrary number, let $c = \operatorname{ess\,sup} f$, and let $\epsilon > 0$ be the solution of $\eta = -3\log(1-(2/c)\epsilon)$ (this will turn out to be a convenient choice for ϵ). Let A be the set $\{x\colon f(x) > c-\epsilon\}$, and let $\delta = \int_A dx$ (which is positive by definition of c). Finally, let $h = h_n$ be the integer part of $(b(n)-\eta)/c(n)$. We let p_i keep its meaning from the introduction, and note that the function f_n on $[0,1]$ defined by

$$f_n(x) = np_i, \qquad x \in \left[\frac{i-1}{n}, \frac{i}{n}\right)$$

is a density. Because $N_1^*, N_2^*, \dots, N_n^*$ are independent Poisson random variables with parameters $np_1, np_2, \dots, np_n$, respectively, we have the following chain of inequalities:

$$\begin{aligned} P\left(M_n^* < h\right) &= \prod_{i=1}^{n} P\left(N_i^* < h\right) \leqslant \prod_{i=1}^{n}\left(1 - P\left(N_i^* = h\right)\right) \\ &\leqslant \exp\left(-\sum_{i=1}^{n} P\left(N_i^* = h\right)\right) \\ &= \exp\left(-\sum_{i=1}^{n}(np_i)^h \frac{e^{-np_i}}{h!}\right) = \exp\left(-n\int_0^1 f_n^h \frac{e^{-f_n}}{h!}\right). \qquad (1) \end{aligned}$$

We need two facts from measure theory: first, $f_n \to f$ for almost all x (Wheeden and Zygmund [12]), and because both f_n and f are densities, this implies $\int |f_n - f| \to 0$ (Scheffé [10]). Thus,

$$\begin{aligned} \int_A \left(\frac{f_n}{c-2\epsilon}\right)^h e^{-f_n} &\geqslant e^{-c}\int_{A,\, f_n > c-2\epsilon} dx \geqslant e^{-c}\int_{A,\, |f_n - f| \leqslant \epsilon} dx \\ &= e^{-c}\left(\delta - \int_{A,\, |f_n - f| > \epsilon} dx\right) \geqslant e^{-c}\left(\delta - \int_A \frac{|f_n - f|}{\epsilon}\right) \\ &= e^{-c}(\delta - o(1)), \end{aligned}$$

and by combining this with (1), we see that

$$P\left(M_n^* < h\right) \leqslant \exp\left(-\frac{n}{h!}(c-2\epsilon)^h e^{-c}(\delta - o(1))\right). \tag{2}$$

Using Stirling's approximation $u! \sim u(\log u - 1) + \frac{1}{2}\log(2\pi u)$ as $u \to \infty$, we see that

$$\begin{aligned}
&\log\left(\frac{n}{h!}(c-2\epsilon)^h e^{-c}(\delta - o(1))\right)\\
&\quad \geqslant \frac{\log n}{\log\log n}(\log(c-2\epsilon) + \eta - \log(c) - o(1))\\
&\quad \geqslant \frac{\log n}{\log\log n}\cdot\frac{\eta}{3}, \qquad \text{all } n \text{ large enough.}
\end{aligned}$$

Thus, for all n large enough,

$$\begin{aligned}
E\left(M_n^*\right) &\geqslant hP\left(M_n^* \geqslant h\right) = h\left(1 - P\left(M_n^* < h\right)\right)\\
&\geqslant h\left(1 - \exp\left(-\exp\left(\frac{\eta}{3}\frac{\log n}{\log\log n}\right)\right)\right)\\
&\geqslant h(1 - \exp(-\exp(\log\log n)))\\
&= h\left(1 - \frac{1}{n}\right) \geqslant \left(\frac{b(n)-\eta}{c(n)} - 1\right)\left(1 - \frac{1}{n}\right)\\
&= \frac{b(n) - \eta - o(1)}{c(n)}.
\end{aligned} \tag{3}$$

Upper Bounds for M_n^*

Again, we let η be an arbitrary positive number, and choose $h = h_n$ as the integer part of $(b(n)+\eta)/c(n)$. Let $k \geqslant h$ be some integer. Then, for $h \geqslant c$,

$$\begin{aligned}
P\left(M_n^* \geqslant k\right) &\leqslant \sum_{i=1}^{n} P\left(N_i^* \geqslant k\right) \leqslant n\sum_{j \geqslant k} c^j \frac{e^{-c}}{j!}\\
&\leqslant nc^k\frac{e^{-c}}{k!}\sum_{j=0}^{\infty}\left(\frac{c}{k+1}\right)^j = nc^k\frac{e^{-c}}{k!}\cdot\frac{k+1}{k+1-c}.
\end{aligned}$$

Thus,

$$\begin{aligned}
E\left(M_n^*\right) &\leqslant h + \sum_{k=h}^{\infty} P\left(M_n^* \geqslant k\right) \leqslant h + \sum_{k=h}^{\infty} nc^k\frac{e^{-c}}{k!}\cdot\frac{k+1}{k+1-c}\\
&\leqslant h + nc^h\frac{e^{-c}}{h!}\cdot\left(\frac{h+1}{h+1-c}\right)^2.
\end{aligned} \tag{4}$$

By some straightforward analysis, one can show that

$$\log\left(nc^h\frac{e^{-c}}{h!}\right) \geqslant -(\eta + o(1))\frac{\log n}{\log\log n}$$

and that

$$\left(\frac{h+1}{h+1-c}\right)^2 = 1 + \frac{2c}{h+1} + o\left(\frac{1}{h}\right).$$

Therefore,

$$\begin{aligned} E(M_n^*) &\leqslant h + \left(1 + \frac{2c}{h} + o\left(\frac{1}{h}\right)\right)\exp\left(-(\eta + o(1))\frac{\log n}{\log\log n}\right) \\ &\leqslant h + \left(\frac{1+o(1)}{\log n}\right)\left(1 + \frac{2c}{h} + o\left(\frac{1}{h}\right)\right) \\ &\leqslant \frac{b(n)+\eta}{c(n)} + \frac{1+o(1)}{\log n} = \frac{b(n)+\eta+o(1)}{c(n)}. \qquad (5) \end{aligned}$$

But η was arbitrary. Thus, by a combination of (3) and (5), $E(M_n^*)c(n) = b(n) + o(1)$. This concludes the proof of the theorem.

3. Proof of Theorem 2

The lower bound of Theorem 2 follows directly from Jensen's inequality. To derive the upper bound, we let $U_i = N_i - np_i$, $U = \max_i U_i$. Note that U is a nonnegative random variable. We have

$$M_n \leqslant \max_i np_i + \max_i U_i = q + U.$$

For $r \geqslant 1$, we can apply Jensen's inequality again:

$$\begin{aligned} E^r(U) &\leqslant E(U^r) = E\left(\max_i U_i^r\right) \qquad (u^r \text{ is considered sign preserving}) \\ &\leqslant n\max_i E((U_i^r)_+) \leqslant n\max_i E\left(\left(\frac{r}{et}\right)^r e^{tU_i}\right), \qquad \text{all } t > 0. \end{aligned}$$

Here we used the inequality $u_+^r \leqslant (r/et)^r e^{tu}$. $t > 0$, where $u_+ = \max(u, 0)$. Also,

$$\begin{aligned} E(e^{tU_i}) &= E(e^{-tnp_i}e^{tN_i}) = e^{-tnp_i}(e^tp - 1 - p)^n \leqslant e^{np_i(e^t-t-1)} \\ &\leqslant e^{q(e^t-t-1)}. \end{aligned}$$

Thus,

$$E(M_n) \leqslant q + \left(\frac{r}{et}\right)n^{1/r}\exp\left(\frac{q}{r}(e^t - t - 1)\right).$$

This bound is minimal with respect to r when $r = \log n + q(e^t - t - 1)$ (just set the derivative of the logarithm of the second term in the bound equal to 0). Resubstitution gives the desired result. The restriction $r \geqslant 1$ forces us to choose $n \geqslant 3$.

References

1. S. G. Akl and H. Meijer, On the average-case complexity of bucketing algorithms, *J. Algorithms* **3** (1982), 9–13.
2. L. Devroye, On the average complexity of some bucketing algorithms, *Comput. Math. Appl.* **7** (1981), 407–412.
3. L. Devroye and T. Klincsek, Average time behavior of distributive sorting algorithms, *Computing* **26** (1981), 1–7.
4. W. Dobosiewicz, Sorting by distributive partitioning, *Inform. Process. Lett.* **7** (1978), 1–6.
5. G. Ehrlich, Searching and sorting real numbers, *J. Algorithms* **2** (1981), 1–14.
6. G. H. Gonnet, Expected length of the longest probe sequence in hash code searching, *J. Assoc. Comput. Mach.* **28** (1981), 289–304.
7. D. E. Knuth, "The Art of Computer Programming," Vol. 3, "Sorting and Searching," Addison–Wesley, Reading, Mass., 1973.
8. P.-A. Larson, Expected worst-case performance of hash files, *Comput. J.* **25** (1982), 347–352.
9. H. Meijer and S. G. Akl, The design and analysis of a new hybrid sorting algorithm, *Inform. Process. Lett.* **10** (1980), 213–218.
10. H. Scheffe, A useful convergence theorem for probability distributions, *Ann. Math. Statist.* **18** (1947), 434–458.
11. G. T. Toussaint, Pattern recognition and geometrical complexity, *in* "5th International Conference on Pattern Recognition, 1980," pp. 1324–1347.
12. R. L. Wheeden and A. Zygmund, "Measure and Integral," Dekker, New York, 1977.