Comp. & Maths with Appls. Vol. 7, No. 5, pp. 407–412, 1981
Printed in Great Britain.

0097-4943/81/050407-06$02.00/0
Pergamon Press Ltd.

ON THE AVERAGE COMPLEXITY OF SOME BUCKETING ALGORITHMS†

LUC DEVROYE

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

Communicated by E. Y. Rodin

(*Received January* 1981)

Abstract—Consider n independent uniform (0, 1) random variables, and let $N_1, \ldots, N_n$ be the cardinalities of the intervals $[(i-1)/n), (i/n)]$, $1 \le i \le n$. Then $E(\max_i N_i) \sim (\log n/\log \log n)$ as $n \to \infty$. This result (proved in the paper) and related results about the asymptotical behavior of $E(g(\max_i N_i))$ for increasing functions g allow us to draw some conclusions about the average complexity of some bucketing algorithms in computational geometry. We illustrate this point by showing that Shamos' unpublished bucketing algorithm for finding the convex hull of n independent identically distributed random vectors $X_1, \ldots, X_n$ in R^2 has an average complexity $0(n)$ whenever the X_i's have a bounded density with compact support.

1. INTRODUCTION

In this paper, we offer a result that may be useful in the average time analysis of certain algorithms in computational geometry. The algorithms considered here operate on a sample of size n from R^d, say, $X_1, \ldots, X_n$:

(1) Find the smallest closed rectangle R covering all the X_i's. (This takes time $0(n)$.)

(2) Divide each side of R into m equal intervals, where m is an integer such that $\alpha n \ge m^d \ge n$ for some constant $\alpha > 1$. We thus obtain m^d rectangles R_i by forming the products of all intervals. Put all the X_i's in the appropriate rectangles. (This takes time $0(n)$.)

(3) Select not more than $a_n \le m^d$ rectangles from $R_1, \ldots, R_{m^d}$ according to an *arbitrary* procedure, taking time $0(n)$.

(4) Let N be the number of X_i's in the selected rectangles. These points are further processed in time bounded by a constant times $g(N)$ where g is some function.

One algorithm that fits this description is an algorithm for finding the convex hull of $X_1, \ldots, X_n$. It was originally suggested by Shamos [1]. Let $d = 2$. In step 3, all the rows and columns of rectangles R_i are considered, and in each row (column), the extremal nonempty rectangles are marked. Thus, per row (column), 0, 1 or 2 rectangles are marked. In addition, the nonempty rectangles immediately adjacent to these (in the same row (column)) are also marked. It is clear that $a_n = 0(\sqrt{(n)})$. In step 4, Graham's convex hull algorithm is applied to the N points in the marked rectangles [2]. The function $g(u)$ is $u \log(u+1) + 1$. Theorem 1 given below shows that this algorithm takes average time $0(n)$ whenever the X_i's are independent identically distributed random variables with density f, where f is bounded and has compact support. One should note, however, that we assume that real numbers can be stored and that the standard operations, including truncation, take constant time.

Theorem

Assume that $X_1, \ldots, X_n$ are independent R^d-valued random vectors with common density f, where f is bounded and has compact support. Let $a_n \ge 1$ for all n. Assume that $g:[0,\infty) \to [0, \infty)$ satisfies: (i) g is nondecreasing; (ii) $g(x) = 0(1 + x^\beta)$ for some $\beta > 0$; (iii) for all $c > 1$ there exists $k(c) > 0$ such that $g(cx) \le k(c)g(x)$, all $x > 0$.

Then

$$E(g(N)) = 0\left(g\left(a_n \frac{\log n}{n \log \log n}\right)\right) \tag{1}$$

where the constant in "0" does not depend upon the selection procedure in step 3.

†Research of the author was supported by Quebec Ministry of Education grant FCAC-1678 and National research Council of Canada Grant A-3456.

Remark 1. (The average time $E(T)$ taken by the algorithm)

The complete algorithm takes time bounded by $0(n+E(g(N)))$. In many circumstances, $E(g(N))=o(n)$, so that $E(T)=0(n)$. For example, in Shamos' convex hull algorithm, we have

$$E(g(N))=0\left(\sqrt{(n)}\frac{\log^2 n}{\log\log n}\right)=o(n).$$

Remark 2. (The optimality of (1))

There are selection procedures in step 3 such that $E(g(N))\geq cg(a_n(\log n/\log\log n))$ for all n and some $c>0$. For example, it sufficies to let $a_n=1$ and pick the rectangle R_i with highest cardinality. The same is true if a_n varies very slowly with n and the a_n rectangles with highest cardinalities are chosen. Thus, without further restrictions in step 3, inequality (1) cannot be improved upon.

Remark 3. (Functions g)

The functions $g(u)=u^a$, $a>0$, and $g(u)=(u+1)\log(u+1)$ satisfy conditions (i), (ii) and (iii) of the theorem.

2. PROOF OF THE THEOREM

Our proof requires a thorough understanding of the Poisson distribution. The properties that are needed here are extracted in Lemmas 1 through 4.

Lemma 1.[3]

If X is a Poisson random variable with parameter λ, then for integer k,

$$P(X\geq k)\geq 1-\Phi\left(\frac{k-\lambda}{\sqrt{(\lambda)}}\right)$$

where Φ is the normal distribution function. In particular, $P(X\geq\lambda)\geq 1/2$.

Lemma 2.

If X is a Poisson random variable with parameter 1, then for integer $k\geq 1$,

$$P(X\geq k)\leq\frac{2}{ek!}\leq\frac{2}{e}\left(\frac{k}{e}\right)^{-k}(2\pi k)^{-1/2}.$$

Proof.

$$P(X\geq k)=e^{-1}\sum_{j=k}^{\infty}j!^{-1}\leq(ek!)^{-1}\sum_{j=0}^{\infty}(k+1)^{-j}$$

$$=(ek!)^{-1}\left(1+\frac{1}{k}\right)\leq 2(ek!)^{-1}\leq 2\left(\frac{k}{e}\right)^{-k}(2\pi k)^{-1/2}.$$

Here we used Stirling's inequality (see, for example, Knopp ([4], pp. 549)).

Lemma 3.[5]

If X is a Poisson random variable with parameter λ, then for integer $k\leq\lambda$,

$$P(X\geq k)\leq\sum_{j=k}^{n}\binom{n}{j}p^j(1-p)^{n-j}$$

where n is an integer and $p\in(0,1)$ is a real number such that $np=\lambda$. In particular, $P(X\geq\lambda)\leq(5/2e)<1$ when λ is integer, $\lambda>0$.

Proof. Take $n=\lambda+1$, $p=\lambda/(\lambda+1)$, $k=np=\lambda$. Then, for $\lambda\geq 2$, the Anderson–Samuels inequality implies that $P(X\geq k)\leq(k+1)p^k(1-p)+p^{k+1}=p^{k+1}((2k+1)/k)\leq(5/2)p^{k+1}\leq(5/2)\exp(-1)$. Also, $P(X\geq 1)=1-P(X=0)=1-e^{-1}$. This concludes the proof of Lemma 3.

Lemma 4.

Consider the solution of the equations

$$\Gamma(x+1)=y \tag{2}$$

or

$$\left(\frac{x}{e}\right)^x \sqrt{(2\pi x)}=y. \tag{3}$$

In both cases, as $y\to\infty$, $x\sim(\log y/\log\log y)$.

Proof. The left-hand sides of (2) and (3) can be written as $\exp(h(x))$ where h is a strictly increasing function of x when $x\geq 1$. It suffices to show that

$$\text{(i) for all } \epsilon>0, \liminf_{y\to\infty}\left[h\left((1+\epsilon)\frac{\log y}{\log\log y}\right)-\log y\right]\geq 0;$$

and that

$$\text{(ii) } \limsup_{y\to\infty}\left[h\left(\frac{\log y}{\log\log y}\right)-\log y\right]\leq 0.$$

Consider (2) (the equation (3) can be treated similarly). As $x\to\infty$, we have by Stirling's formula ([4], p. 549), $\log\Gamma(x+1)=x(\log x-1)+(1/2)\log(2\pi x)+O(1/x)$. Replace x by $(1+\epsilon)(\log y/\log\log y)$. Then

$$\log\Gamma(x+1)=(1+\epsilon)\log y-(1+\epsilon+o(1))\frac{\log y.\log\log\log y}{\log\log y}. \tag{4}$$

The right-hand side of (4) is greater than $(1+\epsilon/2)\log y$ for all y large enough when $\epsilon>0$. It is smaller than $\log y$ for all y large enough when $\epsilon=0$. This concludes the proof of Lemma 4.

We say that a random vector $N_1,\ldots,N_n$ is multinomially distributed with parameter k and equal probabilities when $N_1,\ldots,N_n$ are distributed as the number of X_i's in n intervals $(0, 1/n),\ldots,((n-1)/n,1)$ when the X_i's, $1\leq i\leq k$, are independent random variables with the uniform distribution on $(0,1)$.

Lemma 5.

Let $N_1,\ldots,N_n$ be multinomially distributed with parameter n and equal probabilities. Then, for integer $k\geq 1$,

$$P(\max_i N_i\geq k)\leq\frac{4}{e\sqrt{2\pi}}\frac{n}{\left(\frac{k}{e}\right)^k\sqrt{k}}. \tag{5}$$

Also,

$$E(\max_i N_i)\sim\frac{\log n}{\log\log n}\text{ as } n\to\infty. \tag{6}$$

Furthermore, for integer k,

$$P(\max_i N_i<k)\leq\frac{2e}{2e-5}\exp\left(-\frac{n}{ek!}\right). \tag{7}$$

Proof. Let us first prove (5) and (7). Let $Y_1, \ldots, Y_n$ be independent Poisson (1) random variables with sum S. By Lemmas 1 and 2,

$$\begin{aligned}
&P(\max_i N_i \geq k) = P(\max_i Y_i \geq k \mid S = n)\\
&\leq P(\max_i Y_i \geq k \mid S \geq n) = \frac{P(\max Y_i \geq k,\ S \geq n)}{P(S \geq n)}\\
&\leq P(\max_i Y_i \geq k)/P(S \geq n)\\
&\leq 2\,P(\max_i Y_i \geq k) \text{ (since } S \text{ is Poisson } (n) \text{ distributed)}\\
&\leq 2n\,P(Y_1 \geq k) \leq \frac{4n}{e}\left(\frac{k}{e}\right)^{-k}(2\pi k)^{-1/2}.
\end{aligned} \tag{8}$$

Similarly, by Lemma 3,

$$P(\max_i N_i < k) \leq P(\max_i Y_i < k \mid S \leq n) \leq P(\max_i Y_i < k)/P(S \leq n)$$

$$\leq \frac{2e}{2e-5}P(\max_i Y_i < k) = \frac{2e}{2e-5}\left(1 - \sum_{j=k}^{\infty}\frac{1}{ej!}\right)^n \leq \frac{2e}{2e-5}\exp\left(-\frac{n}{ek!}\right).$$

Finally, (6) follows from (5) and (7) in the following manner. Let r be the largest integer such that $(2e/(2e-5))\exp(-n/er!) \leq (1/n)$, and let x be the solution of $\Gamma(x+1) = (n/e)\log^{-1}(2en/(2e-5)) = y$. Clearly, as $n \to \infty$, $r \sim x \sim (\log y/\log\log y) \sim (\log n/\log\log n)$ because $r \leq x < r+1$. Thus,

$$E(\max_i N_i) = \sum_{k=1}^{\infty} P(\max_i N_i \geq k) \geq \sum_{k=1}^{r} P(\max_i N_i \geq k) \geq \left(1 - \frac{1}{n}\right)r \sim \frac{\log n}{\log\log n}. \tag{9}$$

Let s be the smallest integer such that $(s/e)^s\sqrt{(2\pi s)} \geq 4n/e$, and let x be the solution of $(x/e)^x\sqrt{(2\pi x)} = 4n/e$. Obviously, $x \leq s < x+1$, and $s \sim x$ as $n \to \infty$. Thus, by (5),

$$\begin{aligned}
E(\max_i N_i) &\leq s - 1 + \sum_{k=s}^{\infty}\frac{4n}{e}\left(\frac{k}{e}\right)^{-k}(2\pi k)^{-1/2} < s + \sum_{k=s}^{\infty}\left(\frac{k}{e}\right)^{s-k}\\
&\leq s + \sum_{j=0}^{\infty}\left(\frac{e}{s}\right)^j = s + \frac{s}{s-e} \sim s \sim \frac{\log n}{\log\log n}.
\end{aligned} \tag{10}$$

Remark 4.

Viktorova and Sevastyanov[6] (see also Kolchin *et al.* [7], pp. 96)) have shown that when ρ_n is a sequence of integers such that $(n/e\rho_n!)$ tends to a constant a as $n \to \infty$, then $P(\max_i N_i = \rho_n - 1) \to e^{-a}$ and $P(\max_i N_i = \rho_n) \to 1 - e^{-a}$ as $n \to \infty$, where $N_1, \ldots, N_n$ are as in Lemma 5. Thus, the limit distribution of $\max_i N_i$ is biatomic. Unfortunately, Lemma 5 does not follow from this result without further work.

Remark 5.

By arguments similar to (8) and (10) one can show that if $N_1, \ldots, N_n$ are multinomially distributed with parameter $k = cn$ ($c \geq 1$ is an integer) and with equal probabilities, then

$$E(\max_i N_i) \leq (1 + o(1))\,c\,\frac{\log n}{\log\log n}. \tag{11}$$

For $1 \le i \le c$, let $N_1(i), \ldots, N_n(i)$ be multinomial with parameter n and equal probabilities, and let all c multinomial random vectors be independent. Then $N_1, \ldots, N_n$ is distributed as $\sum_i N_1(i), \ldots, \sum_i N_n(i)$. Therefore,

$$E(\max_j N_j) \le \sum_i E(\max_j N_j(i)) = cE(\max_j N_j(1)) \le (1+o(1))c\frac{\log n}{\log\log n} \text{ (by (10))}.$$

Also,

$$P(\max_j N_j \ge k) \le cP(\max_j N_j(1) \ge k/c), \text{ integer } k. \tag{12}$$

Lemma 6.

Let a_n and g be as in the Theorem. Let $N_1, \ldots, N_n$ be a multinomial random vector with parameter cn ($c \ge 1$ is an integer) and equal probabilities. Then

$$E(g(a_n \max_i N_i)) = 0\left(g\left(a_n\frac{\log n}{\log\log n}\right)\right) + o(1).$$

Proof.

$$E(g(a_n \max_i N_i)) \le g(na_n)P(\max_i N_i > (2\beta+1)\, c\frac{\log n}{\log\log n}\Big) + g\left((2\beta+1)ca_n\frac{\log n}{\log\log n}\right). \tag{13}$$

By condition (iii), the last term of (13) is $0(g(a_n \log n/n \log\log n))$. By combining (5) and (12), it can be checked that the other term on the right-hand-side of (13) is $0(g(na_n)n^{-2\beta}(\log\log n/\log n)^{1/2}) = o(1)$. This proves Lemma 6.

Proof of the Theorem.

Let C be the smallest closed rectangle containing the support of f, and let m be even. Divide each side of C into $(m/2)$ equal intervals and consider the $m' = (m/2)^d$ rectangles T_i thus obtained. (The openness or closedness of the intervals will be irrelevant in the proof that follows.) For $(j_1, \ldots, j_d) \in \{0, 1\}^d$, consider the rectangles $T_i(j_1, \ldots, j_d)$ obtained by translating T_i in the following manner: when $j_k = 0$, do not translate T_i in the kth coordinate direction; when $j_k = 1$, translate T_i in the kth coordinate direction over the distance $(1/m)$ length kth side of C. Clearly, $T_i = T_i(0, 0, \ldots, 0)$. Also, every R_i is contained in some $T_i(j_1, \ldots, j_d)$.

For any set $B \subseteq R^d$, let $N(B)$ be the number of X_i's that fall in B. Then

$$\begin{aligned} E(g(N)) &\le E(g(a_n \max_i N(R_i)) \le E(g(a_n \max_{i,j_1,\ldots,j_d} N(T_i(j_1, \ldots, j_d)))) \\ &\le \sum_{j_1,\ldots,j_d} E(g(a_n \max_i N(T_i(j_1, \ldots, j_d)))). \end{aligned} \tag{14}$$

Without loss of generality, assume that $(j_1, \ldots, j_d) = (0, \ldots, 0)$. Let f_0 be the uniform density on C. There is a positive integer K such that f_0 can be written as a mixture $f_0 = (1/K)F + (1 - 1/K)f'$. Here f' is some residual density function. Let $Y_1, \ldots, Y_{Kn}$ be independent identically distributed random vectors with density f_0, and let the number of Y_i's that correspond to the f-part of the mixture be Z. Clearly, Z is binomial $(Kn, 1/K)$ and $P(Z \ge n) \ge 1/2$ (e.g., see Slud [8], Theorem 2.1)). For any set $B \subseteq R^d$, let $N'(B)$ be the number of Y_i's that belong to B. It is clear that for integer k,

$$P(\max_i N'(T_i) \ge k) \ge P(Z \ge n)P(\max_i N(T_i) \ge k). \tag{15}$$

Now, $N'(T_1), \ldots, N'(T_{m'})$ are multinomially distributed with parameter $Kn \le K2^d m'$ and equal probabilities. Thus, by (15) and Lemma 6,

$$E(g(a_n \max_i N(T_i))) \le 2E(g(a_n \max_i N'(T_i))) = 0\left(g\left(a_n \frac{\log n}{\log\log n}\right)\right) + o(1). \tag{16}$$

The Theorem follows from (16) and (14).

Acknowledgement—The author gratefully acknowledges discussions with Professors Selim Akl and Godfried Toussaint.

REFERENCES

1. M. I. Shamos, Seminar given at McGill University (1978).
2. R. Graham, An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, 1, 132–133 (1972).
3. H. Bohman, Two inequalities for Poisson distributions. *Skand. Aktuarietidskr.*, 46, 47–52 (1963).
4. K. Knopp, *Theorie und Anwendung der unendlichen Reihen*. Springer-Verlag, Berlin (in German) (1964).
5. T. W. Anderson and S. M. Samuels, Some inequalities among binomial and Poisson probabilities. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, Vol. 1, pp. 1–12. University of California Press (1965).
6. I. I. Viktorova and B. A. Sevastyanov, On the limit behavior of the maximum in a polynomial scheme. *Matem. Zametki* (in Russian) 1, 331–338 (1967).
7. V. F. Kolchin, B. A. Sevastyanov and V. P. Chistyakov, *Random Allocations*. Winston, Washington, D.C. (1978).
8. E. V. Slud, Distribution inequalities for the binomial law. *Ann. Probability* 5, 404–412 (1977).