THE COMPLEXITY OF SUMMATION

Yu. Kudryashov, A. Skopenkov
with participation of I. Nikokoshev.

Dedicated to the centenary of A. N. Kolmogorov

Introductory problems.

Задача 1. Let $x_1, \dots, x_{99}$ be real numbers. It is allowed to calculate the sum of any two of numbers obtained and to memorize this result. Find the minimal number of summations sufficient to calculate

(a) the sum $x_1 + \dots + x_{99}$;

(b) the sums $x_1 + \ldots + x_{66}$ and $x_{34} + \ldots + x_{99}$.

Denote $\mathbb{Z}_2 = \{0, 1\}$. Let us introduce the sum (modulo 2) operation on this set as follows:

$$0 + 0 = 1 + 1 = 0, \quad 1 + 0 = 0 + 1 = 1.$$

From now on we shall consider variables from $\mathbb{Z}_2$ (except for Problem 3).

Suppose that we have a set of variables $x_1, \dots, x_n$ from $\mathbb{Z}_2$. It is allowed to calculate the sum of any two of expressions obtained and to memorize this result.

Задача 2. Find the minimal number of summations sufficient to calculate

(a) the sum $x_1 + \dots + x_{99}$;

(b) the sums $x_1 + \ldots + x_{66}$ and $x_{34} + \ldots + x_{99}$.

Задача 3. Suggest a set of sums such that the minimal number of summations sufficient to calculate all sums from this set for *real variables* is not equal to the minimal number of summations sufficient to calculate all sums from this set for $\mathbb{Z}_2$-*variables*.

Задача 4. (a) Prove that any pair of sums $x_{i_1} + \dots + x_{i_p}$ and $x_{j_1} + \dots + x_{j_q}$ made of from variables $x_1, \dots, x_n$ can be calculated using at most $n - 1$ summations.

(b) Prove that any triple of sums $x_{i_1} + \dots + x_{i_p}$, $x_{j_1} + \dots + x_{j_q}$ and $x_{k_1} + \dots + x_{k_r}$ made of variables $x_1, \dots, x_n$ can be calculated using at most $n + 1$ summations.

Задача 5. Prove that any set of sums of the form $x_{i_1} + \dots + x_{i_p}$ made of variables $x_1, \dots, x_n$ can be calculated using at most $2^n - (n + 1)$ summations.

Задача 6. Suppose that any set of m sums made of n variables can be calculated using at most l summations. Prove that

(a) any set of km sums made of n variables can be calculated using at most kl summations;

(b) any set of m sums made of kn variables can be calculated using at most $k(l + m)$ summations.

Задача 7. (a) Prove that for big enough m and n any set of m sums made of n variables can be calculated using at most

$$\frac{mn}{100}$$

summations.

(b) Prove that any set of m sums made of n variables can be calculated using at most

$$2m \left\lceil \frac{n}{\lfloor \log_2 m \rfloor} \right\rceil$$

summations. Here $\lfloor x \rfloor$ is the integer part of x, i.e. the greatest integer not exceeding x, and $\lceil x \rceil$ is the 'upper integer part' of x, i.e. the minimal integer which is greater than or equal to x.

A sequence of calculating sums can be represented as follows. We have a number of 'summation elements', each having two inputs and one output, that is equal to the sum (modulo 2) of the inputs. Using such elements we construct a scheme S realizing the given set of sums.

The number of summations is equal to the number of elements in S. This number is called *the complexity of* S. The minimal complexity of a scheme realizing the given set M of sums is called *the complexity of* M and is denoted by $\mathrm{L}(M)$. (In Problem 2 you have been asked to calculate $\mathrm{L}(\{x_1 + \dots + x_{99}\})$ and $\mathrm{L}(\{x_1 + \ldots + x_{66}, x_{34} + \ldots + x_{99}\})$.)

The minimal number l such that any set of m sums made of variables $x_1, \dots, x_n$ can be calculated using at most l summations is called *the complexity of the pair* (m, n) and is denoted by $\mathrm{L}(m, n)$.

In this notation Problem 4 asserts that $\mathrm{L}(2,n) \leqslant n-1$ and $\mathrm{L}(3,n) \leqslant n+1$, while Problem 7(b) asserts that

$$\mathrm{L}(m,n) \leqslant 2m \left\lceil \frac{n}{\lfloor \log_2 m \rfloor} \right\rceil .$$

A set of m sums can be denoted by a $m \times n$-matrix (i.e. a table) M, in which M_{ij} (the element situated at the intersection of the i-th row and the j-th column) equals 1, if the j-the variable is contained in the i-th sum, and 0 otherwise. Then the i-th sum can be written as follows:

$$s_i = \sum_{j=1}^{n} M_{ij} x_j .$$

In this notation the sets of sums from Problem 2 appear as follows:

$$\begin{pmatrix} 1 & \dots & 1 \end{pmatrix}, \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 1 & \dots & 1 \end{pmatrix},$$

Consider the set of m sums made of variables $x_0, \dots, x_{2^m-1}$, in which all variables are contained in a distinct way. Let B_m be the corresponding matix. The columns of this matrix are formed by binary expansions of all integers from 0 to $2^m - 1$. Formally, denote by $j = b_{m-1,j}2^{m-1} + \dots + b_{0,j}2^0$ binary expansion of number $j < 2^m$. Then $B_m = (b_{i,j})$. E.g.

$$B_1 = \begin{pmatrix} 0 & 1 \end{pmatrix}, B_2 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}, B_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Задача 8. Prove that $\mathrm{L}(B_m) < 2^{m+1}$.

Задача 9. Prove that $\mathrm{L}(m,n) < n + 2^{m+1}$.

Задача 10. Prove that for sufficiently large n

$$\mathrm{L}(n,n) < \frac{n^2}{0,9 \log_2 n}.$$

THE COMPLEXITY OF SUMMATION (Yu. Kudryashov, A. Skopenkov)

Problems suggested after the semifinal.

Задача 11. (a) Prove that there exists at most $(n+k)^{m+2k}$ distinct schemes of k summation elements having n inputs and m outputs.

(b) If $k = \mathrm{L}(m,n)$ then $mn \leqslant (m+2k)\log_2(n+k)$.

(c) Prove that for sufficiently large n

$$\mathrm{L}(n,n) \geqslant \frac{n^2}{5\log_2 n}.$$

Hint. If $k = \mathrm{L}(n,n) < \frac{n^2}{5\log_2 n}$ then $n^2 \leqslant (n + \frac{2n^2}{5\log_2 n})\log_2(n + \frac{n^2}{5\log_2 n})$, which is not true for sufficiently large n.

(d) Prove that for all $\varepsilon > 0$ and for sufficiently large n we have

$$\mathrm{L}(n,n) \geqslant \frac{n^2}{(4+\varepsilon)\log_2 n}$$

For next statements we shall introduce some defintions.

We say that the sequence $a_n > 0$ is *asymptotically less than* a sequence $b_n > 0$ if for all $\varepsilon > 0$ there exists at most finite set of numbers k such that $a_k > (1+\varepsilon)b_k$. Notation $a_n \lesssim b_n$.

In this notaion the statement of the problem 11(d)(d) is $\mathrm{L}(n,n) \gtrsim \frac{n^2}{4\log_2 n}$.

If $a_n/b_n \to 1$ as $n \to \infty$, then we say that sequences $a_n > 0$ and $b_n > 0$ *are equivalent* and write $a_n \sim b_n$.

If $a_n/b_n \to 0$ as $n \to \infty$, then we say that the sequence $a_n > 0$ *is much smaller than* the sequence $b_n > 0$ and write $a_n \ll b_n$.

Задача 12. (a) Suppose that $a_n \lesssim b_n$ and $b_n \lesssim c_n$. Prove that $a_n \lesssim c_n$.

(b) Prove that $a_n \sim b_n$ if and only if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Задача 13. Suppose that $\log_2 n \ll f(n) \ll 2^n$. Then

$$\frac{nf(n)}{\log_2(nf(n))} \lesssim \mathrm{L}(f(n),n).$$

Задача 14. Prove that for $\log_2 n \ll f(n) \ll 2^n$ we have

$$\mathrm{L}(f(n),n) \lesssim \frac{nf(n)}{\log_2 f(n)}.$$

Задача 15. Prove that for $\log_2 n \ll f(n) \ll 2^n$ we have

$$\mathrm{L}(f(n),n) \lesssim \frac{nf(n)}{\log_2 n}.$$

Problems 13, 14 and 15 assert that for $\log_2 n \ll f(n) \ll 2^n$

$$\frac{nf(n)}{\log_2(nf(n))} \lesssim \mathrm{L}(f(n),n) \lesssim \frac{nf(n)}{\max(\log_2 n, \log_2 f(n))}.$$

Задача 16. (a) Consider a matrix A of the size $m \times n$, which contains no zero or unit lines or columns. Is it true that $\mathrm{L}(A) + m = \mathrm{L}(A^t) + n$, where A^t is obtained from A by symmetry with respect to the main diagonal?

(b) Is it true that $\mathrm{L}(m,n) + m = \mathrm{L}(n,m) + n$?

(c) Find $L(B_m)$.

Задача 17. Consider the set of variables $x_{\alpha_1,\dots,\alpha_n}$ numbered by vertices $(\alpha_1,\dots,\alpha_n)$ of n-dimensional cube $\mathbb{Z}_2^n$. It is required to find sums of variables in all those faces of this cube that contain the vertex $(0,\dots,0)$ (in particular, the sum of all variables is to be found). In other words, for each $(\alpha_1,\dots,\alpha_n) \in \mathbb{Z}_2^n$ it is required to find $\sum\limits_{\forall i: \beta_i \leqslant \alpha_i} x_{\beta_1\dots\beta_n}$. What is the minimal number of summations required?

Задача 18. State and prove analogous results vor variables from $\mathbb{R}$ in the following cases:

It is required to find a collection of sums or of linear combinations $a_1x_1 + \dots + a_kx_k$ with integers $a_1,\dots,a_k$.

There could be either only summation elements, or summation elements together with multiplication (by all integer constants) elements.

THE COMPLEXITY OF SUMMATION

Yu. Kudryashov, A. Skopenkov (with participation of I. Nikokoshev).

Solutions (the first part).

Задача 1. (a) The answer is 98. Clearly, the sum $x_1 + \dots + x_{99}$ can be found using at least 98 summations. Suppose that there is a scheme of a summations, realizing our sum. Consider a graph having 99 vertices $P_1, \dots, P_{99}$ and for each summations having the sum $x_{i_1} + \dots + x_{i_p}$ on the first entry and the sum $x_{j_1} + \dots + x_{j_q}$ on the second entry join the vertex P_{i_1} to the vertex P_{j_1} by an edge. Clearly, the obtained graph is connected. It follows that this graph has at least 98 edges, i.e. $a \geqslant 98$.

(b) The answer is 98. The proof is analogous to (a).

Задача 2. The solution is the same as that of 1.

Задача 3. One can take, e.g. the set

$$\{\ x_1 + x_2,\ x_1 + x_2 + x_3,\ x_1 + x_2 + x_3 + x_4,\ x_2 + x_3 + x_4\ \}.$$

Indeed, for the $\mathbb{Z}_2$ case we can calculate these sums in the given order (without calculation of any additional sums), while in the $\mathbb{R}$ case we need to calculate at least one additional sum.

Задача 4. (a) Denote by $S(A)$ the sum of variables indexed by elements of the set A. Let $S(A_1)$ and $S(A_2)$ be the required sums. If all sets $B_1 := A_1 \setminus A_2$, $B_2 := A_1 \cap A_2$ and $B_3 = A_2 \setminus A_1$ are non-empty, then we can calculate the sum for each of these sets using at most

$$|B_1| - 1 + |B_2| - 1 + |B_3| - 1 = |A_1 \cup A_2| - 3 \leqslant n - 3$$

summations. It remains to find $S(A_1) = S(B_1) + S(B_2)$ and $S_2 = S(B_2) + S(B_3)$ using two summations. It is easy to see that the same estimation holds even if one of B_i is empty.

(b) The proof is analogous to (a).

Задача 5. It is easy to see that $2^n - (n+1)$ is the number of all sums of n variables, excluding empty sum and these variables themselves. We can consecutively calculate these sums: first sums of two variables, then sums with three variables, so on, and at last the sum of all n variables. As a result of this process we calculate all sets of sums of variables $x_1, \dots, x_n$.

Задача 6. (c) The set of km sums can be represented as the collection of k sets of m sums. Realizing each element of this collection separately we obtain the required estimation.

(d) Using vertical cuts we can split the matrix corresponding to our set of m sums of kn variables to k matrixes $m \times n$. First using kl summations we can realize all the obtained matrixes $m \times n$. Then each of the required sums is a sum of at most $k - 1$ already calculated sums. It follows that the required set can be constructed using at most $kl + m(k-1) < k(l+m)$.

THE COMPLEXITY OF SUMMATION

Yu. Kudryashov, A. Skopenkov (with participation of I. Nikokoshev).

Solutions (the second part).

Задача 7. (a) By problem 5 for $n = 500$ we obtain that any set of m sums from 500 variables can be calculated using at most $l = 2^{500}$ summations. By problem 6(b) it follows that any set of m sums of $500k$ variables can be calculated using at most $k(l+m)$ summations. Therefore any set of m sums from n variables can be calculated using at most

$$\left\lceil \frac{n}{500} \right\rceil (l+m)$$

summations. Clearly, for big enough m and n this is less than $mn/100$.

(b) By problem 5 we obtain that any set of m sums from $\lfloor \log_2 m \rfloor$ variables can be calculated using at most $l = 2^{\lfloor \log_2 m \rfloor} < m$ summations. By problem 6(b) it follows that any set of m sums of $\lfloor \log_2 m \rfloor k$ variables can be calculated using at most $k(l+m) < 2km$ summations. Therefore any set of m sums from n can be calculated using at most

$$2m \left\lceil \frac{n}{\lfloor \log_2 m \rfloor} \right\rceil$$

summations.

Задача 9. Consider any matrix A of the size $m \times n$. If all columns are distinct, then $\mathrm{L}(A) \leqslant \mathrm{L}(B_m)$. If there are two coinciding columns, then by calculating the sum of variables corresponding to these columns we reduce the problem to realization of some matrix of the size $m \times (n-1)$. So $\mathrm{L}(m,n) \leqslant \mathrm{L}(B_m) + n < n + 2^{m+1}$.

Задача 10. For great enough n we have

$$\mathrm{L}(n,n) < \left\lceil \frac{n}{\lfloor 0.95 \log_2 n \rfloor} \right\rceil \mathrm{L}(\lfloor 0.95 \log_2 n \rfloor, n) < \frac{n}{0.94 \log_2 n} \mathrm{L}(\lfloor 0.95 \log_2 n \rfloor, n) <$$
$$< \frac{n}{0.94 \log_2 n}(n + 2 \cdot 2^{\lfloor 0.95 \log_2 n \rfloor}) < \frac{n}{0.94 \log_2 n}(n + 2n^{0.95}) < \frac{n}{0.9 \log_2 n}.$$

Задача 11. (a) Let *entries* be entries of a scheme and outputs of summation elements (so there are $n+k$ entries). Let *outputs* be outputs of a scheme and entries of summation elements (so there are $m+2k$ outputs). Each scheme is uniquely defined by the information which entries are joined to which outputs. Hence the number of schemes is at most $(n+k)^{m+2k}$.

(b) Since the number of distinct schemes as in the formulation does not exceed the number of $m \times n$ matrices of 0 and 1, it follows that $2^{mn} \leqslant (n+k)^{m+2k}$. This easily implies the solution.

(c) If $k = \mathrm{L}(n,n) < \frac{n^2}{5 \log_2 n}$, then $n^2 \leqslant (n + \frac{2n^2}{5 \log_2 n}) \log_2 (n + \frac{n^2}{5 \log_2 n})$. But this inequality is false for big enough n because

(d) Analogously to (b).

Задача 12. (a) Follows because $a_n < b_n \sqrt{1+\varepsilon}$ and $b_n < c_n \sqrt{1+\varepsilon}$ imply $a_n < (1+\varepsilon) c_n$.

(b) By the definition of '$\sim$' and '$\lesssim$'.

Задача 13. If $k = \mathrm{L}(m,n) < \frac{mn}{(2+\varepsilon)\log_2(mn)}$, then $mn \leqslant (m + \frac{2mn}{(2+\varepsilon)\log_2(mn)}) \log_2(n + \frac{mn}{2 \log_2(mn)})$. But this is false for big enough n because

$$\lim_{n\to\infty} \frac{1}{mn}(m + \frac{2mn}{(2+\varepsilon)\log_2(mn)}) \log_2(n + \frac{mn}{(2+\varepsilon)\log_2(mn)}) =$$
$$= \lim_{n\to\infty} \frac{\log_2(mn)}{mn}(n + \frac{2mn}{(2+\varepsilon)\log_2(mn)}) \cdot \lim_{n\to\infty} \frac{1}{\log_2(mn)} \log_2(n + \frac{mn}{(2+\varepsilon)\log_2(mn)}) =$$
$$= \frac{2}{2+\varepsilon} \cdot 2 = \frac{4}{2+\varepsilon}.$$

Задача 14. Split A into submatrixes of the size $(\lfloor \log_2 \frac{m}{k} \rfloor) \times n$. Each submatrix can be realized using at most $\frac{m}{k}$ summations. Hence A can be realized for $\left\lceil \frac{n}{\lfloor \log_2 \frac{m}{k} \rfloor} \right\rceil \cdot \frac{m}{k} + m \left\lceil \frac{n}{\lfloor \log_2 \frac{m}{k} \rfloor} \right\rceil \lesssim \frac{k+1}{k} \frac{mn}{\log_2 m}$ summations. Therefore for each k we have $\mathrm{L}(m,n) \lesssim \frac{k+1}{k} \frac{mn}{\log_2 m}$. So $\mathrm{L}(m,n) \lesssim \frac{mn}{\log_2 m}$.

Задача 15. Analogously to 10, for each $\varepsilon \in (0,1)$ and sufficiently great n we have $\mathrm{L}(m,n) \leqslant \frac{mn)}{\alpha \log_2 n}$.

Задача 16. (a) A scheme realizing matrix A can be represented as an oriented graph whose vertices are of the following four types:

- entries of the scheme (such vertices have no entries and one output),
- summation elements (such vertices have two entries and one output),
- multiplicator (such vertices have one entry and two outputs)
- outputs of the scheme (such vertices have one entry and no outputs).

Clearly, a_{ij} is the number (mod2) of oriented paths from the j-th entry to the i-й output. Consider the scheme coresponding to the graph, in which all the orientations of the edges are the opposite. This scheme clearly represents A^t. Let l_1 be the complexity of the initial scheme and l_2 the complexity of the second one. The number of edges in our graph is $n + l_1 + 2l_2 = m + l_2 + 2l_1$. Hence $n + l_2 = m + l_1$. Since $l_1 = \mathrm{L}(A)$, we have $\mathrm{L}(A^t) \leqslant m - n + \mathrm{L}(A)$. The equality follows because $(A^t)^t = A$.

(b) Follows from (a).

(c) Let B'_m be B_m without the zero column. Then $(B'_m)^t$ is the matrix corresponding to the set of all sums of m variables. The complexity of $(B'_m)^t$ is known, so by (a) we have $\mathrm{L}(B_m) = 2^{m+1} - 2m - 2$.