COMBINATORICA
Akadémiai Kiadó — Springer-Verlag

A LOWER BOUND FOR FINDING PREDECESSORS IN YAO'S CELL PROBE MODEL

M. AJTAI

Received February 12, 1986
Revised August 31, 1987

Let L be the set consisting of the first q positive integers. We prove in this paper that there does not exist a data structure for representing an arbitrary subset A of L which uses poly $(|A|)$ cells of memory (where each cell holds $c \log q$ bits of information) and which the predecessor in A of an arbitrary $x \leq q$ can be determined by probing only a constant (independent of q) number of cells. Actually our proof gives more: the theorem remains valid if this number is less than $\varepsilon \log \log q$, that is D. E. Willard's algorithm [2] for finding the predecessor in $O(\log \log q)$ time is optimal up to a constant factor.

1. Introduction

Let $L=\{1, \ldots, q\}$ be the set of the first q positive integers and A a subset of L. We want to give a data structure (depending on A) so that for any $x \in L$ we could easily find the greatest element in A not exceeding x $(\text{pred}_A(x))$. Our data structure, together with the operations allowed on it will be within the cell probe model of A. Yao ([3]). In this model the data is stored in m cells each of length b. (In our case b will be $\log q$.) In each step we choose a cell (as a function of the contents of all previously examined cells and x) and examine its content. After l steps as a function of all of the information that we have found in the examined cells and knowing the element x, we have to tell what is $\text{pred}_A(x)$. (We will denote by F the sequence of the functions mentioned here. These functions cannot depend on A.)

If $|L|$ is sufficiently large (compared to $|A|$) and $m=|A|$, then with a constant number of probes it is possible to find $\text{pred}_A(x)$ (see [1]). Actually four probes are sufficient in this case and the algorithm works if $|A|<\log q$. Repeated application of the same algorithm gives the following: for all t there is an l so that if $|A|<(\log q)^t$ then (in a suitable data structure depending on A) $\text{pred}_A(x)$ can be found with l probes where the number of cells $m=c|A|$.

In this paper we show that for certain values of $|A|$ (compared to $q=|L|$) a constant number of probes is not sufficient for finding $\text{pred}_A(x)$ even if the memory size m is a fixed power of $|A|$. Therefore the positive results of [1] are the best possible in the following sense; for all l there is an N_0 so that if $k>N_0$ and q is sufficiently large then there is an A, with $|A|=[(\log q)^k]$ so that $\text{pred}_A(x)$ cannot be found with l probes. (Here we suppose that an F is fixed independently of A.)

AMS subject classification code (1980): 68B15

Actually we prove more. We give A randomly so that for each $S \subseteqq L$ if $|S| > q/\log q$ then there is no F so that with a probability greater than $1/(q^{\log q})$ we find $\mathrm{pred}_A(x)$ in l steps for each $x \in S$. (Here we are speaking about the probability that the algorithm works for all $x \in S$ and not a different probability for each x).

Our data structure will be a table T of size g. T has g cells each containing b bits of information, that is, a 0, 1 sequence of length b. We will denote the contents of the i-th cell by $T(i)$. If we want to answer a query using the table T as a memory, then according to the cellprobe model we go to a cell $i_0 \in \{1, \dots, g\}$ (i_0 depends on the query), then we choose the next cell as a function of the contents of the first cell and the query, we look at the second cell, etc. In each step we choose the next cell as a function of the contents of the previously examined cells and the query. After l steps we give the answer as a function of the contents of all of the examined cells and the query.

We will denote by f_i the function which determines the cell to be examined at the i-th step. So f_0 is a function of one variable and if x is a query then $f_0(x)$ is the cell what we have to examine first; that is $i_0 = f_0(x)$. f_j, $0 < j < l$, is a function of $j+1$ variables so that if x is the query and $a_0, \dots, a_{j-1}$ are the contents of the cells $i_0, \dots, i_{j-1}$ then $i_j = f_j(a_0, \dots, a_{j-1}, x)$. The function f_l gives our answer; that is f_l is a function of $l+1$ variables and if $a_0, \dots, a_{l-1}$ are the contents of the cells $i_0, \dots, i_{l-1}$ and x is the query then $f_l(a_0, \dots, a_{l-1}, x)$ is the returned value of the algorithm. We call the sequence of functions $F = \langle f_0, \dots, f_l \rangle$ a program. F contains all of the rules which describe how will we proceed to get an answer if a table T and a query x is given. We will denote by $F^T(x)$ the returned value that we get using the table T and the program F.

Example. If H is a subset of the ordered set L $x \in L$, then let $\mathrm{pred}_H(x)$ denote the greatest element of H which is not greater than x. (If there is no such element in H then $\mathrm{pred}_H(x)$ is the smallest element of L.) We want to give a program F and for all $H \subseteqq L$ a table $T(H)$ so that our procedure returns $\mathrm{pred}_H(x)$ for any query x, that is $F^{T(H)}(x) = \mathrm{pred}_H(x)$ for all $H \subseteq L$. If the number of the cells in the table can be as large as $|L|$, then it can be done trivially. Indeed for each $i \leqq q$ we may write $\mathrm{pred}_H(x_i)$ in the i-th cell of the table T_H where x_i is the i-th element of L.

We will consider questions depending on a subset H of L (as for example $\mathrm{pred}_H(x)$). If we want to answer such a question using a program which cannot depend on H then all of the information concerning H must be given in the table $T(H)$ which may depend on H but not on the query x. A cellprobe algorithm on L consists of a program F and for all $H \subseteqq L$ a table $T(H)$. If we get a query x concerning the set $H \subseteqq L$ then we go to the table $T(H)$ and perform the program F to get the answer $F^{T(H)}(x)$.

Definition. A cellprobe algorithm of type l, b, g on L is a pair $Q = \langle F, T \rangle$ where F is a program with l steps which works on tables with g cells each containing b bits of information, and T is a function defined on the set of subsets of L so that if $H \subseteqq L$ then $T(H)$ is a table of size $g \times b$. If H is a subset of L and x is an element of L then we define the returned value of Q at H, x by $Q(H, x) = F^{T(H)}(x)$.

We will prove that if the number of steps l in the cellprobe algorithm is fixed and $q=|L|$ is sufficiently large then there is no cellprobe algorithm which gives $\mathrm{pred}_H(x)$ for all $H\subseteqq L$ if the number of cells can be only a polynomial of $|H|$ and b the number of bits in each cell is at most $\log q$. (Actually our proof gives somewhat more: the theorem remains valid if $l<\varepsilon\log\log q$ where ε depends only on $\log_{|H|}g$ where g is the number of cells. So Willard's algorithm ([2]) for finding the predecessor in $O(\log\log q)$ time using $O(\log |H|)$ cells is the best possible up to a constant factor.)

We may try to prove this statement by showing that if there is a cellprobe algorithm which gives $\mathrm{pred}_H(x)$ in l steps for each H and x, then there is another one (not on L but on a smaller universe L') which works in in $l-1$ steps. Since it is easy to show that no cellprobe algorithm can work in one step this would imply our theorem.

The argument given in our proof contains a similar reduction step, but we are able to carry out this reduction only for a weaker statement. (This way we actually prove more.) We substitute the requirement $\forall H,\ x\ \ \mathrm{pred}_H(x)=Q(H,x)$ with a weaker one. Namely instead of considering all $H\subseteqq L$, we will give H randomly according to a suitably defined distribution and require that our algorithm gives $\mathrm{pred}_H(x)$ with a high probability. The other change is that we require the equality $\mathrm{pred}_H(x)=F^T(x)$ not for all x but only on a set S of size at least $q/\log q$. The following definition shows how we will use the two concepts together.

Definition. Let $Q=\langle F,T\rangle$ be a cellprobe algorithm on L, let A be a random variable each of whose values is a subset of L, $S\subseteqq L$, and let p be a real number between 0 and 1. We say that S is p-good with respect to Q, A if $P\big(\forall x\in S,\ F^{T(A)}(x)=$ $=\mathrm{pred}_A(x)\big)\geqq p$.

We will prove that if l is fixed and q is sufficiently large, then there is a random variable A so that there is no $q^{-\log q}$-good set S of size greater than $q/\log q$. (Here we assume that the number of cells is a polynomial in $|A|$ and each cell contains at most $\log q$ bits.)

2. Sketch of the Proof

First we define the random variable A. Let L^i denote the partition of L into intervals of length approximately $q^{2^{-i}}$. We will suppose that L^{i+1} is a refinement of L^i for $i=1,\dots,t$ where t is a large constant. A will depend also on a parameter k. We will suppose that that there is a cellprobe algorithm which works with A_k in l steps and we will prove that on a smaller universe there is another cellprobe algorithm which works for A_{k-1} in $l-1$ steps.

We define A_k by recursion on k. A_1 is a random subset of L with $(\log q)^5$ elements. Suppose that A_{k-1} is defined (for any ordered universe L). For each $i=1,\dots,t$ we pick $(\log q)^5$ random elements of L^i with uniform probability. Let Y be the set of all of these intervals. We will suppose that the elements of Y are pairwise disjoint. If $J\in Y$ then J is an interval and $J\in L^i$ for some $i=1,\dots,t$. The partition L^{i+1} of L induces a partition on J. We will denote this partition by J^1. (The classes of J^1 are intervals of length approximately $|J|^{1/2}$.) Let $L'=J/J^1$ be the set of classes of J^1 with the natural ordering. Now if we consider L' as a universe,

A_{k-1} is already defined on it. (Each element of such an A_{k-1} is an interval of L'.) For each $J\in Y$ let $A_{J,k-1}$ be a random variable with the distribution of A_{k-1} defined on J/J^1 so that these random variables for all $J\in Y$ are independent. Now we define A_k as the set of the smallest elements of the elements of $A_{J,k-1}$ for all $J\in Y$. The number of elements in A_k is $(\log q)^{5k} l^{k-1}$.

In our proof we will suppose that there is a cellprobe algorithm Q of depth l and an $S\subseteqq L$, $|S|\geqq q/\log q$ so that S is $q^{-\log q}$-good with respect to Q, A_k, and we will conclude that in a suitable other universe L', $|L'|=q'$, there is a cellprobe algorithm Q' of depth $l-1$ and an $S'\subseteqq L'$, $|S'|\geqq q'/\log q'$ so that S' is $(q')^{-\log q'}$-good with respect to A_{k-1}. We will reach contradiction by showing that A_1 does not have this property. We will construct Q' from Q using three different operations on cellprobe algorithms which will be denoted by Q_w, $Q|J$, and Q/V respectively. Before we continue the sketch of the proof we briefly describe these operations and their most important properties.

Q_w. Assume that Q is a cellprobe algorithm and S is a subset of L so that $|S|\geqq q/\log q$ and S is p-good with respect to Q, A_k where $p=q^{-\log q}$. If we get a query x then according to our definitions we go to the cell $f_0(x)$ and consider its contents, where f_0 is the function given in the program. The function f_0 induces a partition on the set S, namely $x, y\in S$ are in the same class if $f_0(x)=f_0(y)$. If we substitute S by a class S' of this partition then the first cell to be examined is always the same, however its contents still will depend on the value of A_k. Suppose that an S' is fixed and $z=f_0(x)$ for all $x\in S'$. Denote the contents of this cell by w. (w is a random variable.)

Assume now that a value for w is fixed. We define a new cellprobe algorithm Q_w of length $l-1$. Q_w will work the same way as Q works from the second step if the first examined cell was z and its contents w. That is $Q_w=\langle T, F'\rangle$ with $F'=$ $=\langle g_0, ..., g_{l-1}\rangle$, where $g_i(x)=f_{i+1}(w, a_0, ..., a_{i-1}, x)$ and $a_0, ..., a_{i-1}$ are the contents of the already examined cells. That is at the first step we go to the cell where we would go at the second step performing the original algorithm, in the case where the contents of the cell z is w, etc.

The new algorithm Q_w returns the same value as Q provided that $x\in S'$ and $T(A)(z)=w$. Since there are only q possibilities for the values of w we can pick w so that S' is p/q-good with respect to Q_w, A_k. Since $p=q^{-\log q}$, the parameter of "goodness" remained essentially the same however the size of S' can be essentially smaller than S. Namely if our table has g cells then we known only that it is possible to pick S' with $S/S'\leqq g$. Since we have no control on the location of the set S' (and S) it may happen that S' is an initial segment of L of size $q/(g*\log q)$, which makes the problem trivial since the random set A_k will not intersect this initial segment.

We will overcome this difficulty by considering S' in another universe where S' is comparatively bigger. We will use the other two operations to construct this universe.

Q/J. Suppose that J is an interval of L and Q is a cellprobe algorithm. We want to define a cellprobe algorithm $Q|J$ on the universe J which acts essentially the same way as Q only restricted to J. We will suppose that an $S\subseteqq L$ is given too and we try to define $Q|J$ so that if Q returns the predecessor of x in G for all $x\in S$ (for some $G\subseteqq L$) then $Q|J$ returns the predecessor of y in $G\cap J$ for all $y\in J\cap S$. To define $Q|J$ we have to give a table $T'(H)$ for each $H\subseteqq J$ and a program F'.

If $Q=\langle T, F\rangle$ then F' will be essentially the same as F with the only modification that if a value of f_i is outside J, then the corresponding value of $(f_i)'$ is the smallest element of J. The table $T'(H)$ is defined in the following way. Assume that $H\subseteqq J$. If there is a $G\subseteqq L$ so that for all $x\in S$ $Q(x, G)=\mathrm{Pred}_G(x)$ and $G\cap J=H$ then let $T'(H)=T(G)$ otherwise we can defiine $T'(H)$ arbitrarily. It is easy to see that with this definition $Q|J$ meets the given requirements. Moreover this property of $Q|J$ implies the following (which we actually will use in our proof). (See Lemma 4.)

Suppose that $J_1, \ldots, J_i$ are disjoint intervals of the ordered set L: $R_0\subseteqq L-\bigcup_{j=1}^{i} J_j$, $R_1\subseteqq J_1, \ldots, R_i\subseteqq J_i$, are independent random variables, $R=\bigcup\{R_j|j=0, 1, \ldots, i\}$, Q a cellprobe algorithm on L, and $S\subseteqq L$. If S is p-good with respect to Q, R and we have that for all $j\geqq 1$ $S\cap J$ is p_j-good with respect to $Q|J_j$, R_j (and p_j is maximal with this property) then $\prod_{j=1}^{i} p_j\geqq p$.

Q/V. Suppose that V is a partition of L into intervals (e.g. an L^i) and consider it as an ordered set with the natural ordering. We will denote this ordered set by L/V. The set S/V corresponding to S will be the set of intervals containing at least one point from S and similarly we may define the random variable A/V corresponding to A. It is easy to see that there is a cellprobe algorithm Q' on L/V with the same depth, so that if S is p-good with respect to Q, A then S/V is p-good with respect to Q', A/V. (Lemma 2.)

Now we return to the proof. Let P be the partition on S induced by f_0. We will prove a combinatorial lemma (Lemma 6) which essentially states that if we pick a random $i<t$ and then a random $J\in L^i$ (with uniform distribution in both cases), then with probability of at least $1/(4t\log q)$ there is a class C of P so that C intrseects at least $|J/L^{i+1}|/\sqrt{\log q}$ elements of J/L^{i+1}. The meaning of this statement for us is that if we consider the universe J/L^{i+1} then in this universe we will have a sufficiently, dense S', namely $(C\cap J)/L^{i+1}$ can be S'. We have that, roughly speaking, a positive proportion of the intervals in all of the L^i has the property that at least one class of S is dense in them (in this case we will say that J has the density property). This has the following consequence:

When we pick the value of A, the first step is to choose the intervals of Y. From the more than $(\log q)^5$ elements of Y with high probability at least, $(\log q)^3$ will have the above mentioned density property. Let D be the set of these intervals. In the following we will suppose that D is fixed in a suitable way, and we will consider the random variable A_k with this condition. We will denote this random variable by A_k^D. We will choose D so that S is p-good with respect to Q, A_k^D. (To make our notation simpler we omit the superscript D in the following.) Suppose $D=\{J_1, \ldots, J_s\}$, $s\geqq(\log q)^3$ and for all $1\leqq i\leqq s$ p_i is the maximal number so that $S\cap J_i$ is p_i-good with respect to $Q|J_i$, $A_k\cap J_i$. Since the random variables $A_k\cap J_i$ are mutually independent Lemma 4 (described after the definition of $Q|J$) implies that $\prod_{j=1}^{s} p_j\geqq p$. Since $p\geqq q^{-\log q}$ we have that for at least one i $p_i\geqq 1/2$. Let $J=J_i$.

$S\cap J$ is 1/2-good with respect to $Q|J, A_k\cap J$, so now we can fix a value w of f_0 so that Q_w is still $1/(2q)$-good with respect to $Q|J$, $A_k\cap J$. Let $\tilde{S}=\{x\in S\cap$

$\cap J | f_0(x) = z\}$. J has the density property so we may suppose that $\tilde{S}$ intersects at least $|J_1|\sqrt{q}$ classes of J^1. Therefore if we consider the factor universe J/J^1, then here the corresponding set $\tilde{S}/J$ will be $(q')^{-\log q'}$-good with respect to $(Q|J_w)/J^1$, $A_k \cap J/J^1$ where $q' = |J/J^1|$. According to the definition of A_k the random variable $A_{k-1} = (A_k \cap J)/J^1$, which proves our statement,

3.

Suppose now that S is p-good with respect to Q, R where Q is of type l, b, g and $R \subseteqq L$ is an arbitrary random variable. The following definition and lemma shows how can we deduce from the existence of such a Q the existence of another cellprobe algorithm $Q' (= Q_w)$ of type $l-1$, b, g which still has a p'-good set S' with respect to Q', R, and so that p/p' is not too big.

Assume that $Q = \langle F, T \rangle$, $H \subseteqq L$ and $x \in L$. When we determine the value of $F^{T(H)}(x)$ then at the first step we inspect the contents of a cell i_0 (where i_0 depends only on x). Suppose that the contents of i_0 is w. After this step we examine only $l-1$ cells and if the value of w is fixed it is like performing a cellprobe algorithm of length $l-1$. So for a fixed w we can define a cellprobe algorithm $Q_w = \langle F_w, T \rangle$ of type $l-1$, b, g where T is the same as in Q, and F_ω is the last $l-1$ steps of F using the assumption that the contents of the first examined cell is w. In the following we give a more formal definition.

Definition. Let $Q = \langle F, T \rangle$ be a cellprobe algorithm of type l, b, g on L $F = \langle f_0, \ldots, f_l \rangle$. If w is a 0, 1 sequence of length b then let $F^w = < f_0^w, \ldots, f_{l-1}^w \rangle$ where $f_{j-1}^w(z_0, \ldots, z_j) = f_j(w, z_0, \ldots, z_j)$ for $j = 1, \ldots, l$. Let $Q_w = \langle F_w, T \rangle$. Q_w is a cellprobe algorithm of type $\langle l-1, b, g \rangle$ on L.

The following lemma states that it is possible to fix w so that if $Q' = Q_w$ and S' is a subset of S where the function f_0 is constant then the ratio p/p' is at most 2^b.

Lemma 1. *Let $Q = \langle F, T \rangle$ be a cellprobe algorithm of type l, b, g on L, let $R \subseteqq L$ be a random variable, let $S \subseteqq L$, and $0 \leqq p \leqq 1$, and suppose that S is p-good with respect to Q, R. If $x_0 \in S$ and $S' = \{y \in S | f_0(x_0) = f_0(y)\}$, where $F = \langle f_0, \ldots, f_l \rangle$, then there exists a* 0, 1 *sequence w of length b so that S' is $2^{-b}p$-good with respect to Q_w, R.*

Proof. Since there are only 2^b possibilities for the contents of the cell $i = f_0(x_0)$, there is a w so that $P((T(R))(i) = w$ for all $y \in S' Q(R, y) = \text{pred}_R(y)) \geqq 2^{-b}$. Since $P(\forall y \in S' Q(R, y) = \text{pred}_R(y)) \geqq p$ we have

$$(*) \qquad P((T(R)(i) = w \text{ and for all } y \in S' Q(R, y) = \text{pred}_R(y)) \geqq 2^{-b} p.$$

$y \in S'$ implies $f_0 \equiv i$, hence from $T(R(i)) = w$ follows: for all $y \in S' Q(R, y) = Q_w(R, y)$. Thus $(*)$ is equivalent to the assertion of our Lemma. ▌

In the following we give somewhat more general definitions for $Q|J$ and Q/V than in the sketch of the proof. The following definitions have an algebraic character. If we have only the ordered set L we may define substructures and factor-structures of L. A substructure will be an interval of L with the induced ordering, a factor-structure of L will be a set of intervals which form a partition of L with the natural ordering. In both cases if a cellprobe algorithm is given in L we will define corresponding cellprobe algorithms on substructures and factor-structures in a natural (but not unique) way. We will show that these algorithms preserve and sometimes even amplify the nice properties of Q.

Definitions. 1. If W is a partition of an ordered set into intervals then on the classes of W we may define a linear ordering by $J_1 < J_2$ iff for all $x \in J_1$ and $y \in J_2$ we have $x < y$. We will denote this ordered set by L/W. If $S \subseteq L$ S/W will be the set of those elements of L/W which contain at least one point from S. If $U \subseteq L/W$ then U^{-W} will be the set of the smallest elements of the elements of U.

In the following definition we will give a factor algorithm Q/W on the factor structure. This definition also will depend on a function σ. For each element J of L/W $\sigma(J)$ will be an element of J. We will define Q/W so that $(Q/W)(U, J)$ is the unique interval containing $Q(U^{-W}, \sigma(J))$. When we will use this definition the choice of σ will depend on the set S as is described in Lemma 2 below.

2. Let $Q = \langle F, T \rangle$ be a cellprobe algorithm on L, let W be a partition of L into intervals, and let $\sigma(J) \in J$ for all $J \in W$. We define a cellprobe algorithm $Q^\sigma/W = \langle F', T' \rangle$ on L/W in the following way:

First we give the function T' that tells how we should fill out the table when a subset $U \subseteq L/W$ is given. Let $T'(U) = T(U^{-W})$. Now we have to define the program F' which specifies when a query $J \in L/W$ is given, how will we examine cells in the table $T'(U)$ and at the last step what will be our answer.

We will do everything but the last step in the same way as in the case of the original algorithms Q when the query is $\sigma(J)$ and the given subset is U^{-W}. That is we examine the cells of the table $T(U^{-W}) = T'(U)$ according to the rules given by F if the query is $\sigma(J)$. In the last step we get an answer z which is an element of L. Now our answer will be the unique element of L/W containing z. In a more formal way we can say that $F' = \langle h_0, \ldots, h_l \rangle$ where $F = \langle f_0, \ldots, f_l \rangle$ and for each $i = 0, \ldots, l-1$ we have $h_i(x_0, \ldots, x_{i-1}, J) = f_i(x_0, \ldots, x_{i-1}, \sigma(J))$ and $h_l(x_0, \ldots, x_{l-1}, J)$ is the unique interval in W containing $f_l(x_0, \ldots, x_{l-1}, \sigma(J))$.

Lemma 2. *Let L be an ordered set, W a partition of L into intervals, Q a cellprobe algorithm on L and $S \subseteq L$. There exists a map σ of L/W into L so that, for all $U \subseteq L/W$ we have: $\forall x \in S Q(U^{-W}, x) = \mathrm{Pred}_{U^{-W}}(x)$ implies $\forall J \in S/W (Q^\sigma)(U, J) = \mathrm{Pred}_U(J)$, where S/W is the set of classes of S which contains at least one point from S.*

Proof. For $J \in L/W$ let $\sigma(J)$ be an arbitrary element of $J \cap S$ provided that $J \cap S$ is nonempty; otherwise let $\sigma(S)$ be an arbitrary element of J. Our assertion follows immediately from the definition of Q^σ/W.

Corollary. *Let L be an ordered set, W a partition of L into intervals, Q a cellprobe algorithm on L, $S \subseteq L$, $R \subseteq L$ a random variable, and $0 \leqq p \leqq 1$. Suppose there is a random variable $R' \subseteq L/W$ so that $R'^{-W} = R$. If S is p-good with respect to Q, R then there is a function σ so that S/W is p-good with respect to Q^σ/W, R'.*

Proof. If a value of R satisfies the equation given in the definition of p-goodness, then according to Lemma 2 the corresponding value of R' also satisfies this equation.

If J is an interval of L and Q is a cellprobe algorithm on L then we will define a cellprobe algorithm $Q^\sigma|J$ on J, where σ is a map from the set of all subsets of J into the set of all subsets of L. The choice of σ again will depend on S. Our aim is to define $Q^\sigma|J$ so that $(Q^\sigma|J)(H,x)=Q(\sigma(H),x)$ provided that the latter is an element of J.

Definition. If $Q=\langle F,T\rangle$ is a cellprobe algorithm on L and J is an interval on L and σ is a map from the power set of J into the power set of L, then we define a cellprobe algorithm $Q^\sigma|J=\langle F',T'\rangle$ on J. Let $T'(H)=T(\sigma(H))$ for all $H\subseteqq J$. The program F' is the same as the program F except for its last step. If F gives an answer which is an element of J then this will be the answer of F' too. If the answer returned by F is not in J then the corresponding answer of F' will be the smallest element of J. That is if $F=\langle f_0,\dots,f_l\rangle$ then $F'=\langle f_0,\dots,f_{l-1},h\rangle$ where $h(\tilde{x})=f(\tilde{x})$ if $f_l(\tilde{x})\in J$ and $h(\tilde{x})=\min(J)$ otherwise.

Lemma 3. *Let J be an interval of the ordered set L, $S\subseteqq L$ and Q a cellprobe algorithm on L. There exists a map σ from the power set of J into the power set of L so that for each fixed $Y\subseteqq L$ if $\forall x\in S\,Q(Y,x)=\text{pred}_Y(x)$ then $\forall x\in J\cap S\,(Q^\sigma|J)(Y\cap J,x)=$ $=\text{pred}^J_{Y\cap J}(x)$. (For $H\subseteqq J$, pred^J_H designates the predecessor function for the universe J).*

Proof. Indeed if $H\subseteqq J$ then we will consider two possibilities:
(1) there is a $G\subseteqq L$ with $G\cap J=H$ and $\forall x\in S\cap J\,Q(G,x)=\text{pred}_G(x)$;
in this case let $\sigma(H)=G$ (for some G with this property).
(2) there is no subset of L which satisfies (1),
in that case let $\sigma(H)$ be an arbitrary subset of L.
Now let $Y\subseteqq L$ be fixed with the property $\forall x\in S\,Q(Y,x)=\text{pred}_Y(x)$. $H=Y\cap J$ obviously satisfies (1) (e.g. with $G=Y$), so $\sigma(H)=Y$ which implies our assertion. ∎

Lemma 4. *Let $J_1,\dots,J_i$ be disjoint intervals of the ordered set L; $R_0\subseteqq L-\bigcup_{j=1}^{i} J_j$, $R_1\subseteqq J_1,\dots,R_i\subseteqq J_i$, independent random variables, $R=\bigcup\{R_j|\,j=0,1,\dots,i\}$, Q a cellprobe algorithm on L, and $S\subseteqq L$.*

If S is p-good with respect to Q, R then there are $\sigma_1,\dots,\sigma_i$ so that, if for all $j\geqq 1$ $S\cap J_j$ is p_j-good with respect to $Q^\sigma|J_j$, R_j, $\sigma=\sigma_j$ (and p_j is maximal with this property) then $\prod_{j=1}^{i} p_j\geqq p$.

Proof. Apply Lemma 3 with $J\to J_j$ and let $\sigma_j=\sigma$. Suppose now that for some fixed value of the random variable R we have $\forall x\in S\,Q(R,x)=\text{pred}_R(x)$. Lemma 3. implies that for each $j=1,\dots,i$ we have $\forall x\in J_j\cap S\,(Q^\sigma|J_j)(R\cap J_j,x)=\text{pred}^{J_j}_{R\cap J_j}(x)$ The first equation holds with probability at least p, the second with probability p_j, and the random variables R_j are independent so we have $\prod_{j=1}^{i} p_j\geqq p$.

In the following we give the definition of the random variable A_k. In the sketch of the proof we said that we partition L into intervals of length approximately $|L|^{1/2}$. As we describe in the following definition we will actually partition L into intervals whose length (with one possible exception) is $[|L|^{1/2}]$ and the exceptional interval is not longer. The exceptional intervals make the definitions somewhat cumbersome, but they do not cause any serious problems in the proof.

Definition. Let L be an ordered set. L^1 will denote the partition of L into intervals, where each interval but the greatest one (according to the ordering in L/L^1) contains exactly $[|L|^{1/2}]$ elements and the greatest interval contains at most $[|L|^{1/2}]$ elements. We call the greatest interval exceptional if it contains less than $[|L|^{1/2}]$ elements. For any i we define a partition L^i of L by induction on i. Let $L^i=\{J|\text{there exists } K\in L^{i-1}, J\in K^1\}$. (That is for each interval K of L which is a class of L^{i-1} we form the partition (of K) K^1. The classes of all of these partitions together form the partition L^i.) We say that J is exceptional in L^i if either the corresponding K is exceptional in L^{i+1} or J is exceptional in K^1. Obviously $|\bigcup_i \{J\in L^i|J$ is exceptional in $L^i\}| \leqq \sum_{j=1}^{i} |L|^{1-2^{-j}} \leqq 2|L|^{1-2^{-i}}$.

Before we start the the proof of the theorem we remark that there are a few technical points where the sketch and the actual proof differ:

(1) The definition of the random variable A_k.

Since we define the random variable on different universes we will denote the random variable defined on the universe L by $A_{L,k}$. In the sketch we used a constant t as a parameter in the definition of A_k. Now, to make the formulation of the theorem simpler we will replace t by $\log\log |L|$ but in the actual proof we will use only that part of $A_{L,k}$ which occurred also in the original definition.

(2) In the original formulation of our theorem we said that each cell contains $\log |L|$ bits of information. We prove that given the existence of a cellprobe algorithm with certain properties, another cellprobe algorithm of smaller length exists on another universe L' with the same properties. Since $|L'|<|L|$ if we insist on $b=\log |L|$, then in the new universe the cell length b' can be smaller than the original cell length b, which creates considerable difficulties in the proof. To avoid this situation we will suppose that $b=c\log |L|$ for some constant c. (c does not depend on $|L|$ but it may depend on l or k.) The constant c'($b'=c'\log |L|$) can be bigger than c but it still remains a constant.

Definition. We define a random variable $A_{L,k}\subseteqq L$ for any ordered set L and positive integer k. For $k=1$ let $A_{L,k}$ be a random subset of L with $[(\log |L|)]^5$ elements. Suppose $A_{L,k-1}$ is defined for any ordered set L. Let Z be the set of all systems of intervals $I_{i,j}$ $i=1, \ldots, [\log\log |L|]$ $j=1, \ldots, [(\log |L|)^5]$ so that all $I_{i,j}$ are pairwise disjoint and for a fixed i, $I_{i,j}$ is a nonexceptional element of L_i for all j.

Let $H=\{I_{i,j}\}$ be a random element of Z with uniform distribution on Z. For each $I_{i,j}$ let us consider the ordered set $L_{i,j}=I_{i,j}/(I_{i,j})^1$. According to our inductive hypotheses $A_{L_{i,j},k-1}$ is already defined. Let us suppose that the random variables $A_{L_{i,j},k-1}$ are independent.

Let $A_{L,k}=\bigcup_{i,j} \{\min J|J\in A_{_{ij},k-1}\}$, that is the set of minimal elements of the intervals occurring in some $A_{L_{i,j},k-1}$.

Remarks. 1. In the definition of $I_{i,j}$, i is between 1 and $[\log\log|L|]$. Actually we may substitute $[\log\log|L|]$ by any function F which is not greater than $\log\log|L|$ and increasingly tends to infinity with $|L|$. We will use only values of i which do not depend on $|L|$.
2. $(\log|L|)^{5k-1}\leqq|A_{L,k}|\leqq(\log|L|)^{5k+1}$ if $|L|$ is sufficiently large compared to k.

Definition. Let l be a nonnegative integer and k a positive integer. $P_{l,k}$ will be the following assertion: there exist $c_1, c_2>0$ such that for any N_0 there is an ordered set L, $|L|>N_0$ and a cellprobe algorithm Q on L of type $\langle l, b, g\rangle$, $b=[c_2\log|L|]$, $g=[(\log|L|)^{c_1}]$ and a set $S\subseteqq L$, $|S|\geqq|L|/\log|L|$ so that S is $1/(|L|^{\log|L|})$-good with respect to Q, $A_{L,k}$.

Theorem. (a) *If* $l>0$, $k>1$ *then* $P_{l,k}$ *implies* $P_{l-1,k-1}$.
(b) $P_{0,k}$ *does not hold for any* $k>1$.

Proof. (b) Suppose $P_{0,k}$ holds and Q is an algorithm of type $\langle 0, (\log q)^{c_1}, c_2\log q\rangle$ given in the definition of $P_{0,k}$. The definition of a program $\langle f_0, ..., f_l\rangle$ implies that if $l=0$ then $F^T(x)$ does not depend on T. So we have that $Q(A_{L,k}, x)$ does not depend on $A_{L,k}$ which clearly contradicts the assumptions "S is $q^{-\log q}$-good with respect to Q, $A_{L,k}$; $|S|\geqq q/\log q$".

The following Lemma implies (a).

Lemma 5. *Given* $l>0$, $k>1$, $c_1, c_2>0$ *there exists a positive integer* u *such that if* $|L|$ *is sufficiently large and* Q *is a cellprobe algorithm on* L *of type* $\langle l, b, g\rangle$, $b=[c_2\log|L|]$, $g=[(\log|L|)^{c_1}]$, $S\subseteqq L$, $|L|/\log|L|$, *and* S *is* $|L|^{-\log|L|}$*-good with respect to* Q, $A_{L,k}$ *then there is an* $i\leqq u$ *and a nonexceptional interval* J *in* L^i *and an* $x_0\in S$ *and a* 0, 1 *sequence* w *of length* $[c_2\log|L|]$ *with the following property:*

If $S'=\{y\in S\cap J\,|\,f_0(y)=f_0(x_0)\}$ *where* $Q=\langle F, T\rangle$, $F=\langle f_0, ..., f_l\rangle$, *then for suitable* σ_1, σ_2 S'/J^1 *is* $(1/2)|L|^{-c_2}$*-good with respect to* $(Q^{\sigma_1}|J)_w^{\sigma_2}$, $A_{J,J^1,k-1}$ *and* $|S'/J^1|\geqq|J/J^1|/\log|J/J^1|$.

First we prove a combinatorial lemma necessary for the proof of Lemma 5.

Definition. Suppose $Q_1, ..., Q_{u+1}$ is a sequence of partitions of the finite set L so that for all $i=1, ..., u$ Q_{i+1} is a refinement of Q_i, $S\subseteqq L$ and P is an arbitrary partition of S. We say that P is d-dense in the set $K\in Q_i$ (where $d>1$) with respect to Q_{i+1} if there is a $C\in P$ with $|\bigcup\{K'|K'\in Q_{i+1}$ and $K'\subseteqq K$ and $K'\cap C\neq 0\}|\geqq$ $\geqq(1/d)|K|$.

Lemma 6. *For all natural numbers* t *there exists a natural number* u *such that if* L *is a finite set,* $d>1$, $Q_1, ..., Q_u$ *is a sequence of partitions of* L, Q_{i+1} *is a refinement of* Q_i *for all* $i=1, ..., u$; $S\subseteqq L$, $|S|>(1/d)|L|$ *and* P *is a partition of* S *with at most* d^t *classes, then*

$(*)$ $|\bigcup\{K|$ *there exists an* $i<u$ *with* $K\in Q_i$ *and* P *is* $\sqrt{d}$ *dense in* K *with respect to* $Q_{i+1}\}|\geqq(1/2d)|L|$.

Proof. Let us define a subset R_i of Q_i for all $i=1, ..., u$. For $i=1$ and $K\in Q_1$ let $K\in R_1$ iff P is not $\sqrt{d}$ dense in K and $K\cap(\bigcup P)\neq 0$. For $i>1$ and $K\in Q_i$ let $K\in R_i$ iff P is not $\sqrt{d}$ dense in K, $K\cap(\bigcup P)\neq 0$ and there is a $K'\in Q_{i-1}$ with $K\subseteqq K'$, $K'\in R_{i-1}$. For all $i=1, ..., u$ and $C\in P$ let $R_i(C)=\{K\in R_i|K\cap C\neq 0\}$.

First we prove that for all $C\in P$ and i

$$(1^*)\qquad \left|\bigcup\{K|K\in R_{i+1}(C)\}\right| \leqq (1/\sqrt{d})\left|\bigcup\{K|K\in R_i(C)\}\right|.$$

Indeed for all $K'\in R_i(C)$ there is a $K\in R_{i-1}(C)$ with $K'\subseteqq K$ so it is sufficient to prove that

$$(2^*)\quad \text{for all } K\in R_{i-1}(C) \text{ we have } \left|\bigcup\{K'|K'\in R_i(C) \text{ and } K'\subseteqq K\}\right| \leqq$$
$$\leqq (1/\sqrt{d})|K|.$$

$K\in R_{i-1}$ implies that P is not $\sqrt{d}$ dense in K, therefore $\left|\bigcup\{K\in Q_i|K'\subseteqq K \text{ and } K'\cap C\neq 0\}\right|\leqq 1/\sqrt{d}\,|K|$ which by the definition of R implies $(2*)$ and therefore $(1*)$. According to $(1*)$ if u is sufficiently large compared to t then $\left|\bigcup\{K|K\in R_u(C)\}\right|\leqq\leqq(1/2d^{t+1})|L|$ for all $C\in P$ that is $\left|\bigcup\{K|K\in R_u\}\right|\leqq(1/2d)|L|$.

Let $H=\{K\in Q_u|K\cap(\bigcup P)\neq 0\}$. $\left|\bigcup H\right|\geqq(1/d)|L|$ and for all $K\in H-R_u$ P is $\sqrt{d}$ dense either in K or in a set in some Q_i, $i<u$ containing K, that is we have $(*)$. ▮

Proof of Lemma 5. First we formulate 8 steps of the proof, then we prove each step separately. These steps correspond to the various stages of the proof given in the sketch.

Step 1. We will use the following notation: if i is a positive integer $n(L^i)$ will be the set of all nonexceptional intervals of L^i. Let P be the partition of S induced by f_0, that is x, y are in the same class iff $f_0(x)=f_0(y)$. Let $d(L^i)=\{J\in n(L^i)|P$ is $\sqrt{\log|L|}$ dense in J with respect to $L^{i+1}\}$. Our statement is the following:

There is a $u>0$, where u depends only on c_1 so that for some $i\leqq u$ $|d(L^i)|/|n(L^i)|\geqq 1/(4u\log|L|)$.

Step 2. Let $\{I_{i,j}\}$ be the system of random intervals given in the definition of $A_{L,k}$. There is a $u>0$ depending only on c_1 so that for some fixed $i_0\leqq u$ we have

$$P(|\{I_{i_0,j}\in d(L^{i_0})\}|\geqq(\log|L|)^3)\geqq 1-(1/2)|L|^{-\log|L|}.$$

Step 3. Assume that i_0 is the integer from Step 2. If $D\subseteqq d(L^{i_0})$ then we define a new random variable $A^D_{L,k}$ whose distribution is the same as the distribution of $A_{L,k}$ with the condition $\{I_{i_0,j}\}\cap d(L^{i_0})=D$, that is for any $X\subseteqq L$ $P(X=A^L_{D,k})==P(X=A_{L,k}|\{I_{i_0,j}\}\cap d(L^{i_0})=D)$.

Our assertion is the following: there exists a subset D of $d(L^{i_0})$ with at least $(\log|L|)^3$ elements, so that S is $(1/2)(|L|)^{-\log|L|}$-good with respect to Q, $A^D_{L,k}$.

Step 4. Let $D\subseteqq d(L^{i_0})$ as defined in Step 3. There exists a $J\in D$ so that $S\cap J$ is 1/2-good with respect to $Q^{\sigma_1}|J$, $A^D_{L,k}\cap J$ for a suitable σ_1.

Step 5. If $J\in D$ is the interval from Step 4, then there is an $x_0\in S$ so that if $S'=\{y\in S\cap J|f(x_0)=f(y)\}$, then $\left|\{K\in J^1|K\cap S'\neq 0\}\right|\geqq(\log|L|)^{-1/2}|J^1|$.

Step 6. There exists a 0, 1 sequence w of length b so that S' is $1/2|L|^{-c_2}$-good with respect to $(Q^{\sigma_1}|J)_\omega$, $A^D_{L,,k}\cap J$.

Step 7. S'/J^1 is $1/2|L|^{-c_2}$-good with respect to $((Q^{\sigma_1}|J)_w)^{\sigma_2}/J^1$, $(A^D_{L,k}\cap J)|J^1$ for a suitable σ_2.

Step 8. $(A^D_{L,k}\cap J)/J^1=A_{J/J^1,k-1}$.

We complete the proof of Lemma 5 by observing that Step 5 and Step 7 and Step 8 implies that that the random variable $A_{J/J^1,k-1}$ meets the requirements of the lemma.

Proof. *Step 1.* Let us apply Lemma 6 with $Q_i=L^i$, $d=\log|L|$, $t=c_1$. Since the exceptional intervals in $L^1, \dots, L^u$ cover only $2|L|^{1-2^{-u}}$ elements (see remark after the definition), Lemma 6 implies that $\bigcup_{i=1}^{u} \bigcup d(L^i) \geqq |L|/(3\log|L|)$ so there is an $i\leqq u$ with the required property.

Step 2. We will denote by i_0 the number i whose existence is stated in Step 1. Let U denote the set of those intervals in $\{I_{i,j}I$ which are not of the form $I_{i_0,j}$. It is enough to prove our assertion for the conditional probability with the condition $U=U_0$ for all possible values of U_0.

If $U=U_0$ is fixed then the intervals $I_{i_0,j}$ are taken randomly with uniform distribution from the set of those nonexceptional intervals of L^{i_0} which are disjoint from $\bigcup U_0$. Let B be the set of those elements of $n(L^{i_0})$ which are disjoint from all of the intervals in U_0. According to the definition of $A_{j,k}$ $|B|\geqq \geqq(1-1/(100u\log|L|))n(L^{i_0})$. Step 1 and this inequality implies that $|B\cap d(L^{i_0})|\geqq \geqq|B|/(5u\log|L|)$. Since $\{I_{i_0,j}\}$ is a random subset of B with $(\log|L|)^5$ elements it is easy to see that

$$P\left(\left|\{I_{i_0,j}\in d(L^{i_1})\}\right| \geqq (\log|L|)^3\right) \geqq 1-(1/2)|L|^{-\log|L|}.$$

Actually here we are using only the following probabilistic fact: If H is a finite set and X is a subset of $|H|$ with more than $(1/(c\log|H|))|H|$ elements and Y is a random subset of $|H|$ with $(\log|H|)^5$ elements then $P(|X\cap Y|>(\log|H|)^3)\geqq \geqq 1-(1/2)|H|^{-(\log|H|)^2}$. Here we have $|H|\geqq|L|^{1/2}$.

Step 3. According to Step 2 $P(|\{I_{i_0,j}\}\cap d(L^{i_0})|\geqq(\log|L|)^3)\geqq 1-(1/2)|L|^{-\log|L|}$. Therefore the $|L|^{-\log|L|}$ goodness of S implies that

$$P\left(\forall x\in S Q(x, A_{L,k}) = \mathrm{pred}_{A_{L,k}}(x)\,\middle|\,|\{I_{i_0,j}\}\cap d(L^{i_0})| \geqq (\log|L|)^3\right) \geqq (1/2)|L|^{-\log|L|}.$$

Since for all distinct D $(|D|\geqq(\log|L|)^3)$ the sets $\{X|X=A_{L,k}$ and $\{I_{i_0,j}\}\cap \cap d(L^{i_0})=D\}$ are disjoint we have that for at least one D (with $|D|\geqq(\log|L|)^3$):

$$P\left(\forall x\in S Q(x, A_{L,k}) = \mathrm{pred}_{A_{L,k}}(x)\,\middle|\,\{I_{i_0,j}\}\cap d(L^{i_0}) = D\right) \geqq (1/2)|L|^{-\log|L|}.$$

Step 4. Let $D=\{J_1, \dots, J_s\}$, $s\geqq(\log|L|)^3$. Apply Lemma 4 with $R_i=A^D_{L,k}\cap J_i$, $i=1, \dots, s$ $R_0=A^D_k-\bigcup D$. If for all $i=1, \dots, s$ $p_i\leqq 1/2$ then $p<(1/2)|L|^{-\log|L|}$ in contradiction to our assumption.

Step 5. Our assertion is an immediate consequence of $J\in d(L^{i_0})$.

Step 6. Apply Lemma 1 with $L \to J$, $R \to A^{D}_{L,k} \cap J$, $S \to S'$ and $Q \to (Q^{\sigma_1} | J$.

Step 7. Apply the Corollary of Lemma 2 with $L \to J$, $W \to J^1$, $R \to A^{D}_{L,k} \cap J$, $R' = A^{D}_{L,k} \cap J/J^1$, $S \to S'$.

Step 8. Our assertion follows from the recursive definition of $A_{L,k}$.

References

[1] M. Ajtai, M. Fredman and J. Komlós, Hash Functions for priority Queues, *Proceedings of the 24th Annual Symposium on FOCS*, 1983.

[2] D. E. Willard, Logarithmic worst case range queries are possible in space $O(n)$, *Inform. Proc. Letter*, **17** (1983), 81—89.

[3] A. Yao, Should tables be sorted, *JACM* **28**, 3 (*July* 1981), 615—628.

M. Ajtai

IBM Almaden Research Center