Peter Růžička; Juraj Wiedermann

On the lower bound for minimum comparison selection

Aplikace matematiky, Vol. 23 (1978), No. 1, 1–8

Persistent URL: http://dml.cz/dmlcz/103726

ON THE LOWER BOUND FOR MINIMUM COMPARISON SELECTION

PETER RUŽIČKA, JURAJ WIEDERMANN

(Received June 3, 1976)

1. INTRODUCTION

Given a set X of n distinct objects and an integer k, $0 < k \leqq n$, the selection problem is to determine the minimum number $V_k(n)$ of pairwise comparisons needed to select the k-th largest element of X (shortly $k\theta X$).

This problem has received considerable interest in the past few years [1, 2, 4]. The number $V_k(n)$ has been determined exactly for $k = 1$ and 2. Furthermore, $V_k(n) \leqq n - k + (k - 1) . \lceil \log_2 (n - k + 2) \rceil$ is known by Hadian and Sobel to be an upper bound for the general case, i.e. for all n and k. The search for progressively faster general methods for this problem has culminated in the linear time worst-case algorithm by Blum, Floyd, Pratt, Rivest and Tarjan. They proved an upper bound for $V_k(n)$ to be $5{\cdot}4305n$. Recently a paper by Schönhage, Paterson and Pippenger [5] appeared in which they got an upper bound of $3n$ comparisons for the median asymptotically.

Although this is the most efficient general algorithm known, there is an intuitive feeling that the constant of proportionality can be considerably improved. In order to make it possible to determine how close is the given algorithm to the optimality, lower bounds on the complexity have been examined. Actually, at present there is hardly any mathematical technique available for proving the optimality or even establishing any non-trivial lower bounds of the complexity of some of the most common combinatorial problems and thus ad hoc techniques have been devised for special problems.

Among the few general methods for specifying non-trivial lower bounds for the selection problem, a significant position is occupied by an adversary approach ascribed to Blum et al. and improved by Yao.

The best previously known lower bound up to the quaternian was published by Hyafil [2] who generalized Knuth's intriguing idea of proving the optimality $V_2(n)$ by means of the adversary strategy. In this paper a definition and some basic characteristics are given to handle the general adversary model based on the so called basic strategy and ruled by a certain sequence of constants. This approach yields a general framework for proving correctness of the whole class of adversaries and it allows us to specify a lower bound for this general class considering the sequence of constant $c_1, c_2, \ldots, c_{k-1}$ to be $V_k(n) \geqq n - 1 + \sum_{s=1}^{k-1} \lfloor \log_2 c_s \rfloor$. Further, an optimal adversary for the underlying class is discussed and also a lower bound for the optimal constants is indicated. This new lower bound surpasses the best known lower bounds up to the quaternian for infinitely many values of k and n.

2. GENERAL LOWER BOUND

We consider selection algorithms determining $k0X$ which use only binary comparisons in X. The whole selection process can be formulated as a game between the selection algorithm, the aim of which is to select $k0X$ in as few comparisons as possible, and its adversary, which tries to force the selection algorithm to make as many comparisons as possible. At each step of this game, the selection algorithm poses a comparison request between any two elements of X, say $a : b$, and the adversary responses either $a < b$ or $a > b$. The answer to the request can be completely arbitrary unless it contradicts previous responses to the comparison request. When enough information is extracted from the adversary responses to determine $k0X$, the game is over.

Various adversary strategies are known from literature. For our purposes a particular adversary strategy is described, the so called basic strategy, in which the relation between two elements is determined by means of weights of those elements.

Definition 1. Given a selection algorithm determining $k0X$, the *basic strategy (BS) adversary* $\mathscr{A}(k, n)$ is a deterministic device $\langle X, C \rangle$ where

X is a set of n elements,

C denotes a nondecreasing sequence of positive integer constants $c_1, c_2, \ldots, c_{k-1}$ satisfying $\sum_{s=1}^{k-1} c_s < n$.

The state of an adversary before the t-th comparison of the selection algorithm, $t \geqq 1$, is described by two disjoint sets L and N, where $X = L \cup N$, and by the weights $f_{t-1}(d)$ of all elements $d \in X$. Initially (i.e. for $t = 1$) L is empty and for any $d \in X$, $f_0(d) = 1$. At the t-th comparison of the selection algorithm, the input of the adversary $\mathscr{A}(k, n)$ is a pair of compared elements $a : b$, $a \neq b$, $a, b \in X$. The output of $\mathscr{A}(k, n)$ is the relation between a, b and the next state defined according to the following rules:

1. If $a, b \in N$, then either $f_{t-1}(a) > f_{t-1}(b)$ and then the relation is $a > b$ or $f_{t-1}(a) = f_{t-1}(b)$ and then the relation is arbitrary, but compatible with transitivity. In both cases, if

a) $f_{t-1}(a) + f_{t-1}(b) > c_{|L|+1}$, then $f_t(d) = f_{t-1}(d)$ for all $d \in X$, $N = N - \{a\}$, $L = L \cup \{a\}$, and a becomes the minimal element in the set L;

b) $f_{t-1}(a) + f_{t-1}(b) \leqq c_{|L|+1}$, then $f_t(a) = f_{t-1}(a) + f_{t-1}(b)$, $f_t(b) = 0$, and $f_t(d) = f_{t-1}(d)$ for all $d \neq a$, $d \neq b$, $d \in X$.

2. If $a \in L$, $b \in N$, then $a > b$ and $f_t(d) = f_{t-1}(d)$ for all $d \in X$.

3. If $a, b \in L$, then $a : b$ is known (the adversary maintains a total ordering in L after 1a) and $f_t(d) = f_{t-1}(d)$ for all $d \in X$.

To recapitulate briefly the above definition, as long as the adversary deals with elements from N, it enlarges the weigth of the greater element until the sum of weights of the compared elements is greater than $c_{|L|+1}$. Then the adversary removes the greater element from the set N and places it in the set L and then it continues in the same manner.

It is apparent that in the case $f_{t-1}(a) + f_{t-1}(b) > 0$ of the above definition, the relation between a, b cannot contradict transitivity. Responses not compatible with transitivity in the case $f_{t-1}(a) = f_{t-1}(b) = 0$ are excluded. Therefore the adversary gradually constructs a partial ordering on X.

Further we prove the BS adversary to be correct – i.e. it never terminates before determining $k - 1$ elements greater than $k\theta X$ and $k\theta X$ is not found before the adversary ceases.

Proposition 1. *If* $\sum_{s=1}^{k-1} c_s < n$ *and* $|L| < k - 1$, *then* $\sum_{a \in N} f_t(a) > \sum_{s=|L|+1}^{k-1} c_s$.

Proof. For every $t \geqq 0$ it obviously holds $\sum_{b \in L} f_t(b) \leqq \sum_{s=1}^{|L|} c_s$, and thus

$$n = \sum_{a \in N} f_t(a) + \sum_{b \in L} f_t(b) \leqq \sum_{a \in N} f_t(a) + \sum_{s=1}^{|L|} c_s .$$

Exploiting the assumption $\sum_{s=1}^{k-1} c_s < n$ we get

$$\sum_{a \in N} f_t(a) \geqq n - \sum_{s=1}^{|L|} c_s > \sum_{s=1}^{k-1-|L|} c_{|L|+s} .$$

Proposition 2. *If* $\sum_{s=1}^{k-1} c_s < n$, $|L| < k - 1$, $c_s \leqq c_{s+1}$ *for* $s = 1, \ldots, k - 2$, *then* $|L| + |P| > k - 1$ *where* $P = \{a \in N; f_t(a) > 0\}$.

Proof. Proposition 1 and the assumption imply

$$\sum_{a \in N} f_t(a) = \sum_{a \in P} f_t(a) > \sum_{s=1}^{k-1-|L|} c_{|L|+s} > (k - 1 - |L|) \cdot c_{|L|+1} .$$

Since for all $a \in P$ it holds $f_t(a) \leqq c_{|L|+1}$, for the left-hand side we have $\sum_{a \in P} f_t(a) \leqq$ $\leqq |P| \cdot c_{|L|+1}$. Combining both sides of the relations we get $(k - 1 - |L|) \cdot c_{|L|+1} <$ $< |P| \cdot c_{|L|+1}$ which implies the assertion.

Proof *of the correctnes of the BS adversary.*

Fact 1. We have to prove that as long as $|L| < k - 1$ the elements in N together have enough weight to make such responses possible which promote elements into the set L.

But this is exactly what Proposition 1 claims.

Fact 2. We have to prove that as long as $|L| < k - 1$, $k\theta X$ cannot be determined by the selection algorithm.

Let the selection algorithm find $k\theta X = a$ and let $|L| < k - 1$. Then

- if $a \in L$, then the construction of the adversary implies that a is greater than at least $|N| = n - |L|$ other elements. By the assumption $n - |L| > n - k + 1$ which yields $a \neq k\theta X$, a contradiction;
- if $a \in N$ and $f_t(a) = 0$, then by Proposition 2 there exist more than $k - 1$ elements greater than a, a contradiction;
- if $a \in P$, then since $|P| + |L| > k - 1$ and $|L| \leqq k - 2$ we have $|P| > 1$ which means that there exists also an element $b \in P$, $b \neq a$, uncompared with the element a, and thus the selection algorithm cannot know $a = k\theta X$.

Furthermore, our goal is to determine a general lower bound for *BS* ruled adversaries. Here an important role is played by the notion of the crucial comparison.

Definition 2. The *crucial comparison* for an element $a \in X$, $a \neq k\theta X$, is the first comparison $a : b$ such that $b = k\theta X$ or $a < b < k\theta X$ or $k\theta X < b < a$.

In general, the decision whether a comparison is crucial or not can be made only after performing all comparisons and selecting $k\theta X$. However, for an arbitrary $a \in X$, $a \neq k\theta X$, each algorithm selecting $k\theta X$ must determine whether $a > k\theta X$ or $a < k\theta X$. This proves

Proposition 3. *A selection algorithm has to make precisely $n - 1$ crucial comparisons to select $k\theta X$, where $|X| = n$.*

In establishing the lower bound, we start from the basic estimate $V_k(n) \geqq n - 1 + \sum_{a \in L} \lceil \log_2 f_t(a) \rceil$ following from the fact that $n - 1$ crucial comparisons as well as at least $\lceil \log_2 f_t(a) \rceil$ noncrucial comparisons for each element $a > k\theta X$ are necessary ($\lceil \log_2 f_t(a) \rceil$ is the minimal number of comparisons performed with a whose weight is $f_t(a)$).

First of all we show how the number of comparisons for a given element can be estimated.

Proposition 4. *If $f_t(a) \geqq f_t(b)$, $f_t(a) + f_t(b) > 2^j + \varepsilon$, $0 \leqq \varepsilon < 2^j$, $j > 0$, then $\lceil \log_2 f_t(a) \rceil \geqq j$.*

Proof. Note first that the assumptions yield

$$\lceil \log_2 f_t(a) \rceil > \log_2 (2^{j-1} + \varepsilon/2) .$$

Consider either $\varepsilon = 0$, and then obviously $\lceil \log_2 f_t(a) \rceil \geqq j$, or $0 < \varepsilon < 2^j$, and in that case we obtain

$$\lceil \log_2 f_t(a) \rceil > \log_2 2^{j-1}(1 + \varepsilon/2^j) = j - 1 + \varepsilon_1$$

where $0 < \varepsilon_1 < 1$; so $\lceil \log_2 f_t(a) \rceil \geqq j$.

Assuming $a \in L$ we know that there exists a comparison $a : b$ say, the t-th one, such that $f_t(a) \geqq f_t(b)$ as well as $f_t(a) + f_t(b) > c_{|L|+1} = c_i$. The constant c_i can be written in the form $c_i = 2^{\lfloor \log_2 c_i \rfloor} + \varepsilon$, where $0 \leqq \varepsilon < 2^{\lfloor \log_2 c_i \rfloor}$. In virtue of Proposition 4 it is apparent that the following assertion holds.

Consequence 1. *If $a \in L$, then $\lceil \log_2 f_t(a) \rceil \geqq \lfloor \log_2 c_i \rfloor$ for an appropriate constant c_i, $1 \leqq i \leqq |L|$.*

This consequence enables us to formulate the following important theorem.

Theorem 1. (*general lower bound*). *For any BS adversary $\mathscr{A}(k, n) = \langle X, \{c_1, \ldots, c_{k-1}\} \rangle$, it holds*

$$V_k(n) \geqq n - 1 + \sum_{s=1}^{k-1} \lfloor \log_2 c_s \rfloor .$$

3. OPTIMAL LOWER BOUND

Adversaries constructed by Blum and Hyafil form special cases of our approach˙ Considering sequences of constants $c_s = 2$ or $c_s = 2^{\lceil \log_2 n/(2(k-1)) \rceil}$ for $s = 1, 2, \ldots \ldots, k - 1$, the estimate $V_k(n) \geqq n + k - 2$ or $V_k(n) \geqq n - k + (k - 1) \cdot \lceil \log_2 n/(k - 1) \rceil$ can be reached respectively.

Now we raise the following question: Which sequence of constants maximizes the lower bound estimate? In order to answer this question, the notion of the optimal *BS* adversary is introduced.

Definition 3. *BS* adversary $\mathscr{A}(k, n) = \langle X, \{c_1, \ldots, c_{k-1}\} \rangle$ is *optimal* (with respect to the basic strategy) iff

$$\sum_{s=1}^{k-1} \lfloor \log_2 c_s \rfloor = \max \{ \sum_{s=1}^{k-1} \lfloor \log_2 a_s \rfloor ; \sum_{s=1}^{k-1} a_s < n, a_s > 0 \} . \qquad (A)$$

The following proposition gives a general characterization of the sequence of constants for an optimal *BS* adversary. Our aim is to construct the optimal constants in as simple a manner as possible; so a special case where all constants are balanced (meaning that the difference of their logarithms does not exceed one) is shown.

Proposition 5. *For any n and k there exists a nondecreasing sequence of positive constants $c_1, c_2, \ldots, c_{k-1}$, each of them being a power of two, such that the condition (A) holds iff*

$$\sum_{s=1}^{k-1} c_s = \max \{ \sum_{s=1}^{k-1} a_s ; \sum_{s=1}^{k-1} a_s < n , |\log_2 a_i - \log_2 a_j| \leqq 1, 1 \leqq i, j \leqq k - 1 \}. \qquad (B)$$

Proof.

1. From the left to the right we use an indirect argument. It is sufficient to prove that there exists no sequence of constants $c_1, \ldots, c_{k-1}$ satisfying both the assumptions of the property and the condition (A), but not the condition (B). Three cases when (B) is not fulfilled can be distinguished:

i. $\sum_{s=1}^{k-1} c_s < \max\{\sum_{s=1}^{k-1} a_s ; \sum_{s=1}^{k-1} a_s < n\}$ and all constants are balanced,

ii. $\sum_{s=1}^{k-1} c_s = \max\{\sum_{s=1}^{k-1} a_s ; \sum_{s=1}^{k-1} a_s < n\}$ and the constants are not balanced,

iii. $\sum_{s=1}^{k-1} c_s < \max\{\sum_{s=1}^{k-1} a_s ; \sum_{s=1}^{k-1} a_s < n\}$ and the constants are not balanced.

Let $C \equiv c_1, \ldots, c_{k-1}$ denote a nondecreasing sequence of powers of two satisfying the condition (A).

i. The inequality in (i) guarantees the existence of a constant $\bar{c}_i = 2 \,.\, c_i$ such that the sequence $\bar{C} \equiv c_1, \ldots, \bar{c}_i, \ldots, c_{k-1}$ is nondecreasing and balanced and satisfies $\sum_{c \in \bar{C}} c < n$. Obviously $\log_2 \bar{c}_i > \log_2 c_i$ which means that the sum of logarithms in (A) was not maximal. This leads to a contradiction if we realize that while keeping the constants balanced it is impossible to adapt the sum of logarithms to its previous value by means of decreasing the values of some constants in $\bar{C}$ (excluding the case when we get the original sequence, of course).

ii. Suppose there exists no balanced sequence of constants satisfying both the assumptions and the condition (A). We prove now that using the sequence C it is possible to construct another sequence $\bar{C}$ which either satisfies the condition (A), and thus contradicts the nonexistence of such a sequence, or does not satisfy the condition (A) which contradicts the proposition. Denote by $s \geqq 2$ the maximal integer such that $\log_2 c_i - \log_2 c_j = s$ for some indices $1 \leqq j < i \leqq k - 1$. It follows $\log_2 c_i/c_j = s$, and thus $c_i = 2^s \,.\, c_j$, $s \geqq 2$.

The case $s = 2$:

There exist i, j such that choosing $\bar{c}_i = c_i/2$, $\bar{c}_j = 2 \,.\, c_j$, the sequence $\bar{C} \equiv c_1, \ldots, \bar{c}_j, \ldots, \bar{c}_i, \ldots, c_{k-1}$ is nondecreasing. Because of $\log_2 c_i + \log_2 c_j = \log_2 \bar{c}_i + \log_2 \bar{c}_j$, such choice of constants does not violate the maximality condition for the sum of logarithms and, moreover, this pair of constants is balanced, $\log_2 \bar{c}_i - \log_2 \bar{c}_j \leqq 1$. By iterating this step a balanced sequence of constants satisfying (A) can be constructed, a contradiction.

The case $s > 2$:

Let $\bar{c}_i = c_i/2$, $\bar{c}_j = 4 \,.\, c_j$ and let $\bar{C} \equiv c_1, \ldots, \bar{c}_j, \ldots, \bar{c}_i, \ldots, c_{k-1}$ be a sequence rearranged into the nondecreasing order. Now it holds $c_i + c_j = \bar{c}_i + \bar{c}_j + c_j \,.\, (2^{s-1} - 3) > \bar{c}_i + \bar{c}_j$, $s > 2$ and $\log_2 c_i + \log_2 c_j < \log_2 \bar{c}_i + \log_2 \bar{c}_j$, which contradicts (A).

iii. Balancing the sequence of constants as in the case $s = 2$ of 1ii, the proof can be transformed to that of 1i.

2. The converse assertion is proved similarly as 1i by an indirect argument.

In order to determine constants of an optimal *BS* adversary it is sufficient to consider only those which are balanced and are powers of two; hence $\log_2 c_i/c_j \leqq 1$, and thus $1 \leqq c_i/c_j \leqq 2$ for $j \leqq i$. Starting from the Hyafil constants we get

Consequence 2. *For any n and k, $k \leqq n$, the sequence of constants $c_1, \ldots, c_{k-1}$ for an optimal BS adversary is given by*

$$c_i = 2^{d-1} \qquad i = 1, 2, \ldots, 2 \,.\, (k-1) + \lceil n \,.\, 2^{-d+1} \rceil = r \,,$$
$$c_j = 2^d \qquad j = r+1, \ldots, k-1 \,,$$

where $d = \lceil \log_2 n/(k-1) \rceil$.

Theorem 2. *The minimal number of comparisons necessary to compute the k-th largest element from a linearly ordered set of n elements is bounded from below by the function*

$$B_k(n) = n - 2k + d(k-1) + \lceil n \, 2^{-d+1} \rceil$$

$k \geqq 2$, *where* $d = \lceil \log_2 n/(k-1) \rceil$.

4. ANALYSIS OF THE LOWER BOUND

Denote the upper bound for $V_k(n)$ due to Hadian and Sobel by $HS_k(n)$, and let $\Omega_k(n) = HS_k(n) - B_k(n)$. In the case of the second element, $B_2(n)$ is optimal, i.e. $\Omega_2(n) = 0$ for $n \geqq 2$. In the case of the third element our general lower bound is equal to the result in [4] where the case $k = 3$ was specially analyzed:

$$\Omega_3(n) = \begin{cases} 0 & n = 2^s + 1 \,, \\ 2 & 2^s + 1 < n \leqq 3 \cdot 2^{s-1} \qquad s \geqq 1 \,, \\ 1 & 3 \cdot 2^{s-1} < n \leqq 2^{s+1} \,. \end{cases}$$

By Theorem 2 we get for the case $k = 4$:

$$\Omega_4(n) = \begin{cases} 3 & 3 \cdot 2^{s-1} \leqq n < 2^{s+1} \,, \\ 2 & 2^{s+1} \leqq n \leqq 2^{s+1} + 2 \,, \\ 5 & 2^{s+1} + 2 < n < 5 \cdot 2^{s-1} \,, \\ 4 & 5 \cdot 2^{s-1} \leqq n < 3 \cdot 2^s \,. \end{cases} \qquad s \geqq 1 \,,$$

The best general lower bound previously known for $3 < k < n/4$ is due to Hyafil [2]:

$$H_k(n) = n - k + (k-1) \,.\, \lceil \log_2 n/(k-1) \rceil \,.$$

Proposition 6. *Within the interval* $n/2^{s+1} + 1 < k \leqq n/2^s$, $s = 2, 3, \ldots$ *it holds*

$$1 \leqq B_k(n) - H_k(n) < k - 1$$

where the right-hand side of the inequality cannot be improved.

Thus, the new lower bound $B_k(n)$ is strictly greater than the best lower bounds previously known for an infinite number of values of n and k.

References

[1] *M. Blum, R. W. Floyd, V. Pratt, R. Rivest, R. Tarjan:* Time bounds for selection. JCSS 7 (Aug. 1973), 448—461.

[2] *L. Hyafil:* Bounds for selection. IRIA Rapport de Recherche n° 77. 1974.

[3] *D. E. Knuth:* The art of computer programming. Volume 3. Sorting and Searching. Addison-Wesley publishing company. 1973.

[4] *V. Pratt, F. F. Yao:* On lower bounds for computing the i-th largest elements. Proceedings of the Fourteenth Symposium on Switching and Automata Theory. 1973.

[5] *A. Schönhage, M. Paterson, N. Pippenger:* Finding the median. Theory of Computation Report. The University of Warwick. 1975.

Súhrn

O DOLNOM ODHADE POČTU POROVNANÍ PRE ALGORITMUS VÝBERU

PETER RUŽIČKA, JURAJ WIEDERMANN

Minimálny počet porovnaní nutný na výpočet k-teho najväčšieho z n prvkovej množiny je zdola ohraničený funkciou

$$n - 2 \,.\, k + d \,.\, (k - 1) + \lceil n \,.\, 2^{-d+1} \rceil$$

$k \geqq 2$, kde $d = \lceil \log_2 n/(k - 1) \rceil$. Nový dolný odhad je na intervale $3 < k < n/4$ lepší ako najlepšie známe odhady, a to pre nekonečne mnoho hodnôt k a n.

Authors' addresses: RNDr. *Peter Ružička*, RNDr. *Juraj Wiedermann*, Výskumné výpočtové stredisko, Dúbravská 3, 885 31 Bratislava