INFORMATION AND COMPUTATION **94**, 83–92 (1991)

A Lower Bound for the Integer Element Distinctness Problem*

ANNA LUBIW[†]

Department of Computer Science, University of Waterloo,
Waterloo, Canada N2L 3G1

AND

ANDRÁS RÁCZ

Department of Analysis, Eötvös University, Budapest, Hungary 1088

A lower bound of $\Omega(n \log n)$ is proved for the integer element distinctness problem—Given $(x_1, ..., x_n) \in \mathbf{Z}^n$, are the x_i's distinct—on the bounded-order algebraic decision tree model. © 1991 Academic Press, Inc.

1. INTRODUCTION

The element distinctness problem is to decide for a given $(x_1, ..., x_n) \in \Re^n$ whether all the x_i's are distinct—i.e., is $x_i \neq x_j$ for all $i \neq j$? This is a very basic decision problem easily reducible to many other decision and computation problems, for example, sorting. Thus a lower bound for element distinctness provides lower bounds for other problems.

Ben-Or (1983) proved a lower bound of $\Omega(n \log n)$ for the element distinctness problem on two models of computation both generalizing the comparison tree model. These are the *bounded-order algebraic decision tree model* and the *algebraic computation tree model.* Exact specifications of these models are given below but the main idea is to have for each n a rooted tree with leaves labelled ACCEPT and REJECT, and internal nodes labelled by arithmetic computations or comparisons so that an input $(x_1, ..., x_n)$ is accepted iff the path starting from the root and branching according to the results of the specified comparisons reaches an ACCEPT leaf. The natural measure of complexity is the height of the tree, which

* A preliminary version of this work was presented at the French-Israeli Conference on Combinatorics and Algorithms, 1988.

[†] Research supported in part by NSERC.

0890-5401/91 $3.00

corresponds to the worst case number of computations/comparisons for inputs of size n.

Ben-Or's proof, which uses a theorem of Milnor (1964) and Thom (1965) in algebraic geometry, depends crucially on the topology—specifically the number of connected components— of the subset of $\Re^n$ consisting of the vectors with distinct coordinates. One weakness of his result is the unrealistically large domain for which decision/computation trees are required to work correctly. The more standard RAM model, for example, would be expected to handle only integers. One of the motivations for proving lower bounds on decision/computation trees is the hope of carrying these lower bounds over to RAM's—at least in cases where the set of primitive operations has been restricted so that the RAM is constrained to maintain some mathematical structure. This approach was followed successfully by Paul and Simon (1982) for the problem of sorting. The best known lower bound for element distinctness on a RAM (Dietzfelbinger and Maass, 1986) was obtained by quite different methods, but unduly restricts the RAM. One might hope to get improvements by carrying over Ben-Or's results to a restricted RAM. A main barrier do doing this is the discrepancy in domains: a RAM need only work for integers whereas Ben-Or only bounds the complexity of deciding element distinctness for all real inputs.

Our main result is that for one of the tree models the complexity of element distinctness is the same for integers as for reals—more precisely:

THEOREM 1. *The height of bounded-order algebraic decision trees which correctly decide element distinctness for all integer inputs* $(x_1, ..., x_n)$ *is* $\Omega(n \log n)$.

As we learned after submitting this paper, Theorem 1 has been obtained independently by A. Yao (1989), who also proved an $\Omega(n \log n)$ lower bound for the height of algebraic decision trees deciding integer element distinctness.

Let us define the *real* [respectively *rational, integer*] *element distinctness problem* as follows: Given $(x_1, ..., x_n) \in \Re^n$ [respectively $\mathbf{Q}^n$, $\mathbf{Z}^n$] are all the x_i's distinct?

Our proof is in two parts: In Section 2 we show that the integer element distinctness problem is not easier than the rational one on the algebraic decision tree model. To do this we construct algebraic decision trees which decide the rational problem from ones which decide the integer problem without significantly increasing height or order. In Section 3 we use a modification of Ben-Or's proof for the real case to prove that any bounded-order algebraic decision trees deciding the rational element distinctness problem have height $\Omega(n \log n)$. Combining these two results yields Theorem 1. (The exact constant hidden by the "Ω" is specified in Section 3.) In Section 4 we generalize our methods to a larger class of problems, still staying with the algebraic decision tree model. In Section 5 we turn to the algebraic computation tree model, for which we carry over the second step—a lower bound for rational element distinctness—but not the first step—going from integers to rationals.

We now define precisely the two tree computation models. In these models a problem is solved by a family of trees T_1, $T_2, ...,$ where each T_n handles input vectors of n coordinates. A single tree T_n (of either type) is a rooted tree with leaves labelled ACCEPT or REJECT. The *height* of such a tree is the length of a longest path from the root to a leaf. Each internal node v of an *algebraic decision tree* T_n is labelled by a polynomial p_v in variables $x_1, ..., x_n$. Each node has one incoming edge (on the path from the root) and three outgoing edges labelled $+$, $-$, 0. Branching occurs at the node v according to whether the specified polynomial p_v evaluated at the input x is positive, negative, or zero. The *order* of an algebraic decision tree is the maximum degree of its polynomials, and a family of trees is of bounded order if the orders of its trees T_n are bounded independent of n.

An *algebraic computation tree* T_n has two kinds of internal nodes: (1) computation nodes v of out-degree 1 labelled by instructions of the form $f_v \leftarrow a \circ b$ where $\circ \in \{+, -, \times, /\}$ and where each of a, b may be a real constant, an x_i or an f_u, for u an ancestor of v; (2) comparison nodes v of out-degree 3 labelled by a single x_i or f_u, for u an ancestor of v, and with the outgoing edges labelled $+$, $-$, 0. Branching occurs at comparison nodes according to whether the specified x_i or f_u is positive, negative, or zero for the given input x. Note that we assume no zero division.

At this point it is worth noting that for either of these models there are trees of height $O(n \log n)$ to decide even the real element distinctness problem: either by sorting and then testing consecutive pairs, or, in the case of algebraic computation trees, by computing $\prod_{i \neq j} (x_i - x_j)$ using $O(n \log n)$ multiplications, and comparing the result with 0. Note that this polynomial has (unbounded) degree $\binom{n}{2}$.

Yet another tree model of computation was considered in (Moran *et al.*, 1984): decision trees in which branching at each node may depend on the result of any test of only a bounded number of inputs. Using Ramsey's Theorem an $\Omega(n \log n)$ lower bound was proved for the height of such trees correctly deciding element distinctness even for a finite (very large) set of integers.

Finally, we comment on the possibility of obtaining lower bounds for element distinctness on a RAM. Define a *restricted* RAM to operate on natural numbers; to have an infinite set of registers, indexed by natural numbers and addressable indirectly as well as directly; and to utilize branching based on comparisons, and the arithmetic operations of addition, subtraction (truncating at zero), and multiplication—each at unit cost. Forbidden are Boolean operations (on binary representations of numbers), shift operators, (integer) division, etc.

(The unrestricted use of indirect addressing on a RAM makes the element distinctness problem trivial: For each input x_i store the index i in the register indexed by x_i; but first test whether the current contents of this register provide a j with $x_j = x_i$. Thus for the purpose of lower bounds attention should be restricted to RAM programs for which the addresses of the registers used are bounded by some function of the number of inputs regardless of the actual input values.)

The proof of our present lower bound of $\Omega(n \log n)$ for the height of bounded-order algebraic decision trees deciding integer element distinctness can be shown to imply an $\Omega(n \log n)$ lower bound for element distinctness on a restricted RAM *without multiplication.* This result was obtained earlier by Dietzfelberger and Maass (1986) using entirely different methods. Yao's (1989) proof of an $\Omega(n \log n)$ lower bound for the height of algebraic computation trees deciding integer element distinctness can be used to obtain an $\Omega(n \log n)$ lower bound for element distinctness on a restricted RAM (Lubiw, manuscript).

2. Rational Element Distinctness Reduces to Integer Element Distinctness

THEOREM 2. *If there is an algebraic decision tree T of order d and height h deciding the element distinctness problem for integer inputs $(x_1, ..., x_n)$ then there is an algebraic decision tree T' deciding the element distinctness problem for rational inputs $(x_1, ..., x_n)$, and having order d and height dh.*

This implies that on the bounded-order algebraic decision tree model the integer and rational element distinctness problems have the same complexity within a constant factor.

Proof. Our starting point is the trivial observation that for any rational vector $x = (x_1, ..., x_n)$ there is a positive integer M_x^0 such that $M_x^0 x$ is an integer vector. *Furthermore x has distinct coordinates iff $M_x^0 x$ does.* The only property of the element distinctness problem that the present proof depends on is this property of invariance under integer scaling, and thus the proof applies to any decision problem with this property. See Section 4. We would like T' on a rational input x to imitate the computation of T on the integer input $M_x^0 x$, but we would like to avoid explicitly computing M_x^0 since this seems impossible on an algebraic decision tree.

Consider the tree T'' formed from T by replacing the label $p_v(x)$ at each node v of T by the label $\lim_{M \to \infty} p_v(Mx)$. T'' is no longer an algebraic decision tree, but it can compute in the same way T could: branching now depends on the sign $(+, -$ or $0)$ of $\lim_{M \to \infty} p_v(Mx)$ rather than on the sign of $p_v(x)$. Note that for a given x, $p_v(Mx)$ is a polynomial in M and thus $\lim_{M \to \infty} p_v(Mx)$ exists in the extended reals and has a well-defined sign. We claim that T'' correctly decides the rational element distinctness problem: For each $x \in \mathbf{Q}^n$ there is some $k \in \mathrm{N}$ such that for every polynomial p_v occurring in T, the sign of $p_v(kM_x^0 x)$ is the same as the sign of $\lim_{M \to \infty} p_v(Mx)$. Thus the computation of T'' on input x is the same as the computation of T on input $kM_x^0 x$. But $kM_x^0 x$ is integer-valued so T—and hence T''—correctly decides element distinctness for x.

It remains to eliminate the use of the limit operator in T'' to obtain an algebraic decision tree T'. Let $p(x)$ be a polynomial appearing in T. Rewrite $p(Mx)$ as $p_d(x)M^d + p_{d-1}(x)M^{d-1} + \cdots + p_0(x)$. Each of the p_i's is a polynomial of degree at most d. Then the sign of $\lim_{M \to \infty} p_v(Mx)$ is the sign of $p_d(x)$, or if this is zero, the sign of $p_{d-1}(x)$, or.... Create T' from T'' by (repeatedly) replacing any node with a label of the form $\lim_{M \to \infty} p(Mx)$ by a chain of $d+1$ nodes labelled by $p_d, p_{d-1}, ..., p_0$ as shown in Fig. 1.

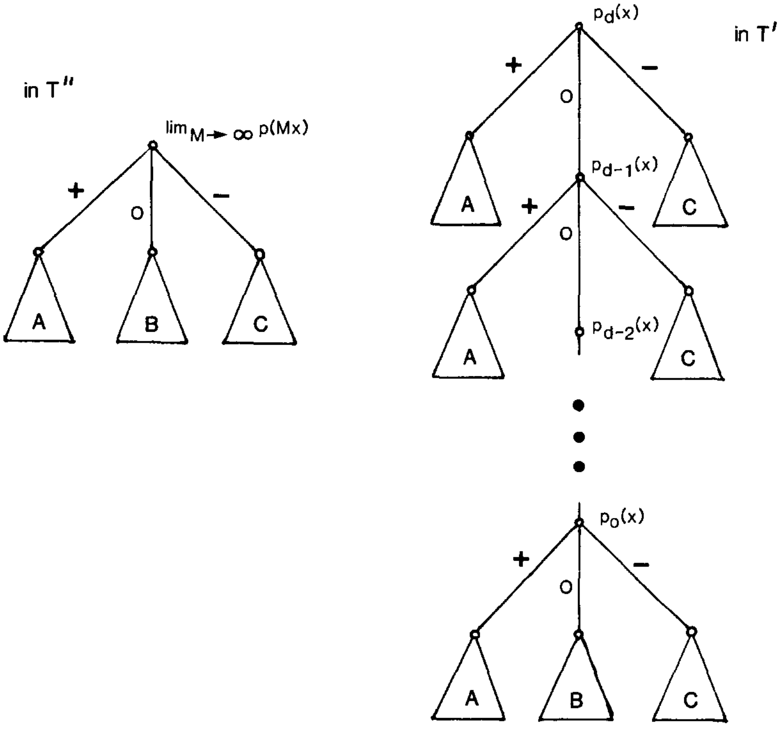

FIGURE 1

Finally, note that the node which tests $p_0(x)$ can be eliminated since there is no point in testing the sign of a constant.

The resulting algebraic decision tree T' correctly decides the rational element distinctness problem. T' has order d and height dh. ▌

The idea used in this proof—modifying an algebraic decision tree by applying limits and then expanding to tests of polynomials once more—is due to Kirkpatrick and Seidel (1986) in a different context.

3. A Lower Bound for Rational Element Distinctness

THEOREM 3. *Any algebraic decision tree of order d which decides element distinctness for all rational inputs $x_1, ..., x_n$ has height at least $k_1 n \log n - k_2 n$, where $k_1 = 1/(1+\log_2(2d-1))$ and $k_2 = 1 + (\log_2 e - 1)/(1+\log_2(2d-1))$.*

Thus any family of bounded-order decision trees deciding rational element distinctness has height $\Omega(n \log n)$.

Proof. We first review Ben-Or's lower bound proof for the real case. The solution space S for the real element distinctness problem is defined to be $\{x \in \Re^n\colon\ x_i \neq x_j \text{ for } i \neq j\}$. This solution space has $n!$ connected components each with non-null interior: specifically, for each permutation π of $\{1, 2, ..., n\}$ the set $S_\pi = \{x \in \Re^n\colon\ x_{\pi(1)} < x_{\pi(2)} < \cdots < x_{\pi(n)}\}$. Then $\bigcup_\pi S_\pi = S$.

Let T_n be an algebraic decision tree of order d solving the real element distinctness problem for input vectors of n coordinates. Let h be the height of T_n. The solution space S can be partitioned another way according to the tree T_n by grouping together vectors accepted at the same leaf of T_n: Let A be the set of accepting leaves of T_n, and for any leaf v of T_n let $S_v = \{x \in \Re^n\colon\ x \text{ ends up at leaf } v \text{ of } T_n\}$. Then $S = \bigcup_{v \in A} S_v$.

One connected component of one S_v can intersect only one S_π. Thus if each S_v consisted of one connected component then since each S_π must be intersected by some S_v, the number of accepting leaves would have to be at least $n!$, implying that h, the height of T_n, is $\Omega(n \log n)$. For $d = 1$ it is true that each S_v consists of one connected component— in fact S_v is convex—but this fails for larger d and Ben-Or applies a powerful algebraic geometry theorem of Milnor and Thom to show that each S_v—being the set of solutions to a set of at most h polynomial equalities and inequalities each of degree d—has at most $d(2d-1)^{n+h-1}$ connected components. This bound is still sufficient to give $h \geqslant \Omega(n \log n)$.

Let us now turn to the rational element distinctness problem. Suppose that the algebraic decision tree T_n only solves the rational element distinctness problem. Still for v a leaf of T_n let $S_v = \{x \in \Re^n$: x ends up at leaf v of $T_n\}$. Then we only have that $\mathbf{Q}^n \cap S = \mathbf{Q}^n \cap \bigcup_{v \in A} S_v$. The difficulty with carrying through the above proof is that one connected component of one S_v may now intersect more than one S_π. In order for this to happen S_v must contain a vector with non-distinct coordinates, and such a vector cannot be in $\mathbf{Q}^n$. What we do is show that S_v's which "cheat" in this way by accepting non-rational vectors with non-distinct coordinates are insignificant in that they do not cover completely any S_π. The remaining S_v's must then have $n!$ connected components altogether, and Ben-Or's argument applies.

We claim first that S_v does not cheat if it is open: Suppose indirectly that $y \in \Re^n$ has $y_i = y_j$ for some $i \neq j$ and $y \in S_v$ for some open S_v. We can approximate y arbitrarily closely by rational vectors $y^{(k)}$ still satisfying $y_i^{(k)} = y_j^{(k)}$, just by staying on the hyperplane $x_i = x_j$. But then the openness of S_v guarantees the existence of a rational $y^{(k)}$ in S_v, which is a contradiction. Note that this result that S_v does not cheat if it is open depends only on the property that any $y \in \Re^n$ which is outside S is a limit point of rational points outside S.

Now observe that S_v is the solution set of the polynomial equalities and inequalities determined by the vertices and edges on the path from the root of T_n to v. If all these tests are *in*equalities then S_v is open. Accordingly let us partition A into two parts: $I = \{v \in A$: the path from the root of T_n to v involves only inequalities$\}$ and $E = \{v \in A$: the path from the root of T_n to v involves at least one equality$\}$. Sets S_v, $v \in I$ do not cheat.

It remains to show that sets S_v, $v \in E$ are insignificant in the sense that no S_π is contained in $\bigcup_{v \in E} S_v$. But $\bigcup_{v \in E} S_v \subseteq \bigcup \{p^{-1}(0)$: p a polynomial in $T_n\} = q^{-1}(0)$ for $q = \prod \{p$: p a polynomial in $T_n\}$, and this latter set has emplty interior so it cannot possibly contain an S_π. (Each S_π has non-null interior.)

Therefore since each S_π intersects some S_v, $v \in I$, and no connected component of an S_v, $v \in I$ intersects more than one S_π, we can use the upper bound of $d(2d-1)^{n+h-1}$ for the number of connected components of one S_v to get

$$n! \leqslant |I| \cdot d(2d-1)^{n+h-1} \leqslant 2^h \cdot d(2d-1)^{n+h-1}.$$

Taking logarithms and using Stirling's formula yields $h \geqslant k_1 n \log n - k_2 n$, where $k_1 = 1/(1 + \log_2(2d-1))$ and $k_2 = 1 + (\log_2 e - 1)/(1 + \log_2(2d-1))$. ∎

Combining Theorems 2 and 3 proves Theorem 1—more precisely, that any algebraic decision tree of order d which decides element distinctness for all $x \in \mathbf{Z}^n$ has height at least $c_1 n \log n - c_2 n$, where $c_1 = (1/d) k_1$ and $c_2 = (1/d) k_2$, and k_1 and k_2 are as above. By noting that the bound in

Theorem 3 depends not on the height of the tree but on the height of leaves v for which S_v is open, and noting that the construction in Theorem 2, though it increases the height by a factor of d, does not increase the height of leaves v with S_v open, we obtain the better bounds $c_1 = k_1$ and $c_2 = k_2$, the same as for the rational or real element distinctness problem.

4. Lower Bounds for Other Integer Problems on the Algebraic Decision Tree Model

Any decision problem can be identified with its solution spaces $S_n \subseteq \mathfrak{R}^n$ for $n \in \mathbf{N}$, so that the decision is: given $x \in \mathfrak{R}^n$ is $x \in S_n$. The *rational* [*integer*] *version* of such a problem is to test given $x \in \mathbf{Q}^n$ [$x \in \mathbf{Z}^n$, respectively] whether $x \in S_n$.

The following two theorems give general conditions on a set $S_n \subseteq \mathfrak{R}^n$ sufficient to allow the proofs in Sections 2 and 3 to carry through.

Theorem 4. *If a decision problem has the property that its solution spaces are invariant under multiplication by positive integers then the integer version and the rational version of the problem have the same complexity* (*within a constant*) *on the bounded-order algebraic decision tree model.*

Theorem 5. *Let $S \subseteq \mathfrak{R}^n$ and denote by c the number of connected components of S which have non-null interior. Suppose that any point outside S is a limit of rational points outside S. Then any algebraic decision tree of order d which decides membership in S for all rational vectors has height at least $k_1 \log c - k_2 n$ for $k_1 = 1/(1 + \log_2(2d-1))$ and $k_2 = (\log_2(2d-1)/(1 + \log_2(2d-1))$.*

As examples of decision problems to which these theorems apply we give a subset of the examples listed by Ben-Or as applications of his general lower bound method for real-input problems. Note that a problem and its complement have the same complexity.

Set Disjointness Problem: Given two sets $A = \{x_1, ..., x_n\}$ and $B = \{y_1, ..., y_n\}$ is their intersection disjoint? In this case $S_{2n} = \{(x_1, ..., x_n, y_1, ..., y_n): x_i \neq y_j \ \forall i, j\}$. Since S_{2n} satisfies the conditions for Theorems 4 and 5, and all $(n!)^2$ connected components of S_{2n} are open and thus have non-null interior, we get a lower bound of $\Omega(n \log n)$ for the integer set disjointness problem on the bounded-order algebraic decision tree model.

Extreme Points Problem: Given a vector in $\mathfrak{R}^{2n}$ specifying n points in the plane does the convex hull of the n points have n distinct vertices? As shown in (Steele and Yao, 1982) the solution space has $(n-1)!$ connected components. These are all open since small perturbations of the vertices of a convex polygon do not change its convexity, nor its number of vertices. The conditions of Theorems 4 and 5 are met, so we get a lower bound of $\Omega(n \log n)$ for the integer extreme points problem on the bounded-order algebraic decision tree model.

Sign of an Ordering Permutation: Given $(x_1, ..., x_n) \in \Re^n$ is there a permutation of odd parity that orders the x_i's? The complementary problem has solution spaces $S_n = \{(x_1, ..., x_n)\colon x_{\sigma(1)} < x_{\sigma(2)} < \cdots < x_{\sigma(n)}$ for some even permutation $\sigma\}$. S_n satisfies the conditions of Theorems 4 and 5, and has $n!/2$ connected components, all open. So we get a lower bound of $\Omega(n \log n)$ for the integer version of the problem on the bounded-order algebraic decision tree model.

Other real-input decision problems which Ben-Or gave lower bounds for are the knapsack problem, which violates the scaling invariance property needed for Theorem 4, and the set equality problem (Is $A = \{x_1, ..., x_n\}$ equal to $B = \{y_1, ..., y_n\}$?) for which it is not clear how to profitably apply Theorem 5.

5. Algebraic Computation Trees

In this section we carry over Theorem 5—lower bounds for rational version of problems—to the algebraic computation tree model.

THEOREM 6. *Let $S \subseteq \Re^n$ and let c be the number of connected components of S which have non-null interior. Suppose that any point outside of S is a limit of rational points outside S. Then any algebraic computation tree which decides membership in S for all rational n-vectors has height at least $k_1 \log c - k_2 n$, where $k_1 = 1/(1 + \log_2 3)$ and $k_2 = \log_2 3/(1 + \log_2 3)$.*

In particular this implies a lower bound of $\Omega(n \log n)$ for the rational element distinctness problem on the algebraic computation tree model.

We note than Ben-Or states a version of this theorem but one which does not provide an $\Omega(n \log n)$ lower bound for rational element distinctness.

Proof. Essentially the same proof works. Define the S_v's, and I and E as before. The only change is that each S_v is now the solution space of a set of equalities and inequalities involving the algebraic functions of the inputs which correspond to the comparison nodes of the tree.

It is still the case that if no equalities are involved—i.e., $v \in I$—then S_v is open and cannot cheat.

If equalities are involved—i.e., $v \in E$—then $S_v \subseteq \bigcup\{f^{-1}(0)\colon f$ an algebraic function of $x_1, ..., x_n$ corresponding to a comparison node of the tree$\} = g^{-1}(0)$ for some algebraic function g. This set has empty interior, so it cannot contain any of the c connected components of S which have non-null interior. Thus the S_v's for $v \in I$ must have c connected components altogether.

To complete the proof we need a bound on the number of connected components of one S_v. Ben-Or gives a bound of $2 \cdot 3^{n+h-1}$. (This bound follows from the Milnor–Thom result though not directly). Then

$$c \leqslant |I| \cdot 2 \cdot 3^{n+h-1} \leqslant 2^h \cdot 2 \cdot 3^{n+h-1}.$$

Taking logarithms yields $h \geqslant k_1 \log c - k_2 n$ for k_1 and k_2 as given. ▌

Theorem 6 provides lower bounds for the rational versions of the problems in Section 4 on the algebraic computation tree model.

One can compute as well as decide on an algebraic computation tree. Again following examples from (Ben-Or, 1983) the present result implies lower bounds of $\Omega(n \log n)$ on the algebraic computation tree model for the problem of computing the *discriminant* $\prod_{i \neq j} (x_i - x_j)$ of rationals $x_1, ..., x_n$, and for the problem of computing the *resultant* $\prod_{i,j} (x_i - x_j)$ of rationals $x_1, ..., x_n$, $y_1, ..., y_n$. In the first case a faster algorithm would provide a fast rational element distinctness test; and in the second case a faster algorithm would provide a fast test for the rational set disjointness problem.

RECEIVED May 22, 1989; FINAL MANUSCRIPT RECEIVED March 16, 1990

REFERENCES

BEN-OR, M. (1983), Lower bounds for algebraic decision trees, *in* "Proceedings, 15th ACM Symposium on Theory of Computing," pp. 80–86.

DIETZFELBINGER, M., AND MAASS, W. (1986), Two lower bound arguments with "inaccesible" numbers, *in* "Structure in Complexity Theory" (A. Selman, Ed.), Lecture Notes in Computer Science, Vol. 223, pp. 163–183, Springer-Verlag, Berlin/New York.

KIRKPATRICK, D. G., AND SEIDEL, R. (1986), The ultimate planar convex hull algorithm? *SIAM J. Comput.* **15**, 287–299.

LUBIW, A., manuscript.

MILNOR, J. (1964), On the Betti numbers of real algebraic varieties, *Proc. Amer. Math. Soc.* **15**, 275–280.

MORAN, S., SNIR, M., AND MANBER, U. (1984), Applications of Ramsey's theorem to decision trees, *in* "Proceedings, 25th IEEE Symposium on Foundations of Computer Science," pp. 332–337.

PAUL, W., AND SIMON, J. (1982), Decision trees and random access machines, *in* "Logic and Algorithmic," Monograph 30, L'Enseignement Mathématique, pp. 331–340.

STEELE, J. M., AND YAO, A. C. (1982), Lower bounds for algebraic trees, *J. Algorithms* **3**, 1–8.

THOM, R. (1965), Sur l'homologie des variétés algébriques réelles, *in* "Differential and Combinatorial Topology" (S. S. Cairns, Ed.), Princeton Univ. Press, Princeton, NJ.

YAO, A. (1989), Lower bounds for algebraic computation trees with integer inputs, *in* "Proceedings, 30th IEEE Symposium on Foundations of Computer Science," pp. 308–313.