Труды Математического института РАН
1995, том 211

УДК 511.12(091)

А.А. Карацуба

Сложность вычислений

Введение

Статья посвящена одному из вопросов истории математики, именно истории возникновения основных арифметических операций и, более точно, возникновению умножения.

Мне неоднократно задавали вопрос о том, как был найден метод быстрого умножения многозначных чисел. В свою очередь, у меня возник вопрос о том, как человечество пришло к тому методу умножения, который был единственным до 1960 г. и который называют "обычным", "известным", "школьным" и т.п. Цель статьи — дать ответ на эти вопросы.

Я буду пользоваться различными математическими понятиями и обозначениями, которых, конечно, не было в те древние времена, когда возникла арифметика. Эти понятия дают возможность правильно и более содержательно с математической точки зрения понять существо дела.

История математики занимается в основном теми вопросами математики, которые достигли достаточно высокой степени абстракции и высокой степени развития. Уже определены математические понятия, уже сформулированы гипотезы и доказаны теоремы, уже построены теории. Это сильно облегчает задачу исследовате́ля, так как он и его древний оппонент в значительной степени уравниваются. Оба могут применять похожие логические рассуждения, оба могут применять буквенные обозначения и т.д. Например, обширные исследования посвящены пифагорейской школе, "Началам" Евклида, трудам Архимеда и Диофанта. Но интересным и существенным является вопрос, когда было придумано (найдено) то или иное понятие (метод, прием) и что послужило поводом для придумывания (нахождения) этого понятия (метода, приема).

Будем далее считать, что числа записываются в двоичной системе счисления. Знаки этой системы 0 и 1 называются битами. Запись знака, сложение, вычитание, умножение двух битов, скобки считаются одной битовой (иногда говорят элементарной) операцией (или просто операцией). Если это ясно из контекста, то говоря об одной операции, всегда имеют в виду битовую операцию.

1. Теория информации, электронные вычислительные машины и их роль в развитии кибернетики

Теория передачи сообщений и создание ЭВМ стимулировали развитие математической кибернетики. Этой областью математики с начала 50-х годов активно занимался А.Н. Колмогоров, которому принадлежат первые постановки задач о сложности вычислений (около 1956 г.). Как подчеркивает сам А.Н. Колмогоров (см. [1, с. 251]) "...цикл моих работ по теории информации был создан под большим влиянием публикаций Норберта Винера и Клода Шеннона (1948) в конце 50-х и 60-х годов".

2. Энтропия и сложность задачи табулирования

К упомянутому времени уже создано и широко применяется, например, в работах А.Н. Колмогорова [2], А.Г. Витушкина [3] понятие энтропии дискретных множеств, введенное К. Шенноном. Так, в монографии [3, с. 18-19] А.Г. Витушкин пишет:

"О п р е д е л е н и е (К. Шеннон). Пусть X есть множество, состоящее из n элементов $x_1, x_2, \ldots, x_n$. Число $H(X) = \log n$ называется энтропией множества X. Таким образом, число $H(X)$, определяемое мощностью множества X, показывает, из скольких (двоичных) разрядов должна состоять наиболее экономная таблица для $x \in X$".

В частности, если мы имеем натуральные числа, меньшие 2^m, их количество равно $2^m - 1$ и достаточно m двоичных знаков для их записи в двоичной системе счисления.

В упомянутой монографии А.Г. Витушкина изложены оригинальные работы автора по оценке сложности табулирования (табличного приближенного задания) различных классов функций. Таким образом, идея сложности уже достаточно четко существовала в конце 50-х годов.

3. Сложность вычислений

Будем рассматривать самую простую ситуацию. Пусть $f = f(x)$ — вещественнозначная функция вещественного переменного x, $a \le x \le b$, и пусть $f(x)$ на (a, b) удовлетворяет условию Липшица порядка α, $0 < \alpha < 1$, т.е. при $x_1, x_2 \in (a, b)$ выполняется неравенство

$$|f(x_1) - f(x_2)| \le |x_1 - x_2|^\alpha.$$

Пусть n — натуральное число.

О п р е д е л е н и е 1. Вычислить $f(x)$ в точке $x = x_0 \in (a, b)$ с точностью до n знаков — это значит найти такое число A, что

$$|f(x_0) - A| \le 2^{-n}.$$

Определение 2. Нижняя граница количества битовых операций, достаточных для вычисления $f(x)$ в точке $x = x_0$ с точностью до n знаков, называется сложностью вычисления $f(x)$ в точке $x = x_0$.

Таким образом, сложность вычисления $f(x)$ в точке $x = x_0$ есть функция n и, конечно, $f(x)$, x_0. Эту функцию будем обозначать символом

$$S_f(n) = S_{f,x_0}(n)$$

и называть сложностью вычисления f. Возникает вопрос о поведении $S_f(n)$ при $n \to +\infty$ для класса f или конкретных f. В такой постановке эта проблема была сформулирована А.Н. Колмогоровым около 1956 г. (может быть не точно так, но, по существу, именно в такой форме; о постановке проблемы см. в [4, 5]). В частности, в статье [5, с. 51] Ю.П. Офман, который одним из первых стал заниматься исследованием $S_f(n)$, говорит: "Изложенные результаты получены при работах над более широкой программой исследований, которая была намечена А.Н. Колмогоровым. Дальнейшее движение в этом направлении оказалось более трудным. Трудности возникают уже при оценке алгоритмической сложности обычного умножения двоичных m-значных чисел".

Для оценки сверху $S_f(n)$ строят алгоритмы, с помощью которых проводится вычисление величины A. Имея конкретный алгоритм, подсчитывают количество битовых операций, которые используются в этом алгоритме. Это количество и будет верхней оценкой $S_f(n)$. Так как при вычислениях в первую очередь используются четыре арифметических действия: сложение, вычитание, умножение и деление, то необходимо знать количество битовых операций, достаточное для выполнения этих действий. Из определений 1 и 2 следует, что числа x_0 и A можно представлять в виде целой части и $m = cn$ двоичных знаков после запятой, т.е.

$$A = [A] + 0, \varepsilon_1 \varepsilon_2 \dots \varepsilon_m,$$

$$x_0 = [x_0] + 0, \delta_1 \delta_2 \dots \delta_m,$$

где $\varepsilon_j, \delta_j = 0$ или 1, $j = 1, 2, \dots, m$, $m = cn$, $c = c(\alpha) > 0$ — постоянная. Так как целые части $[A]$ и $[x_0]$ — фиксированные величины, а $n \to +\infty$, то действия производятся по существу с m-значными числами, или, заменяя m на n, с n-значными числами.

Поэтому первым вопросом в теории сложности вычислений является вопрос о количестве битовых операций, достаточных для вычисления суммы, разности, произведения, частного двух n-значных чисел a и b. Замечу, что деление (с остатком) сводится к сложению, вычитанию, умножению чисел (об этом подробнее будет сказано ниже).

Итак, пусть a и b — два n-значных (для простоты, целых) числа в двоичной системе счисления. Для их записи требуется $2n$ битовых операций. Следовательно, сложность сложения (вычитания) двух n-значных чисел не меньше $3n$. В то же время складывая (вычитая) обычным способом, мы затратим не более $4n$ битовых

операций. Таким образом, порядок количества битовых операций, необходимых и достаточных для выполнения сложения и вычитания, один и тот же.

Следующим вопросом является вопрос о количестве операций, достаточных для вычисления ab. Легко видеть (и это сразу было отмечено А.Н. Колмогоровым), что этот вопрос эквивалентен вопросу о поведении $S_f(n)$, где $f = f(x) = x^2$. Действительно,

$$ab = \frac{1}{4}((a+b)^2 - (a-b)^2),$$

и тем самым сложность вычисления ab сводится к сложности вычисления x^2. Функция $S_f(n)$ при $f = x^2$ обозначается символом $M(n)$. Итак, $M(n)$ — сложность вычисления a^2, где a — n-значное число (или сложность умножения двух n-значных чисел).

4. Алгоритм OML и его сложность

Метод умножения, который является стандартным, это умножение "в столбик". Будем называть его далее OML (Ordinary Multiplication). Этот метод был создан (найден) очень давно. Уже шумеры и египтяне широко применяли подобный метод, а это более четырех тысяч лет тому назад, и есть мнение, что этот метод существует не менее шести тысячелетий. Более подробно вопрос о возникновении OML будет рассмотрен ниже. Сейчас же оценим $M(n)$, пользуясь алгоритмом OML. Пусть число a содержит не менее $n/2$ единиц в двоичной записи. Тогда таблица чисел, отвечающая a^2 и OML, которые еще надо будет сложить, содержит не менее $n^2/2$ бит (и не более $2n^2$ бит). На сложение не более чем $2n$-значных чисел в количестве n требуется не более $8n^2$ операций. Итак, для $M(n)$ получаем оценки:

$$4n \le M(n) \le 8n^2,$$

или, если иметь в виду только верхнюю оценку, $M(n) = O(n^2)$.

5. Гипотеза n^2 Колмогорова

В 1956 г. (может быть несколько раньше) А.Н. Колмогоров высказал гипотезу о том, что нижняя оценка $M(n)$ есть величина порядка n^2. Эту гипотезу естественно назвать гипотезой n^2 Колмогорова. Основанием для возникновения этой гипотезы послужило, по-видимому, то обстоятельство, что человечество всю свою историю пользуется OML, сложность которого есть $O(n^2)$, и если бы был более экономный метод умножения, то он был бы уже найден.

В частности, эта гипотеза обсуждалась на одном из заседаний Московского математического общества в 1956 г. На этом заседании А.Н. Колмогоров рассказал о "чешском" методе записи чисел (чешская система счисления, коротко, CSS) в остаточных классах вычетов по заданным модулям и о сложности умножения в CSS. Сама CSS предложена чешскими учеными А. Свободой и М. Валахом в [6]. Пусть

$p_1 < p_2 < \ldots < p_k$ — простые числа. Тогда каждое натуральное число a, меньшее $p_1 p_2 \ldots p_k$, однозначно представляется в виде (CSS):

$$a \cong (a_1, a_2, \ldots, a_k),$$

где

$$a_j \equiv a \pmod{p_j}, \qquad 0 \le a_j < p_j, \quad j = 1, \ldots, k.$$

Сложение, вычитание, умножение чисел в CSS проводится поразрядно. Оценим сложность умножения в CSS. Пусть a и b — n-значные натуральные числа, т.е. $a < 2^n$, $b < 2^n$, пусть p_j — подряд идущие простые числа, начиная с $p_1 = 2$, и пусть k — наименьшее натуральное число с условием

$$2^n < p_1 p_2 \ldots p_k.$$

Из определения k и известного закона распределения простых чисел следует, что

$$\sum_{j=1}^{k-1} \log p_j \le n \log 2 < \sum_{j=1}^{k} \log p_j,$$

$$n \asymp \sum_{p \le k \log k} \log p \asymp k \log k,$$

т.е. k имеет порядок $n/\log n$. Кроме того, каждое p_j имеет порядок $j \log j$, т.е. каждое a_j имеет порядок $j \log j$, $j \ge 2$. Чтобы найти a^2,

$$a^2 \cong (a_1^2, a_2^2, \ldots, a_k^2),$$

надо найти a_j^2, $1 \le j \le k$. Число двоичных знаков a_j имеет порядок $\log j$, сложность вычисления a_j^2 (пользуемся OML) есть $O(\log^2 j)$ и сложность вычисления a^2 в CSS есть величина порядка

$$\sum_{j=1}^{k} \log^2 j = O\Big(\frac{n}{\log n} \log^2 n\Big) = O(n \log n).$$

По поводу полученной оценки А.Г. Витушкин заметил, что "если бы люди жили в CSS, то не было бы гипотезы n^2". На это замечание А.Н. Колмогоров ответил, что системы счисления (коротко с.с.) появились при измерениях и для измерения величин, и в частности для сравнения измеренных величин (измерение как раз и является сравнением измеряемой величины с некоторым стандартом, масштабом). В CSS нельзя определить, какое из двух чисел $a \cong (a_1, a_2, \ldots, a_k)$ и $b \cong (b_1, b_2, \ldots, b_k)$ больше (меньше) другого до тех пор, пока каждое из них не записано в позиционной с.с. Понятно, что перевод a и b из позиционной с.с. в CSS и обратно требует большого количества операций и никакого улучшения оценки $M(n)$ таким способом получить нельзя.

Естественность предположений, подобных гипотезе n^2, отмечена, например, К.И. Бабенко в [7, с. 5]: "Общеизвестно, сколь большое значение для развития науки имела хорошая система счисления, и великолепную шестидесятеричную систему для целых чисел и дробей мы находим уже в древнем Вавилоне. Казалось бы, за истекшие тысячелетия в чем-чем, а в том, что касается систем счисления и техники счета, наука уже давно получила окончательные и наилучшие решения".

6. Опровержение гипотезы n^2

Осенью 1960 г. в Московском университете на механико-математическом факультете начал работать семинар по математическим вопросам кибернетики под руководством А.Н. Колмогорова, где А.Н. Колмогоровым была сформулирована гипотеза n^2 и поставлен ряд задач об оценке сложности решений линейных систем уравнений и других сходных вычислений. Я активно стал размышлять над гипотезой n^2 и ровно через неделю обнаружил, что алгоритм, которым я надеялся получить нижнюю оценку величины $M(n)$, дает оценку вида

$$M(n) = O(n^{\log_2 3}), \qquad \log_2 3 = 1,5849\ldots$$

После очередного заседания семинара я сообщил А.Н. Колмогорову о новом алгоритме умножения и об опровержении гипотезы n^2. Это сильно взволновало А.Н. Колмогорова, так как противоречило его довольно правдоподобной гипотезе. На следующем заседании семинара мой метод умножения был рассказан самим А.Н. Колмогоровым, и на этом семинар прекратил свою работу. Позднее, в 1962 г., А.Н. Колмогоров написал (может быть, при участии Ю.П. Офмана) небольшую статью и опубликовал ее в Докладах АН СССР. Статья называлась так: А. Карацуба, Ю. Офман, Умножение многозначных чисел на автоматах (Докл. АН СССР. 1962. Т. 145, № 2. С. 293–294). Об этой статье я узнал только тогда, когда мне были даны ее оттиски. Необычность способа публикации подчеркивается и тем, что обе статьи [5] и [8] представлены А.Н. Колмогоровым к опубликованию одновременно 13.II 1962 г.

После этого началась бурная деятельность в области прикладной математики, которая получила название "быстрые вычисления". Она продолжается до сих пор. Несколько подробнее об этом будет сказано ниже.

7. Алгоритм KML и его сложность

В этом параграфе изложен мой алгоритм умножения чисел. В настоящее время он коротко называется алгоритмом KML или просто KML (Karatsuba Multiplication) (см., напр. [9]).

Как было отмечено выше, умножение двух чисел сводится к возведению в квадрат одного числа. Итак, надо возвести в квадрат n-значное число a. Не ограничивая общности, будем считать $n = 2^m$. Запишем a в следующем виде: $a = 2^{n_1} a_1 + a_2$, $2n_1 = n$, где a_1 и a_2 — n_1-значные числа. Имеем

$$a^2 = (2^{n_1} a_1 + a_2)^2 = 2^n a_1^2 + 2^{n_1} 2 a_1 a_2 + a_2^2.$$

К

$$2a_1a_2 = (a_1+a_2)^2 - a_1^2 - a_2^2,$$

т.е.

$$a^2 = 2^n a_1^2 - 2^{n_1}a_1^2 + 2^{n_1}(a_1+a_2)^2 + a_2^2 - 2^{n_1}a_2^2. \tag{1}$$

Так как a_1, a_2 являются n_1-значными числами, то сумма a_1+a_2 будет не более чем (n_1+1)-значным числом. Поэтому ее можно представить так:

$$a_1 + a_2 = \varepsilon + 2a_3,$$

где $\varepsilon = 0$ или 1, a_3 — n_1-значное число. Следовательно,

$$(a_1+a_2)^2 = \varepsilon^2 + 4\varepsilon a_3 + 4a_3^2. \tag{2}$$

Из (1) и (2) получаем

$$a^2 = 2^n a_1^2 - 2^{n_1}a_1^2 + 2^{n_1+2}a_3^2 + 2^{n_1+2}\varepsilon a_3 + 2^{n_1}\varepsilon^2 + a_2^2 - 2^{n_1}a_2^2. \tag{3}$$

Вычисление a^2 будем проводить по (3). Пусть $\varphi(n)$ — количество операций (битовых), достаточное для вычисления квадрата n-значного числа по формуле (3). Из правой части (3) видно, что надо возвести в квадрат три n_1-значных числа, именно a_1, a_2, a_3; для этого потребуется $3\varphi(n_1)$ операций. Затем надо каждое из полученных значений умножить на одно из чисел 2^n, 2^{n_1}, 2^{n_1+2}, что потребует не более $6n$ операций (умножение на степень 2 — это приписывание справа от числа-множимого соответствующее количество нулей). Затем надо будет сложить (имеется в виду алгебраическая сумма) семь не более чем $2n$-значных чисел, для чего потребуется не более чем

$$4\cdot 2n\cdot 4 + 4(2n+2)\cdot 2 + 4(2n+2) = 56n + 24$$

операций. Тем самым для $\varphi(n)$ получаем неравенство

$$\varphi(n) \le 3\varphi(n_1) + 6n + 56n + 24 \le 3\varphi(n_1) + 70n. \tag{4}$$

Полагая далее $2n_{j+1} = n_j$, $j = 1, \ldots, m-1$, получаем

$$\varphi(n_j) \le 3\varphi(n_{j+1}) + 70n_j. \tag{5}$$

Так как $n_{j+1} = 2^{m-j-1}$, то $n_m = 1$, и тривиально $\varphi(1) = 1$. Из (4) и (5) методом математической индукции легко доказать, что при $j \ge 1$

$$\varphi(n) \le 3^j\varphi(n_j) + 3^{j-1}\cdot 70n_{j-1} + \ldots + 3\cdot 70n_1 + 70n. \tag{6}$$

Действительно, при $j = 1$ это так. Предполагая правильность этой формулы при j, докажем ее для $j+1 \le m$. Подставим (5) в (6), найдем

$$\varphi(n) \le 3^{j+1}\varphi(n_{j+1}) + 3^j\cdot 70n_j + 3^{j-1}\cdot 70n_{j-1} + \ldots + 3\cdot 70n_1 + 70n,$$

а это и требовалось доказать.

Полагая теперь в (6) $j = m$, $\varphi(n_m) = \varphi(1) = 1$, $n_j = 2^{m-j}$, получаем

$$\varphi(n) \le 3^m + 3^{m-1}\cdot 70n_{m-1} + 3^{m-2}\cdot 70n_{m-2} + \ldots + 3\cdot 70n_1 + 70n =$$

$$= 3^m + 3^{m-1}\cdot 70\cdot 2 + 3^{m-2}\cdot 70\cdot 2^2 + \ldots + 3\cdot 70\cdot 2^{m-1} + 70\cdot 2^m =$$

$$= 3^m\left(1 + 70\cdot\frac{2}{3} + 70\cdot\left(\frac{2}{3}\right)^2 + \ldots + 70\cdot\left(\frac{2}{3}\right)^{m-1} + 70\cdot\left(\frac{2}{3}\right)^m\right) < 70\cdot 3^m\cdot\left(1-\frac{2}{3}\right)^{-1} = 210\cdot 3^m.$$

Так как $n = 2^m$, $m = \log_2 3$, то

$$\varphi(n) < 210n^{\log_2 3}, \qquad \log_2 3 = 1,5849\ldots$$

Отсюда, в частности, следует, что

$$M(n) = O(n^{\log_2 3}),$$

т.е. опровержение гипотезы n^2.

Замечу, что при оценке $\varphi(n)$ умышленно завышены константы, чтобы все вычисления были как можно проще. Можно более экономно проводить указанные вычисления и от этого константа 210 заменится значительно меньшей. О практическом применении KML будет сказано ниже.

Есть и другой вариант KML — непосредственное умножение двух n-значных чисел a и b. Опять представляя a и b в виде

$$a = 2^{n_1}a_1 + a_2, \qquad b = 2^{n_1}b_1 + b_2, \qquad 2n_1 = n,$$

находим

$$ab = (2^{n_1}a_1 + a_2)(2^{n_1}b_1 + b_2) = 2^n a_1b_1 + 2^{n_1}(a_2b_1 + a_1b_2) + a_2b_2 =$$

$$= 2^n a_1b_1 - 2^{n_1}a_1b_1 + a_2b_2 - 2^{n_1}a_2b_2 + 2^{n_1}(a_1+a_2)(b_1+b_2). \tag{7}$$

В этой формуле мы имеем три произведения вида a_1b_1, a_2b_2, $(a_1+a_2)(b_1+b_2)$. Каждый из сомножителей является не более чем (n_1+1)-значным числом. Обозначая, как и выше, через $\varphi(n)$ количество операций, достаточное для вычисления произведения двух n-значных чисел по формуле (7), получаем неравенство

$$\varphi(n) < 3\varphi(n_1) + cn,$$

где $c > 0$ — абсолютная постоянная, $2n_1 = n$. Из этого неравенства находим оценку

$$\varphi(n) < c_1 n^{\log_2 3},$$

$c_1 > 0$ — абсолютная постоянная.

Совершенно ясно, что разбивая a и b не на два, а на бо́льшее количество слагаемых, можно получить и более точную оценку для $M(n)$. Несколько ниже подробнее будет рассказано о дальнейшем развитии быстрых вычислений и об уточнениях оценки $M(n)$. Сейчас же обратимся к древней истории, связанной с простейшими вычислениями.

8. Арифметика древних

Вычислениями человек разумный занимался с момента своего появления. Вычисления были самые примитивные: сложение, вычитание, умножение и деление небольших чисел. Но арифметика, а вместе с ней и математика, появилась только тогда, когда возникли большие числа. Действительно, если a и b — малые числа (порядка единиц), то сложность операций $a+b$, $a-b$, $ab = a+a+\ldots+a$ в нашей терминологии есть $O(1)$ и эти операции ничем принципиально одна от другой не отличаются. Если же a и b есть n-значные числа, то аддитивные операции, т.е. сложение и вычитание, требуют $O(n)$ операций, а умножение ab, определяемое как результат повторного сложения множимого, требует $O(n2^n)$ операций. Это обстоятельство подчеркивает принципиальное отличие умножения от сложения-вычитания. Такой же по сложности является и операция деления a на b с остатком: $a = bq + r$, $0 \le r < b$ (деление производится последовательным вычитанием из a чисел b, т.е. $a-(b+b+\ldots+b) = r$, $0 \le r < b$).

Отмечу также, что в настоящее время известно, что и некоторые животные могут складывать, вычитать, а следовательно, и умножать малые числа.

Когда люди стали иметь дело с большими числами, в первую очередь, и это вполне естественно, появились позиционные с.с. Вопрос этот довольно подробно исследован историками-математиками (см., напр. [7, 10–13]), и я на нем останавливаться не буду.

Числа люди складывали, вычитали, умножали и делили. Первое деление было достаточно примитивным, делители были либо малыми числами, либо специального вида числами. Все, что связано с делением, также достаточно подробно исследовано в [10, 11]. Пока будем касаться только аддитивных операций и умножения. Замечу, что эти операции дошли до нас в своем первозданном виде. Место и время их появления точно не установлено. Самые древние известные источники — это клинописные таблицы шумеров, египетский папирус Райнда и так называемый московский папирус, который считается на два столетия старше райндовского. Будущие археологические открытия могут сильно отодвинуть в глубь веков время возникновения арифметики. Есть гипотезы о существовании развитых древних цивилизаций в Африке, возраст которых не менее 20 тыс. лет. Косвенным подтверждением этих гипотез служат космические фотосъемки американских астронавтов. Но все это — дело будущего. Сейчас же я имею только одну задачу — пример на умножение из папируса Райнда (см. [10, с. 22]). Во всех других источниках авторы дают свои собственные примеры, демонстрируя методы вычисления древних. Приведу несколько фактов из работ на эту тему.

Сведения о папирусе Райнда из монографии Ван дер Вардена [10, с. 19] таковы. Папирус написан в Египте около 1800 г. до н.э. "Его писец Ахмес уверяет, что он восходит к оригиналу из Среднего царства (2000–1800 гг. до н.э.)". В нем содержится 84 задачи, посвященных технике счета. Система счисления — десятичная. На с. 22 Ван дер Варден пишет: "Сложение этих чисел не составляет трудностей, нужно только сосчитать количество единиц, десятков, сотен и т.д. Удвоение представляет частный случай сложения и также не труден. Однако совершенно своеобразным является

Умножение.

Оно производится при помощи удвоения и сложения полученных результатов. В качестве примера приведем умножение 12×12 по задаче N 32 райндовского папируса сначала в иероглифической записи (которую нужно читать справа налево), а затем в современной". Я привожу здесь только современную запись:

1	12	
2	24	
/4	48	
/8	96	Сумма 144.

"Учетверение и увосьмерение дают вместе двенадцатикратное увеличение заданного числа 12. Числа, которые надо последовательно сложить, отмечаются косой черточкой справа (в "переводе" слева). Перед результатом 144 стоит иероглиф *dmd*, изображающий свиток с печатью". Далее на с. 24 Ван дер Варден пишет: "Этот египетский способ умножения является основой всей техники счета. Он должен быть очень древним, однако в этой форме он удержался до эллинистической эпохи и в греческих школах назывался "египетским" счетом. Еще даже в средние века "duplatio" (удвоение) считалось самостоятельным действием".

Другим источником может служить монография Д.Я. Стройка [13], в которой на с. 36 упоминается папирус Райнда, содержащий 84 задачи, и московский папирус (25 задач), который, может быть, на два столетия старше. В этих папирусах содержится техника счета. На с. 37 своей монографии Д.Я. Стройк пишет: "На основе такой системы египтяне построили арифметику преимущественно аддитивного характера, т.е. ее основное направление состоит в сведении всех умножений к повторным сложениям. Например, умножение на 13 получается умножением сначала на 2, затем на 4, затем на 8 и сложением результатов умножения на 4 и на 8 с первоначальным числом. Например, для вычисления 13×11 писали:

*1	11
2	22
*4	44
*8	88

и складывали все числа, отмеченные звездочкой, что дает 143".

Здесь мы видим не оригинальный пример из папируса, а интерпретацию метода египтян. Приведу еще цитату из статьи И.Г. Башмаковой и А.П. Юшкевича (см. [12, с. 29]): "Египетская система интересна еще по той роли, которую там играет число два. По-видимому, оно служило первоначально основанием системы счисления... Пережитки двоичной системы отразились в способе умножения египтян, которое они производили путем последовательного удвоения и сложения. Например, для умножения некоторого числа n на 15 египтяне поступали (схематически) так: $n \cdot 15 = n(1+2+2^2+2^3) = n \cdot 1 + n \cdot 2 + n \cdot 2^2 + n \cdot 2^3$, т.е. они

представляли множитель по двоичной системе, а затем умножение производилось отдельно на каждый двоичный разряд".

Далее египетский метод умножения коротко будем называть EML.

9. Сложность алгоритма EML

Нетрудно формализовать EML и оценить его сложность. Запишем b в виде двоичного разложения:

$$b = 2^{n_1} + 2^{n_2} + \ldots + 2^{n-1},$$

где

$$0 \le n_1 < n_2 < \ldots < n-1.$$

Тогда ab представится так:

$$ab = a \cdot 2^{n_1} + a \cdot 2^{n_2} + \ldots + a \cdot 2^{n-1}. \tag{8}$$

Начиная с a, последовательно путем сложения получают

$$2a, 2^2a, \ldots, 2^{n_1}a, \ldots, 2^{n_2}a, \ldots, 2^{n-1}a,$$

т.е. получают все слагаемые правой части (8). Складывая эти слагаемые, находят ab. Каждое из слагаемых является не более чем $2n$-значным числом, и число слагаемых не превосходит n. Поэтому количество операций, достаточное для получения суммы, есть $O(n^2)$. Чтобы получить $2a$, надо $O(n)$ операций, 2^2a — также $O(n)$ операций и т.д., наконец, $2^{n-1}a$ — также $O(n)$ операций, так как всегда складываются не более чем $2n$-значные числа. Следовательно, чтобы получить слагаемые правой части (8), достаточно $O(n^2)$ операций. Тем самым сложность EML есть $O(n^2)$ и совпадает со сложностью OML.

10. OML — прямое следствие EML

Легко видеть, что если слагаемые правой части (8) получать не последовательным сложением, а приписыванием к a справа соответствующего количества нулей, то вместо EML мы получим OML. Я думаю, что к OML люди пришли сразу же, как только появился знак 0.

11. EML — теоретическая реализация взвешивания на коромысловых весах

По-видимому, основой появления EML послужило взвешивание на коромысловых весах с двумя чашками весов. Самый быстрый способ получить на коромысловых весах вес $a \cdot 2^{n-1}$ состоит в том, что сначала, имея в левой чашке весов a, справа получают a, тем самым получают вес $2a$; имея $2a$, точно так же получают $2 \cdot 2a = 2^2a$, и так далее до $2^{n-1}a$. Имея все веса вида $a, 2a, 2^2a, \ldots, 2^{n-1}a$, путем сложения получают вес ab, где b — любое n-значное число. Это как раз и есть алгоритм EML, позднейшей модификацией которого стал OML. Замечу, что коромысловые весы и взвешивание на них послужили также появлению двоичной системы счисления.

Дальнейшие догадки в этом направлении приводят к новым выводам. Коромысловые весы и сами коромысла появились вместе с человеком разумным и даже раньше. Самыми простыми коромыслами являются две руки человека, тем самым само устройство человека, именно наличие у него рук, послужило появлению двоичной с.с., EML и OML.

Думаю, что изобретение коромысла, коромысловых весов имеет для человечества такое же большое значение, как изобретение колеса.

12. Современное состояние быстрых вычислений

Коротко остановлюсь на дальнейшем развитии (после 1962 г.) быстрых вычислений и о современном состоянии этого направления. Прежде всего будет показано, как деление сводится к сложению–вычитанию и умножению чисел.

Как было отмечено выше, деление числа a на число b с остатком, т.е. вычисление чисел q и r в формуле

$$a = qb + r, \qquad 0 \le r < b,$$

сводится к сложению–вычитанию и умножению, и если a, b — не более чем n-значные числа, то сложность деления a на b есть величина порядка $O(M(n))$.

Сначала вычисляют $1/b$ с точностью 2^{-n-1}, т.е. находят $\varepsilon_1, \ldots, \varepsilon_{n+1}$ в формуле

$$\frac{1}{b} = 0, \varepsilon_1 \ldots \varepsilon_{n+1} + \theta \cdot 2^{-n-1}, \qquad |\theta| \le 1.$$

Тогда q равно одному из следующих трех чисел: $[a \cdot 0, \varepsilon_1 \ldots \varepsilon_{n+1}] \pm 0, 1$. Следовательно, q найдется за $O(M(n))$ операций. Числа $\varepsilon_1, \ldots, \varepsilon_{n+1}$ находят, пользуясь такой леммой.

Л е м м а. *Пусть* $1/2 < x < 1$, $s = 1/x$. *Тогда если*

$$|s - s_k| < 2^{-k},$$

то для $s_{2k} = xs_k^2 - 2s_k$ *имеем*

$$|s - s_{2k}| < 2^{-2k}.$$

За первое приближение, т.е. за s_1, берут число $3/2$. Таким способом деление осуществляется в [7, с. 335] (см. также [14, 15]).

13. О некоторых быстрых алгоритмах, стимулированных KML

Алгоритм KML является источником и прототипом всех быстрых умножений (короткую историю об этом см. в [15]). В первую очередь сюда относится алгоритм В. Штрассена (1969 г.) умножения матриц, который по существу является

применением KML к умножению матриц (см. [16]). В этом случае за одну операцию считается сложение, вычитание, умножение двух элементов матрицы, запись элемента матрицы и запись арифметической операции.

Действительно, известно, что матрицы перемножаются "блочно", т.е. если A и B имеют вид

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

где A и B — матрицы размера $n \times n$, $n = 2^m$, A_{ij}, B_{ij} — матрицы размера $n_1 \times n_1$, $2n_1 = n$ ("блоки" матриц A и B), то

$$AB = C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

где C_{ij} — матрицы размера $n_1 \times n_1$, причем

$$C_{11} = A_{11}B_{11} + A_{12}B_{21}$$

и т.д. Из формул для C_{ij} видно, что для их получения требуется 8 умножений матриц размером $n_1 \times n_1$ и 4 сложения. Этот алгоритм дает для сложности умножения двух матриц $n \times n$ оценку вида $O(n^3)$. Но обычное умножение матриц также выполняется за $O(n^3)$ операций. В. Штрассен [16] нашел тождество, в котором надо не 8 умножений блоков, а только 7, и 18 сложений. Тождество Штрассена таково.

Пусть

$$\begin{aligned}
\text{I} &= (A_{11} + A_{22})(B_{11} + B_{22}), \\
\text{II} &= (A_{21} + A_{22})B_{11}, \\
\text{III} &= A_{11}(B_{12} - B_{22}), \\
\text{IV} &= A_{22}(-B_{11} + B_{21}), \\
\text{V} &= (A_{11} + A_{12})B_{22}, \\
\text{VI} &= (-A_{11} + A_{21})(B_{11} + B_{12}), \\
\text{VII} &= (A_{12} - A_{22})(B_{21} + B_{22}).
\end{aligned}$$

Тогда

$$\begin{aligned}
C_{11} &= \text{I} + \text{IV} - \text{V} + \text{VII}, \\
C_{21} &= \text{II} + \text{IV}, \\
C_{12} &= \text{III} + \text{V}, \\
C_{22} &= \text{I} + \text{III} - \text{II} + \text{VI}.
\end{aligned}$$

Следовательно, если $\psi(n)$ — сложность умножения двух матриц $n \times n$, то, применяя тождество Штрассена, найдем

$$\psi(n) \le 7\psi(n/2) + cn,$$

$$\psi(n) \le c_1 n^{\log_2 7}, \qquad \log_2 7 = 2,807\ldots,$$

где $c > 0$, $c_1 > 0$ — абсолютные постоянные. Понятно, что если множители A и B разбивать на более мелкие блоки и находить бóльшее количество зависимых умножений, то оценку В. Штрассена можно улучшить. Это сейчас проделано многими авторами, но получить $\psi(n) = O(n^{2+\varepsilon})$, где $\varepsilon > 0$ — любое, пока не удалось (1995 г.).

Отмечу, что В. Штрассен в 1965 г. несколько месяцев был стажером А.Н. Колмогорова, который познакомил его со всей тематикой сложности вычислений и всеми достижениями в ней.

Другим быстрым алгоритмом является алгоритм быстрого преобразования Фурье (дискретного преобразования Фурье), найденный Дж.У. Кули и Дж.У. Таки в 1965 г. [17], который, как я думаю, также был стимулирован KML. В частности, о KML как о прототипе всех быстрых умножений пишут А. Шёнхаге, А.Ф.В. Гротефельд, Е. Феттер в [9].

14. Уточнения KML

Как это было отмечено выше, разбивая a на большое количество слагаемых, т.е. представляя a в виде

$$a = a_0 + 2^m a_1 + 2^{2m} a_2 + \ldots + 2^{rm} a_r,$$

где $a_0, a_1, a_2, \ldots, a_r$ — m-значные числа, $rm = n$, получаем

$$a^2 = \left(\sum_{j=0}^{r} a_j 2^{mj} \right)^2 = \sum_{s=0}^{2r} c_s 2^{ms},$$

где

$$c_s = \sum_{\substack{j+\nu=s \\ 0 \le j,\nu \le r}} a_j a_\nu.$$

Коэффициенты c_s можно найти из системы уравнений

$$(a_0 + a_1 x + a_2 x^2 + \ldots + a_r x^r)^2 = \sum_{s=0}^{2r} c_s x^s,$$

в которой следует положить $x = 0, \pm 1, \ldots, \pm 2r$. Выбирая оптимальное r, получим соответствующую оценку $M(n)$. Таким способом была уточнена оценка $M(n)$ тремя авторами: А.А. Тоомом [18], С.А. Куком [19], А. Шёнхаге [20]. Уточненная оценка $M(n)$ выглядит так:

$$M(n) = O\left(ne^{c\sqrt{\log n}}\right), \tag{9}$$

где $c > 0$ — абсолютная постоянная. Замечу, что А. Шёнхаге в своем алгоритме пользовался модульной арифметикой со специального вида модулями, что дало ему возможность уменьшить константы в оценке (9).

Наконец, в 1971 г. А. Шёнхаге и В. Штрассен [21] построили алгоритм с наилучшей на настоящее время (1995 г.) верхней оценкой $M(n)$:

$$M(n) = O(n \log n \log\log n).$$

При построении своего алгоритма А. Шёнхаге и В. Штрассен существенно пользуются (для вычисления c_s) быстрым преобразованием Фурье.

15. Быстрые вычисления алгебраических и простейших трансцендентных функций

Если $y = f(x)$ — алгебраическая функция, тогда

$$S_f(n) = O(M(n)).$$

При доказательстве этого соотношения существенно используется метод касательных Ньютона (см. [7, с. 335; 14]). Если $y = f(x)$ — простейшая трансцендентная функция ($f(x) \equiv e^x$, обратная к ней, тригонометрические функции, всевозможные суперпозиции названных и алгебраических функций), то

$$S_f(n) = O(M(n)\log n). \tag{10}$$

При доказательстве этого соотношения существенно используются итерационный метод, эллиптические интегралы, АГМ (алгоритм Гаусса средних арифметических и геометрических), преобразования Ландена (см. [22, 23]).

16. Быстрые вычисления высших трансцендентных функций

Если $y = f(x)$ — высшая трансцендентная функция (гамма-функция Эйлера, функция Бесселя, гипергеометрическая функция и т.п.), то

$$S_{f,x_0}(n) = O(M(n)\log^2 n). \tag{11}$$

Целый ряд результатов на эту тему опубликован Дж.М. Борвайном и П.Б. Борвайном в [24] без указаний алгоритмов вычислений, но с упоминанием того, что это итерационные алгоритмы. В последние годы Е.А. Карацуба предложила новый метод быстрого вычисления простейших и высших трансцендентных функций, который не является итерационным, допускает распараллеливание и который назван ею БВЕ (быстрое вычисление функций типа E-функций Зигеля) [25–27]. В частности, в этих работах обнаружен любопытный эффект: оценка (11) получается для упомянутых функций при условии, что параметры вычисляемых функций и x_0 — алгебраические числа (полный аналог известных теорем теории алгебраических чисел).

17. Практическая реализация быстрых алгоритмов

Сколько-нибудь полный обзор практических реализаций быстрых алгоритмов, и в частности KML, дать трудно. Трудно дать по той причине, что, скажем, KML, обладая достаточно простой логической структурой, может быть реализован микросхемой, а проследить такие технические реализации, если они, конечно, не отражены в проекте, невозможно. Отмечу только уже процитированную выше прекрасную монографию А. Шёнхаге, А.Ф.В. Гротефельда, Е. Феттера [9], в которой изложены результаты авторов о возможности практической реализации KML и алгоритма умножения Шёнхаге–Штрассена.

В свою очередь, в мае 1981 г. я дал рукопись короткой статьи "Реальные вычисления" одному из ведущих разработчиков супер-ЭВМ в СССР профессору В.А. Мельникову, в которой предлагались реализации быстрых алгоритмов вычислений элементарных и простейших трансцендентных функций и, конечно, KML. О реальном техническом воплощении этих алгоритмов мне ничего не известно.

18. Нижние оценки

Проблема нижних оценок $S_f(n)$, и в частности $M(n)$, остается нерешенной. Здесь ничего, кроме тривиальных результатов типа

$$n < M(n),$$

нет. Есть много результатов о нижних оценках при ограничениях на используемые алгоритмы (см., напр. [7, с. 334]), но это совсем другое направление исследований. Можно сформулировать гипотезы вида

$$\text{(I)} \quad \sup_n \frac{M(n)}{n} = +\infty;$$

$$\text{(II)} \quad \sup_n \frac{M(n)}{n\log n} > 0;$$

$$\text{(III)} \quad \sup_n \frac{S_f(n)}{n^2} > 0,$$

где $f(x) = \Gamma(x)$, $x_0 = \pi$. Но к доказательству этих гипотез нет пока никаких подходов.

Заключение

В связи с изложенными исследованиями, которые протянули нить от древнейших времен до наших дней, отмечу одно немаловажное обстоятельство, связанное с современным развитием математики. В последние десятилетия появилось много исследователей и большое количество математических работ. Если классики математики воспринимали математическую науку как объективное отражение реальности, то многие новые исследователи фактически не разделяют эту точку зрения. Их лозунг — математика есть продукт чистого вымысла. Их задача — придумать понятие, придумать теорию, придумать доказательство и т.д. Классики же совсем иначе представляли себе работу математика, что находило отражение в их формулировках типа "я нашел решение проблемы", "я нашел доказательство", "я нашел понятие" и т.д. Эти два слова "придумать" и "найти" показывают глубокое различие двух тенденций в математике и двух подходов к занятиям математикой.

Выражаю глубокую благодарность Д.В. Сенченко за полезные замечания, которые улучшили изложение статьи.

Поступило в январе 1995 г.

ЛИТЕРАТУРА

1. *Колмогоров А.Н.* Теория информации и теория алгоритмов. М.: Наука, 1987. 304 с.
2. *Колмогоров А.Н.* О некоторых асимптотических характеристиках вполне ограниченных метрических пространств //Докл. АН СССР. 1956. Т. 108, № 3. С. 385–388.
3. *Витушкин А.Г.* Оценка сложности задачи табулирования. М.: Физматгиз, 1959. 228 с.
4. *Колмогоров А.Н.* Различные подходы к оценке трудности приближенного задания и вычисления функций //Proc. Intern. Congr. Math. Stokholm, 1963. P. 369–376.
5. *Офман Ю.П.* Об алгоритмической сложности дискретных функций //Докл. АН СССР. 1962. Т. 145, № 1. С. 48–51.
6. *Svoboda A., Valach M.* //Stroje zpracov. inform. 1955. Vol. 3. P. 247–295.
7. *Кнут Д.* Искусство программирования для ЭВМ. М.: Мир, 1977. Т. 2. 726 с.
8. *Карацуба А., Офман Ю.* Умножение многозначных чисел на автоматах //Докл. АН СССР. 1962. Т. 145, № 2. С. 293–294.
9. *Schönhage A., Grotefeld A.F.W., Vetter E.* Fast Algorithms //B.I. Manheim etc.: Wissenschaftsverlag, 1994. 297 S.
10. *Ван дер Варден Б.Л.* Пробуждающаяся наука. М.: Физматгиз, 1959.
11. *Нейгебауер О.* Лекции по истории античных математических наук. М.; Л.: ОНТИ, 1937. Т. 1.
12. *Башмакова И.Г., Юшкевич А.П.* Происхождение систем счисления. Энциклопедия элементарной математики. Т. 1. Арифметика. М.; Л.: Гостехтеориздат, 1951. 448 с.
13. *Стройк Д.Я.* Краткий очерк истории математики. М.: Наука, 1990. 254 с.
14. *Бендерский Ю.В.* Быстрые вычисления //Докл. АН СССР. 1975. Т. 223, № 5. С. 1041–1043.
15. *Karacuba A.A.* Berechnungen und die Kompliziertheit von Beziehungen //Elektron. Informationsverarb. und Kybern. 1975. N 10–12. S. 603–606.
16. *Strassen V.* Gaussian elimination is not optimal //Numer. Math. 1969. Vol. 4, N 4. P. 354–356.
17. *Cooley J.W., Tukey J.W.* An algorithm for the machine calculation of complex Fourier series //Math. Comput. 1965. Vol. 19. P. 293–301.
18. *Тоом А.А.* О сложности схемы из функциональных элементов, реализующей умножение целых чисел //Докл. АН СССР. 1963. Т. 150, № 2. С. 496–498.
19. *Cook S.A.* On the minimum computation time of functions: Thesis. New York: Harvard Univ., 1966. P. 51–77.
20. *Schönhage A.* Schnelle Multiplikation großer Zahlen //Computing. 1966. Bd. 1. S. 182–196.
21. *Schönhage A., Strassen V.* Schnelle Multiplikation großer Zahlen //Computing. 1971. Bd. 7, N 3/4. S. 281–292.
22. *Brent R.P.* Fast multiple-precision evaluation of elementary functions //J. Assoc. Comput. Mach. 1976. Vol. 23. P. 242–251.
23. *Borwein J.M., Borwein P.B.* Pi and AGM. A study in analytic number theory and computational complexity. New York: Wiley, 1987.
24. *Borwein J.M., Borwein P.B.* On the complexity of familiar functions and numbers //SIAM Rev. 1988. Vol. 30, N 4. P. 589–601.
25. *Карацуба Е.А.* О быстром вычислении трансцендентных функций //Докл. АН СССР. 1991. Т. 318, № 2. С. 278–279.
26. *Карацуба Е.А.* Быстрые вычисления трансцендентных функций //ППИ. 1991. Т. 27, № 4. С. 87–110.
27. *Karatsuba Catherina A.* Fast evaluation of Bessel functions //Integral Transforms and Special Functions. 1993. Vol. 1, N 4. P. 269–276.

Труды Математического института РАН
1995, том 211

УДК 517.988.54+517.956.223

А.К. Керимов

Обобщенная теорема о неявной функции и задача с препятствием [1]

Введение

Рассматривается следующая хорошо известная задача со свободной границей.

З а д а ч а А. Найти область $\Omega = (\Gamma_\partial, \Gamma)$, ограниченную поверхностями Γ_∂ и Γ изнутри и снаружи соответственно (рис. 1), и функцию $u = u(\Omega)$ такие, что

$$\Delta u = f \text{ в } \Omega, \qquad u\big|_{\Gamma_\partial} = 1, \qquad u\big|_{\Gamma} = \partial_\nu u\big|_{\Gamma} = 0;$$

при этом поверхность Γ_∂ предполагается заданной, а функция f определена на всем пространстве R^n, $n \geq 2$. По поводу этой задачи и аналогичных ей, а также физических ситуаций к ней приводящих см. работы [1, 4–6, и др.]. Ниже доказывается разрешимость этой задачи в подходящих классах регулярных поверхностей и устанавливаются свойства решения как функции заданной части границы Γ_∂ при условии, что бесконечно дифференцируемая функция f всюду положительна. Основным инструментом исследования является глобальный вариант теоремы о неявной функции типа Нэша–Мозера (разд. 2). Полученные результаты обобщают работу [1] (см. также [3]), где рассмотрен двумерный случай с использованием локальной теоремы в варианте Цандера [7]. Предлагаемый метод проверки условий теорем типа Нэша–Мозера несколько отличается от используемого в работах [1, 2], поэтому приводится достаточно подробное (в основных моментах) его изложение, тем более что это не единственная задача, к которой применим такой подход.

Перейдем к точной формулировке полученных результатов. Пусть (α, μ) — пара функций, определенных на единичной сфере S с центром в начале координат, удовлетворяющих условию $\alpha > \mu > -1$ на S. Каждая такая пара определяет двусвязную область $\Omega(\alpha, \mu)$, состоящую из точек y, сферические координаты (ρ, x) ($\rho = |y|$, $x = y/|y|$) которых удовлетворяют условию $1 + \mu(x) < \rho < 1 + \alpha(x)$, $x \in S$ (рис. 2). Такая область ограничена поверхностями $\Gamma(\alpha)$ и $\Gamma(\mu)$, которые являются образами сферы S при взаимно однозначных отображениях

$$h_\alpha\colon x \to (1 + \alpha(x))x, \qquad h_\mu\colon x \to (1 + \mu(x))x$$

соответственно.

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований согласно проекту 95-01-00310а.